УДК 004.832.32

КРАВЕЦЬ І.О., УЗУН Т.Ф., *Миколаївський державний гуманітарний університет ім. П. Могили*

Кравець Ірина Олександрівна – кандидат технічних наук, доцент кафедри інтелектуальних інформаційних систем МДГУ ім. Петра Могили. Коло наукових інтересів: статистичні методи обробки даних, чисельні методи, інтелектуальні системи.

Узун Тетяна Федорівна – магістрант МДГУ ім. Петра Могили. Коло наукових інтересів: статистичні методи обробки даних, інтелектуальні системи, бази даних, WEB-дизайн.

ВИБІР ТА ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ АЛГОРИТМІВ DATA MINING СТОСОВНО АНАЛІЗУ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ПОКАЗНИКІВ

Проведено дослідження ефективності алгоритмів Data Mining для аналізу соціально-економічних показників. Розроблено та реалізовано для аналізу соціально-економічної інформації алгоритми пошуку дерев рішень для задач класифікації, кластерного аналізу, прогнозування даних, алгоритм пошуку асоціативних правил. Запропоновано алгоритм кластеризації з визначенням як центрів кластерів, так і числа кластерів та алгоритм прогнозування часових рядів з автоматичним вибором вигляду прогнозування.

The efficiency of Data Mining algorithms, which are applied for the analysis of social economical parameters, is researched. The algorithm of searching decision's tree for classification, the cluster analyses algorithm, time series forecasting algorithm and the algorithm of searching associated groups are developed.

DATA MINING (розробка даних) є однієї із нових та дуже актуальних інформаційних технологій в організації та аналітичної обробці даних. Розробка даних – це процес витягу статистично достовірної комплексної інформації з баз даних, а також використання цієї інформації для прийняття рішень. При цьому аналіз та виявлення закономірностей повинні проводиться базуючись на науково – обґрунтованих статистичних методах та у режимі ON-LINE і кожного разу з новими даними. Недаром зараз швидкий розвиток мають OLAP-технології та OLTP-технологій (On-Line Transaction Processing), які використовуються в системах керування базами даних та вважаються динамічними системами підтримки прийняття рішень (СППР). Статистичні пакети (SPSS, Statistica, статистичні додатки Excel) використовують науково - обґрунтовані статистичні методи, але не пов'язані з базами даних, працюють в автономному режимі та вимагають від користувача ручного вводу інформації та специфічні знання. Крім того, методи математичної статистики виявляються корисними, головним чином, для перевірки заздалегідь сформульованих гіпотез и добре працюють при достатньому обсязі статистичних даних, заданих у чисельному вигляді Але не працюють при малому обсягу даних особливо у якісному вигляді. У Data Mining тягар формулювання гіпотез і виявлення незвичайних шаблонів перекладено з людини на комп'ютер. Новітні технології інтелектуального аналізу використовують складний статистичний аналіз і моделювання для автоматичного знаходження гіпотез, моделей та відношень, прихованих у базі даних.

Мета даної роботи – вибір алгоритмів Data Mining для аналізу соціально-економічних показників та дослідження їх ефективності.

Була створена інформаційно-аналітична система, яка поєднує базу даних та інформаційний блок, може слугувати демонстраційним прикладом використання методів статистичного аналізу та інтелектуального аналізу з використанням технологій Data Mining.

Для аналізу соціально-економічної інформації було проаналізовано алгоритми розв'язання задач класифікації, кластерного аналізу; прогнозування даних, пошуку асоціативних правил. Також були реалізовані для порівняння традиційні статистичні алгоритми, такі як: однофакторний дисперсійний аналіз; множинний регресійний аналіз, тренд-аналіз часових рядів.

Методами Data Mining розв'язуються три основні задачі : класифікація та регресія, пошук асоціативних правил, кластеризація. [1, 2]. В задачі класифікації і регресії потрібно визначити значення залежної змінної об'єкта на підставі значень інших змінних, що характеризують його. Якщо залежна змінна приймає чисельні значення то це задача регресії, якщо залежна змінна приймає якісні значення то це задача класифікації.

Для розв'язання задач регресії використовуються таки статистичні методи, як кореляційний, однофакторний дисперсійний аналіз (перевірка впливу вхідних факторів) та множинний регресійний аналіз [4, 6]. Однак, ці методи можуть бути використані при достатньому обсязі статистичних даних.

Найбільш розповсюджені моделі, що відображують результати класифікації, – це класифікаційні правила, дерева рішень, розподіляючи математичні (лінійні і нелінійні) функції [1, 2]. Розподіляючи математичні функції використовуються при чисельному представлення вхідних даних, але соціально-економічна інформація часто представлена у якісному вигляді.

Найпоширенішими алгоритмами побудови класифікаційних правил є алгоритми 1R та Баєсовські алгоритми. Однак вони мають суттєві недоліки: алгоритм 1R будує правило по одної змінної, Баєсовські алгоритми працюють з незалежними вхідними змінними. Тому для побудови класифікаційних правил було обрано побудова дерев рішень, а з алгоритмів – алгоритм покриття.

Метою алгоритму покриття є розбивка навчальної вибірки таким чином, щоб одержати підмножини, що відповідають класам. Даний підхід полягає в побудові дерев рішень для кожного класу окремо. На кожному кроці алгоритму обирається значення змінної, яке розділяє всю множину на дві підмножини. Така розбивка робиться доти, поки не буде побудована підмножина, що містить тільки об'єкти одного класу. Блок схема алгоритму представлена на рис. 1. Результат роботи алгоритму на прикладі аналізу рівня міграції у районах області представлено на рис. 2. Отримано правило класифікації виду: „Якщо рівень безробіття високий, заробітна плата низька, товарообіг низький, індекс цін середній, то рівень міграції високий". Алгоритм дозволяє формувати класифікаційні правила при якісному вигляді інформації, не вимагає незалежності змінних та великого обсягу статистичних даних.

Задача кластеризації полягає в пошуку незалежних груп (кластерів) і їхніх характеристик у всій безлічі аналізованих даних. Рішення цієї задачі допомагає краще зрозуміти дані. Крім того, угруповання однорідних об'єктів дозволяє скоротити їхнє число, полегшити аналіз. Так при аналізі соціально-економічної інформації можна згрупувати райони та області по їх показникам, виявити відсталі та провідні районі.

Найпоширенішими алгоритмами кластерізації є дентограми та алгоритми пошуку оптимальної розбивки на кластери по мінімізації функції відстані об'єктів від центрів кластерів [1, 2]. Було проаналізовано декілька алгоритмів ієрархічних та неієрархічних та був обраний неієрархічний алгоритм Fuzzy C-Means [1]. Ідея алгоритму полягає у визначенні центрів кластерів і віднесенні до кожного кластеру об'єктів найбільш близько підходящих до даного кластеру. Його відмінність полягає в тому, що кластери тепер є нечіткими множинами і кожна точка належить різним кластерам з різним ступенем належності. Точка відноситься до того або іншого кластера за критерієм максимуму належності даному кластеру.

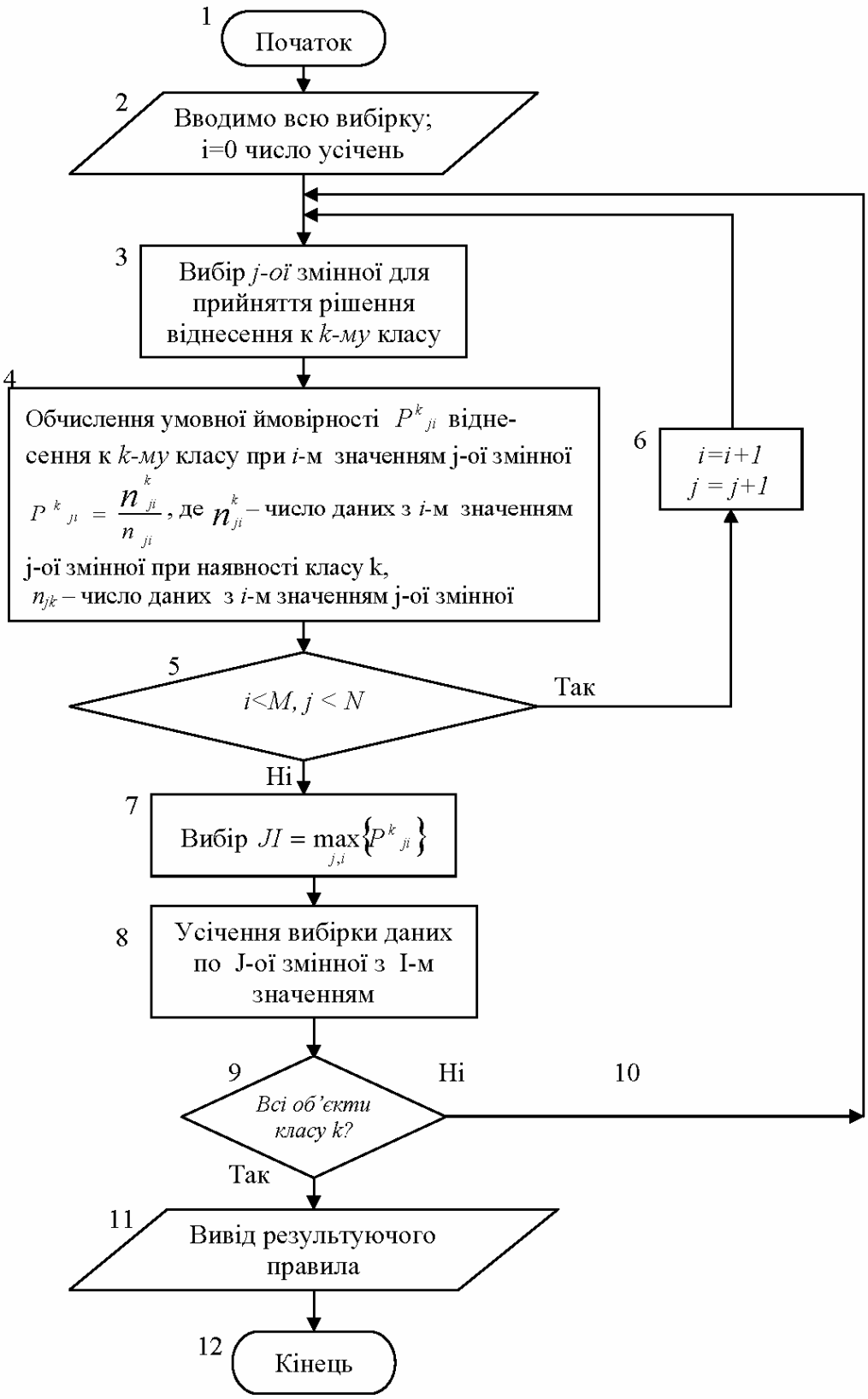
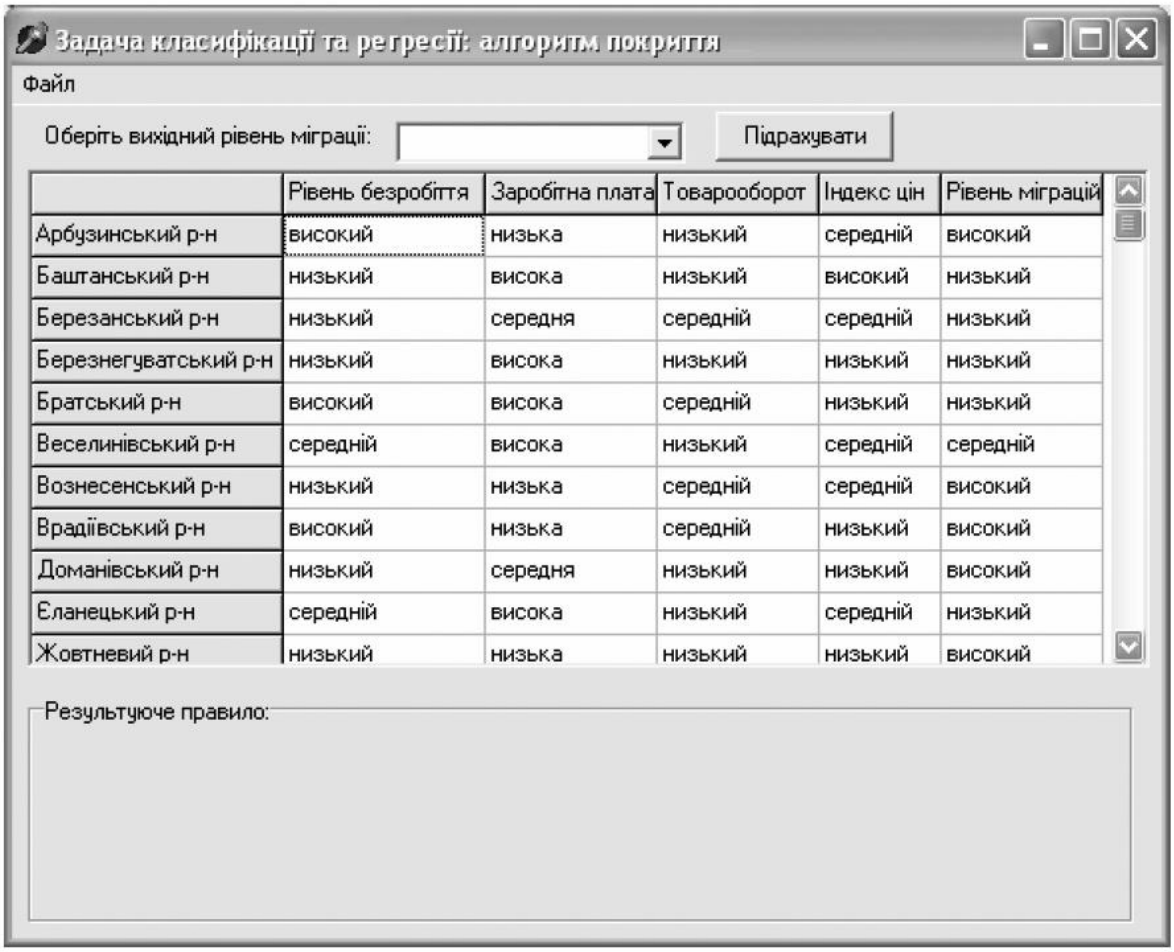
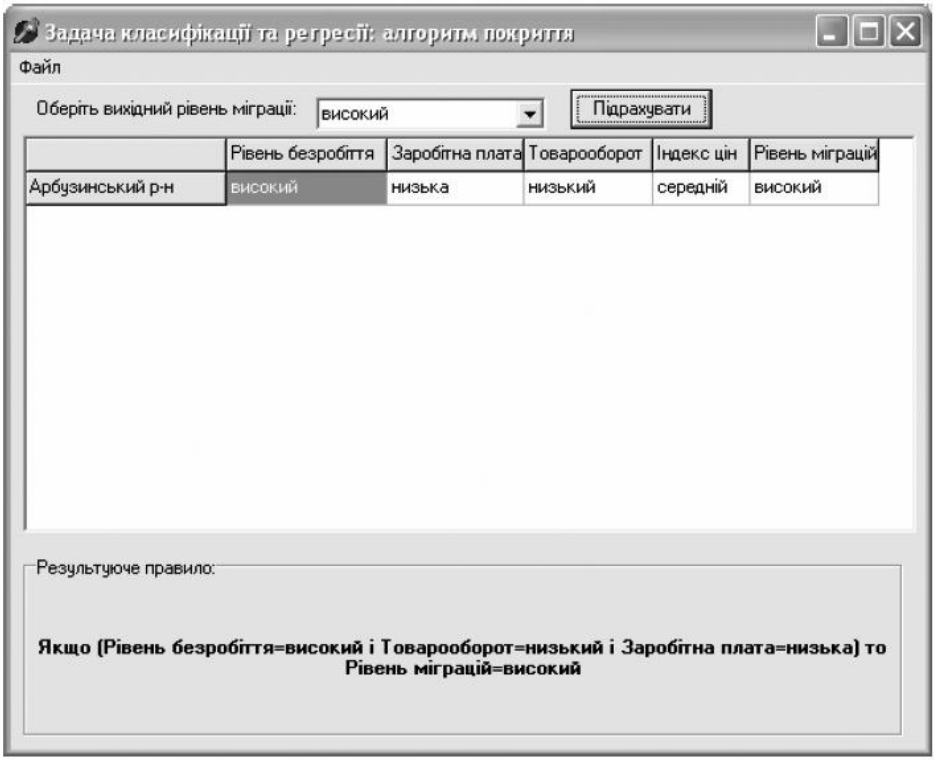


Рис. 1. Блок-схема алгоритму покриття



	Рівень безробіття	Заробітна плата	Товарооборот	Індекс цін	Рівень міграцій
Арбузинський р-н	високий	низька	низький	середній	високий
Баштанський р-н	низький	висока	низький	високий	низький
Березанський р-н	низький	середня	середній	середній	низький
Березнегуватський р-н	низький	висока	низький	низький	низький
Братський р-н	високий	висока	середній	низький	низький
Веселинівський р-н	середній	висока	низький	середній	середній
Вознесенський р-н	низький	низька	середній	середній	високий
Врадіївський р-н	високий	низька	середній	низький	високий
Доманівський р-н	низький	середня	низький	низький	високий
Єланецький р-н	середній	висока	низький	середній	низький
Жовтневий р-н	низький	низька	низький	низький	високий



	Рівень безробіття	Заробітна плата	Товарооборот	Індекс цін	Рівень міграцій
Арбузинський р-н	високий	низька	низький	середній	високий

Рис. 2. Результати роботи алгоритму класифікації.

Запропоновано модифікація алгоритму з визначенням як центрів кластерів, так і числа кластерів. Блок-схему алгоритму наведено на рис. 3. У алгоритмі використані таки позначення:

- навчальна множина $M = \{m_j\}_{j=1}^{d}$, d – кількість точок (векторів) даних об'єктів;
- матриця належності $U = \{u_{ij}\}$ j-го об'єкта і-му кластеру.
- координати центру кластера c_i
- цільова функція:

$$J(M,U,C) = \sum_{i=1}^{c}\sum_{j=1}^{d} u_{ij}^{\omega} d_A^2(m_j, c^{(i)}), \tag{1}$$

де $\omega \in (1,\infty)$ – показник нечіткості (ваговий коефіцієнт), що регулює нечіткість розбивки. Звичайно використовується $\omega = 2$;

- набір обмежень:

$$u_{ij} \in [0,1];\ \sum_{i=1}^{c} u_{ij} = 1;\ 0 < \sum_{j=1}^{d} u_{ij} < d, \tag{2}$$

який визначає, що кожен вектор даних може належати різним кластерам з різним ступенем належності, сума належностей елемента даних усім кластерам простору розбивки дорівнює 1.

Результат роботи алгоритму представлено на рис. 4. Алгоритм дозволяє розбивати районі на області, не вимагаючи незалежності показників та великого обсягу статистичних даних.

Задача прогнозування є ледве не основною задачею великого числа фахівців, що працюють з даними. Запропоновано та розроблено алгоритм аналізу та прогнозування часових рядів з вибором вигляду прогнозування (короткосрокове або довгосрокове), з врахуванням або без врахування сезонної зміни показників при автоматизованому визначенні моделі (мультиплікативна або адитивна). При аналізі часових рядів визначення тренду відбувається за методом найменших квадратів, вигляд моделі автоматично визначається по розмаху коливань часового ряду (якщо розмах коливань збільшується при збільшенні тренду-модель мультиплікативна, якщо розмах коливань не змінюється то модель-адитивна). Перевірка правильності тренд-аналізу відбувається по автокореляційної функції випадкових залишків (якщо автокореляційна функція згасає, тред-аналіз зроблено вірно).

Довгострокове прогнозування відбувається методом декомпозиції: на етапі аналізу визначиються складові випадкового процесу (тренд, сезонна компонента, випадкові залишки) та виконується прогноз на майбутній термін.

Короткосрокове прогнозування виконується за допомогою згладжування ряду методом Холта-Вінтера: з врахуванням сезонної компоненти (метод Вінтера), або без врахування сезонної компоненти (метод Холта). При цьому тренд апроксимується ламаною, перетин та нахил якої перераховуються для кожного інтервалу часу за допомогою рекурентних формул [3, 4, 5]

Визначення коефіцієнтів $b_0(t), b_1(t)$ часового тренду $tr(t+1) = b_0(t) + b_1(t) * \Delta t$, ($\Delta t = 1$) відбувається за такими рекурентними формулами:

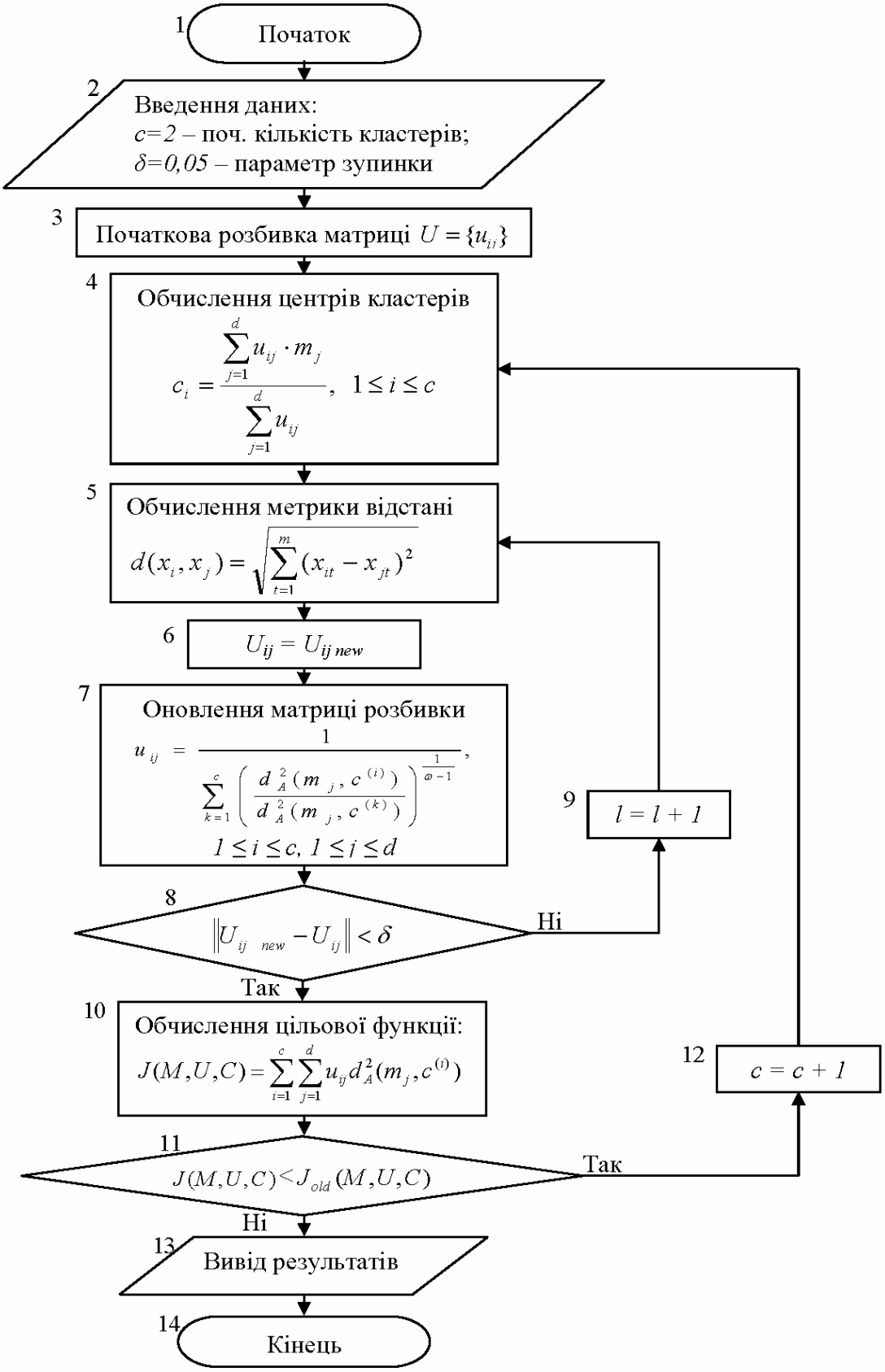


Рис. 3. Блок-схема алгоритму кластеризації

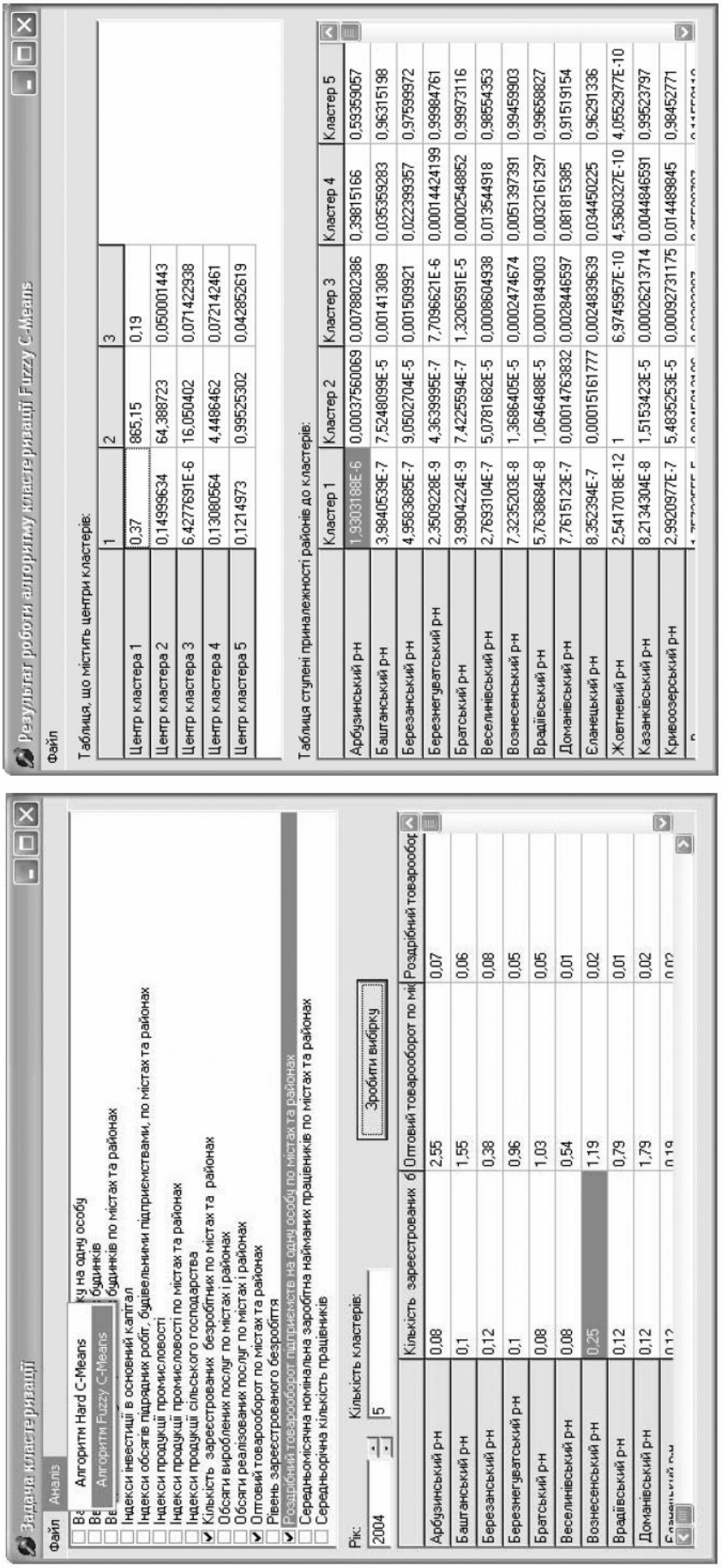


Рис. 4. Результати роботи алгоритму кластеризації

при відсутності сезонного фактора

$$\begin{aligned} b_0(t) &= \alpha * tr(t) + (1-\alpha) * \left[b_0(t-1) + b_1(t-1)\right] \\ b_1(t) &= \gamma\left[b_0(t) - b_0(t-1)\right] + (1-\gamma) * b_1(t-1) \\ tr(t+1) &= b_0(t) + b_1(t), \end{aligned} \tag{3}$$

при наявності сезонного фактора

- адитивна модель

$$\begin{aligned} b_0(t) &= \alpha * \left[X(t) - Sez(t-12)\right] + (1-\alpha) * \left[b_0(t-1) + b_1(t-1)\right] \\ b_1(t) &= \gamma\left[b_0(t) - b_0(t-1)\right] + (1-\gamma) * b_1(t-1) \\ Sez(t) &= \delta\left[X(t) - b_0(t)\right] + (1-\delta) * Sez(t-12) \\ X(t+1) &= b_0(t) + b_1(t) + Sez(t+1) \end{aligned} \tag{4}$$

- мультиплікативна модель

$$\begin{aligned} b_0(t) &= \alpha * \left[\frac{X(t)}{Sez(t-12)}\right] + (1-\alpha) * \left[b_0(t-1) + b_1(t-1)\right] \\ b_1(t) &= \gamma\left[b_0(t) - b_0(t-1)\right] + (1-\gamma) * b_1(t-1) \\ Sez(t) &= \delta\left[\frac{X(t)}{b_0(t)}\right] + (1-\delta) * Sez(t-12) \\ X(t+1) &= \frac{b_0(t) + b_1(t)}{Sez(t+1)} \end{aligned} \tag{5}$$

Коефіцієнти $0.1 < \alpha, \gamma, \delta < 0.3$, якщо усі попередні точки враховуються з однаковими вагомими коефіцієнтами. Коефіцієнти $0.3 < \alpha, \gamma, \delta < 0.5$, якщо усі найближчі попередні точки враховуються з більшими вагомими коефіцієнтами

Запропонований алгоритм аналізу та прогнозування випадкового процесу зміни соціально-економічних показників наведено на рис. 5. Алгоритм було використаю для прогнозування індексу споживчих цін і отримано добре сівпадіння прогнозних та реальних значень показників (рис. 6.)

Алгоритми пошуку асоціативних правил визначають набори об'єктів, які часто зустрічаються. Алгоритми визначають набори, що часто зустрічаються за кілька етапів. На *i*-му етапі визначаються всі *i*-елементні набори, що часто зустрічаються. Кожен етап складається з двох кроків: формування наборів кандидатів і підрахунку підтримки кандидатів. На кроці формування кандидатів *i*-го етапу алгоритм створює множину кандидатів з *i*-елементних наборів, чия підтримка поки не обчислюється. На кроці підрахунку кандидатів *i*-го етапу алгоритм сканує множину транзакцій (аналізуємих наборів), обчислюючи підтримку (Supp) наборів-кандидатів. Після сканування відкидаються кандидати, підтримка яких менше визначеного користувачем мінімуму, і зберігаються тільки *i*-елементні набори, що часто зустрічаються. Під час 1-го етапу обрана множина наборів-кандидатів містить усі 1-елементні часті набори. Алгоритм обчислює їхню підтримку під час кроку підрахунку. Блок-схема алгоритму наведено на рис. 7. Алгоритм було використано для знаходження наборів характеристик осіб, які висповідали на питання анкет (рис. 8).

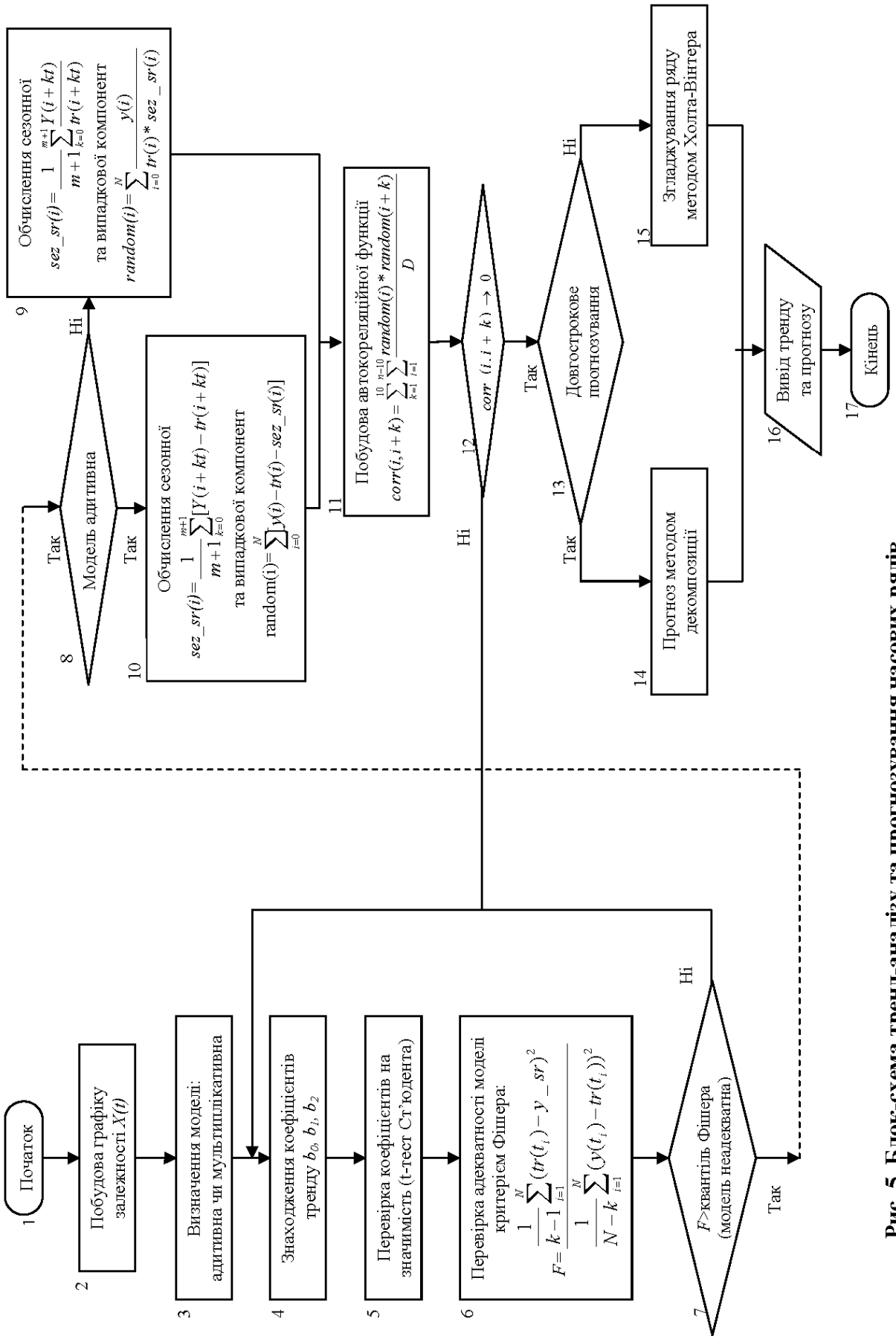


Рис. 5. Блок-схема тренд-аналізу та прогнозування часових рядів

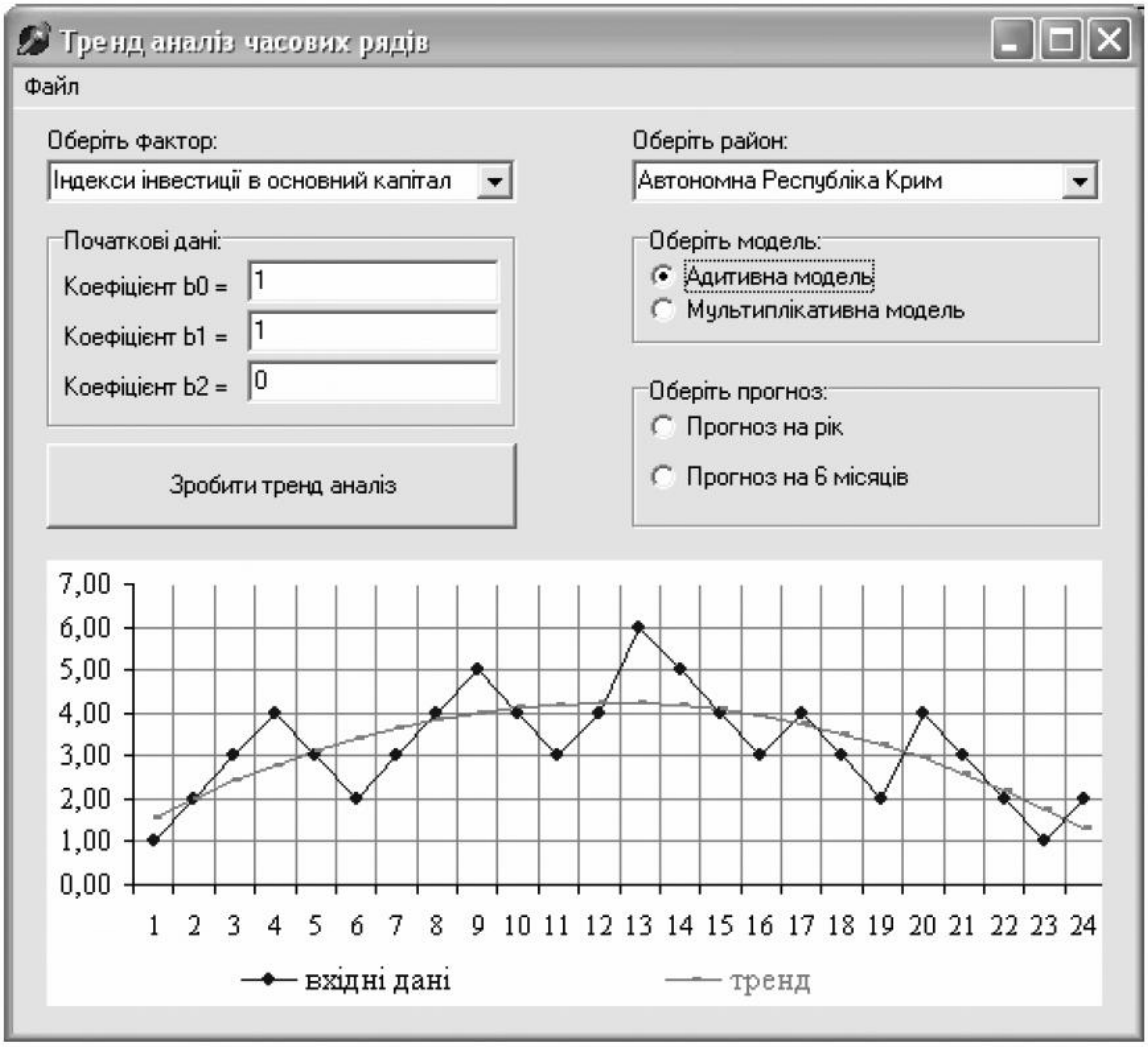


Рис.6. Результати роботи алгоритму тренд-аналізу та прогнозування часових рядів

Таким чином, алгоритми Data Mining дуже ефективно працюють при аналізі соціально-економічних показників. За їх допомогою можливо розв'язувати задачі класифікації та регресії, пошуку асоціативних груп , кластеризації та прогнозування. Для аналізу соціально-економічної інформації найбільш ефективними є таки алгоритми інтелектуального аналізу як:

- алгоритм пошуку дерев рішень для задач класифікації (алгоритм покриття);
- неієрархічний алгоритм Fuzzy C-Means для задачі кластерного аналізу;
- прогнозування часових рядів з автоматичним вибором вигляду прогнозування та врахуванням сезонної зміни показників
- алгоритм пошуку асоціативних правил.

Запропоновані алгоритм кластеризації Fuzzy C-Means з визначенням як центрів кластерів, так і числа кластері та алгоритм прогнозування часових рядів з вибором вигляду прогнозування з врахуванням сезонної зміни показників при автоматизованому визначенні моделі (мультиплікативна або адитивна), які показали також високу ефективність.

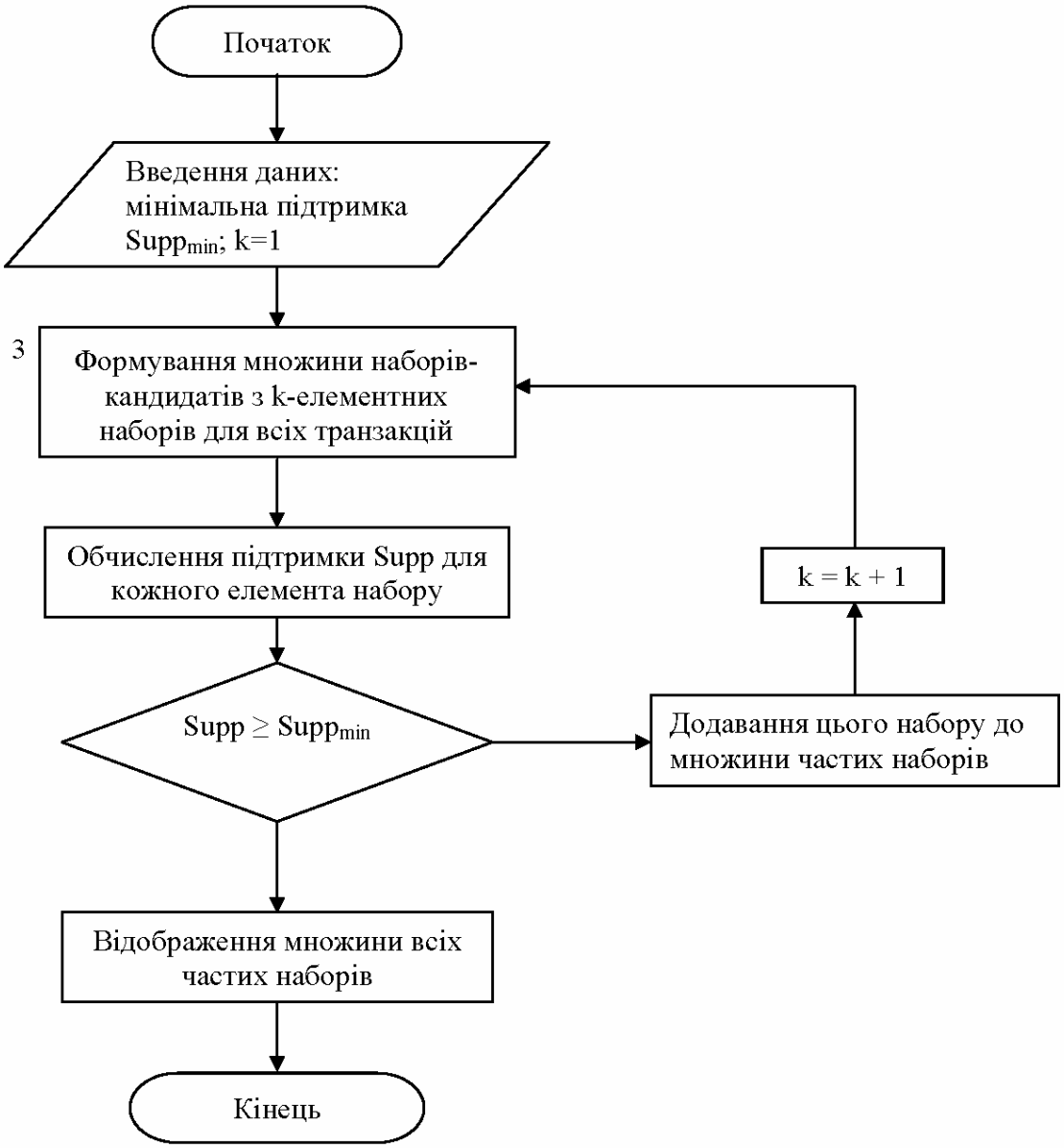


Рис. 7. Блок-схема алгоритму пошуку набору даних

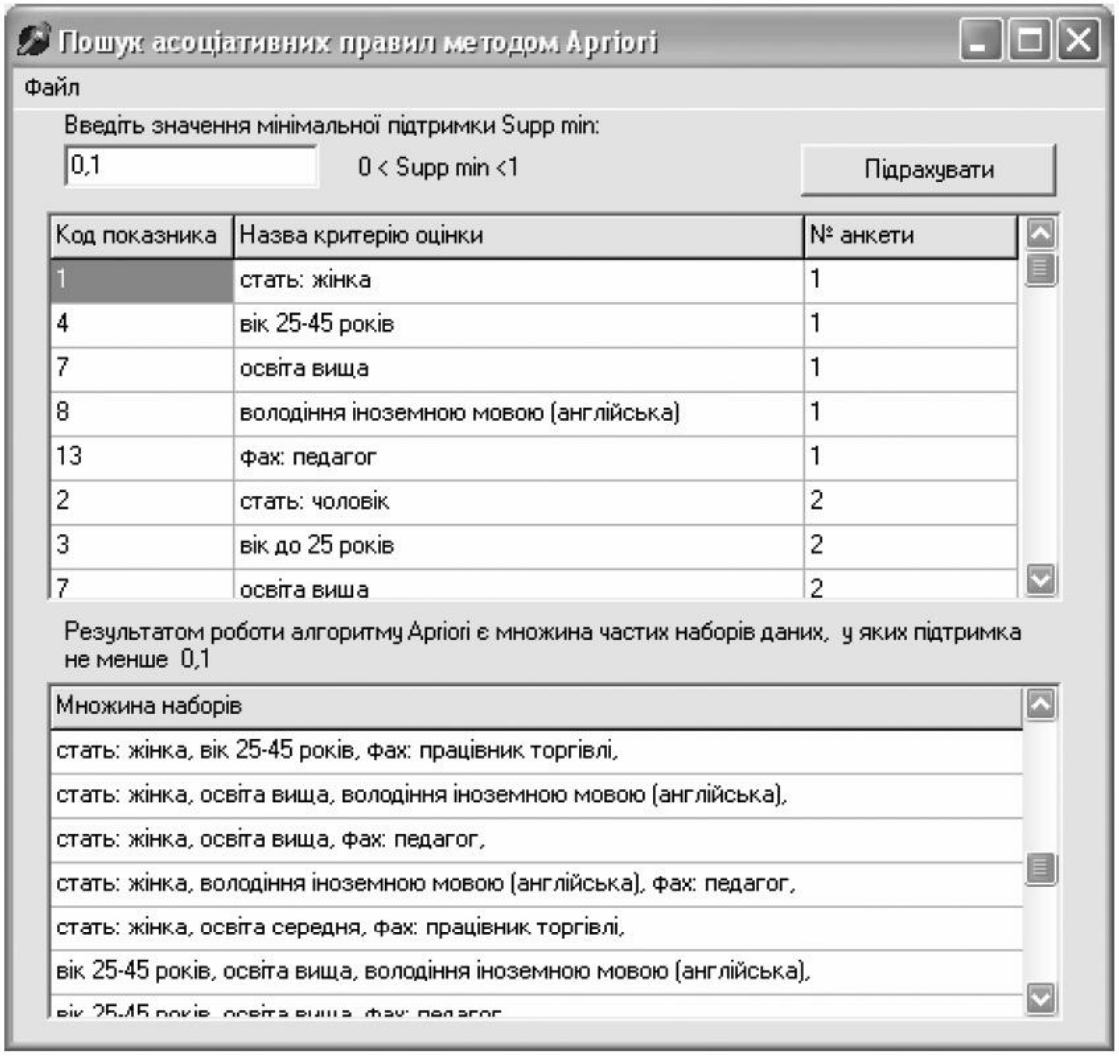


Рис. 8. Результати роботи алгоритму тренд пошуку асоціативних правил (алгоритм Apriori)

Література

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining.– СПб.: БХВ-Петербург, 2004.– 336 с.: ил.
2. Жамбю М. Иерархический кластерный анализ и соответствия. - М.: Финансы и статистика, 1988. – 236 с.
3. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. М.: Наука, 1976. – 736 с.
4. Кравець І.О., Ромакін В.В. Статистичний аналіз даних з використанням статистичних пакетів та MS Excel: Навч. посібник/ Миколаїв: МДГУ, 2002. – 58 с.
5. Лук'яненко І.Г., Краснікова Л.І. Економетрика/ Практикум з використанням комп'ютера. – К.: Товариство "Знання", ККО, 1998. – 231 с.
6. Статистические методы анализа информации в социологических исследованиях. М., 1979, Наука, 194 с.
7. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М: ИНФРА-М. Финансы и статистика, 1995. 384 с.
8. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА-М, 1998. – 528 с.