СУЧАСНІ СТАТИСТИЧНІ МЕТОДИ ОБРОБКИ ЕКОЛОГІЧНИХ ДАНИХ

MODERN STATISTICAL METHODS OF ENVIRONMENTAL DATA PROCESSING

Артеменко Владислав Андрійович, Український гідрометеорологічний інститут Державної служби України з надзвичайних ситуацій та Національної академії наук України, науковий співробітник відділу гідрохімії, e-mail: artemenko@uhmi.org.ua, тел.. 380936011250, Україна, 03028, м. Київ, просп. Науки 37, к.34.

Петрович Володимир Васильович, кандидат технічних наук, професор, старший науковий співробітник, професор кафедри транспортного будівництва та управління майном Національного транспортного університету**.** e-mail: petrovichvv60@ukr.net, тел. +380442807338, Україна, 01010, м. Київ, вул. М. Омеляновича-Павленка, 1, к. 138.,

https://orcid.org/0000-0003-0422-2535

Анотація. Розроблена сучасна парадігма обробки екологічних даних, яка базується на непараметричних методах статистики.

Наведені істотні переваги запропонованої парадігми.

Ключові слова: екологічні дані, обробка даних, часові гідрохімічні ряди, непараметричний метод статистики.

Вступ

Звичайно при аналізі будь-якої екологічної інформації використовують параметричні методи обробки даних.

При застосуванні таких статистичних методів необхідним є виконання слідуючих умов [1]:

-розподілення даних повинно бути близьким до одного із широко відомих розподілень (на практиці застосовують переважно нормальний закон розподілення);

-вибірка повинна бути значною (не менш ніж 30 спостережень);

-всі дані – тільки інтервальні та безперевні.

На сучасному рівні досліджень вихідні дані розглядають переважно із позицій часових рядів.

Застосовуючи нормальне розподілення, тим самим приймають за основу повну випадковість процесу слідкування даних.

Але ігноруючи порядок слідкування, тим самим не враховують фактор сезонності та інші важливі прояви такого типу даних, як природний часовий ряд (наявність аномальних значень та різного роду вибросів, та інше).

Незважаючи на безліч критичних зауважень відносно застосування параметричних методів, сучасна парадігма обробки та аналіза екологічних даних не розроблена.

Для адекватної обробки вихідних даних доцільно використовувати непараметричні методи [2]. В статті наведені основи сучасної парадігми обробки екологічних даних, яка базується переважно на непараметричних методах статистики.

Вихідні дані

Екологічні процеси залежно від типу задач характеризують за допомого різних часових рядів.

Розглянемо у даному випадку часові гідрохімічні ряди.

Для виконання розрахунків були використані дані по річці Десна (створ с. Літки) за період з 1995 по 2010 рік включно (без пропусків). З метою спрощення обчислень значення даних за 29 лютого високосних років із розглянутих рядів виключалися.

Дані являють собою концентрації речовин, які виражають в *мг/л (мг/дм^3)* або *мкг/л (мкг/дм^3)*.

Як відомо, збір та оперативна обробка гідрохімічної інформації у "ручному режимі" може виконуватись переважно один раз на добу.

На підставі такої інформації далі знаходять середньомісячні, або "істинні" показники.

Звісно, що збір інформації можливо виконувати також один раз за сезон.

Однак при такому гранично рідкому заборі річкової води маємо значну втрату корисної інформації. У цьому зв'язку певний оптимум досягається при зборі інформації один раз на місяць (можливо у довільні дні цього місяця).У результаті такого збору інформації одержують так звані "квазісередньомісячні" гідрохімічні показники.

Проведене порівняння "істинних" та "квазісередньомісячних" показників для цих та багатьох інших водних об'єктів дало можливість встановити, що різниця між ними звичайно не перевищує 10% … 20% (середнє значення 15%).

Глобальної втрати точності розрахунків при цьому не спостерігається. В той же час проводити відповідні аналізи стало швидше та дешевше.

На рис. 1а наведений графік ряда "квазісередньомісячних" (далі середньомісячних) концентрацій аммонійних іонів, на рис. 1б – графік ряда середньомісячних концентрацій кисню, що розчинений у річковій воді.

Застосування методів непараметричної статистики при аналізі природних часових рядів.

Звичайно вихідні дані аналізують із позицій природних часових рядів, які проявляють одночасно властивості як детермінованості, так частково також і властивості випадковості [3].

Тобто із практичної точки зору природні часові ряди не є рядами детермінованими, але, в то же час, вони не будуть і повністю випадковими.

На сучасному рівні знань прийнято розглядати такі ряди як "ряди невизначені". Аналізуючи вихідні дані із таких позицій, тим самим допускають застосування до цих рядів спеціальних методів дослідження.

Для адекватної обробки даних у цьому випадку використовують непараметричні методи статистики.

Як відомо, одна із головних переваг непараметричних методів полягає у тому, що встановлювати закон (закони) розподілення вихідних даних не потрібно.

Існує одностадійний підхід, за допомогою якого можливе використання приктично всіх методів непараметричної статистики із єдиних позицій – метод непараметричного бутстрепа [4].

Більш того, це єдиний метод, що дозволяє використання відносно незначних модифікацій при проведенні будь-якого статистичного теста. Таким чином, виконання непараметричних досліджень можливо проводити за єдиною загальною схемою!

Метод непараметричного бутстрепа дозволяє легко виконувати складні числові дослідження, одержуючи при цьому не тільки значення певної статистики, але також знаходити різні довірчі інтервали: для квантиля будь-якого рівня, для коефіцієнтів різного типу регресій, різного виду корреляцій та ін.

Якщо є необхідність, можливе одержання значення "p-value" без застосування спеціальних математичних функій (як це має місце у випадку застосування параметричних методів).

Аналіз на нормальність розподілення гідрохімічних даних.

Перевірялась погодженість гідрохімічних даних із нормальним розподіленням.

Використовувались середньомісячні дані за 20-річний період для крупної рівнинної річки України (р. Десна).

Згідно [1], критерій Шапіро – Уілка є одним із найбільш ефективних критеріїв перевірки нормальності розподілення величин. У дослідженні у першу чергу був використаний цей критрій та значення *p=0.05(p=5%).*

Як показали результати, тільки один показник із 15 досліджуваних можливо вважати нормально розподіленим (кисень, що розчинений у воді).

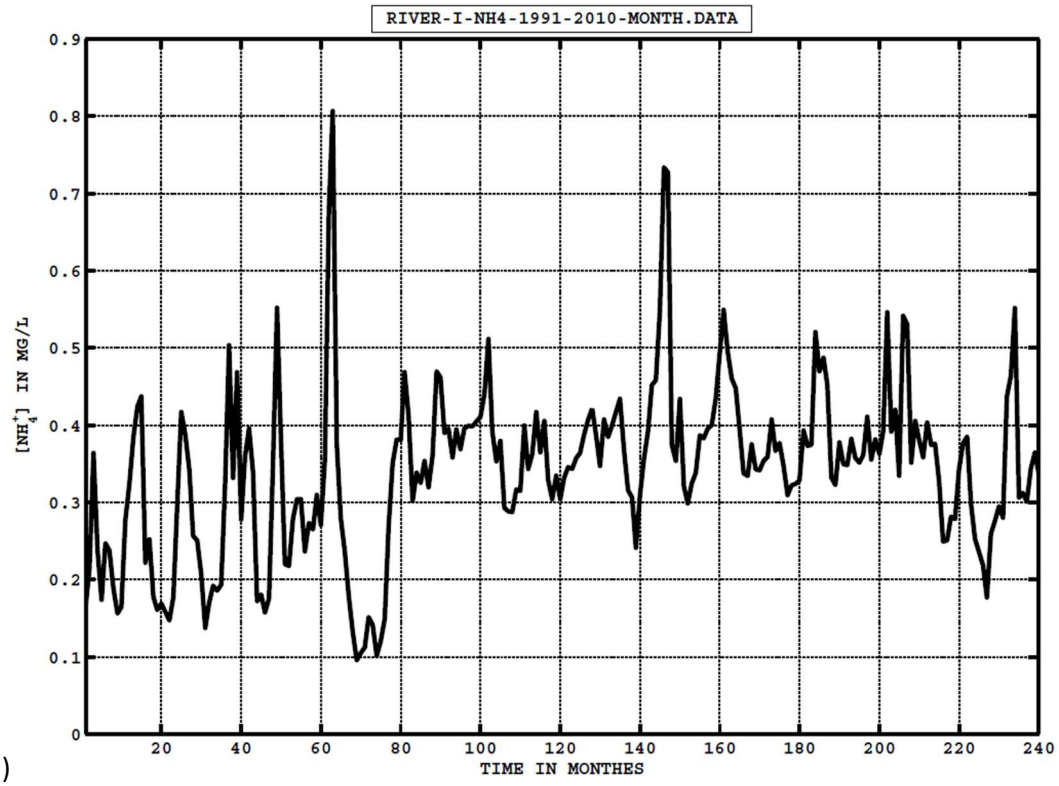
Всі інші показники нормальному розподіленню не підлягали.

Зважаючи на те, що наведений вище аналіз на нормальность буде явно недостатнім, була значно підвищена кількість відповідних тестів [1, 5].

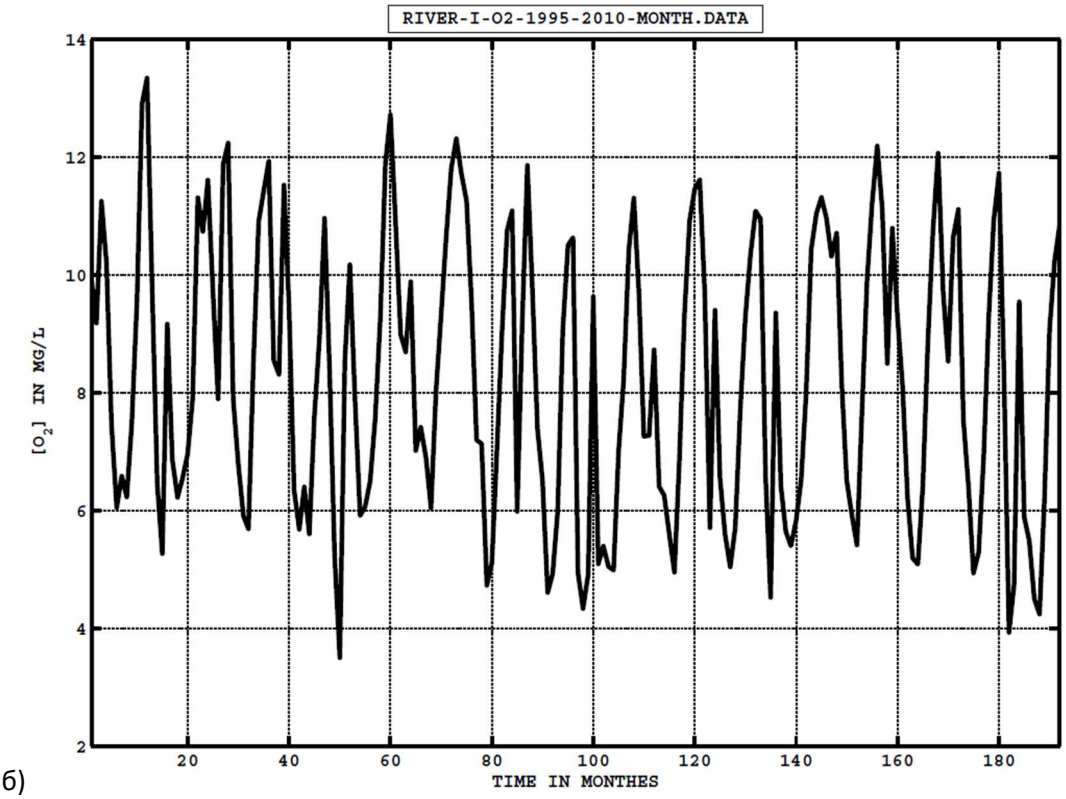
Використовували критерії асиметрії, Колмогорова – Смірнова та інші (порядка 10 критеріїв).

При цьому одержали подібні результати.

Якщо прийняти таку аналогію і для інших водних об'єктів, це може бути достатньо переконливим узагальненням, що має безпосередню практичну цінність.



а)



б)

Рисунок 1 – Графік ряда середньомісячних концентрацій амонійних іонів (а) та кисню, що розчинений у річковій воді (б).

Figure 1 - Graph of a number of average monthly concentrations of ammonium ions (a) and oxygen dissolved in river water (b).

Розглянемо адекватність положення, що гідрохімічний часовий ряд залежить від попередніх складових цього ряда.

Дослідження виконувались за допомогою автокорреляційної функції ряда (ACF).

З цією метою дві копії ряду зсувались на 1;2;3; … точки ряду відносно одна від іншої, отримуваючи таким чином так званий LAG.

Далі для кожного зсуву визначали коефіцієнт корреляції. Маючи достатню кількість точок, що відповідали коефіцієнту корреляції при певному значенні зсуву, з'єднували точки відрізками прямих, одержуючи остаточно графік ACF (див. рис.2а та рис 2б). Як видно, при значеннях LAG, що дорівнює одному – двум місяцям, "падіння" на графіках для нітратних іонів та кисню відносно невелике, тобто зберігаються достатньо високі значення коефіцієнта корреляції. Тому можливо говорити, що для цих рядів існує взаємозв'язок певної частини ряда із попередніми значеннями цього ряда. Тобто дані не можна вважати незалежними у цьому сенсі. Щодо амонійних та нітритних іонів, то такі взаємозв'язки у ряді виражені значно слабше.

У досліджені був використаний рівень значущості 0,05(5%), що є звичайним значенням у гідрохімії. Тому коефіцієнти корреляції, більші по модулю ніж 0,15, розглядались як значущі.

Але, як відомо, така класично виконана автокорреляція базується на лінійній корреляції Пірсона, що припускає нормальне розподілення даних [6].

Тому реалізацію автокорреляційної функції ряду (ACF) визначали також із застосуванням непараметричних корреляцій Кенделла та Спірмена [4].

При порівнянні графіків, що були одержані із використанням метода непараметричної корреляції, із графіками "класичної" корреляції з'ясувалось, що загальний вид графіків та відповідні статистики змінювались мало (значно менше, ніж ми припускали).

У цьому зв'язку оцінка залежності ряду від його певної копередньої складової за допомогою метода непараметричної корреляції та "класичним" методом можуть бути рівнозначно використані при гідрохімічних розрахунках. При цьому "класичний" метод оцінки функції повинен бути попередньо ретельно перевірений, особливо при наявності аномальних значень у вихідних даних.

Враховуючи, що гідрохімічні дані являють собою природні часові ряди, розглянемо далі таку важливу їх властивість як сезонність. Для цього використаємо дискретне перетворення Фур'є.

Фур'є – спектри потужності деяких гідрохімічних даних наведені відповідно на рис. 3а та рис. 3б.

Як показує наш досвід аналіза таких даних, графік Фур'є – спектра стає більш очевидним, якщо по горизонтальній вісі відкладати номер гармоніки. Приймаючи рівень значущості 5% та застосовуючи метод непараметричного бутстрепа, знаходимо значущі гармоніки для $[NH_4^+]$, $[NO_2^-]$, $[NO_3^-]$ та $[O_2]$.

Так, для $[NO_3^-]$ та $[O_2]$ виявляємо три такі гармоніки. Для $[NO_3^-]$ це буде 17-та, 20-та та 40-ва гармоніки, що відповідає періоду коливань

$240/17 \approx 14{,}1$ (місяця),
$240/20 \approx 12$ (місяців),
$240/40 \approx 6$ (місяців).

Для $[O_2]$ це відповідно 16-та, 32-а та 48-а гармоніки, що відповідає періоду коливань

$192/16 \approx 12$ (місяців),
$192/32 \approx 6$ (місяців),
$192/48 \approx 4$ (місяці).

У даному випадку 240 та 192 – довжина відповідних рядів (у місяцях).

Як видно, 12 – місячна та 6 – місячна періодичність відмічається у обох випадках.

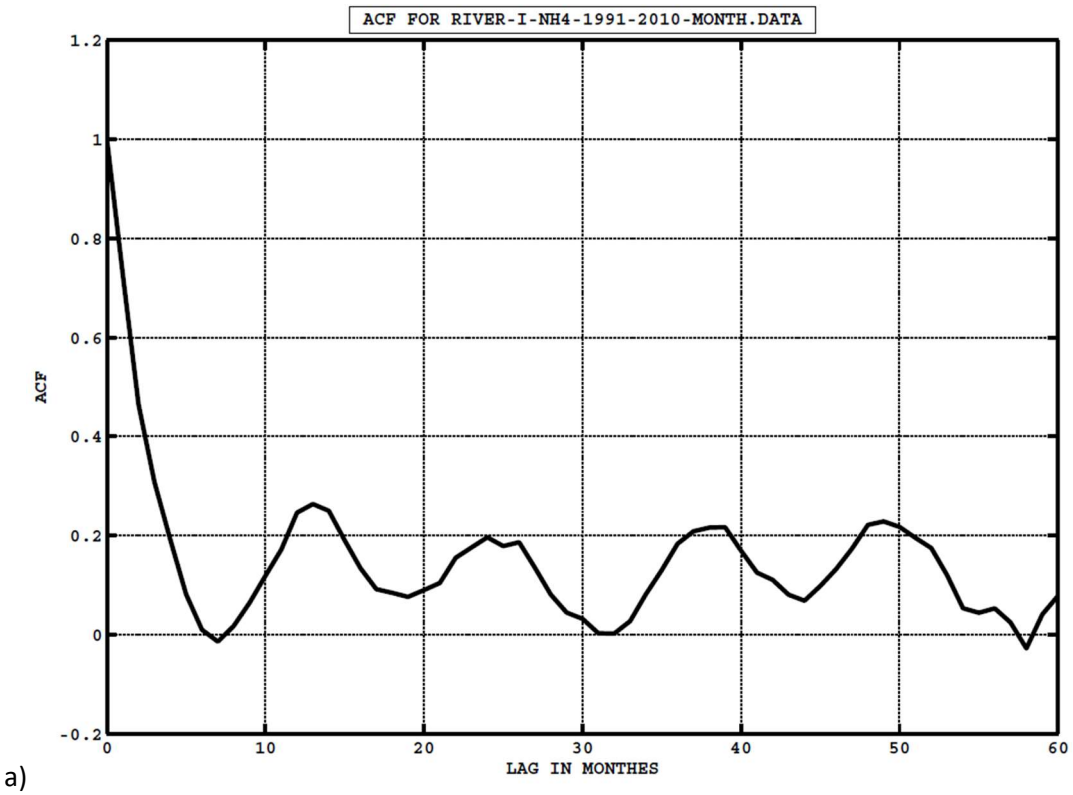
Крім того, для $[O_2]$ фіксується чітка 4 – місячна періодичність.

Для $[NH_4^+]$ та $[NO_2^-]$ значущіх гармонік достатньо багато: для $[NH_4^+]$ буде 9, а для $[NO_2^-]$ – 6 значущіх гармонік.

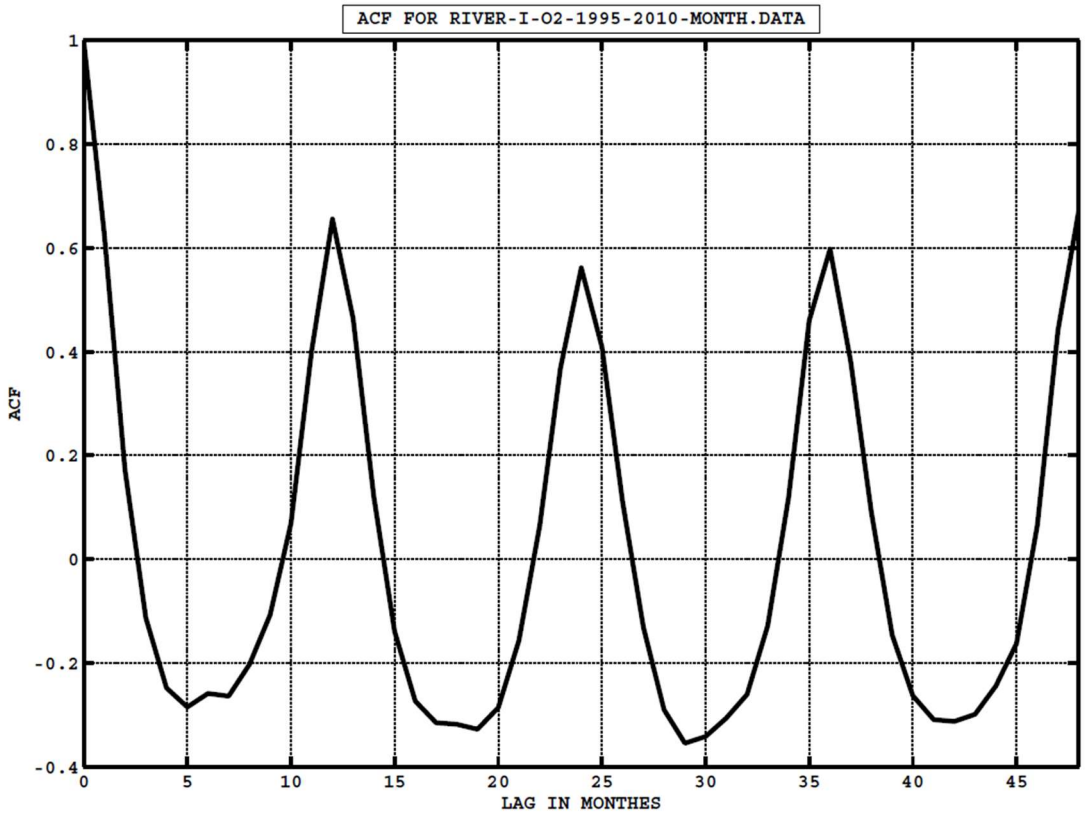
При цьому відмічаємо наявність коливань із періодом 12 місяців (1 рік).

У ряді $[NO_2^-]$ виділяємо також коливання із періодом 6 та 4 місяці.

Таким чином, часові гідрохімічні ряди можуть "розрізняти" сезони календарного року, що узгоджується із нашими попередніми дослідженнями[7].

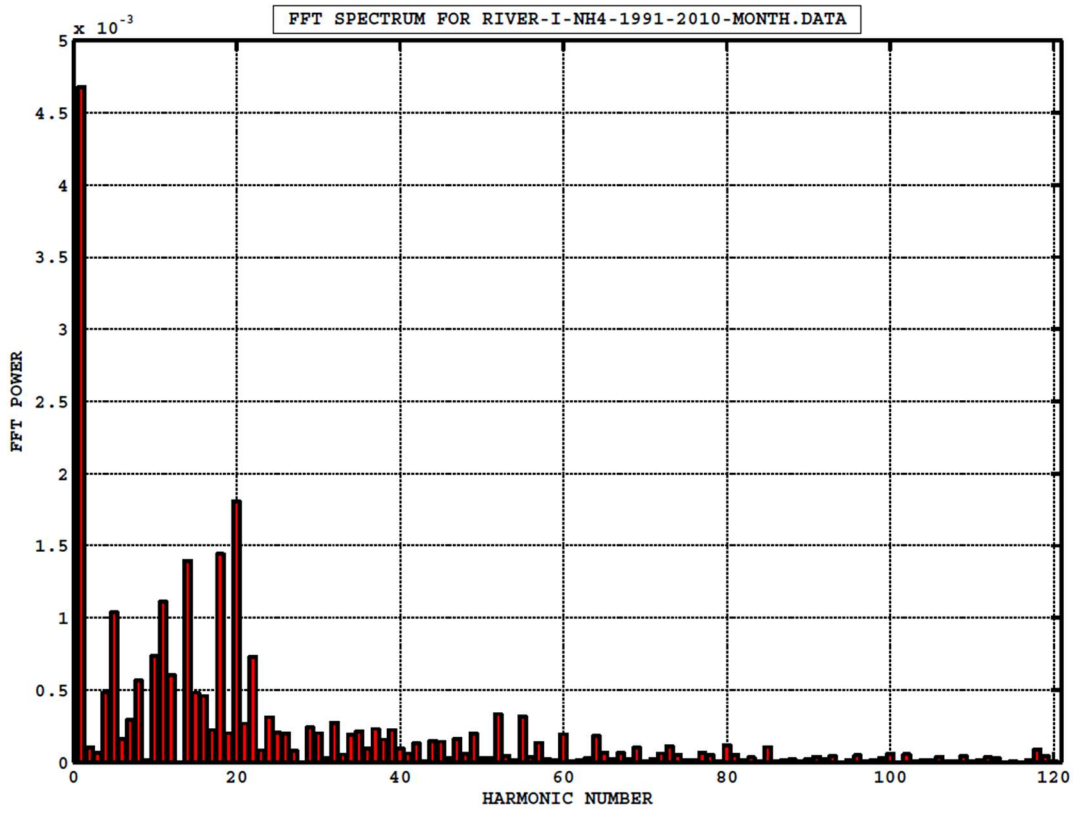


а)

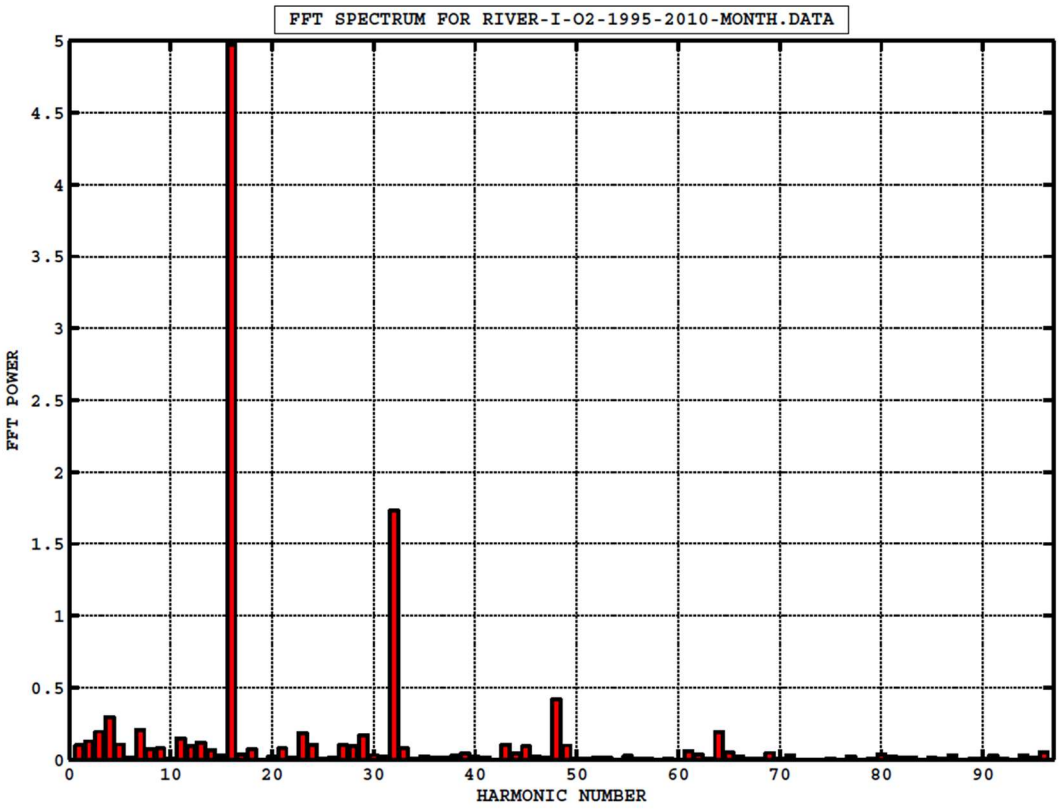


б)

Рисунок 2 – Графік класичної автокореляційної функції гідрохімічного ряда (ACF) для амонійних іонів (а) та кисню, що розчинений у річковій воді (б).
Figure 2 - Graph of the classical autocorrelation function of the hydrochemical series (ACF) for ammonium ions (a) and oxygen dissolved in river water (b).



а)



б)

Рисунок 3 – Фур'є – спектри потужності гідрохімічного ряда амонійних іонів (а) та кисню, що розчинений у річковій воді (б).

Figure 3 - Fourier - power spectra of the hydrochemical series of ammonium ions (a) and oxygen dissolved in river water (b).

Далі розглянемо зміни у поведінці гідрохімічних рядів впродовж кожного місяця. Для цього виберемо із ряда середньомісячних гідрохімічних показників всі значення за 20-річний період, що відповідають січню.

Таким чином, будемо мати вибірку із 20 точок. Далі знаходимо середнє із цих даних та одержуємо одне число (одну точку).

Виконаємо таку операцію відповідно для лютих, березней, … листопадів та грудней. У результаті маємо 12 точок, тобто по одній точці для кожного місяця.

Як відомо, у метеорології існує таке поняття як “середнє багаторічне ”, або “метеонорма”.

Враховуючи, що у метеорології більшість часових рядів підпорядковується закону нормального розподілення, використання значення “середнього арифметичного” буде цілком допустиме.

В той же час гідрохімічні дані, навпаки, мають статистичні розподілення, які досить далекі від нормального.

У цьому зв’язку поняття “метеонорма” для гідрохімічних часових рядів повинне базуватись не на “середньому арифметичному”, а на значенні медіани.

Якщо все ж таки використати значення “середнього арифметичного”, відмінність не буде набувати катастрофічного характеру.

Згідно рис. 4а та 4б, у даному разі спостерігається істотна залежність “середнього значення” (у даному випадку це медіана) від місяця року.

Слід врахувати, що значення окремих показників можуть змінюватись у продовж року у декілька разів (іноді майже на порядок!). Наприклад, для нітратних іонів $[NO_3^-]$ максимум значень спостерігаємо у лютому (біля 4,5 *мг/л*), а мінімум – наприкінці травня (0,6 *мг/л*). Тобто концентрація речовини за цей період змінювалась майже у 8 разів.

Якщо надалі будемо мати у цьому випадку надлишкову концентрацію речовин (за травень -червень), можно вважати, що режим водного об’єкта змінюється у гірший бік.

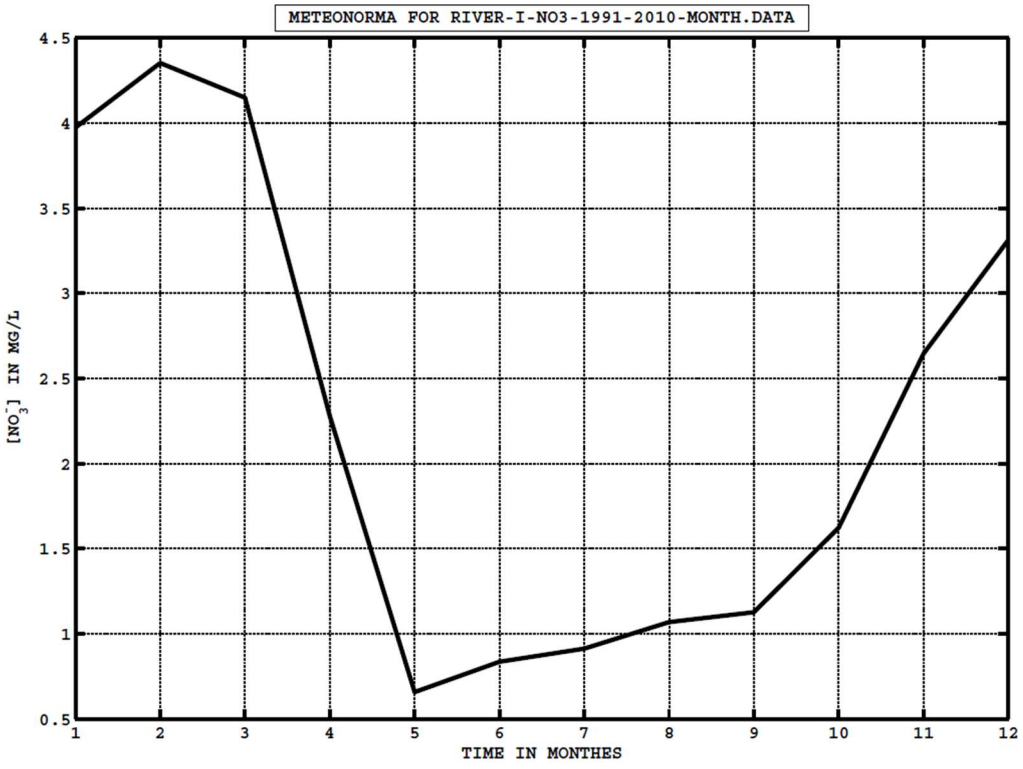
Дослідження, наведені у статті, супроводжувались також визначенням довірчих інтервалів, або значеннями *p-value,* або того чи іншого разом.

Висновки

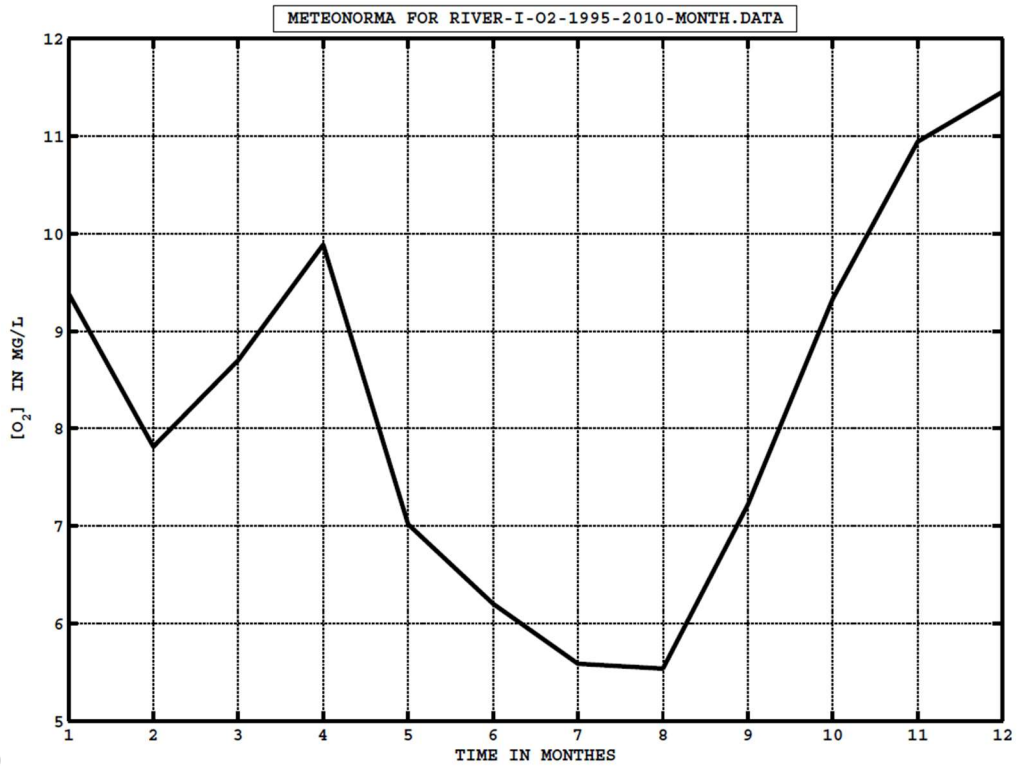
1. Показана неспроможність класичної парадігми обробки гідрохімічної інформації, яка базується на параметричних методах статистики.
2. Гідрохімічні дані слід обов’язково розглядати із позицій природних часових рядів. При цьому для гідрохімічних рядів вкрай важливим є порядок слідкування даних (у більшості випадків – за часом). Тобто окремі складові таких рядів не можна розглядати як повністю незалежні один від одного.

Певні складові гідрохімічних рядів залежать безпосередньо від попередніх складових ряда (спостерігається значна автокорреляція). По-суті, це означає, що є певний закон, який для практичних застосувань можливо вважати детермінованим, і на базі якого по попереднім значенням може бути встановлене конкретне значення ряду. Тобто гідрохімічні часові ряди є більше або менше прогнозованими, що є надзвичайно важливим при їх практичному застосуванні.

3. Гідрохімічні дані можуть бути адекватно описані лише при умові використання непараметричних методів статистики, оскільки у більшості своїй такі дані не підкорюються закону нормального розподілення. Також, як було встановлено експериментально, гідрохімічні дані проявляють нестаціонарність і по відношенню до багатьох інших законів розподілення.
4. На противагу окремим індивідуальним параметричним методам був застосований універсальний підхід при проведенні статистичних досліджень на базі використання непараметричного бутстрепа. З’явилась можливість проведення статистичних досліджень за єдиною методикою, практично із використанням однієї універсальної процедури.
5. Як показали проведені дослідження, застосування непараметричного підходу дозволяє у тому числі розрізняти сезони та зміни у поведінці гідрохімічних часових рядів упродовж календарного року.
6. Одержані у різні роки висновки щодо надійності застосування виключно параметричного підходу при проведенні екологічних досліджень потребують всебічного критичного аналізу та пересмотру застарілої парадігми.



а)



б)

Рисунок 4 – Значення медіани для нітратних іонів (а) та іонів кисню, що розчинений у річковій воді(б).
Figure 4 - Median value for nitrate ions (a) and oxygen dissolved in river water (b).

Перелік посилань

1. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М: ФИЗМАТЛИТ. –2006. – 816 с.
2. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com.
3. Малинецкий Г.Г. Современные проблемы нелинейной динамики/ Г.Г. Малинецкий, А.Б. Потапов -М.: Эдиториал УРСС, –2000. –336с.
4. Шитиков В.К. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R/ В.К. Шитиков, Г.С.Розенберг. – Тольятти: Кассандра. -2013. -314 с.
5. Дворкин В.И. Метрология и обеспечение качества количественного химического анализа. –М.: Химия. -2001. -263 с.
6. Андерсон Т. Статистический анализ временных рядов. –М.: Мир. -1978. -756 с.
7. Артеменко В.А. Сезонна динаміка біогенних речовин крупних водних об'єктів/ В.А. Артеменко, В.В. Петрович// Автомобільні дороги і дорожнє будівництво, вип. 104. –К.: Вид-во Націон. трансп. ун-ту. -2018. – С. 31-43.

Artemenko Vladislav A., Ukrainian Hydrometeorological Institute, State Service on Emergencies of Ukraine and National Academy of Science of Ukraine, Hydrochemical Research, Scientific Enployee, e-mail: artemenko@uhmi.org.ua, tel. 380936011250, Nauki alenue, 37, Kyiv, Ukraine, 03028.
Petrovych Volodymyr V., Candidate of Technical Sciences, Professor, Senior Researcher, Professor of the Transportation Construction and Property Management Department, National Transport University. e-mail: petrovichvv60@ukr.net, tel. +380442807338, Ukraine, 01010, Kyiv, street M. Omelyanovicha-Pavlenka, 1, room 138, https://orcid.org/0000-0003-0422-2535

Abstract. A modern paradigm of environmental data processing is developed, which is based on non-parametric methods of statistics.
The significant advantages of the proposed paradigm are given.
Keywords: ecological data, data processing, time hydrochemical series, nonpaparetic method of statistics.

References

1. Kobzar' A.I. Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov. – M: FIZMATLIT. –2006. – 816 s.
2. Mastitskiy S.E., Shitikov V.K. (2014) Statisticheskiy analiz i vizualizatsiya dannykh s pomoshch'yu R. – Elektronnaya kniga, adres dostupa: http://r-analytics.blogspot.com.
3. Malinetskiy G.G. Sovremennyye problemy nelineynoy dinamiki/ G.G. Malinets-kiy, F.B. Potapov -M.: Editornal URSS, –2000. –336s.
4. Randomizatsiya i budestrep: statisticheskiy analiz v biologii i ekologii s ispol'zovaniyem R/ V.K. Shitikov, G.S.Rozemnsberg. – Tol'yatti: Kassandra. -2013. -314 s.
5. Dvorkin V.I. Metrologiya i obespecheniye kachestva kolichestvennogo khimicheskogo analiza. –M.: Khimiya. -2001. -263 s.
6. Anderson T. Statisticheskiy analiz vremennykh ryadov. –M.: Mir. -1978. -756 s.
7. Artemenko V.A. Sezonna dynamika biohennykh rechovyn krupnykh vodnykh ob″yektiv/ V.A. Artemenko, V.V. Petrovych// Avtomobil′ni dorohy i dorozhnye budivnytstvo, vyp. 104. –K.: Vyd-vo Natsion. Transp. un-tu. -2018. – S. 31-43.