

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ МЕТОДА КВАНТИЛЬНОЇ РЕГРЕСІЇ В ЗАДАЧАХ ГІДРОЕКОЛОГІЇ

FEATURES OF THE APPLICATION OF THE METHOD OF QUANTILE REGRESSION IN PROBLEMS OF HYDROECOLOGY



Артеменко Владислав Андрійович, магістр екології, 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.

<https://orcid.org/0000-0003-0536-5415>



Петрови́ч Володи́мир Васи́льович, 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. Наведені особливості застосування методу квантильної регресії в задачах гідроекології. Визначені переваги кусочної квантильної регресії при обробці вихідних даних.

Ключові слова: гідрохімічні часові ряди, обробка вихідних даних, квантильна регресія, особливості кусочної квантильної регресії.

Вступ

В сучасний період при дослідженнях в області екології найбільш часто використовується регресія на основі методу найменших квадратів (МНК). Однак, як відомо, МНК диктує гранично жорсткі вимоги до властивостей вхідних даних і проявляє невиправдано підвищену чутливість до аномальних значень, які характерні для природних часових рядів. У зв'язку з цим при аналізі і прогнозі різних екологічних часових рядів більш виправдано використання робастних методів регресії, які будуть в тому числі слабо чутливі до різного роду аномалій у вихідних даних.

В роботі [1] на конкретних прикладах були розглянуті особливості застосування простого методу робастної лінійної регресії - методу KENDALL_THEIL_ROBUST_LINE (KTRL) - в гідроекології.

Разом з цим, з точки зору "глибини дослідження", методи лінійної регресії мають той недолік, що з даних видобувається занадто мало корисної інформації.

Перехід від лінійної регресії до регресії нелінійної може в значній мірі збільшити ефективність досліджень різних процесів [2].

У зв'язку з цим для дослідників значно перспективнішими є методи квантильної регресії [3,4].

Головна мета роботи - розглянути, яким чином квантильна регресія може знайти своє застосування на практиці.

Для цього в якості об'єкта дослідження були обрані гідрохімічні ряди, для аналізу яких ефективний метод квантильної регресії практично не використовувався.

Обговорюються переваги розробленого авторами методу кусочної квантильної регресії.

Загальні положення

У статті розглядається аналіз природних даних за допомогою методу квантильної регресії. Фахівці, які працюють з реальними даними, давно відзначали той факт, що аномальні значення (так звані "викиди") адекватно врахувати в моделях, побудованих на основі методу найменших квадратів (МНК), просто неможливо.

При цьому, якщо в даних присутні аномалії, мінімізація абсолютних відхилень буде більш перспективна, ніж метод МНК.

Як відомо, загальна ідея регресійного аналізу - це знаходження деяких "середніх" значень залежних змінних від змінних незалежних.

Будучи залежністю тільки одного параметра розподілу, а саме "середнього", регресійна функція повною мірою не буде відображати вплив всіх незалежних змінних на змінні залежні.

Відповідно, кілька регресійних кривих для різних частин розподілу дозволяють отримувати більш повну картину залежностей.

Вкрай спрощено це можна вважати характерною властивістю квантильної регресії.

На даний час добре відомий такий метод робастної регресії як "медіанна" регресія, де в якості "середнього" виступає медіана (або 0,5 квантиль). За аналогією можна взяти також набір будь-яких інших квантилів.

Зазвичай беруть наступний набір квантилів: 0.05; 0.10; 0.25; 0.50; 0.75; 0.90 і 0.95.

Таким чином, метод квантильної регресії при аналізі трендів дозволяє отримати інформацію по всьому діапазоні значень квантилів від нуля і до одиниці розподілу залежної змінної. Завдяки цьому отримують значно більше інформації, ніж при використанні методів лінійної робастної регресії (таких, наприклад, як метод KTRL [1]).

Як відомо, МНК - регресія, по-суті, є вирішенням задачі мінімізації залишкової суми квадратів:

$$\underset{\mu \in R}{\operatorname{ARGMIN}} \left(\sum_{j=1}^N (Y_j - \mu)^2 \right), \quad (1)$$

де Y_j – значення спостереження з вибірки обсягу N ,

μ – вибіркове середнє, що оцінюється за цією вибіркою.

Відповідно, подібний пошук для медіани може бути здійснений за допомогою мінімізації суми абсолютних залишків.

Тоді говорять про медіанну регресію.

Знаходження квантиля Q заданого порядку τ можна розглядати як пошук аргументу мінімуму спеціальної цільової функції

$$\underset{Q \in R}{\operatorname{ARGMIN}} \left(\sum_{j=1}^N \rho_{\tau}(Y_j - Q) \right), \quad (2)$$

В даному випадку ρ_{τ} – функція, яка забезпечує τ – баланс спостережних значень:

$$\rho_{\tau}(U) = \begin{cases} \tau \cdot U & \text{при } U \geq 0 \\ (\tau - 1) \cdot U & \text{при } U < 0 \end{cases} \quad (3)$$

У звичайному регресійному аналізі оцінка параметрів ρ регресійної функції є вирішенням оптимізаційної задачі

$$\underset{\beta \in R}{\operatorname{ARGMIN}} \left(\sum_{j=1}^N \left(Y_j - \mu(X_j; \beta) \right)^2 \right). \quad (4)$$

За аналогією з регресійним аналізом можливо ввести квантильно-регресійні функції, кожна з яких представляє собою деяку регресію умовної квантилі.

У цьому випадку побудова квантильно-регресійних моделей можливо розглядати як задачу оцінки параметрів функцій $v(X_j; \beta)$ і знаходити рішення мінімізацією

$$\underset{\beta \in R}{\operatorname{ARGMIN}} \left(\sum_{j=1}^N \rho_{\tau} \left(Y_j - v(X_j; \beta) \right) \right). \quad (5)$$

Вирішення даної мінімізаційної проблеми, коли $v(X_j; \beta)$ - лінійна функція з невідомими параметрами, ефективно здійснюється методами лінійного програмування.

Так, задачі із використанням МНК вирішуються через добуток матриць.

Реалізувати квантильну регресію значно складніше, ніж МНК.

Програмно використовується досить складна реалізація із застосуванням процедури мінімізації функції однієї або декількох змінних.

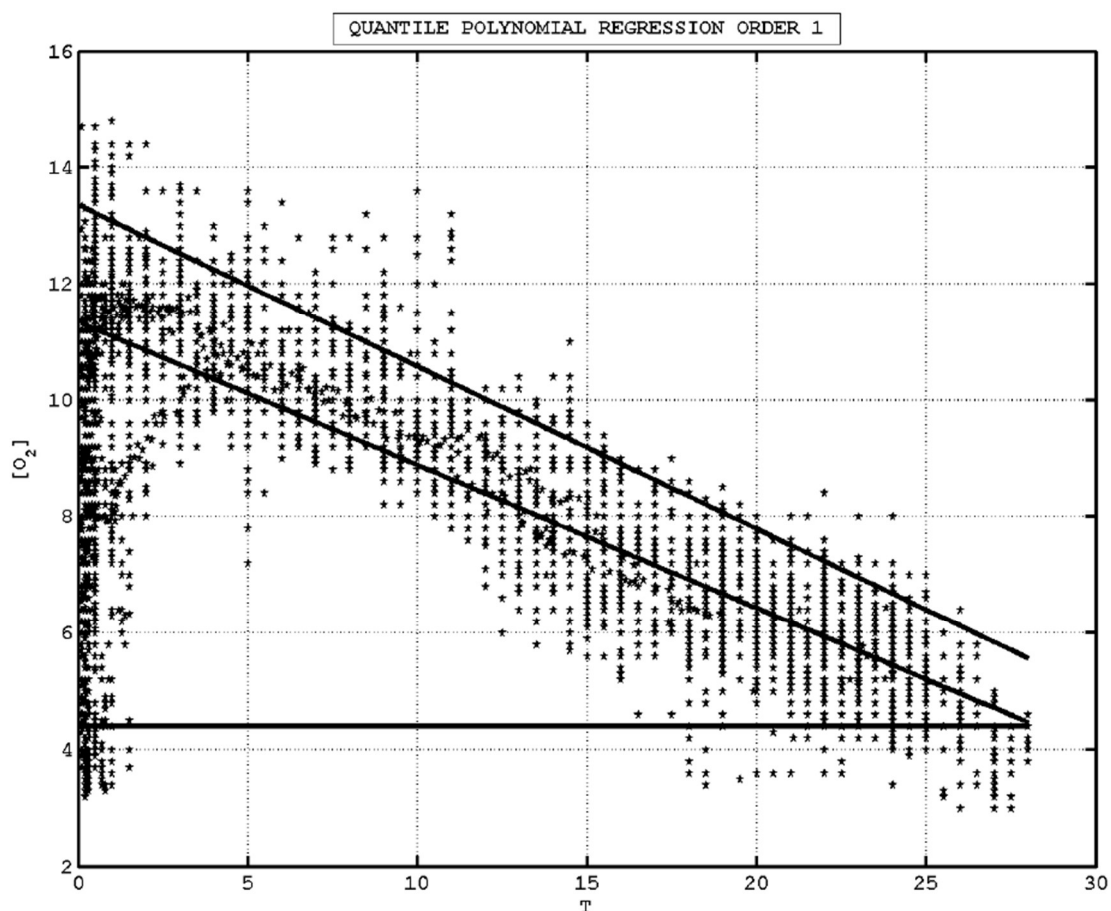


Рисунок 1 - Проведення ліній лінійної квантильної регресії (поліном першого ступеня) через облако експериментальних даних для оцінки залежності концентрації розчиненого кисню від температури річкової води.

Figure 1 - Drawing lines of linear quantile regression (polynomial of the first degree) through the "cloud" of experimental data for the dependence of the concentration of dissolved oxygen on the temperature of river water.

Процедура квантильної регресії реалізована, наприклад, в безкоштовній програмі GRETLM.

При цьому слід зважати на різноманітні “Керівництва” по роботі з GRETЛ.

Автори розробили власну процедуру на мові програмування MATLAB, яка реалізує метод квантильної регресії зі значно більшими можливостями, ніж GRETЛ.

Дослідження поведінки розчиненого кисню у водних об'єктах методом квантильної регресії.

При дослідженні були використані вихідні середньодобові дані за період 1995 ... 2010 роки (без пропусків) концентрації розчиненого кисню в річковій воді і воді водосховища, мг/л , температурі води, $^{\circ}\text{C}$, витрат річкової води, $\text{м}^3/\text{с}$.

Для всіх випадків побудовані графіки для квантилів 0.05, 0.50 і 0.95.

На графіках наведені вихідні дані у вигляді облака точок і ліній регресії для відповідних значень квантилів.

При цьому верхня лінія регресії - для значення $Q = 0.95$, середня лінія - для $Q = 0.50$ і нижня лінія - для $Q = 0.05$.

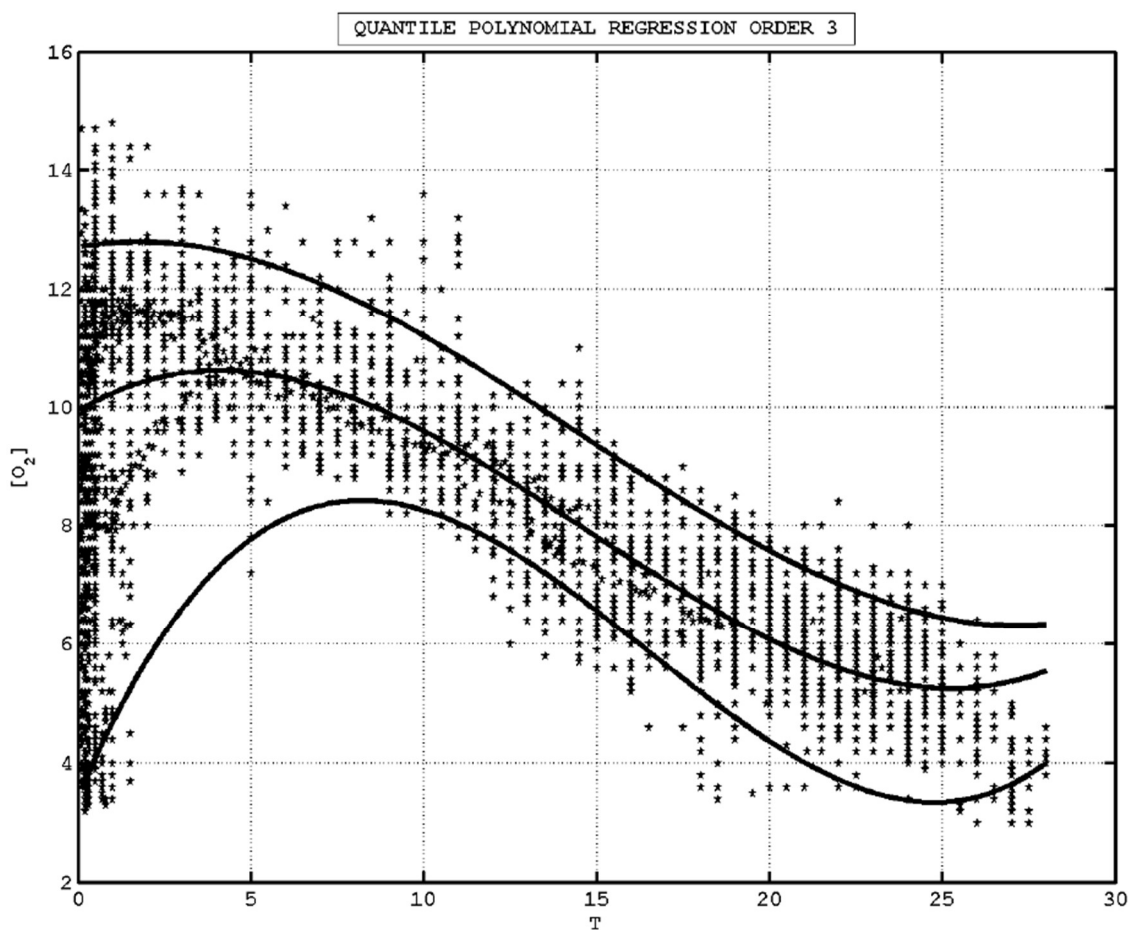


Рисунок 2 - Проведення ліній регресії через облако експериментальних даних для залежності концентрації розчиненого кисню від температури річкової води за допомогою методу квантильної регресії (поліном третього ступеня).

Figure 2 - Drawing the lines of quantile regression (polynomial of the third degree) through the "cloud" of experimental data for the dependence of the concentration of dissolved oxygen on the temperature of river water.

На рис.1 представлені лінії лінійної квантильної регресії (поліном першого ступеня) для залежностей концентрації кисню від температури річкової води. Як видно, якщо при даній температурі

води спостерігається велика ($Q = 0.95$) або навіть середня ($Q = 0.50$) концентрація кисню, використання прямолінійної залежності для ліній регресії буде більш-менш обґрунтованою.

В цьому випадку отримуємо класичну картину поведінки кисню, коли з підвищенням температури води концентрація розчиненого кисню зменшується.

Однак таку картину "псують" аномалії, які мають місце при температурах води від значення 00°C і до приблизно $+2,50^{\circ}\text{C}$.

Тоді спостерігається так званий "кисневий парадокс", коли при збільшенні температури води в цих межах збільшується і кількість розчиненого у воді кисню. Зрозуміло, що насправді парадоксу немає, якщо враховувати можливість утворення льоду на поверхні води.

Про такі явища можна говорити тільки при наявності дуже малих концентрацій розчиненого у воді кисню ($Q = 0.05$), коли кисень не надходить у воду з повітря і витрачається виключно на різні окислювані процеси.

При великих і середніх концентраціях розчиненого кисню льоду на поверхні водного об'єкта, як правило, немає, тобто кисень вільно надходить у воду з повітря.

Таким чином, для малих концентрацій розчиненого кисню пряма лінія, як обвідна облака експериментальних точок, однозначно не підходить.

У зв'язку з цим збільшимо ступінь полінома квантильної регресії до двох.

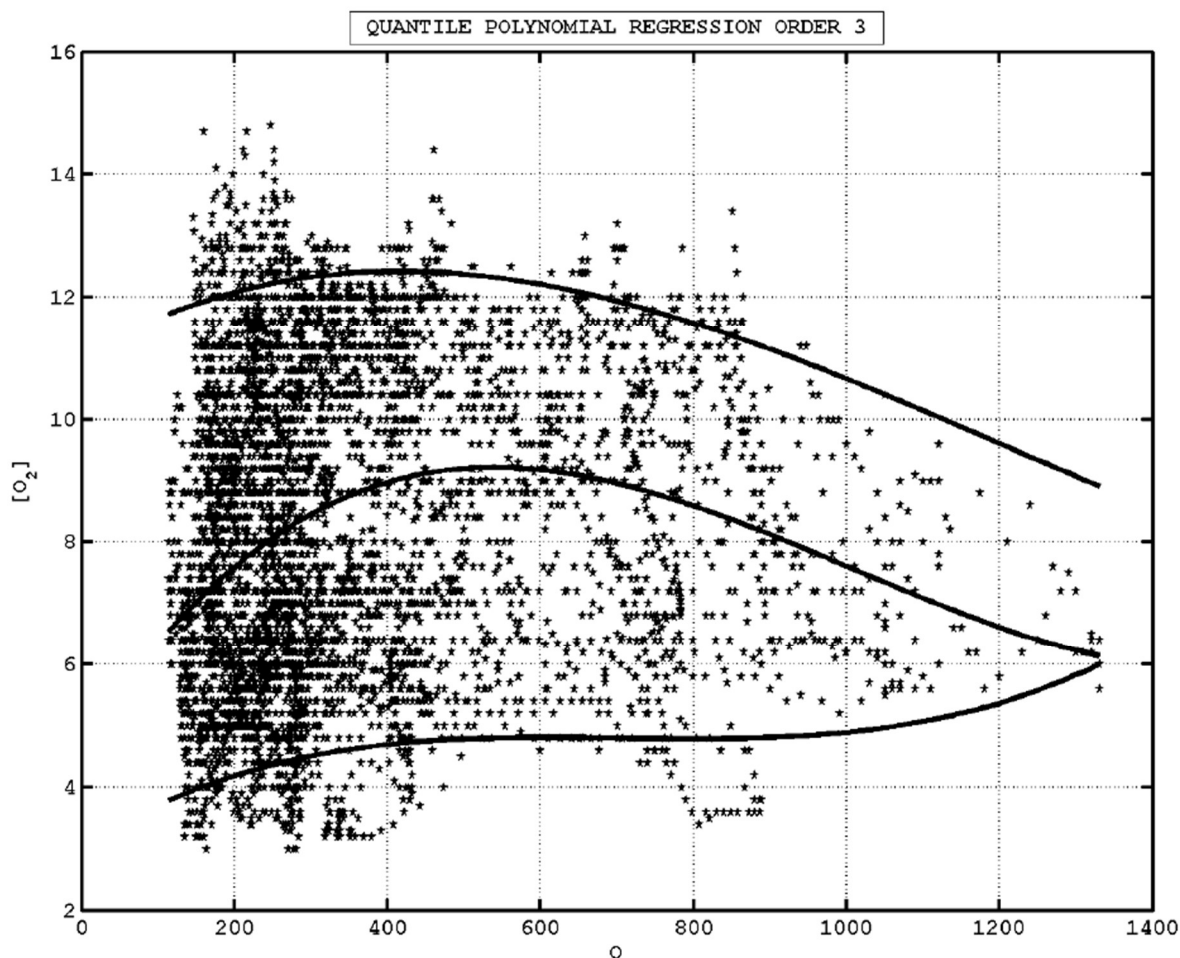


Рисунок 3 - Графік залежності концентрації розчиненого кисню від витрат річкової води, отриманий за допомогою квантильної регресії (поліном ступеня 3).

Figure 3 - A graph of the dependence of the concentration of dissolved oxygen on the flow rate of river water, obtained using quantile regression (regression polynomial 3).

Однак, як виявилось, в разі використання полінома другого порядку нижня обвідна також описується не зовсім коректно.

На рис. 2 наведені дані, отримані за допомогою квантильної регресії для полінома третього порядку.

Як видно, в даному випадку нижня обвідна експериментальних точок описується вже досить адекватно.

Подальше підвищення ступеня полінома не дало принципового поліпшення опису нижньої обвідної.

У зв'язку з цим графіки квантильної регресії для ступеня полінома більше трьох не будували.

Аналіз графіків, що характеризують інший природний водний об'єкт - водосховище, показав, що ми маємо, по-суті, схожі явища і в цьому випадку.

На рис. 3 наведений графік залежності концентрації розчиненого кисню від витрат річкової води, отриманий за допомогою квантильної регресії, для крупної рівнинної річки України (використаний поліном ступеня 3).

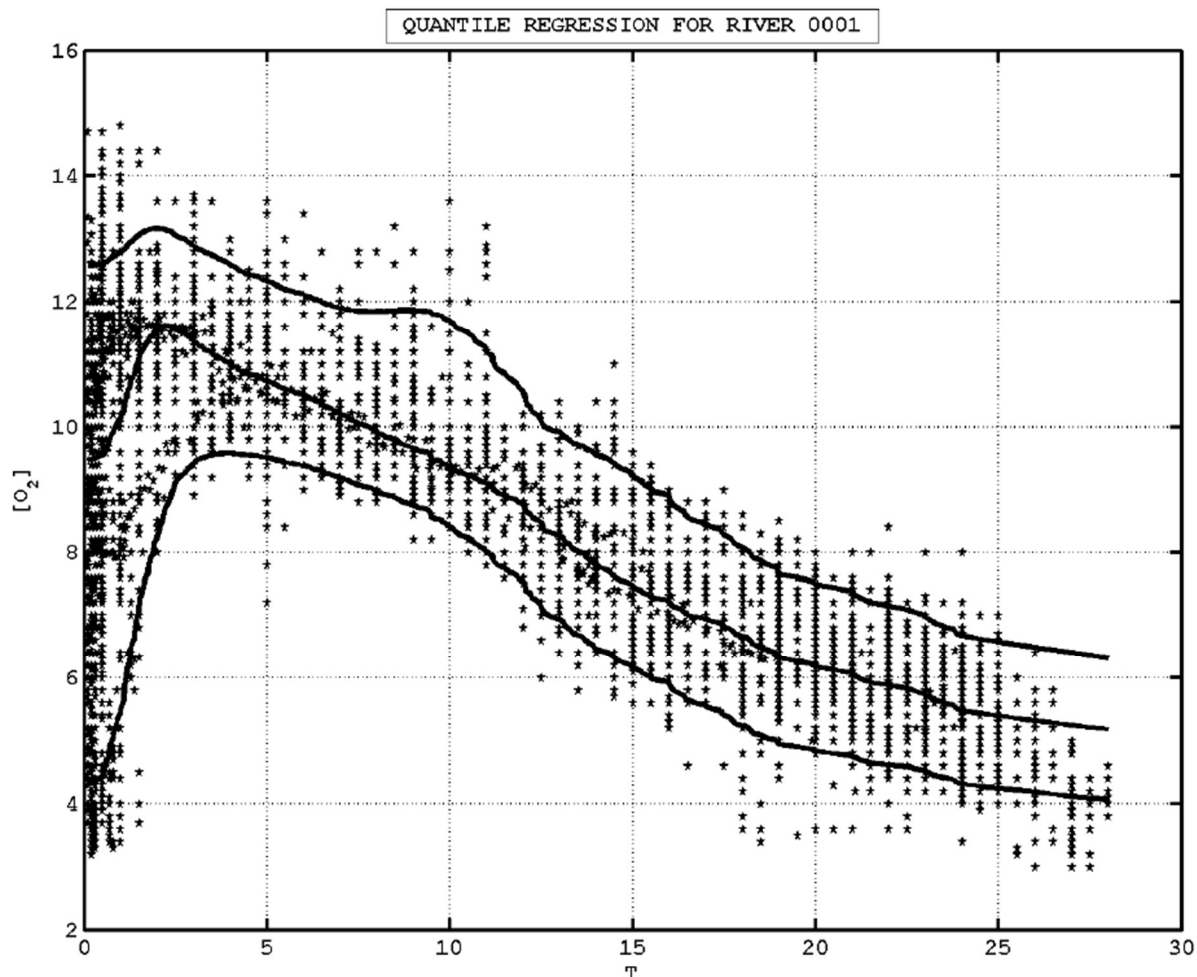


Рисунок 4 - Графік залежності концентрації розчиненого кисню від температури річкової води, отриманий за допомогою методу кусочної квантильної регресії (ступінь полінома 3).

Figure 4 - The graph of the dependence of the concentration of dissolved oxygen on the temperature of river water, obtained using the method of piecewise quantile regression (polynomial degree equal to 3).

У даному разі можливо говорити про те, що для квантилей $Q = 0.50$ та $Q = 0.95$ маємо максимум концентрацій розчиненого кисню при середніх витратах води (див. графік). Для квантилі $Q = 0.05$ спостерігається незначне зростання концентрації розчиненого кисню залежно від зростання витрат води.

Особливості реалізації методу кусочної квантильної регресії.

Далі розглянемо додаткові можливості методу квантильної регресії, не торкаючись самої сутності методу, але які дозволяють найбільш адекватно описувати експериментальні дані.

Раніше ми будували поліноміальну квантильну регресію для всіх наявних експериментальних точок відразу.

Тепер розіб'ємо всю множину значень незалежної змінної на ряд інтервалів рівної довжини. Потім для кожного такого інтервалу застосуємо окремо квантильну регресію з поліномом ступеня 3.

У цьому випадку виникають певні складнощі на стадії програмної реалізації такого підходу. Перша складність полягає у виборі кількості оптимальних сегментів, на які треба "розбити" вихідні дані. Інша складність полягає в "правильному сполученні" різних сегментів на їхніх кордонах.

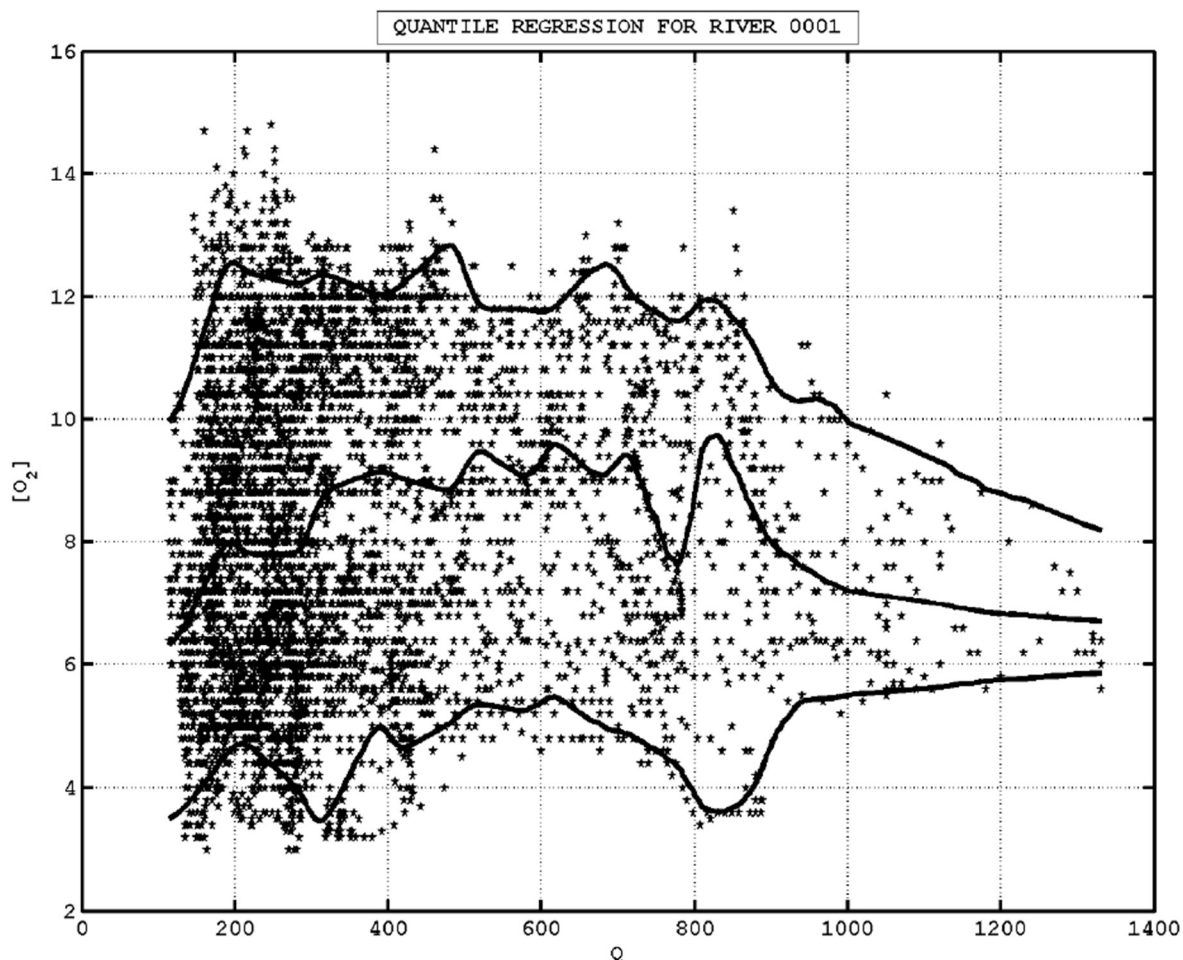


Рисунок 5 - Графік залежності концентрації розчиненого кисню від витрат річкової води, отриманий за допомогою методу кусочної квантильної регресії (ступінь полінома 3).

Figure 5 - The graph of the dependence of the concentration of dissolved oxygen on the flow rate of river water, obtained using the method of piecewise quantile regression (polynomial degree equal to 3).

Ці завдання були успішно вирішені (в статті не розглядаються).

При цьому кожному сегменту незалежної змінної вибираємо відповідні значення змінної залежної.

При дослідженні залежності концентрації розчиненого кисню від температури води для різних квантилів ширина сегмента приймалась рівною 0.50°C , при дослідженні залежності концентрації розчиненого кисню від витрат води ширина сегмента була $25\text{ м}^3/\text{с}$.

На рис. 4 наведені результати залежності концентрації розчиненого кисню від температури річкової води за допомогою методу кусочної квантильної регресії, на рис. 5 - результати залежності концентрації розчиненого кисню від витрат річкової води (ступінь полінома дорівнює 3).

Як видно, обробка експериментальних даних виходить значно краще, ніж при використанні звичайного методу квантильної регресії.

Висновки

1. Коротко розглянута сутність методу квантильної регресії і ті переваги, які вона дає при аналізі природних даних.
2. За допомогою методу квантильної регресії досліджено поведінку розчиненого кисню в залежності від температури та витрат річкової води при квантилях 0,95; 0,50 і 0,05 відповідно.
3. Як показали числові експерименти, одним з перспективних методів аналізу гідроекологічних процесів є розроблений авторами метод кусочної квантильної регресії.

Перелік посилань

1. 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.
2. Хардле В. Прикладная непараметрическая регрессия. – М.: Мир. – 1993. –349 с.
3. Koenker R. Quantile Regression. – Cambridge University Press. –2005. –366 p.
4. Hao L. Quantile Regression/ L. Hao, D. Naiman. – Sage Publications Inc. -2007. – 137 p.

FEATURES OF THE APPLICATION OF THE METHOD OF QUANTILE REGRESSION IN PROBLEMS OF HYDROECOLOGY

Artemenko Vladislav A., Master of Ecology, Ukrainian Hydrometeorological Institute, State Service on Emergencies of Ukraine and National Academy of Science of Ukraine, Hydrochemical Research, Scientific Employee, e-mail: artemenko@uhmi.org.ua, tel. 380936011250, Nauki avenue, 37, Kyiv, Ukraine, 03028, room 34, <https://orcid.org/0000-0003-0536-5415>

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. Such perspective method of the analysis of the natural time series is considered in article as method of quantile regression. They are discussed imperfection existing methods of the analysis of the natural processes (least square method and the method of linear robust regression). It is shown that transition towards quantile regression allows greatly to raise efficiency of the investigations of the natural time series. Main goal of the work consists in practical applications of the quantile regression method for decision of the different problems of the hydroecology.

Hydrochemical time series were considered in article. By means of quantile regressin method was investigated behavior of the dissolved oxygen in water depending on temperature of river water and depending

on discharge of river water (for quantiles of order 0.05; 0,50 und 0,95). When performing the investigations more perfect method of quantile regression was designed such as method piecewisf quantile regression (with uce polynomial degree 1; 2; 3).

Numenical experiments when use the natural time series have shown greater advantage of the designed method of piecewise quantile regression in contrast with classical method of quantile regression.

Keywords: hydrochemical time series, original data processing, quantile regression, features of piecewise quantile regression.

References

1. Artemenko V.A. Sezonna dynamika biohennykh rehovyn 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. – С. 31-43.
2. Khardle V. Prikladnaya neparametricheskaya regressiya. – M.: Mir. – 1993. –349 с.
3. Koenker R. Quantile Regression. – Cambridge University Press. –2005. –366 p.
4. Hao L. Quantile Regression/ L. Hao, D. Naiman. – Sage Publications Inc. -2007. – 137 p.