

Вінницький національний технічний університет
Факультет комп'ютерних систем і автоматики
Кафедра системного аналізу та інформаційних технологій

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему:

“Інформаційна технологія аналізу та прогнозування кількості нових хворих на коронавірус в Україні у 2021 році”

Виконав: студент 2 курсу, групи 2ІСТ-20м спеціальності 126 – «Інформаційні системи та технології»

_____ Гевеленко А.О.

Керівник: д.т.н., проф. каф. САІТ

_____ Мокін В.Б.

«__» _____ 2021 р.

Опонент: доц. каф. АІТ

_____ Іванов Ю.Ю.

«__» _____ 2021 р.

Допущено до захисту

Завідувач кафедри САІТ

_____ д.т.н., проф. Мокін В. Б.

«__» _____ 2021 р.

Вінниця ВНТУ – 2021 рік

Вінницький національний технічний університет
 Факультет комп'ютерних систем і автоматики
 Кафедра системного аналізу та інформаційних технологій
 Рівень вищої освіти – II-й (магістерський)
 Галузь знань – 12 Інформаційні технології
 Спеціальність – 126 Інформаційні системи та технології
 Освітньо-професійна програма – Інформаційні технології аналізу даних та зображень

ЗАТВЕРДЖУЮ

Завідувач кафедри САІТ

_____ д.т.н., проф. Мокін В. Б.

«___» _____ 2021 р.

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
 Гевеленку Артему Олександровичу

1. Тема роботи: «Інформаційна технологія аналізу та прогнозування кількості нових хворих на коронавірус в Україні у 2021 році»,
 керівник роботи: Мокін В.Б., д.т.н., проф. каф. САІТ,

затверджені наказом закладу вищої освіти від «___» _____ 2021 року № ___

2. Строк подання студентом роботи «___» _____ 2021 року

3. Вихідні дані до роботи:

- кількість хворих з відкритих даних РНБО;
- кількість хворих з відкритих даних по країнам світу.

4. Зміст текстової частини:

- загальні методи прогнозування та моделі поширення захворюваності;
- коронавірус COVID-19;
- аналіз та прогнозування кількості нових хворих на коронавірус в Україні у 2021 році
- економічна частина.

5. Перелік ілюстративного матеріалу (з точним зазначенням обов'язкових креслень):

- сумарний за добу Stringency-індекс України з ослабленнями та підсиленнями карантину та дати аномалій-послаблень карантину за допомогою довгострокового прогнозу;
- середньодобова температура в Україні, кількість опадів, мм, дати аномальних «метопаттернів» за допомогою довгострокового прогнозу;
- підтверджені захворювання на коронавірус з відображенням св'ятів за допомогою довгострокового прогнозування;
- підтверджені захворювання на коронавірус в проміжок довгого періоду часу;

- короткостроковий прогноз за спадом хвилі;
- графік трендів найбільш популярних країн Європи з найбільшим негативним прогнозом.

6. Консультанти розділів МКР

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
4	Ратушняк О.Г., к.е.н., доц. каф. ЕПВМ		

7. Дата видачі завдання «__»_____2021 року

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів МКР	Строк виконання етапів роботи	Примітка
1	Аналіз предметної області	09.2021	
2	Огляд методів прогнозування та моделей поширення захворюваності	09.2021	
3	Розробка інформаційної системи	09.2021	
4	Економічна частина	10.2021	
5	Реалізація прогнозу	10.2021	
6	Результат прогнозування в Україні	11.2021	
7	Оформлення матеріалів до захисту МКР	11.2021	

Студент _____

Гевеленко А.О.

Керівник роботи _____

Мокін В.Б.

АНОТАЦІЯ

УДК 004.09

Гевеленко А. О. Інформаційна технологія аналізу та прогнозування кількості нових хворих на коронавірус в Україні в 2021 році. Магістерська кваліфікаційна робота зі спеціальності 126 – інформаційні системи та технології, освітньо-професійна програма – інформаційні технології аналізу даних та зображень. Вінниця: ВНТУ, 2021. 103 с.

На укр. мові. Бібліогр.: 21 назв; рис.: 75; табл.: 7.

В магістерській кваліфікаційній роботі звернено увагу на проблему епідемії коронавірусної інфекції та її прогнозуванню. Проведено ґрунтовний аналіз методів для побудови довгострокових та короткострокових прогнозів поширення інфекційних захворювань. На основі результатів виконаних досліджень було реалізовано моделі для короткострокового та довгострокового прогнозування поширення коронавірусу та розроблено програмний продукт, що може використовуватися для аналізу поширення коронавірусу в Україні та за її межами. Наведено детальний опис обраних методів та результатів отриманих прогнозів.

Об'єкт досліджень – процес аналізу та прогнозування кількості нових хворих на коронавірус в Україні у 2021 р.

Ілюстративна частина складається з 4 плакатів із результатами моделювання.

У розділі економічної частини розглянуто питання про доцільність розробки та впровадження інформаційної технології аналізу та прогнозування кількості нових хворих на коронавірус.

Ключові слова: інформаційна технологія, прогнозування, хворі, передбачення, коронавірус, Prophet.

ABSTRACT

Gevelenko A. O. Information technology of analysis and forecasting the number of new patients with coronavirus in Ukraine in 2021. Master's thesis in specialty 126 - information systems and technologies, educational and professional program - information technology data and image analysis. Vinnytsia: VNTU, 2021. 103 p.

In Ukrainian language. Bibliogr .: 21 titles; fig .: 75; table: 7.

In the master's qualification work attention is paid to the problem of the epidemic of coronavirus infection and its prognosis. A thorough analysis of methods for constructing long-term and short-term forecasts of the spread of infectious diseases. Based on the results of the research, models for short-term and long-term prediction of coronavirus spread were implemented and a software product was developed that can be used to analyze the spread of coronavirus in Ukraine and abroad. A detailed description of the selected methods and the results of the obtained forecasts is given.

The object of research is the process of analysis and forecasting the number of new patients with coronavirus in Ukraine in 2021.

The illustrative part consists of 4 posters with simulation results.

In the section of the economic part the question of expediency of development and introduction of information technology of the analysis and forecasting of quantity of new patients with a coronavirus is considered.

Key words: information technology, forecasting, patients, predictions, coronavirus, Prophet.

ЗМІСТ

ВСТУП	4
1 ЗАГАЛЬНІ МЕТОДИ ПРОГНОЗУВАННЯ ТА МОДЕЛІ ПОШИРЕННЯ ЗАХВОРЮВАНЬ	6
1.1 Основні типи моделей поширення захворюваності	6
1.2 Статистичний підхід до прогнозування захворюваності	6
1.2.1 Регресійні моделі	7
1.2.2 Авторегресійні моделі	8
1.3 Математичний підхід до прогнозування захворюваності	9
1.3.1 Модель SIR	10
1.3.2 Моделі SEIR та SEIR-U	12
1.4 Підхід до прогнозування захворюваності на основі фільтрації	14
1.4.1 Експоненційне згладжування	14
1.4.2 Методи теорії керування	16
1.5 Підхід до прогнозування захворюваності із застосуванням методів машинного навчання	16
1.5.1 Штучні нейронні мережі	17
1.6 Висновки	19
2 КОРОНАВІРУС COVID-19	22
2.1 Особливості захворювання та методи дослідження	22
2.1.1 Огляд захворювання та його основних епідеміологічних характеристик	48
2.1.2 Порівняльний аналіз епідемій, викликаних коронавірусами SARS-CoV-2, SARS-CoV та MERS-CoV	26
2.2 Моделі для аналізу поширення коронавірусної інфекції	29
2.2.1 Прогноз НАН України	30
2.2.2 Прогноз за допомогою нейронної мережі	31
2.3 Опис обраних методів	32
2.4 Висновки	36
3 АНАЛІЗ ТА ПРОГНОЗУВАННЯ КІЛЬКОСТІ НОВИХ ХВОРИХ НА КОРОНАВІРУС В УКРАЇНІ У 2021 РОЦІ	38

	6
3.1 Розроблення алгоритму інформаційної технології.....	43
3.2 Короткостроковий прогноз	46
3.2.1 Короткостроковий прогноз з усіма хвилями.....	46
3.2.2 Короткостроковий прогноз з останньою хвилею	51
3.3 Довгостроковий прогноз	54
3.3.1 Довгостроковий прогноз за допомогою мультиплікативної моделі	54
3.3.2 Довгостроковий прогноз за використання адитивної моделі.....	56
3.4 Короткостроковий прогноз на 7 днів для 70 країн світу	71
3.5 Висновки	75
4 ЕКОНОМІЧНА ЧАСТИНА	76
4.1 Оцінювання комерційного потенціалу розробки.....	76
4.2 Прогнозування витрат на виконання науково-дослідної роботи.....	80
4.3 Розрахунок економічної ефективності науково-технічної розробки	84
4.4 Розрахунок ефективності вкладених інвестицій та періоду їх окупності.	86
4.5 Висновки	88
ВИСНОВКИ.....	89
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	91
Додаток А (обов'язковий). Технічне завдання	94
Додаток Б (обов'язковий). Протокол перевірки кваліфікаційної роботи	96
Додаток В (довідниковий). Лістинг програми	97
Додаток Г (обов'язковий). Ілюстративна частина	99

ВСТУП

Актуальність теми. Пандемія коронавірусу COVID-19, виникла внаслідок інфекції SARS-CoV-2 (Коронавірус) і стала серйозним викликом, який стоїть і зараз перед людством в усьому світі. Актуальність дослідження полягає в тому, щоб одержати реальну картину та прогнозувати на майбутнє більш-менш точний прогноз захворюваності з такими чинниками, як наприклад, свята, невизначеність деяких даних, тощо.

Мета і завдання роботи. Метою даної роботи є підвищення точності короткострокового прогнозування кількості нових хворих на коронавірус у 2021 р. та дослідити його можливість довгострокового прогнозування за відкритими даними. Для того щоб реалізувати мету даної роботи потрібно виконати такі задачі:

- розробка короткострокового прогнозу з усіма хвилями;
- розробка короткострокового прогнозу з останньою хвилею;
- розробка довгострокового прогнозу за допомогою мультиплікативної моделі;
- розробка довгострокового прогнозу за допомогою адитивної моделі;
- розробка короткострокового прогнозу на 7 днів для 70 країн світу.

Об'єктом дослідження магістерської кваліфікаційної роботи є процес аналізу та прогнозування кількості нових хворих на коронавірус в Україні у 2021 р.

Предметом дослідження магістерської кваліфікаційної роботи є технології та програмні засоби аналізу та прогнозування кількості нових хворих на коронавірус в Україні у 2021 р.

Новизна одержаних результатів. Дістала подальший інформаційна технологія прогнозування щоденних прирості кількості нових підтверджених хворих на коронавірус COVID-19 в Україні за відкритими даними шляхом

побудови більш точної короткострокової моделі з песимістичним та оптимістичним сценаріями. Доведено, що з використанням відкритих даних побудова моделей для довгострокового прогнозування на основі моделі Prophet поки не є можливою.

Практичне значення роботи полягає у виявленні закономірностей та у прогнозуванні кількості нових хворих на коронавірус в Україні, для того, щоб можна було завчасно вжити необхідні превентивні заходи.

Апробація результатів магістерської кваліфікаційної роботи. Результати кваліфікаційної роботи доповідались на XX Міжнародної науково-практичної конференції "Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях".

Публікації результатів магістерської кваліфікаційної роботи. В ході виконання магістерської кваліфікаційної роботи опубліковано тези у збірнику матеріалів XX Міжнародної науково-практичної конференції "Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях" [1].

1 ЗАГАЛЬНІ МЕТОДИ ПРОГНОЗУВАННЯ ТА МОДЕЛІ ПОШИРЕННЯ ЗАХВОРЮВАНЬ

1.1 Основні типи моделей поширення захворюваності

В сучасному світі неодноразово вчені знаходять нові віруси, з якими потім бореться людство та економіка. І в цей раз було знайдено вірус, який забрав багато життів та наніс великий удар по економіці світу. І внаслідок цього вірусологи та епідеміологи отримали завдання прогнозувати поширення, розвиток та наслідок вірусних захворювань.

Цифрова епідеміологія – це галузь науки, яка швидко розвивається у ХХ столітті через безперервне зростання кількості даних і обчислювальної потужності, а також прориви в методах аналізу даних і мовах програмування та інформаційних технологіях. На сьогоднішній день обладнання, яке використовується для прогнозування розвитку епідеміологічних процесів, включає багато різних технологій. Враховуючи природу захворювання та довжину прогностичного вікна, це дозволяє підібрати широкий асортимент моделей для кожного конкретного випадку і, за необхідності, звузити до тих захворювань, які виявилися найбільш підходящими для цього типу. моделі. Методи прогнозування захворюваності можна розділити на кілька основних категорій, а саме: статистичні, математичні моделі, моделі на основі фільтрації та моделі на основі методів машинного навчання.

1.2 Статистичний підхід до прогнозування захворюваності

Статистичні методи засновані на теорії ймовірностей і використання випадкових розмірів та процесів для опису часових рядів захворювань. Такі методи є невід'ємною частиною розвитку епідеміології і багато країн світу продовжують їх використовувати для аналізу сезонних захворювань. У цьому розділі є найбільш поширені моделі цього типу, передусім – регресійні.

1.2.1 Регресійні моделі

Регресійний аналіз – популярний сучасний метод математичної статистики, який використовується для вирішення проблеми епідеміологічного прогнозування. Модель регресії є функцією оцінених значень незалежних змінних і невідомих параметрів плюс випадкові величини. Використання цього типу моделі ґрунтується на припущенні, що помилки моделі є незалежними випадковими величинами, мають закон розподілу та мають фіксовану дисперсію при нульовому математичному очікуванні.

Існує два основних типи регресійних моделей: неадаптивні та адаптивні. Неадаптивна модель — це модель, параметри якої оцінюються лише на початку і для цього використовуються всі наявні дані. Загальний тип функціональної залежності цієї моделі зазвичай вибирається на основі атрибутів вхідних даних. Оскільки в більшості випадків захворюваність тісно пов'язана з діяльністю людини протягом року, для моделювання використовується періодична математична функція, яка враховує сезонні зміни [2]. Однією з наступних функцій є функція Серфлінга:

$$\hat{Y} = u + bt + \sum_i a_i \cos(\theta) + \sum_i b_i \sin(\theta), \quad (1.1)$$

де θ – лінійна функція від t .

Його основним недоліком є те, що він не враховує можливі зміни характеристик захворювання з часом. Такі коливання зазвичай можна компенсувати за допомогою ієрархічної регресійної моделі.

На відміну від параметрів неадаптивної моделі, параметри адаптивної моделі розраховуються на основі даних певного періоду часу – так званого «ковзного вікна». Це усуває наслідки застарілих даних, які можуть негативно вплинути на результати прогнозу. Однак у виборі оптимальної ширини такого вікна виникає проблема. Якщо він занадто малий, модель не зможе визначити

приховані закономірності в даних, тому здатність до узагальнення погана. Тому при виборі ширини інтервалу часу необхідно враховувати як складність моделі (кількість невідомих параметрів), так і очікуваний період прогнозування.

Більш популярними є лінійні регресійні моделі.

1.2.2 Авторегресійні моделі

Підвидом регресійних моделей для аналізу часових рядів в задачах прогнозування захворюваності є авторегресія. В авторегресійному методі часовий ряд виражається як послідовність кінцевого лінійного набору попередніх значень процесу та імпульсу, тобто кожен наступний член послідовності залежить від попереднього члена. Форма рівняння авторегресії p -порядку [3]:

$$z_t = \mu + a_1 z_{t-1} + \dots + a_p z_{t-p} + \varepsilon_t, \quad (1.2)$$

де μ – невідомий параметр, що визначає рівень випадкового процесу; ε – білий шум.

Поширеною, також, є модель ковзного середнього, що використовує значення помилок моделі авторегресії. Рівняння ковзного середнього порядку q має вигляд:

$$z_t = \sum_{j=0}^q b_j \varepsilon_{t-j}, \quad (1.3)$$

Модель авторегресії (AR) може бути ефективно поєднана з моделлю ковзного середнього (MA) для формування більш загального класу моделі часового ряду – змішаної авторегресійної моделі з ковзним середнім порядку p, q відповідно (ARMA (p, q)). Поточне значення часового ряду в такій моделі описується рівнянням:

$$\tilde{z}_t = \mu + \phi_1 \tilde{z}_{t-1} + \dots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}, \quad (1.4)$$

де μ – невідомий параметр, що визначає рівень випадкового процесу;

Φ_i та θ_i – невідомі параметри;

a_t – випадкові імпульси з нульовим матсподіванням та сталою дисперсією, яка також є невідомим параметром.

Слід зазначити, що коли дисперсія змінюється з часом, тобто коли послідовність є нестационарною, наведена вище модель не може бути використана. На практиці для такої послідовності використовується перетворення, і в результаті виходить рівняння авторегресії з інтегральним ковзним середнім порядком p, d, q (ARIMA (p, d, q)):

$$\Delta^d z_t = \mu + \phi_1 \Delta^d z_{t-1} + \dots + \phi_p \Delta^d z_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}, \quad (1.5)$$

де d – порядок різниці, що необхідний для стаціонаризації ряду; Δ^d – оператор різниці ряду порядку d .

Існують, також, модифікації формули (1.5) для моделей, що враховують сезонність.

Найбільш загальним випадком регресійних та авторегресійних моделей є комплексна модель авторегресії та проінтегрованого ковзного середнього (АРПКС, англ.: ARIMA) [4].

1.3 Математичний підхід до прогнозування захворюваності

Математичні методи використовують різні диференціальні рівняння для опису динаміки системи. Це дозволяє отримати параметри справжніх характеристик моделі епідеміологічного процесу. Далі розглянемо так звані компартментні моделі – класичні моделі SIR Kermack і McKendrick та їх модифікації.

1.3.1 Модель SIR

Модель SIR заснована на поширенні інфекції, вираженому у вигляді системи диференціальних рівнянь. З цієї причини у випадку класичної моделі SIR було введено метод умовного поділу населення на здорових, хворих і таких, що одужали (чутливі особи (S), інфекційні особи (I) та особи, які одужали (R)) на основі сприйнятливості до захворювання, тому модель отримала назву SIR. На рисунку 1.1 показано, як люди переміщуються між цими групами [5].

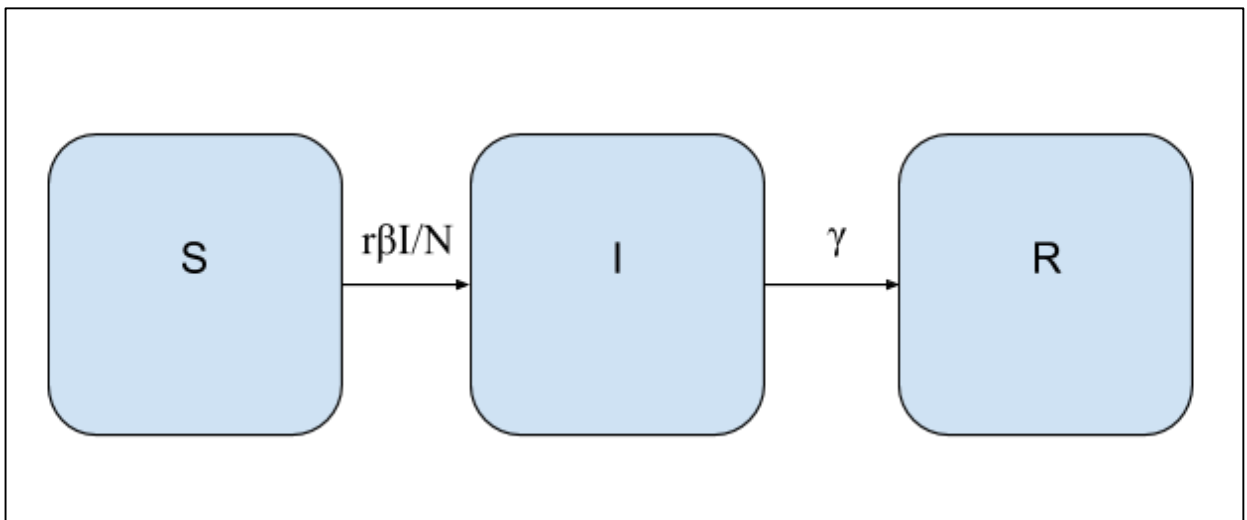


Рисунок 1.1 – Схематичне зображення класичної моделі SIR

Якщо перебіг інфекції в порівнянні з життям людини короткий (наприклад, під час спалаху) і хвороба не є смертельною, то народження і смерть можна ігнорувати. Тоді модель SIR можна записати як систему звичайних диференціальних рівнянь:

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta SI}{N} \\ \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I, \\ \frac{dR}{dt} = \gamma I \end{cases} \quad (1.6)$$

де $N = S + I + R$ – чисельність популяції ;

β – середня кількість здорових людей, яким хворий індивід передає захворювання;

γ – швидкість одужання, величина обернено пропорційна середній тривалості зараження.

У цьому випадку відсутність параметрів у рівняннях, що описують смертність і народжуваність, призведе до поступового ослаблення епідемії, оскільки кількість здорових осіб, сприйнятливих до інфекції, з часом зменшується до нуля (рис. 1.2).

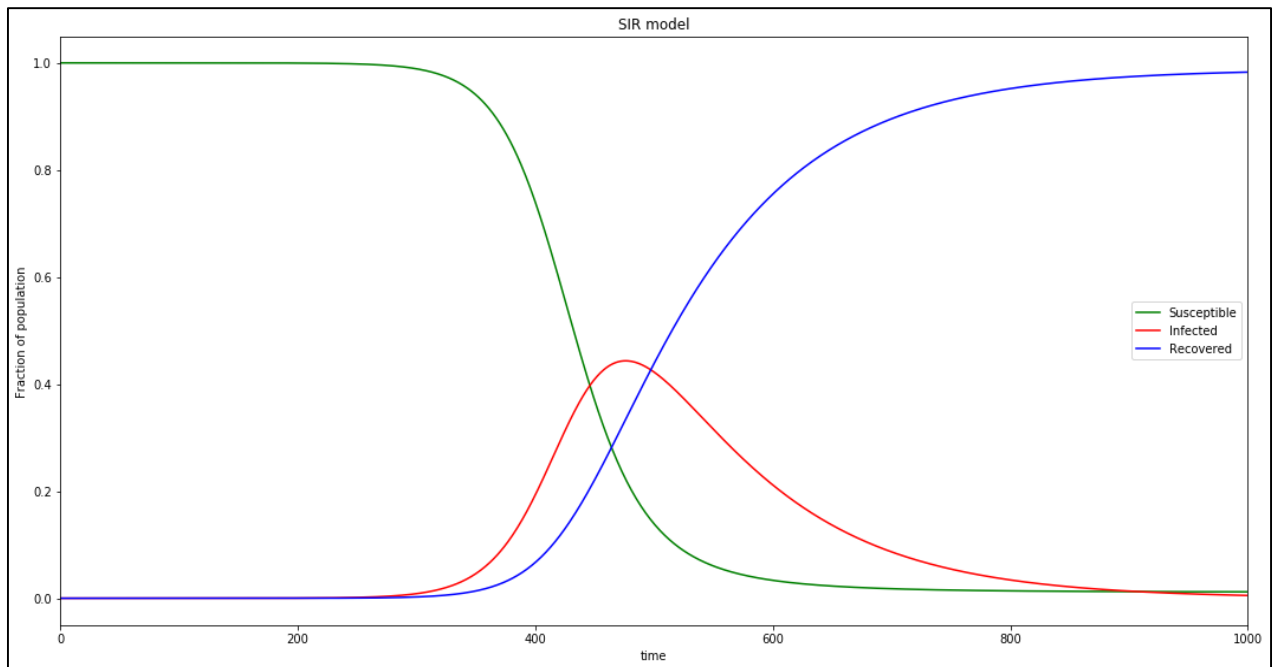


Рисунок 1.2 – Затухання епідемії

Якщо ж враховувати народжуваність і смертність, система матиме вигляд:

$$\begin{cases} \frac{dS}{dt} = \mu N - \frac{\beta SI}{N} - \nu S \\ \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I - \nu I, \\ \frac{dR}{dt} = \gamma I - \nu R \end{cases} \quad (1.7)$$

де ν – смертність;

μ – народжуваність.

Слід зазначити, що класична модель SIR базується на наступних припущеннях:

- Поширеність досліджуваного захворювання подібна до поширеності у щурів;
- Люди однаково вразливі до інфекції;
- У всіх є рівні шанси захворіти;
- Захворюваність, смертність та рівень одужання не зміняться протягом захворювання;
- Усі випадки захворювання закінчуються смертю або виникненням імунітету;
- Чисельність популяції велика, що дозволяє поширювати контакт.

Оскільки в умовах реальних епідемій задовольнити всі припущення неможливо, точність результатів цієї моделі на реальних даних значно знижується, що часто робить її використання недоцільним.

1.3.2 Модель SEIR та SEIR-U

Однією з найвідоміших модифікацій моделі SIR є модель SEIR, яка є результатом додавання іншої групи осіб, які хворіли, але не можуть поширювати хворобу протягом певного періоду часу, тобто під час інкубаційного періоду [6](E).

Тоді система диференціальних рівнянь набуває такого вигляду:

$$\begin{cases} \frac{dS}{dt} = \mu - \frac{\beta SI}{N} - \nu S \\ \frac{dE}{dt} = \frac{\beta SI}{N} - \varepsilon E \\ \frac{dI}{dt} = \varepsilon E - \gamma I - \nu I \\ \frac{dR}{dt} = \gamma I - \nu R \end{cases} \quad (1.8)$$

де $N = S + I + R$ – чисельність популяції ;

β – середня кількість здорових людей, яким хворий індивід передає захворювання;

γ – швидкість одужання – величина, обернено пропорційна середній тривалості зараження;

ε – тривалість латентного періоду.

Структура такої моделі та приклад моделювання за її допомогою зображені на рисунках 1.3 та 1.4 [7].

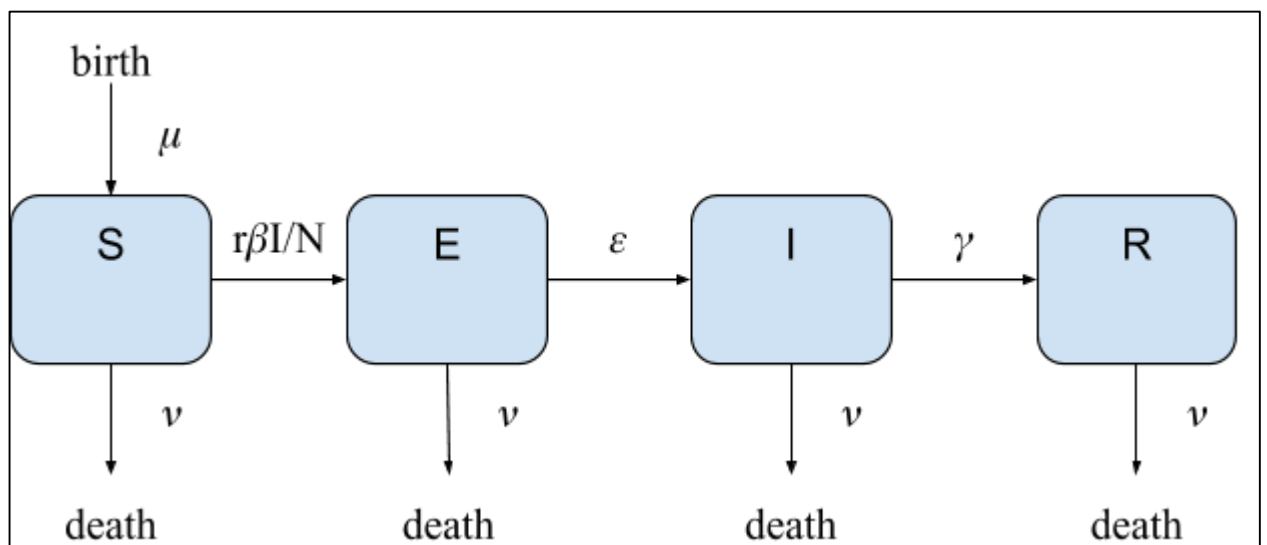


Рисунок 1.3 – Схематичне зображення моделі SEIR

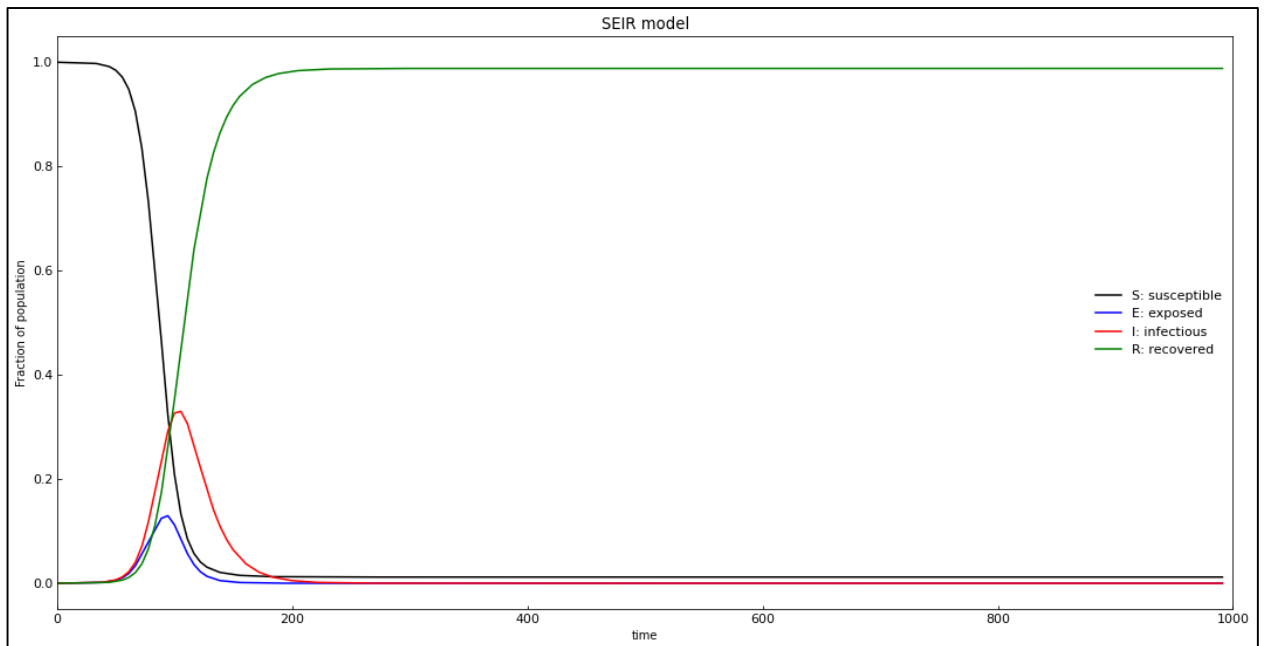


Рисунок 1.4 – Моделювання епідемії за допомогою моделі SEIR

1.4 Підхід до прогнозування захворюваності на основі фільтрації

Будь-який часовий ряд захворюваності можна розглядати як випадковий процес, компонентами якого є сигнал і шум, що відображають реальну епідемічну ситуацію. Фільтр – це алгоритм обробки даних, який може видаляти той самий шум, покращуючи якість сигналу. Фільтрація шуму дозволяє вказувати передбачення і може виконуватися під час попередньої обробки вихідних даних або як частина алгоритму передбачення.

1.4.1 Експоненціальне згладжування

Одним з найпростіших і поширених алгоритмів фільтрації шуму є експоненційне згладжування. Подібно до моделі ковзного середнього (MA), її прогнозом є зважена сума минулих спостережень. Принципова відмінність полягає в тому, що попереднє спостереження відноситься до індексу втрати

ваги, який враховує всі попередні спостереження в серії, а не тільки ті, що потрапляють у ковзне вікно. Загальна формула простого експоненційного згладжування:

$$S_t = \alpha \cdot x_t + (1 - \alpha) \cdot S_{t-1}, \quad (1.9)$$

де α – коефіцієнт згладжування ($0 < \alpha < 1$).

У деяких випадках для прогнозування захворюваності можна використовувати просте експоненціальне згладжування, але істотним недоліком цього методу є те, що він не враховує тенденції або сезонність. Тому частіше використовуються узагальнені моделі, наприклад, модель Холта-Вінтерса, яка підходить для опису сезонних захворювань, таких як грип. Адитивна сезонна модель Холта-Вінтерса описується наступним рівнянням [8]:

$$\begin{aligned} l_t &= \alpha(y_t - s_{t-T}) + (1 - \alpha)(l_{t-1} + r_{t-1}) \\ r_t &= \gamma(l_t - l_{t-1}) + (1 - \gamma)r_{t-1} \\ s_t &= \delta(y_t - l_t) + (1 - \delta)s_{t-T} \\ \hat{y}_{t+h} &= l_t + hr_t + s_{t-T+h} \end{aligned}, \quad (1.10)$$

де r_t відображає тренд;

s_t – сезонний фактор;

γ і δ – коефіцієнти згладжування;

T – період сезонності.

Методи прогнозування, засновані на експоненційному згладжуванні, дуже популярні в економіці і реалізовані практично в усіх пакетах статистичних програм. Узагальнена модель підходить для короткострокового

та довгострокового прогнозування захворюваності, але не так популярна, як регресія та авторегресія.

1.4.2 Методи теорії керування

Будь-які епідемічні процеси можна описати такою системою різницевих рівнянь:

$$\begin{aligned} x_t &= Ax_{t-1} + w_t \\ y_t &= Hx_t + Df_t + v_t \end{aligned} \quad (1.11)$$

де x_t – вектор змінних стану системи в момент часу t ;

y_t – вектор спостережень;

f_t – вектор значень зовнішніх факторів;

w_t і v_t – білий шум.

Матриця параметрів A , H , D представляє модель епідемічного процесу і вибирається відповідно до завдання (короткострокове або довгострокове прогнозування). Така форма запису дозволяє надати універсальну модель передачі захворювання. Для формування прогнозів можна використовувати методи теорії управління, особливо фільтрацію Калмана [9].

1.5 Підхід до прогнозування захворюваності із застосуванням методів машинного навчання

В останні роки інтерес людей до машинного навчання продовжує зростати. Це пов'язано з багатьма факторами, включаючи велику кількість

незалежних дослідницьких спільнот, які займаються різними додатками. У свою чергу, це дає світу робочі приклади використання методів машинного навчання для вирішення проблем, з якими раніше могли впоратися лише люди. Інновації у сфері машинного навчання та епідеміології не оминули.

1.5.1 Штучні нейронні мережі

Основним представником інтелектуальних методів дослідження епідемії на основі машинного навчання є штучні нейронні мережі (АНН). Метод ANN добре відомий і підходить для вирішення неможливих або складних задач в аналізі та дослідженнях. ШНМ – це орієнтований зважений граф (рис. 1.5), вершини якого імітують функцію біологічних нейронів. Вершини отримують вхідні сигнали і перетворюють їх у вихідні при досить великому значенні їх зваженої суми. Навчання ШНМ полягає в поступовому обчисленні коефіцієнта зв'язку між вершинами, які визначають силу вхідного сигналу. Ця процедура проводиться на основі статистики захворюваності і є значення для факторів, які її визначають [10].

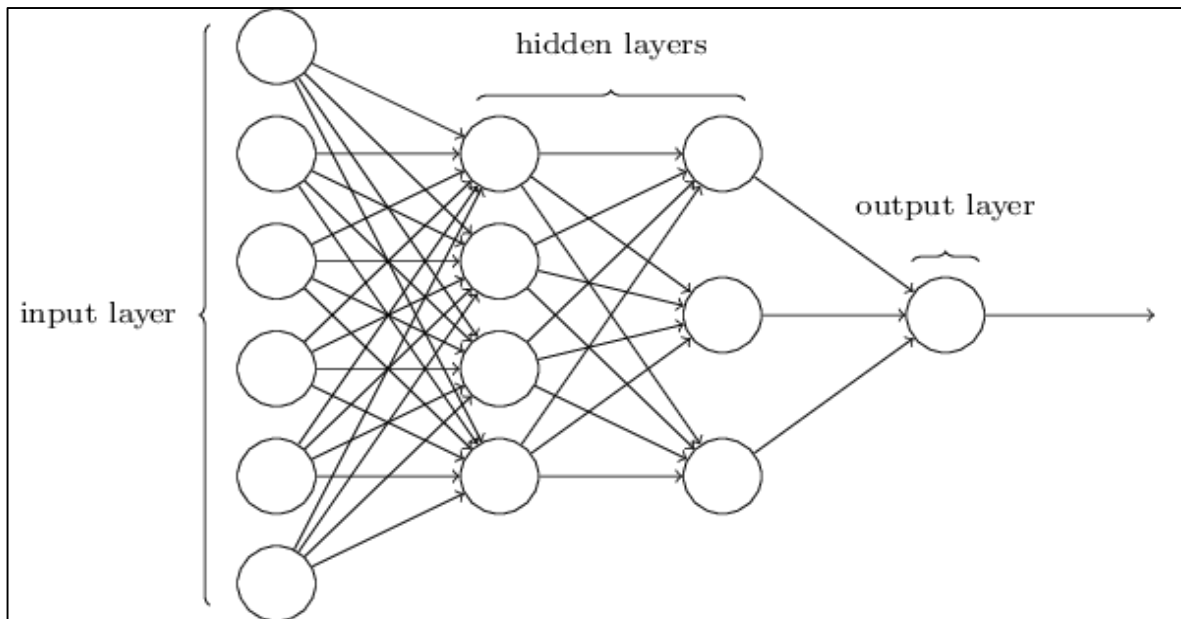


Рисунок 1.5 – ШНМ з двома рівнями прихованих нейронів

Існує багато варіантів використання ANN для прогнозування часового ряду захворюваності. Вони можуть відрізнятися за структурою, кількістю нейронів і типами функцій активації, а також методами навчання. Зокрема, можна розрізнити рекуррентні нейронні мережі та перцептрони. На останньому рівні латентні нейрони є набором лінійних класифікаторів, які, на жаль, не можуть встановити залежність даних від часу. Рекуррентна нейронна мережа може враховувати результати, отримані на попередньому етапі. Цікавою підкатегорією мережі цього типу є мережа довгострокової короткочасної пам'яті (LSTM), елементи якої описуються таким рівнянням:

$$\begin{aligned}
 f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 \tilde{C}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \\
 C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \\
 o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 h_t &= o_t \cdot \tanh(C_t)
 \end{aligned} \tag{1.12}$$

де x_t – вхідний вектор;

h_t – прихований вектор стану;

W – матриця параметрів;

b – вектор параметрів;

\tilde{C}_t – вектор кандидатів станів комірки;

C_t – вектор станів комірки;

f_t – вектор вентиля пам'яті;

i_t та o_t – вектори вентилів входу та виходу відповідно;

σ – сигмоїдальна функція;

\tanh – гіперболічний тангенс.

Перевага мереж LSTM перед традиційними рекурентними нейронними мережами полягає в тому, що вони можуть підтримувати довгострокові зв'язки в даних і робити їх більш ефективними, коли ці зв'язки важливі.

Однак, універсального методу вибору функцій ШНМ не існує — для кожного конкретного завдання вибирається окремий метод. Слід зазначити, що оскільки частота зазвичай зашумлена, то для навчання штучної нейронної мережі потрібна відносно велика кількість випадків. Це, у свою чергу, певною мірою обмежує використання таких моделей у довгостроковому прогнозуванні.

1.6 Висновки

У цьому розділі розглянуті деякі з найбільш поширених методів і методів прогнозування інфекційних захворювань, а також деякі з їх переваг і недоліків. Окрім математичних методів, кожен метод можна використовувати для аналізу та прогнозування інфекційної та неінфекційної захворюваності та навіть смертності. Крім того, вони підходять практично для будь-якої біологічної природи і наступних часових рядів. Основна мета математичних методів вивчення поширення захворювань – аналіз або чисельний аналіз

епідемічного процесу. Тому їх використання як інструментів прогнозування обмежується моделюванням поширення однотипної інфекції.

Регресійний аналіз є основним методом для прогнозування часових рядів будь-якої природи, і його можна легко реалізувати за допомогою будь-якої комп'ютерної математичної системи. Неадаптивні моделі дозволяють прогнозувати захворюваність у будь-який час, але такі моделі ігнорують локальні коливання поширеності та не підходять для короткострокового прогнозування. З іншого боку, адаптивна модель має на меті робити прогнози на найближчі кілька тижнів. При досить великій ширині ковзного вікна також можна розрахувати середню тривалість оцінок захворюваності, але досягти їх високої якості проблематично.

Авторегресійні моделі, як і моделі регресії, широко використовуються та є різноманітними. Вони відносно прості та добре досліджені, що робить їх легшими у використанні. Основним недоліком авторегресійних моделей є те, що вони вимагають великих обсягів даних для побудови точних прогнозів.

Хоча алгоритм фільтрації рідко використовується окремо, він може зменшити шум даних, що може значно покращити продуктивність інших моделей. Мабуть, найбільш гнучкими є обладнання машинного навчання, особливо нейронні мережі. Велика кількість можливих варіантів їх архітектури дозволяє краще налаштувати модель для кожного завдання, що, у свою чергу, зазвичай дає кращі результати. Однак, нейронні мережі важко інтерпретувати та налаштовувати, і вони вимагають великих обчислювальних витрат.

На жаль, не існує універсального методу прогнозування захворюваності – найкращий метод слід вибрати шляхом порівняння результатів, отриманих за допомогою різних методик на основі емпіричних даних. Часто буває важко вибрати метод прогнозування – кілька методів дають схожі результати. Крім того, реалізація одних з описаних методів вимагає значно більших витрат, ніж інших. Для того, щоб сформулювати обґрунтований прогноз із групи конкурентів, рекомендується використовувати засоби теорії прийняття

рішень, особливо мультистандартний аналіз. Нарешті, слід зазначити, що концепція термінового прогнозування певною мірою залежить від кроків вибірки часового ряду. Наприклад, якщо є ряд щорічних захворюваності, прогноз на наступний рік можна віднести до короткострокових.

2 КОРОНАВІРУС COVID-19

2.1 Особливості захворювання та методи дослідження

Коронавірусна інфекція (COVID-19) – це захворювання, що викликається SARS-CoV-2, новим типом вірусу в родині коронавірусів. У грудні 2019 року в Ухані, Китай, було виявлено перший відомий випадок. Згідно з повідомленнями, вірус походить від тварин, можливо, від кажанів. Поширення вірусу може бути пов'язане з впливом ринку морепродуктів, але точне джерело передачі вірусу від людини до людини ще не встановлено [11]. 11 березня Всесвітня організація охорони здоров'я (ВООЗ) підтвердила спалах коронавірусу як пандемію, назвавши його «глобальним поширенням нової хвороби» (рис. 2.1).

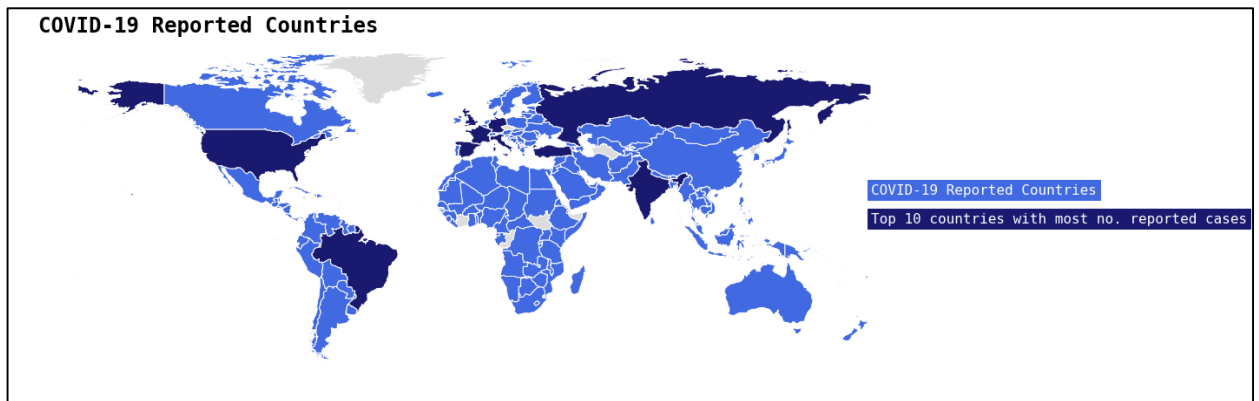


Рисунок 2.1 – Епідемія COVID-19 у світі

2.1.1 Огляд захворювання та його основних епідеміологічних характеристик

Головний біль, кон'юнктивіт, втрата смаку або нюху, висипання або зміна кольору пальців рук або ніг. При складному процесі захворювання

можуть виникати утруднення дихання або задишка, біль у грудях, афазія і навіть рух. У цьому випадку інфекція, викликана вірусом SARS-CoV-2, може викликати дихальну недостатність, септичний шок, органну недостатність і навіть смерть [12]. На рисунку 2.2 показано кількість таких серйозних захворювань у світі.

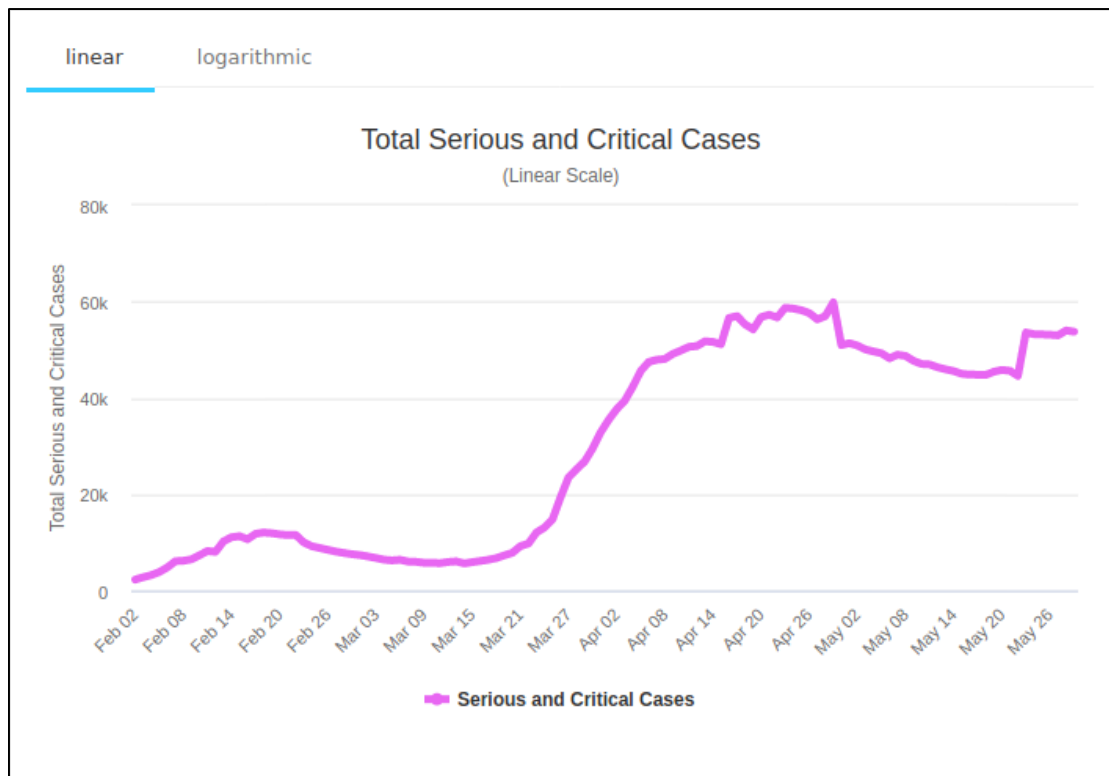


Рисунок 2.2 – Кількість важких перебігів захворювання коронавірусом із часом

Важливими характеристиками захворювання є тривалість інкубаційного періоду, середній час зараження, відсоток смертей та відсоток людей, які одужали. Дослідження показали, що в Китаї кількість померлих становить близько 3,5% від загальної кількості випадків. Проте в Італії це значення набагато вище, наближаючись до 10% протягом кількох тижнів, що вказує на її сильну залежність від демографічного складу. В Україні смертність оцінюється в 2,5%, тоді як рівень смертності у світі в якийсь момент досяг 22% (рис. 2.3). Однак, рівень смертності є різним у різні хвилі захворюваності, залежно від типу штаму, який превалює.

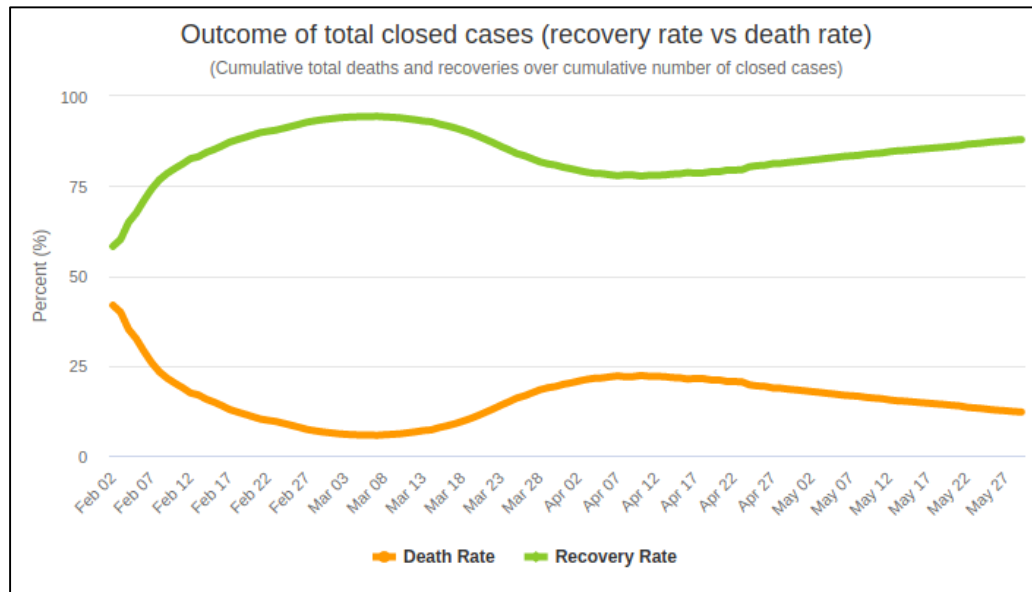


Рисунок 2.3 – Порівняння проценту пацієнтів, що померли та одужали

Тривалість інкубаційного періоду – це період від моменту потрапляння інфекції в організм до появи перших симптомів інфекції. За оцінками, цей пробіл є суттєвим для ефективного впровадження карантинних обмежень. Дослідження показали, що його тривалість в основному становить від 2 до 14 днів, але також можуть бути випадки, коли перші симптоми з’являються від 0 до 27 днів після зараження. Тому цей параметр може відрізнитися від пацієнта до пацієнта. В середньому оцінюється 5,2 дня.

Інфекційний період – це час, протягом якого пацієнт може поширювати інфекцію. Цей параметр істотно впливає на правила ізоляції та виявлення. Середня тривалість COVID-19 становить 9,5 днів.

Знаючи приблизну кількість випадків, що діагностуються щодня, і середній період зараження, можна оцінити два дуже важливі параметри захворювання – основний індекс відтворення, також відомий як R_0 , і ефективний індекс репродукції- R . Обидва ці параметри характеризують, скільки людей кожен буде поширювати інфекцію. Різниця між ними полягає в тому, що R додатково враховує частку населення з імунітетом (Рівняння 2.1).

$$\begin{aligned} R_0 &= \beta \cdot \tau \\ R &= R_0 - R_0 \cdot I \end{aligned} \quad (2.1)$$

де β – кількість контактів з передачею захворювання за одиницю часу;

τ – середній період інфекційності;

I – частка популяції з імунітетом.

За оцінками, індекс R_0 COVID-19 становить від 2 до 3. Індекс R сильно залежить від демографічних факторів та перебігу епідемії. Не менш важливим параметром епідемії є кількість проведених тестів. Чим точніший результат виявлення та більше часу виявлення, тим ближче дані будуть до справжньої кривої популярності. Тому небезпека слабого виявлення полягає в тому, що інформація про кількість захворювань може сильно відрізнятись від фактичного значення, що може призвести до неправильних або запізнених відповідей уряду, як це сталося у Великобританії. На початку епідемії COVID-19 якість тесту була не дуже високою, а масштаб був невеликим, як показано на рисунках 2.4 і 2.5.

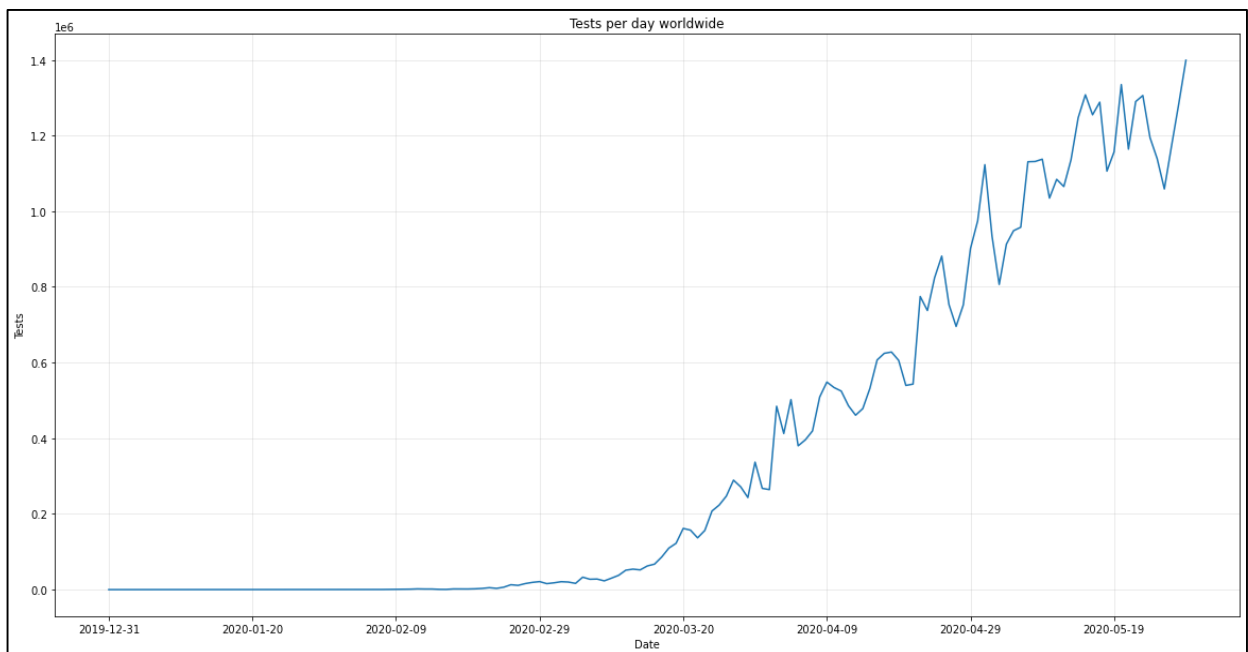


Рисунок 2.4 – Кількість щоденно проведених тестувань у світі

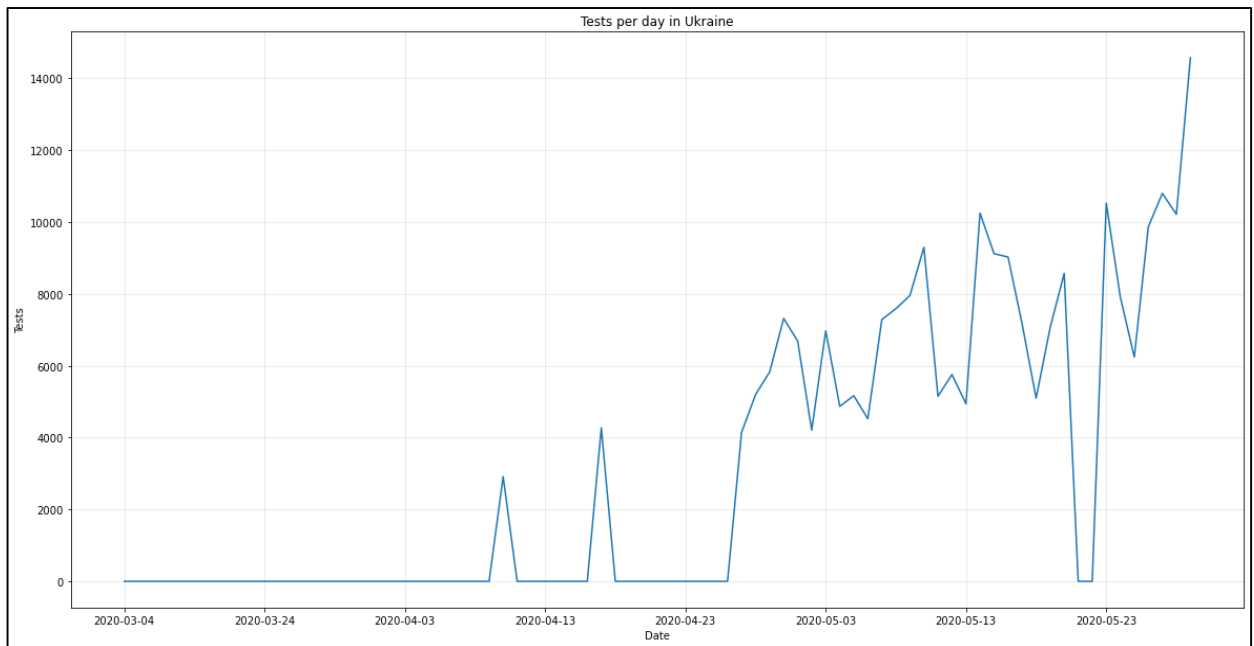


Рисунок 2.5 – Кількість щоденно проведених тестувань в Україні

2.1.2 Порівняльний аналіз епідемій, викликаних коронавірусами SARS-CoV-2, SARS-CoV та MERS-CoV

Подібно до COVID-19, інший коронавірус вперше був виявлений на півдні Китаю в листопаді 2002 року і отримав назву SARS-CoV. Воно спровокувало епідемію SARS, яка призвела до 774 смертей і понад 8000 випадків захворювання, поширилася на 29 країн на всіх континентах (рис. 2.6) і була подолана в липні наступного року. Ряд досліджень і Всесвітня організація охорони здоров'я стверджують, що коронавірус, який викликає це захворювання, є нащадком кажанів, і він передається людині через проміжну тварину – цивету (або віверу). Ринок тварин також вважається центром передачі від господаря до людини, подібно до спалаху COVID-19. Основними симптомами ГРВІ є лихоманка, головний біль і діарея. Порівняно з COVID-19, хоча індекс R_0 оцінюється в межах від 2 до 4 (у COVID-19 R_0 коливається від 2 до 3), епідемію ГРВІ швидко взяли під контроль.

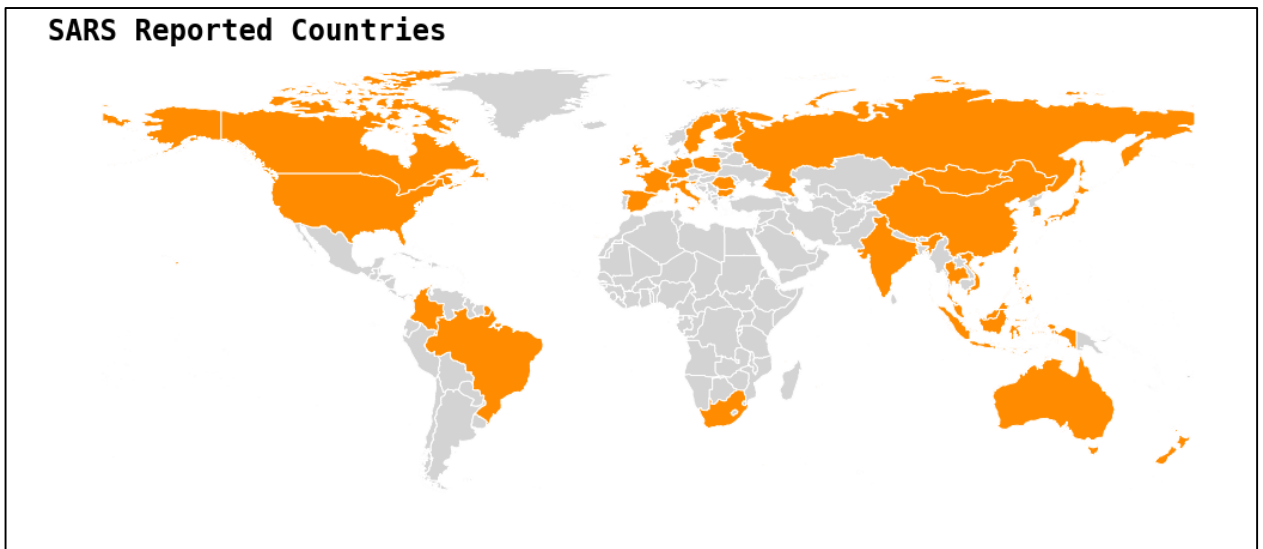


Рисунок 2.6 – Епідемія SARS у світі

MERS-CoV, інший вірус сімейства коронавірусів, спричиняє передачу близькосхідного респіраторного синдрому, такого як COVID-19 та SARS, до людей через тварин. Очевидно, що основними носіями є кажани, або верблюди. Дуже важко передаватися від людини до людини, тільки при тісному контакті. Клінічні симптоми MERS дуже схожі на ГРВІ, але перебіг захворювання більш агресивний, а смертність вища. Проте низький рівень передачі епідемії є сприятливим епідеміологічним фактором. Деякі дослідники вважають, що більшість хворих – не від інших людей, а – від хворих тварин. Станом на кінець січня 2020 року на Близькому Сході було загалом 2519 лабораторно підтверджених випадків респіраторного синдрому, у тому числі 866 смертей (смертність: 34,3%). Більшість із цих випадків (2121) припали на Саудівську Аравію, у тому числі 788 померли. Зона передачі вірусу показана на рисунку 2.7.

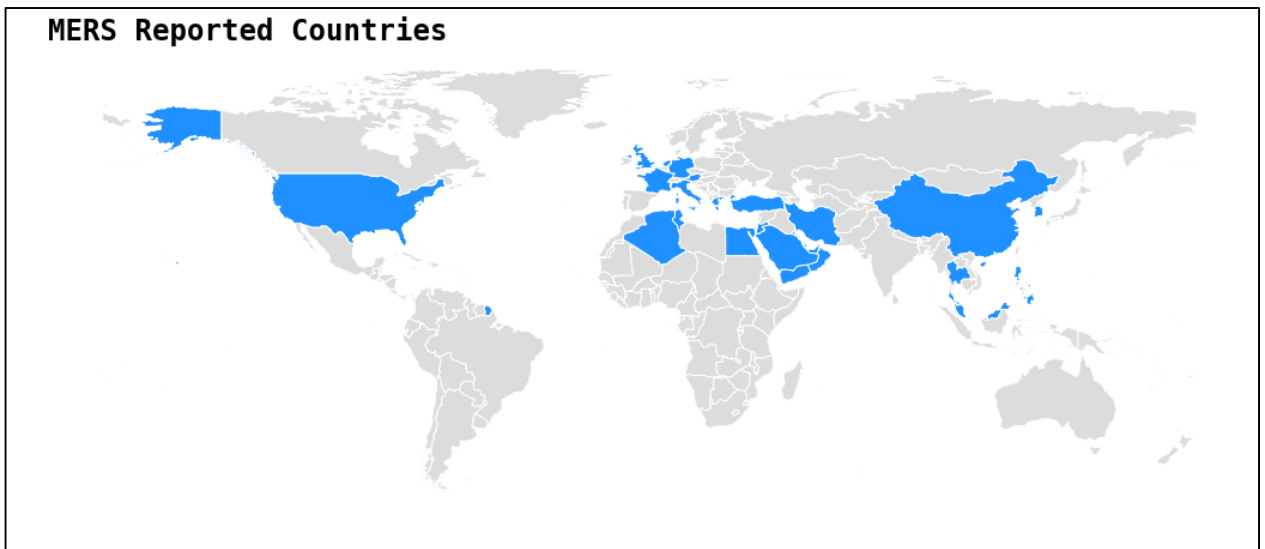


Рисунок 2.7 – Епідемія MERS у світі

Порівнюючи кумулятивні криві поширеності SARS-CoV-2, SARS-CoV та MERS-CoV після 100-го зараження (рис. 2.8), можна зробити висновок, що крива, що описує коронавірус COVID-19 у перші два тижні, дуже схожа на MERS. Крива, але після другого тижня спостерігається стрибок, що наближає його до кривої ГРВІ. На четвертому тижні кількість відомих випадків коронавірусу значно зросла, перевищивши криву ГРВІ більш, ніж у 5 разів. Одним із ключових факторів, що призводять до такого швидкого зростання, є збільшення кількості та якості тестів. З огляду на це, можна зробити висновок, що у 100 випадках COVID-19 справжній поріг був набагато раніше, але не привертав уваги.

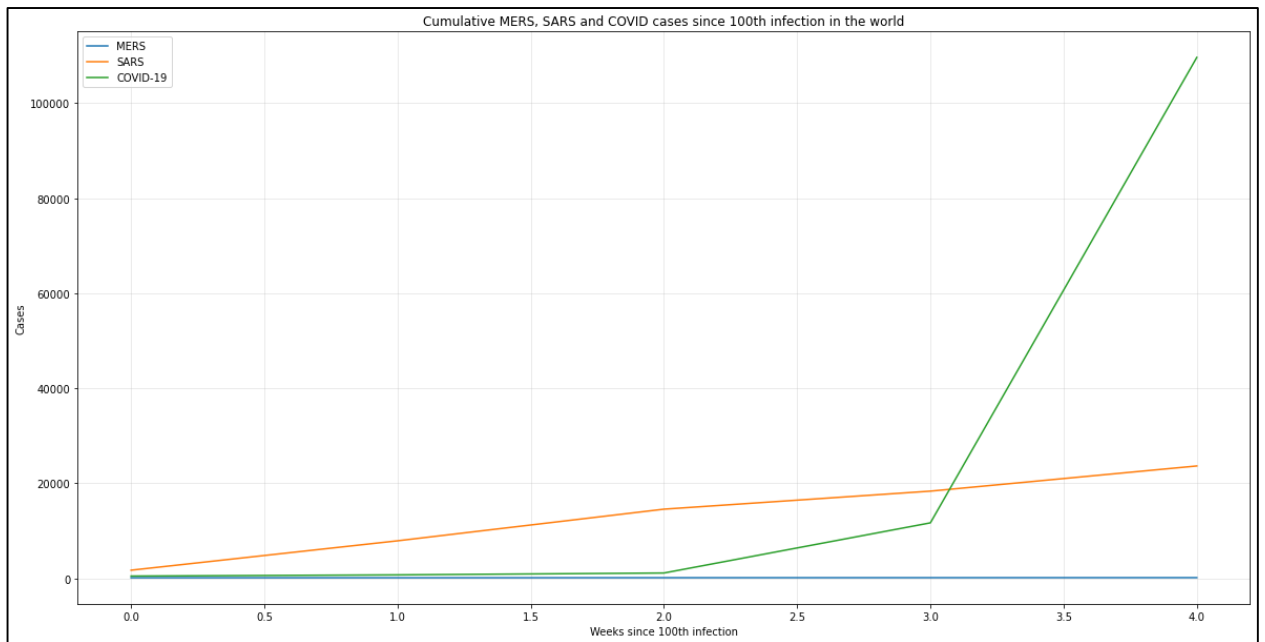


Рисунок 2.8 – Порівняння кумулятивних кривих епідемій SARS-CoV-2, SARS-CoV та MERS-CoV

2.2 Моделі для аналізу поширення коронавірусної інфекції

COVID-19 Коронавірус – це нове захворювання, яке ускладнює вивчення його поширення. Сьогодні ми не знаємо ні його сезонності, ні можливості втрати набутого імунітету. Однак, незважаючи на відносно короткий цикл моніторингу захворювання, вже є відносно велика кількість даних з різних країн, що є дуже цікавим об'єктом дослідження. Їх можна використовувати для відстеження прямих змін у природі коронавірусу та його зв'язку з превентивними заходами, що вживаються в різних країнах/регіонах. Наразі одним із найпростіших способів аналізу поширення коронавірусу є порівняння епідемічних тенденцій із схожою демографією та динамікою зростання кількості інфікованих країн. Польща та Румунія за цими стандартами найбільш схожі на Україну (рис. 2.9). Тому навіть без моделі розвиток української епідемії можна приблизно спрогнозувати.

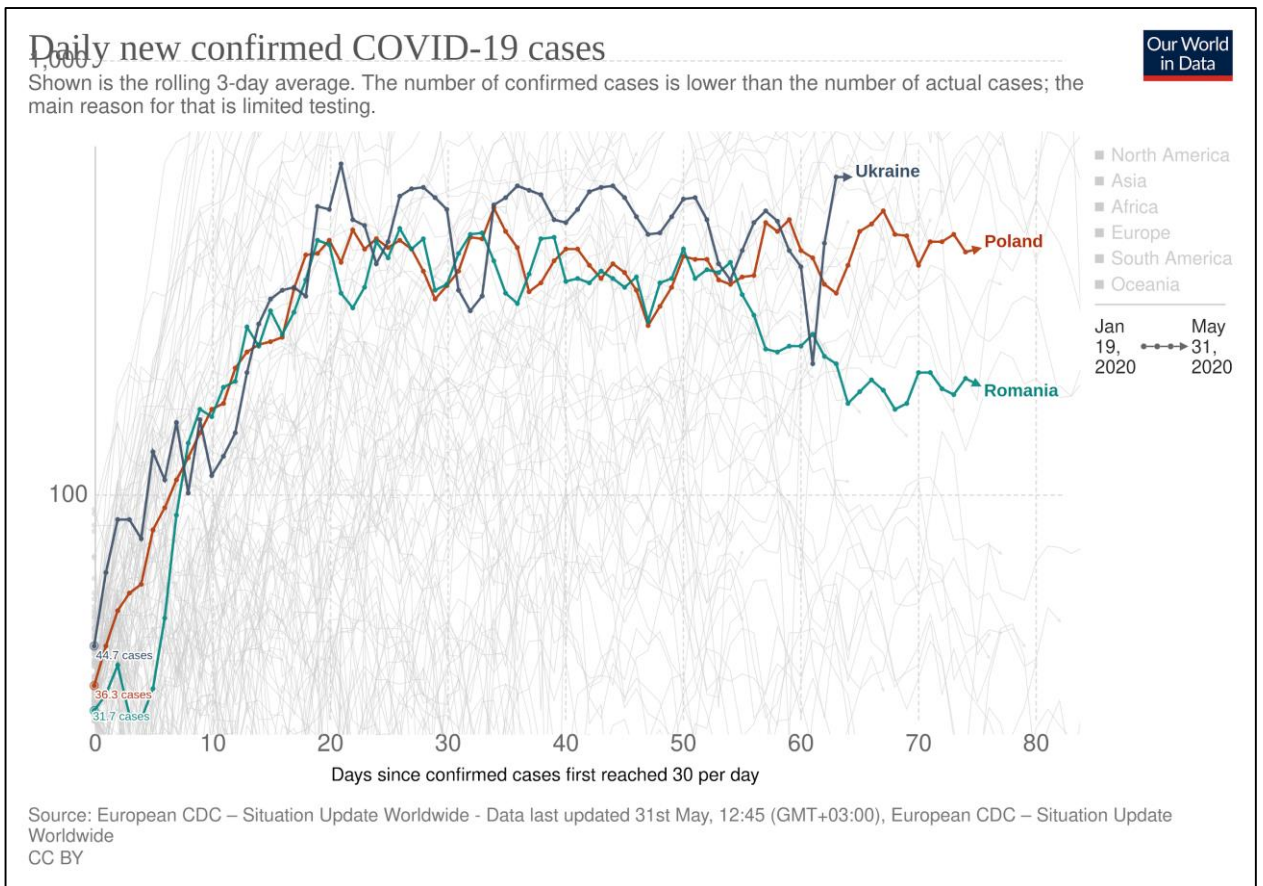


Рисунок 2.9– Порівняння трендів розвитку пандемії в Україні, Польщі та Румунії

2.2.1 Прогноз НАН України

На початку квітня 2020 року створено робочу групу при Президії Національної академії наук України для прогнозування поширення епідемії COVID-19. Команда розробила математичну модель SEIR-U та продемонструвала отримані з її допомогою прогнози.

Модель SEIR-U заснована на моделі SEIR, але додає відсіки для безсимптомних пацієнтів і пацієнтів з-за кордону. Він також може описати три стани: важкий, що вимагає госпіталізації та вентиляції, помірний, що вимагає спеціального медичного нагляду, і легкий, що не вимагає додаткового медичного втручання або місця в інфекційному відділенні. Це робить модель дуже гнучкою, вона може не тільки передбачити масштаби COVID-19, але й кількість необхідних лікарняних ліжок. Приклад прогнозу кількості

захворювань і смертей на добу, отриманий за допомогою цієї моделі, показаний на рисунку 2.10.

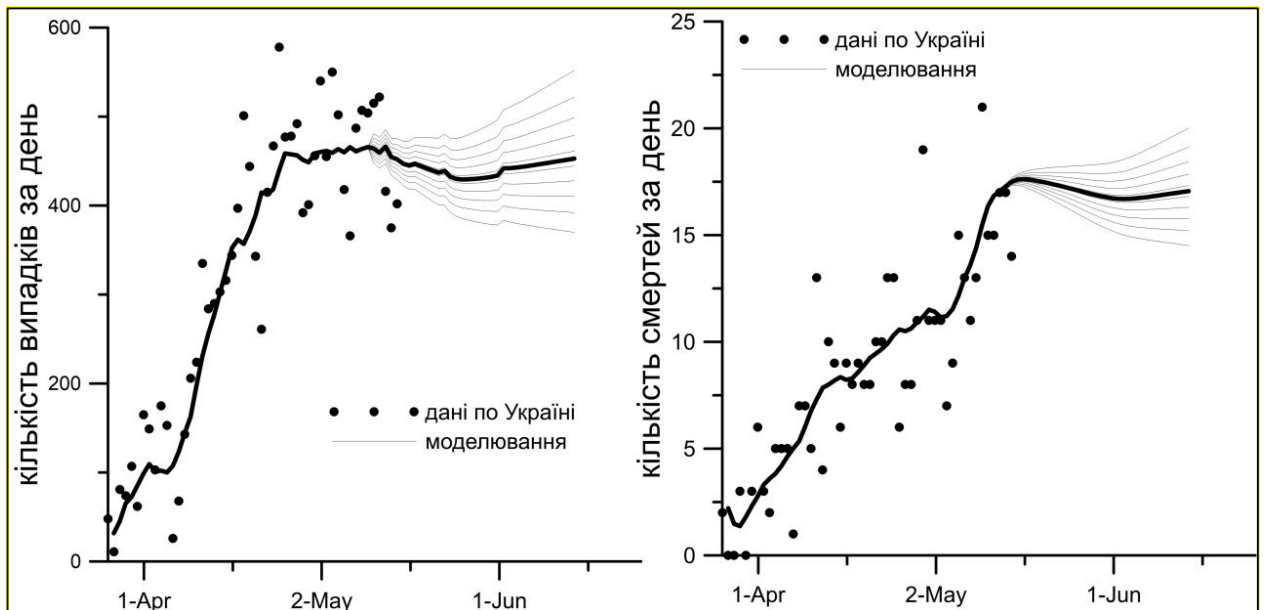


Рисунок 2.10 – Прогноз моделі SEIR-U

2.2.2 Прогноз за допомогою нейронної мережі

Група дослідників з Китаю запропонувала зовсім інший метод прогнозування коронавірусу COVID-19. Вони використовують повторювані нейронні мережі, такі як довгострокова пам'ять (LSTM), структура якої показана на рисунку 2.11.

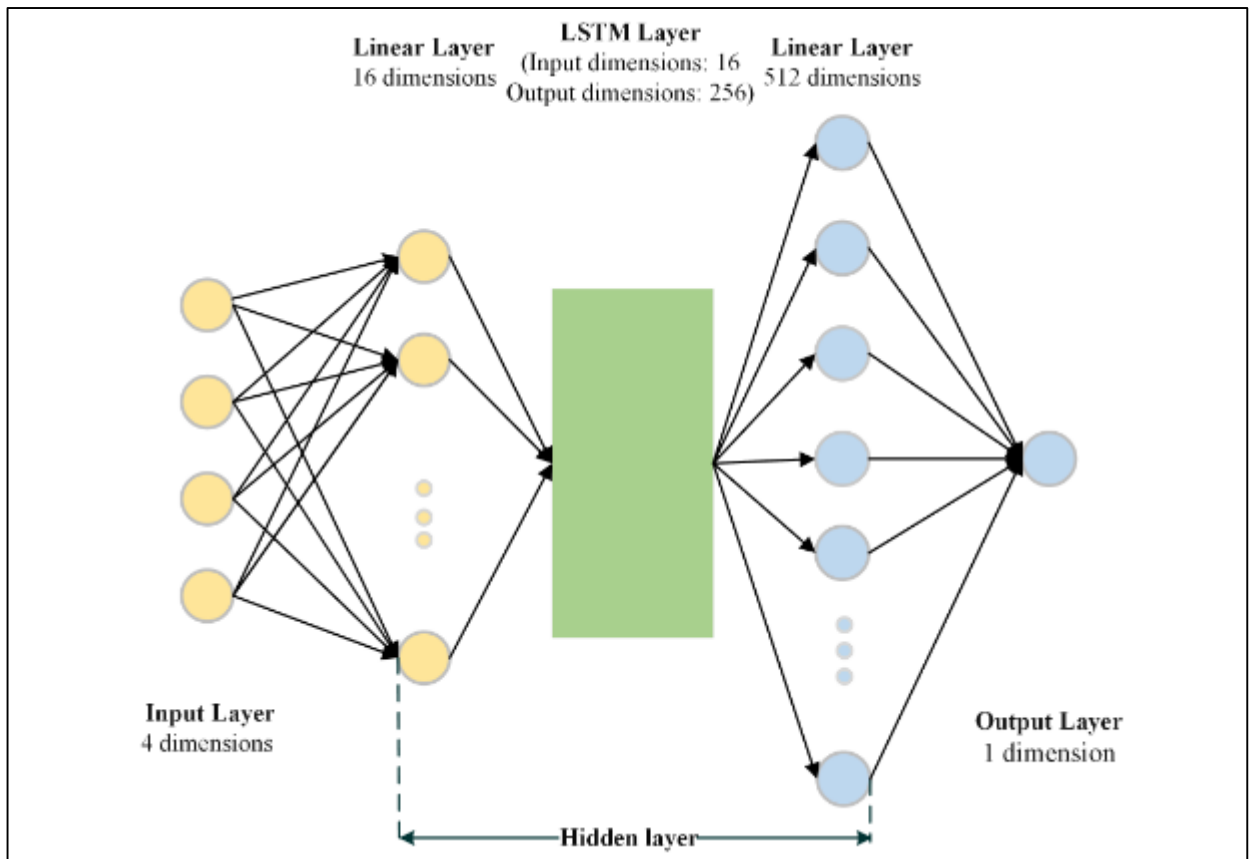


Рисунок 2.11 – Архітектура мережі LSTM

Сама модель будується в кілька етапів. По-перше, слід зібрати дані про випадки з різних країн і регіонів, а саме кумулятивну кількість випадків, кількість нових випадків за день та дані про карантин. Потім використовувати повністю підключену нейронну мережу для стандартизації та обробки даних. Нарешті, передасте перетворені дані в LSTM для введення. Використовуючи отриману модель, дослідники змогли зробити досить точні короткострокові прогнози кількості нових захворювань у різних провінціях Китаю. Однак, для таких моделей треба чимало даних.

2.3 Опис обраних методів

Сьогодні Facebook — це програмне забезпечення з відкритим вихідним кодом Prophet (з англ.: «Пророк»), яке є інструментом прогнозування, який можна використовувати в Python і R. Прогнозування — це завдання

дослідження даних, яке є важливим для багатьох видів діяльності організації. Наприклад, великі організації, такі як Facebook, повинні планувати можливості для ефективного розподілу обмежених ресурсів і встановлювати цілі для вимірювання еталонної ефективності. Для машин або більшості аналітиків створення високоякісних прогнозів не є легким завданням. У практиці створення різноманітних бізнес-прогнозів ми спостерігали дві основні теми: Повністю автоматизовані методи прогнозування можуть бути крихкими, і вони часто занадто негнучкі, щоб включати корисні припущення чи евристики.

Аналітики, які вміють робити якісні прогнози, зустрічаються дуже рідко, тому що прогнозування — це професійна наука про дані, яка вимагає великого досвіду.

Результатом цих тем є те, що попит на високоякісні прогнози часто значно перевищує швидкість, з якою аналітики можуть генерувати прогнози. Це спостереження є нашою мотивацією для створення роботи Пророка: ми хочемо полегшити експертам і неекспертам робити якісні прогнози, які задовольняють попит.

Типові міркування, пов'язані з «масштабом», розрахунками та зберіганням, не мають нічого спільного з прогнозуванням. Ми виявили, що проблеми розрахунку та інфраструктури для прогнозування великої кількості часових рядів є відносно простими – ці процеси налаштування зазвичай легко розпаралелювати, а прогнози не важко зберігати в реляційних базах даних (наприклад, MySQL) або репозиторіях.

Широкомасштабні проблеми, які ми спостерігали на практиці, включають складність, викликану різноманітністю проблем прогнозування та підвищену впевненість у створенні великої кількості прогнозів. Пророцтво є ключовим фактором у покращенні здатності Facebook створювати велику кількість надійних прогнозів для прийняття рішень і навіть функцій продукту.

Не всі проблеми прогнозування можна вирішити за один процес. Prophet оптимізовано для завдань бізнес-прогнозування, які зустрічаються у Facebook. Ці завдання зазвичай мають одну з наступних характеристик:

- Щогодинні, щоденні або щотижневі спостереження, принаймні кілька місяців (бажано рік) історії;
- Потужний багатосезонний «людський масштаб»: день тижня і пора року;
- Важливі свята, що відбуваються час від часу, відомі заздалегідь (наприклад, День Незалежності країни);
- Розумна кількість відсутніх спостережень або великі викиди.

Історичні зміни тенденцій, наприклад, через випуск продуктів або зміни тенденцій журналу, ці тенденції є нелінійними кривими зростання, і тенденція досягає своєї природної межі або насичення.

Ми виявили, що налаштування за замовчуванням Пророка створюють прогнози, які часто є точними, як прогнози досвідчених синоптиків, з набагато меншими зусиллями. З Prophet ви не застрягнете з результатами повністю автоматичної процедури, якщо прогноз незадовільний – аналітик, який не пройшов навчання методам часових рядів, може покращити або налаштувати прогнози, використовуючи різні легко інтерпретовані параметри. Виявлено, що, поєднуючи автоматичне прогнозування з прогнозованими аналітиками для особливих випадків, можна охопити широкий спектр випадків використання бізнесу. Наступна діаграма ілюструє процес прогнозування, який працює у масштабі (рис. 2.12).

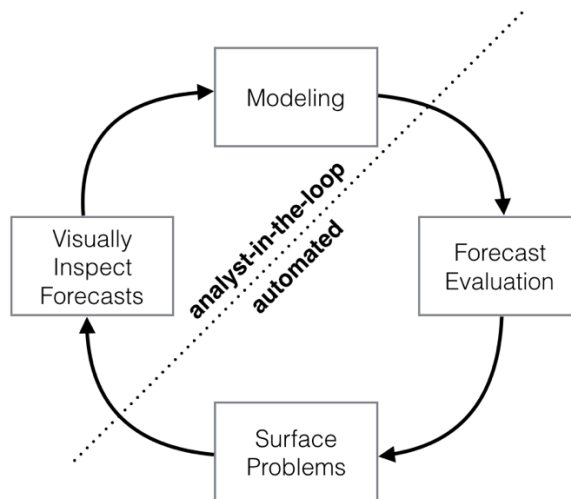


Рисунок 2.12 – Діаграма процесу прогнозування Prophet

Як було зазначено вище, поряд із моделлю Пророка поширення отримала й модель ARIMA.

ARIMA - це модель, яка використовується для прогнозування майбутніх тенденцій на основі даних часових рядів. Ця модель є формою регресійного аналізу.

- AR (авторегресія): Модель, яка показує змінну змінну, яка регресує за своїми власними відстроченими/попередніми значеннями;
- I (інтегрований): Відмінність неопрацьованих спостережень для того, щоб часові ряди стали стаціонарними;
- MA (ковзаюча середня): залежність між спостереженням та залишковою помилкою від моделі ковзного середнього.

Для моделей ARIMA стандартним позначенням буде ARIMA з p , d і q , де цілі значення замінюють параметри, щоб вказати тип використовуваної моделі ARIMA.

- P : кількість спостережень за відставанням у моделі; також відомий як порядок відставання;
- D : кількість разів, коли різняться неопрацьовані спостереження; також відомий як ступінь диференціації;
- Q : розмір вікна ковзної середньої; також відомий як порядок ковзної середньої.

Рівняння прогнозування будується наступним чином. По-перше, нехай y_t позначає d -ту різницю Y , що означає:

$$\text{Якщо } d = 0: y_t = Y_t$$

$$\text{Якщо } d = 1: y_t = Y_t - Y_{t-1}$$

$$\text{Якщо } d = 2: y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

Зауважте, що друга відмінність Y (випадок $d = 2$) – це не відмінність від 2 періодів тому. Швидше це – різниця першої різниці, тобто – першої, яка є дискретним аналогом другої похідної, тобто локального прискорення ряду, а не його локального тренду.

З точки зору y_t , загальне рівняння прогнозування виглядає так:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \dots - \theta_q e_{t-q},$$

де: $\mu \rightarrow$ постійна

$\phi_1 y_{t-1} + \dots + \phi_p y_{t-p} \rightarrow$ терміни AR (відставання значень y)

$-\theta_1 e_{t-1} - \dots - \theta_q e_{t-q} \rightarrow$ Терміни MA (відставання помилок)

Але недоліками ARIMA є те, що немає зручного інструментарію врахування аномалій («свят» та ін.) та він є чутливим до періодичності даних (не повинно бути пропусків у датах). Саме цих недоліків немає у моделі Пророк (Prophet).

Отже, бібліотека Prophet – зручний інструмент для прогнозування часових рядів. Модель працює досить добре з коробки, при цьому є набір гіперпараметрів, які дозволяють поліпшити прогноз. Окремо варто відзначити вбудовані можливості візуалізації.

2.4 Висновки

Завдання прогнозувати коливання хворих – складна. Особливо при такій не точності даних. SEIR-U модель потребує великої кількості вхідних даних, з яких більшість не є у відкритому доступі. Модель ARIMA вимагає регулярних даних, їй складно працювати з даними, на які впливають дати-аномалії та ще й вікном впливу (зараження в усіх проходить по-різному - симптоми мають

місце через 2-10 днів). Prophet добре апроксимує навчальну вибірку. Варто відзначити і розумно виділену річну сезонність.

Отримані передбачення можна використовувати як нову ознаку для інших алгоритмів машинного навчання.

Тому було і вибрано бібліотеку Prophet, так як для наших задач вона сама оптимальна та працює з святами.

3 АНАЛІЗ ТА ПРОГНОЗУВАННЯ КІЛЬКОСТІ НОВИХ ХВОРИХ НА КОРОНАВІРУС В УКРАЇНІ У 2021 РОЦІ

Для моделювання використовувались відкриті офіційні дані РНБО України, які оновлюються щодня і є доступними по API з веб-порталу (<https://covid19.rnbo.gov.ua/>) – саме у такий спосіб їх збирають усі світові веб-сервіси. Для інших країн варто використовувати відомий датасет Google-платформи «COVID- 19 Open Data» [17] 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, але, на жаль, не вся інформація там є достовірною. Аналіз показав, що там мають місце помилки по Україні, тому у дослідженні використовували не ці дані, а – дані РНБО України.

Згідно 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 чи за інших умов, що збільшує ризик поширення вірусу. Проведений авторами кореляційний аналіз даних показав, що доцільно брати запізнення на 7 днів, але з адаптивним вікном (вікно $[-1, 1]$ означає, що аномальними вважаються дати зі зсувом у 6-8 днів).

На етапі ідентифікації (тренування) моделі варто оптимізувати розмір адаптивного вікна навколо зсуву у 7 днів, наприклад, перебирати усі варіанти навколо нього в діапазоні від 4 до 10 днів (прирости $[-3, -2, -1, 0, 1, 2, 3]$) і обирати варіант з найменшою похибкою [13].

Досліджувався можливий вплив таких аномальних дат (свят і псевдосвят):

1. Державні свята (за даними пакету Holidays) [1] із 7-денним зсувом вперед (параметр “ds”) та адаптивним вікном (у першому наближенні беремо вікно $[-3, 3]$, тобто від 4 до 10 днів) (рис. 3.1).

	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

Рисунок 3.1 – Державні свята України (за даними пакету Holidays), 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 «метеопаттерни» (за даними датасету «COVID-19 Open Data», зокрема, за даними National Oceanic and Atmospheric Administration (NOAA) — Національного управління з питань океану та атмосфери США [1]. Пропонуємо відбирати дати зі зсувом 7 днів та адаптивним вікном (рис. 3.2), коли кількість опадів була нульова, а середньодобова температура була більша за квантиль P95, тобто значення, вищі цього, мали місце тільки у 5% випадків.

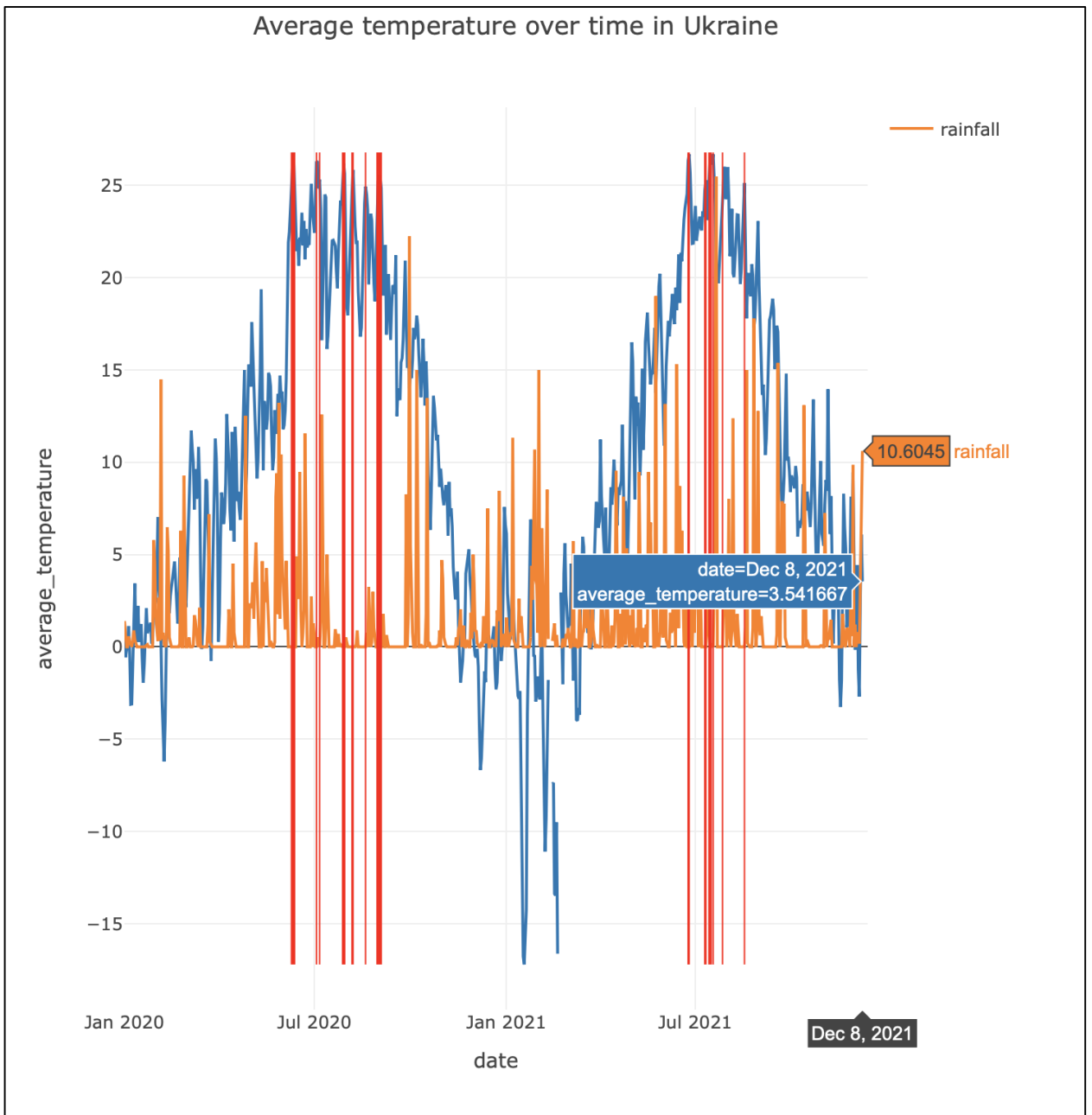


Рисунок 3.2 – Середньодобова температура в Україні (помаранчева крива), кількість опадів, мм (синя крива), дати аномальних «метопаттернів» (червоні вертикальні лінії)

3. Дні послаблення карантину за Stringency-індексом (за даними «Oxford COVID-19 government response tracker» — Оксфордського трекеру коронавірусної діяльності урядів країн світу, 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. 3.3) [14].

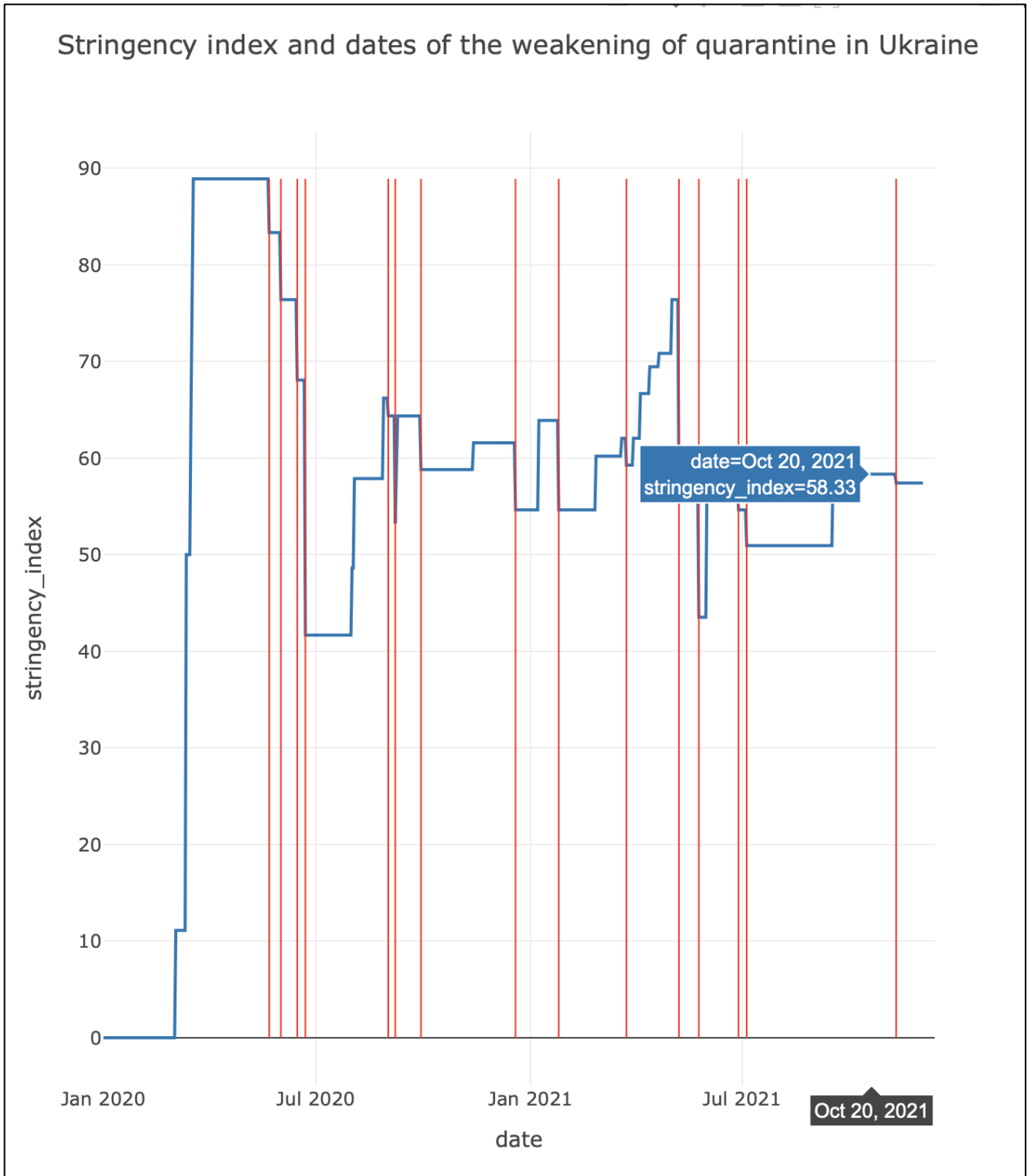


Рисунок 3.3 – Сумарний за добу Stringency-індекс України з ослабленнями та підсиленнями карантину (синя лінія) та дати аномалій-послаблень карантину (червона лінія)

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

Усі відібрані свята (вже з урахуванням зсуву у 7 днів, де він був зроблений) показані на рисунку 3.4 вертикальними лініями.

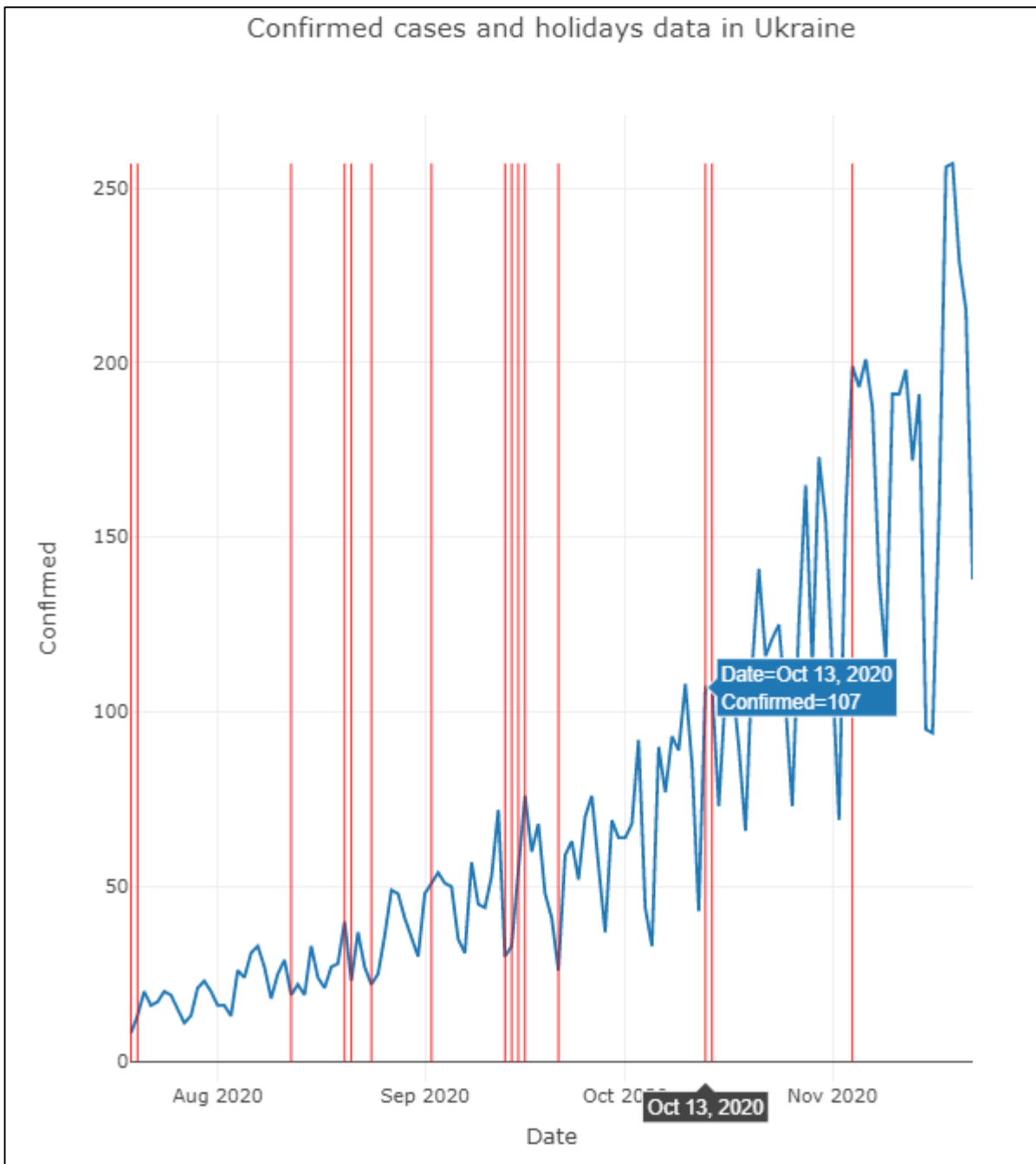


Рисунок 3.4 – Щоденні значення кількості нових випадків захворювань на коронавірус в Україні (синя крива) та дати аномалій (червоні вертикальні лінії)

3.1 Розроблення алгоритму інформаційної технології

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

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

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

- мультиплікативність чи адитивність 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: $n = 2, 3, \dots, 6$ 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 [15].

Як було обґрунтовано вище, для побудови моделі «хвилі», пік якої ще не досягнуто, варто використовувати лінійну регресію Prophet. Така регресія традиційно будується у вигляді кусково-лінійної апроксимації тренду між точками суттєвої зміни значень (за замовчуванням береться 30 таких точок, але модель Prophet їх кількість адаптивно оптимізує – може й усі видалити) [16].

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, яку ще називають як метрика WAPE (Weighted Mean Average Percentage Error), на усіх датах валідаційної вибірки.

Перебір усіх комбінацій таких параметрів – довготривала NP-задача. Тому, для прискорення роботи програми 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 [17].

Для розв'язання поставленої задачі розроблено алгоритм інформаційної технології з опрацювання вхідних даних та налаштування і застосування моделі, яка містить такі етапи (рис. 3.5):

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 $[-3, \dots, 3]$ (усі цілі числа і нуль), сили впливу (масштабу) свят і псевдосвят та варіантів режиму урахування свят (мультиплікативний чи адитивний).

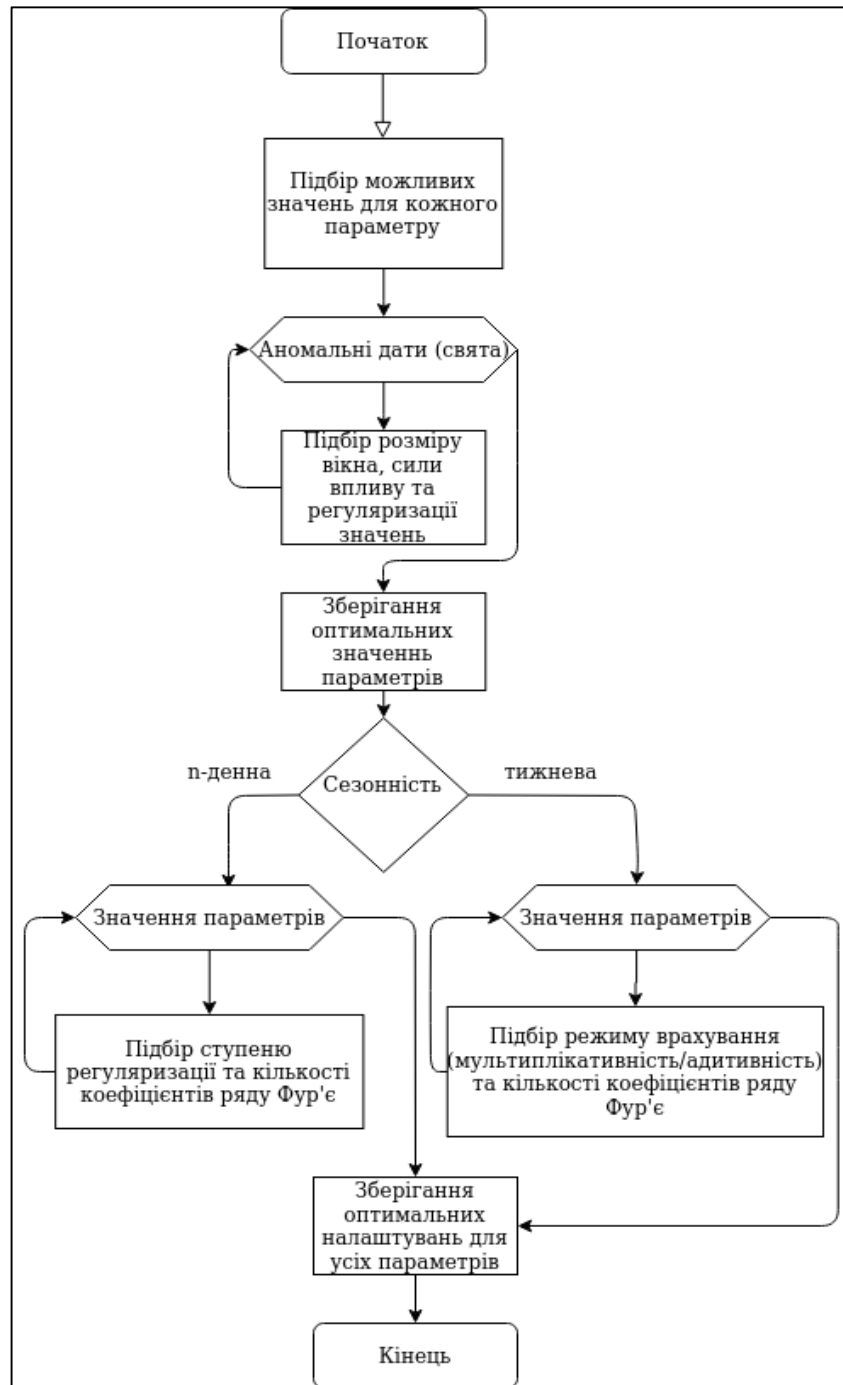


Рисунок 3.5 – Алгоритм розробленої інформаційної технології

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 кількість коефіцієнтів ряду Фур'є для

опису тижневої (7-денного) та, окремо, 4-денного видів 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.

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

- врахування тільки державних свят з 7-денним зсувом та адаптивним вікном;
- врахування тільки тижневої сезонності.

Для такої спрощеної моделі достатньо лише одного етапу ідентифікації, на якому варто оптимізувати тільки такі параметри (усі складові варто одразу враховувати мультиплікативно): розмір вікна, сила впливу (масштаб) і ступінь регуляризації значень в аномальні дати (свята). Інші – задавати фіксованими, на основі попередніх розрахунків. Таку спрощену модель можна застосовувати одразу для досить великої кількості країн, областей [18].

3.2 Короткостроковий прогноз

3.2.1 Короткостроковий прогноз з усіма хвилями

Здійснимо випробування вибірки даних за 06.07.2020 р. по 12.12.2021 р. Дані беремо з Системи моніторингу поширення епідемії коронавірусу Апарату РНБО України.

Відображаємо підтверджені дані хворих на коронавірус за цей період на рисунку 3.6.

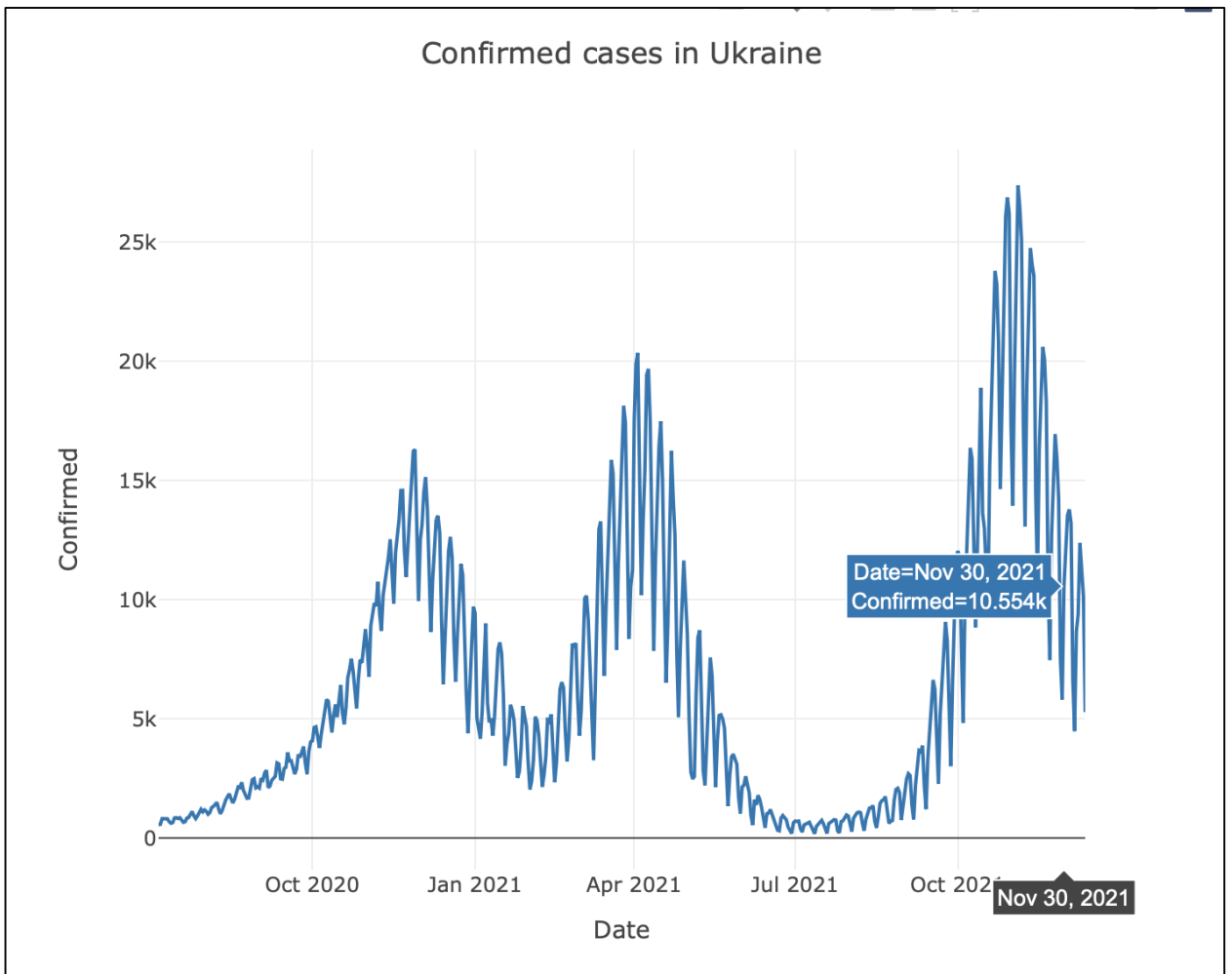


Рисунок 3.6 – Підтверджені дані хворих на коронавірус

Запущена модель, яка прогнозує дані на 14 днів (таблиця 3.1).

Таблиця 3.1 – Прогноз кількості нових підтверджених випадків хворих на коронавірус в Україні за моделлю

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
13.12.2021	0	541	3116
14.12.2021	0	2014	4475
15.12.2021	2125	4529	6869

Продовження таблиці 3.1

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
16.12.2021	2930	5467	7793
17.12.2021	3473	6060	8624
18.12.2021	1171	3767	6245
19.12.2021	0	0	1285
20.12.2021	0	0	0
21.12.2021	0	0	758
22.12.2021	0	553	2935
23.12.2021	0	1528	3925
24.12.2021	0	2048	4529
25.12.2021	0	7	2474
26.12.2021	0	0	0

Щоденна кількість нових хворих на коронавірус з 6 липня 2020р. зображена на рисунку 3.7, де чорні крапки – дані спостережень до 12.12.2021 р., синя лінія – результат моделювання і прогнозування на 2 тижні.

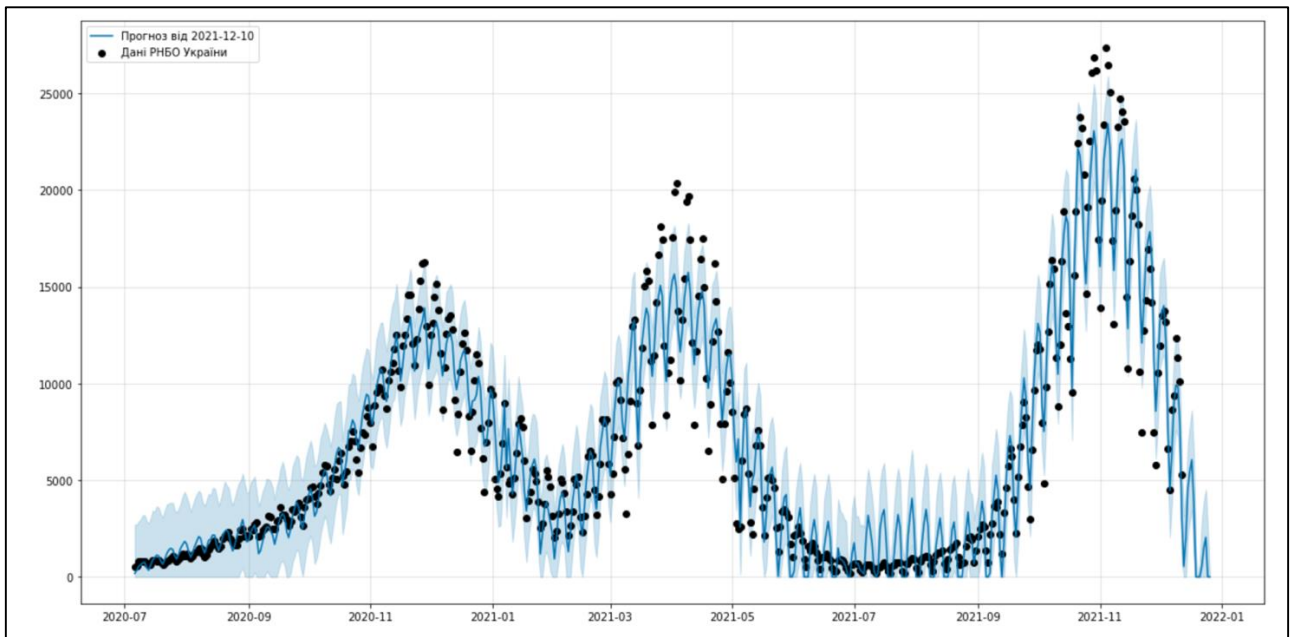
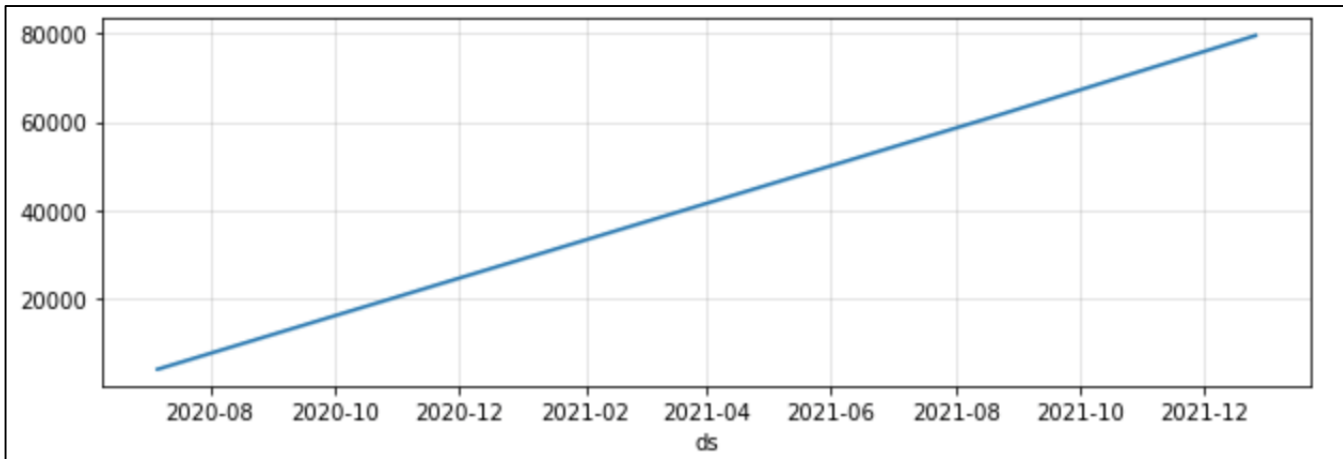
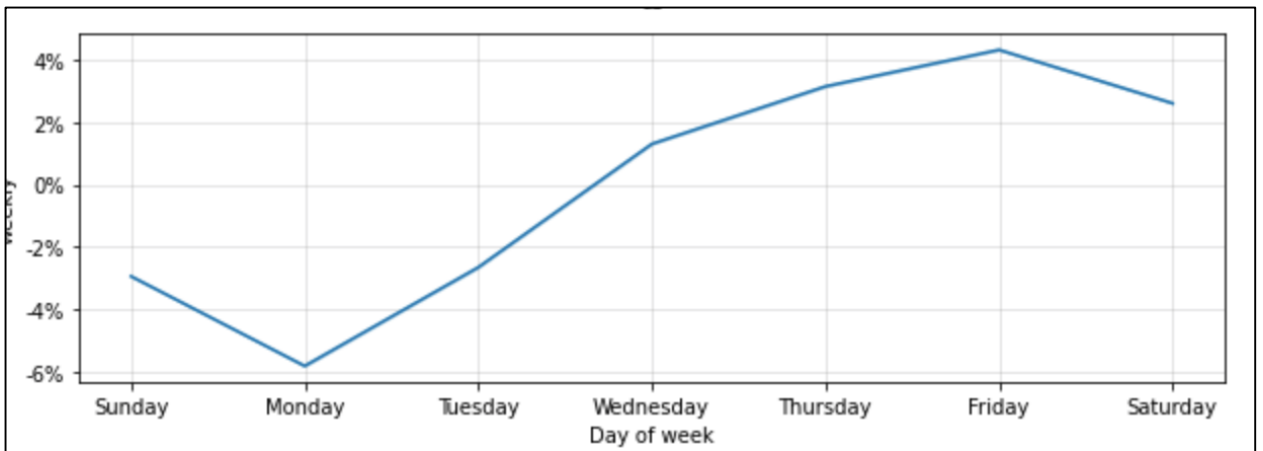


Рисунок 3.7 – Щоденна кількість нових хворих на коронавірус з 6 липня 2020р.

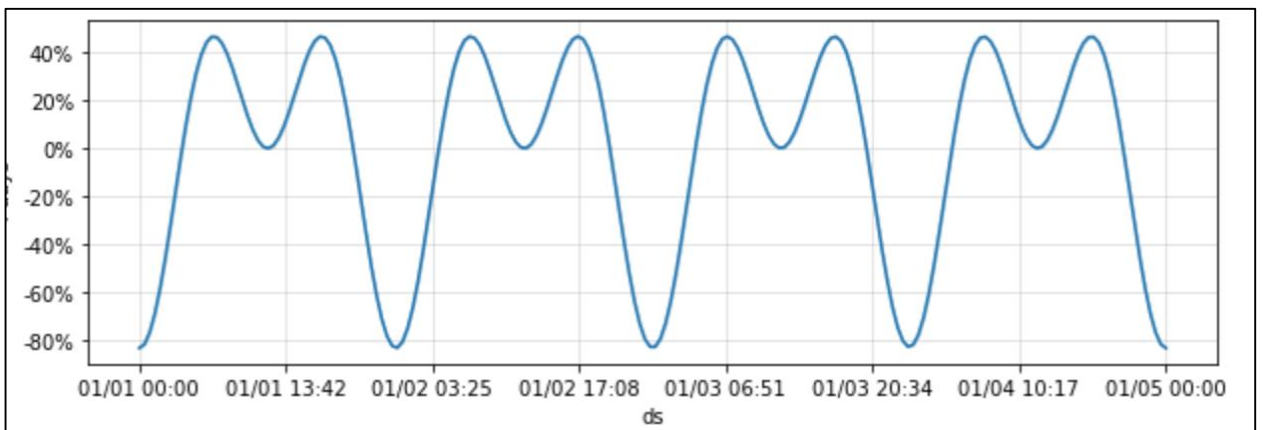
До 12.12.2021 р. за моделлю, відносна похибка прогнозування 2-х останніх тижнів спостережень становила 15%. Складові моделі зображені на рисунку 3.8.



а)



б)



в)

Рисунок 3.8 – Складові моделі 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

Також з цього рисунка 3.8 та отриманих даних випливає висновок про те, що прогноз не є адекватним, тому що прогнозні значення стрімко йдуть до нуля і далі у від'ємну область (такі дані примусово замінюються на нуль, оскільки від'ємні значення неможливі). Цей прогноз виглядає більш оптимістичним, але не реалістичним, тому що досвід моделювання каже, що це – на жаль, неможливо на даний період часу.

3.2.2 Короткостроковий прогноз з останньою хвилиною

Здійснимо випробування вибірки даних за 09.13.2021 р. по 12.12.2021 р. Дані беремо з Системи моніторингу поширення епідемії коронавірусу Апарату РНБО України.

Ідентифікована модель, яка прогнозує дані на 14 днів (таблиця 3.2).

Таблиця 3.2 – Прогноз кількості нових підтверджених випадків хворих на коронавірус в Україні за моделлю

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
13.12.2021	1267	3249	4986
14.12.2021	3540	5484	7327
15.12.2021	6861	8551	10326
16.12.2021	7362	9118	10841
17.12.2021	7551	9360	11135
18.12.2021	5842	7708	9592
19.12.2021	2352	4273	6024
20.12.2021	400	2402	4280

Продовження таблиці 3.2

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
21.12.2021	2898	4658	6630
22.12.2021	4903	6769	8698
23.12.2021	5656	7558	9515
24.12.2021	5753	7638	9606
25.12.2021	3476	5292	7290
26.12.2021	1402	3296	5182

Щоденна кількість нових хворих на коронавірус з 9 вересня 2021р. зображена на рисунку 3.9, де чорні крапки – дані спостережень до 12.12.2021 р., синя лінія – результат моделювання і прогнозування на 2 тижні.

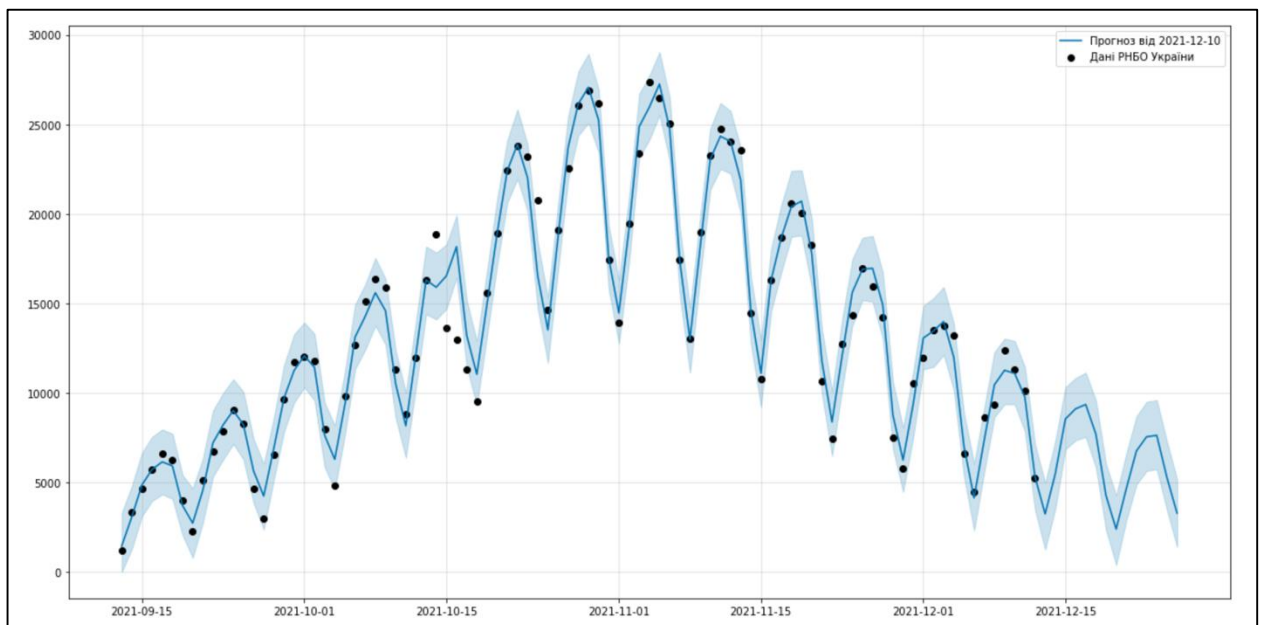
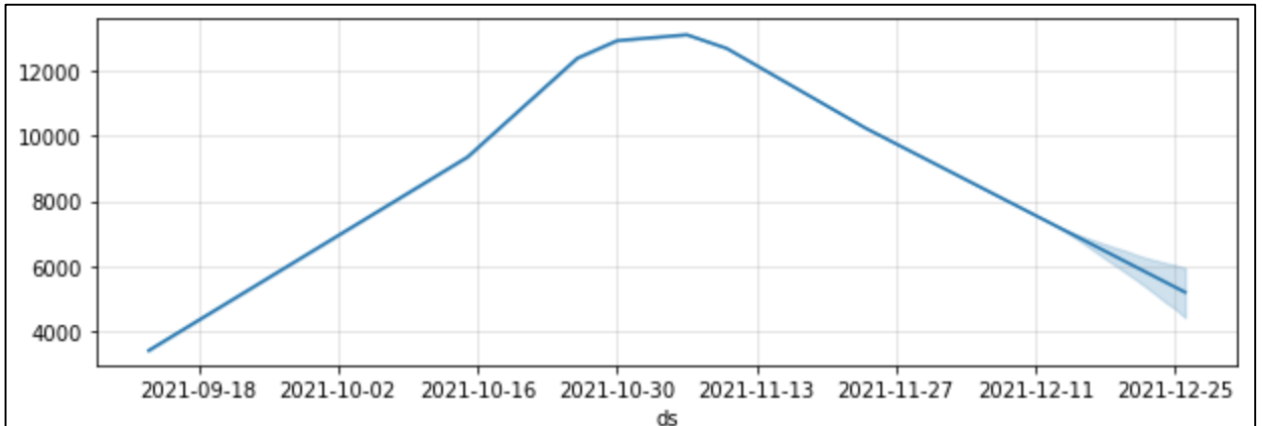
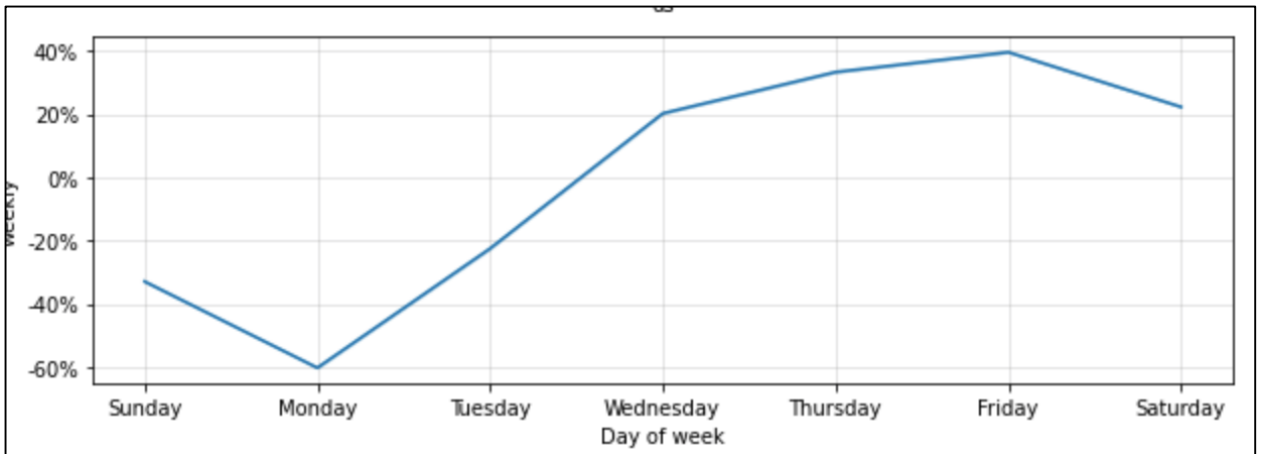


Рисунок 3.9 – Щоденна кількості нових хворих на коронавірус

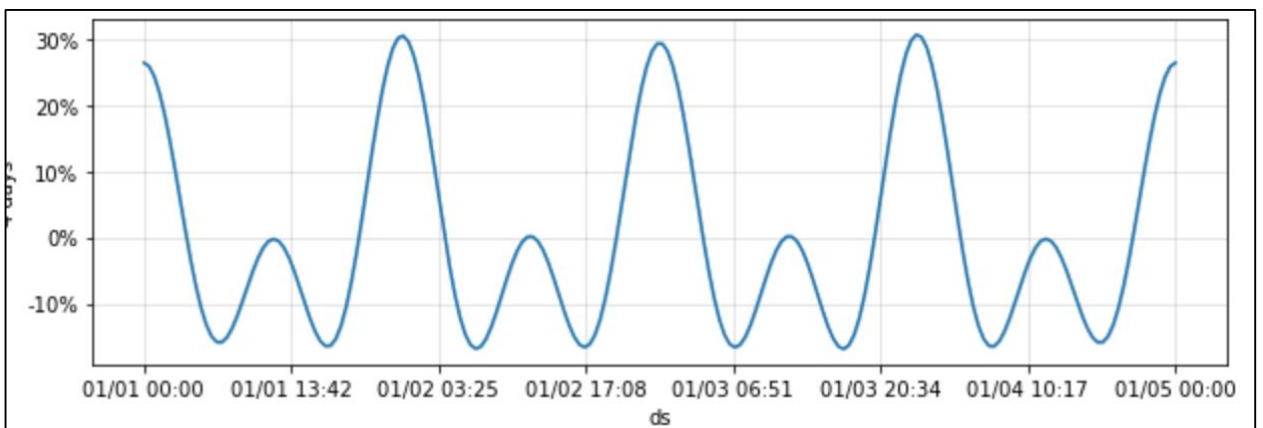
Складові моделі зображені на рисунку 3.10.



а)



б)



в)

Рисунок 3.10 – Складові моделі на для моделювання та прогнозування щоденної кількості нових підтверджених випадків хворих на коронавірус в Україні з 9 вересня 2021 р.: а) основний тренд, б) тижнева сезонність, в) 4-денна сезонність

За моделлю на основі Prophet до 12.12.2021 р. відносна похибка становить 16.38% що на 1.38% більше попереднього прогнозу, але він є песимістичним, але більш адекватний, прогнозовані значення не йдуть до нуля та схоже до реалістичної картини.

3.3 Довгостроковий прогноз

3.3.1 Довгостроковий прогноз за допомогою мультиплікативної моделі

Розглянемо проблему довгострокового прогнозування. Спробуємо спрогнозувати орієнтовну дату завершення так званої «дельта-хвилі» в Україні за даними по жовтень включно, коли хвиля ще тільки розганялась. Для нової хвилі не вистачає адекватних даних, оскільки дельта-хвиля відрізняється від попередніх. Спробуємо зробити прогноз на 2 місяці (61 день) за даними по 31.10.2021 р. за використанням попередньої моделі (рис. 3.11).

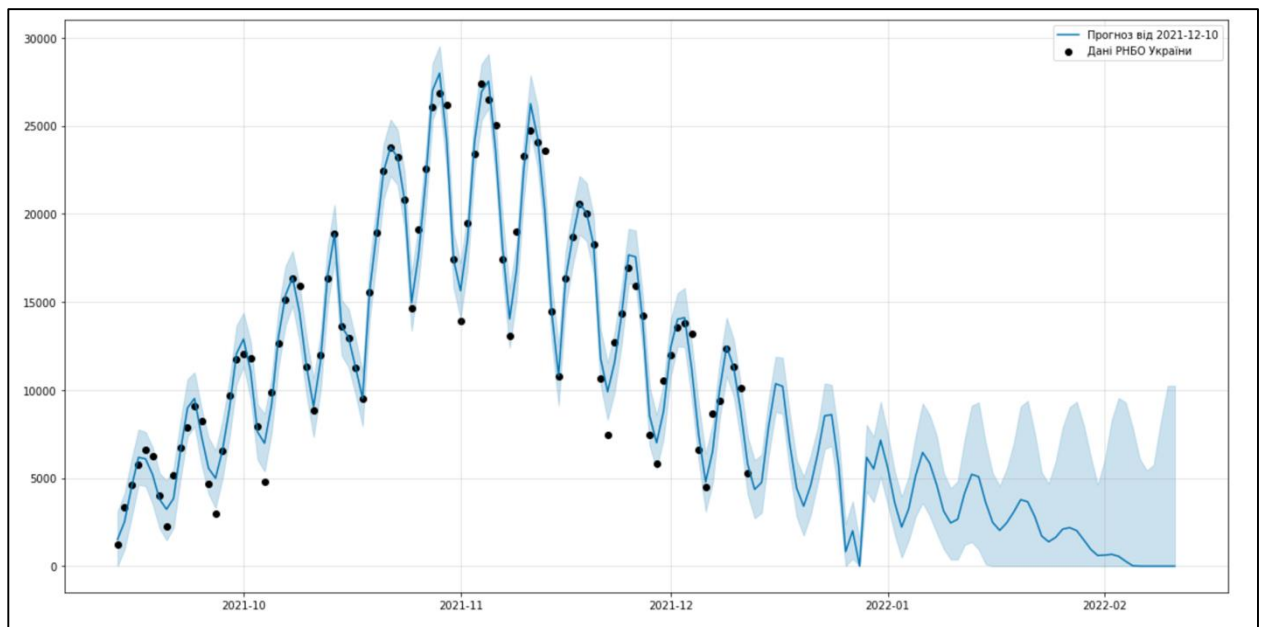
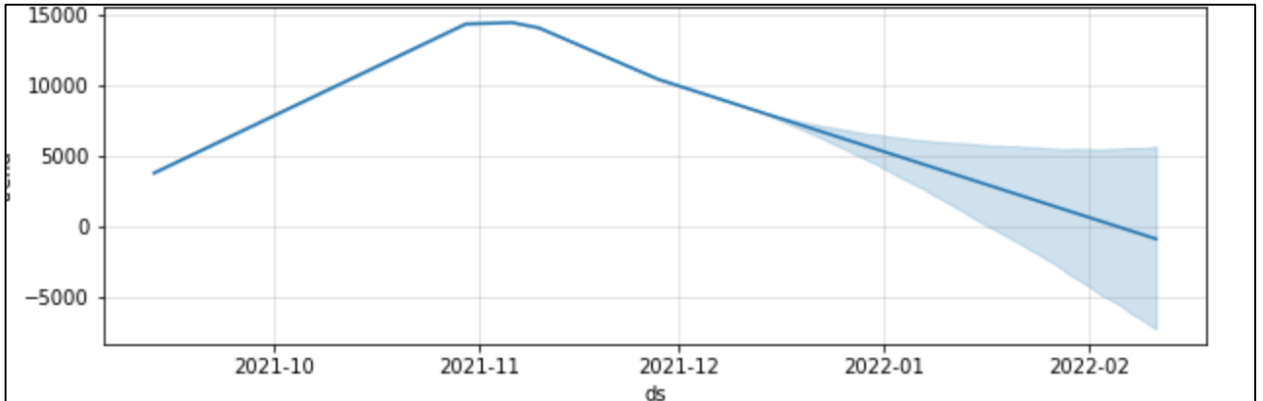
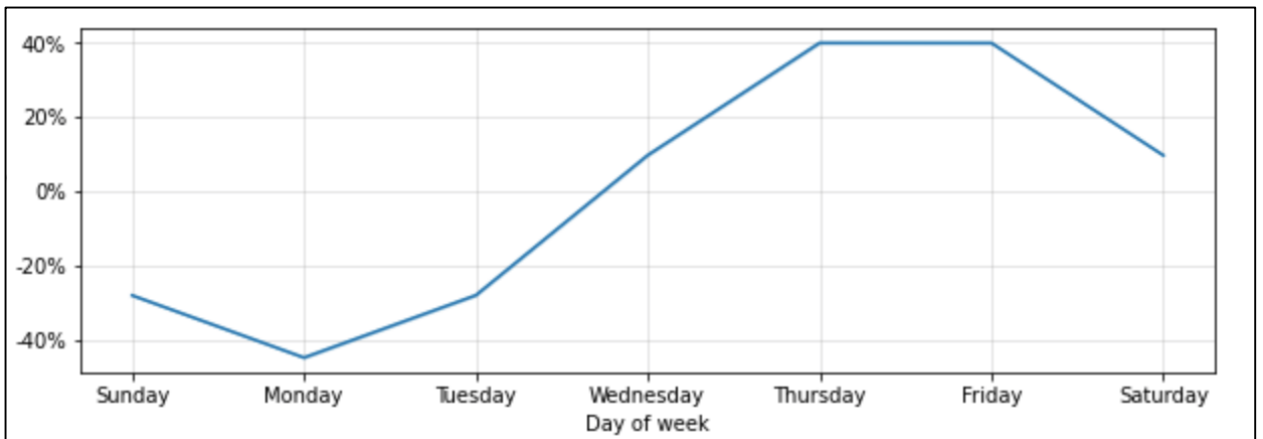


Рисунок 3.11 – Щоденна кількість нових хворих на коронавірус з 9 вересня 2021 р.: чорні крапки – дані спостережень до 11.02.2022 р., синя лінія – результат моделювання і прогнозування на 2 місяця

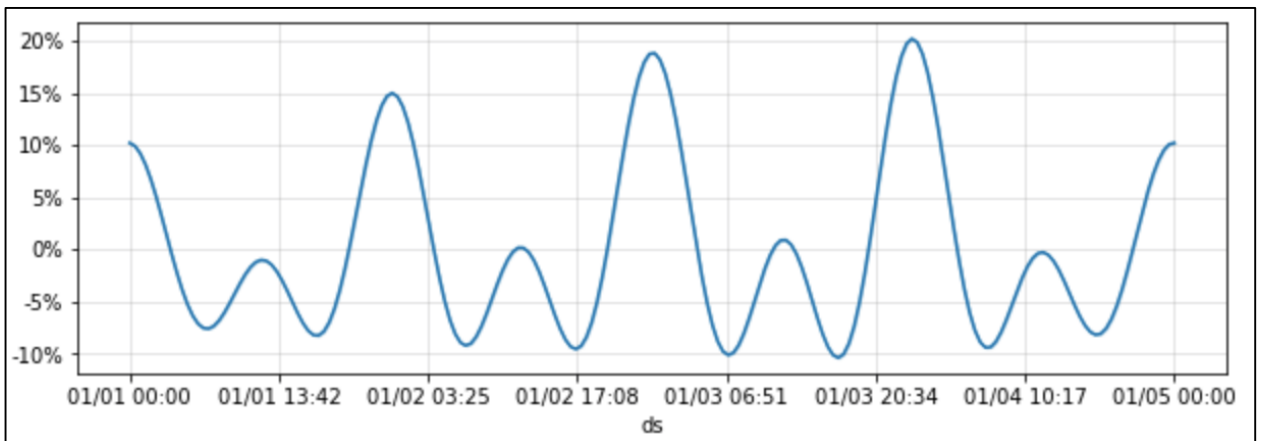
Складові моделі зображені на рисунку 3.12.



а)



б)



в)

Рисунок 3.12 – Складові моделі для моделювання та прогнозування щоденної кількості нових підтверджених випадків хворих на коронавірус в Україні з 9 вересня 2021 р.: а) основний тренд, б) тижнева сезонність, в) 4-денна сезонність

За моделлю на основі Prophet до 11.02.2022 р. відносна похибка становить 152,32%, що є зовсім не адекватно. З цього випливає висновок про те, що мультиплікативну модель недоречно використовувати для довгострокового прогнозування, вона є неточною з великою похибкою та нульовими значеннями. Дослідження показало, що це – через те, що модель є мультиплікативною і містить багато складових, які призводять до її неадекватності у довгостроковій перспективі.

3.3.2 Довгостроковий прогноз за використання адитивної моделі

Завантажимо дані в адитивну модель через API з Системи моніторингу поширення епідемії коронавірусу Апарату РНБО України.

Виставимо налаштування періоду з 180 до 270 та зобразимо нижче на рисунках. Налаштування з періодом 180 зображено на рисунку 3.13.

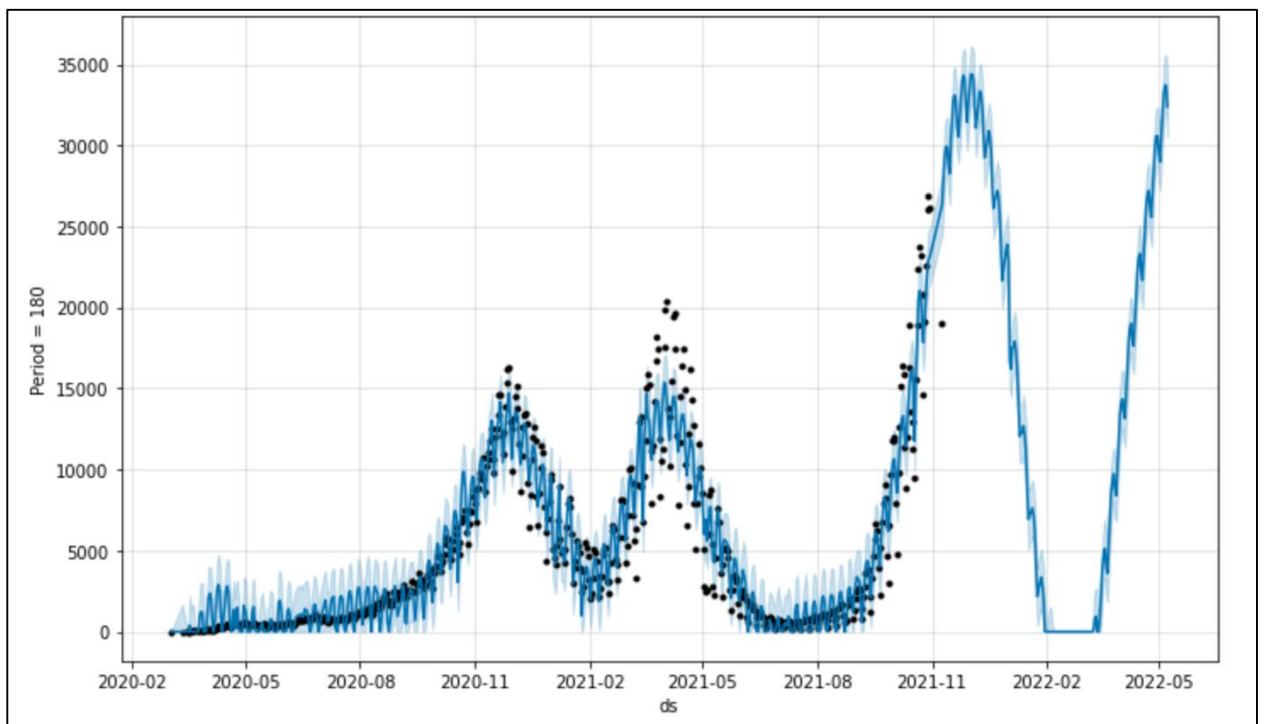


Рисунок 3.13 – Щоденна кількість нових хворих на коронавірус з 6 липня 2020р. за період 180

Результат дельта хвилі зображений на рисунку 3.14.

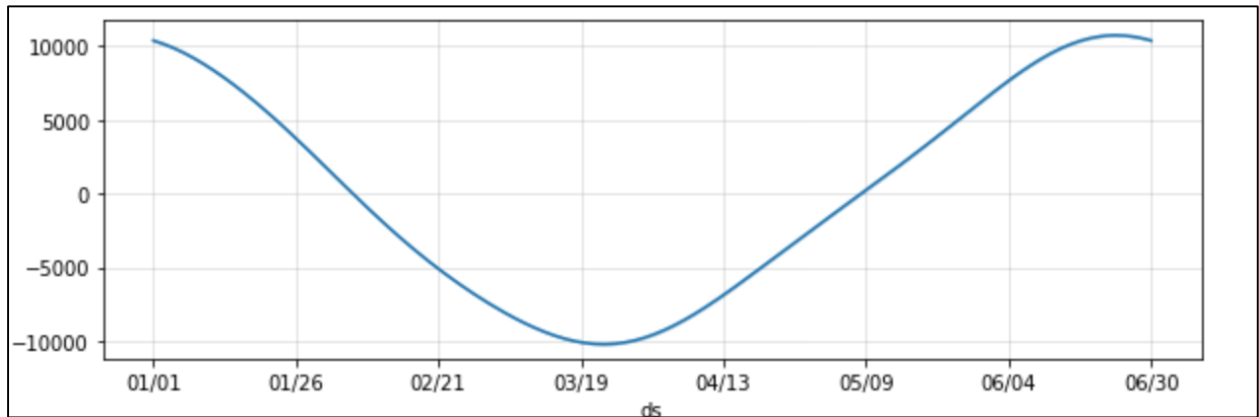


Рисунок 3.14 – Дельта-хвиля за період 180

Прогнозування за параметром 180 днів зображений на рисунку 3.15.

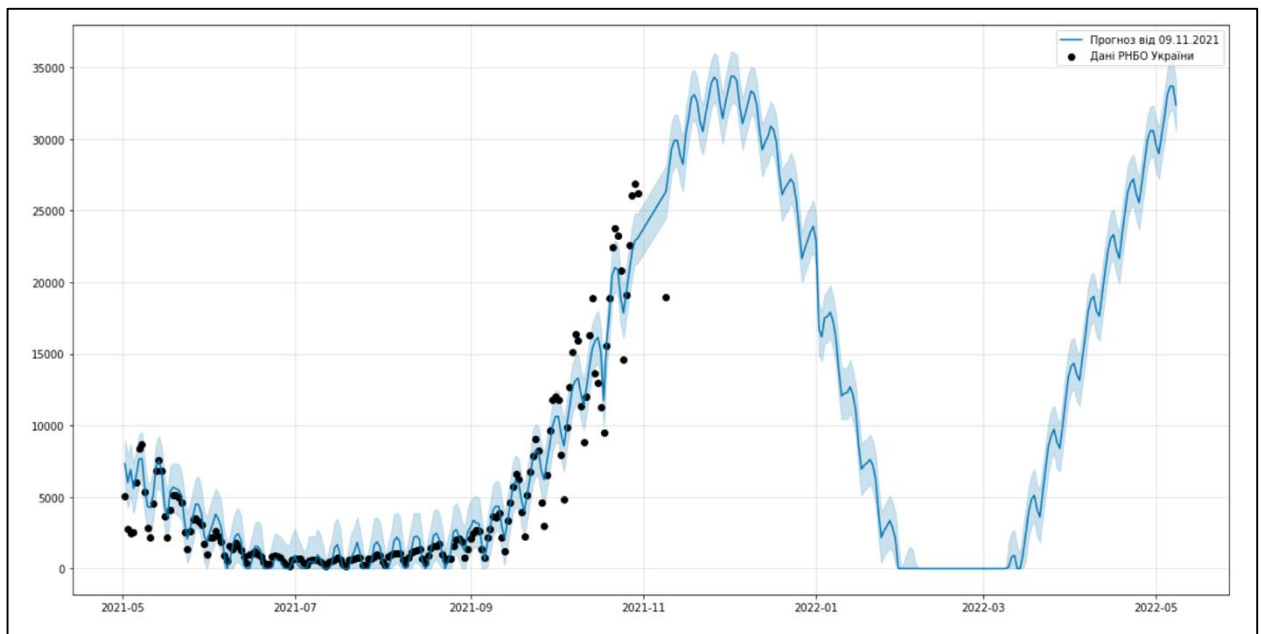


Рисунок 3.15 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021 р. за період 180

Налаштування з періодом 190 зображено на рисунку 3.16.

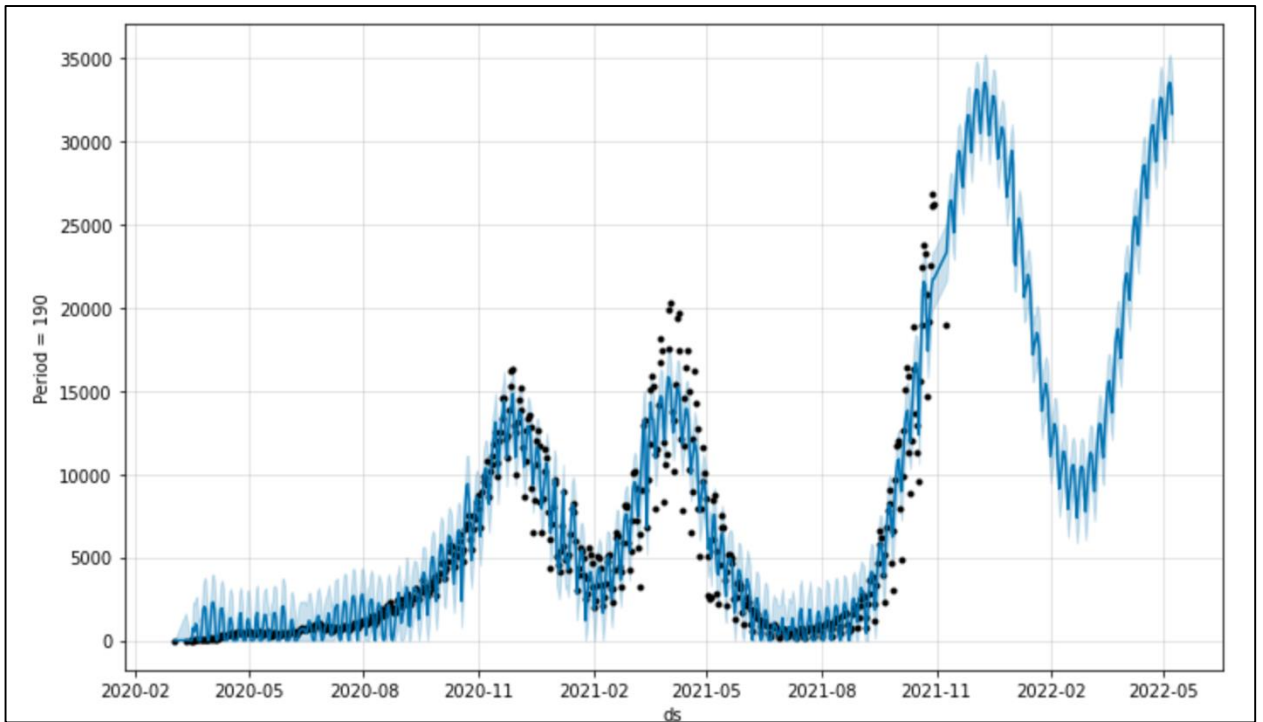


Рисунок 3.16 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020 р. за період 190

Результат дельта хвилі зображений на рисунку 3.17.

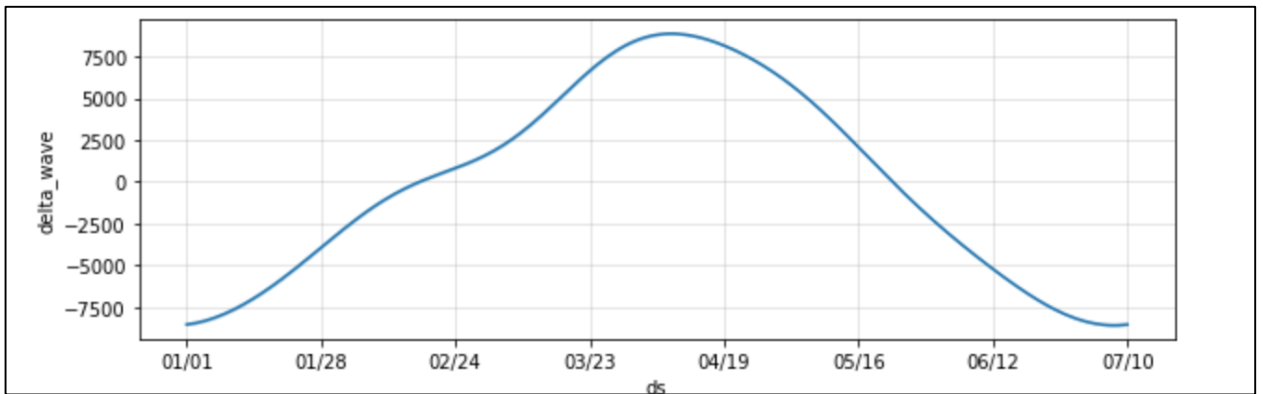


Рисунок 3.17 – Дельта-хвиля за період 190

Прогнозування за параметром 190 днів зображений на рисунку 3.18.

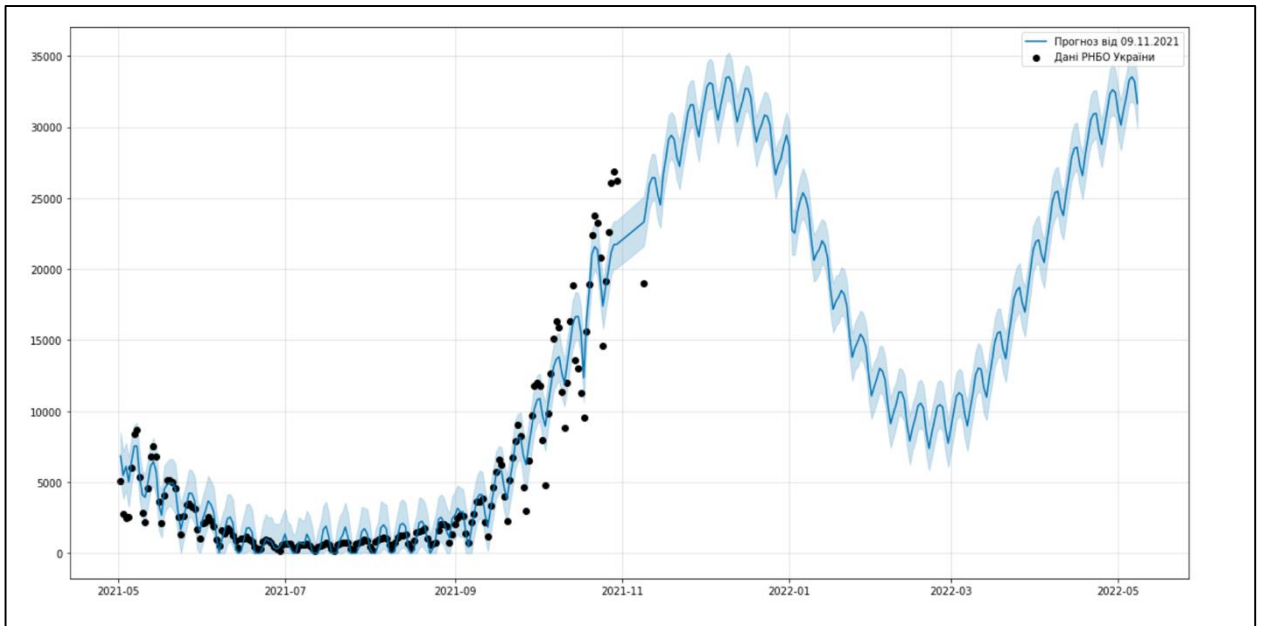


Рисунок 3.18 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 190

Налаштування з періодом 200 зображено на рисунку 3.19.

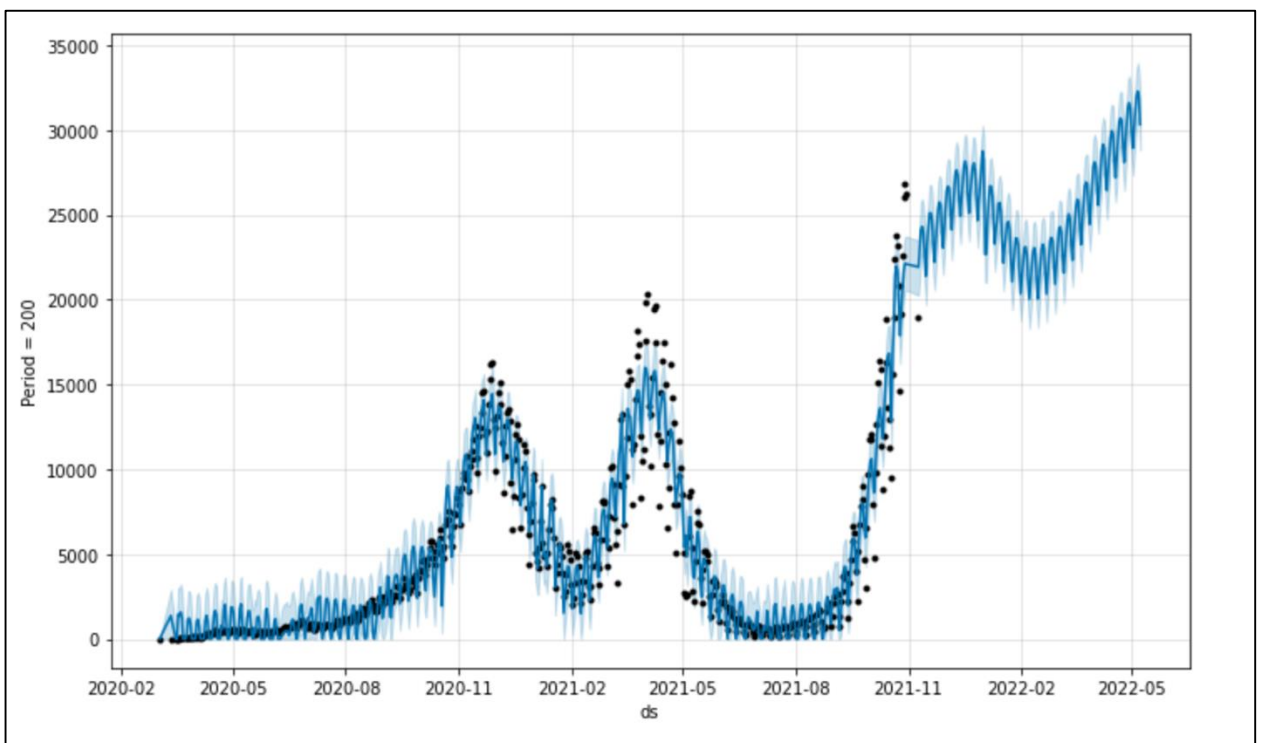


Рисунок 3.19 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020р. за період 200

Результат дельта хвилі зображений на рисунку 3.20.

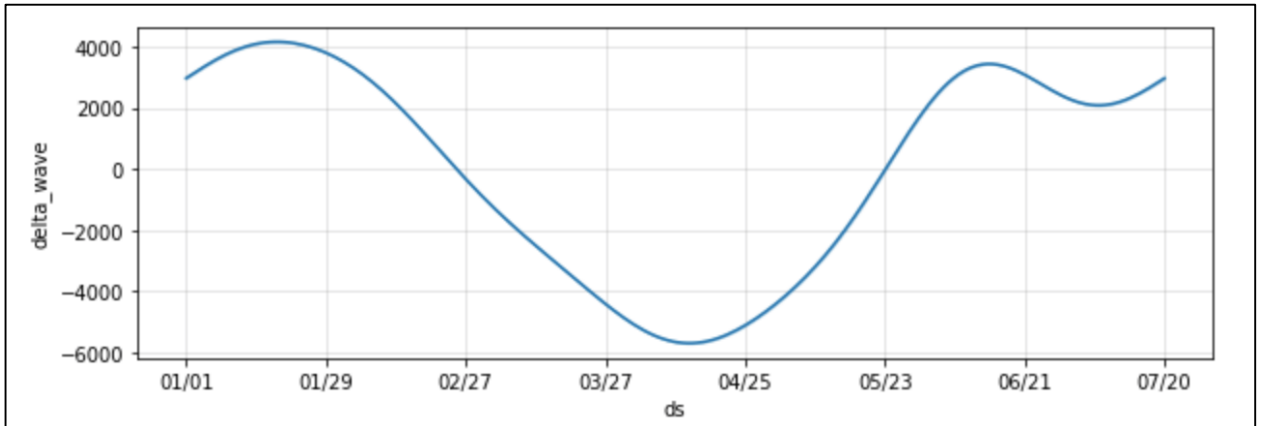


Рисунок 3.20 – Дельта-хвиля за період 200

Прогнозування за параметром 200 днів зображений на рисунку 3.21.

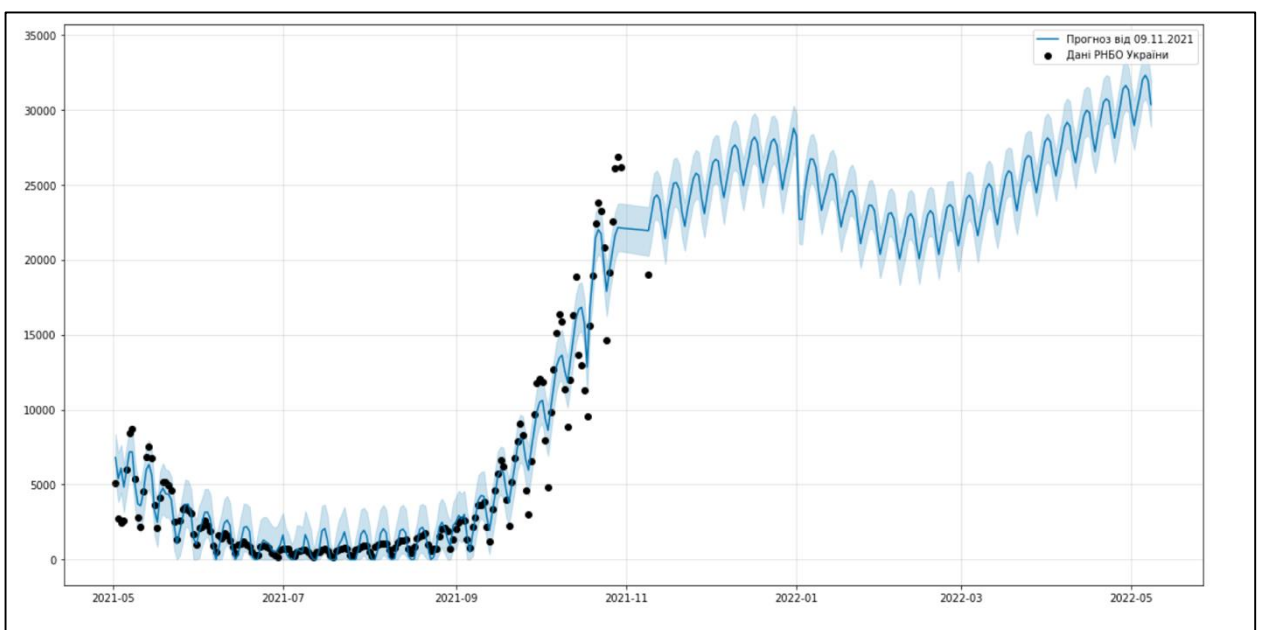


Рисунок 3.21 – Щоденна кількість нових хворих на коронавірус з 1 травня 2021р. за період 200

Налаштування з періодом 210 зображено на рисунку 3.22.

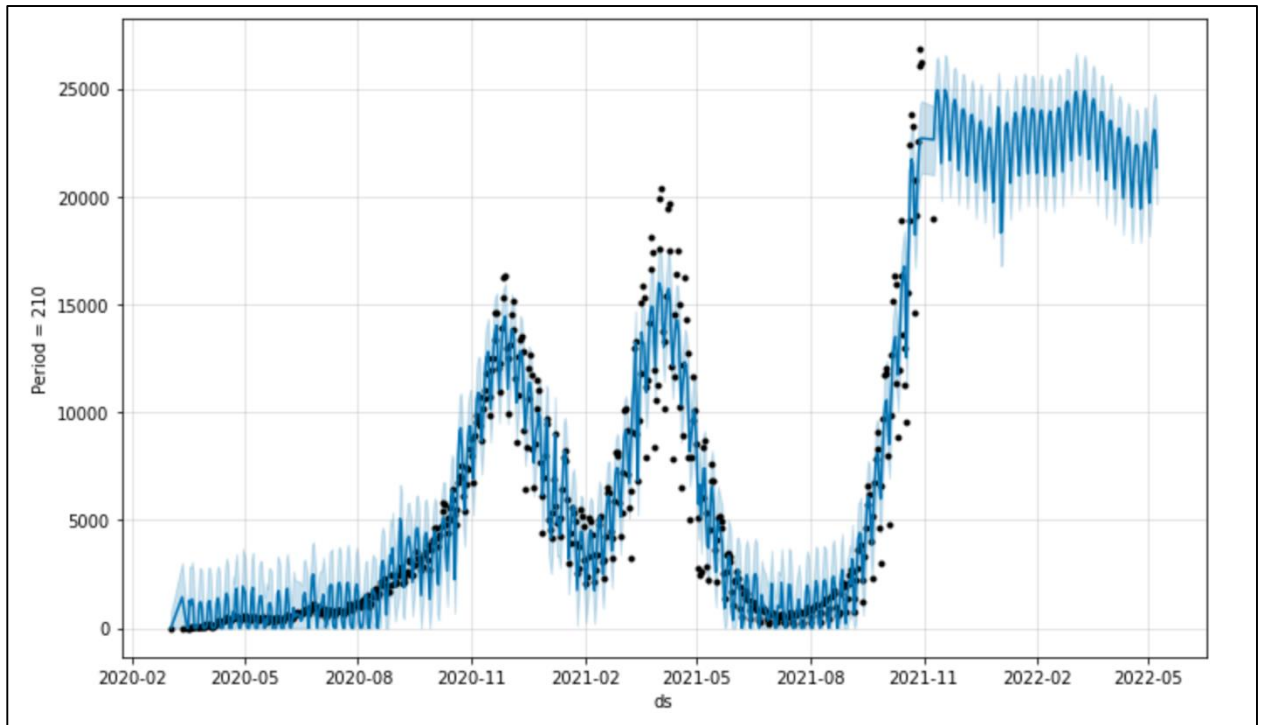


Рисунок 3.22 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020р. за період 210

Результат дельта хвилі зображений на рисунку 3.23.

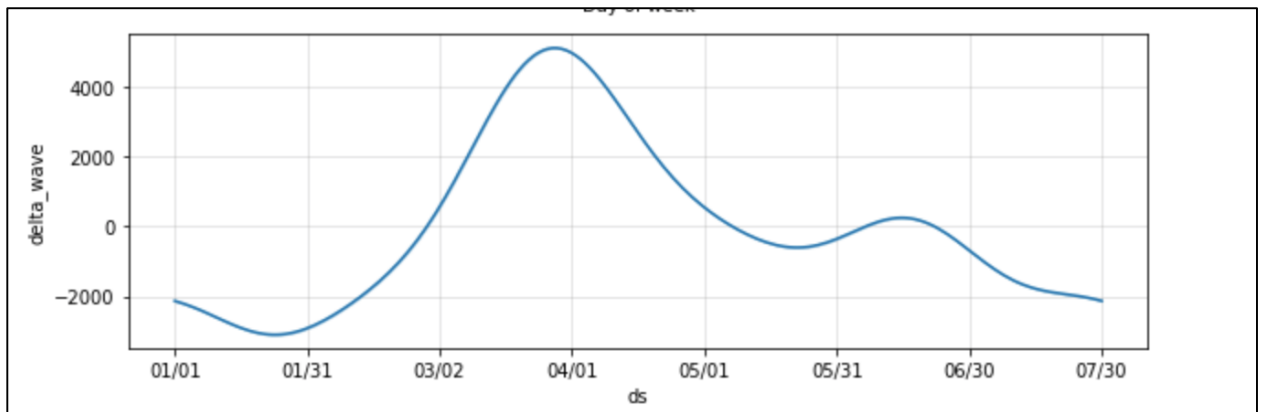


Рисунок 3.23 – Дельта-хвиля за період 210

Прогнозування за параметром 210 днів зображений на рисунку 3.24.

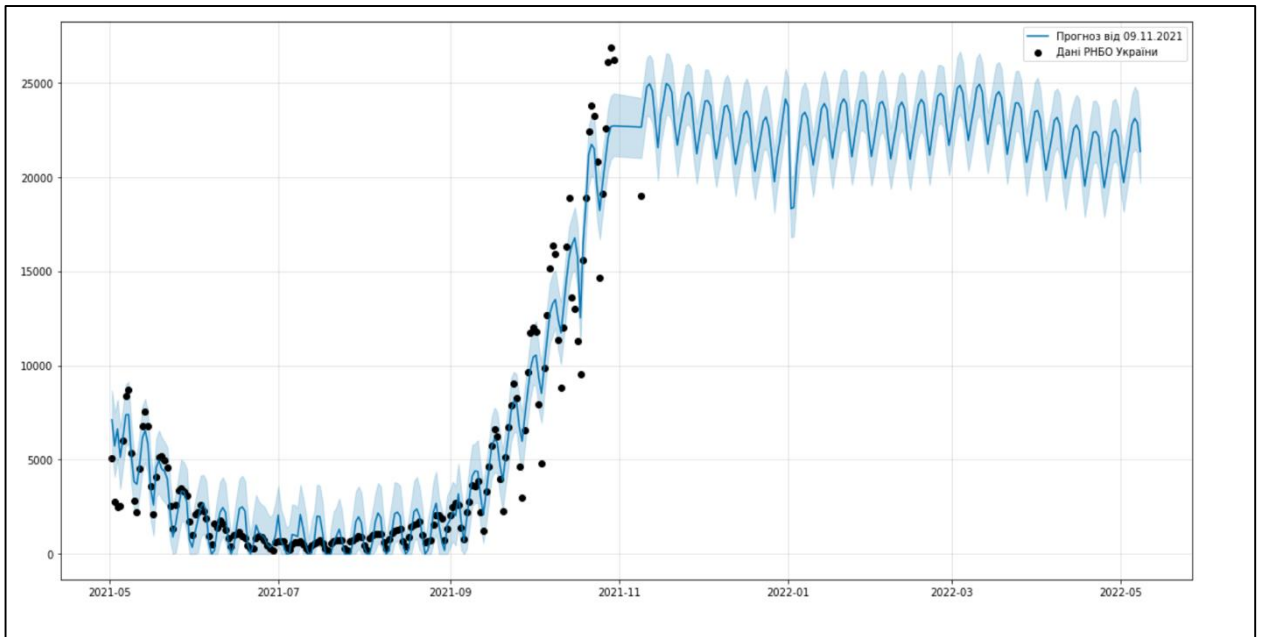


Рисунок 3.24 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 210

Налаштування з періодом 220 зображено на рисунку 3.25.

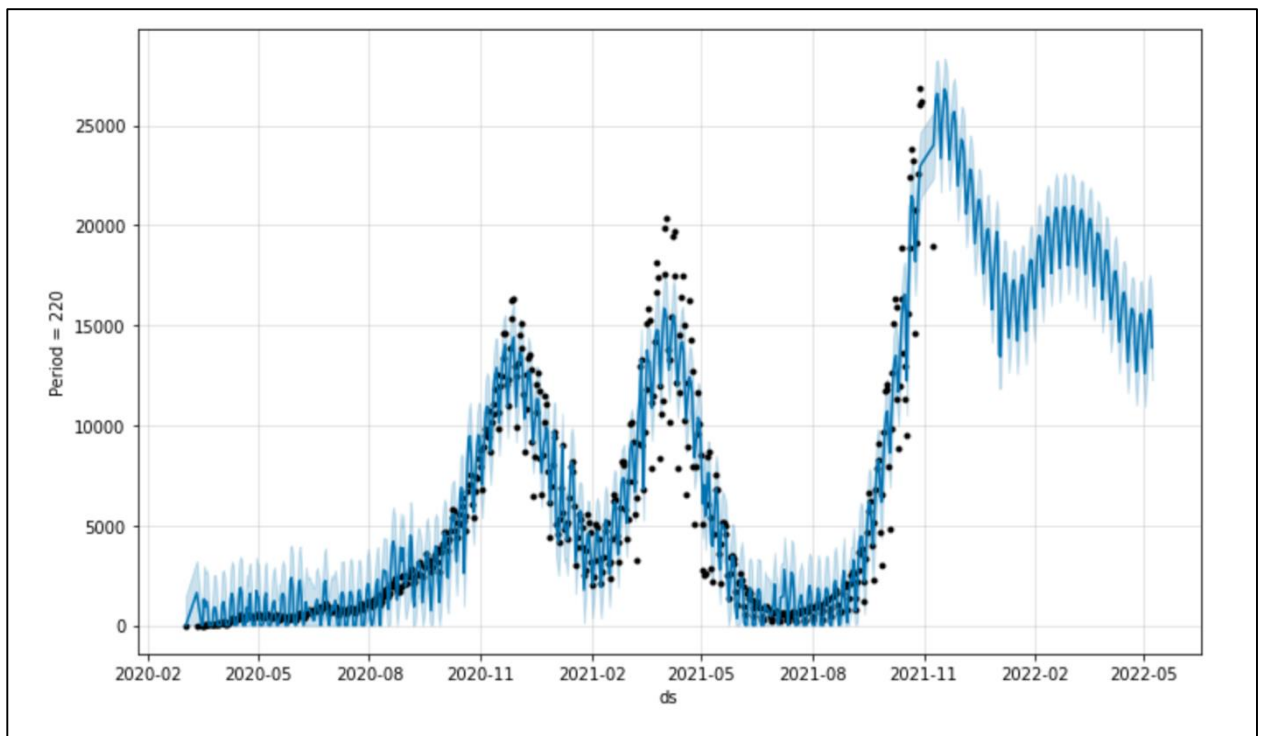


Рисунок 3.25 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020р. за період 220

Результат дельта хвилі зображений на рисунку 3.26.

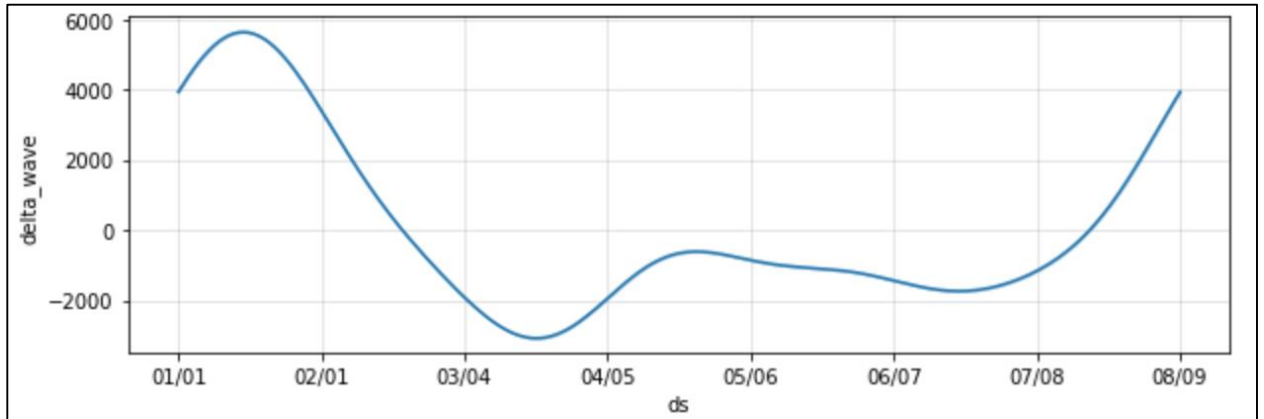


Рисунок 3.26 – Дельта-хвиля за період 220

Прогнозування за параметром 220 днів зображений на рисунку 3.27.

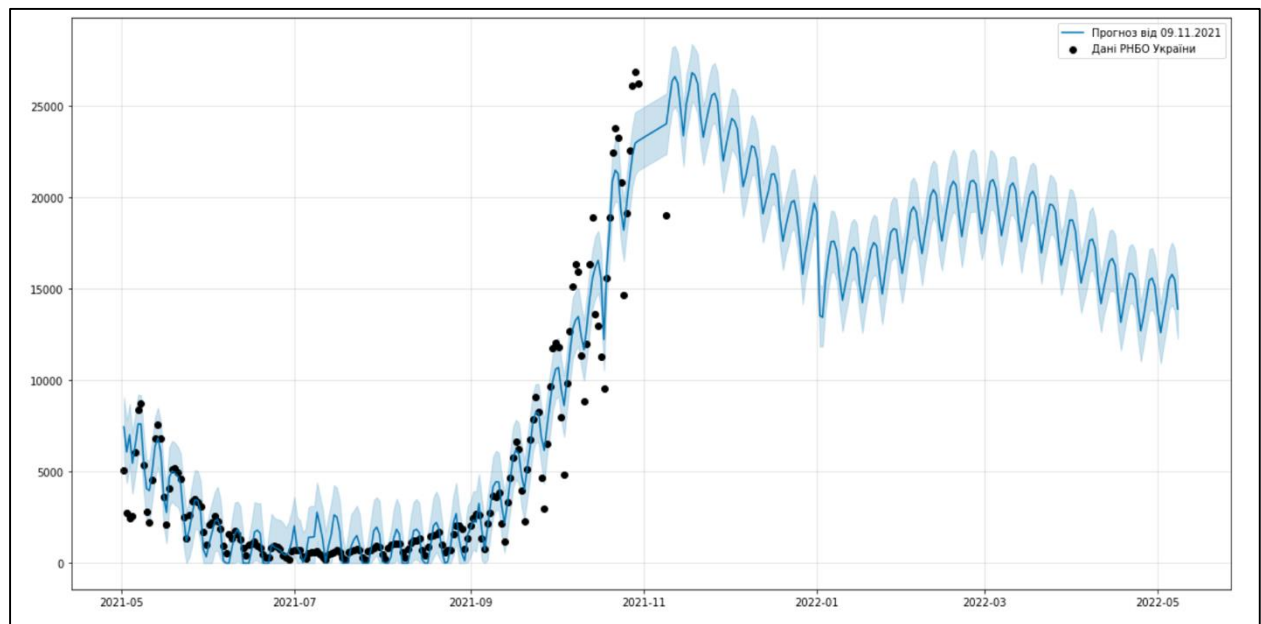


Рисунок 3.27 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 220

Налаштування з періодом 230 зображено на рисунку 3.28.

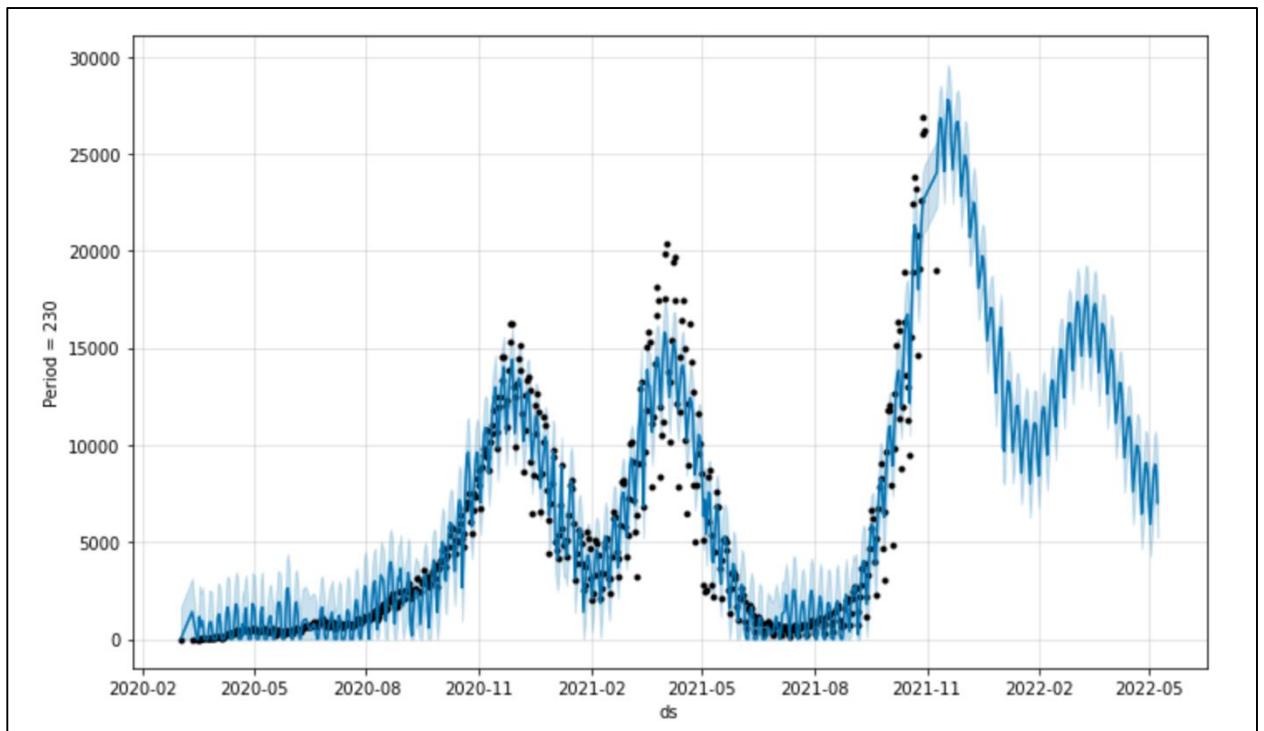


Рисунок 3.28 – Щоденна кількість нових хворих на коронавірус з 6 липня 2020р. за період 230

Результат дельта хвилі зображений на рисунку 3.29.

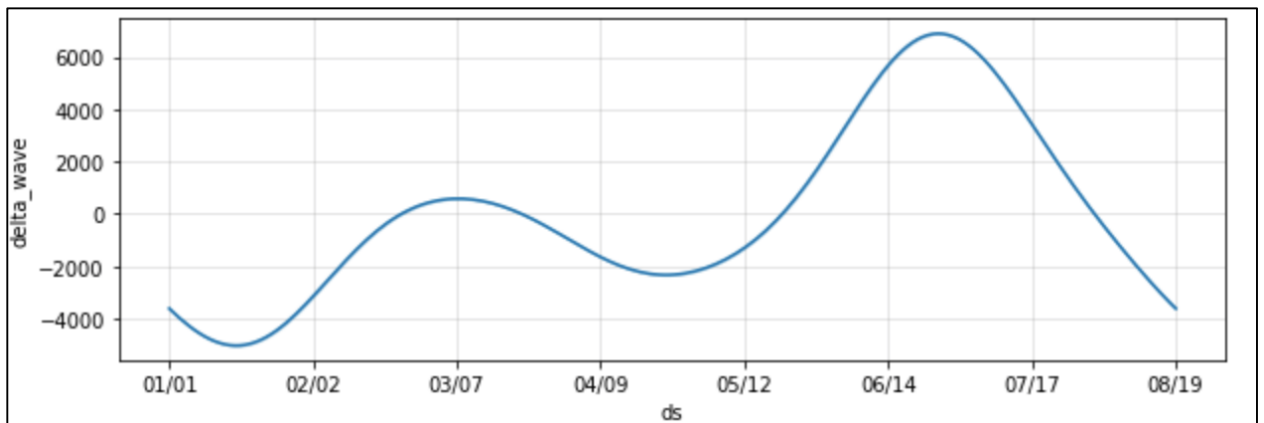


Рисунок 3.29 – Дельта-хвиля за період 230

Прогнозування за параметром 230 днів зображений на рисунку 3.30.

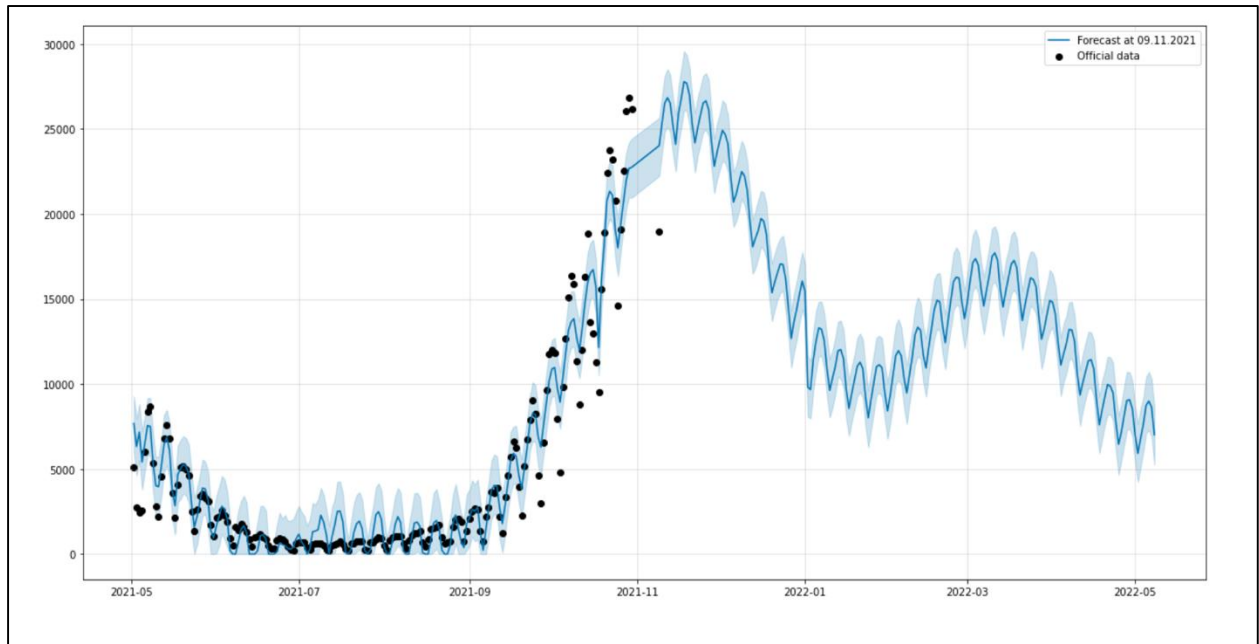


Рисунок 3.30 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 230

Налаштування з періодом 240 зображено на рисунку 3.31.

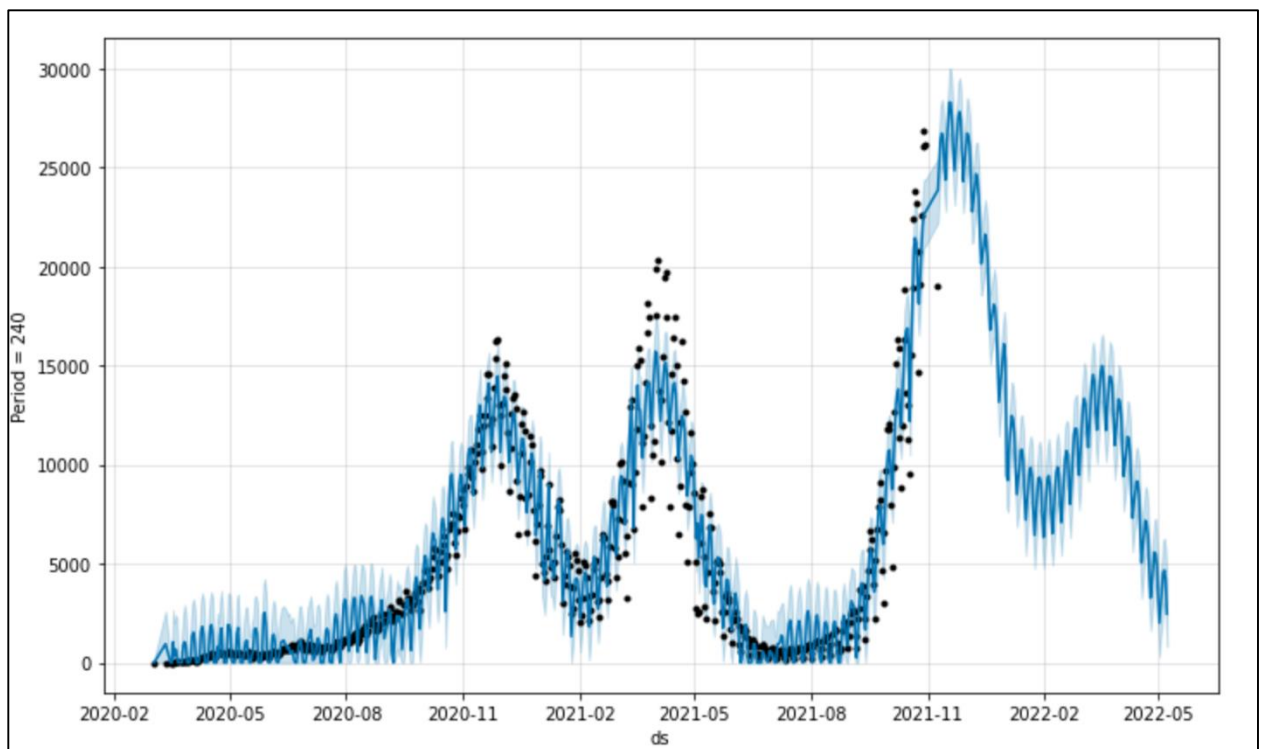


Рисунок 3.31 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020р. за період 240

Результат дельта хвилі зображений на рисунку 3.32.

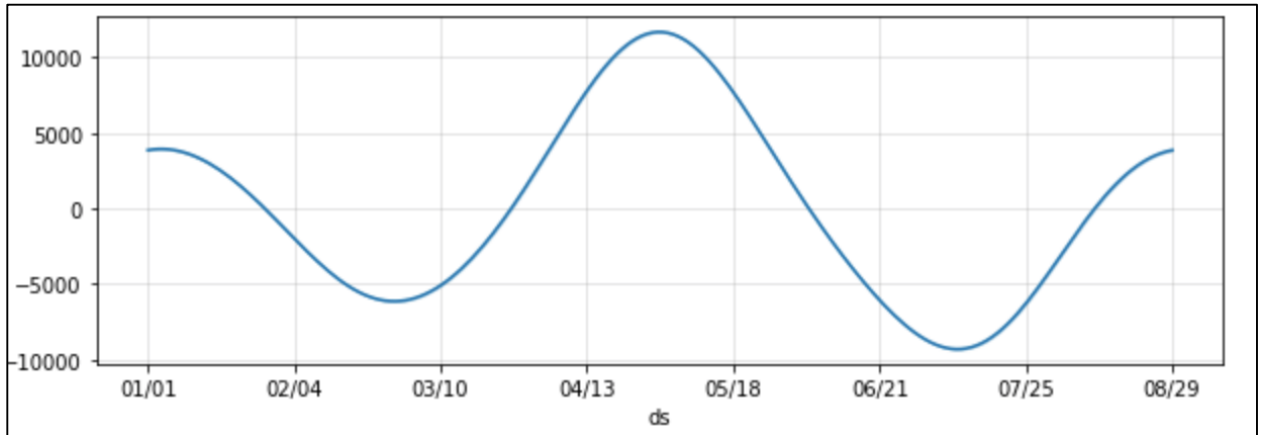


Рисунок 3.32 – Дельта-хвиля за період 240

Прогнозування за параметром 240 днів зображений на рисунку 3.33.

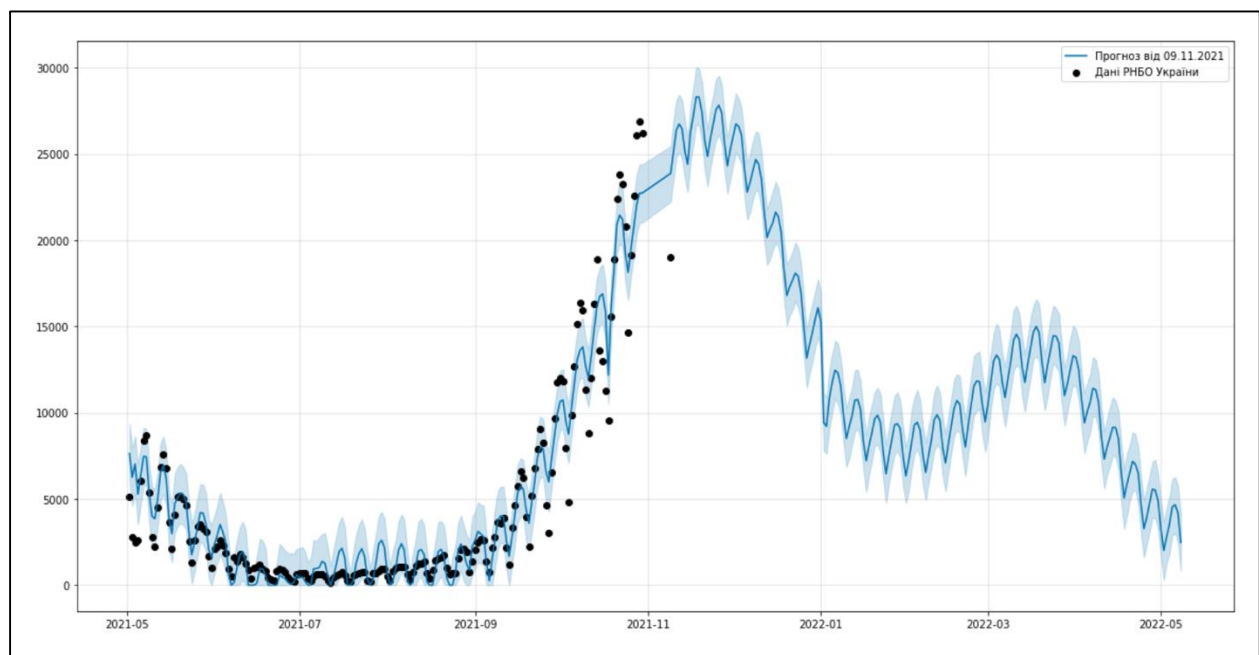


Рисунок 3.33 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 240

Налаштування з періодом 250 зображено на рисунку 3.34.

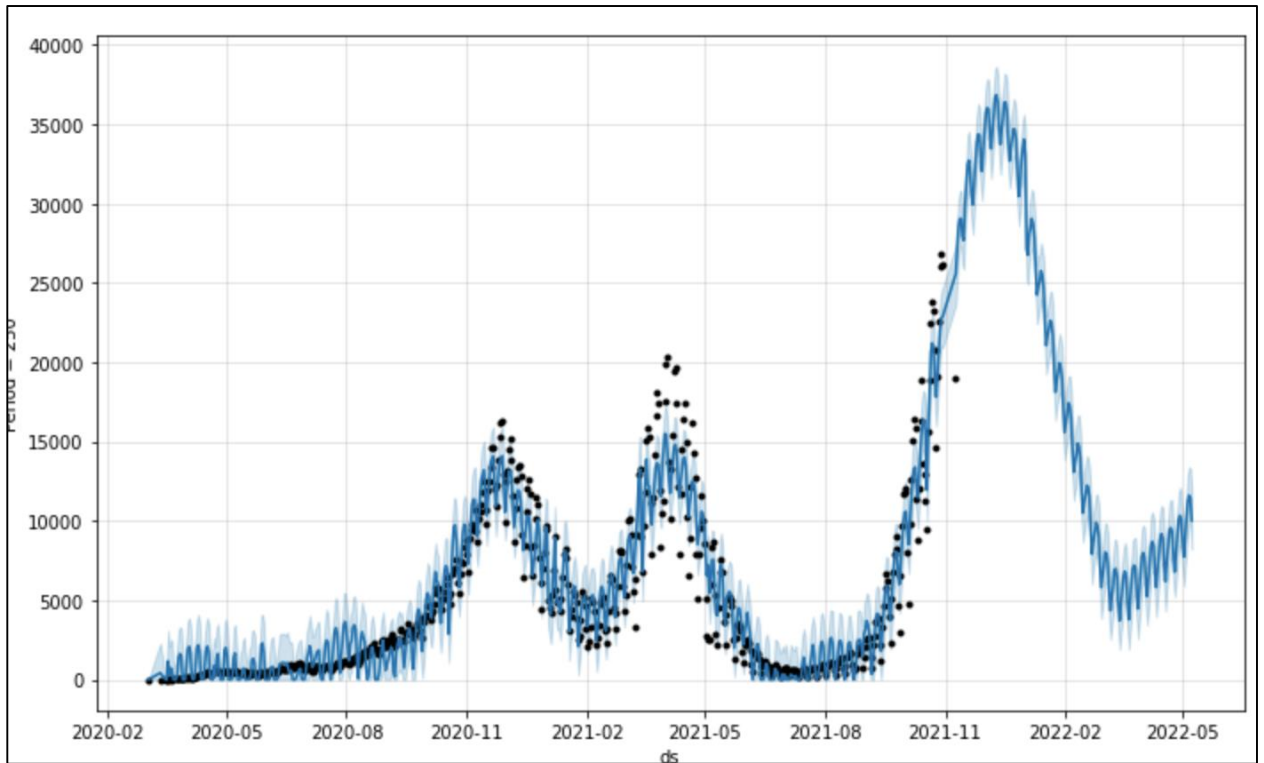


Рисунок 3.34 – Щоденна кількість нових хворих на коронавірус з 6 липня 2020р. за період 250

Результат дельта хвилі зображений на рисунку 3.35.

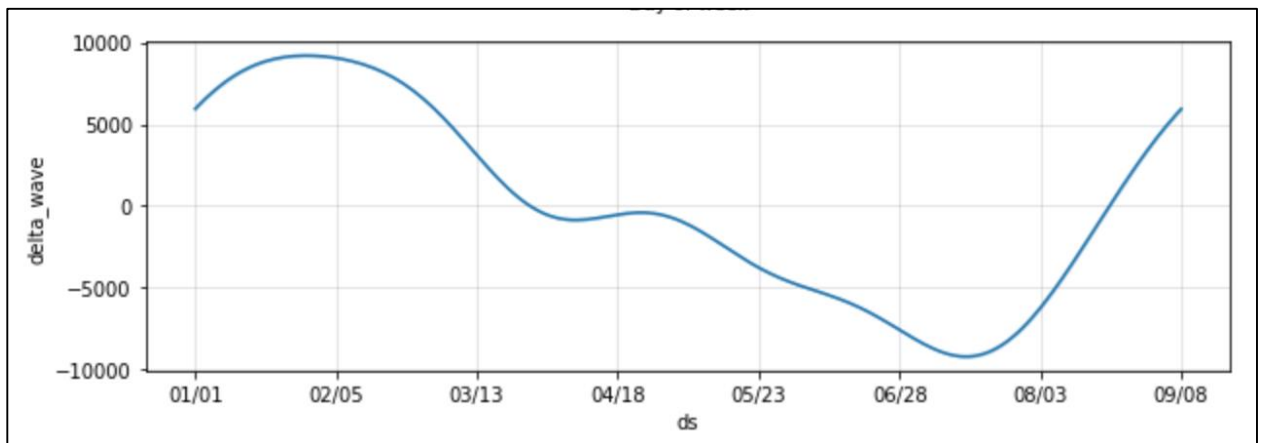


Рисунок 3.35 – Дельта-хвиля за період 250

Прогнозування за параметром 250 днів зображений на рисунку 3.36.

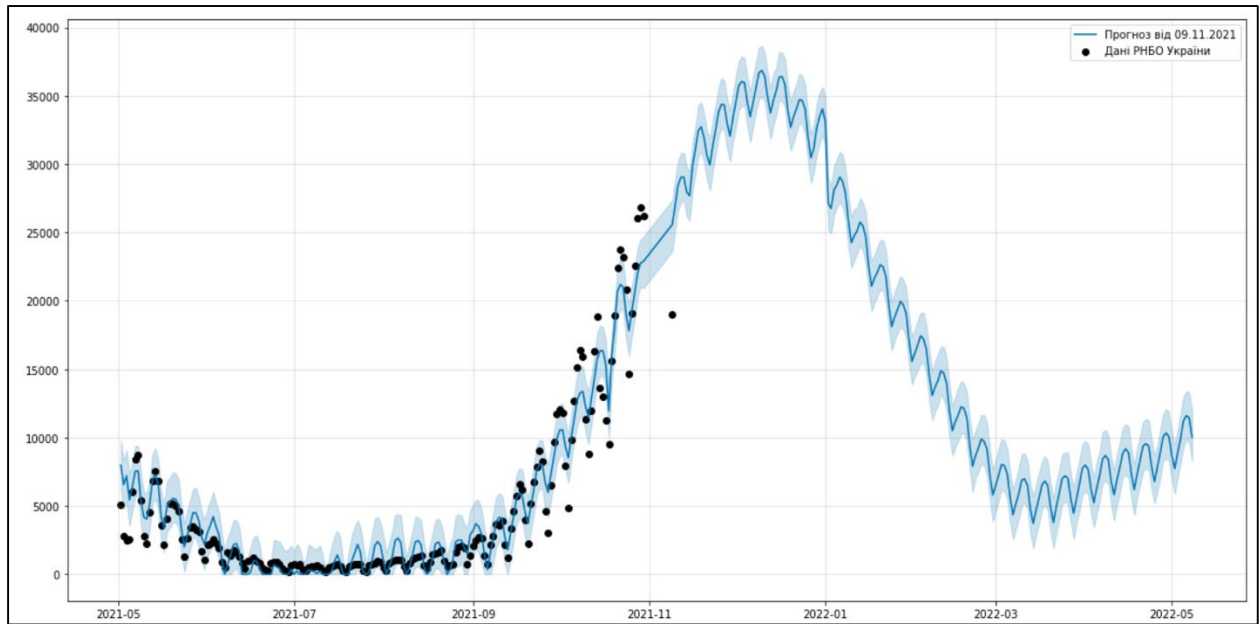


Рисунок 3.37 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 250

Налаштування з періодом 260 зображено на рисунку 3.37.

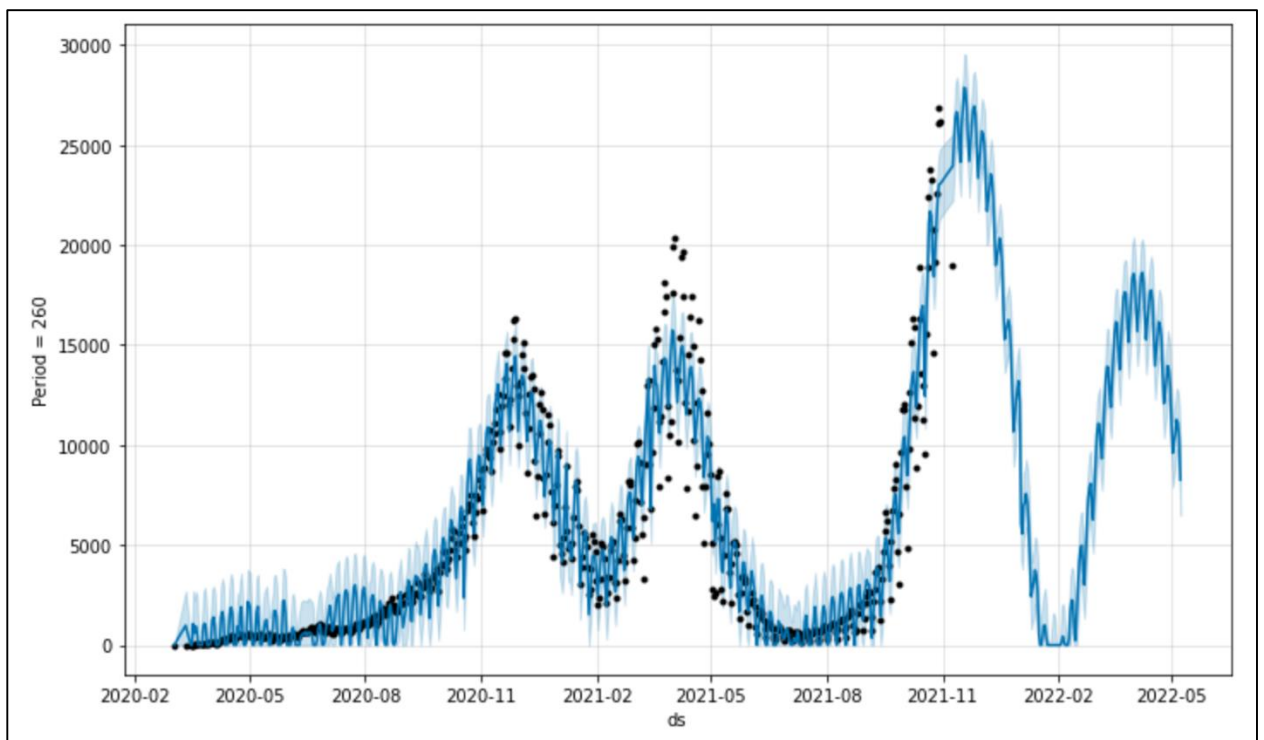


Рисунок 3.37 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020р. за період 260

Результат дельта хвилі зображений на рисунку 3.38.

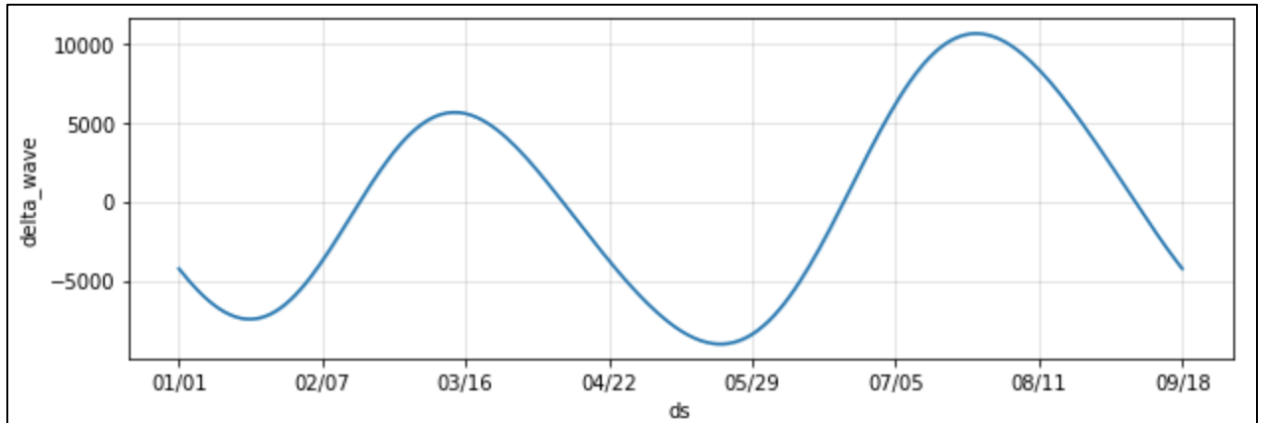


Рисунок 3.38 – Дельта-хвиля за період 260

Прогнозування за параметром 260 днів зображений на рисунку 3.39.

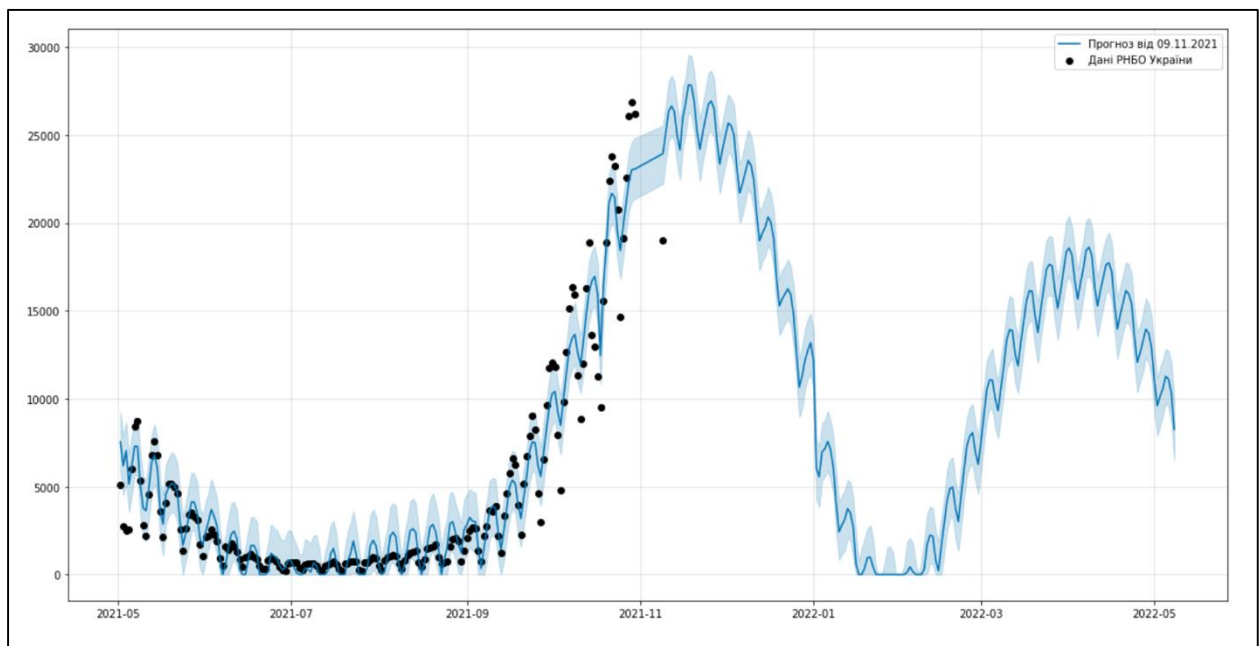


Рисунок 3.39 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 260

Налаштування з періодом 270 зображено на рисунку 3.40.

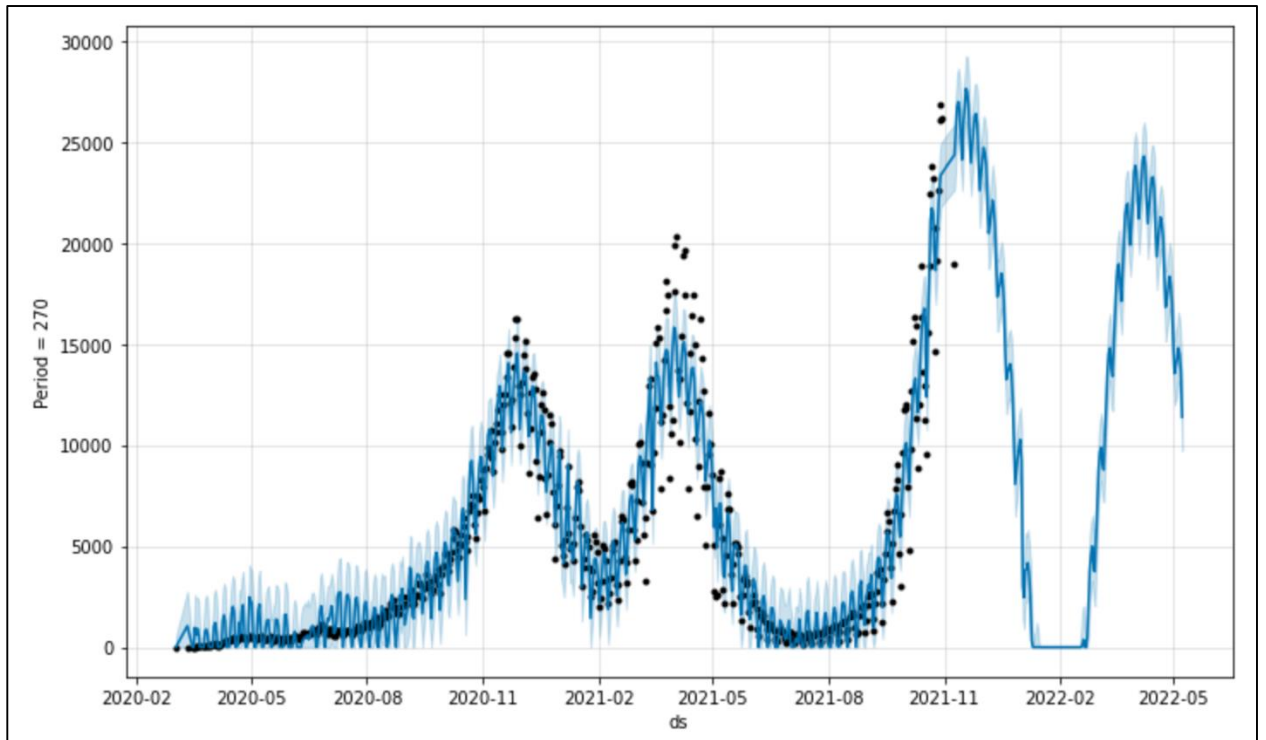


Рисунок 3.40 – Щоденна кількості нових хворих на коронавірус з 6 липня 2020р. за період 270

Результат дельта хвилі зображений на рисунку 3.41.

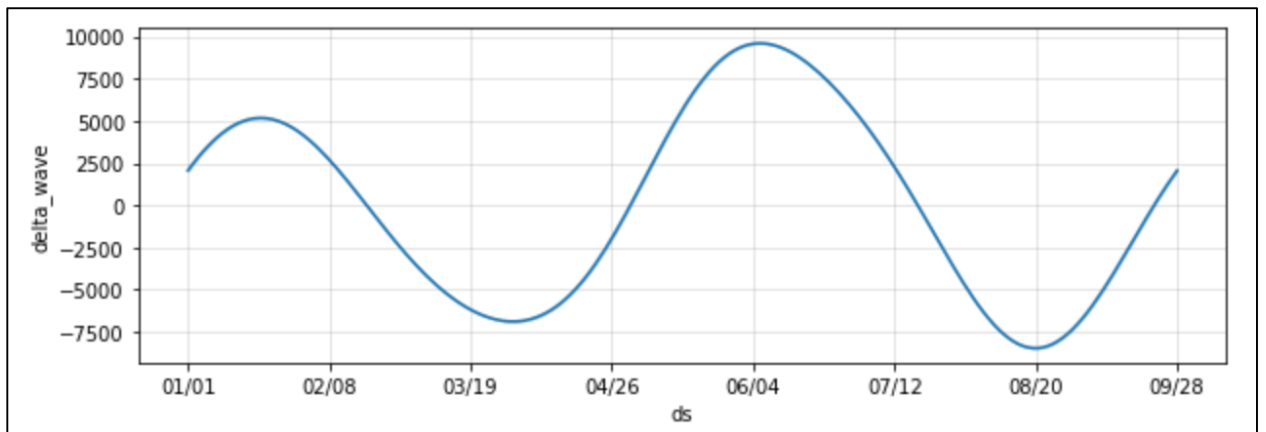


Рисунок 3.41 – Дельта-хвиля за період 270

Прогнозування за параметром 270 днів зображений на рисунку 3.42.

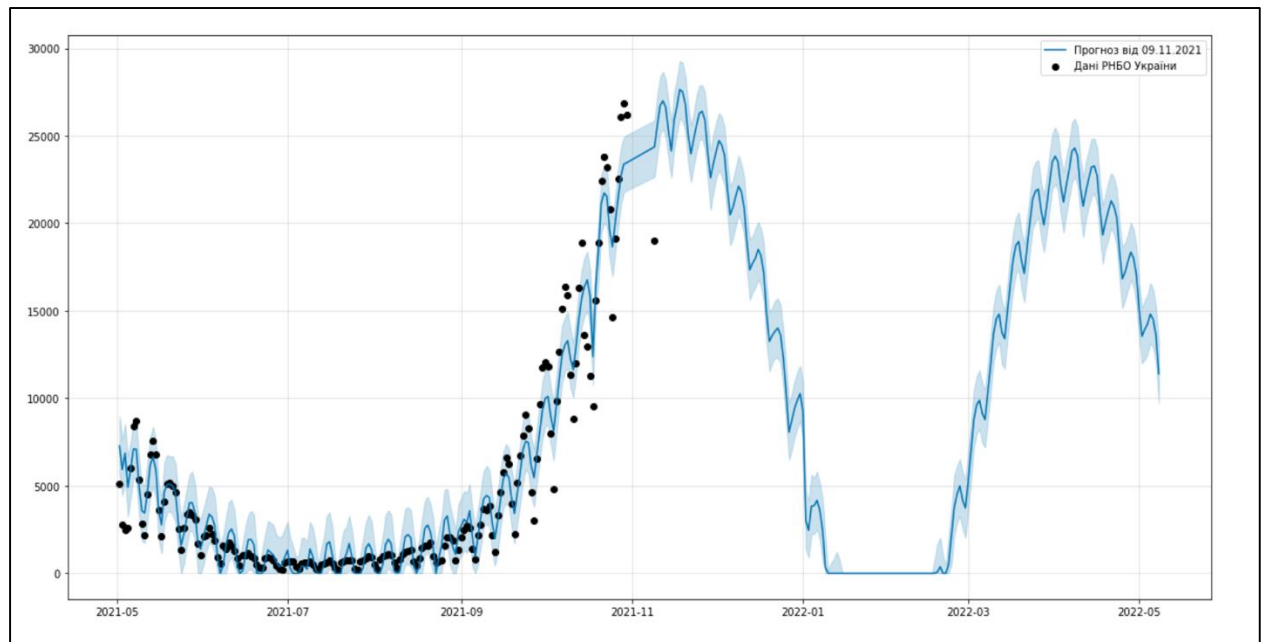


Рисунок 3.42 – Щоденна кількості нових хворих на коронавірус з 1 травня 2021р. за період 270

Експерименти з можливою тривалістю довгострокової дельта-хвилі від 170 до 270 днів, де чверть періоду – це час наростання хвилі у днях, дали отакі графіки, з яких графік з періодом 240 здається дещо адекватним по амплітуді (піку), а графік з періодом 230 – по тривалості, враховуючи сучасні знання про те, який пік і тривалість хвиля має місце в наш час (вона ще не завершилась, але близька до того). Тому і цей підхід не дає значного успіху.

3.4 Короткостроковий прогноз на 7 днів для 70 країн світу

За допомогою розробки мого наукового керівника Мокіна В.Б., який є членом робочої групи при НАН України з моделювання поширення коронавірусу в Україні, яка готує звіти для РНБО кожній 1-2 тижні (у 2020 р. готували щотижня, у 2021 - кожні 2 тижні) було здійснено короткостроковий прогноз на 7 днів для 70 країн світу [20]. За основу взято його короткострокову модель і отримано результат у вигляді csv-файлу – прогноз на 7 днів по 70 країнах світу. Далі ці дані візуалізовано на рисунках нижче (рис. 3.43-3.48).

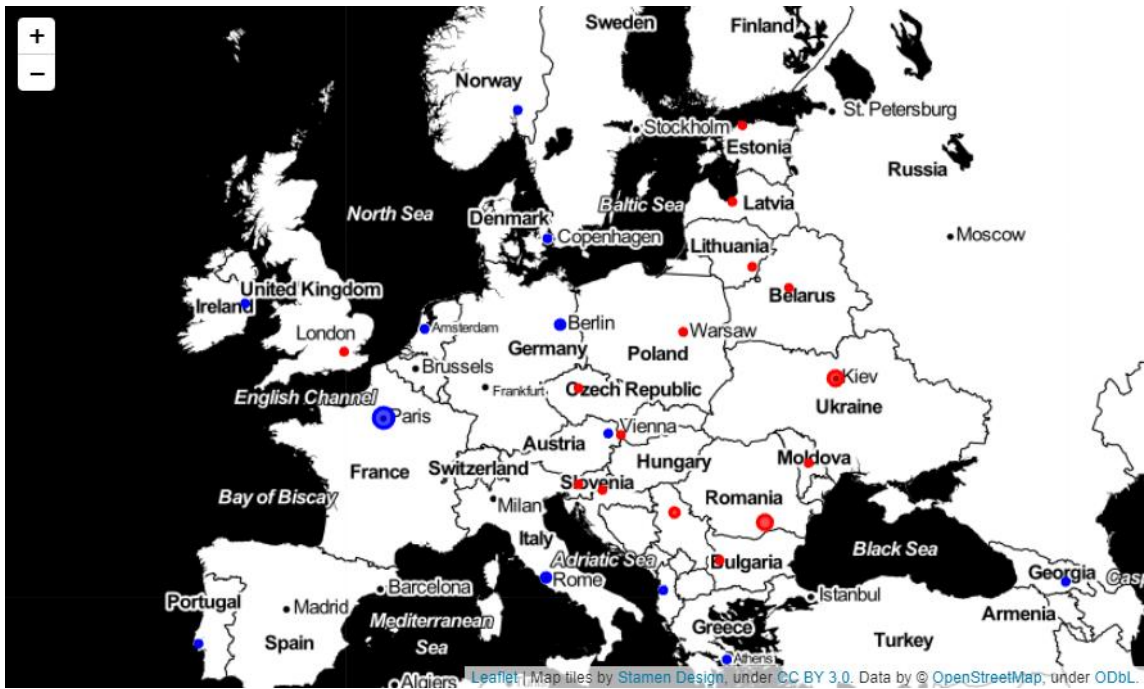


Рисунок 3.43 – Карта трендів тижневих прогнозів на 4.10.2021 р. за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи

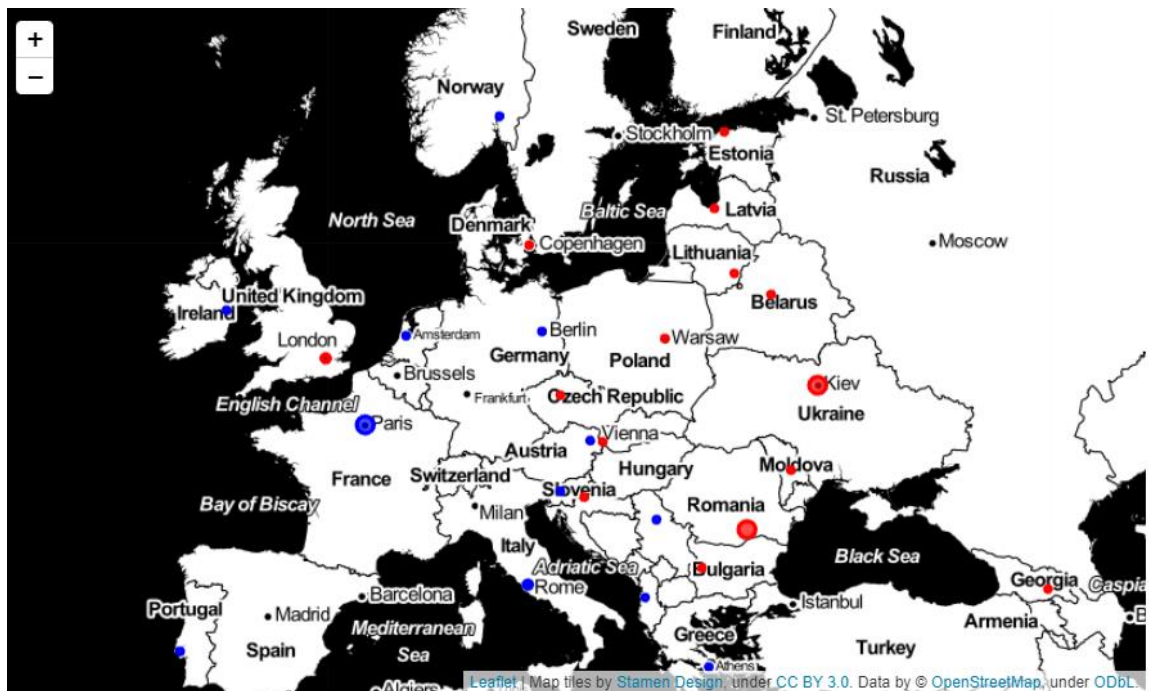


Рисунок 3.44 – Карта трендів тижневих прогнозів на 11.10.2021 р. за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи

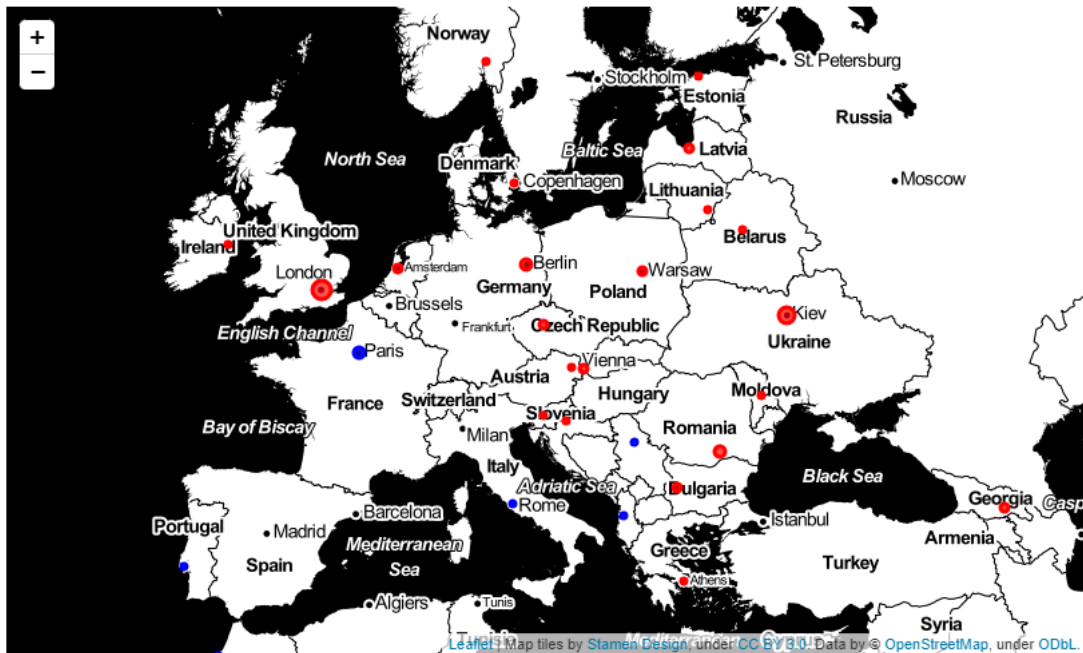


Рисунок 3.45 – Карта трендів тижневих прогнозів на 25.10.2021 р. за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи

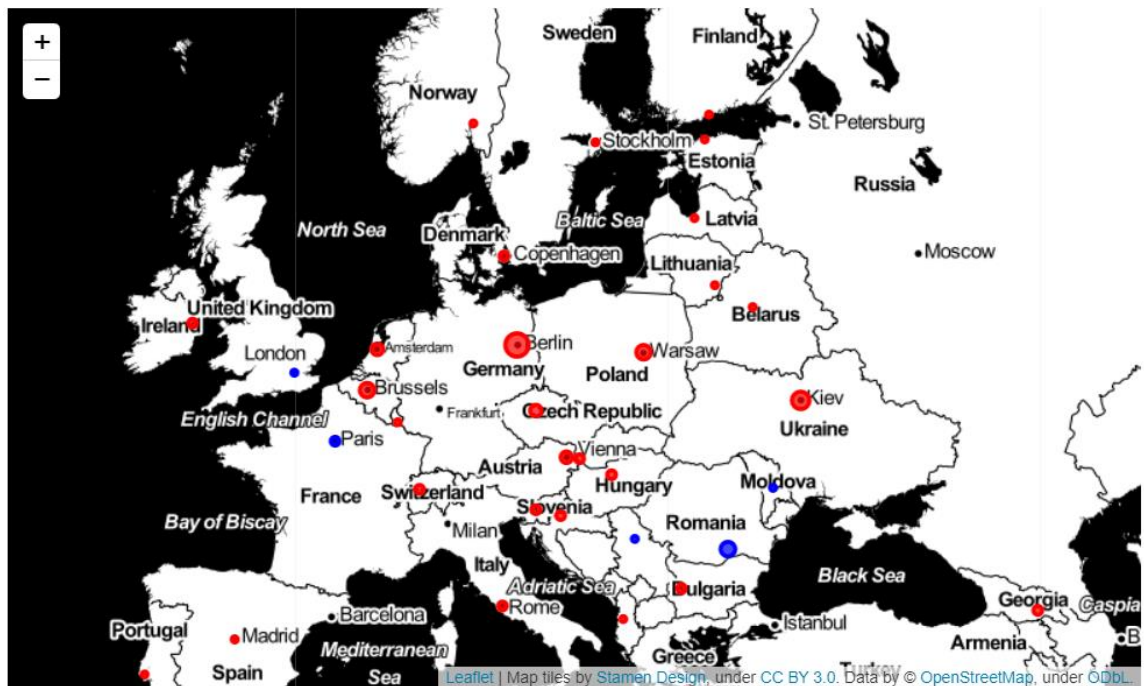


Рисунок 3.46 – Карта трендів тижневих прогнозів на 07.11.2021 р. за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи

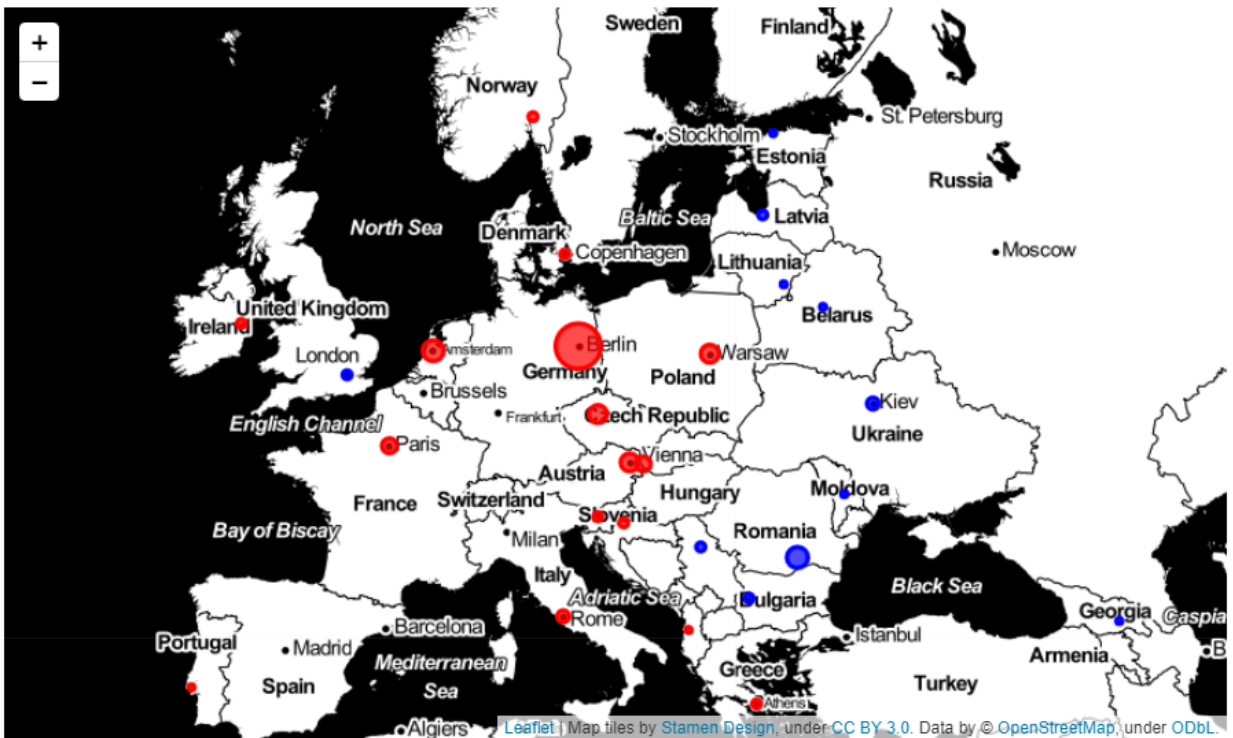


Рисунок 3.47 – Карта трендів тижневих прогнозів на 21.11.2021 р. за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи

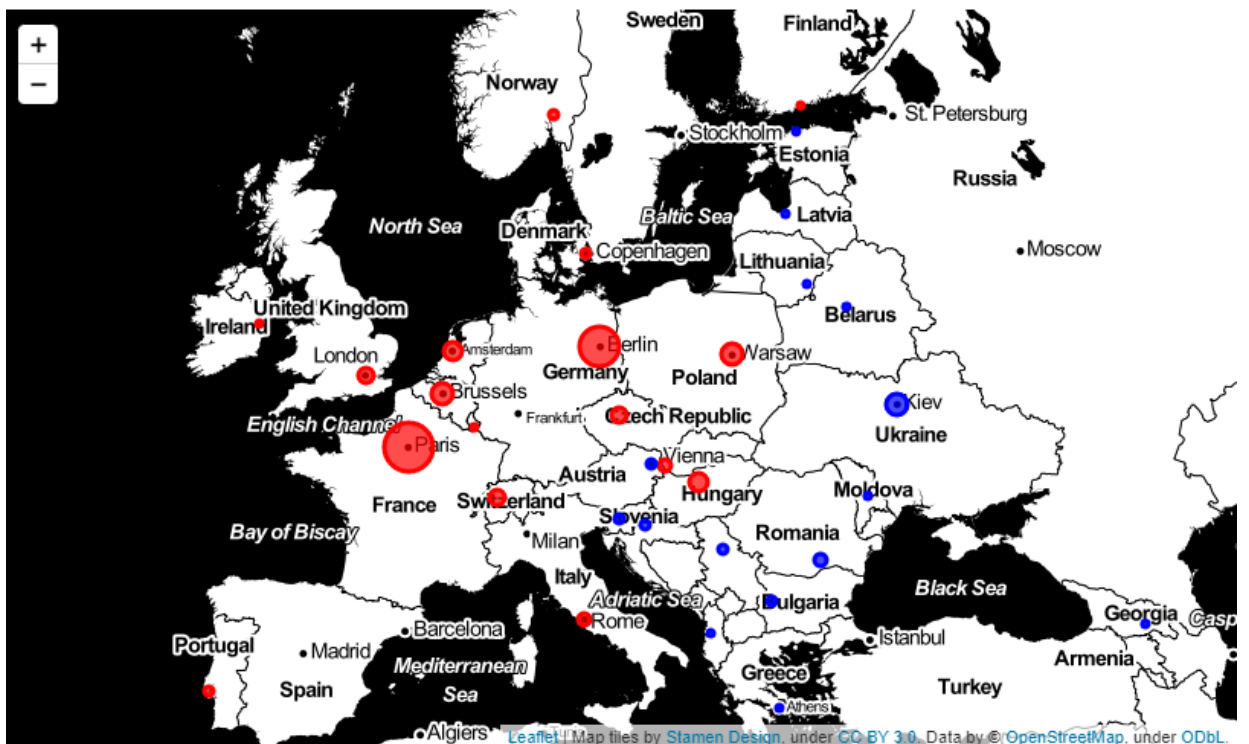


Рисунок 3.48 – Карта трендів тижневих прогнозів на 07.12.2021 р. за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи

Як видно, нова хвиля зароджується в центральній Європі і поширюється на сусідні країни. На даний момент ситуація виглядає як народження нової хвилі, яка незабаром докотиться до нас. Ці картограми показують, що зовнішні фактори (коронавірус у країнах-сусідах) має значний вплив, тому неможливо буде здійснити довгострокове прогнозування, без адекватних даних по такому впливу, але такі дані (кількість іноземців, які прибули в Україну і з яких країн) відсутні у відкритому доступі. Отже, довгострокове моделювання з використанням зазначених підходів та моделей не буде ефективним і варто зосередитись на короткостроковому прогнозуванні.

3.5 Висновки

У даному розділі було проведено аналіз та прогнозування кількості нових підтверджених хворих в Україні у 2021 році. Проведено підготовку даних для подальшого використання в моделях. Також описані та оцінені результати довгострокового та короткострокового прогнозування поширення коронавірусу COVID-19 за допомогою моделі Prophet. Прогнозовано на 7 днів кількість нових хворих по 70 країнам світу та у процесі прогнозування зроблено висновок про те, що адекватно прогнозувати нову хвилю по іншим країнам неможливо з використанням лише відкритих даних. Також довгостроковий варіант з усіх запропонованих не дає чіткої картини про майбутнє і прогнози є не адекватними. Найкраще себе показала мультиплікативна модель на 14 днів з невеликою похибкою, але з песимістичним сценарієм. Тому саме краще використовувати її як розв'язок поставленої задачі.

4 ЕКОНОМІЧНА ЧАСТИНА

4.1 Оцінювання комерційного потенціалу розробки

Метою проведення комерційного та технологічного аудиту є оцінювання комерційного потенціалу розробки інформаційної технології для аналізу та прогнозування кількості нових хворих на коронавірус в Україні.

Для проведення технологічного аудиту було залучено 3-х незалежних експертів Вінницького національного технічного університету кафедри системного аналізу та комп'ютерної інженерії: к.т.н., доц. Жуков С.О., к.т.н., доц. Козачко О.М., к.т.н., доц. Варчук І.В. Для проведення технологічного аудиту було використано таблицю 4.1 [21] в якій за п'ятибальною шкалою використовуючи 12 критеріїв здійснено оцінку комерційного потенціалу.

Таблиця 4.1 – Рекомендовані критерії оцінювання комерційного потенціалу розробки та їх можлива бальна оцінка

Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри- тері й	0	1	2	3	4
Технічна здійсненність концепції:					
1	Достовірність концепції не підтверджена	Концепція підтверджена експертними висновками	Концепція підтверджена розрахунками	Концепція перевірена на практиці	Перевірено роботоздатність продукту в реальних умовах
Ринкові переваги (недоліки):					
2	Багато аналогів на малому ринку	Мало аналогів на малому ринку	Кілька аналогів на великому ринку	Один аналог на великому ринку	Продукт не має аналогів на великому ринку

Продовження таблиці 4.1

Ринкові переваги (недоліки):					
3	Ціна продукту значно вища за ціни аналогів	Ціна продукту дещо вища за ціни аналогів	Ціна продукту приблизно дорівнює цінам аналогів	Ціна продукту дещо нижче за ціни аналогів	Ціна продукту значно нижче за ціни аналогів
4	Технічні та споживчі властивості продукту значно гірші, ніж в аналогів	Технічні та споживчі властивості продукту трохи гірші, ніж в аналогів	Технічні та споживчі властивості продукту на рівні аналогів	Технічні та споживчі властивості продукту трохи кращі, ніж в аналогів	Технічні та споживчі властивості продукту значно кращі, ніж в аналогів
5	Експлуатаційні витрати значно вищі, ніж в аналогів	Експлуатаційні витрати дещо вищі, ніж в аналогів	Експлуатаційні витрати на рівні експлуатаційних витрат аналогів	Експлуатаційні витрати трохи нижчі, ніж в аналогів	Експлуатаційні витрати значно нижчі, ніж в аналогів
6	Ринок малий і не має позитивної динаміки	Ринок малий, але має позитивну динаміку	Середній ринок з позитивною динамікою	Великий стабільний ринок	Великий ринок з позитивною динамікою
7	Активна конкуренція великих компаній на ринку	Активна конкуренція	Помірна конкуренція	Незначна конкуренція	Конкурентів немає
Практична здійсненність					
8	Відсутні фахівці як з технічної, так і з комерційної реалізації ідеї	Необхідно наймати фахівців або витратити значні кошти та час на навчання наявних фахівців	Необхідне незначне навчання фахівців та збільшення їх штату	Необхідне незначне навчання фахівців	Є фахівці з питань як з технічної, так і з комерційної реалізації ідеї

Продовження таблиці 4.1

Практична здійсненність					
9	Потрібні значні фінансові ресурси, які відсутні. Джерела фінансування ідеї відсутні	Потрібні незначні фінансові ресурси. Джерела фінансування відсутні	Потрібні значні фінансові ресурси. Джерела фінансування є	Потрібні незначні фінансові ресурси. Джерела фінансування є	Не потребує додаткового фінансування
10	Необхідна розробка нових матеріалів	Потрібні матеріали, що використовуються у військово-промисловому комплексі	Потрібні дорогі матеріали	Потрібні досяжні та дешеві матеріали	Всі матеріали для реалізації ідеї відомі та давно використовуються у виробництві
11	Термін реалізації ідеї більший за 10 років	Термін реалізації ідеї більший за 5 років. Термін окупності інвестицій більше 10-ти років	Термін реалізації ідеї від 3-х до 5-ти років. Термін окупності інвестицій більше 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій від 3-х до 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій менше 3-х років
12	Необхідна розробка регламентних документів та отримання великої кількості дозвільних документів на виробництво та реалізацію продукту	Необхідно отримання великої кількості дозвільних документів на виробництво та реалізацію продукту, що вимагає значних коштів та часу	Процедура отримання дозвільних документів для виробництва та реалізації продукту вимагає незначних коштів та часу	Необхідно тільки повідомлення відповідним органам про виробництво та реалізацію продукту	Відсутні будь-які регламентні обмеження на виробництво та реалізацію продукту

В таблиці 4.2 наведено рівні комерційного потенціалу.

Таблиця 4.2 – Рівні комерційного потенціалу розробки

Середньоарифметична сума балів СБ, розрахована на основі висновків експертів	Рівень комерційного потенціалу розробки
0-10	Низький
11-20	Нижче середнього
21-30	Середній
31-40	Вище середнього
41-48	Високий

В таблиці 4.3 наведено результати оцінювання експертами комерційного потенціалу розробки.

Таблиця 4.3 – Результати оцінювання комерційного потенціалу розробки

Критерії	Прізвище, ініціали, посада експерта		
	Жуков С.О.	Козачко О.М.	Варчук І.В.
	Бали, виставлені експертами:		
1	4	4	3
2	1	1	1
3	3	4	4
4	2	3	3
5	3	3	3
6	2	3	3
7	2	2	2
8	3	1	1
9	4	4	3
10	4	4	4
11	4	3	3
12	4	4	3
Сума балів	СБ ₁ =36	СБ ₂ =36	СБ ₃ =31
Середньоарифметична сума балів $\overline{СБ}$	$\overline{СБ} = \frac{\sum_1^3 СБ_i}{3} = \frac{36 + 36 + 31}{3} = 34$		

Середньоарифметична сума балів, розрахована на основі висновків експертів склала 34 бали, що згідно таблиці 4.2 вважається, що рівень комерційного потенціалу проведених досліджень є вище середнього.

Результати магістерської кваліфікаційної роботи можуть бути використані в державних та інших установах, як в Україні так і за кордоном.

4.2 Прогнозування витрат на виконання науково-дослідної роботи

Витрати, пов'язані з проведенням науково-дослідної роботи групуються за такими статтями: витрати на оплату праці, витрати на соціальні заходи, матеріали, паливо та енергія для науково-виробничих цілей, витрати на службові відрядження, програмне забезпечення для наукових робіт, інші витрати, накладні витрати.

1. Основна заробітна плата кожного із дослідників Z_0 , якщо вони працюють в наукових установах бюджетної сфери визначається за формулою:

$$Z_0 = \frac{M}{T_p} * t \text{ [грн]}, \quad (4.1)$$

де M – місячний посадовий оклад конкретного розробника (інженера, дослідника, науковця тощо), грн.;

T_p – число робочих днів в місяці; приблизно $T_p=21-23$ дні;

t – число робочих днів роботи дослідника.

Для розробки інформаційної технології аналізу та прогнозування кількості нових хворих на коронавірус в Україні необхідно залучити розробника з посадовим окладом 10000 грн. Кількість робочих днів у місяці складає 22, а кількість робочих днів програміста складає 30. Зведемо сумарні розрахунки до таблиця 4.4.

Таблиця 4.4 – Заробітна плата дослідника в науковій установі бюджетної сфери

Найменування посади	Місячний посадовий оклад, грн.	Оплата за робочий день, грн.	Число днів роботи	Витрати на заробітну плату грн.
Керівник	14000	636,4	5	3182
Розробник	10000	454,5	30	13636
Всього				16818

2. Розрахунок додаткової заробітної плати робітників

Додаткова заробітна плата Z_d всіх розробників та робітників, які приймали участь в розробці нового технічного рішення розраховується як 10 - 12 % від основної заробітної плати робітників.

На даному підприємстві додаткова заробітна плата начисляється в розмірі 10% від основної заробітної плати.

$$Z_d = (Z_o + Z_p) * \frac{N_{\text{дод}}}{100\%}, \quad (4.2)$$

$$Z_d = 0,11 * 16818 = 1635 \text{ (грн.)}.$$

3. Нарахування на заробітну плату $N_{3П}$ дослідників та робітників, які брали участь у виконанні даного етапу роботи, розраховуються за формулою:

$$N_{3П} = (Z_o + Z_d) * \frac{\beta}{100} \text{ [грн]}. \quad (4.3)$$

де Z_o – основна заробітна плата розробників, грн.;

Z_d – додаткова заробітна плата всіх розробників та робітників, грн.;

β – ставка єдиного внеску на загальнообов'язкове державне соціальне страхування, % .

Дана діяльність відноситься до бюджетної сфери, тому ставка єдиного внеску на загальнообов'язкове державне соціальне страхування буде складати 22%, тоді:

$$H_{3П} = (16818 + 1635) * \frac{22}{100} = 4107 \text{ (грн).}$$

4. Витрати комплектуючі К, що були використані під час виконання даного етапу роботи, розраховуються по кожному виду матеріалів за формулою:

$$K = \sum_{i=1}^n H_i \cdot C_i \cdot K_i, \quad (4.4)$$

де H_i – кількість комплектуючих i -го виду, шт.;

C_i – покупна ціна комплектуючих i -го найменування, грн.;

K_i – коефіцієнт транспортних витрат (1,1...1,15).

На таблиці 4.5 показано комплектуючі використані в розробці.

Таблиця 4.5 – Комплектуючі, що використані на розробку

Найменування	Ціна за одиницю, грн.	Витрачено	Вартість витраченого матеріалу, грн.
Папір	125	1	125
Ручка	12	1	12
CD-диск	12	1	12
Флешка	180	1	180
Блокнот	40	1	40
Всього			369
З врахуванням коефіцієнта транспортування			405,9

5. Амортизація обладнання, комп'ютерів та приміщень, які використовувались під час виконання даного етапу роботи

Дані відрахування розраховують по кожному виду обладнання, приміщенням тощо.

$$A = \frac{Ц \cdot T}{T_{кор} \cdot 12} \text{ [грн]}, \quad (4.5)$$

де Ц – балансова вартість даного виду обладнання (приміщень), грн.;

$T_{кор}$ – час користування;

T – термін використання обладнання (приміщень), цілі місяці.

Згідно пункту 137.3.3 Податкового кодекса амортизація нараховується на основні засоби вартістю понад 2500 грн. В нашому випадку для написання магістерської роботи використовувався персональний комп'ютер вартістю 30000 грн.

$$A = \frac{30000 \cdot 1}{2 \cdot 12} = 1250(\text{грн}).$$

6. До статті «Паливо та енергія для науково-виробничих цілей» відносяться витрати на всі види палива й енергії, що безпосередньо використовуються з технологічною метою на проведення досліджень.

$$B_e = \sum_{i=1}^n \frac{W_{yt} \cdot t_i \cdot Ц_e \cdot K_{впi}}{\eta_i}, \quad (4.6)$$

де W_{yt} – встановлена потужність обладнання на певному етапі розробки, кВт;

t_i – тривалість роботи обладнання на етапі дослідження, год;

$Ц_e$ – вартість 1 кВт-години електроенергії, грн;

$K_{впi}$ – коефіцієнт, що враховує використання потужності, $K_{впi} < 1$;

η_i – коефіцієнт корисної дії обладнання, $\eta_i < 1$.

Для написання магістерської роботи використовується персональний комп'ютер для якого розрахуємо витрати на електроенергію.

$$B_e = \frac{0,3 \cdot 190 \cdot 4,1 \cdot 0,5}{0,8} = 146,06(\text{грн}).$$

Витрати на службові відрядження, витрати на роботи, які виконують сторонні підприємства, установи, організації та інші витрати в нашому дослідженні не враховуються оскільки їх не було.

Накладні (загальновиробничі) витрати $B_{взв}$ охоплюють: витрати на управління організацією, оплата службових відряджень, витрати на

утримання, ремонт та експлуатацію основних засобів, витрати на опалення, освітлення, водопостачання, охорону праці тощо. Накладні (загальновиробничі) витрати $V_{\text{нзв}}$ можна прийняти як $(100\dots150)\%$ від суми основної заробітної плати розробників та робітників, які виконували дану МКНР, тобто:

$$V_{\text{нзв}} = (Z_o + Z_p) \cdot \frac{H_{\text{нзв}}}{100\%} [\text{грн}], \quad (4.7)$$

де $H_{\text{нзв}}$ – норма нарахування за статтею «Інші витрати».

$$V_{\text{нзв}} = 16818 \cdot \frac{100}{100\%} = 16818 \text{ (грн)}.$$

Сума всіх попередніх статей витрат дає витрати, які безпосередньо стосуються даного розділу МКНР

$$B = 16818 + 1635 + 4107 + 405,9 + 1250 + 146,06 + 16818 = 41395,3 \text{ (грн)}.$$

Прогнозування загальних витрат ZB на виконання та впровадження результатів виконаної МКНР здійснюється за формулою:

$$ZB = \frac{B}{\eta} [\text{грн}], \quad (4.8)$$

де η – коефіцієнт, який характеризує стадію виконання даної НДР.

Оскільки, робота знаходиться на стадії науково-дослідних робіт, то коефіцієнт $\beta = 0,9$.

Звідси:

$$ZB = \frac{41395,3}{0,5} = 45994,81 \text{ (грн)}.$$

4.3 Розрахунок економічної ефективності науково-технічної розробки

У даному підрозділі кількісно спрогнозуємо, яку вигоду, зиск можна отримати у майбутньому від впровадження результатів виконаної наукової роботи. Розрахуємо збільшення чистого прибутку підприємства $\Delta\Pi_i$, для кожного

із років, протягом яких очікується отримання позитивних результатів від впровадження розробки, за формулою

$$\Delta\Pi_i = \sum_1^n (\Delta\Pi_o \cdot N + \Pi_o \cdot \Delta N)_i \cdot \lambda \cdot \rho \cdot \left(1 - \frac{\nu}{100}\right) \quad (4.9)$$

де $\Delta\Pi_o$ – покращення основного оціночного показника від впровадження результатів розробки у даному році.

N – основний кількісний показник, який визначає діяльність підприємства у даному році до впровадження результатів наукової розробки;

ΔN – покращення основного кількісного показника діяльності підприємства від впровадження результатів розробки:

Π_o – основний оціночний показник, який визначає діяльність підприємства у даному році після впровадження результатів наукової розробки;

n – кількість років, протягом яких очікується отримання позитивних результатів від впровадження розробки:

λ – коефіцієнт, який враховує сплату податку на додану вартість. Ставка податку на додану вартість дорівнює 20%, а коефіцієнт $\lambda = 0,8333$.

ρ – коефіцієнт, який враховує рентабельність продукту. $\rho = 0,25$;

ν – ставка податку на прибуток. У 2021 році – 18%.

Припустимо, що при впровадженні результатів наукової розробки покращується якість отримання прогнозованих даних щодо прогнозу на коронавірус. Припустимо, що ціна зростає на 1000 грн. Кількість одиниць реалізованої продукції також збільшиться: протягом першого року на 30 шт., протягом другого року – на 15 шт., протягом третього року на 10 шт. Реалізація продукції до впровадження розробки складала 1 шт., а її ціна до 15000 грн. Розрахуємо прибуток, яке отримає підприємство протягом трьох років.

$$\Delta\Pi_1 = [1000 \cdot 1 + (15000 + 1000) \cdot 30] \cdot 0,833 \cdot 0,25 \cdot \left(1 + \frac{18}{100}\right) = 82167,55 \text{ (грн).}$$

$$\begin{aligned}\Delta\Pi_2 &= [1000 \cdot 1 + (15000 + 1000) \cdot (30 + 15)] \cdot 0,833 \cdot 0,25 \cdot \left(1 + \frac{18}{100}\right) \\ &= 123995,08 \text{ (грн)}.\end{aligned}$$

$$\begin{aligned}\Delta\Pi_3 &= [1000 \cdot 1 + (15000 + 1000) \cdot (30 + 15 + 10)] \cdot 0,833 \cdot 0,25 \cdot \left(1 + \frac{18}{100}\right) \\ &= 151327,32 \text{ (грн)}.\end{aligned}$$

4.4 Розрахунок ефективності вкладених інвестицій та періоду їх окупності

Розрахуємо основні показники, які визначають доцільність фінансування наукової розробки певним інвестором, є абсолютна і відносна ефективність вкладених інвестицій та термін їх окупності.

Розрахуємо величину початкових інвестицій PV , які потенційний інвестор має вкласти для впровадження і комерціалізації науково-технічної розробки.

$$PV = k_{\text{інв}} \cdot 3B, \quad (4.10)$$

$k_{\text{інв}}$ – коефіцієнт, що враховує витрати інвестора на впровадження науково-технічної розробки та її комерціалізацію. Це можуть бути витрати на підготовку приміщень, розробку технологій, навчання персоналу, маркетингові заходи тощо ($k_{\text{інв}} = 2 \dots 5$).

$$PV = 2 \cdot 45994,81 = 91989,61$$

Розрахуємо абсолютну ефективність вкладених інвестицій $E_{\text{абс}}$ згідно наступної формули:

$$E_{\text{абс}} = (ПП - PV) \quad (4.11)$$

де $ПП$ – приведена вартість всіх чистих прибутків, що їх отримає підприємство від реалізації результатів наукової розробки, грн.;

$$ПП = \sum_1^T \frac{\Delta\Pi_i}{(1+\tau)^t}, \quad (4.12)$$

де $\Delta\Pi_i$ – збільшення чистого прибутку у кожному із років, протягом яких виявляються результати виконаної та впровадженої НДЦКР, грн.;

T – період часу, протягом якою виявляються результати впровадженої НДДКР, роки;

τ – ставка дисконтування, за яку можна взяти щорічний прогнозований рівень інфляції в країні; для України цей показник знаходиться на рівні 0,2;

t – період часу (в роках).

$$ПП = \frac{82167,55}{(1+0,2)^1} + \frac{123995,08}{(1+0,2)^2} + \frac{151327,32}{(1+0,2)^3} = 242561,65 \text{ (грн).}$$

$$E_{abc} = (242561,65 - 91989,61) = 150572,04 \text{ (грн).}$$

Оскільки $E_{abc} > 0$ то вкладання коштів на виконання та впровадження результатів НДДКР може бути доцільним.

Розрахуємо відносну (щорічну) ефективність вкладених в наукову розробку інвестицій E_v . Для цього користуються формулою:

$$E_v = \sqrt[T_{ж}] \left(1 + \frac{E_{abc}}{PV} \right) - 1, \quad (4.13)$$

$T_{ж}$ – життєвий цикл наукової розробки, роки.

$$E_v = \sqrt[3] \left(1 + \frac{150572,04}{91989,61} \right) - 1 = 0,62 = 62\%$$

Визначимо мінімальну ставку дисконтування, яка у загальному вигляді визначається за формулою:

$$\tau = d + f, \quad (4.14)$$

де d – середньозважена ставка за депозитними операціями в комерційних банках; в 2021 році в Україні $d = (0,14 \dots 0,2)$;

f – показник, що характеризує ризикованість вкладень; зазвичай, величина $f = (0,05 \dots 0,1)$.

$$\tau_{\min} = 0,18 + 0,05 = 0,23$$

Так як $E_g > \tau_{\min}$ то інвестор може бути зацікавлений у фінансуванні даної наукової розробки.

Розрахуємо термін окупності вкладених у реалізацію наукового проекту інвестицій за формулою:

$$T_{ок} = \frac{1}{E_g} \quad (4.15)$$

$$T_{ок} = \frac{1}{0,62} = 1,6 \text{ роки}$$

Так як $T_{ок} \leq 3 \dots 5$ -ти років, то фінансування даної наукової розробки в принципі є доцільним.

4.5 Висновки

Було проведено оцінку комерційного потенціалу інформаційної технології для аналізу та прогнозування кількості нових хворих на коронавірус в Україні, який є на вище середньому рівні.

Прогнозування витрат на виконання науково-дослідної роботи по кожній з статей витрат складе 41395,3 грн. Загальна ж величина витрат на виконання та впровадження результатів даної НДР склала 45994,81 грн.

Вкладені інвестиції в даний проект окупляться через 1,6 роки при приведеній вартості всіх чистих прибутків, що їх отримає підприємство від реалізації результатів наукової розробки 242561,65 грн.

ВИСНОВКИ

У роботі вивчено поширення коронавірусу COVID-19 у вигляді щоденного приросту кількості нових підтверджених хворих на нього в Україні у 2021 році та ідентифіковано моделі для короткострокового та довгострокового прогнозування.

Показано актуальність роботи та проаналізовано фактори, що впливають на поширення інфекційних захворювань. Розглянуто та порівняно основні епідеміологічні характеристики трьох коронавірусів: SARS-CoV-2, SARS-CoV та MERS-CoV.

Порівняльний аналіз методів прогнозування інфекційних захворювань висвітлив їх переваги та недоліки. Розглянуто існуючі методи прогнозування поширеності COVID-19 та досліджено часові ряди поширення захворювання в Україні.

На базі отриманих результатів були створені моделі для прогнозування кількості нових хворих на коронавірус в Україні та за її межами на основі відкритих даних. Перевагою є те, що при короткостроковому прогнозуванні можна отримати більш менш зрозумілу картину того, що відбувається.

В ході виконання кваліфікаційної роботи було отримано такі результати:

1. Сформовано висновки щодо наявних методів прогнозування інфекційних захворювань.

2. Розроблена моделі, за допомогою якої здійснено довгострокове та короткострокове прогнозування щоденного приросту кількості нових підтверджених хворих на нього в Україні у 2021 році .

3. Розроблена модель, за допомогою якої побудовано короткостроковий прогноз по 70 країнах світу.

Але причиною невдалого прогнозування на довший період, ніж 14 днів, зумовлено тим, що дані є досить невизначеними, частина – не є у відкритому доступі. Крім того, оскільки перевірити всю сукупність фізично неможливо, в наявній статистиці є серйозні помилки, що знижує якість прогнозів. Щоб

компенсувати цей фактор, необхідно оновлювати дані та проводити більш якісні масштабні випробування. Можливі способи подальшого проведення цього дослідження включають збирання даних про поширення хвороби та створення більшої бази знань. Також можна будувати складні багаторівневі моделі на основі нових, більш точних і вичерпних даних. Зокрема, ці дані можуть бути анонімною інформацією про переміщення населення та соціальну діяльність (також, можна використати дані Google та Apple про мобільність населення), що дозволить ефективніше відстежувати контакти з пацієнтами та розірвати ланцюжок захворювання з самого початку. Оцінка комерційного потенціалу інформаційних технологій, що використовуються для аналізу та прогнозування кількості хворих на коронавірус в Україні, вища за середню.

Інвестиції в проект повернуть протягом 1,6 років виходячи з теперішньої вартості всього чистого прибутку, отриманого компанією від впровадження досягнень наукових розробок.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Мокін В.Б., Лосенко А.В., Ящолт А.Р., Гевеленко А.О. Прогнозування тижневих трендів кількості нових хворих на коронавірус у країнах світу. Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях (матеріали XX Міжнародної науково-практичної конференції, м. Вінниця, жовтень 2021р.), С. 209-212, Електрон. текст. дані, 2021. URL: https://itgip.org/wp-content/uploads/2021/10/1_Збірка_2021.pdf
2. Rouaud M. Probability, Statistics and Estimation : монографія. Англія, 2013. 60 с.
3. Brockwell P. J.; Davis R. A. Time Series: Theory and Methods : монографія. Нью Йорк : Springer New York, 2009. с. 273.
4. Harvey, A. C. Time Series Models. 2nd Edition. : Harvester Wheatsheaf, 1993.
5. Capasso V., The Mathematical Structure of Epidemic Systems : Springer Verlag, 1993.
6. Грешилов А. А., Стакун В. А., Стакун А. А. Математичні методи побудови прогнозів. : монографія. Казахстан : Радио и связь, 1997. 112 с.
7. Melby P. Robustness of Adaptation in Controlled Self-Adjusting Chaotic Systems : монографія. США : Fluctuation and Noise Letters, 2002.
8. Sinitsyn. S.(2013). Safe Protocols for Generating Power Pulses with Heterogeneous Populations of Thermostatically Controlled Loads :монографія. США: Theoretical Computer Science, 2013.
9. Cybenko, G.V. Approximation by Superpositions of a Sigmoidal function : У van Schuppen, Jan H. Mathematics of Control, Signals, and Systems: Springer International, 2006. с. 303–314
10. Wasserman P.D. Advanced methods in neural computing : Van Nostrand Reinhold, 1993.

11. Андрейчин М. Н. Коронавірусна інфекція : Енциклопедія сучасної України : НАН України, НТШ, 2014. 767 с.
12. Голубовська М.А. Інфекційні хвороби : підручник. Київ : ВСВ , 2018. 633 с.
13. Мокін В. Б, Лосенко А.В, Яцолт А.Р. Інформаційна технологія аналізу та прогнозування кількості нових випадків на коронавірус SARS-CoV-2 в Україні на основі моделі Prophet : Вісник Вінницького політехнічного інституту, 2020. 71–83 с.
14. Мокін В. Б, Лосенко А.В, Яцолт А.Р. Інформаційна технологія аналізу та прогнозування багатохвильової кількості нових випадків захворювань на коронавірус COVID-19 на основі моделі Prophet : Вісник Вінницького політехнічного інституту, 2020. 65 с.
15. В. Б. Мокін, А. В. Лосенко, і А. Р. Яцолт, «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, 2020.
16. В. Б. Мокін, А. В. Лосенко, і А. Р. Яцолт, «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.
17. Мокін В.Б. COVID-19 in 70 countries: daily Prophet forecast В. Б. Мокін, А. В. Лосенко Kaggle. URL: <https://www.kaggle.com/vbmokin/covid-19-in-70-countries-daily-prophet-forecast>
18. В. Б. Мокін, А. В. Лосенко, і А. Р. Яцолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків захворювань на коронавірус SARS-COV-2 в Україні на основі моделі prophet», Вісник ВПІ, вип. 5, с. 71–83, Лис 2020.
19. Мокін В.Б. COVID-19: Week trends 70 countries mapping. Version 21. 05.09.2021. Kaggle. URL: <https://www.kaggle.com/vbmokin/covid-19-week-trends-70-countries-mapping?scriptVersionId=74027060>

20. COVID-19 in 70 countries: daily Prophet forecast, 2020 р..URL:
<https://www.kaggle.com/vbmokin/covid-19-in-70-countries-daily-prophet-forecast?scriptVersionId=47433942>

21. В. О. Козловський, О. Й. Лесько, В. В. Кавецький. Методичні вказівки до виконання економічної частини магістерських кваліфікаційних робіт : Уклад. Вінниця : ВНТУ, 2021. – 42 с.

Додаток А
(обов'язковий)

Міністерство освіти і науки України
Вінницький національний технічний університет
Факультет комп'ютерних систем і автоматики

ЗАТВЕРДЖУЮ

Завідувач кафедри САІТ

_____ д.т.н., проф. Мокін В. Б.

«__» _____ 2021 р.

ТЕХНІЧНЕ ЗАВДАННЯ

на магістерську кваліфікаційну роботу

«ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ
КІЛЬКОСТІ НОВИХ ХВОРИХ НА КОРОНАВІРУС В УКРАЇНІ У 2021
РОЦІ»

08-53.МКР.004.02.000.ТЗ

Керівник: д.т.н., проф. каф. САІТ

_____ Мокін В.Б.

«__» _____ 2021 р.

Розробив: студент гр. 2ІСТ-20м

_____ Гевеленко А.О.

«__» _____ 2021 р.

Вінниця 2021

1. Підстава для проведення робіт

Підставою для виконання роботи є наказ № __ по ВНТУ від «__» _____ 2021 р., та індивідуальне завдання на МКР, затверджене протоколом № __ засідання кафедри САІТ від «__» _____ 2021 р.

2. Джерела розробки:

- В. Б. Мокін, А. В. Лосенко, і А. Р. Яцолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків захворювань на коронавірус SARS-COV-2 в Україні на основі моделі prophet», Вісник ВПІ, вип. 5, с. 71–83, 2020.
- Мокін В.Б. COVID-19 in 70 countries: daily Prophet forecast В. Б. Мокін, А. В. Лосенко Kaggle. URL: <https://www.kaggle.com/vbmokin/covid-19-in-70-countries-daily-prophet-forecast>

3. Мета і призначення роботи:

Розробка інформаційної технології аналізу та прогнозування кількості нових хворих на коронавірус в Україні в 2021 році.

4. Вихідні дані для проведення робіт:

- кількість хворих з відкритих даних РНБО;
- кількість хворих з відкритих даних по країнам світу.

5. Методи дослідження:

- позвідувальний аналіз;
- прогнозування даних.

6. Етапи роботи і терміни їх виконання:

1. Аналіз предметної області _____ – _____
2. Розробка інформаційної технології..... _____ – _____
3. Реалізація інформаційної технології..... _____ – _____
4. Оформлення пояснювальної записки. _____ – _____

7. Очікувані результати та порядок реалізації:

Отримання системи моніторингу пасажирських автоперевезень.

8. Вимоги до розробленої документації

Пояснювальна записка оформлена у відповідності до вимог «Методичних вказівок до виконання та оформлення магістерських кваліфікаційних робіт для студентів спеціальності 126 – «Інформаційні системи та технології» денної форми навчання».

9. Порядок приймання роботи

Публічний захист «__» _____ 2021 р.

Початок розробки «__» _____ 2021 р.

Граничні терміни виконання МКР «__» _____ 2021 р.

Розробив студент групи 2ІСТ-20м _____ Гевеленко А.О.

Додаток Б
(обов'язковий)

Протокол перевірки кваліфікаційної роботи

Назва роботи: «Інформаційна технологія аналізу та прогнозування кількості нових хворих на коронавірус в Україні у 2021 році»

Тип роботи: магістерська кваліфікаційна робота

Підрозділ: кафедра САІТ

Науковий керівник: Мокін В.Б. д.т.н., проф. каф. САІТ

Показники звіту подібності

Unicheck	
Оригінальність	81,4 %
Схожість	18,6 %

Аналіз звіту подібності (відмітити потрібне)

- Запозичення, виявлені у роботі, оформлені коректно і не містять ознак плагіату.
- Виявлені у роботі запозичення не мають ознак плагіату, але їх надмірна кількість викликає сумніви щодо цінності роботи і самостійності її автора. Роботу направити на доопрацювання.
- Виявлені у роботі запозичення є недобросовісними і мають ознаки плагіату та/або в ній містяться навмисні спотворення тексту, що вказують на спроби приховування недобросовісних запозичень.

Заявляю, що ознайомлений з повним звітом подібності, який був згенерований системою щодо роботи

Автор _____ Гевеленко А.О.
(підпис)

Опис прийнятого рішення

Робота допускається до захисту _____

Особа, відповідальна за перевірку _____ Жуков С. О.
(підпис)

Додаток В

(ДОВІДНИКОВИЙ)

Лістинг програмного коду.

```

import io
import pandas as pd
import numpy as np
import requests
import seaborn as sns
import matplotlib
from matplotlib import pyplot as plt
import plotly.express as px
import plotly.graph_objects as go

from datetime import date, timedelta, datetime
from fbprophet import Prophet
#from fbprophet.make_holidays import make_holidays_df
#from fbprophet.diagnostics import cross_validation, performance_metrics
#from fbprophet.plot import plot_cross_validation_metric
import holidays
from collections import Counter
import pycountry

from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error

import warnings
warnings.simplefilter('ignore')
print(f'Download confirmed daily data from RNBO of Ukraine')
myfile = requests.get('https://api-covid19.rnbo.gov.ua/charts/main-data?mode=ukraine')
open('data', 'wb').write(myfile.content)
data = pd.read_json('data')
data
data['n_confirmed'] = data['confirmed'].diff()
# A new wave from 6 July
data = data[1:649].reset_index(drop=True)
data = data[:-1]
data['n_confirmed'] = data['n_confirmed'].astype('int')
data.tail(3)
df2 = data[['dates', 'n_confirmed']].dropna()
df2 = df2[df2['n_confirmed'] > 0].reset_index(drop=True)
df2['n_confirmed'].plot()
if rolling_window:
    df2['n_confirmed'] = df2['n_confirmed'].rolling(7).mean()
    df2['n_confirmed'].plot()
def cut_df(date0: str,
           df:pd.DataFrame(),

```

```

    col: str):
# Deletes all rows of dataframe df with df[col] < date0
format0 = '%Y-%m-%d'
df_temp = df.copy()
df_temp['col_dt'] = pd.to_datetime(df_temp[col], format=format0, errors='coerce')
date0_dt = datetime.strptime(date0, format0)
df_temp = df_temp[df_temp['col_dt'] >= date0_dt]

    return df_temp
def plot_with_anomalies(df, cols_y_list, cols_y_list_name, dates_x, col_anomalies, val_a
nomal, log_y=False):
    fig = px.line(df, x=dates_x, y=cols_y_list[0], title=cols_y_list_name[cols_y_list[0]], lo
g_y=log_y, template='gridon',width=700, height=800)
    y_max = df[cols_y_list[0]].max()
    for i in range(len(cols_y_list)-1):
        fig.add_trace(go.Scatter(x=df[dates_x], y=df[cols_y_list[i+1]], mode='lines', name=
cols_y_list_name[cols_y_list[i+1]]))
        max_i = df[cols_y_list[i+1]].max()
        y_max = max_i if max_i > y_max else y_max

    anomal_dates_list = df[df[col_anomalies] == val_anomal][dates_x].tolist()
    y_min = min(df[cols_y_list[0]].min(),0)
    for i in range(len(anomal_dates_list)):
        anomal_date = anomal_dates_list[i]
        fig.add_shape(dict(type="line", x0=anomal_date, y0=y_min, x1=anomal_date, y1=y
_max, line=dict(color="red", width=1)))
    fig.show()
data['stringency_index_jump'] = 0
for i in range(len(data)-1):
    if (data.loc[i+1,'stringency_index'] is not None) and (data.loc[i,'stringency_index'] is n
ot None) and \
    (data.loc[i+1,'stringency_index'] < data.loc[i,'stringency_index']):
        data.loc[i+1, 'stringency_index_jump'] = 1
source_gov = 'https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-gove
rnment-response-tracker'
dates_gov_list = data[data['stringency_index_jump'] == 1]['date'].tolist()
holidays_df = aux_holidays_df_generator(holidays_df, dates_gov_list, 'the weakening of
quarantine', source_gov, 2)
plot_with_anomalies(data, ["stringency_index"], {"stringency_index": "Stringency index
and dates of the weakening of quarantine in " + country_main}, 'date', 'stringency_index_
jump', 1)

```

Додаток Г
(обов'язковий)

ІЛЮСТРАТИВНА ЧАСТИНА

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ
КІЛЬКОСТІ НОВИХ ХВОРИХ НА КОРОНАВІРУС В УКРАЇНІ У 2021
РОЦІ

Виконав: студент гр. 2ІСТ-20м

_____ Гевеленко А.О.

«__» _____ 2021 р.

Керівник: д.т.н., проф. каф. САІТ

_____ Мокін В.Б.

«__» _____ 2021 р.

Нормоконтроль: к.т.н., доцент

_____ Жуков С. О.

«__» _____ 2021 р.

Stringency index and dates of the weakening of quarantine in Ukraine

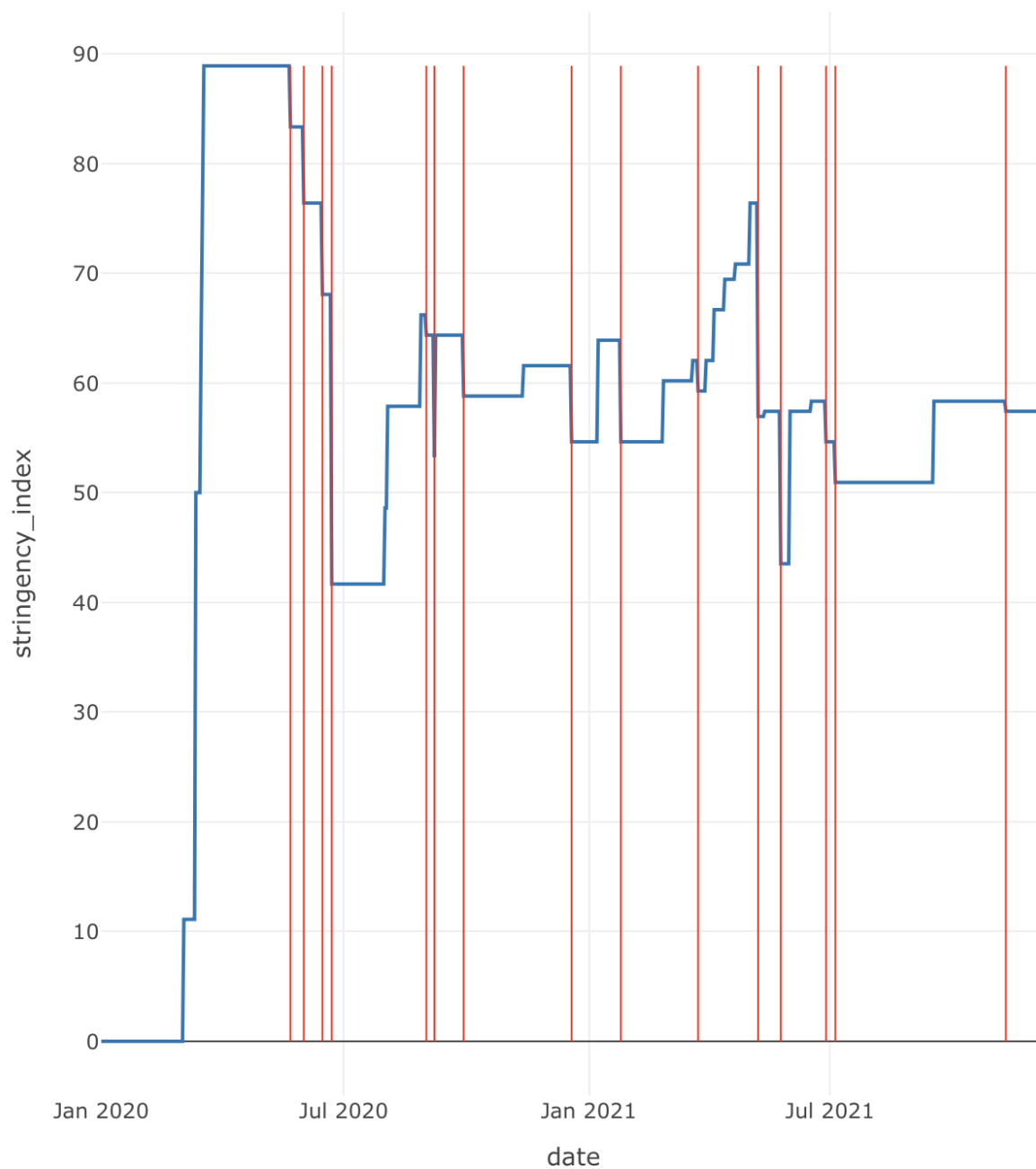


Рисунок Г.1 – Сумарний за добу Stringency-індекс України з ослабленнями та підсиленнями карантину (синя лінія) та дати аномалій-послаблень карантину (червона лінія) за допомогою довгострокового прогнозу

Average temperature over time in Ukraine

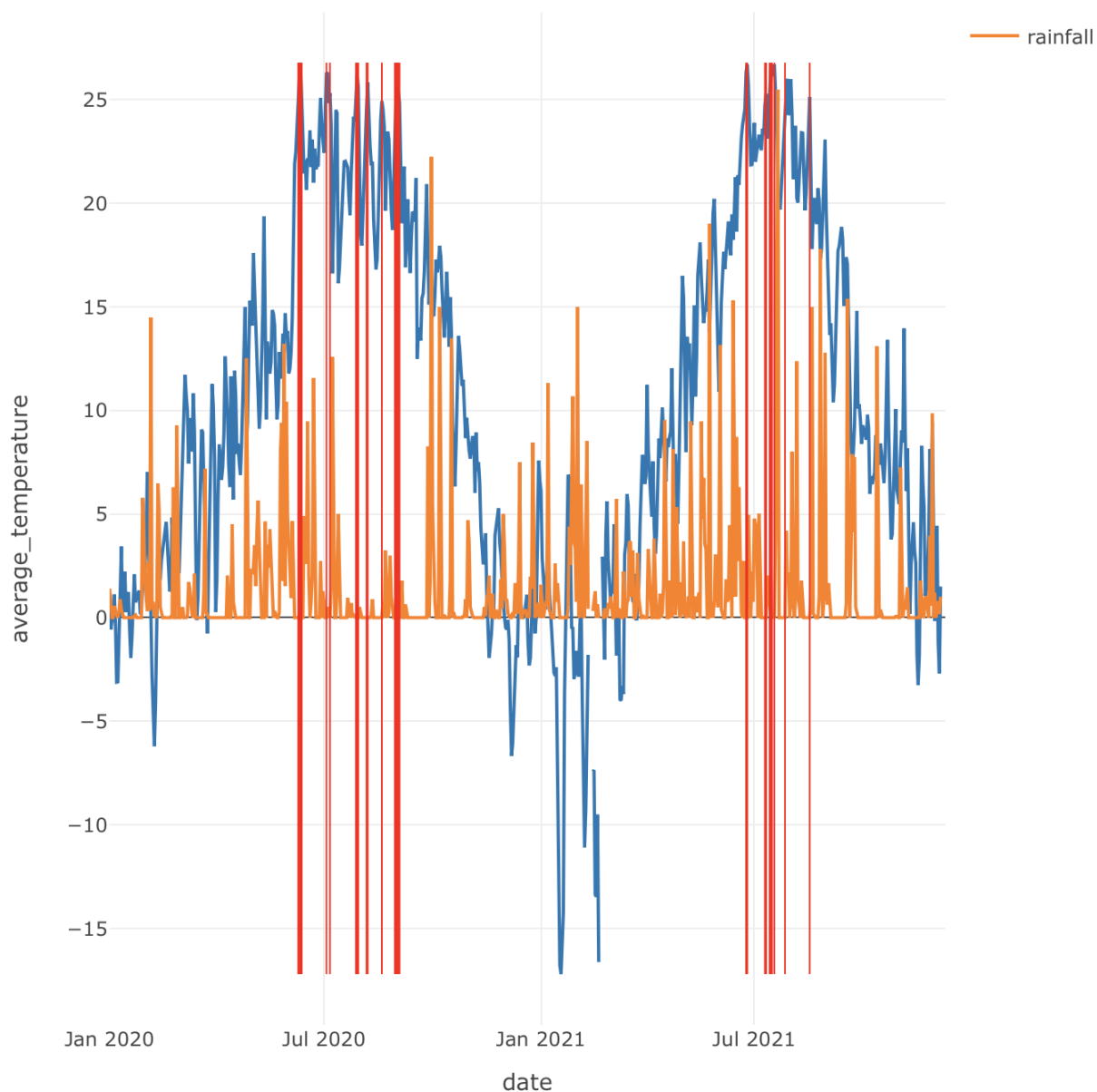


Рисунок Г.2 – Середньодобова температура в Україні (помаранчева крива), кількість опадів, мм (синя крива), дати аномальних «метопаттернів» (червоні вертикальні лінії) за допомогою довгострокового прогнозу

Confirmed cases and holidays data in Ukraine

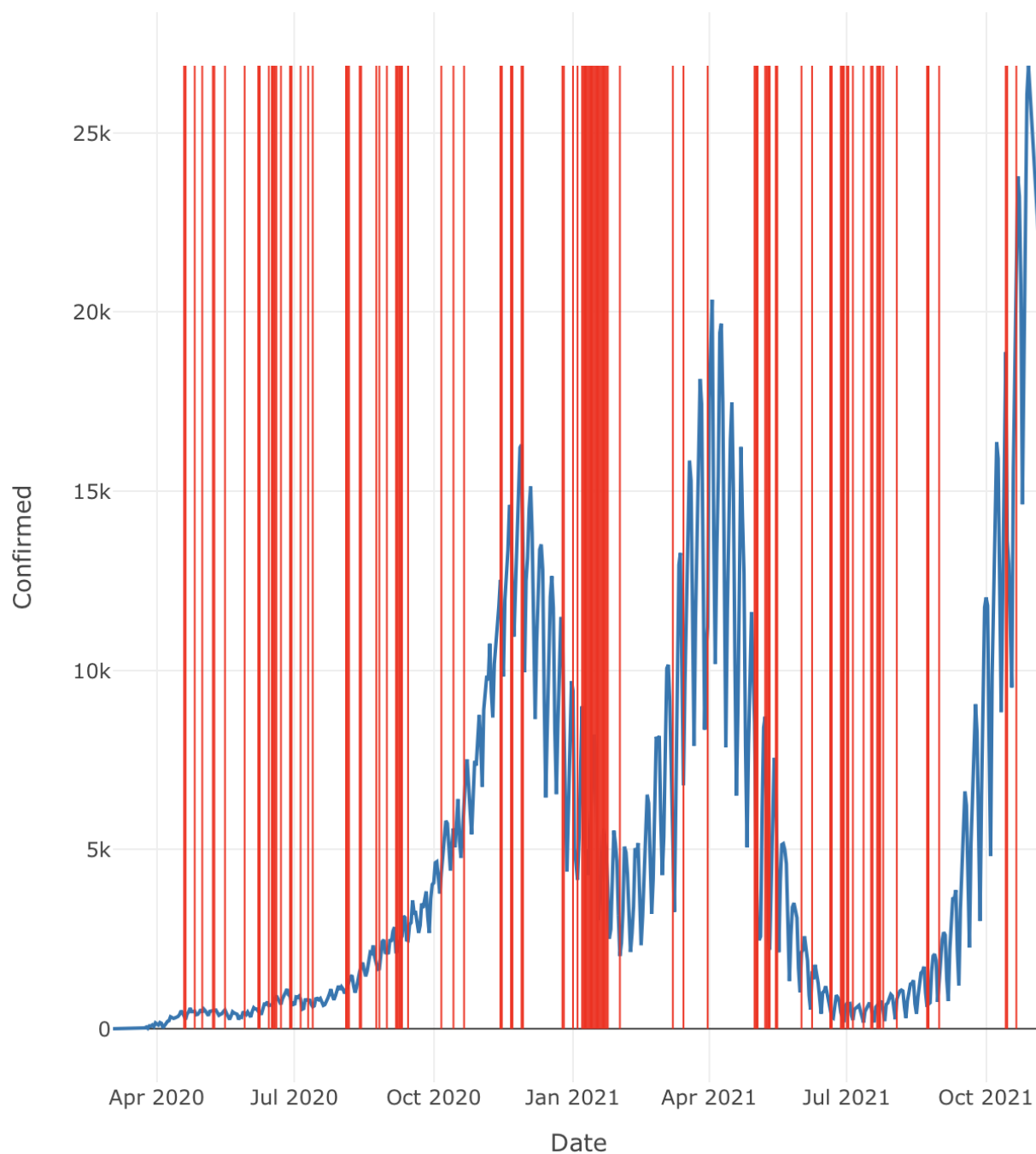


Рисунок Г.3 – Підтвержені захворювання на коронавірус з відображенням св'ятів за допомогою довгострокового прогнозування

Confirmed cases in Ukraine

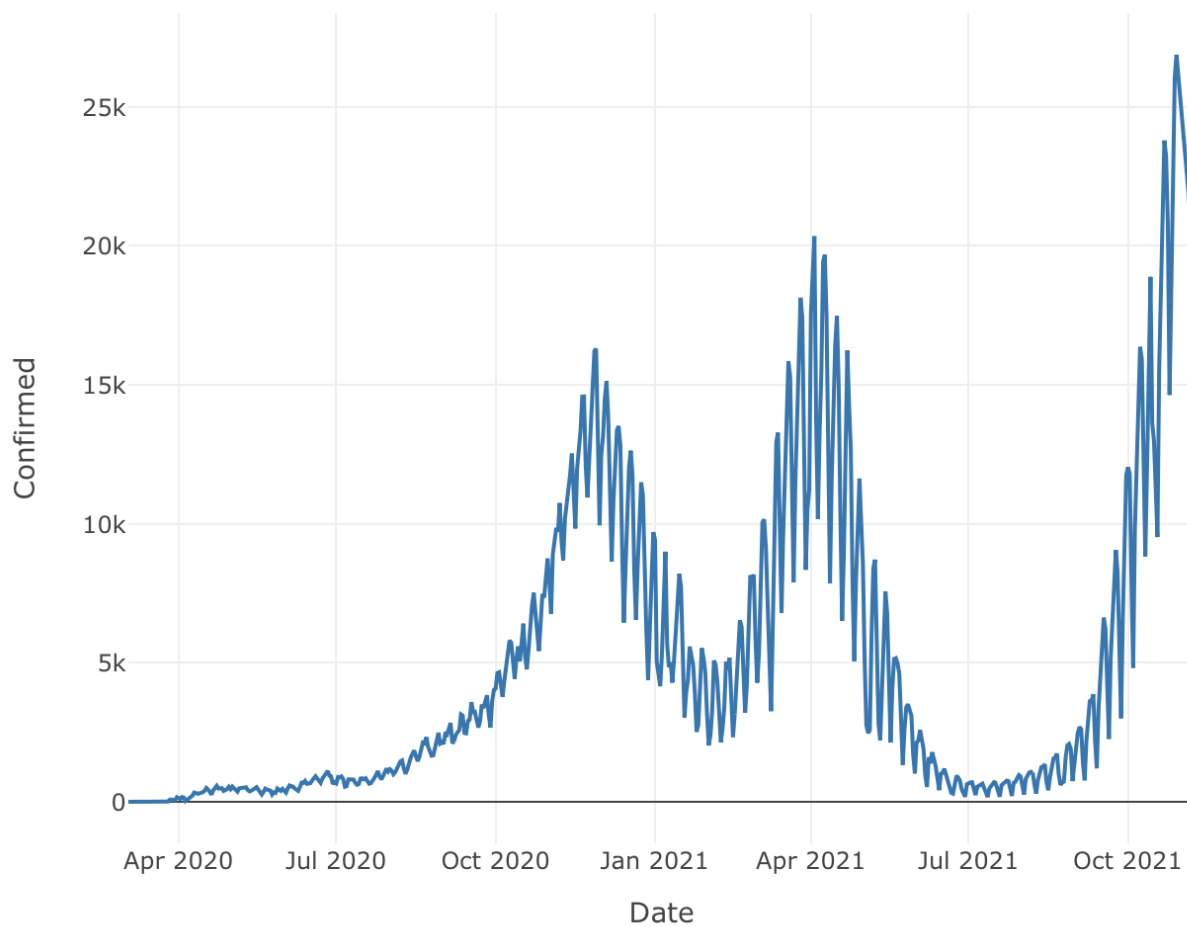


Рисунок Г.4 – Підтверджені захворювання на коронавірус в проміжок довгого періоду часу

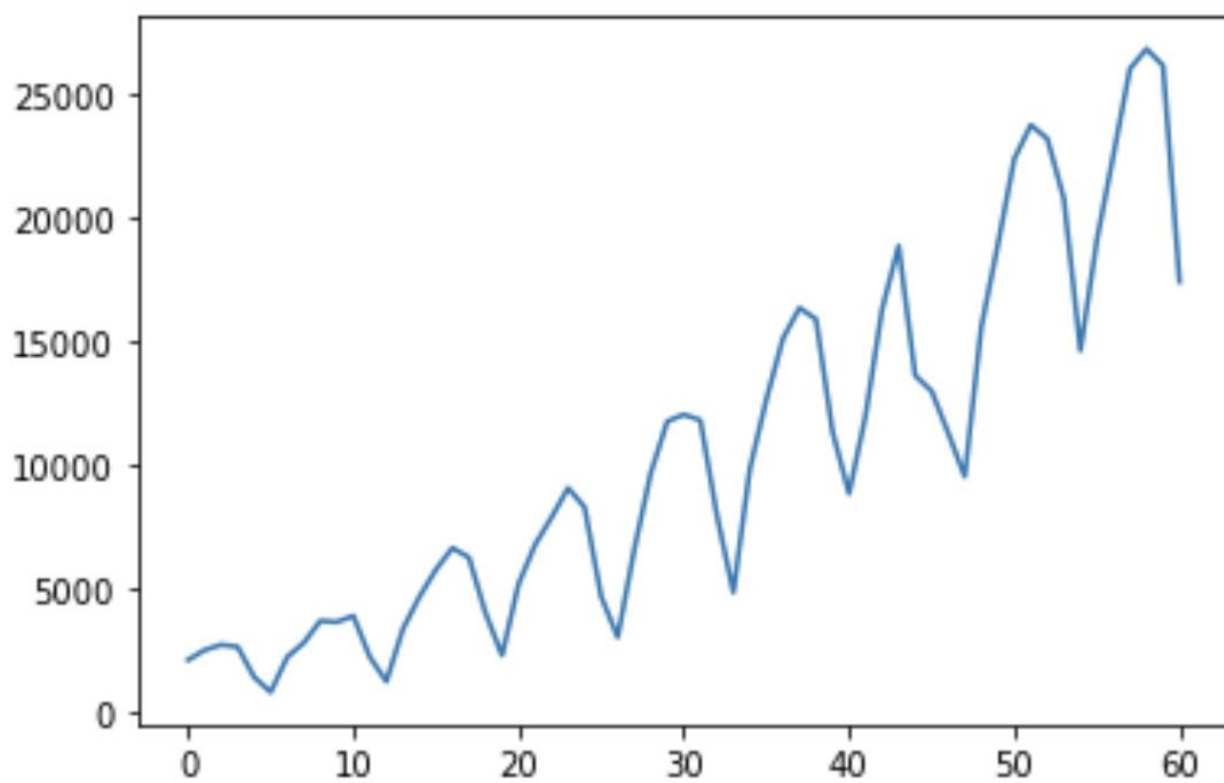


Рисунок Г.5 – Короткостроковий прогноз за спадом хвилі

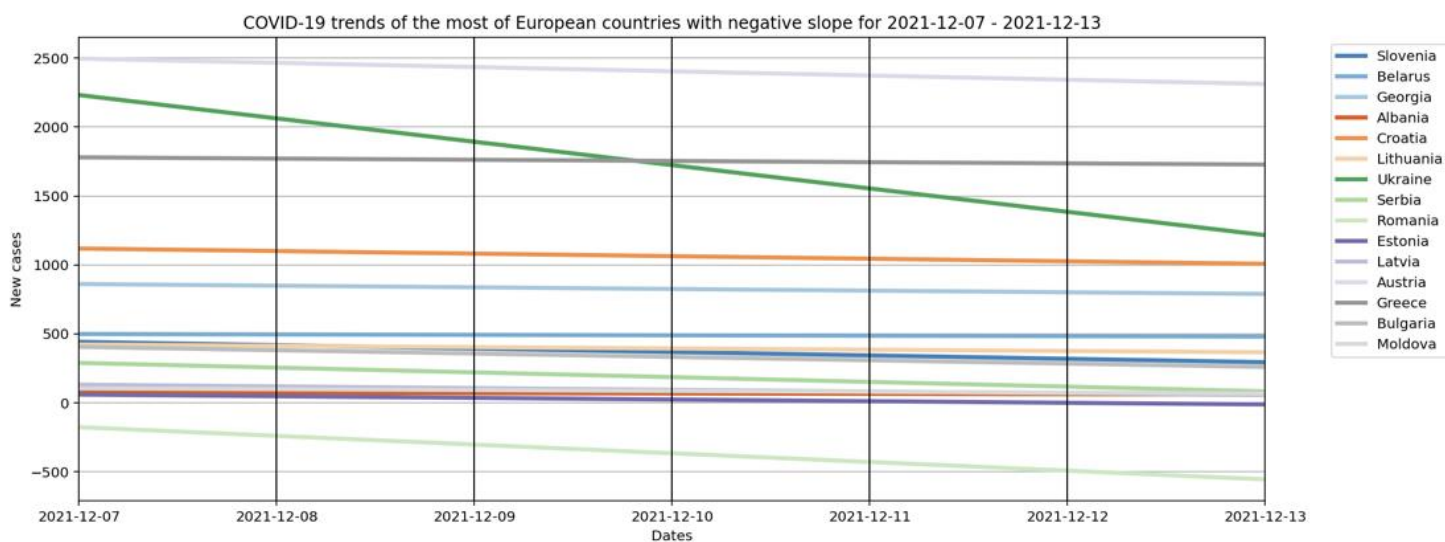


Рисунок Г.6 – Графік трендів найбільш популярних країн Європи з найбільш негативним прогнозом