УДК 004.832.32

Кравець І.О., Афанасьєва Г.О.

ДОСЛІДЖЕННЯ МЕТОДІВ СТАТИСТИЧНОГО ТА ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДЛЯ АВТОРЕГРЕСІЙНИХ МОДЕЛЕЙ

Було проаналізовано методи для регресійного аналізу та прогнозування, придатні для авторегресійних моделей. Розглянуті метод Альмона, метод групового урахування аргументів та нейронні мережі. Запропоновано модифікацію методу групового урахування аргументів для множинного регресійного аналізу авторегресійних та дистрибутивно лагових моделей.

Ключові слова: *авторегресійні моделі, метод групового урахування аргументів, нейронні мережі.*

Проанализированы методы регрессионного анализа и прогнозирования применительно к авторегрессионым моделям. Рассмотрены метод Альмона, метод группового учета аргументов и нейронные сети. Предложена модификация метода группового учета аргументов для множественного регрессионного анализа авторегрессионых и дистрибутивно лагових моделей.

Ключевые слова: *авторегрессионные модели, метод группового учета аргументов, нейронные сети.*

The efficiency of statistical methods for regressive analyzes, which can be applied for autoregressive models is researched. The Almon's method, the groups accounting method and neuron' networks are examined. The modification of groups accounting method for autoregressive models is proposed.

Key words: *авторегресійні models, method of group account of arguments, neuron networks.*

ВСТУП

При статистичному аналізі та прогнозуванні соціально-еконономічних показників необхідно враховувати інерційність та гальмування факторів, тобто необхідно використовувати так звані авторегресійні та дистрибутивно лагові моделі.

У регресійному аналізі, якщо регресійна модель включає не лише поточні, а й попередні (лагові, або затримані) значення незалежних змінних (x), вона має назву ***дистрибутивно-лагова модель (ДЛМ).*** У той же час, якщо до моделі включене одне або більше попередніх значень залежної змінної *(y),* вона має назву ***авторегресійна модель.*** Таким чином, ***дистрибутивно-лагова модель****:*

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \varepsilon_t ; \qquad (1)$$

авторегресійна модель:

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + \varepsilon_t$$

де ε_t – випадкова величина.

Шляхом перетворень дистрибутивно-лагова модель може бути зведена до авторегресійної моделі. Такі авторегресійні моделі також відомі під назвою ***динамічних моделей,*** оскільки вони відображають часові зміни залежної змінної щодо її попереднього (попередніх) значень. Але, нажаль, традиційні методи статистичного аналізу (метод найменших квадратів) дають велики похибки та подібних моделей внаслідок кореляції випадкових похибок від спостереження до спостереження. Тому аналіз та розробка методів аналізу та прогнозування, придатних для

авторегресійних моделей та при недостатньому обсязі статистичних даних є актуальною задачею.

ЗАДАЧІ ДОСЛІДЖЕННЯ

Задачею дослідження у даної роботі є пошук та розробка методів, придатних для аналізу та прогнозування авторегресійних моделей.

Для таких випадків найбільш придатні метод Альмона, метод групового врахування аргументів, та прогнозування нейронимі мережами.

1. МЕТОД АЛЬМОНА ДЛЯ ДИСТРИБУТИВНО-ЛАГОВИХ МОДЕЛЕЙ

Метод Альмона розроблено для простого регрейсійного аналізу економетричних показників [1].

Кінцева ДЛМ у k періодів:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + ... + \beta_k x_{t-k} + \varepsilon_t, \quad (2)$$

може бути записана в такий спосіб:

$$y_t = \alpha + \sum_{i=0}^{k} \beta_i x_{t-i} + \varepsilon_t .$$

Альмоном запропонований підхід, при якому коефіцієнти β_i, згідно з теоремою Вейерштрасса, можна апроксимувати поліномом відповідного ступеня від i – величини тимчасового лагах [2], тобто:

$$\beta_i = a_0 + a_1 i + a_2 i^2 + ... + a_m i^m, \quad (3)$$

де m – ступінь полінома, причому передбачається, що $m < k$.

Після підстановки (3) в (2) одержимо:

$$y_t = \alpha + \sum_{i=0}^{k} (a_0 + a_1 i + a_2 i^2 + ... + a_m i^m) x_{t-i} + \varepsilon_t =$$

$$= \alpha + a_0 \sum_{i=0}^{k} x_{t-i} + a_1 \sum_{i=0}^{k} i x_{t-i} + a_2 \sum_{i=0}^{k} i^2 x_{t-i} + ... + a_m \sum_{i=0}^{k} i^m x_{t-i} + \varepsilon_t . \quad (4)$$

Після заміни

$$Z_{0t} = \sum_{i=0}^{k} x_{t-i} = x_t + x_{t-1} + x_{t-2} + ... + x_{t-k},$$

$$Z_{1t} = \sum_{i=0}^{k} i x_{t-i} = x_{t-1} + 2x_{t-2} + 3x_{t-3} + ... + k x_{t-k},$$

$$Z_{mt} = \sum_{i=0}^{k} i^m x_{t-i} = x_{t-1} + 2^m x_{t-2} + 3^m x_{t-3} + ... + k^m x_{t-k} \quad (5)$$

отримуємо **модель Альмона**

$$y_t = \alpha + a_0 Z_{0t} + a_1 Z_{1t} + a_2 Z_{2t} + ... + a_m Z_{mt} + \varepsilon_t . \quad (6)$$

За умови задоволення ε_t всім припущенням класичної моделі лінійної регресії для оцінки параметрів α і a_i можна використовувати стандартний метод найменших квадратів (МНК). У цьому основна перевагу моделі Альмона перед з звичайною моделлю, тому що остання має серйозні проблеми у зв'язку із присутністю авторегресійної змінної y_{t-1}, яка можливо корелює з випадковою величиною ε.

Таким чином, одержавши оцінки параметрів моделі a_1, знаходимо оцінки параметрів моделі (1) по формулі (3):

$$\beta_0 = a_0,$$

…,

$$\beta_k = a_0 + ka_1 + k^2 a_2 + ... + k^m a_m. \quad (4)$$

Максимальну довжину тимчасового лага k вибирають відносно невеликим залежно від розміру вихідної вибірки. Не менш важливим є вибір ступеня полінома m, який повинна бути хоча б на одиницю більше кількості екстремумів у залежності $\beta_i(i)$. Таким чином, вибір цього параметра також є суб'єктивним. На практиці допускається вибор невеликих значень m (порядку 2-3).

Щоб визначити, чи досить полінома m-й ступені для апроксимації моделі знаходять оцінки параметрів моделі при $m = m + 1$, тобто:

$$y_t = \alpha + a_0 Z_{0t} + a_1 Z_{1t} + a_2 Z_{2t} + ... + a_m Z_{mt} + a_{m+1} Z_{m+1t} + \varepsilon_t. \quad (5)$$

Тоді якщо параметр a_m – статистично значимий, а a_{m+1} – ні, то можна припустити, що ступені полінома m досить для апроксимації. Але можливо наявність мультиколінеарности між Z_i, що може дати неадекватну оцінку параметра a_{m+1}. У цьому випадку оцінка a_{m+1} статистично не значима через те, що початкова вибірка недостатньо інформативна для оцінки впливу змінної Z_{m+1} на y. Оцінки a_m можуть мати помилки внаслідок мультиколінеарности Z_i, тоді частина коефіцієнтів a_i будуть статистично незначущими, але це не означає статистичну незначимість коефіцієнтів β_i.

Недоліки методу:

- Ступінь полінома й максимальний довжина лага вибираються досить суб'єктивно.
- Проблема мультиколінеарності змінних Z_i.
- Метод не пристосовано для множинного регресійного аналізу.

2. МЕТОД ГРУПОВОГО УРАХУВАННЯ АРГУМЕНТІВ

На сьогоднішній день розроблено багато методів статистичного аналізу даних та прогнозування часових рядів, але вони базуються на математичному апараті, якій може бути використаний лише при достатньому обсязі статистичних даних. Часто обсяг даних є недостатнім для пошуку залежностей для великої кількості вхідних факторів.

Найкращим виходом з цього становища є використання метода групового урахування аргументів (МГУА), запропонованого академіком О. Івахненко [2].

У даній роботі запропоновано використання методу групового урахування аргументів для множинного ресійного аналізу авторегресійних та дистрибутивно лагових моделей. Було обрано багаторядні поліноміальні моделі МГУА.

Багаторядні алгоритми МГУА застосовуються для рішення некоректних чи недовизначених задач моделювання, тобто у випадку, коли число точок у таблиці дослідних даних не більше числа аргументів, що входять у синтезовану модель. Методи регресійного аналізу в цьому випадку незастосовні, тому що не дають можливості побудови єдиної моделі, адекватної процесу.

Вважаємо, що початковий склад аргументів, з якого починається процедура багаторядної селекції моделі процесу, будується на так званому нульовому ряді алгоритму, що організується по-різному в поліноміальний і гармонійних алгоритмах.

У випадку авторегресійних моделей до аналогічного виду зводиться модель, отримана за допомогою перетворення усіх вхідних змінних, запізнювань вихідних змінних і заданих нелінійних функцій від них.

$$y_k^{(1)} = a_0 + a_i x_i + a_j x_j + a_{ij} x_i x_j + a_{ii} x_i^2 + a_{jj} x_j^2 + a_{ij} + 1 * y_k - 1. \quad (6)$$

Число часткових описів 1-го ряду дорівнює $M = n(n - 1)/2$.

Багаторядні алгоритми, як правило, працюють за наступною схемою:

1-ий ряд – на основі даних таблиці спостережень будуються часткові описи від усіх попарних комбінацій початкових даних (перепозначених) аргументів, що наближають по МНК вихідну змінну *y*:

$$y_1 = f_1(x_1, x_2),\ y_2 = f_2(x_1, x_3),...,\ y_k = f_k(y(t-1), x_n). \quad (7)$$

З цих моделей вибирається деяке число кращих за зовнішнім критерієм селекції.

2-ий ряд – отримані змінні приймаються як аргументи – входи другого ряду, і знову будуються всі часткові описи від двох аргументів:

$$z_1 = \varphi_1(y_1, y_2),\ z_2 = \varphi_2(y_1, y_3),..., z_l = \varphi_l(y_{F_1-1}, y_F). \quad (8)$$

З них за зовнішнім критерієм відбирається F2 кращих модулів у якості змінних наступного ряду і т.д. Ряди нарощуються доти, поки знижується значення зовнішнього критерію.

Кожний частковий опис може бути лінійною

$$y(t) = a_0 + a_1 x_i(t) + a_2 x_k(t), \text{ або } y(t) = a_0 + a_1 x_i(t) + a_2 y(t-1) \quad (9)$$

чи нелінійною

$$f = a_0 + a_1 x_i + a_2 y(t-1) + a_3 x_i y(t-1) + a_4 x_i^2 + a_5 y^2(t-1) \quad (10)$$

функцією від двох змінних, коефіцієнти яких можна визначити МНК, маючи відповідну кількість точок спостереження в навчальній послідовності. Виключивши проміжні змінні після останову алгоритму, одержимо модель, число коефіцієнтів у якій значно перевищує число точок.

На рис. 1 приведена блок-схема багаторядного алгоритму МГУА, який працює таким чином.

Вибірка даних розбиваються на дві: перевірочну та навчальну вибірки.

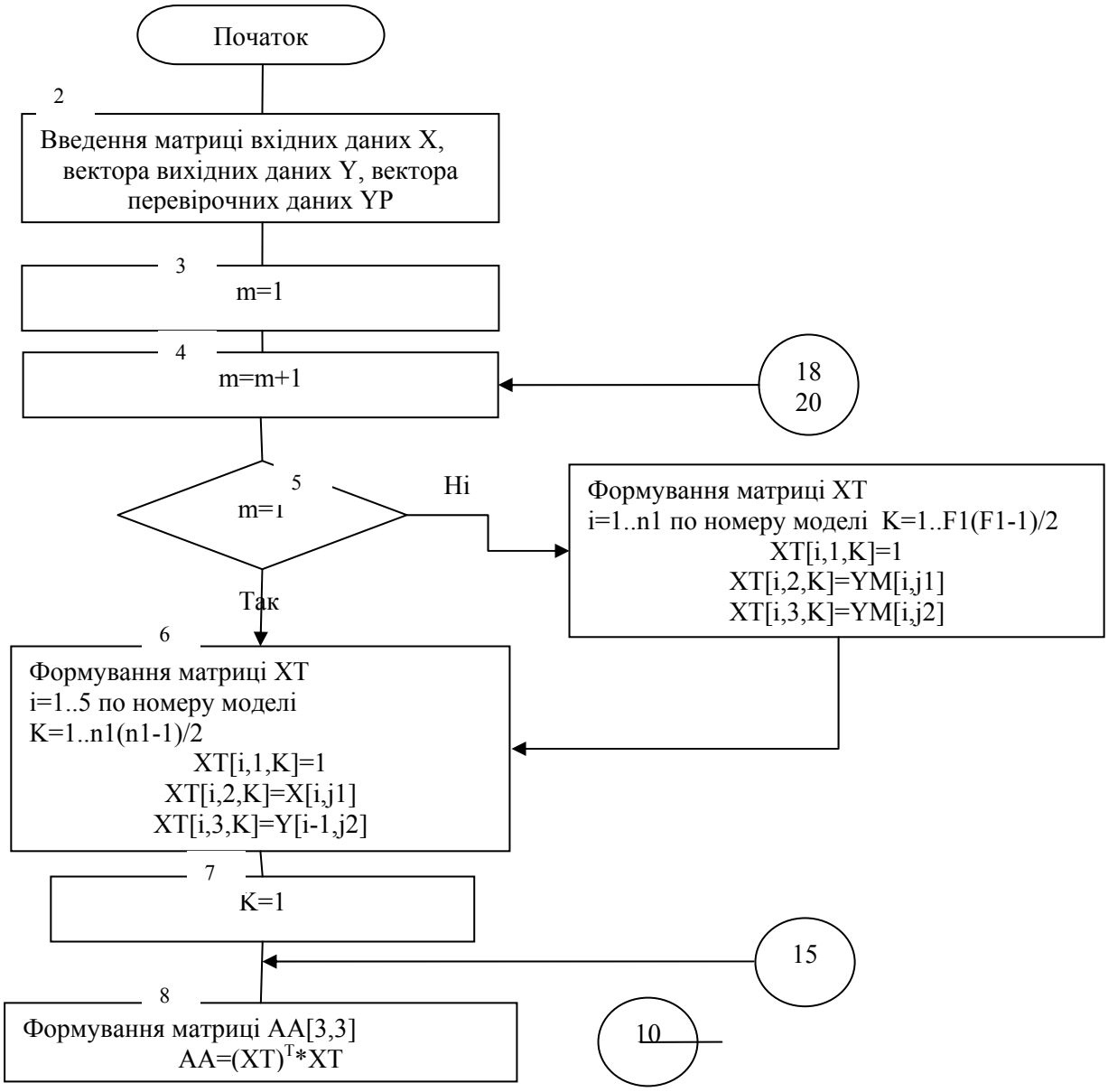
На кожному рівні селекції m відбувається побудова часткових моделей (лінійних або квадратичних) тільки від двох змінних. Причому на першому рівні входами часткових моделей є вхідні фактори навчальної вибірки, а на інших рівнях селекції вхідними даними часткових моделей є вихідні значення часткових моделей попереднього рівня селекції. Знаходження коефіцієнтів часткових моделей відбувається методом найменших квадратів (МНК). Для кожної часткової моделі обчислюється критерій (точнісний або робастний). І за критерієм вибору обирається найкраща модель (або декілька кращих), вихідні значення яких є вхідними даними часткових моделей наступного рівня селекції. Перехід на наступний рівень селекції відбувається до тих пір, поки найкраще значення з критеріїв вибору часткових моделей на цьому рівні зменшується порівнюючи з попереднім рівнем селекції. Коли критерій вибору перестав зменшуватись переходимо на заключний етап алгоритму: рухаючись від кінця до початку і роблячи послідовну заміну перемінних, обчислюються вирази для шуканої моделі у початковому просторі описів.

Існує два підходи при виборі часткових моделей: точнісний та робастний [2]. В першому підході в алгоритмі МГУА при виборі часткових моделей використовується критерій регулярності (або точнісний), який визначається як середньо-квадратична помилка між фактичними значеннями перевірочної вибірки та прогнозними значеннями для точок перевірочної вибірки, але побудованих по моделям навчальної вибірки.

В другому підході в алгоритмі МГУА при виборі часткових моделей використовується критерій незміщеності (або робастний), який визначається як середньо-квадратична помилка між прогнозними значеннями, побудованими за моделями навчальної та перевірочної вибірок. (Робимо 2 експерименти: одна вибірка навчальна, друга перевірочна. Спочатку будуємо часткові моделі по першої. Потім вибірки міняються місцями, будуємо часткові моделі по другій вибірці. Середньо-квадратична помилка між цими видами моделей для відповідних змінних і є критерій часткової моделі.

Не існує точних рекомендацій щодо того, скільки моделей відбирати на поточному рівні селекції за допомогою цих критеріїв. Тому запропоновано на кожному поточному рівні селекції обирати такі моделі, для яких значення критерію селекції менше значення критерію селекції кращої моделі попереднього рівня.

Перехід на наступний рівень селекції відбувається до тих пір, поки найкраще значення з критеріїв вибору часткових моделей на цьому рівні зменшується порівнюючи з попереднім рівнем селекції. Коли критерій вибору перестав зменшуватись переходимо на заключний етап алгоритму: рухаючись від кінця до початку і роблячи послідовну заміну перемінних, обчислюються вирази для шуканої моделі у початковому просторі описів.



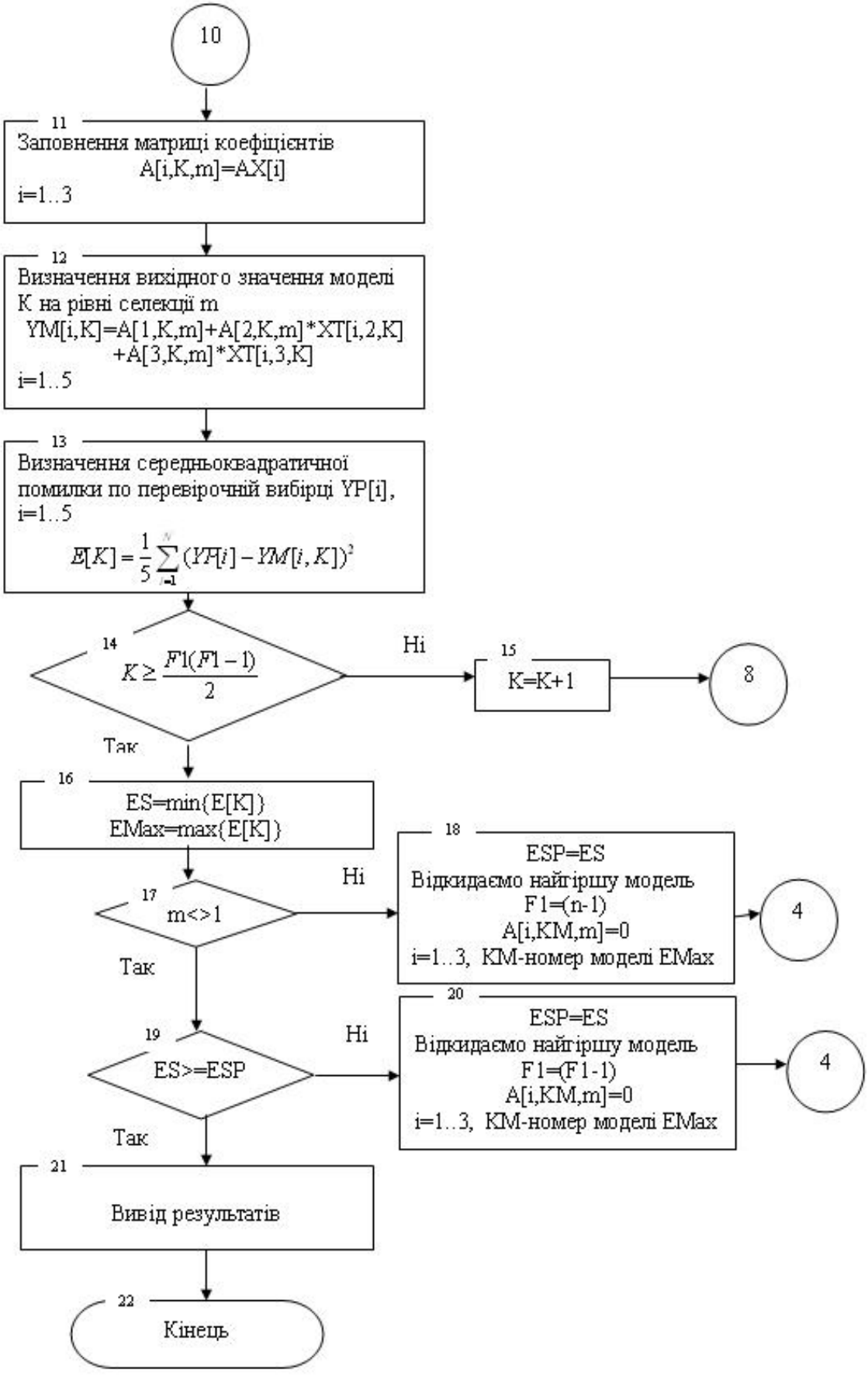


Рис. 1. Блок-схема багаторядного поліноміального алгоритму МГУА для авторегресійнрї моделі

Було розроблено інформаційну аналітичну систему, яка реалізує багаторядний алгоритм МГУА для авторегресійної моделі. Нижче наведено порівняльний графік прогнозованих та реальних значень (для ВВП), Залежність не лише від попередніх значень вихідної змінної, а й від 5 показників. Обсяг навчальної та перевірочної вибірок – 7 точок:

а)

б)

Рис 2. Порівняння прогнозованих та точних значень при прогнозуванні за допомогою МГУА:
а) багаторядний алгоритм; б) багаторядний алгоритм та авторегресійна модель

3. ПРОГНОЗУВАННЯ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Для прогнозування за допомогою нейронних мереж найбільш придатні нейронні мережі зворотнім розповсюдженням похибки. Розроблений алгоритм прогнозування використанням мережі зворотного поширення похибки та плаваючих вікон з врахуванням зв'язків характеристик, що прогнозуються.

Якщо необхідно урахувати кореляцію прогнозованого показника з іншими показниками, то вхідним вектором є значення прогнозованого показника в попередніх часових кроках та значення інших показників в попередніх часових кроках, або у даному часовому кроці.

Якщо потрібно прогнозувати показники далі, то значення вхідного вектора складаються зі значень показників зсунутих у часі вперед, що утворює ефект плаваючого вікна у часі.

Найкраща ширина плаваючого вікна біля 10. Це мережа, яка навчається з учителем. Навчання здійснюється шляхом послідовного пред'явлення вхідних векторів з одночасним налаштуванням ваг відповідно до ітераційної процедури. Даний алгоритм дозволяє виконувати прогнозування декількох показників з урахуванням кореляційних зв'язків між ними. Деякі з параметрів, що приймаються до уваги, справляють незначний вплив на формування виходів і можуть бути відкинуті. Число вхідних нейронів 12, число нейронів прихованого прошарку 6, число вихідних нейронів 2 (2 показника прогнозують одночасно)

Задачі прогнозування з використанням мережі зворотного поширення похибки та плаваючих вікон, який враховує зв'язки прогнозуємих характеристик дає хороший прогноз, який становить до 10 % похибки відповідно до реальних даних. Але число показників обмежено, та ми не отримуємо регресійної моделі. Крім того, добрі результати були отримані для монотонних моделей.

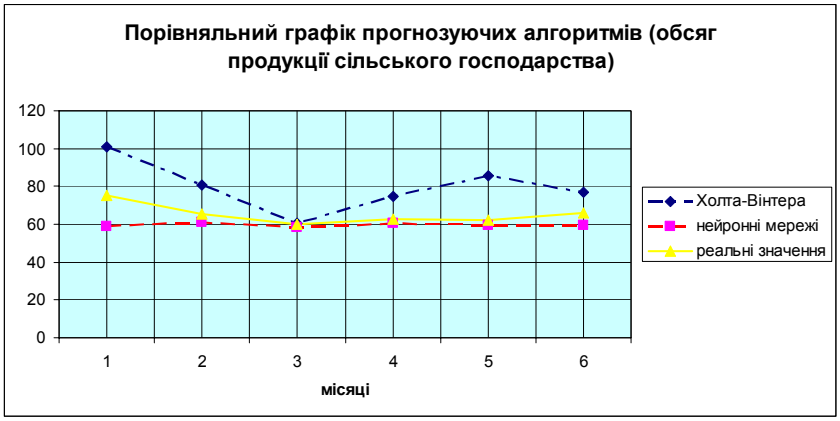


Рис. 3. Порівняльний графік прогнозованих та реальних значень на січень-червень 2008 року («Обсяг продукції сільського господарства»)

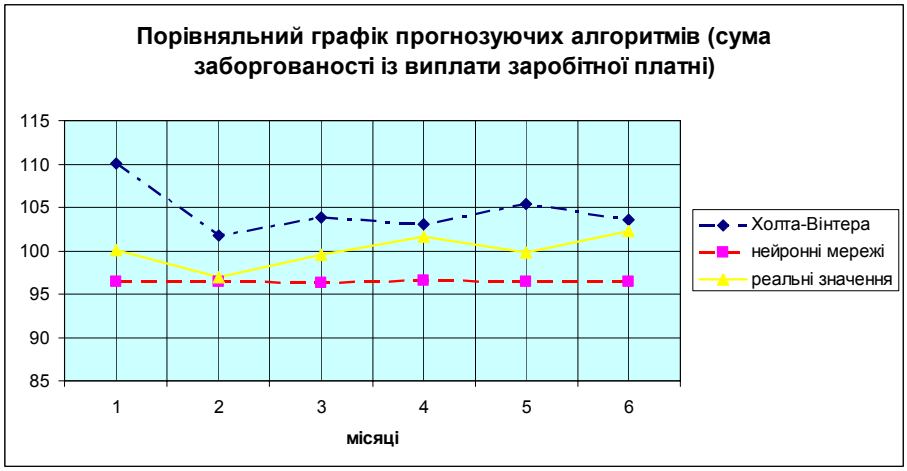


Рис. 4. Порівняльний графік прогнозованих та реальних значень на січень-червень 2008 року («Сума заборгованості із виплати заробітної платні»)

ВИСНОВКИ

Було проаналізовано методи для регресійного аналізу та прогнозування, придатні для авторегресійних моделей. Розглянуті метод Альмона, МГУА та нейронні мережі. Запропоновано модифікація методу групового урахування аргументів для множинного регресійного аналізу авторегресійних та дистрибутивно лагових моделей. Дослідження показали, що найбільш придатним для для авторегресійних моделей є метод групового урахування аргументів, який дозволяє отримати коефіцієнті моделі при недостатньому обсязі даних та має більшу точність прогнозу.

ЛІТЕРАТУРА

1. Лук'яненко І.Г., Краснікова Л.І. Економетрика. – К.: Товариство «Знання», ККО, 1998. – 494 с.
2. Ивахненко А.Г., Зайченко Ю.П., Димитров В.Д. Принятие решений на основе самоорганизации. – М.: «Сов. радио», 1976. – 280 с.
3. Уоссермен З.Ф. Нейрокомпьютерная техника: теория и практика. – М.: Мир, 1992.
4. Саймон Хайкин Нейронные сети: полный курс, 2-е издание: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.
5. Дебок Г., Кохонен Т. Аналіз финансовых данных с помощью самоорганизующихся карт / Пер. с англ. – М.: Издательский Дом «АЛЬПИНА», 2001. – 317 с.

Рецензенти: д.т.н., проф. Данілов В.Я.
д.т.н., проф. Бідюк П.І.

 Стаття надійшла до редколегії 16.01.09