НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
Факультет прикладної математики
Кафедра прикладної математики

Курсовий проект
із дисципліни «Алгоритми і системи комп’ютерної математики»
На тему
«Прогнозування кількості хворих на COVID-19»
Етап №2

Виконав:
студент групи КМ-93
Костенко О. А.

Керівник:
доцент
Олефір О. С.

Київ — 2022

Опис обраного математичного методу

Підхід авторегресійного інтегрованого ковзного середнього (англ. autoregressive integrated moving average, ARIMA) до часових рядах полягає в тому, що в першу чергу оцінюється стаціонарность ряду. Різними тестами виявляються наявність поодиноких коренів і порядок інтегрованості часового ряду (зазвичай обмежуються першим або другим порядком). Далі, при необхідності, (якщо порядок інтегрованості більше нуля) ряд перетворюється взяттям різниці відповідного порядку і вже для перетвореної моделі будується деяка ARMA-модель, оскільки передбачається, що отриманий процес є стаціонарним, на відміну від вихідного нестаціонарного процесу (разностно-стаціонарного або інтегрованого процесу порядку d).

ARIMA - модель і методологія аналізу часових рядів. Є розширенням моделей ARMA для нестаціонарних часових рядів, які можна зробити стаціонарними взяттям різниць деякого порядку від вихідного часового ряду (так звані інтегровані або різносно-стаціонарні часові ряди). Модель $ARIMA(p,d,q)$ означає, що різниці часового ряду порядку d належать моделі $ARMA(p,q)$.

Модель $ARIMA(p,d,q)$ для нестаціонарного часового ряда Xt має вигляд:,

$$\Delta^{d}= c + \varepsilon_t + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}\ ,$$

де X_t – це зпрогнозоване значення на період t;

c – константа, зазвичай для спрощення рівняється нулю;

a_i – параметри моделі, коефіцієнти авторегресії;

θ_i – параметри моделі, ковзного середнього;

Δ^d – оператор різниці часового ряду порядку d (послідовне взяття d раз різниць першого порядку - спочатку від тимчасового ряду, потім від отриманих різниць першого порядку, потім від другого порядку і т.д.)

ε_t – білий шум.

Також, дана модель інтерпретується як $ARMA(p+d,q)$ модель з d одиничними розв'язками. При d = 0 маємо звичайну ARMA модель.

ARIMA-моделі дозволяють моделювати інтегровані або разностностаціонарні часові ряди (DS-ряди, diference stationary). Часовий ряд називається інтегрованим порядку k, якщо різниці ряду порядку Δkxt, тобто є стаціонарними, в той час як різниці меншого порядку (включаючи нульового порядку, тобто сам тимчасової ряд) не є стаціонарними щодо деякого тренда рядами (TS-рядами, trend stationary). В зокрема $I(0)$ – це стаціонарний процес. Порядок інтегрованості часового ряду і є порядок d моделі.

Методологія Бокса-Дженкінса. Методологія Бокса-Дженкінса підбору ARIMA моделі для даного ряду спостережень складається з 5 кроків.

Крок 1. Отримання стаціонарного ряду. Ми тестуємо ряд на стаціонарність, використовуючи описані вище методи: візуальний аналіз графіка, візуальний аналіз ACF і PACF, тести на одиничні коріння. Якщо виходить стаціонарний ряд, то переходимо до наступного пункту, якщо немає стаціонарності ряду, то застосовуємо оператор взяття послідовної різниці і повторюємо тестування. Напрактиці послідовна різниця береться, як правило, не більше двох разів.

Крок 2. Після того, як отримано стаціонарний часовий ряд, будуються його вибіркові ACF і PACF, які, як було показано вище, є своєрідними «відбитками пальців» ARMA(p, q) процесу і дозволяють сформулювати кілька гіпотез про можливі порядки авторегресії p і змінного середнього q. Зазвичай рекомендується використовувати моделі можливо більш низького порядку, як правило, з p q 3 (якщо немає сезонної компоненти).

Крок 3. Для кожної з обраних на першому етапі моделей оцінюються їх параметри і обчислюються залишки.

Крок 4. Кожна з моделей перевіряється, наскільки вона відповідає даним. З моделей, адекватних даними, вибирається найпростіша модель, тобто з найменшою кількістю параметрів.

Крок 5. Прогнозування. Після того, як обрана модель, можна будувати прогноз на один або кілька кроків за часом і оцінювати довірчі інтвервали прогнозованих значень.

Модель $SARIMA(p,d,q$, P, D, Q) - seasonal autoregressive integrated moving average, дуже схожа на модель $ARIMA(p,d,q)$, за винятком того, що вона бере до уваги сезонність даних, з чого випливає додатковий набор компонентів авторегресії та ковзного середнього, що компенсуються частотою сезонності:

$$\Delta^d = c + \varepsilon_t + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{P} \boldsymbol{\varphi_i \Delta^d X_{t-si}} + \sum_{i=1}^{Q} \boldsymbol{\eta_i \varepsilon_{t-si}}$$

Модель $SARIMA$X(p,d,q, r, P, D, Q), окрім сезонності, також враховує й екзогенні змінні, іншими словами, використовує зовнішні данні в прогнозі:

$$\Delta^{d} = c + \varepsilon_t + \sum_{i=1}^{p} a_i \Delta^{d} X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{P} \varphi_i \Delta^{d} X_{t-si} + \sum_{i=1}^{Q} \eta_i \varepsilon_{t-si} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{r}} \boldsymbol{\beta_i \epsilon_{i_t}}$$

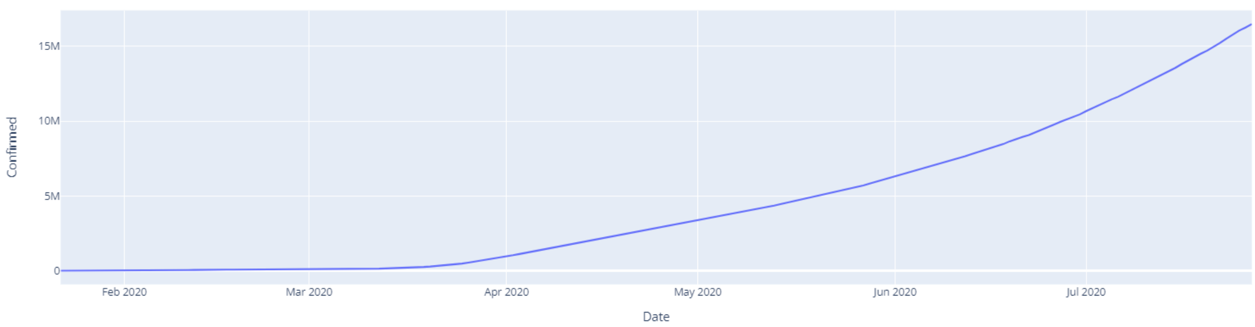
Контрольний приклад

Завантажимо дані про COVID-19 з файлу .csv. Залишивши в ньому тільки інформацію про дату та кількість хворих, отримуємо часовий ряд кількості хворих на COVID-19.

Розглянемо вхідні дані. Перші 10 записів файлу:

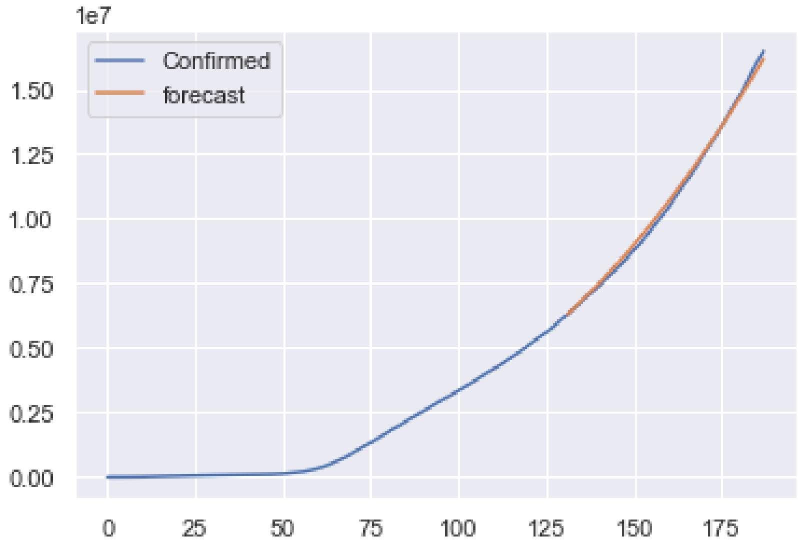
	Date	Confirmed
0	2020-01-22	555
1	2020-01-23	654
2	2020-01-24	941
3	2020-01-25	1434
4	2020-01-26	2118
5	2020-01-27	2927
6	2020-01-28	5578
7	2020-01-29	6166
8	2020-01-30	8234
9	2020-01-31	9927

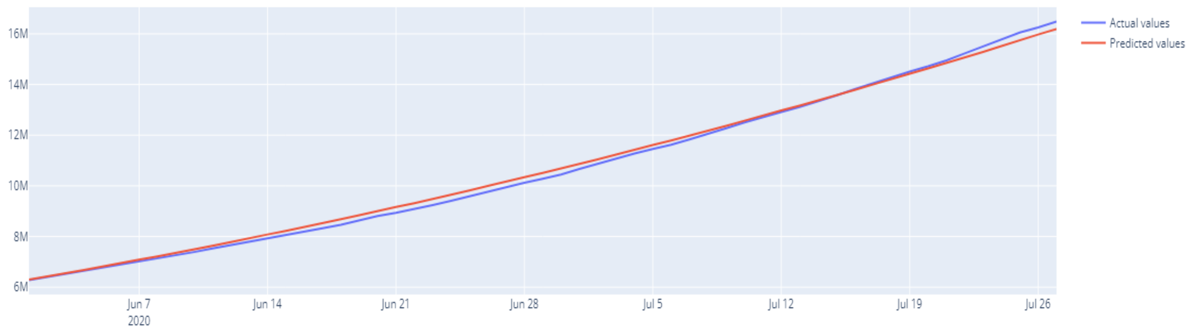
Побудова завантажених даних:



Для прогнозування кількості хворих на COVID-19 використаємо бібліотеку *statsmodels*, а саме модель *SARIMAX*. Розділимо дані на тренувальну та тестову вибірки з відношенням 70:30 та навчимо модель на тренувальних даних.

Результат прогноз моделі з урахуванням тестової виборки:


1e7
Confirmed
forecast
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
125
150
175


16M
14M
12M
10M
8M
6M
Jun 7
2020
Jun 14
Jun 21
Jun 28
Jul 5
Jul 12
Jul 19
Jul 26
Actual values
Predicted values