УДК 6:004.8

ПОБУДОВА АВТОРЕГРЕСІЙНИХ МОДЕЛЕЙ НА ОСНОВІ КОМБІНАТОРНО-ГЕНЕТИЧНОГО АЛГОРИТМУ

О.Г.Мороз

Міжнародний науково-навчальний центр інформаційних технологій та систем (МННЦ ІТС)НАН *та* МОН *України,*

olhahryhmoroz@gmail.com

Розроблено теоретичні основи перебірного гібридного алгоритму КОМБІ-ГА для моделювання складних процесів у класі авторегресійних моделей. Ефективність розробленого алгоритму продемонстровано на тестових задачах моделювання.

Ключові слова: індуктивне моделювання, МГУА, перебірні алгоритми, генетичні алгоритми, алгоритм КОМБІ-ГА, модель авторегресії.

The theoretical foundations of a COMBI-GA hybrid algorithm for complex processmodeling in the autoregressive models class have been developed. The effectiveness of the developed algorithm is demonstrated by a test modeling tasks.

Keywords: inductive modeling, GMDH, sorting-out algorithms, genetic algorithm, COMBI-GA algorithm, autoregressive model.

Разработаны теоретические основы переборного гибридного алгоритма КОМБИ-ГА для моделирования сложных объектов в классе авторегрессионных моделей. Эффективность разработанного алгоритма продемонстрирована на тестовых задачах моделирования.

Ключевые слова: индуктивное моделирование, МГУА, переборные алгоритмы, генетические алгоритмы, алгоритм КОМБИ-ГА, модель авторегрессии.

Вступ

Загальну модель авторегресії – ковзного середнього (autoregression-moving-average models, ARMA), частковим випадком якої є модель авторегресії (Autoregression, AR [1]), було описано в 1951 році в дисертації Пітера Уіттла «Перевірка гіпотез в аналізі часових рядів» і популяризовано в книзі Джорджа Боксата Ґвилима Дженкінса в 1970 році.

Будується ARMA за стаціонарними часовими рядами. Це система рівнянь, в якій кожна змінна (компонента багатовимірного часового ряду) в даний момент часу є лінійною комбінацією всіх змінних у попередні моменти. Порядок такої моделі визначається порядком запізнюваних значень (лагів).

Авторегресійна модель використовується також для побудови залежності окремо взятої змінної від самої себе, вірніше від того, якою вона була в минулі періоди (день, місяць, рік тощо). Можна припустити практично для будь-якого показника, що його поточний рівень в якійсь мірі залежить від того який він був раніше, тобто значення досліджуваного процесу виражається через скінченну лінійну сукупність попередніх значень процесу і деякої похибки. Саме пошук цієї залежності часто дозволяє будувати досить точні моделі, за якими можна зробити прогноз. Така модель використовується в багатьох економічних та фінансових задачах, де необхідно прогнозувати різні показники, наприклад, прогнозування значень ВВП, обсягу продажів товарів на підприємстві, вартості цінних паперів тощо, а також для опису стаціонарних часових рядів різної природи.

Одними з найбільш ефективних засобів для побудови моделі за експериментальними даними є алгоритми на основі методу групового урахування аргументів (МГУА) [2], які поділяються на дві групи: ітераційні та перебірні алгоритми. Серед перебірних алгоритмів індуктивного моделювання ефективним та перспективним є гібридний комбінаторно-генетичний алгоритм КОМБІ-ГА [3], який одночасно поєднує основні переваги та усуває недоліки комбінаторного алгоритму МГУА (КОМБІ) [4] та генетичного алгоритму(ГА) [5]. У роботах [3,6,7] цей алгоритм успішно застосовано для побудови лінійних та нелінійних моделей тестових і реальних статичних об'єктів у лінійному за параметрами класі.

В [8] прогнозування складних процесів у класі моделей векторної авторегресії на основі рекурентно-паралельного алгоритму.

В цій роботі алгоритм КОМБІ-ГА вперше застосовується для моделювання динамічних процесів за допомогою авторегресійної моделі на тестових даних.

1 Модифікація алгоритму КОМБІ-ГА для побудови авторегресійної моделі

Початковим етапом роботи алгоритму КОМБІ-ГА для розв'язання задачі індуктивного моделювання [2] є її зведення до задачі, лінійної за параметрами, введенням узагальнених «лінійних» аргументів. Спосіб формування цих аргументів залежить від класу функцій, в якому моделюється процес чи об'єкт. Так, наприклад, при моделюванні у лінійному класі алгоритм не потребує додаткових перетворень (початкові змінні є вхідними, тобто тими, з якими працює алгоритм) [3], при моделюванні нелінійних об'єктів необхідно здійснити відповідне перетворення початкових змінних [6].

Розглянемо процедуру лінеаризації для побудови авторегресійної моделі. В загальному випадку авторегресійна або різницева модель порядку L є такою:

$$z_k = f(z_{k-1}, z_{k-2},\ldots,z_{k-L}, \theta) = \theta_1 z_{k-1}+\ldots\theta_L z_{k-L}=\sum_{j=1}^{L}\theta_j\, z_{k-j}\;, k=L+1,\, L+2,\, \ldots,N,$$

де θ_j – коефіцієнти авторегресії, тобто оцінювані параметри, які описують ступінь залежності формальної «вихідної» змінної z від впливаючих факторів,в даному разі від її значень в минулі періоди регресії;

$z_1,\ldots,z_L$ – задані початкові значення (початкові умови цієї різницевої моделі);

z_{k-j}, $j=1,\ldots,L$ – попередні значення часового ряду як фактори впливу на поточні значення z_k;

z_k – формальна «залежна» змінна в моменти часу $k = 1,\ldots,N$.

Нехай наявні початкові дані – це вектор-стовпець $z=[z_1,z_2,\ldots,z_N]^T$, що містить N значень часового ряду. Припускаємо, що цей часовий ряд описується авторегресійною моделлю невідомого порядку L з невідомим складом попередніх (запізнюваних) значень цього часового ряду та з невідомими параметрами. Задача полягає у знаходженні оптимальної структури моделі та оцінки її параметрів. Оцінювання параметрів авторегресії в цьому випадку здійснюється аналогічно оцінюванню параметрів множинної лінійної регресії.

Ввівши узагальнені вектори аргументів

$$x_j = \left[z_{L-j+1} \quad z_{L-j+2} \quad \cdots \quad z_{N-j}\right]^T,\; j = 1,\ldots,L,$$

одержимо матрицю $X[n{\times}m]$ еквівалентних вхідних аргументів та відповідний вихідний вектор $y[n{\times}1]$, де $n = N - L$, $m = L$:

$$X = \left[x_1 \quad x_2 \quad \cdots \quad x_m\right] = \begin{bmatrix} z_L & z_{L-1} & \cdots & z_1 \\ z_{L+1} & z_L & \cdots & z_2 \\ \cdots & \cdots & \cdots & \cdots \\ z_{N-1} & z_{N-2} & \cdots & z_{N-L} \end{bmatrix},\; y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} z_{L+1} \\ z_{L+2} \\ \cdots \\ z_N \end{bmatrix}.$$

Взагалі кажучи, до таким чином перетворених початкових даних можна формально застосовувати будь-який алгоритм МГУА. Зокрема, відповідна модифікація алгоритму КОМБІ-ГА дозволяє розв'язувати за його допомогою задачі моделювання динаміки процесів, заданих у формі часових рядів.

Якщо k – період прогнозування, індекс точки прогнозу k=N+1,…,N+k та задано початкові умови k=N-L+1, N, .., N, то решта N+1 .. N+k обчислюються за формулою:

$$\hat{y}_{N+p} = \theta_1 y_{L+p-1} + \ldots + \theta_L y_p,\; p = 1, 2, \ldots, k$$

Оригінальність побудови моделей за алгоритмом КОМБІ-ГА у вказаному класі полягає в тому, що знаходиться не тільки оптимальний порядок авторегресії, але й оптимальний склад запізнюваних (лагових) змінних цієї моделі. Тим самим виявляються приховані внутрішні закономірності інерційних ефектів післядії в цьому процесі, що підвищує прогнозні можливості такої моделі.

2 Тестування ефективності модифікованого алгоритму КОМБІ-ГА

Опис початкових даних. Нехай вектор-стовпець початкових даних має розмірність $N = 60$, тобто $z[60{\times}1]$, а порядок авторегресії L = 10 випадково згенерованих значень,відповідно перетворена матриця X матиме розмірність [50×10], а вектор $y[50{\times}1]$. Щоб згенерувати тестові дані для перевірки коректності роботи модифікованого алгоритму в задачі побудови моделі авторегресії, слід задати перші 10 значень вектора z (вони ж будуть у першому рядку матриці X), які визначають порядок авторегресії L – в даному разі ми використали випадково згенеровані числа в інтервалі [–5,5]. Після цього, використовуючи задану тестову «істинну» модель авторегресії, розраховуються 50 точних значень виходу тестової моделі.

Авторегресійні моделі будувалися алгоритмом КОМБІ-ГА з еволюційним ускладненням структур моделей [3] з модифікацією для побудови авторегресійних моделей для збіжного і розбіжного часового ряду з шумом та без нього.

Побудова моделі для збіжного ряду. Шукана тестова модель за відсутності шуму має 5 «істинних аргументів», тобто окремих запізнюваних значень з усіх 10 можливих:

$$y = -0.2x_1 - 0.1x_4 + 0.5x_7 + 0.4x_8 + 0.3x_{10}$$

Матрицю X перетворених початкових даних y векторі z подано в Табл. 1.

Табл. 1. Матриця X перетворених початкових даних y векторі z

Номер точки	x_1	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	2.4	-1.3	-0.6	-2.2	-0.6	-4.9	-0.5	-4.1	0.7	1.9
2	-1.3	-0.6	-2.2	-0.6	-4.9	-0.5	-4.1	0.7	1.9	-1.6
3	-0.6	-2.2	-0.6	-4.9	-0.5	-4.1	0.7	1.9	-1.6	-1.9
…	…	…	…	…	…	…	…	…	…	…
49	0.2	0.2	-0.1	-0.1	0.4	0.1	-0.1	-0.1	-0.01	0.01
50	0.2	-0.1	-0.1	0.1	0.1	-0.1	-0.1	-0.01	0.01	-0.1

На Рис. 1. представлено порівняння значень точного часового ряду, модельного із зашумленим вектором y та модельним із зашумленим вектором z. З нього видно, що повне зашумлення має більший вплив на процес.

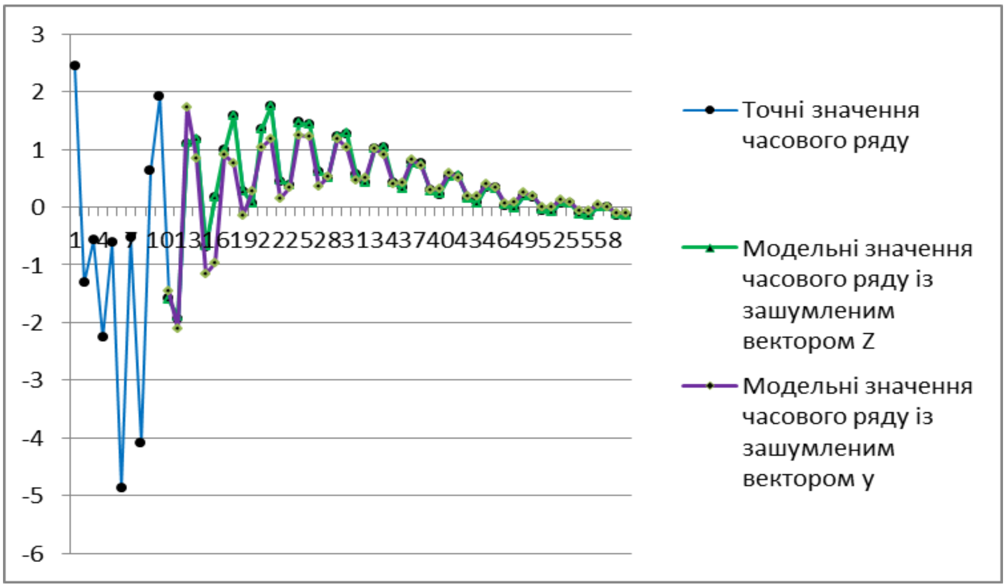

Рис. 1 Значення часового ряду: точні, модельні з зашумленим вектором y, модельні з зашумленим вектором z.

Враховуючи специфіку задачі при побудові моделі в авторегресійному класі, що полягає у наявності залежності між значеннями початкової змінної z_i, i=L+1,...,N та відповідно між елементами матриці X_i вектора y традиційний спосіб зашумлення даних набуває некоректного характеру. Надалі для більш коректних експериментів було здійснено зашумлення усього вектора початкових даних z, а не лише формальної «вихідної» змінної y, одержаної після перетворення.

Комбінаторно-генетичним алгоритмом було отримано модель:

$$y = -0.193x_1 - 0.117x_4 + 0.524x_7 + 0.413x_8 + 0.327x_{10}$$

Результат, отриманий алгоритмом КОМБІ-ГА при наявності шуму 30% у перетвореному векторі вихідної змінної *y* спрощує складність моделі з втратою точності у коефіцієнтах при значущих аргументах $\boldsymbol{x_1}$, $\boldsymbol{x_8}$, $\boldsymbol{x_{10}}$:

$$y = -0.144\,x_1 + 0.601\,x_8 + 0.73\,x_{10}$$

На Рис.2 представлено графік зашумленого збіжного ряду, який загалом зберігає характер істинного процесу.

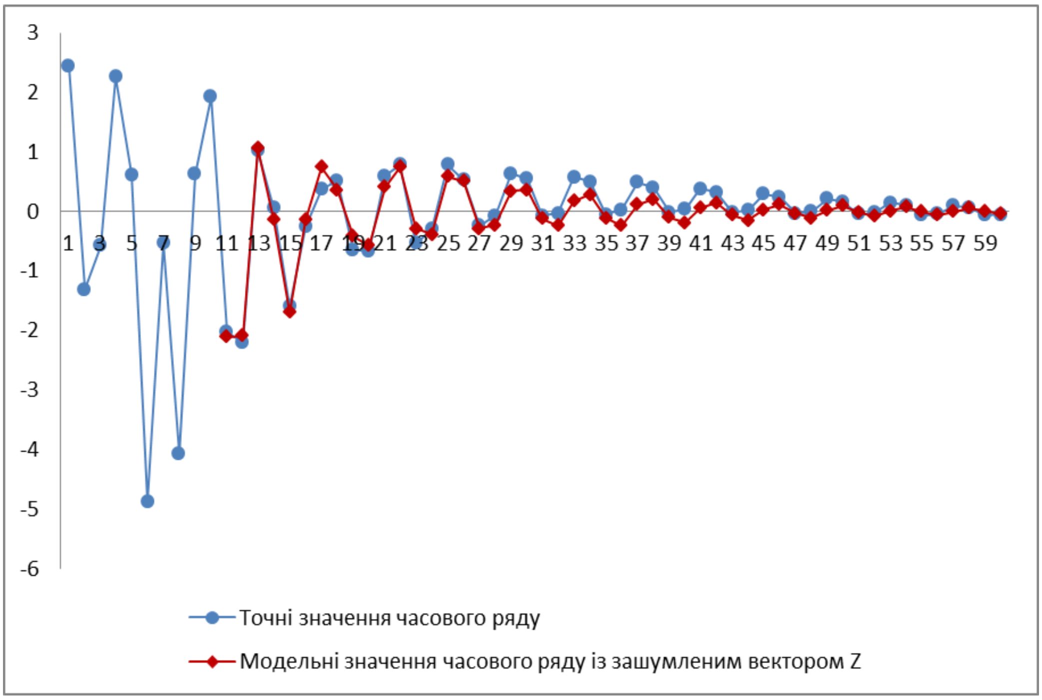

Рис. 2. Значення часового ряду: точні, модельні з зашумленим вектором Z.

Побудова моделі для розбіжного ряду. У чисельному експерименті за відсутності шуму тестова модель має 5 «істинних аргументів», тобто окремих запізнюваних значень з усіх 10 можливих:

$$y = -0.3x_1 + 0.9x_3 + 0.5x_5 + 0.6x_7 - 0.3x_9$$

Комбінаторно-генетичний алгоритм знайшов модель:

$$y = -0.312x_1 + 0.886\,x_3 + 0.502x_5 + 0.624x_7 - 0.311x_9$$

При зашумленні усього вектора початкових даних *z*, алгоритм КОМБІ-ГА зменшує складність моделі:

$$y = -0.569\,x_1 + 0.545\,x_7$$

Ступінь зашумленості початкових даних візуалізована на Рис.3. На Рис.4 показано незначна різниця між точним і модельним значенням часового ряду без шуму.

Модель за зашумленими даними розбіжного процесу хоч і неточна, але зберігає розбіжний характер, що представлено на Рис.5.

Для перевірки достовірності результату роботи адгоритму було здійснено порівняння знайдених ним оптимальних моделей з результатом повного перебору варіантів частинних моделей комбінаторним алгоритмом при чому для всіх випадків результати КОМБІ та КОМБІ-ГА співпадали.

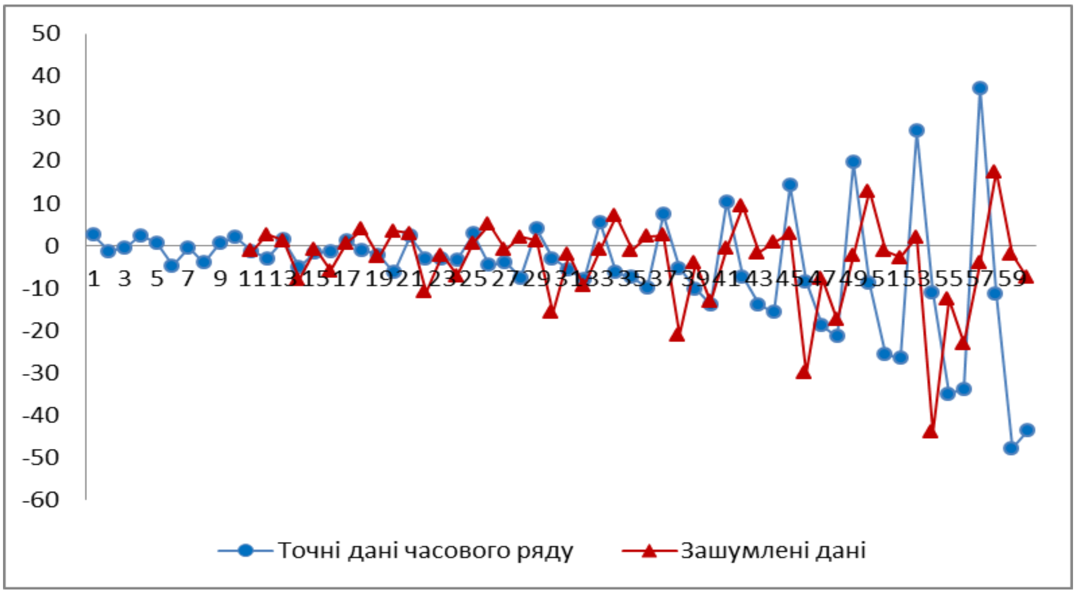

Рис. 3. Значення часового ряду: точні, зашумлені.

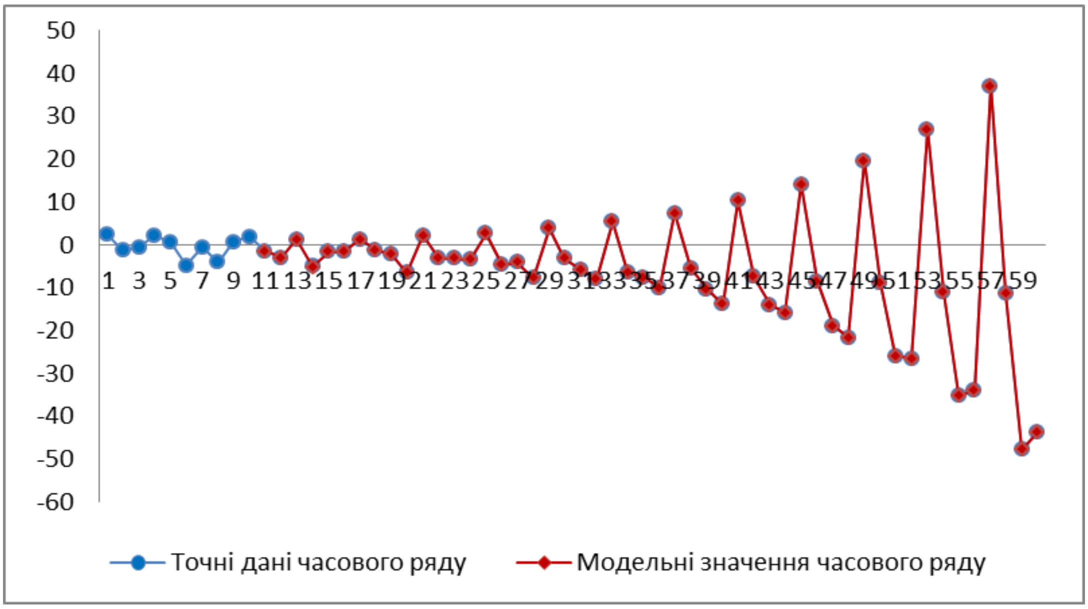

Рис. 4. Значення часового ряду: точні, модельні.

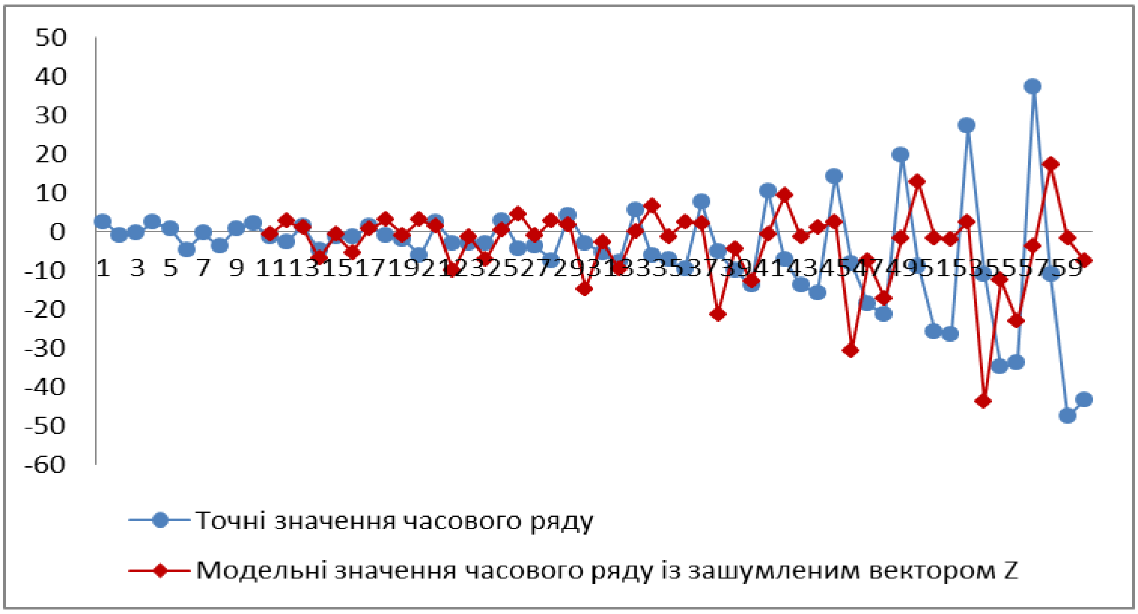

Рис. 5. Значення часового ряду: точні, модельні з зашумленим вектором z.

Висновки

При розв'язанні тестової задачі індуктивного моделювання як зі збіжним так і розбіжним часовим рядом алгоритмами КОМБІ та КОМБІ-ГА у авторегресійному класі було отримано однакові оптимальні моделі як з шумом, так і без нього при малих часових витратах та вимог до вхідних даних. КОМБІ-ГА знаходив модель більш ніж у 2 рази швидше.

При наявності шуму 30% у векторі z початкових змінних оптимальні моделі для обох рядів спрощувалися. Побудована КОМБІ-ГА модель розбіжного часового ряду втратила більше істинних значень, тому застосовування авторегресії при невеликому рівні шуму є доцільним. Точність відтворення істинного процесу в авторегресійному класі сильно залежить від рівня шуму через наявність залежності між значеннями початкового вектору.

Проведені тестові експерименти показують, що алгоритм КОМБІ-ГА можна ефективно використовувати при розв'язанні задач індуктивного моделювання в авторегресійному класі навіть у випадку зашумлених даних і, отже, доцільним є його використання при розв'язанні задач з реальними даними.

Література

[1] Льюнг Л. Идентификация систем. Теория для пользователей. М.: Наука,1991. 492 с.

[2] Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. Киев: Наукова думка, 1985. 216

[3] Moroz O., Stepashko V. Hybrid Sorting-Out Algorithm COMBI-GA with Evolutionary Growth of Model Complexity / N. Shakhovska, V. Stepashko. Advances in Intelligent Systems and Computing II Editors, AISC book series. Vol. 689. Cham: Springer, 2017. P. 346–360.

[4] Степашко В.С., Єфіменко С.М., Савченко Є.А. Комп’ютерний експеримент в індуктивному моделюванні. Київ: Наукова думка, 2014. 222 с.

[5] Мороз О.Г. Аналіз застосування генетичних алгоритмів у задачах глобальної оптимізації. Управляющие системи и машины. 2018. № 2. С. 73–79.

[6] Мороз О.Г., Степашко В.С. Тестування комбінаторно-генетичного алгоритму в задачі пошуку оптимальної нелінійної моделі. Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту». Залізний Порт, 2018. С. 256–258.

[7] Moroz O. Stepashko V. Data reconstruction of seasonal changes of amylolytic microorganisms amount in copper polluted soils. Proceedings of the XIII IEEE International Conference CSIT-2018 & International Workshop on Inductive Modeling. Lviv, Ukraine, 2018. P. 492–495.

[8] Єфіменко С.М. Прогнозування складних процесів у класі моделей векторної авторегресії на основі рекурентно-паралельного алгоритму COMBI МГУА. Індуктивне моделювання складних систем: Зб. наук. пр. К.: МННЦ ІТС НАН та МОН України, 2015. Вип. 7. С. 129–139.