УДК 51.74:004.942(043.2)

О. В. Дергунов, Ю. В. Куц, Л. М. Щербак

ПЕРВИННА ОБРОБКА ЧАСОВИХ РЯДІВ ПРИ ОБМЕЖЕНИХ АПРІОРНИХ ДАНИХ ПРО ЇХ МОДЕЛЬ

The purpose of this article is analysis of priori uncertainty elimination methods in poorly studied process experimental results interpretation under conditions of limited priori knowledge about research process models. Two modern adaptive methods that can be used at experimental data preprocessing stage: empirical mode decomposition and singular spectral analysis (caterpillar) are presented. Comparative analysis of these two methods by power consumption analysis example is performed.

Key words: empirical mode decomposition, singular spectral analysis (“Caterpillar”), time-series preprocessing.

Проаналізовано шляхи зменшення апріорної невизначеності досліджень слабо вивчених фізичних явищ, процесів, математичні моделі яких не визначені. Наведено результати використання двох сучасних адаптивних методів які можуть бути використані на етапі первинної обробки експериментальних даних: емпірична модова декомпозиція та сингулярний спектральний аналіз («Гусениця»). Проведено порівняльний аналіз результатів використання цих методів на конкретному прикладі статистичної обробки часового ряду процесу електроспоживання.

Ключові слова: емпірична модова декомпозиція, сингулярний спектральний аналіз, первинна обробка часових рядів.

Розвиток земної цивілізації потребує постійного розширення і поглиблення знань про навколишній світ, стимулює проведення досліджень і вивчення явищ та процесів у різних предметних областях – від кліматології до палеоокеанографії, від медицини до ядерної фізики. Методологія наукового пізнання світу охоплює інформаційне забезпечення (гіпотези, методики проведення експериментальних досліджень, дані експериментів тощо), математичного забезпечення (математичні моделі, алгоритми обробки даних) і програмного забезпечення (прикладні програми, які реалізують алгоритми обробки даних і візуалізацію результатів на всіх етапах). Слід зазначити, що саме вимірювальні експерименти дають змогу отримати об’єктивну первинну інформацію, яка необхідна для створення і удосконалення математичних моделей досліджуваних процесів і явищ [1]. З метою більш досконалого проведення експериментів і отримання емпіричних даних створюється новий інструментарій вимірювальної техніки, розробляються методики вимірювань. У деяких випадках такі дані містять корисну інформацію у прихованому, «замаскованому» виді. Інтерпретація отриманих результатів експериментальних досліджень та виявлення прихованих закономірностей суттєво ускладнюється внаслідок відсутності апріорної інформації про динаміку чи навіть про загальний характер досліджуваних процесів і явищ, що дозволило би обґрунтувати застосування для їх аналізу певного математичного апарату. Водночас відомі методи дослідження сигналів, як правило, при використанні вимагають виконання певних умов і мають відповідні обмеження, що в цілому призводить до звужування класу досліджуваних процесів. Виконання таких умов і обмежень іноді пов’язані з труднощами, а часом і неможливістю їх обґрунтування, що не дає змоги обрати методику вимірювального експерименту. За цієї обставини коректний аналіз та обґрунтування фізичного тлумачення отриманих емпіричних даних перетворюється на ключовий етап пізнання об’єкта дослідження.

Одним із можливих типів емпіричних даних, які отримуються за результатами експериментальних досліджень, є часові ряди. Часові області визначення таких рядів коливаються у значному діапазоні – від декількох секунд до мільйонів років.

Особливістю аналізу часових рядів є важливість збереження послідовності отриманих даних, що забезпечує збереження інформації про розвиток досліджуваних процесів у часі. Слід зазначити, що за результатами аналізу важливо не тільки розкрити загальні тенденції перебігу і розвитку динамік досліджуваного процесу, але й виявити його локальні особливості.

Останнім часом в питанні дослідження процесів за відсутності достатньої апріорної інформації все більш застосовується нова тенденція, яка базується на методах обчислювальної математики та новітніх інформаційних технологій. Такі методи не мають чіткого математичного обґрунтування, але їх методологія виглядає логічно і несуперечливо і приводять до результатів, які узгоджуються з фізичною сутністю процесів. До таких методів належать методи, відомі як «Емпірична модова декомпозиція» (EMD) [2] та «Сингулярний спектральний аналіз» (SSA), відомий у вітчизняній літературі під назвою «Гусениця» [5]. Їх основними перевагами є здатність представляти аналізований ряд сумою адитивних компонент, які зазвичай містять чітку частотно-часову інформацію про фактори, що впливають на досліджуваний процес. Від інших методів аналізу сигналів, наприклад, від перетворення Фур'є, їх відрізняє відсутність зв'язку з апріорно заданим базисом розкладу при представленні часових рядів у виді ортогонального розкладу.

Метою статті є порівняльний аналіз результатів використання методів EMD та SSA на прикладі дослідження процесу електроспоживання, заданого часовим рядом потужності електроспоживання на річному інтервалі спостереження.

Постановка задачі. Доступною опрацюванню є вибірка $P[j], j=\overline{1,J}$, обсягу $J=8760$ значень потужності електроспоживання організації. Вибірка отримана за результатами спостереження показів лічильника електроенергії протягом року з періодом дискретизації 1 година.

Необхідно виконати аналіз вибірки $P[j]$ (рис. 1) методами EMD та SSA, провести аналіз отриманих результатів з подальшою інтерпретацією обчислених компонент ряду, що в сукупності зменшить апріорну невизначеність під час створення математичної моделі процесу електроспоживання.

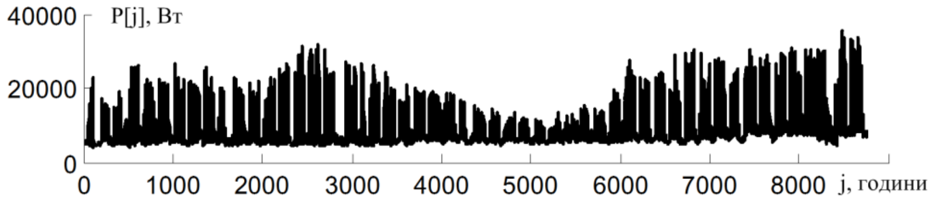

Рис. 1. Графік часового ряду $P[j]$ електроспоживання

Емпірична модова декомпозиція (EMD). Метод вперше був описаний та впроваджений Н. Е. Хуангом [2]. Його сутність полягає в адаптивному алгоритмі представлення нестаціонарних даних у вигляді суми власних модових функцій (intrinsic mode functions). Запропонована технологія була визнана ефективною, проте сьогодні строге математичне обґрунтування та обмеження щодо її застосування повністю не встановлені. Тому основним методом її дослідження є проведення комп'ютерних модельних експериментів [3].

В цьому методі власну модову функцію розглядаємо як циклічний процес, що має нульовий середній рівень та характеризується такою властивістю: кількість нуль-переходів функції дорівнює кількості екстремумів або відрізняється не більше ніж на одиницю.

Нижче розглянуто методику знаходження власних модових функцій для вибірки $P[j]$:

1. Процес обчислень «Власної модової функції».

1.1. Для досліджуваного часового ряду $P[j]$ знаходять екстремуми – максимуми і мінімуми та проводять їх інтерполяцію кубічними сплайнами. Для отриманих таким чином обвідних знаходять криву середнього $m[i]$. Різниця між отриманим середнім і вхідною послідовністю – це прообраз власної модової функції:

$$h[j] = P[j] - m[j].$$

Операцію 1.1 повторюють до $h[j]$ і отримують послідовність наступних прообразів власних модових функцій.

1.2. Операцію 1.1 виконують k разів до моменту, коли прообраз $h[j]$ відповідатиме визначенню власної модової функції, що дає змогу прийняти його за першу власну модову функцію, тобто $c_1[j] = h[j]$. В прийнятому позначенні $c_i[j]$ – i-та власна модова функція за межами процесу її обчислення, а $h[j]$ – позначення прообразу власної модової функції в процесі її обчислення. Критерії зупинки процесу відсіювання більш детально розглянуто в [4].

2. Процес розкладу ряду на суму власних модових функцій.

2.1. Отриману власну модову функцію $c_1[j]$ віднімають від початкових даних і отримують ряд залишків $r_1[j]$:

$$r_1[j] = P[j] - c_1[j].$$

Ряд $r_1[j]$ використовують як вхідні дані для повторення операцій 1.1, 1.2, 2.1. Такого роду обчислювальні операції продовжуються доти, коли графік залишків ряду можна описати такою математичною моделлю:

- постійна складова;
- монотонна зростаюча або спадаюча функція (монотонний тренд);
- функція з одним екстремумом.

В результаті застосування методу емпіричної декомпозиції вхідний процес розкладається на суму n власних модових функцій та кінцевий залишок:

$$P[j] = \sum_{i=1}^{n} c_i + r_n .$$

Властивість алгоритму полягає в тому, що визначенні власні модові функції з меншими індексами відповідають більш високочастотним складовим досліджуваного сигналу, а з більшими індексами – відповідають трендам сигналу.

Аналіз результатів обробки часового ряду. В результаті декомпозиції річного досліджуваного числового ряду, графік якого наведено на рис. 1, за описаним вище алгоритмом отримаємо 13 власних модових функцій, наведених на рис. 2 та рис. 3, на інтервалі за січень місяць річного часового ряду.

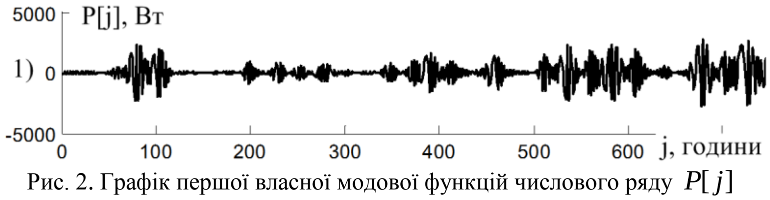

Рис. 2. Графік першої власної модової функцій числового ряду $P[j]$

За своїм характером першу та другу власну модову функцію можна віднести до реалізації нестаціонарного випадкового процесу.

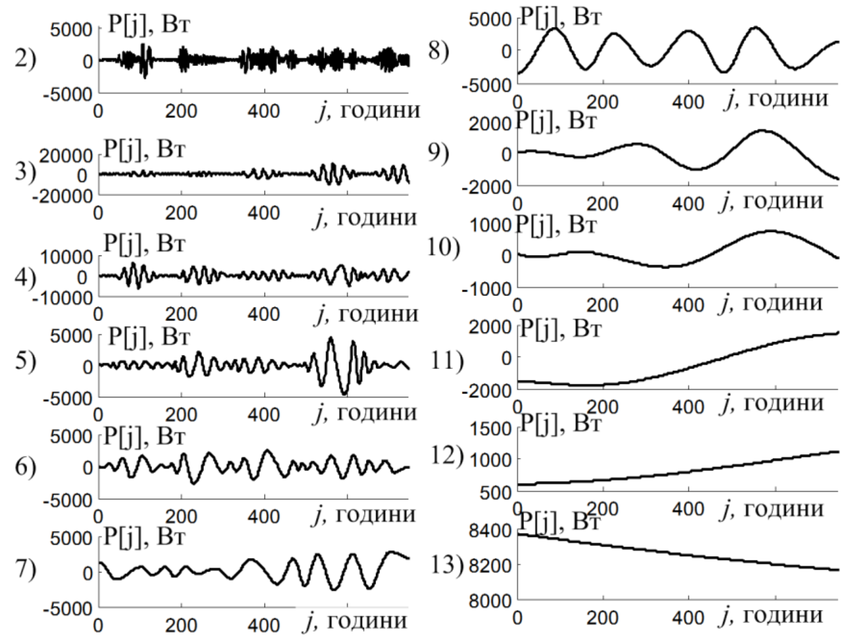

Рис. 3. Графіки власних модових функцій числового ряду $P[j]$

Якщо проаналізувати отримані функції, можна чітко виділити циклічні компоненти, що відповідають за добові, тижневі, місячні цикли, нестаціонарну шумову складову: 4,5,6 власні модові функції мають добову циклічність, а 8-ма – тижневу. Якщо ж скласти власні модові функції з 11-ї по 13-ту, отримаємо тренд часового ряду (рис. 4).

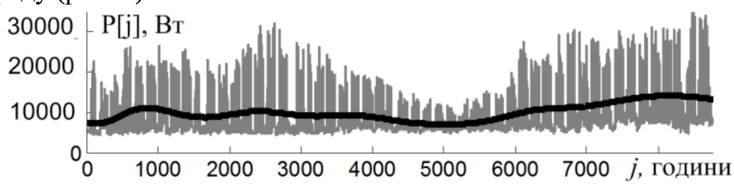

Рис. 4. Графік часового ряду $P[j]$ та його тренду, отриманого за методом ЕМД

Отримані результати дають можливість робити висновки про характер перебігу досліджуваного процесу на різних рівнях деталізації і зменшити апріорну невизначеність.

Метод «Гусениця» - SSA.

Значний внесок в питання дослідження та практичного використання цього методу зробив В.М. Бухштабер [5] та Н.Е. Голяндіна [6]. Алгоритм методу складається з двох етапів: декомпозиції та відновлення.

На етапі декомпозиції відбувається перетворення часового ряду на траєкторну матрицю та розкладання її на елементарні матриці. На етапі відновлення отримані елементарні матриці групують для розділення адитивних компонент часового ряду. Потім до згрупованих елементарних матриць застосовують операцію діагонального усереднення і отримують виділені адитивні компоненти. Розглянемо ці етапи більш детально.

На початку етапу декомпозиції необхідно перевести наш вихідний часовий ряд $P[j]$ у послідовність багатомірних векторів довжиною в L точок ($1 < L < J$). Число L необхідно обґрунтувати, базуючись на фізичній природі формування часового ряду. Тоді ця компонента буде виділена з більшою точністю. В результаті отримаємо $K = J - L$ векторів $X_i = (p_i, \ldots, p_{i+L})^{\mathrm{T}}, 1 \le i \le K$, які утворюють траєкторну матрицю $\mathbf{X} = [X_1 : \ldots : X_K]$, або:

$$\mathbf{X} = \begin{pmatrix} p_1 & p_2 & \cdots & p_K \\ p_2 & p_3 & \cdots & p_{K+1} \\ \vdots & \vdots & \ddots & \\ p_L & p_{L+1} & \cdots & p_J \end{pmatrix}$$

Далі до отриманої траєкторної матриці застосовують сингулярну декомпозицію: розкладання на елементарні матриці. В результаті сингулярної декомпозиції матриці $\mathbf{X}$, отримаємо власні числа матриці $\lambda_1, \ldots, \lambda_L$, лівий U_i та правий V_i сингулярні вектори матриці $\mathbf{X}$. Сингулярна декомпозиція матриці $\mathbf{X}$ може бути записана як

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i, \; d = \max\{i, \; \lambda_i > 0\}.$$

Етап відновлення починається з групування елементарних матриці за певними характеристиками. Множину індексів елементарних матриць $i = \{1, \ldots, d\}$ ділять на N непересічних підмножин $I_1, \ldots, I_N$. Нехай $I_1 = \{1,2,3\}$, тоді $\mathbf{X}_{I_1} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$. Суми відповідних груп елементарних матриць відповідають за адитивні компоненти вихідного числового ряду $P[j]$. Для перетворення групи елементарних матриць в числовий ряд – компоненту вихідного ряду $P[j]$ – застосовують операцію діагонального усереднення.

Нехай $\mathbf{X}_{I_1}$ – $L \times K$ матриця. Діагональне усереднення переводить матрицю $\mathbf{X}_{I_1}$ в ряд $R[j]$ довжиною J за формулою

$$R[j] = \begin{cases} \dfrac{1}{j} \sum\limits_{m=1}^{j} \mathbf{X}_{I_1}[m, j-m+1], & \text{для } 1 \le j < L, \\ \dfrac{1}{L} \sum\limits_{m=1}^{L} \mathbf{X}_{I_1}[m, j-m+1], & \text{для } L \le j < K, \\ \dfrac{1}{J-j} \sum\limits_{m=j-K+1}^{J-K} \mathbf{X}_{I_1}[m, j-m+1], & \text{для } K \le j < J. \end{cases}$$

Застосовуючи діагональне усереднення до всіх груп елементарних матриць $\mathbf{X}_{I_k}$, отримаємо вихідні ряди $R^{(k)}[j]$, тобто вихідний ряд $P[j]$ розкладається на суму N компонент:

$$P[j] = \sum_{k=1}^{N} R^{(k)}[j].$$

Вихідний числовий ряд $P[j]$ був проаналізований за описаним алгоритмом і в результаті були отримані компоненти, ключові з яких подані на інтервалі за січень на рис. 5 , 6.

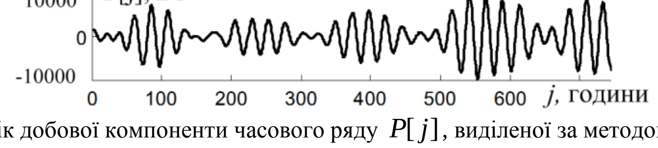

Рис. 5. Графік добової компоненти часового ряду $P[j]$, виділеної за методом «Гусениця»

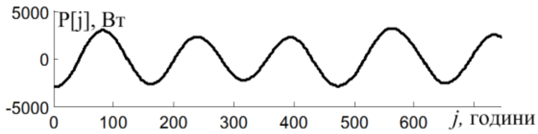

Рис. 6. Графік тижневої компоненти часового ряду $P[j]$, виділеної за методом «Гусениця»

Аналогічно до результатів, отриманих за методом емпіричної модової декомпозиції, в отриманих компонентах за методом «Гусениця» чітко прослідкується добова та тижнева циклічність. Так само можемо виділити тренд часового ряду, зображений на рис. 7.

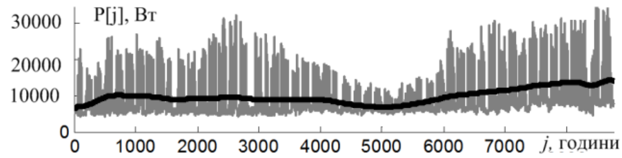

Рис. 7. Графік часового ряду $P[j]$ та його тренду, виділеного за методом «Гусениця»

Метод «Гусениця» вимагає певної підготовки оператора і потребує деяку апріорну інформацію про процес (наприклад, припущення про періодичність компонент та їх кількість). Проте за наявності таких даних отриманий результат буде точніше описувати локально-глобальні особливості обраних компонент.

Порівняльний аналіз результатів. Серед отриманих результатів обробки часового ряду електроспоживання організації на річному інтервалі спостереження можна чітко виявити добову, тижневу циклічність, а комбінації повільно змінних компонент в сумі дадуть тренд. Графіки компонент з добовою та тижневою циклічністю наведені на рис. 8.

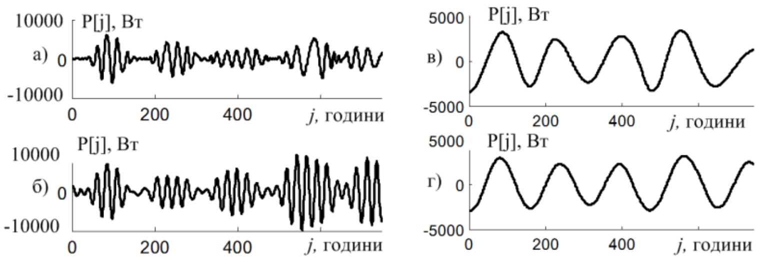

Рис. 8. Графіки циклічних компонент часового ряду $P[j]$: а) – сумарний добовий цикл, отриманий за методом EMD; б) – добовий цикл, отриманий за методом «Гусениця»; в) – тижневий цикл за методом EMD; г) – тижневий цикл за методом «Гусениця»;

Компоненти, отримані за різними методами, але такі що відповідають за однакову циклічність мають певні відмінності. Але їх аналіз дає змогу отримати інформацію про загальний характер досліджуваного процесу, а той факт, що періоди компонент, отриманих за різними методами, збігається, підтверджує наявність коливань із таким періодом в досліджуваному процесі. Слід зазначити, що для підвищення точності аналізу за методом «Гусениця» число L необхідно обирати виходячи із знань про період компоненти. В цьому і полягає головна складність використання методу «Гусениця». В той же час метод EMD не передбачає втручання користувача в процес дослідження часового ряду. Звідси випливає ідея послідовної методики первинного дослідження часових рядів за відсутності апріорної інформації про їх модель:

1. Провести аналіз часового ряду, використовуючи метод EMD, та отримати інформацію про періоди основних компонент.

2. Використовуючи отримані дані про періоди компонент, провести аналіз часового ряду за методом «Гусениця».

Слід також зазначити, що така послідовність використання вказаних методів обробки часових рядів дала можливість порівняти найбільш інтегральні характеристики – тренди досліджуваного часового ряду. Порівнюючи числові значення трендів часового ряду, графіки яких наведені на рис. 4 та рис. 7, отримано інтегральні характеристики, наведені в таблиці.

Інтегральні характеристики порівняння трендів часового ряду, отриманих за різними методами

Характеристика	Метод EMD	Метод «Гусениця»
Математичне сподівання (М.С.)	$\bar{x}_{EMD}$ =9717 Вт	$\bar{x}_{CAT}$ =9913 Вт
Середнє М.С. $\bar{x} = 0.5(\bar{x}_{EMD} + \bar{x}_{CAT})$	$\bar{x}$ =9815 Вт	
Середнє квадратичне відхилення С.К.В.	σ_{EMD} =1941 Вт	σ_{CAT} =2034 Вт
М.С. різниці трендів ($\overline{\Delta}$)	$\overline{\Delta}$ =195 Вт	
С.К.В. різниці трендів (σ_{Δ})	σ_{Δ} =324 Вт	
Відносна різниця трендів $\delta = \overline{\Delta}/\bar{x}$	δ =1.9%	

Різниця між трендами, отриманими різними методами, не суттєва порівняно із самими трендами. Але в задачах виділення трендів великих вибірок метод емпіричної модової декомпозиції більш вигідний через менший час обчислення.

Висновки. Розглянуті методи дослідження часових рядів є потужними засобами аналізу в умовах апріорної невизначеності характеру досліджуваного процесу. Вони можуть бути використані в багатьох галузях науки, зокрема як засоби дослідження сигналів в неруйнівному контролі та діагностиці.

Запропоновано методику дослідження часових рядів за умов повної відсутності апріорної інформації про їх модель, яка складається з двох етапів:

1. За допомогою методу EMD отримати приблизну інформацію о періодах основних компонент досліджуваного числового ряду.
2. Використовуючи отримані дані, провести ретельний аналіз часового ряду за допомогою методу «Гусениця».

1. *Graham P. Weedon.* Time-series analysis and cyclostratigraphy.// Cambridge: Cambridge University Press 2003, – 276p.
2. *Norden E. Huang.* Hilbert-Huang transform and its Applications// World Scientific Publishing Co. Pte.Ltd, 2005 – 324p.
3. *Dergunov O. V., Trots V. M., Kuts Y. V., Shcherbak L. M.* Empirical mode decomposition in signal analysis [“Aviation in the XXI-st centurx 2010”], (Kyiv, 21-23 September 2010) [etc.]. – K.: NAU, 2010. – P. 12.21 – 12.26.
4. *Patrick Flandrin.* On Empirical mode decomposition and its algorithms// IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (I), 2003
5. *Бухштабер В. М.* Многомерные развертки временных рядов. Теоретические основы и алгоритмы. // Обозрение прикл. промышл. матем., сер. Вероятн. и статист. – 1997. – Т. 4. Вып. 4. – С 629-645.
6. *Голяндина Н. Э.* Метод «Гусеница» - SSA: Анализ временных рядов// СПб.: С.-Петербургский государственный университет, 2004. –76 с.

Національний авіаційний університет, Київ