УДК 519.163 + 681.5.015

ПРОГНОЗУВАННЯ СКЛАДНИХ ПРОЦЕСІВ У КЛАСІ МОДЕЛЕЙ ВЕКТОРНОЇ АВТОРЕГРЕСІЇ НА ОСНОВІ РЕКУРЕНТНО-ПАРАЛЕЛЬНОГО АЛГОРИТМУ COMBI МГУА

С.М.Єфіменко
Міжнародний науково-навчальний центр інформаційних технологій та систем (МННЦ ІТС) НАН та МОН України,
syefim@ukr.net

Розроблено теоретичні основи рекурентно-паралельних обчислень у комбінаторному алгоритмі МГУА для моделювання та прогнозування складних багатовимірних взамозв'язаних процесів у класі моделей векторної авторегресії. Ефективність розробленого алгоритму продемонстровано на прикладі розв'язання задачі прогнозування показників енергетичної сфери України з метою інформаційної підтримки управлінських рішень.
Ключові слова: індуктивне моделювання, МГУА, комбінаторний алгоритм, рекурентно-паралельні обчислення, кластерна система.

Theoretical grounds of recurrent-and-parallel computing in combinatorial GMDH algorithm for modeling and forecasting of complex multidimensional interrelated processes in the class of vector autoregression models are developed. The effectiveness of constructed algorithm is demonstrated by forecasting of interrelated processes in the field of power safety of Ukraine with the purpose of information support of administrative decisions.
Keywords: inductive modeling, GMDH, combinatorial algorithm, recurrent-and-parallel computing, cluster system.

Разработаны теоретические основы рекуррентно-параллельных вычислений в комбинаторном алгоритме МГУА для моделирования и прогнозирования сложных многомерных взаимосвязанных процессов в классе моделей векторной авторегрессии. Эффективность разработанного алгоритма продемонстрирована на примере решения задачи прогнозирования показателей энергетической сферы Украины с целью информационной поддержки управленческих решений.
Ключевые слова: индуктивное моделирование, МГУА, комбинаторный алгоритм, рекуррентно-параллельные вычисления, кластерная система.

Вступ

Розглядається задача математичного моделювання і прогнозування багатовимірних взаємозв'язаних часових рядів, яка знаходить своє застосування передусім в економічній, екологічній, соціологічній сферах [1, 2]. Якщо моделюванню одновимірних часових рядів у науковій літературі приділяється багато уваги, то досвід моделювання багатовимірних часових рядів є недостатнім.

У випадку прогнозування векторного процесу, представленого у вигляді сукупності часових рядів (багатовимірного часового ряду) природно орієнтуватися на такий клас моделей, як векторна авторегресія [3]. Далі

розглядається один з можливих підходів до структурно-параметричної ідентифікації такого процесу, коли параметри для кожної моделі оцінюються незалежно. Недоліком такого підходу є те, що для взаємозв'язаних процесів параметри окремих моделей взаємозалежні. Для усунення цього недоліку у роботі використовується алгоритм, за яким для кожного з процесів обирається не одна найкраща модель, а декілька. Це робиться для того, щоб із обраних найкращих моделей скомбінувати всі можливі варіанти систем і за додатковим критерієм обрати найкращу.

1 Моделі векторної авторегресії для прогнозування багатовимірних взаємозв'язаних процесів

Модель векторної авторегресії (Vector AutoRegression, VAR) була запропонована Крістофером Сімсом в 1980-му році та узагальнює модель авторегресії на багатовимірний випадок. Будується вона за стаціонарними часовими рядами. Це система рівнянь, в якій кожна змінна (компонента багатовимірного часового ряду) є лінійною комбінацією всіх змінних у попередні моменти часу. Порядок такої моделі визначається порядком запізнюваних значень (лагів). Для найпростішого випадку двох часових рядів із одним лагом модель VAR має такий вигляд:

$$\begin{aligned} x_1(t) &= \theta_{11} x_1(t-1) + \theta_{12} x_2(t-1); \\ x_2(t) &= \theta_{21} x_1(t-1) + \theta_{22} x_2(t-1), \end{aligned} \tag{1}$$

де $\theta_{ij},\ i, j = 1,2$ – параметри моделі.

У загальному випадку для m часових рядів та k лагів така модель має вигляд системи m рівнянь

$$\begin{aligned} & x_1(t) = \theta_{11} x_1(t-1) + \ldots + \theta_{1k} x_1(t-k) + \theta_{1,k+1} x_2(t-1) + \ldots + \\ & + \theta_{1,2k} x_2(t-k) + \ldots + \theta_{1,mk} x_m(t-k); \\ & \ldots \\ & x_m(t) = \theta_{m1} x_1(t-1) + \ldots + \theta_{mk} x_1(t-k) + \theta_{m,k+1} x_2(t-1) + \ldots + \\ & + \theta_{m,2k} x_2(t-k) + \ldots + \theta_{m,mk} x_m(t-k), \end{aligned} \tag{2}$$

або в матричному вигляді

$$X(t) = \sum_{j=1}^{k} \Theta_j X(t-j), \tag{3}$$

де $\Theta_j,\ j = \overline{1,k}$ – матриці коефіцієнтів моделі (3) розмірності $m \times m$.

Структуру моделі (3) та її коефіцієнти будемо визначати за комбінаторним алгоритмом COMBI МГУА [4]. Моделі формуються у вигляді системи лінійних різницевих рівнянь. Число аргументів для кожного з m взаємозв'язаних процесів становить $m \cdot k$.

2 Методика структурно-параметричної ідентифікації моделей векторної авторегресії

Загальна структура моделі у вигляді системи m різницевих рівнянь визначається в результаті виконання таких операцій:

1. Виходячи з кількості m взаємозв'язаних процесів та запізнюваних значень k, формується масив даних з $m \cdot k$ аргументів.

2. Визначається максимальна складність для перебору та виконується моделювання за алгоритмом COMBI МГУА з послідовним ускладненням структур моделей на основі рекурентно-паралельних обчислень. Для кожної вихідної змінної відбирається F кращих (за значенням критерію регулярності) моделей. Загалом на наступний крок передається $F \cdot m$ моделей.

3. Виконується перебір $G = F^m$ варіантів систем моделей та відбирається найкраща за значенням системного інтегрального критерію якості векторних моделей, який обчислюється на заданій частині початкової вибірки даних у режимі прогнозу процесу на задане число кроків n_S:

$$B = \sum_{i=1}^{n_S} \sum_{j=1}^{m} (x_{ij} - x_{ij}^*)^2 \ , \qquad (4)$$

де x_{ij}^* – результат покрокового інтегрування системи з m рівнянь.

При визначенні свободи вибору F (кількості кращих моделей) слід враховувати, що, час виконання третього кроку (перебору всіх варіантів систем моделей) швидко зростає та може перевищити прийнятне значення.

На рис. 1 показано експериментальні результати для тестової задачі з m=11 часовими рядами та k=2 запізнюваними значеннями, згідно з якими час перебору варіантів систем моделей (з обчисленням значення системного інтегрального критерію) зростає експоненційно.

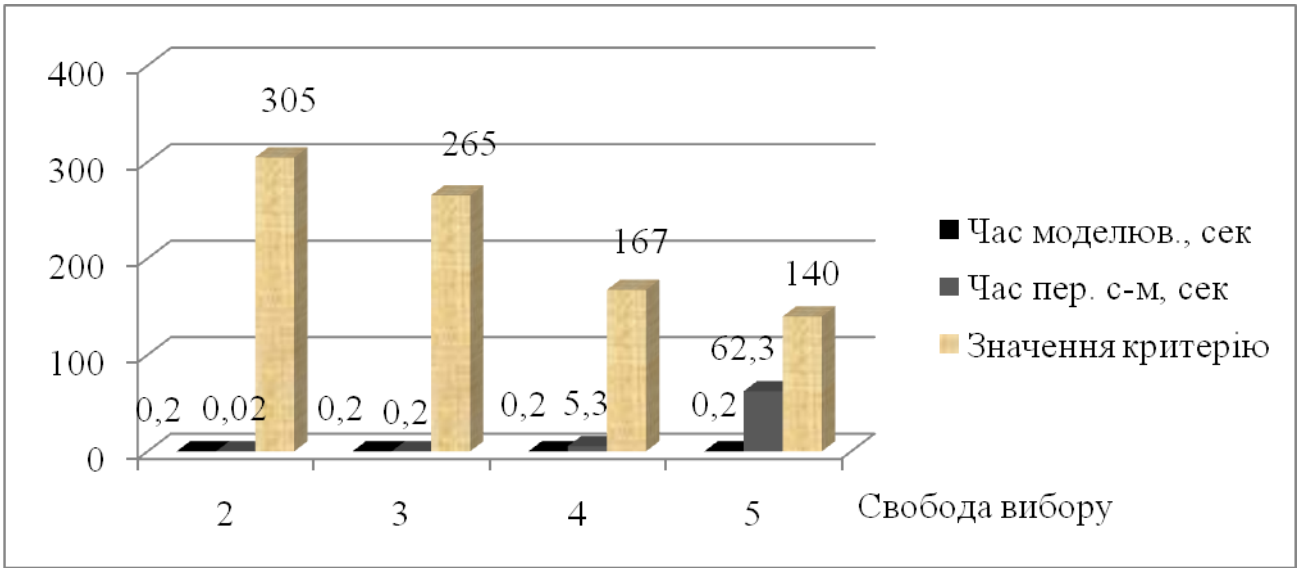

Рис. 1 Експериментальні результати в залежності від свободи вибору F

3 Побудова моделей векторної авторегресії на основі рекурентно-паралельного алгоритму

Для побудови моделей векторної авторегресії використовується комбінаторний алгоритм COMBI МГУА [5] з послідовним ускладненням структур моделей, рекурентним оцінюванням параметрів моделей за допомогою модифікованого алгоритму Гаусса та розпаралелюванням обчислень на кластерних багатопроцесорних системах. Використання алгоритму COMBI передбачає повний перебір всіх можливих моделей та вибір найкращої за значенням критерію. В процесі перебору для кожного з m взаємозв'язаних процесів порівнюються лінійні моделі з $g = m \cdot k$ (у нашому випадку) аргументами (входами)

$$\hat{y}_\nu = X_\nu \hat{\theta}_\nu,\ \nu = 1,\ldots,2^g - 1 \qquad (5)$$

де десятковому числу ν ставиться у відповідність двійковий структурний вектор $d_\nu = \{d_i\},\ \ i = \overline{1, g}$ з елементами 0 або 1 (включення або не включення в модель відповідного аргументу). Послідовність генерації двійкових векторів організовано таким чином, що спочатку утворюються всі сполучення з однією одиницею у складі структурного вектора (усього генерується $C_g^1 = g$ можливих варіантів), потім – з двома одиницями ($C_g^2 = \frac{g(g-1)}{2}$ можливих варіантів), і т.д. до одного варіанту ($C_g^g = 1$) включення в модель усіх аргументів.

Спосіб рівномірного розбиття загальної кількості моделей на всі процесори кластерної системи (за умови, що однакова загальна кількість аргументів припадатиме на кожен процесор) розроблено в [5]. Схему алгоритму COMBI з послідовним ускладненням структур для побудови моделей векторної авторегресії вибрано у зв'язку з тим, що вона дозволяє частково розв'язувати задачу повного перебору у випадку, коли такий перебір за прийнятний час моделювання стає неможливим навіть із розпаралелюванням (орієнтовно при g>50). У такому випадку повний перебір доцільно виконувати не серед усіх можливих моделей, а лише моделей обмеженої складності. Іншою причиною застосування обмеження на складність може бути недостатня кількість точок навчальної частини вибірки n_A. При $n_A < g$ перебір буде виконуватися серед моделей зі складністю не більше, ніж n_A. Якщо ж одночасно $g < 50$ та $g < n_A$, то структури моделей будуть ускладнюватися від 1 до g, тобто реалізовуватиметься повний перебір усіх серед усіх можливих моделей.

Оскільки час побудови моделей векторної авторегресії фактично визначається часом перебору варіантів систем моделей (це чітко видно за двома діаграмами для свободи вибору F=5 з рис. 1), то ефективність розпаралелювання цього етапу алгоритму є навіть більш пріоритетною, аніж ефективність розпаралелювання етапу побудови F кращих моделей для

кожного з m виходів (часових рядів). Тому розглянемо спосіб розпаралелювання цього етапу.

Оскільки виконується перебір F^m варіантів систем моделей, то доцільно використовувати структурний вектор у вигляді числа з основою F, елементи $\{f_i\}, i=\overline{1,m}$ якого змінюються від 1 до F та вказують, яка з кращих (за значенням критерію) моделей для i–го часового ряду входитиме в модель VAR. Покажемо на простому прикладі (m=2, F=3) всі варіанти систем моделей (їх буде $G=3^2=9$):

$\{1,1\}$ (для першого та другого виходів беруться перші з кращих моделей)

$\{1,2\}$ (для першого виходу береться перша модель, для другого - друга)

$\{1,3\}$ (і т.д.)

$\{2,1\}$

$\{2,2\}$

$\{2,3\}$

$\{3,1\}$

$\{3,2\}$

$\{3,3\}$ (для першого та другого виходів беруться треті за зростанням значення критерію моделі).

Таку схему можна досить легко застосувати для розпаралелювання на задану кількість процесорів. Ідея рівномірного розбиття загальної кількості системних моделей на p процесорів кластерної системи полягає у наступному:

– визначаємо загальну кількість системних моделей $G=F^m$;

– кількість моделей G рівномірно розподіляємо між усіма p процесорами кластера;

– якщо G не ділиться націло на p, то на перший процесор приходиться більше навантаження (але не більше, ніж на p додаткових моделей), аніж на інші процесори.

4 Тестування ефективності розпаралелювання алгоритму

Для експериментального визначення ефективності розпаралелювання розробленого алгоритму було виконано тестовий експеримент по розв’язанню задачі структурно-параметричної ідентифікації. Будувалася системна модель для m=11 часових рядів з k=2 запізнюваними значеннями. Обчислення було розподілено на п’ять потоків і послідовно виконано на персональному комп’ютері з процесором Intel Pentium M з частотою 1.73 ГГц. Таким чином отримуємо результат, наближений до теоретичного через виключення втрат, пов’язаних із міжпроцесорною взаємодією.

Для кожного часового ряду відбиралося F=5 кращих моделей за критерієм регулярності [4]. Вимірювався час виконання кожного з п’яти

потоків та час роботи програми без розпаралелювання. Результат експерименту у вигляді діаграми часу виконання представлено на рис. 2.

Результати експерименту можна використати для обчислення ефективності розпаралелювання

$$E = \frac{T_1}{5 \times T_{5\max}} \times 100\% \qquad (6)$$

та рівномірності навантаження

$$P = (1 - \frac{T_{5\max} - T_{5\min}}{T_{5\max}}) \times 100\%, \qquad (7)$$

де T_1 – час виконання алгоритму з одним потоком (тобто без розпаралелювання), T_{5max} – час виконання алгоритму з розпаралелюванням на 5 потоків (визначається як максимальний серед п’яти потоків час виконання програми), T_{5min} – мінімальний серед п’яти потоків час виконання програми.

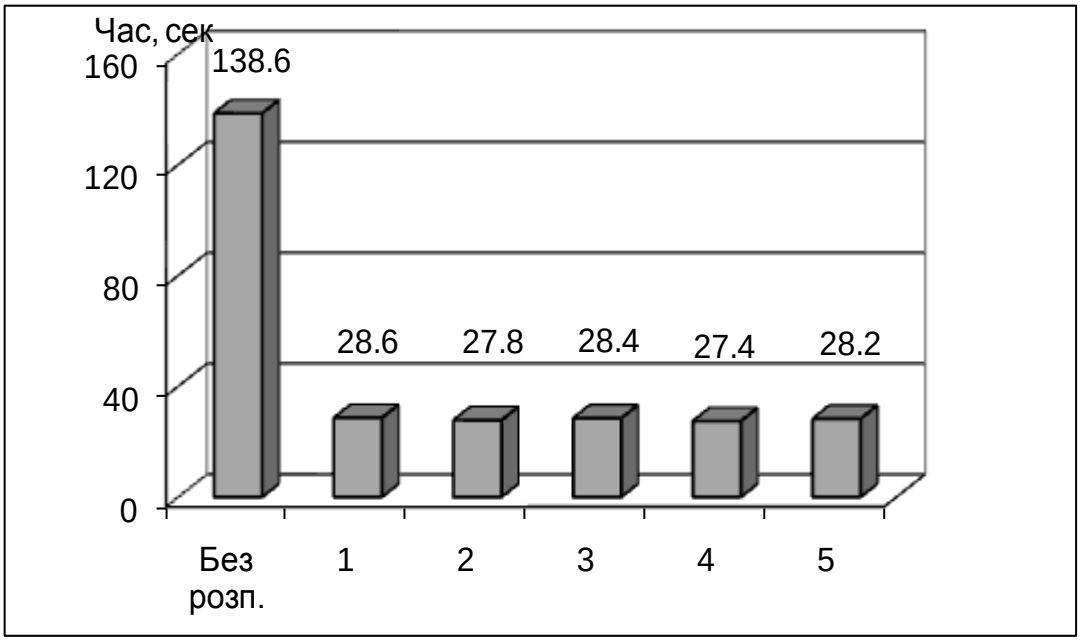

Рисунок 2 – Час виконання алгоритму

Суть формули (6) полягає в тому, що, якщо при використанні розпаралелювання на n потоків (у даному випадку п’яти) час моделювання зменшується у n разів, то ефективність розпаралелювання становить 100%. Відповідно до (7), для забезпечення стовідсоткової рівномірності навантаження всі обчислювачі (ядра, процесори, процеси) мають виконувати моделювання за один і той же час.

В таблиці 1 представлено експериментальні значення показників ефективності алгоритму

Таблиця 1.

Експериментальні результати

Ефективність розпаралелювання, %	97
Рівномірність навантаження, %	96

5 Моделювання та прогнозування показників енергетичної сфери України

Вибір для дослідження показників енергетичної сфери України обумовлений стратегічною важливістю енергетичного комплексу для нашої держави в сучасних умовах. Метою дослідження є розроблення економіко-математичної моделі для використання задля підтримки прийняття ефективних управлінських рішень.

5.1 Постановка задачі та опис початкових даних

Було використано дані Міністерства економіки за 1996 – 2005 роки (усього по n=10 точок) для 11 показників енергетичної сфери України:

x_1 – частка власних джерел в балансі паливно-енергетичних ресурсів (ПЕР), %;

x_2 – частка домінуючого паливного ресурсу у споживанні ПЕР, %;

x_3 – частка імпорту палива з однієї країни (компанії) у загальному обсязі його імпорту, %;

x_4 – знос основних виробничих фондів підприємств ПЕК, %;

x_5 – відношення інвестицій в підприємства ПЕК до ВВП;

x_6 – енергоємність ВВП, кг нафтового еквіваленту/$;

x_7 – обсяг видобутку вугілля, млн. т.;

x_8 – транзит нафти, млн. т.;

x_9 – транзит газу, млрд. куб. м.;

x_{10} – обсяг видобутку природного газу, млрд. куб. м.;

x_{11} – обсяг видобутку нафти і газового конденсату, млн. т.

з метою побудови моделі векторної авторегресії цих показників у вигляді (3) та отримання прогнозних значень на 2006 рік. Для перших десяти показників відомі реальні дані за 2006 рік (табл. 2), що дозволяє оцінити точність прогнозування.

5.2 Побудова прогнозних моделей векторної авторегресії

Для отримання моделі векторної авторегресії (3) для кожного з 11 показників відбиралися по 5 кращих моделей (у вигляді системи лінійних різницевих рівнянь), структура та коефіцієнти яких визначалися за алгоритмом COMBI МГУА з послідовним ускладненням. Значення свободи вибору F=5

було встановлено з метою отримання прийнятного часу моделювання. Вибірку спостережень було поділено у співвідношенні 2/1: навчальна підвибірка n_A містила 5 точок, перевірочна n_B – 3. Кількість запізнюваних значень k=2 було вибрано, виходячи з невеликої кількості точок спостережень. Тож загалом для кожного з 11 показників виконувався перебір серед 22 аргументів з обмеженням складності 5 (n-k-n_B).

Після відбору 55 кращих моделей виконувався перебір $G = 5^{11}$ варіантів систем моделей та відібрано найкращу за значенням системного інтегрального критерію якості векторних моделей (4), який обчислювався у режимі прогнозу процесу на 8 кроків.

Таблиця 2.

Дані показників енергетичної сфери України

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
x_1	50	50	50	49	49.9	50.26	49.9	46.75	47.4	47.34	48.78
x_2	55	55	55	55	43	43.3	41.5	43.2	42.2	41.93	40.64
x_3	56.7	56.7	56.7	56.7	56.7	56.7	56.7	62.7	76.17	68.45	54.59
x_4	60	60	60	60	61.8	62	62	62	62.00	60.6	60.9
x_5	2	2	2	2	2	3.1	2.6	2.6	2.30	2.62	2.74
x_6	1.08	1.1	1.07	1.05	0.99	0.93	0.89	0.84	0.79	0.79	0.76
x_7	74.8	75.9	76.2	62.84	62.4	61.6	59.4	64.61	60.68	60.35	61.63
x_8	53.5	53	53.5	54	27.4	27.4	27.4	33	31.4	25.28	33.21
x_9	139.9	133.2	141.1	133.3	123.6	124.4	121.4	129.3	137.1	136.4	128.5
x_{10}	18.4	18.1	18	18.1	17.4	17.6	17.8	18.7	19.2	20.5	
x_{11}	4.1	4.13	3.9	3.8	3.7	3.71	3.74	3.98	4.18	4.357	

5.3 Результати прогнозування та їх аналіз

Побудовано таку системну модель:

$$x_1(t) = 0.74x_3(t-2) + 42.93x_6(t-2) - 0.06x_9(t-2) + 7.96x_{11}(t-1) - 15.16x_{11}(t-2)$$
$$x_2(t) = 0.27x_2(t-1) - 0.93x_3(t-2) - 0.59x_9(t-1) - 0.44x_9(t-2) + 57.04x_{11}(t-2)$$
$$x_3(t) = -5x_1(t-1) + 4.28x_1(t-2) - 0.29x_7(t-2) - 0.5x_9(t-2) + 9.98x_{10}(t-2)$$
$$x_4(t) = 1.46x_4(t-2) + 0.08x_7(t-2) + 0.14x_8(t-2) - 0.04x_9(t-2) - 8.68x_{11}(t-2)$$
$$x_5(t) = 0.03x_2(t-2) - 0.003x_7(t-1) - 0.05x_8(t-1) + 0.01x_9(t-1) + 0.01x_9(t-2)$$
$$x_6(t) = 0.02x_1(t-1) - 0.03x_4(t-1) + 0.75x_6(t-2) + 0.004x_{10}(t-1) + 0.15x_{11}(t-1)$$
$$x_7(t) = -1.04x_2(t-1) + 0.78x_4(t-2) + 288.53x_6(t-1) - 249.19x_6(t-2) + 9.33x_{11}(t-2)$$
$$x_8(t) = 1.81x_2(t-1) - 4.71x_5(t-1) - 1.96x_9(t-1) - 1.86x_9(t-2) + 118.35x_{11}(t-2)$$
$$x_9(t) = 7.38x_1(t-2) - 13.94x_4(t-2) - 1.46x_8(t-2) - 0.6x_9(t-2) + 41.94x_{10}(t-2)$$
$$x_{10}(t) = 0.32x_4(t-2) - 11.34x_6(t-2) + 0.04x_8(t-2) - 0.04x_9(t-2) + 3.35x_{11}(t-2)$$
$$x_{11}(t) = 0.08x_4(t-2) - 0.21x_5(t-1) - 2.06x_6(t-2) + 0.01x_7(t-1) + 0.27x_{11}(t-1)$$

Значення відносних помилок у відсотках для перших дев'яти показників представлені в таблиці 3.

Таблиця 3.

Точність моделей на екзаменаційній вибірці (2006-й рік)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Відносна помилка, %	9.3	11	38.9	5.7	6.2	4.3	22	4.3	30.6

Результати моделювання та табличні значення показників x_3 та x_8, для яких отримано відповідно найгіршу та найкращу точність на екзаменаційній вибірці, представлено на рисунках 3 та 4. На рисунках 5 та 6 показано результати моделювання для показників x_{10} та x_{11}, для яких побудовано чистий прогноз на 2006 рік. Для кожного показника модельні значення отримано шляхом інтегрування процесу з початкових умов (двох перших експериментальних точок)

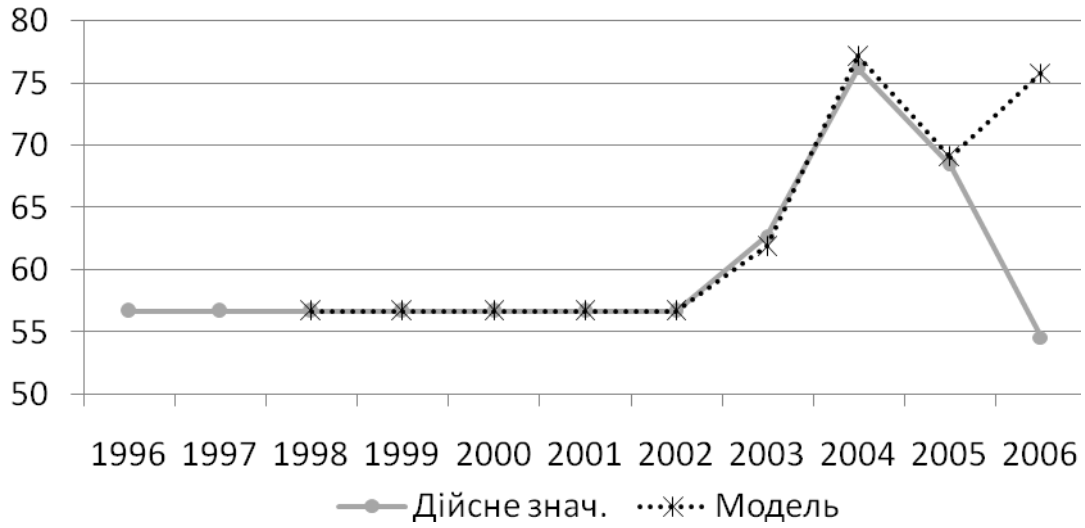

Рисунок 3 – Результати моделювання для показника x_3

Рисунок 4 – Результати моделювання для показника x_8

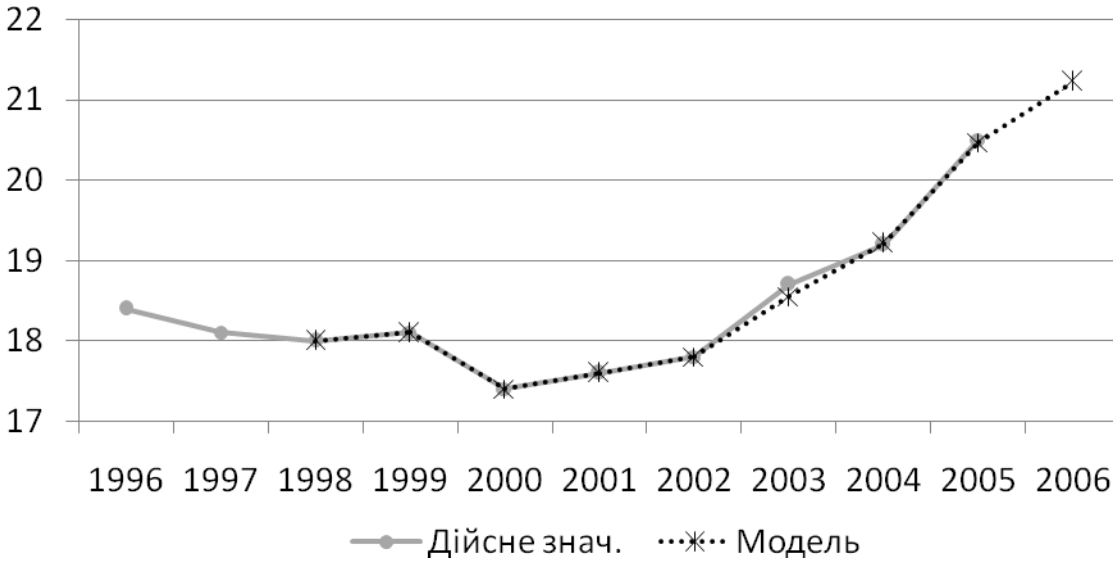

Рисунок 5 – Результати моделювання для показника x_{10}

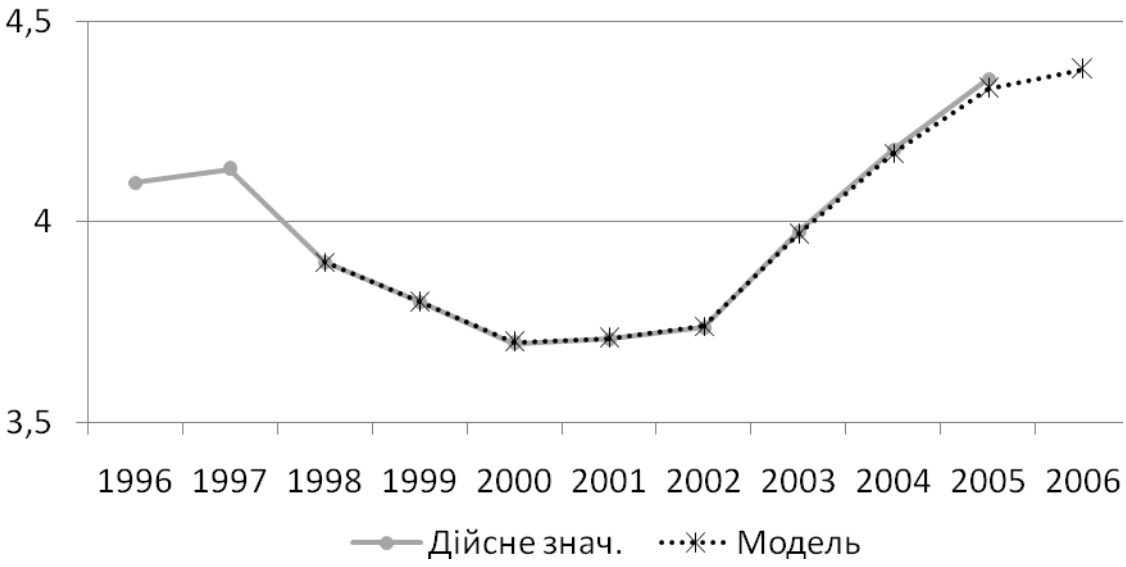

Рисунок 6 – Результати моделювання для показника x_{11}

Побудована системна модель точно описує експериментальні дані для всіх 11-ти показників енергетичної сфери. Для більшості показників модель VAR має прийнятні прогностичні властивості. Розходження реальних та прогнозних значень для деяких показників можна пояснити занадто короткою вибіркою та, можливо, не врахуванням деяких економічних та інших факторів.

Таким чином, побудовану системну модель можна застосовувати для середньострокового прогнозу, адже інтегрування оптимальною системою має високу точність на 8 точок, а для більшості показників – навіть на 9.

Висновки

Розроблено принцип розпаралелювання обчислень в комбінаторному алгоритмі COMBI МГУА з рекурентним оцінюванням параметрів моделей для побудови дискретних прогнозних моделей динаміки складних багатовимірних взаємозв'язаних процесів.

Важливою особливістю розробленої схеми розпаралелювання алгоритму COMBI є те, що вона дозволяє частково розв'язувати задачу повного перебору у випадку, коли кількість аргументів для перебору перевищує можливості схеми алгоритму з повним перебором.

Розроблено програмні засоби для моделювання та прогнозування складних багатовимірних взаємозв'язаних процесів на основі високопродуктивного рекурентно-паралельного алгоритму МГУА у класі дискретних динамічних моделей векторної авторегресії, які застосовано для моделювання і прогнозування взаємозв'язаних процесів у сфері енергетичної безпеки України з метою інформаційної підтримки управлінських рішень.

Література

1. Єфіменко С.М., Кваша Т.К., Степашко В.С. Системне прогнозування динаміки взаємозалежних показників енергетичної сфери України // Індуктивне моделювання складних систем. Збірник наукових праць. – Київ: МННЦ ІТС. – 2009.– С. 54 – 60.

2. Костенко Ю.В. Моделювання багатовимірних циклічних процесів за дворівневим алгоритмом МГУА // Там же, вип. 3. – 2011. – С. 99-109.

3. Гурский, С. К. Адаптивное прогнозирование временных рядов в электроэнергетике / С. К. Гурский. – Минск: Наука и техника, 1983. – 271 с.

4. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. - Киев: Наук. думка, 1985. - 216с.

5. Єфіменко С.М. Комбінаторний алгоритм МГУА з послідовним ускладненням структур моделей на основі рекурентно-паралельних обчислень // Індуктивне моделювання складних систем: Зб. наук. пр. – К.: МННЦ ІТС НАН та МОН України, 2014. – Вип. 6. – С. 64-71.