

УДК 004

Колосов К.Є.¹, Подковаліхіна О.О.²

¹ студ. гр. КНТ-221 НУ «Запорізька політехніка»

² доц. НУ «Запорізька політехніка»

ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДІВ ВІДНОВЛЕННЯ ВІДСУТНІХ ДАНИХ МОНІТОРИНГУ ЯКОСТІ ПОВІТРЯ НА ОСНОВІ ДАНИХ КОНЦЕНТРАЦІЇ PM_{2.5}

Об'єкт дослідження – процес обробки та аналізу даних моніторингу якості повітря.

Предмет дослідження – методи та алгоритми машинного навчання та статистичного аналізу для відновлення пропущених даних у часових рядах концентрації PM_{2.5}.

Мета роботи – дослідження та порівняльний аналіз методів відновлення пропущених даних у часових рядах концентрації PM_{2.5} з метою підвищення точності відновлення.

Матеріали, методи та технічні засоби: структурне та об'єктно-орієнтоване програмування, поліноміальна регресія, інтерполяція сплайнами (PCHIP та згладжуючий сплайн), ARIMA, SARIMA, Random Forest, мова програмування Python, бібліотеки pandas, numpy, matplotlib, scikit-learn, statsmodels, scipy, tkinter, персональний комп'ютер з процесором Intel Core i7-12700F під управлінням операційної системи Microsoft Windows 10.

У межах дослідження було реалізовано й порівняно шість методів відновлення пропущених значень у часових рядах концентрацій PM_{2.5} на основі реальних даних моніторингу повітря у Запоріжжі, Дніпрі та Львові: поліноміальна регресія третього ступеня, інтерполяція сплайнами (PCHIP і згладжуючий сплайн), ARIMA, SARIMA, Random Forest. Якість відновлення оцінювалася за метрикою RMSE.

На основі проведеного аналізу створено програмний застосунок для Windows із графічним інтерфейсом. Він дозволяє завантажувати та обробляти часові ряди, обирати метод відновлення, візуалізувати результати (графіки, таблиці) та експортувати їх для подальшого аналізу. Реалізацію виконано на Python із використанням бібліотек scikit-learn, statsmodels, tkinter та Matplotlib.

Було проведено систематичне порівняння ефективності різних методів відновлення. Для цього в даних штучно моделювали пропуски, що дозволило оцінити точність кожного підходу на основі RMSE. Найкращі результати в усіх наборах продемонструвала поліноміальна регресія третього ступеня, що дозволяє вважати її найбільш ефективною для досліджуваних даних. А саме відновлення пропущених значень має вирішальне значення для подальшого аналізу, зокрема:

- підвищення точності прогнозування – забезпечує безперервність ряду для достовірних моделей;
- коректне виявлення трендів і сезонності – сприяє точному аналізу динаміки концентрацій;
- краща візуалізація та інтерпретація – забезпечує цілісне уявлення про дані;
- можливість застосування складних моделей – є необхідною умовою для роботи більшості сучасних алгоритмів.

У межах проведеного дослідження було здійснено глибокий аналіз та порівняння ефективності різних методів відновлення пропущених значень у часових рядах концентрацій твердих частинок PM_{2.5}, що є критично важливим аспектом екологічного моніторингу. Завдяки використанню реальних даних із трьох міст України (Запоріжжя, Дніпро, Львів) вдалося досягти високої репрезентативності результатів, що забезпечує релевантність висновків для практичного застосування.

Використання таких методів, як поліноміальна регресія, інтерполяція сплайнами, ARIMA, SARIMA та Random Forest, дало змогу отримати комплексну оцінку їхньої ефективності. За результатами розрахунків середньоквадратичної помилки (RMSE), найкращі показники точності продемонстрував метод поліноміальної регресії третього ступеня. Це дає підстави вважати його найбільш доцільним і практичним для задач відновлення даних у даній галузі.

Практичним результатом дослідження стало створення спеціалізованого програмного забезпечення для Windows, яке поєднує зручність графічного інтерфейсу та потужні інструменти аналізу. Застосунок дозволяє обирати метод відновлення, обробляти користувацькі набори даних, візуалізувати результати та експортувати їх для подальшого використання. Розробка реалізована мовою Python із використанням сучасних бібліотек, що гарантує гнучкість і масштабованість рішення.

Загалом, результати дослідження підтверджують важливість якісного заповнення пропущених значень у часових рядах, зокрема через:

- підвищення точності прогнозів;
- точніше виявлення трендів і сезонності;
- покращену інтерпретацію результатів;

– розширення можливостей для застосування сучасних методів машинного навчання.

Таким чином, представлений підхід має не лише теоретичну цінність, а й вагоме практичне значення для фахівців у галузі екологічного моніторингу, аналітики даних та розробки програмного забезпечення.

Галузь використання – сфера екологічного моніторингу, аналітики якості повітря, урбаністики, а також у наукових дослідженнях, пов’язаних із аналізом часових рядів та екологічним прогнозуванням