

РЕГРЕСІЙНІ МОДЕЛІ ДЛЯ ДОСЛІДЖЕННЯ ЯКОСТІ ПОВІТРЯ

О. М. Залеська¹, Г. О. Яйлимова¹

¹Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
ІН Фізико-технічний інститут

Анотація

За допомогою моделей авторегресії і ковзного середнього побудовано прогноз концентрації забруднюючих частинок в атмосфері на основі даних про якість повітря у місті Києві. Для цього було досліджено залежності між різними ознаками набору даних, здійснено попередню обробку та аналіз даних, обрано модель і реалізовано її, а також оцінено якість цієї моделі.

Ключові слова: машинне навчання, авторегресія, аналіз якості повітря, часові ряди

Вступ

У сучасному світі забруднення повітря є однією із глобальних проблем, адже розвиток технологій, розростання міст і збільшення виробництва певної продукції сприяють більшим викидам шкідливих речовин у повітря. За останні роки спостерігалося зростання кількості автомобілів у м. Києві, що теж погіршувало стан повітря. Через забруднення повітря відбуваються зміни клімату, збільшується кількість хворих на ішемічну хворобу серця і людей, що страждають на різні хвороби органів дихання. Тому майже в неперервному режимі відбувається моніторинг концентрації шкідливих речовин у повітрі. Важливо також аналізувати зміни показників, що відповідають за концентрацію, із часом, щоб передбачити стан повітря у майбутньому. Так можна буде запобігти погіршенню стану або мінімізувати можливі погані наслідки.

1. Авторегресійні моделі для прогнозування часових рядів

1.1. Основні поняття, що стосуються часових рядів

Часовим рядом називається послідовність значень, які вимірювалися через певні однакові проміжки часу [1].

Виділяють декілька компонент, із яких складаються часові ряди:

- Тренд – зміна рівня часового ряду із часом, тобто плавне зростання або спадання ряду.
- Сезонність – періодична зміна рівня часового ряду.
- Помилка – випадкова величина, яка виникає через похибки вимірювання, зовнішні непередбачувані фактори.

Як визначено у статті [2], часовий ряд називається стаціонарним в широкому сенсі, якщо математичне сподівання та дисперсія є сталими і коваріація між значеннями ряду у момент часу t і у момент часу

$(t + k)$ не залежить від часу, тобто залежить лише від значення k .

Для перевірки ряду на стаціонарність можна подивитись на зміну математичного сподівання та дисперсії та наочно оцінити, чи є ряд стаціонарним, або користуватися тестами на стаціонарність. Одним з таких тестів є тест Дікі-Фуллера [3]. Цей тест використовує статистику Дікі-Фуллера. Рівняння авторегресії для часового ряду y має вигляд:

$$y_t = a * y_{t-1} + \varepsilon_t. \quad (1)$$

Перепишемо 1 у вигляді:

$$y_t - y_{t-1} = a * y_{t-1} + \varepsilon_t - y_{t-1} = (a - 1) y_{t-1} + \varepsilon_t.$$

Перевіряється гіпотеза $H_0 : (a - 1) = 0$. Якщо ця гіпотеза підтверджується, то ряд не є стаціонарним. Якщо $|a| < 1$, то ряд стаціонарний.

Автокореляційною функцією (ACF – autocorrelation function), згідно з [4], називається функція кореляції між значеннями часового ряду та його зміщенням на k кроків за часом:

$$acf(y, k) = corr(y_t, y_{t-k}).$$

Значення ACF рахується за допомогою коефіцієнтів парної кореляції Пірсона, які визначаються так:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Іншою важливою функцією для аналізу часових рядів є часткова автокореляційна функція (PACF – partial autocorrelation function), що врахує усі проміжкові кореляції. Формально PACF записується так [5]:

$$pacf = corr(y_t, y_{t-k} | y_{t-1}, \dots, y_{t-k+1}).$$

1.2. Моделі AR, MA, ARMA, ARIMA, SARIMA

Одним із підходів до прогнозування часових рядів є використання моделей авторегресії (AR –

autoregressive) і ковзного середнього (MA – moving average), а також їх модифікації [6].

Для зручності запису у роботі [7] вводиться поняття оператора лагу L , який визначається так:

$$L^k y_t = y_{t-k}. \quad (2)$$

AR базується на звичайній лінійній регресії, де для пошуку параметрів w_j використовується метод найменших квадратів. Проте замість залежності цільової ознаки від інших використовують залежність значення у момент часу t від значень у моменти часу $t-1, t-2, \dots, t-p$, де p – параметр моделі.

$$y_t = \sum_{j=1}^p w_j y_{t-j} + \varepsilon_t = \sum_{j=1}^p w_j L^j y_t + \varepsilon_t.$$

MA використовує для прогнозування значення функції через помилки, які були зроблені на попередньому кроці. Рівняння для цієї моделі виглядає так:

$$y_t = \mu + \sum_{j=1}^q b_j \varepsilon_{t-j} + \varepsilon_t = \mu + \sum_{j=1}^q b_j L^j \varepsilon_t + \varepsilon_t.$$

Суть моделі ковзного середнього полягає в тому, що для оцінки значення функції на певному кроці часу ми беремо середнє значення μ і додаємо значення помилок на попередній кроках, помножені на коефіцієнти, таким чином поліпшуючи модель.

ARMA (autoregressive moving average) перекладається як «модель авторегресійного ковзного середнього» і базується на побудові функції як поєднання авторегресії і ковзного середнього [8]. Це виражається наявністю у рівнянні двох поліномів.

$$y_t = \sum_{j=1}^p w_j y_{t-j} + \sum_{j=1}^q b_j \varepsilon_{t-j} + \varepsilon_t + c,$$

де p – порядок першого поліному, q – порядок другого поліному, w_i – параметри авторегресійного поліному, b_j – параметри поліному ковзного середнього, ε_t – випадкова похибка, c – деяка константа.

Якщо у ряду є тренд, то зробити його стаціонарним можна за допомогою диференціювання. Це враховує модель ARIMA (autoregressive integrated moving average) [6]. Крім параметрів, що приймає попередня модель, на вхід ARIMA також подається порядок диференціювання, необхідний для того, щоб позбутися тренду і щоб ряд став стаціонарним.

$$u_t = y_t - y_{t-1}. \quad (3)$$

Модель SARIMA (Seasonal ARIMA) використовується для рядів із наявними сезонністю і трендом. Для того, щоб позбавитися сезонності, використовують різниці часового ряду і його зсуву на m кроків по часу, де m – довжина сезону.

$$u_t = y_t - y_{t-m}. \quad (4)$$

Для того, щоб описати модель SARIMA, скористаємося лаговим оператором L . Модель приймає 7 параметрів: p, d, q, P, D, Q, m . Перші три відповідають параметрам ARIMA, наступні 3 – аналоги цих параметрів для сезонності, а m – довжина сезону. Нехай u_t – ряд, вже приведений до стаціонарного,

тобто:

$$u_t = (1 - L)^d (1 - L^m)^D y_t.$$

Тоді загальний вигляд моделі буде таким [7]:

$$(1 - \sum_{i=1}^p w_i L^i)(1 - \sum_{i=1}^P W_i L^{mi})(u_t - \mu) = (1 + \sum_{i=1}^q \theta_i L^i)(1 + \sum_{i=1}^Q \vartheta_i L^{mi})\varepsilon_t.$$

2. Прогнозування показника PM2.5 із використанням моделі SARIMA

2.1. Постановка задачі

Задача полягає у прогнозуванні на деякий час у майбутнє рівня забрудненості повітря, тобто потрібно визначити залежності даних від попередніх і на основі цього побудувати модель, яка б враховувала ці залежності. Для вирішення таких задач існує багато підходів, зокрема побудова авторегресійних моделей. Дані взято із сайту [9]. У датасеті містяться такі дані:

- phenomenon – показник, який вимірюється;
- value – значення вимірюваної величини;
- logged_at – точний час, коли відбувалися вимірювання.

У стовпці phenomenon містяться 6 характеристик: вологість повітря (humidity), температура (temperature), тиск (pressure_pa), концентрація PM1 (pm1), PM10 (pm10), PM2.5 (pm25) у повітрі.

PM (particulate matter) – маленькі частинки, які є забруднювачами повітря (пил, бруд, дим та інші). Вони відрізняються своїм діаметром.

Індекс якості повітря (ІЯП) є зручним способом для надання інформації про забрудненість повітря [10]. Це перетворення показників, що характеризують якість повітря, у більш зручний вигляд. Частинки з діаметром 2.5 мікрметри (PM2.5), як зазначається у [11], є найшкідливішими для організму і зазвичай ІЯП розраховується якраз для цього показника. Тому його можна вважати цільовою ознакою.

2.2. Попередня обробка і аналіз даних

У таблиці 1 наведено дані у початковому вигляді.

Табл. 1. Перші 3 рядки з набору даних

	phenomenon	value	logged at
0	pm1	17.0000	2020-12-01 00:08:10
1	pm25	22.6842	2020-12-01 00:08:10
2	pm10	25.4737	2020-12-01 00:08:10

Попередня обробка даних є необхідною для можливості подальшої роботи з даними. Вона включає у себе переведення даних у зручний формат, приведення до необхідних типів даних, масштабування/стандартизацію, пошук та заміну або видалення пропущених значень у стовпцях, пошук та заміну або видалення викидів у даних, а також візуалізацію. У таблиці 2 наведено дані, для яких здійснено вищеперелічені перетворення.

Табл. 2. Перші 3 рядки з нового (відформатованого) набору даних

logged at	humidity	pm1	pm10	pm25	pressure pa	temperature
2020-12-01 00:08:10	-0.842644	0.789140	0.415845	0.518428	1.344273	-0.605259
2020-12-01 00:13:04	-0.838620	0.573316	0.760338	0.729244	1.340463	-0.597894
2020-12-01 00:16:25	-0.831112	0.699994	0.694579	0.770710	1.340346	-0.600399

Дослідимо залежність між значенням ознаки PM2.5 та часом, так як ця ознака є цільовою. Спробуємо виявити якісь залежності. Для цього зобразимо середнє по годинах та по днях тижня значення ознаки PM2.5 (рис. 1).

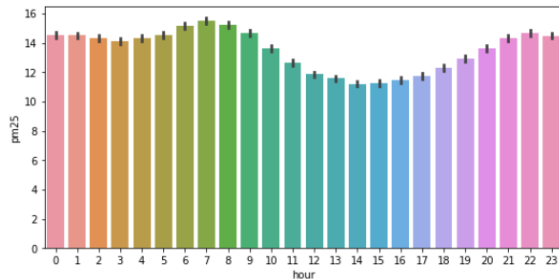


Рис. 1. Середні значення PM2.5 по годинах

Бачимо, що середнє значення приблизно о 14:00-15:00 найменше, найвище значення – о 7:00 ранку, і ще один пік випадає на 22:00. Ймовірно, що це пов'язано із зростанням кількості машин у години, коли люди їдуть на роботу або з роботи.

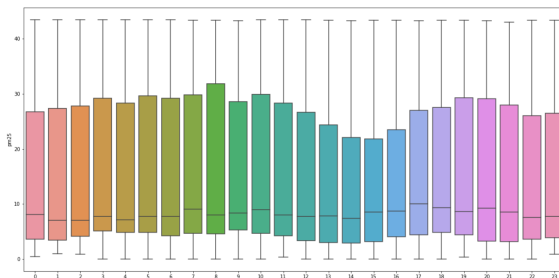


Рис. 2. Boxplot для PM2.5 по годинах

Максимальні і мінімальні значення, а також медіани, для кожної години приблизно однакові. Для 14:00-15:00 значення PM2.5 75% даних знаходяться нижче 20, у той час як для 8:00 та 20:00 – нижче 30. Власне, бачимо ту саму тенденцію, яку виявили із середніх значень.

2.3. Дослідження часового ряду

Будемо розглядати як часовий ряд усереднені значення стовпця PM2.5 за кожні 2 години. Графік залежності показника від часу зображено на рисунку 3.

Можна помітити присутність аномальних значень. Знайдемо їх за допомогою розкладу ряду на тренд, сезонність та помилку. Будемо вважати аномаліями ті значення, які значно відхиляються від поєднаних сезонної компоненти та тренду, тобто спостерігається велика помилка.

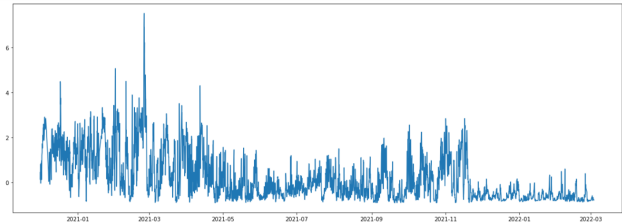


Рис. 3. Графік залежності PM2.5 від часу

Нехай res_i – значення помилки вимірювання i . Тоді аномаліями будуть вважатися всі вимірювання i , для яких $res_i \notin [m - 3\sigma, m + 3\sigma]$, де m – математичне очікування помилки, σ – стандартне відхилення (правило сігм). На рисунку 4 зображені точки, які за даним правилом визначилися як аномальні, помаранчевим кольором – ряд після видалення цих точок.

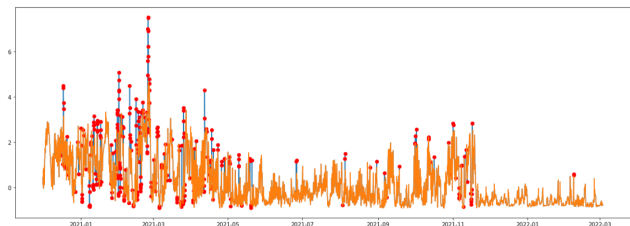


Рис. 4. Аномальні точки і ряд після їх заміни

Так як наявний нелінійний тренд, то ряд не є стаціонарним. Щоб спробувати привести цей ряд до стаціонарного, від даних цього ряду віднімемо ряд, зміщений на крок назад за формулою 3. Зобразимо отриманий ряд на рисунку 5.

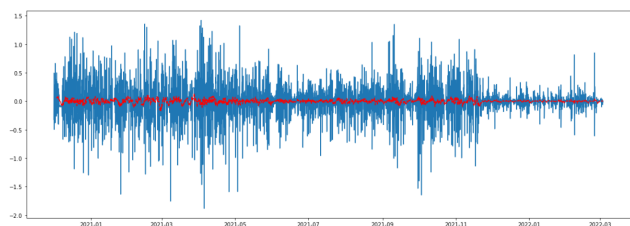


Рис. 5. Графік залежності PM2.5 від часу після першого диференціювання

Для того, щоб зрозуміти, чи наявна сезонність, та для вибору параметрів моделей використовують ACF та PACF. Зобразимо значення цих функцій для нового ряду (рис. 6).

Бачимо велику кореляцію на значенні 12. Це означає, що дані корелюють з тим, що було 24 години тому. На 48-и, 72-х, 96-и годинах кореляція зменшується, але все ще залишається досить високою. Тому

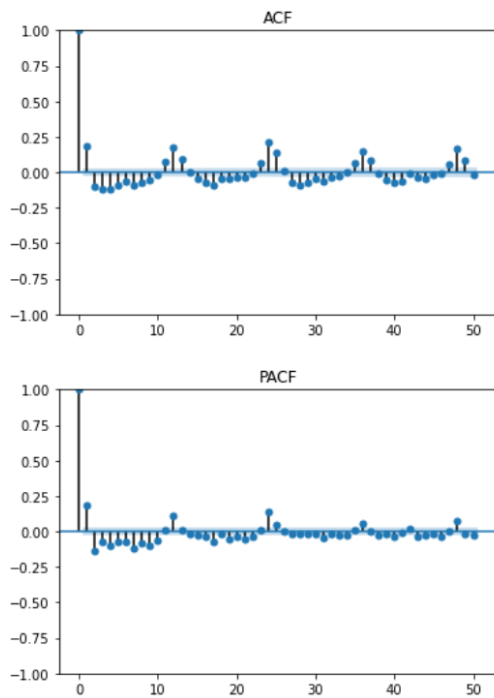


Рис. 6. Значення ACF та PACF для ряду після першого диференціювання

потрібно також позбутися сезонності за формулою 4. Для цього ряду гіпотеза у тесті Дікі-Фуллера про стаціонарність ряду не відхиляється, тобто справді ряд є стаціонарним.

Отже, для прогнозування значень ряду будемо використовувати SARIMA ((p, 1, q)(P, 1, Q), 12).

2.4. Побудова моделі

Будемо будувати моделі на значеннях PM2.5 до 2021-07-01 16:00:00 і перевіряти на даних з 2021-07-01 18:00:00 по 2021-07-03 0:00:00, тобто спробуємо побудувати прогноз на 30 годин вперед.

У мовах програмування Python та R наявні функції, що дозволяють перебрати можливі параметри моделі і визначити найкращі. Визначена таким чином модель – SARIMA ((5,1,0), (2,1,0), 12). Побудуємо прогноз і зобразимо на 7 спрогнозовані і реальні дані. Для оцінки якості побудова-

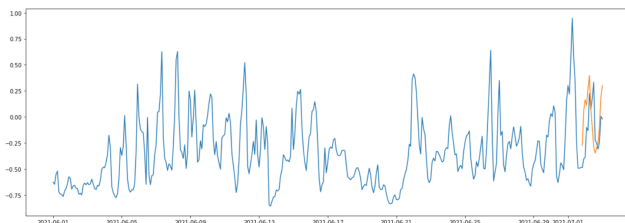


Рис. 7. Прогноз показника PM2.5 на 30 годин

ної моделі визначимо середньоквадратичну помилку (MSE) за формулою $MSE = 1/n \sum_{i=1}^n (y_{test} - y_{pred})^2$ на тестовій та навчальній вибірках: на тестовій – 0.09663392974962698, на навчальній –

0.10378716299643699. Тобто модель побудувала доволі гарний прогноз.

Висновки

Для дослідження було обрано дані з однієї зі станцій моніторингу якості повітря в м. Києві. Для цього набору даних виявлено залежності між різними ознаками, а також проаналізовано часовий ряд зі значеннями концентрації забруднюючих частинок PM2.5 в атмосфері. На основі цього аналізу обрано авторегресійну модель для прогнозування значень цього показника у майбутнє.

За допомогою моделі SARIMA побудовано прогноз PM2.5 у атмосфері на 30 годин. Модель виявилася досить якісною, проте існують способи покращення моделі та інші підходи (наприклад, адаптивні методи побудови авторегресійних моделей, нейронні мережі тощо).

Перелік використаних джерел

1. Complete Guide on Time Series Analysis in Python[Електронний ресурс]. — Режим доступу: <https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python/notebook>.
2. Стаціонарні часові ряди [Електронний ресурс]. — Режим доступу: https://stud.com.ua/179685/prirodoznastvo/statsionarni_chasovi_ryadi.
3. Dickey-Fuller test[Електронний ресурс]. — Режим доступу: https://uk.upwiki.one/wiki/Dickey\T2A\textendashFuller_test.
4. Time Series Analysis[Електронний ресурс]. — Режим доступу: https://www.youtube.com/playlist?list=PLvcbyUQ5t0UHOlnBz146_Q6QKtFgfMGc3.
5. Understanding Partial Auto-correlation And The PACF [Електронний ресурс]. — Режим доступу: <https://timeseriesreasoning.com/contents/partial-auto-correlation/>.
6. Towards Data Science - Time Series Models[Електронний ресурс]. — Режим доступу: <https://towardsdatascience.com/time-series-models-d9266f8ac7b0>.
7. Seasonal ARIMA models[Електронний ресурс]. — Режим доступу: <https://online.stat.psu.edu/stat510/lesson/4/4.1>.
8. Statistics How To - ARMA model[Електронний ресурс]. — Режим доступу: <https://www.statisticshowto.com/arma-model/>.
9. SaveEcoBot[Електронний ресурс]. — Режим доступу: <https://www.saveecobot.com>.
10. Beijing Air Pollution: Real-time Air Quality Index (AQI) [Електронний ресурс]. — Режим доступу: <https://aqicn.org>.
11. Undark - The Weight of Numbers: Air Pollution and PM2.5[Електронний ресурс]. — Режим доступу: <https://undark.org/breathtaking/>.