ISSN 2306-4412 (Print), ISSN 2306-4455 (CD-ROM), ISSN 2708-6070 (Online)

УДК 004.67:519.25

[0000-0001-6302-565X] **А. Г. Батурінець,**
e-mail: baturinets.anastasiya@gmail.com
[0000-0001-6611-4543] **С. В. Антоненко,** *канд. техн. наук, доцент*
e-mail: svitlanav.antonenko@gmail.com
Дніпровський національний університет імені Олеся Гончара
просп. Гагаріна, 72, м. Дніпро, 49010, Україна

ПОДОВЖЕННЯ РЯДІВ ДАНИХ ЗА ЗНАЧЕННЯМИ ПОКАЗНИКІВ СХОЖИХ РЯДІВ

Проблема недостатності інформації суттєво впливає на вибір підходів та методів аналізу рядів даних, а також на якість отримуваних результатів. Зважаючи на таку проблему, автори роботи вважають, що актуальним є питання розробки й аналізу підходів та моделей для подовження рядів даних. Основною задачею є описання та реалізація технології подовження рядів даних. В основу реалізації технології закладено використання значень схожих рядів даних як ознак для подовження певного ряду даних, представленого тими ж показниками, що й схожі ряди даних. В роботі описано схему визначення схожих рядів даних. Згідно з цією схемою найбільш схожими рядами даних є такі, що мають найменше значення відстані та сильний прямий кореляційний зв'язок, обчислені між потенційно схожим рядом та рядом, для якого буде відбуватися подовження. Для подовження ряду розглядаються сім моделей. За результатами обчислювального експерименту встановлено, що найкращі результати отримано при використанні двох моделей: суми зважених значень по групі схожих рядів та середньозважених значень по групі схожих рядів, з коригуванням на середнє значення ряду, для якого виконується подовження. В результаті проведеного аналізу можна дійти висновку про можливість використання розробленої технології для вирішення задачі подовження рядів даних. При подальших дослідженнях планується використання отриманих результатів для розробки та аналізу методів поповнення пропущених значень у часових рядах.

Ключові слова: *часові ряди, регресія, поповнення даних, машинне навчання, sklearn, недостатність даних, гідрологія.*

Вступ. Своєчасний, точний та якісний аналіз вхідних даних є важливою складовою успішного розв'язання великої кількості прикладних задач у різних галузях.

Все частіше для розв'язання прикладних задач використовуються більш складні регресійні моделі та методи машинного навчання. Приклади використання можна побачити в роботах: [1] – прогнозування вартості легкових авто, [2] – прогнозування повеней, [3] – прогнозування потоків та рівнів води, [4] – прогнозування вартості криптовалют, в [5] використовуються для прогнозування цін на нерухомість. В роботі [6] представлено огляд та класифікацію моделей прогнозування. Роботи [7-9] присвячені питанням прогнозування гідрологічної інформації.

Проблема коротких часових рядів та малих вибірок трапляється достатньо часто в задачах аналізу даних, зокрема висвітлена в роботі [10]. Недостатність інформації суттєво впливає на вибір методів аналізу та якість отримуваних результатів. Окрім цього, наявність рядів даних різної довжини викликає необхідність або обирати методи, пристосовані для аналізу рядів різної довжини, або скорочувати довжину всіх рядів даних до однакового розміру, або моделюванням подовжувати ряди спостережень. Слід також зазначити, що неможливо однозначно вказати, якої саме довжини ряди даних вважати короткими, а які – ні. Наприклад, у гідрологічних розрахунках короткими рядами вважаються всі ряди, що не відповідають принципам репрезентативності та точності статистичних оцінок, тобто короткими рядами спостережень можуть вважатися часові ряди з показниками як за десять років, так і за двадцять.

Зважаючи на вищевикладене, актуальним є питання не лише розробки методів та підходів аналізу коротких рядів даних, але й методів та моделей подовження рядів даних та їх програмної реалізації.

DOI: 10.24025/2306-4412.3.2021.244266

Мета та задачі дослідження. Основною метою роботи є дослідження можливостей подовження рядів даних за значеннями показників схожих рядів, реалізація та аналіз технології подовження рядів.

Для досягнення поставленої мети необхідно вирішити наступні задачі:

1) визначити схожі ряди, значення яких будуть використовуватися в моделях. Для вирішення цієї задачі пропонується застосувати обчислювальну схему, де визначення схожості рядів даних проводиться на підставі обчислення значень відстані та коефіцієнта кореляції;

2) реалізувати моделі, які будуть використовуватися для подовження ряду. Для виконання поставленої задачі варто реалізувати сім моделей: лінійної множинної регресії; суми зважених значень по групі схожих рядів; середньозважених значень по групі схожих рядів, з коригуванням на середнє значення ряду, для якого виконується подовження; випадкового лісу; *k*-найближчих сусідів; методу опорних векторів; градієнтного бустінгу;

3) забезпечити представлення результатів роботи технології у вигляді графіків, таблиць, значень розрахованих оцінок, що надасть можливість провести аналіз роботи технології.

Виклад основного матеріалу. Ідея застосування цього підходу полягає в тому, що моделям регресії як ознаки, за якими обчислюється відповідне значення для подовження ряду, передаються значення схожих рядів. Таким чином, першочергово для подовження деякого ряду T необхідно визначити найбільш схожі ряди даних.

Визначення схожих рядів виконується за схемою:

1) обчислюються відстані між рядом, для якого відбуватиметься подовження, та рядами, значення яких потенційно можуть бути використані як факторні ознаки для моделі регресії;

2) на підставі отриманих результатів у п. 1 за кожним запропонованим типом відстані обираються k найближчих рядів (мають найменші значення відстані). Таким чином, отримуємо кількість множин, що дорівнює кількості використовуваних відстаней;

3) далі знаходимо перетин множин, отриманих на етапі виконання п. 2;

4) для набору рядів, що залишилися після виконання п. 3, та ряду, для якого буде виконуватися подовження, розраховуємо коефіцієнти кореляції Пірсона та відкидаємо ряди, для яких значення коефіцієнта кореляції є меншим ніж 0,75.

Незалежно від кількості представлених для порівняння рядів, в результаті буде отримано набір рядів у кількості, що не перевищує k. Всі відібрані ряди даних матимуть відносно невеликі, порівняно з іншими, значення відстаней та сильний прямий кореляційний зв'язок з рядом, для якого визначаються схожі ряди.

Якщо в результаті визначення схожих рядів за наведеною схемою отримано порожню множину, тоді необхідно повторити обчислення, змінивши один або декілька наступних параметрів:

– набір використовуваних відстаней;

– період спільних спостережень (зменшити/збільшити);

– значення k.

У роботі для подовження ряду використовуються моделі випадкового лісу (далі RFR) [11], k-найближчих сусідів (далі KNR) [12], методу опорних векторів (далі SVR) [13] та градієнтного бустінгу (далі GBR) [14], реалізованих з використанням бібліотеки sklearn на python [15]. Перераховані моделі є досить відомими в задачах прогнозування та групування рядів даних.

Крім того, реалізовано й більш прості в обчисленні моделі: множинна лінійна регресія, сума зважених значень по групі схожих рядів та модель, побудована за середньозваженими значеннями по групі, з коригуванням на середнє значення ряду, для якого виконується подовження.

Введемо позначення: T – ряд, для якого проводиться подовження; S – схожий ряд даних; $k = \overline{1,N}$, де N – кількість схожих рядів даних; $\bar{T}$ – середнє значення ряду T на навчальному відрізку; $\bar{S}_k$– середнє значення певного k-го ряду даних на навчальному відрізку; $i = \overline{1,M}$, де M – кількість значень, на які буде подовжено ряд T.

Множинна лінійна регресія (далі LR) представлена формулою

$$T_i = a_0 + S_{1,i} * a_1 + S_{2,i} * a_2 + \cdots + S_{N,i} * a_N,$$

де a_k – параметри регресії, $k = \overline{0,N}$.

DOI: 10.24025/2306-4412.3.2021.244266

Модель суми зважених значень по групі схожих рядів (далі SWG):

$$T_i = S_{1,i} * w_1 + S_{2,i} * w_2 + \cdots + S_{N,i} * w_N,$$

де $w_i \in [0; 1]$ – вага складової кожного схожого ряду, розрахована за формулою

$$w_i = \frac{q_i}{\sum_{i=1}^{N} q_i}, \quad (1)$$

де $\sum_{i=1}^{N} w_i = 1$, а q_i розраховано за формулою

$$q_i = \sum_{k=1}^{N} MSE - MSE_k, \quad (2)$$

де $\sum_{k=1}^{N} MSE$ – середнє значення похибки MSE, розрахованої між кожним порівнюваним рядом S та рядом T, MSE_k – похибка MSE на періоді спільних спостережень між S_k та T; похибка MSE, розрахована за формулою

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(T_i - S_i)^2. \quad (3)$$

Завдяки ваговим коефіцієнтам w_i з використанням формул (1-3) вплив значень кожного ряду на результат є обернено пропорційним значенню розрахованої оцінки MSE на періоді спільних спостережень між обраними схожими рядами та рядом, для якого відбувається подовження.

Для випадку, коли подовження рядів проводиться за значеннями двох схожих рядів даних, формула розрахунку w_i може бути спрощена та представлена в наступному вигляді:

$$w_i = \frac{\sum_{i=1}^{N} MSE - MSE_i}{\sum_{i=1}^{N} MSE}.$$

Модель, побудована на середньозважених значеннях по групі схожих рядів, з коригуванням на середнє значення ряду, для якого виконується подовження (далі SWGC), має наступний вигляд:

$$T_i = \left(\frac{S_{1,i}}{\bar{S_1}} * w_1 + \frac{S_{2,i}}{\bar{S_2}} * w_2 + \cdots + \frac{S_{N,i}}{\bar{S_N}} * w_N\right) * \bar{T}.$$

Для оцінки отриманих результатів використовуються значення похибок MSE, MAE, MAPE та коефіцієнта детермінації R^2.

Результати досліджень. Для проведення обчислювального експерименту обрано ряди гідрологічних даних, що представлені показниками рівнів води на постах спостережень басейну річки Дніпро. Потрібно виконати подовження ряду даних посту 79545, розташованого на річці Случ, м. Новоград-Волинський, Житомирської обл., за період з 01.01.2010 р. по 31.12.2010 р., тобто виконати подовження на 365 значень.

Період спільних спостережень – з 01.01.2000 р. по 31.12.2009 р. На рисунку 1 представлено 94 ряди даних, серед яких проводиться пошук схожих, та ряд посту 79545.

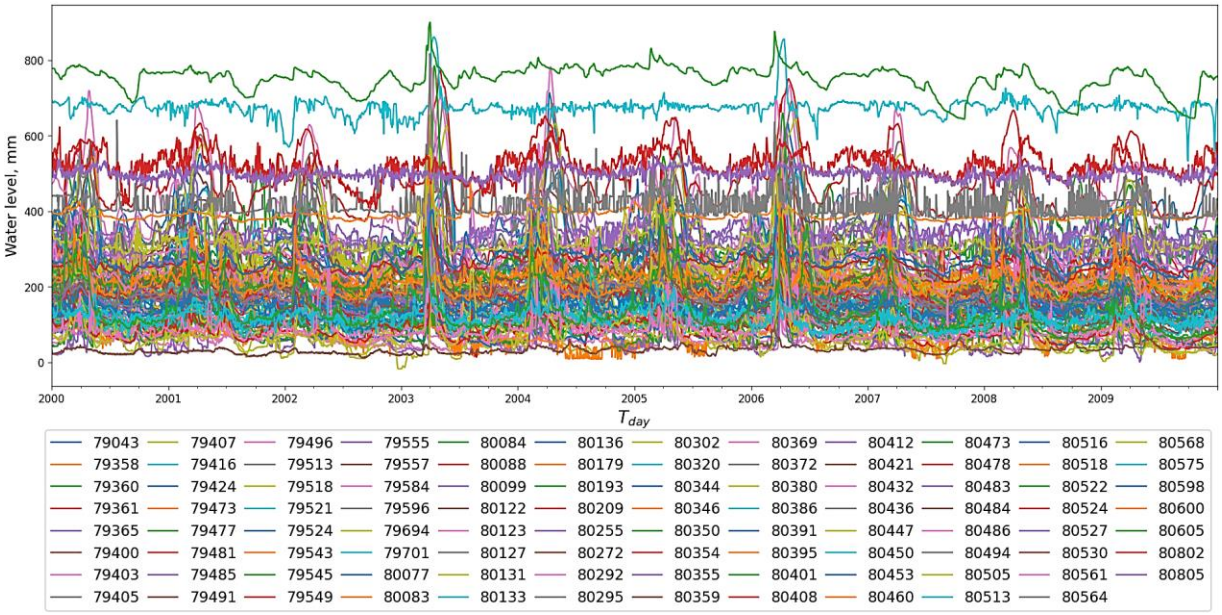

Рисунок 1 – Показники рівнів води, зафіксовані на гідрологічних постах басейну р. Дніпро

DOI: 10.24025/2306-4412.3.2021.244266

Першочергово необхідно визначити ряди, які мають найменші значення розрахованих відстаней, і лише для обраних рядів встановлювати наявність та силу кореляційного зв'язку.

За значеннями евклідової, манхеттенської та евклідової зваженої відстаней, розрахованих на вихідних значеннях рядів, визначено, що схожими гідрологічними рядами є дані постів 79694, 79555, 80344, 79596. Період спільних спостережень заданий інтервалом з 01.01.2000 р. по 31.12.2009 р. Далі, за розрахованими значеннями коефіцієнтів кореляції визначено, що ряди показників постів 80344 та 79596 не мають сильного кореляційного зв'язку із рядом показників посту 79545, водночас значення коефіцієнтів кореляції для постів 79694 і 79555 становлять 0,77 та 0,82 відповідно. Таким чином, для подовження ряду даних посту 79545 будуть використовуватися значення постів 79694 та 79555. Слід також зазначити, що пости 79545, 79694 і 79555 розташовані на території Житомирської області, тобто мають відносно близьке географічне розташування.

На рисунку 2 відображено ряди посту 79545 та постів, визначених як найбільш схожі.

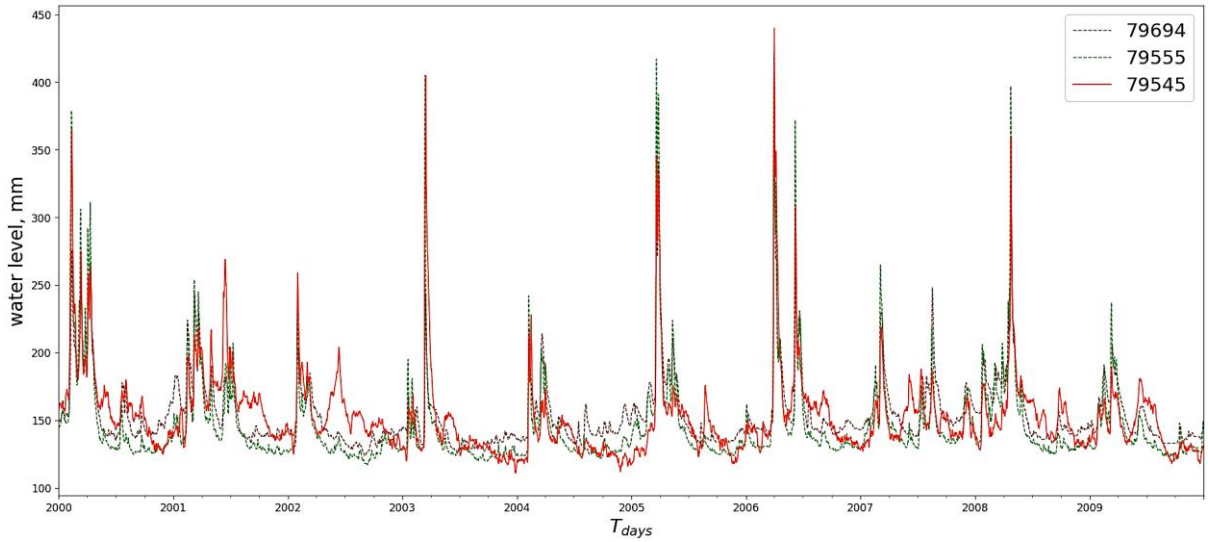

Рисунок 2 – Значення рівнів води за період спільних спостережень

В таблиці 1 наведено параметри налаштування моделей з бібліотеки sklearn [11-14].

Таблиця 1 – **Параметри моделей**

Модель	Параметри	Значення
RFR	max_features; criterion; n_estimators	sqrt; mse; 100
SVR	kernel; degree; cache_size	linear; 1; 200
GBR	learning_rate; n_estimators; validation_fraction	0.1; 100; 0.05
KNR	algorithm ; metric; n_neighbors	auto; euclidean; k^2

Таблиця 2 містить оцінки якості наведених моделей, де ■ – найкращі значення за відповідною оцінкою, ■ – найгірші значення за відповідною оцінкою (по стовпцях), *min* *max* – шкала допустимих значень.

Таблиця 2 – **Оцінки якості моделей**

Модель\ Оцінка	Оцінки			
	MSE	$MAPE, \%$	MAE	R^2
LR	184,57	6,50	10,27	0,77
RFR	429,76	11,17	17,04	0,46
KNR	454,52	11,28	17,37	0,43
SVR	192,11	6,29	10,10	0,76
GBR	306,54	9,41	14,43	0,62
SWG	169,40	5,83	9,63	0,79
SWGC	160,31	5,98	9,72	0,80

На рисунку 3 представлено дані ряду посту 79545 та результати подовження за всіма використовуваними моделями за період з 01.01.2010 р. по 31.12.2010 р.

DOI: 10.24025/2306-4412.3.2021.244266

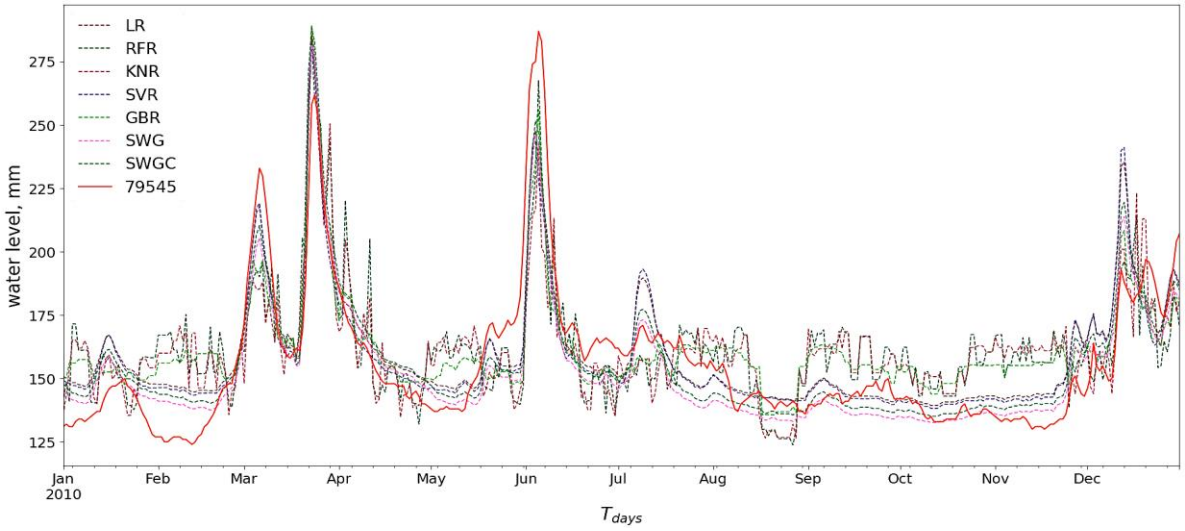

Рисунок 3 – Результати подовження ряду показників посту 79545

Рівняння моделей LR, SWG, SWGC мають наступний вигляд:

– модель множинної лінійної регресії:

$$T_i = 34.83 + S_{1,i} \times 0.1145 + S_{2,i} \times 0.6897,$$

де $S_{1,i}$ – значеннння ряду посту 79694, $S_{2,i}$ – значення ряду посту 79555;

– модель SWG:

$$T_i = S_{1,i} * 0{,}5168 + S_{2,i} * 0{,}4832;$$

– модель SWGC:

$$T_i = \left(\frac{S_{1,i}}{154{,}38} \times 0{,}5168 + \frac{S_{2,i}}{145{,}68} \times 0{,}4832\right) \times 152.99.$$

В таблиці 3 наведено характеристики за кожною моделлю на періоді подовження.

В таблиці 4 подано інформацію щодо змінення характеристик за кожною моделлю порівняно зі значеннями ряду 79545 на періоді подовження.

Дані, представлені в таблицях 3-4, дають краще розуміння можливих змін у характеристиках вихідного ряду при виборі певної моделі для подовження. В таблиці 4 відображено зміну характеристик отриманих рядів на періоді подовження до значень ряду 79545, виражені у відсотках, при використанні кожної з моделей. Для чотирьох моделей з найкращими оцінками (таблиця 1) побудовано діаграми розсіювання відносно лінійної регресії та за зазначеним коефіцієнтом кореляції Пірсона (рисунок 4).

Таблиця 3 – **Характеристики рядів на періоді подовження**

Показник	79545	LR	RFR	KNR	SVR	GBR	SWG	SWGC
Середнє значення	156,4	159,7	162,1	161,6	158,2	161,6	153,5	156,6
Стандартне відхилення	28,4	23,4	21,9	20,8	23,1	19,8	24,4	25,1
Мінімальне значення	124,0	139,4	123,8	126,5	138,2	135,7	132,6	135,1
Максимальне значення	287,0	285,4	283,7	285,8	282,4	289,1	281,7	288,4

Таблиця 4 – **Зміна значень характеристик рядів на періоді подовження**

Показник	LR	RFR	KNR	SVR	GBR	SWG	SWGC
Середнє значення	2,1%	3,6%	3,3%	1,2%	3,4%	-1,8%	0,1%
Стандартне відхилення	-17,4%	-22,8%	-26,6%	-18,7%	-30,1%	-14,0%	-11,7%
Мінімальне значення	12,4%	-0,2%	2,0%	11,5%	9,4%	6,9%	9,0%
Максимальне значення	-0,5%	-1,2%	-0,4%	-1,6%	0,7%	-1,9%	0,5%

DOI: 10.24025/2306-4412.3.2021.244266

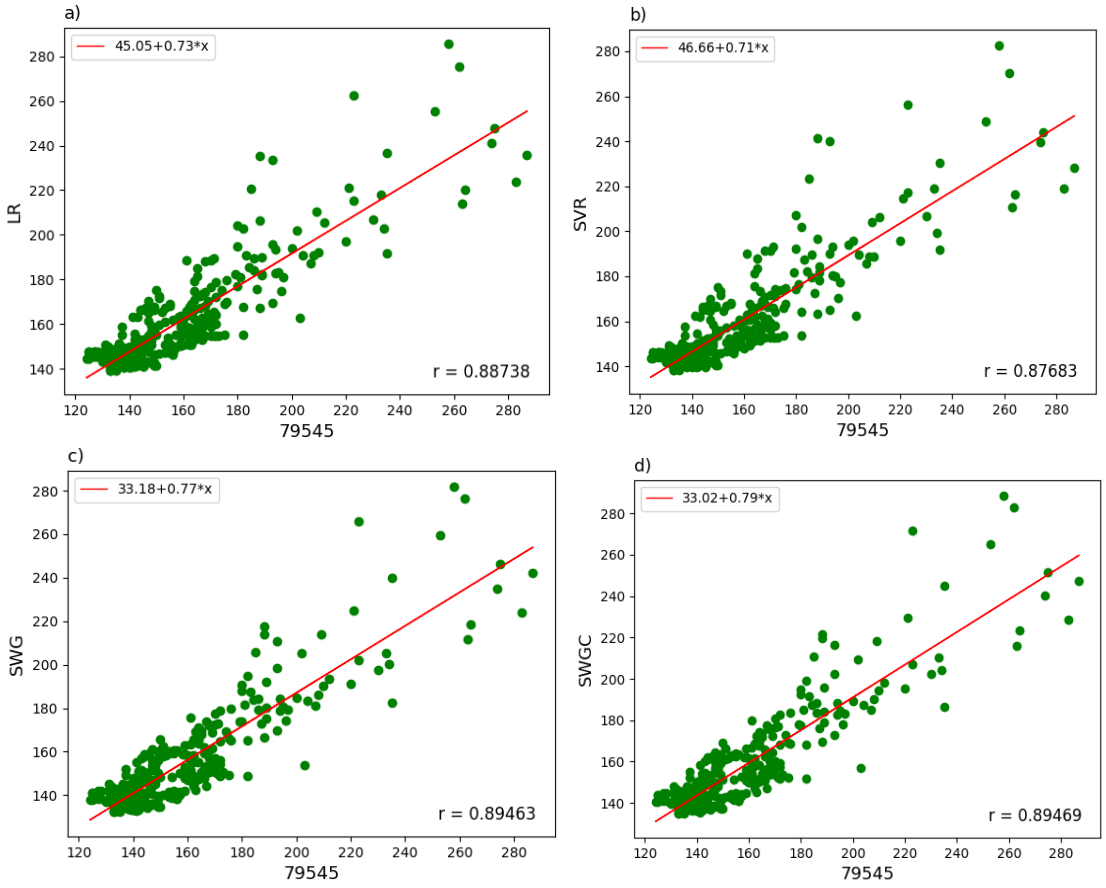

Рисунок 4 – Діаграми розсіювання для найкращих моделей відносно лінійної регресії для ряду 79545

Обговорення результатів. Серед усіх рядів даних, що обрано для аналізу як потенційно схожі на ряд показників посту 79545 (рисунок 1), визначено лише два ряди даних (рисунок 2), які за значеннями розрахованих відстаней та коефіцієнта кореляції можуть бути використані для подовження ряду показників посту 79545. За результатами проведеного подовження ряду показників посту 79545 за сьома представленими моделями встановлено, що найкращі оцінки (таблиця 2) отримано при використанні моделей SWG та SWGC. Найгірші результати при подовженні ряду за розрахованими оцінками показали моделі випадкового лісу та *k*-найближчих сусідів. За результатами, представленими в таблицях 3 та 4, можна зробити висновок, що кожна з моделей подовження по-різному впливає на характеристики модельованих значень порівняно з реальними значеннями ряду посту 79545 на періоді подовження. Наприклад, серед розглянутих моделей середнє значення та стандартне відхилення найменших змін зазнають при подовженні ряду за моделлю SWGC, а розмах значень вибірки зменшився приблизно на 6%. Якщо ж розглянути значення характеристик за моделями RFR та KNR (таблиця 4), то ситуація прямо протилежна моделі SWGC. За розрахованими значеннями коефіцієнтів кореляції Пірсона, що відображені на графіках рисунка 3, найсильніший прямий кореляційний зв'язок між змодельованими і реальними даними на періоді подовження отримано з використанням моделей SWG (рисунок 3, c) та SWGC (рисунок 3, d).

Слід зазначити, що набір визначених схожих рядів суттєво впливає на отримувані результати подовження ряду. При визначенні схожих рядів із застосуванням значень відстаней можлива ситуація, коли ряди з сильним кореляційним зв'язком можуть бути відкинуті через не найменші значення відстаней. Залежно від використовуваних підходів до визначення схожих рядів, періоду спільних спостережень, періоду прогнозування, періоду навчання моделей, застосовуваних моделей то-

DOI: 10.24025/2306-4412.3.2021.244266

що можливо отримати різні результати та якість подовження ряду. Проте вже на етапі підгонки моделей можна отримати попередні дані щодо можливості отримання якісних результатів при прогнозуванні з використанням конкретної моделі.

Висновки. В ході виконання дослідження реалізовано різні за складністю моделі для подовження рядів даних на основі значень схожих рядів. Перевагою такого підходу є те, що використання значень схожих рядів забезпечує збереження довгострокових та короткострокових тенденцій та коливань. Окремо слід зазначити, що найкращі моделі виявилися найпростішими в обчисленні серед розглянутих у цій роботі.

Наукова новизна полягає в застосуванні регресійного аналізу для подовження часових рядів за значеннями показників схожих рядів. Застосування такого підходу дозволяє підвищити якість побудованих прогнозів навіть при використанні простих регресійних моделей завдяки збереженню та відображенню поведінки схожих рядів на прогнозованих значеннях.

Практичне значення полягає в можливості подовження часових рядів. З використанням отриманих результатів можливо вирішити питання недостатності кількості спостережень, заповнення великих проміжків пропущених значень, приведення рядів даних до однакової довжини.

Подальші дослідження планується спрямувати на використання отриманих результатів для розробки та аналізу методів відновлення пропущених значень різної природи в часових рядах.

Список використаних джерел

[1] Е. Ковпак, Ф. Орлов, "Порівняльний аналіз моделей машинного навчання і регресій для прогнозування ціни легкового авто", *Вісник Харківського національного університету імені В. Н. Каразіна. Серія: Економічна*, вип. 97, с. 31-40, 2019. doi: 10.26565/2311-2379-2019-97-04.

[2] M. S. Khan, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach", *Journal of King Saud University-Computer and Information Sciences*, 2021. doi: 10.1016/j.jksuci.2021.06.003.

[3] Assem Haytham et al., "Urban water flow and water level prediction based on deep learning", *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, Springer, Cham, pp. 317-329, 2017. doi: 10.1007/978-3-319-71273-4_26.

[4] В. Д. Дербенцев, Г. І. Великоіваненко, та Н. В. Даценко, "Застосування методів машинного навчання до прогнозування часових рядів криптовалют", *Нейронечіткі технології моделювання в економіці*, № 8, с. 66-99, 2019. doi: 10.33111/nfmte.2019.065.

[5] Ю. Л. Хлевна, та Ю. С. Бура, "Інформаційне забезпечення прогнозування цін на нерухомість методами машинного навчання", *Sciences of Europe*, вип. 71-1, с. 54-62, 2021. doi: 10.24412/3162-2364-2021-71-1-54-62.

[6] Ю. О. Андрусенко, "Аналіз основних моделей прогнозування часових рядів", *Збірник наукових праць Харківського національного університету Повітряних Сил*, вип. 3 (65), с. 91-96, 2020. doi.org/10.30748/zhups.2020.65.14.

[7] В. А. Артеменко, и В. В. Петрович, "Повышение качества прогнозирования гидрологических временных рядов", *Автомобільні дороги і дорожнє будівництво*: наук.-техн. зб., вип. 92, Київ: Вид-во НТУ, с. 146-127, 2014.

[8] C. Chen, Q. Hui, Q. Pei, Y. Zhou, B. Wang, N. Lv, and J. Li, "CRML: A convolution regression model with machine learning for hydrology forecasting", *IEEE Access*, vol. 7, pp. 133839-133849, 2019. doi:10.1109/ACCESS.2019.2941234.

[9] S. K. Jain at al., "A brief review of flood forecasting techniques and their applications", *Int. J. River Basin Manag.*, vol. 16, pp. 329–344, 2018. doi:10.1080/15715124.2017.1411920.

[10] Д. А. Тамбиева, Е. В. Попова, и Ш. Х. Салпагарова, "К проблеме недостаточности информации. Малые выборки или "очень короткие" временные ряды", *Политематический сетевой электрон-*

DOI: 10.24025/2306-4412.3.2021.244266

ный научный журнал Кубанского государственного аграрного университета, № 107, с. 126-141, 2015.
[11] Random Forest Regressor [Online]. Available: http://surl.li/aidsj.
[12] K-Neighbors Regressor. [Online]. Available: http://surl.li/aidsk.
[13] SVR. [Online]. Available: http://surl.li/aidsm.
[14] Gradient Boosting Regressor. [Online]. Available: http://surl.li/aidsn.
[15] scikit-learn. [Online]. Available: https://scikit-learn.org/stable/index.html.

References

[1] E. Kovpak, F. Orlov, "Comparative analysis of machine learning models and regressions for prediction the car price", *Visnyk Kharkivskoho natsionalnoho universytetu imeni V. N. Karazina. Seriia: Ekonomichna*, iss. 97, pp. 31-40, 2019 [in Ukrainian]. doi: 10.26565/2311-2379-2019-97-04.
[2] M. S. Khan, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach", *Journal of King Saud University-Computer and Information Sciences*, 2021. doi: 10.1016/j.jksuci.2021.06.003.
[3] Assem Haytham et al., "Urban water flow and water level prediction based on deep learning", *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, Springer, Cham, pp. 317-329, 2017. doi: 10.1007/978-3-319-71273-4_26.
[4] V. D. Derbentsev, H. I. Velykoivanenko, and N. V. Datsenko, "Machine learning approach for forecasting cryptocurrencies time series", *Neiro-nechitki tekhnolohii modeliuvannia v ekonomitsi*, no. 8, pp. 65-93, 2019 [in Ukrainian]. doi: 10.33111/nfmte.2019.065.
[5] Yu. L. Khlevna, and Yu. S. Bura, "Information software for real estate prices prediction by machine learning", *Sciences of Europe*, iss. 71-1, pp. 54-62, 2021 [in Ukrainian]. doi: 10.24412/3162-2364-2021-71-1-54-62.
[6] Yu. O. Andrusenko, "Analysis of the basic models for forecasting time series", *Zbirnyk naukovykh prats Kharkivskoho natsionalnoho universytetu Povitrianykh Syl*, iss. 3 (65), pp. 91-96, 2020 [in Ukrainian]. doi.org/10.30748/zhups.2020.65.14.
[7] V. A. Artemenko, and V. V. Petrovich, "Improving the quality of forecasting of hydrological time series", *Avtomobilni dorohy i dorozhnie budivnytstvo*: sci. and techn. coll., iss. 92, pp. 114-127, 2014 [in Russian].
[8] C. Chen, Q. Hui, Q. Pei, Y. Zhou, B. Wang, N. Lv, and J. Li, "CRML: A convolution regression model with machine learning for hydrology forecasting", *IEEE Access*, vol. 7, pp. 133839-133849, 2019. doi:10.1109/ACCESS.2019.2941234.
[9] S. K. Jain at al., "A brief review of flood forecasting techniques and their applications", *Int. J. River Basin Manag*., vol. 16, pp. 329–344, 2018. doi:10.1080/15715124.2017.1411920.
[10] Д. А. Тамбиева, Е. В. Попова, и Ш. Х. Салпагарова, "К проблеме недостаточности информации. Малые выборки или "очень короткие" временные ряды", *Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета*, № 107, с. 126-141, 2015.
[11] Random Forest Regressor [Online]. Available: http://surl.li/aidsj.
[12] K-Neighbors Regressor. [Online]. Available: http://surl.li/aidsk.
[13] SVR. [Online]. Available: http://surl.li/aidsm.
[14] Gradient Boosting Regressor. [Online]. Available: http://surl.li/aidsn.
[15] scikit-learn. [Online]. Available: https://scikit-learn.org/stable/index.html.

DOI: 10.24025/2306-4412.3.2021.244266

A. H. Baturinets,
e-mail: baturinets.anastasiya@gmail.com
S. V. Antonenko, *Ph. D., Associate Professor*
e-mail: svitlanav.antonenko@gmail.com
Oles Honchar Dnipro National University,
Gagarin ave., 72, Dnipro, 49010, Ukraine

LENGTHENING THE DATA SERIES BY VALUES OF SIMILAR DATA SERIES SAMPLES

The problem of insufficient information essentially influences the choice of approaches and methods of data series analysis, as well as the quality of the obtained results. Considering this problem, the authors believe that the development of such approaches and models for data series lengthening is relevant. The main task of this work is to describe and implement the technology of data series lengthening. The basis for the implementation of the technology is the use of values of similar data series as a signs for the lengthening of a certain data series represented by the same indicators, as well as similar data series. The work describes a scheme for identifying similar data series. According to this scheme, the most similar data series are those that have the smallest distance value and the strongest direct correlation, calculated between the potentially similar series and the series for which the lengthening will take place. For lengthening of the series, the work considers seven models: linear regression; sum of weighted values for a group of similar series; average weighted values for a group of similar series, with a correction to the average value of the series for which the lengthening is performed; random forest; k-nearest neighbors; support vector regression; gradient busting. The calculation experiment was carried out on the series represented by the values of water level indicators recorded at hydrological stations located in the water objects of the Dnieper River basin. For the data series of post 79545, located on the river Sluch, Novograd-Volynsky, Zhytomyr region, a lengthening by one year is carried out, i.e. the length of the series increases by 365 values. As a result, it was found that the most similar are the data series of values by the posts 79555 and 79694, which have the lowest values of the calculated distances and the value of the correlation coefficient greater than 0.75. When the series is lengthened, the best results are obtained with the use of two models: the sum of weighted values for a group of similar series and average weighted values for a group of similar series, with a correction to the average value of the series for which the lengthening is performed. In future research it is planned to use the obtained results for the development and analysis of methods for replenishment of missing values in time series.

Keywords: *time series, regression, data replenishment, machine learning, sklearn, insufficient data, hydrology.*

Стаття надійшла 22.09.2021
Прийнято 11.10.2021

DOI: 10.24025/2306-4412.3.2021.244266