«Системні технології» 5 (136) 2021 «System technologies»
DOI 10.34185/1562-9945-5-136-2021-10
УДК 004.67:519.25

А.Г. Батурінець, С.В. Антоненко

ВИЗНАЧЕННЯ СХОЖИХ ГІДРОЛОГІЧНИХ РЯДІВ ДАНИХ З ВИКОРИСТАННЯМ КОЕФІЦІЄНТІВ КОРЕЛЯЦІЇ

Анотація. В роботі описано розроблену складову програмного забезпечення для визначення схожих рядів даних, що представлені певними показниками. Основою обчислювальної схеми є обчислення коефіцієнтів кореляції Спірмена та Пірсона. При проведені аналізу передбачається поетапна оцінка результатів, що дозволяє контролювати хід аналізу та інтерпретувати отримані результати, а також широкий спектр відображення результатів. Представлено результати проведення обчислювального експерименту проведеного за показниками рівня води в період з 01.01.2000р. по 31.12.2014р., в якості досліджуваного ряду обрано дані пункту спостереження розташованого на р. Турія в м. Ковель Волинської обл. За результатами аналізу знайдено три гідрологічні пункти спостережень, дані яких мають сильний кореляційний зв'язок з даними досліджуваного посту, проте лише два з них за значеннями оцінок вважаються дуже схожими.

Ключові слова: коефіцієнти кореляції, схожі ряди даних, гідрологічний моніторинг, інформаційна технологія.

Постановка проблеми. В умовах сучасності важливими є питання дослідження стану водних ресурсів, а тому все більша увага приділяється питанням моніторингу та аналізу отриманих гідрологічних даних для попередження та своєчасного реагування на можливі негативні наслідки. Проте, вже на початку аналізу даних дослідники стикаються з низкою проблем, серед яких недостатня довжина рядів даних спостережень, наявність пропущених значень, тощо. Одним із підходів, що використовується для вирішення зазначених проблем, є використання методів на основі визначення об'єктів із схожими характеристиками та показниками.

Кореляційний аналіз знайшов широке використання при дослідженні зв'язків між рядами даних при розв'язанні задач в різних прикладних областях, зокрема і в гідрологічних дослідженнях. Здебільшого це пов'язано з тим, що з

© Батурінець А.Г., Антоненко С.В., 2021

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

однієї сторони коефіцієнти кореляції не складні в обчислені, а з іншої – їх легко інтерпретувати.

Зазвичай, гідрологічні дослідження проводяться з використанням засобів MS Excel, Statistica, Mathlab, Matcad і лише в невеликій кількості досліджень для проведення розрахунків частково використовуються готові спеціалізовані програмні рішення. Ще менша кількість робіт проводиться саме з програмною реалізацією обчислювальних схем та алгоритмів на мовах програмування високого рівня.

Аналіз останніх досліджень і публікацій. Використання коефіцієнтів кореляції в гідрологічних дослідженнях можна побачити, наприклад, в роботах [1-5]. Аналіз зв'язків гідрологічних показників, зокрема з використанням кореляційного аналізу представлено в роботі [6].

Використання кореляційного аналізу є важливою складовою при визначенні річок-аналогів в гідрологічних розрахунках. Використання значень коефіцієнтів кореляції при відновлення значень стоку за пунктом-аналогом розглядається в роботах [7-8], а в роботі [9] при відновлені середнього річного стоку р. Дніпро обчислюється коефіцієнт кореляції між пунктами-відновлення і пунктами-аналогами. Зокрема і в роботі [10] автором при проведені аналізу гідрологічних даних та відновлені рядів гідрологічних даних за даними річок-аналогів використовується обчислення значень лінійної кореляції.

Мета роботи. Розробити, описати та реалізувати технологію пошуку схожих гідрологічних постів за рядами даних, що представлені певними показниками, забезпечити оцінку результатів, відображення даних та провести аналіз отриманих результатів.

Основний матеріал. Сутність кореляційного аналізу полягає в розрахунку коефіцієнтів кореляції, які в свою чергу можуть приймати додатні та від'ємні значення, а їх абсолютне значення визначає силу зв'язку між досліджуваними величинами. Знак коефіцієнта кореляціє дозволяє визначити направлення зв'язку: при додатних значеннях зв'язок вважається прямим, при від'ємних – зворотнім.

Для розв'язання поставленої задачі обчислюються коефіцієнти кореляції Пірсона та Спірмена. Коефіцієнт кореляції Спірмена є непараметричним аналогом коефіцієнта парної кореляції Пірсона. Зазвичай, в гідрологічних дослідженнях проводяться розрахунки коефіцієнта кореляції Пірсона, що між іншим, зумовлено саме простотою обчислення значень, проте гідрологічні дані зазвичай не є нормально розподіленими.

Від користувача необхідні наступні налаштування: обрати пост, для якого відбуватиметься пошук схожих рядів даних; визначити показник, за яким відбуватиметься аналіз; період, за який дані підлягають аналізу; допустиму кількість пропущених значень ряду (у відсотках, за замовчуванням – 0.25%); параметр для визначення межі відбору коефіцієнтів кореляції (за замовчуванням – 5%); необхідність перетворення вихідних даних: працювати з вихідними даними ряду, середньотижневими або середньомісячними значеннями ряду (за замовчуванням – без перетворень).

На рис.1 представлено схему роботи технології пошуку схожих гідрологічних об’єктів: від отримання вихідних даних до візуалізації результатів.

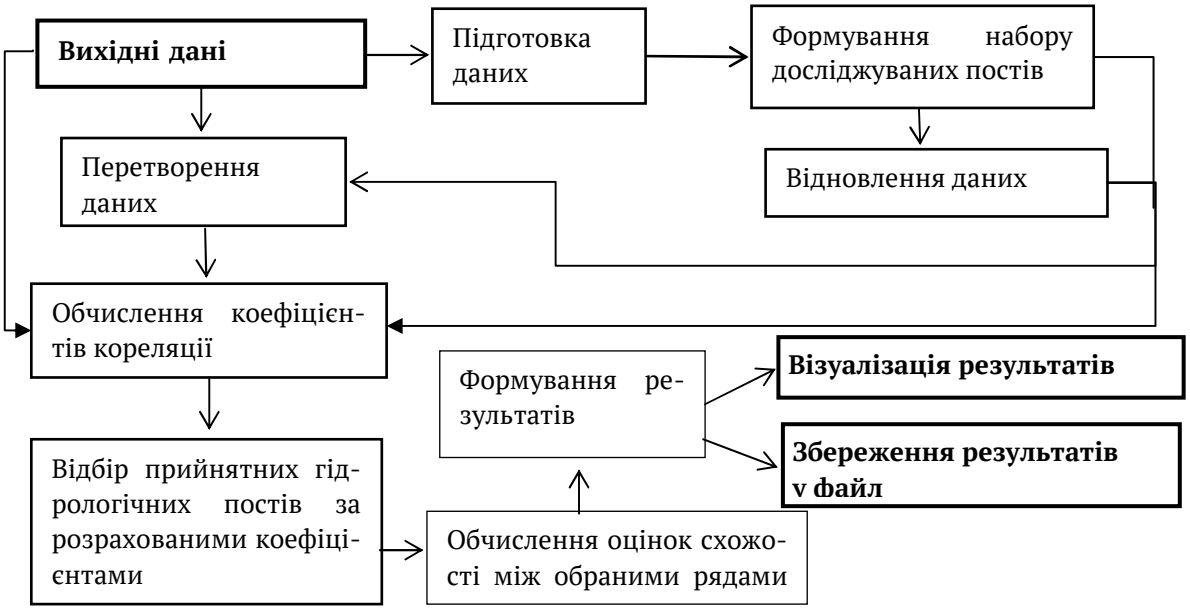

Рисунок 1 – Схема роботи технології визначення схожих рядів даних

Першочергово перевіряються ряди на наявність пропущених в них значень за обраний проміжок часу. Ряди, що містять пропущені значення в кількості, більшій вказаного користувачем відсотку, видаляються із розгляду, а у випадку наявності пропущених значень, менше визначеного користувачем відсотку, відсутні значення відновлюються за допомогою ковзного середнього. *Необхідність включення* такої можливості зумовлена ситуацією, якщо, наприклад, ряд має декілька пропущених значень на тисячу наявних показників, тобто відновлені таким чином показники несуттєво впливають на основні характеристики ряду.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Далі, якщо це визначено користувачем, проводиться перетворення даних (взяття середньотижневих або середньомісячних значень), а в іншому випадку – даний етап пропускається.

Окрім цього, одним з етапів аналізу є повторне обчислення коефіцієнтів кореляції на перетворених рядах даних. Формула перетворення має наступний вигляд:

$$ts_{ik} = \frac{ts_{ij}}{\mu},$$

де ts_{ik} – i значення k ряду даних, ts_{ij} – j значення i вихідного ряду, μ_i – середнє значення i вихідного ряду.

Дослідження отриманих рядів після перетворення надає можливість оцінити безпосередньо поведінку рядів та не зважати на їх рівні.

Інтервал відбору найбільш схожих постів з досліджуваним визначається за наступною нерівністю:

$$r_{max} \leq r \leq r_{max} - r_{max} * p,$$

де r_{max} – максимальне значення серед отриманих коефіцієнтів кореляції, p – задається користувачем. За замовчуванням значення параметра p встановлено 0.05, оскільки саме 5% рівня достатньо для визначення нижньої межі коефіцієнтів кореляції, що задовольнить відбір рядів даних, які мають вищу силу зв'язку з досліджуваним рядом в порівнянні з іншими.

Для оцінки схожості двох рядів даних обрано оцінки, що використовуються при оцінці регресійних моделей, серед яких $MAPE$ та R^2. Можливість їх застосування зумовлена тим, що їх розрахунок проводиться з використанням наявних даних двох рядів.

- $MAPE$ (середня абсолютна похибка у відсотках) розрахована за формулою:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - \hat{x}_i|}{x_i} * 100,$$

де x_i – фактичне значення ряду, $\hat{x}_i$ – значення ряду даних, що є потенційного схожим з досліджуваним.

Будемо вважати, що ряд даних є достатньо схожим на досліджуваний, якщо значення даного коефіцієнта не перевищує 8-10%.

- R^2 – вибірковий коефіцієнт детермінації розрахований за формулою:

$$R^2 = 1 - \frac{\hat{\delta}^2}{\hat{\delta}_y^2} = 1 - \frac{RSS}{TSS},$$

де – сума квадратів залишків, TSS – загальна сума квадратів, розраховані за наступними формулами:

$$TSS = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

де $\bar{x}$ – середнє значення елементів часового ряду, x_i – фактичне значення ряду, $\hat{x}_i$ – значення ряду даних, що є потенційно схожим з досліджуваним.

При цьому вважається, що чим ближче значення коефіцієнта детермінації до 1, тим більш схожими є ряди даних. На достатньому рівні будемо вважати досягнені коефіцієнта детермінації не менше значення 0,5, а при досягненні значення 0,8 будемо вважати знайдений ряд даних схожим на досліджуваний на високому рівні.

Також в даній технології обчислюються оцінки: *MSE* – середня квадратична похибка, *MAE* – середня абсолютна похибка. *MSE* та *MAE* використовуються як допоміжні оцінки, що не дозволяють однозначно оцінити схожість рядів, але дозволяють визначити найбільш схожий ряд даних з набору.

Як результат користувач отримує представлення результатів на різних етапах аналізу в табличному вигляді (із збереженням у файл) та графічному (діаграма розсіювання, графіки рядів даних).

Обчислювальний експеримент. В якості досліджуваного посту обрано пост 79407, розташований на р. Турія в м. Ковель Волинської області. Для аналізу обрано щоденні значення рівнів води, період спільних спостережень з 01.01.2000р. по 31.12.2014р., тобто для аналізу кожен ряд даних: як досліджуваний, так і ті, що потенційно схожі з ним, повинні мати довжину в 5480 значень. Допустиму кількість пропущених значень встановлено на рівні 0,25%, що для обраного періоду допускає наявність не більше 13 пропущених в кожному ряді даних.

В результатів відбору для аналізу представлено 93 потенційно схожі за обраним показником гідрологічні пости. З графіку рівнів води, що представлений на рис.1, видно, що обрані пости є різними в першу чергу за середніми значенням даних.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Рисунок 2 – Графік рівнів води з 01.01.2000р. по 31.12.2014р.

В табл.2 наведено значення коефіцієнтів кореляції за вихідними даними для відібраних пунктів спостережень, які обрано в результаті аналізу. Як видно із отриманих значень, ряди даних відібраний пунктів спостережень мають сильний зв'язок з досліджуваним.

Таблиця 1

Результати обчислень коефіцієнтів кореляції

Пост	Коефіцієнт кореляції	
	Пірсона	Спірмена
79403	0,84694	0,87559
79405	0,85654	0,86038
79416	0,84474	0,81638

На рис. 3 представлено вихідні дані досліджуваного посту та постів, з якими наявний сильний кореляційний зв'язок. Оскільки застосування коефіцієнтів кореляції має на меті визначення наявності геометричного зв'язку, то як можна побачити на рис.3, пост 79416, на відміну від постів 79407,79403 та 79405, відрізняється рівнями значень ряду.

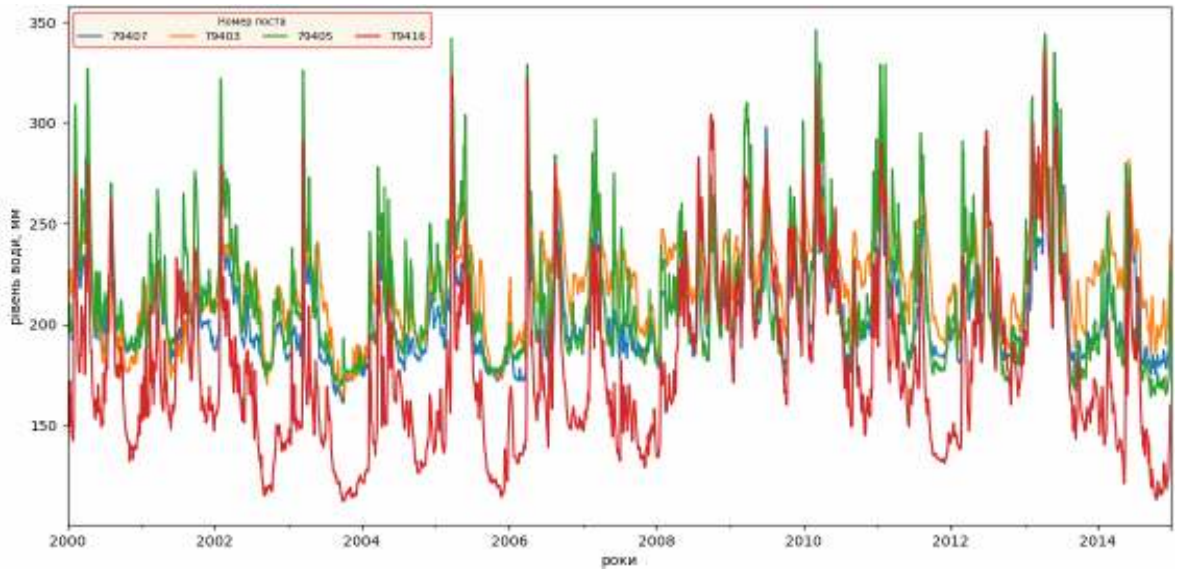

Рисунок 3 – Графік результатів відбору постів по вихідним даним

В табл. 2 представлено оцінки схожості на досліджуваний ряд рядів даних, обраних в результаті аналізу за вихідними даними. Розглядаючи зазначені оцінки, можна сказати, що дані досліджуваного поста та поста 79405 відрізняються на приблизно 6,2%, що є прийнятним результатом, окрім цього коефіцієнт детермінації складає приблизно 64,7% , що є достатньо прийнятним значенням.

Таблиця 2

Оцінки схожості рядів за вихідними даними

Пост	Оцінки			
	MSE	MAE	MAPE, %	R^2
79403	442,8137	18,06352	8,019547	0,255678
79405	384,654	14,27213	6,205114	0,647618
79416	1314,676	31,87315	20,39622	0,255936

Далі проводиться перетворення значень показників ряду діленням на середні значення кожного відповідного ряду. Середні значення досліджуваних рядів складає 205,28 мм для досліджуваного поста, для поста №79403 – 221,41 мм, поста №79405 – 214,18 мм, а для поста №79416 – 179,08 мм.

Як видно на рис.4, після перетворення ряди графічно є більш наближеними один до одного, на відміну від представлення рядів у вихідному вигляді (рис. 3).

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

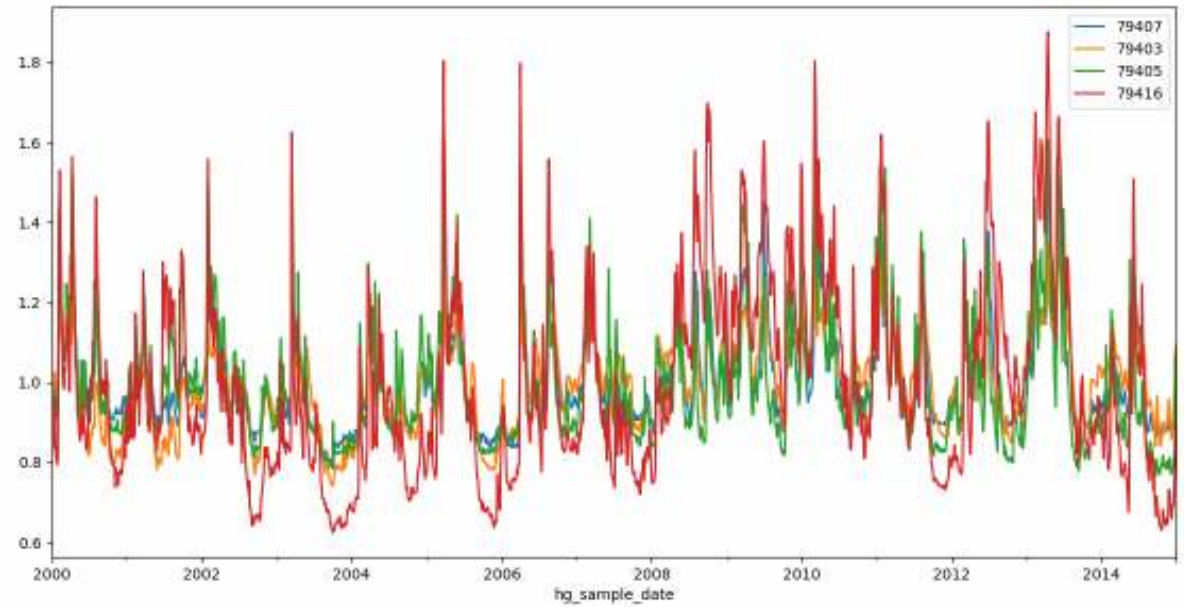

Рисунок 4 – Перетворені ряди даних гідрологічних показників

Як видно з табл.3, розраховані оцінки підтверджують, що ряди, обрані в результаті аналізу, є дуже схожими на показники досліджуваного посту.

Таблиця 3

Оцінки схожості рядів після ділення на середнє значення

Пост	Оцінки			
	MSE	MAE	MAPE, %	R^2
79403	0,0041	0,0499	4,9673	0,6621
79416	0,0220	0,1258	12,4752	0,6000

За результатами аналізу формується файл формату excel, що містить інформацію про всі етапи аналізу: таблиці значень коефіцієнтів кореляції, інформацію про пости (географічні координати, населений пункт, назва річки, на якій розміщено пост, номер гідрологічного об'єкта), характеристики рядів даних, оцінки схожості постів на досліджуваний, тощо. Збереження зазначеної інформації в такому форматі надає змогу використовувати їх для подальшого аналізу та поза межами розроблюваного програмного забезпечення.

Оцінка результатів. Аналізуючи отримані результати, можна дійти висновку, що пост 79405 р. Турія в с. Ягідне Волинської обл. має досить високу схожість за рівнями води з досліджуваним постом за всіма розрахованими оцінками. При цьому, отримані в результаті аналізу значення оцінок свідчать, що пост 79403 р. Вижівка в смт. Стара Вижівка має досить непогані показники середньої похибки апроксимації. Проте після того як ряди було приведено до ну-

льового середнього значення, коефіцієнт детермінації збільшився більше ніж в 2,5 рази (0,2556 до перетворення та 0,6621 після).

Результати даного аналізу свідчать, що показники рівнів води поста 79416 р. Стохід в с. Малинівка Волинської обл. хоча й мають сильний кореляційний зв'язок з рівнями води досліджуваного поста, проте потребують додаткових досліджень, оскільки значення MAPE і коефіцієнта R^2 (табл. 2-3) мають неприйнятні значення.

Варто також зазначити, що за результатами визначення схожих гідрологічних постів за рівнями води, пости, відібрані в результаті аналізу та досліджуваний пост розміщені на території однієї області.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Гопченко Е.Д., Овчарук В.А., Тодорова Е.И. Максимальный сток дождевых паводков рек Горного Крыма // Вісник Одеського державного екологічного університету. – 2014. – Вип.17. – С.133-140.

2. Ободовський О.Г. та ін. Узагальнення середнього річного стоку води річок відповідно до гідрографічного районування України// Вісник Харківського національного університету імені ВН Каразіна, серія «Геологія. Географія. Екологія» – 2019. – Вип. 51. – С. 158-170.

3. Мозговий А. О. Дослідження кореляційної залежності максимальної товщини льоду за статистичними даними спостережень у водосховищах гідровузлів Дніпровського каскаду // Науковий вісник будівництва – 2018. – Т.– 3 (93). – С. 149-155.

4. Ободовський О.Г. та ін. Середній річний стік води в межах районів річкових басейнів України. Гідрологія, гідрохімія і гідроекологія – 2019. – Т. 3. – С. 65-66.

5. Ошурок Д.О., Скриник О.Я., Осадчий В.І. Приведення виміряних значень швидкості вітру до умов відкритого горизонту // Гідрологія, гідрохімія і гідроекологія – 2019. – Т.3. – С. 139-141.

6. Chen Lu, Singh Vijay P., Guo Shenglian. Measure of correlation between river flows using the copula-entropy method. Journal of Hydrologic Engineering, 2013, 18.12: p. 1591-1606.

7. Мудра К. В. Відновлення стоку води на гідрологічних постах річки Дністер з метою вивчення його довгоперіодних коливань // Гідрологія, гідрохімія і гідроекологія. – 2017. – Т. 2. – С. 30-39.

8. Артеменко В. А., Петрович В. В. Метод восстановления сверхбольших пропусков в гидрологических временных рядах // Автомобільні дороги і дорожне будівництво, 2015, 93: 150-156.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

9. Чорноморець Ю. О., Павленко П. О.; Лук'янець О. І. Відновлення середнього річного стоку води річки Дніпро. Гідрологія, гідрохімія і гідроекологія – 2017. – Т. 4 (47). – С. 36-47.
10. Бажанова Л. В. Оценка гидрологического монторинга и восстановление стока рек методом парной корреляции // Наука, новые технологии и инновации Кыргызстана, 2018, 3: 134-140.

REFERENCES

1. Hopchenko E.D., Ovcharuk V.A., Todorova E.Y. Maksymalnyi stok dozhdevykh pavodkov rek Hornoho Kryma // Visnyk Odeskoho derzhavnoho ekolohichnoho universytetu. – 2014. – Vyp.17. – S.133-140.
2. Obodovskyi O.H. ta in. Uzahalnennia serednoho richnoho stoku vody richok vidpovidno do hidrohrafichnoho raionuvannia Ukrainy// Visnyk Kharkivskoho natsionalnoho universytetu imeni VN Karazina, ceriia «Heolohiia. Heohrafiia. Ekolohiia» – 2019. – Vyp. 51. – S. 158-170.
3. Mozghovyi A. O. Doslidzhennia koreliatsiinoi zalezhnosti maksymalnoi tovshchyny lodu za statystychnymy danymy sposterezhen u vodoskhovyshchakh hidrovuzliv Dniprovskoho kaskadu // Naukovyi visnyk budivnytstva – 2018. – T.– 3 (93). – S. 149-155.
4. Obodovskyi O.H. ta in. Serednii richnyi stik vody v mezhakh raioniv richkovykh baseiniv Ukrainy. Hidrolohiia, hidrokhimiia i hidroekolohiia – 2019. – T. 3. – S. 65-66.
5. Oshurok D.O., Skrynyk O.Ia., Osadchyi V.I. Pryvedennia vymirianykh znachen shvydkosti vitru do umov vidkrytoho horyzontu // Hidrolohiia, hidrokhimiia i hidroekolohiia – 2019. – T.3. – S. 139-141.
6. Chen Lu, Singh Vijay P., Guo Shenglian. Measure of correlation between river flows using the copula-entropy method. Journal of Hydrologic Engineering, 2013, 18.12: r. 1591-1606.
7. Mudra K. V. Vidnovlennia stoku vody na hidrolohichnykh postakh richky Dnister z metoiu vyvchennia yoho dovhoperiodnykh kolyvan // Hidrolohiia, hidrokhimiia i hidroekolohiia. – 2017. – T. 2. – S. 30-39.
8. Artemenko V. A., Petrovych V. V. Metod vosstanovlenyia sverkhbolshykh propuskov v hydrolohycheskykh vremennыkh riadakh // Avtomobilni dorohy i dorozhne budivnytstvo, 2015, 93: 150-156.
9. Chornomorets Yu. O., Pavlenko P. O.; Lukianets O. I. Vidnovlennia serednoho richnoho stoku vody richky Dnipro. Hidrolohiia, hidrokhimiia i hidroekolohiia – 2017. – T. 4 (47). – S. 36-47.

10. Bazhanova L. V. Otsenka hydrolohycheskoho montorynha i vosstanovlenye stoka rek metodom parnoi korreliatsyy // Nauka, novye tekhnolohyy i innovatsii Kurhuzstana, 2018, 3: 134-140.

Received 03.02.2021.
Accepted 30.04.2021.

Определение похожих гидрологических рядов данных с использованием коэффициентов корреляции

В работе описано вычислительную схему определения похожих гидрологических рядов, представленных некоторыми показателями, основой которой является расчет коэффициентов корреляции Спирмена и Пирсона. При проведении анализа предусмотрено поэтапную оценку результатов, что дает возможность контролировать ход анализа и интерпретировать результаты. По результатам представленного в работе эксперимента определены три гидрологических пункта наблюдений, ряды данных которых имеют сильную корреляционную связь с рядом данных исследуемого пункта, но при этом только два из них по значениям рассчитанных оценок можно считать похожими на ряд данных исследуемого пункта.

Search similar hydrological data series using correlation coefficients

Water resources are an important part of the social and economic development of countries. A large number of hydrological studies are carried out using MS Excel, Statistica, Mathlab, Matcad and only a small number of studies for calculations partially used specialized software solutions. An even smaller amount of research is carried out on the software implementation of computational schemes and algorithms in high-level programming languages.

Therefore, the development of software for the analysis of hydrological data is relevant. The aim of this work is to develop, describe and implement a technology to search for similar hydrological posts on the series of data represented by certain samples.

The paper describes the developed software component for determining similar series of data represented by hydrological samples. The computational scheme is based on the determination of the geometric similarity of data series, and the correlation coefficients of Spearman and Pearson are chosen for the calculations. The analysis involves a step-by-step calculation of the scores used in the regression analysis, which allows controlling the progress of the analysis and interpreting the results. Among the selected scores are MAPE, R2, MSE and MAE.

The paper presents the results of a computational experiment conducted on water level samples in the period from 01.01.2000 to 31.12.2014 years, the data of the post wich located on the river Turia in the city of Kovel, Volyn region, were selected as the studied series. According to the results of the analysis, three hydrological posts were found, the data of which have a strong correlation with the data of the studied post. However, the MAPE, R2, MSE and MAE scores show that only two of them can be considered sufficiently similar to the studied series.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Батурінець Анастасія Геннадіївна – аспірантка, Дніпровський національний університет імені Олеся Гончара.
Антоненко Світлана Валентинівна – к.т.н., доцент, Дніпровський національний університет імені Олеся Гончара.

Батуринец Анастасия Геннадьевна – аспирантка, Днепровский национальный университет имени Олеся Гончара.
Антоненко Светлана Валентиновна – к.т.н., доцент, Днепровский национальный университет имени Олеся Гончара.

Baturinets Anastasiia – graduate student, Oles Honchar Dnipro National University.
Antonenko Svetlana –Candidate of Technical Sciences, Associate Professor, Oles Honchar Dnipro National University.