КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Кафедра прикладної статистики

Кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 124 Системний аналіз

на тему:

АНАЛІЗ НАДЛИШКОВОЇ СМЕРТНОСТІ ВІД COVID-19 В УКРАЇНІ ЗА ДОПОМОГОЮ МОДЕЛЕЙ ПРОГНОЗУ ЧАСОВИХ РЯДІВ

Виконала студентка 4 курсу

Плющ Дар'я Олександрівна

(підпис)

Керівник дипломної роботи

Доцент, кандидат фізико-математичних наук

Лівінська Ганна Володимирівна

(підпис)

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

Студент

(підпис)

Роботу розглянуто й допущено до захисту на засіданні кафедри прикладної статистики

«06» червня 2022 р.,

протокол № 11

Завідувач кафедри

Розора І. В.

(підпис)

Київ - 2022

ЗМІСТ

ВСТУП

На сьогоднішній день в якісних моделях прогнозу відчувають потребу багато галузей, такі як економіка, кліматологія, біологія, епідеміологія та інші. Останні роки весь світ зіткнувся з такою проблемою як пандемія COVID-19, спричинена вірусом SARS-CoV-2, що відноситься до сімейства коронавірусів.

Вперше цей вірус був зафіксований в місті Ухань, провінції Хубей, що розташована в Китаї. І хоча перше повідомлення китайської влади про «пневмонію невідомого походження» було зроблене лише 30 грудня 2019 року, а Всесвітня організація охорони здоров'я отримала сповіщення про це на наступний день, насправді перші клінічні прояви в хворих з'явилися ще 8 грудня. Вже 13 січня був зафіксований перший випадок за межами Китаю, а вже 11 березня 2020 року масштаб поширення хвороби охарактеризували як пандемію. Цього ж дня в Україні було оголошено про перший карантин, який продовжувався декілька разів з посиленням та послабленням карантинних обмежень.

Наприкінці 2020 року в Україні була зафіксована найвища смертність за останні п'ять років. За офіційною статистикою, в Україні за одинадцять місяців 2020 року померли 549 тисяч осіб, що майже на 3% більше, ніж за аналогічний період 2019 року. Смертність, яка перевищує усереднені показники минулих років називають «надлишковою». Така «надлишкова» смертність фіксується щороку, але зазвичай вона незначна, і з року в рік залежить від різних показників: від економічного спаду чи погодних умов. Офіційна статистика смертності каже, що частка смертей від коронавірусу досить невелика, натомість причиною більшості смертей вказано серцево-судинні захворювання, хвороби органів дихання, онкологічні захворювання, тощо. Проте, науковці вважають, що причина такої ситуації – «недовиявленість» хворих на COVID-19. І справді, через складність проведення тестування, часто літнім людям або людям, що мають хронічні хвороби простіше вказати іншу причину смерті, ніж діагностувати COVID-19. Так науковці припускають, що «недовиявлені» хворі на COVID-19 становлять 90

відсотків всієї «надлишкової» смертності останніх місяців 2020 року. Можливими причинами високої надлишкової смертності можуть, проте, бути відсутність планово-профілактичного лікування людей з хронічними захворюваннями, несвоєчасне надання медичної допомоги, критичні стани, які виникають у людей як наслідок перенесеної ковід-інфекції.

АКТУАЛЬНІСТЬ РОБОТИ ТА ПІДСТАВИ ДЛЯ ЇЇ ВИКОНАННЯ

Під час поширення пандемії COVID-19 перед владою кожної держави щодня постають важливі питання. Наприклад, чи послаблювати або чи посилювати карантинні обмеження, на яких умовах це робити, яке навантаження лікарень очікувати, яка тенденція захворюваності, одужання, смертності, тощо. Так в усьому світі прогнозування епідемічної ситуації стало дуже актуальним та вагомим питанням, як необхідний інструмент для прийняття рішень щодо запобігання або виходу з такої критичної ситуації як пандемія.

Завдання прогнозу має мету за отриманими раніше даними передбачити майбутні значення досліджуваного процесу або системи, скласти прогноз на певний відрізок часу. В даний час, для прогнозування та вивчення складних систем, широко використовується підхід, заснований на аналізі сигналів, вироблених системою в різні моменти часу - аналізі часових рядів.

Щомісяця оприлюднюються дані щодо кількості смертей від хвороби COVID-19. З математичної точки зору ці дані є часовим рядом. Дані епідеміологічного типу, що поєднують в собі малу довжину ряду і складність породжуваного процесу, є одними з найважчих для прогнозування. Але саме завдяки аналізу та прогнозу часових рядів ми можемо правильно описати дані, створити математичну модель, отримати орієнтовні показники за майбутній період, а все це може служити гарною основою для прийняття необхідних рішень.

В даній роботі з використанням методів прогнозування часових рядів проводиться аналіз різниці реальної щомісячної смертності в Україні під час пандемії хвороби COVID-19 в 2020-2021 роках, зі смертністю, що за прогнозами могла би бути в цей самий період за звичних обставин.

Метою цієї роботи є дослідження надлишкової смертності від COVID-19 за допомогою методів прогнозування часових рядів.

Об'єктом дослідження є сам процес прогнозування часового ряду, що базується на даних про смертність в Україні в період з січня 2015 р. по грудень 2021 р.

Предмет дослідження – методи та алгоритми побудови моделі для прогнозування часових рядів.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ВІДОМОСТІ ПРО ЧАСОВІ РЯДИ

1.1. Визначення часових рядів

Часовий ряд (ЧР) − це набір спостережень, отриманих в послідовні моменти часу. Або це послідовність впорядкованих в часі числових спостережень (показників) x_t, що характеризують рівень стану та зміни досліджуваного явища в послідовні моменти часу t. Якщо час змінюється дискретно ($t = 0,1,2, \dots$), часовий ряд називається дискретним. Якщо ж $t \geq 0$, часовий ряд є неперервним.

Характерною властивістю часових рядів є те, що дані зазвичай не генеруються незалежно, їх розсіювання може залежати від часу, вони часто слідують певній тенденції (*тренду*) та можуть мати циклічні компоненти. Статистичні процедури, які припускають незалежність та однакову розподіленість даних, таким чином, виключаються з аналізу часових рядів. Дослідження часових рядів вимагає відповідних методів, які об'єднуються в «Аналіз часових рядів».

Нехай процес x_t триває в часі, тобто $(x_t)_{t=0,\pm1,\pm2,\dots}$, але ми маємо спостереження лише в моменти часу $t = 0,1,2, \dots, n$. Таким чином ми спостерігаємо послідовність $x_1, x_2, \dots, x_n$. Теоретичні властивості цієї послідовності залежать від базового випадкового процесу $(x_t)_{t\in\mathbb{Z}}$. Будь-який ЧР містить два обов'язкових елемента: час t та відповідне йому значення показника або рівень ряду x_t.

1.2. Цілі, завдання та етапи аналізу часових рядів

1.2.1. Цілі аналізу часових рядів

1) Модель часового ряду може бути використана просто для того, щоб компактно описати та продемонструвати дані (дескриптивні статистики, графічні зображення).

2) Аналіз та інтерпретація (визначення та опис природи ряду, розуміння

механізму, що породжує часовий ряд, наявність сезонних факторів, зв’язок з іншими змінними, побудова математичної моделі часового ряду).

3) Виділення сигналу за наявності шуму (тенденції за наявності випадкових збурень).

4) Прогноз (передбачення майбутніх значень (одного чи кількох) часових рядів за поточним та попередніми значеннями).

5) Керування (підгонка різних параметрів керування, щоб зробити часовий ряд ближчим до мети).

Тощо.

В даній роботі аналіз часового ряду виконується саме з ціллю прогнозування майбутніх значень часового ряду за попередніми значеннями, та аналізу їх відмінностей від реальних показників.

Першим кроком дослідження будь-якого часового ряду завжди має бути надзвичайно ретельний аналіз процесу, системи, даних, що спостерігаються.

Це вивчення часто призводить до певного вибору статистичного аналізу та статистик (даних), які будуть використані для сумування інформації по даних та при побудові адекватної ймовірнісної моделі для представлення даних.

Після того, як вибрана відповідна модель, можна оцінювати параметри моделі, перевіряти модель на якість та на відповідність даним, та, можливо, використати побудовані моделі для покращення нашого розуміння механізму, що породжує часовий ряд.

1.2.2. Завдання аналізу часових рядів

- виокремлення та опис основних характерних особливостей ряду; підбір статистичної моделі, що найкращим у певному розумінні способом відображає ряд;

- прогнозування майбутніх значень показників, що утворюють ряд, за попередніми спостереженнями;

- підготовка рекомендацій з управління процесом, що породжує досліджуваний часовий ряд.

Аналіз часових рядів, як правило, передбачає проведення таких основних етапів:

- графічне подання й попередній аналіз поведінки часового ряду; - виокремлення і видалення закономірних складових ряду (тренду, сезонних та циклічних компонент);

- виокремлення і видалення низько- та високочастотних складових (фільтрація);

- дослідження випадкової складової часового ряду, що залишилася після видалення вищезазначених компонент;

- побудова і перевірка адекватності моделі випадкової складової; - побудова загальної моделі досліджуваного ряду;

- дослідження отриманої моделі і прогнозування майбутньої поведінки об’єкта що вивчається;

- вивчення взаємодії між різними часовими рядами, що характеризують певну систему або процес.

1.2.3 Етапи аналізу часових рядів

1) Дескриптивні статистики, зображення.
2) Побудова математичної моделі.
3) Прогноз.
4) Контроль (оцінка якості).
5) Підгонка моделі.

РОЗДІЛ 2. Прогнозування часових рядів

2.1. Концепція стаціонарності

Стаціонарні процеси – випадкові процеси, характеристики якого не залежать від часу, а саме:

1) Математичні сподівання однакові. Для часових рядів це означає, що ряд не має тенденції рости чи спадати: може випадково піти вгору чи вниз, але в середньому залишається на тому ж рівні.

2) Дисперсії рівні, тобто зміни (розсіювання) ряду приблизно однакове з часом.

3) Автоковаріаційна функція (сила зв'язку між сусідніми значеннями процесу) не змінюється з часом, тобто зв'язок між показниками однаковий. Те саме можна сказати про показники, віддалені між собою на будь-яку кількість моментів часу.

Для формального визначення слабкої стаціонарності достатньо двох передумов: 1) та 3).

Зазвичай в наявності є лише одна реалізація часового ряду, тобто кожна з випадкових величин (значення часового ряду в певний момент часу) представлена лише одним значенням. Це приблизно те саме, що по одному представнику певної популяції казати про середні значення та варіацію якихось характерних ознак для всієї популяції. Проте внаслідок рівності математичних сподівань та дисперсій значень стаціонарного часового ряду для всіх моментів часу, з'являється можливість використати всі доступні значення часового ряду для того, щоб оцінити середнє значення процесу та його дисперсію.

Послідовність $\gamma_k = cov(x_{t+k}, x_t)$ називається *автоковаріаційною функцією стаціонарного процесу*.

Автокореляційна функція (ACF) задається таким чином:

$$\rho_k = corr(x_{t+k}, x_t) = \frac{\gamma_k}{\gamma_0}.$$

Зображення ACF називається *корелограмою*.

З графіка функції автокореляції можна визначити стаціонарність. Нестаціонарні процеси часто мають графік з повільно спадаючою кривою. Сезонність або періодичність також може бути визначена з графіка функції автокореляції. Якщо дуже важко визначити, чи є ряд стаціонарним, то будується корелограма, графік якої не спадає до 0 для нестаціонарних рядів.

Часткова корелограма використовується для визначення кількості лагових змінних у AR(p) процесі. З теоретичної точки зору, оцінка коефіцієнта автокореляції k-того порядку при $k > p$ повинна дорівнювати нулю. Таким чином, якщо процес згенерований AR(p) процесом, то на частковій корелограмі графік повинен спадати до нуля, тобто належати довірчому інтервалу, при $k > p$.

На практиці графіки автокореляції та часткової автокореляції не завжди є доволі ясними, тому іноді є досить важким завданням визначити параметри моделі. Часто коли на діаграмах не можна побачити наближення до нуля, використовують змішані процеси ARMA. Іноді, оцінюють декілька можливих моделей і обирають найкращу за допомогою третього етапу – діагностики моделі.

2.2 Наївні методи прогнозу

2.2.1 Метод прогнозу за середнім значенням

Метод прогнозу за середнім значенням - це один з найпростіших методів, що дозволяє виділити тренд. Для застосування цього методу треба мати доволі довгий ряд спостережень. Формально метод описується виразом:

$$\overline{y_t} = \frac{1}{k} \sum_{j=-k_1}^{k_2} y_{t+j}, \;\; k = k_1 + k_2 + 1.$$

Члени нового часового ряду є середніми значеннями певної кількості сусідніх до відповідного члена даного ряду. З формули видно, що нова кількість

спостережень становить *(T – k)*. Єдиною складністю є визначення чисел k_1 та k_2. Як правило, їх сума дорівнює повному циклу сезонності.

Існує також метод подвійного усереднення. Цей метод двічі використовує усереднення часового ряду. При цьому кількість спостережень зменшується на два повних цикли сезонності, тому для використання методу необхідно мати часовий ряд, який складається щонайменше з 3-х повних циклів сезонності.

Прогноз за середнім значенням спирається на всі спостереження в даних:

$$\widehat{y_t} = \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t,$$

де *T* це розмір вибірки. Модель прогнозу будується виходячи з припущення, що ряд може бути описаний за допомогою рівняння:

$$X_t = \mu + \varepsilon_t,$$

ε_t – випадкова похибка, яка має нормальний розподіл з нульовим математичним сподіванням, μ – середній рівень ряду.

Прогноз будується по середньому значенню ряду:

$$X_{T+h} = \mu,$$

де μ оцінюється за допомогою вибіркового середнього.

$$\hat{X}_{T+h|T} = \bar{X} = \frac{(X_1 + \cdots + X_T)}{T}$$

Тут $\hat{X}_{T+h|T}$ позначає оцінку прогнозного значення $X_{T+h|T}$ з горизонтом прогнозу h на базі наявних даних спостережень $X_1, \ldots, X_T$.

Цей метод прогнозу зазвичай використовується для стаціонарних рядів.

2.2.2. Наївний прогноз (прогноз випадкового блукання)

Наївний — один із найпростіших методів прогнозування. Відповідно до нього, прогноз на крок вперед дорівнює останньому фактичному значенню:

$$X_{T+h|T} = X_T.$$

Використання цього підходу може здатися наївним, але є випадки, коли дуже важко перевершити цей метод. Наприклад для прогнозуванням температури, якщо ми хочемо знати, яка температура буде на вулиці через 5 хвилин, то метод наївного прогнозу, як правило, буде дуже точним: температура через 5 хвилин буде такою ж, як зараз. Статистична модель, що лежить в основі наївного методу, називається «випадкове блукання» і записується так:

$$y_t = y_{t-1} + \epsilon_t.$$

Варіантність ϵ_t вплине на швидкість зміни даних: чим вона вища, тим швидше змінюватимуться значення.

Якщо часові ряди демонструють зміщення рівнів або інші типи несподіваних змін у динаміці, наївний метод буде швидко оновлюватися і миттєво досягти нового рівня. Однак, оскільки він має пам’ять лише одного (останнього) спостереження, він не буде відфільтровувати шум у даних, а копіювати його в майбутнє. Таким чином, він має обмежену корисність у прогнозуванні. Однак, будучи найпростішим з можливих методів прогнозування, він вважається одним з основних критеріїв прогнозування. Цей метод добре працює для багатьох економічних та фінансових часових рядів.

2.2.3. Наївний прогноз з трендом

В попередніх методах відсутній один з основних компонентів часового ряду – тренд. Найпростіша модель, яка може бути використана з його врахуванням, називається «випадкове блукання зі зміщенням», яке формулюється так:

$$y_t = y_{t-1} + a_0 + \epsilon_t.$$

де a_0 – постійний член, введення якого призводить до збільшення або зменшення траєкторій залежно від значення a_0. Точковий прогноз цієї моделі розраховується як:

$$\hat{y}_{t+h} = y_t + a_0 h,$$

маючи на увазі, що прогноз з моделі є прямою лінією з кутом нахилу a_0.

2.2.4. Наївний сезонний прогноз

У випадку сезонних даних існує простий метод прогнозування, який можна вважати хорошим орієнтиром у багатьох ситуаціях. Як і простий наївний прогноз, сезонний наївний прогноз спирається лише на одне спостереження, але замість останнього значення він використовує значення за той самий період сезону тому. Наприклад, для створення прогнозу на січень 2021 року використовується січень 2020 року. Математично це записується так:

$$\hat{y}_t = y_{t-m},$$

де m – сезонна частота. Цей метод має базову модель, сезонне випадкове блукання:

$$y_t = y_{t-m} + \epsilon_t.$$

Подібно до наївного методу, чим вища варіантність члена помилки ϵ_t, тим швидше змінюються дані. Сезонний наївний метод не вимагає оцінки будь-яких параметрів, і тому вважається одним із популярних контрольних показників для використання з сезонними даними.

2.3. Прогноз з використанням експоненційного згладжування

Використання експоненційної віконної функції вперше приписується Пуассону як розширення техніки чисельного аналізу з 17 століття, а пізніше прийнято спільнотою обробки сигналів у 1940-х роках. Тут експоненціальне згладжування є застосуванням експоненційної або Пуассонівської віконної функції. Експоненціальне згладжування було вперше запропоновано в статистичній літературі Робертом Г. Брауном в 1956 році , а потім розширено Чарльзом С. Хольтом у 1957 році. Усі методи Хольта, Вінтерса та Брауна можна розглядати як просте застосування рекурсивної фільтрації, вперше знайденої в

1940-х роках для перетворення фільтрів кінцевої імпульсної характеристики у фільтри нескінченної імпульсної характеристики.

Експоненційне згладжування використовується зазвичай для короткострокових прогнозів. Належить до групи так званих адаптивних методів, які дозволяють швидко оновити прогноз на основі свіжих даних.

Кожному значенню ряду надається ваговий коефіцієнт, величина якого експоненційно спадає з часом, що відділяє це значення від останнього відомого значення ряду. Таким чином, «найстаріші» значення ряду практично не впливають на результати прогнозу, в той час як останні відомі величини мають найбільшу вагу.

Цим метод експоненційного згладжування наближається до локальних методів, оскільки головний вклад при прогнозі фактично дає лише невелика кількість останніх по часу значень ряду.

Виявлення і аналіз тенденції динамічного ряду часто робиться за допомогою його вирівнювання або згладжування. Експоненційне згладжування - один з найпростіших і поширених прийомів вирівнювання ряду. В його основі лежить розрахунок експоненційних середніх.

2.3.1. Початкові умови експоненційного згладжування

Експоненційне згладжування завжди вимагає попереднього значення експоненційної середньої. Коли процес тільки починається, повинна бути деяка величина S_0, яка може бути використана в якості значення, що передує S_1. Якщо є минулі дані до моменту початку вирівнювання, то в якості початкового значення S_0 можна використовувати арифметичну середню всіх наявних точок або якоїсь їх частини. Коли для такого оцінювання S_0 немає даних, потрібно прогнозування початкового рівня ряду.

Передбачення може бути зроблено виходячи з апріорних знань про процес або на основі його аналогії з іншими процесами. Після k кроків вага, яка надається

початковому значенню, дорівнює $(1-\alpha)^{\mathrm{k}}$. Якщо є впевненість в справедливості початкового значення S_0, то можна коефіцієнт α взяти малим. Якщо такої впевненості немає, то параметру α слід дати велике значення, з таким розрахунком, щоб вплив початкового значення швидко зменшилася. Однак велике значення α, як це випливає з останнього рівняння попереднього розділу, може з'явитися причиною великої дисперсії коливань S_t. Якщо необхідно придушення цих коливань, то після достатнього віддалення від початкового моменту часу величину α можна відняти.

2.3.2. Вибір постійного згладжування

Вибору величини постійної згладжування слід приділяти особливу увагу. Пошуки повинні бути спрямовані на відшукання підстав для вибору найкращого значення. Потрібно враховувати умови, при яких ця величина повинна приймати значення, близькі то одному крайньому значенню, то іншому. Неважко помітити, що при $\alpha = 0,\ S_t = S_0$ представляє випадок абсолютної фільтрації і повної відсутності адаптації, а при $\alpha = 1$ приходимо до так званої наївної моделі $\widehat{x_\tau}(t) = S_t = x_t$ відповідно до якої прогноз на будь-який термін дорівнює поточному фактичним значенням ряду. На практиці ця модель через простоту користується особливою популярністю.

Постійне згладжування характеризує швидкість реакції моделі $\widehat{x_\tau}(t) = S_t$ на зміни рівня процесу, але одночасно визначає і здатність системи згладжувати випадкові відхилення. Тому величиною α слід давати ту чи іншу проміжне значення між 0 і 1 в залежності від конкретних властивостей динамічного ряду.

В якості задовільного компромісу Браун рекомендував брати α в межах від 0,1 до 0,3. Ця рекомендація повторена в ряді робіт. Тим часом показано, що навіть при прогнозуванні ряду, використаного Брауном для ілюстрації, найкращі результати виходять при $\alpha = 0{,}9$. Найбільша точність прогнозування може бути досягнута при будь-яких допустимих значеннях α. Однак, як правило, якщо в

результаті випробувань виявлено, що оптимальне значення константи близько α до 1, слід перевірити законність вибору моделі даного типу. Часто до більших значень α призводить наявність в досліджуваному ряді яскраво виражених тенденцій або сезонних коливань. У цьому випадку для отримання ефективних прогнозів потрібна інша модель.

Ясно, що оптимальне значення α в загальному випадку має залежати від терміну прогнозування τ. Для кон'юнктурних прогнозів в більшій мірі повинна враховуватися свіжа інформація. При збільшенні періоду попередження τ пізніша інформація, яка відображає останню кон'юнктуру, повинна, очевидно, мати в декілька раз меншу вагу, ніж в разі малих τ. Для того щоб згладити кон'юнктурні коливання, слід в більшій мірі враховувати інформацію за минулі періоди часу. Для проведення такого аналізу вводять поняття середнього віку даних.

Вік поточного спостереження дорівнює 0, вік попереднього спостереження дорівнює 1 і т. д. Середній вік - це сума зважених віку даних, використаних для підрахунку згладженої величини. Причому віку мають ті ж ваги, що і відповідна інформація. При експоненційному вирівнюванні вага, що дається точці з віком k, дорівнює $\alpha\beta^k$, де $\alpha - 1$ і середній вік інформації дорівнює:

$$k = 0 \cdot \alpha + 1 \cdot \alpha\beta + 2 \cdot \alpha\beta^2 + \cdots = \alpha \sum_{k=0}^{\infty} k\beta^k = \frac{\beta}{\alpha}.$$

Таким чином, чим менше α, тим більший середній «вік» інформації.

2.3.3. Просте експоненційне згладжування (метод Брауна та Хольта)

Найкраще цей метод працює, коли дані мають дуже гладкий, або навіть горизонтальний тренд. Нова послідовність будується за правилом:

$$S_t = \alpha y_t + (l - \alpha)S_{t-1}, \;\; 0 < \alpha < 1.$$

Початкове значення S_1, треба вибирати доволі обережно, бо від нього залежить як саме буде поводити себе згладжена послідовність. Найчастіше вибирають $S_1 = y_1$, або $S_1 = \mu$. Однак з деяких міркувань може бути обране й інше значення S_1.

Ваговий коефіцієнт α може обиратися кількома шляхами. По-перше, якщо обирається значення близьке до 1, то будуть більш важливими при прогнозуванні останні дані часового ряду, при виборі α близьким до 0, більш впливовими будуть минулі значення. По-друге, можна покласти $\alpha = \frac{2}{T+1}$. По-третє – вибір α, при якому мінімізується один з критеріїв точності прогнозів на *n* періодів. При цьому розрахунки повинні проводитися лише по перших *(T – n)* значеннях часового ряду, а отримані прогнози повинні бути порівняні з реальними даними. При використанні цієї методики відкидається обмеження $0 < \alpha < 1$. Прогноз значень часового ряду дорівнює останньому члену послідовності S_t:

$$\hat{y}_{T+p} = S_T, \qquad p = 1,2, \dots$$

2.3.4. Подвійне експоненційне згладжування (метод Хольта)

Просте експоненціальне згладжування погано працює, коли в даних є тенденція. У таких ситуаціях було розроблено кілька методів під назвою «подвійне експоненціальне згладжування» або «експоненціальне згладжування другого порядку», що є рекурсивним застосуванням експоненційного фільтра двічі, тому названо «подвійне експоненціальне згладжування». Основна ідея подвійного експоненціального згладжування полягає в тому, щоб ввести термін, щоб врахувати можливість ряду, що демонструє певну форму тенденції. Цей компонент нахилу сам оновлюється за допомогою експоненційного згладжування.

$$S'_t = \alpha y_t + (1 - \alpha) S'_{t-1}$$

$$S''_t = \alpha S'_t + (1 - \alpha) S''_{t-1}, \;\; 0 < \alpha < 1.$$

Метод використовується, коли дані часового ряду нестаціонарні. Прогноз будується як останнє значення другої послідовності:

$$\hat{y}_{T+p} = S_T'', \qquad p = 1,2, \dots$$

2.3.5. Потрійне експоненційне згладжування (метод Хольта-Вінтерса)

Потрійне експоненціальне згладжування тричі застосовує експоненційне згладжування, яке зазвичай використовується, коли є три високочастотні сигнали, які потрібно видалити з досліджуваного часового ряду. Існують різні типи сезонності: «множинна» та «адитивна» за своєю природою, так само, як додавання і множення є основними операціями в математиці.

Якщо кожного місяця в серпні ми продаємо на 5 000 квартир більше, ніж у липні, то сезонність носить адитивний характер. Однак, якщо в осінні місяці ми продаємо на 5% більше квартир, ніж у весняні місяці, сезонність носить мультиплікативний характер. Мультиплікативна сезонність може бути представлена як постійний фактор, а не абсолютна величина.

Потрійне експоненціальне згладжування було вперше запропоноване учнем Хольта, Пітером Вінтерсом, у 1960 році після прочитання книги з обробки сигналів 1940-х років про експоненційне згладжування. Нова ідея Хольта полягала в тому, щоб повторити фільтрацію непарну кількість разів більше 1 і менше 5, що було популярно серед вчених попередніх епох. Хоча раніше використовувалася рекурсивна фільтрація, вона застосовувалася двічі і чотири рази, щоб збігтися з гіпотезою Адамара , тоді як потрійне застосування вимагало більш ніж вдвічі більше операцій сингулярної згортки. Використання потрійного згладжування вважається практичним правилом, а не таким, що ґрунтується на теоретичних основах.

Це дозволяє прогнозувати нестаціонарш часові ряди з великими перепадами мінімального та максимального значень. Нові послідовності будуються за правилом:

$$S_t' = \alpha y_t + (1 - \alpha) S_{t-1}'$$

$$S_t'' = \alpha S_t' + (1-\alpha) S_{t-1}''$$

$$S_t''' = \alpha S_t'' + (1-\alpha) S_{t-1}''', \;\; 0 < \alpha < 1.$$

Прогноз на наступні періоди має вигляд:

$$\hat{y}_{T+p} = S_T''', \qquad p = 1,2, \ldots$$

2.4. Моделі AR, MA, ARMA, ARIMA та SARIMA

Стаціонарні часові ряди можна представити широким класом лінійних параметричних моделей. Найпоширенішими є моделі авторегресії AR (autoregression), ковзної середньої MA (moving average) та змішані ARMA (autoregression moving average). Царина застосування цих моделей не обмежується стаціонарними процесами. Так, ряди зі специфічною однорідною нестаціонарністю можна звести до стаціонарних і описувати модифікованою формою моделі ARMA, відомої як модель Бокса-Дженкінса.

2.4.1. MA

Прогноз за моделлю MA (q): $y_t = \theta + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots + b_q\varepsilon_{t-q}$.

Якщо коефіцієнти моделі точно відомі, і є значення y_t для $t \in [1, n]$, то безумовним точковим прогнозом для будь-якого моменту часу буде математичне сподівання процесу, тобто θ. Умовним прогнозом для моменту часу $t+1$ буде умовне математичне сподівання:

$$\hat{y_t}(1) = M\{\theta + \varepsilon_{t+1} + b_1\varepsilon_t + b_2\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q+1} | y_1, \ldots, y_t\}.$$

Серед випадкових величин ε, що знаходяться ліворуч, є такі, що пов'язані зі спостереженнями. Оскільки спостереження складаються із «модельного значення» й похибки, умовні математичні сподівання усіх складових, окрім ε_{t+1}, не дорівнюють нулю.

Наприклад, $M\{\varepsilon_t|y_1,\dots,y_t\}$ є залишком між спостереженням і розрахунком (прогнозом) за моделлю, тобто $e_t = y_t - \widehat{y_t}$. Тому умовні математичні сподівання від усіх минулих значень випадкової складової треба замінити відповідними залишками. Так само будується прогноз на 2 й більше кроків уперед. Усі майбутні замінюються нулями, а минулі — залишками, які можна обчислити. Отже, для моделі MA(q) прогноз залежить від того, які похибки були на попередніх кроках. Починаючи із кроку *(q + 1)* умовний прогноз є математичним сподіванням θ, тобто умовний прогноз збігається з безумовним.

Умовна дисперсія помилки прогнозу на 1 крок випередження становить:

$$D(y_{t+1} - \widehat{y_t}(1)|y_1,\dots,y_t) = M\{\big(\big(\theta + \varepsilon_{t+1} + b_1\varepsilon_t + b_2\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q+1} - \theta - b_te_t - - \cdots - b_qe_{t-q+1}\big)|y_1,\dots,y_t\big)^2\} = \sigma_\varepsilon^2.$$

Аналогічно дисперсія прогнозу на 2 кроки випередження дорівнює:

$$D(y_{t+2} - \widehat{y_{t+2}}(1)|y_1,\dots,y_t) = (1 + b_1^2)\sigma_\varepsilon^2,$$

а дисперсія на τ кроків становить

$$(1 + b_1^2 + \cdots + b_\tau^2)\sigma_\varepsilon^2 \quad \text{для } \tau < q.$$

Якщо $\tau \geq q$, дисперсія помилки умовного прогнозу стає такою самою як і для безумовного прогнозу, тобто дорівнює дисперсії випадкового процесу y_t.

2.4.2. AR

Прогноз за моделлю AR(p): $y_t = \theta + \varepsilon_t + a_1y_{t-1} + a_2y_{t-2} + \cdots + a_py_{t-p}$.

Для прогнозу на один крок уперед можна записати:

$$\widehat{y_t}(1) = M\{y_{t+1}|y_1,\dots,y_t\} = M\{\theta + a_1y_t + \cdots + a_py_{t-p+1} + \varepsilon_{t+1}|y_1,\dots,y_t\} = \\ = +a_1y_t + \cdots + a_py_{t-p+1}.$$

Тобто у рівняння моделі підставляють p минулих значень реалізації часового ряду. Для прогнозу на два кроки вперед отримують:

$$\widehat{y_t}(2) = M\{y_{t+2}|y_1, \dots, y_t\} = M\{\theta + a_1 y_{t+1} + \dots + a_p y_{t-p+2} + \varepsilon_{t+2}|y_1, \dots, y_t\}.$$

Математичне сподівання від випадкової похибки ε знов дасть 0, умовне математичне сподівання від $y_1, y_2, \dots, y_t$ дорівнює цим самим значенням, але до цього виразу входить умовне математичне сподівання від y_{t+1}, отримане на попередньому кроці. Можна підставити його вираз і отримати розгорнуту формулу через значення реалізації. Насправді зручніше розглядати рекурентне співвідношення, яке пов'язує послідовні значення прогнозу. Це співвідношення є лінійним різницевим рівнянням порядку p, і його розв'язок прагне, якщо збільшується t, до величини $\frac{\theta}{1-a_1-\dots-a_p}$, тобто знов таки до безумовного прогнозу.

Умовну дисперсію помилки прогнозу розраховують аналогічно до випадку моделі ковзної середньої, але доведення стають досить громіздкими навіть для моделей невеликого порядку. Наприклад, для моделі AR(2) без вільного члена прогноз на один крок випередження становить: $\widehat{y_t}(1) = a_1 y_t + a_2 y_{t-1}$ та $y_{t+1} = a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1}$. Очевидно, що дисперсія помилки прогнозу на 1 крок дорівнює:

$$D(y_{t+1} - \widehat{y_t}(1)|y_1, \dots, y_t) = \\ = M\{\big(\big(\theta + \varepsilon_{t+1} + b_1\varepsilon_t + \dots + b_q\varepsilon_{t-q+1} - \theta - b_t\varepsilon_t - \dots \\ - b_q\varepsilon_{t-q+1}\big)|y_1, \dots, y_t)^2\} = \sigma_\varepsilon^2.$$

Для прогнозу на 2 кроки відповідно отримуємо:

$$\widehat{y_t}(2) = a_1\widehat{y_t}(1) + a_2 y_t = a_1(a_1 y_t + a_2 y_{t-1}) + a_2 y_t,$$

$$y_{t+2} = a_1 y_{t+1} + a_2 y_t + \varepsilon_{t+2} = a_1(a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1}) + a_2 y_t + \varepsilon_{t+2}$$

Дисперсія помилки прогнозу на 2 кроки випередження дорівнює:

$$(1 + b_1^2)\sigma_\varepsilon^2.$$

Для прогнозу на 3 кроки отримуємо:

$$\widehat{y_t}(3) = a_1\widehat{y_t}(2) + a_2\widehat{y_t}(1) = a_1(a_1(a_1 y_t + a_2 y_{t-1}) + a_2 y_t) + a_2(a_1 y_t + a_2 y_{t-1}),$$

$$y_{t+3} = a_1 y_{t+2} + a_2 y_{t+1} + \varepsilon_{t+3}$$
$$= a_1(a_1(a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1}) + a_2 y_t + \varepsilon_{t+2})$$
$$+ +a_2(a_1 y_t + a_2 y_{t-1} + \varepsilon_{t+1}) + \varepsilon_{t+3}.$$

Дисперсія помилки прогнозу на 3 кроки дорівнює:

$$(1 + b_1^2 + b_2^2 + 2b_1^2 b_2 + b_1^4)\sigma_\varepsilon^2.$$

Очевидно, що дисперсія помилки прогнозу збільшується з кожним кроком. Значно простішими виходять вирази для дисперсії помилки прогнозу, якщо перейти від AR(p) представлення до еквівалентного MA представлення:

$y_t = \theta + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \cdots + \psi_q \varepsilon_{t-q} + \cdots$ із необмеженою кількістю складових. Тоді дисперсію помилки прогнозу на τ кроків можна виразити формулою:

$$\sigma_\varepsilon^2 \sum_{j=0}^{\tau-1} {\psi_j}^2 \quad (\psi_0 = 1).$$

2.4.3. ARMA

Для загальної моделі ARMA(p, q) потрібно об’єднати все те, про що говорилося вище. За моделлю, підставляючи туди для часу t спостереження y_t та розраховані значення залишків, обчислюють прогнозовані значення y_t, а для майбутніх моментів часу — замінюють залишки нулями і замість y_t підставляють їхні прогнозовані значення. Дисперсію помилки прогнозу обчислюють за формулою:

$$\sigma_\varepsilon^2 \sum_{j=0}^{\tau-1} {\psi_j}^2 \quad (\psi_0 = 1).$$

В усіх розглянутих випадках умовний точковий прогноз асимптотично наближається до математичного сподівання ряду, а дисперсія помилки прогнозу — до дисперсії ряду. Це означає, що для стаціонарного процесу вплив наявної

інформації на прогноз та його точність асимптотично спадає до нуля. До того ж за збільшення горизонту прогнозування дисперсія помилки не перевищує дисперсії часового ряду. Цей висновок, на жаль, є наслідком нереалістичного припущення про те, що коефіцієнти моделі відомі точно.

2.4.4. ARIMA

Моделі ARMA можна використовувати лише для стаціонарних часових рядів. Однак на практиці багато часових рядів, наприклад, пов'язані з соціально-економічною і бізнес сферами демонструють нестаціонарну поведінку. Часові ряди, які містять тренд та сезонні компоненти, також мають нестаціонарний характер. Таким чином, з точки зору застосування моделі ARMA є неадекватні для належного опису нестаціонарних часових рядів, які часто зустрічаються на практиці. З цієї причини запропоновано модель ARIMA, яка є узагальненням моделі ARMA, щоб також включити випадок нестаціонарності.

У моделях ARIMA нестаціонарний часовий ряд стає стаціонарним шляхом застосування оператора диференціювання (різницевого оператора) часового ряду. Математичне формулювання моделі ARIMA(p, d, q) з використанням поліномів відставання наведено нижче:

$$\varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t$$

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d y_t = \left(1+\sum_{j=1}^{q}\theta_j L^j\right)\varepsilon_t$$

- Тут *p, d* і *q* є цілими числами, більшими або рівними нулю і відносяться до авторегресійної, інтегрованої моделям та до моделі ковзного середнього відповідно.
- Ціле число *d* контролює рівень диференціювання. Загалом у більшості випадків достатньо *d = 1* (або 2).
- Коли *d = 0*, модель ARIMA(p, 0, q) зводиться до моделі ARMA(p, q).

- ARIMA(p, 0, 0) — це не що інше, як модель AR(p), а ARIMA(0, 0, q) — модель MA(q).
- ARIMA(0, 1, 0), тобто $y_t = y_{t-1} + \varepsilon_t$ є спеціальною і відома як модель випадкового блукання. Вона широко використовується для нестаціонарних даних, таких як економічні ряди та ряди цін на акції.

Корисним узагальненням моделей ARIMA є модель авторегресійного дробового інтегрованого ковзного середнього (ARFIMA), яка допускає нецілі значення диференціювання (параметру *d*). ARFIMA має корисне застосування при моделюванні часових рядів з довгою пам'яттю. У цій моделі розширення $(1 - L)^d$ має бути виконано нами із застосуванням теореми загального бінома.

2.4.5 SARIMA

Модель ARIMA призначена для несезонних нестаціонарних даних. Бокс і Дженкінс узагальнили цю модель для рядів з вираженою сезонною компонентою. Запропонована ними модель відома як сезонна модель ARIMA (SARIMA – seasonal ARIMA). У цій моделі сезонне диференціювання відповідного порядку використовується для видалення з ряду нестаціонарності, пов’язаної з наявністю сезонності. Сезонна різниця першого порядку – це різниця між спостереженням і відповідним спостереженням за попередній період (наприклад, рік) і розраховується як $z_t = y_t - y_{t-S}$. Для місячних часових рядів $s = 12$ і для квартальних рядів $s = 4$. Цю модель зазвичай позначають *SARIMA(p, d, q) × (P, D,Q)*^{s}.

Математичне формулювання моделі *SARIMA(p, d, q) × (P, D,Q)*^{s} в умовах поліномів з затримкою наведено нижче:

$$\Phi_p(L^s)\varphi_p(L)(1 - L)^d(1 - L^s)^D y_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t,$$

$$\Phi_p(L^s)\varphi_p(L)z_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t.$$

Тут z_t — ряди з сезонною різницею.

РОЗДІЛ 3. ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ ПРОГНОЗУ ЧАСОВИХ РЯДІВ

3.1. Використані інструменти

Для побудови часового ряду і його прогнозу було обрано мову програмування R та програмне середовище RStudio.

R — мова програмування для статистичної обробки даних та роботи з графікою, а також вільне програмне середовище обчислень з відкритим вихідним кодом у рамках проекту GNU. Мова створювалася як аналогічна мові S, розробленій в Bell Labs, і є її альтернативною реалізацією, хоча між мовами є суттєві відмінності, але здебільшого код мовою S працює в середовищі R. Спочатку R був розроблений співробітниками статистичного факультету Оклендського університету Росом Айхекою (англ. Ross Ihaka) та Робертом Джентлменом (англ. Robert Gentleman) (перша літера їх імен - R); мова та середовище підтримуються та розвиваються організацією R Foundation. Широко використовується як статистичне програмне забезпечення для аналізу даних та фактично став стандартом для статистичних програм.

RStudio — вільне та відкрите інтегроване середовище розробки (IDE) для R, мови програмування обчислювальної статистики та візуалізації даних. RStudio була започаткована Джозефом Аллером, творцем мови програмування ColdFusion. Хадлі Вікхем головний науковець RStudio. RStudio написана на мові C++ за допомогою Qt framework для графічного інтерфейсу користувача.

Серед пакетів мови R для роботи було обрано наступні:

- dplyr (розширення граматичних конструкцій для маніпуляцій із даними)
- readr (покращений імпорт текстових даних у R)
- tsibble (призначений для створення об'єктів з даними часових рядів відповідно до принципів "охайних даних", тобто даних, з якими легко

працювати: їх легко перетворити, візуалізувати та використовувати для побудови моделі)

- feasts (являє собою колекцію функцій, призначених для візуалізації часових рядів та розрахунку цілого набору показників, що узагальнюють їх властивості)
- lubridate (пакет, який дозволяє проводити арифметичні обчислення між датами)
- ggplot2 (розширення граматичних конструкцій для візуалізації даних)
- tidyr (являє собою колекцію функцій, призначених для візуалізації часових рядів та розрахунку цілого набору показників, що узагальнюють їх властивості)
- forecast (надає методи та інструменти для відображення та аналізу одновимірних прогнозів часових рядів, включаючи експоненціальне згладжування за допомогою моделей простору станів і автоматичне моделювання ARIMA)
- astsa (супроводжує аналіз часових рядів та його застосування: із прикладами R і часовими рядами та підходом до аналізу даних із використанням R)
- bsts (інструмент для моделювання та прогнозування часових рядів, в основі цього пакета лежить методологія, відома в літературі під декількома назвами: "Байєсівські структурні моделі часових рядів", "моделі простору станів", "динамічні лінійні моделі", моделі на основі фільтра Калмана та інші. bsts спеціалізується на прогнозуванні часових рядів, представлених денними даними, і дозволяє включати в моделі сторонні предиктори. Пакет використовує принципи байєсівської статистики для оцінювання параметрів моделей. У пакеті bsts відбувається байєсовське усереднення передбачень набору, що складається з великої кількості моделей)

3.2. Постановка задачі прогнозу

Для застосування методів прогнозу часових рядів візьмемо офіційні дані смертності в Україні в період з 2015 по 2021 роки, отримані на сайті Державної служби статистики України. Дані прикріплені в Додатку А.

Проаналізуємо кількості смертей за січень 2015 – квітень 2020 і спрогнозуємо смертність на травень 2020 – грудень 2021. Потім порівняємо отриманий прогноз з реальними показниками смертності за цей же період і оцінимо вплив захворюваності на COVID-19 на надлишкову смертність.

3.3. Розбиття даних на вибірки, розвідувальний аналіз та підготовка даних

Розіб’ємо вхідні дані на наступні три частини:

1. Навчальна вибірка (з січня 2015 року по вересень 2019 року) – для навчання альтернативних прогнозних моделей.

2. Перевірочна вибірка (з жовтня 2019 року по березень 2020 року) – для перевірки якості прогнозів моделей, підігнаних до даних з навчальної вибірки.

3. Тестова вибірка (з квітня 2020 року по грудень 2021 року) – дані за період пандемії COVID-19.

Так як ми маємо справу з щомісячними даними, то певна відмінність в показниках ряду буде обумовлена різною кількістю днів у відповідному місяці. Наприклад, протягом лютого показник смертності буде меншим ніж у січні з тої простої причини, що в лютому на три дні менше. Для того, щоб інформація відображалася більш коректно, пронормуємо ряд відносно кількості змінних і отримаємо середньоденні показники протягом відповідного місяця.

Код для побудови моделей, моделювання відгуку, специфікації компонент моделей, підгонки моделі та метрики якості моделей описані в Додатку Б.

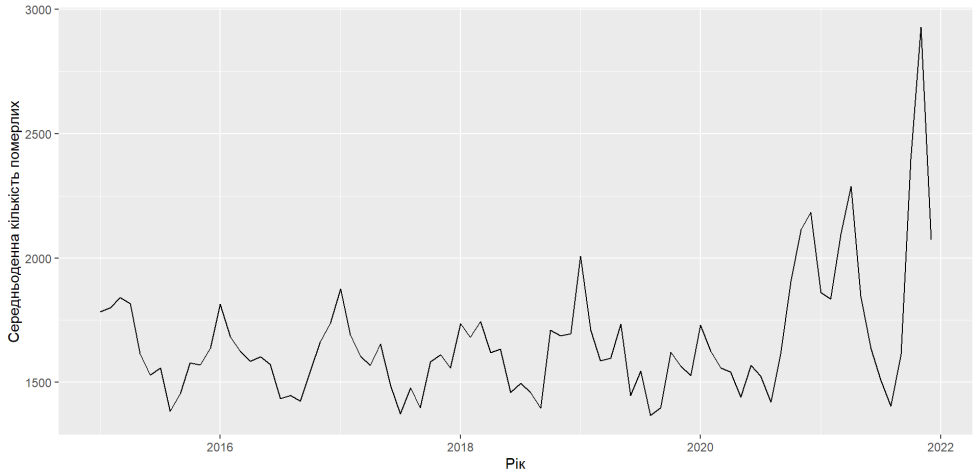


Рис. 1. Середньоденна динаміка місячної смертності в Україні в період з січня 2015 р. по грудень 2021 р. (включно)

Виконаємо невеликий розвідувальний аналіз даних з навчальної вибірки, який допоможе визначитись зі структурою прогнозної моделі та методологією її побудови.

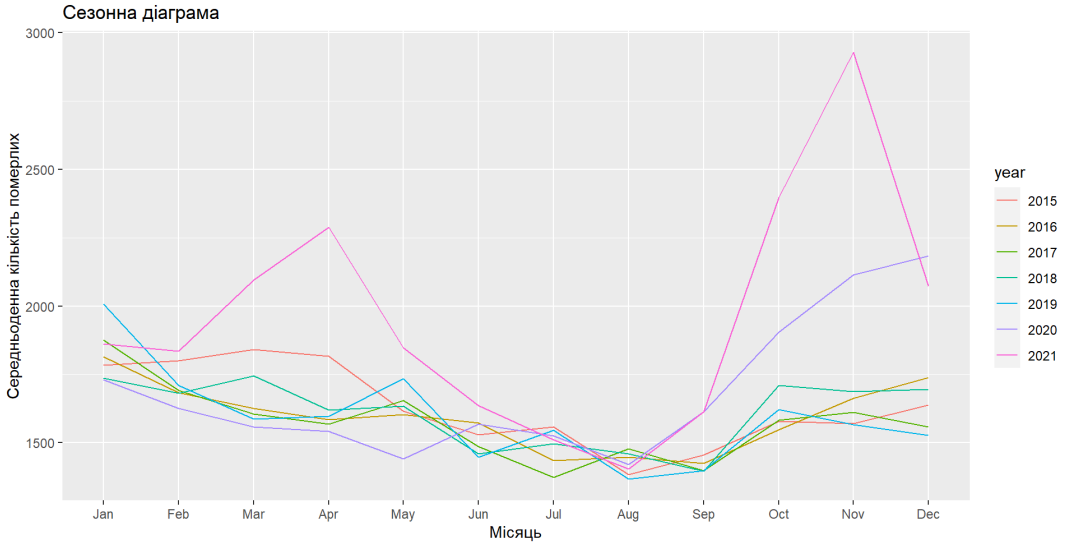


Рис. 2. Середньоденні рівні місячної смертності в Україні в різні роки спостережень

Як бачимо на Рис. 2 , помісячна смертність в Україні до пандемії має чітко виражену сезонність з піковими рівнями, що приходяться на літні місяці (мінімум) та січень, жовтень, листопад та грудень (максимум).

Також побудуємо корелограму та часткову корелограму для майбутнього вибору параметрів для моделей.

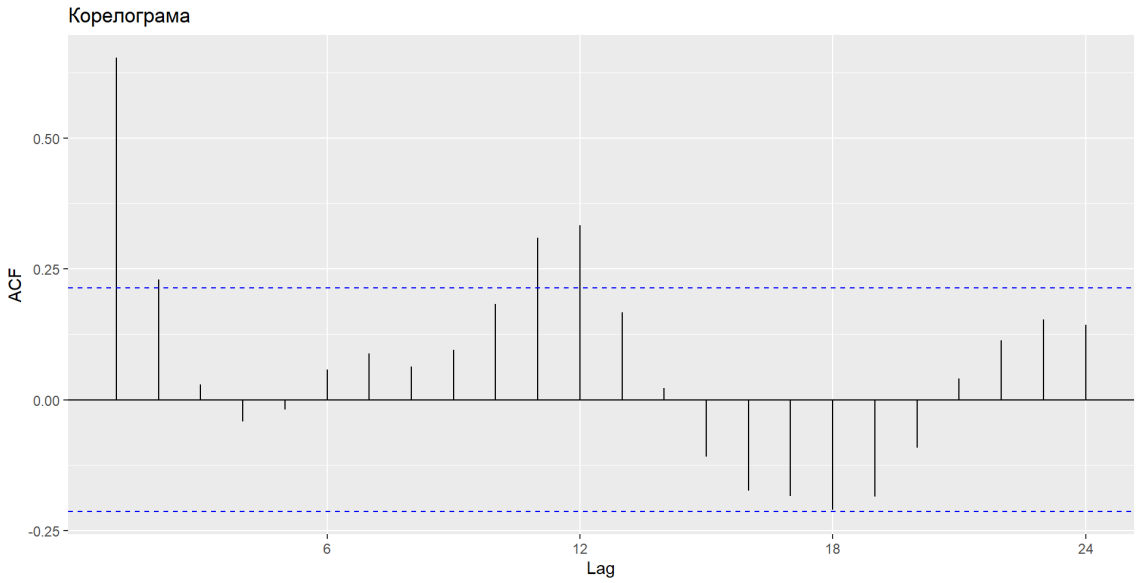


Рис. 3. Корелограма з лагом від 1 до 24

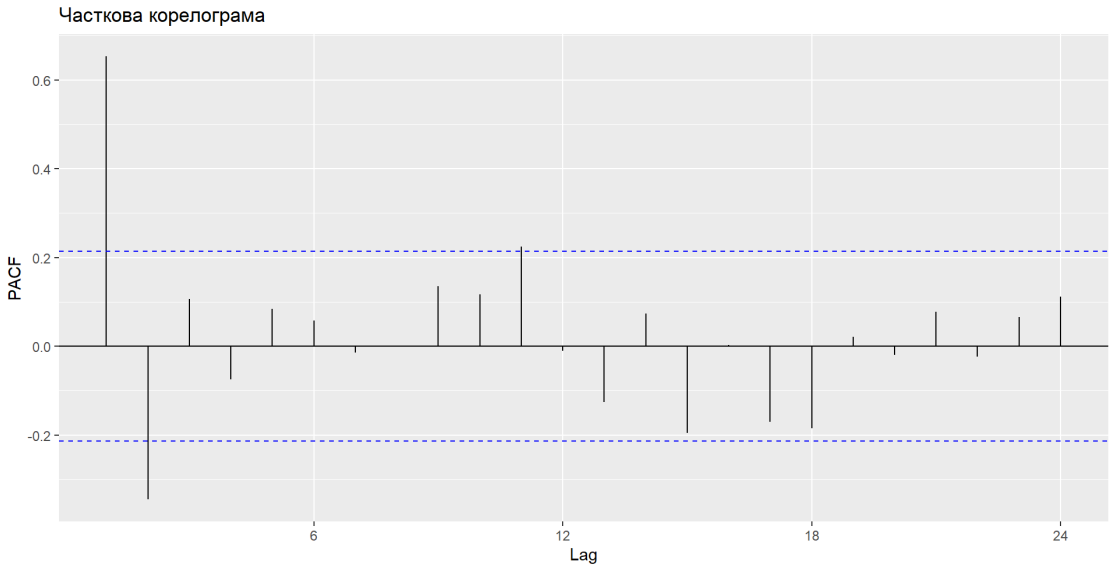


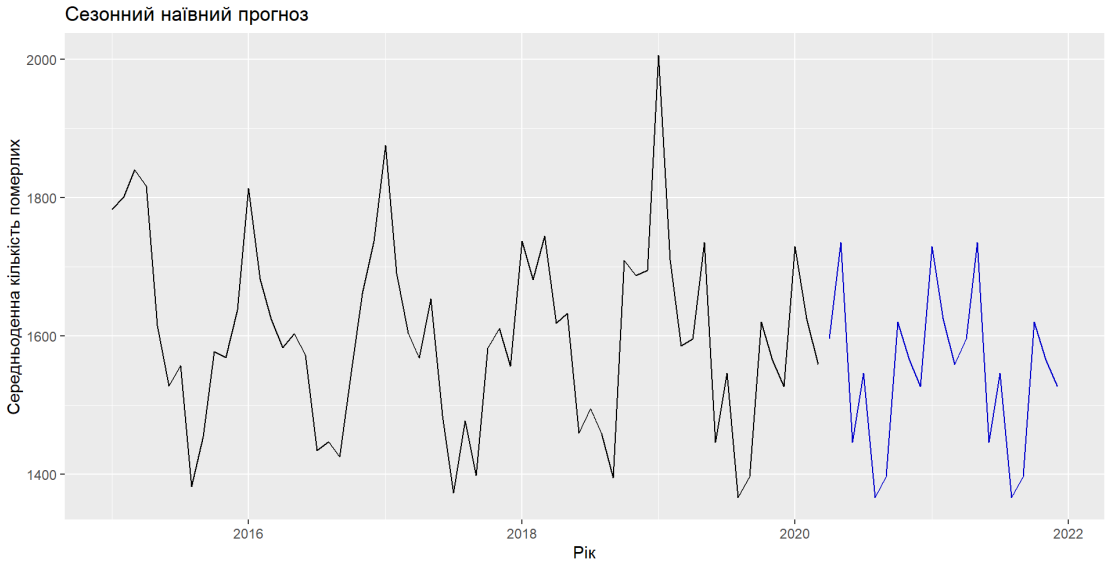
Рис. 4. Часткова корелограма з лагом від 1 до 24

На Рис.3. та Рис. 4. бачимо, що значущими є 1, 2, 12 та можливо 11 лаги. Цю інформацію використаємо пізніше при побудові моделі прогнозу ARIMA.

3.4. Побудова різних типів моделей

3.4.1. Побудова моделі сезонного наївного прогнозу

Реалізуємо цей метод за допомогою функції snaive( ).



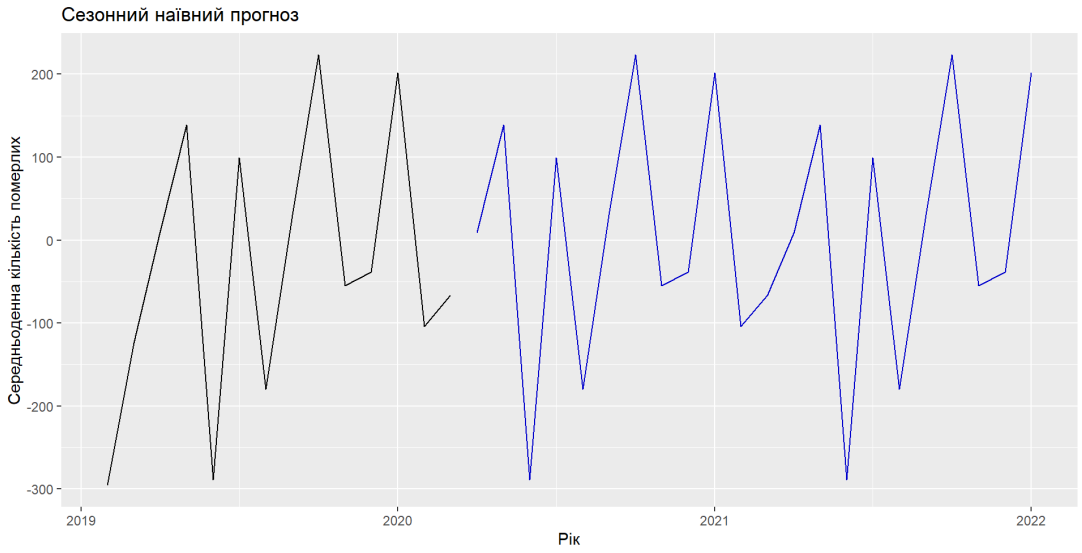


Рис. 5. та Рис. 6. Прогноз природного рівня місячної смертності в Україні в квітні 2020 – грудні 2021 року (синя лінія), який мав би місце за відсутності спалаху COVID-19

Отже, бачимо, що один з найпростіших методів може бути досить непоганим методом побудови прогнозу незважаючи на свою простоту. Для нашої

моделі прогнозу цей метод буде слугувати певними орієнтиром, з якими ми зможемо порівнювати інші моделі.

Також зробимо аналіз залишків моделі побудованої за допомогою сезонного наївного прогнозу.

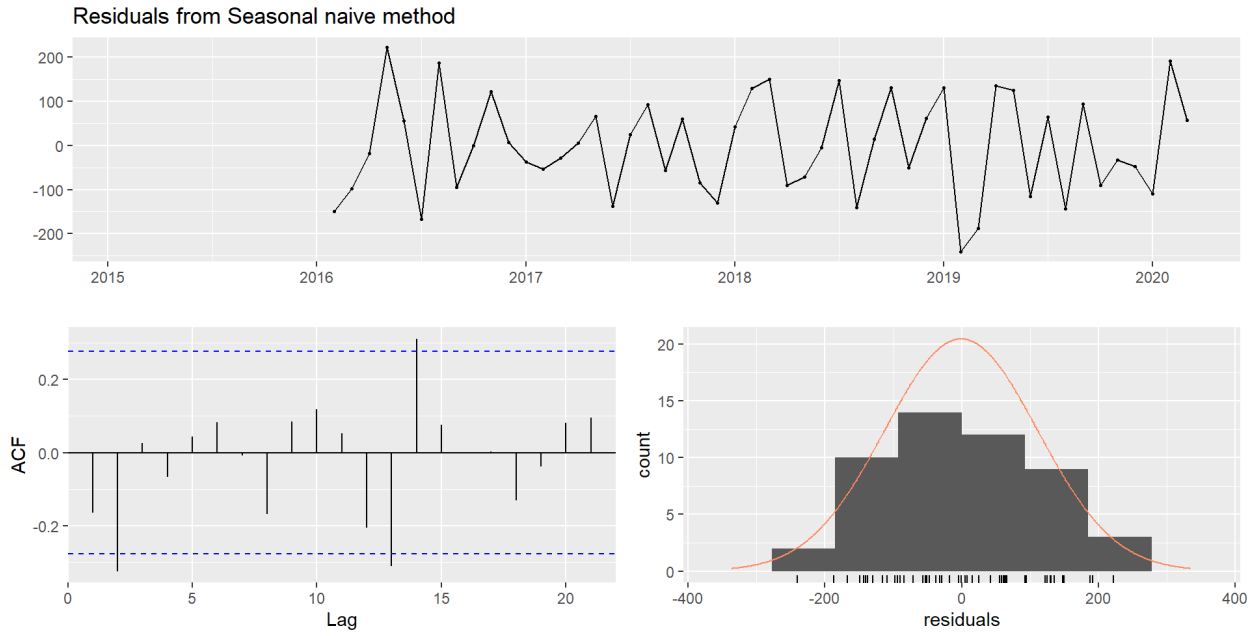


Рис. 7. Залишки сезонної наївної моделі, корелограма та гістограма залишків

Бачимо, поведінка залишків не схожа на білий шум. За гістограмою можна бачити, що, можливо, спостерігається певне відхилення від нормального розподілу. Наявність трьох значущих лагів в корелограмі свідчить про те, що модель не увібрала в себе всі наявні залежності.

3.4.2. Побудова ARIMA моделі

Параметри моделі варто визначати за корелограмою і частинною корелограмою. Раніше в пункті 3.3. ми побачили, що значущими є перші дві автокореляції ряду, перші дві частинні автокореляції ряду, і по одній сезонній (річній) автокореляції. Це означає, що параметрами як AR так і MA моделей можуть бути 0, 1, 2, а параметрами сезонності можуть бути 0 або 1 (як AR так і MA). Щоб обрати найкращу специфікацію для ARIMA моделі скористаємося процедурою Хенона-Рісанена. Згідно з цією процедурою оптимальними лагами для включення в модель вважають такі, за яких досягається мінімальне значення інформаційного критерію Акаїке (AIC). Зауважимо, що є інші інформаційні критерії (наприклад, BIC – інформаційний критерій Баєса), які аналогічним чином визначають оптимальну «економічну» модель, мінімізуючи середню похибку, при цьому накладаючи «штраф» на зайві параметри.

За допомогою функції auto.arima( ) пакету forecast, визначаємо (вказавши AIC в якості критерію вибору оптимальної моделі), що доцільно розглядати специфікації ARIMA(1, 1, 1), ARIMA(0, 1, 2) та ARIMA(1, 0, 0).

Для ARIMA(1, 1, 1) маємо:

```
Call:
arima(x = y, order = c(1, 1, 1), method = "ML")

Coefficients:
         ar1      ma1
      0.5678  -1.0000
s.e.  0.1097   0.0509

sigma^2 estimated as 12986:  log likelihood = -383.04,  aic = 772.09
```

Для ARIMA(0, 1, 2) маємо:

```
Call:
arima(x = y, order = c(0, 1, 2), method = "ML")

Coefficients:
          ma1      ma2
      -0.2436  -0.0689
s.e.   0.1299   0.3040

sigma^2 estimated as 15581:  log likelihood = -387.28,  aic = 780.56
```

Для ARIMA(1, 0, 0) маємо:

```
Call:
arima(x = y, order = c(1, 0, 0), method = "ML")

Coefficients:
         ar1  intercept
      0.5426  1604.3983
s.e.  0.1059    30.5804

sigma^2 estimated as 12777:  log likelihood = -387.41,  aic = 780.82
```

Менше значення AIC маємо для моделі ARIMA(1, 1, 1), а похибка є трохи меншою для моделі ARIMA(1, 0, 0). Тому подивимося на поведінку залишків обох моделей:

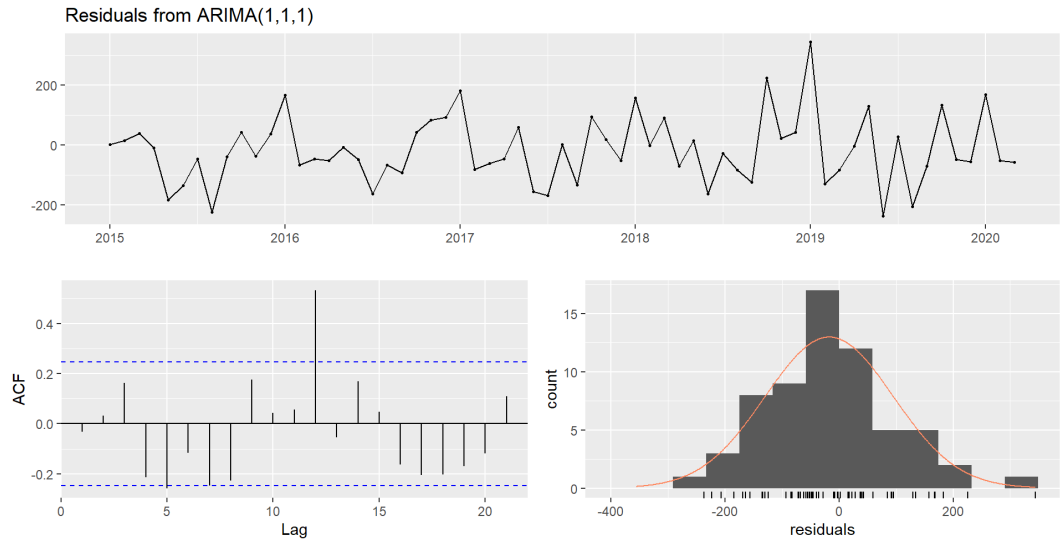


Рис. 8. Поведінка залишків моделі ARIMA(1, 1, 1)

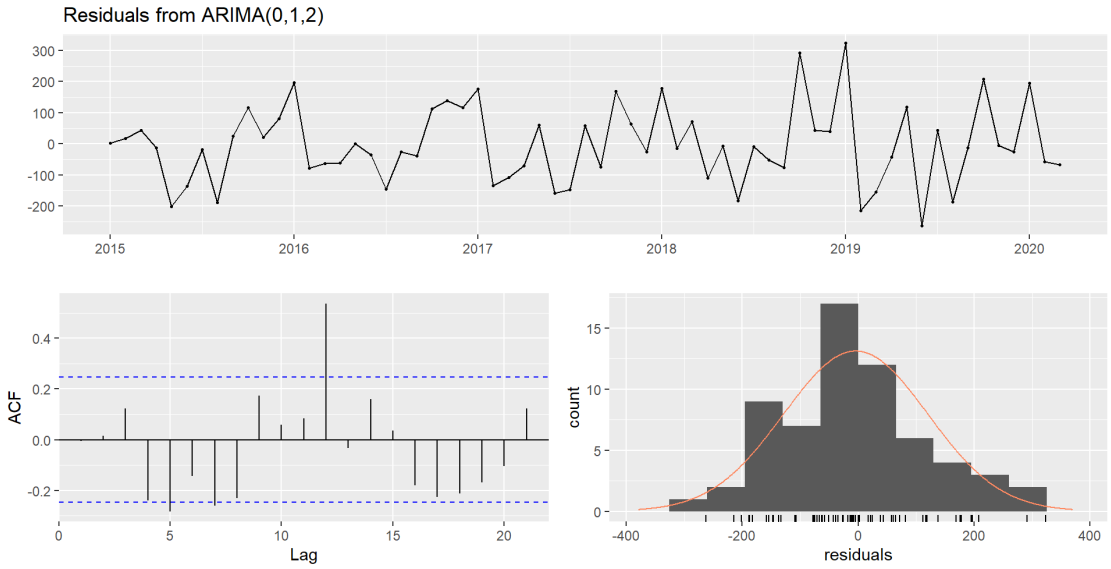


Рис. 9. Поведінка залишків моделі ARIMA(0, 1, 2)

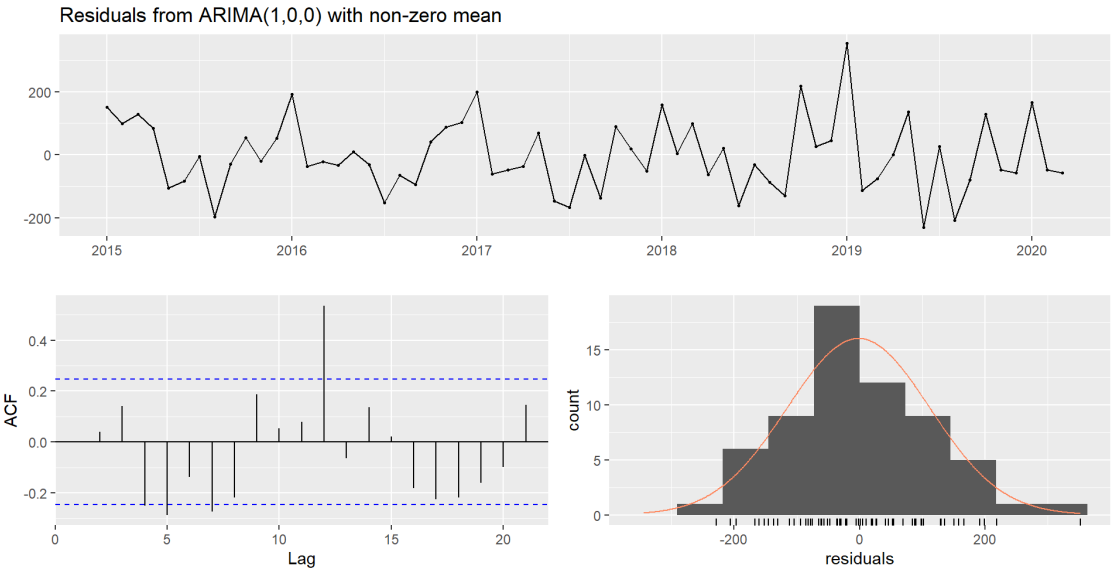


Рис. 10. Поведінка залишків моделі ARIMA(1, 0, 0)

Бачимо, що поведінка залишків обох моделей майже однакова, хоча для третьої дані на гісторгамі мають розподіл більш подібний до нормального.

Також бачимо, що в корелограмах для залишків є значущий 12 лаг, тобто спостерігається періодичність тривалість в рік. Тому побудуємо ще сезонну ARIMA модель – SARIMA(0, 0, 1)(1, 1, 0)^{12}.

Для неї отримаємо наступні результати:

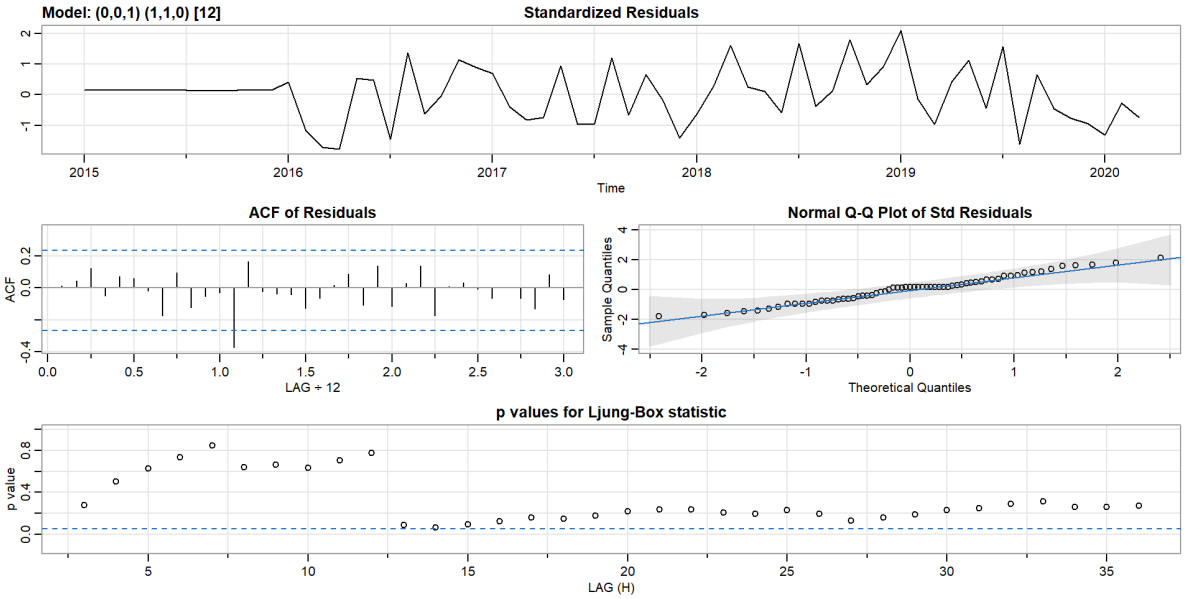


Рис. 11. Результати для сезонної авторегресійної моделі

Бачимо, що для сезонної моделі поведінка залишків краща, ніж для звичайної. Тому за нульову модель візьмемо специфікацію $SARIMA(0,0,1)(1,1,0)^{12}$.

Скористаємось методологією «баєсівських структурних моделей часових рядів», яка в даному випадку є представленням ARIMA моделі.

Враховуючи результати розвідувального аналізу даних, наша базова прогнозна модель буде включати тренд, а також сезонну та авторегресійну компоненти. Крім того, оскільки дисперсія в даних має достатньо стабільний характер, залишимо прийняту за замовчуванням адитивну сезонність.

Як слідує з наведених нижче кількісних показників якості отриманої моделі, вона добре описує навчальні дані (про це говорить високий коефіцієнт детермінації rsquare):

```
> summary(m0)
$residual.sd
[1] 0.09128429

$prediction.sd
[1] 0.3443769

$rsquare
[1] 0.9579429

$relative.gof
[1] 0.5452379
```

Рис. 12 ілюструє оцінений вплив окремих компонент в прогнозовані значення з навчальної вибірки. Видно, що основний вплив належить тренду. Впливи сезонної та авторегресійної компонент набагато нижчі, але в більшості спостережень вони відмінні від нуля.

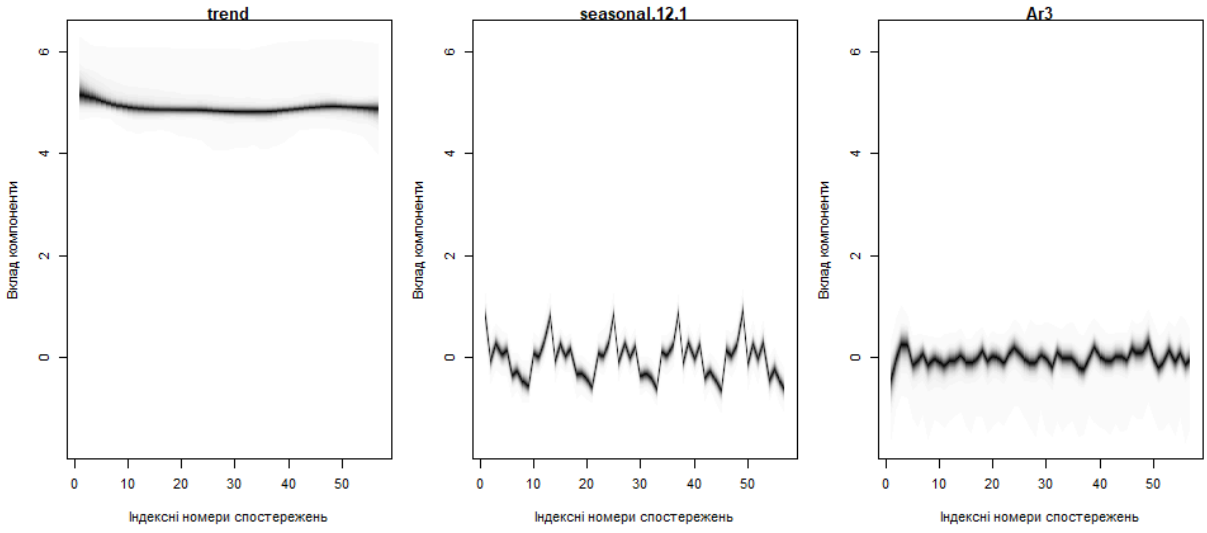


Рис. 12. Апостеріорні розподіли компонент моделі m0 (зліва направо: тренд, сезонна та авторегресійна компоненти)

3.4.2. Побудова MA моделі

Як слідує з Рис. 3 з корелограмою, автокореляція в даних має помірний характер і є сенс побудувати ще одну, простішу модель, виключивши авторегресійну компоненту, тобто MA(1) модель:

```
> summary(m1)
$residual.sd
[1] 0.2014163

$prediction.sd
[1] 0.3529647

$rsquare
[1] 0.7952439

$relative.gof
[1] 0.526466
```

3.5. Порівняння результатів моделей

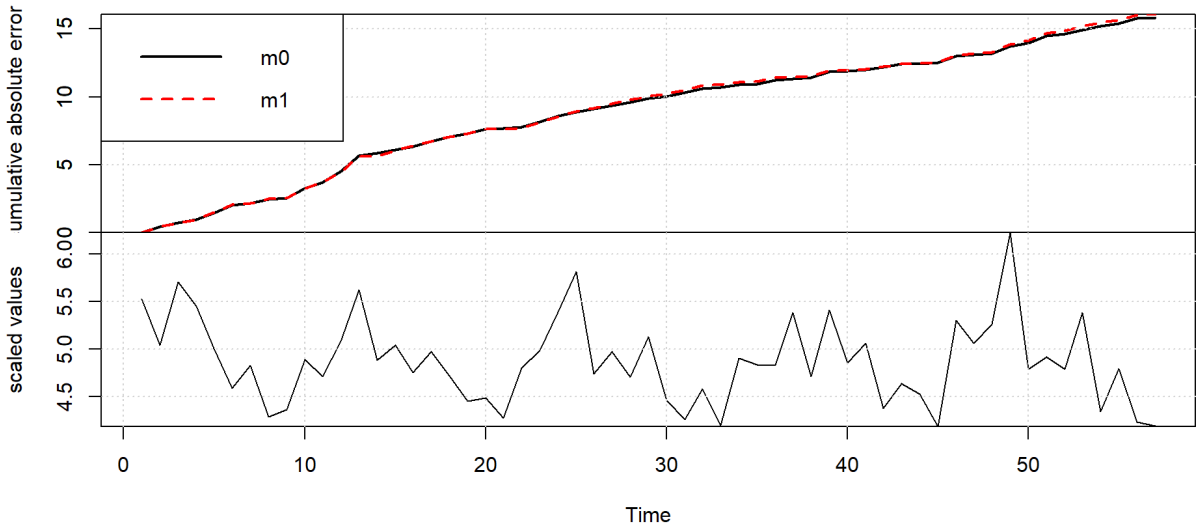


Рис. 13. Порівняння якості апроксимації навчальних даних моделями m0 і m1 за допомогою «Помилок наступного кроку».

Вгорі: накопичені середні абсолютні помилки наступного кроку. Знизу: навчальні дані. Більш низький рівень помилок, накопичених моделлю m0, говорить про її більш високу якість.

До цього моменту ми оцінювали якість побудованих моделей по тому, наскільки добре вони описували навчальні дані. Безумовно, такий підхід супроводжується високим ризиком вибрати в якості оптимальної перенавчену модель.

Також порівняємо коефіцієнти детермінації і побачимо, що в моделі m0 він вищий:

```
> summary(m0)
$residual.sd
[1] 0.09128429

$prediction.sd
[1] 0.3443769

$rsquare
[1] 0.9579429

$relative.gof
[1] 0.5452379
```

```
> summary(m1)
$residual.sd
[1] 0.2014163

$prediction.sd
[1] 0.3529647

$rsquare
[1] 0.7952439

$relative.gof
[1] 0.526466
```

Діагностика з використанням помилок наступного кроку (Рис. 13.) лише частково допомагає застрахуватися від цього і єдиним об'єктивним тестом якості моделі завжди буде точність її пророкувань на незалежному наборі даних.

Щоб виконати такий незалежний тест скористаємося моделями m0 і m1 і розрахуємо прогнозні значення смертності на наступні 6 місяців, які відповідають тимчасовому періоду створеної нами раніше перевірочної вибірки (з жовтня 2019 року по грудень 2021 року). Отримані прогнозні значення далі порівняємо з фактичними в цей період рівнями смертності. Для вимірювання якості прогнозів скористаємося середньою абсолютною помилкою:

```
  m0   m1
2094 1929
```

Модель m1 помилилася в своїх прогнозах для перевірочної вибірки в середньому на 1929 смертей, що набагато менше помилки, допущеної моделлю

m0. За браком додаткових предикторів, які потенційно могли б ще більше підвищити якість прогнозів m1, будемо вважати, що структура цієї моделі оптимальна для апроксимації аналізованих даних.

3.6. Підгонка моделі

Об'єднаймо тепер дані з навчальної та перевірочної вибірок і підгонимо остаточну модель для прогнозування смертності. Обчислимо прогнозні рівні смертності на період з квітня 2020 по грудень 2021 року з допомогою отриманої моделі m_final і зобразимо результат графічно (Рис. 14):

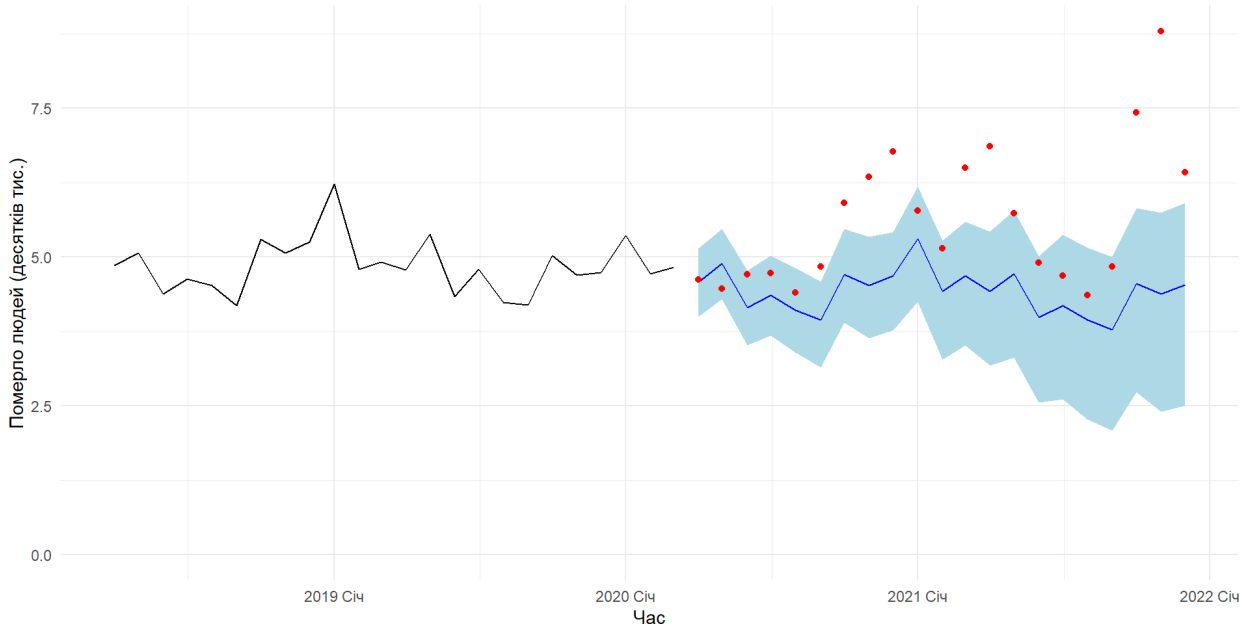


Рис. 14. Прогноз природного рівня місячної смертності в Україні в квітні 2020 - грудні 2021 року (синя лінія), який мав би місце за відсутності спалаху COVID-19. Світло-блакитна смуга навколо синьої лінії відповідає 95% -вій довірчій області отриманих прогнозів. Чорна крива зображує історичні спостереження. Червоними крапками показані фактично спостережені рівні смертності під час пандемії.

На Рис. 10. добре видно, що зареєстровані в дійсності рівні смертності (показані червоними крапками), особливо в періоди спалаху захворюваності на COVID -19 набагато перевищують прогнозні значення (синя лінія).

Віднявши від загальної кількості смертей офіційні дані по смертності саме від COVID-19 (Додаток В), на Рис. 15 зможемо побачити «чисту» надлишкову смертність, яка може бути пов’язана з «недовиявленістю» захворювань а також з рядом інших причин таких як: ненадання вчасної медичної допомоги, зміни способу життя (стреси, зменшення фізичної активності, зміни харчової поведінки, тощо) та відсутності вчасного профілактичного огляду.

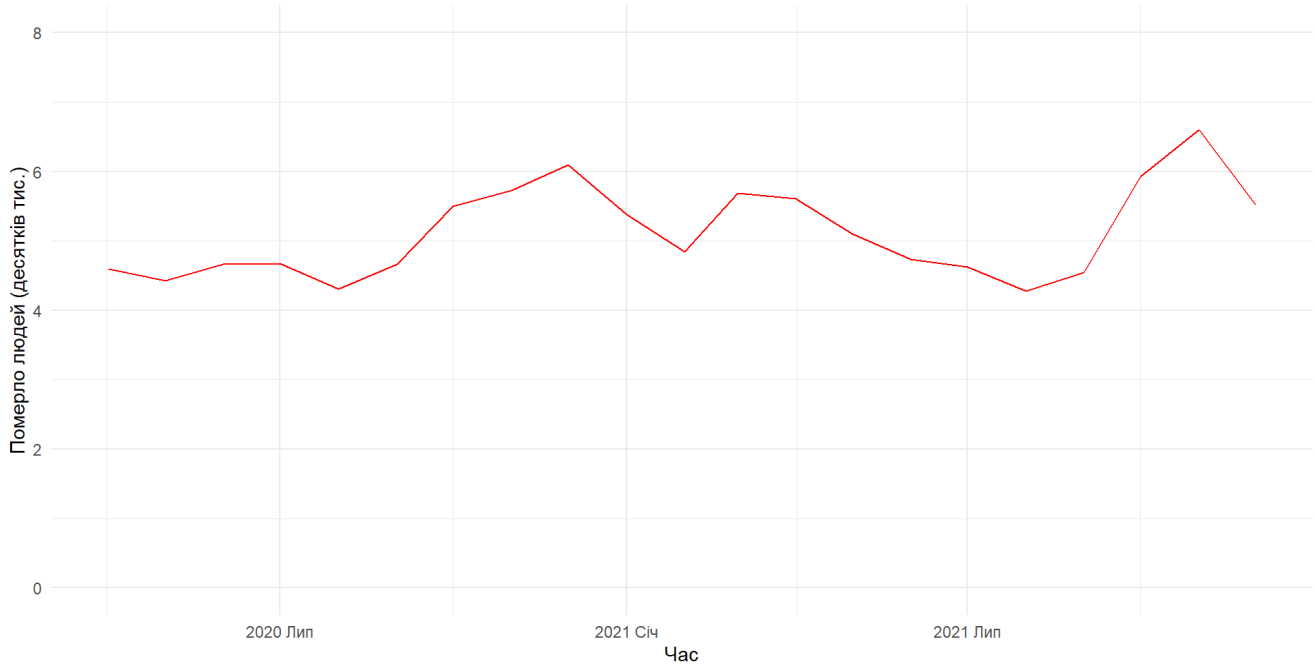


Рис. 15. «Чиста» надлишкова смертність за період пандемії COVID-19, з квітня 2020 по грудень 2021

Висновок. Пандемія COVID-19 сильно вплинула на збільшення смертності в Україні в період з квітня 2020 року по грудень 2021 року. Особливо помітний цей вплив в період з вересня по грудень 2020, з березня по травень 2021 та з жовтня по грудень 2021, там показники реальної смертності дуже суттєво перебільшують дані, які ми отримали від прогнозу. Також бачимо високий рівень «чистої» надлишкової смертності, що несе за собою тяжкі наслідки для України в майбутньому, враховуючи демографічну кризу в країні протягом останніх десятиліть.

ВИСНОВКИ

Пандемія коронавірусу COVID-19, яка викликала сильну епідеміологічну кризу і застала зненацька всі країни світу, безумовно ввійде в історію як одна з найвпливовіших подій початку XXI сторіччя. Ця криза виявила тотальну непідготовленість світової спільноти до подібного роду ситуацій. Протягом останніх двох років перед усіма країнами стоїть задача забезпечити себе актуальними даними та навчитися ними оперувати, і звісно - прогнозувати для більш ефективного та швидкого реагування та прийняття правильних рішень. І хоча більшість країн, завдяки дотриманню карантинних обмежень та вакцинації населення, вже нормалізували епідеміологічну ситуацію, проте ми все ще можемо спостерігати нові спалахи захворювань та появу нових штамів вірусу.

Отже, питання оцінки епідеміологічної ситуації та її прогнозування, питання надання медичної допомоги з інших проблем, організація планово-профілактичного лікування під час пандемії все ще залишаються актуальними. Особливо прогноз цих даних, як часового ряду, є актуальним для України, через демографічну кризу, що спостерігається протягом останніх десятиліть. Знання методів аналізу та вміння їх застосувати є необхідною складовою підготовки аналітиків.

Метою цієї роботи було дослідження надлишкової смертності від COVID-19 за допомогою методів прогнозування часових рядів.

В процесі виконання були виконані наступні завдання:

- Вивчення та опрацювання літератури, присвяченої часовим рядам та методам їх аналізу і прогнозу;
- Пошук релевантні даних щодо кількості померлих в Україні за 2015-2021 роки та кількість померлих саме від COVID-19;
- Створення програмну реалізацію прогнозування часових рядів на реальних даних за допомогою мови програмування R;
- Оцінка «чистої» надлишкової смертності в Україні під час пандемії COVID-19 та аналіз причин її виникнення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Avril Coghlan, A little book of R for Time series, release 0.2, 2018.
2. Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with Exponential Smoothing.. Springer Berlin Heidelberg.
3. Ord, K., Fildes, R., Kourentzes, N., 2017. Principles of Business Forecasting, 2nd ed.. Wessex Press, Inc, New York, New York, USA.
4. Hyndman, R.J., & Athanasopoulos, G. 2018 Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia.
5. G.E.P. Box, G. Jenkins, “Time Series Analysis, Forecasting and Control”, Holden-Day, San Francisco, CA, 1970.
6. Браун, Роберт Г. (1956). Експоненціальне згладжування для прогнозування попиту . Кембридж, Массачусетс: Arthur D. Little Inc.
7. Калехар, Праякта С. «Прогнозування часових рядів за допомогою експоненціального згладжування Холта–Вінтерса», 2014.
8. Гамільтон, Джеймс (1994). Аналіз часових рядів . Видавництво Прінстонського університету.
9. Чумаченко Д.І., Чумаченко Т.О. Математичні моделі та методи прогнозування епідемічних процесів, Харків, 2020.
10. Майборода Р. Є. Комп’ютерна статистика – професійний старт : підручник. Київ : Університетська книга, 2020.
11. Юрченко М. Є. Прогнозування та аналіз часових рядів. Методичні вказівки до практичних занять та самостійної роботи студентів, 2018.
12. Шипунов А.Б., Балдін Е.М. … Наглядная статистика. Используем R!, 2014.
13. Мастицький С. Е. Анализ временных рядов с помощью R, 2020.
14. Афанасьєв В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование, 2001.

ДОДАТОК А.

Year	Month	Value
2021	January	57721
2021	February	51391
2021	March	64938
2021	April	68621
2021	May	57293
2021	June	49077
2021	July	46851
2021	August	43538
2021	September	48425
2021	October	74282
2021	November	87862
2021	December	64264
2020	January	53610
2020	February	47128
2020	March	48326
2020	April	46223
2020	May	44680
2020	June	47043
2020	July	47297
2020	August	44008

2020	September	48372
2020	October	59047
2020	November	63438
2020	December	67663
2019	January	62196
2019	February	47891
2019	March	49177
2019	April	47879
2019	May	53791
2019	June	43396
2019	July	47930
2019	August	42359
2019	September	41914
2019	October	50245
2019	November	46986
2019	December	47350
2018	January	53851
2018	February	47088
2018	March	54088
2018	April	48567
2018	May	50627
2018	June	43796

2018	July	46359
2018	August	45236
2018	September	41870
2018	October	52999
2018	November	50623
2018	December	52561
2017	January	58151
2017	February	47356
2017	March	49737
2017	April	47050
2017	May	51274
2017	June	44565
2017	July	42566
2017	August	45806
2017	September	41966
2017	October	49054
2017	November	48332
2017	December	48266
2016	January	56236
2016	February	48796
2016	March	50367
2016	April	47510

2016	May	49717
2016	June	47181
2016	July	44485
2016	August	44871
2016	September	42760
2016	October	47994
2016	November	49849
2016	December	53865
2015	January	55287
2015	February	50428
2015	March	57059
2015	April	54514
2015	May	50070
2015	June	45851
2015	July	48291
2015	August	42867
2015	September	43642
2015	October	48919
2015	November	47087
2015	December	50781

ДОДАТОК Б.

```
require(dplyr)
require(readr)
require(tsibble)
require(feasts)
require(lubridate)
require(ggplot2)
require(bsts)
require(tidyr)
require(RColorBrewer)
require(forecast)
require(astsa)

dat <- read_csv("D:/курсовая/дані/smertnist2.csv") %>%
  setNames(., c("y"))

dat_ts <- ts(dat, start = c(2015, 1), frequency = 12)

daily_avg_dat <- dat_ts/monthdays(dat_ts)

autoplot(daily_avg_dat)+
    xlab("Рік") + ylab("Середньоденна кілікість померлих")
```

```
ggseasonplot(daily_avg_dat)+
  ggtitle("Сеонна діаграма")+
  xlab("Місяць") + ylab("Середньоденна кілікість померлих")

ggAcf(daily_avg_dat)+
  ggtitle("Корелограма")

ggPacf(daily_avg_dat)+
  ggtitle("Часткова корелограма")

train.dat <- window(daily_avg_dat, start = c(2015, 1), end = c(2020, 3))
valid.dat <- window(daily_avg_dat, start = c(2020, 4))

y <- train.dat

snaive <- snaive(y, h=21)

autoplot(snaive, PI=FALSE)+
  ggtitle("Сезонний наївний прогноз") +
  xlab("Рік") + ylab("Середньоденна кілікість померлих")
```

```
snaive.last <- window(y, start = c(2019, 1))


autoplot(diff(snaive.last)) +
  autolayer(snaive(diff(snaive.last), h=22),
   PI=FALSE) +
  ggtitle("Сезонний наївний прогноз") +
  xlab("Рік") + ylab("Середньоденна кількість померлих")


checkresiduals(snaive(diff(y)))


auto.arima(daily_avg_dat)


model1 <- arima(x=y, order=c(0,1,2), method="ML")
model1


model2 <- arima(x=y, order=c(1,0,0), method="ML")
model2


checkresiduals(model1)
checkresiduals(model2)


model3 <- arima(x=y, order=c(1,1,1), method="ML")
```

```
model3

checkresiduals(model3)

model4 <- sarima(log(y), 0, 0, 1, P=1, D=1, Q=0, S=12)


dat <- read_csv(«D:/курсовая/дані/smertnist1.csv») %>%

  setNames(., c(«year», «month», «y»))


dat <- dat %>%

  mutate(dm = paste(year, substr(month, 1, 3)) %>% yearmonth(.)) %>%

  as_tsibble(., index = dm) %>%

  fill_gaps() %>%

  mutate(y = y /10000) %>%

  dplyr::select(y, dm)


pre_covid <- dat %>%

  filter(dat$dm < yearmonth(«2020 Apr»))


n_pre_covid <- nrow(pre_covid)

train <- pre_covid[1 :(n_pre_covid – 6), ]

valid <- pre_covid[(n_pre_covid – 5):n_pre_covid, ]

test <- dat %>% dplyr::filter(dat$dm >= yearmonth(«2020 Apr»))
```

```
ggplot(dat, aes(dm, y)) +
  geom_line() +
  geom_smooth(se = FALSE) +
  theme_minimal() +
  ylim(c(0, 8)) +
  labs(x = «Час», y = «Померло людей (десятків тис.)»)

train %>% gg_season(y, pal = brewer.pal(6, «OrRd»)) +
  ylim(c(0, NA)) +
  theme_minimal() +
  labs(x = «Місяць року», y = «Померло людей (десятків тис.)»)

train %>%
  mutate(y = y) %>%
  gg_lag(geom = «point», lags = 1:6,
         col = «black», alpha = 0.4) +
  theme_minimal()
y <- train$y

ss <- list()
ss <- AddLocalLinearTrend(ss, y)
ss <- AddSeasonal(ss, y, nseasons = 12)
```

```
ss <- AddAutoAr(ss, y, lag = 3)
m0 <- bsts(y, ss, niter = 2000, ping = 0, seed = 42)

summary(m0)

par(mar = c(5.1, 4.1, 1, 1))
plot(m0, ylab = «Померло людей (десятків тис.)»,
     xlab = «Індексні номери спостережень»)

plot(m0, y = «components», same.scale = TRUE,
     ylab = «Вклад компоненти»,
     xlab = «Індексні номери спостережень»)

ss <- list()
ss <- AddLocalLinearTrend(ss, y)
ss <- AddSeasonal(ss, y, nseasons = 12)
m1 <- bsts(y, ss,
           niter = 2000, ping = 0, seed = 42)
summary(m1)

models_to_compare <- list(«m0» = m0, «m1» = m1)
CompareBstsModels(models_to_compare, colors = c(«black», «red»))
```

```
mae <- function(observed, predicted){
  mean(abs(observed – predicted))
}

sapply(list(«m0» = (predict(m0, horizon = 6) %>% .$median)*10000,
            «m1» = (predict(m1, horizon = 6) %>% .$median)*10000),
       mae, observed = valid$y * 10000) %>% round()

y_full <- c(y, valid$y)

ss <- list()
ss <- AddLocalLinearTrend(ss, y_full)
ss <- AddSeasonal(ss, y_full, nseasons = 12)
ss <- AddAutoAr(ss, y_full, lag = 3)
m_final <- bsts(y_full, ss,
                niter = 2000, ping = 0, seed = 42)

m_final_pred <- predict(m_final, horizon = 21)
point_predictions <- tibble(dm = test$dm,
                            y = m_final_pred$median,
                            ll95 = m_final_pred$interval[1, ],
```

```
                    ul95 = m_final_pred$interval[2, ]) %>%
  as_tsibble(., index = dm)

bind_rows(train, valid) %>%
  dplyr::filter(dm >= yearmonth(«2018 Apr»)) %>%
  ggplot(., aes(dm, y)) + geom_line() +
  geom_ribbon(data = point_predictions,
              aes(ymin = ll95, ymax = ul95),
              fill = «lightblue») +
  geom_line(data = point_predictions, col = «blue») +
  geom_point(data = test, aes(dm, y), col = «red») +
  ylim(c(0, NA)) +
  theme_minimal() +
  labs(x = «Час», y = «Померло людей (десятків тис.)»)

data <- read_csv(«D:/курсовая/дані/smert_covid2020-2021.csv») %>%
  setNames(., c(«year», «month», «y»))

data <- data %>%
  mutate(dm = paste(year, substr(month, 1, 3)) %>% yearmonth(.)) %>%
  as_tsibble(., index = dm) %>%
  fill_gaps() %>%
```

```
  mutate(y = y /10000) %>%
  dplyr::select(y, dm)

nadlyshok <- data.frame(test$dm, test$y – data$y)

ggplot(nadlyshok, aes(test.dm, test$y – data$y)) +
  geom_line(col = «red») +
  theme_minimal() +
  ylim(c(0, 8)) +
  labs(x = «Час», y = «Померло людей (десятків тис.)»)
```

ДОДАТОК В.

Year	Month	Value
2021	January	3922
2021	February	3014
2021	March	8104
2021	April	12545
2021	May	6327
2021	June	1762
2021	July	599
2021	August	804
2021	September	2937
2021	October	15068
2021	November	21867
2021	December	9066
2020	April	243
2020	May	420
2020	June	449
2020	July	525
2020	August	928
2020	September	1735
2020	October	4050
2020	November	6184
2020	December	6750