УДК 519.254

А.Г. Батурінець, С.В. Антоненко

Дніпровський національний університет ім. О. Гончара

ІДЕНТИФІКАЦІЯ СКЛАДОВИХ ЧАСОВОГО РЯДУ ГІДРОЛОГІЧНИХ ДАНИХ

Розглядається обчислювальна схема розкладання часового ряду на компоненти та її реалізація на прикладі даних гідрологічного моніторингу.

Ключові слова: *компоненти часового ряду; декомпозиція; тренд; сезонність*.

Рассматривается вычислительная схема разложения временного ряда на компоненты и ее реализация на примере данных гидрологического мониторинга.

Ключевые слова: *компоненты временного ряда; декомпозиция, тренд; сезонность*.

Conducting research on the hydrological data objects is an important step in monitoring their condition. Changes in the quality and quantity of surface waters has social, environmental and economic consequences and require careful monitoring and, if necessary, quick response to adverse changes. This article discusses the computational scheme for the allocation of components in the time series, as well as the implementation and analysis of the above scheme using the example of hydrological monitoring data.The developed computational scheme processes the data presented in the form of time series.The first step in the analysis of hydrological data is the determination of the primary statistical characteristics of the object under study. Among these characteristics: average, minimum and maximum values of the time series; median, variance, coefficient of variation, kurtosis and asymmetry of the time series. The preliminary assumptions about the presence of deterministic components in the time series are made on the basis of a correlogram analysis based on the initial data.To test the hypothesis of the presence in the time series of the trend component, the Fosters-Stewarts method is used. When confirming the hypothesis that a trend in the time series is present, the trend component is removed by the described linear regression model, the parameters of which are calculated by the least squares method. Identification of periodic component of a time series is carried out using the methods of the Fourier analysis. The dynamics of the series are determined with using the methods of linear and nonlinear regression, in particular: polynomial (3 degrees), power, exponential, Phillips curve and Engel curve, whose parameters are calculated by the least

squares method; modified exponential, the parameters of which are calculated by the method of three points. The analysis of the adequacy of the regression models is checked using the coefficient of determination and the adjusted coefficient of determination, also as the average approximation error. The Akaike information criterion and Bayesian information criterion Schwartz are used to determine the best model. The results of the study are accompanied by appropriate graphs and tables.

Keywords: *time series components; decomposition; trend; seasonality.*

Вступ. Сучасні умови різкого зростання інженерно-господарського освоєння та зміна кліматичних умов викликає низку несприятливих факторів, що в свою чергу має суттєвий вплив на соціальний та економічний стан країни. Актуальним є питання моніторингу водних об'єктів, оскільки являє собою систему регулярних спостережень за різними їхніми характеристиками.

Дані моніторингу представлені у вигляді деяких замірів протягом певного часу, тобто являють собою часові ряди.

Аналіз та прогнозування змін характеристик гідрологічних об'єктів дає можливість попереджувати або швидко та ефективно реагувати на негативні явища.

Надзвичайно актуальним є питання створення сучасних обчислювальних схем для ідентифікації складових в часових рядах даних гідрологічного моніторингу. Це спрощує розуміння структури та поведінки часового ряду, що, у свою чергу, надає можливості для більш ефективного підбору та використання методів аналізу, використання або створення/модифікації вже наявних алгоритмів прогнозування (оскільки дозволяє більш точно описати поведінку процесу в майбутньому) тощо.

Ускладнюється процес аналізу часових рядів даних гідрологічного моніторингу не тільки великими об'ємами накопиченої інформації, але тим, що дані зазвичай не є стаціонарними.

Аналіз літературних даних. Значний внесок у розробку теоретичних та методичних підходів аналізу даних, зокрема часових рядів, здійснили такі науковці: Т. Андерсон, С. Айвазян, В. Глушков, Дж. Джонстон, Г. Дженкінс, Дж. Кендалл, А. Стюарт та багато інших.

Серед робіт, що описують використання різних методик визначення та аналізу компонент часового ряду, можна визначити наступні:

1) в роботі [1] розглядається застосування до часового ряду таких методів, як сезонна декомпозиція, спектральний аналіз Фур'є, регресійний аналіз;

2) авторами роботи [2] пропонується методика декомпозиції часових рядів, що не підпорядковуються нормальному закону

розподілу, характеризуються відсутністю видимого тренду та невиконанням умови незалежності значень часового ряду;

3) використання методів кореляційного та спектрального аналізу для дослідження структури часових рядів наведено в роботах [3-6] .

Мета статті. Метою цієї роботи є розробка обчислювальної схеми, яка дозволяла б на сучасному рівні виділяти та аналізувати компоненти часового ряду даних гідрологічного моніторингу.

Основний матеріал. Під час аналізу даних гідрологічного моніторингу передбачається опрацювання великої кількості інформації, накопиченої протягом років та навіть десятиліть. Дані моніторингу представлені у вигляді замірів деяких показників протягом певного часу, тобто являють собою часові ряди.

Часовий ряд може, зазвичай, може бути описано:

– адитивною моделлю, тобто сумою всіх компонент часового ряду та мати вигляд: $x(t) = T(t) + S(t) + \xi_t$,

– мультиплікативною моделлю, тобто добутком всіх компонент часового ряду та мати вигляд: $x(t) = T(t) * S(t) * \xi_t$,

де $T(t)$ – трендова компонента; $S(t)$ – сезонна компонента; ξ_t – випадкова компонента.

В деяких випадках можливе використання моделі часового ряду змішаної форми, оскільки реальні дані можуть бути хаотично зашумлені, містити в собі як адитивні, так і мультиплікативні компоненти.

Вказані моделі є досить простими за структурою, яку можна використовувати для аналізу та прогнозування часових рядів.

При проведенні аналізу часового ряду необхідною умовою є однорідність даних. У даному випадку однорідність розуміється з точки зору структури даних, тобто дані повинні бути представлені однаковими показниками, в однакових одиницях виміру, а у випадку часових рядів – з однаковими проміжками часу. За наявності пропущених значень вони повинні бути відновлені.

Обчислювальна схема декомпозиції часового ряду:

1. Перевірка часового ряду на стаціонарність. Для перевірки ряду використовується метод Фостера-Стюарта. Цей метод має великі можливості ї дає більш надійні результати в порівнянні з, наприклад, перевіркою різниць середнього між рівнями ряду. Крім тренду самого ряду (тренду в середньому), цей метод дозволяє встановити наявність тренду дисперсії часового ряду. Реалізація методу проводиться наступним чином:

Порівнюємо кожен рівень вихідного часового ряду, починаючи другого рівня з усіма попередніми та обчислюємо величини *s* і *d*, при цьому: *s* – застосовується для виявлення тенденцій зміни дисперсії, а *d* – для виявлення тенденції у середній. Ці величини розраховуються за наступними формулами:

$$s = \sum_{i=2}^{n} (k_i + l_i), \tag{1}$$

$$d = \sum_{i=2}^{n} (k_i - l_i), \tag{2}$$

де k_i та l_i розраховуються наступним чином:

$$k_i = \begin{cases} 1, \text{якщо } x_i \text{ більше від усіх попередніх рівнів} \\ 0 - \text{ у протилежному випадку} \end{cases} \tag{3}$$

$$l_i = \begin{cases} 1, \text{якщо } x_i \text{ менше від усіх попередніх рівнів} \\ 0 - \text{ у протилежному випадку} \end{cases} \tag{4}$$

де x_i – значення відповідного рівня ряду, $i = \overline{2,n}$, n – довжина часового ряду.

За допомогою t-критерію Стьюдента перевіряємо наявність тенденцій:

$$t_1 = \frac{d}{f}\ , \tag{5}$$

$$t_2 = \frac{s - f^2}{m}\ , \tag{6}$$

де f та m мають розподіл Стьюдента з $v = n$ степенями вільності та розраховуються за наступними формулами:

$$f = \sqrt{2 \ln n - 3.4253} \tag{7}$$

$$m = \sqrt{2 \ln n - 0.8456} \tag{8}$$

Порівнюємо отримані значення t_1 та t_2 зі значеннями t-критерію $(t_{0,975})$ з *n-1* степенями вільності. Якщо $|t_1| > t_{0,975}$ та $|t_2| > t_{0,975}$, то гіпотеза про наявність тренду в часовому ряді приймається з імовірністю 95 %.

Якщо за результатами перевірки вихідний часовий ряд не є стаціонарним – переходимо на наступний пункт, якщо ряд стаціонарний – пункт 3.

Якщо часовий ряд є нестаціонарним, зведення до стаціонарного вигляду відбувається видаленням тренду. Для видалення тренду з часового ряду будуємо та видаляємо лінійну регресію, що визначається наступним рівнянням:

$$y_t = a + bx_t + e_t, \quad (9)$$

де a, b – параметри моделі, що розраховуються методом найменших квадратів (МНК), x_t – регресори, e_t– випадкова похибка моделі.

Оскільки лінійний тренд відображає не лише направлення динаміки часового ряду, але й середні значення, при видаленні його з початкового ряду отримаємо ряд, на підставі якого зручно виділити сезонну компоненту.

2. Визначення періодичної складової та виключення її з початкового часового ряду. Визначення періодичних коливань проводиться з використанням методів аналізу Фур'є [8].

Часовий ряд з виключеним трендом розкладається в ряд Фур'є за наступною формулою:

$$y = A_0 + \sum_{i=1}^{n} A_i \cos\left(i * \frac{2\pi i}{l}\right) + B_i \sin\left(i * \frac{2\pi i}{l}\right), \quad (10)$$

де l- період коливань, A_0, A_i, B_i, $i = \overline{1,n}$ – коефіцієнти Фур'є, n-довжина часового ряду, i – визначає номер гармоніки.

Параметри A_0, A_i, B_i розраховується за наступними формулами:

$$A_0 = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad (11)$$

$$A_i = \frac{2}{n}\sum_{j=1}^{n} x_j \cos\left(j * \frac{2\pi i}{l}\right), \quad (12)$$

$$B_i = \frac{2}{n}\sum_{j=1}^{n} x_j \sin\left(j * \frac{2\pi i}{l}\right), \quad (13)$$

Період l визначається за виглядом корелограми, що побудована за рядом з виключеним трендом або за допомогою амплітудного спектру [7].

3. Припускається, що в часовому ряді присутній тренд. Визначаємо моделі тренду в початковому часовому ряді без сезонної компоненти. Для опису моделей введемо наступні позначення:

y – описувана модель регресії, a, b, c, d, γ – розраховувані параметри,

e – випадкова похибка моделі, x – регресори,

t – момент часу відповідної складової, $t = \overline{1,n}$, де n – кількість елементів ряду.

Для виділення трендової компоненти розглядаються наступні моделі регресії:

– модель лінійної регресії є найпростішою з усіх та описується наступним рівнянням:

$$y = a + bx + e_t \quad (14)$$

– поліноміальна (3 степені) модель якої має вигляд:

$$y = ax^3 + bx^2 + cx + d + e_t \quad (15)$$

– степенева визначається наступним рівнянням:

$$y = a_0 * t^{a_1} + e^t \quad (16)$$

– показникова описується наступним рівнянням:

$$y = a_0 + a_1^t + e^t \quad (17)$$

– крива Філіпса описується рівнянням:

$$y = a_0 + \frac{a_1}{t} + e_t \quad (18)$$

– крива Енгеля задана в наступному вигляді:

$$y = \frac{1}{a_0 + a_1 t} + e_t \quad (19)$$

Зазначені вище моделі, окрім лінійної, є нелінійними за виглядом, але лінійними за параметрами. За допомогою логарифмічних перетворень вказані моделі зводяться до лінійного вигляду, а їх параметри розраховуються за МНК.

– модифікована показникова крива описана наступним рівнянням:

$$\mathrm{y} = \mathrm{a_0 a_1^t} + \gamma + \mathrm{e_t} \quad (20)$$

Виконуючи логарифмічні перетворення до рівняння модифікованої показникової кривої, зводимо рівняння до лінійного вигляду. При обчисленні параметрів першочергово необхідно визначити параметр γ. Оцінка параметрів проводиться з використанням МНК, проте обчислення значення параметру γ проводиться методом трьох точок.

4. Проаналізувати якість описання визначеними моделями тренду вихідного часового ряду. Для визначення якості побудованих моделей регресії використовуються наступні оцінки:

– Середня похибка апроксимації, що визначається наступним чином:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100, \quad (21)$$

де y_i – фактичне значення ряду, $\hat{y}_i$ – значення ряду за відповідною моделлю.

При значенні цього коефіцієнту не більше 8–10 %, традиційного вважається, що побудована модель є досить точною, в протилежному випадку говорять про недостатню якість побудованої моделі.

– R^2 – коефіцієнт детермінації, який визначає долю дисперсії залежної змінної, що пояснюється моделлю залежності (пояснювальними змінними). Цей коефіцієнт розраховується за формулою:

$$R^2 = 1 - \frac{\hat{\delta}^2}{\hat{\delta}_{\hat{y}}^2} = 1 - \frac{RSS}{TSS}, \quad (22)$$

Де RSS – сума квадратів залишків, TSS – загальна сума квадратів, $\bar{y}$ – середнє значення елементів часового ряду, розраховані за наступними формулами:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad (23)$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad (24)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad (25)$$

y_i – фактичне значення ряду, $\hat{y}_i$ – значення ряду за відповідною моделлю.

Значення коефіцієнта детермінації для моделі з константою приймає значення [0;1], чим ближче до 1, тим краще модель відповідає даним. Для прийнятних моделей значення коефіцієнта детермінації повинно бути не менше 0,5, а моделі зі значенням коефіцієнта детермінації більше 0,8 визнаються досить ефективними для описання ряду, значення коефіцієнта детермінації рівному 1 визначає функціональну залежність між змінними.

– скориговане R^2 використовується для порівняння моделей з різним числом факторів, визначаємо за формулою:

$$R^2{}_{\text{зкор}} = 1 - \frac{\delta^2}{\delta_{\hat{y}}^2} = 1 - (1 - R^2)\frac{(n-1)}{(n-k-1)}, \quad (26)$$

де n – кількість спостережень, k – кількість параметрів.

Цей показник завжди менший відодиниці. Теоретично при досить малому значенні коефіцієнта детермінації та великій кількості факторів може набувати значення менше нуля, тому втрачається

інтерпретація показника як «частини», але тим не менш показник застосовується для порівняння.

5. Використовуючи інформаційні критерії, визначаємо найкращу модель регресії для описання відповідного часового ряду. Реалізовано наступні інформаційні критерії:

– AIC – інформаційний критерій Акаіке визначає міру якості статистичних моделей для заданого набору даних, визначається наступним чином:

$$AIC = \frac{2 * k}{n} + \ln\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right), \quad (27)$$

де y_i – фактичне значення ряду, $\hat{y}_i$ – значення ряду за побудованою моделлю, n – довжина часового ряду, k – кількість параметрів, що описуються

– BIC – байєсовський інформаційний критерій Шварца, в термінах залишкової суми квадратів розраховується за формулою:

$$BIC = \frac{k * \ln n}{n} + \ln\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right), \quad (28)$$

де y_i – фактичне значення ряду, $\hat{y}_i$ – значення ряду за побудованою моделлю, n – довжина часового ряду, k – кількість параметрів, що описуються.

При порівнянні моделей перевага віддається тій, для якої значення BIC та AIC є меншими.

BIC зазвичай штрафує вільні параметри сильніше за – інформаційний критерій Акаіке, хоча це залежить від розміру n і відносної величини n і k.

6. Будуємо тренд-періодичну компоненту (адитивно) та визначаємо, наскільки якісно вона описує початковий часовий ряд, використовуючи показник середньої похибки апроксимації.

Обчислювальний експеримент. Розглядається часовий ряд, представлений щоденними показниками рівня води в р. Дніпро в період з 1.01.2010 р. по 31.12.2014 р.

Графічне відображення часового ряду представлено на рис. 1.

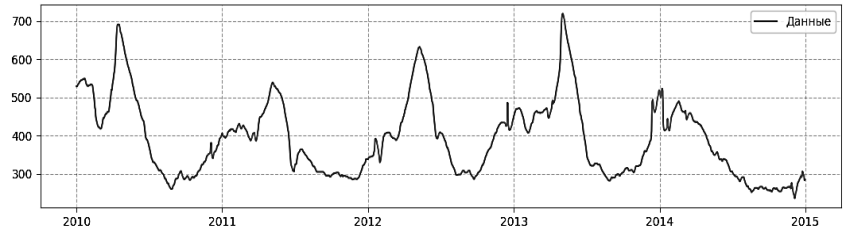

Рисунок 1 – Рівень води в р. Дніпро (1.01.2010 р. - 31.12.2014 р.)

Використовуючи описову статистику даних, отримаємо наступні показники:

Середнє значення рівня води – 390,77 (см);

медіана часового ряду – 375,5 (см);

дисперсія часового ряду – 0,26;

мінімальне значення рівня води – 236,0 (см);

максимальне значення часового ряду – 719,0 (см);

коефіцієнт варіації часового ряду – 0,26;

ексцес часового ряду – 0,06;

асиметрія часового ряду – 0,78.

З корелограми (рис. 2), побудованої на значеннях вхідних даних, можна зробити припущення про залежність рівнів ряду, наявність в ряді тенденції та сезонної компоненти з періодом приблизно 365 днів.

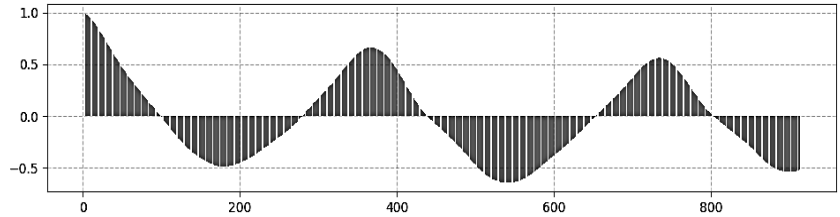

Рисунок 2. – Корелограма показників рівня води в р. Дніпро

Для отримання інформації щодо наявності тенденції в часовому ряді перевіряємо наявність тенденції середнього та дисперсії, використовуючи метод Фостера-Стюарта.

Для заданого часового ряду отримуємо значення:

s =125,0; d = -49,0; t_1= 35,5; t_2 = -13,02; $t_{0,975}$= 1,96.

Оскільки $|35{,}5| > 1{,}96$ та $|-13{,}02| > 1{,}96$, то гіпотеза про наявність у часовому ряді тренду дисперсії та середнього приймається з імовірністю 95 %.

У зв’язку з підтвердженням нестаціонарності часового ряду

видаляємо лінійну складову та отримуємо часовий ряд наступного вигляду (рис. 3)

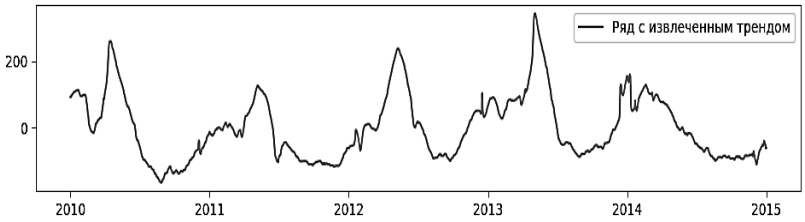

Рисунок 3 – Стаціонарний часовий ряд

Використовуючи Фур'є аналіз, отримуємо періодичну компоненту, що зображено на рис. 4.

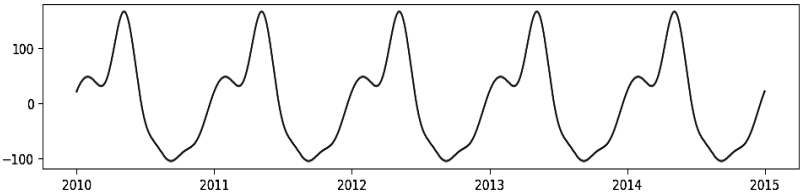

Рисунок 4 – Графічне відображення періодичної складової

Період коливань, визначений за спектром ряду, складає 365,2 днів, кількість використаних гармонік для опису періодичної складової – 5.

Видаляємо з початкового ряду отримані значення періодичної складової, після чого часовий ряд приймає наступний вигляд:

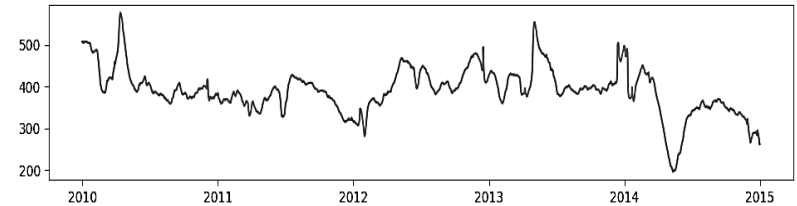

Рисунок 5 – Вихідний часовий ряд без періодичної складової

До часового ряду з видаленою періодичною компонентою будуємо моделі регресії та аналізуємо якість описання обраними моделями часового ряду на підставі визначення показників, визначених в табл. 1.

Таблиця 1

Зведена таблиця показників якості моделей регресії

Модель регресії	Коефіцієнт кореляції, R	R^2	$R^2_{(зкор)}$	Середня похибка апроксимації, %
Лінійна	0,29	0,09	0,08	10,8
Поліноміальна (3 степені)	0,63	0,40	0,39	8,61
Степенева	0,33	0,11	0,11	10,61
Показникова	0,28	0,08	0,08	10,8
Крива Філіпса	0,18	0,03	0,03	10,7
Крива Енгеля	0,23	0,05	0,05	10,88
Модифікована показникова	0,29	0,08	0,08	10,78

Аналізуючи результати, представлені у зведеній таблиці показників якості моделей регресії, визначаємо, що за коефіцієнтами детермінації ($R^2_{(зкор)}$, R^2) всі моделі не є прийнятними, оскільки їхнє значення менше 0,5. Значення коефіцієнта кореляції (R) означає, що для більшості моделей характерний слабкий зв'язок з досліджуваними даними та досить високий коефіцієнт похибки апроксимації. Для поліноміальної кривої третього порядку визначається середня сила зв'язку з досліджуваними даними та значення похибки апроксимації складає менше 10 %, що є достатньо непоганими показникам.

Оскільки за визначеними вище характеристиками неможливо адекватно визначити найкращу регресійну модель, особливо звертаючи увагу на різну кількість параметрів, за якими побудовано відповідні моделі, звертаємось до інформаційних критеріїв AIC та BIC. Отримані результати наведено в табл. 2.

Таблиця 2

Зведена таблиця значень інформаційних критеріїв

Модель регресії	Кількість параметрів регресії	Значення AIC	Значення BIC
Лінійна	2	7,93	15,45
Поліноміальна (3 ступеня)	4	7,51	15,04
Степенева	2	7,91	15,42
Показникова	2	7,94	15,46
Крива Філіпса	2	7,99	15,5
Крива Енгеля	2	7,97	15,48
Модифікована показникова	3	7,94	15,45

Проаналізувавши отримані результати, можемо зробити висновок, що серед реалізованих моделей регресії трендова складова часового ряду найкраще описується за допомогою поліноміальної кривої третього порядку; зобразимо графічно побудовану лінію регресії та значення даних вихідного часового ряду (рис. 6).

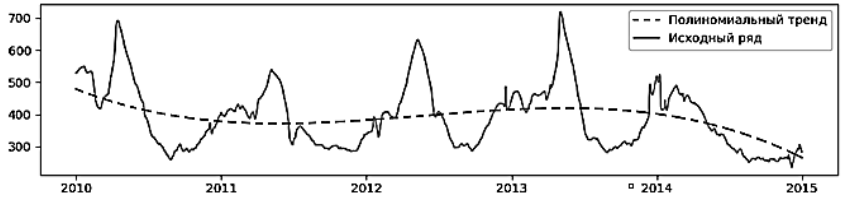

Рисунок 6 – Графічне зображення поліноміальної кривої та вихідного ряду

Побудуємо тренд-періодичну модель наступного вигляду: $y = (S(t) + T(t))$, де $S(t)$ – періодична складова, $T(t)$ – трендова компонента, t – відповідний момент часу, $t = \overline{1,n}$, n – довжина часового ряду. Не важко помітити, що ця тренд-періодична компонента загалом повторює поведінку часового ряду (рис. 7). Ряд А відображає значення вихідного часового ряду, а ряд Б – тренд-періодичної моделі.

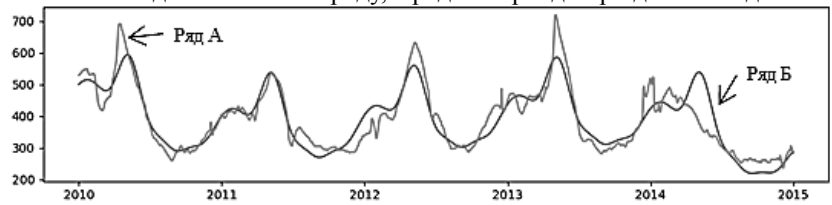

Рисунок 7 – Графічне зображення тренд-періодичної моделі

Наступним етапом є видалення з часового ряду тренд-періодичної компоненти. В цій обчислювальній схемі тренд-періодична модель є адитивною, тобто обчислення залишків відбувається за формулою:

$$y = X(t) - (S(t) + T(t)),$$

де X – значення часового ряду, t – відповідний момент часу, $t = \overline{1,n}$, n – довжина часового ряду.

Отримуємо залишки наступного вигляду (рис.8).

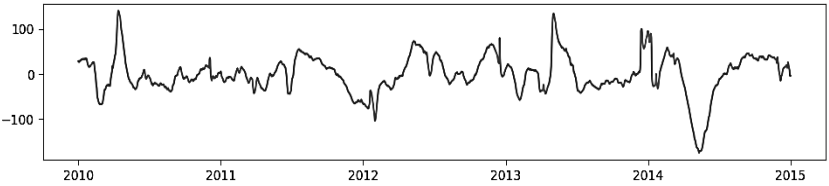

Рисунок 8 – Залишки часового ряду

Значення середньої похибки апроксимації для цієї тренд-періодичної моделі складає 8,11 %, що означає досить високу точність побудованої моделі.

Висновки та перспективи подальшої роботи. Розроблена в цій статті обчислювальна схема дозволяє ефективно ідентифікувати складові часового ряду. Це сприяє спрощенню розуміння поведінки ряду, що в свою чергу полегшує вибір методів прогнозування, групування та подальшого аналізу гідрологічних даних. Планується подальша розробка програмної системи з реалізацією методів прогнозування поведінки показників гідрологічних даних.

Бібліографічні посилання

1. Михалап С. Г., Мингалёв Д. Э., Евдокимов С. И. Использование анализа временных рядов в изучении многолетних температурных измененний // Вестник Псковского государственного университета Серия «Естественные и физико-математические науки», 2014. Вип. 4. С.17–24.

2. Тебуева Ф. Б., Перепелица В. А., Кабиняков М. Ю. Декомпозиция и прогнозирование временных рядов с долговременными корреляциями // Известия ЮФУ. Технические науки. 2013. № 1 (138). С.111–120.

3. Балабух В. О. Мінливість дуже сильних дощів та сильних злив в Україні. // Наукові праці УкрНДГМ. 2008. Вип.1. С. 61–72.

4. Сусідко М. М., Лук’янець О. І. Багаторічні коливання водності в Україні. // Гідрологія, гідрохімія і гідроекологія: наук. збірник. К.: ВГЛ «Обрії», 2010. Т.4 (21). С. 34–40.

5. Токмакова А. А. Выделение периодической компоненты из временного ряда // Машинное обучение и анализ данных. 2011. № 1. С. 40–50.

6. Протасов Ю. М., Юров В. М. Гармонический анализ периодических колебаний объёмов продаж компании на основе

инструмента «регрессия» MS Excel // Вестник Московского государственного областного университета. Серия: Экономика.

7. Витязев В. В. Спектрально-корреляционный анализ равномерных временных рядов: уч. пособие. СПб.: Изд-во С.–Петерб. ун-та, 2001. 48 С.

Надійшла до редколегії 11.11.2018