УДК 621.382.2

А.С. КАШТАЛЬЯН
Хмельницький національний університет

ВИЗНАЧЕННЯ ОСНОВНИХ ПЕРІОДИЧНИХ КОМПОНЕНТ ЧАСОВОГО РЯДУ

Стаття присвячена дослідженню структури часового ряду, зокрема виділенню періодичних складових ряду за методом сингулярного спектрального аналізу. Для групування основних періодичних складових використано дискретне перетворення Фур'є та кластеризацію за методом k середніх. Розв'язок даної задачі необхідний для побудови моделі часового ряду та виявлення прихованих залежностей. В статті приведено аналіз структури фінансового часового ряду денних цін закриття та ряду середньодобових температур.

Ключові слова: сингулярний спектральний аналіз, період, перетворення Фур'є, кластеризація.

A.S. KASHTALIAN
Khmelnitsky National University, Khmelnitsky, Ukraine

THE DEFINITION OF TIME SERIES MAIN PERIODICAL COMPONENTS

Abstract. The investigation of time series structure is important issue, because firstly, it gives capability to analyze the processes which conducting in caused this time series phenomenon. In addition, the knowledge of time series structure allows performing adequate prediction of further values. The article is devoted to research of time series structure and extraction of time series periodical components by singular spectrum analysis method. Discrete Fourier transform and k-means clustering are used for grouping main periodical components. The solution of this task is necessary for time series model creating and identification of hidden dependences. Structures analysis of the financial time series of day closing prices and time series of daily average temperatures is considered in article.

Key words: singular spectral analysis, period, Fourier transform, clustering.

Постановка задачі, аналіз досліджень та публікацій. В даний час для вивчення властивостей складних систем, в тому числі і при експериментальних дослідженнях, широко використовується підхід, що ґрунтується на аналізі сигналів системи. Це особливо актуально в тих випадках, коли математично описати досліджуваний процес практично неможливо, але наявною є деяка величина, що спостерігається, наприклад, метрологічні або сейсмологічні дані, дані дослідження фізичного стану людини, дані фінансових ринків. Наявність тільки часового ряду значно обмежує знання про досліджувану систему і накладає значні обмеження [1]. Головною метою дослідження часового ряду є знаходження закономірностей, прихованих в емпіричних даних, що відображають внутрішню структуру ряду. Виявлена внутрішня структура часового ряду дозволяє також в подальшому виконувати прогнозування поведінки ряду.

Методи аналізу часових рядів поділяють на два загальні класи: методи частотної області і методи часової області, серед яких можна виділити такі методи: спектральний аналіз (дозволяє знаходити періодичні складові часового ряду); кореляційний аналіз (дозволяє знаходити існуючі періодичні залежності та відповідні їм затримки як всередині одного ряду, так і між декількома рядами); моделі авторегресії та рухомого середнього (моделі орієнтовані на опис процесів, що проявляють однорідні коливання, збуджені випадковими впливами); сезонна модель Бокса-Дженкінса (застосовується, якщо часовий ряд має явно виражений лінійний тренд та сезонні складові); модель експоненційно–зваженого рухомого середнього (одна з найпростіших моделей прогнозування часових рядів) тощо [2, 3]. Основна проблема полягає в тому, що більшість існуючих методів виявлення періодичних складових ряду передбачає наявність повторюваності в часовому ряду, що часто не відповідає дійсності, зокрема для фінансових рядів, і як наслідок, призводить до неадекватних результатів.

В даній роботі застосований апарат сингулярного спектрального аналізу, розроблений та обґрунтований наприкінці минулого століття [4, 5]. В основу методу покладений аналіз головних компонент, метод дозволяє досліджувати стаціонарні та нестаціонарні часові ряди, які можуть мати виражену періодичність або періодичність яких є невиявленою. Особливістю методу є те, що він не накладає на ряд попередніх обмежень, таких як наявність явно виражених періодичних коливань, наявність лінійного тренду, наявність повторюваності.

Матеріал і результати досліджень. Розглянемо задачу дослідження структури часового ряду $\left(f_0, f_1, \ldots, f_{n-1}\right)$ довжини *N*. Під структурою розуміють деякі характеристики або властивості ряду, які зберігаються протягом часу. Звичайно, постає задача знаходження тренду або виділення періодичних складових. Як правило, для знаходження структури певного явища необхідна повторюваність цього явища. Якщо ж розглядається часовий ряд, який існує в єдиному екземплярі, то повторюваність відсутня. В якості способу виходу з цієї ситуації в стандартних методах аналізу часових рядів часто припускається справедливість параметричної моделі ряду або його стаціонарність. Однак ці обмеження можуть виявитися надто жорсткими. Тому необхідно отримати повторюваність ряду, не накладаючи на ряд попередніх жорстких обмежень. Розглянемо множину відрізків часового ряду заданої достатньо великої довжини *L*, (довжини вікна). Розглядатимемо ці відрізки послідовно, з першої по *L*-ту точку, з другої по (*L*+1)-у і так далі. Дані відрізки (вектори *L*-вкладення) будуть володіти властивостями ряду. Якщо ряд містить тренд, то вектори вкладення теж його міститимуть, якщо в ряді присутня періодична складова, то вона буде присутня також в векторах вкладення. Таким чином можна дослідити всю сукупність векторів для виявлення їх загальної структури. Складемо з векторів вкладення так звану траєкторну матрицю розмірності $L \times K$, де $K=N-L+1$ – число векторів вкладення. Вона матиме такий вигляд

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & \cdots & f_{K-1} \\ f_1 & f_2 & \cdots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \cdots & f_{N-1} \end{bmatrix} = \left[X_1, \ldots, X_K\right], \qquad X_j = \begin{bmatrix} f_{j-1} \\ \vdots \\ f_{j+L-2} \end{bmatrix}. \tag{1}$$

В цій матриці є повторюваність і можна спробувати побачити структуру векторів вкладення. Для цього розкладемо всю траєкторну матриці на елементарні частини (в суму елементарних матриць), в певному сенсі незалежні (ортогональні) і впорядковані за їх вкладом в розклад. Якщо розкласти їх вдало, то можна буде згрупувати ці елементарні матриці таким чином, щоб одна група відповідала трендовій складовій, друга – періодичній і так далі. Потім просумуємо матриці всередині кожної групи і повернемося від розкладу матриць до розкладу ряду на тренд, періодичні складові та залишок.

Способом розкладу траєкторної матриці, який виявляється добре узгодженим з розглянутою вище задачею, є сингулярний розклад. В зарубіжній літературі метод найбільше відомий як SSA (Singular Spectrum Analysis), причому під сингулярним спектром мається на увазі набір власних чисел сингулярного розкладу траєкторної матриці, що розумується як сингулярний спектр відповідного матриці оператора. В Росії метод отримав назву «Гусениця» через послідовну процедуру нарізання векторів вкладення з вихідного ряду.

Початково ідентифікація складових ряду на основі сингулярного розкладу його траєкторної матриці проводилась інтерактивно, в основному візуальним способом, користуючись графічним представленням результатів та ґрунтуючись на теоретичних відомостях. З одного боку, інтерактивність є позитивною стороною методу, оскільки дає можливість свідомого його застосування з боку користувача та приводить до більш глибокого та якісного аналізу. З іншого боку, зокрема за необхідності аналізу великої кількості однотипних даних) виникає необхідність в автоматизації процедури ідентифікації складових ряду.

Базовий алгоритм SSA. Алгоритм сингулярного спектрального аналізу можна розбити на чотири кроки: вкладення, сингулярний розклад, групування та діагональне усереднювання. Перші два в сукупності називаються розкладом, останні – відновленням. Основним параметром алгоритму є так звана довжина вікна L, $1<L<N$. результатом алгоритму є розбиття часового ряду на адитивні складові. Нехай $N>2$. Розглядаємо часовий ряд $F=\left(f_0, f_1, \ldots, f_{n-1}\right)$ довжини N. Вважаємо, що ряд F – ненульовий, тобто існує як мінімум одне i, таке що $f_i \neq 0$. Числа 0, …, N-1 можуть визначатися не тільки як дискретні моменти часу, але також як деякі мітки, що мають лінійно-впорядковану структуру.

Вкладення. Процедура вкладення переводить вихідний часовий ряд в послідовність декількох багатомірних векторів. Якщо довжина вікна L, $1<L<N$, то в результаті процедури вкладення буде утворено К векторів вкладення, що мають розмірність L. Результатом виконання є траєкторна матриця (1). Очевидно, що $x_{ij} = f_{i+j-2}$ і матриця $\mathbf{X}$ має однакові елементи на «діагоналях» $i+j=const$. Таким чином, траєкторна матриця є ганкелевою. Існує відповідність між ганкелевими матрицями розмірності $L \times K$ та рядами довжини $N=L+K-1$.

Сингулярний розклад. Результатом цього кроку є сингулярний розклад траєкторної матриці ряду. Нехай $\mathbf{S}=\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Позначимо $\lambda_1, \ldots, \lambda_L$ - власні числа матриці $\mathbf{S}$, взяті в не зростаючому порядку $\left(\lambda_1 \geq \ldots \geq \lambda_L \geq 0\right)$ та $U_1, \ldots, U_L$ - ортонормовану систему власних векторів матриці $\mathbf{S}$, що відповідають власним числам. Нехай $d=\max\{i : \lambda_i > 0\}$. Якщо позначити $V_i = \mathbf{X}^{\mathrm{T}} U_i \big/ \sqrt{\lambda_i}$, $i=1,\ldots,d$, то сингулярний розклад матриці $\mathbf{X}$ може бути записаний як

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d, \tag{2}$$

де $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$. Кожна з матриць $\mathbf{X}_i$ має ранг 1, тому їх можна назвати елементарними матрицями. Набір $\left(\sqrt{\lambda_i}, U_i, V_i\right)$ називається i-ою власною трійкою сингулярного розкладу (2).

Групування. На основі розкладу (2) процедура групування поділяє всю множину індексів $\{1,\ldots,d\}$ на m множин, що не перетинаються $I_1, \ldots, I_m$. Нехай $I=\{i_1, \ldots, i_p\}$. Тоді результуюча матриця $\mathbf{X}_I$, що відповідає групі I, визначається як

$$\mathbf{X} = \mathbf{X}i_1 + \ldots + \mathbf{X}_{ip}.$$

Такі матриці визначаються для $I = I_1, \ldots, I_m$, тим самим розклад (2) може бути записаний в згрупованому вигляді

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_m}, \tag{3}$$

Процедура вибору множини $I_1, \ldots, I_m$ називається групуванням власних трійок.

Діагональне усереднення. На останньому кроці базового алгоритму кожна матриця згрупованого розкладу (3) переводиться в новий ряд довжини N. Нехай $\mathbf{Y}$ – деяка $L \times K$ матриця з елементами y_{ij}, де

$1 \le i \le L$, $1 \le i \le K$. Вважаємо $L^* = \min(L,K)$, $K^* = \max(L,K)$ та $N = L + K - 1$. Нехай $y^*_{ij} = y_{ij}$, якщо $L<K$ та $y^*_{ij} = y_{ji}$, в іншому випадку. Діагональне усереднення переводить матрицю $\mathbf{Y}$ в ряд $g_0,\ldots,g_{N-1}$ за формулою

$$g_k = \begin{cases} \dfrac{1}{k+1}\sum\limits_{m=1}^{k+1} y^*_{m,k-m+2} & \text{для } 0 \le k < L^* - 1, \\ \dfrac{1}{L^*}\sum\limits_{m=1}^{L^*} y^*_{m,k-m+2} & \text{для } L^* - 1 \le k < K^*, \\ \dfrac{1}{N-k}\sum\limits_{m=k-K^*+2}^{N-K^*+1} y^*_{m,k-m+2} & \text{для } K^* \le k < N. \end{cases} \quad (4)$$

Вираз (4) відповідає усередненню елементів матриці вздовж «діагоналей» $i + j = k + 2$: для $k = 0$ отримуємо $g_0 = y_{11}$, для $k = 1$ отримуємо $g_0 = (y_{12} + y_{21})/2$ і т. д.

Застосовуючи діагональне усереднення (4) до результуючих матриць $\mathbf{X}_{I_k}$, отримуємо ряди $\widetilde{F}^{(k)} = \left(\widetilde{f}_0^{(k)},\ldots,\widetilde{f}_{N-1}^{(k)}\right)$, відповідно вихідний ряд розкладається у суму m рядів.

$$f_n = \sum_{k=1}^{m} \widetilde{f}_n^{(k)}. \quad (5)$$

Виділення основних періодів. Для виділення основних періодичних компонент ряду необхідно згрупувати ряди відповідно до їх частот. Для цього виконаємо для кожного отриманого ряду перетворення Фур'є

$$S(k) = \sum_{n=0}^{N-1} s(n)\cdot \exp\left(-j \cdot \frac{2\pi}{N} \cdot nk\right),\ k = 0,\ldots,N-1,$$

N – кількість значень сигналу, виміряних за період, а також компонент розкладу; $s(n)$, $n = 0,\ldots,N-1$, – виміряні значення сигналу (в дискретних часових точках з номерами $n = 0,\ldots,N-1$, які є вхідними даними для прямого перетворення Фур'є; $S(k)$, $k = 0,\ldots,N-1$, - комплексні амплітуди синусоїдальних сигналів, що складають вихідний сигнал, є вихідними даними для прямого перетворення Фур'є. Вираз для дискретного перетворення Фур'є ставить у відповідність N відлікам сигналу $s(n)$ в загальному випадку N відліків спектру $S(k)$.

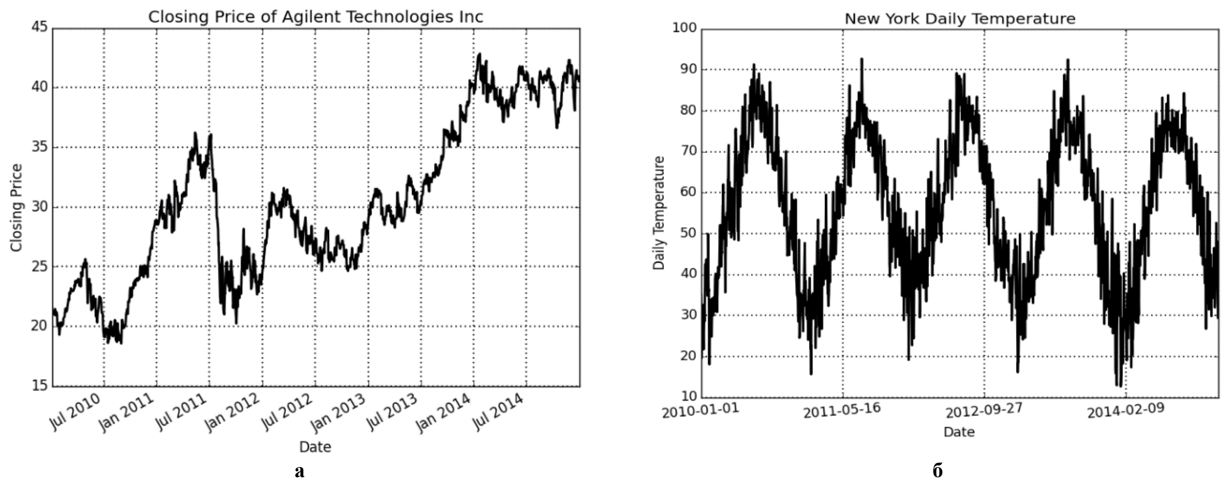


Рис. 1. Часові ряди: а) денних цін закриття компанії Agilent Technologies; б) середньодобових температур міста Нью-Йорк

За дискретним перетворенням Фур'є виділяємо визначаємо найвагомішу частоту найвагомішої гармоніки кожного ряду, в результаті отримуємо ряд частот $F = (f_0,\ldots,f_{N-1})$, що відповідає рядам, отриманим при спектральному сингулярному розкладі (5). Відповідно до цих частот отримуємо ряд основних періодів $T = (T_0,\ldots,T_{N-1})$, які також відповідають ряду (5).

Для групування періодів з близькими значеннями використаємо кластеризацію за метод k середніх.

Отримана множина $T = (T_0, \ldots, T_{N-1})$ періодів перетворюється в k $(\leq N-1)$ множину $S = (S_0, \ldots, S_k)$ таким чином, щоб мінімізувати сумарне квадратичне відхилення точок кластерів від центрів цих кластерів

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i\right)^2 ,$$

де k – число кластерів, S_i – отримані кластери, i=1,…, k, μ_i – центри мас векторів $x_j \in S_i$.

Експериментальний аналіз. Було проведено аналіз різних типів часових рядів, зокрема нестаціонарного часового ряду з невиявленою періодичністю та часового ряду з вираженою періодичністю. В роботі представлено ряд денних цін закриття для компанії Agilent Technologies з 1 січня 2010 року до 1 січня 2015 року (рис. 1а) та ряд середньодобових температур міста Нью-Йорк з 1 січня 2010 року до 1 січня 2015 року (рис. 1б).

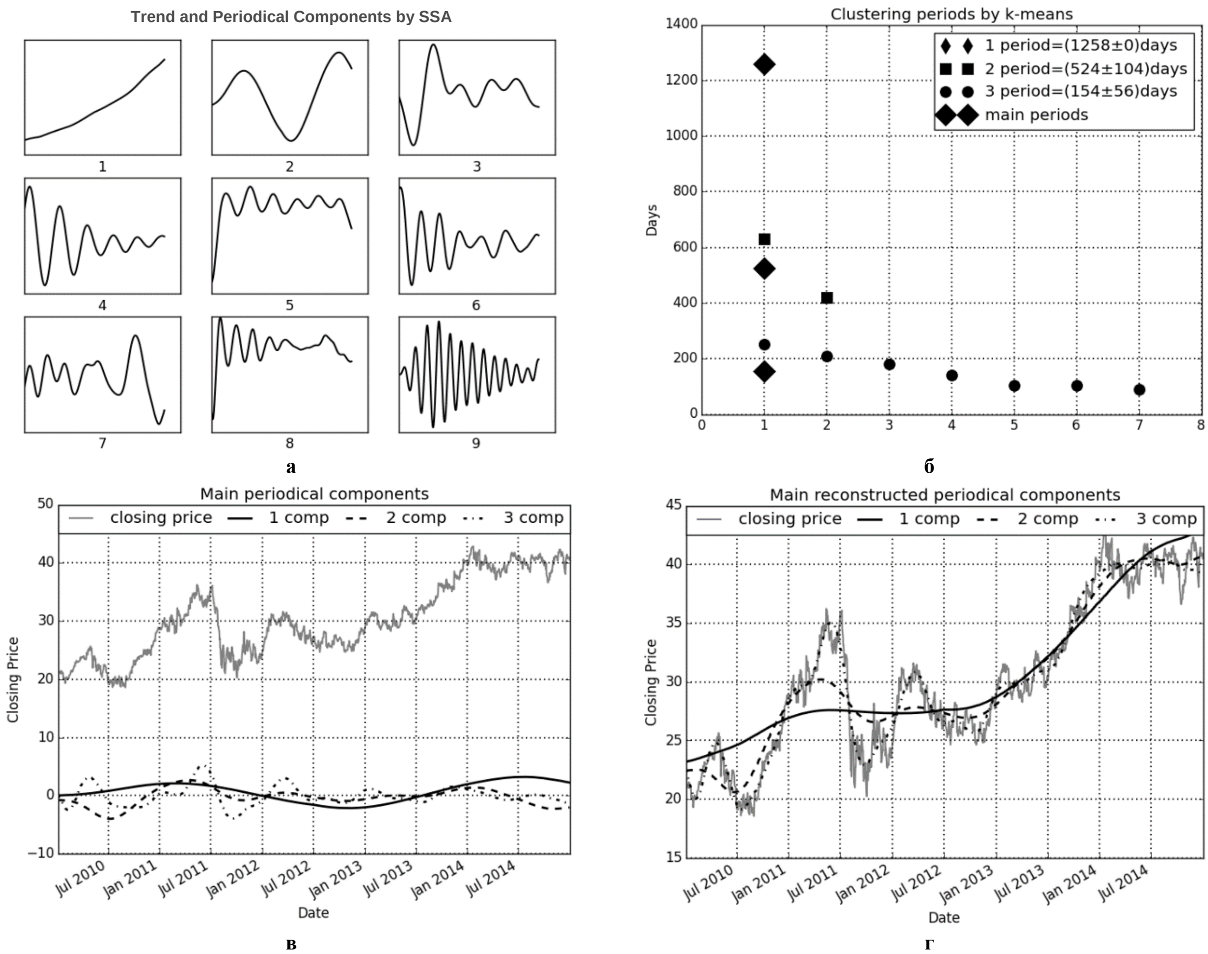
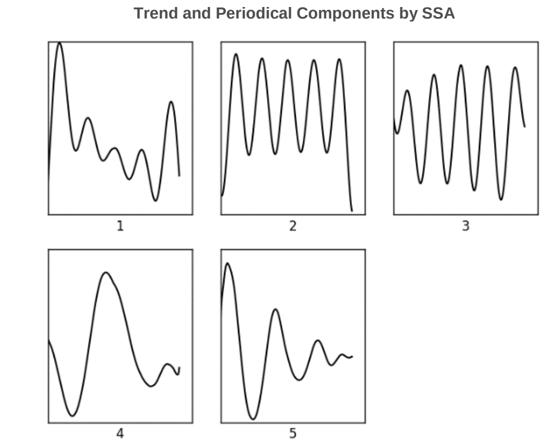


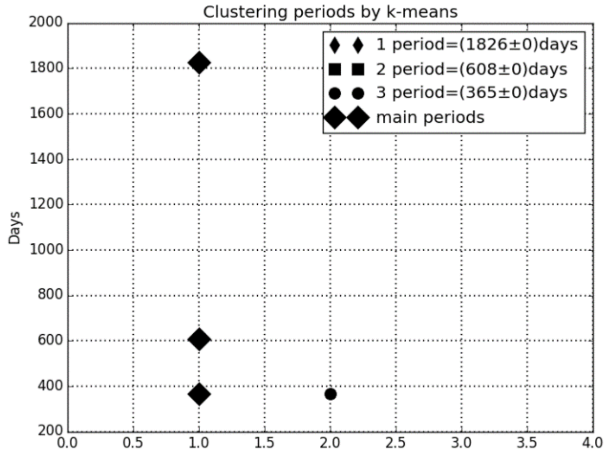
Рис. 2. Результати аналізу часового ряду денних цін закриття компанії Agilent Technologies:
а) трендова складова та виділені періодичні компоненти за методом SSA;
б) групування періодичних компонент за методом k середніх;
в) основні періодичні компоненти часового ряду
г) основні періодичні компоненти часового ряду з врахуванням трендової складової

Наведений часовий ряд цін закриття не має вираженої періодичності, як і більшість фінансових рядів, таких як ціна на акції компаній, курси валют і т. п. Однак за методом сингулярного спектрального аналізу можна виділити приховані гармоніки, присутні в таких часових рядах. В даному часовому ряді було виділено перших 10 складових ряду, перші 9 з яких відображені на рисунку 2 а. Перша складова є трендовою і відображає загальну тенденцію ряду, всі наступні є періодичними з періодами [1258, 629, 251, 209, 179, 419, 139, 104, 104, 89] днів. Деякі періодичні складові мають близькі за значенням періоди. Прослідкувати поведінку всіх десятьох складових важко і це може ускладнити розуміння структури ряду та його прогнозування, тому вони згруповані у три основних періодичних складових (рис. 2 б): [1258±0, 524±104, 154±56] днів. Згруповані основні періодичні складові зберігають періодичність (рис. 2 в) і накладені на трендову складову (рис. 2 г) більш повно відображають поведінку ряду, ніж тільки трендова складова, тому що враховують не тільки загальне зростання за п'ять років, а також проміжні падіння цін.

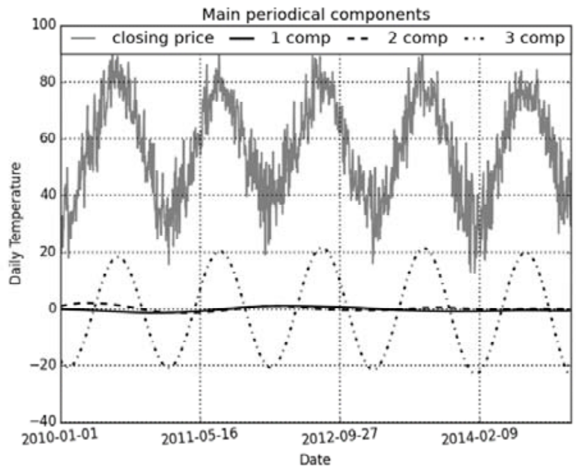
Температурний часовий ряд є виражено періодичним, що добре видно на його графіку. При його аналізі за мету було поставлено виділення річної періодичності, тому початково було виділено 5 періодичних складових: [1826, 608, 365, 365, 365] днів (рис. 3 а), які в подальшому згруповані в 3 основних періодичних складових [1826±0, 608±0, 365±0] днів (рис. 3 б, рис. 3 в). Із врахуванням постійної складової температурного ряду періодичні складові досить повно відображають загальну поведінку ряду (рис. 3 г). Виділена третя періодична складова чітко вловила річні коливання температур, інші дві відображають незначні зміни, які відбуваються з температурою з року в рік.



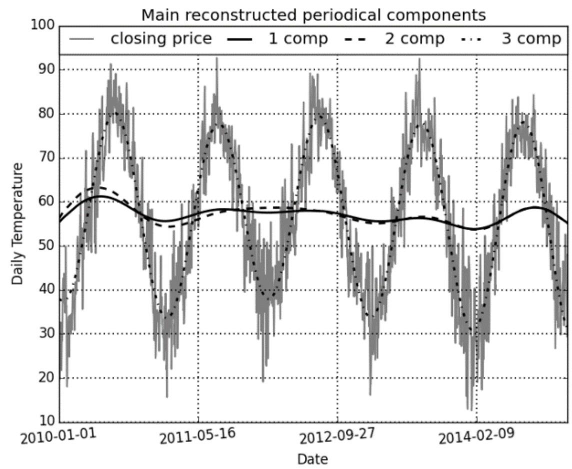
а



б



в



г

Рис. 3. Результати аналізу часового ряду середньодобових температур міста Нью-Йорк:
а) виділені періодичні компоненти за методом SSA;
б) групування періодичних компонент за методом k середніх;
в) основні періодичні компоненти часового ряду
г) основні періодичні компоненти часового ряду з врахуванням постійної складової

Висновки. Запропонована в статті методика аналізу структури часового ряду дозволяє виділити гармоніки та отримати основні періодичні компоненти часового ряду, що характеризують явища та процеси, які породжують часовий ряд. Попередні знання про структуру часового ряду при цьому не вимагаються, метод дозволяє проаналізувати періодичні ряди різних типів, з невиявленою або вираженою періодичністю. Запропонований метод кластеризації k середніми виправдовує себе і об'єднує близькі або рівні періодичні складові, що дозволило в температурному ряді чітко виділити річний період коливань, а в ціновому ряді об'єднати близькі періоди і узагальнити структуру. Варто відмітити, що особливої уваги потребує вибір такого параметра, як довжина вікна для аналізу за методом сингулярного спектрального аналізу. Довжина вікна залежить від того, якого рівня періодичні складові необхідно виділити; більш чітке визначення цього параметра потребує подальшого дослідження.

Література

1. Лоскутов А.Ю. Основы теории сложных систем./ А.Ю. Лоскутов, А.С. Михайлов. – М. – Ижевск: Институт компьютерных исследований, 2007. — 620 с.

2. Enders W. Applied Econometric Time Series, 4th edition/ Enders W. – John Wiley & Sons Inc, USA, 2014. – 460p.

3. Ruey S. Tsay Analysis of Financial Time Series/ Ruey S. Tsay. – John Wiley & Sons Inc., Hoboken, New Jersey, USA, 2005. – 640p.

4. Данилов Д.Л. Главные компоненты временных рядов: метод «Гусеница»./ Д.Л. Данилов, А.А. Жиглявский. – СПб.: Пресском, 1997. – 308с.