Вінницький національний технічний університет
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

ЛОСЕНКО АРСЕН ВОЛОДИМИРОВИЧ

УДК 004.9+578.834.1

ДИСЕРТАЦІЯ
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Спеціальність 126 – «Інформаційні системи та технології»
Галузь знань 12 – «Інформаційні технології»

Подається на здобуття ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело _____ А. В. Лосенко

Науковий керівник: Мокін Віталій Борисович,
доктор технічних наук, професор

Вінниця – 2023

АНОТАЦІЯ

Лосенко А.В. Інформаційна технологія прогнозування часових рядів кількості хворих на коронавірус методами машинного навчання – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 126 «Інформаційні системи та технології». – Вінницький національний технічний університет, Вінниця, 2023.

Метою дисертаційної роботи є підвищення точності прогнозування кількості хворих на коронавірус у короткостроковій перспективі за допомогою методів машинного навчання.

Дисертація присвячена вирішенню актуальної науково-технічної задачі, що полягає у підвищенні точності прогнозування кількості нових хворих на коронавірус за допомогою методів машинного навчання.

Основні наукові та практичні результати полягають в наступному:

1. Охарактеризовано основні проблеми, пов’язані з прогнозуванням кількості хворих на коронавірус в Україні та інших країнах, обумовлені, передусім, проблемами з даними, пов’язаними зі складностями діагностування медичного та організаційного характеру, мутаціями хвороби, недотриманням населенням карантинних умов тощо. Проведено розвідувальний аналіз даних. Проаналізовано якість даних про інші фактори, які впливають на поширення коронавірусу (пересування жителів країни, метеодані, рішення уряду (на основі Оксфордського трекера антикоронавірусної діяльності урядів країн світу), вчасність оприлюднення підтверджених тестів на коронавірус тощо. Доведено, що вплив цих факторів є дуже мінливим, дані є неточними і зашумленими, надходять із різним запізненням, а отже, їх варто враховувати не як додаткові ознаки для побудови множинної регресії, а краще дати їх різкої зміни чи аномальних значень формалізувати як дати аномальної поведінки основного часового ряду, а тоді ці аномалії можна враховувати параметрично і звести задачу

до однофакторної, тобто – до побудови моделі лише часового ряду кількості нових випадків захворювання на коронавірус. Проведено огляд відомих моделей і методів прогнозування цієї ознаки. Зазначено, що через проблеми з даними, перевагу варто віддавати моделям машинного навчання, а не – диференціальним рівнянням, наприклад SIR, SEIR чи SEIR-U. Здійснено порівняння точності різних моделей машинного навчання за даними 2020 року (лінійна регресія, метод опорних векторів, багатошарова нейромережа, дерева рішень та їх ансамблі, ARIMA, Facebook Prophet з декількома варіантами параметрів ряду Фур'є для опису сезонності). Для формування додаткових ознак багатофакторних моделей використовувалась бібліотека Tsfresh, для оптимізації параметрів – техніка GridSearchCV. Аналіз довів беззаперечну перевагу моделі Facebook Prophet. Однак, зазначено, що традиційні методи налаштування параметрів цієї моделі є недостатньо ефективними для цієї задачі, а отже необхідно розробити нові методи чи удосконалити існуючі, щоб підвищити точність прогнозування.

2. Декомпозиція ряду з різним періодом сезонності показали, що має місце тижнева періодичність і ще мінімум 1-2 з іншим періодом, але це потребує окремого дослідження. Формалізовано модель Facebook Prophet для розв'язання поставленої задачі, з урахуванням виявлених видів сезонностей та аномалій. Охарактеризовано яку кількість і яких параметрів слід визначити для ідентифікації дійсно адекватної моделі та зазначено, що повний перебір їх варіантів займе забагато часу, а тому потрібні нові методи для їх ідентифікації. Запропоновано метод оцінювання порядку ряду Фур'є багатохвильового періодичного процесу, в якому оцінювання здійснюється лише по 10% верхівки однієї хвилі з використанням емпіричних співвідношень, що дозволяє здійснювати таке оцінювання, навіть, за умов, коли одна хвиля перекриває іншу і їх важко відокремити одну від іншої. Результат цього оцінювання варто використовувати як початкове наближення для подальшого уточнення методами машинного навчання. Також, розроблено ряд співвідношень для оцінювання періоду багатохвильового періодичного процесу по відстані між сусідними піками

та мінімальними значеннями, навіть, якщо хвилі не спадають до нуля. Запропоновано новий метод паралельно-послідовної багатопараметрової ідентифікації моделі Facebook Prophet для короткострокового прогнозування часового ряду кількості хворих на коронавірус в заданому регіоні, який відрізняється від існуючих більшою кількістю параметрів, що ідентифікуються: сила і розмір вікна впливу дат аномалій, ступінь регуляризації і тип моделі (адитивна чи мультиплікативна), порядок ряду Фур'є (уточнення початкового наближення) і ступінь регуляризації 3-х різних періодичних складових, які враховують внутрішньотижневі, тижневі і багатотижневі закономірності, характерні для інфекційних хвороб, у т.ч. коронавірусу, що дозволяє суттєво підвищити точність прогнозів та більш глибоко дослідити закономірності, які впливають на цей часовий ряд. Запропоновано удосконалений варіант цього методу з багатоітераційною оптимізацією параметрів, який дозволяє підвищити його точність. Для пошуку оптимальних значень параметрів використовується техніка HyperOpt та метод байєсівської оптимізації. Застосування цього удосконаленого багатоітераційного методу дозволило суттєво підвищити точність прогнозування, зокрема, за даними по Україні у 2021-2022 роках відносна похибка зменшилась у 3-10 разів, у порівнянні з методом, де була тільки одна така ітерація. Також, запропоновано робити прогнози для найбільш песимістичного та найбільш оптимістичного сценаріїв розвитку явища.

3. Удосконалено метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах, який відрізняється від існуючих використанням в якості інтегрального показника прогнозу на тиждень нахилу тижневої ділянки шматково-лінійної апроксимації тренду цього прогнозу, отриманого за моделлю, ідентифікованою із застосуванням запропонованого методу паралельно-послідовної багатопараметрової ідентифікації, що дозволяє аналізувати закономірності міжрегіонального поширення хвороби, ігноруючи місцеві особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень, тощо). Запропоновано здійснювати картографічну

візуалізацію прогнозів динаміки приросту кількості хворих на коронавірус, в якій радіус кругових діаграм у геометричному центрі чи центральному місті регіону буде пропорційним нахилу, визначеному за запропонованим методом узагальнення прогнозів, що дозволяє більш точно виявити закономірності розповсюдження захворювання на цій карті. Доцільно обробляти одразу декілька сусідніх регіонів. Продемонстровано роботу методу та картографічної візуалізації його результатів на рівні країн на карті світу та карті Європи. Колір кругової діаграми свідчить про зростання (червоний) чи зменшення (синій) кількості нових хворих. Зазначено, що ці нахили можна й кластеризувати, а потім за ними побудувати байєсівську мережу для виявлення закономірностей у кожному кластері окремо, що, також, дасть нові цікаві результати (такий підхід здобувач вже успішно випробував на іншій задачі з аналізу медичних даних).

4. Розроблена структура запропонованої інформаційної технології прогнозування часових рядів кількості хворих на коронавірус та охарактеризовано її блоки і модулі. Побудовано та описано UML-діаграми послідовності роботи модулів технології, UML-діаграма діяльності модулю прогнозування захворюваності та UML-діаграма модулю створення картограм нахилів трендів інфекційного впливу. Архітектурно інформаційна технологія реалізовується у вигляді комплексу програм-ноутбуків на Python та ряду датасетів для збереження вхідних, проміжних і результуючих даних.

5. Створено програмно-інформаційне забезпечення для автоматизації запропонованих методів і складових інформаційної технології. На основні його складові подано заявки на отримання свідоцтв про реєстрацію авторських прав на твір (комп’ютерну програму). Для реалізації вибрано мову Python та безкоштовну платформу датасайтнистів Kaggle, де зручно розміщати і Python-ноутбуки, і датасети. Універсальність редактору Kaggle Code дозволяє доволі швидко його адаптувати і до інших відомих середовищ для роботи з Python-ноутбуками: AWS SageMaker, Google Colab, Jupyter Notebook або JupyterLab пакету Anaconda, Microsoft Azure Notebooks, PyCharm, Visual Studio Code та ін. У Google-платформі

дасайнтистів Kaggle опубліковано у відкритий доступ 10 програм-ноутбуків у співавторстві з науковим керівником, які за 2020-2023 роки були переглянуті більше 36 тисяч разів (станом на 07.12.2023 р.). Здійснено випробування розробленого програмно-інформаційного забезпечення на реальних даних. Протягом 2020-2022 років здобувач брав участь у прогнозуванні щодобового приросту кількості хворих на коронавірус в 70 країнах світу, у т.ч. в Україні, – ці результати передавались в Робочу групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, офіційно створену при НАН України і членом якої є науковий керівник здобувача. Ці результати увійшли у звіти, в яких прямо вказано «Обчислення за допомогою моделі Facebook Prophet і аналіз отриманих результатів виконали завідувач кафедри системного аналізу та інформаційних технологій (САІТ) Вінницького національного технічного університету (ВНТУ) доктор технічних наук, професор В.Б. Мокін і аспірант кафедри САІТ ВНТУ А.В. Лосенко», що підтверджується актом впровадження результатів роботи, підписаним в.о. директора Інституту проблем математичних машин і систем НАН України. Усі ці звіти опубліковані на сайті Президії НАН України (https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx), а перед тим надсилались в РНБО України та МОЗ України для використання під час прийняття рішень щодо керування ситуацією в Україні.

Продемонстровано ефективність роботи запропонованої технології не тільки за даними для України, а й для інших 70 країн світу, що забезпечило доволі високу точність прогнозування. Проведено порівняння відносної похибки точності прогнозування моделі Prophet, що використовується в інформаційній технології, та моделі SEIR-U, розробленої Робочою групою НАНУ з прогнозування коронавірусу. Аналіз показав, що запропонована технологія здійснює прогнозування на 2 тижні вперед за даними 2021-2022 років з відносною похибкою, іноді в 1,2-2 разів меншу, ніж модель SEIR_U. Охарактеризовано впровадження у навчальний процес.

Розроблену інформаційну технологію можна, за певного адаптування, застосовувати і для інших нестаціонарних часових рядів багатохвильової природи техногенного чи природного характеру.

Результати дисертаційного дослідження опубліковані у 9 роботах, у т.ч. у 5 статтях у наукових фахових періодичних виданнях, 2 матеріалах доповідей на міжнародних конференціях, що увійшли у колективні монографії, опублікованими за результатами цих конференцій, 2 тезах доповідей на науково-практичних конференціях. Крім того, результати увійшли у 25 звітів Робочої групи НАНУ з прогнозування коронавірусу (2020-2022 рр.).

Наукова новизна дисертаційної роботи полягає в наступних положеннях:

1. Уперше запропоновано метод оцінювання порядку ряду Фур'є багатохвильового періодичного процесу, який відрізняється від існуючих тим, що оцінювання здійснюється лише по 10% верхівки однієї хвилі з використанням емпіричних співвідношень, що дозволяє здійснювати таке оцінювання, навіть, за умов, коли одна хвиля перекриває іншу і їх важко відокремити одну від іншої.

2. Уперше запропоновано метод паралельно-послідовної багатопараметрової ідентифікації моделі Facebook Prophet для короткострокового прогнозування часового ряду кількості хворих на коронавірус в заданому регіоні, який відрізняється від існуючих більшою кількістю параметрів, що ідентифікуються: сила і розмір вікна впливу дат аномалій, ступінь регуляризації і тип моделі (адитивна чи мультиплікативна), порядок ряду Фур'є і ступінь регуляризації 3-х різних періодичних складових, які враховують внутрішньотижневі, тижневі і багатотижневі закономірності, характерні для інфекційних хвороб, у т.ч. коронавірусу, що дозволяє суттєво підвищити точність прогнозів та більш глибоко дослідити закономірності, які впливають на цей часовий ряд. Запропоновано варіант цього методу з багатоітераційною оптимізацією параметрів, який дозволяє підвищити його точність.

3. Подальшого розвитку отримав метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах, який відрізняється від

існуючих використанням в якості інтегрального показника прогнозу на тиждень нахилу тижневої ділянки шматково-лінійної апроксимації тренду цього прогнозу, отриманого за моделлю, ідентифікованою із застосуванням запропонованого методу паралельно-послідовної багатопараметрової ідентифікації, що дозволяє аналізувати закономірності міжрегіонального поширення хвороби, ігноруючи місцеві особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень, тощо).

Практична цінність одержаних результатів полягає в наступному:

1. Отримані під час написання дисертаційної роботи наукові положення сприяли розробленню і створенню комплексу програм на Python, які автоматизують розрахунки за усіма запропонованими методами та підходами і реалізують побудову та використання розроблених моделей. У Google-платформі Kaggle опубліковано у відкритому доступі 10 програм-ноутбуків у співавторстві з науковим керівником, які за 2020-2023 роки були переглянуті більше 36 тисяч разів (станом на 07.12.2023 р.). Основні наукові результати та практичні розробки дисертаційної роботи пройшли апробацію на 4 наукових конференціях, у т.ч. на 2-х міжнародних у НАН України у м. Київ, та у нарадах Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні: протягом 2020-2022 рр., в результаті яких їх було включено у 25 звітів цієї робочої групи, що підтверджено актом впровадження з Інституту проблем математичних машин і систем НАН України.

2. Результати дослідження впроваджені у навчальний процес Вінницького національного технічного університету під час викладання дисципліни «Інформаційні технології моніторингу та аналізу даних» для студентів, які навчаються за освітньою програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності 126 Інформаційні системи та технології, а також – дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньою програмою «Інформаційні системи та технології» рівня «доктора філософії» цієї ж спеціальності 126 для навчання

знанням і навичкам застосування методів і моделей машинного навчання та інформаційних технологій оброблення і прогнозування даних часових рядів захворюваності на інфекційні захворювання. Результати дослідження можуть використовуватися як додатковий матеріал для навчання та поглибленого розуміння концепцій, пов'язаних із цією областю.

3. Досягнуто підвищення точності моделі. У період з листопада 2020 р. по лютий 2022 р. здійснювалось прогнозування кількості нових хворих на коронавірус в Україні. Метод дозволив в 1,2-2 рази зменшити відносну похибку прогнозування, у порівнянні з прогнозами за моделлю SEIR-U команди з НАН України у складі Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні. Запропоновано застосовувати ідентифіковану за цією технологією модель для прогнозування найбільш песимістичного та найбільш оптимістичного сценаріїв розвитку явища, тобто зміни кількості нових підтверджених випадків хвороби на коронавірус у майбутньому у заданій країні чи регіоні. Удосконалено цей метод, за рахунок використання багатоітераційного послідовно-паралельного визначення параметрів, що дозволяє отримувати більш адекватну модель. Застосування цього удосконаленого багатоітераційного методу дозволило суттєво підвищити точність прогнозування, зокрема, за даними по Україні у 2021-2022 роках відносна похибка зменшилась у 3-10 разів, у порівнянні з методом паралельно-послідовної ідентифікації параметрів багатохвильової моделі.

Отримані результати є корисними для:

- Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні (створена Розпорядженням Президії НАН України від 3 квітня 2020 р. № 198, базова установа - Інститут проблем математичних машин і систем НАН України), до складу якої входить науковий керівник здобувача – д.т.н., проф. Віталій Мокін. За час співпраці з листопаду 2020 р. по лютий 2022 р. було взято участь у написанні 25 звітів (під час найбільших зростань кількості хворих) і де в кожному вказано в тексті, що

співавтором розділу «Прогноз розвитку епідемії в Україні з використанням статистичної моделі часових рядів Facebook Prophet» був «аспірант кафедри САІТ ВНТУ А.В. Лосенко» (цей факт, також, підтверджується актом впровадження результатів роботи, підписаний в.о. директора Інституту проблем математичних машин і систем НАН України) (акт впровадження від 18 жовтня 2023 р.);

- навчального процесу Вінницького національного технічного університету під час викладання дисципліни «Інформаційні технології моніторингу та аналізу даних» для студентів, які навчаються за освітньою програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності 126 Інформаційні системи та технології, а також – дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньою програмою «Інформаційні системи та технології» рівня «доктора філософії» цієї ж спеціальності 126 для навчання знанням та навичкам застосування методів і моделей машинного навчання та інформаційних технологій оброблення і прогнозування даних часових рядів захворюваності на інфекційні захворювання (акт впровадження від 4 грудня 2023 р.).

Також необхідно зазначити, що розроблену інформаційну технологію прогнозування часових рядів кількості хворих на коронавірус методами машинного навчання можна, за певного адаптування, застосовувати і для інших нестаціонарних часових рядів багатохвильової природи техногенного чи природного характеру.

Ключові слова: оцінювання параметрів математичної моделі, моделювання та прогнозування, інтелектуальна інформаційна технологія, машинне навчання, коронавірус, штучний інтелект.

ABSTRACT

Losenko A.V. Information technology for predicting time series of the number of patients with coronavirus using machine learning methods - Qualifying scientific work with manuscript rights.

Dissertation for obtaining the scientific degree of Doctor of Philosophy in specialty 126 "Information systems and technologies". – Vinnytsia National Technical University, Vinnytsia, 2023.

The aim of the dissertation is to increase the accuracy of forecasting the number of coronavirus patients in the short-term using machine learning methods.

The dissertation is devoted to the solution of an actual scientific and technical problem, which consists in increasing the accuracy of forecasting the number of new coronavirus patients using machine learning methods.

The main scientific and practical results are as follows:

1. The main problems associated with forecasting the number of coronavirus patients in Ukraine and other countries are characterized, primarily due to problems with data related to the complexities of medical and organizational diagnosis, disease mutations, non-observance of quarantine conditions by the population, etc. An exploratory data analysis was carried out. The quality of data on other factors that affect the spread of the coronavirus (movement of the country's residents, weather data, government decisions (based on the Oxford tracker of the anti-coronavirus activity of the governments of the world), timeliness of publicizing confirmed tests for the coronavirus, etc.) has been analyzed. It has been proven that the influence of these factors is highly variable , the data are imprecise and noisy, arrive with different delays, and therefore, they should be considered not as additional signs for constructing a multiple regression, but it is better to formalize their sharp changes or abnormal values as data of anomalous behavior of the main time series, and then these anomalies can be take into account parametrically and reduce the problem to a one-factor one, that is, to building a model of only the time series of the number of new cases of the coronavirus. An overview of known models and methods of forecasting this symptom was conducted.

It was noted that due to problems with the data, preference should be given to machine learning models, rather than – a differential equation, for example SIR, SEIR or SEIR-U. A comparison of the accuracy of different machine learning models based on 2020 data was made (linear regression, support vector method, multilayer neural network, decision trees and their ensembles, ARIMA, Facebook Prophet with several options of Fourier series parameters to describe seasonality). The Tsfresh library was used to form additional features of multivariate models, and the GridSearchCV technique was used to optimize parameters. The analysis proved the undoubted superiority of the Facebook Prophet model. However, it is stated that the traditional methods of adjusting the parameters of this model are not efficient enough for this task, and therefore it is necessary to develop new methods or improve the existing ones in order to improve the forecasting accuracy.

2. Decomposition of the series with different periods of seasonality showed that there is a weekly periodicity and at least 1-2 with another period, but this requires a separate study. The Facebook Prophet model was formalized for solving the given task, taking into account the identified types of seasonality and anomalies. It is characterized how many and which parameters should be determined in order to identify a truly adequate model, and it is stated that a complete review of their options will take too much time, and therefore new methods for their identification are needed. A method for estimating the order of the Fourier series of a multi-wave periodic process is proposed, in which the estimation is carried out only for 10% of the top of one wave using empirical ratios, which allows such estimation to be carried out, even under conditions when one wave overlaps another, and it is difficult to separate them from one another. The result of this evaluation should be used as an initial approximation for further refinement using machine learning methods. Also, several relations have been developed for estimating the period of a multiwave periodic process by the distance between adjacent peaks and minimum values, even if the waves do not decrease to zero. A new method of parallel-serial multi-parameter identification of the Facebook Prophet model is proposed for short-term forecasting of the time series of the number of

coronavirus patients in a given region, which differs from the existing ones by a larger number of identified parameters: the strength and size of the window of influence of anomaly dates, the degree of regularization and the type of model (additive or multiplicative), the order of the Fourier series (refinement of the initial approximation) and the degree of regularization of 3 different periodic components, which take into account intra-weekly, weekly and multi-week regularities characteristic of infectious diseases, including coronavirus, which makes it possible to significantly increase the accuracy of forecasts and to more deeply investigate the patterns that affect this time series. An improved variant of this method with multi-iteration optimization of parameters is proposed, which allows to increase its accuracy. Techniques are used to search for optimal parameter values HyperOpt and the Bayesian optimization method. The application of this improved multi-iteration method made it possible to significantly increase the accuracy of forecasting according to data for Ukraine in 2021-2022, the relative error decreased by 3-10 times, compared to the method where there was only one such iteration. Also, it is proposed to make forecasts for the most pessimistic and the most optimistic scenarios of the development of the phenomenon.

3. The method of generalizing forecasts of the number of coronavirus patients in different regions has been improved, which differs from the existing ones by using as an integral indicator of the forecast for a week the slope of the weekly section of the piecewise linear approximation of the trend of this forecast, obtained according to the model identified using the proposed method of parallel-serial multiparameter identification, which allows analyzing patterns of interregional spread of the disease, ignoring local features (holidays and lockdowns of countries or regions, unique work schedule of laboratories and hospitals, etc.). It is proposed to carry out cartographic visualization of forecasts of the dynamics of the increase in the number of coronavirus patients, in which the radius of circular diagrams in the geometric center or the central city of the region will be proportional to the slope determined by the proposed method of generalizing forecasts, which allows more precisely to identify the patterns of the spread of the disease on this map. It is advisable to process several neighboring regions

at once. The operation of the method and cartographic visualization of its results at the level of countries on the world map and the map of Europe are demonstrated. The color of the pie chart indicates an increase (red) or decrease (blue) in the number of new patients. It is noted that these slopes can also be clustered, and then a Bayesian network can be built based on them to identify regularities in each cluster separately, which will also give interesting new results (the acquirer has already successfully tested this approach on another task of medical data analysis).

4. The structure of the proposed information technology for forecasting time series of the number of patients with coronavirus is developed and its blocks and modules are characterized. Constructed and described UML diagrams of the sequence of work of technology modules, UML diagram of the activity of the morbidity forecasting module and UML diagram of the module for creating cartograms of trends of infectious influence. Architectural information technology is implemented in the form of a set of Python notebook programs and several datasets for saving input, intermediate and resulting data.

5. Software and information support was created to automate the proposed methods and components of information technology. Applications for obtaining certificates of copyright registration for the work (computer program) have been submitted for its main components. For implementation, the Python language and Kaggle, a free datasite platform, were chosen, where it is convenient to host both Python notebooks and datasets. The versatility of the Kaggle Code editor allows you to quickly adapt it to other well-known environments for working with Python notebooks: AWS SageMaker, Google Colab, Jupyter Notebook or JupyterLab of the Anaconda package, Microsoft Azure Notebooks, PyCharm, Visual Studio Code, etc. In the Google platform of Dascientists Kaggle, 10 notebook programs co-authored with a scientific supervisor have been published for open access, which were viewed more than 36 thousand times in 2020-2023 (as of 07.12.2023). The developed software and information support was tested on real data. During 2020-2022, the winner took part in forecasting the daily increase in the number of coronavirus patients in 70 countries of the world, including in

Ukraine, these results were submitted to the Working Group on Mathematical Modeling of Problems Related to the SARS-CoV-2 Coronavirus Epidemic in Ukraine, officially created at the National Academy of Sciences of Ukraine and of which the scientific supervisor of the acquirer is a member. These results were included in the reports, which explicitly state that "Calculations using the Facebook Prophet model and analysis of the obtained results were performed by the head of the Department of System Analysis and Information Technologies (SAIT) of the Vinnytsia National Technical University (VNTU), Doctor of Technical Sciences, Professor V.B. Mokin and graduate student of the department of SAIT VNTU A.V. Losenko", which is confirmed by the act of implementation of work results, signed by acting director of the Institute of Problems of Mathematical Machines and Systems of the National Academy of Sciences of Ukraine. All these reports are published on the website of the Presidium of the National Academy of Sciences of Ukraine (https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx), and before that they were sent to the NSDC of Ukraine and the Ministry of Health of Ukraine for use during the adoption decisions on managing the situation in Ukraine.

The effectiveness of the proposed technology was demonstrated not only according to data for Ukraine, but also for other 70 countries of the world, which ensured a fairly high accuracy of forecasting. A comparison of the relative error of forecasting accuracy of the Prophet model, used in information technology, and the SEIR-U model, developed by the Working Group of the National Academy of Sciences on Corona Virus Prediction su. The analysis showed that the proposed technology makes forecasts 2 weeks ahead based on the data of 2021-2022 with a relative error sometimes 1.2-2 times smaller than the SEIR_U model. The introduction into the educational process is characterized.

The developed information technology can, with certain adaptation, be applied to other non-stationary time series of a man-made or natural multi-wave nature.

The results of the dissertation research were published in 9 works, including in 5 articles in scientific professional periodicals, 2 materials of reports at international conferences included in collective monographs published as a result of these

conferences, 2 abstracts of reports at scientific and practical conferences. In addition, the results were included in 25 reports of the Working Group of the National Academy of Sciences on the prediction of the coronavirus (2020-2022).

The scientific novelty of the dissertation consists in the following provisions:

1. For the first time, a method for evaluating the order of the Fourier series of a multi-wave periodic process is proposed, which differs from the existing ones in that the evaluation is carried out only for 10% of the top of one wave using empirical ratios, which allows such an evaluation, even under conditions when one wave overlaps the other and it is difficult to separate them one from the other.

2. For the first time, a method of parallel-sequential multi-parameter identification of the Facebook Prophet model is proposed for short-term forecasting of the time series of the number of coronavirus patients in a given region, which differs from the existing ones by a larger number of parameters that can be identified: the strength and size of the window of influence of anomaly dates, the degree of regularization and the type of model (additive or multiplicative), the order of the Fourier series and the degree of regularization of 3 different periodic components, which take into account intra-weekly, weekly and multi-week regularities characteristic of infectious diseases, including coronavirus, which makes it possible to significantly increase the accuracy of forecasts and to more deeply investigate the patterns that affect this time series. A variant of this method with multi-iteration optimization of parameters is proposed, which allows to increase its accuracy.

3. The method of generalizing forecasts of the number of coronavirus patients in different regions received further development, which differs from the existing ones by using as an integral indicator of the forecast for the week the slope of the weekly section of the piecewise linear approximation of the trend of this forecast, obtained according to the model identified using the proposed method in parallel consistent multi-parameter identification, which allows analyzing patterns of interregional spread of the disease, ignoring local features (holidays and lockdowns of countries or regions, unique work schedule of laboratories and hospitals, etc.).

The practical value of the obtained results is as follows:

1. The scientific provisions obtained during the writing of the dissertation contributed to the development and creation of a set of Python programs that automate calculations using all the proposed methods and approaches and implement the construction and use of the developed models. In the Google platform Kaggle, 10 notebook programs co-authored with a scientific supervisor were published in open access, which were viewed more than 36 thousand times in 2020-2023 (as of 07.12.2023). The main scientific results and practical developments of the dissertation were approved at 4 scientific conferences, including at the 2nd international meetings at the National Academy of Sciences of Ukraine in Kyiv, and at meetings of the Working Group on Mathematical Modeling of Problems Related to the SARS-CoV-2 Coronavirus Epidemic in Ukraine: during 2020-2022, as a result of which they were included in 25 reports of this working group, which is confirmed by the act of implementation from the Institute of Problems of Mathematical Machines and Systems of the National Academy of Sciences of Ukraine.

2. The results of the research are implemented in the educational process of the Vinnytsia National Technical University during the teaching of the discipline "Information technologies of monitoring and data analysis" for students studying under the educational program "Information technologies of data and image analysis" at the "master's" level, specialty 126 Information systems and technologies , as well as the disciplines "Internet of Things and Intelligent Data Analysis" for graduate students studying under the educational program "Information Systems and Technologies" at the "Doctor of Philosophy" level of the same specialty 126 for teaching knowledge and skills in the application of methods and models of machine learning and information technologies processing and forecasting of time series data on the incidence of infectious diseases. The results of the study can be used as additional material for learning and in-depth understanding of concepts related to this field.

3. Improved accuracy of the model was achieved. In the period from November 2020 to February 2022, the number of new coronavirus patients in Ukraine was forecast.

The method made it possible to reduce the relative error of forecasting by 1.2-2 times, compared to forecasts based on the SEIR-U model of the team from the National Academy of Sciences of Ukraine (Research Group on Mathematical Modeling of Problems Related to the SARS-CoV-2 Coronavirus Epidemic in Ukraine). It is proposed to use the model identified by this technology to forecast the most pessimistic and the most optimistic scenarios of the development of the phenomenon, i.e. the change in the number of new confirmed cases of the coronavirus disease in the future in a given country or region. This method has been improved, due to the use of multi-iterative serial-parallel determination of parameters, which allows obtaining a more adequate model. The application of this improved multi-iteration method made it possible to significantly increase the accuracy of forecasting according to the data for Ukraine in 2021-2022, the relative error decreased by 3-10 times, compared to the method of parallel-serial identification of the parameters of the multi-wave model.

The obtained results are useful for:

- Reasearch group on mathematical modeling of problems related to the SARS-CoV-2 coronavirus epidemic in Ukraine (created by Order of the Presidium of the National Academy of Sciences of Ukraine dated April 3, 2020, No. 198, base institution - Institute of Problems of Mathematical Machines and Systems of the National Academy of Sciences of Ukraine), to which includes the recipient's scientific supervisor - Doctor of Technical Sciences, Prof. Vitaly Mokin. During the period of cooperation from November 2020 to February 2022, 25 reports were written (at the time of the greatest increase in the number of patients) and where it is indicated in the text of each one that the co-author of the section "Forecast of the development of the epidemic in Ukraine using a statistical model of time ranks of Facebook Prophet" was a "graduate student of the department of SAIT VNTU A.V. Losenko" (this fact is also confirmed by the act of implementation of work results, signed by the acting director of the Institute of Mathematical Machines and Systems Problems of the National Academy of Sciences of Ukraine) (act of implementation dated October 18, 2023);

- the educational process of the Vinnytsia National Technical University during the teaching of the discipline "Information technologies of monitoring and data analysis" for students studying under the educational program "Information technologies of data and image analysis" at the "master's" level, specialty 126 Information systems and technologies, as well as - disciplines "Internet of Things and Intelligent Data Analysis" for graduate students studying under the educational program "Information Systems and Technologies" at the "Doctor of Philosophy" level of the same specialty 126 for training knowledge and skills in the application of methods and models of machine learning and information technologies for processing and forecasting time data incidence rates for infectious diseases (implementation act of December 4, 2023).

It should also be noted that the developed information technology for predicting time series of the number of coronavirus patients using machine learning methods can, with certain adaptation, be applied to other non-stationary time series of a man-made or natural multi-wave nature.

Keywords: estimation of mathematical model parameters, modeling and forecasting, intelligent information technology, machine learning, coronavirus, artificial intelligence.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

[1] В. Мокін, А. Лосенко, М. Дратований, «Інтелектуальна технологія аналізу та передбачення цін на вживані автомобілі», *Вісник Вінницького політехнічного інституту*, № 6, с. 62-72 (Груд 2019.) **(Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[2] В. Мокін, А. Лосенко, А. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків на коронавірус SARS-CoV-2 в Україні на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*. – 2020. – № 5. – С. 71–83. **(Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[3] В. Мокін, А. Лосенко, А. Ящолт, «Інформаційна технологія аналізу та прогнозування багатохвильової кількості нових випадків захворювань на коронавірус COVID-19 на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*. - Вип. 6, С. 65–75, 2020. **(Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[4] В. Мокін, М. Дратований, А. Лосенко, С. Жуков, «Прогнозування хвиль коронавірусу на основі відновленої когнітивної карти міжрегіонального впливу», *Інформаційні технології та комп'ютерна інженерія*, 2021, Том 52, Вип. 3, с. 86–94, Груд 2021. **(Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[5] А. Лосенко, «Інформаційна технологія прогнозування часового ряду кількості хворих на коронавірус на основі моделі Facebook Prophet», *Вісник Вінницького політехнічного інституту*, № 5, с. 50-59 (Жовтень 2023.) (**Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126**)

[6] В. Мокін, А. Лосенко, «Картування тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи протягом січня-березня 2021 року», *на L науково-технічній конференції підрозділів ВНТУ*, Вінниця, 10-12 березня 2021 р. – Електрон. текст. дані. – 2021. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2021/paper/view/12849

[7] В. Мокін, А. Лосенко, А. Ящолт, А. Гевеленко, «Прогнозування тижневих трендів кількості нових хворих на коронавірус у країнах світу», Колективна монографія за матеріалами XX Міжнародної науково-практичної конференції «Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях», Жовтень 2021, с. 209-212. Електрон. текст. дані, 2021. – Режим доступу: https://itgip.org/wp-content/uploads/2021/10/1_zbirka_2021.pdf.

[8] Д. Шмундяк, А. Лосенко, В. Мокін, «Огляд підходів до визначення порядку Фур'є у моделі Facebook Prophet для моделювання сезонної складової часового ряду», *на LII Науково-технічній конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету*, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/17200/14329.

[9] В. Мокін, А. Лосенко «Інформаційна технологія короткострокового прогнозування кількості нових хворих на коронавірус на основі моделі Facebook Prophet», *Інформаційно-комунікаційні технології для перемоги та відновлення, Колективна монографія за матеріалами XXII Міжнародної науково-практичної конференції «Інформаційно-комунікаційні технології та сталий розвиток»*, Київ, 14-15 листопада 2023 р. За заг. ред. С.О. Довгого. – К.: ТОВ «Видавництво «Юстон», 2023. – С. 27-30. URL: https://itgip.org/wp-content/uploads/2023/11/1_zbirka_08_11_23-1-1.pdf

[10] І. Бровченко, Р. Беженар, В. Мокін , А. Лосенко та ін. 25 звітів Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні при НАН України «Прогноз РГ-29» (Опубл. 23.11.2020), «Прогноз РГ-30» (Опубл. 30.11.2020), «Прогноз РГ-31» (Опубл. 07.12.2020), «Прогноз РГ-32» (Опубл. 14.12.2020), «Прогноз РГ-33» (Опубл. 21.12.2020), «Прогноз РГ-34» (Опубл. 28.12.2020), «Прогноз РГ-35» (Опубл. 11.01.2021), «Прогноз РГ-36» (Опубл. 25.01.2021), «Прогноз РГ-37» (Опубл. 08.02.2021), «Прогноз РГ-38» (Опубл. 22.02.2021), «Прогноз РГ-39» (Опубл. 10.03.2021), «Прогноз РГ-40» (Опубл. 23.03.2021), «Прогноз РГ-41» (Опубл. 06.04.2021), «Прогноз РГ-42» (Опубл. 20.04.2021), «Прогноз РГ-51» (Опубл. 14.09.2021), «Прогноз РГ-52» (Опубл. 28.09.2021), «Прогноз РГ-53» (Опубл. 12.10.2021), «Прогноз РГ-54» (Опубл. 26.10.2021), «Прогноз РГ-55» (Опубл. 09.11.2021), «Прогноз РГ-56» (Опубл. 23.11.2021), «Прогноз РГ-57» (Опубл. 07.12.2021), «Прогноз РГ-58» (Опубл. 21.12.2021), «Прогноз РГ-60» (Опубл. 26.01.2022), «Прогноз РГ-61» (Опубл. 08.02.2022), «Прогноз РГ-62» (Опубл. 22.02.2022). Режим доступу на сайті Президії НАН України: https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ 26
ВСТУП..... 27
РОЗДІЛ 1 АНАЛІЗ ПРОБЛЕМ ПРОГНОЗУВАННЯ ЗАХВОРЮВАНОСТІ ЛЮДЕЙ НА КОРОНАВІРУС 35
1.1 Огляд основних проблем прогнозування захворюваності на коронавірус... 35
1.2 Аналіз відомих методів та моделей динаміки захворюваності на коронавірус…..... 37
1.2.1. Моделі SIR, SEIR, SEIR-U на основі диференціальних рівнянь..... 38
1.2.2. Моделі часових рядів на основі АРПКС (ARIMA) 39
1.2.3. Моделі часових рядів на основі вейвлет-перетворень 41
1.2.4. Моделі часових рядів на основі Facebook Prophet..... 42
1.2.5. Інші моделі машинного навчання 45
1.2.6. Метрика моделей машинного навчання 47
1.3 Аналіз впливу різних факторів і проблем на кількість хворих на коронавірус в Україні 49
1.4 Висновки до розділу та постановка задач дослідження 54
РОЗДІЛ 2 РОЗРОБЛЕННЯ МЕТОДІВ ТА ТЕОРЕТИЧНИХ ОСНОВ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС57
2.1 Розвідувальний аналіз даних 57
2.1.1. Аналіз стаціонарності ряду. 58
2.1.2. Аналіз сезонності ряду..... 59
2.1.3. Побудова автокореляційної та часткової автокореляційної функцій. .. 61
2.1.4. Прогнозування даних за різними моделями. 64
2.1.5. Розширення множини дат аномалій. 68
2.1.6. Висновки розвідувального аналізу даних. 72

2.2 Розроблення методу ідентифікації структури та параметрів моделі для прогнозування кількості нових випадків на коронавірус на основі моделі FB Prophet 73
2.3 Розроблення методу оцінювання порядку ряду Фур'є для апроксимації багатоденного періодичного процесу зміни кількості нових випадків на коронавірус на основі моделі FB Prophet 87
2.4 Розширення методів на багатоітераційний випадок 97
2.5 Порівняльний аналіз трендів прогнозу сусідніх регіонів 101
2.6 Висновки до розділу 104
РОЗДІЛ 3 РОЗРОБЛЕННЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС 110
3.1 Проєктування структури інформаційної технології 110
3.2 Розроблення UML-діаграм інформаційної технології 114
3.2.1 UML-діаграма послідовності роботи модулів технології 114
3.2.2 UML-діаграма діяльності модулю прогнозування захворюваності 116
3.2.3 UML-діаграма діяльності модулю створення картограм нахилів трендів інфекційного впливу 118
3.3 Порівняння ефективності створеної інформаційної технології з аналогами 120
3.4 Висновки до розділу 123
РОЗДІЛ 4 ПРИКЛАДНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ 124
4.1 Прогнозування динаміки приросту щодобової кількості хворих на коронавірус в Україні 124
4.2 Прогнозування динаміки приросту щодобової кількості хворих на коронавірус у 69 країнах світу за одноетапним методом ідентифікації моделі FB Prophet . 132
4.3 Картографічна візуалізація прогнозів динаміки приросту щодобової кількості хворих на коронавірус у провідних країнах світу 136
4.4 Впровадження розробленої технології у навчальний процес 141
4.5 Висновки до розділу 142

ВИСНОВКИ 144
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ..... 149
ДОДАТКИ 163
ДОДАТОК А АНАЛІТИЧНІ ЗВІТИ НАДАНІ НАНУ ПРОТЯГОМ 2020-2022 РР. 164
ДОДАТОК Б ТАБЛИЦЯ ГРАФІКІВ ПРОГНОЗУ ЗАХВОРЮВАНОСТІ НА КОРОНАВІРУС З 2020 ПО 2022 РР.. 189
ДОДАТОК В РОЗШИРЕНА ПОРІВНЯЛЬНА ТАБЛИЦЯ ПОХИБОК МОДЕЛЕЙ FACEBOOK PROPHET ТА SIER-U ПРОТЯГОМ 2020-2022 РР. 197
ДОДАТОК Г СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ 198
ДОДАТОК Д АКТ ПРО ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНОЇ РОБОТИ..... 201

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

АРПКС – модель авторегресії та проінтегрованого ковзного середнього

ВНТУ – Вінницький національний технічний університет

НАН України – Національна академія наук України

РГ або РГ по ковіду – Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, створена Розпорядженням Президії НАН України від 3 квітня 2020 р. № 198, базова установа – Інститут проблем математичних машин і систем НАН України, термін функціонування: 03.04.2020 р. – 23.02.2022 р. (24.02.2022 р. тимчасово призупинила свою діяльність), склад Робочої групи: https://www.nas.gov.ua/UA/Colegial/Pages/Default.aspx?CID=000000188

Кафедра САІТ – кафедра системного аналізу та інформаційних технологій ВНТУ

ARIMA – autoregressive and integrated moving average model

COVID-19 – коронавірус

FB Prophet чи Prophet – модель Facebook Prophet для моделювання і прогнозування часових рядів

UML – Unified Modeling Language

ВСТУП

Обґрунтування вибору теми дослідження. Пандемія COVID-19 зробила актуальною задачу прогнозування захворюваності інфекційних хвороб та коронавірусу в цілому. Але, враховуючи кількість чинників, які впливають на розповсюдження хвороби та хаотичність їх значень, з'явилась потреба в оптимізації методів прогнозування. Вже існують моделі прогнозування часових рядів, створені для вирішення такої типової задачі, але дане інфекційне захворювання створює нові виклики, такі як оптимальна ідентифікація параметрів моделі прогнозування, а також здатність характеризувати багатохвильову природу часового ряду захворюваності. Крім, власне прогнозування захворюваності в тому чи іншому регіоні (або країні), варто також звернути увагу на аналіз можливого інфекційного впливу сусідніх регіонів, тому що це безпосередньо впливає на об'єкт дослідження. Всі вище описані аспекти дослідження вимагають глибшого розуміння вже наявних моделей прогнозування часових рядів, а також інтеграції таких моделей в єдину інформаційну технологію, що повинна використовуватись для прогнозування кількості хворих на коронавірус, оскільки дана модель буде лише основою системи, а методи цієї технології будуть ідентифікувати та оптимізувати параметри моделі, здійснювати розвідувальний аналіз даних, а також визначати тренди розповсюдження захворювання та використовувати зміни тренду для апроксимації інфекційного впливу, що дозволить вчасно приймати науково обґрунтовані рішення з мінімізації негативного впливу цього захворювання чи його локалізації. Подібними дослідженнями займаються такі вчені, як: Бровченко І.О., Беженар Р.В., Скрипниченко М. І., Althaus C.L., Peipei Wanga, Мокін В.Б., Квєтний Р.Н., Голуб С.В., Бідюк П.І., та ін. [1-8]. Для розв'язання таких задач можуть використовуватись як класичні моделі часових рядів, так і різні інтелектуальні моделі машинного навчання, якими займаються вчені в усіх країнах світу, у т.ч.

такі вчені, як Мокін Б.І., Davies N.G., Мокін О.Б., Бісікало О.В., Іванов Ю.Ю. та ін [9-13].

Одними з найважливіших проблем створення ефективної інтелектуальної інформаційної технології прогнозування даних є вибір оптимальної структури моделі та розроблення оптимального алгоритму ідентифікації її параметрів та аналізу і співставлення отриманих за нею прогнозів, а також пошук та оброблення існуючих наборів даних, за допомогою яких буде здійснюватися розроблення цих моделей і верифікація технології в цілому. У випадку з такою актуальною проблемою, як пандемія COVID-19, набори даних захворюваності знаходяться у публічному доступі на міжнародному рівні, що спрощує пошук та їх використання. До російського вторгнення, також, через API можна було завантажити з порталу Ради національної безпеки та оборони України (РНБО) добре структуровані різні дані по Україні, навіть у розрізі адміністративних областей, а на сайті МОЗ Україні були доступними Excel-таблиці з даними по окремих лікарнях, лабораторіях, населених пунктах та ін.

Отже, виникає потреба у розробленні нової інтелектуальної інформаційної технології прогнозування кількості хворих на коронавірус, яка може здійснювати точне прогнозування та ефективний аналіз та візуалізацію прогнозів по різних регіонах для виявлення певних закономірностей та в автоматизації методів та алгоритів цієї технології у вигляді програмного забезпечення для її прикладного застосування.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у межах НДР кафедри САІТ 28-К7 «Інформаційні технології та системи моніторингу, моделювання та системного інтелектуального аналізу й оптимізації даних у складних об’єктах» протягом 2020-2023 рр. Але основна робота велась у межах співпраці з Робочою групою з математичного моделювання проблем, пов’язаних з епідемією коронавірусу SARS-CoV-2 в Україні (створена Розпорядженням Президії НАН України від 3 квітня 2020 р. № 198, базова установа - Інститут проблем математичних машин і

систем НАН України), до складу якої входив науковий керівник здобувача – д.т.н., проф. Віталій Мокін. За час співпраці з листопаду 2020 р. по лютий 2022 р. було взято участь у написанні 25 звітів (під час найбільших зростань кількості хворих) і де в кожному вказано в тексті, що співавтором розділу «Прогноз розвитку епідемії в Україні з використанням статистичної моделі часових рядів Facebook Prophet» був «аспірант кафедри САІТ ВНТУ А.В. Лосенко» (цей факт, також, підтверджується актом впровадження результатів роботи, підписаний в.о. директора Інституту проблем математичних машин і систем НАН України).

Мета і завдання дослідження. Метою дисертаційної роботи є підвищення точності прогнозування кількості хворих на коронавірус у короткостроковій перспективі за допомогою методів машинного навчання.

У результаті проведеного аналізу для досягнення поставленої мети сформульовано нижчевказані задачі дослідження.

1. Здійснити аналіз відомих методів та моделей прогнозування часових рядів, які можуть бути використані для створення інформаційної технології прогнозування захворюваності на коронавірус та проаналізувати основні проблеми, які виникають під час такого прогнозування.

2. Розробити методи ідентифікації структури та параметрів моделі для прогнозування кількості нових випадків захворювання на коронавірус, які дозволять підвищити точність прогнозування та допоможуть виявляти нові закономірності щодо даних.

3. Розробити метод картографічної візуалізації прогнозів динаміки приросту кількості хворих на коронавірус, що дозволить більш точно виявити закономірності розповсюдження захворювання на карті обраного регіону.

4. Розробити структуру запропонованої інформаційної технології та її компонентів.

5. Створити програмно-інформаційне забезпечення для автоматизації запропонованих методів і складових інформаційної технології та випробувати його на реальних даних.

Об'єкт дослідження – прогнозування кількості хворих на коронавірус у різних регіонах.

Предмет дослідження – інформаційна технологія, методи, алгоритми та програми для прогнозування кількості хворих на коронавірус.

Методи дослідження містять загальнонаукову методологію проведення досліджень і принципи системного підходу, а саме: аналіз літературних джерел, відкритих даних, експериментальні дослідження, об'єктно-орієнтоване програмування, системний аналіз, методи теорії часових рядів, методи машинного навчання та інші інтелектуальні методи.

Наукова новизна отриманих результатів.

1. Уперше запропоновано метод оцінювання порядку ряду Фур'є багатохвильового періодичного процесу, який відрізняється від існуючих тим, що оцінювання здійснюється лише по 10% верхівки однієї хвилі з використанням емпіричних співвідношень, що дозволяє здійснювати таке оцінювання, навіть, за умов, коли одна хвиля перекриває іншу і їх важко відокремити одну від іншої.

2. Уперше запропоновано метод паралельно-послідовної багатопараметрової ідентифікації моделі Facebook Prophet для короткострокового прогнозування часового ряду кількості хворих на коронавірус в заданому регіоні, який відрізняється від існуючих більшою кількістю параметрів, що ідентифікуються: сила і розмір вікна впливу дат аномалій, ступінь регуляризації і тип моделі (адитивна чи мультиплікативна), порядок Фур'є і ступінь регуляризації 3-х різних періодичних складових, які враховують внутрішньотижневі, тижневі і багатотижневі закономірності, характерні для інфекційних хвороб, у т.ч. коронавірусу, що дозволяє суттєво підвищити точність прогнозів та більш глибоко дослідити закономірності, які впливають на цей часовий ряд. Запропоновано варіант цього методу з багатоітераційною оптимізацією параметрів, який дозволяє підвищити його точність.

3. Подальшого розвитку отримав метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах, який відрізняється від існуючих

використанням в якості інтегрального показника прогнозу на тиждень нахилу тижневої ділянки шматково-лінійної апроксимації тренду цього прогнозу, отриманого за моделлю, ідентифікованою із застосуванням запропонованого методу паралельно-послідовної багатопараметрової ідентифікації, що дозволяє аналізувати закономірності міжрегіонального поширення хвороби, ігноруючи місцеві особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень, тощо).

Практичне значення отриманих результатів. Створено комплекс програм на Python, які автоматизують розрахунки за усіма запропонованими методами та підходами і реалізують побудову та використання розроблених моделей. У Google-платформі дасайнтистів Kaggle опубліковано у відкритий доступ 10 програм-ноутбуків у співавторстві з науковим керівником, які за 2020-2023 роки були переглянуті більше 36 тисяч разів (станом на 07.12.2023 р.)

Протягом 2020-2022 років здобувач брав участь у прогнозуванні щодобового приросту кількості хворих на коронавірус в 70 країнах світу, у т.ч. в Україні, – ці результати передавались в Робочу групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, офіційно створену при НАН України і членом якої є науковий керівник здобувача. Ці результати увійшли у звіти, в яких прямо вказано «Обчислення за допомогою моделі Facebook Prophet і аналіз отриманих результатів виконали завідувач кафедри системного аналізу та інформаційних технологій (САІТ) Вінницького національного технічного університету (ВНТУ) доктор технічних наук, професор В.Б. Мокін і аспірант кафедри САІТ ВНТУ А.В. Лосенко» (це підтверджується актом впровадження з НАН України). Усі ці звіти опубліковані на сайті Президії НАН України [14], а перед тим надсилались в РНБО України та МОЗ України для використання під час прийняття рішень щодо керування ситуацією в Україні.

Також, результати роботи впроваджено у навчальний процес і наукову діяльність кафедри системного аналізу та інформаційних технологій ВНТУ.

Особистий внесок здобувача

Усі теоретичні результати, що виносяться на захист, отримані здобувачем особисто. У роботі [15] здобувачеві належать усі теоретичні результати. В автоматизації розроблених методів на Python на базі платформи Kaggle та їх застосуванні за даними Робочої групи по ковіду брав участь науковий керівник, проф. Віталій Мокін, який був членом тієї Робочої групи. У роботах, опублікованих у співавторстві, автору дисертаційної роботи належать такі результати: [16] – визначені параметри моделі, які впливають на точність прогнозування захворюваності, запропоновано удосконалений метод ідентифікації параметрів моделі машинного навчання прогнозування кількості нових хворих на коронавірус, здійснено порівняльний аналіз моделей машинного навчання, та обрана модель Facebook Prophet, за допомогою якої здійснювався прогноз захворюваності; [17] – проведено дослідження багатохвильового прогнозування кількості нових випадків на коронавірус; [18], [19] – розроблено метод картування трендів динаміки захворюваності на коронавірус в певному регіоні світу, в якому було формалізовано припущення про вплив країн регіону на розповсюдження захворювання; [20] – запропонований удосконалений метод паралельно-послідовної ідентифікації параметрів моделі Facebook Prophet, з урахуванням практичного застосування моделі; [21] – запропонований новий багатоітераційний метод ідентифікації параметрів моделі Facebook Prophet; автоматизація і застосування багатоітераційного методу здійснювались здобувачем повністю самостійно.

Апробація матеріалів дисертації. Основні наукові результати та практичні розробки дисертаційної роботи пройшли апробацію на 4 наукових конференціях, у т.ч. на 2-х міжнародних у НАН України у м. Київ, та у нарадах Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні: протягом 2020-2022 рр., в результаті яких їх було включено у 25 звітів цієї РГ (в усі звіти за періоди під час найбільшого зростання кількості хворих):

- L науково-технічна конференція підрозділів ВНТУ, Вінниця, 10–12 березня 2021 р.;

- XX Міжнародна науково-практична конференція "Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях" (НАН України, жовтень 2021);

- LII Науково-технічна конференція факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, Вінниця, 21 – 23 червня 2023 р;

- XXII Міжнародна науково-практична конференція "Інформаційно-комунікаційні технології та сталий розвиток" (НАН України, листопад 2023);

- 25 звітів Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні «Прогноз РГ-29» (Опубл. 23.11.2020), «Прогноз РГ-30» (Опубл. 30.11.2020), «Прогноз РГ-31» (Опубл. 07.12.2020), «Прогноз РГ-32» (Опубл. 14.12.2020), «Прогноз РГ-33» (Опубл. 21.12.2020), «Прогноз РГ-34» (Опубл. 28.12.2020), «Прогноз РГ-35» (Опубл. 11.01.2021), «Прогноз РГ-36» (Опубл. 25.01.2021), «Прогноз РГ-37» (Опубл. 08.02.2021), «Прогноз РГ-38» (Опубл. 22.02.2021), «Прогноз РГ-39» (Опубл. 10.03.2021), «Прогноз РГ-40» (Опубл. 23.03.2021), «Прогноз РГ-41» (Опубл. 06.04.2021), «Прогноз РГ-42» (Опубл. 20.04.2021), «Прогноз РГ-51» (Опубл. 14.09.2021), «Прогноз РГ-52» (Опубл. 28.09.2021), «Прогноз РГ-53» (Опубл. 12.10.2021), «Прогноз РГ-54» (Опубл. 26.10.2021), «Прогноз РГ-55» (Опубл. 09.11.2021), «Прогноз РГ-56» (Опубл. 23.11.2021), «Прогноз РГ-57» (Опубл. 07.12.2021), «Прогноз РГ-58» (Опубл. 21.12.2021), «Прогноз РГ-60» (Опубл. 26.01.2022), «Прогноз РГ-61» (Опубл. 08.02.2022), «Прогноз РГ-62» (Опубл. 22.02.2022). Режим доступу на сайті Президії НАН України: https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx

Публікації.

Всього за тематикою дисертаційного дослідження опубліковано 9

наукових праць. Опубліковано 5 статей у фахових журналах України з технічних наук [15], [16], [17], [20], [22], 2 тези доповідей на міжнародних конференціях, що увійшли у колективні монографії [19], [21], та 2 тез доповідей на науково-практичних конференціях [18], [23], також матеріали дисертаційного дослідження увійшли у 25 аналітичних звітів Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні [24]-[45].

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг дисертації становить 203 сторінок, у тому числі: 121 сторінок основного тексту, 63 рисунків, 4 таблиці, список використаних джерел із 100 найменувань, кількість додатків – 5.

РОЗДІЛ 1
АНАЛІЗ ПРОБЛЕМ ПРОГНОЗУВАННЯ ЗАХВОРЮВАНОСТІ ЛЮДЕЙ НА КОРОНАВІРУС

1.1 Огляд основних проблем прогнозування захворюваності на коронавірус

Пандемія коронавірусу COVID-19, викликана інфекцією SARS-CoV-2, призвела до ряду серйозних проблем та складнощів в Україні. Однією з головних проблем є те, що хвороба проявляється зі значним запізненням, що ускладнює вчасне виявлення та контроль її поширення. Ця запізненість є результатом низки факторів, включаючи недостатню доступність діагностичних тестів для всіх громадян та довгі терміни очікування результатів аналізів. Це призводить до нестачі об'єктивної інформації про розповсюдження хвороби та ускладнює прийняття ефективних заходів для боротьби з нею чи мінімізації наслідків. Окрім того, в Україні існує проблема з недостатньою кількістю доступних аналізів, оскільки не всі громадяни мають можливість пройти їх в обов'язковому порядку. Це створює нерівність у доступі до медичної допомоги та може призвести до недооцінювання справжніх масштабів епідемії. Додатковою проблемою є відмінність динаміки поширення хвороби в Україні та за кордоном. Різниця у відсотку вакцинованого населення та інші фактори можуть призвести до відмінностей у розповсюдженні та контролі хвороби. Нелегко, також, населенню дотримуватися карантинних заходів в Україні, і ця проблема є актуальною. Також, складності додає наявність різних штамів хвороби та їхнє одночасне розповсюдження в популяції. Офіційні дані про захворюваність можуть надходити із затримкою, лабораторії не завжди встигають внести дані вчасно, особливо під час стрімкого зростання захворюваності що ускладнює моніторинг та аналіз ситуації. Всі ці чинники по-різному впливають на динаміку розповсюдження коронавірусу у світі, це можна спостерігати, порівнявши графіки

захворюваності в порівнянні з епідеміологічною ситуацією в Україні. На рис. 1.1 наведено відповідний графік порівняння.

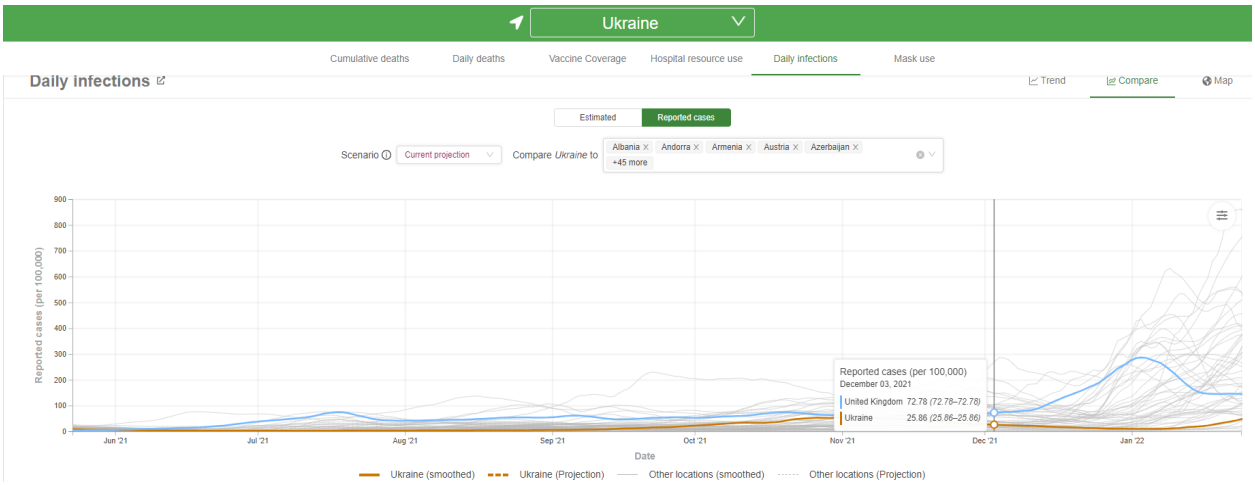


Рис. 1.1. Графік порівняння розповсюдження COVID-19 у країнах світу та в Україні [46]

Усі ці проблеми потребують комплексного підходу до аналізу самих наборів даних та застосування моделей машинного навчання, які в подальшому інтегруються в інформаційні технології прогнозування захворюваності та смертності на коронавірус.

Зважаючи на все описане вище, першочерговим завданням стає моделювання і прогнозування щоденної кількості нових підтверджених випадків хвороби за тестами, основаними на полімеразній ланцюговій реакції (ПЛР), оскільки інші типові показники (кількість госпіталізованих, кількість тих, хто видужав, кількість смертельних випадків, кількість позитивних експрес-тестів та ін.), вже є наслідками розвитку хвороби у тих, у кого вона підтвердилась. Другим важливим завданням є аналіз нових закономірностей, які дозволяє виявити ідентифікована математична модель за даними різних країн, які допомогли б краще зрозуміти ситуацію і вчасно вжити необхідних заходів з її покращення.

Усі ці задачі, звичайно, були актуальними під час активної фази пандемії коронавірусу у світі, яка, в кінці 2023 року, вже – в минулому, але важливо навчитись будувати більш адекватні математичні моделі за даними минулих

років, оскільки, по-перше, згодом, на жаль, матимуть місце нові епідемії чи пандемії, а по-друге, може з’явитись якийсь новий ще більш заразний та небезпечний штам коронавірусу, тому треба буде бути вже готовими до вирішення можливих проблем.

1.2 Аналіз відомих методів та моделей динаміки захворюваності на коронавірус

Враховуючи характер критичності поширення коронавірусу у світі, особливо у 2020-2022 роках, для вирішення задачі прогнозування захворюваності почали застосовувати різноманітні моделі машинного навчання, найбільш популярними серед яких є такі моделі [47-60]:

- моделі на основі диференціальних рівнянь SIR, SEIR, SEIR-U;
- моделі часових рядів ARIMA;
- моделі часових рядів на основі вейвлет-перетворень;
- моделі часових рядів FB Prophet;
- інші моделі машинного навчання на основі регресій, дерев рішень, нейронних мереж тощо.

Варто зазначити, що більшість аналітиків прогнозує накопичувальну криву, тобто сумарне за усі попередні дати значення [47, 49, 51, 54]. Такий прийом дозволяє штучно зменшити відносну похибку, адже вона ділиться на значно більші числа, ніж, якщо брати тільки щоденні прирости. Але прийняття рішень про карантинні обмеження робиться виключно за щоденними приростами, тобто скільки випадків є лише за одну добу, отже, більш цінним є прогнозування саме їх. Так само, через це, не варто використовувати інший популярний прийом, коли аналізують і прогнозують ковзне середнє показників, взяте з 7-денним вікном, що теж робить щоденні дані, суттєво відмінними від вихідних. Модель, яка будується

на необроблених щоденних даних є найбільш чутливою до їх змін, а отже, більш цінною для практичного застосування.

Розглянемо типи моделей більш розлого.

1.2.1. Моделі SIR, SEIR, SEIR-U на основі диференціальних рівнянь

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]), основана на системі диференціальних рівнянь, яка використовується для формування прогнозів і прийняття на їх основі управлінських рішень вченими з НАН України у складі Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, створеній Розпорядженням Президії НАН України від 3 квітня 2020 р. № 198 (базова установа – Інститут проблем математичних машин і систем НАН України, термін функціонування: 03.04.2020 р. – 23.02.2022 р., 24.02.2022 р. тимчасово призупинила свою діяльність, склад Робочої групи [61] (далі – «РГ по ковіду»). Основні положення моделі SIR описані у статті [1].

Основними при роботі з такою моделлю є такі припущення:

- про пропорційність швидкості захворюваності кількості інфікованих;
- про те, що хворі люди одужують через певний середній проміжок часу;
- про те, що люди, які одужали, не можуть захворіти повторно.

Дану модель можна представити за допомогою наступних рівнянь:

$$\frac{dS}{dt} = -\frac{R_0}{T_{inf}}\frac{1}{N} S \cdot I , \tag{1.1}$$

$$\frac{dI}{dt} = -\frac{R_0}{T_{inf}}\frac{1}{N} S \cdot I - \frac{I}{T_{inf}} , \tag{1.2}$$

$$\frac{dR}{dt} = \frac{I}{T_{inf}} , \tag{1.3}$$

де S – кількість здорових осіб, I – кількість інфікованих осіб, R – кількість тих осіб, що одужали або померли, N – загальна кількість населення; R_0 – коефіцієнт

репродукції, який визначає середню кількість заражень, що спричиняє одна хвора людина; T_{inf} – період часу, за який хвора людина залишається заразною.

Як було зазначено раніше, також використовувалась варіація SEIR, що бере до уваги також хворих в інкубаційному періоді. Наведемо додаткове рівняння, що враховує дану умову:

$$\frac{dE}{dt} = -\frac{R_0}{T_{inf}}\frac{1}{N} S \cdot I - \frac{E}{T_{inc}}, \quad (1.4)$$

де T_{inc} – тривалість інкубаційного періоду захворювання, діб.

Знаючи загальний вигляд та принцип застосування моделі, вчені з НАН України РГ по ковіду розробили модель з модифікованою структурою для України під назвою «SEIR_U», де були враховані різні типи пацієнтів, базуючись на суворості перебігу хвороби, а також за потребою в госпіталізації. Таким чином, модель стала більш гнучкою у визначенні прогнозів захворюваності, а також дає змогу опиратись на результати моделі при прийнятті рішень про карантинні заходи в короткостроковій перспективі.

Але, навіть з урахуванням наведеної вище інформації, головним недоліком цієї моделі є те, що вона – дуже чутлива до якості вхідних даних, яких вимагає чималу кількість, тому, тому, як видно з прогнозів РГ по ковіду у звіті «Прогноз РГ-58» на 22.12.2021-04.01.2022 рр., відносна похибка прогнозу на 2 тижні вперед склала 34% [24], у «Прогноз РГ-42» на 21.04.2021-03.05.2021 рр. – 33% [33], у «Прогноз РГ-30» на 01.12.2020-13.12.2020 рр. – 31% [44], хоча бували і малі значення.

1.2.2. Моделі часових рядів на основі АРПКС (ARIMA)

Вже тривалий час найбільш популярною моделлю для часових рядів є модель на основі авторегресії та проінтегрованого ковзного середнього (АРПКС) [62]. Ця модель враховує наступні впливи на значення ряду z_t від часу *t* [62]:

- вплив *l* попередніх значень ряду (складова під назвою «авторегресія»);

- вплив інших факторів, який враховується як регресія q-го порядку значень «білого шуму» з нульовим середнім і постійною диспресією a_t (складова «ковзне середнє»);
- нестаціонарність ряду, яка усувається взяттям різниці першого, другого чи вищих d порядків (∇^d), тобто коли модель АРКС будується не для самого значення z_t, а – для w_t («проінтегрованість»):

$$w_t = \nabla^d z_t. \tag{1.5}$$

Тоді модель АРПКС(l, q, d) для z_t записується у вигляді:

$$w_t = \phi_1^* w_{t-1} + \phi_2^* w_{t-2} + \cdots + \phi_l^* w_{t-l} + a_t - \Theta_1 a_{t-1} - \Theta_2 a_{t-2} - \cdots - \Theta_q a_{t-q}. \tag{1.6}$$

У Python-бібліотеці для автоматизації ідентифікації цієї бібліотеки вживається позначення ARIMA(p, q, d) [63], де p – це порядок авторегресії, q – порядок різниці, d – порядок моделі ковзного середнього. Є й варіант більшої кількості параметрів (з урахуванням сезонності): SARIMAX, для якого є функція автоматичного визначення параметрів (рис. 1.2).

```
Performing stepwise search to minimize aic
 ARIMA(4,1,4)(0,0,0)[0] intercept   : AIC=3084.743, Time=0.71 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=3243.452, Time=0.01 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=3245.316, Time=0.02 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=3245.056, Time=0.06 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=3244.282, Time=0.01 sec
 ARIMA(3,1,4)(0,0,0)[0] intercept   : AIC=3165.577, Time=0.46 sec
 ARIMA(4,1,3)(0,0,0)[0] intercept   : AIC=3136.683, Time=0.52 sec
 ARIMA(5,1,4)(0,0,0)[0] intercept   : AIC=3063.079, Time=0.62 sec
 ARIMA(5,1,3)(0,0,0)[0] intercept   : AIC=3091.233, Time=0.58 sec
 ARIMA(5,1,5)(0,0,0)[0] intercept   : AIC=3064.929, Time=0.77 sec
 ARIMA(4,1,5)(0,0,0)[0] intercept   : AIC=3131.373, Time=0.84 sec
 ARIMA(5,1,4)(0,0,0)[0]             : AIC=3081.960, Time=0.58 sec

Best model:  ARIMA(5,1,4)(0,0,0)[0] intercept
Total fit time: 5.207 seconds
                               SARIMAX Results
==========================================================================
Dep. Variable:                      y   No. Observations:                221
Model:               SARIMAX(5, 1, 4)   Log Likelihood               -1520.540
Date:                Sun, 08 Oct 2023   AIC                          3063.079
Time:                        21:28:07   BIC                          3100.409
Sample:                             0   HQIC                         3078.154
                                - 221
Covariance Type:                  opg
```

Рис. 1.2 Результат застосування Python-функції SARIMAX для визначення параметрів моделі ARIMA щоденної кількості нових хворих на коронавірус в Україні за даними 2020 року

Як видно на рис. 1.2, оптимальною моделюю є АРПКС(5,1,4). Сезонні складові не виявлені – їх параметри (0,0,0).

Як було зазначено вище, ARIMA вимагає ряд значень з фіксованим кроком, де відсутні пропущені значення. Усі пропущені значення треба інтерполювати. Є певні складнощі в урахуванні сезонності, аномальних дат та ін.

1.2.3. Моделі часових рядів на основі вейвлет-перетворень

Вейвлет-перетворення дають змогу виконати декомпозицію часового ряду, а саме – отримати розподіл часового ряду на окремі складові різної частоти, з

наступним аналізом локальних змін цих складових. Під час такого аналізу визначаються коефіцієнти, які враховують зростання або спадання часового ряду [64].

В основі таких перетворень лежить застосування вейвлет-функцій, які повинні задовольняти наступній умові [64]:

$$\int_{-\infty}^{\infty} \frac{\hat{\psi}(\omega)}{|\omega|} d\omega < \infty, \tag{1.7}$$

де $\hat{\psi}(\omega)$ описує перетворення Фур'є часового ряду $\psi(\omega)$.

Загальний вигляд вейвлет-перетворення для часового ряду $\psi(t)$, який відповідає умові (1.7) [64]:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right); a, b \in R; a \neq 0, \tag{1.8}$$

де a – параметр зростання функції, b – зсув, який визначає початок того чи іншого вейвлету.

Дана функція в подальшому застосовується для дискретної декомпозиції часового ряду, після чого визначається апроксимаційний коефіцієнт $\alpha_{j,k}$ для кожного вейвлету, за допомогою такої згортки [64]:

$$\alpha_{j,k} = \int_{-\infty}^{\infty} f(t)\psi_{j,k} dt. \tag{1.9}$$

Отримані значення коефіцієнта використовуються в основній моделі:

$$f(t) = \sum_{k=-\infty}^{2^{J-k}-1} \alpha_{J,k}\varphi_{j,k}(t) + \sum_{j=1}^{J} \sum_{k=-\infty}^{2^{J-k}-1} \alpha_{J,k}\psi_{j,k}(t) \tag{1.10}$$

Вейвлет-перетворення рідко використовуються самостійно. Частіше їх застосовують для опису періодичних складових, разом з іншими моделями, наприклад Wavelet-ARIMA [64], а тому їм властиві усі ті ж недоліки, як і для моделей ARIMA, однак, їх можна використовувати й з іншими моделями.

1.2.4. Моделі часових рядів на основі Facebook Prophet

Основні аспекти моделі FB Prophet вперше описані у 2017 р. у статті [65] та у документації відповідної Python-бібліотеки [66]:

1. Автоматичне визначення сезонності та трендів: Prophet визначає задані параметри сезонних складових $s(t)$ та тренду $g(t)$ у вхідних часових рядах автоматично. Є можливість автоматичного виявлення щорічної, щотижневої і денної сезонності, а також визначення тренду у вигляді шматково-лінійної апроксимації між так званими точками зміни «changepoint». Це дозволяє моделі адаптуватися до доволі складних нестаціонарних часових рядів.

2. Додаткові гармоніки: якщо часовий ряд має сезонні залежності, які не можуть бути адекватно описані типовими щорічною, щотижневою або денною сезонностями, він може використовувати додаткові гармоніки для кращої апроксимації сезонності на додаток чи замість цих трьох типових видів сезонностей. Однак, для додаткових видів сезонності слід вручну задавати відповідні гіперпараметри, про які буде нижче.

3. Урахування святкових днів та важливих подій (або дат аномалій): Prophet надає можливість вручну задати і врахувати вплив важливих подій та святкових днів чи інших аномалій, які можуть вплинути на часовий ряд. Модель автоматично враховує ці точки у прогнозах. Важливо, що можна регулювати «вікно» впливу, тобто за скільки часових кроків цей вплив починається і коли завершується. Крім того, можна задавати ступінь цього впливу (ступінь регуляризації).

4. Оброблення пропущених даних і «викидів» (сильних відхилень від даних спостережень): Prophet розроблений з урахуванням можливої відсутності даних у певні моменти часу і «викидів». Він може адекватно працювати з часовими рядами, де є пропуски в даних або незвичайно великі чи малі значення. Викиди, які сильно відрізняються від основного тренду, ігноруються – більш правильно їх враховувати як аномальні дані і тоді для їх урахування є більш ефективне рішення.

5. Байєсівська апроксимація: Prophet використовує байєсівську апроксимацію для оцінювання параметрів моделі та зони їхньої невизначеності.

Це дозволяє отримувати діапазони надійних прогнозів з урахуванням можливої зони невизначеності із заданим рівнем.

6. Гнучкість параметрів: Prophet дозволяє користувачам вручну налаштовувати деякі параметри моделі, такі як режим урахування складових (мультиплікативний чи адитивний) та ступінь регуляризації значень в аномальні дати, ступінь регуляризації кожного виду сезонності та період і порядок ряду Фур'є для опису кожної сезонності окремо тощо.

7. Є можливість формування логістичної моделі (із насиченням) та урахування впливу додаткових ознак, тобто – побудови множинної (багатофакторної) моделі.

Математично найпростіша модель Prophet для моделювання та прогнозування значень ряду $y(t)$, в залежності від часу *t*, записується таким чином [65]:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \tag{1.11}$$

де $g(t)$ – тренд ряду, $s(t)$ – сезонна складова (їх може бути й декілька одночасно), $h(t)$ – складова, яка відображає вплив свят чи інших аномалій, які відбуваються за потенційно нерегулярними графіками протягом одного або кількох днів, ϵ_t – похибка («шум») з нульовим середнім, розподілена за нормальним законом.

Тренд є шматково-лінійною апроксимацією, яка описується таким чином [65]:

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma), \tag{1.12}$$

де k – темп зростання, δ – величина корегування темпу, m – параметр зсуву, а γ застосовується для перетворення функції у неперервну.

При моделюванні тренду за допомогою моделі Prophet потрібно явно вказати значення точок зміни (англ. «changepoints»), в яких темп зростання може змінюватись. Припустимо, що є S точок зміни в моменти часу s_j, $j = 1, \ldots, S$. Визначимо вектор корегувань темпу $\delta \in R^s$, де δ_j – зміна швидкості, що відбувається в момент часу s_j. У такому разі, темп зростання k в будь-який момент

часу t, враховуючи всі зміни до цього моменту, можна представити таким чином: $k + \sum_{j:t>s_j} \delta_j$. А тоді темп зростання в момент часу *t* може бути визначений таким чином [65]:

$$k + a(t)^T\delta, \quad a(t) = \begin{cases} 1, \\ 0 \end{cases} if\ t \geq s_j. \tag{1.13}$$

Коли значення k визначене, також повинен бути визначений параметр зсуву m для певного значення точки зміни *j* [28]:

$$\gamma_j = \left(s_j - m - \sum_{l<j} \gamma_l\right)\left(1 - \frac{k+\sum_{l<j}\delta_l}{k+\sum_{l\leq j}\delta_l}\right). \tag{1.14}$$

Сезонна складова $s(t)$ апроксимується рядом Фур'є [28 - стаття 2018 р., наша стаття у Віснику № 6, 2020]:

$$s(t) = \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2\pi\, nt}{P}\right) + b_n \sin\left(\frac{2\pi\, nt}{P}\right)\right), \tag{1.15}$$

де *P* – тривалість періоду, a_n та b_n – амплітуди гармонік.

Для визначення складової $s(t)$ треба, передусім, задати порядок *N* та тривалість періоду *P* сезонності ряду Фур'є, решта параметрів формули (1.15) ідентифікується автоматично.

Адитивна модель (1.11) ще має ряд варіацій:

- логістична модель (коли тренд прямує в насичення);

- модель множинної регресії з урахуванням додаткових рядів ознак (регресорів);

- мультиплікативна модель:

$$y(t) = g(t)\, s(t)\, h(t)\, \epsilon_t. \tag{1.16}$$

1.2.5. Інші моделі машинного навчання

У статті [22] продемонстровано можливості розв'язання задачі передбачення з використанням моделей машинного навчання: регресори на основі дерев рішень чи їх ансамблів (DecisionTreeRegressor, ExtraTreesRegressor, RandomForestRegressor, AdaBoost Regressor, XGBoost, LGBM), регресійні моделі

(LinearRegression, Ridge), методу опорних векторів та ансамблі моделей на основі BaggingRegressor та VotingRegressor.

Популярними для прогнозування часових рядів, також, є рекурентні нейромережеві моделі (RNN), зокрема їх підвид з довготерміновою короткочасовою пам’яттю (LSTM), але, особливістю нейромережевих моделей є вимога щодо значної кількості даних, інакше узагальнюючі можливості моделей не будуть виражені належним чином. Тому, якщо здійснюється моделювання даних тільки за 2020 р., а саме за той період отримані основні теоретичні результати цієї роботи, тоді застосування нейромережевих моделей не є доцільним.

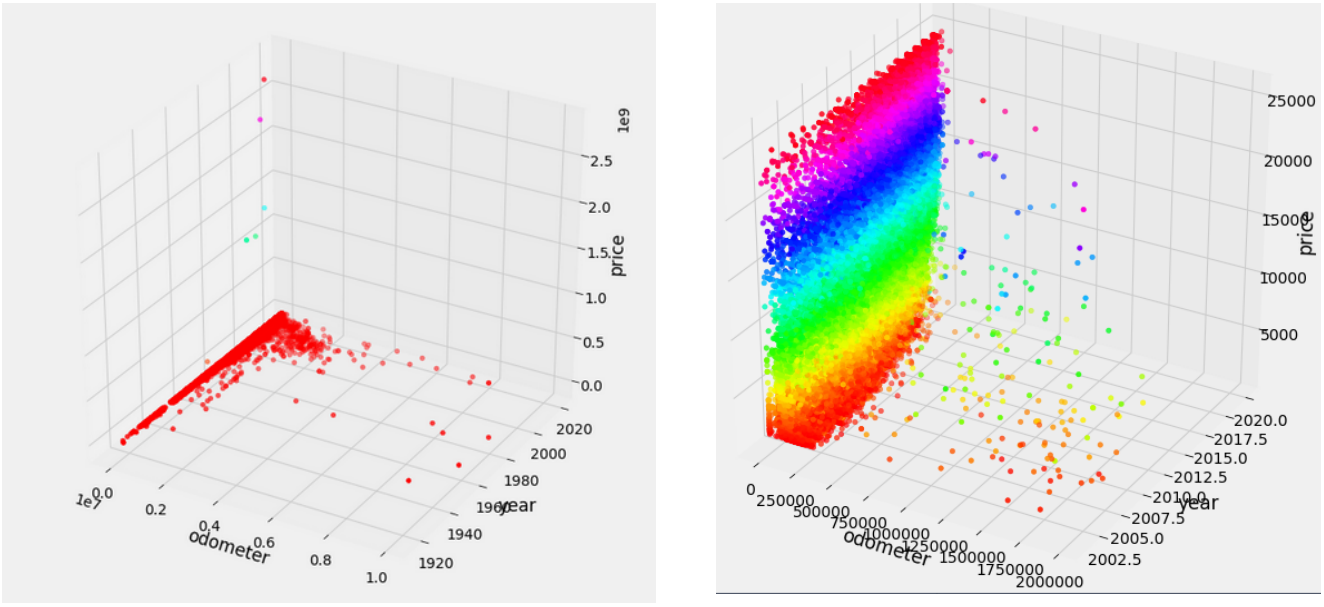
У таких задачах важливо правильно здійснити передоброблення даних:

- стаціонаризація, тобто перехід до першої, другої чи третьої різниці, щоб ряд став стаціонарним;

- стандартизація, коли від усіх значень віднімається середнє і вони діляться на середньоквадратичне відхилення ряду (на Python цю операцію виконує команда preprocessing.StandardScaler бібліотеки sklearn);

- фільтрування аномалій – для цього використовуються різні прийоми.

У статті [22] продемонстрована ефективність попереднього фільтрування аномалій для розв’язання задачі передбачення вартості вживаних авто на датасеті з 525 тисяч авто у США методами машинного навчання (рис. 1.2).



а) б)

Рис. 1.3 Фільтрування аномальних значень за рядом параметрів: а) до фільтрування; б) після фільтрування аномалій

Фільтрування по перцентилях, використане для отримання рис. 1.3б, не завжди є ефективним. Іноді важливо врахувати аномалії більш гнучко. Ще однією проблемою є ситуація з коронавірусом, коли у 2020 р. аномаліями були значення, які у 2021 р. були вже звичайними, а у 2022 р. – взагалі, «шумом» між «хвилями». Крім того, цікаво мати можливість дослідити ступінь впливу аномалій на інші значення, а не просто відкинути їх.

1.2.6. Метрика моделей машинного навчання

Точність моделей пропонується оцінювати за трьома критеріями (метриками):

- RMSE або rmse – «root-mean-square-error» або «середньоквадратична похибка»;

- R^2 або r2_score чи R-squared, чи R2 – коефіцієнт детермінації;

- відносна похибка δ або метрика WAPE (англ. «Weighted Absolute Percentage Error») – зважена абсолютна похибка.

Пояснимо яким чином обчислюються ці методи оцінювання точності. RMSE обчислюється у вигляді квадратного кореня із середнього значення квадратів поелементної різниці результатів натренованої моделі, та даних, що містяться у тренувальному наборі даних [48]:

$$RMSE = \sqrt{\frac{\sum_{t=1} T(\hat{y}_t - y_t)^2}{T}}, \tag{1.17}$$

де T — загальна кількість результатів (рядків даних); $\hat{y}$ — результат обчислення моделі; y — відповідна величина з тренувального набору даних (так званий «таргет», з англ. – «ціль»).

Коефіцієнт детермінації використовується у статистичних моделях та у регресійних задачах машинного навчання як показник залежності варіації залежної змінної від варіації незалежних змінних, що вказує наскільки отримані спостереження підтверджують модель [53]:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{1.18}$$

де y — істинне значення з тренувального набору даних (таргет); $\hat{y}$— передбачене значення відповідного набору ознак; $\bar{y}$ — середнє арифметичне істинних значень з тренувального масиву даних.

Відносна похибка обчислюється за відомою формулою [51]:

$$\delta = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i|} \cdot 100, \%. \tag{1.19}$$

В машинному навчанні їй відповідає метрика WAPE з однаковими вагами 1, яку слід помножити на кількість даних, оскільки у ній відносна похибка, по суті, ділиться на цю кількість [51]:

$$WAPE = \frac{1}{T}\sum_{t=1}^{T}\frac{|y_t - \hat{y}_t|}{y_t}. \quad (1.20)$$

Як критерій можуть використовуватись ще, наприклад: MAE — середня абсолютна похибка (з англ. «Mean Average Error») чи MSE — середньоквадратична похибка (з англ. «Mean Square Error»).

$$MAE = \frac{\sum_{i=1}^{T}|y_i - x_i|}{T} = \frac{\sum_{i=1}^{T}|e_i|}{T}. \quad (1.21)$$

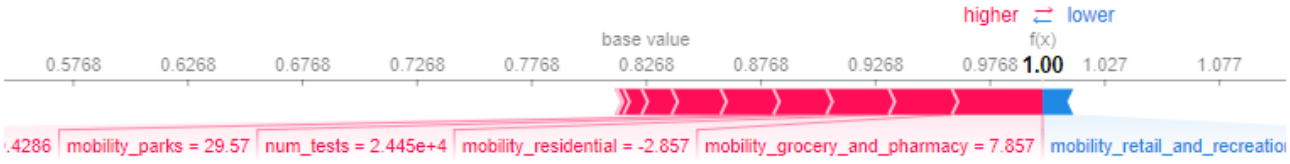
1.3 Аналіз впливу різних факторів і проблем на кількість хворих на коронавірус в Україні

Під час пандемії на коронавірус силами багатьох вчених та дослідників з багатьох країн вдалось зібрати і постійно оновлювати датасет, розміщений на базі Оксфордської лабораторії («Oxford COVID-19 government response tracker» — Оксфордський трекер коронавірусної діяльності урядів країн світу), в якому зберігається інформація про державні заходи регулювання епідеміологічної ситуації, охарактеризовані у вигляді Stringency Index (SI) [67]. Також використовувались дані про Google-тренди пересувань жителів країн світу, у т.ч. України [68], внаслідок відслідковування мобільних телефонів, які дозволили у відносних одиницях співставити активність перебування людей (з усередненням по усій країні) в зонах відпочинку, удома, на роботі, в парках, аптеках, місцях торгівлі, транспортних станціях різного виду тощо. У 2020 р. був проведений аналіз впливу цих трендів на кількість нових хворих на коронавірус в Україні (рис. 1.4).

Weight[?]	Feature
+6899.726	<BIAS>
+3979.037	mobility_retail_and_recreation
+3875.869	mobility_transit_stations
-2948.260	mobility_workplaces
-4416.882	mobility_grocery_and_pharmacy
-5655.193	mobility_parks
-5686.202	mobility_residential

Рис. 1.4. Аналіз впливу Google-трендів щодо пересувань жителів України протягом 16.05-25.10.2020 р. [69]

Як видно на рис. 1.4, на зростання кількості хворих «позитивно» впливає перебування людей у зонах торгівлі, відпочинку та транспортних станціях, а на зменшення цієї кількості – перебування удома, на роботі, в аптеках та парках. Подібні висновки було одними з обгрунтувань введення так званих «локдаунів», особливо під час стрімкого наростання «хвиль». Однак, проведені дослідження показали, що ці висновки сильно відрізняються, в залежності від використаного методу визначення впливу та від періоду даних, які беруться до уваги, у т.ч. через особливості методики Google для узагальнення результатів спостережень, що ускладнює їх співставлення. На рис. 1.5 наведено діаграми важливості та напрямку впливу ознак, отримані з використанням Python-бібліотеки SHAP (метод аналізу на основі теорії ігор) [70].



а)

б)

в)

Рис. 1.5. Діаграми важливості та напрямку впливу ознак на щоденну кількість хворих в Україні на коронавірус, отримані з використанням Python-бібліотеки SHAP для заданих дат: а) 20.10.2020 р.; б) 25.10.2020 р.; в) 26.10.2020 р.

Як видно з рис. 1.5, на збільшення значень кількості хворих то впливає перебування у місцях торгівлі чи відпочинку (рис. 1.5а,б), то – ще й перебування на роботі (рис. 1.5б) або наявність дощу чи перебування в парках (рис. 1.5в), а на збільшення – перебування в парках (1.5а,б), доволі висока, якдля жовтня, середня температура (рис. 1.5б) та ін., отже, є певна невизначеність. Стабільно позитивно впливає тільки перебування у місцях торгівлі чи відпочинку.

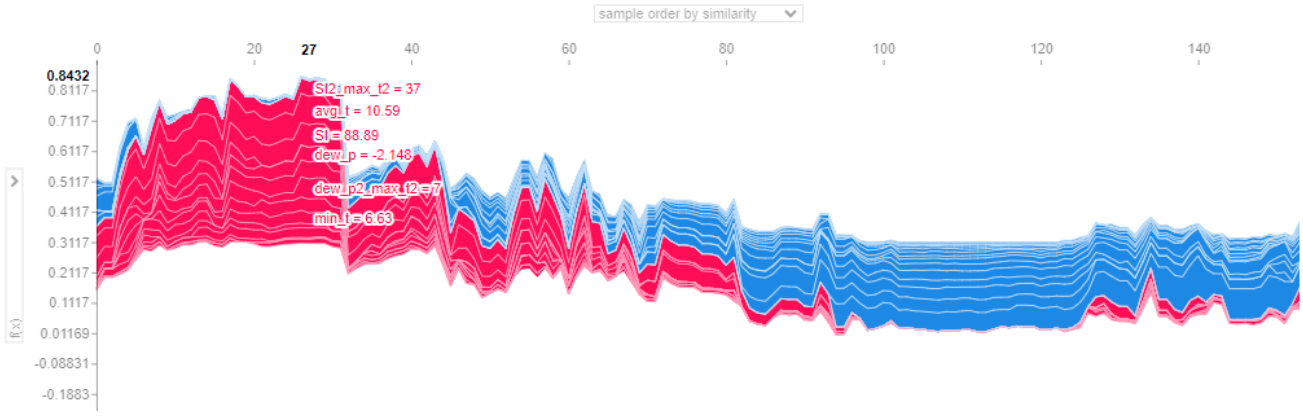
Аналогічне дослідження проводилось і для інших ознак згаданого вище датасету: індекс урядового трекера stringency_index (SI) та усереднені для країни метеодані (середня, мінімальна та максимальна температура в країні, наявність опадів, точка роси, відносна вологість). Для отримання більшої кількості закономірностей був синтезований ряд похідних ознак [70] (рис. 1.6).

```
def fe_creation(df):
    df['SI2'] = df['SI']//5
    df['avg_t2'] = df['avg_t']//2
    df['min_t2'] = df['min_t']//2
    df['max_t2'] = df['max_t']//2
    df['rain2'] = df['rain']//1
    df['dew_p2'] = df['dew_p']//2
    df['r_hum2'] = df['r_hum']//10
    for i in ['SI2', 'rain2', 'dew_p2', 'r_hum2']:
        for j in ['max_t2']:
            df[i + "_" + j] = df[i].astype('str') + "_" + df[j].astype('str')
    return df
```

Рис. 1.6 Синтез вторинних ознак на основі урядового трекера stringency_index та метеоданих щодо України

На рис. 1.7 наведені SHAP-діграми цих ознак [70].

а)



б)

Рис. 1.7 SHAP-діаграми первинних і вторинних ознак урядового трекера stringency_index та метеоданих щодо України за даними 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.

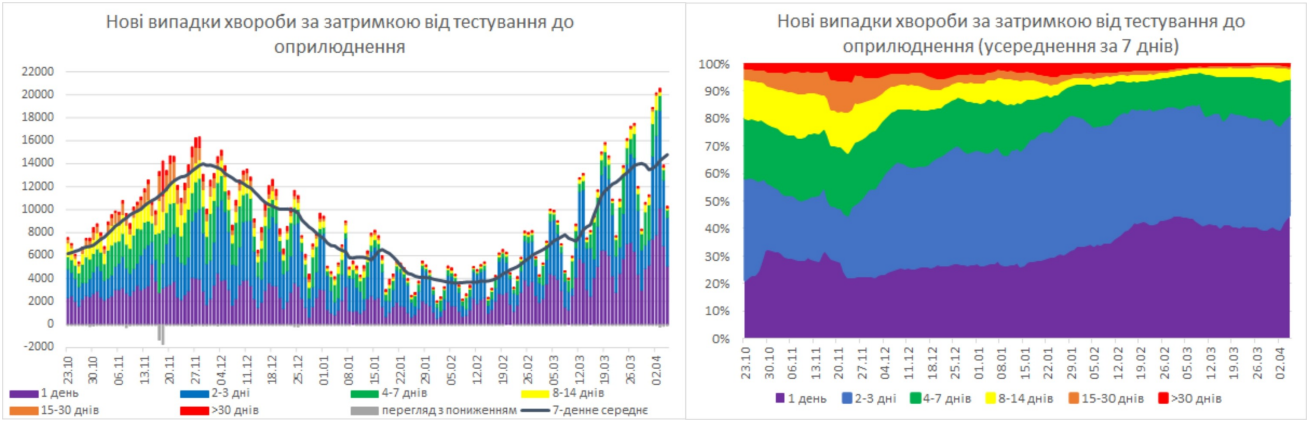
Аналіз показав, що ці ознаки впливають нерівномірно. Крім того, проблемою є запізнення на 4-7 днів оновлення цих даних, що суттєво ускладнює їх використання під час короткострокового прогнозування. Отже, до їх врахування слід ставитись з обережністю, наприклад враховувати не як додатковий регресор для побудови множинної регресії, а на їх основі згенерувати вторинні ознаки і їх ураховувати більш ефективно.

Хоча, щодо специфіки в Україні, доцільно спробувати дослідити чи матиме вплив ще й кількість тестів. Показники захворюваності суттєво збільшились, коли налагодили систему тестування, але багато аналітиків та медиків зазначали, що раніше випадків захворюваності теж було багато, просто вони не фіксувались, в силу специфіки хвороби. У багатьох інфікування протікало безсимптомно, а хворобливий стан був обумовлений наслідками захворювання, однак, на тому етапі вже важко було точно діагностувати причини такого стану – це коронавірус чи інша хвороба. Такі тести є, але їх точність – невисока.

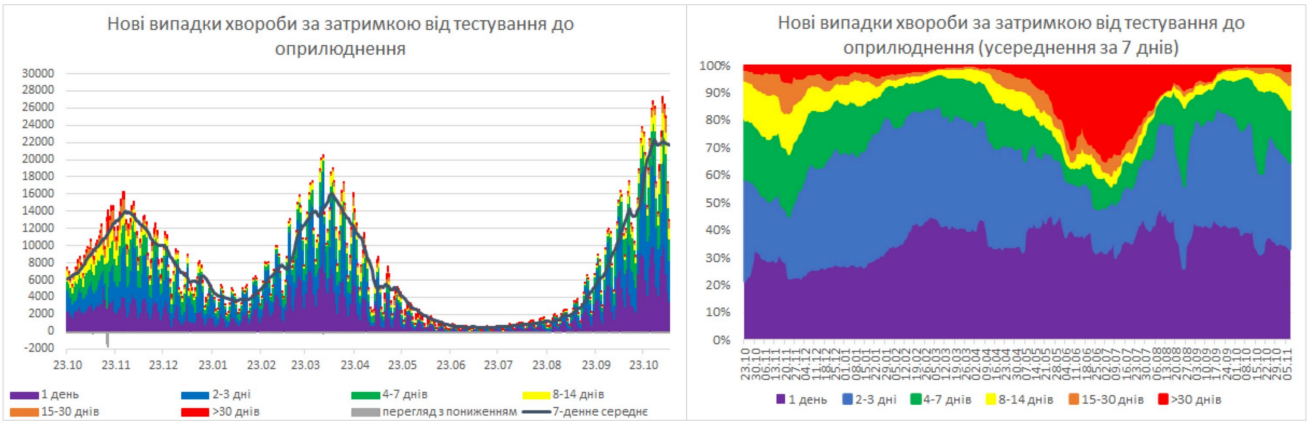
Важливо усвідомлювати високий ступінь невизначеності даних про кількість хворих в Україні, особливо у 2020 році. РГ по ковіду досліджувала

ступінь затримок між датою проведення тестування і датою оприлюднення результату та його потраплянням в офіційну статистику.

На рис. 1.8 наведено такі графіки під час двох різних хвиль наростання кількості хворих, коли система тестування виходила в насичення.



а)



б)

Рис. 1.8 Еволюція в часі затримок оприлюднення кількості нових випадків захворювання на коронавірус в Україні: а) 7–20 квітня 2021 року («Прогноз РГ-41») [34, рис. 36] б) 10–23 листопада 2021 року («Прогноз РГ-55») [27, рис. 37]

Як видно з рис. 1.8б (на 1.8а подано цей же рисунок, але за менший період і краще видно більш ранні закономірності), в окремі періоди часу мали місце дуже суттєві затримки (значний відсоток затримок був на 30 і більше днів), що суттєво

псувало статистику та укладнювало моделювання і прогнозування даних. Саме цей аспект дуже суттєво ускладнює роботу моделей типу SEIR_U, але дає перевагу статистичним моделям типу FB Prophet, в силу певної монотонності графіку затримок тестів в часі, особливо у деякі періоди часу.

1.4 Висновки до розділу та постановка задач дослідження

У першому розділі охарактеризовано основні проблеми прогнозування захворюваності на коронавірус:

- мутації хвороби (різні штами) та вплив інших хвороб;
- різна динаміка кількості тестів;
- недотримання населенням карантинних умов;
- зміни у режимі роботи лабораторій протягом тижня і під час та після святкових днів,

що, в цілому, призводить до суттєвої волатильності часового ряду та вносить складнощі в його прогнозування.

Здійснено огляд відомих моделей для прогнозування поширення коронавірусу:

- на основі диференціальних рівнянь SIR, SEIR, SEIR-U;
- моделі часових рядів ARIMA;
- моделі часових рядів на основі вейвлет-перетворень;
- моделі часових рядів FB Prophet;
- інші моделі машинного навчання на основі дерев рішень та ін.

Охарактеризовано їх переваги і недоліки щодо перспектив використання для розв’язання поставлених задач.

Наведено метрики (критерії точності) моделей, які є поширеними в задачах машинного навчання.

Проаналізовано вплив різних факторів та можливості прогнозування поширення ковіду, зокрема Google-трендів щодо пересувань жителів України,

метеоданих, даних Оксфордського трекера коронавірусної діяльності урядів країн світу stringency_index (SI) та ознак, отриманих на їх основі. Досліджено які з них викликають зростання, а які – зменшення кількості нових хворих. Однак, аналіз показав, що ці ознаки впливають нерівномірно. Отже, до їх урахування слід ставитись з обережністю, наприклад, не враховувати як додатковий регресор для побудови множинної регресії, а натомість, на їх основі згенерувати вторинні ознаки і їх враховувати більш ефективно.

Продемонстровано, що в Україні під час проходження піків хвиль по кількості хворих на ковід, з'являються суттєві (навіть більші за 30 діб) затримки в оприлюденні результатів тестів – зазначено, що саме цей аспект дуже суттєво ускладнює роботу моделей типу SEIR_U на основі диференціальних рівнянь, але дає перевагу статистичним моделям типу Facebook Prophet, в силу певної монотонності графіку затримок тестів у часі, особливо у деякі періоди часу.

Отже, для досягнення поставленої у дисертації мети щодо підвищення точності прогнозування кількості хворих на коронавірус у короткостроковій перспективі за допомогою методів машинного навчання необхідно розв'язати такі задачі:

1. Здійснити аналіз відомих методів та моделей прогнозування часових рядів, які можуть бути використані для створення інформаційної технології прогнозування захворюваності на коронавірус та проаналізувати основні проблеми, які виникають під час такого прогнозування.

2. Розробити методи ідентифікації структури та параметрів моделі для прогнозування кількості нових випадків захворювання на коронавірус, які дозволять підвищити точність прогнозування та допоможуть виявляти нові закономірності щодо даних.

3. Розробити метод картографічної візуалізації прогнозів динаміки приросту кількості хворих на коронавірус, що дозволить більш точно виявити закономірності розповсюдження захворювання на карті обраного регіону.

4. Розробити структуру запропонованої інформаційної технології та її

компонентів.

5. Створити програмно-інформаційне забезпечення для автоматизації запропонованих методів і складових інформаційної технології та випробувати його на реальних даних.

Перша задача вже була частково розв’язана у цьому розділі, але варто ще провести розвідувальний аналіз даних та спробувати ідентифікувати та порівняти різні види моделей машинного навчання з типовою структурою та вибрати серед них оптимальний вид для подальшого удосконалення точності його прогнозів.

РОЗДІЛ 2
РОЗРОБЛЕННЯ МЕТОДІВ ТА ТЕОРЕТИЧНИХ ОСНОВ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС

2.1 Розвідувальний аналіз даних

Як відомо, першим етапом застосування моделей машинного навчання є розвідувальний аналіз (EDA) та відбір ознак (FE) [22], [72], тому проведемо такий аналіз даних.

Перші теоретичні результати даного дослідження були отримані у 2020 р. Більш-менш стабільні дані щодо кількості хворих в Україні почали з'являтись з 1 квітня 2020 р. Проаналізуємо закономірності, які мали місце у той час протягом 01.04.2020-06.12.2020 рр. (рис. 2.1), тобто до момента опублікування перших результатів прогнозування у статтях [16], [17], [45].

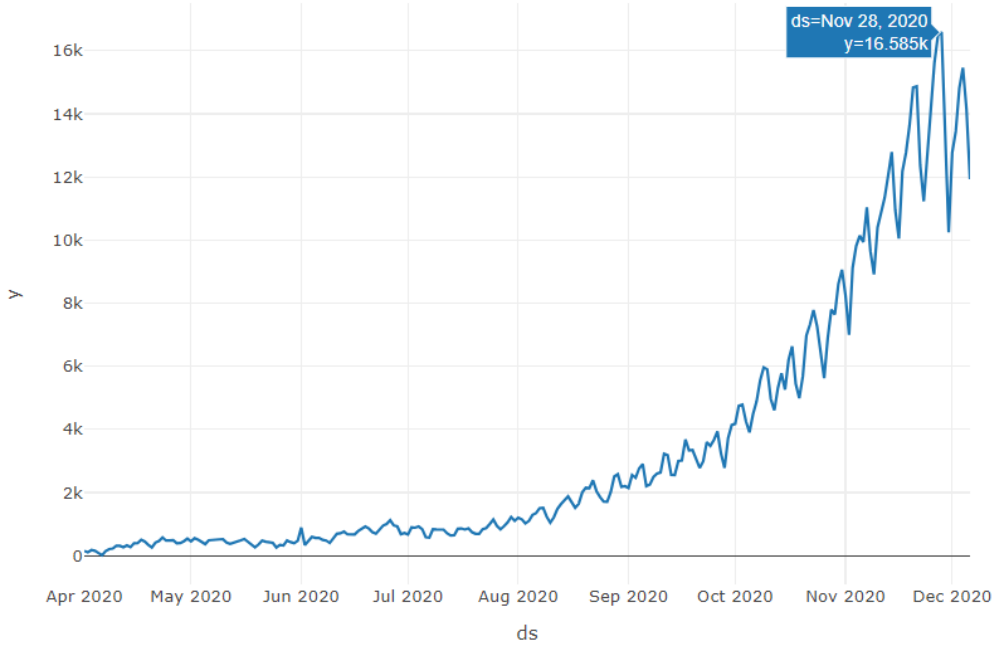


Рис. 2.1 Щоденна кількість нових хворих на коронавірус в Україні за період 01.04.2020-06.12.2020 [72]

2.1.1. Аналіз стаціонарності ряду.

Аналіз даних за критерієм Діка-Фуллера показав, що тільки друга різниця ряду є стаціонарною, а сам ряд і його перша різниця є нестаціонарними [72] (рис. 2.2).

```
# Stationarity check
check_stationarity(df['y'])
```

```
ADF Statistic: -2.762176
p-value: 0.063883
Critical Values:
        1%: -3.459
        5%: -2.874
        10%: -2.573
Non-stationary
```

```
# Stationarity check of the first difference of time series
check_stationarity(df['y'].diff().dropna())
```

```
ADF Statistic: -1.969920
p-value: 0.299916
Critical Values:
        1%: -3.459
        5%: -2.874
        10%: -2.573
Non-stationary
```

```
# Stationarity check of the second difference of time series
check_stationarity(df['y'].diff().diff().dropna())
```

```
ADF Statistic: -4.813857
p-value: 0.000051
Critical Values:
        1%: -3.459
        5%: -2.874
        10%: -2.573
Stationary
```

Рис. 2.2 Аналіз ряду з рис. 2.1 на стаціонарність [72]

Класичні моделі часових рядів вміють враховувати таку специфіку, але у разі застосування моделей машинного навчання – дерев рішень, нейронних мереж тощо краще їх застосовувати саме до другої різниці ряду.

2.1.2. Аналіз сезонності ряду.

Здійснимо декомпозицію ряду і виділимо в ньому тренд («Trend»), сезонну складову («Seasonal») із заданим періодом та залишки («Resid») (рис. 2.3).

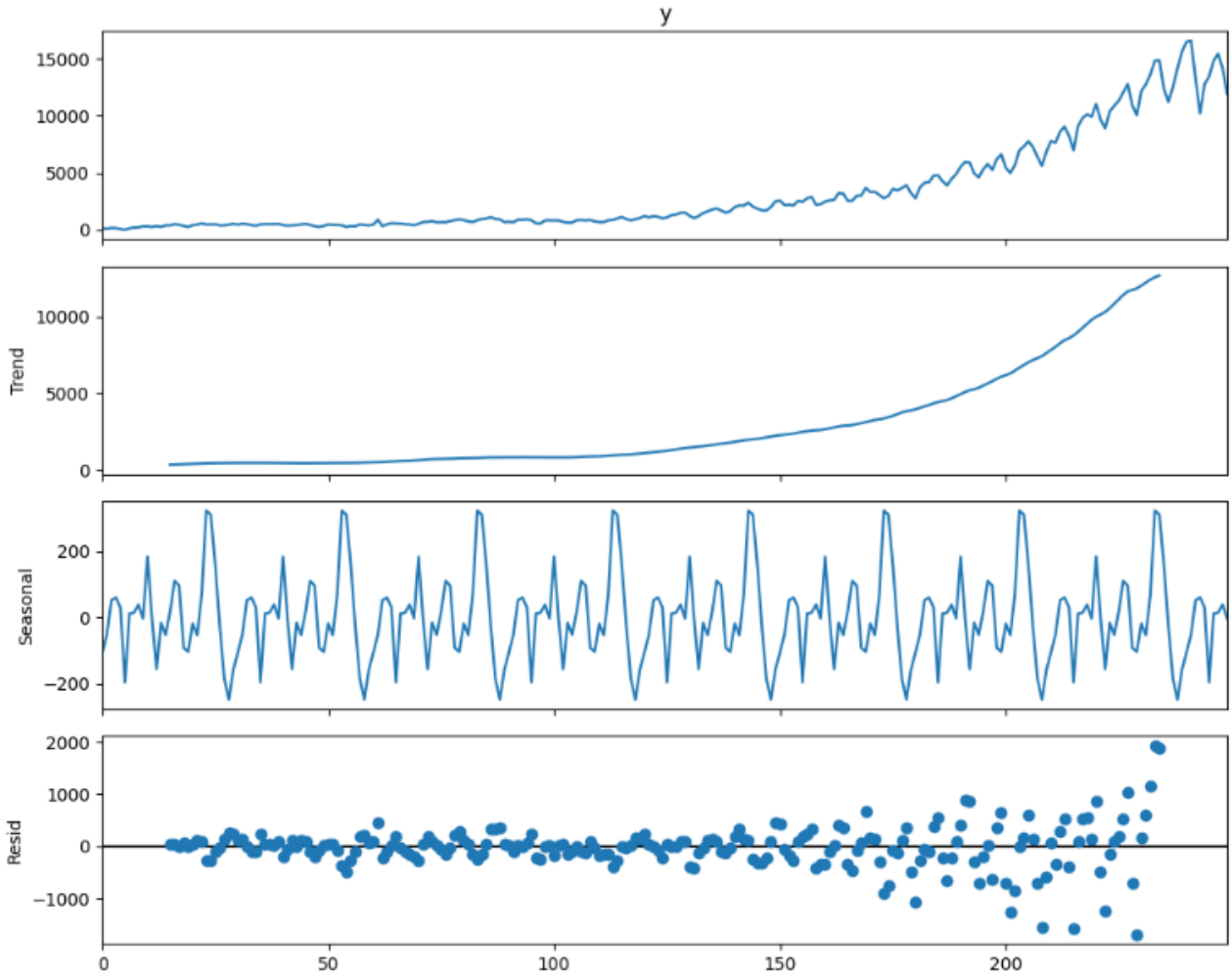


Рисунок 2.3 – Результат декомпозиції ряду кількість хворих на ковід в Україні у 2020 р. на складові для сезонності з періодом 7 діб [72]

При цьому цікаво аналізувати, яку саме частку має амплітуда сезонної складової «Seasonal» від амплітуди (по суті, від максимального значення, оскільки кількість хворих не є від'ємною) основного ряду.

На рис. 2.4 наведено графік зміни частки сезонної складової відносно періоду цієї складової, згладжений з 7-денним вікном [50].

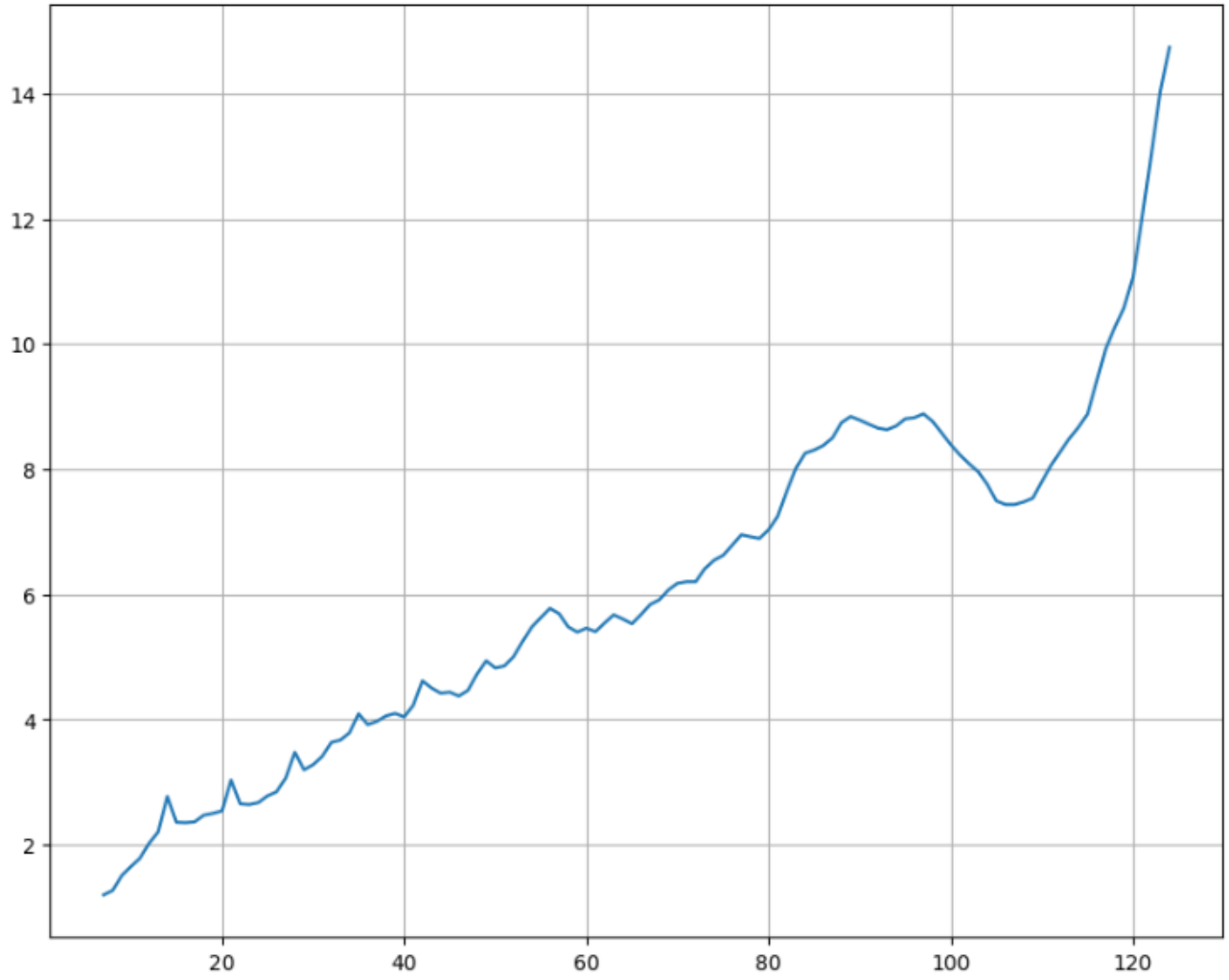


Рисунок 2.4 – Частка сезонної складової, залежно від періоду *P* у добах, до амплітуди основного ряду кількості нових хворих на коронавірус в Україні у 2020 р., зі згладжуванням у 7 днів [71]

Як видно на рис. 2.4:

1) частка сезонних складових очікувано монотонно зростає зі збільшенням періоду;

2) перший сильний стрибок має місце у перші 7 діб, що доволі очевидно, бо це – цикл роботи лабораторій, які проводять тестування;

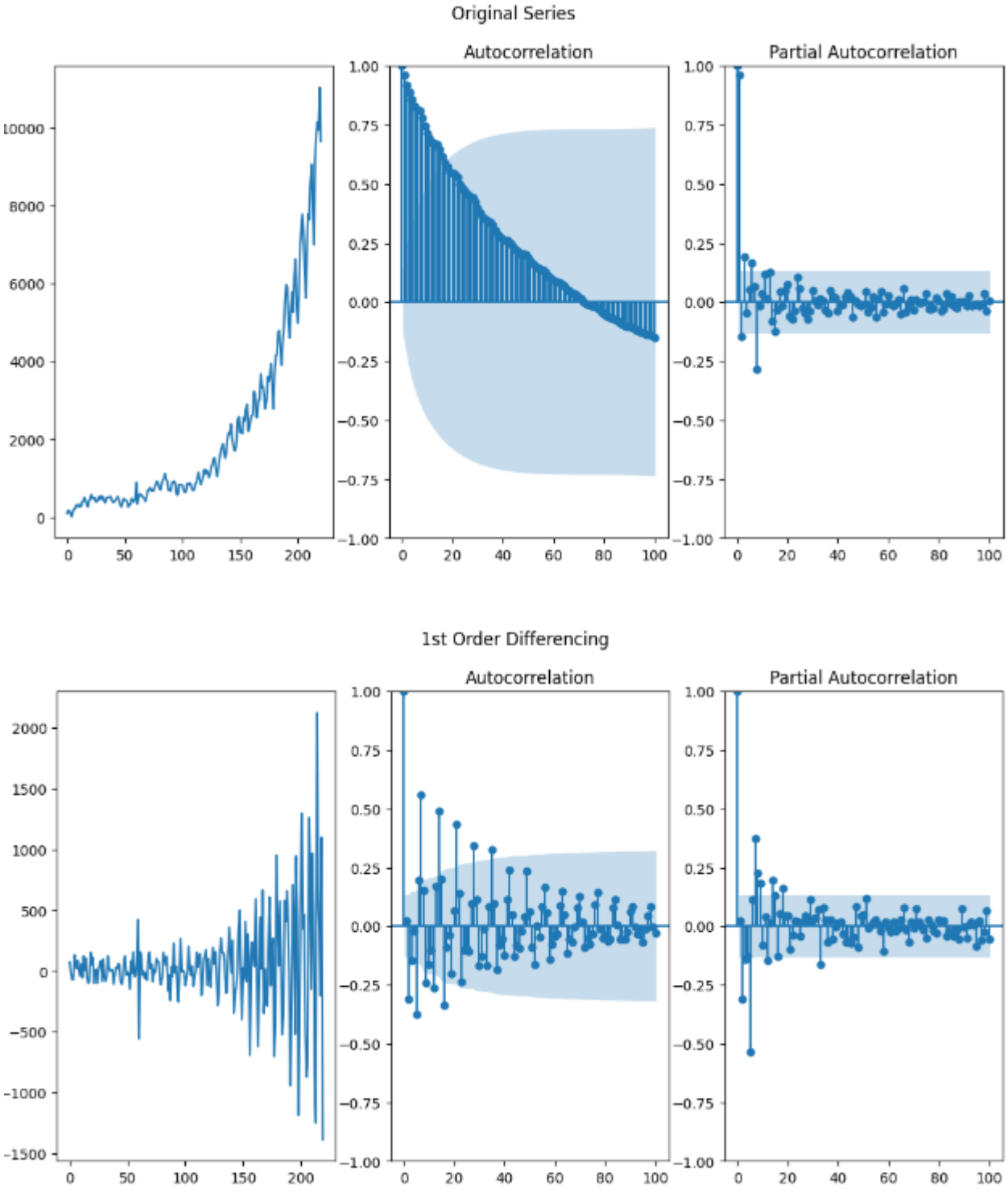
3) в подальшому теж мають місце стрибки, але – не чітко виражені і це питання потребує окремого дослідження;

4) суттєва нелінійність кривої свідчить про наявність багатьох видів сезонності одночасно.

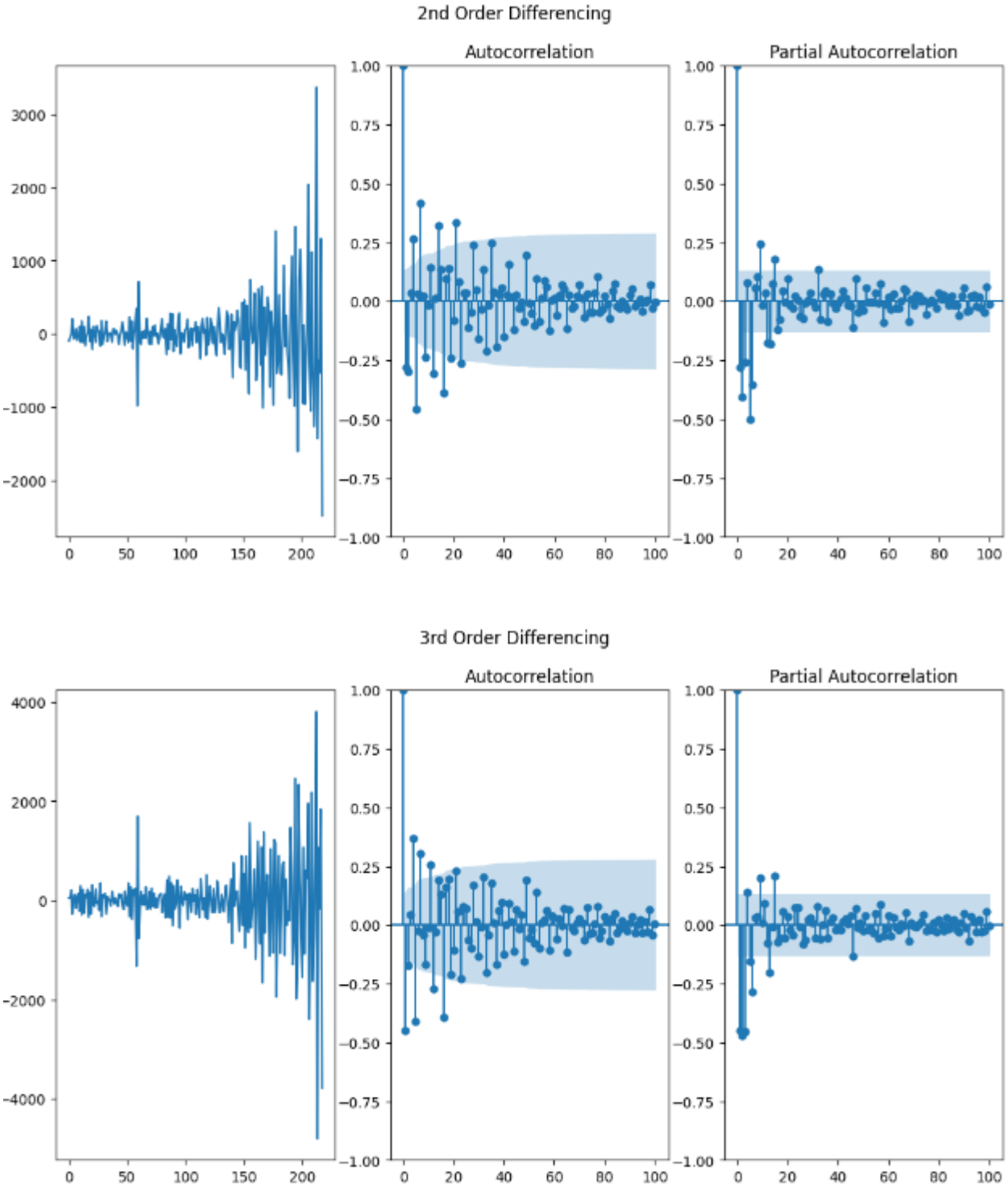
2.1.3. Побудова автокореляційної та часткової автокореляційної функцій.

Як відомо, по автокореляційній (АКФ) та частковій автокореляційній функціях (ЧАКФ) можна, у першому наближенні, оцінити порядок p авторегресійної AR(p) моделі та порядок q моделі ковзного середнього КС(q), якщо вони доволі швидко спадають і потім поводять себе плавно, не виходячи з «коридору» статистичної незначущості.

На рис. 2.5 наведені АКФ та ЧАКФ для ряду значень за 2020 р. [72]


Original Series
Autocorrelation
Partial Autocorrelation
1st Order Differencing
Autocorrelation
Partial Autocorrelation

a)



б)

Рис. 2.5 АКФ та ЧАКФ ряду кількості хворих на коронавірус в Україні за даними 01.04-06.12.2020 р. а) у першому наближенні; б) у другому та третьому наближенні

Як видно з рис. 2.5, значення АКФ та ЧАКФ спадають різко, але потім довго ще не залишаються повністю у зоні статистичної незначущості. Найбільш швидко спадає ЧАКФ для першої різниці, але і там є певні коливання, тобто визначення параметрів АРПКС може бути проблематичним, а сама модель може й не бути достатньо адекватною.

2.1.4. Прогнозування даних за різними моделями.

Однією зі складових розвідувального аналізу є пробна побудова моделей з мінімальною (базовою) ідентифікацією структури та параметрів. Побудуємо такі моделі, використавши досвід статті [22] та додавши моделі часових рядів ARIMA і FB Prophet з різними періодичними складовими (відповідно до рекомендацій п. 2.1.2).

З урахуванням викладеного вище, побудуємо моделі для ряду кількості хворих на коронавірус в Україні у 2020 р. без урахування інших ознак як регресорів.

Але для використання багатофакторних моделей машинного навчання потрібні різні ознаки, а не – лише одна. Для цього скористаємось Python-бібліотекой Tsfresh, яка автоматично генерує 1200 ознак для заданого часового ряду, а потім прибирає маловажливі. З відібраних нею ознак ще приберемо такі, які містять багато пропущених значень чи є незмінними для усіх дат. Результуючі 77 ознак наведені у публічному Python-ноутбуці у Kaggle [72] (рис. 2.6).

	ds	y	level_0__sum_values	level_0__abs_energy	level_0__median	level_0__mean	level_0__root_mean_square	level_0
0	2020-04-01	149	0.0	0.0	0.0	0.0	0.0	0.0
1	2020-04-02	103	1.0	1.0	1.0	1.0	1.0	1.0
2	2020-04-03	175	2.0	4.0	2.0	2.0	2.0	2.0
3	2020-04-04	153	3.0	9.0	3.0	3.0	3.0	3.0
4	2020-04-05	83	4.0	16.0	4.0	4.0	4.0	4.0
...	...	...	...	...	...	...	...	...
245	2020-12-02	13443	245.0	60025.0	245.0	245.0	245.0	245.0
246	2020-12-03	14812	246.0	60516.0	246.0	246.0	246.0	246.0
247	2020-12-04	15456	247.0	61009.0	247.0	247.0	247.0	247.0
248	2020-12-05	14157	248.0	61504.0	248.0	248.0	248.0	248.0
249	2020-12-06	11928	249.0	62001.0	249.0	249.0	249.0	249.0

250 rows × 79 columns

Рис. 2.6. Синтезовані бібліотекою Tsfresh ознаки для часового ряду кількості нових щодобових хворих на коронавірус в Україні у 2020 р.

По графіку часового ряду 2020-2022 рр. визначаємо дати аномалій, коли наростання кількості хворих змінювалось на спадання, і – навпаки. Повторюємо цю процедуру для першої різниці ряду. Результатом є такі дати [63]: 2020-11-28, 2021-04-03, 2021-11-04, 2021-04-01, 2021-04-04, 2021-11-14, але для 2020 р. (як на рис. 2.6) підходить тільки 2020-11-28.

Для ARIMA оптимальними параметрами є (5,1,4), тобто AR(5), перша (1) різниця (див. висновки у п. 2.1.3) і MA(4). Результат діагностування цієї моделі наведено на рис. 2.7.

```
Optimal parameters are [5, 1, 4]
```

```
if is_ARIMA:
    # Best model from AutoARIMA
    fig = model_auto.plot_diagnostics(figsize=(12,10))
    plt.show()
```

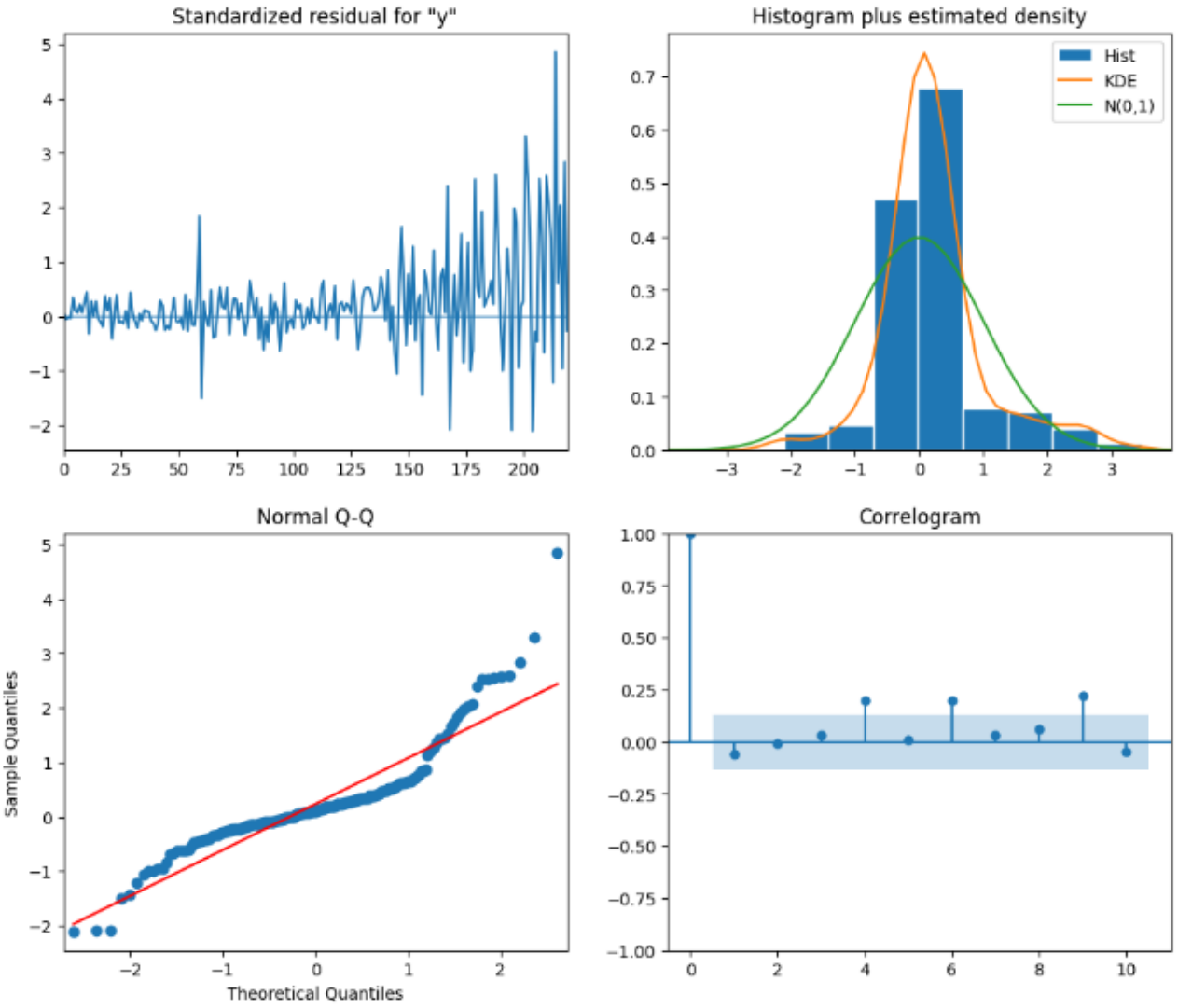


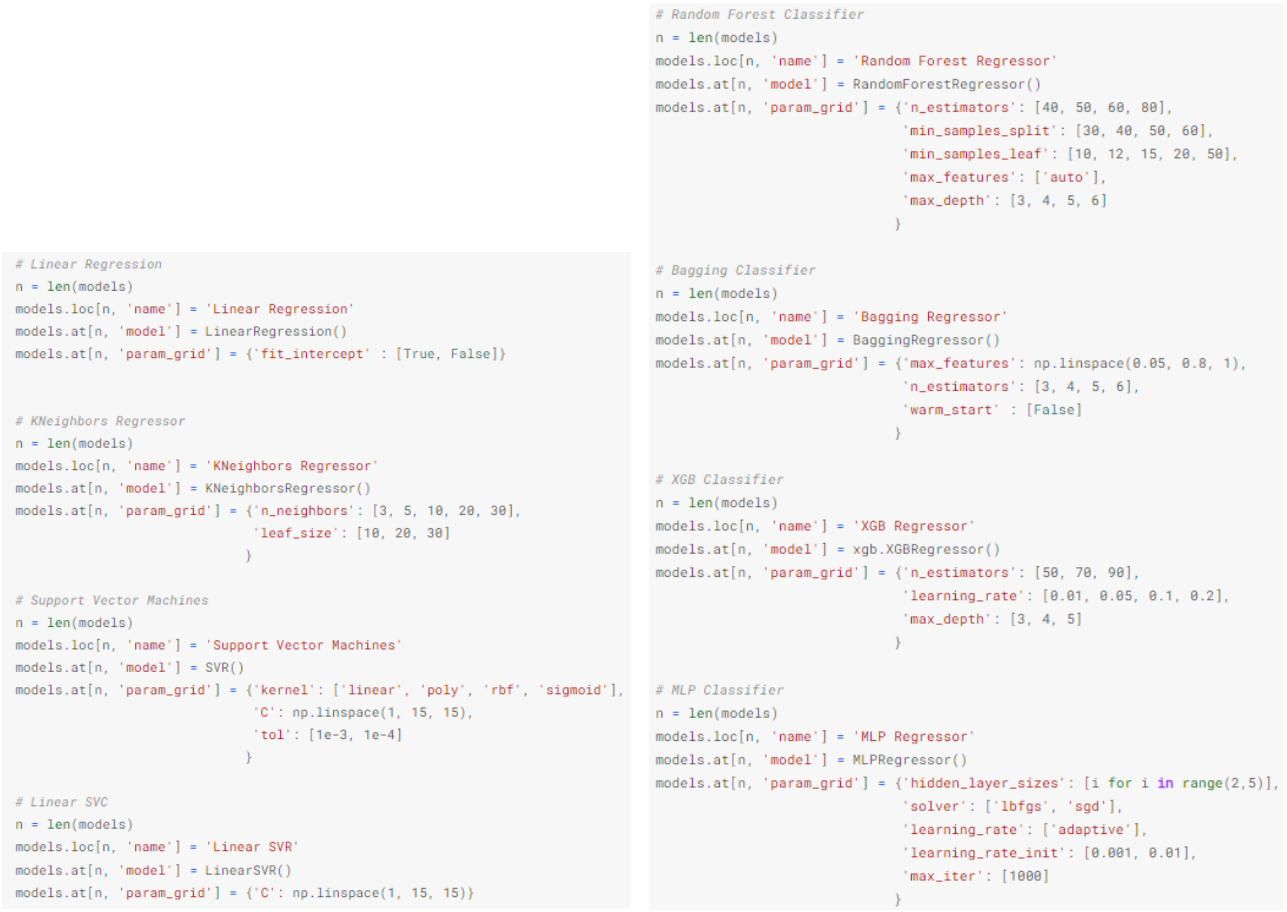
Рис. 2.7. Діагностування моделі ARIMA (5,1,4) за даними 2020 р.

Як видно на рис. 2.7, результат моделювання є задовільним, але не дуже добрим, оскільки графік похибок на правому верхньому рисунку не повністю відповідає нормальному закону, а крива на лівому нижньому рисунку – не зовсім схожа на пряму, а значення у правому нижньому рисунку не залишаються у зоні

статистичної значущості після першого потрапляння у неї і носять коливальний характер. Отже, така модель не є достатньо адекватною.

Щодо моделі FB Prophet враховуємо дату аномалії та здійснюємо в циклі перебір усіх комбінацій періоду у 4, 7, 30 та 365 діб з порядком 3 або 12 для моделювання сезонності рядом Фур'є.

Щодо моделей машинного навчання будуємо моделі з параметрами, підібраними на основі практичного досвіду (рис 2.8).



```
# Linear Regression
n = len(models)
models.loc[n, 'name'] = 'Linear Regression'
models.at[n, 'model'] = LinearRegression()
models.at[n, 'param_grid'] = {'fit_intercept' : [True, False]}


# KNeighbors Regressor
n = len(models)
models.loc[n, 'name'] = 'KNeighbors Regressor'
models.at[n, 'model'] = KNeighborsRegressor()
models.at[n, 'param_grid'] = {'n_neighbors': [3, 5, 10, 20, 30],
                              'leaf_size': [10, 20, 30]
                             }

# Support Vector Machines
n = len(models)
models.loc[n, 'name'] = 'Support Vector Machines'
models.at[n, 'model'] = SVR()
models.at[n, 'param_grid'] = {'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
                              'C': np.linspace(1, 15, 15),
                              'tol': [1e-3, 1e-4]
                             }

# Linear SVC
n = len(models)
models.loc[n, 'name'] = 'Linear SVR'
models.at[n, 'model'] = LinearSVR()
models.at[n, 'param_grid'] = {'C': np.linspace(1, 15, 15)}
```

```
# Random Forest Classifier
n = len(models)
models.loc[n, 'name'] = 'Random Forest Regressor'
models.at[n, 'model'] = RandomForestRegressor()
models.at[n, 'param_grid'] = {'n_estimators': [40, 50, 60, 80],
                              'min_samples_split': [30, 40, 50, 60],
                              'min_samples_leaf': [10, 12, 15, 20, 50],
                              'max_features': ['auto'],
                              'max_depth': [3, 4, 5, 6]
                             }

# Bagging Classifier
n = len(models)
models.loc[n, 'name'] = 'Bagging Regressor'
models.at[n, 'model'] = BaggingRegressor()
models.at[n, 'param_grid'] = {'max_features': np.linspace(0.05, 0.8, 1),
                              'n_estimators': [3, 4, 5, 6],
                              'warm_start' : [False]
                             }

# XGB Classifier
n = len(models)
models.loc[n, 'name'] = 'XGB Regressor'
models.at[n, 'model'] = xgb.XGBRegressor()
models.at[n, 'param_grid'] = {'n_estimators': [50, 70, 90],
                              'learning_rate': [0.01, 0.05, 0.1, 0.2],
                              'max_depth': [3, 4, 5]
                             }

# MLP Classifier
n = len(models)
models.loc[n, 'name'] = 'MLP Regressor'
models.at[n, 'model'] = MLPRegressor()
models.at[n, 'param_grid'] = {'hidden_layer_sizes': [i for i in range(2,5)],
                              'solver': ['lbfgs', 'sgd'],
                              'learning_rate': ['adaptive'],
                              'learning_rate_init': [0.001, 0.01],
                              'max_iter': [1000]
                             }
```

Рис. 2.8. Допустимі множини значень для ідентифікації параметрів моделей машинного навчання

Для ідентифікації параметрів моделей машинного навчання використовуємо функцію GridSearchCV, яка здійснює повний перебір можливих варіантів і вибір найкращого серед них – довго, але максимально точно.

Результат ідентифікації усіх цих видів моделей та оцінювання точності прогнозу за валідаційними даними за метриками r2_score із п. 1.2.6 наведено на рис. 2.9.

name_model	type_data	r2_score	rmse	mape
Prophet_7_days_3_order	valid	0.948961	377.283507	2.724119
Prophet_7_days_12_order	valid	0.932915	432.541405	3.095432
Linear Regression	valid	0.546218	1124.963248	7.598622
MLP Regressor	valid	0.591935	1066.791733	7.632433
Prophet_4_days_3_order	valid	0.56331	1103.574012	7.692877
Prophet_4_days_12_order	valid	0.494738	1187.060858	8.28186
KNeighbors Regressor	valid	0.37277	1322.599121	8.970068
Random Forest Regressor	valid	0.321687	1375.402431	9.175315
Prophet_365_days_3_order	valid	0.043321	1633.419954	11.046927
Prophet_30_days_3_order	valid	-0.034667	1698.693218	11.329504
Prophet_30_days_12_order	valid	-0.508503	2051.104402	14.480213
ARIMA_auto	valid	-1.043162	2387.075372	15.823972
Support Vector Machines	valid	-1.222799	2489.801995	16.240231
Linear SVR	valid	-1.434648	2605.75055	17.124106
Prophet_365_days_12_order	valid	-2.839393	3272.246318	18.751056
XGB Regressor	valid	-9.462468	5401.716501	37.091622
Bagging Regressor	valid	-74.651107	14525.208186	101.13883

Рис. 2.9. Результат ідентифікації моделей машинного навчання за валідаційними даними за 2 тижні в кінці 2020 р., відсортований за погіршенням значення метрики r2_score

Як видно, на рис. 2.9, найбільш перспективною є модель FB Prophet, але для удосконалення її точності варто більш ретельно визначати дати аномалій та ідентифікувати більше параметрів, для чого необхідно розробити більш досконалі методи ідентифікації, ніж простий перебір варіантів усіх можливих значень.

2.1.5. Розширення множини дат аномалій.

Проведений аналіз показав, що варто враховувати такі дати аномалій (свята і псевдосвята) [63]:

1. Державні свята (за даними пакету Holidays [73]) із 7-денним зсувом вперед (параметр “ds”) та адаптивним вікном (у першому наближенні беремо вікно [-3, 3], тобто від 4 до 10 днів) (рис. 2.10).

	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. 2.10. Державні свята України (за даними пакету Holidays [73]), дата «ds» - це справжня дата, посунута на 7 днів вперед

2. Дати, коли одночасно було дуже тепло і без опадів, коли різко збільшувалась кількість людей у місцях відпочинку – назвемо їх «метеопатерни» (за даними датасету «COVID-19 Open Data», зокрема, за даними National Oceanic and Atmospheric Administration (NOAA) — Національного управління з питань океану та атмосфери США [74]). Пропонуємо відбирати дати зі зсувом 7 днів та адаптивним вікном (рис. 2.4), коли кількість опадів була нульова, а середньодобова температура була більша за перцентиль P95, тобто значення, вищі цього, мали місце тільки у 5% випадків (рис. 2.11).

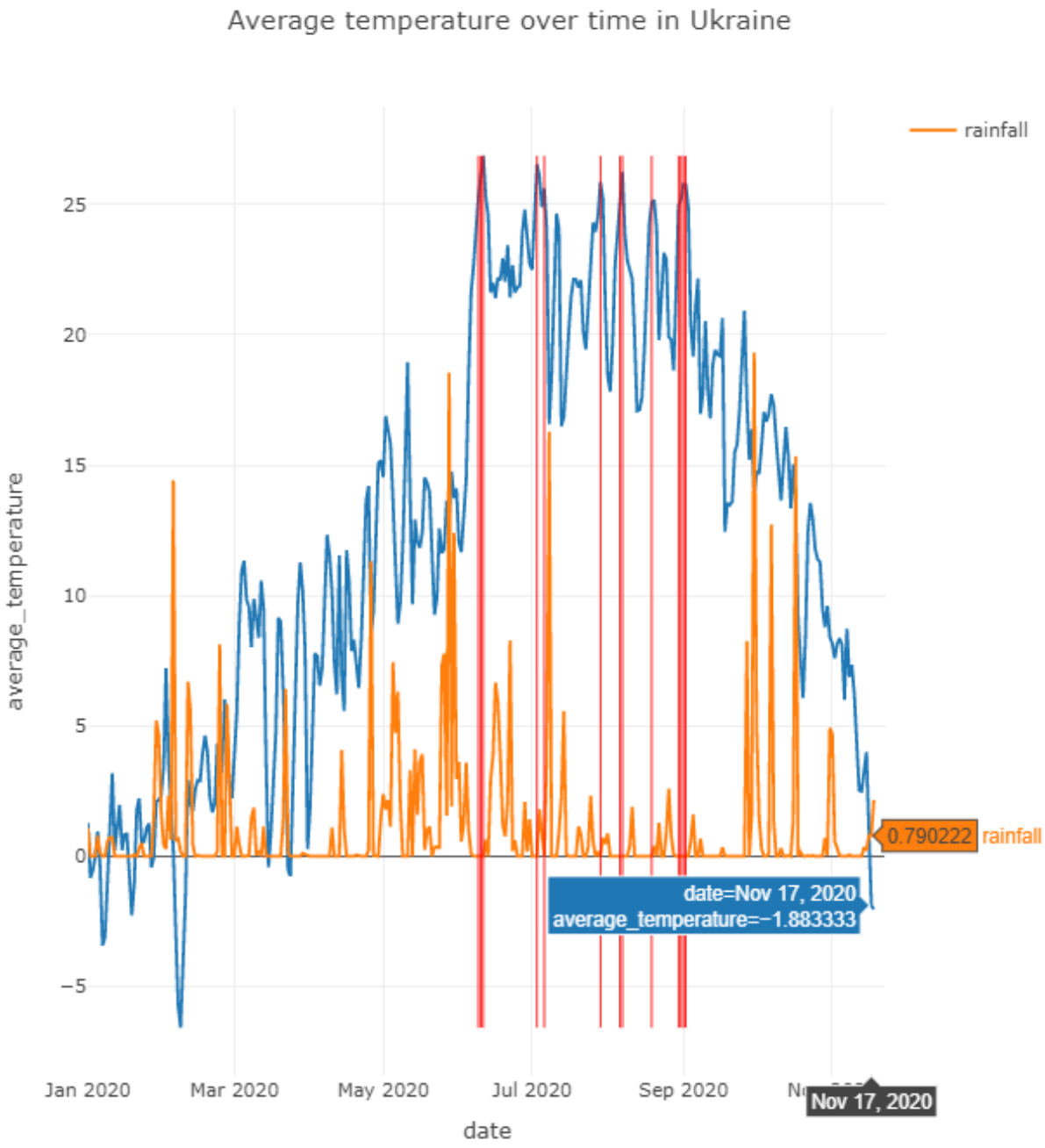


Рис. 2.11. Середньодобова температура в Україні (помаранчева крива), кількість опадів, мм (синя крива), дати аномальних «метопатернів» (червоні вертикальні лінії)

3. Дні послаблення карантину за Stringency-індексом (за даними «Oxford COVID-19 government response tracker» — Оксфордського трекеру коронавірусної діяльності урядів країн світу [67], які містяться у згаданому вище датасеті «COVID-19 Open Data»), котрий відображає усі послаблення карантину, згідно

рішень уряду України, за 17 критеріями. Дати, коли ця сума зменшувалась, формалізовано як дати послаблення карантину (рис. 2.12).

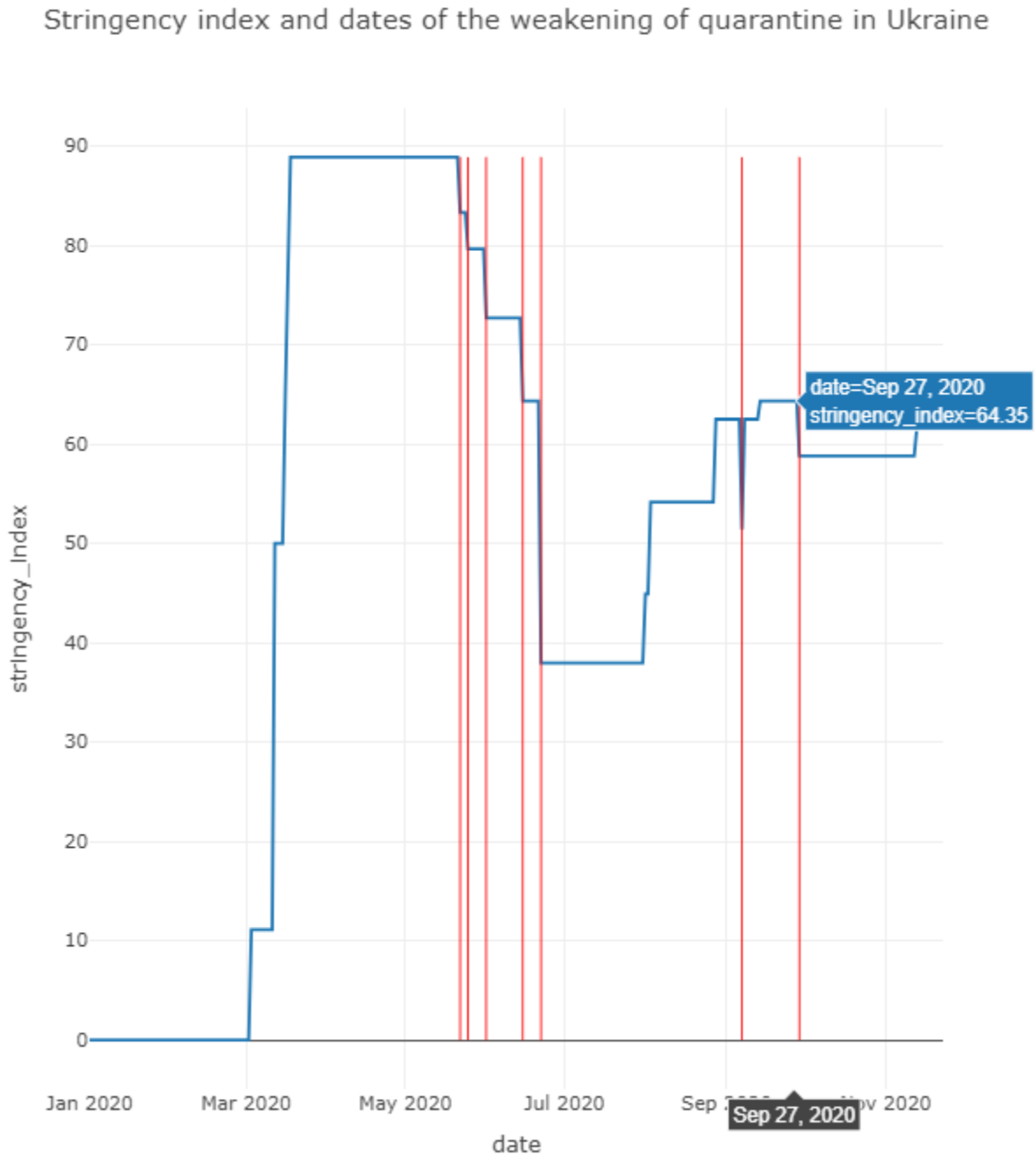


Рис. 2.12. Сумарний за добу Stringency-індекс України з ослабленнями та підсиленнями карантину (синя лінія) та дати аномалій-послаблень карантину (червона лінія)

2.1.6. Висновки розвідувального аналізу даних.

Проведений у підрозділі 1.1, 1.3 та пунктах 2.1.1-2.15 розвідувальний аналіз дозволяє зробити такі висновки:

1. Аналіз впливу різних факторів та можливості прогнозування поширення ковіду, зокрема Google-трендів щодо пересувань жителів України, метеоданих, даних Оксфордського трекера коронавірусної діяльності урядів країн світу та ознак, отриманих на їх основі, показав, що ці ознаки впливають нерівномірно. Отже, їх краще враховувати не як додатковий регресор для побудови множинної регресії, а – на їх основі згенерувати вторинні ознаки, що буде більш ефективним.

2. Аналіз матеріалів РГ по ковіду показав, що мають місце значні затримки в оприлюдненні результатів тестів (під час піків хвиль - більше 30 днів) і вони мають нерівномірний характер у часі, а отже, за цими даними важко буде здійснювати довгострокове прогнозування і варто обмежитись короткостроковим прогнозуванням з горизонтом в 1-2 тижні.

3. Доведено, що ряд кількості нових хворих на коронавірус в Україні та його перша різниця є нестаціонарними, а стаціонарною є тільки його друга різниця, що варто враховувати під час моделювання методами машинного навчання.

4. Декомпозиція ряду з різним періодом сезонності показали, що має місце тижнева періодичність і мінімум 1-2 з іншим періодом, але це потребує окремого дослідження.

5. Аналіз різних видів аномалій, які впливають на кількість хворих на коронавірус, показав, що їх слід враховувати з різним запізненням та різним вікном (початок і кінець впливу у добах) і силою впливу, що потребує окремого дослідження, тобто це – параметри, які потребують ідентифікації.

6. Порівняльний аналіз відомих методів та моделей для прогнозування кількості хворих на коронавірус (ARIMA, Facebook Prophet та інших моделей машинного навчання на основі дерев рішень, регресій та ін.) довів, що більш

точною є модель Facebook Prophet, отже варто ідентифікувати саме її структуру та параметри.

7. Модель Facebook Prophet з урахуванням декількох видів сезонних складових, загалом, має десятки параметрів. Повний перебір усіх комбінацій значень для пошуку оптимального варіанту займе забагато часу, а тому потрібні більш швидкі методи.

2.2 Розроблення методу ідентифікації структури та параметрів моделі для прогнозування кількості нових випадків на коронавірус на основі моделі FB Prophet

Для розв'язання поставленої задачі, з урахуванням Для моделювання використовувались відкриті офіційні дані РНБО України, які оновлювались щодня і протягом 2020-2022 рр. до вторгнення росії в Україну були доступними по API з веб-порталу (https://covid19.rnbo.gov.ua/) – саме у такий спосіб їх збирали й усі світові веб-сервіси. Для інших країн варто використовувати відомий датасет Google-платформи «COVID-19 Open Data»[68], де доступно багато статичної і динамічної, у т.ч. щоденної, інформації (показники захворюваності на коронавірус, дати карантинних обмежень, мобільність населення за даними Google, погодні умови тощо) по більшості країн світу.

Згідно поставленої задачі варто дослідити гіпотезу того, що свята можуть впливати на динаміку захворюваності, але із запізненням. Люди, особливо у карантинних умовах, звикли святкувати та контактувати під час свят, у громадському транспорті, в магазинах, на роботі тощо, що збільшує ризик поширення вірусу. Проведений авторами кореляційний аналіз даних показав, що доцільно брати запізнення на 7 діб, але з адаптивним вікном (наприклад, вікно [-1, 1] означає, що аномальними вважаються дати зі зсувом у 6-8 діб).

На етапі ідентифікації (тренування) моделі варто оптимізувати розмір адаптивного вікна навколо зсуву у 7 діб, наприклад, перебирати усі варіанти

навколо нього в діапазоні від 4 до 10 діб (прирости -3, -2, -1, 0, 1, 2, 3) і обирати варіант з найменшою похибкою.

Дні свят без зсуву на 7 діб 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 діб, де він був зроблений) показані на рис. 2.13 вертикальними лініями.

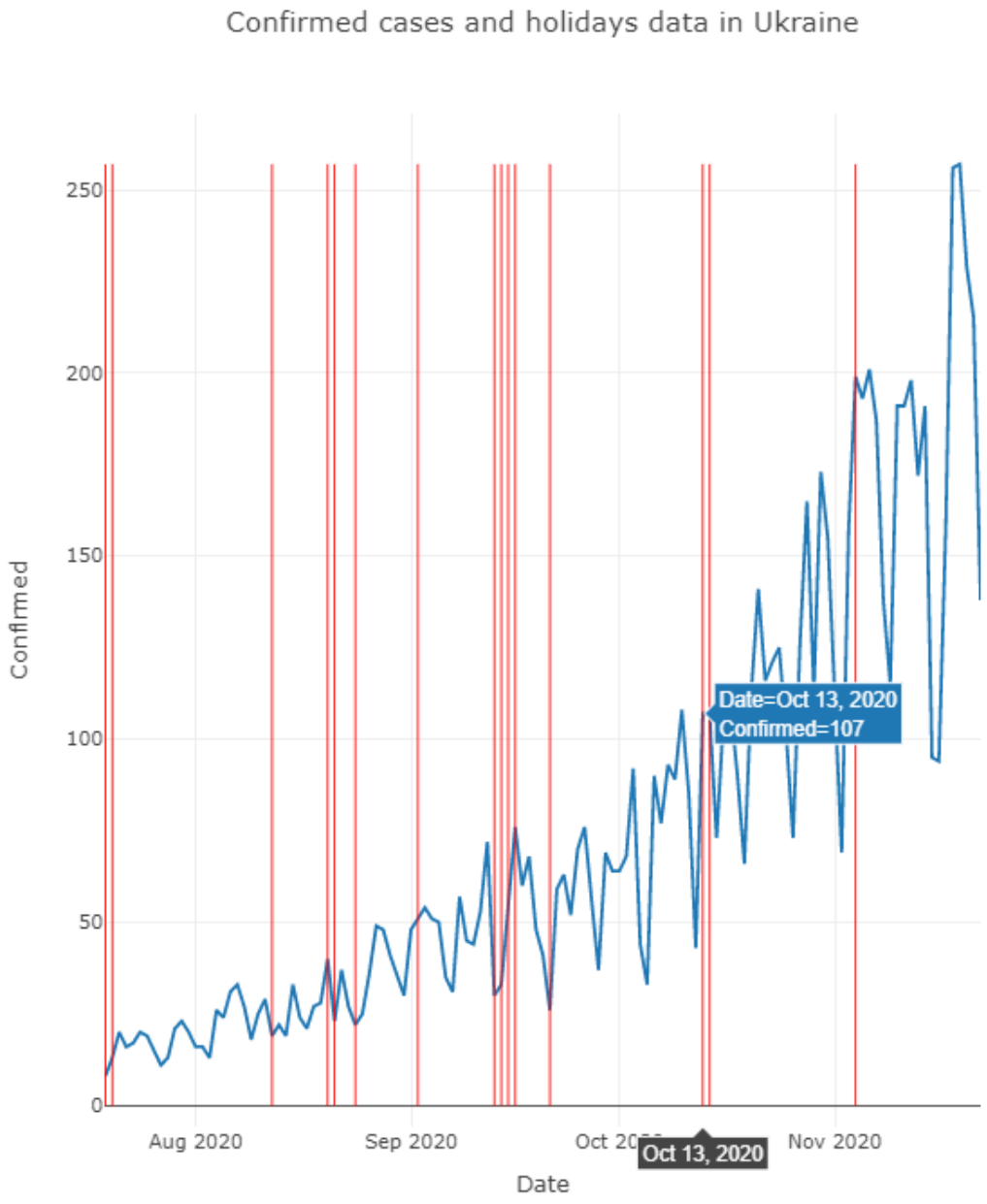


Рис. 2.13. Щоденні значення кількості нових випадків захворювань на коронавірус в Україні (синя крива) та дати аномалій (червоні вертикальні лінії)

Аналіз показав, що, параметрів Prophet, які задаються за замовчуванням, недостатньо для побудови достатньо точної моделі – часто саме цю помилку роблять інші дослідники [75-77]. Дослідження показали, що, враховуючи складний нелінійний характер зміни даних, варто здійснювати налаштування таких параметрів моделі [63]:

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

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

3. 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: $n = 2, 3, \ldots 6$ 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 було обґрунтовано вище, для побудови моделі «хвилі», пік якої ще не досягнуто, варто використовувати лінійну регресію Prophet. Така регресія традиційно будується у вигляді шматково-лінійної апроксимації тренду між точками суттєвої зміни значень (за замовчуванням береться 30 таких точок, але модель Prophet їх кількість адаптивно оптимізує) [69].

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, у т.ч. моделей часових рядів, прийнято наявні дані розділяти на тренувальну вибірку, на якій налаштовуються параметри, і – валідаційну вибірку, на якій перевіряється яка із 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, якій відповідає метрика WAPE («Weighted Mean Average Percentage Error» – англ. «зважена середня абсолютна відносна помилка»), на усіх датах валідаційної вибірки.

Перебір усіх комбінацій таких параметрів – довготривала задача. Тому, для прискорення роботи алгоритму пропонується задавати обмежену кількість варіантів можливих значень. Наприклад, значення кожного параметра вибирати тільки із 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 із них (наприклад, регуляризацію урахування свят), а інші (різні види сезонності) – вираховувати із нього шляхом поділу на певний коефіцієнт.

Для розв’язання поставленої задачі розроблено алгоритм інформаційної технології з опрацювання вхідних даних та налаштування і застосування моделі, яка містить такі етапи (рис. 2.5):

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

Етап 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).

Етап 3. Другий етап побудови моделі та оптимізація її параметрів із вже оптимізованими на першому етапі параметрами, зокрема перебір значень таких параметрів (усі види сезонності, як було зазначено вище, враховуються мультиплікативно): варіанти режиму урахування (мультиплікативний чи адитивний), ступінь регуляризації та порядок ряду Фур'є для опису тижневої (7-денної) та, окремо, 4-денної сезонності. Для спрощення можна вибирати однаковий варіант режиму врахування обох видів сезонності.

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

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

Алгоритм запропоноваго методу наведений на рисунку 2.14:

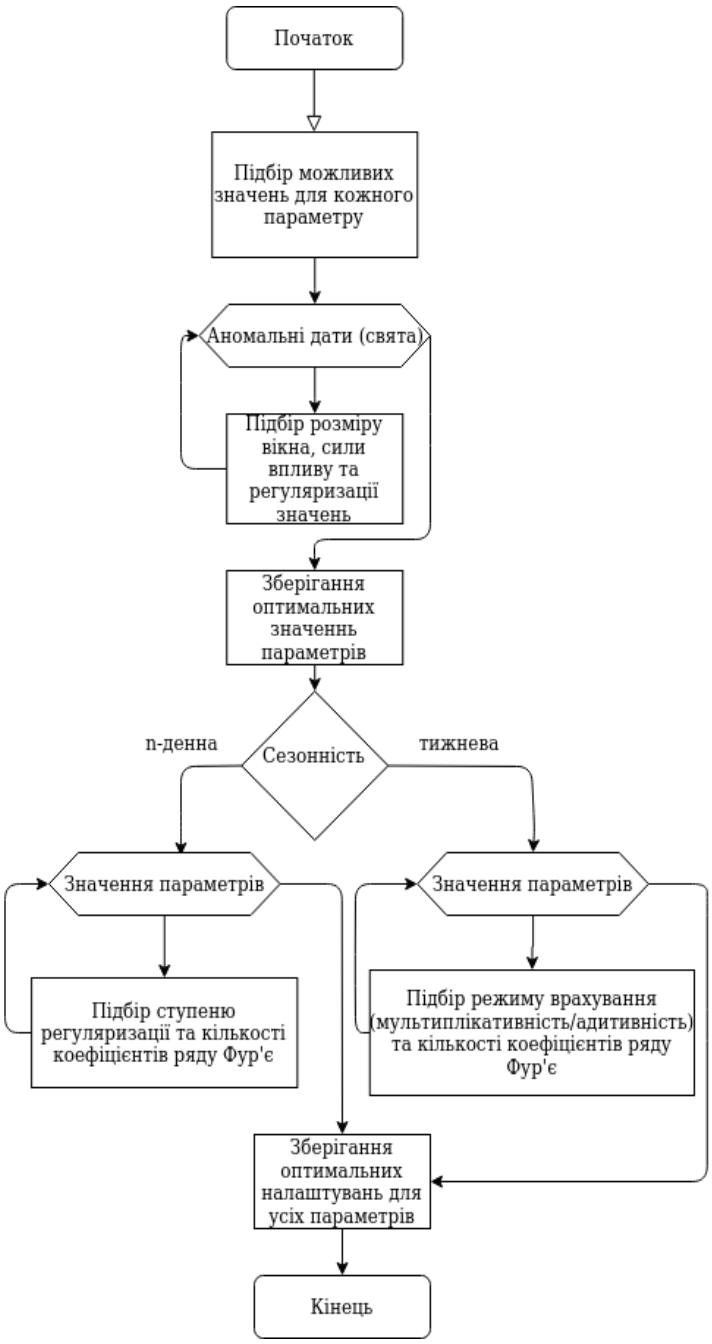


Рис. 2.14. Блок-схема алгоритму запропоноваго методу

Математично запропонований метод ідентифікації полягає в наступному. Замість типового запису моделі FB Prophet (1.16) пропонується такий:

$$y(t) = g\left(R_g, t\right) \cdot \prod_{i=1}^{\Phi} s(R_{si}, P_i, n_i, t) \cdot h\left(\bigcup_{j=1}^{\Psi} H_j\left(R_{hj}, \Delta t_j, t_{0j}, t_{1j}, t\right)\right) \cdot \epsilon_t, \quad (2.1)$$

де R_g – ступінь регуляризації тренду (як правило, від 0,01 до 100), тобто наскільки гнучко тренд має припасовуватись до різких змін значень; Φ – кількість сезонних складових, які одночасно враховуються, для кожної i-ої сезонної складової окремо задаються такі параметри: R_{si} – ступінь регуляризації, P_i – період у добах, n_i – порядок ряду Фур'є, яким буде описуватись ця складова; Ψ – кількість видів аномалій, дати яких враховуються як «свята», для кожної j-ої аномалії («свята») окремо задаються такі параметри: R_{hj} – ступінь регуляризації, Δt_j – зсув дати відносно справжнього значення для врахування 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 вже практично не проявляється її вплив; $H_j(\cdot)$ – матриця з датами аномалій та їх параметрами.

Важливо, що одна аномалія в (2.1) може враховуватись двома способами одночасно, наприклад, офіційні свята враховуються двічі: і як аномалія у роботі лабораторії, 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 їх контагіозності, результат якого проявиться через 4-7 діб.

В загальному випадку, кількість N_k гіперпараметрів моделі (2.1), тобто параметрів, які слід ідентифікувати, складає:

$$N_k = 1 + 3\Phi + 4\Psi. \tag{2.2}$$

За умови ідентифікації параметрів методом повного перебору, коли для кожного гіперпараметра вибирається одне з m варіантів значень, тоді кількість N_M можливих моделей, які слід застосувати, що теж займе час, дорівнює:

$$N_M = m^{N_K}. \quad (2.3)$$

Наприклад, для випадку $\Phi = 3$ видів сезонності та $\Psi = 5$ видів аномалій $N_k = 30$. Якщо кожен параметр має $m = 4$ варіанти значень, тоді це складе $N_M = 1.15 \cdot 10^{18}$ комбінацій, що – задовго. А тому слід шукати більш оптимізовані підходи.

Проведений аналіз показав, що мають місце $\Phi = 3$ види сезонності [78]:

- внутрішньотижнева (найкраще себе проявила 4-добова: $P_1 = 4$);
- тижнева, обумовлена циклом роботи лабораторій (така особливість має місце в усьому світі): $P_2 = 7$;
- багатоденна, яка враховує багатохвильову природу процесу (період P_3 потребує окремого визначення).

Доцільно враховувати $\Psi = 5$ види «свят» [78]:

- дати офіційних свят як складова, що впливає на контагіозність людей (“holiday”) ($\Delta t_1 = \Omega$);
- дати офіційних свят як складова, що впливає на перерву у роботі багатьох лабораторій, які аналізують тести на коронавірус (“lab”) ($\Delta t_1 = 0$);
- системні урядові заходи та карантинні обмеження щодо транспорту, освіти тощо (див. статтю [11]) як дати зміни сумарного значення урядового трекера (“SI”) ($\Delta t_1 = \Omega$);
- дати «локдаунів» – «карантини вихідного дня» (“weekend”) ($\Delta t_1 = 0$) – береться з нульовим зсувом, що враховує самі ці дні, а вплив зі зсувом цих днів урахований в аномалії типу “SI”;
- дати метеорологічних аномалій, описаних у статті [16] (“meteo”) ($\Delta t_1 = \Omega$),

де Ω – гіперпараметр, який, як правило, приймає значення 4-7, але його треба щоразу уточняти, через вплив різних штамів коронавірусу та їх сукупність (у першому наближенні, можна вважати, що він є однаковим для різних видів сезонності, хоча, в загальному випадку, це може і не мати місце). Також, у

першому наближенні, інші параметри для усіх «свят» можуть ідентифікуватись як однакові:

$$R_{hj} = R_h,\ t_{0j} = t_0,\ t_{1j} = t_1,\quad j = \overline{1,\Psi}. \tag{2.4}$$

Для ефективного застосування моделі (2.1), з урахуванням спрощення (2.4), її гіперпараметрами є такі 12:

$$K = [R_g, R_{s1}, n_1, R_{s2}, n_2, R_{s3}, P_3, n_3, R_h, t_0, t_1, \Omega]. \tag{2.5}$$

У разі застосування методу повного перебору, наприклад, з 4 варіантами, за формулою (2.3) це – 4 194 304 можливих комбінацій, що теж забагато. Звичайно, є можливим застосування байєсівських методів та Python-бібліотеки HyperOpt, але такий підхід дає менш достовірні оцінки параметрів, аніж повний перебір варіантів.

Для пришвидшення ідентифікації параметрів (2.5) моделі (2.1) у статті [16] 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 поодиноких аномалій у певні моменти часу, що дозволяє їх параметри ідентифікувати окремо у 2 етапи:

$$\text{Stage 1: } \begin{cases} R_g = R_{g0},\ R_{s1} = R_{s10},\ n_1 = n_{10},\ R_{s2} = R_{s20},\ n_2 = n_{20}, \\ R_{s3} = R_{s30},\ P_3 = P_{30},\ n_3 = n_{30}, \\ R_h \to R_{hopt},\ t_0 \to t_{0opt},\ t_1 \to\ t_{1opt},\ \Omega \to \Omega_{opt}, \end{cases} \tag{2.6}$$

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

де змінні з індексом «0» – це початкове наближення їх значень, отримане на основі розвідувального аналізу, а змінні з індексом «opt» – це оптимальні значення, обчислені на відповідному етапі ідентифікації.

Введення певних додаткових співвідношень може ще зменшити кількість параметрів. Зокрема, варто врахувати, що сезонні складові з меншим періодом доцільно робити більш гнучкими (з меншою регуляризацією), аніж складові з більшим періодом (з більшою регуляризацією), а тоді можна ввести додаткові співвідношення:

$$R_{si} = \frac{R_g}{k_{si}}, i = \overline{1, \Phi}\,, \tag{2.7}$$

де $k_{si}, i = \overline{1, \Phi}$ – числові коефіцієнти (для відносно малих значень періодів вони є меншими 1, наприклад: 0.5, для більших – більшими 1, наприклад: 2.0).

Крім того, порядки рядів Фур'є окремих видів сезонності теж можна вибирати однаковими або такими, які пов'язані певними співвідношеннями з іншими значеннями, з відповідним округленням до цілого.

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

1. урахування тільки державних свят з 7-денним зсувом та адаптивним вікном;
2. урахування тільки тижневої сезонності.

З урахуванням цього модель (2.1) упроститься до такої:

$$y(t) = g\big(R_g, t\big) \cdot \prod_{i=1}^{\Phi} s\big(R_{si}, \theta_{wsi,} n_i, t\big) \cdot h\Big(\bigcup_{j=1}^{\Psi} H_j\big(R_{hj}\,, \theta_{hj}, t\big)\Big) \cdot \epsilon_t, \tag{2.8}$$

де θ_{wsi} – параметр тижневої сезонності, а θ_{hj} – державні свята з 7-денним зсувом.

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 – задавати фіксованими, на основі попередніх розрахунків:

$$\text{Stage 1:} \begin{cases} R_{gopt},\ R_{sopt},\ n_{opt}, \\ R_{hopt},\ \theta_{ws}, \theta_h, \\ R_h = R_{h0},\ t_0 \to t_{0opt},\ t_1 \to\ t_{1opt},\ \Omega \to \Omega_{opt}, \end{cases} \tag{2.9}$$

$$\text{Stage 2:} \begin{cases} R_{gopt},\ R_{sopt},\ n_{opt},\ ,\ n_2 = n_{20}, \\ R_{sopt}, \theta_{ws}, \theta_h, \\ R_h = R_{h0},\ t_0 \to t_{0opt},\ t_1 \to\ t_{1opt},\ \Omega \to \Omega_{opt}. \end{cases}$$

Таку спрощену модель можна застосовувати одразу для досить великої кількості країн, областей та ін.

Було здійснено випробування запропонованого методу для України за даними 6.07-22.11.2020 р. Дані про захворювання на коронавірус брались у 2020 році по API з Системи моніторингу поширення епідемії коронавірусу Апарату РНБО України, решта даних – там, де було рекомендовано вище.

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

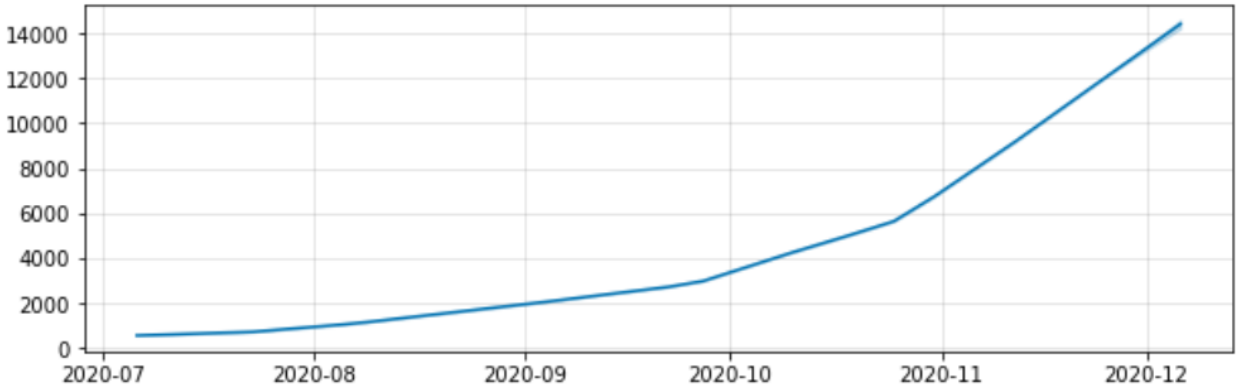
Станом на 22.11.2020 р. оптимальною моделлю з N = 14 днів, тобто з прогнозом на 2 тижні вперед, є модель з такими параметрами: свята слід ураховувати з вікном від 5 до 7 днів із силою впливу 2,5, регуляризацією 0,15, мультиплікативним урахуванням тижневої сезонності із регуляризацією 0,12, яка описується рядом Фур'є порядку 8, та мультиплікативним урахуванням 4-денної сезонності із регуляризацією 0,075, яка описується рядом Фур'є порядку 1 (рис. 7, 8) [10]. Тоді модель забезпечує відносну похибку 2,2% за останні 14 днів і дозволяє оцінити прогноз на наступні N = 14 днів (табл. 2.1), за умови, що

збережеться така сама динаміка (карантинний режим, сумарна за добу кількість тестів та ін.).

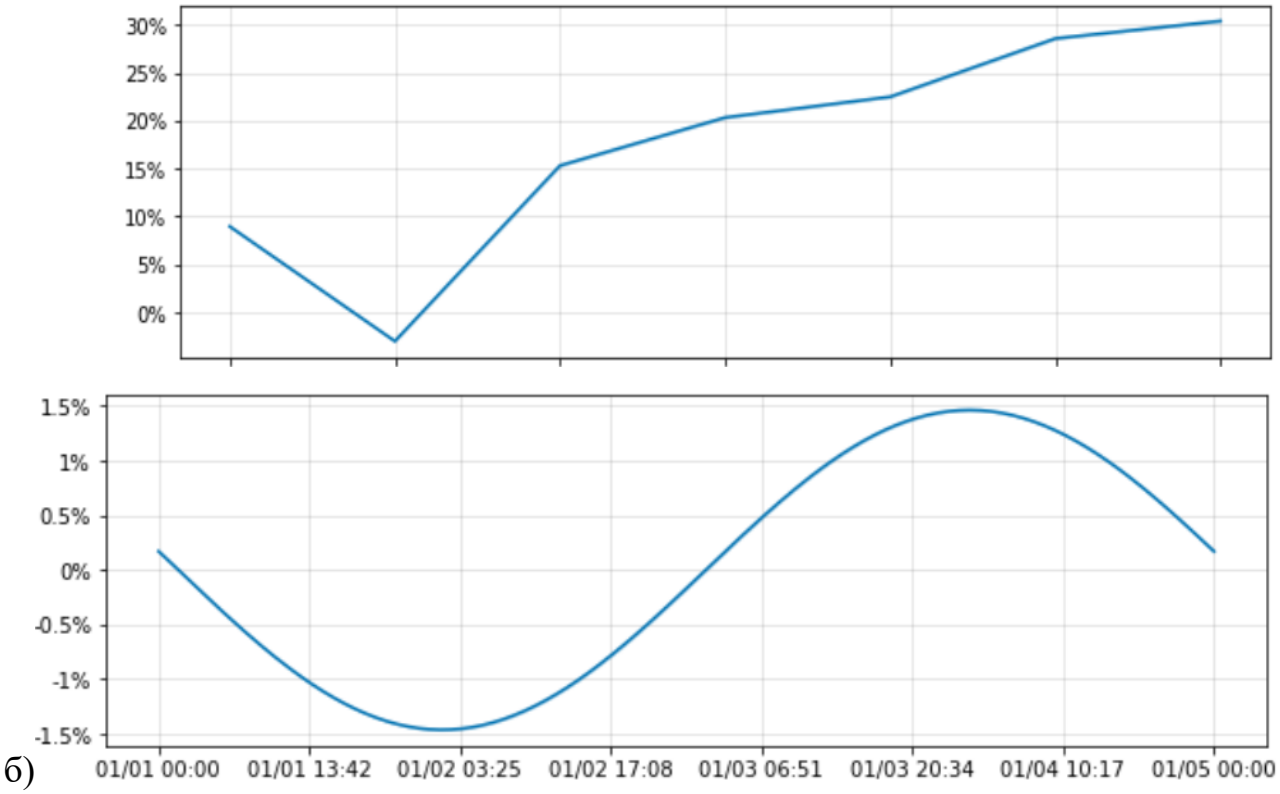
Таблиця 2.1. Прогноз кількості нових підтверджених випадків хворих на коронавірус в Україні за моделлю з урахуванням впливу аномальних дат з довірчим інтервалом 0,8

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. 2.15. 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. 2.16. Складові моделі на рис. 2.6 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) 4-денна 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 4-денна сезонність, яка теж демонструє нелінійне зростання, причому, саме зростання має місце кожні 2 доби – її урахування 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 швидше, ніж за кожні 7 днів.

Отже, були розроблені методи багатопараметрової ідентифікації параметрів моделі машинного навчання для прогнозування часового ряду кількості хворих чи померлих від коронавірусу в заданому регіоні.

2.3 Розроблення методу оцінювання порядку ряду Фур'є для апроксимації багатоденного періодичного процесу зміни кількості нових випадків на коронавірус на основі моделі FB Prophet

Описаний у підрозд. 2.2 метод ідентифікації моделі FB Prophet для прогнозування кількості нових випадків на коронавірус був розроблений восени 2020 р. і спочатку давав дуже високу точність (див. рис. 2.15 – відносна похибка 2,2%), поки мала місце перша чверть періоду найбільшої хвилі цієї кількості. З часом, з переходом до ділянки зменшення щодобової кількості нових хворих (рис. 2.17), а тим більше, з появою наступних хвиль, модель перестала бути адекватною.

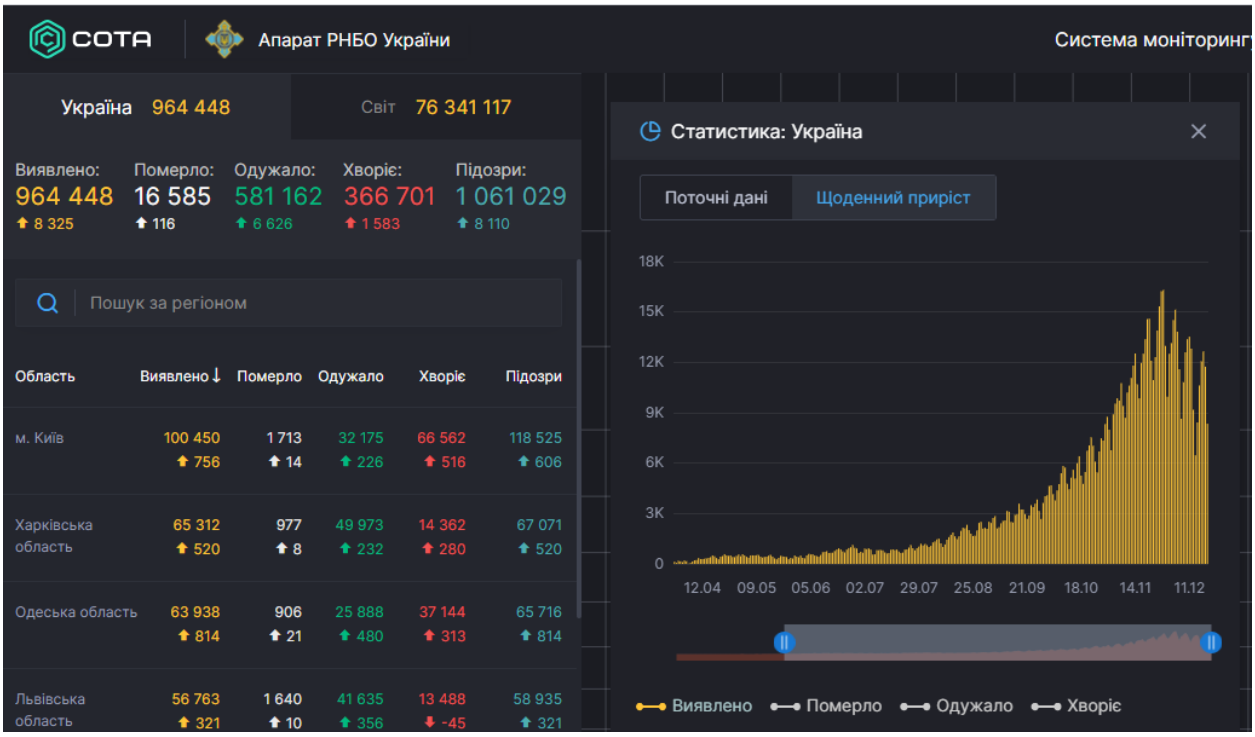


Рис. 2.17. Кількість нових підтверджених випадків хвороби на коронавірус «COVID-19» у 2020 р.
за даними РНБО України (https://covid19.rnbo.gov.ua/)

Виникла необхідність додавання багатоденної чи, навіть, багатотижневої сезонності, наявність якої передбачав й аналіз графіка на рис. 2.4.

Подібна ситуація мала місце й під час моделювання щодобової кількості нових хворих на коронавірус в інших країнах світу. Але, в багатьох випадках, задача ускладнювалась тим, що хвилі часто накладались одна на іншу. Тобто нова починалась, коли попередня ще тільки-но пішла на спад. За таких умов традиційні методи визначення параметрів ряду Фур'є для опису таких періодичних складових не працюють і потрібний новий метод, який міг би хоча б оцінити перше наближення основних гіперпараметрів.

Як було зазначено вище, для опису довільного виду сезонності у моделі Prophet у вигляді ряду Фур'є слід задати такі головні параметри (гіперпараметри) [12]:

1. «period» (T) – період сезонності у добах;

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

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 0,3–0,5);

4. 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:

$$T = 2(t_1 - t_0). \tag{2.10}$$

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.

Як відомо, ряд Фур'є має вигляд [23]:

$$y(t) = \frac{a_0}{2} + \sum_{i=1}^{n} a_i \cos\left(\frac{2\pi}{T} it\right) + \sum_{i=1}^{n} b_i \sin\left(\frac{2\pi}{T} it\right), \quad (2.11)$$

де коефіцієнти Фур'є знаходяться за допомогою виразів [8-11]:

$$a_0 = \frac{1}{T}\int_0^T y(t)dt,$$

$$a_i = \frac{2}{T}\int_0^T y(t)\cos\left(\frac{2\pi}{T} it\right)dt, \quad (2.12)$$

$$b_i = \frac{2}{T}\int_0^T y(t)\sin\left(\frac{2\pi}{T} it\right)dt, \quad i = \overline{1,n},$$

де, у разі, якщо ряд починається не з нуля, а – з деякого часу t_0, тоді час t замінюється на $(t - t_0)$.

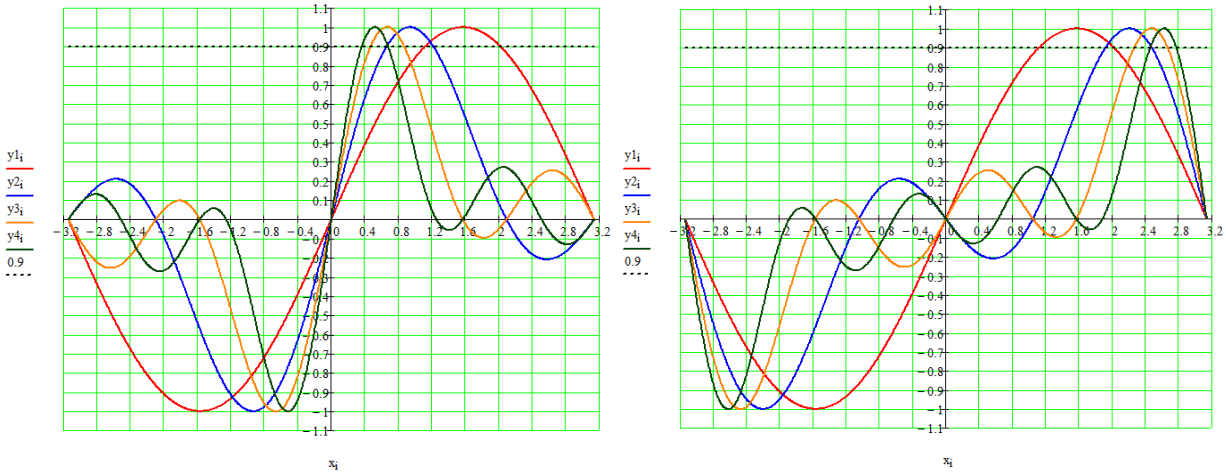
Як видно з виразів (2.11), (2.12), перші гармоніки мають найбільший період коливання *T* при *i* = 1, а усі інші – менший у 2, 3, … *n* разів.

Основна проблема застосування відомих методів для знаходження параметрів ряду Фур'є у вигляді (2.12) полягає в тому, що немає окремих значень часового ряду, обумовлених тільки заданою сезонністю, до яких можна застосувати ці методи. 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 пропонується певним шляхом оцінити можливі значення параметрів *T* і *n*, а потім запропонувати вбудованому алгоритму ідентифікації моделі Prophet вибрати з них найкращий варіант. 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. 2.18) є сукупністю так званих «гармонік» (синусоїд і косинусоїд певної амплітуди і частоти) [17]. Побудуємо графіки таких рядів Фур'є:

$$yn_i = \sum_n b_i \sin(nx_i), \quad (2.13)$$

де коефіцієнти b_i підбираються таким чином, щоб максимумом функції було значення 1. Графіки таких функцій для n = 1, 2, 3 показано на рис. 2.18.



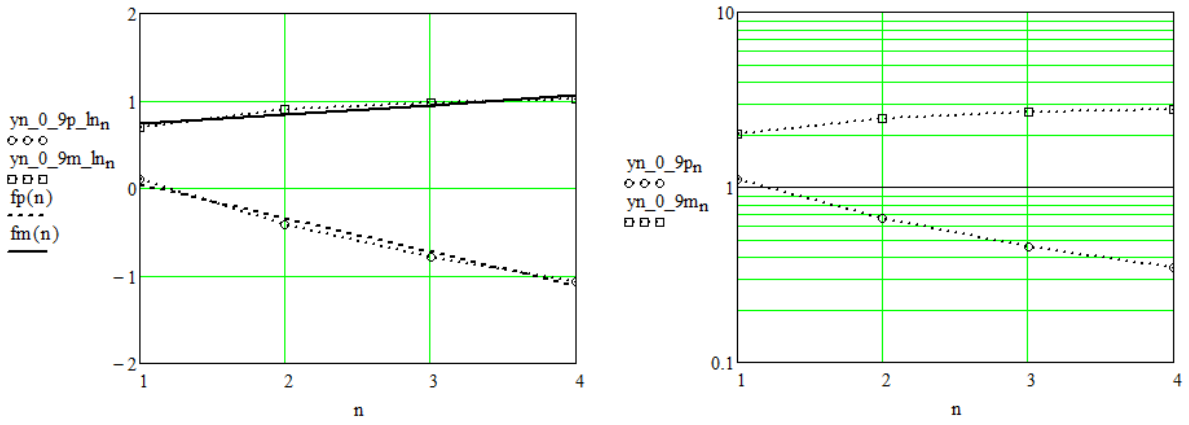
а) б)

Рис. 2.18. Сума різної кількості n гармонік (лише синусоїд виду (2.13)) 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 на рис. 2.18, при збільшенні кількості гармонік n пік (максимальне значення) ряду зсувається дедалі від піку ряду, побудованого при 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 до тих пір, поки 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 [17]. Однак, 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 має місце доволі часто, уся крива відсутня і висновок про порядок ряду Фур'є слід зробити тільки по частині хвилі, тому

пропонується інший підхід. Як правило, висновок про наявність хвилі робиться за наявності її піка і хоча б 10% значень верхівки кривої, у протилежному випадку, відповідну ділянку можна вважати частиною поточної (чи попередньої) хвилі. Проведемо на рис. 2.18 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. 2.18а, відстань першої точки перетину кожною кривою цієї лінії до середини напівперіоду $\frac{\pi}{2}$ прямо пропорційна значенню n: чим більшим є це значення n, тим більшою є відстань. Для графіку на рис. 2.18б 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. 2.19). Отже, можна припустити, що по координаті однієї цієї точки можна оцінити порядок ряду Фур'є.



а) б)

Рис. 2.19. Залежність від порядку рядів Фур'є n значень нормованого часу $t = {2\pi}/{T}$ для точок 90% від максимального для рядів з усіма додатними коефіцієнтами (yn_0_9p) і з від'ємними коефіцієнтами у парних гармоніках (yn_0_9m): а) у логарифмічних координатах по осі абсцис; б) для логарифмованих варіантів цих залежностей ln(yn_0_9p), ln(yn_0_9m) та їх апроксимація прямими

Апроксимація прямими залежностей на рис. 2.19 дала такий результат:

$$n_p = floor\left(\frac{0{,}424-\ln\left(t_{\text{yn\_0\_9\_p}}\right)}{0{,}385}\right), \qquad n_m = floor\left(\frac{\ln\left(t_{\text{yn\_0\_9\_m}}\right)-0{,}644}{0{,}104}\right), \qquad (2.14)$$

де floor() — це функція округлення до цілого значення у менший бік (принамні, на Python), а – $t_{\text{yn\_0\_9\_p}}$ та $t_{\text{yn\_0\_9\_m}}$ моменти часу, коли крива досягає значення у першій чверті періоду (наростаюча частина хвилі) за випадку додатніх та знакозмінних значень, відповідно.

Звичайно, формули (2.14) 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 (2.13). За інших коефіцієнтів Фур'є вона буде іншою, але виведення її у більш загальному вигляді — це тема окремого дослідження.

Для оцінювання довжини і періоду кожної хвилі, тобто параметрів T, t_0, t_1, пропонуємо застосовувати такий алгоритм (рис. 2.20):

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 погіршення в подальшому прогнозуванні.

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:

$$\frac{y_{min1}}{y_{max}} = sin\left((t_{min1} - t_0)\frac{2\pi}{4(t_{max}-t_0)}\right), \quad (2.15)$$

де y_{min1} — мінімальне значення кількості нових хворих на початку хвилі у дату t_{min1}, 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:

$$t_0 = \frac{t_{min}-at_{max}}{1-a}, \quad a = \frac{2}{\pi}arcsin\left(\frac{y_{min}}{y_{max}}\right). \quad (2.16)$$

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:

$$\frac{y_{min1}}{y_{max}} = sin\left(\left(t_{min2} - \left(t_1 - \frac{T}{2}\right)\right)\frac{2\pi}{T}\right), \quad T = 4(t_1 - t_{max}), \quad (2.17)$$

де y_{min2} — мінімальне значення кількості нових хворих в кінці хвилі у дату t_{min2}, знайдене у п.2 алгоритму.

Нескладно показати, що:

$$t_1 = \frac{t_{min2}-(a-2)t_{max}}{1-a}, \quad a = \frac{2}{\pi}arcsin\left(\frac{y_{min}}{y_{max}}\right). \quad (2.18)$$

4. Знаходимо період хвилі за формулою (2.17).

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

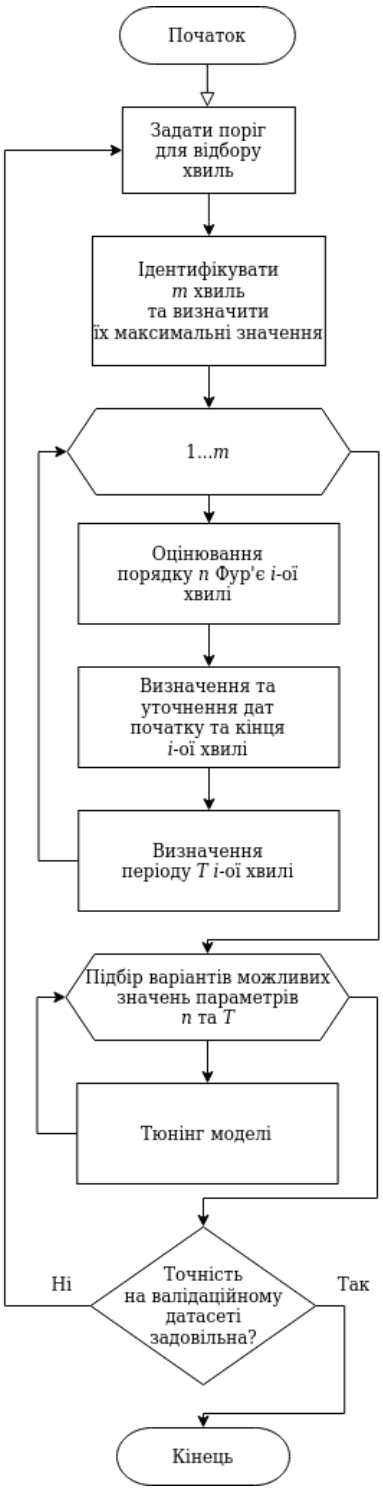


Рис. 2.20. Алгоритм методу оцінювання порядку Фур’є для опису багатотижневої сезонності ряду

Здійснено удосконалення запропонованого у підрозд. п.2.2 методу ідентифікації моделі з використанням запропонованого у цьому підрозділі методу оцінювання порядку Фур’є для опису багатотижневої сезонності.

Узагальнюючи викладене, математично запропонований метод можна записати таким чином. Запропоновано ряд співвідношень, у т.ч. емпіричних, яким чином можна оцінювати порядок n ряду Фур'є по 10% верхньої частини однієї хвилі і по ¼ періоду (тобто аналізувати достатньо тільки верхівку однієї половини хвилі, яку, як правило, завжди можна виокремити на графіку), а період P визначається по умовному початку і кінцю хвилі, які екстраполюються на основі мінімального значення кількості нових хворих на коронавірус y_{min} на початку хвилі, максимального значення хвилі на піку y_{max} та моментів часу для цих точок $t_{y_{min}}$ та $t_{y_{max}}$, відповідно:

$$n_{opt} = \varphi_n\left(y_{0.9max}, t_{y_{0.9max}}\right),\ y_{0.9max} = 0.9y_{max}, \tag{2.19}$$

$$P = \varphi_P(y_{min}, t_{y_{min}}, y_{max}, t_{y_{max}}), \tag{2.20}$$

де $t_{y_{0.9max}}$ – дата, коли крива перетинає рівень $y_{0.9max}$, який дорівнює 90% значення від максимального значення піку y_{max} хвилі, яка аналізується, у лівій (наростаючій) частині першої чверті її періоду; φ_n та φ_P – деякі функції, детально описані у статті [17].

На практиці вираз (2.19) не завжди дає гарні результати, якщо хвилі мають надто різний період – в таких випадках, порядок Фур'є краще задавати як гіперпараметр, який треба ідентифікувати. Оптимально, для окремих видів сезонності цей порядок варто оцінювати за (2.19), а для частини – робити гіперпараметром. Крім того, практичний досвід показав, що після визначення P за виразом (2.20) його ще можна уточнити, з урахуванням порядку Фур'є n_{opt}:

$$P_{opt} = \varphi_{nP}\left(n_{opt}, P\right), \tag{2.21}$$

де φ_{nP} – деяка емпірична функція, наприклад, в програмі [79] використано таке емпіричне співвідношення:

$$P_{opt} = P - 7(n - 3), \tag{2.22}$$

де n є гіперпараметром, а $P = 620$ діб.

У такий спосіб запропоновано визначати періоди усіх основних хвиль і далі або усереднювати і брати одну сезонність з довготривалим періодом, або враховувати їх усі окремо. У множині параметрів це дозволяє оцінити параметр P_{3opt}.

Проведені попередні дослідження для ряду кількості нових хворих в Україні дозволили звести кількість параметрів (2.5) на кожному з етапів (2.6) до трьох:

$$\text{Stage 1:} \begin{cases} R_g = R_{g0},\ n_2 = n_{20},\ R_{s2} = \frac{R_g}{k_{s2}},\ n_1 = n_3 = n_{30},\ R_{s3} = \frac{R_g}{k_{s3}}, \\ R_{s1} = 0.5,\ \Omega = 7,\ P_3 = 620, \\ R_h \to R_{hopt},\ t_0 \to t_{0opt},\ t_1 \to\ t_{1opt}, \end{cases} \tag{2.23}$$

$$\text{Stage 2:} \begin{cases} R_h = R_{hopt},\ t_0 = t_{0opt},\ t_1 =\ t_{1opt}, \\ R_{s1} = 0.5,\ \Omega = 7, \\ R_g \to R_{gopt},\ n_2 \to n_{2opt,}\ \ n_1 = n_3 \to n_{1opt} = n_{3opt}, \\ R_{s2} = \frac{R_g}{k_{s2}},\ R_{s3} = \frac{R_g}{k_{s3}}, P_3 = 620 - 7(n_3 - 3). \end{cases}$$

А отже, кількість варіантів N_M зі співвідношення (2.3) моделей звелась до такої [17]:

$$N_M = 2m^3. \tag{2.24}$$

За умови використання 4-х допустимих варіантів для кожного параметра, це складає 128 варіантів, що працює відносно швидко. З урахуванням візуалізації та збереження проміжних результатів у файли, програма [79] працювала у Kaggle, зазвичай, 6-20 хв., залежно від волатильності ряду (за більшої волатильності тривалість виконання, іноді, була більшою – у вересні 2021 р. працювала 1 годину і 47 хвилин).

Застосування моделей, побудованим за цим удосконаленим методом ідентифікації параметрів багатохвильової природи часового ряду кількості хворих чи померлих від коронавірусу в заданому регіоні дало суттєве зростання точності прогнозів, але з часом і ця модель перестала бути задовільною, тому було здійснено ще одне удосконалення.

2.4 Розширення методів на багатоітераційний випадок

Запропоновані вище методи та математичні вирази передбачають ідентифікацію параметрів в 1-2 етапи. Але, в загальному випадку, можливим є й їх багатоітераційне застосування. Формалізуємо цей процес.

Для цього, пропонуємо записати модель (2.1) у більш лаконічній формі [21]:

$$y_t = g(K_g, t) \cdot \prod_{i=1}^{\Phi} s(K_{si}, t) \cdot h\left(\bigcup_{j=1}^{\Psi} H_j\left(K_{hj}, t\right)\right) \cdot \epsilon_t \qquad (2.25)$$

або

$$y_t = f\left(K_g, K_{si}, K_{hj}\right), i = \overline{1, \Phi}, j = \overline{1, \Psi}, \qquad (2.26)$$

де K_{si} – параметри кожної сезонної складової (період, порядок ряду Фур'є, ступінь регуляризації); K_{hj} – параметри складових, які враховують вплив дат аномалій (зсув – тривалість від моменту зараження до моменту проявлення симптомів інфекції (для коронавірусу це – 4-7 діб), початок і кінець вікна впливу та ступінь регуляризації цього впливу).

Основною задачею є ідентифікація структури (визначення Φ і Ψ) та параметрів K_g, K_{si}, K_{hj} моделі (2). У разі повного перебору варіантів для пошуку оптимальної моделі це може бути надто довготривалий процес.

Як було запропоновано та пояснено вище, Φ визначається на основі розвідувального аналізу даних шляхом декомпозиції ряду на складові з різним періодом. А параметр Ψ вибирається на основі вивчення процесу протікання хвороби та факторів, які на неї можуть впливати (медичні дослідження, ретроспективні рішення, аналіз метеоданих, урядових рішень, Google-трендів тощо) – як правило, чим більше буде ідентифіковано і враховано обґрунтованих дат аномалій, тим точнішим буде розв'язок.

Ідентифікацію параметрів моделі пропонуємо здійснювати багатоітеративно паралельно-послідовно: спочатку, як і пропонувалось вище у підрозд. 2.2 та 2.3, визначаються параметри дат аномалій K_{hj}, задавши початкові значення K_{g0}, K_{si0}, потім – параметри тренду та сезонних складових K_g, K_{si}, а далі

пропонується на цьому не зупинятись і повторювати ці операції в циклі до досягнення достатньої точності, тобто наближеності до даних спостережень $J = F(y_t, \hat{y}_t)$ за цією моделлю або певного обмеження на кількість W таких циклів ($w = \overline{1, W}$):

$$\begin{gathered} J_{opt11} = F\left(f\left(K_{hj} \to K_{hjopt1}, K_{g0}, K_{si0}\right), \hat{y}_t\right), \\ J_{opt12} = F\left(f\left(K_{hjopt1}, K_g \to K_{gopt1}, K_{si} \to K_{siopt1}\right), \hat{y}_t\right), \ldots \\ J_{optw1} = F\left(f\left(K_{hj} \to K_{hjoptw}, K_{goptw-1}, K_{sioptw-1}\right), \hat{y}_t\right), \\ J_{optw2} = F\left(f\left(K_{hjoptw}, K_g \to K_{goptw}, K_{si} \to K_{sioptw}\right), \hat{y}_t\right). \end{gathered} \tag{2.27}$$

На рис. 2.21 наведено алгоритм застосування цього прийому для методу, запропонованого у підрозд. 2.2 (з урахуванням удосконалення, зробленого у підрозд. 2.3).

Для пошуку оптимальних значень параметрів варто використовувати відому техніку HyperOpt на основі методу байєсівської оптимізації, яка дозволяє швидко генерувати різні варіанти рішень і потім вибирати із них оптимальне значення.

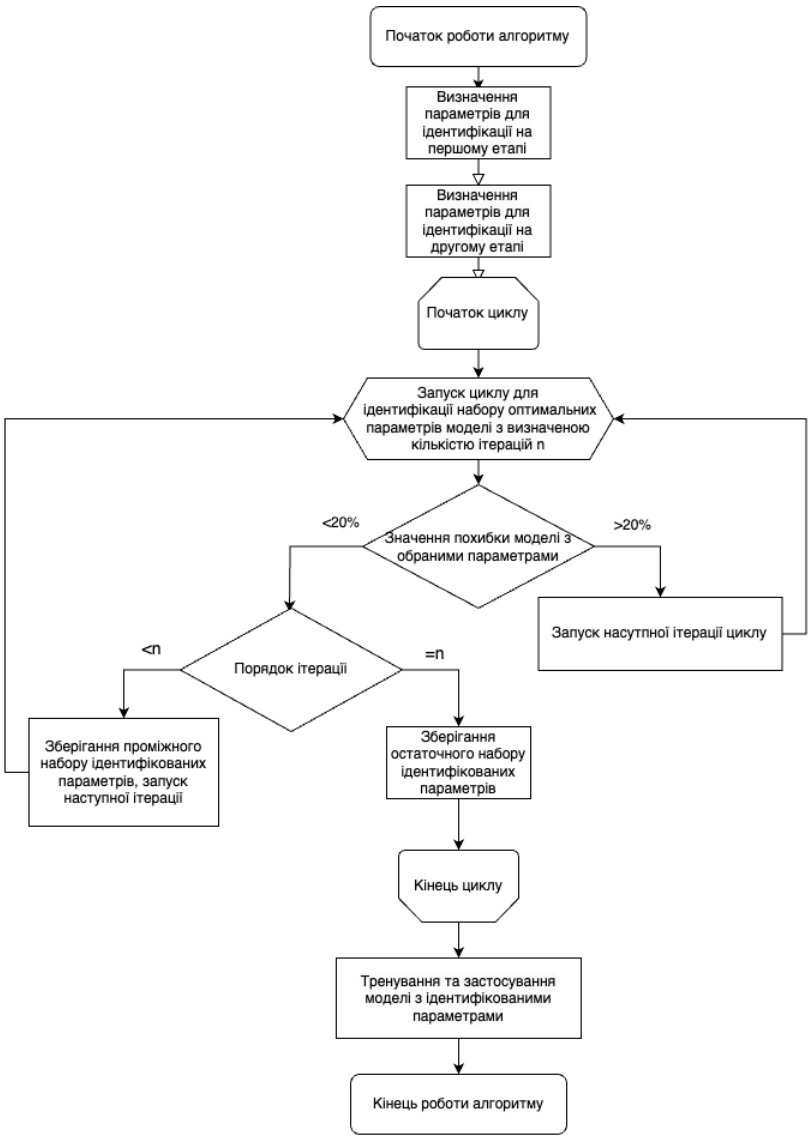


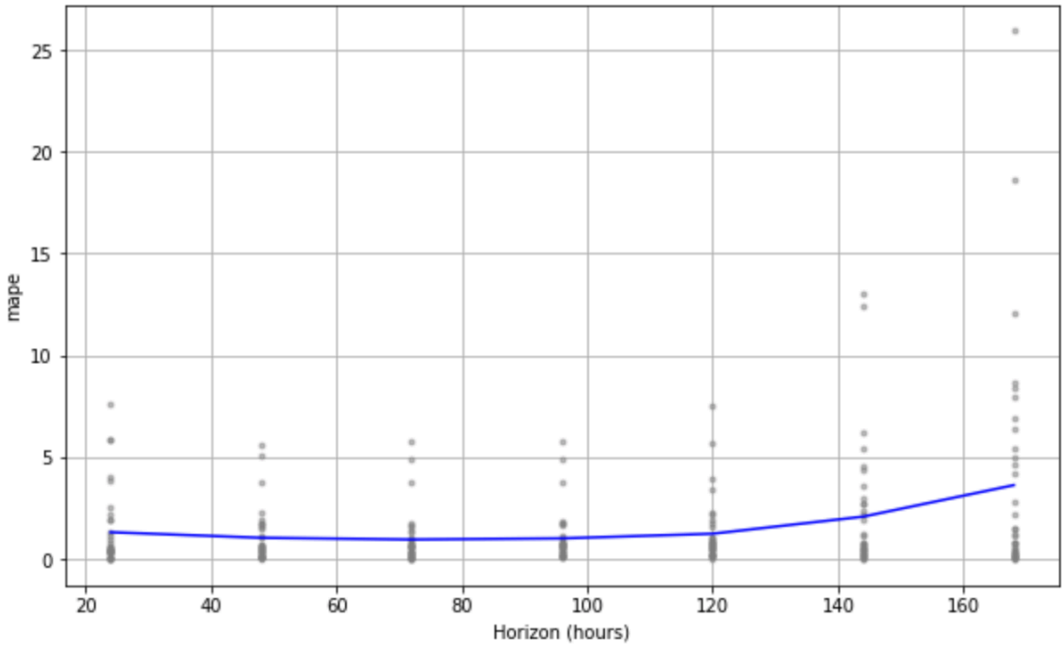
Рис. 2.21. Алгоритм методу ідентифікації параметрів моделі, удосконаленого шляхом введення багатоітераційності

У підрозд. 2.2 був реалізований подібний алгоритм побудови моделі за умови $W = 1$.

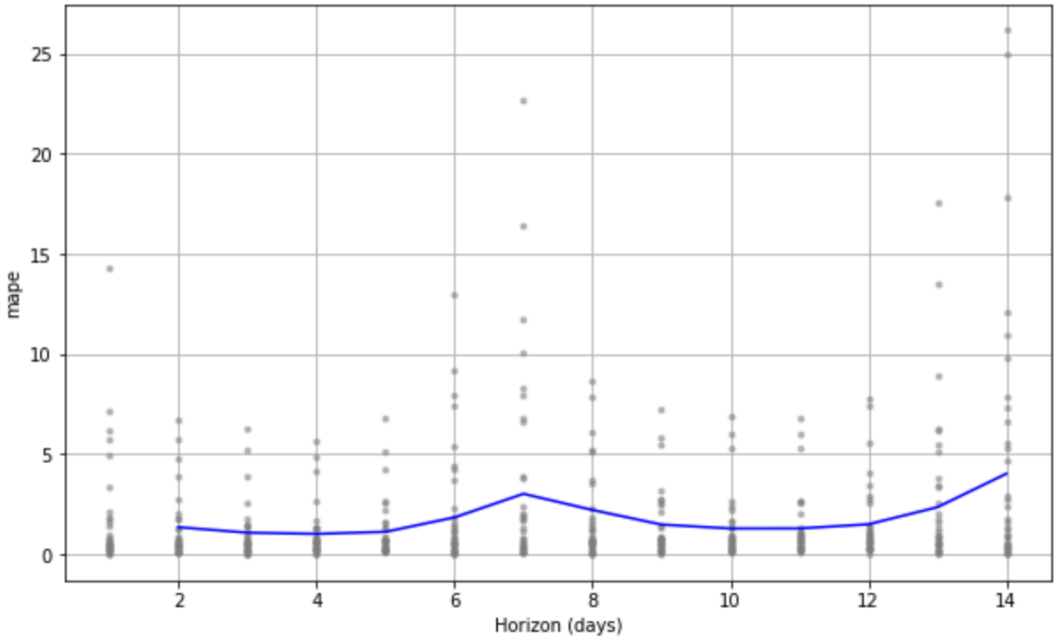
Таке удосконалення дозволяє підвищити точність прогнозування даних, навіть за умов суттєвої волатильності часового ряду.

При застосуванні даного методу не менш корисним буде здійснити діагностування часового ряду захворюваності, щоб отримати додаткову оцінку точності моделі при прогнозуванні з горизонтом в 7 або 14 днів, за допомогою

інструментів бібліотеки Facebook Prophet. Результати такого діагностування наведені на рисунку 2.22:



а)



б)

Рис. 2.22. Результати діагностування часового ряду захворюваності: а) за з горизонтом у 7 днів;

б) з горизонтом у 14 днів;

Дані результати свідчать про те, що горизонт прогнозування в 7 днів потенційно дасть меншу похибку прогнозування, ніж горизонт у 14 днів, але тим не менш горизонт в 14 днів є прийнятним, оскільки похибка в середньому знаходиться в межах 15%.

2.5 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 прийняття рішень. Наприклад, в задачі прогнозування за багато тижнів ряду 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 зі зсувом на певну кількість тижнів дозволить прогнозувати зростання кількості хворих в Україні. Традиційно для такої задачі роблять лінійний прогноз та картують нахил відповідних ліній. Але, як було показано вище, дані про кількість хворих на коронавірус у різних регіонах, наприклад країнах, є суттєво зашумленим з боку впливу регіональних факторів (свята, метеопатерни тощо), а це суттєво ускладнює співставлення прогнозів та пошук важливих закономірностей. Для уникнення цього, пропонується співставляти тренди прогнозів із результатів застосування моделей, побудованих з використанням наведених у підрозд. 2.2-2.4 методів. Обчислений прогноз з використанням побудованої за тими методами моделі FB Prophet дозволяє виокремити у прогнозі періодичні складові і складову, обумовлену впливом дат аномалій, та, окремо – основний тренд, ігноруючи регіональні особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень тощо), що, у свою чергу, дозволяє більш точно аналізувати основні закономірності динаміки процесу [18]. Математично, цей підхід виглядає таким чином. Прогноз моделі FB Prophet, за методологією роботи

їх алгоритму, складається з прогнозів усіх складових: $g(t), s_i(t), h(t)$. А тоді пропонується брати до уваги тільки $g(t)$ та апроксимувати шматково-лінійною функцією з кроком в 1 тиждень (загалом, може бути й інший інтервал узагальнення):

$$g\left(R_{gopt,}t_q\right) = b_g + k_g t_q, \tag{2.28}$$

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

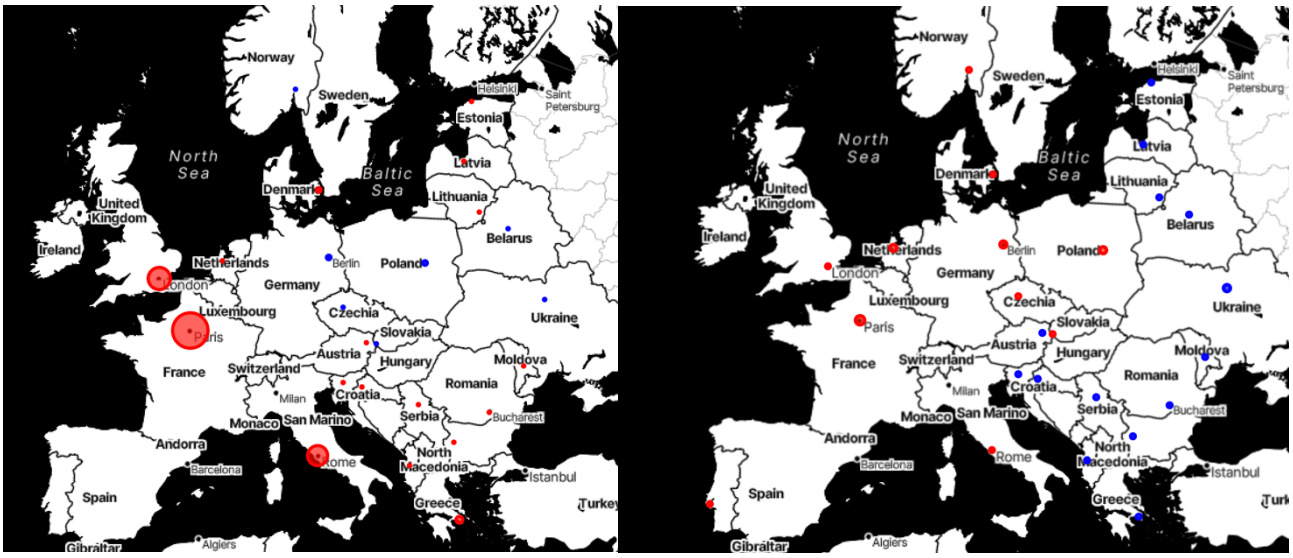
З (14) знаходимо:

$$k_g = \frac{g\left(R_{gopt,}t_q\right)}{t_q}. \tag{2.29}$$

Далі відображаємо цей нахил колами на карті. Наприклад, для країн світу (рис. 2.22) [80]:

- центр кола збігається з координатами столиці країни;
- колір кола береться червоним, якщо коефіцієнт (15) є додатнім (наростання кількості хворих на коронавірус) чи дорівнює 1, або – синім, якщо коефіцієнт є додатнім;
- радіус кола r_g лінійно пропорційний значенню k_g.

Розроблено програму «Картографування прогнозів зміни кількості нових хворих у країнах світу на основі трендів моделі Facebook Prophet» [80]. Результатом роботи цієї програми є інтерактивна карта, яка дозволяє візуально оцінити розповсюдження коронавірусу і прослідкувати напрямок розповсюдження захворювання протягом часу дослідження. Приклади створених на Python картограм наведені на рисунках 2.23–2.25.



а) б)

Рис. 2.23. Картограми прогнозів 10.01.2022 – 16.01.2022 за даними датасету Університету Джона Гопкінса: а) за формулою $r_g = 0.75 + 0.005k_g$; б) за формулою $r_g = 2 + 0.004k_g$ [80]

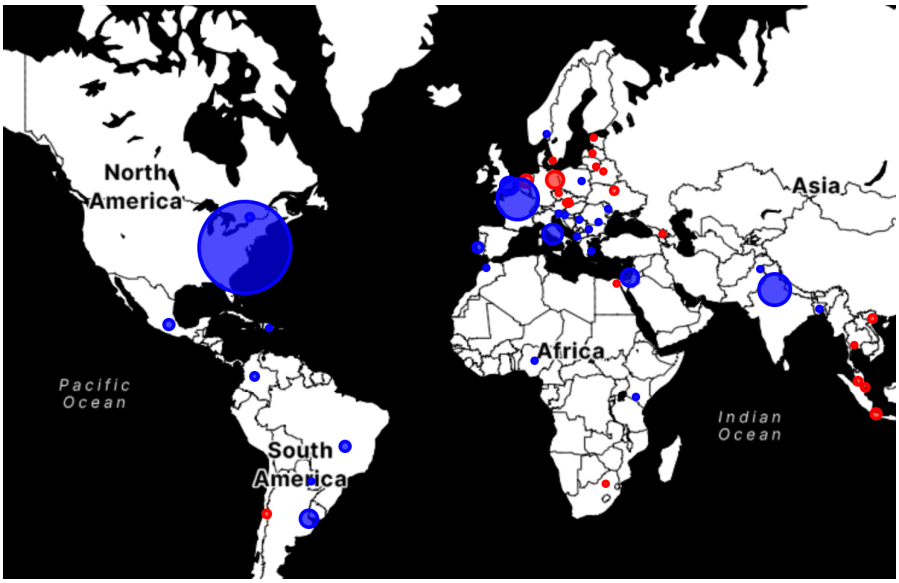


Рис. 2.24. Картограма трендів інфекційного впливу, нанесених на карту світу

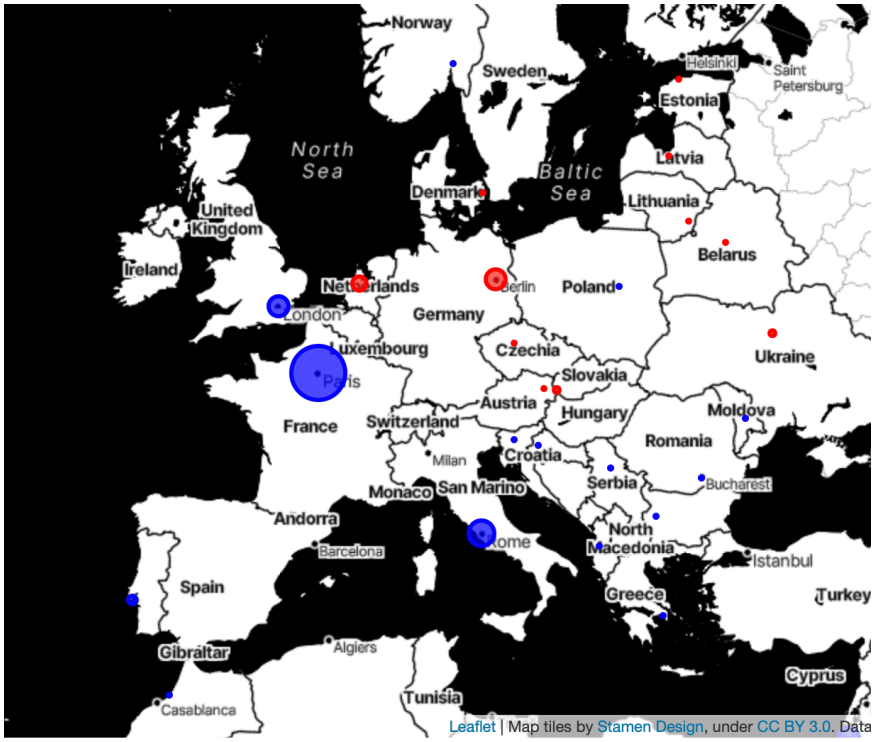


Рис. 2.25 Картограма трендів інфекційного впливу, нанесених на карту Європи

Побудований датасет нахилів прогнозів з ноутбуками [81], який може бути використаний і для інших задач, наприклад, як у статті [9], де було досліджено задачу моніторингу захворюваності на COVID-19 під впливом зміни захворюваності в інших країнах.

Крім того, по нахилах за декілька послідовних прогнозів, наприклад по прогнозах за ряд послідовних тижнів) можна легко визначити в який тиждень мав місце початок, пік (чи «плато») та кінець хвилі. А тоді до таких значень по багатьох сусідніх регіонах можна застосувати математичний апарат синтезу стійких когнітивних карт, розроблений Дратованим М.В. та Мокіним В.Б. у статті [наша з тобою по КК], що дозволить здійснювати завчасне прогнозування цих дат в одних країнах – по інших, де динаміка схожа, але хвилі проходять раніше.

2.6 Висновки до розділу

Розділ присвячений розробленню методів та теоретичних основ технології прогнозування кількості хворих на коронавірус.

Здійснено розвідувальний аналіз часового ряду кількості хворих на коронавірус в Україні. Причому, оскільки основні теоретичні результати роботи були отримані у 2020 р., то й аналіз проводився за даними саме 2020 р. (з 1 квітня по 6 грудня 2020 р. – діапазон даних, використаний у статтях 2020 року). Звичайно, методологія аналізу легко може бути поширена і на інші діапазони дат – усі результати наведені з публічних Python-ноутбуків в Kaggle, де співавторами є здобувач і його науковий керівник.

Проведено аналіз стаціонарності ряду за критерієм Діка-Фуллера, який показав, що тільки друга різниця ряду є стаціонарною, а сам ряд і його перша різниця є нестаціонарними.

Здійснено декомпозицію ряду на шматково-лінійний тренд (Trend), сезонну складову (Seasonal) із заданим періодом та залишки (Resid). Досліджено як змінюється частка сезонної складової, залежно від періоду у добах, зі згладжуванням у 7 днів, який показав, що:

1) частка сезонних складових очікувано монотонно зростає зі збільшенням періоду;

2) перший стрибок має місце на 1 тиждень, що доволі очевидно, бо це – цикл роботи лабораторій, які проводять тестування;

3) в подальшому теж мають місце стрибки, але не чітко виражені і це питання потребує окремого дослідження, з урахуванням більшої кількості даних.

Побудовано автокореляційну та часткову автокореляційні функції. Аналіз показав, що їх значення спадають різко, але потім довго ще не залишаються повністю у зоні статистичної незначущості. Найбільш швидко спадає ЧАКФ для першої різниці, але і там є певні коливання. Це показує, що ряд має складну стохастичну природу та різні сезонності.

Побудовано ряд моделей машинного навчання:

- модель ARIMA(5,1,4);

- моделі FB Prophet з аномаліями та ідентифікацією варіантів періодичності у [4, 7, 30, 365] діб з порядком 3 або 12 рядів Фур'є в усіх комбінаціях;

- моделі машинного навчання з використанням бібліотеки Tsfresh, за якою відібрано 77 інформативних статистичних ознак, із 1200, синтезованих по значеннях основному ряду.

Аналіз точності за 3 різними метриками показав, що найбільш точною за усіма ними є модель FB Prophet.

Отже, етапи розвідувального аналізу та вибору моделі показали, що для досягнення поставленої у роботі задачі:

1. Варто будувати інформаційну технологію на базі моделі FB Prophet.
2. Варто більш ретельно визначати дати аномалій та вікно їх впливу.
3. Варто ідентифікувати більше параметрів моделі, для чого необхідно розробити більш досконалі методи ідентифікації, ніж простий перебір варіантів усіх можливих значень, оскільки це займатиме забагато часу.

Відповідно до цих пропозицій, запропоновано і продемонстровано на прикладах розширення множини дат аномалій шляхом урахування офіційних свят, умов метеорологічних аномалій та дат зміни сумарного значення урядового трекера SI Оксфордської лабораторії.

Загалом, проведений аналіз даних, аналіз аналогів та спроби розв'язання поставленої задачі з використаннях відомих методів і технологій, дозволили сформулювати основні висновки, які в подальшому були використані для створення нових методів і нової інформаційної технології:

1. Аналіз впливу різних факторів та можливості прогнозування поширення ковіду, зокрема Google-трендів щодо пересувань жителів України, метеоданих, даних Оксфордського трекера коронавірусної діяльності урядів країн світу та ознак, отриманих на їх основі, показав, що ці ознаки впливають нерівномірно. Отже, їх краще враховувати не як додатковий регресор для побудови множинної регресії, а – на їх основі згенерувати вторинні ознаки, що буде більш ефективним.

2. Аналіз матеріалів РГ по ковіду показав, що мають місце значні затримки в оприлюдненні результатів тестів (під час піків хвиль - більше 30 днів) і вони мають нерівномірний характер у часі, а отже, за цими даними важко буде здійснювати довгострокове прогнозування і варто обмежитись короткостроковим прогнозуванням з горизонтом в 1-2 тижні.

3. Доведено, що ряд кількості нових хворих на коронавірус в Україні та його перша різниця є нестаціонарними, а стаціонарною є тільки його друга різниця, що варто враховувати під час моделювання методами машинного навчання.

4. Декомпозиція ряду з різним періодом сезонності показали, що має місце тижнева періодичність і мінімум 1-2 з іншим періодом, але це потребує окремого дослідження.

5. Аналіз різних видів аномалій, які впливають на кількість хворих на коронавірус, показав, що їх слід враховувати з різним запізненням та різним вікном (початок і кінець впливу у добах) і силою впливу, що потребує окремого дослідження, тобто це – параметри, які потребують ідентифікації.

6. Порівняльний аналіз відомих методів та моделей для прогнозування кількості хворих на коронавірус (ARIMA, Facebook Prophet та інших моделей машинного навчання на основі дерев рішень, регресій та ін.) довів, що більш точною є модель Facebook Prophet, отже варто ідентифікувати саме її структуру та параметри.

7. Модель Facebook Prophet з урахуванням декількох видів сезонних складових, загалом, має десятки параметрів. Повний перебір усіх комбінацій значень для пошуку оптимального варіанту займе забагато часу, а тому потрібні більш швидкі методи.

Запропоновано метод оцінювання порядку ряду Фур’є багатохвильового періодичного процесу, який відрізняється від існуючих тим, що оцінювання здійснюється лише по 10% верхівки однієї хвилі з використанням емпіричних співвідношень, що дозволяє здійснювати таке оцінювання, навіть, за умов, коли

одна хвиля перекриває іншу і їх важко відокремити одну від іншої. 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.

Запропоновано метод паралельно-послідовної багатопараметрової ідентифікації моделі Facebook Prophet для короткострокового прогнозування часового ряду кількості хворих на коронавірус в заданому регіоні, який відрізняється від існуючих більшою кількістю параметрів, що ідентифікуються: сила і розмір вікна впливу дат аномалій, ступінь регуляризації і тип моделі (адитивна чи мультиплікативна), порядок Фур'є і ступінь регуляризації 3-х різних періодичних складових, які враховують внутрішньотижневі, тижневі і багатотижневі закономірності, характерні для інфекційних хвороб, у т.ч. коронавірусу, що дозволяє суттєво підвищити точність прогнозів та більш глибоко дослідити закономірності, які впливають на цей часовий ряд. Запропоновано варіант цього методу з багатоітераційною оптимізацією параметрів, який дозволяє підвищити його точність.

Подальшого розвитку отримав метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах, який відрізняється від існуючих використанням в якості інтегрального показника прогнозу на тиждень нахилу тижневої ділянки шматково-лінійної апроксимації тренду цього прогнозу, отриманого за моделлю, ідентифікованою із застосуванням запропонованого методу паралельно-послідовної багатопараметрової ідентифікації, що дозволяє аналізувати закономірності міжрегіонального поширення хвороби, ігноруючи місцеві особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень, тощо). Ці нахили, також, можна кластеризувати, а потім за ними побудувати байєсівську мережу для виявлення закономірностей – здобувач має такий досвід з оброблення медичних даних іншого типу.

Запропоновано застосовувати ідентифіковану за цією технологією модель для прогнозування найбільш песимістичного та найбільш оптимістичного

сценаріїв розвитку явища, тобто зміни кількості нових підтверджених випадків хвороби на коронавірус «COVID-19» у майбутньому у заданій країні.

Наведено алгоритми цих методів та деякі приклади їх застосування на практиці.

РОЗДІЛ 3
РОЗРОБЛЕННЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС

Відповідно до поставлених задач та з урахуванням наведених у розділі 2 методів ідентифікації параметрів моделі розробимо інформаційну технологію прогнозування кількості хворих на коронавірус у заданому регіоні на прикладі України.

3.1 Проєктування структури інформаційної технології

В цілому, процес роботи інформаційної технології з прогнозування часового ряду повинен складатись з етапів оброблення даних, самого прогнозування та аналізу результатів прогнозування. Розіб'ємо кожний з цих етапів на окремі самостійні модулі, згруповані у блоки (рис. 3.1), та опишемо їхні функції та вхідні і вихідні дані більш детально.

Звертаємо увагу, що, оскільки результатом роботи має бути картограма для аналізу закономірностей для N_R усіх регіонів, то етапи ідентифікації параметрів та прогнозування даних теж слід проходити N_R разів, тобто для кожного регіону окремо.

I. Блок «Оброблення первинних даних». Може застосовуватись як один раз для усіх N_R регіонів одразу, так і для кожного регіону окремо.

1. У модулі оброблення набору даних здійснюється першочергове оброблення набору даних із датасету первинних даних захворюваності на коронавірус. Також у ньому здійснюється оброблення типів даних, що містяться у датасеті, і його очищення від пустих та неактуальних даних. На виході цього модулю – датасет оброблених даних, який використовується наступним блоком та його модулями.

2. Модуль розвідувального аналізу даних реалізує глибший аналіз датасету. Аналізується інформація про захворюваність в тих чи інших країнах світу, метеодані тощо. В результаті такого аналізу отримуємо список дат аномалій та першие наближенням їх параметрів, а також – вхідні дані про хвилі для застосування методу оцінювання порядку ряду Фур'є для опису багатотижневої сезонності. Результати роботи цього модулю, також, зберігаються у датасеті оброблених даних.

Здійснюється візуалізація роботи обох модулів для ілюстрування виявлених закономірностей та обґрунтування вибору початкових наближень параметрів.

II. Блок «Ідентифікація параметрів та прогнозування».

3. У модулі ідентифікації параметрів застосовується метод оцінювання порядку ряду Фур'є для опису багатотижневої сезонності за даними модулю розвідувального аналізу даних, а потім у взаємозв'язку з модулем тренування моделі здійснюється побудова оптимальної моделі з використанням запропонованого методу багатоітераційної багатопараметрової ідентифікації параметрів моделі FB Prophet з багатьма видами сезонності (підрозд. 2.2-2.4). Результатом є модель на основі FB Prophet з оптимальною структурою (видами сезонності та аномалій) і параметрами.

4. Модуль прогнозування здійснює прогнозування кількості хворих на строк від 7 до 14 днів з використанням оптимальної моделі, побудованої у модулі тренування моделі, в залежності від запиту та епідеміологічної ситуації. Результатом є як графіки прогнозів, так і – складові цього прогнозу (тренд, кожна сезонність – окремо та складова із впливом аномалій).

Цей блок запускається N_R разів, тобто для кожного регіону – окремо. Усі результати прогнозів щоразу зберігаються у датасеті згенерованих прогнозів.

Здійснюється візуалізація прогнозів та їх складових.

III. Блок «Визначення нахилів трендів прогнозів для їх аналізу».

5. Модуль визначення нахилів трендів реалізовує метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах з обчисленням нахилу тижневої ділянки шматково-лінійної апроксимації тренду прогнозу для кожного регіону із датасету згенерованих прогнозів.

6. Модуль генерації картограм. За обчисленими у попередньому модулі нахилами для заданих дат і регіонів визначаються радіуси і колір відповідних кругових діаграм та їх координаті на карті.

Результати модулів цього блоку відображаються у різний спосіб:

- тренди та нахили їх лінійної апроксимації – у вигляді графіків, крім того, усі ці значення зберігаються у датасеті тенденцій;

- картограми нахилів – у вигляді карт, у т.ч. інтерактивних з різними шарами.

7. Модуль аналізу трендів регіонів здійснює імпорт нахилів трендів та їх подальше оброблення одним із відомих способів. Наприклад, як було зазначено в кінці підрозд. 2.5 та показано у статті [20], можна визначити в який тиждень мав місце початок, пік (чи «плато») та кінець хвилі, а тоді по когнітивній карті здійснити завчасне прогнозування цих дат в одних країнах – по інших, де динаміка схожа, але хвилі проходять раніше.

8. Модуль візуалізації містить основні способи візуалізації результатів роботи різних блоків.

9. Модуль зберігання результатів компонує всі отримані результати та зберігає їх у відповідних форматах: прогнозовані дані – у форматі CSV, графіки – у форматі TIFF, а картограми – у форматі SVG чи ін. Таким чином, результати роботи інформаційної технології можуть бути надані для подальшого дослідження, або презентування в якості аналітичного звіту чи публікації.

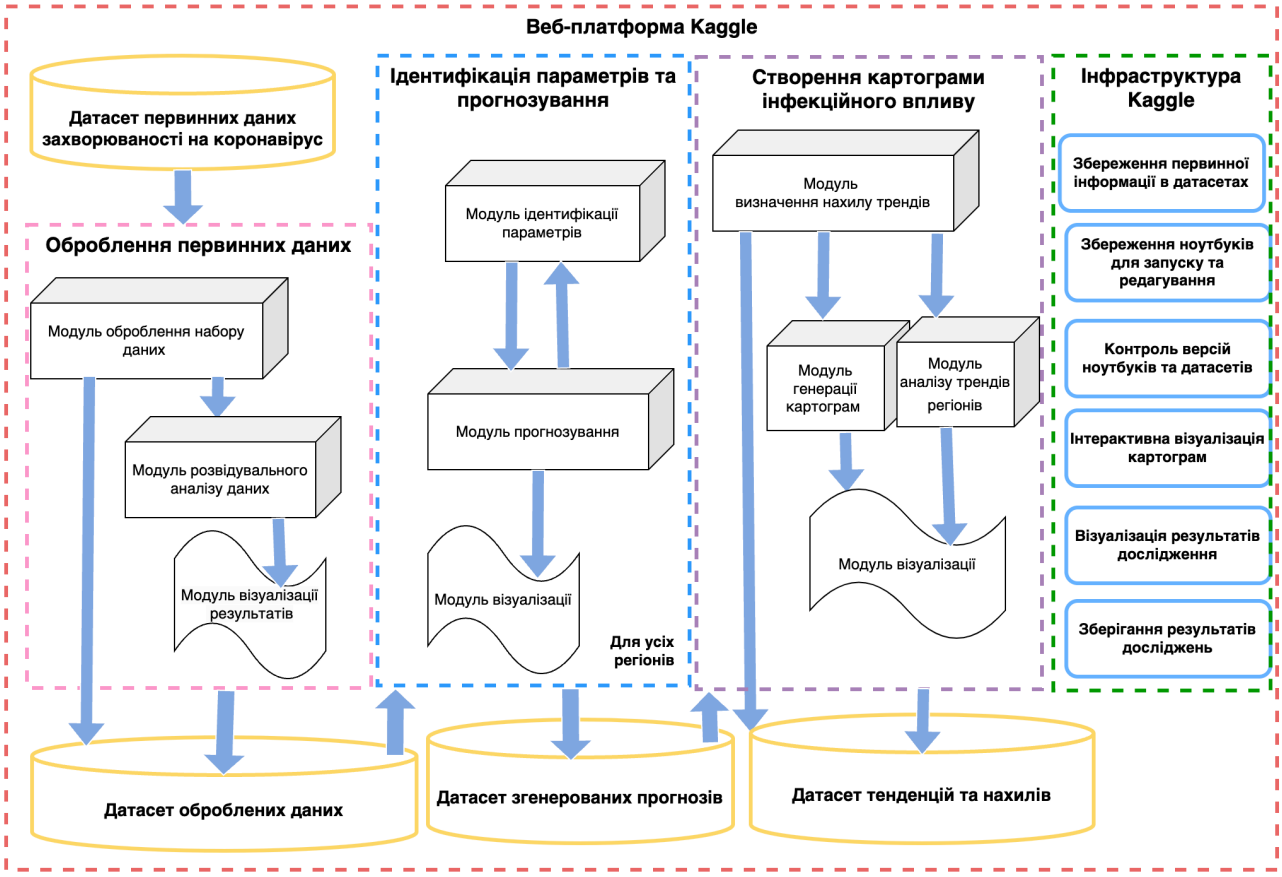


Рис. 3.1 Структура розробленої інформаційної технології

Для реалізації розробленої інформаційної технології було вирішено скористатись популярною веб-платформою Kaggle, яка дозволяє здійснювати роботу над інформаційною технологією в режимі онлайн, за допомогою програмного середовища Jupyter Notebook, адаптованого для використання в межах цієї веб-платформи. Дане програмне середовище зберігає програмний код у форматі інтерактивних «ноутбуків», які зберігаються на Kaggle, при цьому кожна нова версія ноутбуку зберігається окремо, і, у разі потреби, можна обрати конкретну його версію. Також даний веб-сервіс надає хмарні ресурси для запуску програмного коду, що дає можливість ефективно тестувати та тренувати модель в основі даної ІТ. Крім цього, даний веб-ресурс надає доступ до різноманітних

наборів даних у зручному форматі, що спрощує задачу їх зчитування. В цілому, Kaggle надає весь спектр інструментів, потрібних для продуктивної розробки інформаційних технологій в сфері машинного навчання та аналізу даних.

Важливою перевагою програм-ноутбуків редактору Kaggle Code є їх універсальність. Їх легко можна перенести й в інше середовище:

- AWS SageMaker;
- Google Colab;
- Jupyter Notebook або JupyterLab пакету Anaconda;
- Microsoft Azure Notebooks;
- PyCharm;
- Visual Studio Code

та інші популярні засоби роботи з Python. Головною відмінністю буде тільки вирішення питання де зберігати первинний і проміжні датасети та перепідключити функції зчитування і збереження даних до них у модулі зберігання результатів досліджень.

3.2 Розроблення UML-діаграм інформаційної технології

3.2.1 UML-діаграма послідовності роботи модулів технології

Врахувавши вище наведену інформацію, побудуємо UML-діаграму послідовності роботи модулів інформаційної технології, яка пропонується:

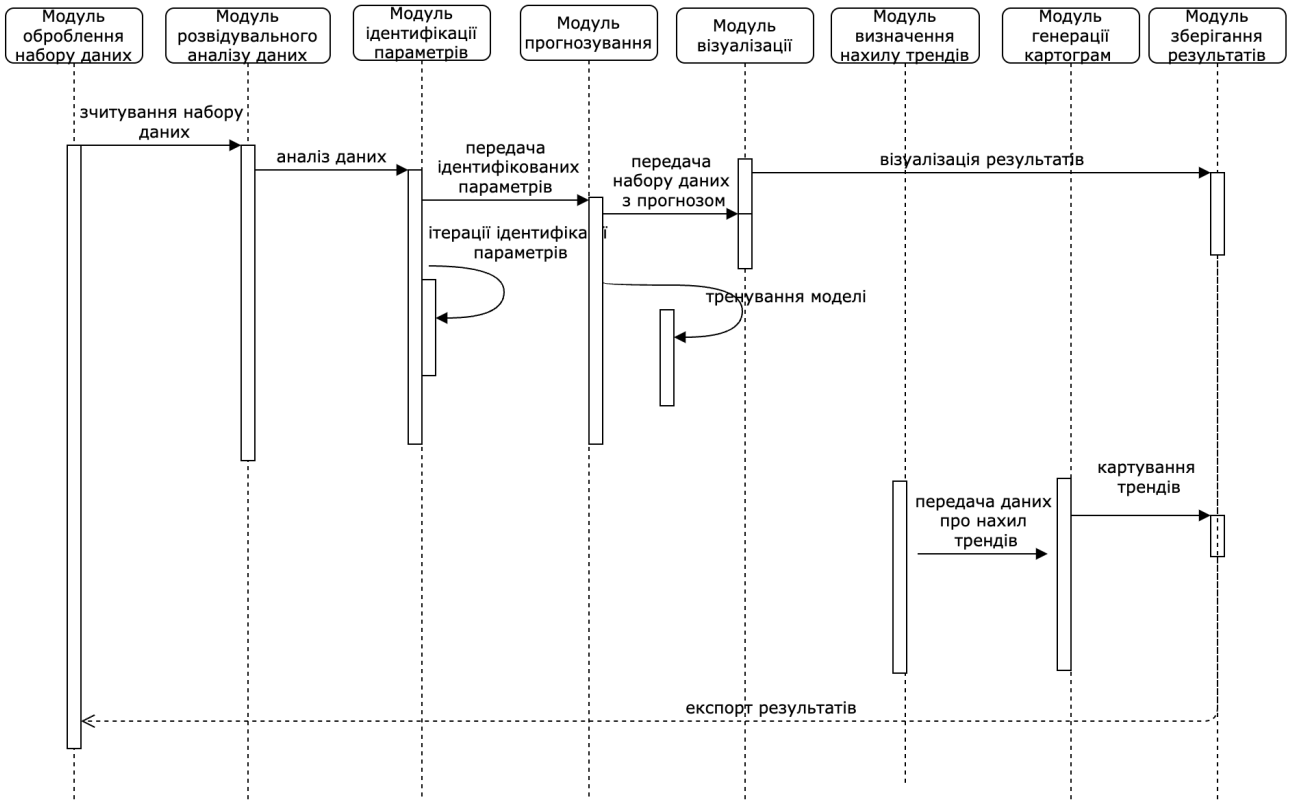


Рис.3.2 UML-діаграма послідовності роботи модулів інформаційної технології

Рис. 3.2 враховує наступну хронологію процесів:

- першим спрацьовує модуль оброблення набору даних, в ньому здійснюється зчитування набору даних та перетворення у формат датафрейму, за допомогою Python-бібліотеки Pandas;

- далі, в модулі розвідувального аналізу даних відбувається додатковий аналіз датасету, результатом якого є отримання списку дат аномалій та першим наближенням їх параметрів, а також – вхідні дані про хвилі для застосування методу оцінювання порядку ряду Фур’є для опису багатотижневої сезонності;

- у модулі ідентифікації параметрів застосовуються методи, описані в підрозділах 2.2-2.4;

- модуль прогнозування обчислює прогноз за натренованою у попередньому модулі моделлю, який в подальшому використовується в наступних модулях;

- модуль візуалізації результатів використовується для візуалізації прогнозу, а також генерування порівняльного графіка з прогнозами, зроблених в минулому;

- у свою чергу, в модулі визначення трендів відбувається відокремлення трендів захворюваності в країнах певного регіону, після чого дані про ці тренди передаються на наступний модуль;

- далі, в модулі генерації картограм здійснюється візуалізація отриманих трендів в інтерактивній картограмі, приклади якої наведені в підрозділі 2.5;

- після відпрацювання усіх модулів всі результати дослідження зберігаються за допомогою модулю зберігання результатів, який зберігає результати прогнозування, картограми, та графіки у відповідних форматах.

3.2.2 UML-діаграма діяльності блоку ідентифікації параметрів та прогнозування

Як зазначено вище, блок ідентифікації параметрів та прогнозування реалізує методи, описані у підрозд. 2.2-2.4, та за оптимальною моделлю здійснює прогнозування даних для заданого регіону на задану кількість діб, а потім – візуалізацію та збереження цього прогнозу чи прогнозів (наприклад, на 7 і 14 діб).

Як зазначено у підрозділі 2.3, в загальному випадку, пропонується здійснювати налаштування таких параметрів моделі FB Prophet у певній послідовності у два етапи, базуючись на гіпотезі про те, що вплив тренду і періодичних складових ряду характерно відрізняється від впливу окремих аномалій у певні моменти часу [17]:

1. розмір вікна, сила впливу (масштаб), режим (мультиплікативний чи адитивний) урахування та ступінь регуляризації значень в аномальні дати (у т.ч. свята та дати локдаунів);

2. мультиплікативність чи адитивність урахування, ступінь регуляризації та кількість коефіцієнтів ряду Фур’є для опису тижневої (7-денної) сезонності або n-денної сезонності;

Такий паралельно-послідовний підхід дозволяє отримати діапазони наближених значень параметрів, які вже можна застосовувати для тренування моделі, але слід також провести оптимізацію цих значеннь за допомогою багатоітераційного методу ідентифікації, задля подальшого зменшення похибки при тренуванні моделі.

Все описане вище можна описати за допомогою UML-діаграми діяльності блоку ідентифікації параметрів та прогнозування:

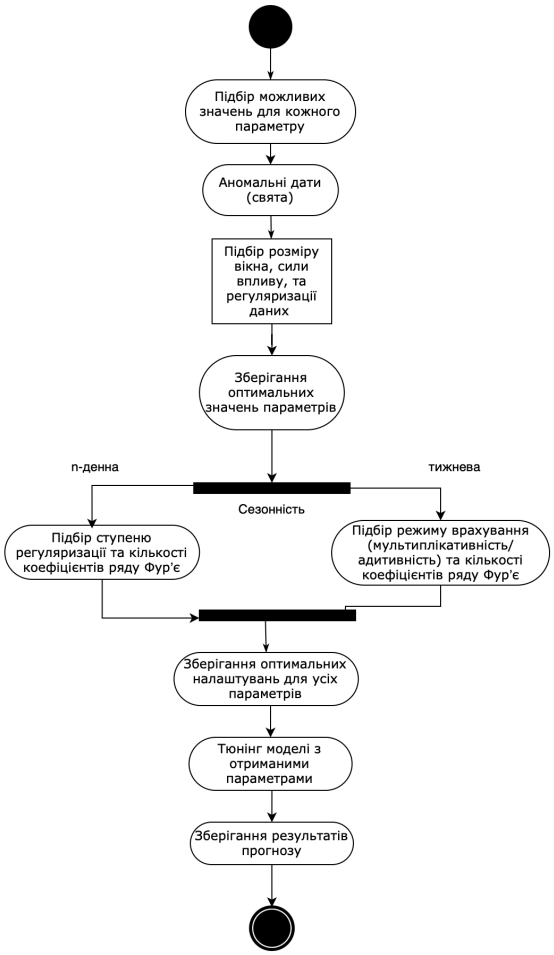


Рис. 3.3 UML-діаграма діяльності блоку ідентифікації параметрів та прогнозування інформаційної технології

3.2.3 UML-діаграма діяльності блоку створення картограми інфекційного впливу

Як зазначено вище, блок створення картограми інфекційного впливу реалізує метод, описаний у підрозд. 2.5, у т.ч. обчислення нахилів трендів, визначення способу їх візуалізації (колір і радіус кругової діаграми та координати її центру на карті для кожного регіону) і здійснює цю візуалізацію за заданий період для заданих користувачем регіонів (країн або областей чи штатів країни тощо).

Блок створення картограм інфекційного впливу потребує детальнішного розгляду, оскільки в собі він містить модулі, що виконують різні задачі. Опишемо ці модулі детальніше.

1. Результати роботи модулю прогнозування відфільтровуються, та створюється вибірка даних по країнах, для яких потрібно побудувати картограму.

2. Модуль обчислення нахилу тренду інфекційного впливу здійснює опрацювання набору даних, відфільтрованого в попередньому модулі, та здійснює апрокисмацію трендів лініями та обчислює їх нахил, після чого генерує новий набір даних, що містить в собі значення нахилів інтерпольованих ліній трендів інфекційного впливу для кожного кроку (як правило, 1-2 тижні), для кожної країни, що досліджується.

3. Далі, отримані дані про нахили трендів наносяться на графік в модулі візуалізації нахилів трендів інфекційного впливу, даючи можливісь наочно дослідити зміни тренду протягом часу дослідження.

4. Наступним модулем вже є модуль побудови картограми, на якому, в залежності від набору країн, що досліджуються, наносяться тренди або на карту світу, або на карту Європи. На карту наносяться одразу позитивні та негативні тренди, при чому позитивні треди кодуються кругами синього кольору, а

негативні – червоного. Радіус нанесених кругів відповідає абсолютному значенню куту нахилу лінії, якою інтрепольовано тренд за один крок.

Враховуючи все описане вище, була побудована UML-діаграма діяльності блоку створення картограми інфекційного впливу наведена на рисунку 3.4:

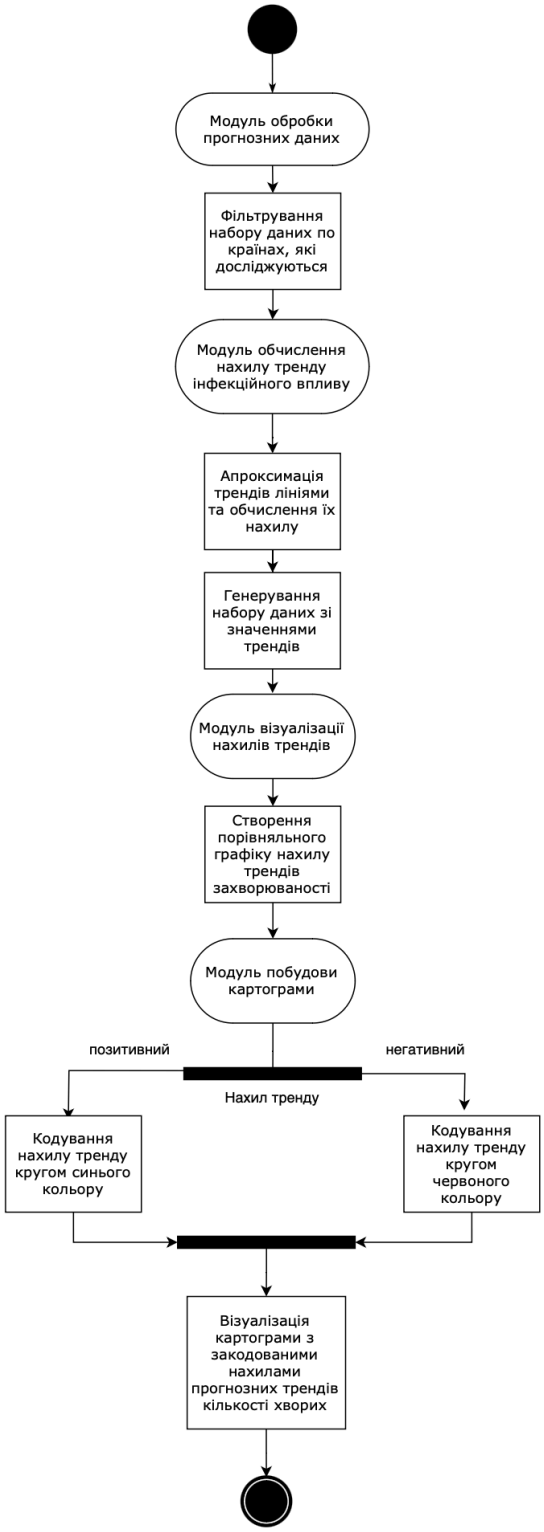


Рис. 3.4 UML-діаграма діяльності блоку створення картограми інфекційного впливу

3.3 Порівняння ефективності створеної інформаційної технології з аналогами

Упродовж співпраці з РГ по ковіду протягом 2020-2022 рр. були зроблені численні удосконалення в моделі, яка використовувалась для прогнозування захворюваності на COVID-19, причому, в залежності від типу найбільш поширених в Україні штамів, на момент моделювання, створювались окремі моделі, як, наприклад у випадку зі штамом Омікрон або Дельта. Команда вчених НАН України у складі РГ по ковіду, у свою чергу, використовувала спеціально розроблену для України згадану у пункті 1.2.1 модель SEIR-U, результати якої також додавались в аналітичні звіти протягом 2020-2022 рр. В кожному аналітичному звіті, в який включені результати цього дисертаційного дослідження, містилась порівняльна таблиця прогнозів моделей – приклад такої таблиці наведений на рисунку 3.5.

Дата	Середні прогнозні значення статистичної моделі Prophet	Середні прогнозні значення компартментної моделі SEIR-U
22.12.2021	5937	6731
23.12.2021	6752	8136
24.12.2021	6311	8066
25.12.2021	4965	7308
26.12.2021	3433	4687
27.12.2021	1739	2092
28.12.2021	3607	4609
29.12.2021	4131	5558
30.12.2021	4468	6603
31.12.2021	4084	6486
1.01.2022	3570	5850
2.01.2022	2166	3750
3.01.2022	1357	1478
4.01.2022	2027	3709

Рис. 3.5 Приклад порівняльної таблиці результатів прогнозування моделей Prophet та SEIR-U, взятий з аналітичного звіту «Прогноз РГ-58»

Для порівняння динаміки точності прогнозування обох моделей внесемо значення точності у порівняльну таблицю, використовуючи критерій WAPE, що використовувася для оцінювання точності моделі Prophet та був описаний в підрозділі 2.2. Для порівняння використаємо дані про захворюваність з відкритих джерел, а саме набір даних Університету Джона Гопкінса [82], похибка ж обчислюватиметься на основі прогнозів, наведених у звітах, таким чином порівняння відбуватиметься у рівних умовах. В ряді випадків розроблені за запропонованою технологією моделі дали кращі прогнози, аніж прогнози за моделлю SEIR-U (таблиця 3.1).

Таблиця 3.1 Порівняння точності прогнозування за запропонованою ІТ на базі моделі FB Prophet та моделі SEIR-U вчених із НАН України на прикладі даних, опублікованих в аналітичних звітах Робочої групи НАНУ

№ Звіту	Дати прогнозу	WAPE-критерій SEIR-U	WAPE-критерій Prophet
РГ-62	23.02.2022-08.03.2022	87,14	131,2
РГ-61	09.02.2022-22.02.2022	36,39	29,93
РГ-60	27.01.2022-08.02.2022	36,22	32,95
РГ-59	12.01.2022-25.01.2022	63,60	49,61
РГ-58	**22.12.2021-04.01.2022**	**34,09%**	**24,5%**
РГ-57	08.12.2021-21.12.2021	25,55%	28,54%
РГ-56	**24.11.2021-07.12.2021**	**23,21%**	**19.51%**
РГ-55	10.11.2021-23.11.2021	24,28%	26,57%
РГ-54	**27.10.2021-9.11.2021**	**23,43%**	**21,95%**
РГ-53	**13.10.2021-26.10.2021**	**24,01%**	**21,86%**

Таблиця 3.1 Порівняння точності прогнозування за запропонованою ІТ на базі моделі FB Prophet та моделі SEIR-U вчених із НАН України на прикладі даних, опублікованих в аналітичних звітах Робочої групи НАНУ

№ Звіту	Дати прогнозу	WAPE-критерій SEIR-U	WAPE-критерій Prophet
РГ-52	29.09.2021-12.10.2021	21,61%	27,67%
РГ-51	15.09.2021-27.09.2021	27,81%	33,52%
РГ-42	**21.04.2021-03.05.2021**	**32,63%**	**24,17%**
РГ-41	07.04.2021-19.04.2021	24,1%	43,39%
РГ-40	24.03.2021-05.04.2021	25,61%	30,93%
РГ-39	11.03.2021-22.03.2021	22,22%	23,28%
РГ-38	23.02.2021-1.03.2021	17,75%	18,41%
РГ-37	09.02.2021-15.02.2021	18,1%	19,9%
РГ-36	26.01.2021-01.02.2021	13,73%	18,95%
РГ-35	12.01.2021-18.01.2021	15,11%	21,85%
РГ-34	29.12.2020-11.01.2021	19,61%	30,75%
РГ-32	15.12.2020-28.12.2020	11,65%	22,12%
РГ-31	**8.12.2020-21.12.2020**	**17,87%**	**9,21%**
РГ-30	01.12.2020-13.12.2020	30,81%	32,47%
РГ-29	**24.11.2020-06.12.2020**	**21,29%**	**15,27%**

Як бачимо із результатів проведеного ретроспективного аналізу точності моделей, запропонована технологія на основі моделі Prophet в порівнянні з моделлю SEIR-U дає іноді точність, у 1,2-2 рази кращу, ніж модель SEIR-U.

3.4 Висновки до розділу

У третьому розділі розроблена структура запропонованої інформаційної технології прогнозування часових рядів кількості хворих на коронавірус методами машинного навчання. Описані основні складові інформаційної технології.

Побудована та охарактеризована UML-діаграма послідовності роботи модулів технології та охарактеризовано основні етапи її роботи. Побудована та охарактеризована UML-діаграма діяльності блоку ідентифікації параметрів та прогнозування. Сформована та охарактеризована UML-діаграма діяльності блоку створення картограми інфекційного впливу .

Проведено порівняння точності прогнозування моделі Prophet, що ідентифікувалась протягом 2020-2022 рр. з використанням результатів даного дослідження, та моделі SEIR-U, розробленої командою вчених НАН України у складі Робочої групи з математичного моделювання проблем, пов’язаних з епідемією коронавірусу SARS-CoV-2 в Україні. Усі результати скомпоновані у порівняльну таблицю. Для порівняння був використаний критерій WAPE. Аналіз показав, що запропонована технологія на основі моделі Prophet в порівнянні з моделлю SEIR-U дає іноді точність, у 1,2–2 рази кращу, ніж модель SEIR-U.

РОЗДІЛ 4
ПРИКЛАДНЕ ЗАСТОСУВАННЯ РОЗРОБЛЕНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ

4.1 Прогнозування динаміки приросту щодобової кількості хворих на коронавірус в Україні

Для оцінювання прикладного застосування розробленої інформаційної технології зібрано всі результати роботи моделі в межах співпраці з РГ по ковіду. Це дає можливість спостерігати зміни в динаміці прогнозування протягом 2020-2022 років, тобто розроблена інформаційна технологія удосконалювалась весь час протягом трьох років. Перші теоретичні результати, опубліковані у статтях [16][17] (з урахуванням досвіду роботи з моделями машинного навчання, отриманого під час роботи над статтею [22]), вже у 2020 р. дали цікаві результати, які увійшли у звіт РГ по ковіду «Прогноз розвитку епідемії COVID-19 в Україні на 23 листопада – 7 грудня 2020 року («Прогноз РГ-29») [45]. В подальшому в кожному наступному звіті ці результати включались як окремий розділ «Прогноз розвитку епідемії в Україні з використанням статистичної моделі часових рядів Facebook Prophet». Винятки робились тільки для періодів між хвилями, коли статистична модель FB Prophet була неефективною і прогнозування РГ по ковіду здійснювалось тільки на основі моделі SEIR-U. Однак, в інші (найбільш складні для України періоди) щоразу проводилось прогнозування і за моделлю SEIR-U, і за моделлю FB Prophet. Розділ, співавторами якого були здобувач і його науковий керівник, часто розширявся – додавався аналіз динаміки по 69 країнах світу на додаток до прогнозу по Україні. Обмеженням була кількість країн, по яких бібліотека FB Prophet має списки офіційних свят, хоча ці списки були власноруч доповнені деякими європейськими країнами, яких не було в бібліотеці FB Prophet – Мокін В.Б. створив новий датасет «COVID-19: Holidays of countries» [83] (у FB Prophet тоді було тільки 62 країни).

Мутації інфекційної хвороби вимагали постійного доопрацювання методу ідентифікації параметрів моделі Prophet, особливо при розповсюдженні таких штамів коронавірусу як Delta та Omicron. Ці штами характерні більшою частотою розповсюдження в порівнянні зі штамом Alpha, який був поширений в 2020 році. Враховуючи наявність цих трьох основних штамів, закодуємо версії моделі Prophet по назвах штамів, а саме – модель M_α, що використовувалась для прогнозування захворюваності на штам Alpha, та M_δ і M_o, які використовувались для штаму Delta та Omicron. Крім цього, додамо в таблицю характер кривої, яка згенерована на основі прогнозу моделі, закодуємо характер наступним чином:

- D_1 відповідає наростаючому графіку,

- D_{-1} – спаданню,

- D_0 відповідає графіку між хвилями зростання захворюваності.

На відміну від таблиці 3.1, де наведена відносна похибка, обчислена на тестових даних, наведемо похибку, обчислену на валідаційному датасеті (як правило, це – 2 останні тижні ряду, які не використовувались для тренування моделі) за оптимальною моделлю. Саме вона свідчить про апроксимаційну можливість побудованих моделей. Результати роботи цих моделей за реальними даними наведені в таблиці 4.1:

Таблиця 4.1 Результати прогнозування часового ряду кількості хворих на коронавірус в Україні протягом 2020-2022 р.

Дата прогнозу	Тривалість прогнозу	Точність прогнозу за оптимальною моделлю на валідаційному датасеті	Версія моделі та штамму	Характер кривої
РГ-62 (22.02.2022)	23.02.2022-08.03.2022	4.67%	M_o	D_{-1}

Таблиця 4.1 Результати прогнозування часового ряду кількості хворих на коронавірус в Україні протягом 2020-2022 р.

Дата прогнозу	Тривалість прогнозу	Точність прогнозу за оптимальною моделлю на валідаційному датасеті	Версія моделі та штамму	Характер кривої
РГ-61 (08.02.2022)	09.02.2022-22.02.2022	1.81%	M_o	D_1
РГ-60 (26.01.2022)	27.01.2022-08.02.2022	0.98%	M_o	D_1
РГ-59 (11.01.2022)	12.01.2022-25.01.2022	16.53%	M_o	D_0
РГ-58 (21.12.2022)	22.12.2021 - 04.01.2022	7.56%	M_o	D_{-1}
РГ-57 (07.12.2021)	08.12.2021 - 21.12.2021	5.07%	M_o	D_{-1}
РГ-56 (23.11.2022)	24.11.2021 - 07.12.2021	16.9%	M_o	D_{-1}
РГ-55 (09.11.2021)	10.11.2021 - 23.11.2021	9.3%	M_o	D_1
РГ-54 (26.10.2021)	27.10.2021 - 9.11.2021	7.28%	M_o	D_1
РГ-53 (12.10.2021)	13.10.2021 - 26.10.2021	7.75%	M_o	D_1
РГ-52 (29.09.2021)	29.09.2021 - 12.10.2021	5.41%	M_o	D_1

Таблиця 4.1 Результати прогнозування часового ряду кількості хворих на коронавірус в Україні протягом 2020-2022 р.

Дата прогнозу	Тривалість прогнозу	Точність прогнозу за оптимальною моделлю на валідаційному датасеті	Версія моделі та штамму	Характер кривої
РГ-51 (14.09.2021)	15.09.2021 - 27.09.2021	25.68%	M_o	D_1
РГ-42 (20.04.2021)	21.04.2021 - 03.05.2021	8.76%	M_δ	D_{-1}
РГ-41 (06.04.2021)	07.04.2021 - 19.04.2021	9.3%	M_δ	D_1
РГ-40 (23.03.2021)	24.03.2021 - 05.04.2021	7.2%	M_δ	D_1
РГ-39 (10.03.2021)	11.03.2021 - 22.03.2021	5.4%	M_δ	D_1
РГ-38 (22.02.2021)	23.02.2021 - 1.03.2021	20.15%	M_δ	D_0
РГ-37 (08.02.2021)	09.02.2021 - 15.02.2021	12.7%	M_δ	D_0
РГ-36 (25.01.2021)	26.01.2021 - 01.02.2021	19.9%	M_δ	D_{-1}
РГ-35 (11.01.2021)	12.01.2021 - 18.01.2021	18.4%	M_δ	D_{-1}
РГ-34 (28.12.2020)	29.12.2020 - 11.01.2021	7.48%	M_δ	D_{-1}

Таблиця 4.1 Результати прогнозування часового ряду кількості хворих на коронавірус в Україні протягом 2020-2022 р.

Дата прогнозу	Тривалість прогнозу	Точність прогнозу за оптимальною моделлю на валідаційному датасеті	Версія моделі та штамму	Характер кривої
РГ-32 (14.12.2020)	15.12.2020 - 28.12.2020	3.5%	M_α	D_{-1}
РГ-31 (07.12.2020)	8.12.2020 - 21.12.2020	6.4%	M_α	D_1
РГ-30 (30.11.2020)	01.12.2020 - 13.12.2020	3.2%	M_α	D_1
РГ-29 (23.11.2020)	24.11.2020 - 06.12.2020	2.2%	M_α	D_1

У додатку В наведено об’єднану таблицю і таблиць 3.1 та 4.1, де видно усі основні закономірності.

Як видно з табл. 4.1, найкращі результати інформаційна технологія прогнозування кількості хворих на коронавірус, тобто ідентифікована за її допомогою оптимальна модель, давала під час наростання хвилі, у т.ч. з урахуванням попередньої історії спостережень.

Виконувались спроби удосконалення моделі:

1. Врахування додаткових ознак як регресорів у моделі (множинна регресія FB Prophet), але така спроба станом на 07.12.2020 р. дала похибку в 7,4% [84], у той час, як похибка без урахування такого регресора дала похибку в 6,4% (див. табл. 4.1).

2. Моделювання наростання хвилі з використанням логістичної регресії для опису тренда у моделі FB Prophet [85], адже звичайною моделлю важко передбачити де хвиля «вийде на плато», тобто припинить своє стрімке наростання. Експерименти показали, що для задачі з кількістю хворих на коронавірус такий підхід не є ефективним, хіба що використовувати його для можливих сценаріїв розвитку ситуації у кожний момент такого наростання – як один із ймовірних сценаріїв (рис. 4.1).

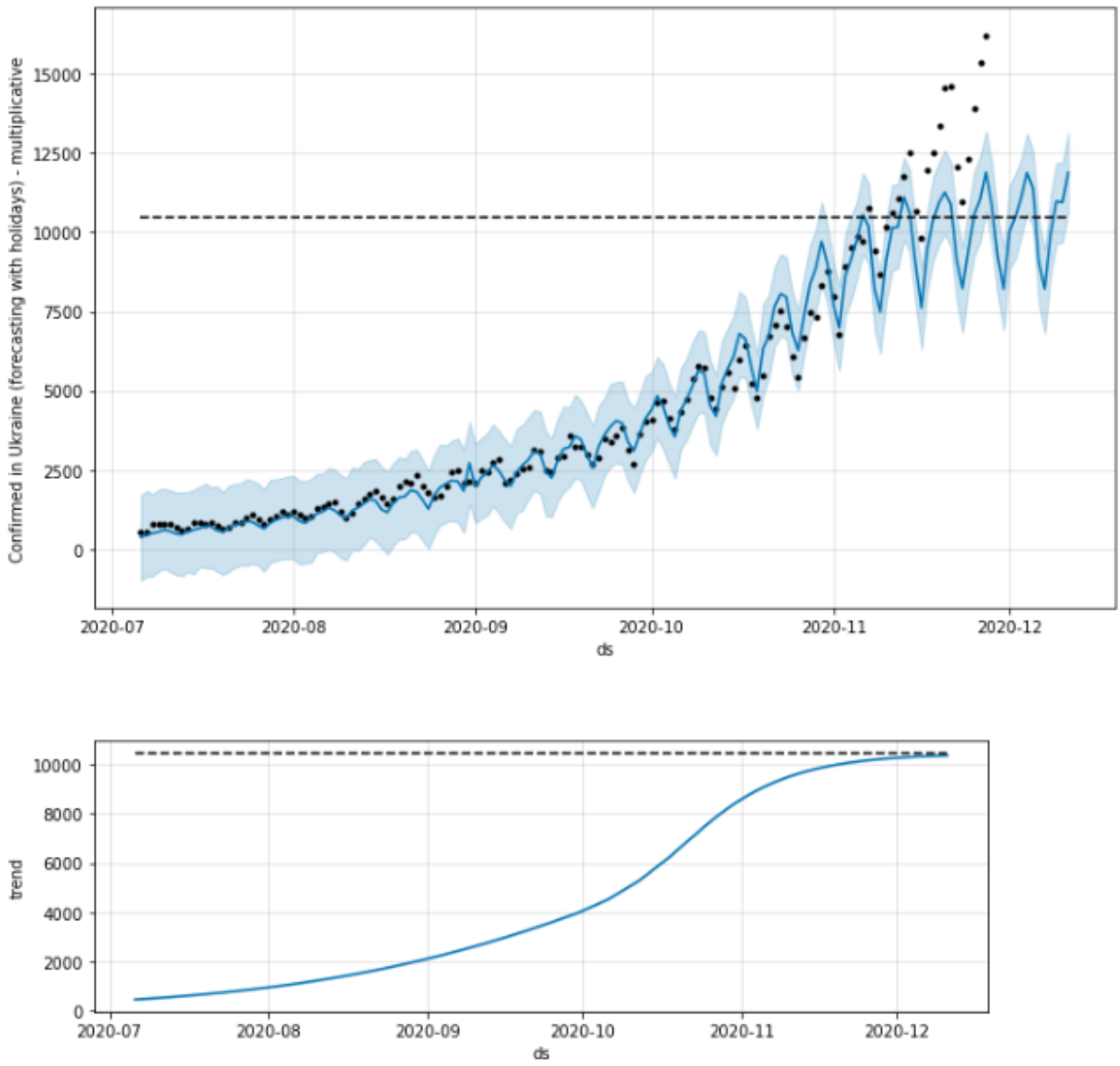


Рис. 4.1. Один з можливих сценаріїв зростання кількості хворих на коронавірус в Україні на етапі наростання хвилі з використанням логістичної регресії у моделі FB Prophet для здійснення прогнозування

Як видно з таблиці 4.1, відносна похибка моделі дещо погіршилась від 2,2% у 2020 р. до 16,9-25,68% у 2021-2022 рр., через значну волатильність процесу, але, використовуючи удосконалення алгоритму на багатоітераційний випадок (див. підрозд. 2.4), цю точність можна підвищити.

Наведемо найбільш наочні графіки порівняння прогнозів моделей SEIR-U та моделі Prophet, ідентифікованої за запропонованою інформаційною технологією, на рисунках 4.2-4.6:

Рис. 4.2 Графік зі звіту «Прогноз РГ-31» [43, рис. 24]

Рис. 4.3 Графік зі звіту «Прогноз РГ-42» [33, рис. 39]

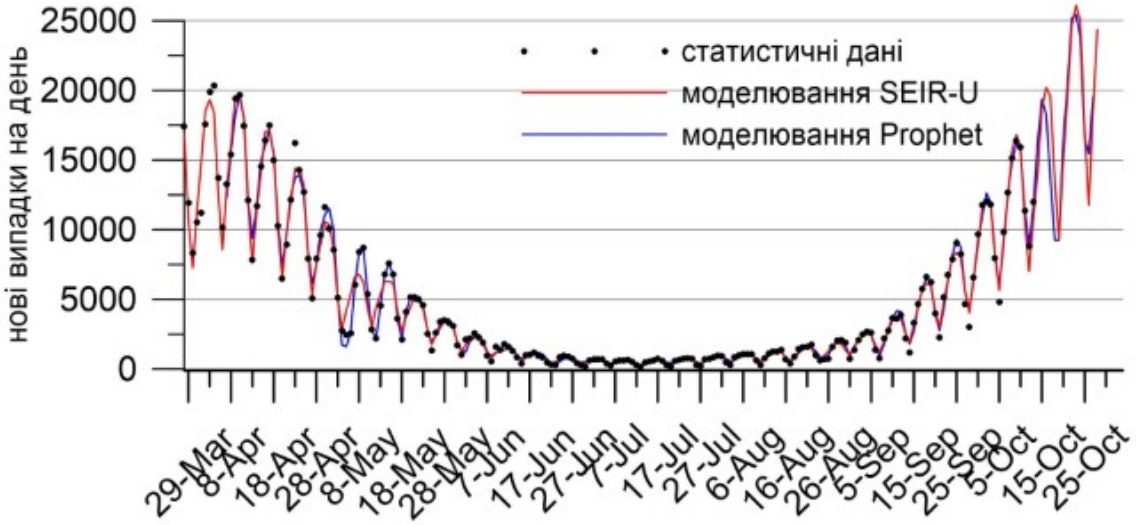


Рис. 4.4 Графік зі звіту «Прогноз РГ-53» [29, рис. 41]

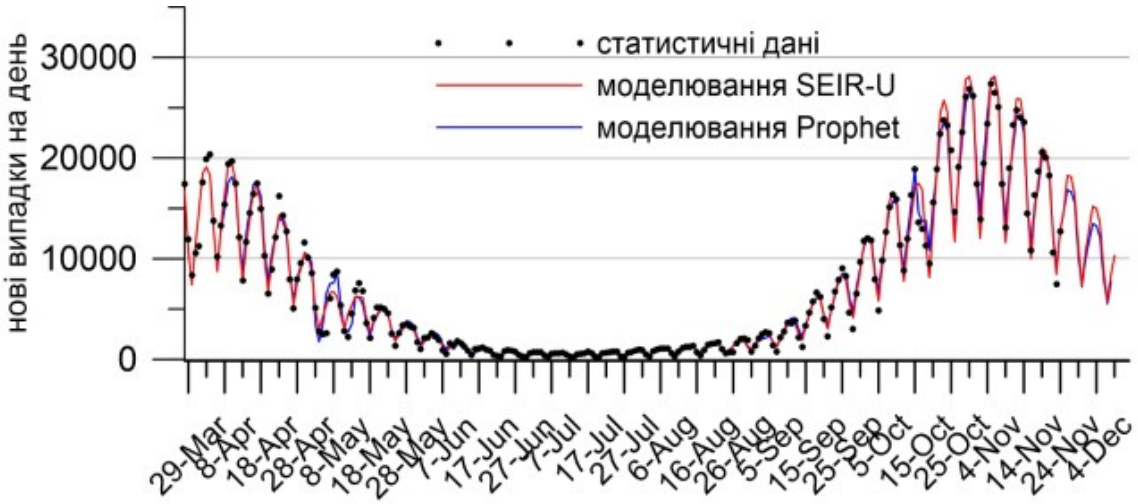


Рис. 4.5 Графік зі звіту «Прогноз РГ-56» [26, рис. 49]

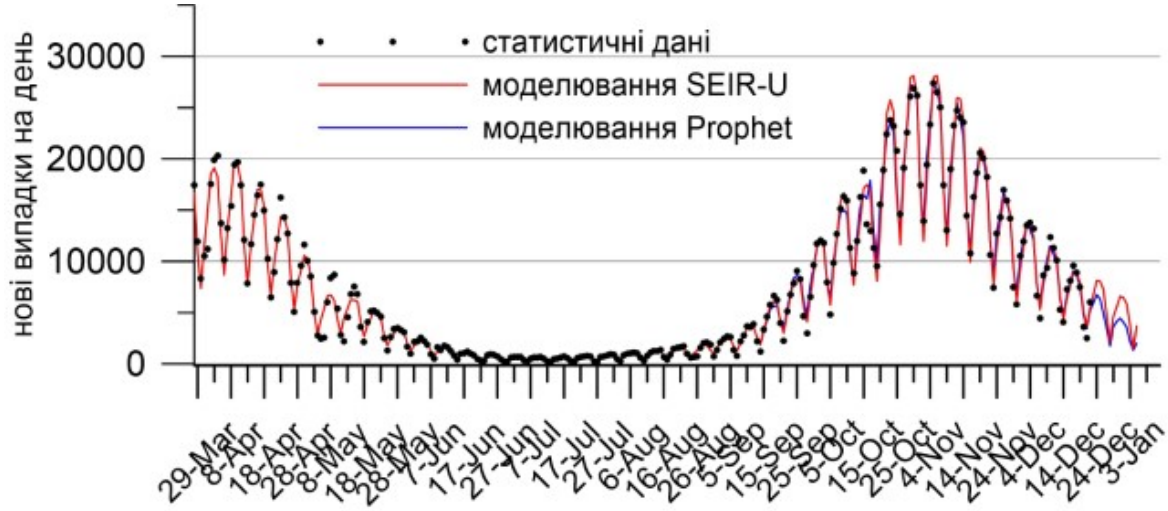


Рис. 4.6 Графік зі звіту «Прогноз РГ-58» [24, рис. 49]

Рисунок 4.2 показує початок хвилі приросту захворюваності восени 2020 року і показує початок приросту захворюваності, обумовленої штамом Alpha, а рис. 4.3 ілюструє прогнозування хвиль, через поширення штаму Delta, причому показує сезонне спадання захворюваності на початку травня 2021 року. У свою чергу, рисунки 4.4-4.6 репрезентують прогрес хвилі під час стрімкого розповсюдження штаму Omicron.

Наведені порівняльні графіки, також, показують різницю у деяких прогнозах моделі Prophet та моделі SEIR-U.

4.2 Прогнозування динаміки приросту щодобової кількості хворих на коронавірус у 68 країнах світу за одноетапним методом ідентифікації моделі FB Prophet

Протягом 2021-2022 років здійснювалось прогнозування щодової кількості хворих на коронавірус у майже 70 країнах світу за спрощеною моделлю. Результати увійшли у звіти РГ по ковіду [24-45], пораховані у Python-ноутбуку [86]. Оскільки розрахунки займали від 4 до 13 годин, в залежності від тривалості та волатильності рядів, застосовувався одноетапний метод ідентифікації, тобто ідентифікувались тільки параметри вікна дат аномалій.

Результат такого підходу, наприклад у 06.03.2022 р. наведено у таблиці 4.5. Як видно (див. останній стовпець з різницею відносної похибки прогнозів з урахуванням дат аномалій і без них), по більшості країн урахування дат аномалій забезпечує збільшення точності прогнозу. Крім того, по більшості країн досягнуто доволі високу точність прогнозування, яку можна збільшити, врахувавши ще й другий етап методів ідентифікації та багатоітеративність, запропоновані у підрозділах 2.3 та 2.4.

Таблиця 4.2 Параметри вікна аномальних дат та відносна похибка моделей FB Prophet прогнозів на 07-13.03.2022 р. для 68 країн світу з урахуванням дат аномалій і без них

Країна	y_t	$\hat{y}_t$	Ψ	$\delta_0, \%$	$\delta_h, \%$	t_0	t_1	R_h
Албанія	180	0	28	329	314.35	-3	3	0.5
Аргентина	2274	2474	53	34.55	31.811	-1	1	0.5
Австралія	22182	12990	26	26.14	19.739	-3	1	0.5
Австрія	28463	25153	37	12.16	11.18	-1	1	0.5
Бангладеш	529	0	21	83.36	69.494	0	3	0.05
Білорусь	1662	2414	25	58.52	54.463	0	2	1
Бельгія	14713	1051	35	69.76	64.416	-3	0	0.5
Бразилія	15725	15822	23	46	45.672	-3	2	0.05
Болгарія	515	186	44	28.69	26.269	-2	3	15
Бурунді	16	0	34	674.7	154.02	-3	3	15
Канада	2281	720	29	61.18	42.013	-2	2	1
Чилі	20113	21656	44	16.85	16.761	0	1	0.05
Китай	31335	59313	23	63.8	63.794	-3	0	15
Колумбія	1088	0	51	66.07	58.292	-3	3	1
Хорватія	875	0	40	86.58	70.245	-3	2	15
Чехія	4578	132	38	62.75	57.447	0	2	1
Данія	10827	8298	35	40.35	40.294	0	2	1
Домініканська республіка	66	244	34	30.79	31.898	0	0	0.05
Єгипет	1093	1321	37	20.43	19.367	-1	3	15
Естонія	1641	3355	35	26.61	26.304	-1	1	1

Таблиця 4.2 Параметри вікна аномальних дат та відносна похибка моделей FB Prophet прогнозів на 07-13.03.2022 р. для 68 країн світу з урахуванням дат аномалій і без них

Країна	y_t	$\hat{y}_t$	Ψ	$\delta_0, \%$	$\delta_h, \%$	t_0	t_1	R_h
Фінляндія	7185	6107	43	93.03	91.173	0	0	0.05
Франція	45328	4805	32	62.53	50.355	-3	2	1
Грузія	1700	0	24	51.25	48.384	-3	0	0.05
Німеччина	67466	57830	26	34.72	34.084	-2	3	1
Греція	9213	6681	31	11.97	17.394	0	0	0.05
Гондурас	966	2290	52	156.1	152.55	0	1	15
Угорщина	3064	1568	43	28.1	17.617	-2	2	0.05
Ісландія	2648	4207	47	44.18	43.894	0	0	0.05
Індія	4362	0	67	69.95	52.812	-3	3	0.5
Індонезія	24867	24401	37	14.5	13.328	-1	1	15
Ірландія	3995	5502	31	41.36	15.357	-3	2	0.5
Ізраїль	3991	0	118	157.5	125.94	-2	1	15
Італія	35889	18911	34	24.55	14.873	-3	3	1
Японія	53897	47948	48	11.61	7.9309	-2	3	0.5
Кенія	4	0	30	165.5	29.655	-1	1	1
Республіка Корея	210713	189988	50	11.08	11.075	-3	0	15
Латвія	3992	5704	38	10.83	10.062	-1	1	0.5
Литва	3212	2496	47	24.81	23.82	-3	0	15
Люксембург	653	833	32	16.15	14.875	0	0	0.5
Малайзія	27435	25860	39	9.033	8.8702	-1	1	1
Мексика	10593	1465	33	52.24	31.128	0	1	15

Таблиця 4.2 Параметри вікна аномальних дат та відносна похибка моделей FB Prophet прогнозів на 07-13.03.2022 р. для 68 країн світу з урахуванням дат аномалій і без них

Країна	y_t	$\hat{y}_t$	Ψ	$\delta_0, \%$	$\delta_h, \%$	t_0	t_1	R_h
Молдова	518	0	21	71.16	64.862	-3	0	15
Марокко	60	0	25	227.5	162.33	-3	1	15
Нідерланди	66799	0	30	118.6	109.28	0	3	0.5
Нова Зеландія	17552	26302	39	54.24	55.266	-1	0	15
Нікарагуа	101	114	24	12.87	12.871	0	0	0.05
Нігерія	3	0	20	59.13	44.348	-1	1	15
Норвегія	3336	5602	35	22.51	22.301	0	0	0.05
Пакистан	756	0	51	105.7	94.641	0	3	0.5
Парагвай	260	475	35	33.5	32.056	-2	0	0.05
Перу	3763	0	41	149.4	139.32	-1	3	1
Філіппіни	864	0	34	86.8	74.804	-3	2	0.05
Польща	7698	3566	37	38.29	31.881	-3	3	0.5
Португалія	10066	2974	38	47.57	46.057	-3	0	15
Румунія	3092	475	43	43.22	35.225	-3	3	0.05
Сербія	1641	0	37	95.77	53.938	-1	2	15
Сінгапур	13158	16158	38	29.84	12.666	-3	3	15
Словаччина	7380	9373	43	7.601	6.8886	-2	2	1
Словенія	942	0	37	144.2	127.36	0	0	15

Таблиця 4.2 Параметри вікна аномальних дат та відносна похибка моделей FB Prophet прогнозів на 07-13.03.2022 р. для 68 країн світу з урахуванням дат аномалій і без них

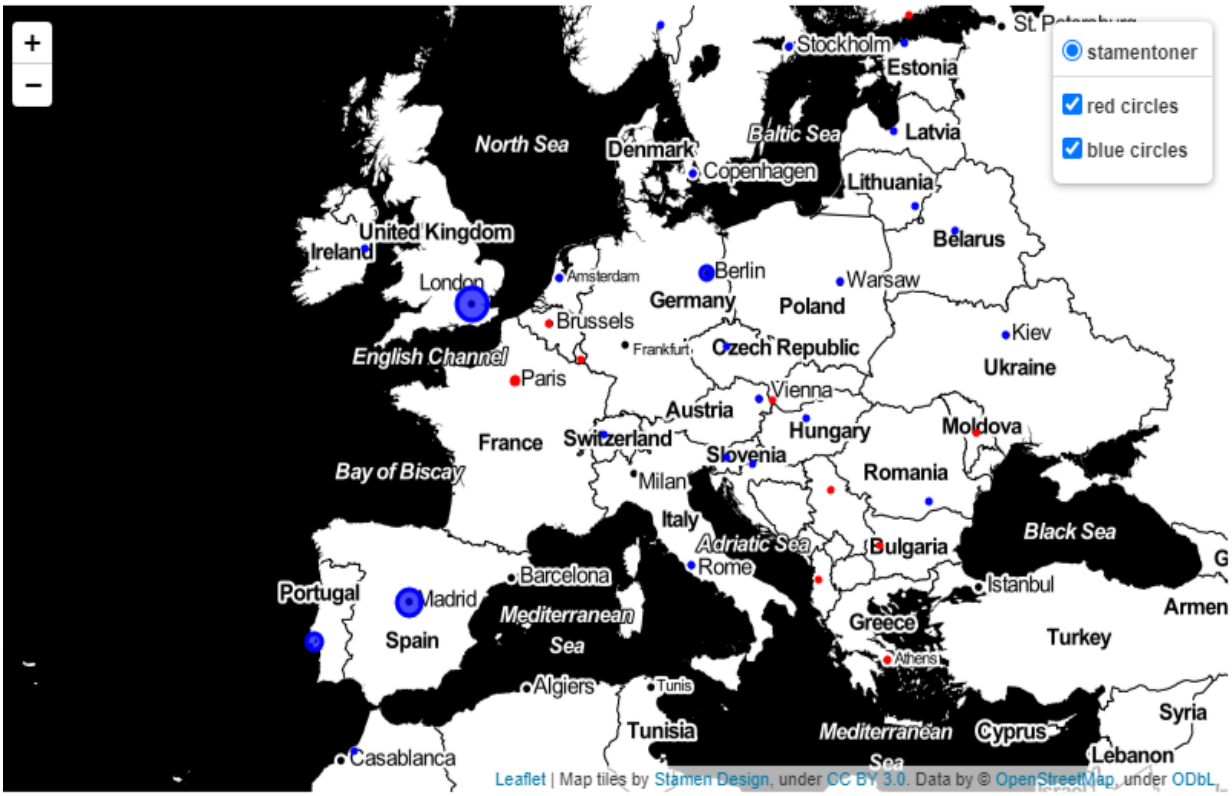
Країна	y_t	$\hat{y}_t$	Ψ	$\delta_0, \%$	$\delta_h, \%$	t_0	t_1	R_h
Південна Африка	1147	1398	40	10.24	9.6818	0	0	0.05
Іспанія	22400	7770	25	70.13	39.202	-3	3	0.05
Швеція	2030	5217	46	76.52	150.13	0	2	0.05
Швейцарія	25132	8463	26	49.11	47.165	-3	3	0.05
Таїланд	21162	28132	36	29.84	29.717	0	0	0.5
Туреччина	27671	37855	23	16.34	15.187	-3	3	0.05
Об'єднане Королівство	44891	0	56	90.66	93.097	0	0	0.05
Сполучені Штати	5740	0	33	99.21	65.887	0	3	15
В'єтнам	202180	142552	32	24.95	24.497	-2	3	15

4.3 Міжрегіональний аналіз прогнозів динаміки приросту щодобової кількості хворих на коронавірус у майже 70 країнах світу

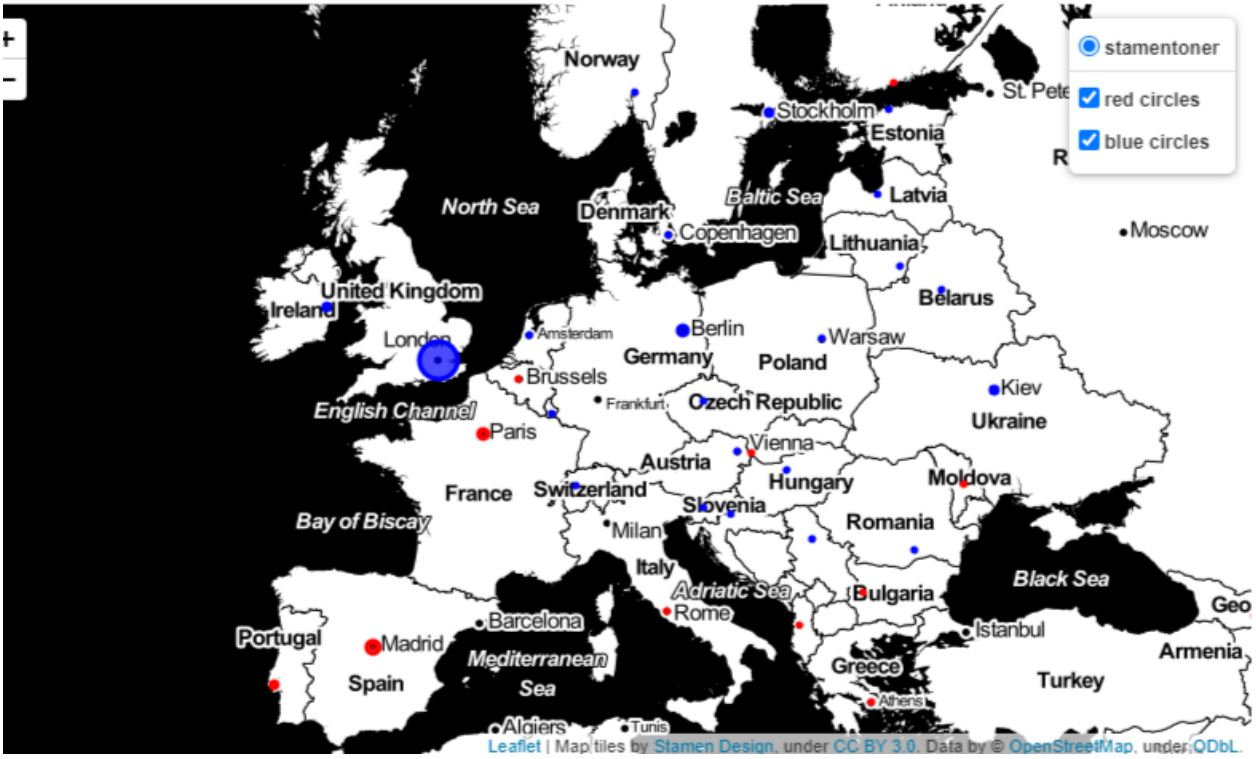
Протягом січня-березня 2021 року проводилось прогнозування для майже 70 країн світу [19], описане у попередньому підрозділі. Потім було застосовано метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах з використанням в якості інтегрального показника прогнозів на тиждень нахилу тижневої ділянки шматково-лінійної апроксимації тренду прогнозів по майже 70

країнах світу, описаних у попередньому підрозділі, що дозволяє аналізувати закономірності міжрегіонального поширення хвороби, ігноруючи місцеві особливості. Зокрема, виділено тренд прогнозів, потім здійснено апроксимацію їх прямими, після чого, здійснено картування нахилів цих прямих на карті країн світу. Ця операція проведена для прогнозів за багато тижнів протягом декількох місяців 2021 року.

Картування здійснювалось за допомогою Python-бібліотеки Folium. Дана бібліотека дає можливість генерувати інтерактивні карти, які можна компонувати та конфігурувати під власні потреби. Зокрема, побудовано серію карт для 7 послідовних тижнів протягом січня-березня 2021 року [87-96]. Ці карти для Європи наведені на рис. 1-4.



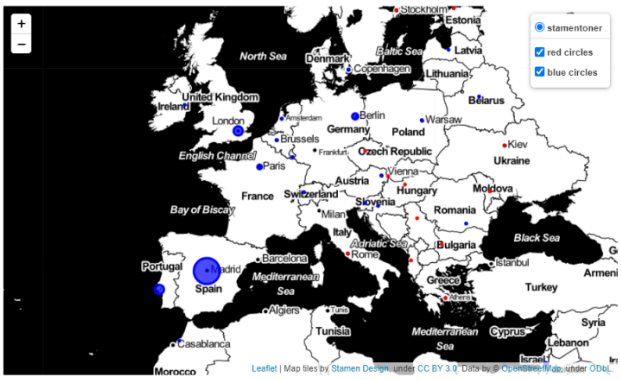
а)



б)

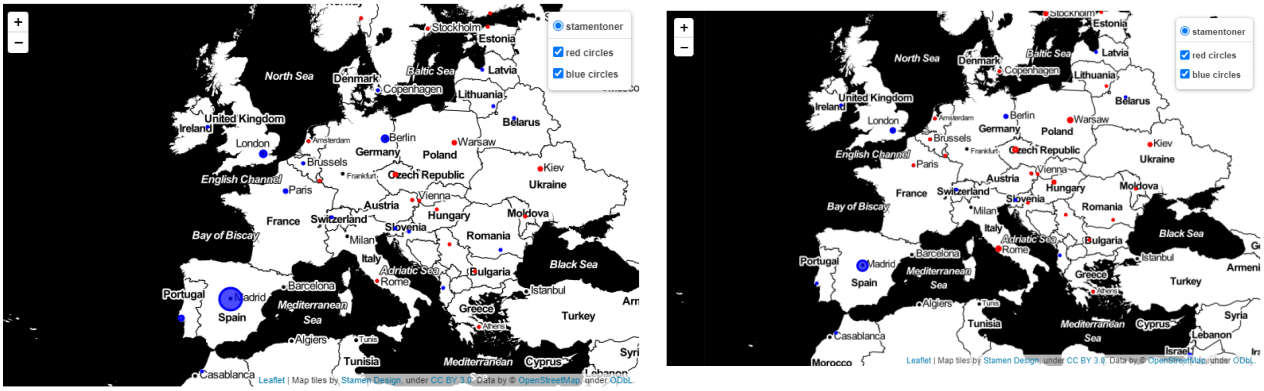
Рис. 4.7 Карти тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи, побудовані у 2021 році: а) 24.01.2021 [87, 88]; б) 31.01.2021 [89]

а)



б)

Рис. 4.8 Карти тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи, побудовані у 2021 році: а) 07.02.2021 [90, 91]; б) 14.02.2021 [92]



а) б)

Рис. 4.9 Карти тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи, побудовані у 2021 році: а) 21.02.2021 [93, 94]; б) 01.03.2021 [95]

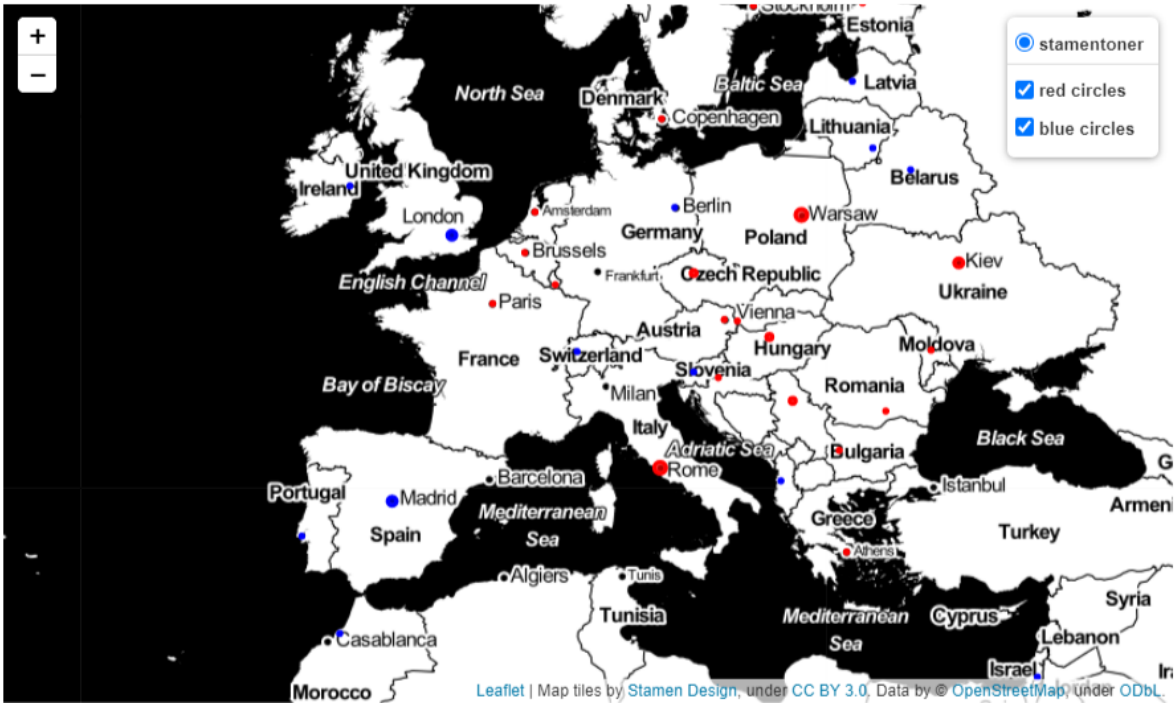


Рис. 4.10 Карта тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи, побудована 07.03.2021 р. [96]

Як видно з доданих рисунків, такий підхід дає можливість оцінювання потенційного поширення панпідемії в більш зручному вигляді. Така візуалізація дає змогу аналізувати стан епідеміологічної ситуації в набагато ширшому масштабі та отримати попереднє розуміння того, що відбувається на даний момент. Після звернення до такої карти можна переходити до наступного огляду більш детальних графіків задля поглиблення дослідження та аналізу.

Крім того, виділені нахили були проаналізовані з використанням когнітивних карт, як було зазначено у підрозд. 2.5. У статті [20] зазначено, що характери динаміки кількості хворих на коронавірус в Україні є схожим на протікання цього процесу в Румунії, але в Румунії хвилі, а особливо їх пікові значення, мають місце раніше (рис. 4.11).

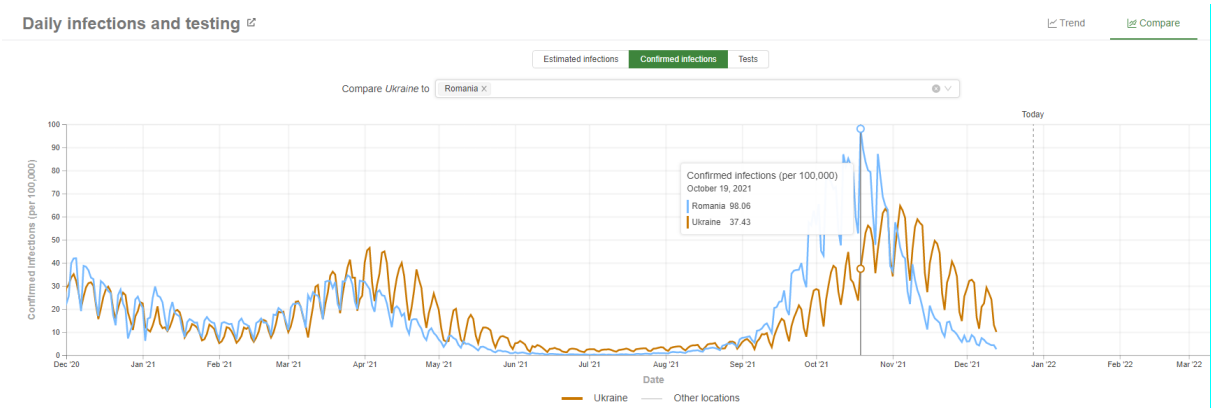


Рис. 4.11. Хвилі з щодобовими приростами кількості нових підтверджених хворих в Україні та Румунії за 2020-2021 рр.

А отже, можна реалізувати прогнозування. По нахилах були визначені дати початку хвиль, по них, з використанням математичного апарату, розробленого Дратованим М.В. та Мокіним В.Б., побудована стійка когнітивна карта, а по ній – здійснено когнітивне моделювання нахилів трендів. Тобто по трендах прогнозів для Румунії спрогнозовані тренди нахилів для України (рис. 4.12), щоб дало задовільну точність. А це, додатково підтвердило ефективність та практичну цінність розробленої технології.

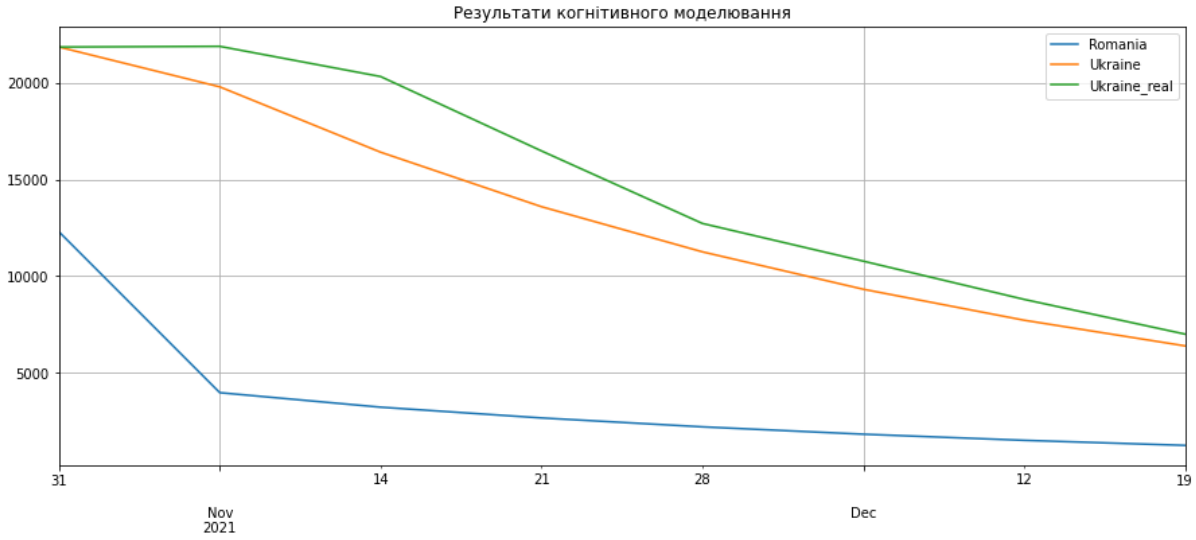


Рис. 4.12. Прогнозування нахилів трендів прогнозів в Україні по нахилам трендів прогнозів в Румунії [20]

4.4 Впровадження розробленої технології у навчальний процес

Результати дослідження впроваджені у навчальний процес Вінницького національного технічного університету під час викладання дисципліни «Інформаційні технології моніторингу та аналізу даних» для студентів, які навчаються за освітньою програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності 126 Інформаційні системи та технології, а також – дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньою програмою «Інформаційні системи та технології» рівня «доктора філософії» цієї ж спеціальності 126 для навчання знанням і навичкам застосування методів і моделей машинного навчання та інформаційних технологій оброблення і прогнозування даних часових рядів захворюваності на інфекційні захворювання. Ці результати можуть використовуватися як додатковий матеріал для навчання та поглибленого розуміння концепцій, пов’язаних із цією областю.

Зокрема, при викладанні цих дисциплін використовуються такі результати досліджень, отримані у дисертаційній роботі здобувача (підтверджується актом впровадження ВНТУ від 4 грудня 2023 року):

- розвідувальний аналіз даних часових рядів: декомпозиція ряду на різні види сезонності, перевірка на стаціонарність, пошук аномалій та візуалізація результатів такого аналізу, у т.ч. побудова інтерактивних карт;

- тренування моделей машинного навчання (лінійна регресія, метод опорних векторів, дерева рішень та їх ансамблі, багатошарова нейронна мережа, ARIMA, Facebook Prophet з типовою структурою) з використанням технік GridSearchCV, HyperOpt і байєсівської оптимізації для пошуку оптимальних параметрів) та порівняльний аналіз їх точності за різними метриками і вибір оптимальної моделі;

- багатоітераційний метод ідентифікації параметрів та структури моделі нестаціонарного часового ряду з різними видами сезонності на основі Facebook Prophet.

Використання зазначених результатів дозволило підвищити якість навчального процесу із згаданих дисциплін.

4.5 Висновки до розділу

У розділі 4 наведені приклади застосування основних складових розробленої інформаційної технології прогнозування часового ряду кількості хворих на коронавірус.

Продемонстровано можливість аналізу інфекційного впливу сусідніх регіонів по нахилу елементів шматково-лінійної апроксимації кривих прогнозів зміни тренду на весь горизонт цього прогнозу (як правило, на 1 чи 2 тижні), який виділяє запропонована Prophet-модель та відокремлює від нього 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 тощо), що дозволяє більш точно аналізувати основні закономірності динаміки процесу.

Дослідження датасету захворюваності на коронавірус дало змогу перевірити інформаційну технологію та оцінити точність прогнозування не лише в абсолютних значеннях, а й – у порівнянні з моделлю SEIR-U Робочої групи з математичного моделювання проблем, пов’язаних з епідемією коронавірусу SARS-CoV-2 в Україні. Випробування розробленої інформаційної технології дало результати прийнятної точності.

Продемонстровано ефективність роботи запропонованої технології не тільки за даними для України, а й для інших майже 70 країн світу, що забезпечило доволі високу точність прогнозування.

Охарактеризовано впровадження у навчальний процес.

ВИСНОВКИ

У дисертаційній роботі розв’язано актуальну науково-прикладну задачу підвищення точності прогнозування кількості хворих на коронавірус у короткостроковій перспективі шляхом розроблення інформаційної технології за допомогою методів машинного навчання та моделей часових рядів.

1. Охарактеризовано основні проблеми, пов’язані з прогнозуванням кількості хворих на коронавірус в Україні та інших країнах, обумовлені, передусім, проблемами з даними, пов’язаними зі складностями діагностування медичного та організаційного характеру, мутаціями хвороби, недотриманням населенням карантинних умов тощо. Проведено розвідувальний аналіз даних. Проаналізовано якість даних про інші фактори, які впливають на поширення коронавірусу (пересування жителів країни, метеодані, рішення уряду (на основі Оксфордського трекера антикоронавірусної діяльності урядів країн світу), вчасність оприлюднення підтверджених тестів на коронавірус тощо. Доведено, що вплив цих факторів є дуже мінливим, дані є неточними і зашумленими, надходять із різним запізненням, а отже, їх варто враховувати не як додаткові ознаки для побудови множинної регресії, а краще дати їх різкої зміни чи аномальних значень формалізувати як дати аномальної поведінки основного часового ряду, а тоді ці аномалії можна враховувати параметрично і звести задачу до однофакторної, тобто – до побудови моделі лише часового ряду кількості нових випадків захворювання на коронавірус. Проведено огляд відомих моделей і методів прогнозування цієї ознаки. Зазначено, що через проблеми з даними, перевагу варто віддавати моделям машинного навчання, а не – диференціальним рівнянням, наприклад SIR, SEIR чи SEIR-U. Здійснено порівняння точності різних моделей машинного навчання за даними 2020 року (лінійна регресія, метод опорних векторів, багатошарова нейромережа, дерева рішень та їх ансамблі, ARIMA, Facebook Prophet з декількома варіантами параметрів ряду Фур’є для опису сезонності). Для формування додаткових ознак багатофакторних моделей

використовувалась бібліотека Tsfresh, для оптимізації параметрів – техніка GridSearchCV. Аналіз довів беззаперечну перевагу моделі Facebook Prophet. Однак, зазначено, що традиційні методи налаштування параметрів цієї моделі є недостатньо ефективними для цієї задачі, а отже необхідно розробити нові методи чи удосконалити існуючі, щоб підвищити точність прогнозування.

2. Декомпозиція ряду з різним періодом сезонності показали, що має місце тижнева періодичність і ще мінімум 1-2 з іншим періодом, але це потребує окремого дослідження. Формалізовано модель Facebook Prophet для розв'язання поставленої задачі, з урахуванням виявлених видів сезонностей та аномалій. Охарактеризовано яку кількість і яких параметрів слід визначити для ідентифікації дійсно адекватної моделі та зазначено, що повний перебір їх варіантів займе забагато часу, а тому потрібні нові методи для їх ідентифікації. Запропоновано метод оцінювання порядку ряду Фур'є багатохвильового періодичного процесу, в якому оцінювання здійснюється лише по 10% верхівки однієї хвилі з використанням емпіричних співвідношень, що дозволяє здійснювати таке оцінювання, навіть, за умов, коли одна хвиля перекриває іншу і їх важко відокремити одну від іншої. Результат цього оцінювання варто використувавати як початкове наближення для подальшого уточнення методами машинного навчання. Також, розроблено ряд співвідношень для оцінювання періоду багатохвильового періодичного процесу по відстані між сусідними піками та мінімальними значеннями, навіть, якщо хвилі не спадають до нуля. Запропоновано новий метод паралельно-послідовної багатопараметрової ідентифікації моделі Facebook Prophet для короткострокового прогнозування часового ряду кількості хворих на коронавірус в заданому регіоні, який відрізняється від існуючих більшою кількістю параметрів, що ідентифікуються: сила і розмір вікна впливу дат аномалій, ступінь регуляризації і тип моделі (адитивна чи мультиплікативна), порядок ряду Фур'є (уточнення початкового наближення) і ступінь регуляризації 3-х різних періодичних складових, які враховують внутрішньотижневі, тижневі і багатотижневі закономірності,

характерні для інфекційних хвороб, у т.ч. коронавірусу, що дозволяє суттєво підвищити точність прогнозів та більш глибоко дослідити закономірності, які впливають на цей часовий ряд. Запропоновано удосконалений варіант цього методу з багатоітераційною оптимізацією параметрів, який дозволяє підвищити його точність. Для пошуку оптимальних значень параметрів використовується техніка HyperOpt та метод байєсівської оптимізації. Застосування цього удосконаленого багатоітераційного методу дозволило суттєво підвищити точність прогнозування, зокрема, за даними по Україні у 2021-2022 роках відносна похибка зменшилась у 3-10 разів, у порівнянні з методом, де була тільки одна така ітерація. Також, запропоновано робити прогнози для найбільш песимістичного та найбільш оптимістичного сценаріїв розвитку явища.

3. Удосконалено метод узагальнення прогнозів кількості хворих на коронавірус у різних регіонах, який відрізняється від існуючих використанням в якості інтегрального показника прогнозу на тиждень нахилу тижневої ділянки шматково-лінійної апроксимації тренду цього прогнозу, отриманого за моделлю, ідентифікованою із застосуванням запропонованого методу паралельно-послідовної багатопараметрової ідентифікації, що дозволяє аналізувати закономірності міжрегіонального поширення хвороби, ігноруючи місцеві особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень, тощо). Запропоновано здійснювати картографічну візуалізацію прогнозів динаміки приросту кількості хворих на коронавірус, в якій радіус кругових діаграм у геометричному центрі чи центральному місті регіону буде пропорційним нахилу, визначеному за запропонованим методом узагальнення прогнозів, що дозволяє більш точно виявити закономірності розповсюдження захворювання на цій карті. Доцільно обробляти одразу декілька сусідніх регіонів. Продемонстровано роботу методу та картографічної візуалізації його результатів на рівні країн на карті світу та карті Європи. Колір кругової діаграми свідчить про зростання (червоний) чи зменшення (синій) кількості нових хворих. Зазначено, що ці нахили можна й кластеризувати, а потім за ними

побудувати байєсівську мережу для виявлення закономірностей у кожному кластері окремо, що, також, дасть нові цікаві результати (такий підхід здобувач вже успішно випробував на іншій задачі з аналізу медичних даних).

4. Розроблена структура запропонованої інформаційної технології прогнозування часових рядів кількості хворих на коронавірус та охарактеризовано її блоки і модулі. Побудовано та описано UML-діаграми послідовності роботи модулів технології, UML-діаграма діяльності модулю прогнозування захворюваності та UML-діаграма модулю створення картограм нахилів трендів інфекційного впливу. Архітектурно інформаційна технологія реалізовується у вигляді комплексу програм-ноутбуків на Python та ряду датасетів для збереження вхідних, проміжних і результуючих даних.

5. Створено програмно-інформаційне забезпечення для автоматизації запропонованих методів і складових інформаційної технології. На основні його складові подано заявки на отримання свідоцтв про реєстрацію авторських прав на твір (комп’ютерну програму). Для реалізації вибрано мову Python та безкоштовну платформу датасайтнистів Kaggle, де зручно розміщати і Python-ноутбуки, і датасети. Універсальність редактору Kaggle Code дозволяє доволі швидко його адаптувати і до інших відомих середовищ для роботи з Python-ноутбуками: AWS SageMaker, Google Colab, Jupyter Notebook або JupyterLab пакету Anaconda, Microsoft Azure Notebooks, PyCharm, Visual Studio Code та ін. У Google-платформі дасайнтистів Kaggle опубліковано у відкритий доступ 10 програм-ноутбуків у співавторстві з науковим керівником, які за 2020-2023 роки були переглянуті більше 36 тисяч разів (станом на 07.12.2023 р.). Здійснено випробування розробленого програмно-інформаційного забезпечення на реальних даних. Протягом 2020-2022 років здобувач брав участь у прогнозуванні щодобового приросту кількості хворих на коронавірус в 70 країнах світу, у т.ч. в Україні, – ці результати передавались в Робочу групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, офіційно створену при НАН України і членом якої є науковий керівник здобувача. Ці результати

увійшли у звіти, в яких прямо вказано «Обчислення за допомогою моделі Facebook Prophet і аналіз отриманих результатів виконали завідувач кафедри системного аналізу та інформаційних технологій (САІТ) Вінницького національного технічного університету (ВНТУ) доктор технічних наук, професор В.Б. Мокін і аспірант кафедри САІТ ВНТУ А.В. Лосенко», що підтверджується актом впровадження результатів роботи, підписаним в.о. директора Інституту проблем математичних машин і систем НАН України. Усі ці звіти опубліковані на сайті Президії НАН України (https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx), а перед тим надсилались в РНБО України та МОЗ України для використання під час прийняття рішень щодо керування ситуацією в Україні.

Продемонстровано ефективність роботи запропонованої технології не тільки за даними для України, а й для інших 70 країн світу, що забезпечило доволі високу точність прогнозування. Проведено порівняння відносної похибки точності прогнозування моделі Prophet, що використовується в інформаційній технології, та моделі SEIR-U, розробленої Робочою групою НАНУ з прогнозування коронавірусу. Аналіз показав, що запропонована технологія здійснює прогнозування на 2 тижні вперед за даними 2021-2022 років з відносною похибкою, іноді в 1,2-2 разів меншу, ніж модель SEIR_U. Охарактеризовано впровадження у навчальний процес.

Розроблену інформаційну технологію можна, за певного адаптування, застосовувати і для інших нестаціонарних часових рядів багатохвильової природи техногенного чи природного характеру.

Результати дисертаційного дослідження опубліковані у 9 роботах, у т.ч. у 5 статтях у наукових фахових періодичних виданнях, 2 матеріалах доповідей на міжнародних конференціях, що увійшли у колективні монографії, опублікованими за результатами цих конференцій, 2 тезах доповідей на науково-практичних конференціях. Крім того, результати увійшли у 25 звітів Робочої групи НАНУ з прогнозування коронавірусу (2020-2022 рр.).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

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

[2] Althaus C.L., "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 https: ispmbern.github.io/covid-19/swiss-epidemic-model

[3] Mokin V. B. «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. Червень, 12. 2020.

[4] Dr. Shikha Gaur. "Global forecasting of COVID-19 USING ARIMA BASED FB-PROPHET". *International Journal of Engineering Applied Sciences and Technology*, 2020 Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467.

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

[6] Indhuja M and Sindhuja PP. "Prediction of COVID-19 cases in India using prophet". *International Journal of Statistics and Applied Mathematics* 2020; 5(4): 103-106.

[7] Р.Н. Квєтний, Д. А. Ткачик, «Аналіз емоційного забарвлення тексту для прогнозування даних на фінансових ринках», *Інформаційні технології та комп'ютерна інженерія*, № 3, с. 51–58, 2022. Режим доступу: (фахове видання) https://doi.org/10.31649/1999-9941-2022-55-3-51-58

[8] Levenchuk L., Tymoshchuk O., GuskovaV., Bidyuk P., «Uncertainties in data processing, forecasting and decision making», *Системні дослідження та інформаційні технології: міжнародний науково-технічний журнал*, №3, 2023. Available online: 10.20535/SRIT.2308-8893.2023.3.05

[9] С. В. Голуб, С. Ю. Куницька, «Побудова ешелонів у поліагентних функціоналах для прогнозування кількості захворювань на Covid-19 в Україні», *Математичні машини і системи*, № 2, с. 45–51, 2021. Режим доступу: (фахове видання) https://doi.org/34121/1028-9763-2020-4-11-19 http://www.immsp.kiev.ua/publications/articles/2021/2021_2/02_21_Holub.pdf

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

[11] О.В. Бісікало, О. В. Кудрик, «База знань інтелектуальної інформаційної системи прогнозування фазової стабільності твердих розчинів», *Інформаційні технології та комп'ютерна інженерія*, № 1, 2023. Режим доступу: (фахове видання) http://ir.lib.vntu.edu.ua//handle/123456789/36554

[12] В. Півошенко, М. Кулик, Ю. Іванов, та А. Васюра «Аналіз та експериментальне дослідження методу безмодельного навчання з підкріпленням», *Вісник Вінницького політехнічного інституту*, 2019. № 3, с. 40 – 49 Режим доступу: (фахове видання) http://nbuv.gov.ua/UJRN/vvpi_2019_3_7.

[13] Davies N.G., Kucharski A.J., Eggo R.M., Gimma A. (2020) “The effect of non-pharmaceutical interventions on COVID-19 cases, deaths and demand for hospital services in the UK: a modelling study”. *CMMID COVID-19 Working Group, W. John Edmunds medRxiv* 2020.04.01.20049908; doi: https://doi.org/10.1101/2020.04.01.20049908.

[14] Президія НАН України - Робоча група з математичного моделювання проблем, пов’язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx

[15] А. Лосенко, «Інформаційна технологія прогнозування часового ряду кількості хворих на коронавірус на основі моделі Facebook Prophet», *Вісник Вінницького політехнічного інституту*, № 5, с. 50-59, 2023. Режим доступу: https://doi.org/10.31649/1997-9266-2023-170-5-50-59

[16] 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*, вип. 5, с. 71–83, Лис 2020.

[17] 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*, вип. 6, с. 65–75, 2020. https://doi.org/10.31649/1997-9266-2020-153-6-65-75.

[18] В. Мокін, А. Лосенко, «Картування тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи протягом січня-березня 2021 року», *на L науково-технічній конференції підрозділів ВНТУ*, Вінниця, 10-12 березня 2021 р. – Електрон. текст. дані. – 2021. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2021/paper/view/12849

[19] В. Мокін, А. Лосенко, А. Ящолт, А. Гевеленко, «Прогнозування тижневих трендів кількості нових хворих на коронавірус у країнах світу», *Колективна монографія за матеріалами XX Міжнародної науково-практичної конференції «Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях»*, Жовтень 2021, с. 209-212. Електрон. текст. дані, 2021. – Режим доступу: https://itgip.org/wp-content/uploads/2021/10/1_zbirka_2021.pdf.

[20] В. Мокін, М. Дратований, А. Лосенко, С. Жуков, «Прогнозування хвиль коронавірусу на основі відновленої когнітивної карти міжрегіонального впливу», *Інформаційні технології та комп'ютерна інженерія*, № 3(52), с. 86–94, 2021. doi: 10.31649/1999-9941-2021-52-3-86-94.

[21] В. Мокін, А. Лосенко «Інформаційна технологія короткострокового прогнозування кількості нових хворих на коронавірус на основі моделі Facebook

Prophet», Інформаційно-комунікаційні технології для перемоги та відновлення, *Колективна монографія за матеріалами XXII Міжнародної науково-практичної конференції «Інформаційно-комунікаційні технології та сталий розвиток»*, Київ, 14-15 листопада 2023 р. За заг. ред. С.О. Довгого. – К.: ТОВ «Видавництво «Юстон», 2023. – С. 27-30. URL: https://itgip.org/wp-content/uploads/2023/11/1_zbirka_08_11_23-1-1.pdf

[22] В. Мокін, А. Лосенко, М. Дратований, «Інтелектуальна технологія аналізу та передбачення цін на вживані автомобілі», *Вісник Вінницького політехнічного інституту*, № 6, с. 62-72, 2019. doi: 10.31649/1997-9266-2019-147-6-62-72.

[23] Д. Шмундяк, А. Лосенко, В. Мокін, «Огляд підходів до визначення порядку Фур'є у моделі Facebook Prophet для моделювання сезонної складової часового ряду», *на LII Науково-технічній конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету*, Вінниця, 2023. Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/17200/14329.

[24] Прогноз розвитку епідемії COVID-19 в Україні в період 22 грудня 2021 р. – 4 січня 2022 р. «Прогноз РГ-58» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8551

[25] Прогноз розвитку епідемії COVID-19 в Україні в період 8–21 грудня 2021 р. «Прогноз РГ-57» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8502

[26] Прогноз розвитку епідемії COVID-19 в Україні в період 24 листопада – 7 грудня 2021 р. «Прогноз РГ-56» / В. Б. Мокін, А. В. Лосенко //

Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8446

[27] Прогноз розвитку епідемії COVID-19 в Україні в період 10 - 23 листопада 2021 р. «Прогноз РГ-55» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8386

[28] Прогноз розвитку епідемії COVID-19 в Україні в період 27 жовтня – 9 листопада 2021 р. «Прогноз РГ-54» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8334

[29] Прогноз розвитку епідемії COVID-19 в Україні в період 13–26 жовтня 2021 р. «Прогноз РГ-53» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8277

[30] Прогноз розвитку епідемії COVID-19 в Україні в період 10–23 листопада 2021 р. «Прогноз РГ-55» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8386

[31] Прогноз розвитку епідемії COVID-19 в Україні в період 29 вересня – 12 жовтня 2021 р. «Прогноз РГ-52» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8220

[32] Прогноз розвитку епідемії COVID-19 в Україні в період 15–28 вересня 2021 р. «Прогноз РГ-51» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8165

[33] Прогноз розвитку епідемії COVID-19 в Україні в період 20 квітня – 4 травня 2021 р. «Прогноз РГ-42» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7738

[34] Прогноз розвитку епідемії COVID-19 в Україні в період 7 квітня – 20 квітня 2021 р.«Прогноз РГ-41» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7694

[35] Прогноз розвитку епідемії COVID-19 в Україні в період 24 березня – 6 квітня 2021 р.«Прогноз РГ-40» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7641

[36] Прогноз розвитку епідемії COVID-19 в Україні в період 11–24 березня 2021 р. «Прогноз РГ-39» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7581

[37] Прогноз розвитку епідемії COVID-19 в Україні в період 23 лютого – 8 березня 2021 р. «Прогноз РГ-38» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією

коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7526

[38] Прогноз розвитку епідемії COVID-19 в Україні в період 9 – 22 лютого 2021 р. «Прогноз РГ-37» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7480

[39] Прогноз розвитку епідемії COVID-19 в Україні в період 26 січня – 8 лютого 2021 р. «Прогноз РГ-36» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7422

[40] Прогноз розвитку епідемії COVID-19 в Україні в період 11 – 25 січня 2021 р. «Прогноз РГ-35» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7356

[41] Прогноз розвитку епідемії COVID-19 в Україні в період 28 грудня 2020 р. – 11 січня 2021 р. «Прогноз РГ-34» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7322

[42] Прогноз розвитку епідемії COVID-19 в Україні в період 14–28 грудня 2020 р. «Прогноз РГ-32» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7277

[43] Прогноз розвитку епідемії COVID-19 в Україні в період 7–21 грудня 2020 р. «Прогноз РГ-31» / В. Б. Мокін, А. В. Лосенко // Робоча група

з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7238

[44] Прогноз розвитку епідемії COVID-19 в Україні в період 1–14 грудня 2020 р. «Прогноз РГ-30» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7215

[45] Прогноз розвитку епідемії COVID-19 в Україні в період 23 листопада – 7 грудня 2020 р. «Прогноз РГ-29» / В. Б. Мокін, А. В. Лосенко // Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні – Режим доступу: https://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7187

[46] COVID-19 Projections, Електронний ресурс. Режим доступу: https://covid19.healthdata.org/ukraine

[47] Ferguson Neil M et. al (2020) Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. 16 March 2020 Imperial College COVID-19 Response Team.

[48] Flaxman Seth et. al. (2020) Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. 30 March 2020 Imperial College COVID-19 Response Team.

[49] Hartemink N. A., Randolph 1, S. E., Davis S. A., Heesterbeek 1, J. A. P. (2008) The Basic Reproduction Number for Complex Disease Systems: Defining R0 for Tick-Borne Infections. The american naturalist june 2008 vol. 171, no. 6.

[50] Jiwei J., Jian D., Siyu L., Guidong L., Jingzhi L., Ben D., Guoqing W., Ran Z. (2020) Modeling the Control of COVID-19: Impact of Policy Interventions and Meteorological Factors. https: arxiv.org/abs/2003.02985.

[51] Lin Q., Zhaob S., Gao D., Lou Y., Yang S., Musa S. S., Wang M.H., Cai Y., Wang W., Yang L., He D. (2020) A conceptual model for the coronavirus disease 2019

(COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. International Journal of Infectious Diseases, 93, 211–216.

[52] Liu Q., Liu Z, Li D., Gao Z., Zhu J., Yang J., Wang Q. (2020) Assessing the Tendency of 2019-nCoV (COVID-19) Outbreak in China. medRxiv preprint doi: https: doi.org/10.1101/2020.02.09.20021444.

[53] Nicholas G. Davies, Petra Klepac, Yang Liu , Kiesha Prem , Mark Jit (2020) Age-dependent effects in the transmission and control of COVID-19 epidemics https: doi.org/10.1101/2020.03.24.20043018.

[54] Nishiuraa H., Lintona N.M., Akhmetzhanov A, R.(2020) Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*, 93, 284–286.

[55] Kiss I.Z., Miller J.C., Simon P.L. (2017) *Mathematics of Epidemics on Networks*. Springer International Publishing AG 2017 DOI 10.1007/978-3-319-50806-1.

[56] Wu J.T., Leung K., Leung G.M. (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet, V 395, Issue 10225,P689-697.

[57] Zhang G., Pang H., Xue Y., Zhou Y., Wang R. (2020) Forecasting and Analysis of Time Variation of Parameters of COVID-19 Infection in China Using An Improved SEIR Model DOI: 10.21203/rs.3.rs-16159/v1.

[58] Zhu G., Fu X., Chen G. (2012) Spreading dynamics and global stability of a generalized epidemic model on complex heterogeneous networks. Applied Mathematical Modelling, 36, 5808–5817.

[59] M.H.D.M. Ribeiro, R.G. da Silva, J.H.K. Larcher, V.C. Mariani, L..S. Coelho, “Ensemble Learning Models Coupled with Urban Mobility Information Applied to Predict COVID-19 Incidence Cases,” Modeling, Control and Drug Development for COVID-19 Outbreak Prevention. Studies in Systems, Decision and Control, vol 366, pp. 821-858, 2021. https://doi.org/10.1007/978-3-030-72834-2_24.

[60] Cao Z., Zhang Q., Lu X., Pfeiffer D., Wang L. Song H., Pei T. Jia Z., Zeng D.D. “Incorporating Human Movement Data to Improve Epidemiological Estimates for 2019-nCoV”. 2020. medRxiv preprint doi: https://doi.org/10.1101/2020.02.07.20021071.

[61] Робоча група з математичного моделювання проблем, пов’язаних з епідемією коронавірусу SARS-CoV-2 в Україні. Електронний ресурс. Режим доступу: https://www.nas.gov.ua/UA/Colegial/Pages/Default.aspx?CID=000000188

[62] Б. І. Мокін, В. Б. Мокін, і О. Б. Мокін, Математичні методи ідентифікації динамічних систем, навч. посіб. Вінниця, Україна: ВНТУ, 2010, 260 с. Режим доступу: http://mokin.com.ua/files/articles/59/34/Mokin_MMIDS_2010.pdf

[63] Мокін В.Б , Лосенко А.В. «COVID-19 Ukraine daily cases - EDA & forecasting» – Версія 7 – 12.10.2020 р.: [Електронний ресурс]. Режим доступу: https://www.kaggle.com/code/vbmokin/covid-19-ukraine-daily-cases-eda-forecasting

[64] Sarbjit Singh, Kulwinder Singh Parmar, Jatinder Kumar, Sidhu Jitendra Singh Makkhan, “Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19”, (https://www.sciencedirect.com/science/article/pii/S0960077920302666), ISSN 0960-0779, https://doi.org/10.1016/j.chaos.2020.109866.

[65] Sean J. Taylor & Benjamin Letham (2018) “Forecasting at Scale”, *The American Statistician*, 72:1, 37-45, DOI: 10.1080/00031305.2017.1380080

[66] Facebook Prophet Quick Start. Електронний ресурс. Режим доступу: https://facebook.github.io/prophet/docs/quick_start.html

[67] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow. (2021). “A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker).” *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01079-8

[68] COVID-19 Open Data. Електронний ресурс. Режим доступу: https://github.com/GoogleCloudPlatform/covid-19-open-data

[69] В. Мокін, А. Лосенко, «COVID-19-UA: Regression with Google mobility» / В. Б. Мокін, А. В. Лосенко // Kaggle. – Режим доступу: https://www.kaggle.com/code/vbmokin/covid-19-ua-regression-with-google-mobility

[70] Мокін В.Б., Лосенко А.В., «COVID-19 in Ukraine: Explanation of patterns» / В. Б. Мокін, А. В. Лосенко // Kaggle. – Режим доступу: https://www.kaggle.com/code/vbmokin/covid-19-in-ukraine-explanation-of-patterns

[71] Мокін В.Б , Лосенко А.В. «COVID in UA: Prophet with 4, 7d seasonality» – Версія 2 – 22.11.2020 р.: [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/covid-in-ua-prophet-with-4-7d-seasonality/output?scriptVersionId=47484394 Листопад, 22. 2020.

[72] Мокін В.Б, Лосенко А.В. «COVID-19 Ukraine daily cases - EDA & forecasting» – Версія 6 – 6.10.2020 р.: [Електронний ресурс]. Режим доступу: https://www.kaggle.com/code/vbmokin/covid-19-ukraine-daily-cases-eda-forecasting

[73] python-holidays. Електронний ресурс. Режим доступу: https://github.com/dr-prodigy/python-holidays

[74] NCEI Data Service API User Documentation Електронний ресурс. Режим доступу: https://www.ncei.noaa.gov/support/access-data-service-api-user-documentation

[75] Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., ... & Januschowski, T. (2022). “Deep learning for time series forecasting: Tutorial and literature survey”. *ACM Computing Surveys*, Surv. 55, 6, Article 121 (June 2023), 36 pages. https://doi.org/10.1145/3533382

[76] Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). “Time series modelling to forecast the confirmed and recovered cases of COVID-19”. *Travel medicine and infectious disease*, 37, 101742. DOI: https://doi.org/10.1016/j.tmaid.2020.101742

[77] Papastefanopoulos, V., Linardatos, P., & Kotsiantis, S. (2020). “COVID-19: a comparison of time series methods to forecast percentage of active cases per population”. *Applied sciences*, 10(11), 3880. DOI: https://doi.org/10.3390/app10113880

[78] Мокін В.Б , Лосенко А.В. «COVID-19 in 70 countries: daily Prophet forecast» – Версія 18 – 21.11.2020 р.: [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/covid-19-in-70-countries-daily-prophet-forecast?scriptVersionId=47433942 Листопад, 21. 2020.

[79] Мокін В.Б., Лосенко А.В., «COVID in UA: Prophet logistic with 4, 7d» / В. Б. Мокін, А. В. Лосенко // Kaggle. – Режим доступу: https://www.kaggle.com/code/vbmokin/covid-in-ua-prophet-logistic-with-4-7d/

[80] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 10. 31.01.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=53290019

[81] Мокін В.Б., Лосенко А.В., «COVID-19: Forecast trends for the many countries» / В. Б. Мокін, А. В. Лосенко // Kaggle. – Режим доступу: https://www.kaggle.com/datasets/vbmokin/covid19-forecast-trends-for-the-many-countries

[82] Anthony Goldbloom, “COVID-19 data from John Hopkins University” Kaggle. – Режим доступу: https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university

[83] Мокін В.Б , Лосенко А.В. «COVID-19: Holidays of countries» – версія датасета – 21.11.2020 р.: [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/covid19-holidays-of-countries Листопад, 21. 2020.

[84] Мокін В.Б., Лосенко А.В., «COVID in UA: Prophet - 4, 7d & tests-regr» / В. Б. Мокін, А. В. Лосенко // Kaggle. – Режим доступу: https://www.kaggle.com/code/vbmokin/covid-in-ua-prophet-4-7d-tests-regr

[85] Мокін В.Б., Лосенко А.В., «COVID in UA: Prophet logistic with 4, 7d» / В. Б. Мокін, А. В. Лосенко // Kaggle. – Режим доступу: https://www.kaggle.com/code/vbmokin/covid-in-ua-prophet-logistic-with-4-7d/

[86] Мокін В.Б , Лосенко А.В. «COVID-19 in 70 countries: daily Prophet forecast» – Версія 18 – 21.11.2020 р.: [Електронний ресурс]. Режим доступу:

https://www.kaggle.com/vbmokin/covid-19-in-70-countries-daily-prophet-forecast?scriptVersionId=47433942 Листопад, 21. 2020.

[87] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 7. 24.01.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/c/cassava-leaf-disease-classification/discussion/221957

[88] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 11. 07.02.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=53752993

[89] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 13. 14.02.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping/output?scriptVersionId=54907516

[90] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 15. 21.02.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56125897

[91] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 16. 01.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126354

[92] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 17. 07.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126983

[93] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 18. 21.02.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56125897

[94] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 19. 01.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126354

[95] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 20. 07.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126983

[96] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 21. 07.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126983

[97] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 22. 21.02.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56125897

[98] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 23. 01.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126354

[99] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 24. 07.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126983

[100] Мокін В.Б., Лосенко А.В. «COVID-19: Week trends 70 countries mapping». Version 25. 07.03.2021 // Kaggle. – Режим доступу: https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=56126983

ДОДАТКИ

ДОДАТОК А
АНАЛІТИЧНІ ЗВІТИ, НАДАНІ НАН УКРАЇНИ ПРОТЯГОМ 2020-2022 РР.

«Прогноз РГ-29» (23 листопада – 7 грудня 2020 р.)
3. Прогноз розвитку епідемії в Україні з використання статистичної моделі часових рядів Facebook Prophet.

У цьому документі представлено прогноз, обчислений двома різними підходами. Крім традиційної компартментної моделі було використано статистичну модель, яка, хоч і не має закладених фізичних механізмів розповсюдження епідемії, але дозволяє неявно враховувати багато інших факторів. В умовах стабільного поширення епідемії, а також зменшення обсягів тестування та збільшення відсотку позитивності тестів, а також при зміні погодних умов і карантинних обмежень використання такої моделі виглядає доцільним.

За допомогою методів статистичного аналізу було досліджено динаміку щоденної кількості нових хворих із липня 2020 року для виявлення закономірностей поширення епідемії, для дослідження впливу свят і псевдосвят (аномальних дат на кшталт державних свят, теплих днів без опадів тощо), впливу тижневої та інших видів сезонної мінливості та виявлення їхнього характеру.

Аналізувалися дані щодо нових виявлень на день і нових летальних випадків для України в цілому, коли спостерігалося невпинне зростання з 7-денною періодичністю – з 6 липня 2020 року. Було використано найсучаснішу модель Facebook Prophet, яка демонструє високу ефективність для моделювання часових рядів, що містять аномальні дати, різні види сезонності та лінійну чи нелінійну динаміку впливу різних складових моделі. Розроблено й застосовано алгоритм налаштування багатьох параметрів цієї моделі, який прогнозує дані на задану кількість днів вперед, але дані наявних спостережень за останні дні

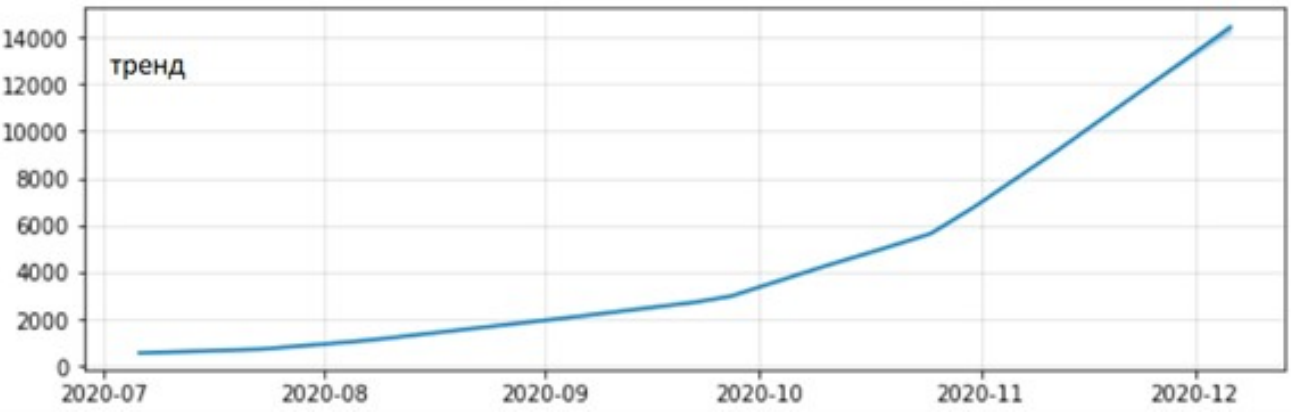
використовувалися для вибору найкращої моделі з налаштованих. Проведено дослідження для періоду прогнозування 14 днів.

Як аномальні дати (свята і псевдосвята) розглядалися дати державних свят, дати потенційного зростання кількості відпочивальників (коли було дуже тепло і не було опадів) та дати послаблення карантину за відкритими даними датасету Google-платформи «COVID-19 Open Data».

Спрощена модель застосовувалась як для України, так і для інших майже 70-ти країн світу, щодо яких у наборі даних «Facebook holidays» є інформація про державні свята (див. результати). Для України окремо було застосовано ефективнішу модель, яка за даними 6.07-8.11.2020 р. дала прогноз на 9.11-22.11 із сумарною відносною похибкою за всі 14 днів – 2,2% (рис.30, 31, прогноз даних на 2 тижні вперед див. у таблиці 7). Цей же, ефективніший, алгоритм налаштування моделі було застосовано для моделювання щоденної кількості смертельних випадків в Україні, але він поки не забезпечив такої ж високої точності – лише 13,8% (див. рис.32, 33, прогноз даних на 2 тижні вперед див. у таблиці 8). Це говорить про те, що в динаміці смертності фактор випадковості є суттєвішим, ніж для кількості нових випадків.

Рис.30. Щоденна кількість нових підтверджених випадків хворих на COVID-19 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%)



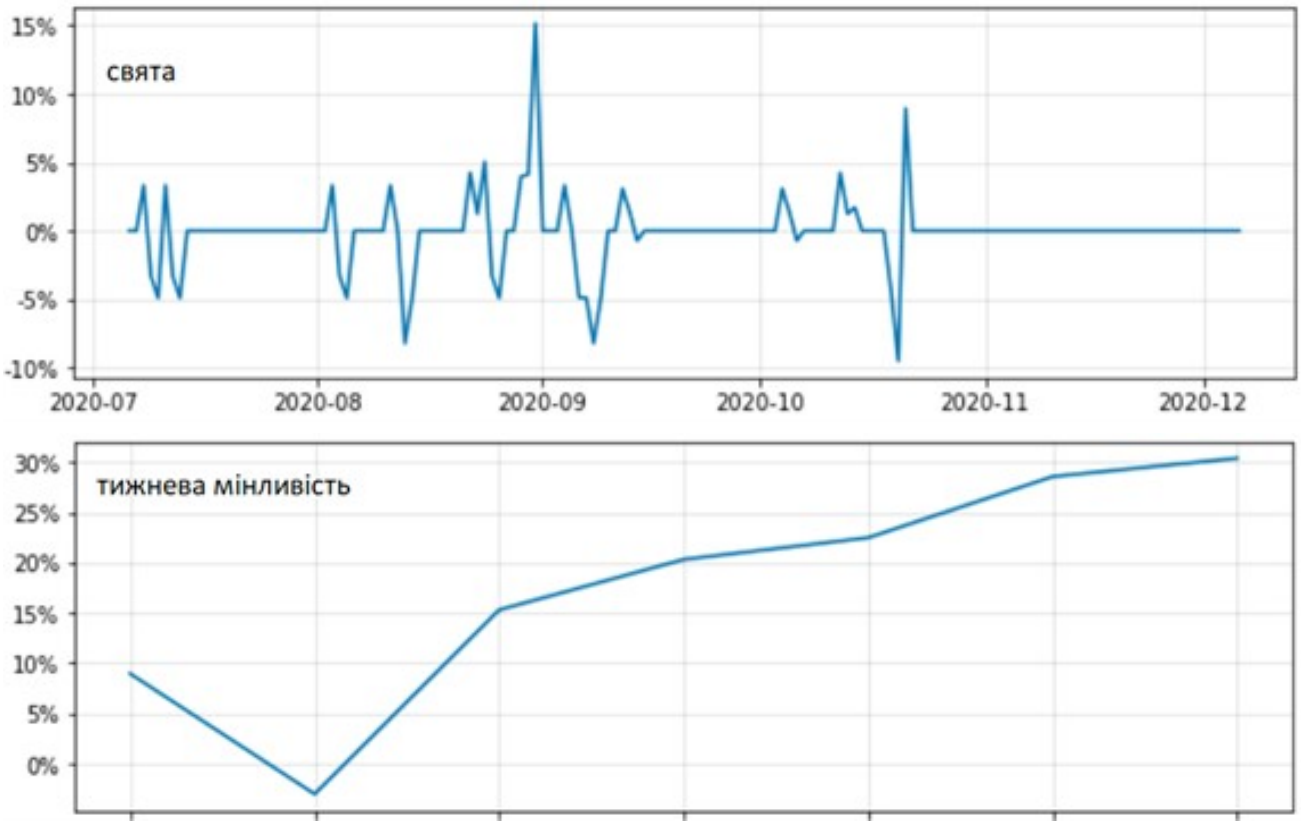


Рис.31. Складові моделі на рис.30 для моделювання та прогнозування щоденної кількості нових підтверджених випадків хворих на COVID-19 в Україні з 6 липня 2020 р.:

вгорі – основний тренд; посередині – динаміка впливу аномальних дат (державних свят, дуже теплих днів без опадів і дат послаблення карантину) зі зсувом в 1 тиждень; унизу – тижнева мінливість

Таблиця 7. Прогноз кількості нових підтверджених випадків хворих на COVID-19 в Україні за моделлю з урахуванням впливу аномальних дат

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

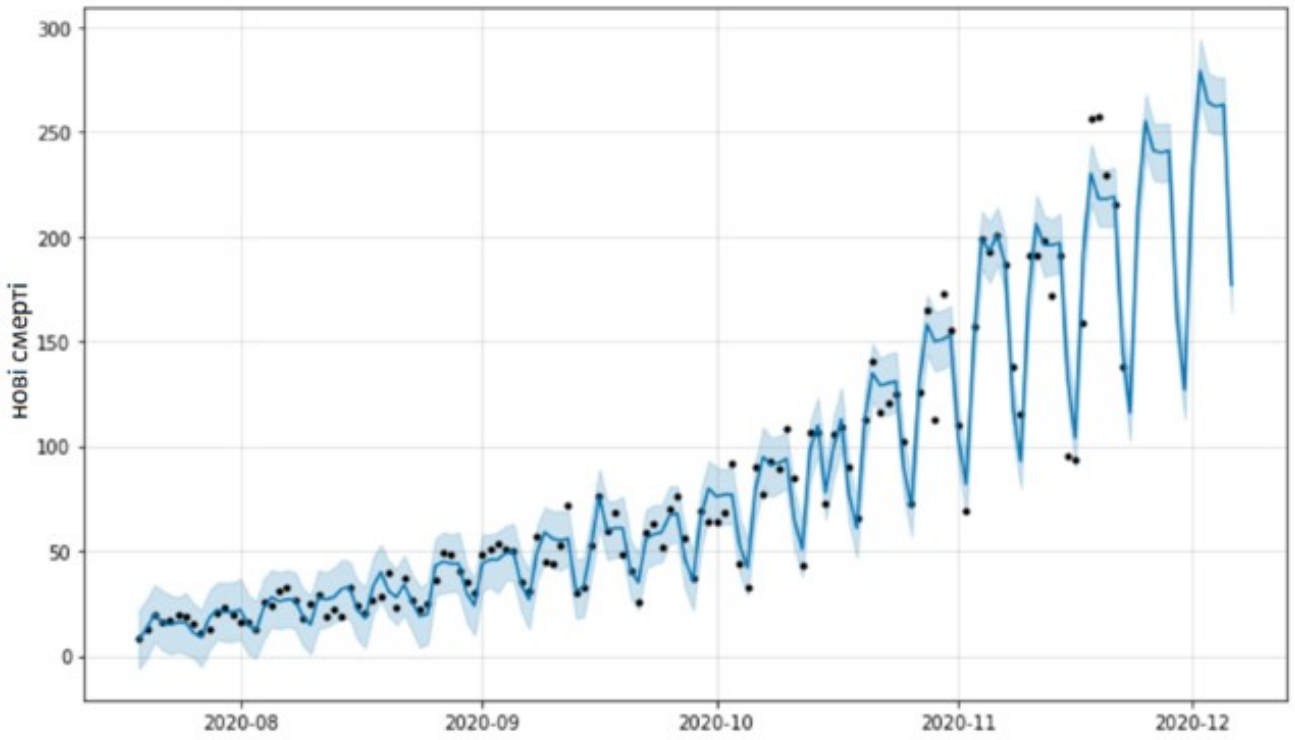


Рис.32. Щоденна кількість смертельних випадків хворих на COVID-19 в Україні з 6 липня 2020 р.:

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 – 13,8%)

Рис.33. Складові моделі на рис.32 для моделювання та прогнозування щоденної кількості смертельних випадків хворих на COVID-19 в Україні з 6 липня 2020 р.:

вгорі – основний тренд; посередині – динаміка впливу аномальних дат (державних свят, дуже теплих днів без опадів і дат послаблення карантину) зі зсувом у 3 тижні (додаються до основного тренду); внизу – тижнева мінливість

Таблиця 8. Прогноз кількості смертельних випадків хворих на COVID-19 в Україні за моделлю з урахуванням впливу аномальних дат

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
23.11.2020	103	116	129
24.11.2020	199	213	228
25.11.2020	240	255	268
26.11.2020	227	241	254
27.11.2020	226	240	254
28.11.2020	227	241	254
29.11.2020	149	163	177
30.11.2020	113	127	141
01.12.2020	220	233	248
02.12.2020	265	279	294
03.12.2020	250	264	278
04.12.2020	249	262	276
05.12.2020	249	263	276
06.12.2020	164	177	191

Для України вплив свят виявився не досить впливовим, порівняно, наприклад, із такими країнами, як Швеція, Франція, Іспанія, де модель з урахуванням свят і псевдосвят має похибку майже вдвічі меншу, ніж модель без

урахування таких аномальних дат. В Україні ж спрощена модель з урахуванням аномальних дат дає похибку 2,55%, а без їх урахування – 2,57%.

Виявлено, що, крім тижневої мінливості, спостерігається ще й 4-денна сезонність, яка теж демонструє нелінійне зростання, причому саме зростання стається що дві доби – її врахування дозволило знизити похибку з 2,38% до 2,2%. Це означає, що існує додаткова сукупність факторів, яка прискорює вплив іще швидше, ніж за кожні 7 днів.

Присутнє нелінійне зростання кількості летальних випадків із кожним тижнем, але часто трапляються аномальні значення, які не можна пояснити впливом аномальних дат, навіть узятих зі значно більшим зсувом, ніж у випадку прогнозування нових підтверджених випадків захворювань. Вплив аномальних дат (із 3-тижневим зсувом) може бути досить суттєвим.

Отримана похибка 13,8% є доволі значною для не достатньої впевненості в отриманих результатів і потребує подальшого уточнення. 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 і що отримані результати дають остаточні відповіді на поставлені питання.

Обчислення за допомогою моделі Prophet і аналіз отриманих результатів виконали завідувач кафедри системного аналізу та інформаційних технологій (САІТ) Вінницького національного технічного університету (ВНТУ) доктор технічних наук професор В.Б. Мокін і аспірант кафедри САІТ ВНТУ А.В. Лосенко.

«Прогноз РГ-30» (1–14 грудня 2020 р.)

Прогноз розвитку епідемії в Україні з використання статистичної моделі часових рядів Facebook Prophet.

У цьому документі представлено прогноз, обчислений двома різними підходами. Крім традиційної компартментної моделі, було використано статистичну модель, яка, хоч і не має закладених фізичних механізмів розповсюдження епідемії, але дозволяє неявно враховувати багато інших факторів. В умовах стабільного поширення епідемії, а також зменшення обсягів тестування та збільшення відсотка позитивності тестів, а також при зміні погодних умов і карантинних обмежень використання такої моделі видається доцільним.

За допомогою методів статистичного аналізу було досліджено динаміку щоденної кількості нових хворих із липня 2020 року для виявлення закономірностей поширення епідемії, для дослідження впливу свят і псевдосвят (аномальних дат на кшталт державних свят, теплих днів без опадів тощо), впливу тижневої та інших видів сезонної мінливості і виявлення їхнього характеру.

Аналізувалися дані щодо нових виявлень на день і нових летальних випадків для України загалом, коли спостерігалося невпинне зростання з 7-денною періодичністю – з 6 липня 2020 року. Було використано найсучаснішу модель Facebook Prophet, яка демонструє високу ефективність для моделювання часових рядів, що містять аномальні дати, різні види сезонності та лінійну чи нелінійну динаміку впливу різних складових моделі. Розроблено й застосовано алгоритм налаштування багатьох параметрів цієї моделі, який прогнозує дані на задану кількість днів, але дані наявних спостережень за останні дні використовувалися для вибору найкращої моделі з налаштованих. Проведено дослідження для періоду прогнозування 14 днів.

Як аномальні дати (свята і псевдосвята) розглядалися дати державних свят, дати потенційного зростання кількості відпочивальників (коли було дуже тепло і не було опадів) та дати послаблення карантину за відкритими даними датасету Google-платформи «COVID-19 Open Data».

Спрощена модель застосовувалась як для України, так і для інших майже 70-ти країн світу, щодо яких у датасеті Facebook holidays є інформація про державні свята цих країн (див. результати). Для України окремо було застосовано ефективнішу модель, яка за даними 6.07–15.11 дала прогноз на 16.11–29.11 із сумарною відносною похибкою за всі 14 днів – 3,2% (рис.17), прогноз даних на 2 тижні див. у таблиці 3).

Рис.17. Щоденна кількість нових підтверджених випадків хворих на COVID-19 в Україні з 6 липня 2020 р.: чорні крапки – дані спостережень до 29.11.2020 р., синя лінія – результат моделювання і прогнозування на 2 тижні до 13.12.2020 р. за моделлю на основі Facebook Prophet

Таблиця 3. Прогноз кількості нових підтверджених випадків хворих на COVID-19 в Україні за моделлю з урахуванням впливу аномальних дат

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
30.11.2020	12070	12360	12656
01.12.2020	14345	14611	14925
02.12.2020	15468	15738	16024
03.12.2020	16195	16508	16812
04.12.2020	17311	17619	17900
05.12.2020	17769	18067	18360
06.12.2020	14734	15031	15328
07.12.2020	13305	13613	13900
08.12.2020	15759	16060	16376
09.12.2020	17107	17418	17732
10.12.2020	17947	18262	18580
11.12.2020	18980	19319	19652
12.12.2020	19424	19768	20130
13.12.2020	16180	16555	16885

Щодо нових підтверджених випадків здійснено порівняння прогнозу на 2 тижні, зробленого рівно тиждень тому, і нового прогнозу.

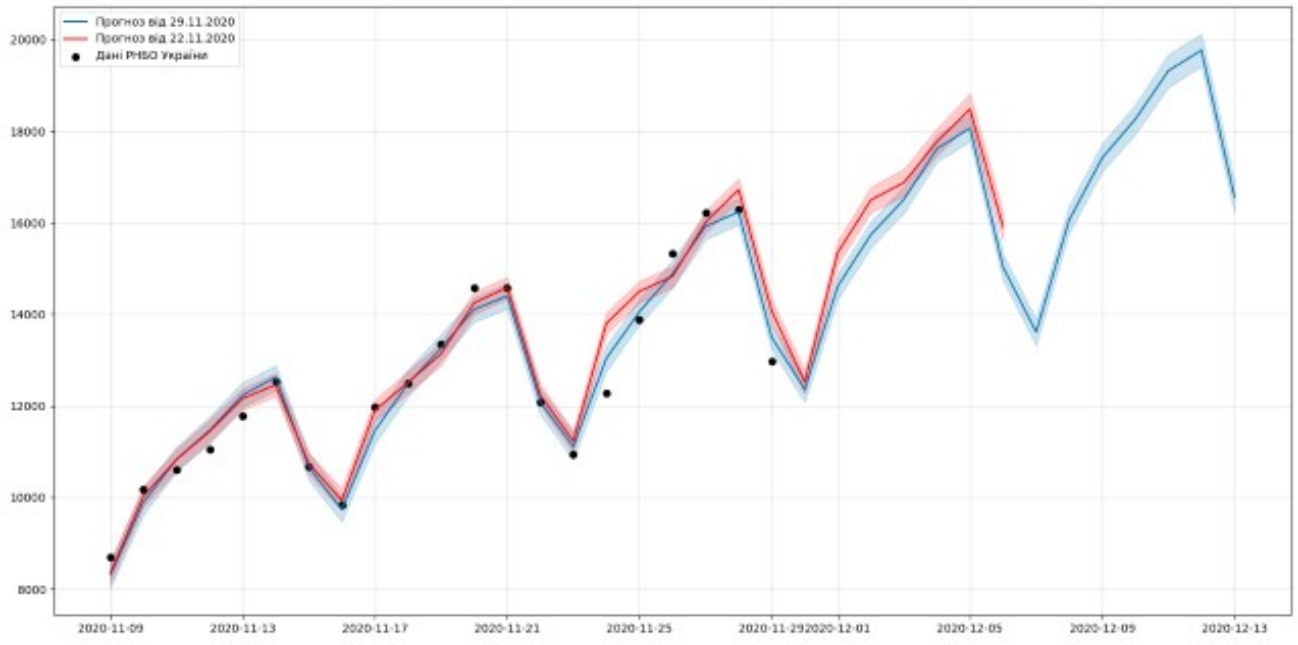


Рис.18. Останні 3 тижні спостережень і 2 тижні прогнозу

Оскільки дані про летальні випадки демонструють аномальну поведінку і не мають стабільної залежності від днів тижня, то було застосовано згладжування шляхом використання ковзного середнього з 7-денним вікном (рис.19, таблиця 4).

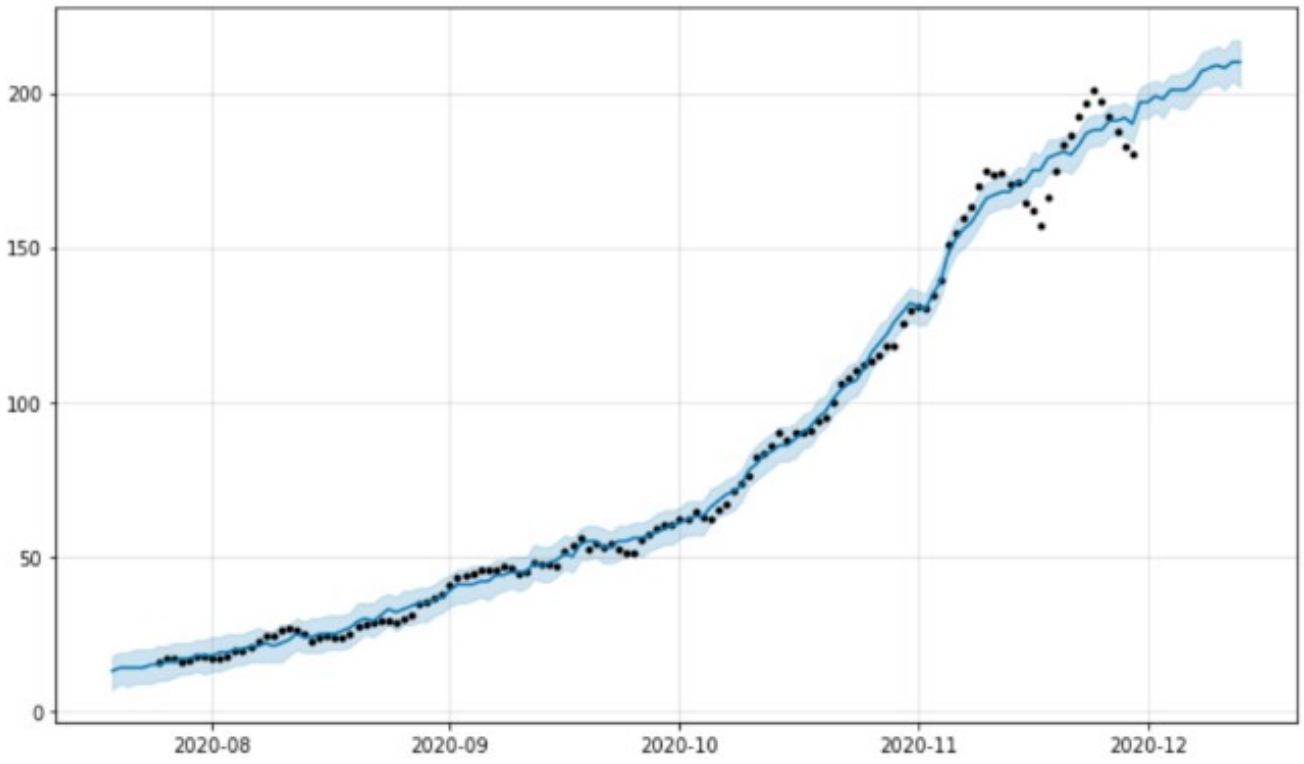


Рис.19. Усереднена за 7-днів щоденна кількість летальних випадків хворих на коронавірус в Україні з 6 липня 2020 р.: чорні крапки – дані спостережень до 29.11.2020 р., синя лінія – результат моделювання і прогнозування на 2 тижні до 6.12.2020 р. за моделлю на основі Facebook Prophet

Таблиця 4. Прогноз кількості смертельних випадків хворих на COVID-19 в Україні за моделлю з урахуванням впливу аномальних дат

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
30.11.2020	192	197	202
01.12.2020	192	197	203
02.12.2020	194	199	204
03.12.2020	192	198	203
04.12.2020	196	201	206
05.12.2020	195	201	206
06.12.2020	105	201	206
07.12.2020	198	203	209
08.12.2020	201	207	213
09.12.2020	202	208	214
10.12.2020	203	209	215
11.12.2020	201	208	214
12.12.2020	204	210	217
13.12.2020	202	210	217

Аналіз щодо кількості нових підтверджених і летальних випадків захворювань на коронавірус показав таке:

- Наявна модель іще мінімум 4 тижні не зможе оцінити вплив карантину вихідного дня, оскільки аномальні дати беруться зі зсувом на тиждень, а потім необхідно ще мінімум 2 тижні, щоб налаштувати модель за даними, де враховується ця аномальна дата (точність розраховується за даними цих 2-х тижнів), і ще потім необхідно 2 тижні, щодо яких є дані спостережень, – для валідаційного датасету, за допомогою якого відбирається найкраща модель.
- Для України вплив свят та інших аномальних дат, як і раніше, не є досить суттєвим, порівняно, наприклад, із такими країнами, як Сінгапур, Франція, Ізраїль, Фінляндія, Сербія, Швеція, де модель з урахуванням свят і псевдосвят має похибку майже у 1,5–2 рази меншу, ніж модель без урахування таких аномальних дат. В Україні ж спрощена модель з

урахуванням аномальних дат дає похибку 3,63%, а без їх урахування – 3,65%. Ефективніша ж модель дає похибку 3,20% проти 3,28%

- Модель для летальних випадків з 7-денним ковзним середнім демонструє постійне нелінійне зростання, яке, однак, дещо уповільнилось останнім часом. Відбувається явно нелінійне зростання даних із кожним тижнем, але часто трапляються щодобові аномальні значення, які не можна пояснити впливом аномальних дат.
- Вплив аномальних дат на летальні випадки може бути досить значним, оскільки модель з їх урахуванням дає точність 11,5%, а модель без їх урахування – 17,6%

Отримана похибка 11,5% для летальних випадків є досить значною, що не дає достатньої впевненості в отриманих результатах і потребує подальшого уточнення.

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 що отримані результати дають остаточні відповіді на поставлені питання.

Обчислення за допомогою моделі Prophet і аналіз отриманих результатів виконали завідувач кафедри системного аналізу та інформаційних технологій (САІТ) Вінницького національного технічного університету (ВНТУ) доктор технічних наук професор В.Б. Мокін і аспірант кафедри САІТ ВНТУ А.В. Лосенко.

«Прогноз РГ-31» (7–21 грудня 2020 р.)

Прогноз розвитку епідемії в Україні з використанням статистичної моделі часових рядів Facebook Prophet.

За допомогою методів статистичного аналізу було досліджено динаміку щоденної кількості нових хворих із липня 2020 року для виявлення закономірностей поширення епідемії, для дослідження впливу свят і псевдосвят (аномальних дат на кшталт державних свят, теплих днів без опадів тощо), впливу тижневої та інших видів сезонної мінливості і виявлення їхнього характеру.

Аналізувалися дані про нові виявлення на день і нові летальні випадки для України загалом, коли спостерігалося невпинне зростання з 7-денною періодичністю – з 6 липня 2020 року. Було використано найсучаснішу модель Facebook Prophet, яка демонструє високу ефективність для моделювання часових рядів, що містять аномальні дати, різні види сезонності та лінійну чи нелінійну динаміку впливу різних складових моделі. Розроблено й застосовано алгоритм налаштування багатьох параметрів цієї моделі, який прогнозує дані на задану кількість днів, але дані наявних спостережень за останні дні використовувалися для вибору найкращої моделі з налаштованих. Проведено дослідження для періоду прогнозування 14 днів.

Як аномальні дати (свята і псевдосвята) розглядалися дати державних свят, дати потенційного зростання кількості відпочивальників (коли було дуже тепло і не було опадів) та дати послаблення карантину за відкритими даними датасету Google-платформи «COVID-19 Open Data».

Побудовано модель, яка за даними 6.07–23.11 дала прогноз на 24.11-07.12 із сумарною відносною похибкою за всі 14 днів – 15,6% (рис.17, прогноз даних на 2 тижні – до 21.12 див. у таблиці 3).

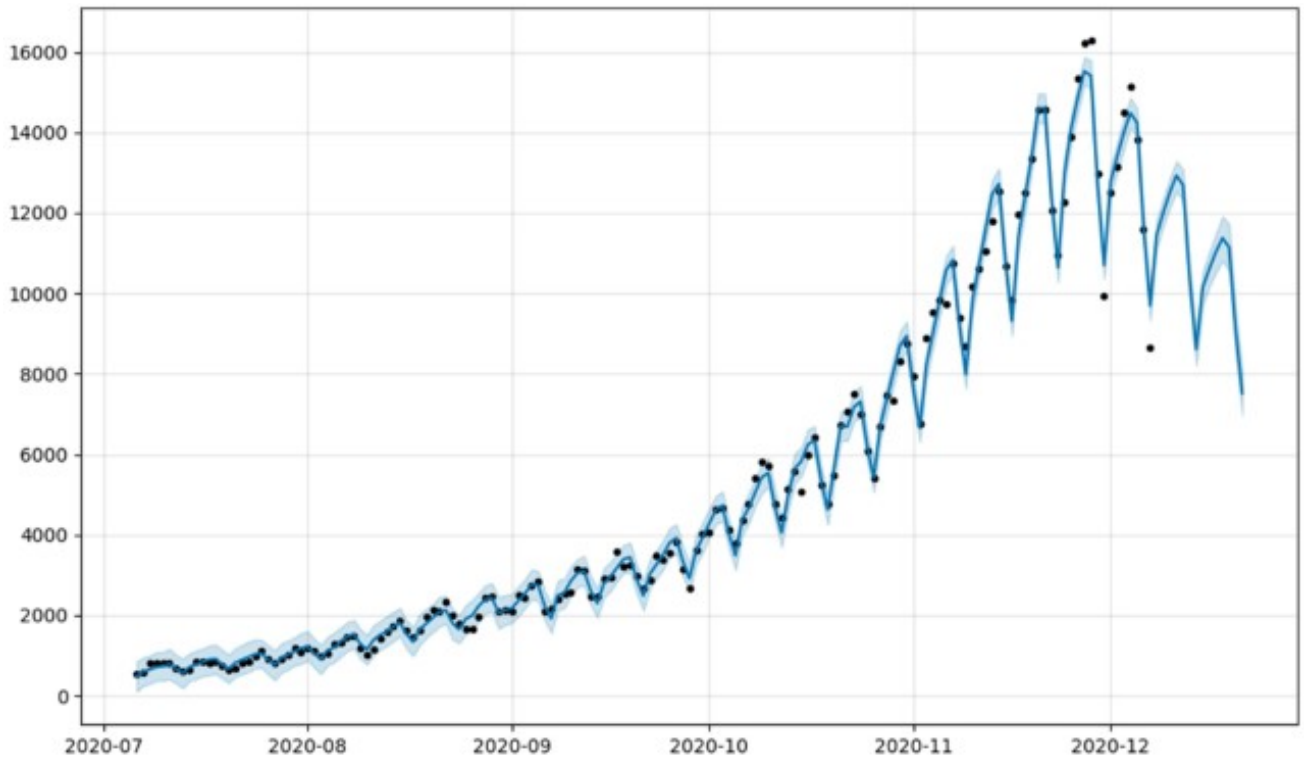


Рис.17. Щоденна кількість нових підтверджених випадків хворих на COVID-19 в Україні з 6 липня 2020 р.: чорні крапки – дані спостережень до 07.12.2020 р., синя лінія – результат моделювання і прогнозування на 2 тижні до 21.12.2020 р. за моделлю на основі Facebook Prophet

Таблиця 3. Прогноз кількості нових підтверджених випадків хворих на COVID-19 в Україні за моделлю з урахуванням впливу аномальних дат

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
08.12.2020	11122	11460	11845
09.12.2020	11677	12005	12357
10.12.2020	12141	12513	12878
11.12.2020	12534	12932	13296
12.12.2020	12289	12695	13076
13.12.2020	9966	10364	10761
14.12.2020	8206	8610	8996
15.12.2020	9705	10150	10543
16.12.2020	10186	10611	11042
17.12.2020	10523	11034	11494
18.12.2020	10818	11378	11917
19.12.2020	10550	11143	11717
20.12.2020	8483	9075	9593
21.12.2020	6973	7519	8038

Щодо нових підтверджених випадків порівняно прогноз на 2 тижні, зроблений за допомогою моделі Facebook Prophet 1 і 2 тижні тому, та новий прогноз (рис.18).

Рис.18. Останні 3 тижні спостережень і 2 тижні прогнозу

Вивчено також проблему великої похибки у 15,6%. Припускається, що на це могло вплинути зниження кількості тестів останнім часом. Тому для аналізу динаміки кількості тестувань на ПЛР теж було застосовано аналогічну модель Prophet з точністю 7,4% – і це дозволило виокремити чіткий тренд (рис.19).

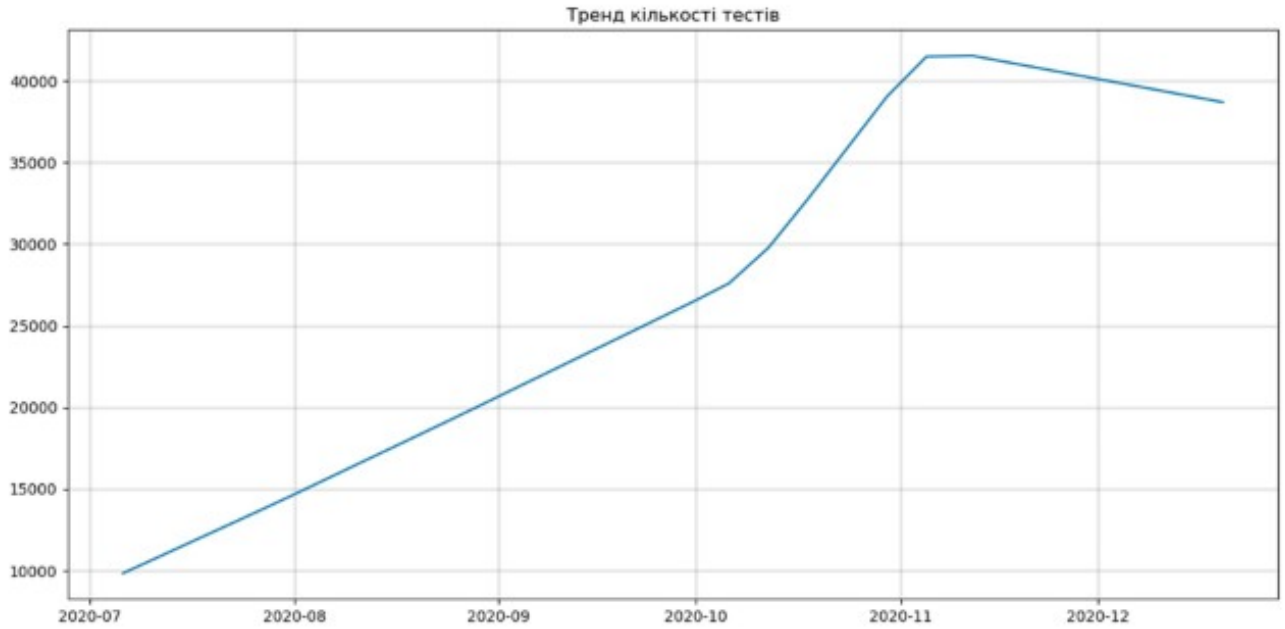


Рис.19. Тренд щоденної кількості ПЛР-тестів на COVID-19 в Україні з 6 липня 2020 р. – результат моделювання з 07.12.2020 р. і прогнозування на 2 тижні до 21.12.2020 р. за моделлю на основі Facebook Prophet

На рис.19 видно мінімум 3 характерні ділянки зміни тренду, зокрема, в цей час спостерігається ділянка, що розпочалася 06.11.2020 р. Тому було здійснено повторне моделювання вже тільки для цієї ділянки з прогнозуванням на 1 тиждень. Можливо, слід було би брати такі дані зі зсувом, оскільки результати тестів можуть з'являтися дещо пізніше, ніж реєструються, але для врахування цього – замало надійної інформації. За даними 6.11–30.11 ця нова модель дала прогноз на 01.12–07.12 із сумарною відносною похибкою за всі 7 днів – 6,4% (рис.20), що майже втричі точніше за прогноз даних за моделлю для всієї цієї хвилі (прогноз на 1 тиждень див. у таблиці 4).

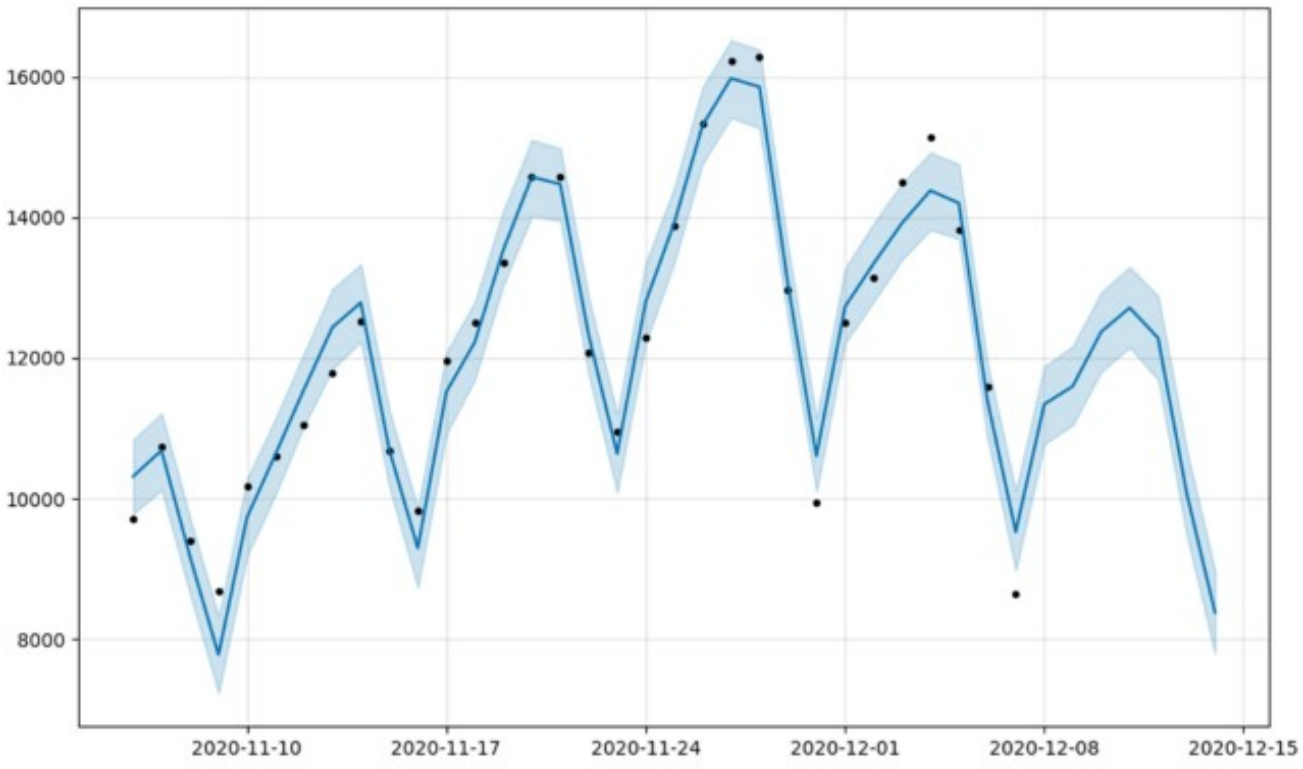


Рис.20. Щоденна кількість нових підтверджених випадків хворих на COVID-19 в Україні з 6 листопада 2020 р.: чорні крапки – дані спостережень до 07.12.2020 р., синя лінія – результат моделювання і прогнозування на 1 тиждень до 14.12.2020 р. за моделлю на основі Facebook Prophet

Таблиця 4. Прогноз кількості нових підтверджених випадків хворих на COVID-19 в Україні за моделлю Prophet за даними з 6 листопада 2020 р.

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
08.12.2020	10777	11341	11883
09.12.2020	11050	11598	12167
10.12.2020	11807	12373	12937
11.12.2020	12151	12715	13296
12.12.2020	11691	12284	12881
13.12.2020	9488	10082	10699
14.12.2020	7791	8379	8964

Аналогічно щодо цих нових підтверджених випадків порівняно прогноз, підготовлений за допомогою моделі Facebook Prophet 1 і 2 тижні тому, та новий прогноз (рис.21).

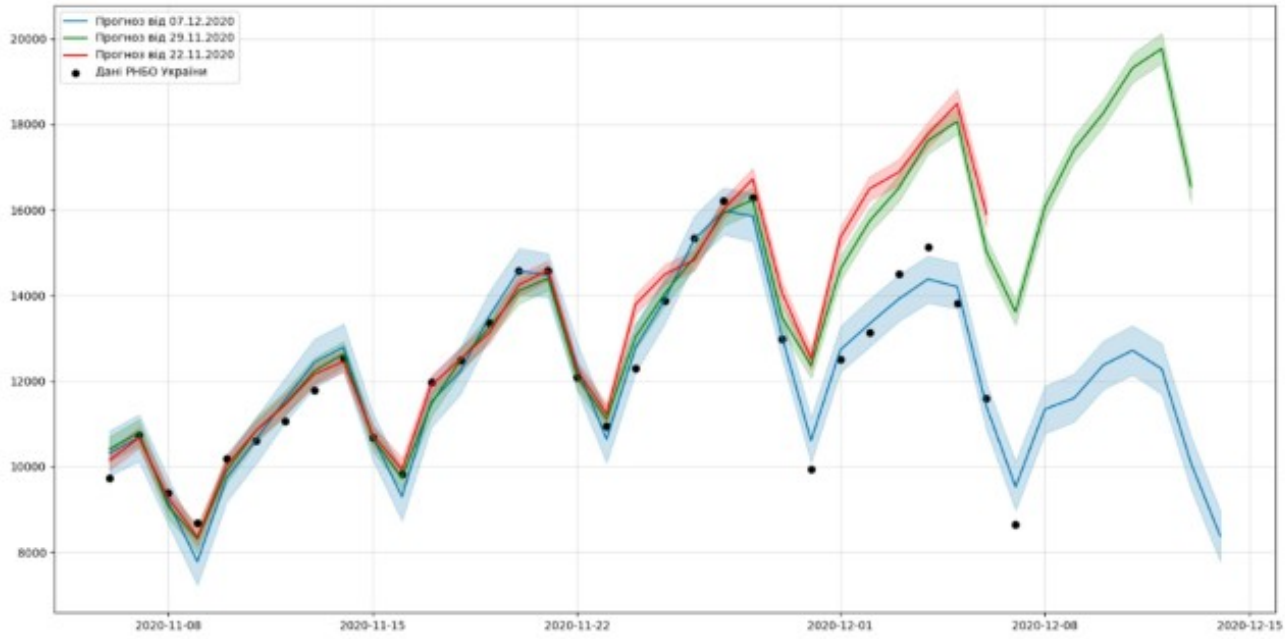


Рис.21. Останні 4 тижні спостережень і 1 тиждень прогнозу

Оскільки дані про летальні випадки демонструють аномальну поведінку і не

узалежнені стабільно від днів тижня, то було застосовано згладжування шляхом використання ковзного середнього з 7-денним вікном (рис.22, таблиця 5).

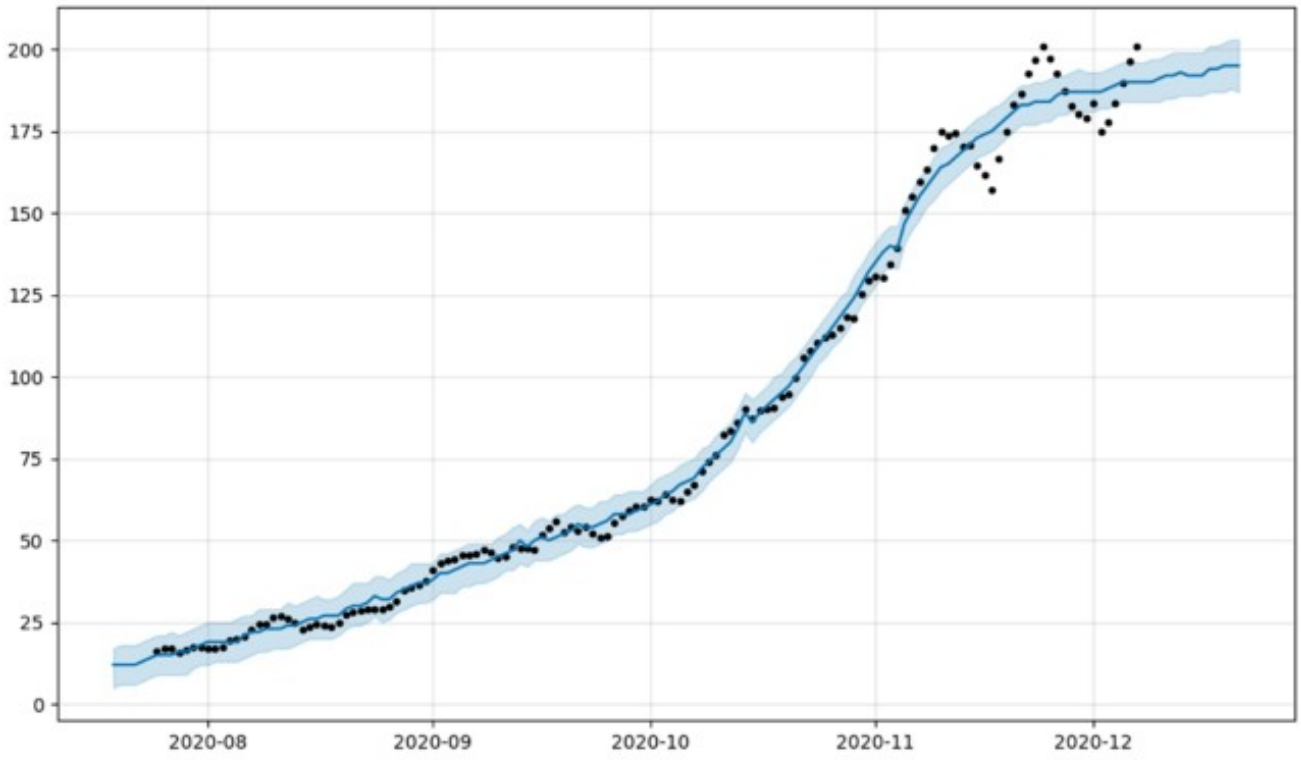


Рис.22. 7-денне ковзне середнє щоденної кількості летальних випадків в Україні з 6 липня 2020 р.: чорні крапки – 7-денне ковзне середнє даних спостережень до 07.12.2020 р., синя лінія – результат моделювання і прогнозування на 2 тижні до 21.12.2020 р. за моделлю на основі Facebook Prophet

Таблиця 5. Прогноз 7-денного ковзного середнього кількості смертельних випадків хворих на COVID-19 в Україні за моделлю з урахуванням впливу аномальних дат

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
08.12.2020	184	190	196
09.12.2020	184	190	197
10.12.2020	184	191	197
11.12.2020	185	192	198
12.12.2020	185	192	199
13.12.2020	186	193	199
14.12.2020	186	192	199
15.12.2020	186	192	199
16.12.2020	186	192	199
17.12.2020	187	194	201
18.12.2020	187	194	201
19.12.2020	187	195	202
20.12.2020	188	195	203
21.12.2020	187	195	203

Аналіз щодо кількості нових підтверджених і летальних випадків захворювань показав таке:

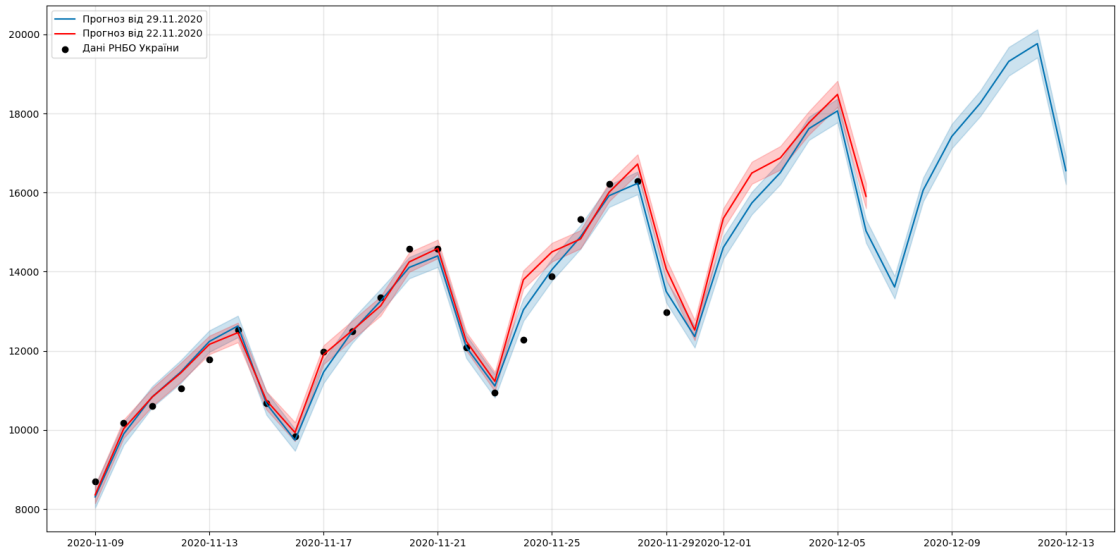
- для України вплив свят та інших аномальних дат, як і раніше, не є досить значним;
- доцільно моделювати дані за окремими характерними ділянками зміни кількості тестів; такий алгоритм моделювання, застосований для ділянки, яка почалася з 6 листопада, дав вищу точність – 6,4%, що означає, що кількість виявлених нових хворих може залежати від обсягів тестування;
- модель для летальних випадків із 7-денним ковзним середнім продовжує демонструвати стале нелінійне зростання, котре дещо уповільнилось останнім часом; спостерігається зростання даних із кожним тижнем, але часто трапляються щодобові аномальні значення, які не можна пояснити впливом аномальних дат.

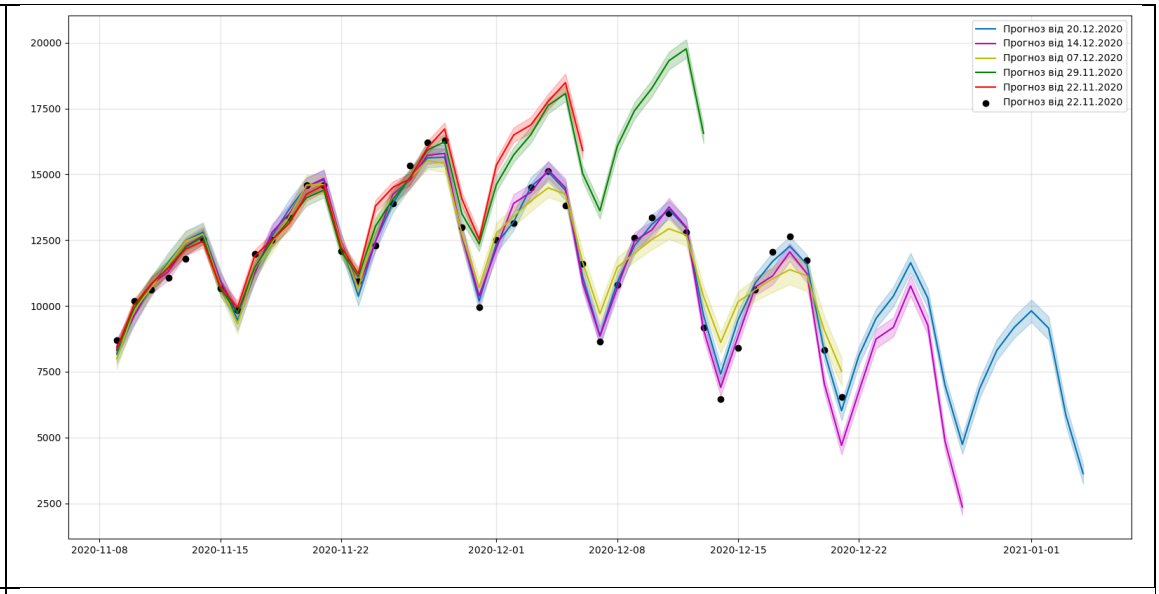
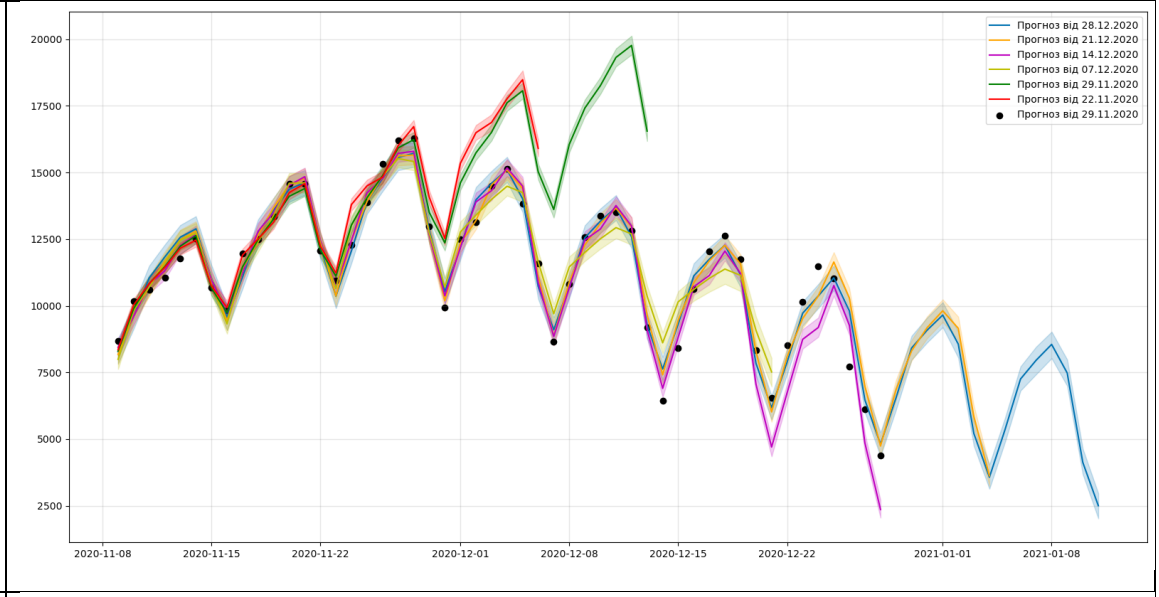
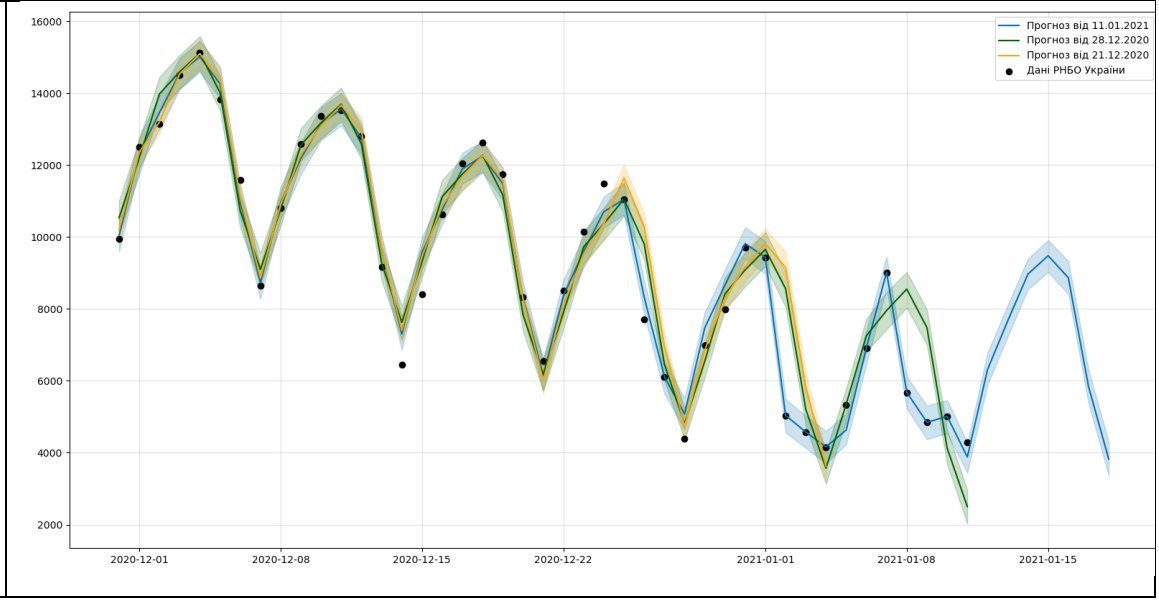
Отримана похибка 9,3% для 7-денного ковзного середнього кількості летальних випадків є досить суттєвою – це не дає достатньої впевненості в отриманих результатах, тому потрібне подальше уточнення.

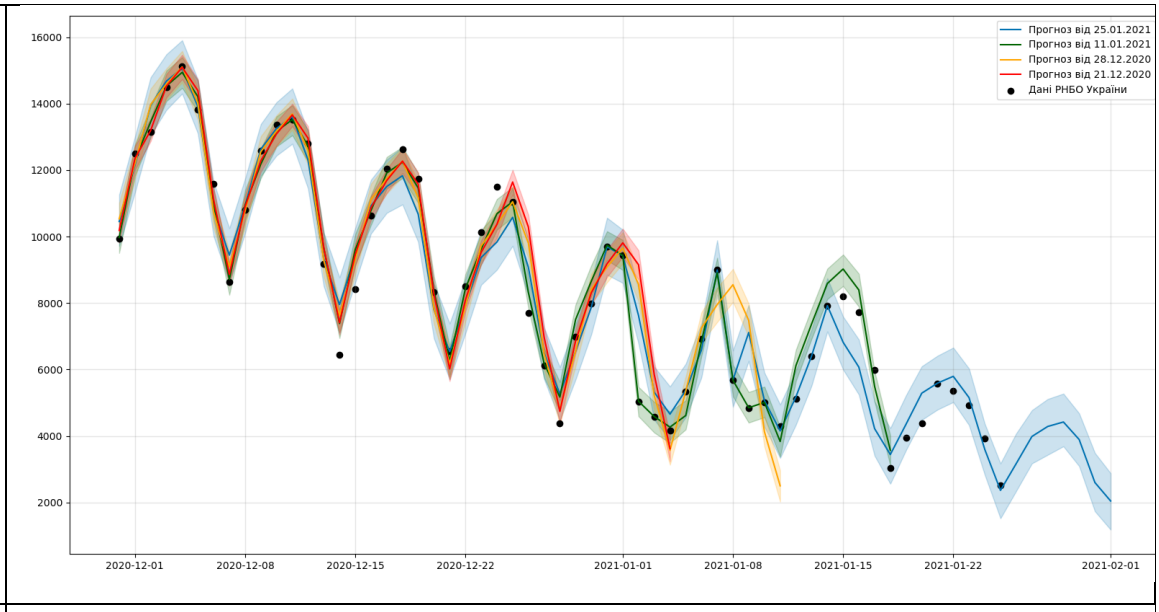
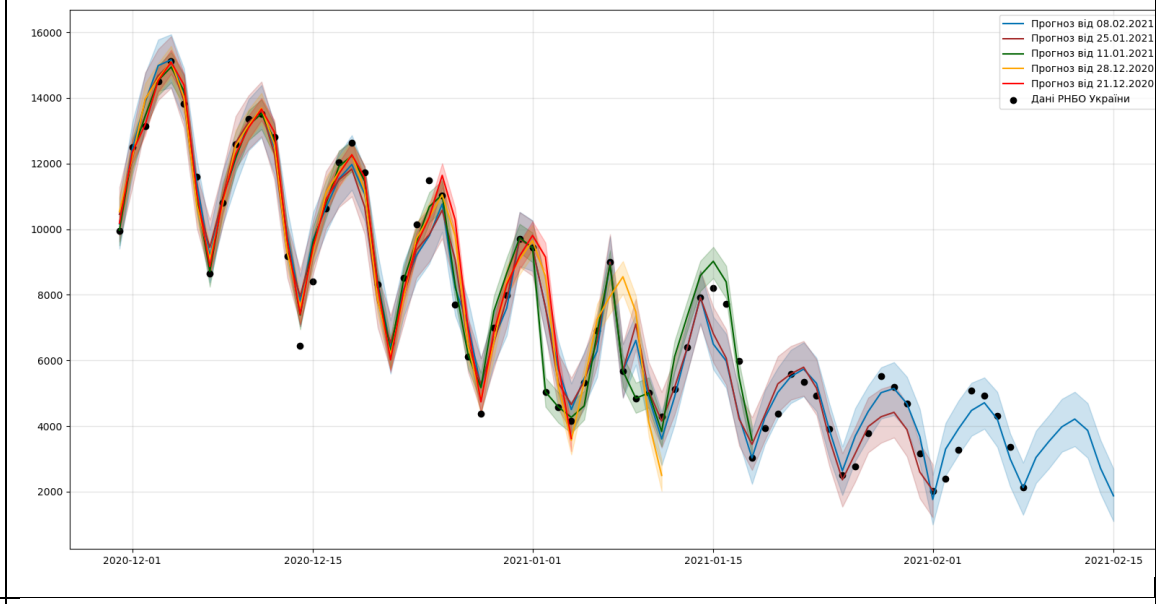
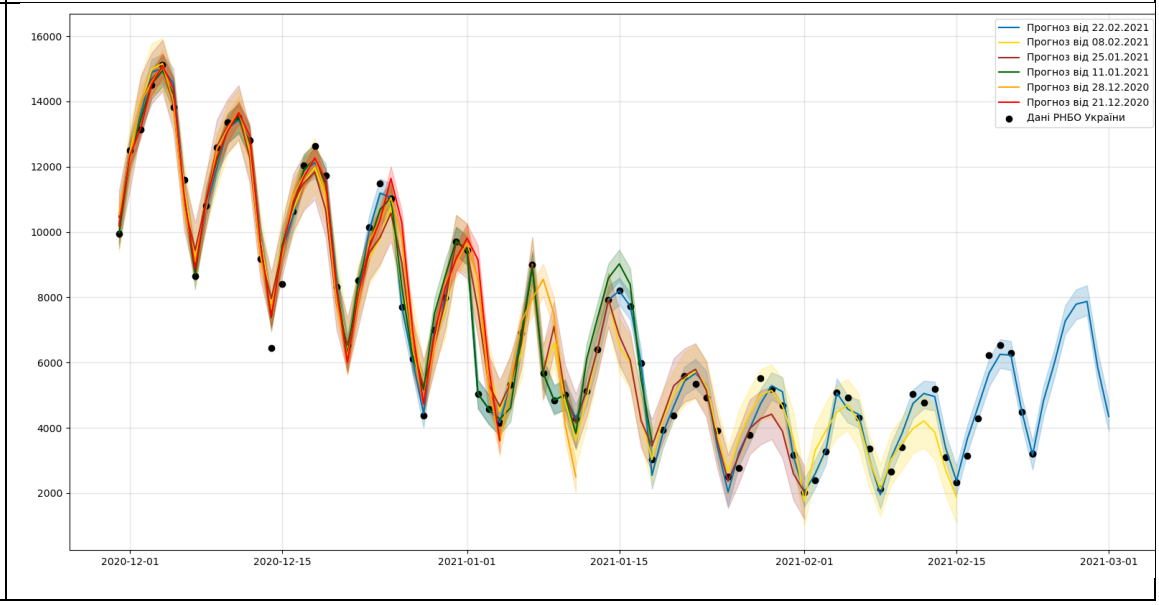
Те, що в моделях не перебиралися повністю всі можливі значення параметрів, не враховувалась явно динаміка інших факторів, на жаль, не дає впевненості в тому, що ці моделі можна використовувати для довгострокового прогнозування та що отримані результати остаточно відповідають на поставлені питання.

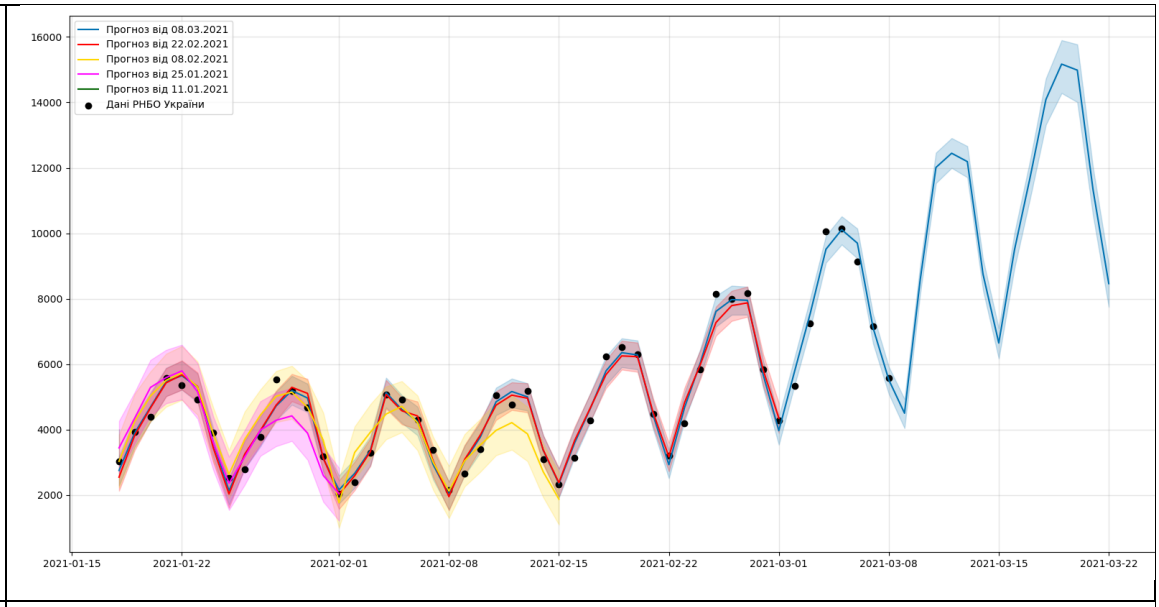
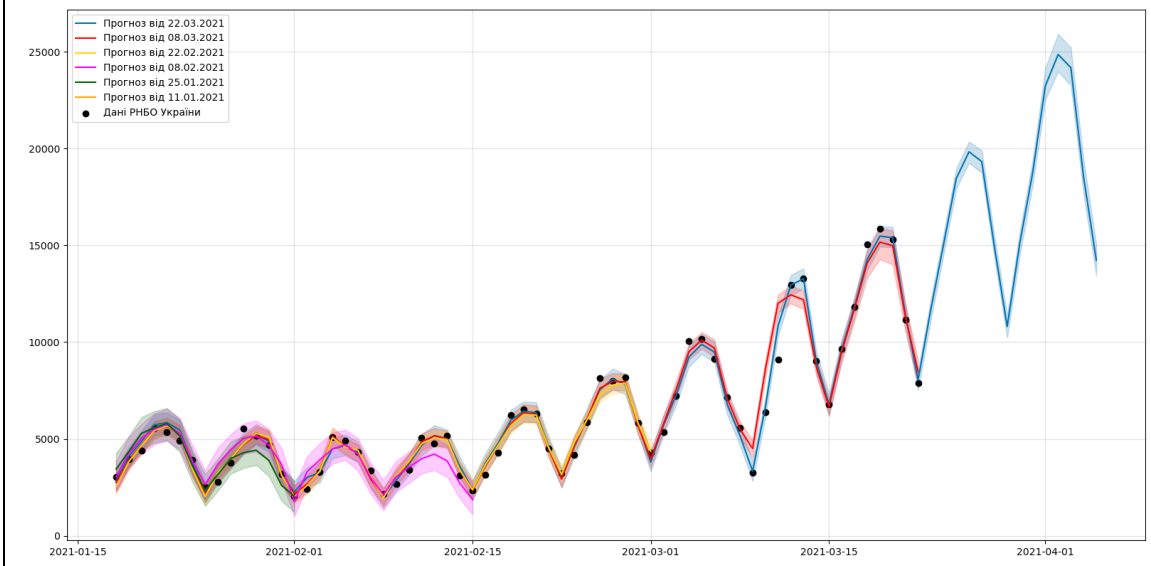
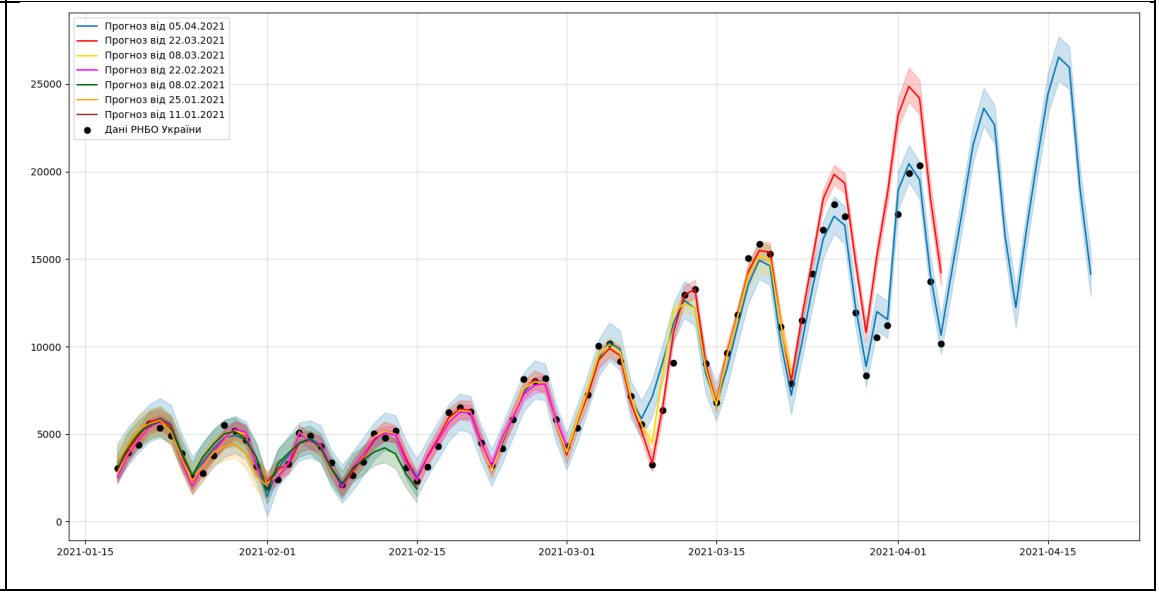
Обчислення за допомогою моделі Prophet і аналіз отриманих результатів виконали завідувач кафедри системного аналізу та інформаційних технологій (САІТ) Вінницького національного технічного університету (ВНТУ) доктор технічних наук, професор В.Б. Мокін і аспірант кафедри САІТ ВНТУ А.В. Лосенко.

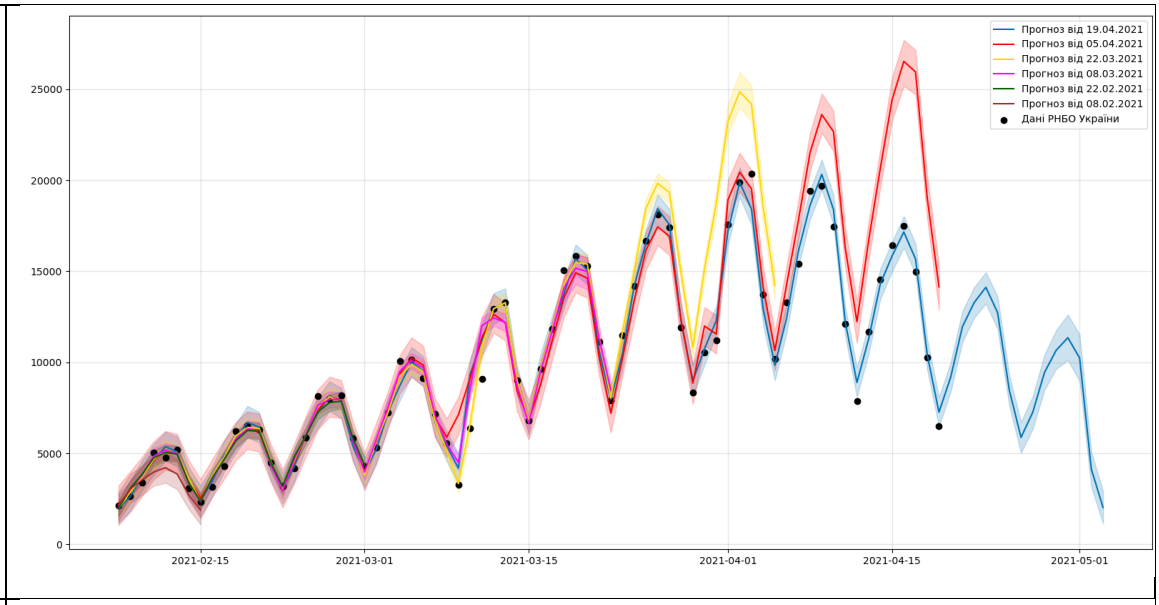
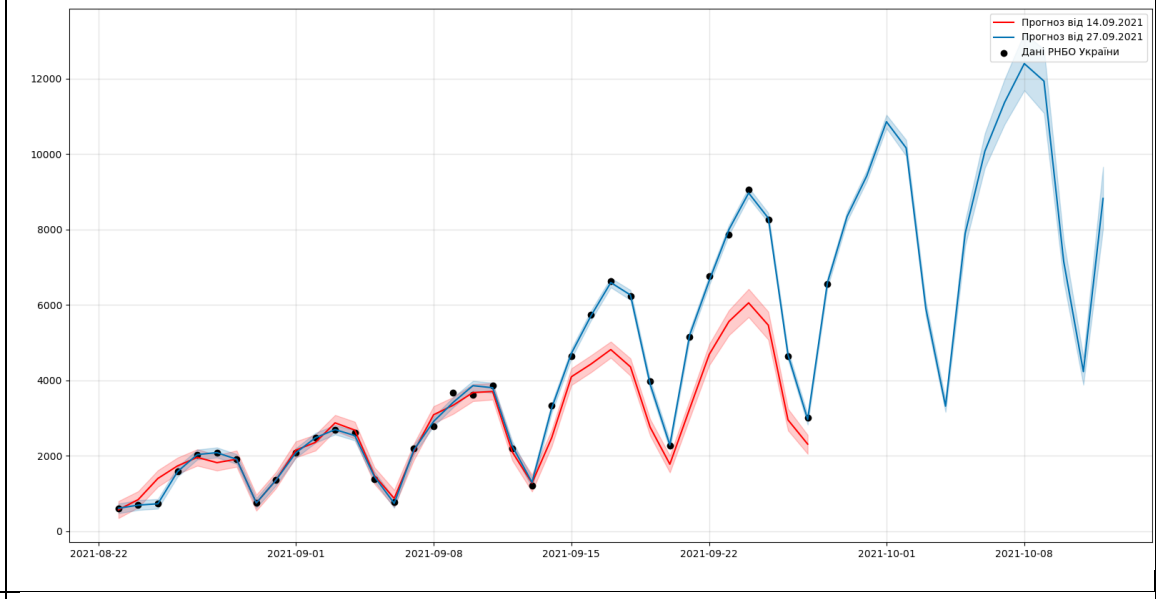
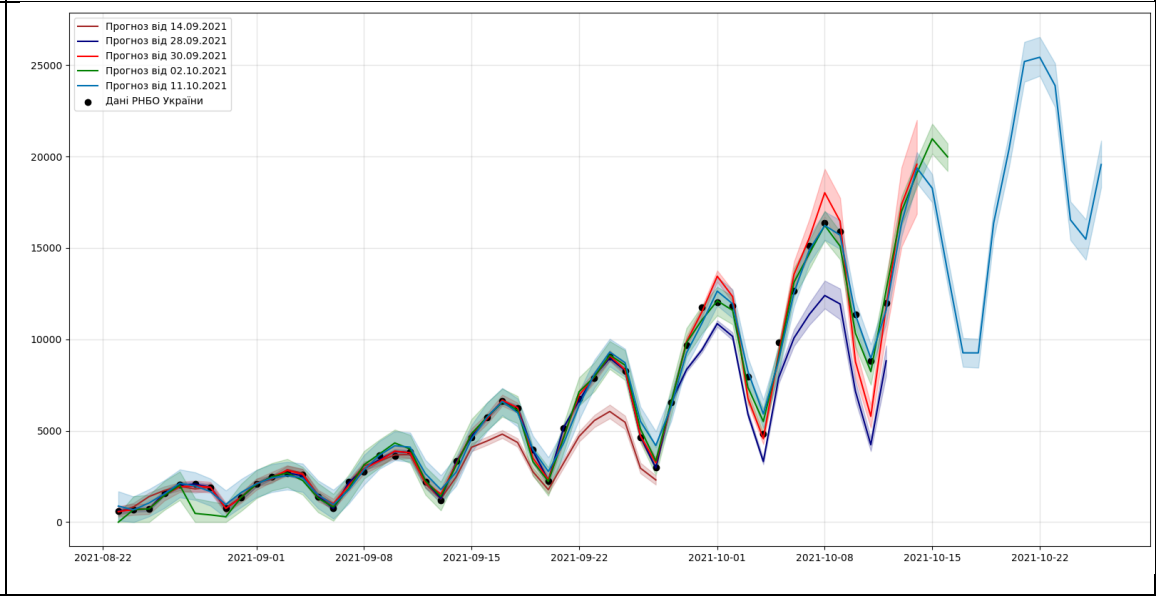
ДОДАТОК Б
ТАБЛИЦЯ ГРАФІКІВ ПРОГНОЗУ ЗАХВОРЮВАНОСТІ НА КОРОНАВІРУС З 2020 ПО 2022 РР..

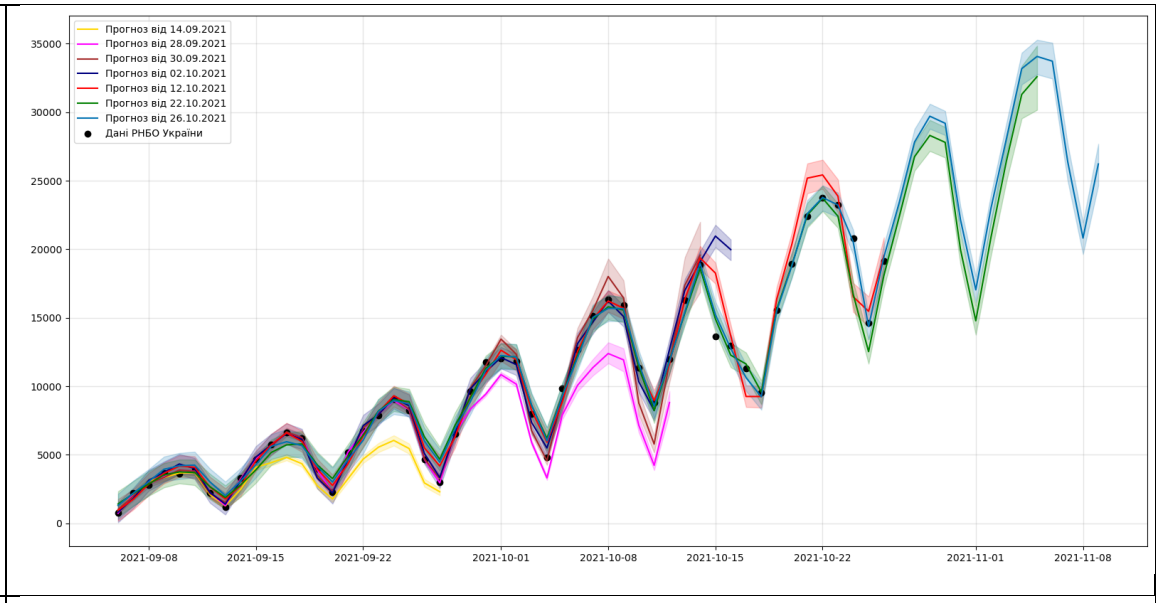
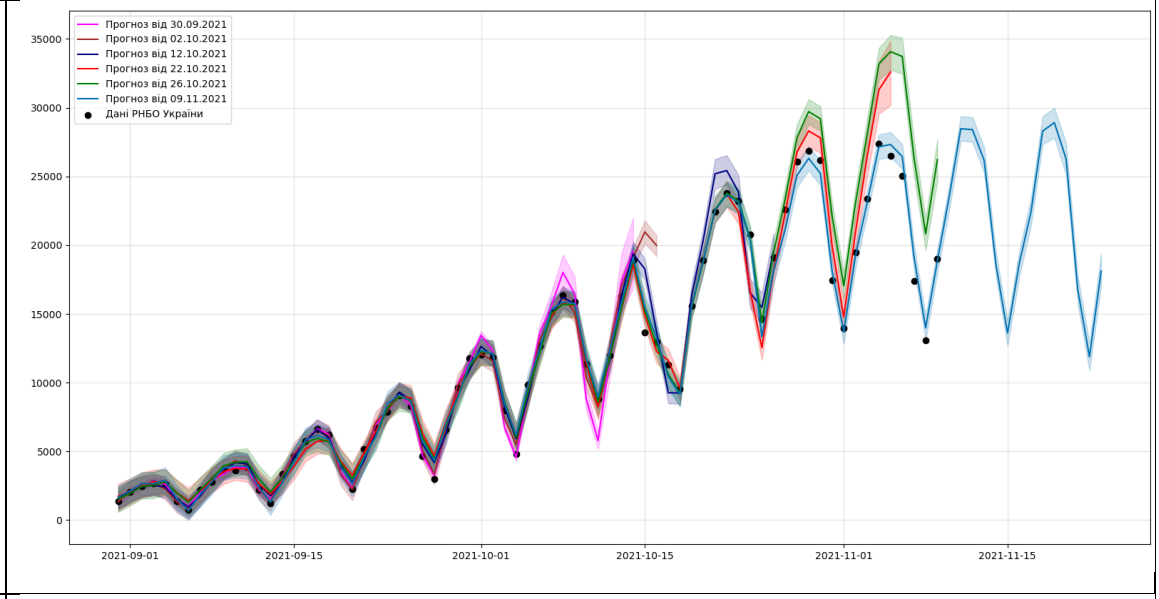
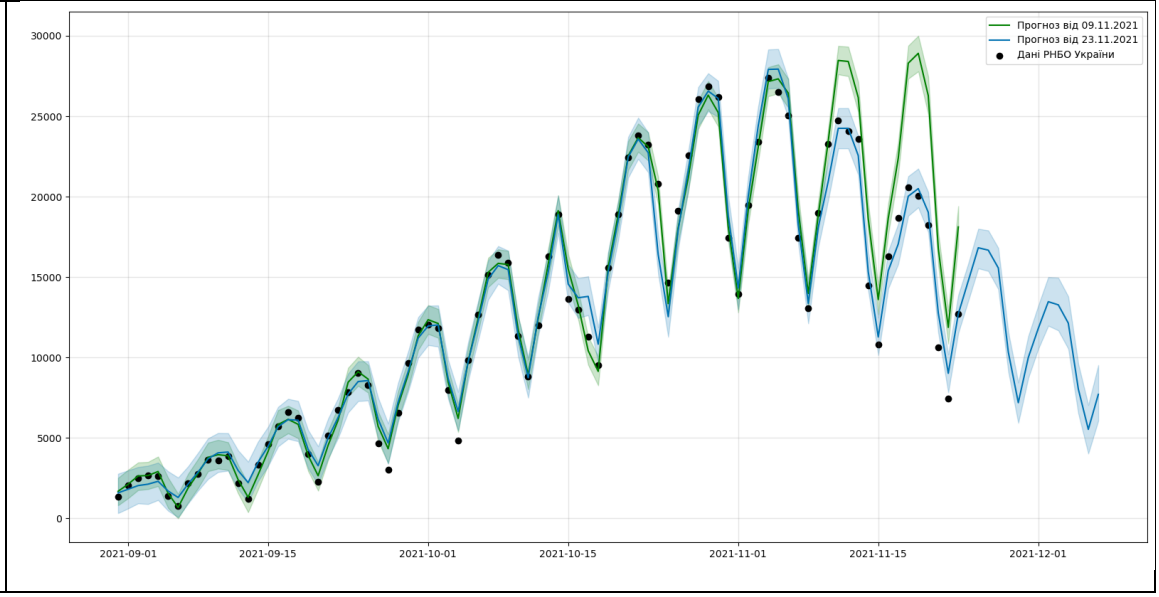
Дата звіту	Графік прогнозу
29.11.20	
14.12.20	

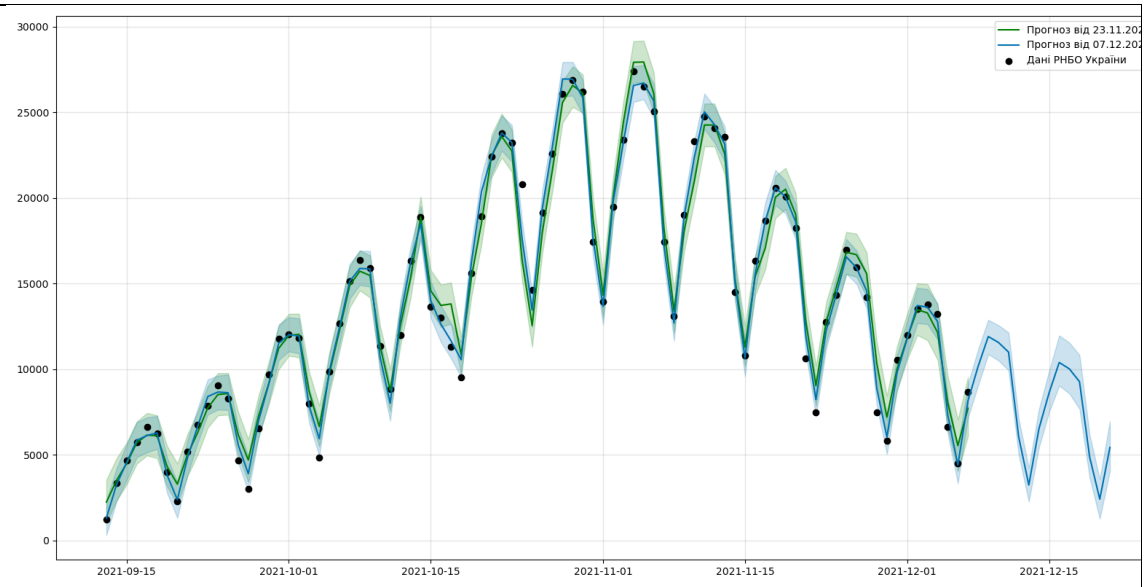
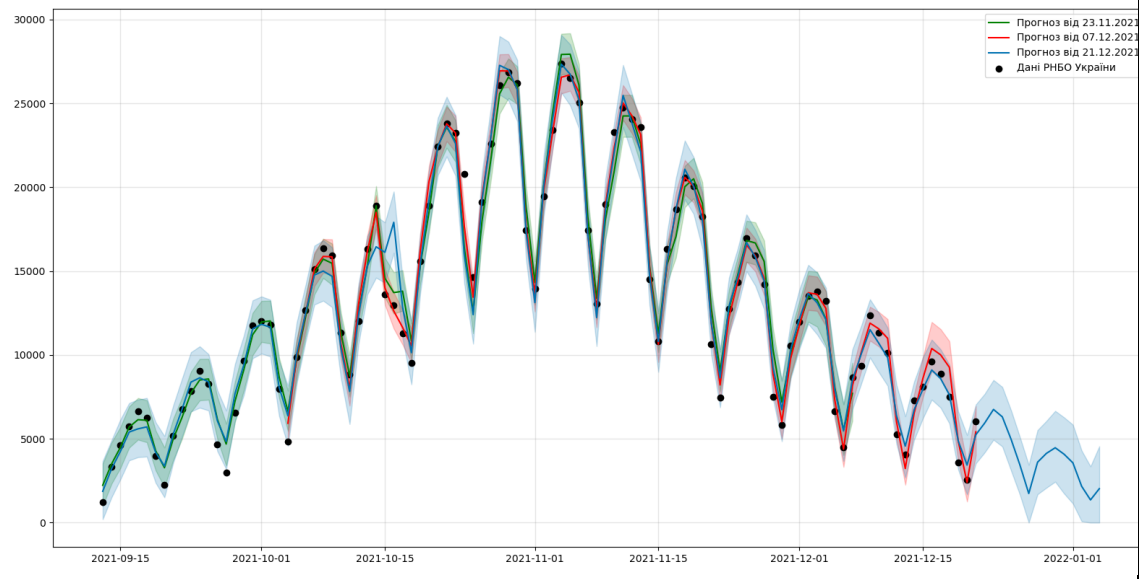
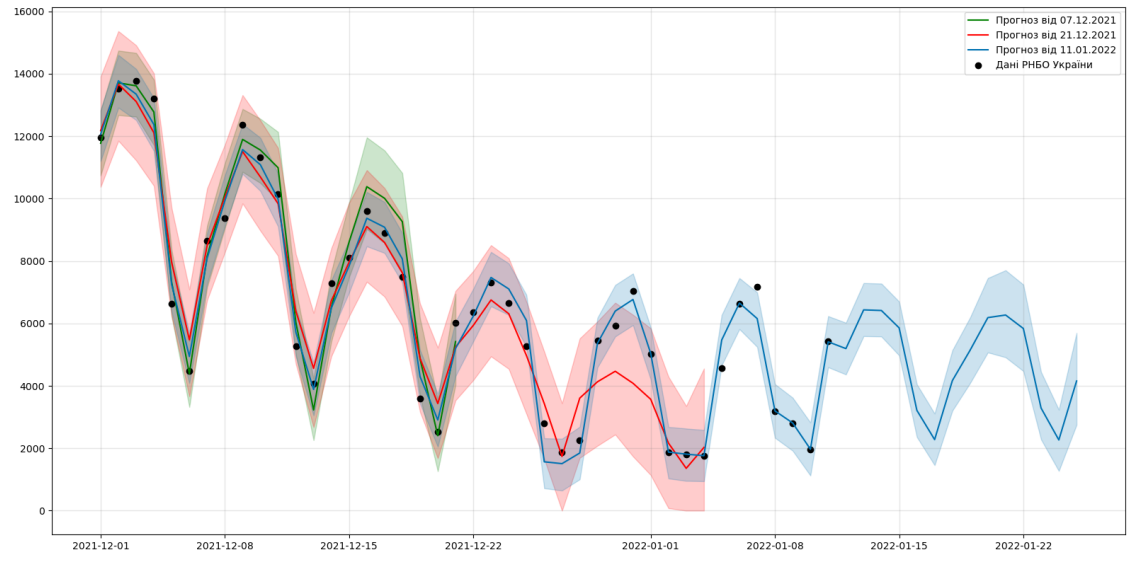
21.12.20	
28.12.20	
11.01.21	

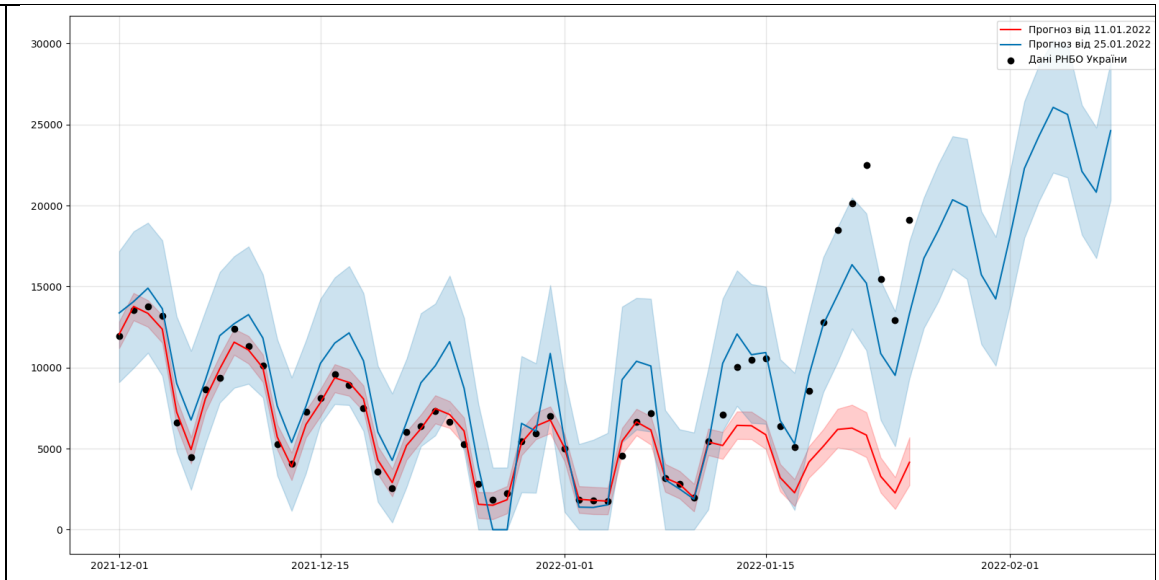
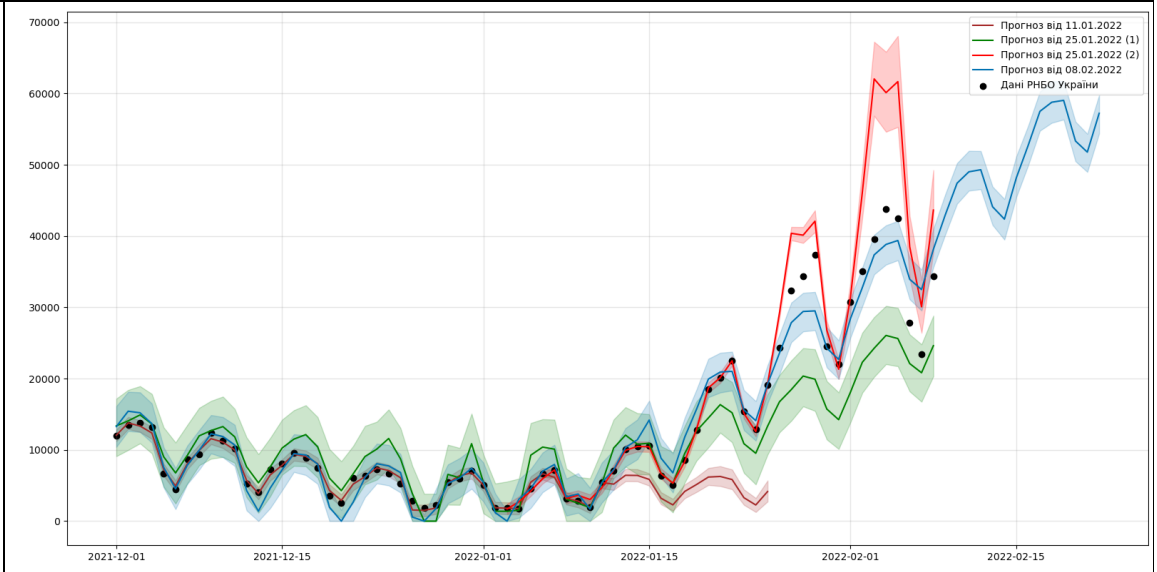
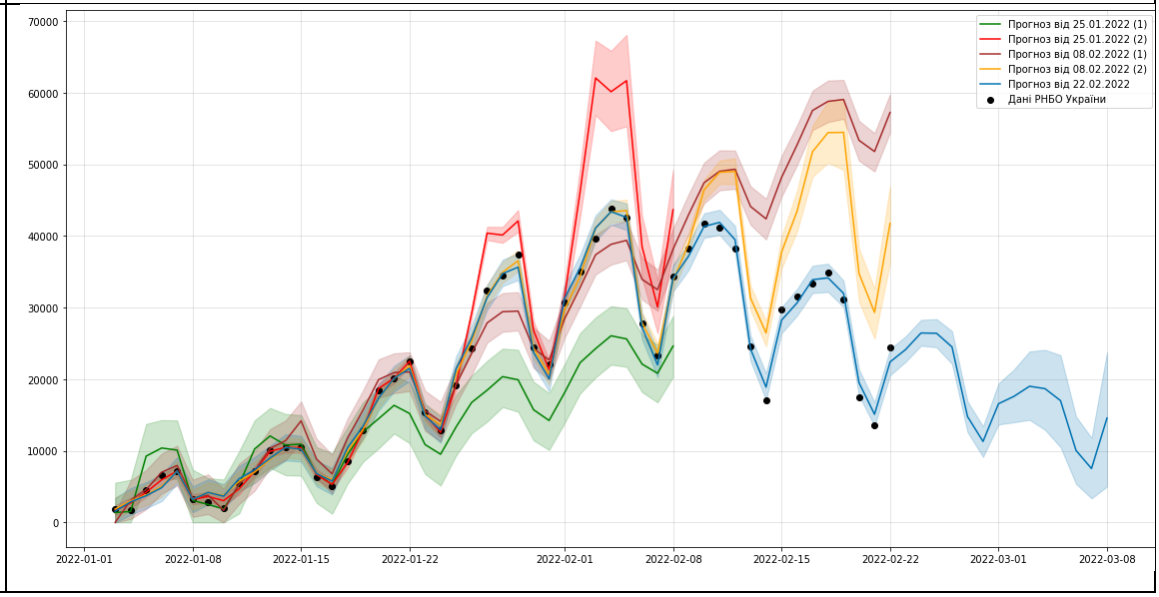
25.01.21	
08.02.21	
22.02.21	

8.03.21	
22.03.21	
05.04.21	

19.04.21	
28.09.21	
12.10.21	

26.10.21	
09.11.21	
23.11.21	

07.12.21	
21.12.21	
11.01.22	

25.01.22	
08.02.22	
22.02.22	

ДОДАТОК В

РОЗШИРЕНА ПОРІВНЯЛЬНА ТАБЛИЦЯ ПОХИБОК МОДЕЛЕЙ FACEBOOK PROPHET ТА SIER-U 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

ДОДАТОК Г
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

[1] В. Мокін, А. Лосенко, М. Дратований, «Інтелектуальна технологія аналізу та передбачення цін на вживані автомобілі», *Вісник Вінницького політехнічного інституту*, № 6, с. 62-72 (Груд 2019.) **(Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[2] В. Мокін, А. Лосенко, А. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків на коронавірус SARS-CoV-2 в Україні на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*. – 2020. – № 5. – С. 71–83. **(Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[3] В. Мокін, А. Лосенко, А. Ящолт, «Інформаційна технологія аналізу та прогнозування багатохвильової кількості нових випадків захворювань на коронавірус COVID-19 на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*. - Вип. 6, С. 65–75, 2020. **(Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[4] В. Мокін, М. Дратований, А. Лосенко, С. Жуков, «Прогнозування хвиль коронавірусу на основі відновленої когнітивної карти міжрегіонального впливу», *Інформаційні технології та комп'ютерна інженерія*, 2021, Том 52, Вип. 3, с. 86–94, Груд 2021. **(Наукове фахове видання України, категорія «Б» зі спеціальності 126)**

[5] А. Лосенко, «Інформаційна технологія прогнозування часового ряду кількості хворих на коронавірус на основі моделі Facebook Prophet», *Вісник Вінницького політехнічного інституту*, № 5, с. 50-59 (Жовтень 2023.) (**Index Copernicus, Наукове фахове видання України, категорія «Б» зі спеціальності 126**)

[6] В. Мокін, А. Лосенко, «Картування тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи протягом січня-березня 2021 року», *на L науково-технічній конференції підрозділів ВНТУ*, Вінниця, 10-12 березня 2021 р. – Електрон. текст. дані. – 2021. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2021/paper/view/12849

[7] В. Мокін, А. Лосенко, А. Ящолт, А. Гевеленко, «Прогнозування тижневих трендів кількості нових хворих на коронавірус у країнах світу», *на XX Міжнародній науково-практичній конференції "Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях"* (Жовтень 2021), С. 209-212, Електрон. текст. дані, 2021. – Режим доступу: https://itgip.org/wp-content/uploads/2021/10/1_%D0%97%D0%B1%D1%96%D1%80%D0%BA%D0%B0_2021.pdf

[8] Д. Шмундяк, А. Лосенко, В. Мокін, «Огляд підходів до визначення порядку Фур'є у моделі Facebook Prophet для моделювання сезонної складової часового ряду», *на LII Науково-технічній конференції факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету*, Вінниця, 21 – 23 червня 2023 р. – Електрон. текст. дані. – 2023. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/view/17200/14329.

[9] В. Мокін, А. Лосенко, «Інформаційна технологія короткострокового прогнозування кількості нових хворих на коронавірус на основі моделі Facebook Prophet», *на XXII Міжнародній науково-практичній конференції "Інформаційно-комунікаційні технології та сталий розвиток"* (Листопад 2023), С. 209-212, Електрон. текст. дані, 2023. – Режим доступу: https://itgip.org/wp-content/uploads/2023/11/1_%D0%97%D0%B1%D1%96%D1%80%D0%BA%D0%B0_2023.pdf

[10] І. Бровченко, Р. Беженар, В. Мокін , А. Лосенко та ін. 25 звітів Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні при НАН України «Прогноз РГ-29» (Опубл. 23.11.2020), «Прогноз РГ-30» (Опубл. 30.11.2020), «Прогноз РГ-31» (Опубл. 07.12.2020), «Прогноз РГ-32» (Опубл. 14.12.2020), «Прогноз РГ-33» (Опубл. 21.12.2020), «Прогноз РГ-34» (Опубл. 28.12.2020), «Прогноз РГ-35» (Опубл. 11.01.2021), «Прогноз РГ-36» (Опубл. 25.01.2021), «Прогноз РГ-37» (Опубл. 08.02.2021), «Прогноз РГ-38» (Опубл. 22.02.2021), «Прогноз РГ-39» (Опубл. 10.03.2021), «Прогноз РГ-40» (Опубл. 23.03.2021), «Прогноз РГ-41» (Опубл. 06.04.2021), «Прогноз РГ-42» (Опубл. 20.04.2021), «Прогноз РГ-51» (Опубл. 14.09.2021), «Прогноз РГ-52» (Опубл. 28.09.2021), «Прогноз РГ-53» (Опубл. 12.10.2021), «Прогноз РГ-54» (Опубл. 26.10.2021), «Прогноз РГ-55» (Опубл. 09.11.2021), «Прогноз РГ-56» (Опубл. 23.11.2021), «Прогноз РГ-57» (Опубл. 07.12.2021), «Прогноз РГ-58» (Опубл. 21.12.2021), «Прогноз РГ-60» (Опубл. 26.01.2022), «Прогноз РГ-61» (Опубл. 08.02.2022), «Прогноз РГ-62» (Опубл. 22.02.2022). Режим доступу на сайті Президії НАН України: https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx

ДОДАТОК Д
АКТ ПРО ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНОЇ РОБОТИ

ЗАТВЕРДЖУЮ
В. о. директора Інституту проблем
математичних машин та систем НАН
України
_____ д.ф.-м.н. Віталій КЛИМЕНКО
«___» _____2023 р.

АКТ
про впровадження результатів дисертаційної роботи
«Інформаційна технологія прогнозування часових рядів кількості хворих на коронавірус методами машинного навчання»
Лосенка Арсена Володимировича

Комісія у складі:

голова комісії – заступник директора з наукової роботи Інституту проблем математичних машин та систем НАН України (ІПММС), член-кореспондент НАН України, д.ф.-м.н., Ігор БРОВЧЕНКО та члени комісії – зав. відділом, проф., д.ф.-м.н Володимир Мадерич, старший науковий співробітник, д.т.н. Роман БЕЖЕНАР та старший науковий співробітник, к.ф.-м.н. Ігор Іванов, склала цей акт про підтвердження того, що дійсно у 25-ти звітах 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, при базовій установі – Інституті проблем математичних машин і систем НАН України, використано результати роботи аспіранта Арсена ЛОСЕНКА, який є здобувачем наукового ступеню доктора філософії зі спеціальності 126 Інформаційні системи та технології у Вінницькому національному технічного університеті (ВНТУ). Зокрема, він, спільного з його науковим керівником д.т.н., професором ВНТУ, Віталієм МОКІНИМ, є співавтором розділу «Прогноз розвитку епідемії в Україні з використання статистичної моделі часових рядів Facebook Prophet» звітів «Прогноз РГ-29» (Опубл. 23.11.2020), «Прогноз РГ-30» (Опубл. 30.11.2020), «Прогноз РГ-31» (Опубл. 07.12.2020), «Прогноз РГ-32» (Опубл. 14.12.2020), «Прогноз РГ-33» (Опубл. 21.12.2020), «Прогноз РГ-34» (Опубл. 28.12.2020), «Прогноз РГ-35» (Опубл. 11.01.2021), «Прогноз РГ-36» (Опубл. 25.01.2021), «Прогноз РГ-37» (Опубл. 08.02.2021), «Прогноз РГ-38» (Опубл. 22.02.2021), «Прогноз РГ-39» (Опубл. 10.03.2021), «Прогноз РГ-40» (Опубл. 23.03.2021), «Прогноз РГ-41» (Опубл. 06.04.2021), «Прогноз РГ-42» (Опубл. 20.04.2021), «Прогноз РГ-51» (Опубл. 14.09.2021), «Прогноз РГ-52» (Опубл.

28.09.2021), «Прогноз РГ-53» (Опубл. 12.10.2021), «Прогноз РГ-54» (Опубл. 26.10.2021), «Прогноз РГ-55» (Опубл. 09.11.2021), «Прогноз РГ-56» (Опубл. 23.11.2021), «Прогноз РГ-57» (Опубл. 07.12.2021), «Прогноз РГ-58» (Опубл. 21.12.2021), «Прогноз РГ-60» (Опубл. 26.01.2022), «Прогноз РГ-61» (Опубл. 08.02.2022), «Прогноз РГ-62» (Опубл. 22.02.2022), опублікованих на сайті Президії НАН України: https://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx

Зокрема, у цих звітах використано такі результати застосування розробленої Арсеном ЛОСЕНКОМ та його науковим керівником Віталієм МОКІНИМ інформаційної технології прогнозування часових рядів кількості хворих на коронавірус на базі статистичної моделі Facebook Prophet:

- прогнози на 2 тижні щоденної кількості нових хворих на коронавірус в Україні;
- прогнози на 1 тиждень щодо 69 країн світу;
- висновки щодо закономірностей, які мають місце та можуть очікуватись в майбутньому, отримані на основі аналізу складових цих прогнозів та на основі аналізу картограм, побудованих з урахуванням цих прогнозних даних.

Про використання результатів, що отримано Арсеном ЛОСЕНКОМ зазначено в опублікованих звітах.

Голова комісії _____ Ігор БРОВЧЕНКО

Члени комісії _____ Роман БЕЖЕНАР

_____ Володимир Мадерич

_____ Ігор Іванов

ДОДАТОК Д
АКТ ПРО ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ У НАВЧАЛЬНИЙ ПРОЦЕС ВНТУ

ЗАТВЕРДЖУЮ
Проректор з науково-педагогічної роботи та організації освітнього процесу Вінницького національного технічного університету

_____ доц. Олександр ПЕТРОВ
«_04_» ____12____ 2023 р.

АКТ
про впровадження результатів дисертаційної роботи
Лосенка Арсена Володимировича
«Інформаційна технологія прогнозування часових рядів кількості хворих на коронавірус методами машинного навчання», представленої до захисту на здобуття наукового ступеня доктора філософії, в навчальному процесі

Комісія Вінницького національного технічного університету у складі:

голова комісії – декан Факультету інтелектуальних інформаційних технологій та автоматизації (ФІІТА), к.т.н, доцент Севастьянов В. М. та члени комісії – заступник декана з навчально-методичної роботи ФІІТА, ст. викладач каф. САІТ Присяжнюк В. В., заступник завідувача кафедри системного аналізу та інформаційних технологій (САІТ) ФІІТА, к.т.н., доцент Крижановський Є. М., склала цей акт про підтвердження того, що дійсно у Вінницькому національному технічному університеті під час викладання дисципліни «Інформаційні технології моніторингу та аналізу даних» для студентів, які навчаються за освітньою програмою «Інформаційні технології аналізу даних та зображень» рівня «магістр» спеціальності 126 Інформаційні системи та технології, а також – дисципліни «Інтернет речей та інтелектуальний аналіз даних» для аспірантів, які навчаються за освітньою програмою «Інформаційні системи та технології» рівня «доктора філософії» цієї ж спеціальності 126 впроваджено результати дисертаційної роботи на здобуття доктора філософії аспіранта кафедри САІТ Лосенка Арсена Володимировича у ряді лекцій та лабораторних робіт з питань інформаційних технологій.

При викладанні цих дисциплін використовуються такі результати досліджень, отримані у дисертаційній роботі Лосенка А. В.:

- розвідувальний аналіз даних часових рядів: декомпозиція ряду на різні види сезонності, перевірка на стаціонарність, пошук аномалій та візуалізація результатів такого аналізу, у т.ч. побудова інтерактивних карт;
- тренування моделей машинного навчання (лінійна регресія, метод опорних векторів, дерева рішень та їх ансамблі, багатошарова нейронна мережа, ARIMA, Facebook Prophet з типовою структурою) з використанням технік GridSearchCV, HyperOpt і байєсівської оптимізації для пошуку оптимальних параметрів) та порівняльний аналіз їх точності за різними метриками і вибір оптимальної моделі;
- багатоітеративний метод ідентифікації параметрів та структури моделі нестаціонарного часового ряду з різними видами сезонності на основі Facebook Prophet.

Використання зазначених результатів дозволило підвищити якість навчального процесу із згаданих дисциплін.

Голова комісії _____ Володимир СЕВАСТЬЯНОВ

Члени комісії _____ Василь ПРИСЯЖНЮК

_____ Євгеній КРИЖАНОВСЬКИЙ