Київський національний університет
імені Тараса Шевченка
Факультет комп'ютерних наук та кібернетики
Кафедра моделювання складних систем

Кваліфікаційна робота
на здобуття ступеня бакалавра

за спеціальністю 113 Прикладна математика
на тему:

Прогнозування розповсюдження COVID-19 з використанням аналізу часових рядів та нейронних мереж

Виконала студентка 4-го курсу
Логвіна Ангеліна Вікторівна

_____

Науковий керівник:
кандидат технічних наук, доцент
Кулян Віктор Романович

_____

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент
_____

Роботу розглянуто й допущено до захисту на засіданні кафедри моделювання складних систем
«___» _____ 202_ р.,
протокол №___
Завідувач кафедри
Д.І.Черній

_____

Київ — 2021

ЗМІСТ

ВСТУП

З моменту спалаху коронавірусу COVID-19 на початку 2020 року, вірус уразив увесь світ і забрав життя мільонів людей. У березні 2020 р. Всесвітня організація охорони здоров'я (ВООЗ) оголосила пандемію, першу у своєму роді в нашому поколінні [17]. На сьогоднішній день багато країн та регіонів були закриті, а також застосовувались суворі заходи соціального дистанціювання для припинення розповсюдження вірусу.

З точки зору стратегічного управління та управління охороною здоров'я, схема розповсюдження хвороби та прогнозування її розповсюдження з часом має велике значення, щоб врятувати життя і мінімізувати соціальні та економічні наслідки захворювання. Всередині всього наукового співтовариства дана проблема викликає інтерес до дослідження, включаючи такі області як математична епідеміологія, моделювання біологічних систем, обробка сигналів та управління технікою. Проблема моделювання пандемії має важливе практичне значення для урядів та осіб, які приймають рішення. Нефармацевтичні втручання(НФВ) є одним з найкращих способів боротьби з пандемічними захворюваннями при відсутності вакцини та ліків. НФВ стосуються дій та політики, прийнятих окремими особами, органами влади чи урядами з метою сприяння уповільнення поширення епідемічних захворювань.

Під час пандемії COVID-19 було зроблено кілька спроб класифікувати та кількісно визначити різні НФВ різних регіонів і націй. Вважається, що кількісна оцінка НФВ корисна для порівняння ефективності регіональної політики спрямованої на зменшеняня швидкості розповсюдження пандемії. За допомогою методів машинного навчання, кількісно визначені НФВ можуть бути використані для прогнозування майбутніх тенденції пандемії та моделювання сценаріїв для кращого управління людьми та медичними ресурсами та врешті-решт призначення відповідних НФВ для контролю пандемії [18], [19].

Оскільки ще не знайдено ліків від коронавірусу, актуальність данної роботи полягає у моделюванні поширення хвороби, і передбачені її впливу для оптимізації планування управління різними послугами та ресурсами.

Моделювання та прогнозування поведінки щоденного розповсюдження вірусу може допомогти системам охорони здоров'я бути готовими до обслуговування майбутньої кількості пацієнтів. Точне прогнозування хвороби є важливим, оскільки від нього залежать протиепідемічні заходи, які потрібно буде запровадити локальним урядам задля безпечного соціального життя.

Метою даної роботи є дослідження та порівняльна оцінка моделей часових рядів, які можуть бути використані для прогнозування активних, підтверджених, смертельних та одужаних випадків. Було розглянуто моделі ARIMA та FBProphet,

які широко використовуються завдяки широким можливостям прогнозування.

Також представлений короткий вступ до математичного моделювання біологічних систем, щоб висвітлити сферу застосування цього дослідження та відкрити перспективи для зацікавлених дослідників, які можуть бути менш обізнані з контекстом прикладної проблеми.

Задля повного розкриття теми прогнозування згадується метод нейронних мереж. Ми описали алгоритм побудови мережі для прогнозування епідемії.

РОЗДІЛ 1. МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ЕПІДЕМІОЛОГІЧНИХ ХВОРОБ

1.1 Математичне моделювання поширення інфекції

Спалах заразної хвороби у великій популяції є стохастичною подією. Починаючи з однієї зараженої особини, інфекція передається іншим стохастично безпосереднім контактом, близькістю або слідами навколишнього середовища (інфікованими предметами). Нові заражені особини по черзі передають інфекцію (з певною ймовірністю) здоровим особам, з якими вони контактують.

На первинних етапах спалаху епідемії здорові заражені особини статистично незалежні. Як результат, шанс того, що декілька інфікованих людей можуть зустріти одну здорову особу є ймовірнісно низькою. Припустимо, що кожна заражена особина інфікує R_0 нових людей в середньому[1]. Якщо $R_0 > 1$ хвороба поширюється експоненційно від початку до кінця часового інтервалу (наприклад, щодня). Однак при скінченній кількості населення експоненціальне зростання не може тривати вічно. Залежно від чисельності популяції та механізму передачі, ймовірність того, що декілька заражених людей зустрінуть незалежно здорових осіб зменшується. Тому після первинного спалаху інфекція експоненціально поширюється серед населення, але через деякий час заражених стає все більше і більше і пересічному зараженому складно зустріти неінфікованих осіб R_0, а отже, кількість розмноження падає і кількість нових інфекцій експоненційно зменшується. Отже, стохастична модель зараження розмноження, якось насичується. Імовірнісні моделі зазвичай використовуються для моделювання такого поширення епідемії про процес розгалуження та розподіл Пуассона для ймовірність контакту між інфекційними та здоровими суб'єктами.

Існує евристичний підхід до формування моделі, який менш ретельний, але однаково точний при великих популяціях.

Припустимо, що $x(t)$ позначає кількість заражених особин популяції в момент часу t. Далі, припускаючи, що ймовірність зараження зростає із збільшенням кількості заражених особин, ми припускаємо, що варіація кількості заражених популяції між часом t і $t+\Delta$ (понад відносно невеликих інтервалів Δ) пропорційна кількості інфікованих осіб, тобто

$$\frac{dx(t)}{dt} \approx \frac{x(t+\Delta) - x(t)}{\Delta} = \phi(t)x(t). \tag{1.1}$$

Назвемо $\phi(t)$ функцією відтворення, яка моделює еволюцію зараженого населення з часом. Ця функція враховує різні імовірнісні фактори, такі як швидкість

[1]Базове репродукційне число (R_0) — середня кількість осіб, що безпосередньо інфікуються хворим упродовж усього заразного періоду хвороби за умови потрапляння хворого до повністю незараженої популяції.[5]

передачі інфекції, густота населення та схеми контактів. $\phi(t)$ можна інтерпретувати як експоненціальну швидкість, з оберненими одиницями часу.

Позначимо поширення $k-$го покоління інфекції $x_k \stackrel{\Delta}{=} x(k\Delta)$,

Тоді (1.1) можна дискретизувати наступним чином:

$$x_{k+1} = [1 + \Delta\phi(k\Delta)]x_k. \tag{1.2}$$

Тепер визначимо репродукційне число

$$r_k \stackrel{\Delta}{=} [1 + \Delta\phi(k\Delta)]. \tag{1.3}$$

Очевидно, що населення за дискретизованим індексом часу k можна рекурсивно знайти з початкової умови x_0:

$$x_k = (r_{k-1}r_{k2}\dots r_0)x_0. \tag{1.4}$$

Виявляється, якщо для всіх k, $r_k < 1$ (або еквівалентно $\phi(t) < 0$), інфекція зникає, інакше, якщо $r_k > 1$ (або $\phi(t) > 0$) вона поширюється. У найпростішому випадку, коли функція відтворення - це константа $\phi(t) = \lambda$, маємо постійне репродукційне число $\mathcal{R} = 1 + \lambda\Delta$, що призводить до експоненційного зростання / спаду:

$$x_k = x_0\mathcal{R} \tag{1.5}$$

або в неперервному випадку:

$$x_k = x_0e^{\lambda t}. \tag{1.6}$$

Більш загально, значення функції відтворення $\phi(t)$ (або r_k в дискретному випадку) змінюється в часі та залежить від загальної сприйнятивої популяції, популяції осіб, що зазнали впливу (носії хвороби, але без симптомів), моделі контакту та таких заходів, як соціальне дистанціювання та карантин. Функція відтворення дозволяє проводити аналіз стійкості епідемічних моделей.

1.2 Математична епідеміологія

З метою моделювання розповсюдження епідемічних захворювань у населення, необхідні певні припущення щодо конкретних захворювань та популяцій, а саме:

– Хвороби заразні і передаються через контакт.

– Хвороба може призвести до летального результату

– У період епідемії можуть народжуватися нові особини, при чому хвороба може бути передана від матері до дитини.

– Хвороба може мати інкубаційний період, протягом якого збуднювачі переносять і поширюють хворобу, але не мають видимих симптомів.

– Підхоплення та перенесення хвороби може призвести до короткочасного або тривалого здобутого імунітету. Залежно від випадку, одужані особини можуть захворіти знову

– Сторонні втручання, такі як ліки, вакцинація, карантин та соціальне дистанціювання можуть змінити схему розвитку розмноження хвороби.

Розглянемо приклад, який є основною моделлю, яка буде пізніше вдосконалена до схеми поширення вірусу COVIC-19.

В найпростішому випадку населення поділяють на дві групи: сприйнятливих до захворювання осіб (позначають, як S — від англ. susceptible), та осіб інфікованих патогеном (позначають, як I — від англ. infected). Таким чином, патогенна взаємодія базується на феноменологічних припущеннях, на основі яких побудована математична модель. Для дослідження цих моделей використовують звичайні диференціальні рівняння (які є детермінованими), проте можна розглядати й стохастичні моделі (наприклад, модель Ґіллеспі). В подальшому використанні цих моделей, також описується кількість осіб, які одужали (позначають, як R — від англ. recovered).

1) Базовою моделлю, що використовується для моделювання епідемічних захворювань без пожиттєвого імунітету називається моделлю SIR (модель сприйнятливого-зараженого-одужаного) [10], [11], [12].

У цій моделі загальна чисельністьнаселення в N осіб, що піддавалися епідемічному захворюванню на кожен момент часу t ділиться на три групи: сприйнятлива частка $s(t)$, частка інфікованих $i(t)$, і частка одужавших $r(t)$. Відповідно, система є замкненою.

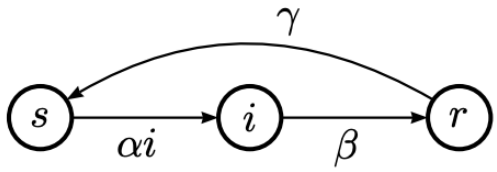


Рис. 1: Базова модель сприйнятливого-зараженого-одужаного (SIR)

Маємо:

$$s(t) + i(t) + r(t) = 1. \tag{1.7}$$

Модель розповсюдження хвороби на рис. 1 еквівалентна наступному набору диференціальних рівнянь:

$$\begin{aligned}
\frac{ds(t)}{dt} &= -\alpha s(t)i(t) + \gamma r(t) \\
\frac{di(t)}{dt} &= \alpha s(t)i(t) - \beta i(t) \\
\frac{dr(t)}{dt} &= \beta i(t) - \gamma r(t).
\end{aligned} \tag{1.8}$$

Відповідно, перехід від сприйнятливої групи до зараженої групи відбувається зі швидкістю, пропорційною популяції інфікованих та сприйнятливих груп з параметром α. У той же час передбачається, що заражені особини одужують з постійною швидкістю β. Нарешті, не передбачається, що хвороба призводить до довічного імунітету в осіб, які одужали, і вони знову повертаються до сприйнятливої групи з фіксованою швидкістю γ. З (1.8) очевидно, що:

$$\frac{ds(t)}{dt} + \frac{di(t)}{dt} + \frac{dr(t)}{dt} = 0. \tag{1.9}$$

Дана система вважається закритою (не розглядається жодного народження та смерті).

Припускаючи початкові умови для кожної групи, набір нелінійних рівняннь (1.8) можна (чисельно) вирішити, щоб знайти еволюцію популяції кожної групи з часом. Чисельним розв'язком базової (не фатальної) моделі SIR є криві на рис. 2 та рис. 3, із довічним імунітетом та без нього. Дисктретизація часу диференціальних рівнянь береться за $\Delta = 0,1$ дня. Варто зауважити, що спалах хвороби, яка не викликає довічного імунітету (наприклад, типовий грип) може призвести до постійної швидкості захворювання, після перехідного періоду. У випадку поширених епідемічних хвороб, стратегів охорони здоров'я цікавлять нахили (кутові коефіцієнти) $s(t), i(t)$ та $r(t)$, а не загальна кількість заражених особин (як це зараз відбувається для коронавірусу COVID-19). Подовження розповсюдження захворювання забезпечує краще управління ресурсами охорони здоров'я такими як госпіталізація, медикаменти, медичний персонал тощо.

Великий інтерес представляюсь фіксовані точки моделі SIR (де $\dot{s}(t) = \dot{i}(t) = \dot{r}(t) = 0$) [15]. Прирівнючи ліві сторони (1.8) до нуля, можна алгебраїчно показати, що якщо $\alpha, \gamma \neq 0$ (безімунний випадок), модель SIR має лише дві нерухомі точки:

$$\begin{aligned}
&(s^*(t), i^*(t), r^*(t)) = (1, 0, 0), \\
&(s^*(t), i^*(t), r^*(t)) = (\frac{\beta}{\alpha}, I_0, \frac{\beta}{\gamma} I_0),
\end{aligned} \tag{1.10}$$

де $I_0 \triangleq \frac{\gamma(\alpha - \beta)}{\alpha(\gamma + \beta)}$. Перша нерухома точка відповідає відсутності заражених випадків, а друга - стійкому захворюванню серед населення, як показано на рис. 3.

Ця ситуація досягається лише у випадку, коли $\beta < \alpha$, тобто коли рівень зараження перевищує рівень одужання.

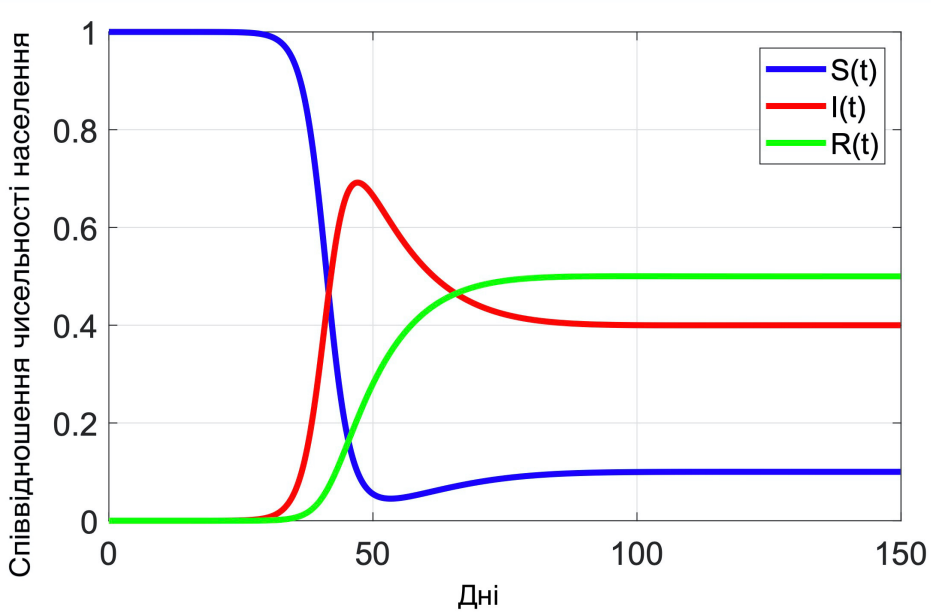


Рис. 2: Базова модель SIR з імунітетом

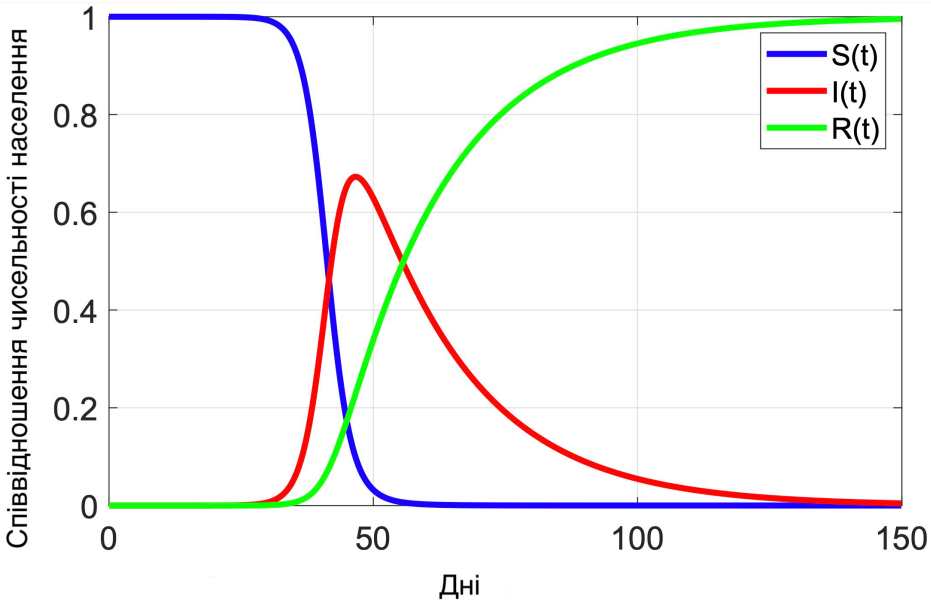


Рис. 3: Базова модель SIR без імунітету

Ми також можемо перевірити, чи є фіксовані точки стійкими.[16] Для цього можна використовувати різні методи. Можливо, найбільш підходящий підхід заснований на теорії збурень: до фіксованих точок системи додаються невеликі збурення та перевіряється чи компенсуються збурення динамікою системи коли вектор стану збігаєтьсяд до фіксованої точки. Відповідно, перша фіксована точка в (1.10) може бути збурена до:

$$(s(t), i(t), r(t)) = (1 - \epsilon, \epsilon, 0), \tag{1.11}$$

де $0 < \epsilon \ll 1$ - невелике збурення (наприклад, еквівалентно одиничному випадку спалаху хвороби у великій популяції). Тепер замінивши збурену точку в (1.8) і нехтуючи другим і членами вищого порядку, що містять ϵ, отримуємо:

$$\begin{aligned} \frac{ds(t)}{dt} &= \alpha(1-\epsilon)\epsilon \approx -\alpha\epsilon < 0 \\ \frac{di(t)}{dt} &= \alpha(1-\epsilon)\epsilon - \beta\epsilon \approx (\alpha-\beta)\epsilon \\ \frac{dr(t)}{dt} &= \beta\epsilon > 0. \end{aligned} \tag{1.12}$$

Як результат, перша фіксована точка нестабільна, оскільки через знак похідних збуреної системи, динаміка системи відводить вектор стану від фіксованої точки (оскільки популяція сприйнятливої групи має негативну похідну). Однак, залежно від того, правильності нерівності $\alpha > \beta$, спалах може призвести або не призвести до збільшення кількості заражених. Простіше кажучи, якщо рівень зараження більше ніж швидкість одужання $\alpha > \beta$ хвороба буде поширюватися; але якщо швидкість одужання швидша за коефіцієнт розповсюдження інфекції $\alpha < \beta$, то відсоток зараженого населення залишиться близьким до нуля. У будь-якому випадку для несмертельної хвороби, проти якої не виробляється імунітет, всі особи, які інфікуються та одужують через деякий час і переходять до групи одужавших знову повертаються до сприйнятливої групи зі швидкістю γ. Варто зауважити, що модель SIR з ненульовою часткою інфікованої популяції в стаціонарному стані, вказує на те, що існує постійний потік між групами, тобто люди постійно хворіють, одужують і знову стають сприйнятливими до хвороби.

Збурення другої фіксованої точки призводить до

$$\begin{aligned} \frac{ds(t)}{dt} &= \alpha(\frac{\beta}{\alpha} - \epsilon)(I_0 + \epsilon) \approx \epsilon(\alpha I_0 - \beta) \\ \frac{di(t)}{dt} &= \alpha(1-\epsilon)\epsilon - \beta\epsilon \approx -\epsilon(\alpha I_0 - \beta) \\ \frac{dr(t)}{dt} &= \beta\epsilon > 0. \end{aligned} \tag{1.13}$$

У цьому випадку, залежно від знаку нерівності $(\alpha I_0 - \beta) > 0$, фіксована точка може бути або стійкою, або нестійкою. Також можна показати, що під час спалаху моделі SIR ($s(t) \approx 1$), кількість заражених випадків підпорядковується наступній експоненційній моделі:

$$i(t) \approx i(0)e^{(\alpha-\beta)t}. \tag{1.14}$$

2) Фатальна модель SIR [9]: фатальна версія моделі SIR із показником народжуваності μ^*, рівнем смертності сприйнятливої (μ_s), інфікованої (μ_i) та одужаної

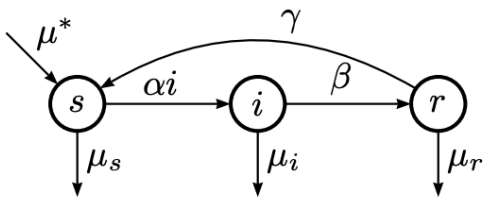


Рис. 4: Модель (SIR) з показниками народжуваності і смертності

(μ_r) групи (рис. 4). Ця система вже не є закритою, і її рівняння стану можна записати так:

$$\begin{aligned}
\frac{ds(t)}{dt} &= \gamma r(t) - \alpha s(t)i(t) - \mu_s s(t) + \mu^* \\
\frac{di(t)}{dt} &= \alpha s(t)i(t) - \beta i(t) - \mu_i i(t) \\
\frac{dr(t)}{dt} &= \beta i(t) - \gamma r(t) - \mu_r r(t).
\end{aligned} \tag{1.15}$$

Подальше дослідження цієї системи наведено у наступному параграфі.

1.3 Епідемічна модель COVID-19

Для багатьох інфекційних захворювань характерний інкубаційний період між зараженням та проявом клінічних симптомів. Суб'єкти, схильні до зараження, набагато більше небезпечні для громадськості порівняно з тими, у кого проявляються клінічні симптоми. Стан стає все більше і більше небезпечний, із збільшенням рівня інкубації. Добре відомий випадок - вірус ВІЛ у стадії клінічної латентності. Досвід COVID-19 показує, що двотижневий інкубаційний період може поширити вірус по всьому світу і майже на будь-якому рівні суспільства. Згідно теорії шести рукостискань, кожна людина опосередковано знайома з будь-яким іншим жителем планети через недовгий ланцюжок спільних знайомих. У середньому цей ланцюжок складається з шести чоловік. З цієї причини додається додаткова група між двома етапами моделі SIR: сприятливістю та інфікованістю. Дана група буде пояснювати безсимптомне протікання інфекції в особин. Більше того, оскільки ми також зацікавлені в мінімізації смертності від захворювання, додамо також групу померлого від інфекції населення. Отже, змінними моделі є:

1. $s(t)$: сприйнятлива частка населення (кількість осіб, яким загрожує зараження, поділених на загальну кількість населення).

2. $e(t)$: частка населення, що зазнала впливу вірусу, частка вражених (кількість осіб, що зазнали впливу вірусу, але без симптомів, поділена на загальну

чисельність населення).

3. $i(t)$: частка зараженого населення (кількість заражених особини з симптомами, поділена на загальну кількість населення).

4. $r(t)$: частка осіб, що одужали (кількість осіб, що одужали, поділена на загальну популяцію).

5. $p(t)$: кількість померлих (кількість особин, які померли через хворобу, поділена на загальну популяцію).

Вважаючи, що

$$s(t) + e(t) + i(t) + r(t) + p(t) = 1 \tag{1.16}$$

маємо наступну модель [9]:

$$\begin{aligned}
\frac{ds(t)}{dt} &= -\alpha_e s(t)e(t) - \alpha_i s(t)i(t) + \gamma r(t) \\
\frac{de(t)}{dt} &= \alpha_e s(t)e(t) + \alpha_i s(t)i(t) - \kappa e(t) - \rho e(t) \\
\frac{di(t)}{dt} &= \kappa e(t) - \beta i(t) - \mu i(t) \\
\frac{dr(t)}{dt} &= \beta i(t) + \rho e(t) - \gamma r(t) \\
\frac{dp(t)}{dt} &= \mu i(t).
\end{aligned} \tag{1.17}$$

Також вона зображена на рис 5:

У (1.17), подібно до класичної моделі SIR, нелінійні доданки, включаючи $s(t), e(t)$ та $s(t)i(t)$, можна трактувати наступним чином: швидкість впливу вірусу пропорційна чисельності населення як сприйнятливих, так і вражених інфекцією/заражених.

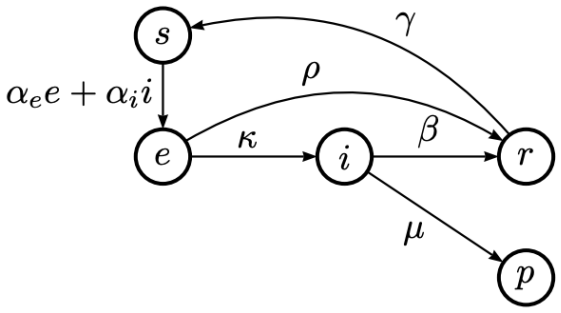


Рис. 5: Базова модель сприйнятливого-зараженого-одужаного (SIR)

Варто зауважити, що через те, що система (1.16) замкнена, з'являється надлишковий степінь свободи. Таким чином можна зменшити порядок моделі, замінивши $s(t) = 1-e(t)-i(t)-r(t)-p(t)$.

Це спрощує полігамну модель наступним чином:

$$
\begin{aligned}
\frac{de(t)}{dt} &= [1 - e(t) - i(t) - r(t) - p(t)][\alpha_e e(t) + \alpha_i i(t)] - \kappa e(t) - \rho e(t) \\
\frac{di(t)}{dt} &= \kappa e(t) - \beta i(t) - \mu i(t) \\
\frac{dr(t)}{dt} &= \beta i(t) + \rho e(t) - \gamma r(t) \\
\frac{dp(t)}{dt} &= \mu i(t).
\end{aligned} \tag{1.18}
$$

1.3.1 Виміри моделі

Серед змінних стану запропонованої моделі всі крім $e(t)$ можна безпосередньо виміряти (з потенційними помилками). Виміри можна записати у матричній формі наступним чином:

$$
\begin{bmatrix} I(t) \\ R(t) \\ P(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e(t) \\ i(t) \\ r(t) \\ p(t) \end{bmatrix} + \begin{bmatrix} v_i(t) \\ v_r(t) \\ v_p(t) \end{bmatrix}, \tag{1.19}
$$

де $I(t)-$ частка зареєстрованих інфекцій, $R(t)-$ частка повідомлень про одужання (як після симптомного так і безсимтомного протікання хвороби), $P(t)-$ частка повідомлень про кількість загиблих, і $v(t) = [v_i(t), v_r(t), v_p(t)]^T$ - це шум вимірювання. Очевидні джерела шумів вимірювання включають: недоступну інформацію щодо точної кількості населення, навмисні на ненавмисні помилкові повідомлення, неправильно класифіковані причини смерті (особливо для людей похилого віку або осіб, які мають проблеми зі здоров'ям), а також незначні випадки, які можуть бути невідомими або неправильно класифікованими системою охорони здоров'я. Рівняння (1.19) можна записати в більш компактній формі наступним чином:

$$
\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{v}(t), \tag{1.20}
$$

де $\mathbf{x}(t) = [e(t), i(t), r(t), p(t)]^T$ - зведений вектор стану.

У наведеній вище моделі вимірювань передбачається що $R(t)-$ загальна частка одужавших, які хворіли і симптомно і безсимптомно, припускаючи, що безсимптомні одужання можна виміряти за допомогою (випадкових або систематичних) публічних тестів серед населення, наприклад, аналізи антитіл, які були проведений деякими країнами під час спалаху COVID-19.

1.3.2 Припущення та параметри моделі

В запорпонованій моделі лежать наступні припущення:

1. Змінні моделі неперервні по часу

2. Народження та природні випадки смерті нехтуються. Отже, інші параметри, що призводять до змін у популяції, не враховується. Також допускається нехтування народжуваністю. За сучасними висновками, немовлята не сприйнятливі до цього вірусу і, наскільки відомо, немає вродженої предачі вірусу від матері до дитини.

3. Поточна модель не розрізняє чоловіків та жінок; хоча нинішні глобальні дані свідчать про те, що чоловіки більш схильні до вірусу, ніж жінки.

4. Вікові діапазони не враховуються; хоча ми знаємо що особи старшого віку вразливіші до вірусу, і піраміди розподілу віку населення різні для усіх країн.

5. Не розглядається можливість вакцинації

6. Геополітичні фактори, такі як відстань, кордони країн та континентальні відмінності також ігноруються. Враховуючи, що різні країни прийняли індивідуальні контрзаходи проти розповсюдження вірусу, параметри моделі відповідають даним на рівні країни.

Сформувавши модель, ми тепер пояснимо її параметри та їх взаємозв'язок із реальними факторами та клінічними протоколами.

– κ: швидкість, з якою симптоми проявляються у відкритих випадках, внаслідок чого відбувається перехід від враженого до зараженого населення.

  Тут є багато спрощень: віковий діапазон, особливості імунної системи кожного організму, тяжкість перенесення вірусу та багато інших факторів які нехтуються. Але це дає уявлення про те, як можна налаштувати параметри на практиці. Слід додати, що політики широкого скрінінгу, прийняті певними країнами є зовнішніми факторами, які можуть суттєво прискорити виявлення заражених випадків. У цьому випадку, скринінг є фактором, який збільшує κ.

– α_i: показник зараження між інфікованою та сприйнятливою популяціями, який пов'язаний із заразністю вірусу та соціальними факторами, такими як особиста гігієна, густота населення, та рівень взаємодії людей. Для того, щоб знайти діапазон цього параметра, ми можемо розпочати з дослідження факторів зараження більш відомих вірусів, таких як простуда та грип, які більшою чи меншою мірою зазнають впливу тих самих факторів поширення.

– α_e: показник зараження між враженою та сприйнятливою популяціями. Цей параметр є набаго логічнішим за α_i, оскільки в звичайних умовах (до карантину) люди рідко уникають контакту з безсимптомно хворіючими особами; так само і сам індивід не уникає взаємодії з іншими особинами.

– γ: швидкість повторного зараження або швидкість повернення з групи одужавшиз осіб до сприйнятливої групи. Це трапляється у випадках, коли організм не виробляє довічний імунітет після одужання або сам вірус починає мутувати з часом. Цей параметр є оберненим достепеня імунітету вірусу. Наразі ще рано коментувати характеристики імунітету до коронавірусу COVID-19. Хоча було виявлено, що принаймні один раз реінфекція може трапитись після одужання, короткочасний набутий імунітет становить до чотирьох місяців.

– β: Швидкість одужання заражених випадків. Розглядаючи четверте рівнянням у (1.17) ми можемо позначити кількість госпіталізованих (включаючи осіб, які знаходяться під контролем в будь-якій формі, наприклад, під домашнім наглядом) за r_h, в результаті чого $r_h(t+\Delta) - r_h(t) \approx \Delta\beta i(t)$, де Δ - одиниця часу наближення (наприклад, 1 день). Тому параметр β можна апроксимувати діленням добової кількості одужавших на загальну кількість заражених випадків в той же день. У реальному світі, крім здатності інфікованого суб'єкта протидіяти вірусу, цей параметр залежить від інфраструктури охорони здоров'я країни (закладів госпіталізації, наявності ліків, кількість відділень інтенсивної терапії тощо).

– ρ: коефіцієнт одужання відкритих випадків (випадків, які піддалися впливу, але одужують без будь-яких симптомів). Цей параметр не піддається безпосередньому вимірюванню з чистих спостережень і вимагає лабораторних експериментів. Однак ми логічно очікуємо, що порядок цього параметра однаковий або більший за параметр β (коефіцієнт одужання зараженої популяції з симптомами).

– μ: Рівень смертності від заражених випадків. Наближаючи останнє рівняння у (1.17) до $p(t+\Delta)-p(t) \approx \Delta\mu i(t)$, де $\Delta-$ одиниця часу наближення (наприклад 1 день), параметр μ можна апроксимувати діленням щоденної кількості загиблих на загальну кількість інфікованих випадків у той самий день. Як і у випадку з β, смертність від самого вірусу, набуття імунітету осіб та медична інфраструктура є важливими факторами, що впливають на параметр.

– e_0: Початкова кількість людей, яка зазнала впливу вірусом.

Вивчаючи вищезазначені фактори, ми можемо побачити, що єдині параметри моделі, які можуть бути змінені в короткостроковій перспективі (до розробки дов-

гострокових рішень таких як вакцинація, ліки, поліпшення госпіталізації закладів тощо), це у зменшенні рівня зараження шляхом мінімізації людських контактів (соціальне дистанціювання), або застосування публічного скрінінгу. Це дві політики, які були введені у світі.

1.3.3 Аналіз фіксованих точок

Як і для базової моделі SIR, представленої в розділі 1.2 - 1) фіксовану точку (точки) моделі можна шукати, прирівнюючи ліві частини рівняння (1.17) до нуля. Припускаючи, що всі параметри моделі є ненульовими, єдиною фіксованою точкою є випадок відсутності захворювання ($i(t) = e(t) = r(t) = 0$):

$$(s^*(t), e^*(t), i^*(t), r^*(t), p^*(t)) = (1 - p_0, 0, 0, 0, p_0), \tag{1.21}$$

де $0 \leq p_0 \leq 1-$ загальна частка смертості у стаціонарному стані. Стабільність цієї нерухомої точки можна дослідити шляхом збурення нерухомої точки з незначним збуренням ϵ (що може відповідати одному новому відкритому випадку у реальному світі):

$$(s(t), e(t), i(t), r(t), p(t)) = (1 - p_0 - \epsilon, \epsilon, 0, 0, p_0), \tag{1.22}$$

Підставивши цю точку в рівняння динаміки стану (1.17), маємо:

$$\begin{aligned}
\frac{ds(t)}{dt} &= -\alpha_e(1 - p_0 - \epsilon)\epsilon \approx -\alpha_e(1 - p_0)\epsilon < 0 \\
\frac{de(t)}{dt} &= \alpha_e(1 - p_0 - \epsilon)\epsilon - \kappa\epsilon - \rho\epsilon \approx (\alpha_e - \alpha_e p_0 - \kappa - \rho)\epsilon \\
\frac{di(t)}{dt} &= \kappa\epsilon > 0 \\
\frac{dr(t)}{dt} &= \rho\epsilon > 0 \\
\frac{dp(t)}{dt} &= 0,
\end{aligned} \tag{1.23}$$

Дана сукупність рівнянь нестабільна, тобто динаміка системи відводить її від фіксованої точки у напрямку зменшення здорових випадків, що призводить до подальшого зараження.

1.3.4 Аналіз моделі під час спалаху

Дослідимо модель під час початкового спалаху захворювання епідемії, коли кількість заражень все ще значно менша, ніж загальна чисельність населення. Наприклад, припустимо, що країна має 100000 відкритих або заражених випадків,

що насправді багато для будь-якої країни, оскільки це далеко за межами доступної кількості ліжок інтенсивної терапії навіть у найрозвиненіших країнах. Але для країни зі 100 мільйонами населення такий рівень зараження становить лише $0,1\%$ від загальної кількості населення. Тому під час первинних фаз поширення захворювання модель можна спростити, припустивши, що сприйнятлива до ураження хворобою кількість населення майже постійна ($s(t) \approx 1$) і $\frac{ds(t)}{dt} \approx 0$, незалежно від інших параметрів моделі. Це припущення практично означає, що загальна чисельність населення не є важливою під час спалаху епідемії (із низьким відсотком зараження), в результаті чого маємо наступні висновки:

Результат 1. За низьких відсотків зараження, ефективність політики яка контролює епідемію країни чи регіонів не повинна оцінюватися нормалізацією зараження / виздоровлення / смертності до загальної чисельності населення; треба порівнювати абсолютні значення. Іншими словами, поширення хвороби не залежить від загальної чисельності населення.

Згідно з цим припущенням (1.17) спрощується до лінійної системи рівнянь:

$$\begin{bmatrix} \frac{de(t)}{dt} \\ \frac{di(t)}{dt} \\ \frac{dr(t)}{dt} \\ \frac{dp(t)}{dt} \end{bmatrix} \approx \begin{bmatrix} \alpha_e - \kappa - \rho & \alpha_i & 0 & 0 \\ \kappa & -\beta - \mu & 0 & 0 \\ \rho & \beta & -\gamma & 0 \\ 0 & \mu & 0 & 0 \end{bmatrix} \begin{bmatrix} e(t) \\ i(t) \\ r(t) \\ p(t) \end{bmatrix}. \tag{1.24}$$

Позначимо $\mathbf{x}(t) = [e(t), i(t), r(t), p(t)]^T$. Тоді (1.24) можна переписати у матричній формі:

$$\frac{d}{dt}\mathbf{x(t)} = \mathbf{A}\mathbf{x}(t), \tag{1.25}$$

де $\mathbf{A}-$ матриця стану 4×4 у правій частині (1.24). Рівняння (1.25) можна розв'язати для довільної початкової умови, наприклад $x(0) = (e_0, 0, 0, 0)$. Характеристична функція цієї лінійної системи буде мати вигляд:

$$|\lambda\mathbf{I} - \mathbf{A}| = \lambda(\lambda + \gamma)[\lambda^2 + (\beta + \mu - \delta)\lambda - \delta(\beta + \mu) - \kappa\alpha_i] = 0, \tag{1.26}$$

де $\delta \stackrel{\Delta}{=} (\alpha_e - \kappa - \rho)$.

Маємо наступні власні значення системи:

$$\lambda_1 = \frac{\delta - \beta - \mu + \sqrt{(\delta + \beta + \mu)^2 + 4\kappa\alpha_i}}{2}, \tag{1.27}$$

$$\lambda_2 = \frac{\delta - \beta - \mu - \sqrt{(\delta + \beta + \mu)^2 + 4\kappa\alpha_i}}{2}, \tag{1.28}$$

$$\lambda_3 = 0, \quad \lambda_4 = -\gamma. \tag{1.29}$$

Усі вони є дійснозначими. Більше того, $\lambda_1 > \delta > \lambda_2$.

Знайдемо власні вектори, що відповідають кожному власному значенню:

$$\mathbf{v}_1 = k_1 \left[1, \frac{\lambda_1 - \delta}{\alpha_i}, \frac{\rho\alpha_i + \beta(\lambda_1 - \delta)}{\alpha_i(\lambda_1 + \gamma)}, \frac{\mu(\lambda_1 - \delta)}{\alpha_i\lambda_1}\right]^T, \tag{1.30}$$

$$\mathbf{v}_2 = k_2 \left[1, \frac{\lambda_2 - \delta}{\alpha_i}, \frac{\rho\alpha_i + \beta(\lambda_2 - \delta)}{\alpha_i(\lambda_2 + \gamma)}, \frac{\mu(\lambda_2 - \delta)}{\alpha_i\lambda_2}\right]^T, \tag{1.31}$$

$$\mathbf{v}_3 = [0, 0, 0, k_3]^T, \quad \mathbf{v}_4 = [0, 0, k_4, 0]^T. \tag{1.32}$$

де k_1, k_2, k_3 і k_4- довільні константи. Загальну форму розв'язку змінних можна знайти шляхом підсумовування показникових виразів із зазначеним вище власними значеннями і власними векторами:

$$\mathbf{x}(t) = \sum_{k=1}^{4} a_k e^{\lambda_k t}\mathbf{v}_k. \tag{1.33}$$

Зокрема, після деяких алгебраїчних спрощень, ми можемо порахувати інфіковану та сприйнятливу до вірусу популяції наступним чином:

$$\begin{aligned} i(t) &= \frac{e_0(\lambda_1 - \delta)(\delta - \lambda_2)}{\alpha_i(\lambda_1 - \lambda_2)}[e^{\lambda_1 t} - e^{\lambda_2 t}], \\ e(t) &= \frac{e_0}{\lambda_1 - \lambda_2}\left[(\lambda_1 - \delta)e^{\lambda_2 t} + (\delta - \lambda_2)e^{\lambda_1 t}\right]. \end{aligned} \tag{1.34}$$

З останнього рівняння (1.17) видно, що люди будуть гинути до тих пір поки кількість інфікованих людей не буде дорівнювати нулю: $i(t) = 0$. Це також випливає з (1.34): оскільки λ_1 є домінуючим власним значенням, стаціонарна поведінка та відповідь на питання розбігаються чи збігаються до нуля $i(t)$ та $e(t)$, залежить від знака λ_1. Необхідною і достатньою умовами стабільності лінеаризованої системи (зупинка смертності) є $\lambda_1 < 0$, що спрощується до $\kappa\alpha_i + \delta(\beta + \mu) < 0$, або:

$$\kappa\alpha_i < (\kappa + \rho - \alpha_e)(\beta + \mu). \tag{1.35}$$

Достатньою умовою, яка гарантує цю властивість, є $\alpha_i = \alpha_e = 0$. З умови $\alpha_i = 0$ випливає, що сприйнятлива група уникає контакту з зараженими. Однак

другу умову ($\alpha_e = 0$) важко виконати в реальному світі, оскільки у групи, яка зазнала впливу, немає симптомів. Ось чому для забезпечення необхідного соціального дистанціювання треба $\alpha_e \approx 0$ і дозволити всім об'єктам, що потрапляють на оголення, перейти до інфікована група без зараження нових особин, після чого безсимптомну групу можна вважати вільною від захворювання. Інший практичний випадок - коли $\alpha_i \approx 0$ (здорові люди уникати контакту із зараженим) та $\kappa + \rho > \alpha_e$ (коефіцієнт відновлення оголених або поява їх симптомів швидше, ніж швидкість нових експозицій). Ця умова є здійснюється соціальним дистанціюванням і блокуванням (ізоляція навіть безсимптомні випадки протягом певного періоду). Однак, якщо жодна з перерахованих вище умов не виконується і $\lambda_1 > 0$, кількість підданих та інфікованих випадків збільшується експоненційно зі швидкістю λ_1. У цьому випадку з фіксованою системою параметрів, рівень зараження зростає в геометричній прогресії аж до позначки при якому лінійне наближення вже не виконується. Це практично перекладається на:

Результат 2. Під час експоненціального спалаху епідемії ($\lambda_1 > 0$), система нестабільна і без застосування тимчасових блокувань, соціального дистанціювання і карантину заражених випадків (в результаті чого змінюються параметри моделі), експоненціальне зростання кількості заражених суб'єктів призведе до того, що значний відсоток населення буде інфікований.

Можна показати, що для епідемічної моделі (1.17) репродукційне число дорівнює:

$$\mathcal{R}_0 = \frac{\alpha_e(\beta + \mu) + \kappa\alpha_i}{(\kappa + \rho)(\beta + \mu)}. \tag{1.36}$$

Результат 3. Через введення контрзаходів змінюються власні значення моделі. λ_1 (домінуюче власне значення лінеаризованої динамічної моделі) - єдиний параметр, за відстеженням якого можна зробити оцінку того, наскільки ефективними є контрзаходи, такі як соціальне дистанціювання та карантин.

Також цікаво спостерігати з (1.34), що кількість населення з різних груп моделі лише лінійно пропорційна початковому розміру враженої вірусом популяції e_0. Тому для великої популяції (на рівні населеного міста чи країни) початковий розмір зараженого населення не настільки важливий, як інші параметри моделі, що впливають на експоненційну поведінку моделі (наприклад, частота та степінь соціальних контактів).

Таким чином:

Результат 4. Початкова кількість зараженого населення не є найбільш критичним параметром для управління епідемією. У регіонах з меншою початковою кількістю заражених/незахищених випадків в кінцевому підсумку може бути більша кількість заражених та кількість загиблих від зараження, що визначається

такими факторами як степінь контакту людей між собою та особиста гігієна.

Існує ще одна цікава властивість - співвідношення між кількістю заражених (що вимірюється в реальному світі) та кількістю вражених вірусом (що не можна безпосередньо виміряти). З (1.34) ми можемо знайти:

$$\frac{i(t)}{e(t)} = \frac{e^{\widetilde{\lambda}_1 t} - e^{-\widetilde{\lambda}_2 t}}{\alpha_i \left[\widetilde{\lambda}_1^{-1} e^{\widetilde{\lambda}_1 t} + \widetilde{\lambda}_2^{-1} e^{-\widetilde{\lambda}_2 t}\right]}, \tag{1.37}$$

де $\widetilde{\lambda}_1 \triangleq \lambda_1 - \delta$ та $\widetilde{\lambda}_2 \triangleq \delta - \lambda_2$ додатні. Отже, коли вирази, що містять $e^{-\widetilde{\lambda}_2 t}$ (спадна експонента), зникають та епідемічна модель все ще знаходиться у лінійній фазі ($i(t) \ll s(t)$ або $s(t) \approx 1$) співвідношення може бути наближено наступним чином:

$$\frac{i(t)}{e(t)} \to \frac{\widetilde{\lambda}_1}{\alpha_i}, \tag{1.38}$$

для $t \gg \widetilde{\lambda}_2^{-1}$ і $i(t) \ll s(t)$. Це дає наступний практичний результат:

Результат 5. Під час первинних фаз спалаху епідемії (коли кількість заражених випадків має експоненціальне зростання, але частка заражених особин від загальної чисельності ще невелика), кількість вражених вірусом може бути апроксимована до $e(t) \approx \alpha_i \widetilde{\lambda}_1^{-1} i(t)$, що дозволяє оцінити $e(t)$ з $i(t)$.

1.3.5 Повторювані хвилі епідемії

Піки зараженої групи населення та її потенціал повторення в часі важливі зі стратегічної точки зору. Ці пункти відповідають локальним або глобальним екстремумам $i(t)$, які математично відповідає тому, що $di(t)/dt = 0$ в (1.17), тобто де $i(t) = \kappa e(t)/(\beta + \mu)$. Можна показати що це призводить до зменшеного порядку набору нелінійної динаміки рівняння, які можна вирішити для решти змінних $[(s(t), e(t), r(t), p(t)]^T$. Шляхом подальшого моделювання можна показати, що заражене населення може мати декілька локальних піків з часом, із періодичною поведінкою, що доводить, що:

Результат 6. Епідемічна хвороба може повторюватись псевдоперіодично з часом (у пізніші сезони чи роки) і в довгостроковій перспективі перетворюється на стійке захворювання. Амплітуда та часовий розрив інфекції досягають піку залежно від параметрів моделі.

Така поведінка спостерігалася в попередніх пандеміях, такі як пандемічний грип 1918 року, відомий як іспанський грип, де спостерігалися три пандемічні хвилі інфекції протягом інтервалу в кілька місяців.

РОЗДІЛ 2. МОДЕЛІ ТА МЕТОДИ АНАЛІЗУ ЧАСОВИХ РЯДІВ ДЛЯ ПРОГНОЗУВАННЯ РОЗПОВСЮДЖЕННЯ ПАНДЕМІЇ

Моделі прогнозування часових рядів використовуються для прогнозування результатів на основі історичних даних. Ми адаптували модель ARIMA та Facebook Prophet (FBProphet) у нашому оціночному та прогнозному дослідженні. Огляд цих моделей разом з методом рухомого середнього наведено в даному розділі.

2.1 Метод рухомого середнього

Метод рухомого середнього (*англ.* Moving average) – це метод розрахунку для аналізу точок даних шляхом створення серії середніх значень різних підмножин певного набору даних [20]. Рухоме середнє може обчислюватись для довільних даних, однак, найчастіше його використовують в аналізі часових рядів для згладжування раптових коливань та підкреслення довготермінових трендів або циклів.

Простіше кажучи, у нашому випадку прогнозована кількість інфікованих людей на кожен день буде середнім множини раніше спостережуваних значень. Замість обчислення простого середнього, ми будемо використовувати техніку рухомого середнього, яка для кожного наступного прогнозу використовує набір останніх N значень. При цьому "застарілі значення"(ті, що були перед $N-$м) при обчисленні більше до уваги не беруться.

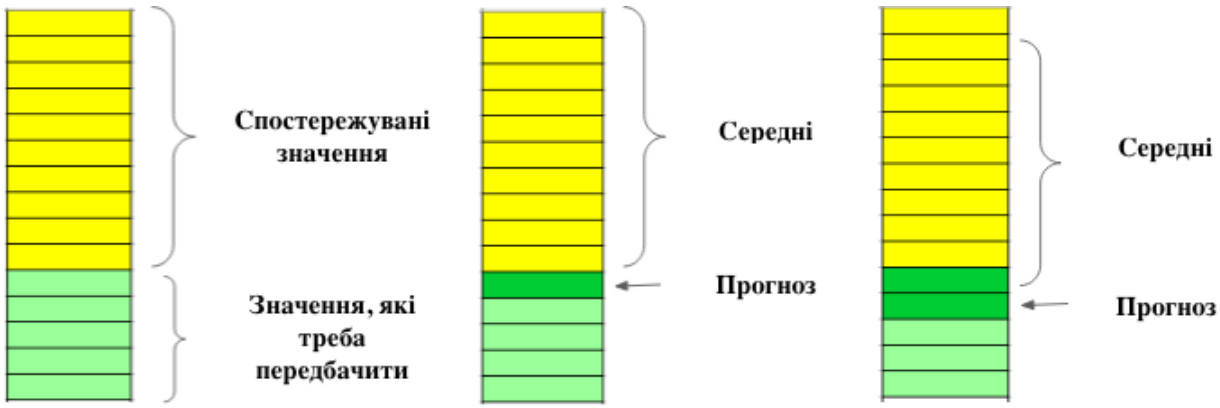


Рис. 6: Наочна демонстрація алгоритму moving average

Є декілька підходів цього методу. Будемо розгядати часовий ряд $\{p_t\}$, рухоме середнє $\{\overline{p}\}$ і вибірку ціни за попередні n днів.

1) **Просте рухоме середнє(*англ.* Simple Moving Average – SMA)** - це незважене середнє множини значень.

$$\overline{p}_{\text{SMA}} = \frac{p_t + p_{t-1} + \cdots + p_{t-(n-1)}}{n} = \frac{1}{n}\sum_{i=0}^{n-1} p_{t-i}.$$

При обчисленні послідовних значень $\overline{p}_{\text{SM}}$ нове значення надходить до суми, а найдавніше значення випадає, це означає, що повне підсумовування можна звести до наступної формули:

$$\overline{p}_{\text{SMA}} = \overline{p}_{\text{SMA,prev}} + \frac{1}{n}(p_t - p_{t-n}).$$

Вибраний період n залежить від виду прогнозування: короткостроковий або довгостроковий. Цей період також називають згладжуючим інтервалом. Чим довшим він є, тим більш плавним виходить графік функції. Чим коротшим є згладжуючий інтревал, тим метод швидше визначає нову тенденцію, а й одночасно робить більше помилкових коливань, і навпаки чим більше параметр n, тим повільніше визначається новий тренд, але надходить менше помилкових коливань.

2) Ковзне середнє може мати вагові коефіцієнти, наприклад, для посилення впливу новіших даних у порівнянні зі старішими.

Тому виділяють **зважене ковзне середнє(*англ.* Weighted moving average – WMA)**.

$$\overline{p}_{\text{WMA}} = \sum_{i=0}^{n-1} w_{t-i} \cdot p_{t-i},$$

де w_{t-i}–нормовані ваги, тобто $\sum_{i=0}^{n-1} w_{t-i} = 1$.

Найчастіше, в якості ваг використовують або 1 (для простого рухомого середнього – SMA), або формальні ряди, наприклад, арифметична прогресія (LWMA) або експоненційна функція (EMA). Але у якості вагового коефіцієнта можуть виступати і значення часового ряду.

3) **Лінійне зважене рухоме середнє (*англ.* Linear Weighted Moving Average – LWMA)**

– частковий випадок WMA; рухоме середнє, при обчисленні якого вага кожного члена вихідної функції, починаючи з меншого, дорівнює відповідному члену арифметичній прогресії. Тобто, при обчисленні LWMA для часового ряду, ми вважаємо останні значення більш значущими ніж попередні, причому функція значущості лінійно спадна.

Наприклад, для арифметичної прогресії з початковим значенням і кроком, рівним 1, формула обчислення рухомого середнього набуде вигляду:

$$\overline{p}_{\text{LWMA}} = \frac{n \cdot p_t + (n-1) \cdot p_{t-1} + \cdots + (n-i) \cdot p_{t-i} + \cdots + 2 \cdot p_{t-n+2} + 1 \cdot p_{t-n+1}}{n + (n-1) + \cdots + (n-i) + \cdots + 2 + 1} =$$
$$= \frac{2}{n \cdot (n+1)} \sum_{i=0}^{n-1} (n-i) \cdot p_{t-i}.$$

4) **Експоненційне зважене рухоме середнє (*англ.* Exponential Wei-**

ghted Moving Average – EMA) – різновид WMA, ваги якого зменшуються експоненційно і ніколи не дорівнюють нулю. Визначається наступною формулою:

$$\overline{p}_{\text{EMA,t}} = \begin{cases} p_1, & t = 1 \\ \alpha \cdot p_t + (1-\alpha) \cdot \overline{p}_{\text{EMA,(t-1)}}, & t > 1 \end{cases}, \tag{2.1}$$

де $\overline{p}_{\text{EMA,t}}$– значення EMA у точці t, α(згладжуюча константа) – коефіцієнт що характеризує швидкість зменшення ваг, приймає значення від 0 і до 1, чим менше його значення тим більше вплив попередніх значень на поточну величину середнього.

Перше значення EMA, зазвичай приймається рівним першому значенню у числовому ряді:

$\overline{p}_{\text{EMA},0} = p_0$

А значення коефіцієнту α може бути виражене через величину вікна усереднення: $\alpha = \dfrac{2}{n+1}$

Формула (2.1) також може бути виражена наступним чином:

$$\text{EMA}_{\text{сьогодні}} = \text{EMA}_{\text{вчора}} + \alpha \left[\text{p}_{\text{сьогодні}} - \text{EMA}_{\text{вчора}}\right].$$

Розгортання $\text{EMA}_{\text{вчора}}$ щоразу призводить до наступного ряду потужностей, показуючи як коефіцієнт зважування на кожній точці p_1, p_2 тощо зменшується експоненційно:

$$\text{EMA}_{\text{today}} = \alpha \left[p_1 + (1-\alpha)p_2 + (1-\alpha)^2 p_3 + (1-\alpha)^3 p_4 + \cdots\right].$$

Недоліки методу рухомого седнього:

1. Як правило, у WMA значне запізнення на вході в тренд і на виході з тренда, але менше ніж у SMA, так як завдяки надання пізнішим значенням більших ваг, WMA швидше реагує на зміни.

2. Для кожного періоду, який прогнозується необхідна історія минулих періодів.

3. MA нехтує складними зв'язками в даних

4. MA не реагує на коливання

2.2 Модель ARIMA

ARIMA (*англ.* AutoRegressive Integrated Moving Average.) – дуже популярний статистичний метод для прогнозування часових рядів. ARIMA має три компоненти - AR (авторегресивний терм), I (диференціальний терм) та MA (рухомий середній терм).

Нехай маємо часовий ряд X_t, де $t-$цілий індекс. Введемо наступний лаговий оператор:

$$L : Lx_t = x_{t-1}.$$

Розглянемо дві моделі - складові моделі ARIMA.

Авторегресійна (AR-) модель (англ. Autoregressive model)[13]— модель часових рядів, в якій значення часового ряду в даний момент лінійно залежать від попередніх значень цього ж ряду. Тоді дана модель визначається наступним чином:

$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t,$$

де c— константа, α_i— параметри моделі та ε_t— білий шум оцінені моделлю.

Використовуючи лаговий оператор, модель можна представити так:

$$X_t = c + \sum_{i=1}^{p} a_i L^i X_t + \varepsilon_t.$$

Модель MA(q) рухомого середнього порядку q (англ. Moving Average model) — модель часових рядів, яка вказує, що поточне значення ряду залишків ε_t залежить від q попередніх випадкових помилок:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

де μ - середнє значення ряду, $\theta_1, \theta_2, \ldots, \theta_q$— параметри моделі, а $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$— помилки білого шуму авторегресійної моделі відповідних попередніх значень часового ряду. Використовуючи лаговий оператор модель MA(q) можна визначити наступним чином:

$$X_t = \mu + \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t.$$

Тоді модель ARMA в загальному випадку можна записати наступним чином:

$$X_t = c + a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

або

$$\left(1 - \sum_{i=1}^{p} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t.$$

Інтегрувавши в модель множник нестаціонарної сезонності періоду d, $(1-L)^d$, ми отримаємо модель ARIMA(p,d,q):

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t.$$

З цього маємо такі параметри моделі:

- Терм AR означає минулі значення, що використовуються для прогнозування наступного значення. Термін AR визначається параметром 'p', який в свою чергу визначається за допомогою графіку PACF(часткової автокореляції).

- Терм MA використовується для визначення кількості помилок минулого прогнозу, які використовуються для прогнозування майбутніх значень. Його представляє параметр 'q', який розраховується по діаграмі ACF(повна функціця автокореляції).

- d − розмір вікна рухомого середнього, що також називається порядком різниці часового ряду, що необхідний для того, щоб зробити часовий ряд стаціонарним.

Перш ніж використовувати модель ARIMA, необхідно переконатися, що часовий ряд задовольняє наступним припущенням:

1. Ряд має бути стаціонарним, тобто середня величина та дисперсія не повинні змінюватися з часом.

2. Вхідні дані, повинні бути одновимірними рядами, оскільки ARIMA використовує минулі значення для прогнозування майбутніх значень.

Загальна реалізація моделі ARIMA

1. Завантаження набору даних.

2. Попередня обробка: залежно від набору даних будуть визначені етапи попередньої обробки. Наприклад, створення часових позначок, перетворення багатовимірних рядів на одновимірні тощо.

3. Перевірити числовий ряд на стаціонарність.

4. Визначити значення d: Для того, щоб зробити ряд стаціонарним необхідно певну кількість разів виконати наступну операцію: відняти попереднє значення від поточного значення. Мінімальна кількість таких операцій і є параметр d. Очевидно, що коли ряд вже стаціонарний, то d = 0.

5. Побудувати графіки ACF та PACF для визначенна вхідних параметрів моделі p і q.

6. Використовуючи значення даних та параметрів, які ми обчислили на попередніх кроках, будуємо модель ARIMA.

7. Прогнозування значень на валідаційному наборі даних.

8. Обчислення похибки моделі, наприклад RMSE.

Налаштування параметрів для ARIMA вимагає багато часу. Тож існує модель auto ARIMA, яка автоматично вибирає найкращу комбінацію (p, q, d) та забезпечує найменшу помилку, тобто кроки 3-5 можна пропустити.

2.3 Модель FBProphet

Тейлор та ін. [6] запропонували модель Facebook Prophet (FBProphet), яка використовує кілька нелінійних та лінійних методів як складові з часом як регресор. FBProphet розвинений і випущений як програмне забезпечення з відкритим кодом командою з обробки даних з Facebook. Модель ігнорує тимчасову залежність даних, і у наборі даних допускаються неправильні спостереження. Модель має багато переваг, зокрема, вона може вмістити декілька періодів сезонності, розрізняє події, пов'язані зі звичаями та традиційними святами (для нас це дуже важливо, оскільки шанс заразитися коронавірусом вищий при більшій концентрації людей, наприклад під час новорічних закупок у магазині чи освіщення паски у церкві). Дана модель є гнучкою, пропонуючи два варіанти для тренду: 1. кусково-задану лінійну модель, 2. модель зі зростанням, що насичується; і модель навчається дуже швидко. Крім вже відомих компонент, таких як тренд, сезонність та функція помилки, модель включає в себе ще одну складову - свята. Тому часовий ряд можна визначити таким рівнянням:

$$z(t) = T(t) + S(t) + H(t) + \epsilon(t). \tag{2.2}$$

$T(t)-$ це функція тренду, яка моделює неперіодичні зміни у значенні часового ряду, $S(t)$ представляє періодичні зміни (наприклад, щотижнева та річна сезонність), а $H(t)-$ вплив свят, які відбуваються під час потенційно нерегулярного графіку подій, як правило, більше одного дня. Компонента помилки $\epsilon(t)$ передбачає будь-які своєрідні зміни, які не враховуються моделлю.

РОЗДІЛ 3. ЗАСТОСУВАННЯ МЕТОДУ НЕЙРОННИХ МЕРЕЖ ДЛЯ КОРОТКОСТРОКОВОГО ПРОГНОЗУ РОЗВИТКУ ПАНДЕМІЇ

3.1 Актуальність використання нейронних мереж

Багато з використовуваних методів мають окремий недолік – лінійність, що означає можливість описувати процеси лише лінійною залежністю. Ще один недолік – використання лише одного стаціонарного рішення для системи лінійних рівнянь, що дає змогу отримувати результати, але з похибкою. Проте методи регресійного аналізу та відомі статистичні підходи не дають точних результатів, а часто неправильно прогнозують навіть тренди.

Інтелектуальні системи на основі штучних нейронних мереж дозволяють з успіхом вирішувати завдання прогнозування. Штучні нейромережі є нелінійним, непараметричним підходом до управління даними. Це дає змогу користувачу повністю використовувати наявну інформацію, яка і визначає структуру та параметри моделі без жодних наперед заданих умов чи обмежень. Нейронні мережі є нестандартними підходами у задачах прогнозування. Відмінність цього підходу від стандартних полягає в тому, що він дозволяє зробити систему самонавчальною. Завдяки можливості роботи з «зашумленими» даними система виходить гнучкою і зі 100% точністю, може оцінити майбутню кількіть інфікованих, в залежнсті від заходім, які впроваждуються. Штучні нейронні мережі дають багатообіцяючі альтернативні рішення. По-перше, вони мають здатність до генерування гнучких нелінійних функцій, близьких до будь-яких неперервних функцій із бажаною точністю. По-друге, нейронні мережі не потребують кількості змінних порівняно з такими лінійними моделями, як поліноміальна, сплайн чи розклади в ряд Тейлора тригонометричних функцій.

3.2 Процес побудови нейронної мережі

З точки зору нейромереж завдання короткострокового прогнозування зводиться до наближення функції багатьох змінних. Нейромережа використовується для відновлення наступних значень цієї функції по набору прикладів з історії тимчасового ряду. Доведено, що будь-яку дійсну функцію декількох змінних можна як завгодно точно наблизити за допомогою нейромережі.

Модель є чутливою до наявності у даних шуму. Сама нейромережа, являє собою багатошарову мережну структуру однотипних елементів - нейронів, з'єднаних між собою і згрупованих у шари.

Є вхідний шар, на нейрони якого подається інформація, а також вихідний, з якого береться результат. При проходженні по мережі вхідні сигнали посилюються або послаблюються, що визначається вагами міжнейронних зв'язків. Кажучи про нейромережу типу навчання з учителем, перед застосуванням на реальній вибірці, її необхідно навчити на прикладах - за допомогою корекції ваг міжнейронних зв'язків, тобто за відомими вхідними параметрами і результатами мережу змушують видавати відповідь, максимально близьку до правильної. Проблему оцінки зовнішніх умов, що постійно змінюються і відповідно ступені зараження залежно від тих чи інших параметрів нейромережа вирішує завдяки самому принципу роботи.

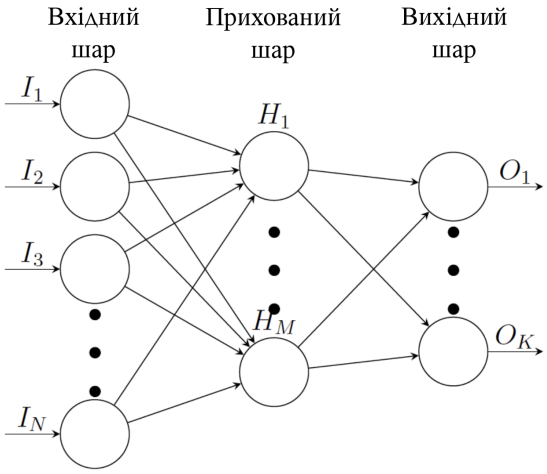


Рис. 7: Модель простої нейронної мережі.

Існує три основних етапи створення моделі нейромережі для вирішення задачі прогнозування часових рядів [1].

1. *Підготовка даних.* На цьому етапі вхідні дані кодуються: приводяться до єдиного масштабу, підвищується інформативність вхідних даних. Але виникає проблема «обмеженості» вікна. Тобто для передбачення значення ряду в наступний момент часу, потрібно N попередніх значень цього ряду подати на вхід нейромережі, при цьому більш ранні дані ніяк не впливають на прогноз. Якщо збільшити величину вікна, то знижується точність передбачення. Вирішенням цієї проблеми є використання інструментів аналізу часового ряду. Різні методи створення вибірки застосовуються щоразу перед початком тренування, наприклад, метод фільтрації. Інші методи базуються на визначенні значимості вхідних даних у процесі тренування, наприклад, за допомогою оцінювання чутливості вихідної інформації відносно вхідної доки не закінчиться потік неістотної інформації. Ці групи методів використовуються для вирішення складних проблем з численними вхідними даними та кількістю схованих шарів нейронів, так як вони допомагають спростити архітектуру нейронної мережі.

2. *Побудова моделі і навчання нейромережі.* Найпростіший варіант застосування штучних нейронних мереж - використання звичайного персептрона з одним, двома або трьома прихованими шарами. При цьому на вхід нейронної мережі зазвичай подається набір параметрів, на основі якого можна успішно прогнозувати. Виходом зазвичай є прогноз мережі на майбутній момент часу. При використанні багатошарових нейронних мереж необхідно також

пам'ятати про те, що потрібно акуратно робити нормування, і що для вихідного нейрона краще використовувати лінійну функцію активації. Узагальнюючі властивості від цього трохи погіршуються, але мережа буде набагато краще працювати з даними, що містять тренд. Ще однією часто використовуваною нейромережевою архітектурою є нейронна мережа із загальною регресією. Це сумісні архітектури в тому сенсі, що в працюючій системі прогнозування можна замінити роботу персептрона на мережу із загальною регресією. При цьому не потрібно проводити ніяких додаткових маніпуляцій з даними.

Завдання навчання - знайти параметри моделі (ваги мережі) шляхом мінімізації середньоквадратичної помилки виходу мережі. Процедура навчання нейромереж стандартна. Для побудови моделі прогнозування дані поділяються на три частини: для тренування (training), підтвердження (validation) та перевірки (testing). Тренувальна вибірка відповідно включає 70% зібраної інформації, причому вибірки для підтвердження та перевірки – 20% та 10% відповідно. Перша використовується для навчання, друга - для вибору оптимальної архітектури мережі і / або для вибору моменту зупинки навчання. Нарешті, третя, яка взагалі не використовувалася в навчанні, служить для контролю якості прогнозу навченої нейромережі. Після групування даних задається структура мережі з вибором кількості схованих шарів нейронів, вхідних нейронів та функції перетворення, що впливає на результативність функціонування нейронної мережі. Вибір параметрів нейронної мережі здійснюється на основі певних критеріїв: кількості вхідних та вихідних змінних, складності та структури наявної інформації, теоретичних знань та фактів про діяльність, яка прогнозується, тощо. Сигнали з вхідного шару нейронів передаються до прихованих шарів, які їх опрацьовують та перетворюють зазвичай за допомогою логістичної апроксимації на ступеневу або порогову функції. Потім одержаний сигнал передають до вихідного шару нейронів, де інформація обробляється знову для одержання фінального результату. Повнозв'язна нейронна мережа є такою, де кожен вхідний нейрон зв'язаний зі всіма схованими нейронами, і кожен зі схованих, у свою чергу, зв'язаний з наступними вихідними нейронами.

У літературі є свідчення поліпшення якості прогнозів за рахунок використання нейромереж із зворотними зв'язками. Такі мережі можуть мати локальну пам'ять, що зберігає інформацію про більш далеке минуле.

3. *Вибір функції помилки.* Для навчання нейромережі недостатньо сформувати навчальні набори входів - виходів. Необхідно також визначити помилку передбачення. Помилка мережі представляється у вигляді функції від синаптичних коефіцієнтів і мінімізується одним з градієнтних методів.

Як приклад, розглянемо багатошаровий персептрон, який може бути використано як модель для прогнозування кількості інфікованих. Цей вид нейронної мережі має дві основні переваги – простота у застосуванні та забезпечення необхідних узагальнюючих властивостей.

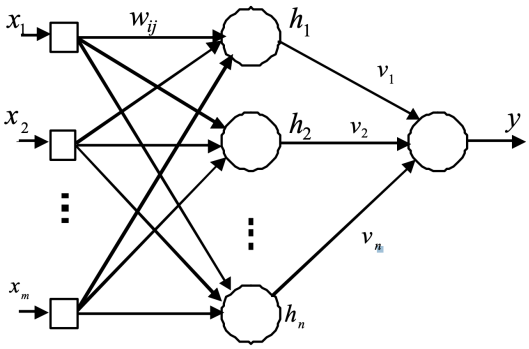


Рис. 8: Структура тришарового персептрона.

Вихідне значення тришарового персептрона (див. Рис 7) обчислюється за виразом:

$$y = F_3\left(\sum_{j=1}^{n} \nu_j h_j - b_3\right), \tag{3.1}$$

де $n-$ кількість нейронів у схованому шарі;

ν_j- вага синапсу нейрона j схованого шару до вихідного нейрона;

h_j- вихідне значення нейрона j схованого шару;

b_3- поріг вихідного нейрона;

F_3- функція активації вихідного нейрона.

Вихідне значення нейрона j схованого шару визначається за формулою:

$$h_j = F_2\left(\sum_{i=1}^{m} w_{ij} x_i - b_{2j}\right), \tag{3.2}$$

де $w_{ij}-$ вага від i-го вхідного нейрона до j-го нейрона схованого шару;
x_i- вхідні значення;
$b_{2j}-$ поріг j-го нейрона схованого шару.

При моделюванні нейромереж існує багато функцій активації, але найбільш поширеною є сигмоїд:

$$f(x) = \frac{1}{1 + exp(-\alpha x)}. \tag{3.3}$$

Основна перевага цієї функції в тому, що вона диференційована на всій осі абсцис і має дуже просту похідну:

$$f(x) = \alpha f'(x)(1 - f(x)). \tag{3.4}$$

Сигмоїдну функцію активації використовують для нейронів схованого шару і вихідного нейрона.

Середньоквадратична помилка для навчальної ітерації t розраховується за формулою:

$$E^p(t) = \frac{1}{2}\left(y^p(t) - d^p(t)\right)^2, \tag{3.5}$$

де для навчання вектора p:

$y^p(t)$– обчислене вихідне значення багатошарового персептрона на ітерації t;

$d^p(t)$– бажане вихідне значення багатошарового персептрона на навчальній ітерації t.

Під час навчання узагальнена помилка навчання для всіх навчальних векторів обчислюється за формулою:

$$E(t) = \sum_{i=1}^{p} E^p(t). \tag{3.6}$$

Помилка вихідного нейрона для навчання вектора p розраховується за формулою:

$$\gamma_3^p(t) = y^p(t) - d^p(t), \tag{3.7}$$

а помилка нейрона і схованого шару за формулою:

$$\gamma_i^p(t) = \sum_{j=1}^{n} \gamma_3^p(t) \cdot \nu_i(t) \cdot h_j^p(t) \cdot (1 - h_j^p(t)). \tag{3.8}$$

Для навчання багатошарового персептрона використовується алгоритм зворотного поширення помилки, що складається з таких кроків [13]:

1. Задати мінімальну середньоквадратичну помилку навчання багатошарового персептрона E_{min} , яку необхідно досягти в процесі навчання.

2. Ініціювати ваги та пороги нейронів випадковими величинами з діапзону (-0.5...0.5) .

3. Для навчального вектора p обчислити вихідне значення багатошарового персептрона y, використовуючи формули (3.2–3.3)

4. Обчислити помилку вихідного нейрона за формулою (3.6)

5. Модифікувати ваги і пороги вихідного нейрона згідно з формулами:

$$v_j^p(t+1) = v_j^p(t) - \alpha \cdot \gamma_3^p(t) \cdot h_j^p(t) \cdot y^p(t) \cdot (1 - y^p(t)),$$
$$b_3^p(t+1) = b_3^p(t) + \alpha \cdot \gamma_3^p(t) \cdot y^p(t) \cdot (1 - y^p(t)).$$

6. Обчислити помилку нейронів схованого шару $\gamma_i^p(t)$ згідно з виразом (3.7).

7. Модифікувати ваги і пороги нейронів схованого шару відповідно до формул:

$$w_{ij}^p(t+1) = w_{ij}^p(t) - \alpha \cdot \gamma_i^p(t) \cdot h_j^p(t) \cdot x_i^p(t) \cdot (1 - h_j^p(t)),$$
$$b_{2j}^p(t+1) = b_{2j}^p(t) + \alpha \cdot \gamma_j^p(t) \cdot h_j^p(t) \cdot (1 - h_j^p(t)).$$

8. Розрахувати середньоквадратичну помилку для тренувальної ітерації t , використовуючи формулу (3.4).

9. Повторити кроки 3–8 для всіх векторів тренувальної вибірки.

10. Розрахувати узагальнену середньоквадратичну помилку E(t) багатошарового персептрона за допомогою (3.5).

11. Якщо $E(t)$ є все ще більшою за бажану мінімальну помилку E_{min}, необхідно почати знову.

На даний момент відоме використання декількох нейронних мереж для прогнозування розвитку пандемії. Модель довгої короткочасної пам'яті (LSTM) застосовується в різних темах часових рядів, такі як прогнозування акцій, погоди тощо. Проте висновки щодо точних її проявів під час задачі прогнозування COVID-19 все ще обмежені. LSTM була використана для прогнозування кінця пандемії в Китаї з використанням невеликої вибірки, яка лише представляла місцеву характеристику спалаху [8]. Більше того, їх навчальним набором є статистика епідемії ГРВІ 2003 року, що відрізняється від епідемії COVID-19.

При великих навчальних даних, LSTM фіксує закономірність динамічного зростання графіків з мінімальною похибкою RMSE порівняно з RNN (рекурентими нейроннимим мережами). Результати дослідження [7] свідчать про те, що LSTM є перспективним інструментом для прогнозування пандемії COVID-19 і потенційно може передбачити майбутні спалахи, маючи велику кількість даних для навчання.

РОЗДІЛ 4. ОБЧИСЛЮВАЛЬНИЙ ЕКСПЕРИМЕНТ ТА АНАЛІЗ РЕЗУЛЬТАТІВ ПРОГНОЗУВАННЯ

Цей розділ містить практичну частину роботи. Задача полягала у порівнянні двох моделей прогнозування кількості підстверджених, активних, одужаних та померлих людей: ARIMA та FBProphet та їх оцінці за допомогою вибраних функцій похибок. Цей розділ також описує дані [14], які ми використовували для прогнозування випадків COVID-19 деяких вибраних країн та у всьому світі, а також структуру моделі, яку ми використовували. Розділ завершується результатами та висновками.

Спочатку нами було зроблено візуалізацію даних кожної групи для вибраних країн. З цього було зроблено висновок про те, що швидкіть зарегестрованих випадків COVID-19 у кожній країні з часом збільшується та через деякий час вирівнюється. Це відбувається через масштабне тестування, дотримання карантинних заходіи та масочного режиму. Для нашого дослідження, ми взяли наявні дані вибраних 10 країн на період початку розвитку пандемії. Ми відкинули початкові дані за 5 днів для кожної країни в нашому дослідженні, оскільки спочатку повідомлені результати не були точними через відсутність тестування та контролю і не відображали справжню швидксіть розповсюдження. У таблиці на рис.9 наведені використані дані до 20 травня 2020 включно. Для даного дослідження ми відокремили тренувальну та тестову вибірки у віднощенні 80 % та 20 % відповідно для кожної країни.

Регіон	Розмір вибірки (дні)	Дата початку	Підтверджені	Одужані	Смерті	Активні
У світі	120	2020-01-22	4996472	1897466	328115	2770891
США	115	2020-01-27	1551853	294312	93439	1164102
Іспанія	105	2020-02-06	232555	150376	27888	54291
Італія	106	2020-02-05	227364	132282	32330	62752
Франція	113	2020-01-29	181700	63472	28135	90093
Німеччина	110	2020-02-01	178473	156966	8144	13363
Росія	106	2020-02-05	308705	85392	2972	220341
Іран	87	2020-02-24	126949	98808	7183	20958
Британія	106	2020-02-05	249619	1116	35786	212717
Туреччина	66	2020-03-16	152587	113987	4222	34378
Індія	107	2020-02-04	112028	45422	3434	63172

Рис. 9: Загальна кількість випадків COVID-19 до 20 травня 2020 р., використаних для моделювання.

4.1 Структура прогнозування та вибір функції помилки

На рис. 10 описується прийнята структура прогнозування та аналіз випадків COVID-19 за допомогою двох моделей: ARIMA та FBProphet. Для аналізу ми розділили набори даних підтверджених, активних, одужавших та смертельних випадків на навчальну та тестову вибірки. В усіх можливих випадках з даних було усунуто тренд, після чого ми виконали прогнозування та використали статистичні міри вимірювання для оцінки моделей.

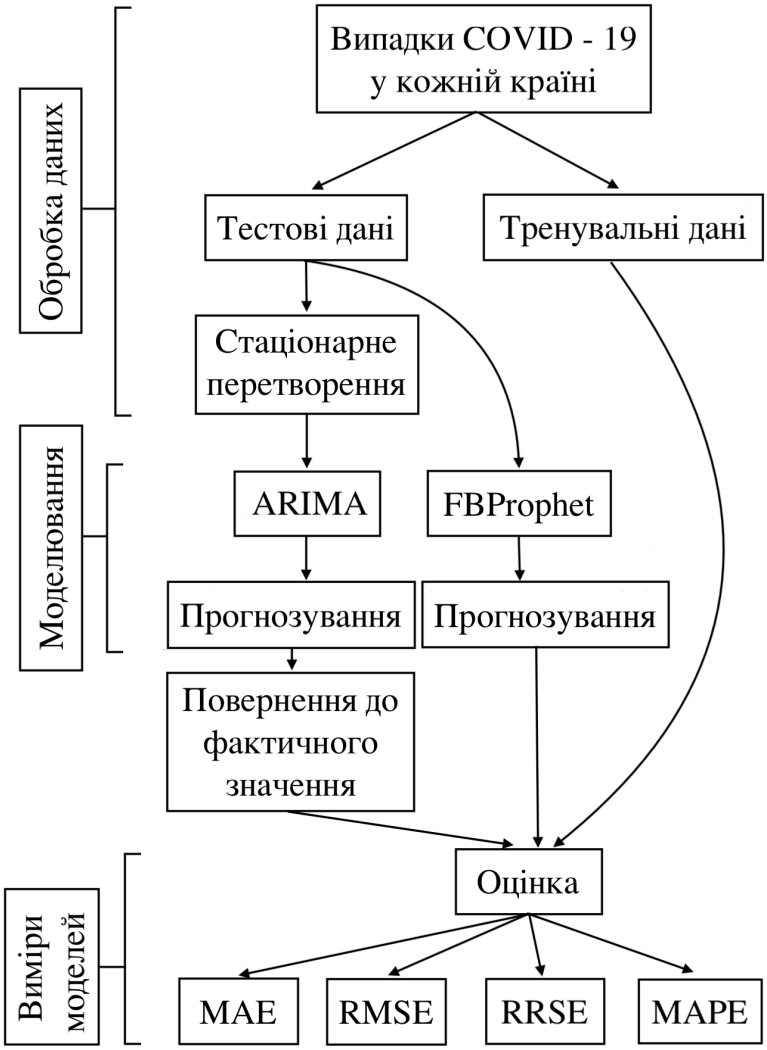


Рис. 10: Структура для оцінки моделей прогнозування.

Для оцінки моделей прогнозування ми використали наступні статистичні показники [4].

Середня абсолютна похибка (MAE):

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|z_k - \hat{z_k}|. \tag{4.1}$$

Середньоквадратичка похибка (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(z_k - \hat{z_k})^2}. \tag{4.2}$$

Відносна середньоквадратичка похибка (RRSE):

$$RRSE = \sqrt{\frac{\sum_{k=1}^{N}(\hat{z_k} - z_k)^2}{\sum_{k=1}^{N}(\overline{z} - z_k)^2}}, \tag{4.3}$$

де $\overline{z} = \frac{1}{N}\sum_{k=1}^{N} z_k$.

Середня абсолютна похибка у процентах (MAPE):

$$MAPE = \frac{100}{N}\sum_{k=1}^{N}\left|\frac{z_k - \hat{z_k}}{z_k}\right|, \tag{4.4}$$

де z_k позначає фактичне значення, а $\hat{z^k}$ позначає передбачуване значення для $k-$ го екземпляра. $\overline{z}$ позначає середнє значення z, а $N-$загальну кількість випробувальних екземплярів.

4.2 Прогнозування активних випадків

Як зазначалося раніше, активні випадки - це кількість заражених, які знаходяться під наглядом лікаря. Будемо визначати активні випадки наступним чином:

$$\text{Активні} = \text{Підтверджені} - \text{Одужані} - \text{Смертельні}$$

Ми використовували моделі ARIMA та FBProphet для прогнозування кількості майбутніх випадків. Як вже зазначалося у розділі 2.2 метод ARIMA можна використовувати для прогнозування, якщо дані є нерухомими. Як стало відомо при дослідженні, дані, що відображають активні випадки не стаціонарні. Для перевірки вибірки на стаціонарність ми застосували тест Дікі-Фуллера. Тому ми застосували наступні техніки перетворення даних у стаціонарну форму для оцінки ARIMA: нормування квадратним коренем та метод lag-1 difference. Також для визначення значень q та p моделі ARIMA було проаналізовано графіки PACF та ACF відповідно. FBProphet застосовувався безпосередньо на фактичних даних. Результати точності прогнозування для активних випадків 10 окремих країн та в усьому світі наведені у наступній таблиці:

Регіон	Модель	MAE	RMSE	RRSE	MAPE
У світі	ARIMA(9,1,2)	19141.89	21377.14	0.086	0.816
	FBProphet	168452.05	182230.63	0.706	6.943
США	ARIMA(10,1,3)	5732.16	8050.31	0.079	0.586
	FBProphet	95766.22	108424.76	1.07	9.12
Іспанія	ARIMA(8,1,4)	2191.68	2603.02	0.346	3.293
	FBProphet	67132.86	69748.42	9.274	109.40
Італія	ARIMA(9,1,3)	3197.25	4266.60	0.320	3.411
	FBProphet	26934.34	30963.76	2.325	35.55
Франція	ARIMA(5,1,4)	10974.15	11489.85	6.166	11.75
	FBProphet	44596.16	48195.48	25.864	48.340
Німеччина	ARIMA(11,1,4)	2114.09	2597.193	0.407	9.052
	FBProphet	50902.42	52259.90	8.197	277.26
Росія	ARIMA(10,1,2)	6456.26	6786.96	0.158	4.238
	FBProphet	36430.36	40232.57	0.936	20.748
Іран	ARIMA(4,1,2)	328.28	379.79	0.147	2.202
	FBProphet	12856.19	12902.11	5.009	82.503
Британія	ARIMA(4,1,2)	8090.84	8637.25	0.375	4.66
	FBProphet	2954.65	4649.43	0.202	1.481
Туреччина	ARIMA(8,1,2)	3631.37	3655.74	0.884	9.485
	FBProphet	59801.55	60725.11	14.678	158.59
Індія	ARIMA(11,1,5)	7007.09	7330.06	0.61	16.74
	FBProphet	10245.17	12085.37	1.005	21.429

Рис. 11: Результати роботи моделей активних випадків COVID-19 у вибраних країнах.

У даній таблиці наведено порядок ARIMA, при якому модель була підібрана найкраще та точність результатів. Аналізуючи оцінку MAPE можемо побачити, що вона була найкраща для даних США та Великобританії у розмірі 0,586 та 1481 у моделей ARIMA та FBProphet відповідно. З результатів можна чітко сказати, що ARIMA має набагато кращі показники порівняно з моделлю FBProphet стосовно всіх обраних функцій помилок.

4.3 Прогнозування одужання

Для прогнозування та аналізу темпів одужання ми виконали оціночне дослідження прийнятих моделей з використанням даних про одужання 10 вибраних країн та окремо розглянули ситуацію у всьому світі. Як і у попередньому випадку ми впевнилися, що дані про одужання також є нестаціонарним. Тому ми використали методи переведення часового ряду до стаціонарного, подібні до тих, що обговорюються в минулому пункті для оцінки моделі ARIMA. Ми застосували FBProphet безпосередньо на фактичних даних, що відповідають моделі та генерують прогнозування результатів. У таблиці 12 наведені результати точності моделей для одужаних випадків. Маємо наступні результати: для MAE найкращі показники 78,19 та 69,11 для Великобританії моделей ARIMA та FBProphet

відповідно.

Регіон	Модель	MAE	RMSE	RRSE	MAPE
У світі	ARIMA(9,1,2)	34932.99	36992.53	0.128	2.523
	FBProphet	185584.71	214741.61	0.712	12.49
США	ARIMA(5,1,2)	31899.89	33109.68	0.667	15.635
	FBProphet	53970.19	57816.45	1.165	24.174
Іспанія	ARIMA(8,1,4)	9683.45	9774.06	0.786	7.361
	FBProphet	3021.53	3766.69	0.303	2.22
Італія	ARIMA(9,1,3)	12910.06	13078.23	0.693	12.78
	FBProphet	8721.87	10057.88	0.533	7.881
Франція	ARIMA(3,1,1)	5780.87	5853.29	1.21	10.574
	FBProphet	7323.90	8362.88	1.729	12.613
Німеччина	ARIMA(5,1,3)	13702.61	13901.04	1.287	9.808
	FBProphet	25017.26	28763.20	2.664	16.969
Росія	ARIMA(4,1,0)	2376.69	3212.50	0.141	5.103
	FBProphet	26988.80	33858.56	1.484	60.964
Іран	ARIMA(1,1,1)	4213.14	4496.75	0.736	4.933
	FBProphet	5638.72	6037.87	0.988	6.267
Британія	ARIMA(4,1,2)	78.19	91.12	1.177	8.311
	FBProphet	69.11	79.44	1.026	7.326
Туреччина	ARIMA(8,1,2)	4242.09	4333.57	0.44	4.321
	FBProphet	45986.27	46211.42	4.688	45.536
Індія	ARIMA(2,1,0)	721.17	1066.65	0.096	2.911
	FBProphet	11395.90	14381.55	1.295	42.882

Рис. 12: Результати роботи моделей одужаних у вибраних країнах.

Результати показують що прогнозування за допомогою моделі ARIMA майже відповідає фактичним значенням, тоді як модель FBProphet показала себе значно гірше. Бачимо, що мінімальне та максимальне значення оцінки MAPE для ARIMA 15,6 та 2,5 відповідно. Дані показники є прийнятними для отримання результатів прогнозування. Для FBProphet маємо протилежну ситуацію: мінімальна MAPE дорівнює 31,822, а максимальна - 3,759, що не є гарними показниками в сукупності.

4.4 Прогнозування смертельних випадків

Коронавірус забрав багато життів. Отже, аналіз та прогнозування смертності від вірусу є необхідним для того, щоб зрозуміти план дій на майбутнє, який допоможе протидіяти ситуації. У цьому підрозділі ми оцінили моделі прогнозування випадків смертності обраних 10 країн та світу в цілому. Ми перетворили нестаціонарні дані в стаціонарну форму, щоб застосувати модель ARIMA. Модель FBProphet застосована до фактичних даних для прогнозування результатів. Таблиця (рис. 13) показує точність прогнозування моделей для смертельних випадків. Ми бачимо, що помилок у прогнозуванні за допомогою методу ARIMA

набагато менше, тоді як прогнозування з використанням FBProphet мають високий коефіцієнт помилки в результатах.

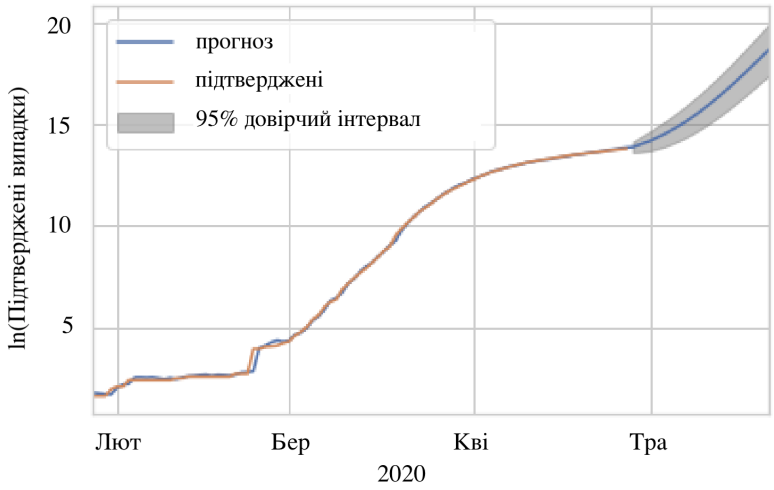
Регіон	Модель	MAE	RMSE	RRSE	MAPE
У світі	ARIMA(9,1,2)	661.98	821.20	0.026	0.257
	FBProphet	21666.12	24874.45	0.735	7.465
США	ARIMA(2,1,0)	1924.08	1988.71	0.19	2.571
	FBProphet	4799.16	5856.54	0.56	5.751
Іспанія	ARIMA(2,1,0)	940.85	953.08	0.907	3.577
	FBProphet	2573.67	2961.76	2.818	9.525
Італія	ARIMA(2,1,0)	1240.10	1254.32	0.892	4.128
	FBProphet	3008.94	3433.46	2.443	9.703
Франція	ARIMA(3,1,1)	1335.79	1355.02	0.983	5.139
	FBProphet	6545.93	7270.98	5.274	24.382
Німеччина	ARIMA(1,1,0)	318.04	341.57	0.668	4.382
	FBProphet	1446.64	1668.89	3.262	18.761
Росія	ARIMA(2,1,0)	43.31	48.98	0.082	2.252
	FBProphet	628.39	709.50	1.184	30.597
Іран	ARIMA(1,1,1)	836.66	836.86	2.929	12.487
	FBProphet	257.44	291.70	1.021	3.759
Британія	ARIMA(2,1,0)	959.53	984.02	0.343	3.119
	FBProphet	4171.84	4867.84	1.699	12.639
Туреччина	ARIMA(8,1,2)	113.54	117.61	0.619	2.909
	FBProphet	280.83	312.96	1.647	6.945
Індія	ARIMA(2,1,0)	48.94	60.75	0.085	2.704
	FBProphet	771.35	897.58	1.26	31.822

Рис. 13: Результати роботи моделей смертельних випадків COVID-19 у вибраних країнах.

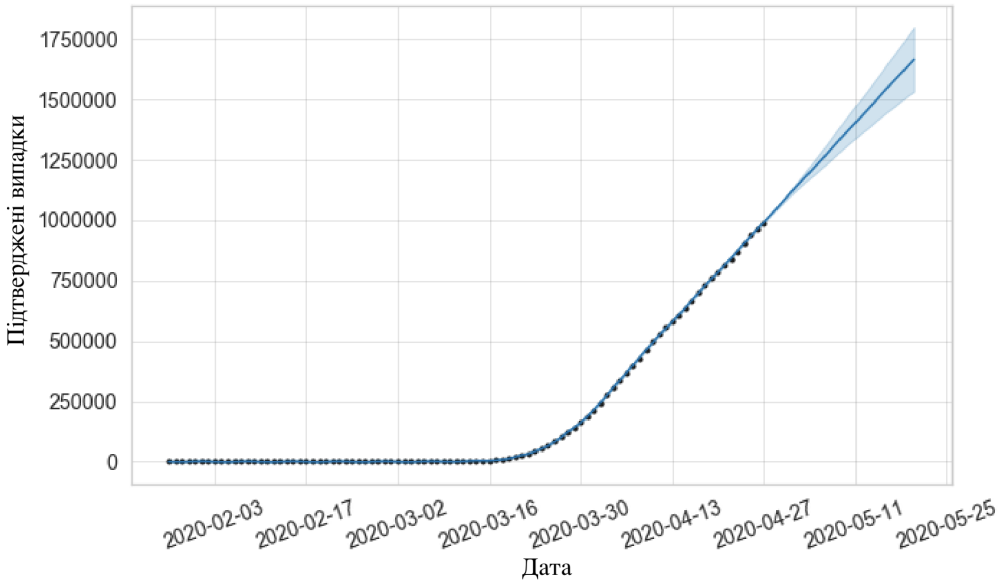
Результати свідчать про те, що ARIMA можна використовувати для фактичного прогнозування смертельних випадків та відповідно планувати заходи.

4.5 Прогнозування підтверджених випадків

У цьому підрозділі ми визначили точність моделей для підтверджених випадків. Для цього аналізу ми обрали дві країни: прогресивні США та густо населену Індію. Результати для даних регіонів для моделей ARIMA та FBProphet показано на рис. 14 та рис. 15 відповідно. Для візуалізації точністості підбору моделі ми зобразили обидва результати на одному графіку. Як можемо побачити, прогнозовані та фактичні дані збігаються. Варто помітити, що модель FBProphet працює краще у випадку даних США, як показано на рис. 14, тоді як ARIMA добре прогнозує дані Індії, як показано на рис. 15.

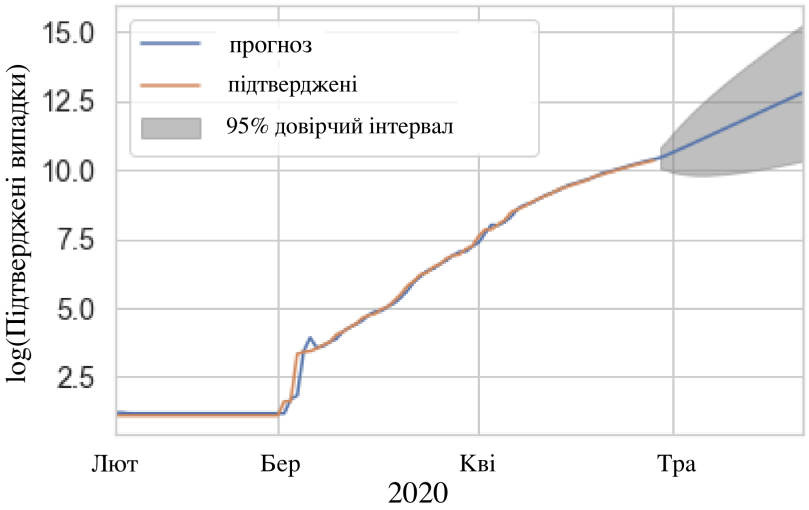


(а)

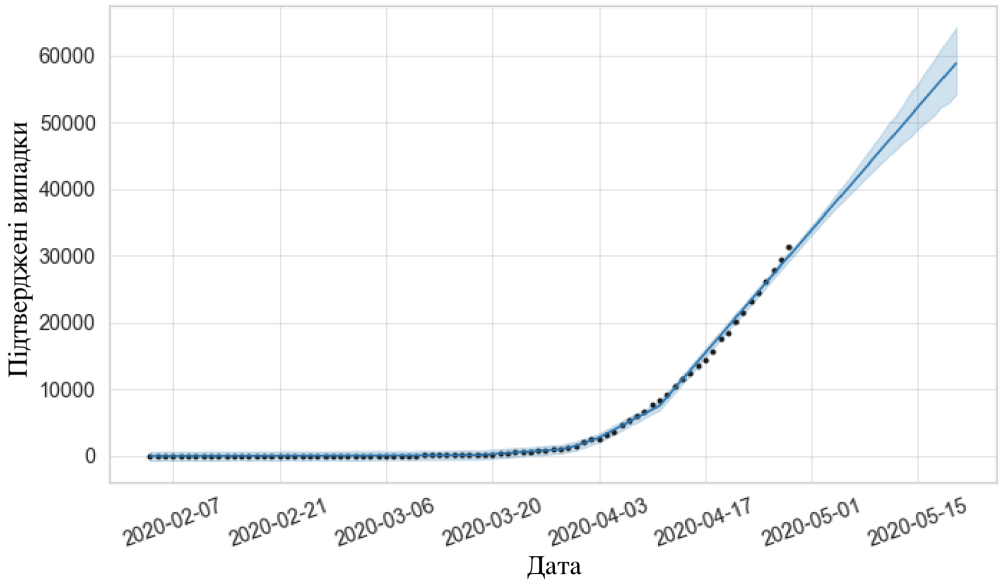


(б)

Рис. 14: Фактичні на прогнозовані значення графіків (а) - ARIMA та (б) - FBProphet підтверджених випадків у США



(а)



(б)

Рис. 15: Фактичні на прогнозовані значення графіків (а) - ARIMA та (б) - FBProphet підтверджених випадків в Індії

Бачимо, що FBProphet уникає викидів під час моделювання та прогнозування. Результати також показують, що FBProphet може добре прогнозувати у випадку меншої кількості даних, тоді як використання моделі ARIMA вимагає достатньо великої кількості даних для моделювання та прогнозування результатів.

ВИСНОВКИ

ВООЗ оголосила COVID-19 пандемією, оскільки вона інфікувала більшість країн, і на сьогодні це є однією з основних загроз для людей. Важливіть даної роботи полягає до досліджені аналізу часових рядів для задачі прогнозування кількості інфікованих, загиблих, померлих та тих, що одужали.

В рамках дипломної роботи виконано наступні завдання та зроблено відповідні висновки:

- розглянуто та досліджено епідемічну модель COVID-19, виконано аналіз моделі під час спалаху та отримано висновки стосовно розвитку хвоби під час спалаху;
- розглянуто методи для прогнозування та аналізу часових рядів: Moving Average, Arima, FBProphet;
- під час обчислювального експерименту виконано аналіз та прогнозування кількості активних, підвтверджених, померлих та одужаних людей за допомогою широко прийнятих моделей прогнозування ARIMA та FBProphet. Зібрано дані про COVID-19 з 10 сильно постраждалих країн на початку пандемії: США, Іспанії, Італії, Франції, Німеччини, Росії, Ірану, Великобританії, Туреччини, Індії та в усьому світі до 20 травня 2020 р.;
- прогнозне дослідження показало значне зростання випадків у кожній з чотирьох досліджувальних груп для кожної країни та в усьому світі. Однак, карантинні заходи і політика дистанціювання може вплинути на результати прогнозування;
- проведено детальний аналіз похибок двох моделей для кожної з обраних 10 країн. Результати показали, що для більшості даних країн, модель ARIMA має кращі показники порівняно з моделлю FBProphet, порівнюючи похибки MAE, RMSE, RRSE та MAPE;
- дане дослідження може бути використане Міністерством Охорони Здоров'я задля розуміння і перспективної оцінки впливу заходів для поліпшення епідеміологічної ситуації в країні;
- отримані результати дослідження в подальшому можна покращити, беручи до уваги такі фактори як густота населення, погода, стан стистеми здоров'я, історія пацієнта тощо, застосовуючи методи штучного інтелекту, наприклад нейромереж LSTM;

– продовження даного дослідження доцільно провести у напрямку нейронних мереж, а саме LSTM, яка наразі є перспективною. Нейромережевий аналіз набирає все більше популярності, оскільки, на відміну від методів аналізу часових рядів не передбачає будь-яких обмежень щодо характеру вхідної інформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. А. Г. Ивахненко. Индуктивный метод самоорганизации моделей сложных систем / Ивахненко А. Г.— Киев: Наук. думка, 1981 — 296 с.

2. Aurélien Géron. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.

3. Haykin S. Neural Networks and Learning Machines / S. Haykin // New Jersey : Prentice Hall, 2008. — 936 p.

4. Steel, R.G.D, and Torrie, J. H., Principles and Procedures of Statistics with Special Reference to the Biological Sciences., McGraw Hill, 1960

5. Guerra FM, Bolotin S, Lim G, Heffernan J, Deeks SL, Li Y, Crowcroft NS (December 2017). "The basic reproduction number (R0) of measles: a systematic review". The Lancet. Infectious Diseases. 17 (12): e420–e428

6. Sean J Taylor and Benjamin Letham. Forecasting at scale. The American Statistician, 72(1):37–45, 2018

7. Novanto Yudistira COVID-19 growth prediction using multivariate long short term memory

8. Yang, Zifeng, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease 12.3 (2020): 165.

9. H. W. Hethcote, “The mathematics of infectious diseases,” SIAM review, vol. 42, no. 4, pp. 599–653, 2000.

10. F. Brauer, C. Castillo-Chavez, and C. Castillo-Chavez, Mathematical models in population biology and epidemiology. Springer, 2012, vol. 2.

11. R. M. Anderson and R. M. May Population biology of infectious diseases: Part 1 Nature, vol. 280, no. 5721, pp. 361–367, 1979.

12. R. M. May and R. M. Anderson, Population biology of infectious diseases: Part 2, Nature, vol. 280, no. 5722, pp. 455–461, 1979.

13. Shaun S Wulff. Time series analysis: Forecasting and control. Journal of Quality Technology, 49(4):418, 2017.

14. Github Inc. Covid-19 cases. https://github.com/cssegisanddata/covid-19 (accessed in 21 may, 2020).

15. Kermack, W. O.; McKendrick, A. G. (1927). "A Contribution to the Mathematical Theory of Epidemics". Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character. 115 (772): 700–721

16. Harko, Tiberiu; Lobo, Francisco S. N.; Mak, M. K. (2014). "Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates". Applied Mathematics and Computation. 236: 184–194

17. World Health Organization. 2020. Coronavirus Disease (COVID-19) Pandemic. URL: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/country-overviews

18. Rasheed Omobolaji Alabi, Akpojoto Siemuri, and Mohammed Elmusrati. Covid-19: Easing the coronavirus lockdowns with caution. medRxiv, 2020

19. Nalini Chintalapudi, Gopi Battineni, and Francesco Amenta. Covid19 disease outbreak forecasting of registered and recovered cases after sixty day lockdown in italy: A data driven model approach. Journal of Microbiology, Immunology and Infection, 2020.

20. Грешилов А. А., Стакун В. А., Стакун А. А. Математические методы построения прогнозов. — М.: Радио и связь, 1997.- 112 с.