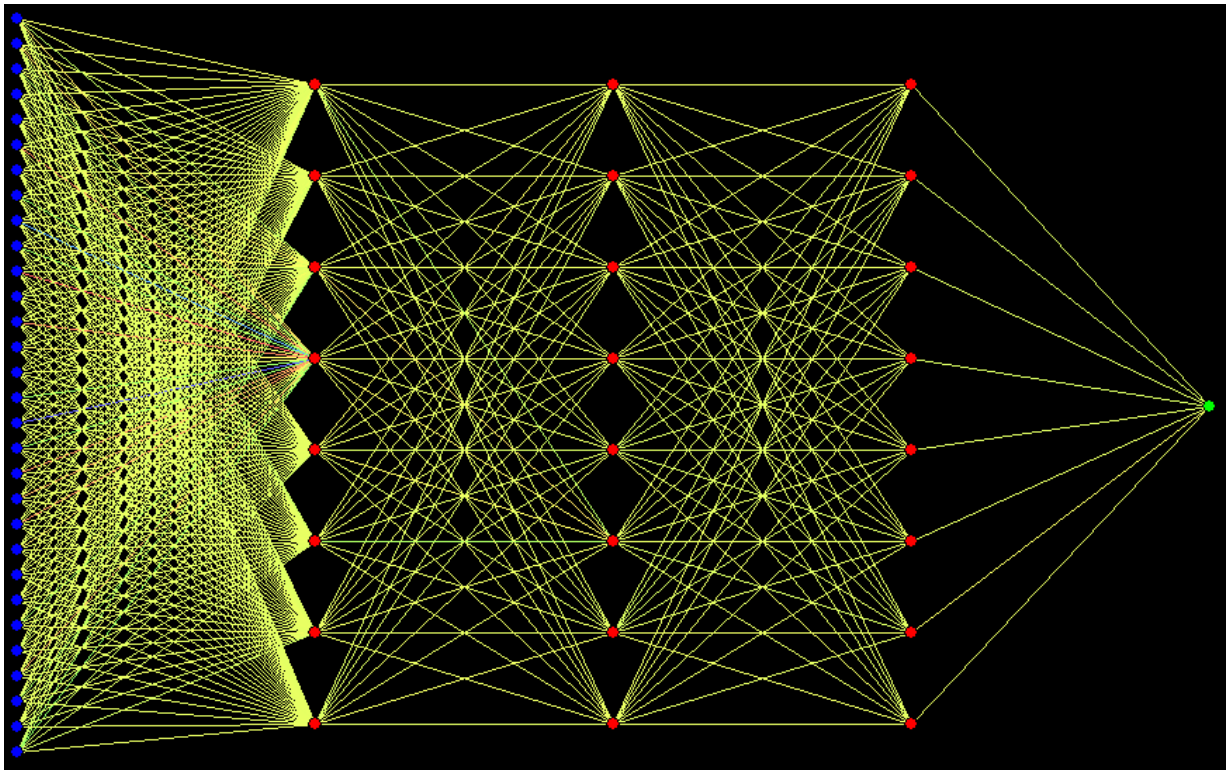
Акіменко В.В.

ПРИКЛАДНІ ЗАДАЧІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ (DATA MINING)



Київський національний університет
Імені Тараса Шевченка

Акіменко В.В.

ПРИКЛАДНІ ЗАДАЧІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ (DATA MINING)

Київ 2018

УДК 004.451.53

ББК 32.973.26.-018.2

Рекомендовано до друку Вченою Радою факультету комп'ютерних наук та кібернетики Київського національного університету ім. Тараса Шевченка (протокол № 13 від 29 травня 2018 р.)

Рецензенти:
д.ф.м.-н., професор Наконечний О.Г., д.ф.м.-н., професор Глибовець М.М.

Акіменко В.В.
Прикладні задачі інтелектуального аналізу даних (DATA MINING). – К.: КНУ ім. Тараса Шевченко, 2018. – 152 с.

В посібнику розглянуті теоретичні та практичні питання з методів розв'язування різного типу прикладних задач що відносяться до інтелектуального аналізу даних. Особливість викладення матеріалу пов'язана з його практичною спрямованістю – основу посібника складають розділи із покроковим розв'язуванням різних типів прикладних задач аналізу даних наборами різних методів. Приклади розв'язування задач розглянуті за допомогою аналітичної програмної платформи Deductor Studio. Посібник розраховано на студентів спеціальностей "Системний аналіз", "Прикладна математика", "Комп'ютерні науки".

ЗМІСТ

ВСТУП

В основу даного посібника покладено необхідний мінімум теоретичного та практичного матеріалу, який використовується для вивчення методів інтелектуального аналізу даних (Data mining). Матеріал посібника орієнтований перш за все на вивчення методів та дослідження властивостей різних підходів та алгоритмів щодо розв'язування основного набору прикладних задач Data mining.

Даний посібник складається з набору тем в яких розглядається певний клас прикладних задач, необхідний мінімум теоретичного матеріалу стосовно методів та алгоритмів розв'язування даних задач а також надається покроковий розв'язок кожної задачі за допомогою аналітичної платформи із детальними поясненнями. Основу матеріалу складають лабораторні роботи, кожна з яких змістовно складається з трьох основних частин. Перша частина містить теоретичний матеріал, який є основою, базисом для побудови алгоритму розв'язку за допомогою даного методу. Теоретичний матеріал містить головні ідеї методу, опис моделі, основні математичні формули та рівняння для його реалізації. Друга частина матеріалу містить покроковий приклад реалізації методу для розв'язування на комп'ютері достатньо простих, невеликих за обсягом учбових задач. Користувачу надається детальний матеріал із скріншотами, коментарями та поясненнями що до його використання. Третя частина матеріалу містить завдання для виконання лабораторної роботи. Основна спрямованість посібника полягає у виборі сучасних кібернетичних методів та дослідженні їх властивостей при розв'язуванні набору прикладних задач, які найбільш часто розглядаються у Data mining.

Для опанування матеріалу посібника користувачу необхідні знання з основ програмування, баз даних, математичного аналізу, лінійної алгебри, чисельних методів, нейронних мереж, баз знань, основ штучного інтелекту, основ теорії ймовірності та математичної статистики. У посібнику використовується безкоштовна академічна версія аналітичної платформи Deductor Studio 5.3 (2013 р.).

Даний матеріал використовується в учбовому процесі на факультеті комп'ютерних наук та кібернетики Київського національного університету ім. Тараса Шевченко.

Тема 1. Імпорт даних різного формату та парціальна предобробка даних.

Задачі: Завантаження даних різного формату. Візуалізація даних. Комплексна попередня обробка даних: відновлення пропущених даних, виявлення та видалення дублікатів та протиріч.

Методи: Інтерполяція даних, графічне представлення даних та їх первинних статистичних покажчиків, відновлення пропущених даних, виявлення та видалення дублікатів та протиріч.

1. Теоретичні основи роботи з масивами даних у Deductor Studio

Майстер імпорту дає можливість в покроковому режимі вибрати тип джерела даних і налаштувати відповідні параметри. Для виклику майстра імпортування можна скористатися кнопкою "Майстер імпорту" на панелі інструментів "Сценарії" або вибрати відповідну команду з контекстного меню. У середовищі Deductor Studio існують можливості імпорту файлів даних з різних джерел, а саме [1]:

1. Сховища даних:
   - Virtual Warehouse – імпорт даних з Virual Warehouse.
   - Deductor Warehouse – імпорт даних Deductor Warehouse.
   - Microsoft AS 2005 – імпорт даних з MS Analysis Services 2005.
2. Бізнес-додатки:
   - MS Dynamics CRM - імпорт даних з CRM-системи MS Dynamics.
   - Terrasoft CRM - імпорт даних з CRM-системи Terrasoft.
3. Налаштовані підключення:
   - База даних – імпорт даних з баз даних різних типів.
   - Універсальний драйвер – імпорт даних через підключення до універсального драйвера.
   - Файлова база даних – імпорт даних з файлової бази даних.

Deductor Studio підтримує імпорт даних із файлів наступних форматів:

- Deductor Data файл (Direct) – це власний формат файлу даних. Він є самим швидким джерелом і приймачем даних при виконанні імпорту або експорту.
- Текстовий файл (Direct) – текстовий файл у форматі, в якому стовпці даних розділені однотипними символами-роздільниками.
- DBF файл (Direct) – прямий доступ до файлів плоских баз даних типу DBF, який підтримується такими програмами, як dBase, FoxBase, FoxPro.
- XML – це ієрархічна структура, призначена для зберігання будь-яких даних.
- MS Excel (OLE) – книга Microsoft Excel (*.xlsx, *.xlsxx, *.xlsxb).
- HTML - сторінка (OLE) – HTML - сторінка з таблицями.

Число кроків майстра імпорту, а також його налаштування відрізняється для різних типів джерел та форматів. На кожному кроці майстра імпорту можна використовувати кнопки "Далі" і "Назад", які відповідно дозволяють перейти до наступного кроку або повернутися на попередній для редагування параметрів.

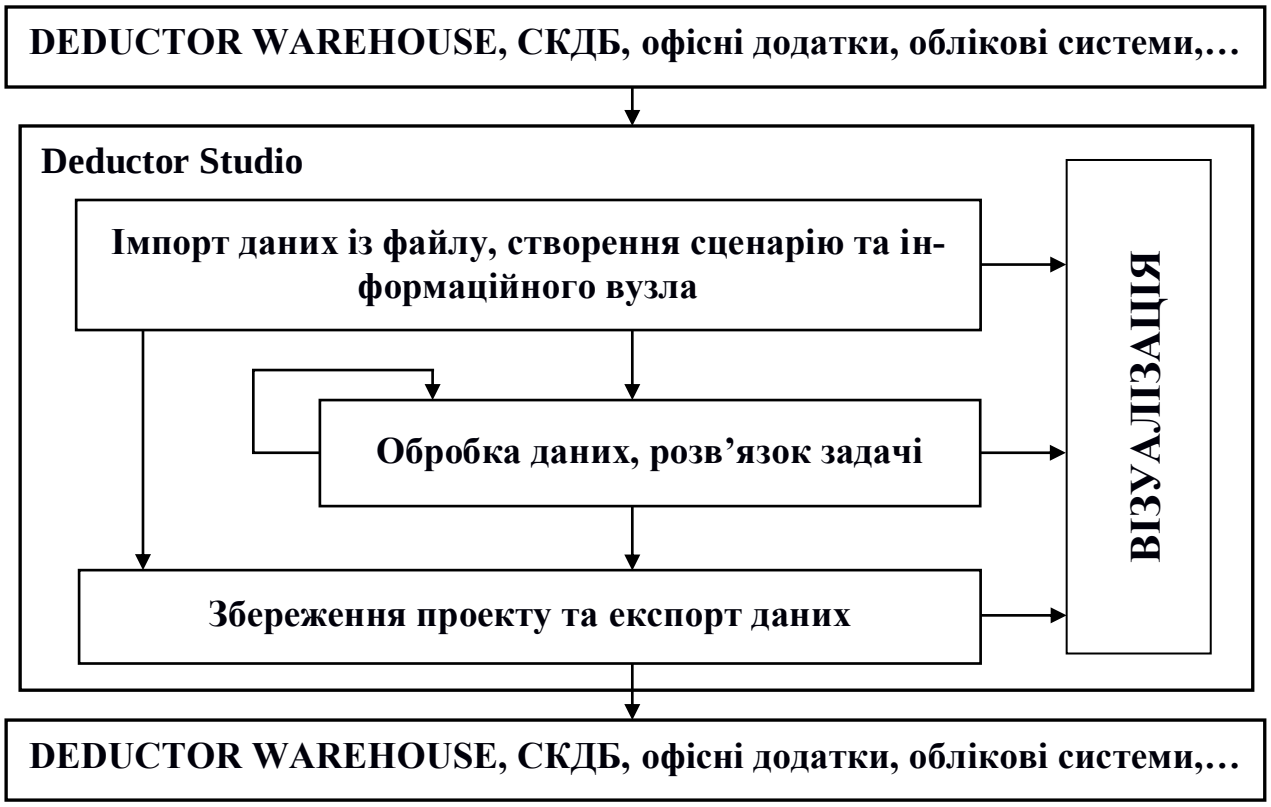


Рис.1. Схема функціонування Deductor Studio [1].

На Рис.1. показана загальна схема функціонування Deductor Studio із [1]. Алгоритм починається з процедури імпорту даних, створення сценарію та інформаційного вузла для цих даних. Імпортовані з зовнішніх джерел данні обробляються далі процедурами (методами) Deductor Studio які необхідно вибрати у вікні майстра обробки. Результатом обробки даних є також масив даних, який в свою чергу, може оброблятися далі. Результати виконання кожної дії можна візуалізувати різними способами. Спосіб можливих відображень залежить від обраного методу обробки даних. Імпортований набір даних, а також дані, отримані на кожному етапі обробки, можуть бути експортовані у зовнішні файли.

2. Графічне представлення даних (методи візуалізації)

Deductor Studio містить багато різних спеціалізованих засобів візуалізації даних для всіх методів обробки даних, такі як таблиці, графіки, діаграми або гістограми, тощо. Розглянемо більш детально вбудовані засоби візуалізації даних.

Таблиця. Стандартне табличне подання з можливістю фільтрації даних, сортування і швидкого розрахунку статистики (он-лайн статистика). Використовуємо файл з даними task11.txt.

Відкриваємо "Мастер импорта" за допомогою кнопки на панелі інструментів (Рис.2). У відкритому вікні обираємо формат файлу даних "Text" та "Далее" (Рис.3). У новому вікні вказуємо повне ім'я текстового файлу що містить таблицю значень функції (Рис.4). Оскільки текстовий файл побудовано на основі шаблону, залишаймо усі значення за замовчуванням (Рис.5). Натискаємо "Далее".

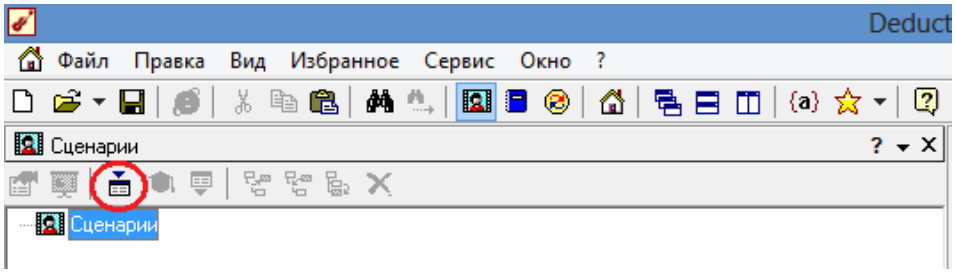


Рис.2. Кнопка відкриття меню «Мастер импорта».

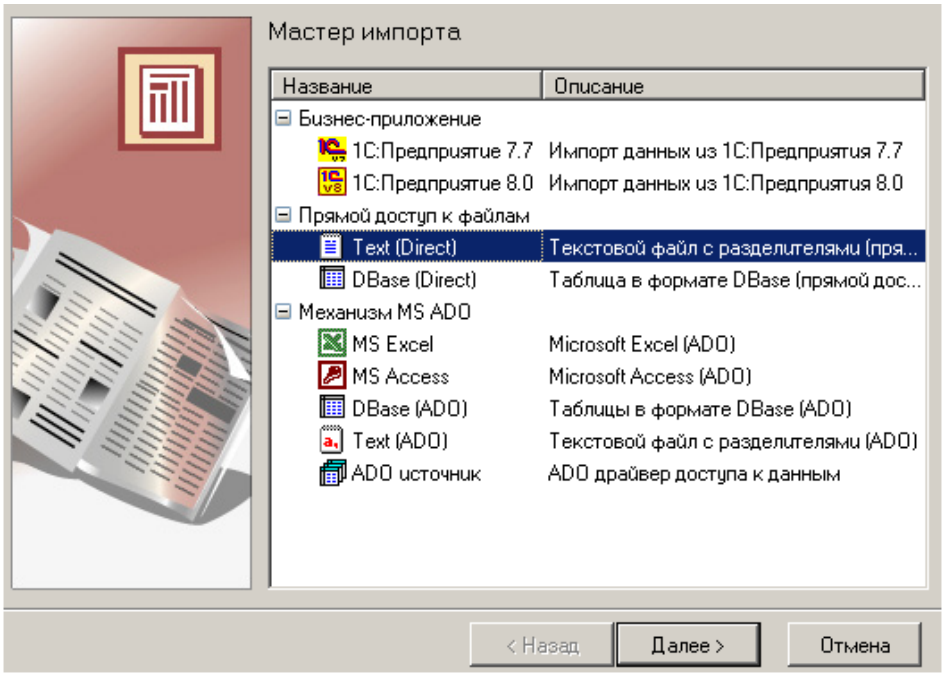


Рис.3. Меню «Мастер импорта».

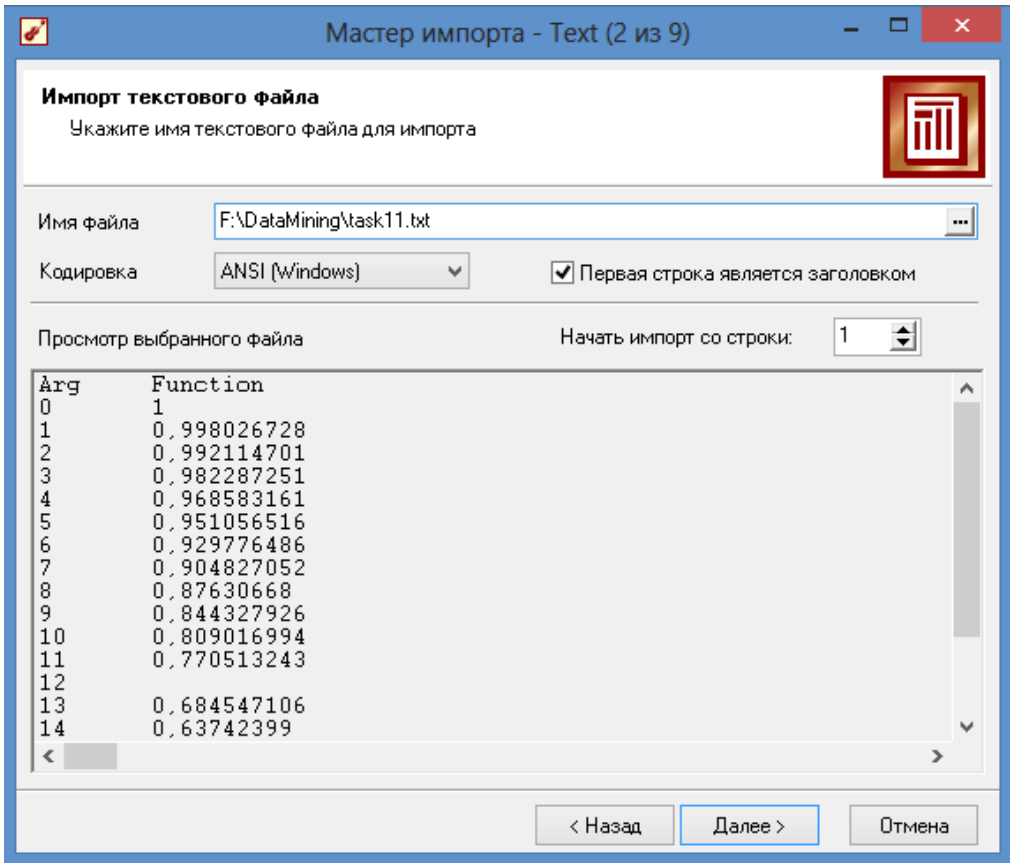


Рис.4. Початковий перегляд даних файлу.

У наступному вікні відображається інформація про імпортовані дані з файлу. Зліва знаходяться ім’я змінних, праворуч — інформація о зчитаних змінних. Хоча Deductor автоматично знаходить тип даних, рекомендується перевірити кожну змінну на наявність помилок. Задаємо кожну змінну як дійсну та неперервну. Натискаємо “Далее”. Імена змінних (Рис.6, 7):

- Arg — аргумент функції;
- Function — функція косинуса у точках arg із пропущеними значенями;

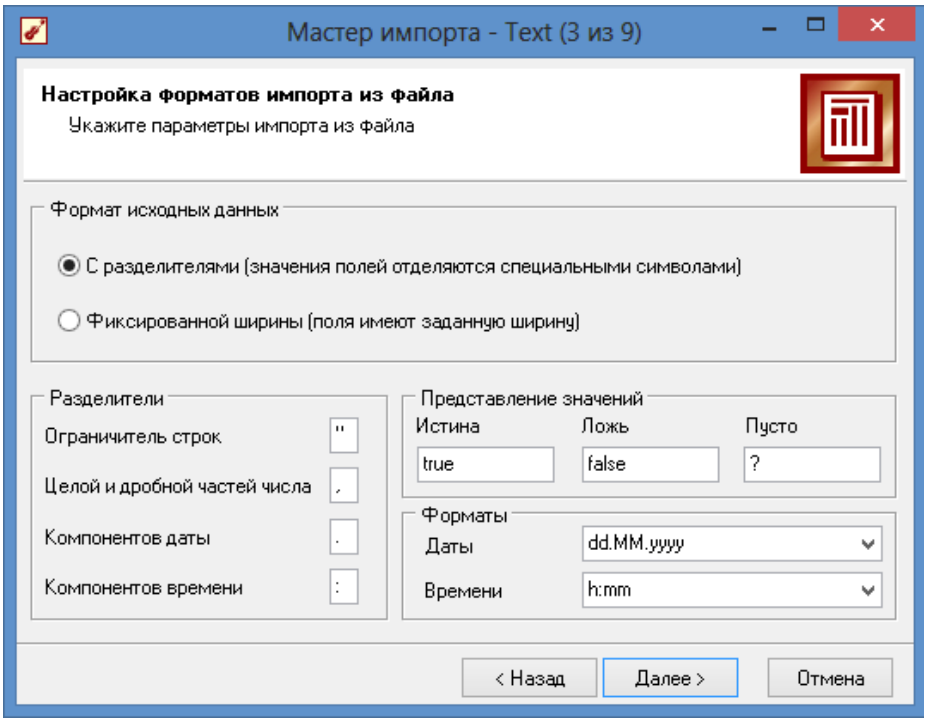


Рис.5. Вікно налаштування імпорту

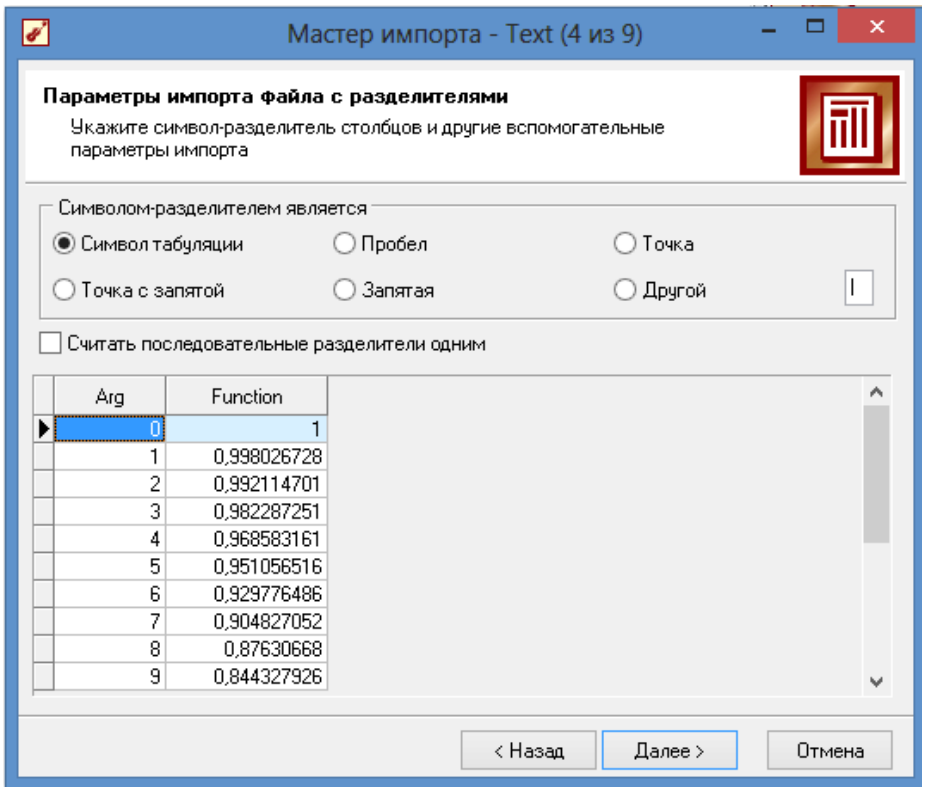


Рис.6. Знайдені змінні аргументу та функції.

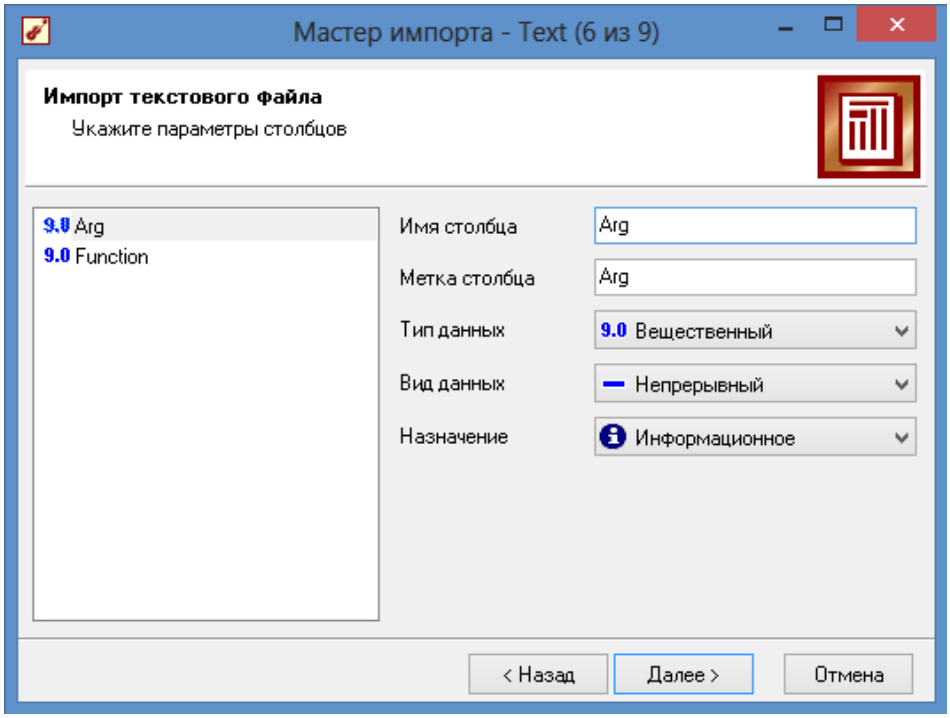


Рис.7. Імена імпортованих змінних та їх атрибути.

У наступному вікні (Рис.8) натискаємо кнопку “Пуск” та чекаємо закінчення імпорту. Натискаємо “Далее”.

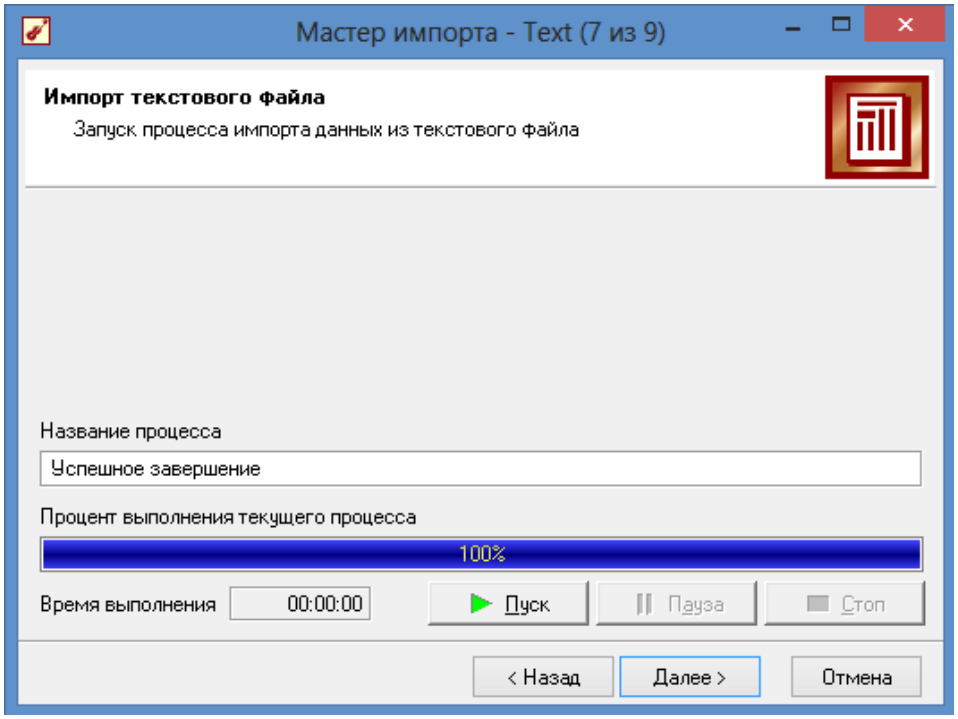


Рис.8. Вікно процесу імпортування даних.

У наступному вікні необхідно обрати тип візуалізації даних. У даному випадку демонстративного приклада ми обираємо усі типи візуалізації крім OLAP-кубу.

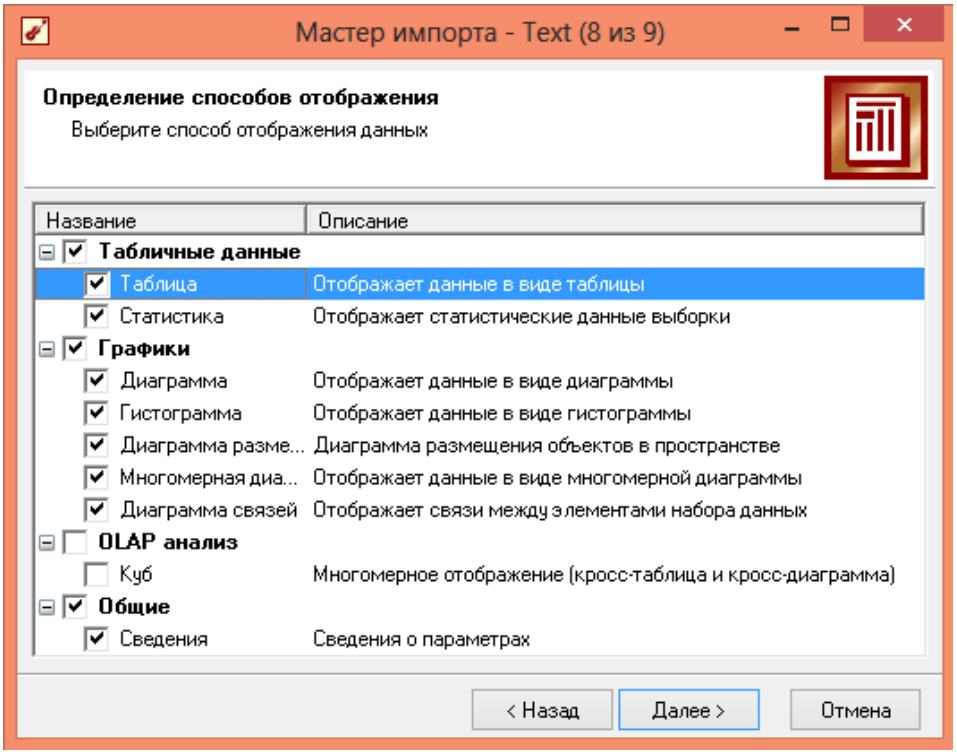


Рис.9. Вікно вибору типу візуалізації.

У наступному вікні відображені налаштування діаграми. Можна задавати тип діаграми, кольори графіків функцій, значення та підписи по осі X. Виділяємо Function та переходимо далі.

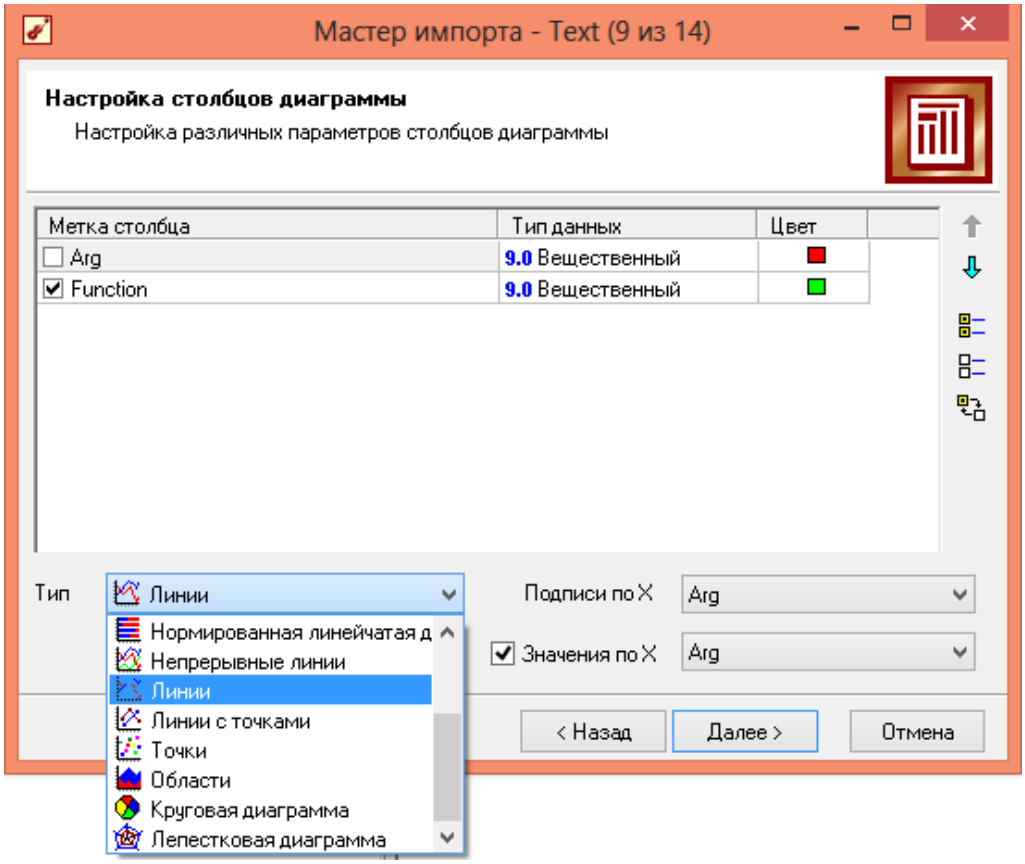


Рис.10. Вікно налаштування діаграми.

У наступних вікнах (Рис.11 - 14) налаштовуємо гістограму, діаграму розміщення, багатовимірну діаграму, діаграму зв’язків та переходимо далі.

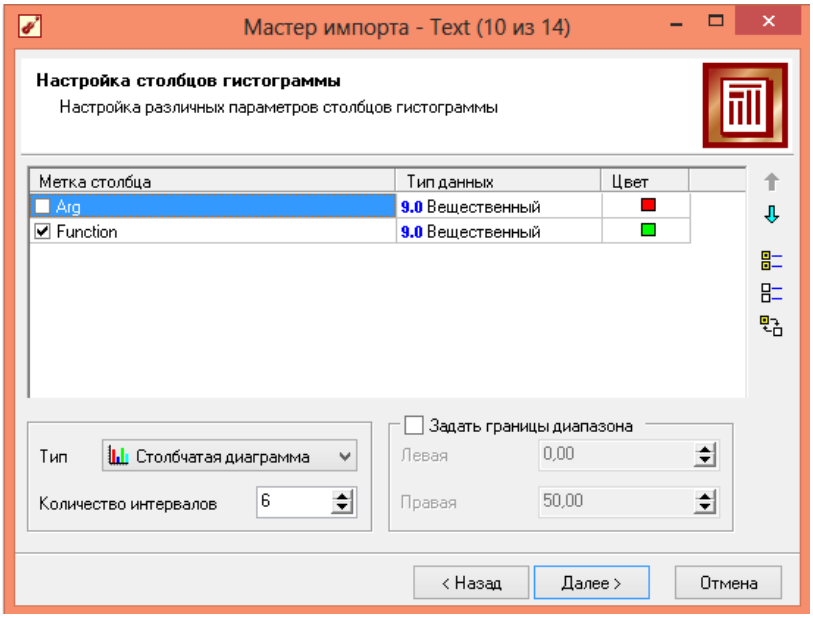


Рис.11. Вікно налаштування гістограми.

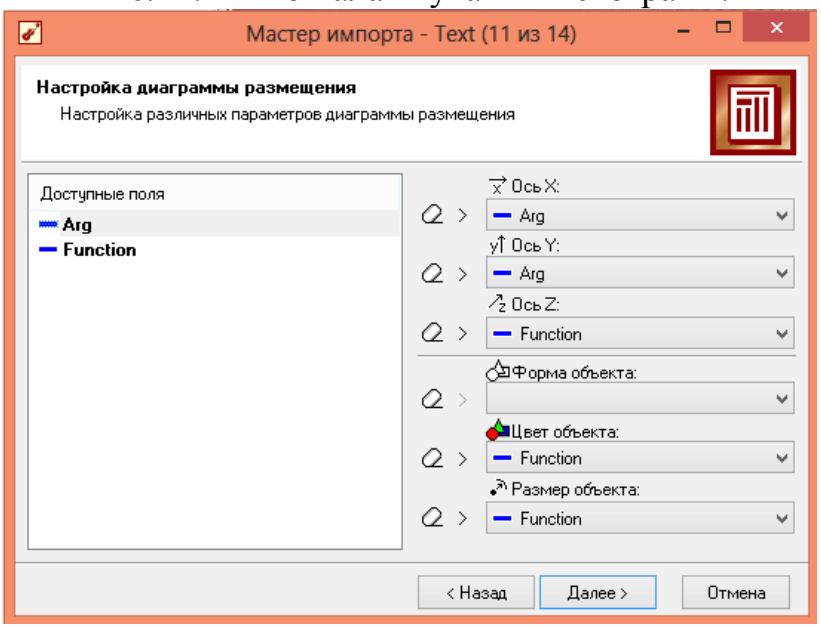


Рис.12. Вікно налаштування діаграми розміщення.

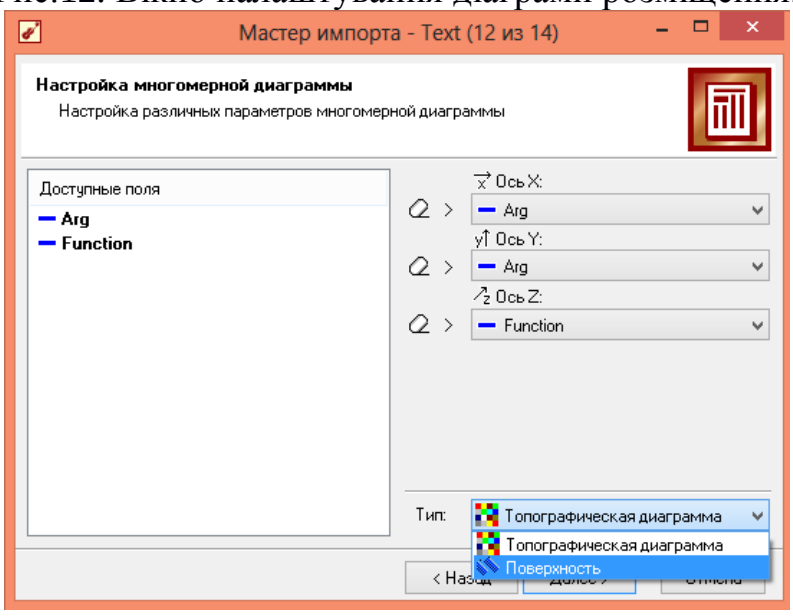


Рис.13. Вікно багатовимірної діаграми (поверхня або топографічна діаграма).

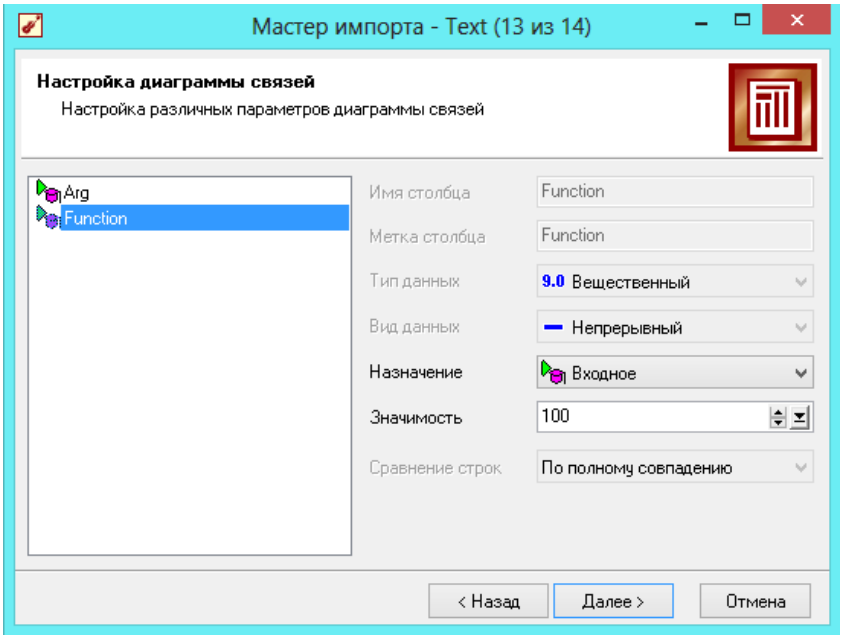


Рис.14. Вікно налаштування діаграми зв’язків.

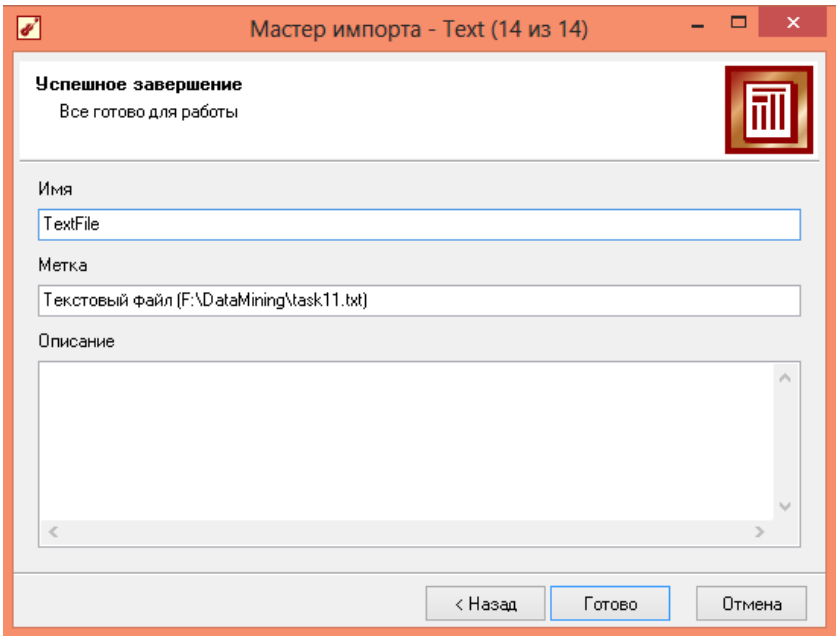
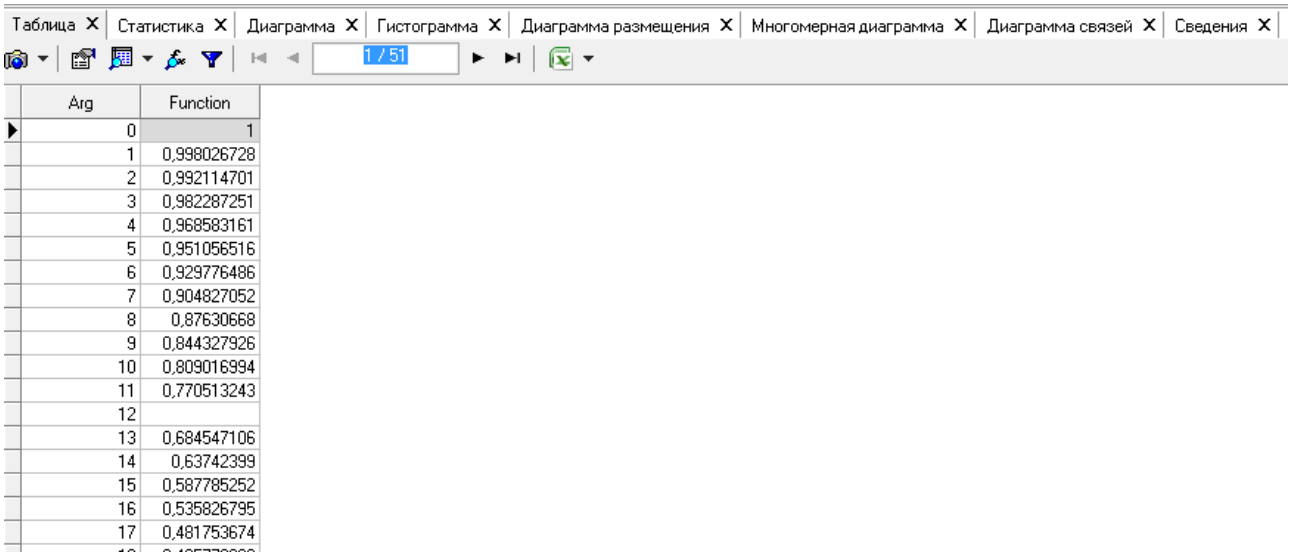


Рис.15. Повідомлення про успішне завершення налаштувань візуалізації даних.



Arg	Function
0	1
1	0,998026728
2	0,992114701
3	0,982287251
4	0,968583161
5	0,951056516
6	0,929776486
7	0,904827052
8	0,87630668
9	0,844327926
10	0,809016994
11	0,770513243
12	
13	0,684547106
14	0,63742399
15	0,587785252
16	0,535826795
17	0,481753674

Рис.16. Таблиця даних.

Тепер у вікні що розташовано ліворуч, в дереві сценаріїв з'явиться новий інформаційний вузол з необхідними даними. У головному вікні зверху представлені вкладки усіх обраних типів візуалізації даних.

Статистика. Статистичні покажчики вибірки.

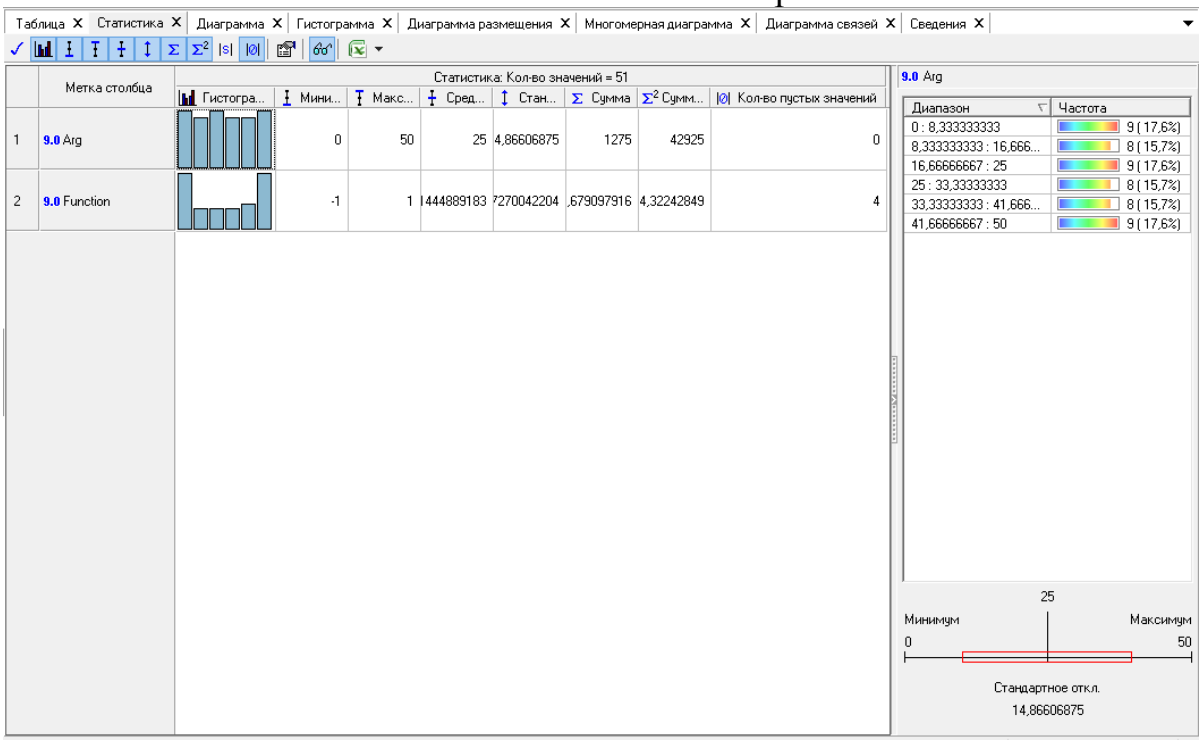


Рис.17. Статистичні покажчики вибірки.

Діаграма. Графік зміни будь-якого показника. Є можливість вибору різних варіантів діаграм: стовпчасті, лінійні, кругові та інше.

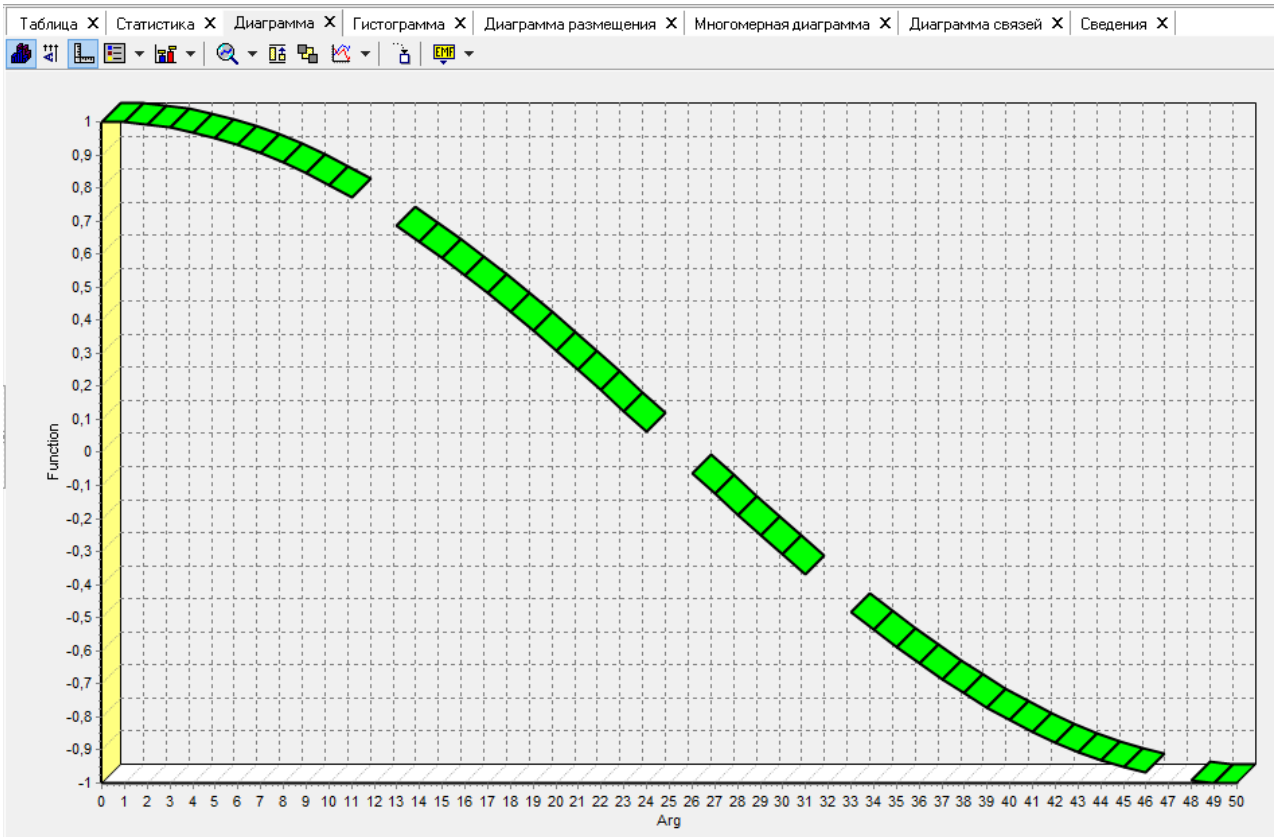


Рис.18. Графік функції побудований по табличних значеннях (діаграма).

Багатовимірна діаграма. Відображає дані в багатовимірному вигляді – поверхня або топографічним способом.

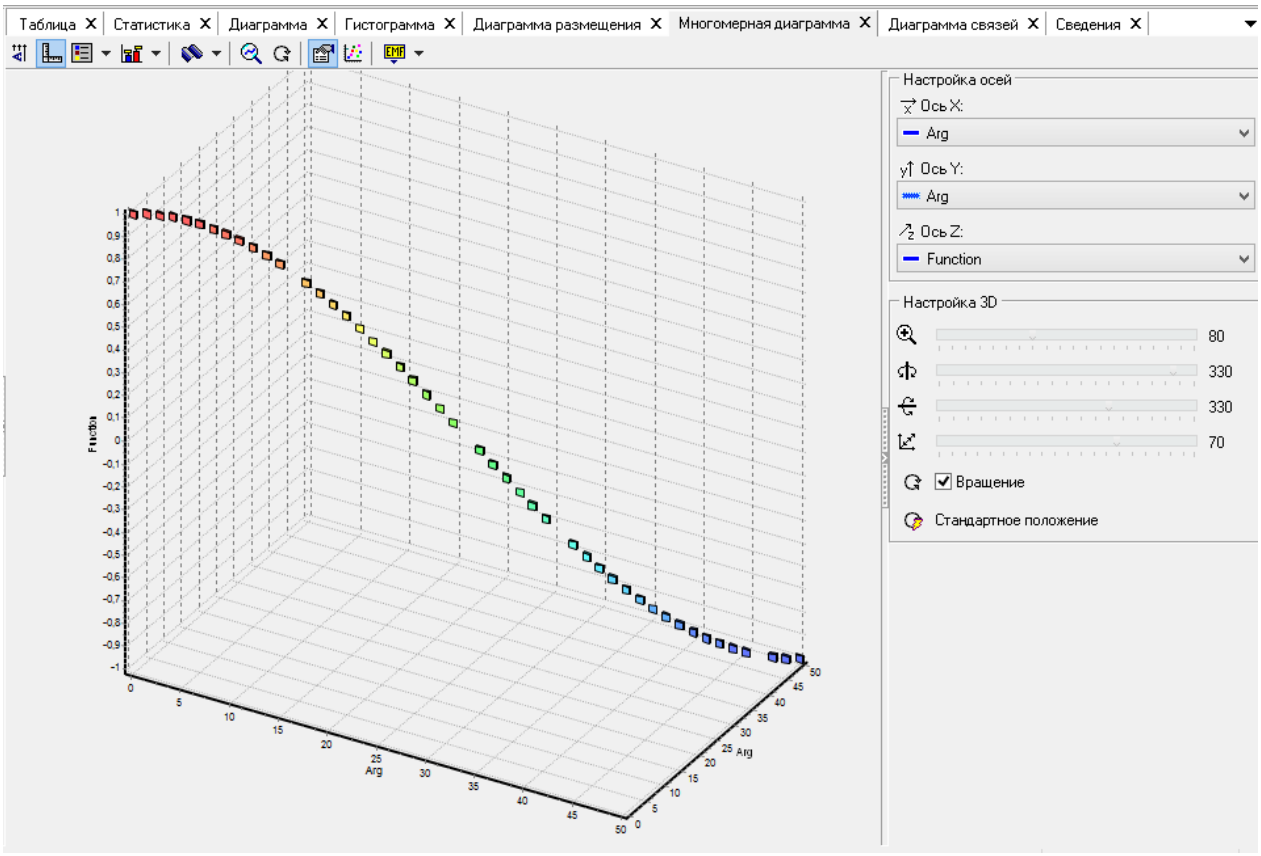


Рис.19. Відображення багатовимірної діаграми, режим поверхня.

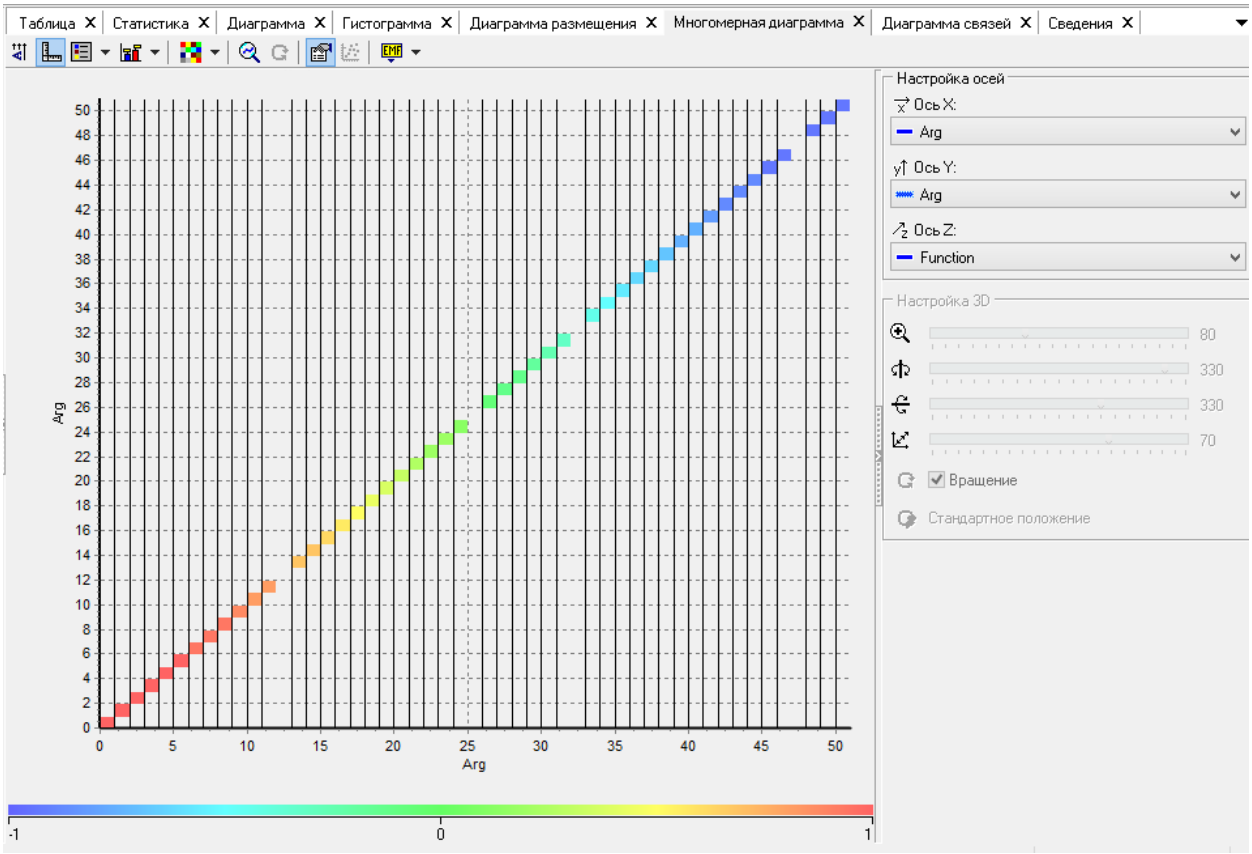


Рис.20. Відображення багатовимірної діаграми, режим топографічна діаграма (значення функції відображено кольором згідно легенді).

Діаграма розміщення. Діаграма показує багатовимірні дані у різних тривимірних зрізах, що розміщені у просторі параметрів даних.

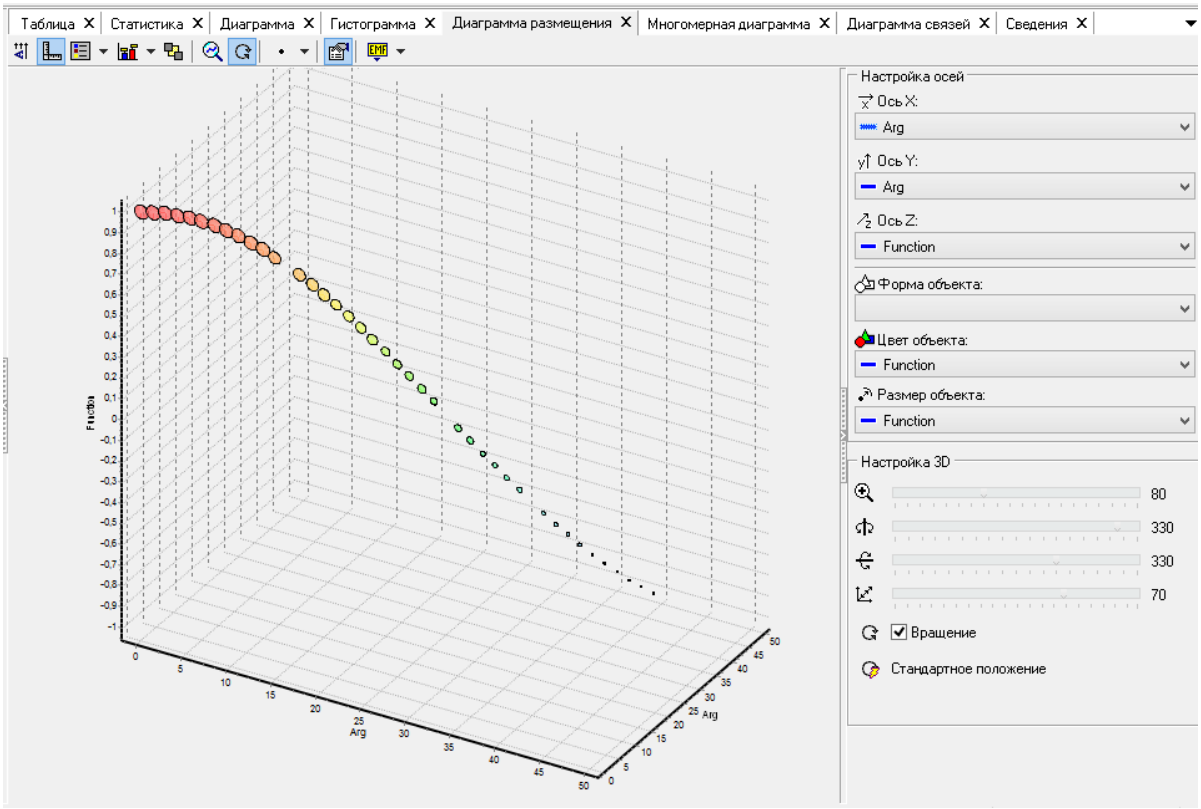


Рис.21. Діаграма розміщення для багатовимірного масиву даних.

Діаграма зв’язків. Візуалізація об'єктів та зв'язків між ними.

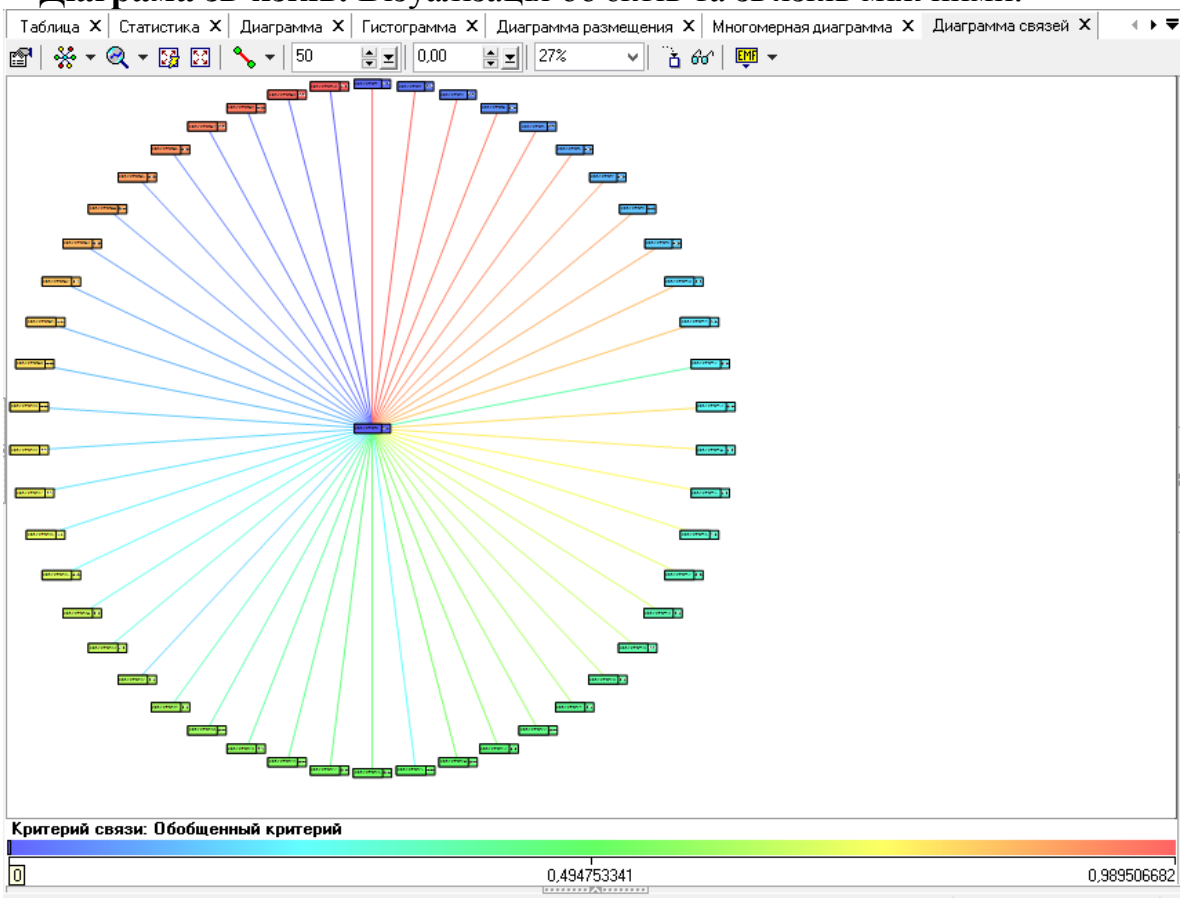


Рис.22. Діаграма зв’язків між усіма 50 значеннями масиву даних.

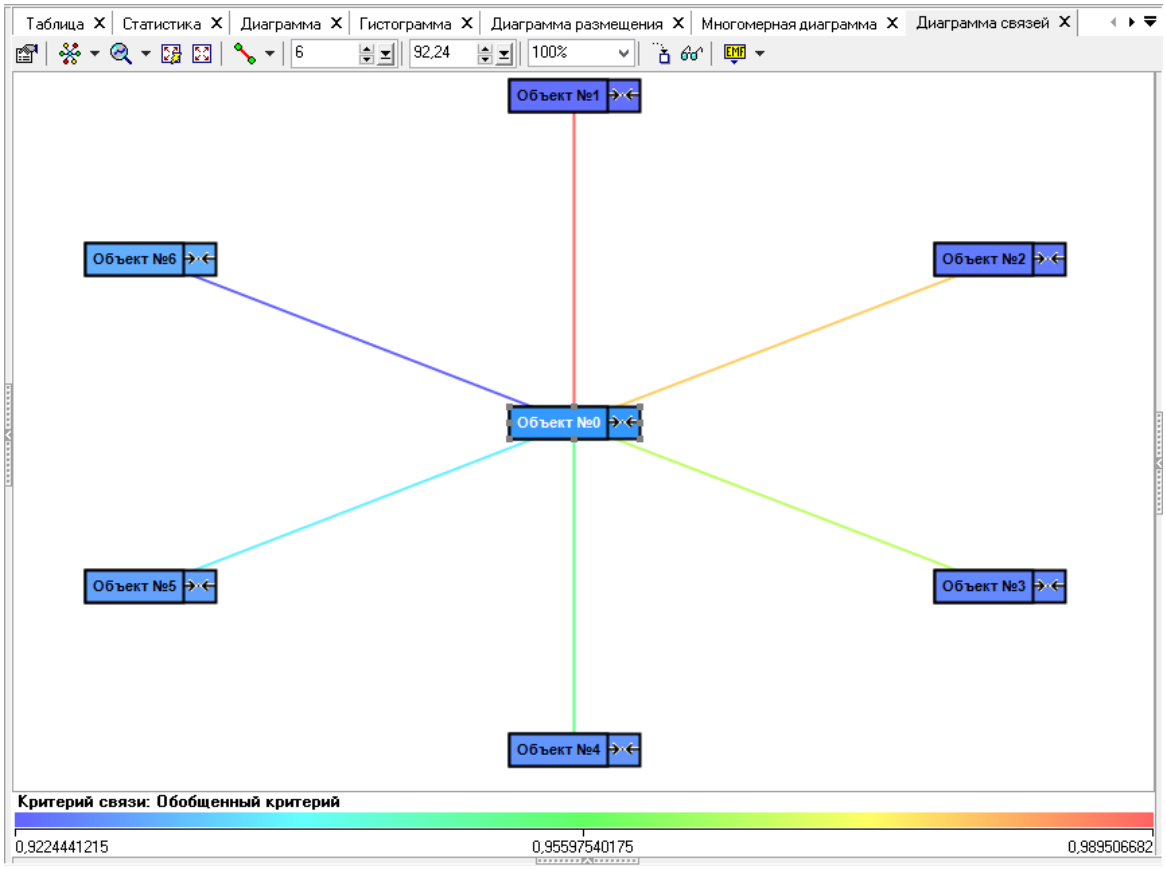


Рис.23. Діаграма зв’язків між 6 першими значеннями масиву даних.

Гістограма. Графік розкиду показників. Гістограма призначена для візуальної оцінки моделі розподілу даних.

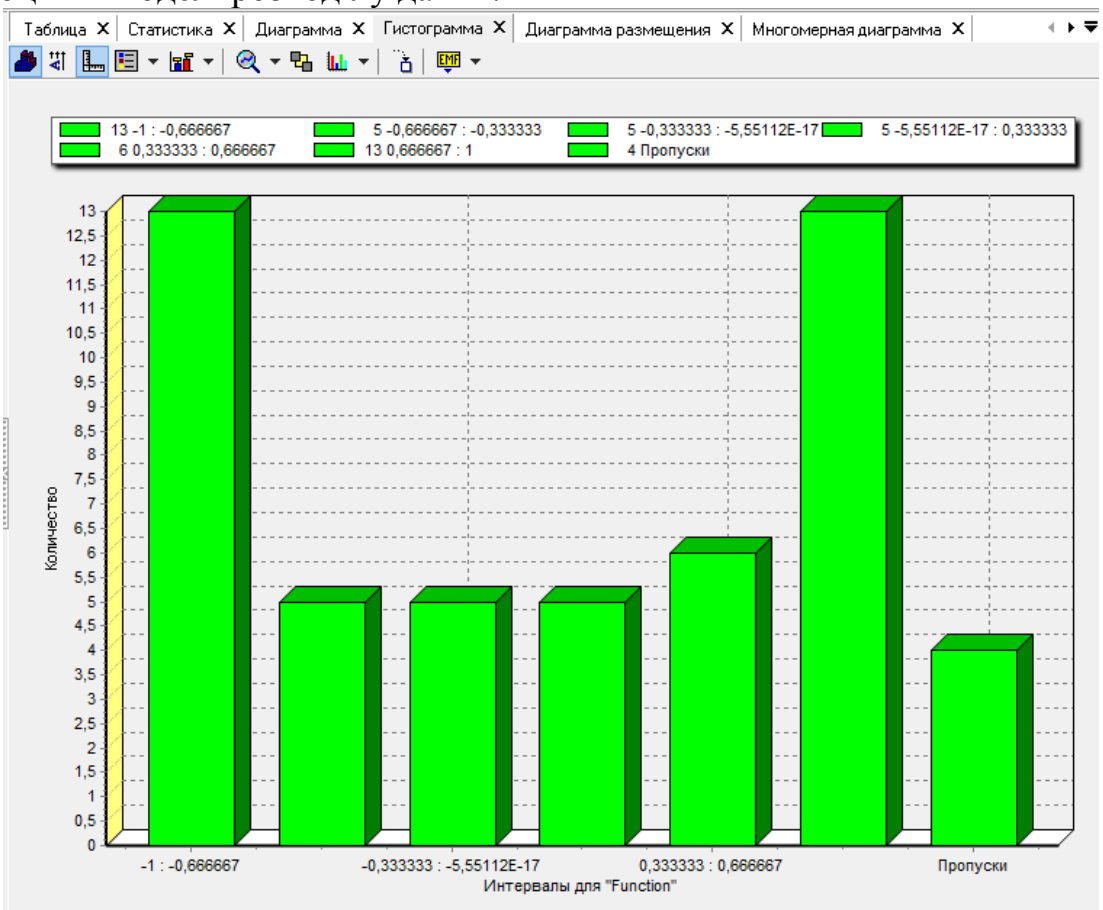


Рис.24. Об’ємна гістограма даних.

Усі засоби візуалізації дозволяють експортувати результати у таблиці (MS Excel, MS Word, HTML, текстові файли, тощо) та в графічні файли (формати GIF, BMP, EMF, тощо).

3. Інтерполяція даних

Іноді, в силу певних причин, в зв'язаних масивах даних деякі дані можуть бути відсутні (наприклад, дані невідомі, або відсутні за рахунок технічного збою). Видалення усіх рядків, що містять пропущені дані не завжди є способом вирішення проблеми, так як втрачається інформація по заповненим клітинам, або в результаті видалення даних інформації для аналізу може залишитися занадто мало. Одним із способів поновлення зв'язаних даних (часових рядків, дискретних значень неперервної функції, тощо) є інтерполяція функцій за допомогою різних поліномів [2], [16]. Найбільш поширеним класом інтерполяційних функцій є сплайни. Сплайн-функція - це кусково-поліноміальна функція, область визначення якої розбита на скінчене число відрізків, на кожному з яких вона співпадає з деяким поліномом. Слово сплайн (spline, англ.) означає гнучке лекало, спеціальну лінійку для креслення гладких кривих ліній. Основною перевагою сплайнів є побудова достатньо гладкої інтерполяційної функції що не містить великої кількості осциляцій на відміну від інтерполяційних поліномів Лагранжа високого порядку. Поліном Лагранжа n-го порядку (n- кількість точок інтерполяції які з'єднує поліном, $n \geq 3$) може містити $(n-1)$ точку локального екстремуму в наслідок коливань на інтерполяційних відрізках. При досить великих значеннях n коливання інтерполяційного полінома приводять до його значних відхилень від первинної функції та, як наслідок, до великої похибки інтерполяції. Такий ефект називають феноменом Рунге.

Тому, більш практичним у багатьох випадках є застосування сплайн-інтерполяції, при якій апроксимуюча функція складається із окремих поліномів однакового порядку – сплайнів (як правило, поліномів третього порядку – кубічна сплайн-інтерполяція). Кожен сплайн визначений на своєму відрізку та гладко з'єднаний з іншими у точках інтерполяції таким чином що інтерполяційна функція є неперервно-диференційованою на всій множині визначення. Для обчислення інтерполяційних значень в Deductor Studio передбачено спеціальний вузол "Заповнення пропусків", який може читати налаштування, зроблені у вузлі "Якість даних", або бути самостійним зі своїми параметрами. Метод інтерполяції доступний тільки для впорядкованих даних (наприклад, часові ряди). Відновлення пропусків у стовпці, значення яких упорядковані, можна розглядати як інтерполяцію значень функції в точках, де вона невизначена. Ця задача реалізується за допомогою А-сплайнів. Розглянемо приклад.

Задано текстовий файл task11.txt з таблицею аргументу та функцією $y = \cos(t)$ із пропущеними значеннями. Необхідно імпортувати данні у Deductor та визначити пропущені значення функції. Після імпорту даних отримуємо графік функції з пропущеними даними, зображений на Рис.18.

Для отриманого вузла завантажуємо вікно "Мастер обработки" за допомогою кнопки  на панелі інструментів (Рис.25, 26). Обираємо метод "Заполнение пропущенных даннях".

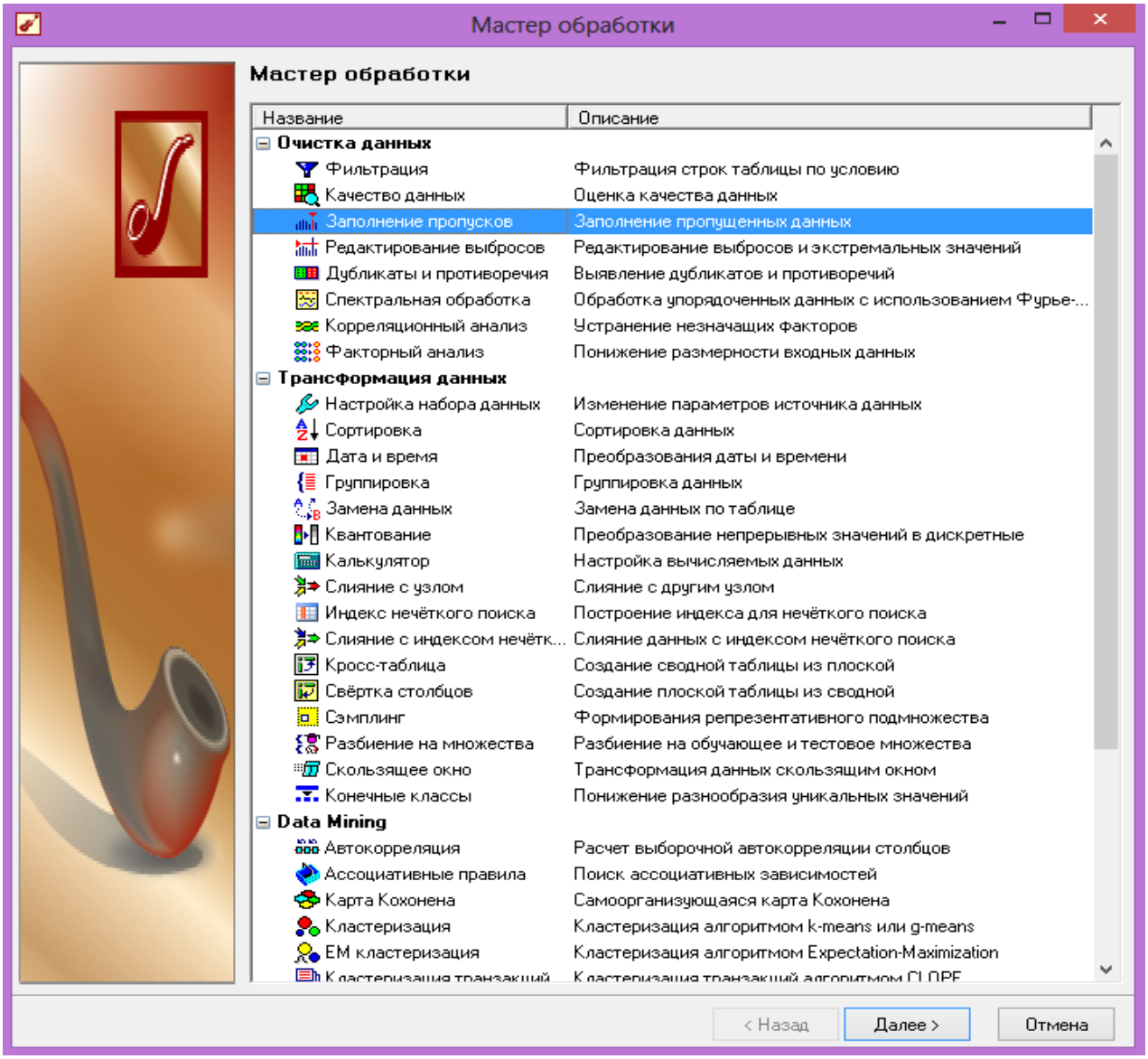


Рис.25. Вікно "Мастер обработки"

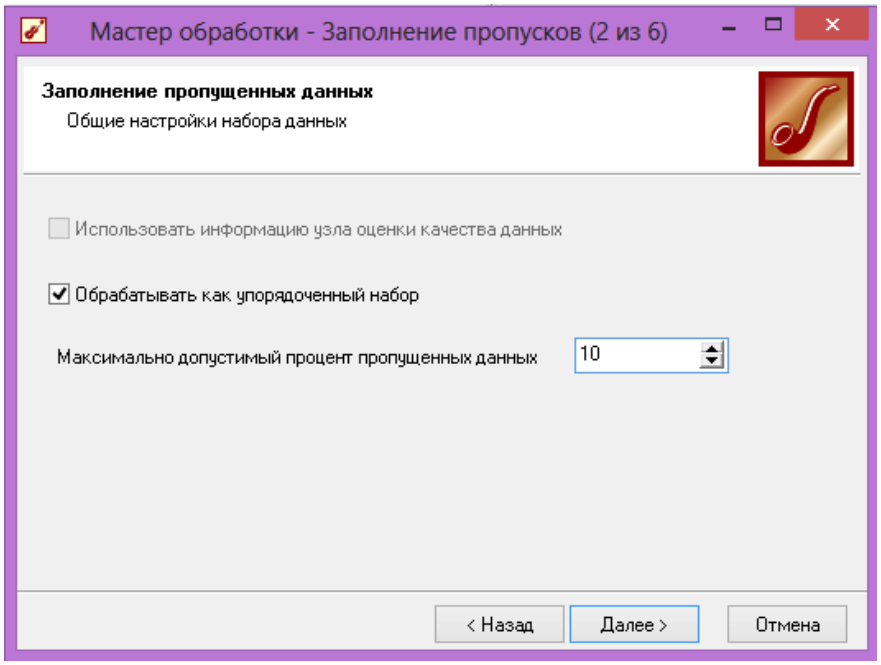


Рис.26. Крок 2 апроксимації функції.

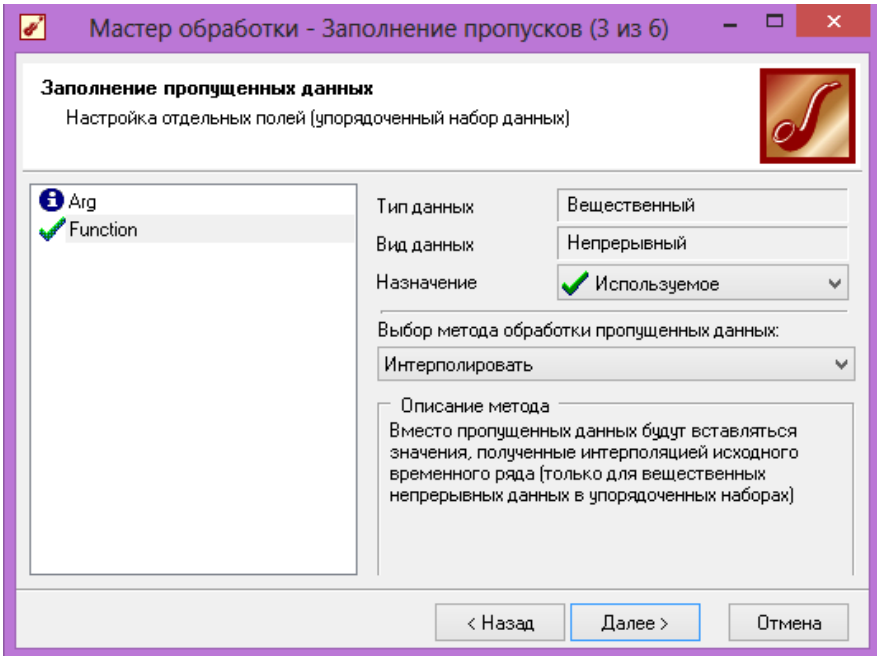


Рис.27. Вибір методу інтерполяція функції.

Для того, що б відновити пропущені дані в стовпчику Function потрібно побудувати інтерполяційну функцію. Переходимо на 3-й крок методу та обираємо “Інтерполяція”. В меню присутня також функції “Заменять средним”, “Заменять медианой”, “Заменять наиболее вероятным”, “Заменять случайными значениями”, що використовується для відновлення даних у неупорядкованих масивах.

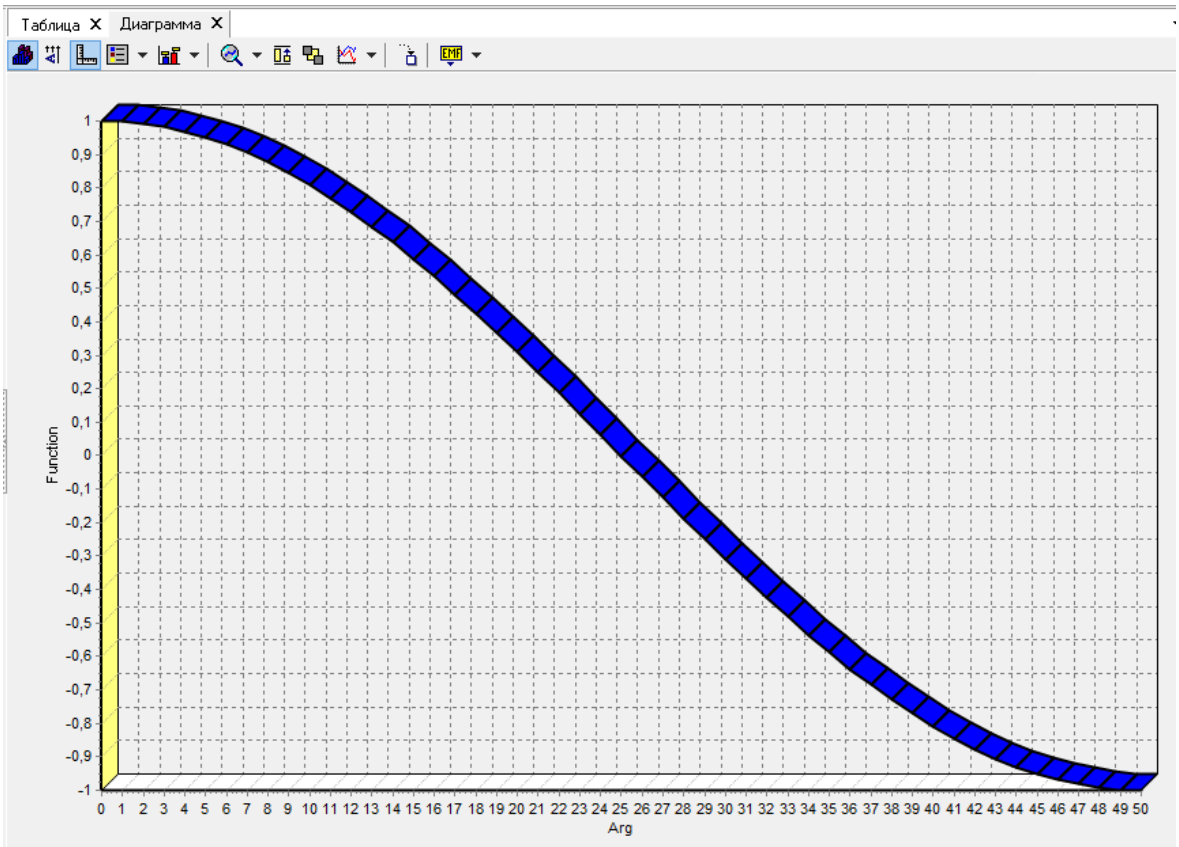


Рис.28. Результат апроксимації (інтерполяційна функція).

У порівнянні з графіком на Рис.18, графік інтерполяційної функції на Рис.28 фактично ідеально відтворює пропущені дані гладкої функції $y = \cos(t)$.

4. Дублікати та протиріччя

Deductor Studio надає інструментальні засоби для виявлення у початковій вибірці даних дублюючих та суперечливих записів. Дублікати – це данні (записи) що повторюють вже існуючі дані, збільшуючи обсяг вибірки, її надмірність та не підвищують інформативність даних. Протиріччя – це записи вибірки даних, що містять однакові значення вхідних змінних та різні значення вихідних даних. Протиріччя створюють умови для суперечливих висновків при аналізі даних та знижують якість моделей побудованих на масивах даних.

Алгоритм Deductor Studio виявляє в масивах даних записи, для яких однаковим наборам вхідних змінних відповідають однакові (дублікати) або різні (протиріччя) вихідні поля. В результаті в таблиці даних створюються два додаткових стовпчика – Дублікати та Протиріччя, що містять відмітки на відповідних рядках таблиці з дублікатами та протиріччями, та додаткові стовпчики – Група дублікатів і Група протиріч що позначують номери груп дублікатів та протиріч. Рядки таблиці, що не входять до обох груп не містять ніяких відміток.

Розглянемо приклад виявлення дублікатів та протиріч для даних з файлу task12.txt. Після імпорту даних обираємо спосіб відображення Таблиця (Рис.29). У файлі знаходяться дані з кредитування групи осіб: код анкети, Прізвище, Ім’я, По-батькові та Сума кредиту, грн. (Рис.29) Особи, які мають однакові П.І.Б., код анкети та суму кредиту – це дублікати. Люди, які мають однакові П.І.Б., але різні коди анкет та суми кредиту – протиріччя.

Таблица | Статистика | Диаграмма

1 / 20

КанцНомер	Прізвище	Ім'я	По-батькові	Сума премії, грн.
140	Смотіков	Петро	Миколайович	67000
161	Волошин	Микола	Петрович	82000
229	Волошин	Микола	Петрович	63000
212	Лисенко	Максим	Сергійович	125000
212	Лисенко	Максим	Сергійович	125000
153	Сердючка	Ігнат	Володимирович	64000
279	Білов	Юрій	Володимирович	89500
279	Білов	Юрій	Володимирович	89500
202	Бокай	Сергій	Іванович	57000
228	Варановська	Ганна	Олександрівна	18500
243	Гаврилов	Василій	Абрамович	99000
209	Гаврилов	Василій	Абрамович	88000
275	Гуміцина	Гаврило	Степанович	29300
242	Дружковка	Ігор	Миколайович	61500
252	Олексійченко	Гаврило	Викторович	75000
252	Олексійченко	Гаврило	Викторович	75000
283	Захарова	Надія	Павловна	88450
248	Загородній	Олексій	Павлович	93000
254	Забродський	Марк	Петрович	78350
230	Горбачова	Вікторія	Максимівна	81200

Рис.29. Таблиця з імпортованими даними.

Завантажуємо “Мастер обработки” та обираємо “Дубликаты и противоречия” (Рис.25). На другому кроці налаштовуємо призначення полів. Вхідними полями є “Прізвище”, “Ім’я”, “По-Батькові”, вихідним полем є “КанцНомер”, поле “Сума премії” - інформаційне (Рис.30).

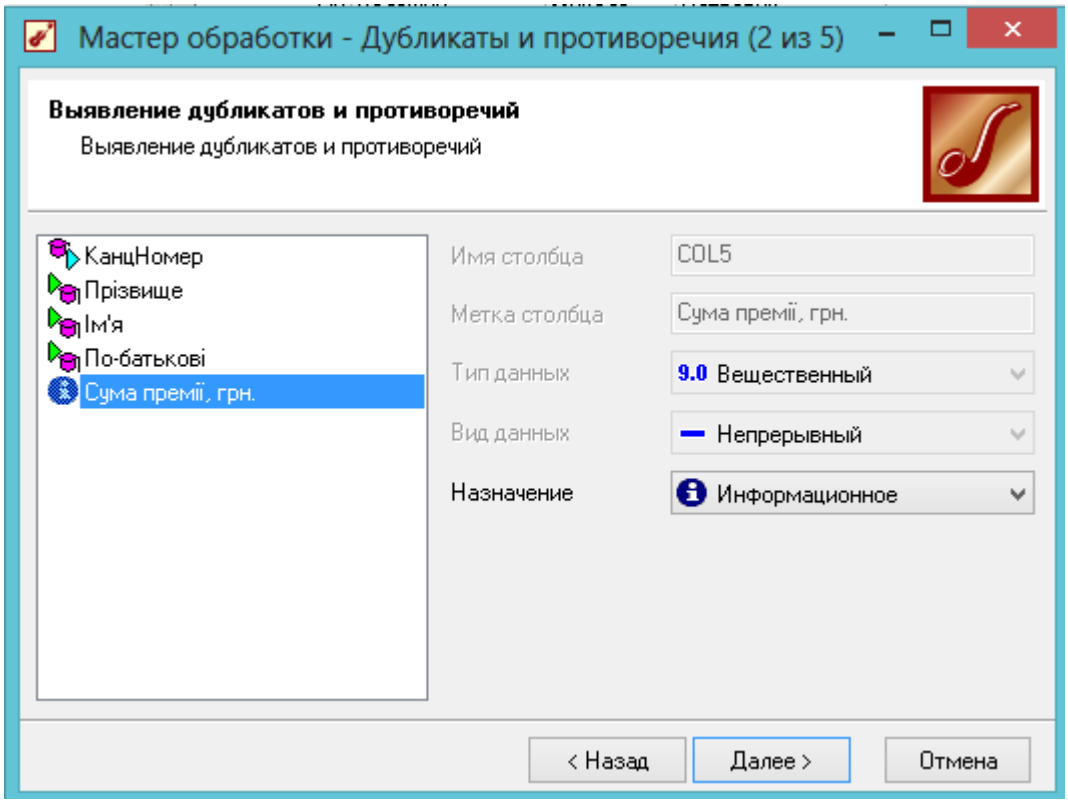


Рис.30.Налаштування призначення полів даних.

На наступному кроці запускаємо обробку. Дані відображаємо у вигляді таблиці. Після обробки виявлення дублікатів та протиріч нам видає результат.

Дубликаты и противоречия | Таблица | Диаграмма

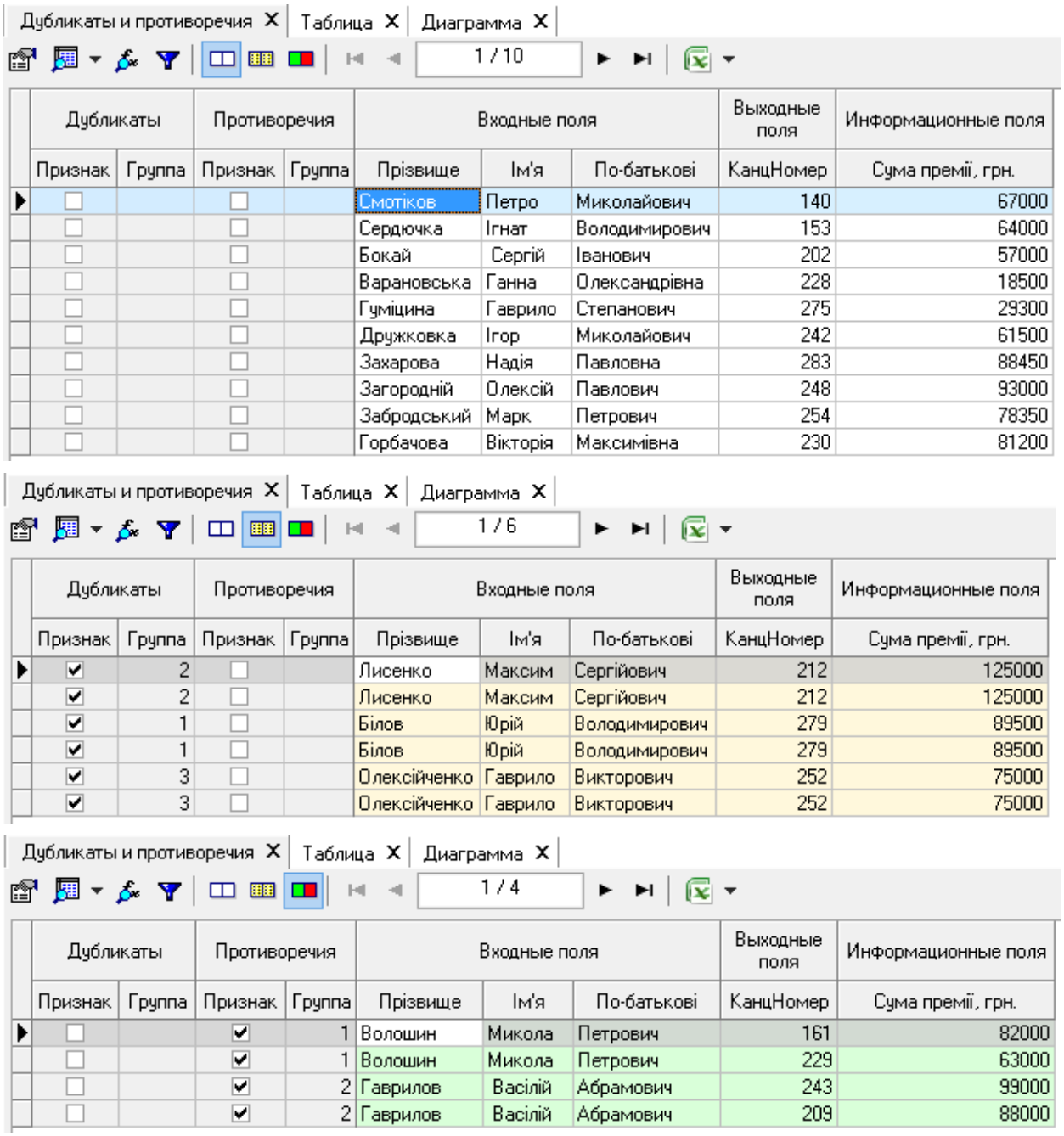
1 / 20

Дубликаты		Противоречия		Входные поля			Выходные поля	Информационные поля
Признак	Группа	Признак	Группа	Прізвище	Ім'я	По-батькові	КанцНомер	Сума премії, грн.
☐		☐		Смотіков	Петро	Миколайович	140	67000
☐		☑	1	Волошин	Микола	Петрович	161	82000
☐		☑	1	Волошин	Микола	Петрович	229	63000
☑	2	☐		Лисенко	Максим	Сергійович	212	125000
☑	2	☐		Лисенко	Максим	Сергійович	212	125000
☐		☐		Сердючка	Ігнат	Володимирович	153	64000
☑	1	☐		Білов	Юрій	Володимирович	279	89500
☑	1	☐		Білов	Юрій	Володимирович	279	89500
☐		☐		Бокай	Сергій	Іванович	202	57000
☐		☐		Варановська	Ганна	Олександрівна	228	18500
☐		☑	2	Гаврилов	Василій	Абрамович	243	99000
☐		☑	2	Гаврилов	Василій	Абрамович	209	88000
☐		☐		Гуміцина	Гаврило	Степанович	275	29300
☐		☐		Дружковка	Ігор	Миколайович	242	61500
☑	3	☐		Олексійченко	Гаврило	Викторович	252	75000
☑	3	☐		Олексійченко	Гаврило	Викторович	252	75000
☐		☐		Захарова	Надія	Павловна	283	88450
☐		☐		Загородній	Олексій	Павлович	248	93000
☐		☐		Забродський	Марк	Петрович	254	78350
☐		☐		Горбачова	Вікторія	Максимівна	230	81200

Рис.31. Таблиця з результатами виявлених 3 груп дублікатів та 2 груп протиріч.

Таким чином виявляємо протиріччя – Волошин М.П. і Гаврилов В.А., бо ПІБ співпадає, але код анкети та сума кредиту різні. Дублікати – це Лисенко

М.С., Білов Ю.В. та Олексійченко Г.В., бо мають однакові всі поля. Для більш зручного вигляду результатів відфільтруємо дані, за допомогою значка фільтр на панелі: показати дані без дублікатів та протиріч – , показати дані з дублікатами - , показати дані з протиріччями - (Рис.32).



Дубликаты и противоречия | Таблица | Диаграмма — 1 / 10

Дубликаты		Противоречия		Входные поля			Выходные поля	Информационные поля
Признак	Группа	Признак	Группа	Прізвище	Ім'я	По-батькові	КанцНомер	Сума премії, грн.
☐		☐		Смотіков	Петро	Миколайович	140	67000
☐		☐		Сердючка	Ігнат	Володимирович	153	64000
☐		☐		Бокай	Сергій	Іванович	202	57000
☐		☐		Варановська	Ганна	Олександрівна	228	18500
☐		☐		Гуміцина	Гаврило	Степанович	275	29300
☐		☐		Дружковка	Ігор	Миколайович	242	61500
☐		☐		Захарова	Надія	Павловна	283	88450
☐		☐		Загородній	Олексій	Павлович	248	93000
☐		☐		Забродський	Марк	Петрович	254	78350
☐		☐		Горбачова	Вікторія	Максимівна	230	81200

Дубликаты и противоречия | Таблица | Диаграмма — 1 / 6

Дубликаты		Противоречия		Входные поля			Выходные поля	Информационные поля
Признак	Группа	Признак	Группа	Прізвище	Ім'я	По-батькові	КанцНомер	Сума премії, грн.
☑	2	☐		Лисенко	Максим	Сергійович	212	125000
☑	2	☐		Лисенко	Максим	Сергійович	212	125000
☑	1	☐		Білов	Юрій	Володимирович	279	89500
☑	1	☐		Білов	Юрій	Володимирович	279	89500
☑	3	☐		Олексійченко	Гаврило	Викторович	252	75000
☑	3	☐		Олексійченко	Гаврило	Викторович	252	75000

Дубликаты и противоречия | Таблица | Диаграмма — 1 / 4

Дубликаты		Противоречия		Входные поля			Выходные поля	Информационные поля
Признак	Группа	Признак	Группа	Прізвище	Ім'я	По-батькові	КанцНомер	Сума премії, грн.
☐		☑	1	Волошин	Микола	Петрович	161	82000
☐		☑	1	Волошин	Микола	Петрович	229	63000
☐		☑	2	Гаврилов	Василій	Абрамович	243	99000
☐		☑	2	Гаврилов	Василій	Абрамович	209	88000

Рис.32. Результати фільтрації даних без дубликатів та протиріч, тільки з дубликатами, тільки з протиріччями.

Таким чином, програмні засоби Deductor Studio дозволяють виявити некоректні набори даних з дублікатами та протиріччями.

Завдання для лабораторної роботи 1.

1. Створити файли Lab11.xlsxx, Lab12.xlsxx.

2. У файлі Lab11.xlsx перший стовпчик A з назвою N містить цифри 1, 2, ..., 50 - номери точок.
3. Другий стовпчик B з назвою Arg містить формулу - значення комірки ліворуч поділити на 10, тобто N/10 - 0.1, 0.2, ... - аргумент функції.
4. Третій стовпчик C з назвою Func1 містить формулу $=\text{Arg}+\text{SIN(Arg)}$, де Arg - значення комірок стовпчику B.
5. Після закінчення вводу даних у Lab11.xlsx необхідно видалити значення комірок C7, C12, C22, C32, C42.
6. Зберігаємо файл Lab11.xlsx. Зберігаємо цей файл також як Lab11.txt.
6. Завантажуємо Deductor. Робимо імпорт даних з файлу Lab11.txt. Налаштовуємо візуалізацію даних у різних форматах та демонструємо результат викладачу.
7. Майстром обробки даних відтворюємо пропущені дані у стовпчику C.
8. Копіюємо таблицю файлу Lab11.xlsx у Lab12.xlsx
9. Третій стовпчик C з назвою Func1 у Lab12.xlsx містить формулу $=\text{Arg}+\text{SIN(Arg)}$, четвертий стовпчик D з назвою Func2 містить формулу $=\text{Arg}+\text{COS(Arg)}$, де Arg - значення комірок стовпчику B (без пропусків).
10. Копіюємо значення (“Специальная вставка” – “значение”) рядків
1) A5, B5, C5, D5 у A15, B15, C15, D15 та у A19, B19, C19, D19;
2) A18, B18, C18, D18 у A35, B35, C35, D35 та у A45, B45, C45, D45.
11. У комірку A19 вносимо значення 19, у A45 значення 45.
12. Зберігаємо файл Lab12.xlsx. Зберігаємо цей файл також як Lab12.txt.
13. Імпорт даних з файлу Lab12.txt. Налаштовуємо візуалізацію даних у різних форматах.
14. Завантажуємо майстром обробки даних та обираємо дублікати та протиріччя. Встановлюємо стовпчик N як вхідний, решта – вихідні.
15. Майстром обробки даних знаходимо дублікати та протиріччя.

Тема 2. Комплексна передобробка даних. Зменшення кількості вхідних факторів, видалення незначущих факторів.

Задачі: Аналіз розмірності набору даних одного аргументу, виявлення та видалення незначних факторів, зменшення кількості факторів.

Методи: Багатовимірний факторний аналіз.

1. Теоретичні положення факторного аналізу

Факторний аналіз – це статистичний метод аналізу впливу набору факторів на результативний (вихідний) показник. В результаті аналізу кожен фактор отримує кількісну та якісну оцінку. Кожен показник, в свою чергу, може виступати як факторний так і результативний. Можна виділити чотири типи задач, в яких використовується факторний аналіз:

- зменшення кількості змінних (редукція даних);
- групування даних;
- класифікація та компактна візуалізація даних;
- пошук прихованих змінних.

Для формування набору статистично незалежних змінних (необхідного для опису даних мінімуму факторів) факторний аналіз використовує обчислення коефіцієнтів кореляції між змінними (факторами). Тобто, зменшення розмірності вхідних даних відбувається за рахунок виявлення значущих і незначних змінних Саме таку задачу ми розглянемо у даному розділі (Рис.1).

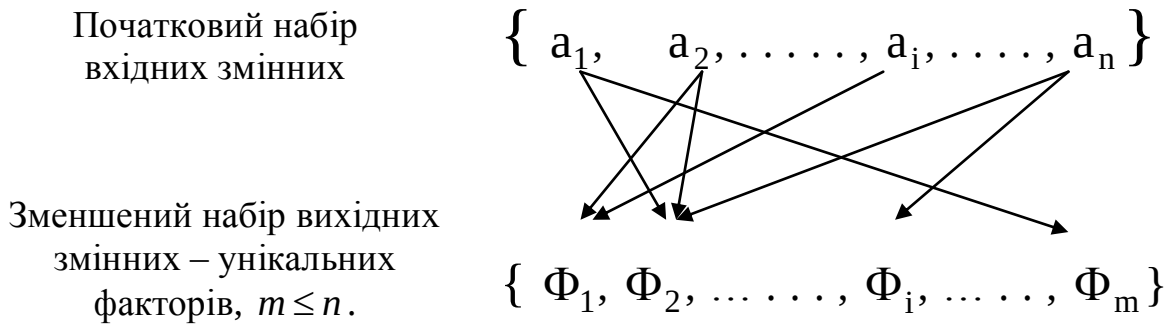


Рис.1. Схема моделі факторного аналізу.

Дані для факторного аналізу підготовлюються у вигляді двовимірного масиву даних - матриці. Стовпчики матриці містять фактори, а рядки - кількісні характеристики атрибутів по кожному фактору. Усі комірки матриці обов’язково мають бути заповнені числовими значеннями. Як правило, кількість рядків не менш за кількість стовпчиків, а кількість факторів є достатньо великою (більше за 10).

2. Процедури факторного аналізу

Однією з проблем факторного аналізу є проблема повороту факторних структур. Поворот дозволяє зробити факторні навантаження контрастнішими за рахунок певного зменшення навантажень по одних змінних і збільшення по інших. Це сприяє більш чіткому прояву ознак, які формують той чи інший фактор. Існує близько 13 методів повороту, найбільш поширені серед яких є варімакс, квартимакс, еквімакс (реалізують ортогональний поворот), промакс та прямий облімін. Найчастіше в прикладних дослідженнях використовується метод варімакс, що максимізує розмах квадратів навантажень для кожного фактора та призводить до збільшення великих та зменшення малих значень факторних навантажень [1], [3], [4], [16]. Методи факторного аналізу:

- Метод головних компонент;
- Незважений метод найменших квадратів**;**
- Узагальнений метод найменших квадратів**;**
- Метод максимальної правдоподібності;
- Альфа-факторний метод;
- Метод розпізнавання образів.

Ми обираємо найбільш розповсюджений Метод головних компонент який реалізовано у середовищі Deductor Studio.

У більшості аналітичних платформ факторний аналіз складається з наступних етапів. Спочатку здійснюється імпорт початкового набору даних вхідних змінних у вигляді двовимірного масиву із підготовленого файлу. Далі користувач обирає метод факторного аналізу та встановлює значення початкової дисперсії що утворюється вхідними факторами. На наступному кроці користувач може уточнювати кількість вихідних факторів за рахунок корегування порога значущості. Далі обирається методу повороту факторів та проводиться процедура факторного аналізу із поворотом. Після цього користувач аналізує та інтерпретує результати факторного аналізу. В деяких аналітичних платформах (SPSS Statistics) передбачено обчислення характеристик з оцінки якості побудованої моделі факторного аналізу. Йдеться про відтворену кореляційну матрицю, тобто кореляційну матрицю вхідних факторів, яку б ми отримали з припущенням, що обчислені вихідні фактори є правильними. Нажаль, у Deductor Studio не передбачено обчислення таких характеристик.

3. Факторний аналіз у середовищі Deductor Studio

Для проведення факторного аналізу даних використовуємо пакет Deductor Studio та метод головних компонент (МГК) [1], [3], [4], [16]. Розглянемо приклад обробки даних експертного опитування стосовно побудови у центрі великого міста паркової зони із туристичною інфраструктурою. Учасники опитування виставляли бали від 1 (повної незгоди) до 10 (повної підтримки) стосовно:

1. Необхідно поліпшити екологію міста.
2. Необхідно м'якше ставитися до витрат на соціальні програми.

Таблица | Статистика

1 / 100

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
8	7	5	3	8	7	2	7	8	2	10	5	4	2	1
8	8	5	4	10	10	3	10	10	5	10	5	1	1	1
7	10	10	8	8	8	8	7	7	10	7	5	10	5	1
10	8	4	2	10	10	2	7	2	1	2	1	1	1	1
10	7	7	2	10	3	2	8	10	1	10	1	1	1	1
10	10	4	1	10	10	4	10	10	5	10	2	7	1	2
8	10	10	1	10	5	4	10	10	5	10	1	1	1	1
10	7	1	5	10	8	3	10	10	2	10	1	1	1	1
10	10	10	4	10	10	4	10	10	5	10	5	10	4	10
10	8	5	5	10	5	10	10	10	5	5	1	5	1	1
10	10	10	5	10	10	1	5	1	5	10	1	1	1	1
8	10	10	7	10	7	3	7	10	10	10	1	1	1	1
2	8	10	10	5	10	10	4	2	10	2	8	8	7	5
7	7	7	8	8	8	5	8	8	5	10	4	5	1	1
5	8	8	3	7	2	2	5	5	10	4	5	8	4	5
5	7	10	10	8	10	8	3	7	10	7	8	7	5	5
1	1	5	10	3	5	1	2	5	5	8	10	10	10	10
7	1	8	1	10	10	4	4	8	10	10	2	7	1	7
10	10	5	8	10	8	7	10	8	7	4	1	8	2	2
7	7	4	2	7	7	2	4	8	7	8	2	2	1	2
7	10	7	10	10	10	8	10	10	8	10	8	10	8	8
7	8	7	5	10	8	7	10	10	7	8	2	8	1	2
7	10	7	5	10	7	4	8	10	8	10	1	7	1	1
7	8	10	5	10	7	7	7	10	8	7	10	8	4	8
10	10	7	1	10	4	1	10	10	10	10	1	1	1	1
4	1	10	10	1	10	1	1	7	10	1	1	5	1	1
5	5	8	1	10	10	5	8	8	8	4	2	10	1	2
8	1	8	10	8	7	10	5	8	8	10	7	10	1	8

Рис.2. Фрагмент таблиці з вхідними даними експертного опитування.

Таблица | Статистика

Статистика: Кол-во значений = 100

	Метка столбца	Гистограмма	Минимум	Максимум	Среднее	Стандартное откл.	Сумма	Сумма квадратов	Кол-во пустых значений
1	9.0 Q1		1	10	7,54	1,956186767	754	6064	0
2	9.0 Q2		1	10	7,85	2,350048355	785	6709	0
3	9.0 Q3		1	10	6,65	2,579542673	665	5081	0
4	9.0 Q4		1	10	5,52	2,972873996	552	3922	0
5	9.0 Q5		1	10	8,6	1,901620787	860	7754	0
6	9.0 Q6		1	10	7,14	2,71553704	714	5828	0
7	9.0 Q7		1	10	4,59	2,843013766	459	2907	0

Рис.3. Статистичні характеристики вхідних даних.

3. Гроші країни мають бути витрачені на економічні програми.
4. Країна повинна стрімко нарощувати ВНП.
5. Необхідно намагатися наближати людину з природою.
6. Бюджет екологічних програм має бути урізаним.
7. Економіка країни не витримає такого навантаження соціальних витрат.
8. Програми озеленення міст мають бути впроваджені у всій Європі.
9. Негативне ставлення до програм озеленення міст та розвитку туризму завдає шкоду економіці країни.

10. Спочатку необхідно створити гарні умови розвитку економіки країни.
11. Ми усі теж, по суті, діти природи та полюбляємо розважатися.
12. Розвиток розважальних закладів у центрі міста сприяє зростанню криміналу.
13. Заборона видачі дозволу на використання землі у центрі міста для великих парково-розважальних комплексів.
14. Обмеження в правах громадських екологічних організацій.
15. Використання міського бюджету та земельних ділянок у центрі міста для масштабних парково-розважальних центрів – це вбивство економіки та інфраструктури міста.

Усім положенням присвоєно ідентифікатори $Q_1,\ldots,Q_{15}$. У опитуванні взяло участь 100 експертів. Фрагмент таблиці імпорту даних з файлу task2.txt показано на Рис.2. На Рис.3. показано статистичні характеристики даних. Для проведення факторного аналізу у вікні Мастер обработки (Рис.4) треба обрати метод Факторный анализ.

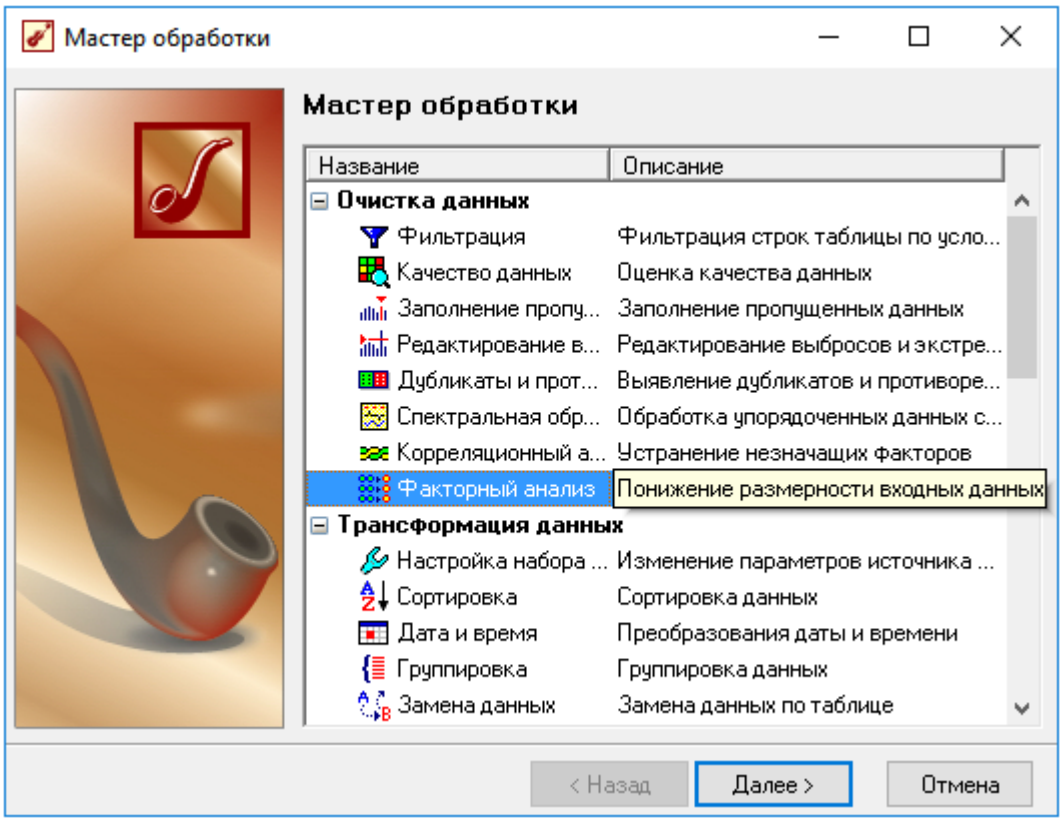


Рис.4. Вибір методу для аналізу даних.

На другому кроці (Рис.5) позначаємо всі змінні як вхідні. Далі серед запропонованих методів отримання кінцевого розв'язку обираємо спочатку метод "Без вращения" (Рис.6). Задаємо дисперсію 90%. Кількість факторів наперед не вказуємо, бо програма сама обчислює їх кількість за допомогою вбудованих критеріїв підбору.

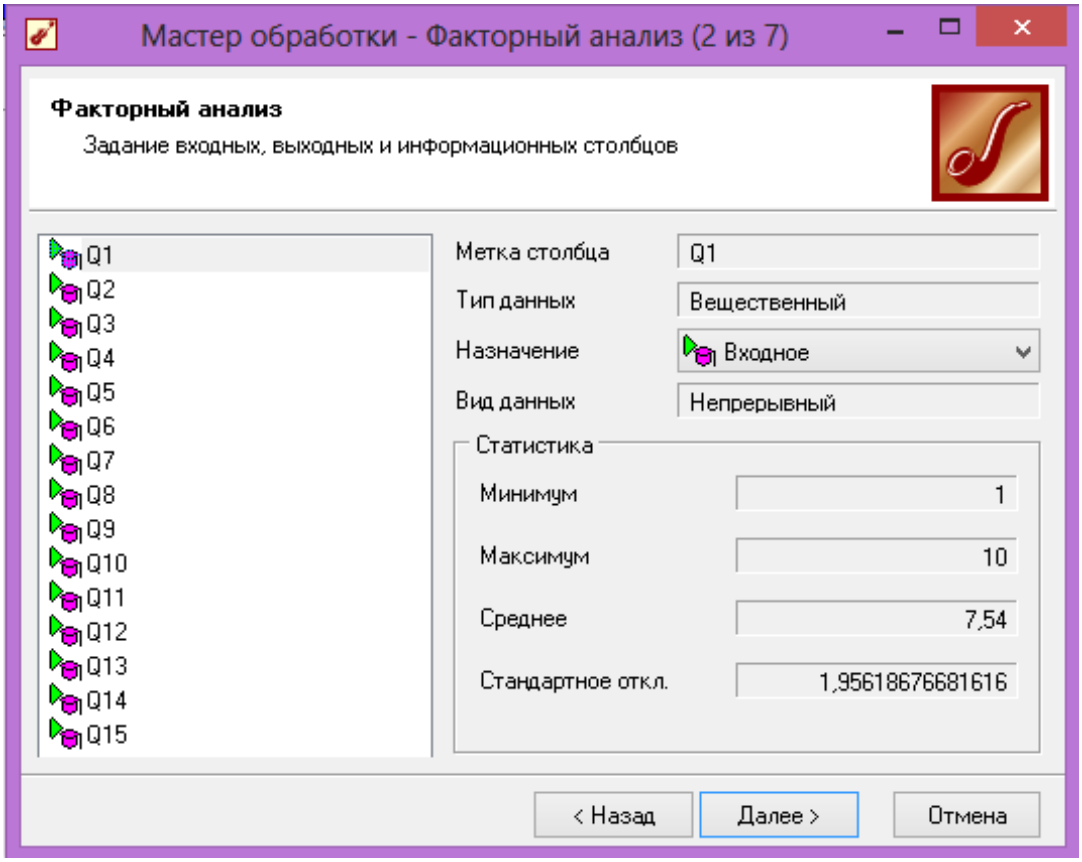


Рис.5. Вікно налаштування вхідних та вихідних стовпчиків

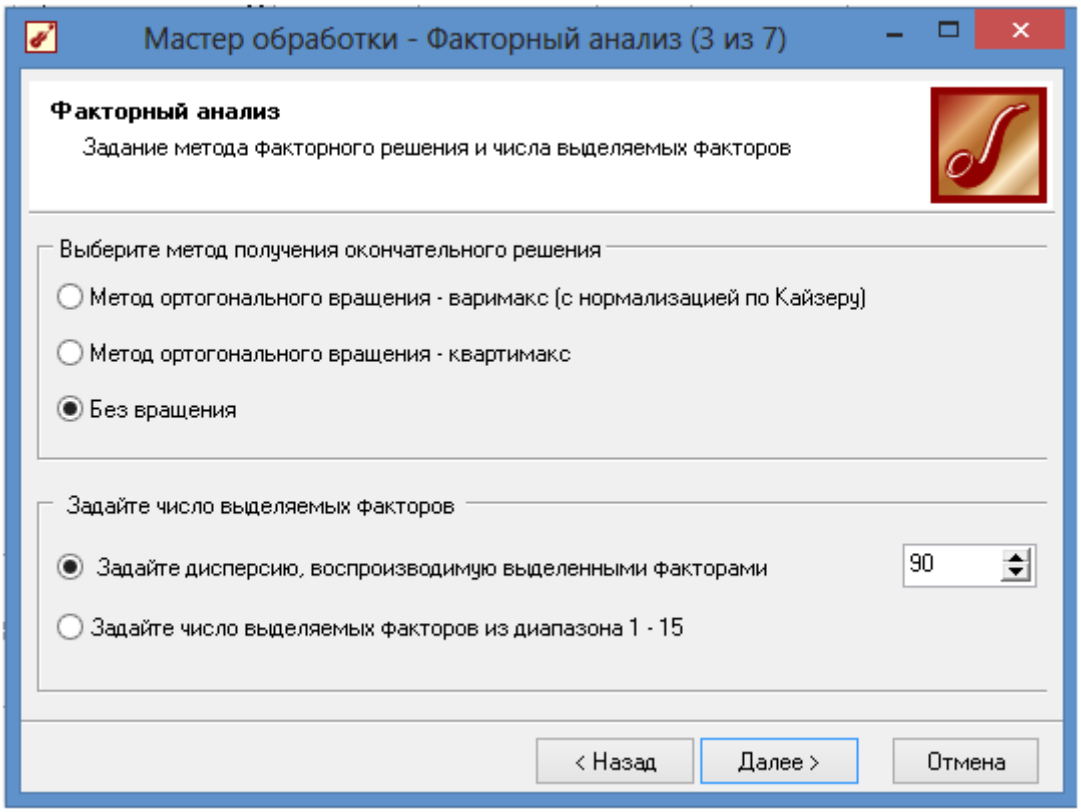


Рис.6. Вікно вибору методу повороту та числа обраних факторів.

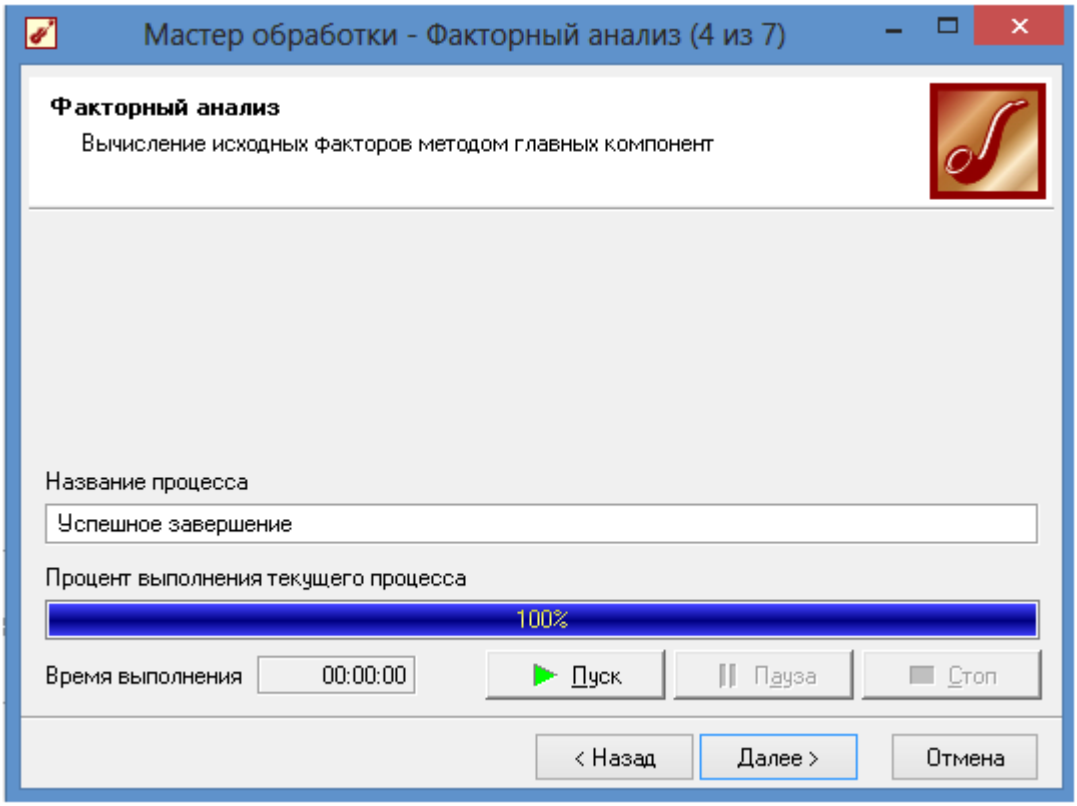


Рис.7. Запуск процедури факторного аналізу.

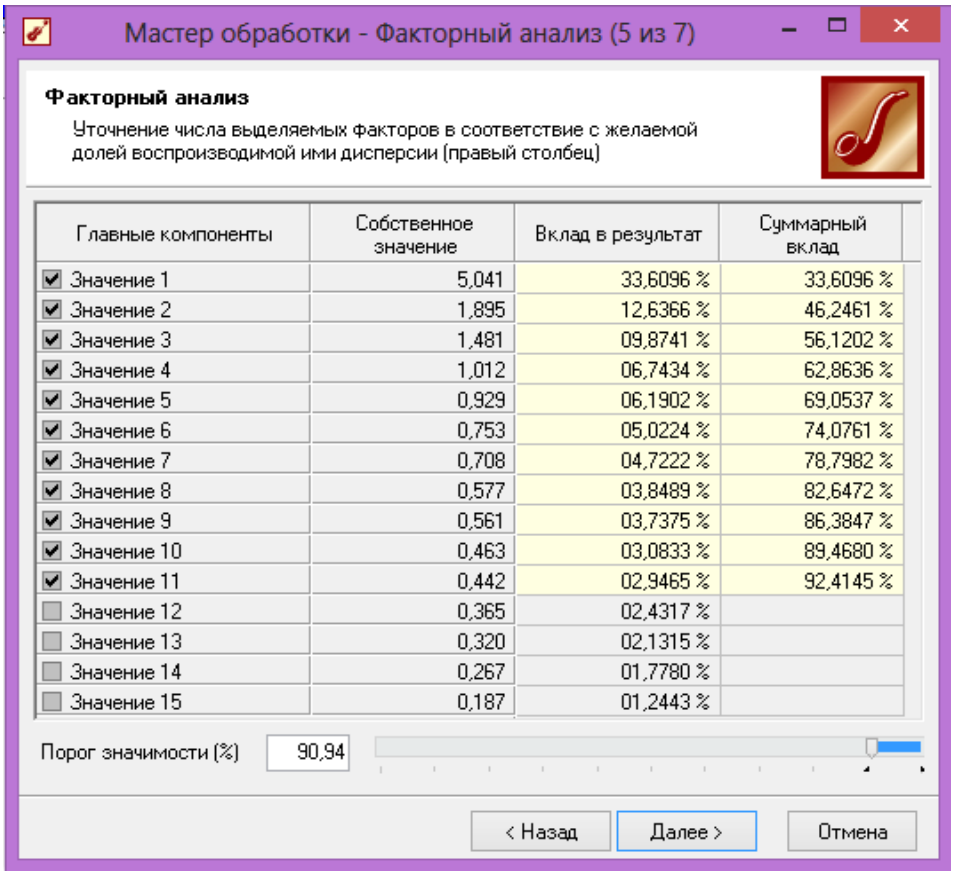
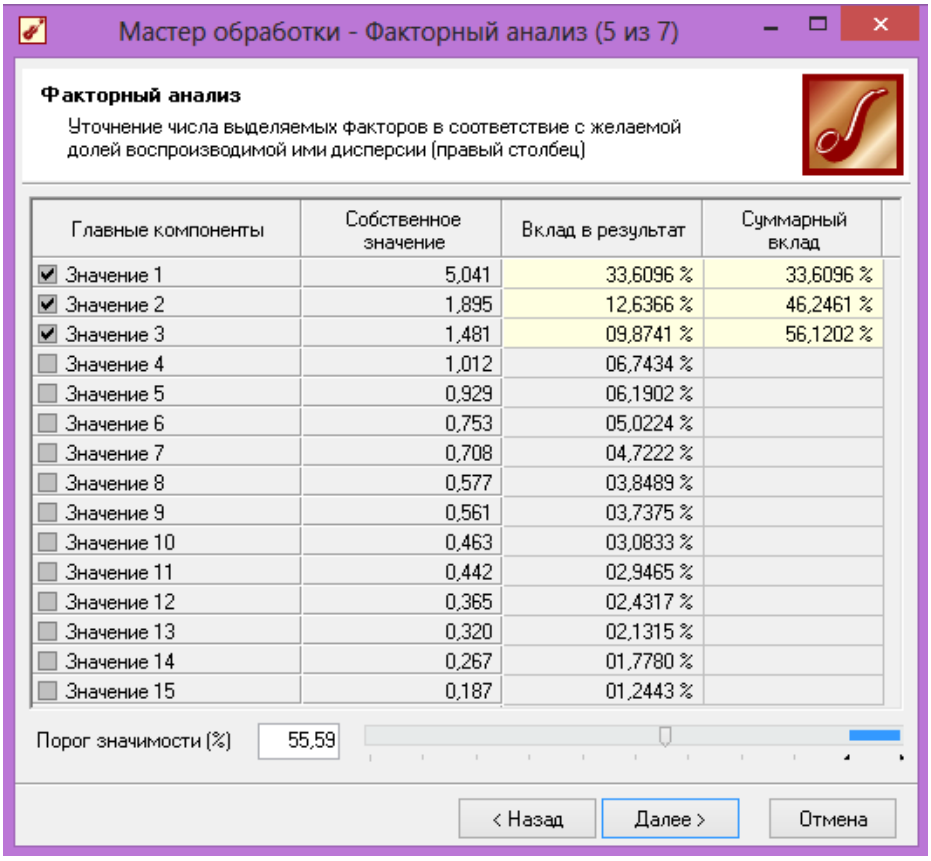


Рис.8. Одинадцять визначених факторів що відповідають рівню значущості 90%.



Главные компоненты	Собственное значение	Вклад в результат	Суммарный вклад
Значение 1	5,041	33,6096 %	33,6096 %
Значение 2	1,895	12,6366 %	46,2461 %
Значение 3	1,481	09,8741 %	56,1202 %
Значение 4	1,012	06,7434 %	
Значение 5	0,929	06,1902 %	
Значение 6	0,753	05,0224 %	
Значение 7	0,708	04,7222 %	
Значение 8	0,577	03,8489 %	
Значение 9	0,561	03,7375 %	
Значение 10	0,463	03,0833 %	
Значение 11	0,442	02,9465 %	
Значение 12	0,365	02,4317 %	
Значение 13	0,320	02,1315 %	
Значение 14	0,267	01,7780 %	
Значение 15	0,187	01,2443 %	

Рис.9. Уточнення кількості виділених факторів до 3 за допомогою встановлення повзунка з порогом значущості 56,58%.

Після запуску процедури (Рис.7) отримуємо визначену програмою 11 факторів з рівнем значущості 88,43%. Аналіз власних значень та внеску кожного фактору в результат (другий та третій стовпчики на Рис.8) виявляє що основний внесок в результат належить першим трьом фактором, що відповідають порогу значущості ~57%. Для оцінки кількості значущих факторів можна також залучити візуальний метод графіку “кам’яного насипу” (scree plot) (Рис.10), або графік власних значень кожного фактору (графік побудовано в Excel за другим стовпчиком таблиці Рис.8). Даний засіб оцінки кількості значущих факторів не має строгого статистичного обгрунтованя, але часто використовується як допоміжний критерій.

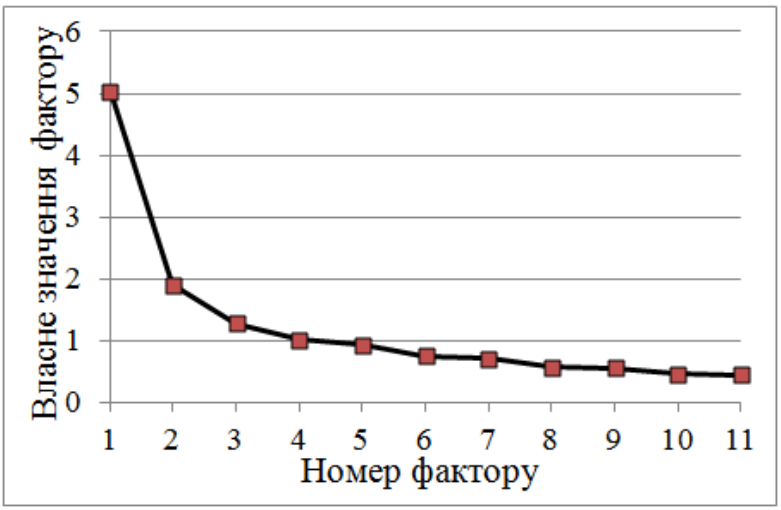


Рис. 11. Графік “кам’яного насипу”.

Вперше Р.Кеттел (R.Cattell) запропонував знайти на графіку такий фактор, для якого зменшення власних значень (похідна) максимально сповільнюється. Із графіку на Рис.10 слідує що таким фактором є четвертий фактор і ми маємо залишити тільки три значущі для аналізу фактори. Виходячи із наведених критеріїв аналітик може змінити поріг значущості за допомогою повзунка (Рис.8) що, приводить до уточнення кількості факторів до 3 за рахунок зміни порога значущості до 56,58% (Рис.9).

4. Інтерпретація результатів факторного аналізу

Налаштування параметрів візуалізації результатів факторного аналізу показано на Рис.11. На Рис.12, 13 показано результати факторного аналізу без обертання (“Без вращения”) у вигляді таблиці та гістограми. На таблиці вказуємо поріг значущості 

0,5. За допомогою цього параметру відбувається фільтрація менш значущих вхідних змінних що приймають участь у формуванні факторів. Верхня таблиця на Рис.12. показує, яку частину дисперсії кожної зі змінних аналізу пояснює запропонована факторна модель, тобто для яких змінних факторний аналіз спрацював краще чи гірше. Наприклад, змінна Q_1 (що відповідає положенню «Необхідно поліпшити екологію міста») на 69,46% пояснюється запропонованою моделлю. Якщо значення менш порогу значущості, то відповідні змінні виключаються із аналізу як незначущі. Такими змінними є Q_2 (“Необхідно м'якше ставитися до витрат на соціальні програми”), Q_6 (“Бюджет екологічних програм має бути урізаним”), Q_9 (“Негативне ставлення до програм озеленення міст та розвитку туризму завдає шкоду економіці країни”), Q_{11} .(“Ми усі теж, по суті, діти природи та полюбляємо розважатися”). У нижній таблиці (кнопка ) на Рис.12 показано, яких величин набувають власні значення кожного фактору та відсоток дисперсії (%), обумовлений кожним із факторів. Відібрано таку мінімальну кількість факторів, яка може пояснити більшу частину дисперсії. Відібрано три фактори, які пояснюють 56,12% сукупної дисперсії. Перший пояснює 33,61%, другий - 12,637 % і третій - 9,874 %. Використовується метод головних компонент.

Необхідною процедурою факторного аналізу є поворот факторів. Після повороту відбувається перерозподіл навантаження на кожен фактор таким чином, щоб зменшити різницю між навантаженнями усіх факторів. Як бачимо з верхньої таблиці на Рис.12, основне навантаження вхідних змінних приходиться на Фактор 1, деякі компоненти можна неоднозначно віднести до певного фактору: вхідні змінні Q_5, Q_{14} приймають участі у формуванні двох факторів одночасно. Після повороту навантаження на усі фактори розподіляються більш рівномірно, що спрощує подальшу їх інтерпретацію. В Deductor Studio застосовуються методи повороту варімакс (поворот, що максимізує дисперсію) та квартимакс (мінімізує кількість факторів, що є зручним для подальшої інтерпретації). Для повороту матриці факторних навантажень необхідно на третьому кроці факторного аналізу (Рис.6) обрати відповідний метод варімакс або квартимакс.

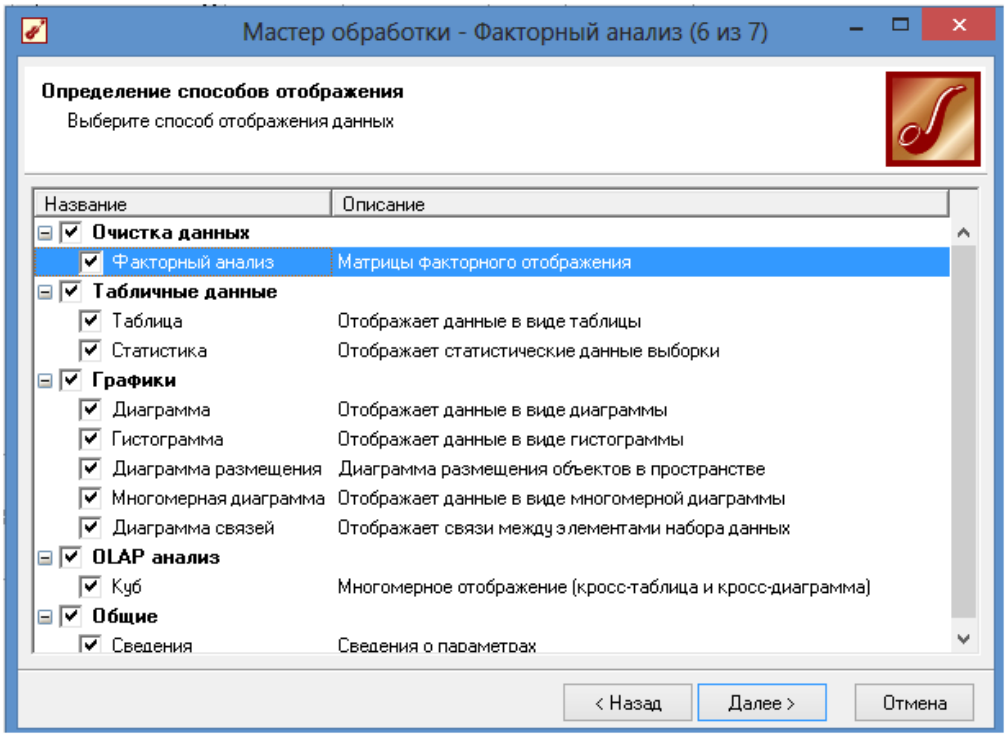


Рис.11.Налаштування параметрів візуалізації результату.

Факторный анализ | Таблица | Статистика | Диаграмма | Гистограмма

Порог значимости 0,50

Переменные	Окончательные факторы (Без вращения)		
	Фактор 1	Фактор 2	Фактор 3
Q1	✔ -0,6946	0,3772	0,0187
Q2	-0,2495	✔ 0,6771	-0,0896
Q3	✔ 0,6829	0,2404	-0,3401
Q4	✔ 0,6165	0,2226	0,0640
Q5	✔ -0,5945	✔ 0,5647	-0,1121
Q6	0,4175	0,2980	✔ -0,5844
Q7	✔ 0,6602	0,3321	-0,0995
Q8	✔ -0,5526	0,4359	0,2696
Q9	-0,2488	0,4830	0,2096
Q10	✔ 0,5368	0,2144	-0,3831
Q11	-0,2114	0,2735	0,3944
Q12	✔ 0,7152	0,0331	0,4247
Q13	✔ 0,7927	0,2689	0,0593
Q14	✔ 0,6723	0,0815	✔ 0,5098
Q15	✔ 0,6283	0,2127	0,3551

Факторы	Собственные значения матрицы ковариаций			Сумма квадратов нагрузок		
	Значение	Значение в %	Сумма в %	Значение	Значение в %	Сумма в %
1	5,0414	33,610	33,610	5,0414	33,610	33,610
2	1,8955	12,637	46,246	1,8955	12,637	46,246
3	1,4811	9,874	56,120	1,4811	9,874	56,120
4	1,0115	6,743	62,864			
5	0,9285	6,190	69,054			
6	0,7534	5,022	74,076			
7	0,7083	4,722	78,798			
8	0,5773	3,849	82,647			
9	0,5606	3,738	86,385			
10	0,4625	3,083	89,468			
11	0,4420	2,946	92,414			
12	0,3647	2,432	94,846			
13	0,3197	2,132	96,978			
14	0,2667	1,778	98,756			
15	0,1867	1,244	100,000			

Рис.12. Результат факторного аналізу без обертання (таблиця).

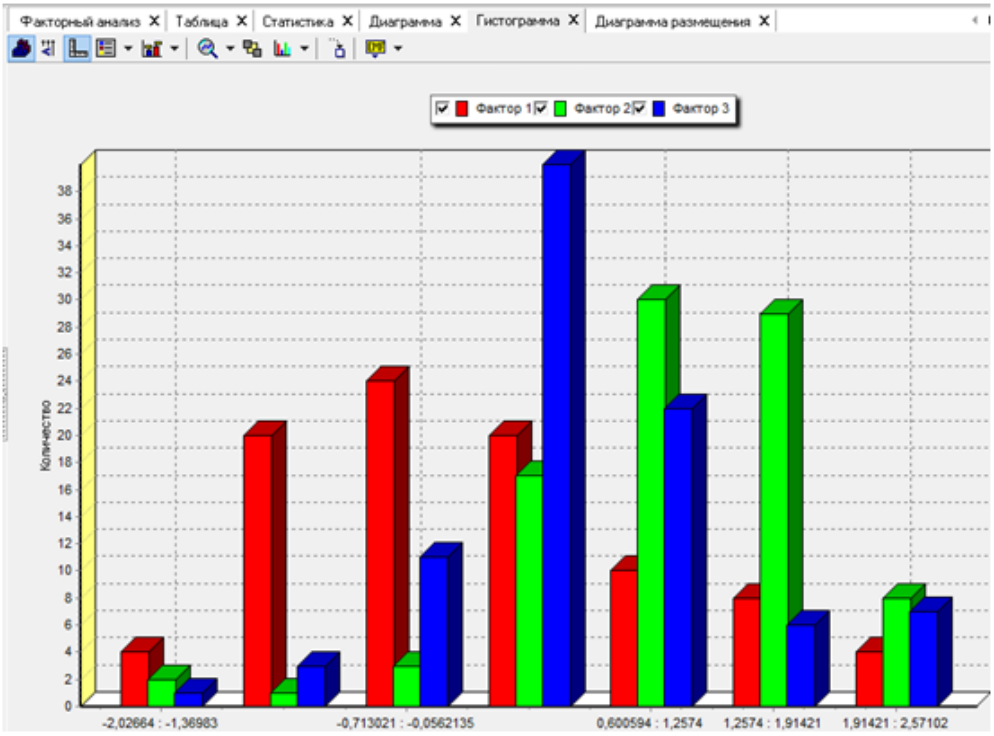


Рис.13. Результат факторного аналізу без обертання (гістограма).

Повернуті матриці факторних навантажень за допомогою методу варімакс та квартимакс а також гістограми зображено на Рис.14 - 17. За даними матриць можна зробити висновок, що незалежно від обраного методу обертання 1) кожен з трьох факторів сформований практично однаковою сукупністю вхідних змінних; 2) на кожен фактор приходиться більш рівномірне навантаження вхідних змінних (від 4 до 6 у прикладі на Рис.14, 16, дивись також “Сумма квадратов нагрузок после вращения” у нижній таблиці) ніж у випадку без обернення (від 2 до 11 на Рис.12).

Роль факторних навантажень грають кореляційні коефіцієнти між вхідними змінними та вихідними факторами. У верхній таблиці на Рис.15 змінна Q_1 (“Необхідно поліпшити екологію міста”) більш за все корелює з Фактором 2 (0,6358); змінна Q_3 (“Гроші країни мають бути витрачені на економічні програми”) з Фактором 3 (0,7105); змінна Q_{15} (“Використання міського бюджету та земельних ділянок у центрі міста для масштабних парково-розважальних центрів – це вбивство економіки та інфраструктури міста”) з Фактором 1 (0,7358) тощо.

У більшості випадків окрема змінна входить тільки в один фактор. У деяких випадках змінна може входити до двох факторів одночасно, наприклад, змінна Q_7 (“Економіка країни не витримає такого навантаження соціальних витрат”) потрапила у Факторі 1 та Факторі 3 (Рис.17). Деякі змінні можуть не входити за обраними критеріями до жодного з факторів, наприклад, Q_{11} (“Ми усі теж, по суті, діти природи та полюбляємо розважатися”). Таким чином, за результатами методу варімакс, на основі оцінок експертів можна сформувати три вихідних фактори (Рис.15, 17):

Фактор 1.

1. Країна повинна стрімко нарощувати ВНП.
2. Розвиток розважальних закладів у центрі міста сприяє зростанню криміналу.

3. Заборона видачі дозволу на використання землі у центрі міста для великих парково-розважальних комплексів.
4. Обмеження в правах громадських екологічних організацій.
5. Використання міського бюджету та земельних ділянок у центрі міста для масштабних парково-розважальних центрів – це вбивство економіки та інфраструктури міста.

Фактор 2.

1. Необхідно поліпшити екологію міста.
2. Необхідно м'якше ставитися до витрат на соціальні програми.
3. Необхідно намагатися наближати людину з природою.
4. Програми озеленення міст мають бути впроваджені у всій Європі.
5. Негативне ставлення до програм озеленення міст та розвитку туризму завдає шкоду економіці країни.

Фактор 3.

1. Гроші країни мають бути витрачені на економічні програми.
2. Бюджет екологічних програм має бути урізаним.
3. Економіка країни не витримає такого навантаження соціальних витрат.
4. Спочатку необхідно створити гарні умови розвитку економіки країни.

Факторный анализ | Таблица | Статистика | Диаграмма | Гистограмма

Порог значимости 0,50

Переменные	Окончательные факторы (Варимакс метод)		
	Фактор 1	Фактор 2	Фактор 3
Q1	-0,4169	✔ -0,6358	-0,2167
Q2	-0,0951	✔ -0,6854	0,2235
Q3	0,3345	0,1519	✔ 0,7105
Q4	✔ 0,5420	0,0581	0,3696
Q5	-0,3872	✔ -0,7314	0,0119
Q6	-0,0053	0,0351	✔ 0,7768
Q7	0,4924	0,0134	✔ 0,5599
Q8	-0,1385	✔ -0,6761	-0,3029
Q9	0,0562	✔ -0,5741	-0,0797
Q10	0,1939	0,1202	✔ 0,6549
Q11	0,1574	-0,4106	-0,2857
Q12	✔ 0,8058	0,1955	0,0741
Q13	✔ 0,6783	0,0946	0,4849
Q14	✔ 0,8395	0,1174	0,0094
Q15	✔ 0,7358	0,0139	0,1564

Факторы	Собственные значения матрицы ковариаций			Сумма квадратов нагрузок до вращения			Сумма квадратов нагрузок после вращения		
	Значение	Значение в %	Сумма в %	Значение	Значение в %	Сумма в %	Значение	Значение в %	Сумма в %
1	5,0414	33,610	33,610	5,0414	33,610	33,610	3,4212	22,808	22,808
2	1,8955	12,637	46,246	1,8955	12,637	46,246	2,4677	16,451	39,259
3	1,4811	9,874	56,120	1,4811	9,874	56,120	2,5292	16,861	56,120
4	1,0115	6,743	62,864						
5	0,9285	6,190	69,054						
6	0,7534	5,022	74,076						
7	0,7083	4,722	78,798						
8	0,5773	3,849	82,647						
9	0,5606	3,738	86,385						
10	0,4625	3,083	89,468						
11	0,4420	2,946	92,414						
12	0,3647	2,432	94,846						
13	0,3197	2,132	96,978						
14	0,2667	1,778	98,756						
15	0,1867	1,244	100,000						

Рис.14. Результат факторного аналізу з методом обернення варімакс (таблиці).

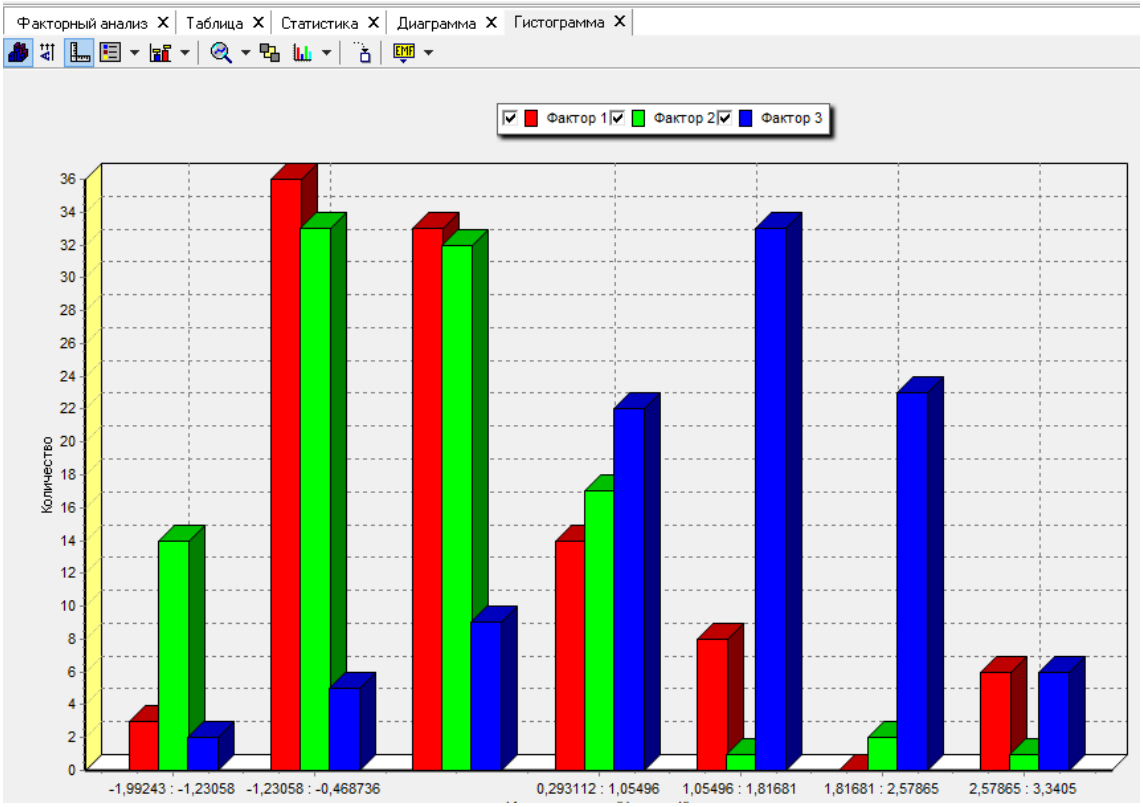


Рис.15. Результат факторного аналізу з методом обернення варімакс (гістограма).

Факторный анализ | Таблица | Статистика | Диаграмма | Гистограмма

Порог значимости 0,50

Переменные	Окончательные факторы (Квартимакс метод)		
	Фактор 1	Фактор 2	Фактор 3
Q1	-0,4334	✔ -0,6297	-0,2016
Q2	-0,0726	✔ -0,6938	0,2052
Q3	0,3977	0,1272	✔ 0,6822
Q4	✔ 0,5734	0,0475	0,3205
Q5	-0,3827	✔ -0,7335	0,0196
Q6	0,0658	0,0061	✔ 0,7748
Q7	✔ 0,5416	-0,0046	✔ 0,5126
Q8	-0,1640	✔ -0,6651	-0,3143
Q9	0,0500	✔ -0,5704	-0,1061
Q10	0,2527	0,0968	✔ 0,6385
Q11	0,1316	-0,3987	-0,3142
Q12	✔ 0,8087	0,1973	0,0075
Q13	✔ 0,7196	0,0804	0,4242
Q14	✔ 0,8365	0,1219	-0,0629
Q15	✔ 0,7470	0,0124	0,0890

Факторы	Собственные значения матрицы ковариаций			Сумма квадратов нагрузок до вращения			Сумма квадратов нагрузок после вращения		
	Значение	Значение в %	Сумма в %	Значение	Значение в %	Сумма в %	Значение	Значение в %	Сумма в %
1	5,0414	33,610	33,610	5,0414	33,610	33,610	3,6644	24,429	24,429
2	1,8955	12,637	46,246	1,8955	12,637	46,246	2,4310	16,206	40,636
3	1,4811	9,874	56,120	1,4811	9,874	56,120	2,3226	15,484	56,120
4	1,0115	6,743	62,864						
5	0,9285	6,190	69,054						
6	0,7534	5,022	74,076						
7	0,7083	4,722	78,798						
8	0,5773	3,849	82,647						
9	0,5606	3,738	86,385						
10	0,4625	3,083	89,468						
11	0,4420	2,946	92,414						
12	0,3647	2,432	94,846						
13	0,3197	2,132	96,978						
14	0,2667	1,778	98,756						
15	0,1867	1,244	100,000						

Рис.16. Результат факторного аналізу з методом обернення квартимакс (таблиці).

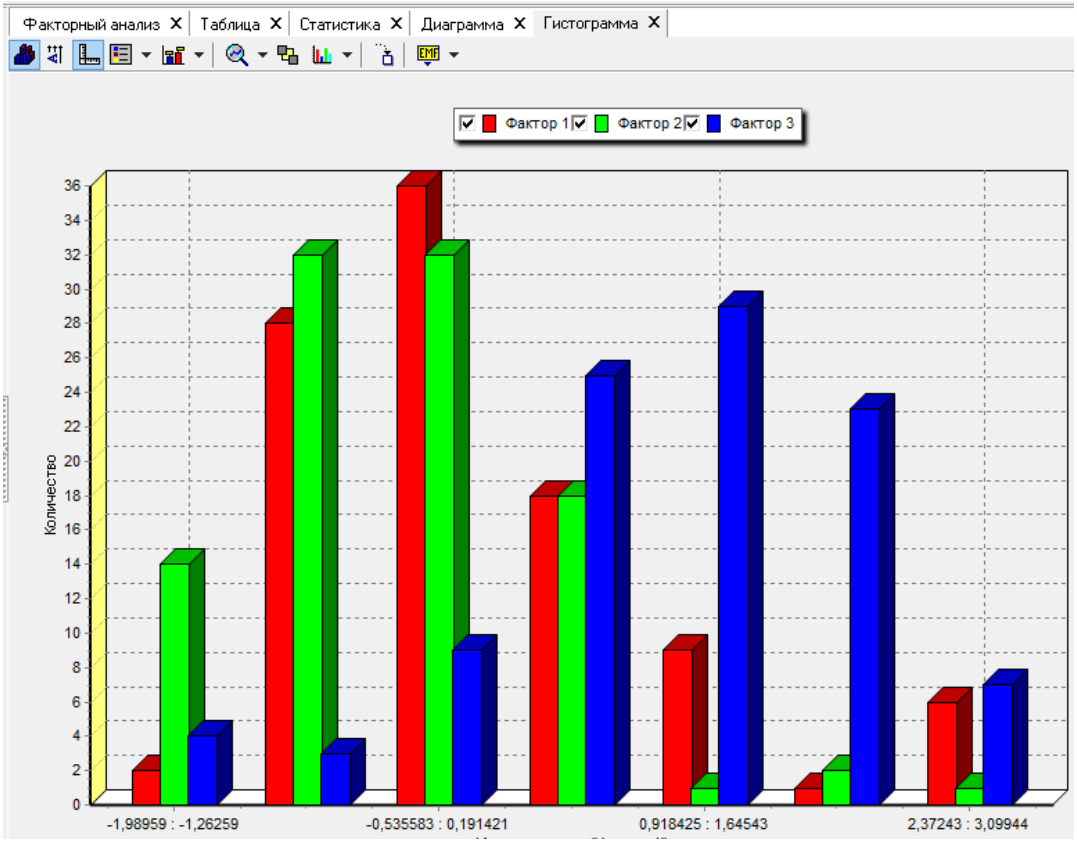


Рис.17. Результат факторного аналізу з методом обернення квартимакс (гістограма).

Через мале навантаження на кожен з факторів, положення “Ми усі теж, по суті, діти природи та полюбляємо розважатися” не включено до жодного фактору. Тепер можна зробити висновки та описати зв’язок між факторами. Перший фактор містить усі положення, що описують негативне ставлення щодо бюджетних витрат на будівництво паркової туристичної зони у центрі міста. Високе значення цього фактора (10 балів у таблиці) говорить про відсутність підтримки таких проектів з боку опитуваних. До другого фактора входять положення, які вказують на підтримку експертами проектів озеленення міста. До третього фактора увійшли положення, що відповідають обережному ставленню експертів до таких проектів з урахуванням балансу бюджетних коштів.

Відповідно до трьох отриманих факторів система згенерувала три нові змінні – значення Фактору 1, Фактору 2 та Фактору 3 для кожного експерта (Рис.18), що містять обчислені значення факторів по методу обернення квартимакс. По кожному з відібраних факторів для кожного опитаного було розраховано спеціальне факторне значення. Статистичні покажчики для кожного з факторів показані у таблиці на Рис.19.

Таким чином, факторний аналіз є потужним інструментом аналізу даних на предмет зменшення кількості параметрів (змінних моделі) та визначення структури взаємозв’язків між параметрами (змінними), тобто класифікації вхідних факторів. Можливості використання факторного аналізу дуже широкі у технічних, економічних, соціальних системах, демографічних, психологічних дослідженнях тощо.

Факторный анализ | Таблица | Статистика | Диаграмма

1 / 100

	Фактор 1	Фактор 2	Фактор 3
▶	-0,172319110196539	0,316516847140947	-1,18552650615296
	-0,470671625041337	-0,732959921203302	-0,298303630310979
	0,817101304705434	-0,214019134673866	1,20484678330845
	-1,98959118992464	0,934617774368832	0,182574416654324
	-0,793131198899983	-0,397903260867303	-1,40067388579097
	-0,501350971996375	-1,37497154445621	0,0438191826890533
	-0,760135799480085	-0,938788795490939	-0,117888312659529
	-0,785914016537669	-0,777331414150097	-1,18726935228956
	1,40325229192504	-1,865959870708	0,130228191948759
	-0,517163751052467	-0,810796490912717	0,155357301928372
	-1,46121273250058	0,245385050218188	0,924514907804687
	-0,754660294062784	-0,707453154842997	0,942701156091717
	1,38367004796796	2,15356580724547	1,40194573868134
	-0,0396968399895201	0,0728496040778041	-0,0368180209133943
	0,412173820346914	1,49296717784519	-0,345445370951191
	1,23958874005367	0,813152376277526	1,1524620285744
	3,09943709957951	3,19036230873418	-2,57668584149964
	-0,36571130523581	0,984006263038838	0,60070275201862
	-0,33730416470439	-0,954231723268414	0,897774585785362
	-0,981975553935077	1,18106147767111	-0,348776743496676
	2,67949974781342	-1,77396487062554	-0,0330748729905909
	-0,133493512830097	-0,765579183934802	0,629586282874881
	-0,391946833592005	-0,801084240078486	0,53506606734292
	1,64568370627809	-0,586035532998322	0,165218650547241

Рис.18. Фрагмент таблиці значень факторів для кожного експерта (квартимакс).

Факторный анализ | Таблица | Статистика | Диаграмма | Гистограмма

	Метка столбца	Статистика: Кол-во значений = 100						
		Гистограмма	Минимум	Макси...	Среднее	Стандартное откл.	Сумма	Σ^2 Сумма квадратов
1	9.0 Фактор 1		-1,98959119	3,0994371	-7,21644966E-18	1	-7,21644966E-16	99
2	9.0 Фактор 2		-1,865959871	4,128719456	-1,371125435E-16	1	-1,371125435E-14	99
3	9.0 Фактор 3		-3,087392686	2,040850474	1,298960939E-16	1	1,298960939E-14	99

Рис.19. Статистичні дані по кожному фактору (квартимакс).

Завдання для лабораторної роботи 2.

1. Створити файл Lab2.xlsx.
2. Занести у файл таблицю із 10 показниками соціально-економічного та демографічного розвитку України по областям за 2009-2012 р.р. (показники довільні, на вибір студента).
3. Перевести дані за кожен окремий рік у безрозмірний формат (нормалізація даних на відрізку [0,1]).
4. Провести факторний аналіз даних із використанням різних методів повороту та без повороту.
5. Провести порівняльний аналіз результатів факторного аналізу.

Тема 3. Багатовимірні моделі даних та алгебраїчні функції різного типу

Задачі: Розбиття гіперкубу даних на групи, розбиття даних по датах, квантування даних по інтервалах, фільтрація даних, групування даних. Ковзаюче вікно. Обчислення неперервних, гладких, кусково-гладких, розривних функцій одного та двох аргументів.

Методи: Процедури роботи з OLAP-кубом даних. Побудова двох- та трьох- вимірних графіків функцій та діаграм різного типу.

1. Основні поняття багатовимірних моделей даних

Однією з більш важливих задач Data Mining є робота з багатовимірними масивами даних [1]. Масив даних – це впорядкований набір фіксованої кількості даних одного типу, що мають порядковий номер та спільне ім’я. **Багатовимірним масивом даних** називають масив кожним елементом якого є масив. **Багатовимірна модель даних** [5], [16] характеризується багатовимірним представленням структур даних. Основними елементами моделі є гіперкуб даних, вимір, атрибут, комірка та значення. OLAP (On-Line Analytic Processing) - куб даних містить один або декілька вимірів та складається з набору впорядкованих комірок що можуть містити дані (значення) певного типу. Вимір – множина атрибутів що утворюють одну грань гіперкуба. Прикладами багатовимірних масивів даних можуть бути сукупність даних про клієнтів банку, сукупність даних про характеристики та продаж певної групи товарів, та ін. На відміну від звичайних однорідних за типом багатовимірних масивів даних доступ до елементів гіперкубу здійснюється як за вказівкою набору індексів-вимірів так і за їх підмножиною, що надає можливість отримати інформації у вигляді множини елементів. Розглянемо основні операції з гіперкубами.

2. Операції з багатовимірними масивами даних

2.1. Розбиття даних на групи

Для швидкої обробки даних їх розбивають на групи за допомогою критеріїв встановлених користувачем. Така процедура може виконуватися, наприклад, коли аналітик хоче продивитися не всю інформацію, а тільки по окремих групах. Наприклад, продаж групи товарів тільки національного виробництва, кредити надані особам старше 50 років, та ін. Відповідно, моделі прогнозу побудовані на таких даних будуть враховувати тільки дану групу об’єктів. Таким чином, по-перше, визначається вплив зміни деякої групи факторів на загальну модель; по-друге, розбиття на групи може виявити внесок кожної групи вихідних параметрів у кінцевий результат.

Імпорт даних здійснюється з файлу task31.txt. Після завершення процесу необхідно перевірити правильність типів усіх стовпчиків та обрати спосіб відображення даних “Таблиця”, показаний на Рис.1. Після імпорту даних за допомогою кнопки  завантажуємо вікно “Мастер обработки”. Обираємо розділ “Настройка набора данных”, Рис.2.

Таблица | Статистика

1 / 30

КанНомер	Дата прийняття	Вік	Стать	Розряд(1-6)	Заробіт.пл.(гр.)	Премія(гр.)
2104	13.09.2010	42	М	5	14000	5000
2105	10.11.2010	35	Ж	5	14000	2000
2106	14.12.2010	45	Ж	2	10000	4000
2107	16.09.2011	33	М	5	12000	5000
2108	17.10.2011	52	М	5	15000	5000
2109	18.10.2011	41	М	2	14000	3000
2110	10.09.2012	51	Ж	5	12000	5000
2111	20.09.2012	53	Ж	5	10000	4000
2112	21.11.2012	45	Ж	2	10000	3000
2113	22.09.2013	35	М	2	15000	4000
2114	12.11.2013	28	Ж	2	13000	4000
2115	24.09.2014	42	М	4	9000	2000
2116	25.03.2015	51	М	5	9000	5000
2117	16.04.2015	30	М	6	12000	5000
2118	21.04.2015	34	М	6	11000	5000
2119	28.09.2015	39	Ж	2	10000	2000
2120	16.10.2015	46	М	6	12000	2000
2121	10.12.2015	42	М	4	12000	2000
2122	01.03.2016	29	Ж	5	12000	2000
2123	02.05.2016	54	М	5	14000	5000
2124	11.05.2016	40	Ж	5	10000	4000
2125	04.02.2017	55	М	4	15000	4000
2126	05.02.2017	52	М	3	12000	4000
2127	06.05.2017	30	Ж	4	15000	3000
2128	04.07.2017	32	Ж	5	12000	4000
2129	08.07.2017	35	М	5	13000	3000
2130	19.10.2017	55	М	6	15000	4000
2131	10.02.2018	31	М	3	9000	5000
2132	11.02.2018	42	М	3	9000	3000
2133	03.04.2018	46	Ж	2	15000	4000

Рис.1. Таблиця з імпортованими даними.

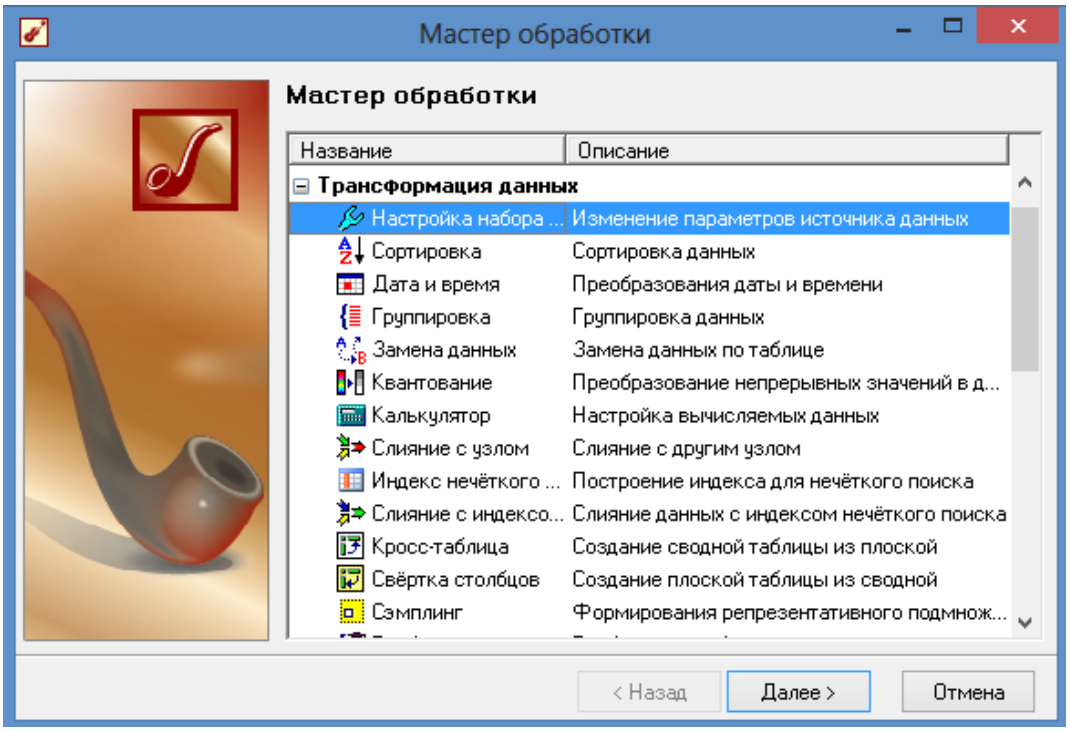


Рис.2. Вікно майстру обробки. Налаштування обробки даних.

Далі стовпчикам “Вік” та “Розряд(1-6)” встановлюємо Назначение – “Измерения”, “Премія” – Факт, решта стовпчиків – “Информационное” (Рис.3).

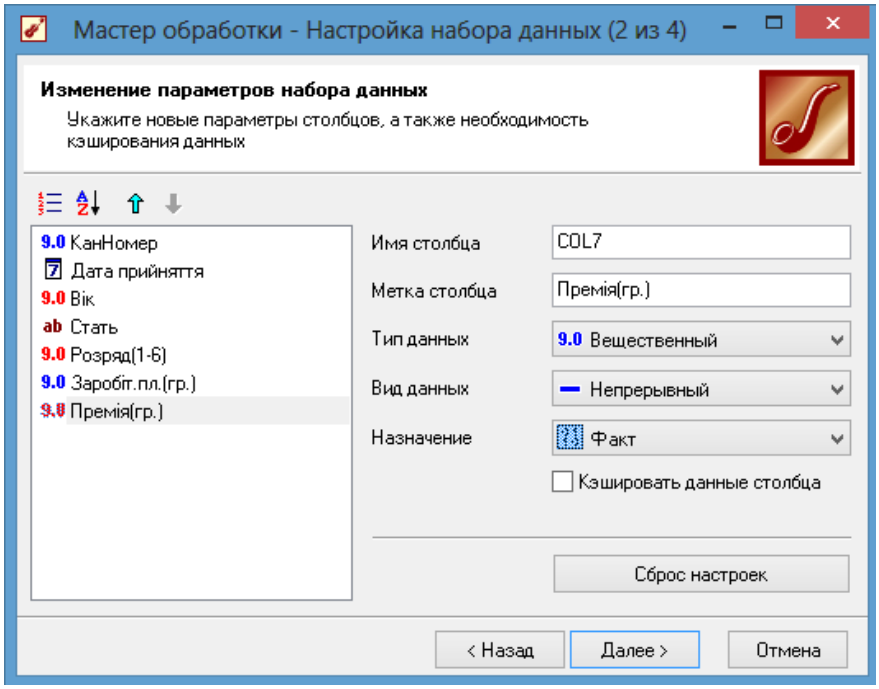


Рис.3. Налаштування атрибутів стовпчиків.

Обираємо засоби візуалізації “Табличные данные”, “Графики” та “OLAP – анализ” (Рис.4).

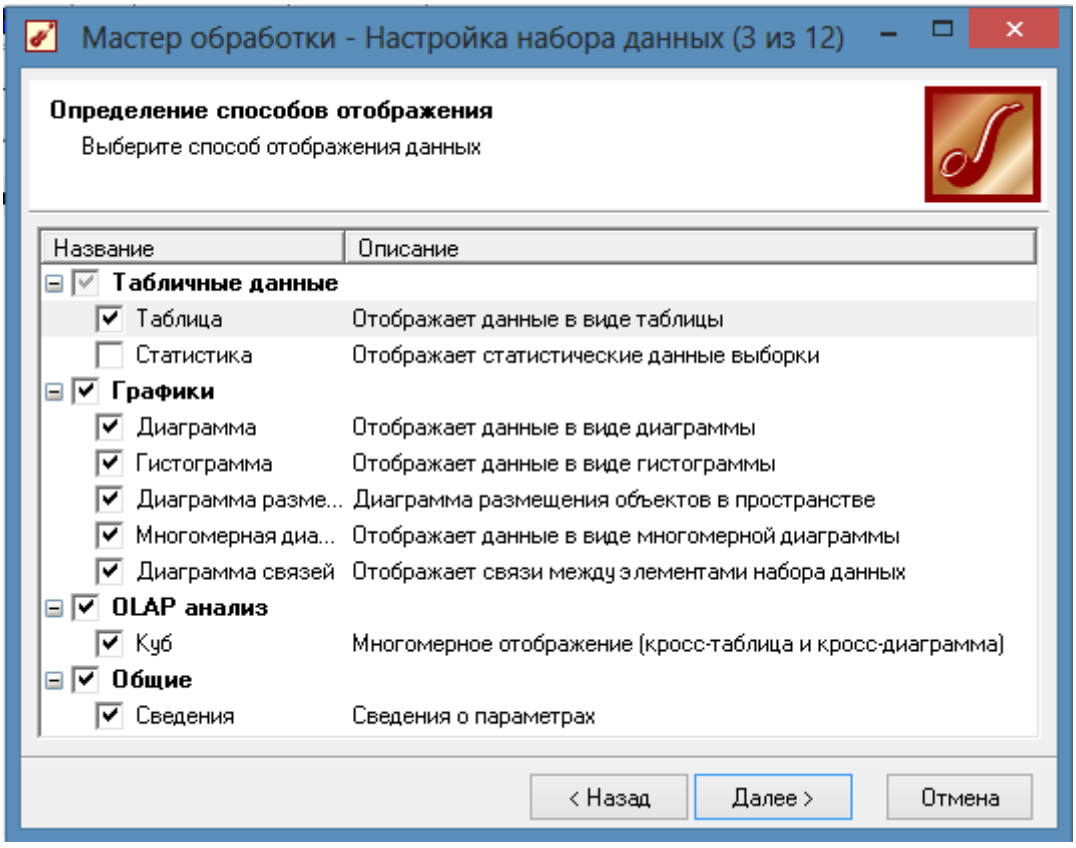


Рис.4. Вибір засобів візуалізації даних.

Далі на 9 кроці з’явиться вікно налаштування полів кубу, Рис.5.

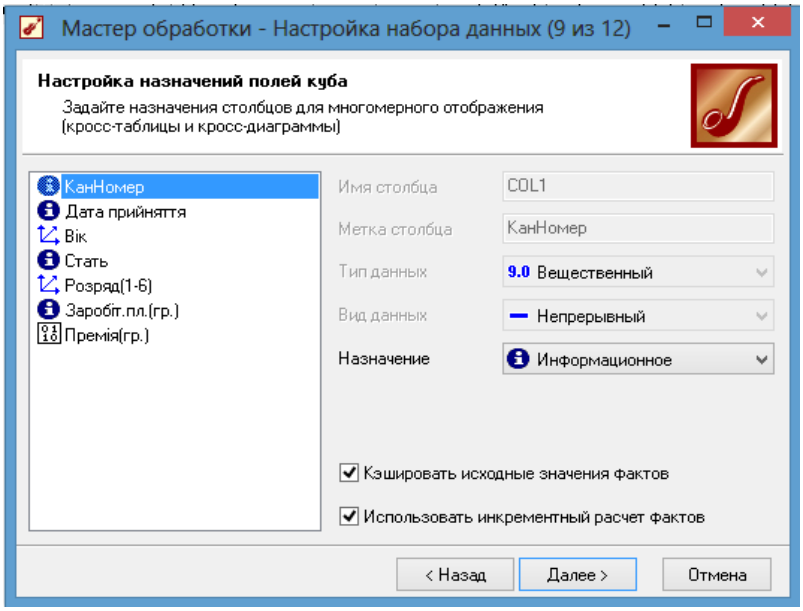


Рис.5. Налаштування полів кубу.

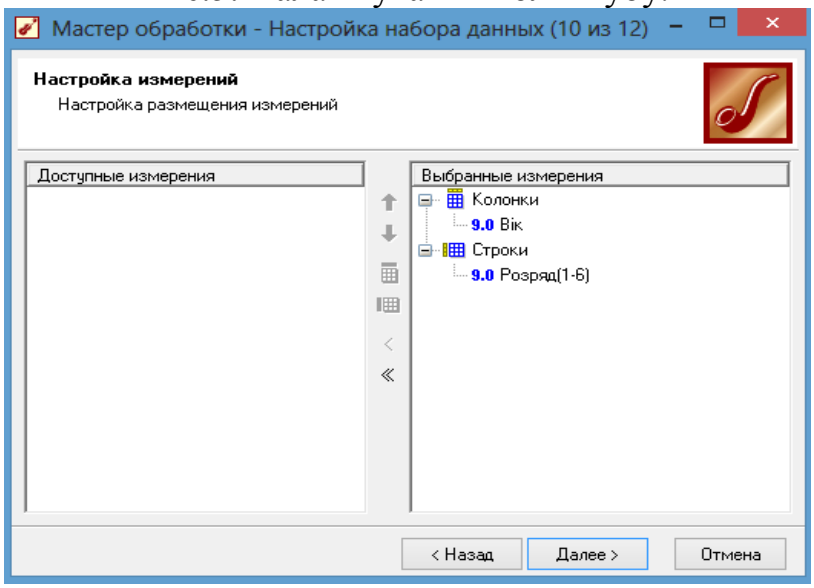


Рис.6. Налаштування вимірів кубу: Колонки – “Вік”, Строки – “Розряд(1-6)”.

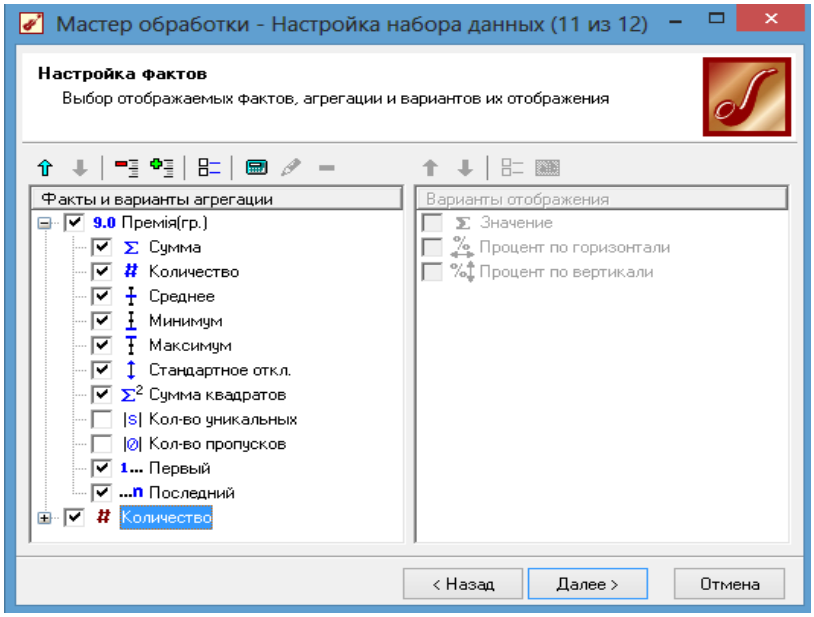


Рис.7. Налаштування фактів для визуализации: Сумма, Количество, Среднее, Минимум, Максимум, Стандартное отклонение, Сумма квадратов.

Результат буде відображений на вкладці Куб (верхня частина меню). У випадаючих списках “Вік” та “Розряд(1-6)” можна встановлювати зрізи по різним полям. Також доступні різні опції налаштувань на панелі інструментів.

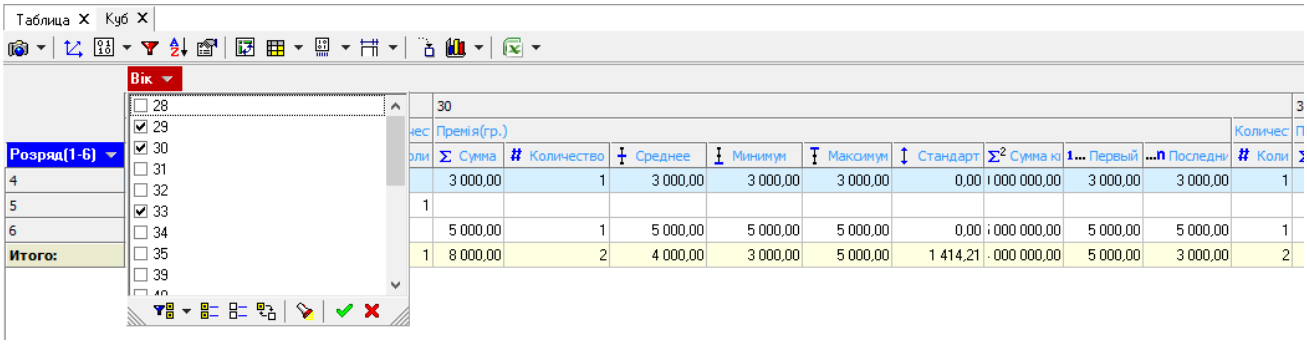


Рис.8. Крос-таблиця Куб. У випадаючих списках “Вік” та “Розряд(1-6)” можна встановлювати зрізи по різним полям.

2.2 Розбиття даних по датах

Розбиття даних по датах дозволяє провести аналіз зміни груп параметрів за повний період часу по певним інтервалам (наприклад, за день, за місяць, за квартал, тощо). Це означає що для стовпчика, що містить інформацію з датою, формується другий стовпчик, в якому вказується, до якого заданого інтервалу часу належить рядок даних. Тип часового інтервалу задається аналітиком.

Для проведення розбиття по датах встановлюємо курсор на вузлі з імпортованими з файлу task31.txt даними. Далі у вікні “Мастер обработки” обираємо “Дата и Время” (Рис.9).

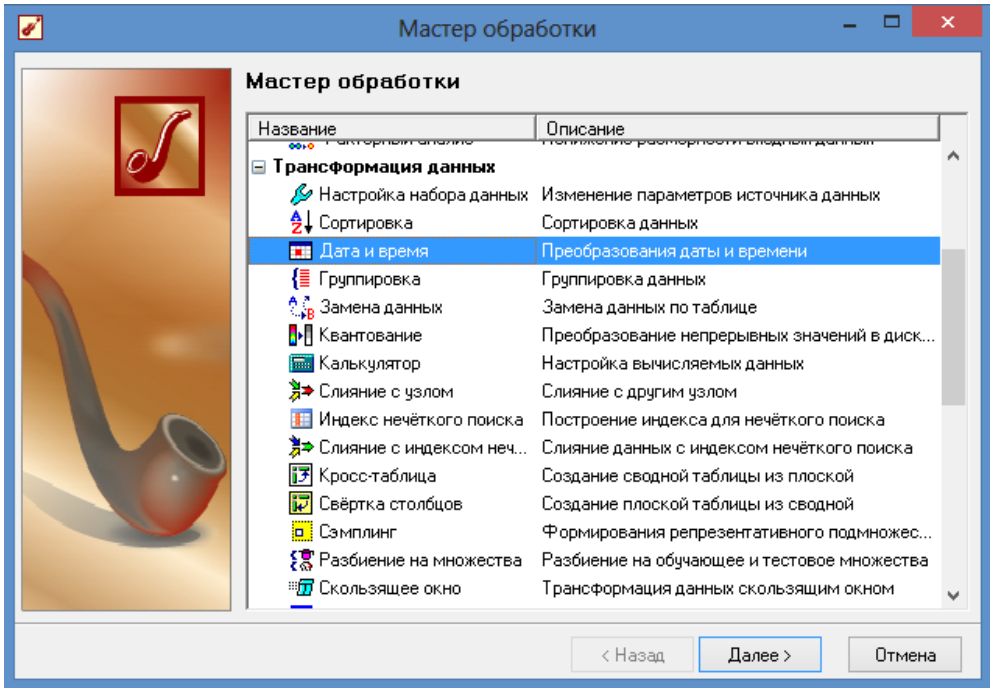


Рис.9. “Дата и Время” в майстра обробки.

У наступному вікні (Рис.10) обираємо “Год + Месяц”, вказавши рядок з датою як “Используемое”.

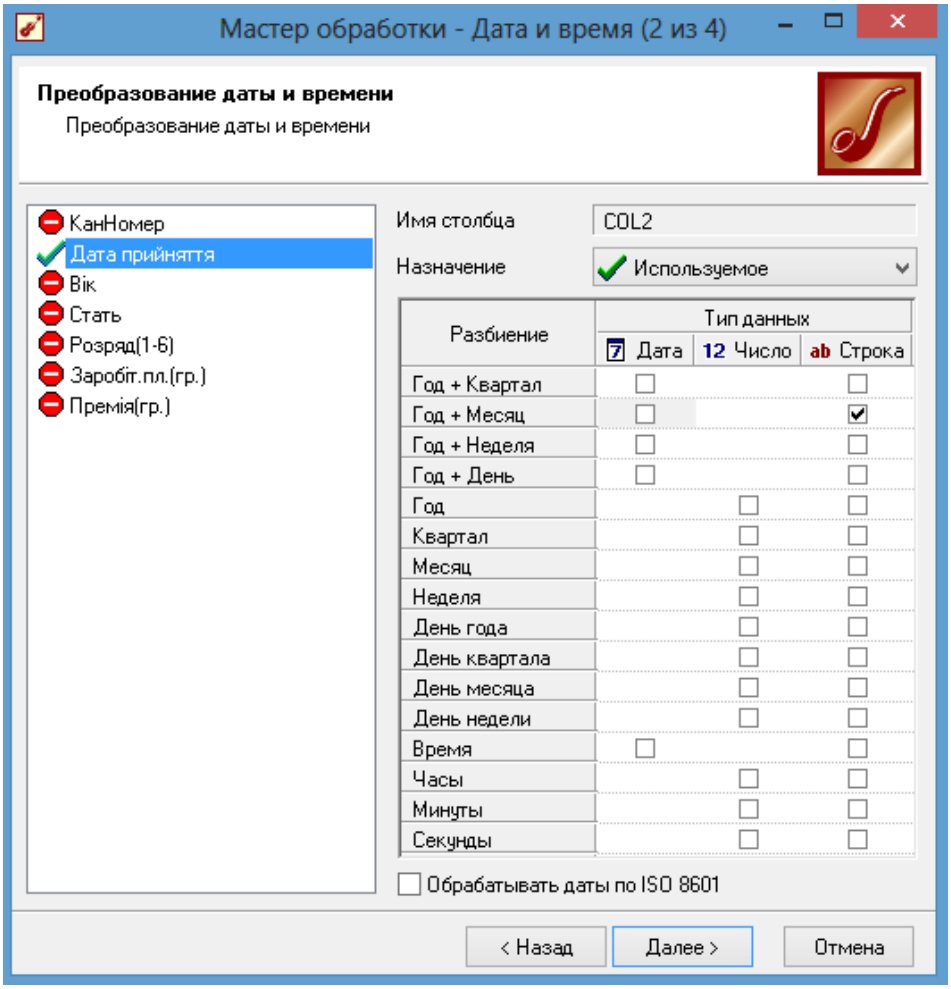


Рис.9. Перетворення дати та часу.

Обираємо тип візуалізації OLAX-куб. Відмічаємо “Стаж работы” і “Дата принятия (Год+Месяц) ” як виміри, а “Заробіт.пл.(грн.)” як факт.

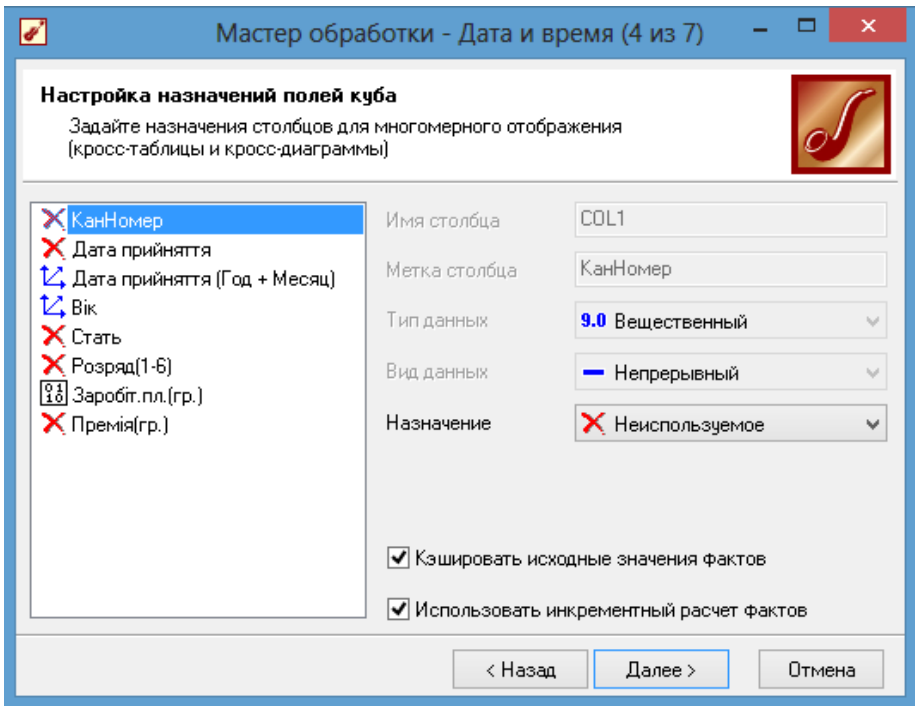


Рис.10. Налаштування полів OLAX-куба.

Далі встановлюємо “Дата прийняття (Год+Месяц) ” виміром по стовпцях, а “Вік” - по рядках (Рис.10). Налаштування функцій для фактів у наступному вікні показано на Рис.11.

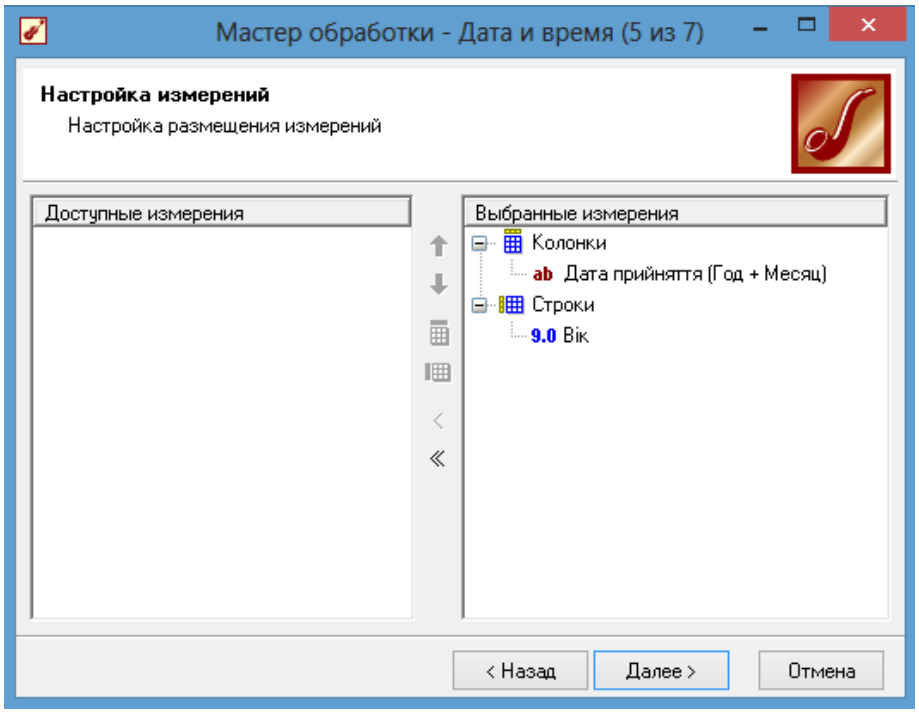


Рис.10. Налаштування вимірів OLAX-куба

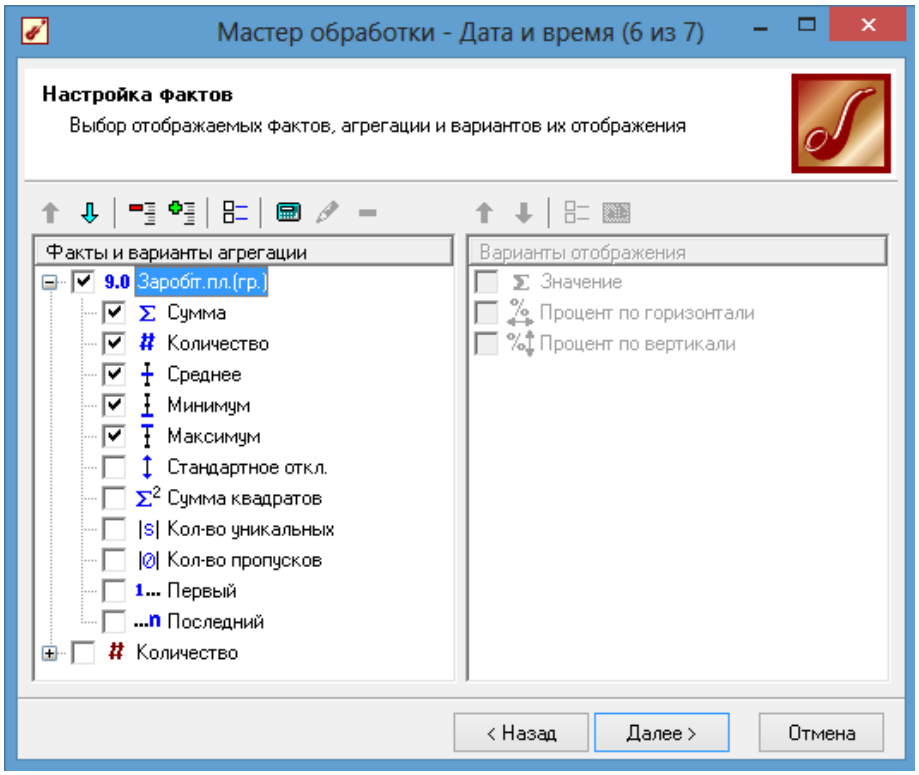


Рис.11. Налаштування функцій для фактів.

Кінцевий результат – розбиття даних по датах (по місяцях) показано на Рис.12 у вигляді куба.

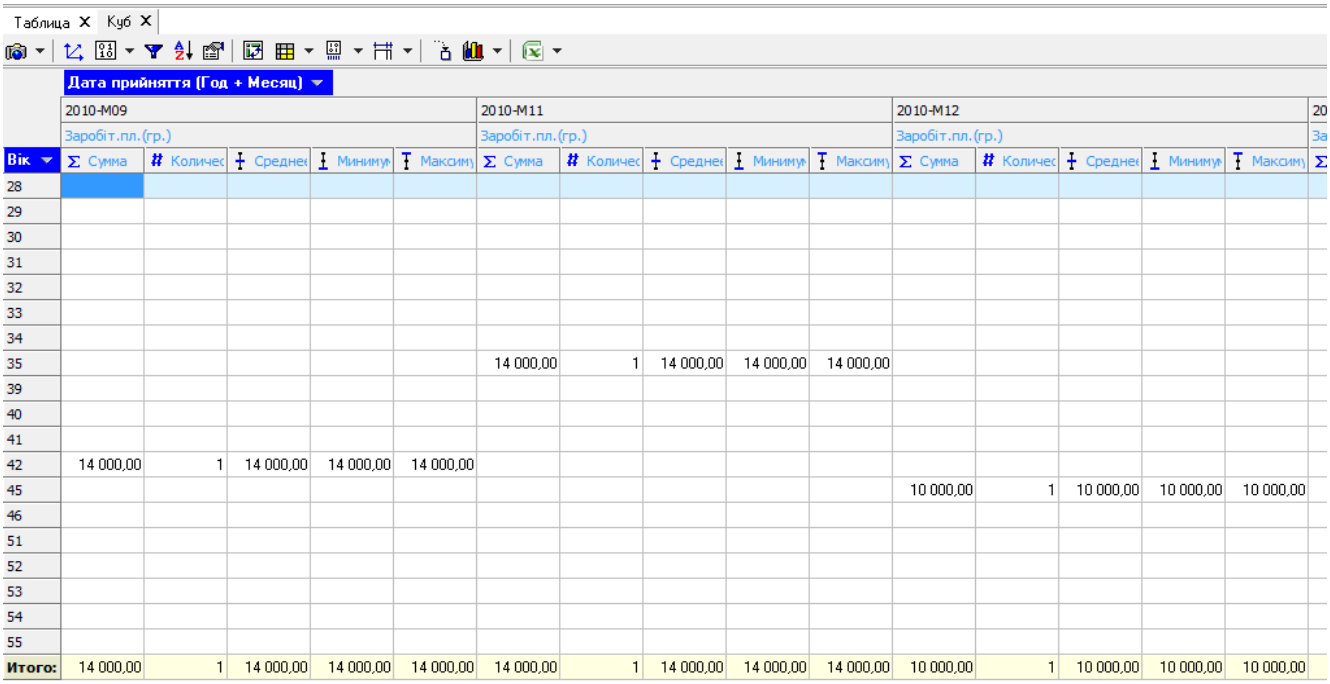


Таблица ✕ Куб ✕

Дата прийняття (Год + Месяц)

Вік	2010-M09 Заробіт.пл.(гр.) Сумма	Количес	Среднее	Минимум	Максим	2010-M11 Заробіт.пл.(гр.) Сумма	Количес	Среднее	Минимум	Максим	2010-M12 Заробіт.пл.(гр.) Сумма	Количес	Среднее	Минимум	Максим
28															
29															
30															
31															
32															
33															
34															
35						14 000,00	1	14 000,00	14 000,00	14 000,00					
39															
40															
41															
42	14 000,00	1	14 000,00	14 000,00	14 000,00										
45											10 000,00	1	10 000,00	10 000,00	10 000,00
46															
51															
52															
53															
54															
55															
Итого:	14 000,00	1	14 000,00	14 000,00	14 000,00	14 000,00	1	14 000,00	14 000,00	14 000,00	10 000,00	1	10 000,00	10 000,00	10 000,00

Рис.12. Результат розбиття даних по місяцях у вигляді куба.

2.3. Квантування даних по інтервалах

Іноді виникає необхідність розбити данні неперервного або дискретного типу на скінченні набори даних, наприклад, виконати квантування по інтервалах. Така процедура означає перехід від даних неперервного (дискретного) типу до дискретного. Квантування може бути інтервальним або квантильним. У першому випадку діапазон значень вхідної змінної розбивається на певну кількість однакових інтервалів (їх кількість задається аналітиком). Після чого кожному значенню вхідного параметру ставиться у відповідність ідентифікатор - номер інтервалу, значення нижньої або верхньої границь інтервалу розбиття, середнє значення інтервалу розбиття або мітка інтервалу (у ручному або автоматичному режимах). Інтервали розбиття містять в собі тільки нижню границю, окрім останнього інтервалу, який містить обидві границі. Наприклад, якщо значення неперервного параметру належать відрізку [1,100], то при розбитті на 10 інтервалів отримуємо відрізки від 1 до 10, від 10 до 20, …, від 90 до 100. Значення 1 буде відноситися до 1-го інтервалу, значення 10 – до 2-го інтервалу, …, значення 90 та 100 – до 10-го інтервалу.

У випадку квантильного квантування інтервал значень неперервної вхідної змінної розбивається на інтервали, які містять у собі однакову (або приблизно однакову) кількість значень вхідних даних. С точки зору теорії ймовірності такі інтервали можна розглядати як множини, куди потрапляють значення даних з майже рівною ймовірністю. Якщо кількість вхідних даних не кратна числу інтервалів розбиття, то кількість елементів що потрапляють в кожен інтервал може відрізнятися на одиницю від кількості елементів в деяких інших інтервалах. Тоді такі інтервали мають невеличку перевагу у кількості елементів перед іншими інтервалами.

Налаштування процедури квантування полягає у виборі методу - інтервального або квантильного, задані кількості інтервалів, виборі (із випадаючого списку) ідентифікатору вихідного значення змінної – номеру інтервалу, мітки інтервалу (автоматичної мітки інтервалу), його нижньої (верхньої) границі або середини ін-

тервалу та визначенні типу даних (дискретний або неперервний). У разі вибору мітки інтервалу необхідно визначити також назву (мітку) кожного інтервалу розбиття. Визначення границь інтервалів відбувається в автоматичному режимі після запуску методу. Але у разі необхідності, наприклад, коли початковий набір даних далеко не покриває усю множину значень вхідної змінної, аналітик може скорегувати границі інтервалів в ручному режимі з клавіатури.

Обираємо знову вузол з імпортованими з файлу task31.txt даними. У вікні майстра обробки обираємо “Квантование” (Рис.9). Далі у вікні квантування (Рис.13) встановлюємо поле “Премія” як “Используемое”, інтервальний спосіб квантування, кількість інтервалів 6 та значення “Середина интервала”.

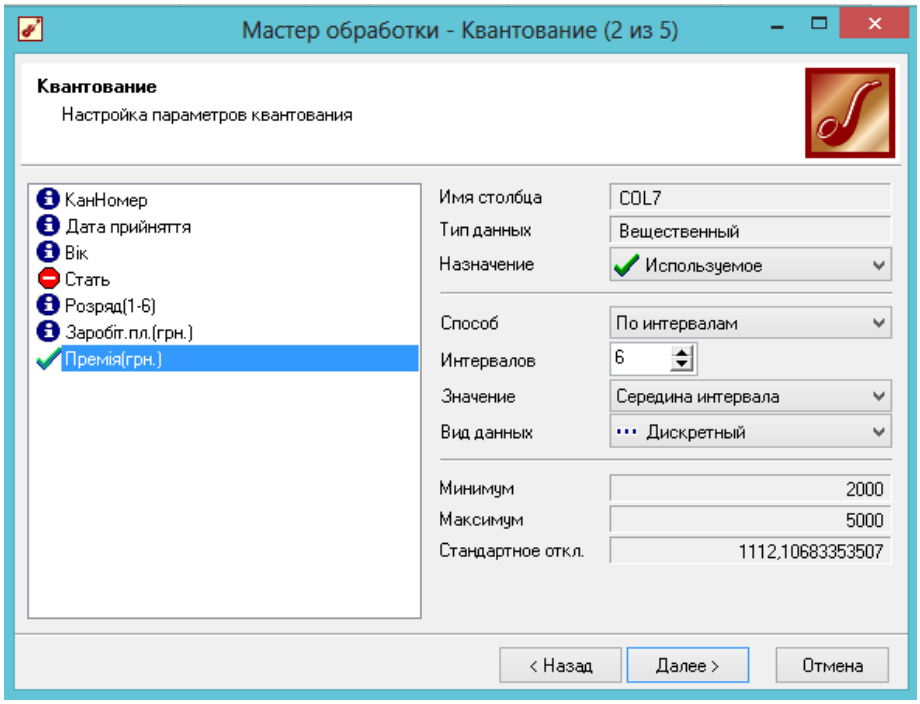


Рис.13. Налаштування параметрів квантування по інтервалам.

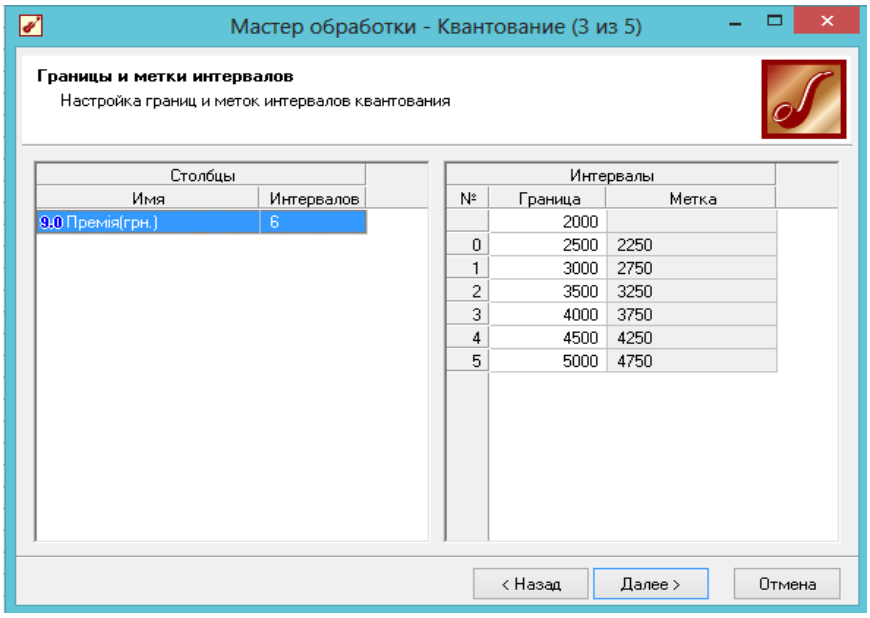


Рис.14. Корегування границь інтервалів в ручну.

Далі необхідно обрати тип візуалізації - “Таблица” та “Куб”. У майстру налаштувань полів куба обираємо “Премія” та “Вік” як виміри, а “Заробіт.пл.(грн.)” як

факт (Рис.14, 15). Далі “Вік” встановлюємо як стовпчики, “Премія” - як рядки (Рис. 16). Отримуємо результат у вигляді OLAX-куба (Рис.17).

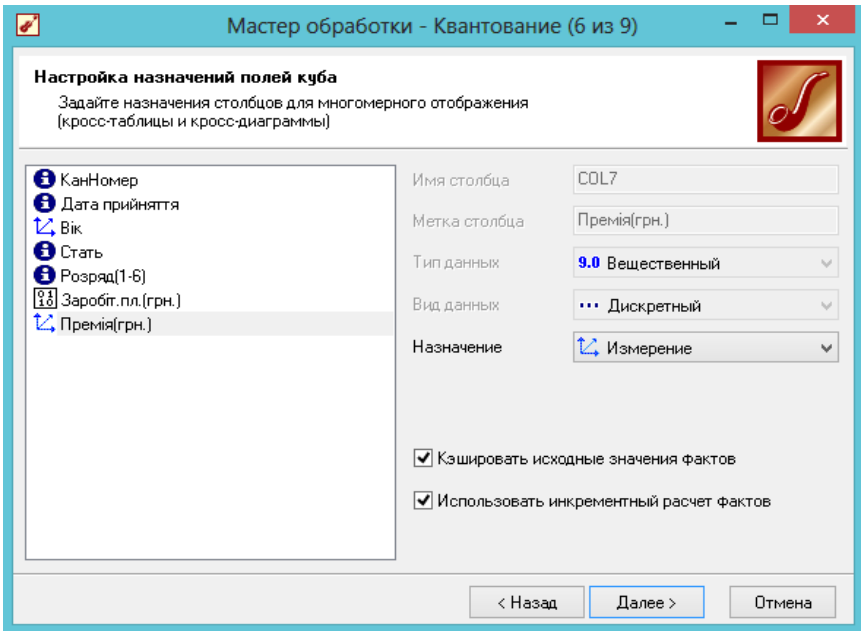


Рис.15. Налаштування полів кубу.

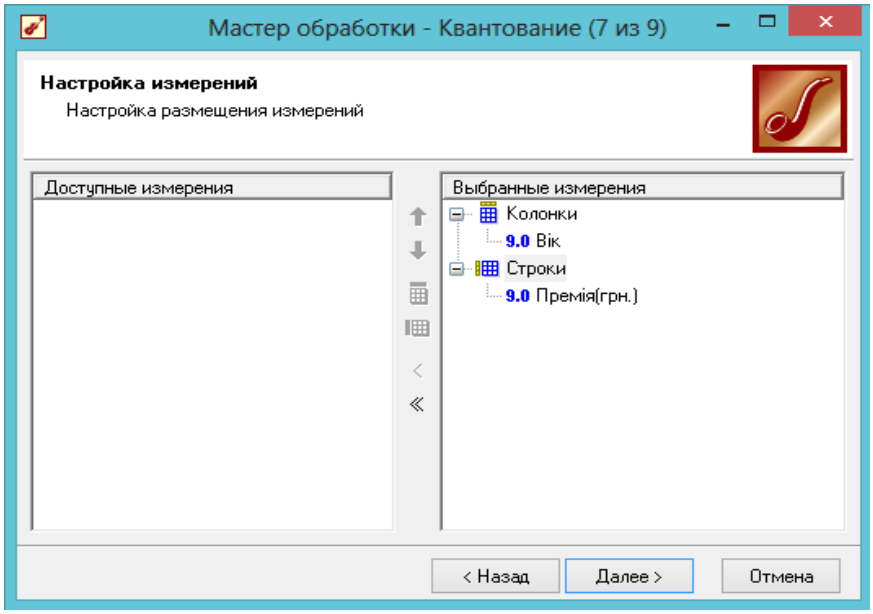


Рис.16. Налаштування вимірів OLAX-куба.

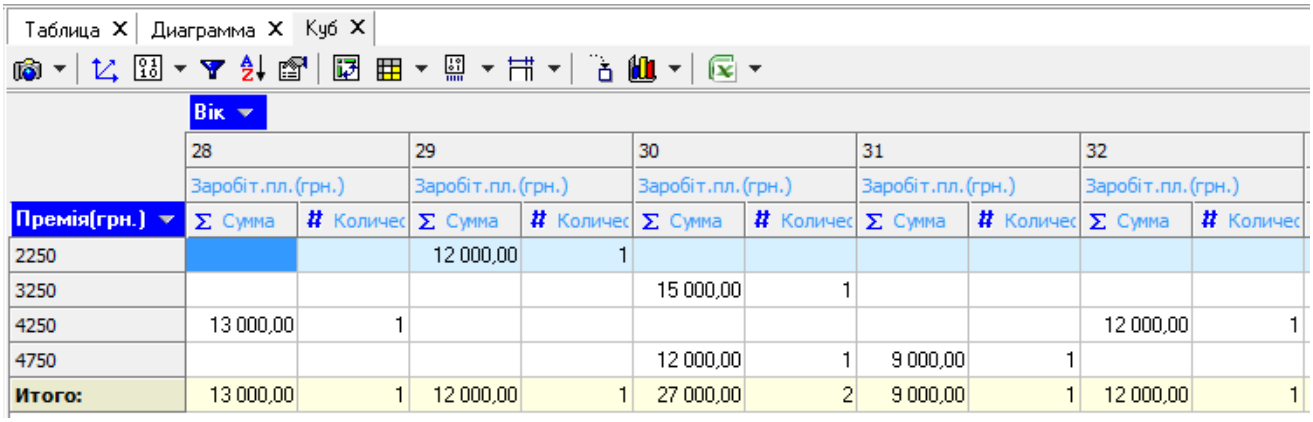


Рис.17. Квантування на інтервали, OLAX-куб.

2.4. Фільтрація даних

Фільтрація даних – це процедура пошуку та відбору записів в таблицях баз даних відповідно до встановлених аналітиком критеріїв. За допомогою фільтрації можна розбивати дані таблиць на групи по інтервалам, по датам, тощо. Фільтри також спрощують процес доступу до даних та їх редагування. Слід відзначити, що фільтровані дані не видаляються з таблиць, а залишаються прихованими на своїх місцях.

У Deductor Studio для налаштування процесу фільтрації необхідно задати ім’я полів із випадаючого списку (назви змінних) та за необхідністю обрати одну із логічних операцій (“АБО”, “ТА”) між полями, для яких застосовується фільтрація. Для фільтрації даних можна використовувати налаштування за декількома операціями визначеними для декількох пар полів одночасно. Крім того необхідно вказати математичну умову (>, <, =, і т.д.) та значення, що використовується разом з цією умовою. В результаті усіх налаштувань створюється логічний вираз, що служить критерієм фільтрації. В результаті фільтрації отримуємо таблиці із значеннями даних, для яких логічний вираз приймає значення істина.

Безумовно, процедура фільтрації є дуже важливим інструментом обробки даних і може успішно використовуватися для групування даних, відсіювання не актуальних даних, допомагає пришвидшити доступ до певних наборів даних, тощо. Але для складних процедур аналізу даних в задачах Data Mining рекомендовано проводити попередню фільтрацію даних та імпортувати вже підготовлені набори даних для їх подальшої обробки. Такий підхід заощаджує час та обчислювальні ресурси програми.

Для імпортованих даних запускаємо майстер обробки та обираємо “Фильтрация”. У вікні прописується умова фільтрації. Встановлюємо “Вік” не менш 40, а “Заробіт.пл.(грн.)” – не менш, ніж 10000. Користуємось випадаючим списком для вибору поля та кнопками навігатора на панелі праворуч, Рис.18. Встановлюємо відповідні значення в “Условие и “Значение.

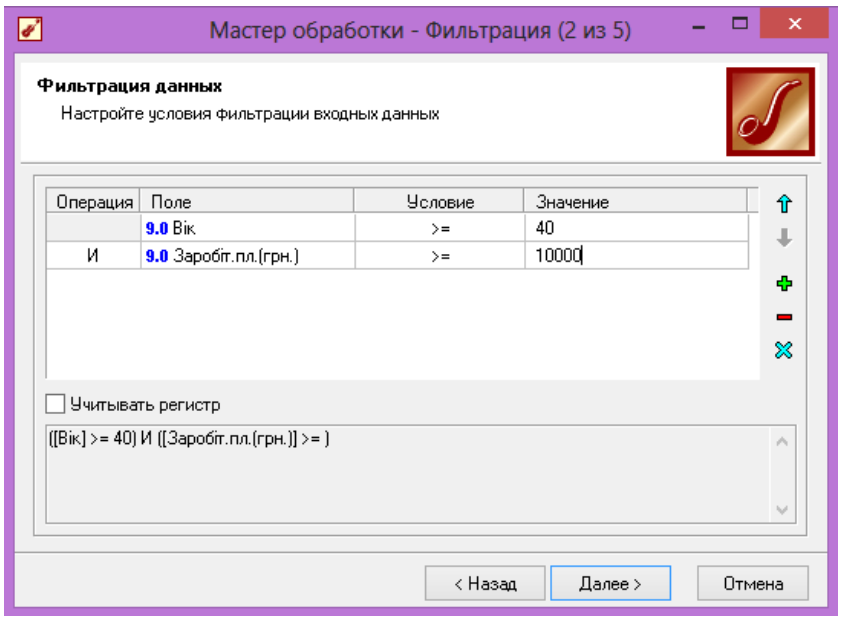


Рис.18. Налаштування фільтру даних.

Для візуалізації результатів використовується Таблиця та Статистика. Результати представлені на Рис.19, 20. В наборі даних виявлено 15 записів, що задовольняють умовам фільтрації.

Таблица | Статистика | Диаграмма

1 / 15

КанНомер	Дата прийняття	Вік	Стать	Розряд(1-6)	Заробіт.пл.(грн.)	Премія(грн.)
2104	13.09.2010	42	М	5	14000	5000
2106	14.12.2010	45	Ж	2	10000	4000
2108	17.10.2011	52	М	5	15000	5000
2109	18.10.2011	41	М	2	14000	3000
2110	10.09.2012	51	Ж	5	12000	5000
2111	20.09.2012	53	Ж	5	10000	4000
2112	21.11.2012	45	Ж	2	10000	3000
2120	16.10.2015	46	М	6	12000	2000
2121	10.12.2015	42	М	4	12000	2000
2123	02.05.2016	54	М	5	14000	5000
2124	11.05.2016	40	Ж	5	10000	4000
2125	04.02.2017	55	М	4	15000	4000
2126	05.02.2017	52	М	3	12000	4000
2130	19.10.2017	55	М	6	15000	4000
2133	03.04.2018	46	Ж	2	15000	4000

Рис. 19. Таблиця з фільтрованими даними.

Таблица | Статистика | Диаграмма

	Метка столбца	Статистика: Кол-во значений = 15					
		Гистогра...	Минимум	Максимум	Среднее	Сумма	Сумма кв...
1	9.0 КанНомер		2104	2133	2117,466667	31762	67256198
2	Дата прийняття		13.09.2010	03.04.2018	19.07.2014 8:00:00		
3	9.0 Вік		40	55	47,93333333	719	34875
4	ab Стать		1	1	1	15	15
5	9.0 Розряд(1-6)		2	6	4,066666667	61	279
6	9.0 Заробіт.пл.(грн.)		10000	15000	12666,66667	190000	2464000000
7	9.0 Премія(грн.)		2000	5000	3866,666667	58000	238000000

Рис.20. Статистичні покажчики фільтрованих даних.

2.5. Групування даних

Метод групування даних має на меті розділення усіх даних на групи за встановленими користувачем ознаками. Для ефективної обробки великих об’ємів інформації доцільно проводити аналіз не усіх даних одночасно, а деяких відокремлених за певними критеріями відбору груп даних. Тоді можна зробити

висновки про локальні властивості даних у групах та виокремити в них певні закономірності. Так же само аналіз окремих груп даних може виявити закономірності та нові властивості усього набору даних. Наприклад, групування студентів одного курсу факультету за рівнем успішності навчання може виявляти вплив високої успішності окремих студентів на середню успішність усього курсу і виявити резерви, що приховані у зростанні успішності усіх студентів курсу до рівня найкращих.

Для імпортованих даних запускаємо майстер обробки та обираємо “Группировка” (Рис.9). Вказуємо “Стать” як вимір, “ Заробіт. пл.(грн.)” як факт з функцією агрегації “Среднее”. Решта даних не використовується.

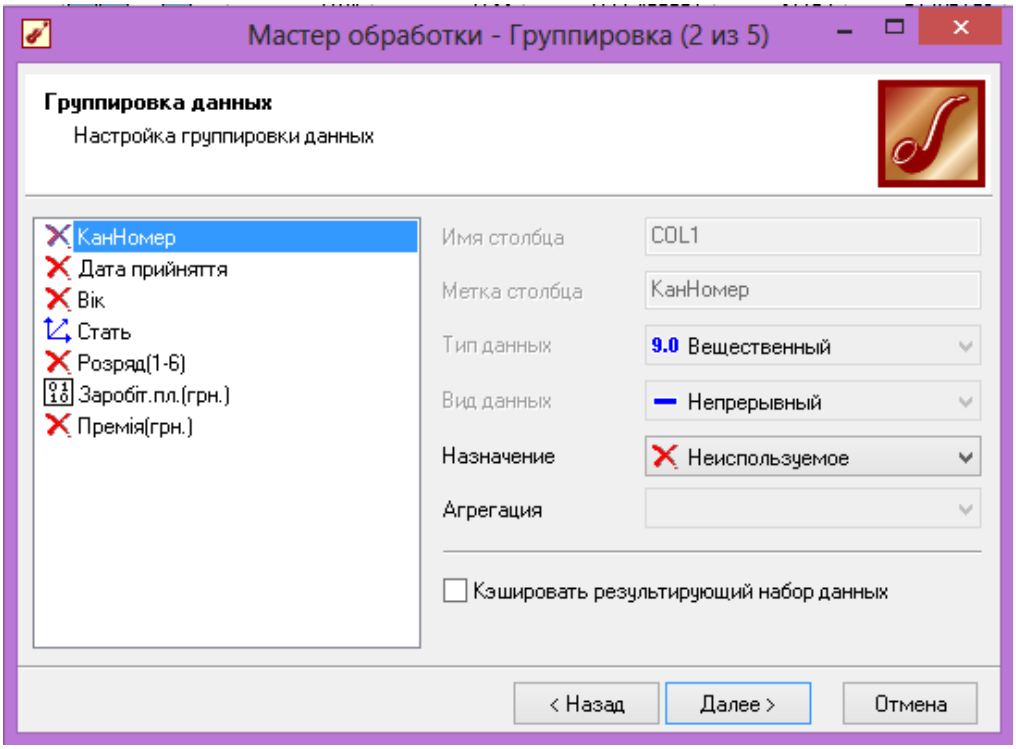


Рис.21. Налаштування групування даних.

Результат групування у вигляді таблиці показаний на Рис.22.

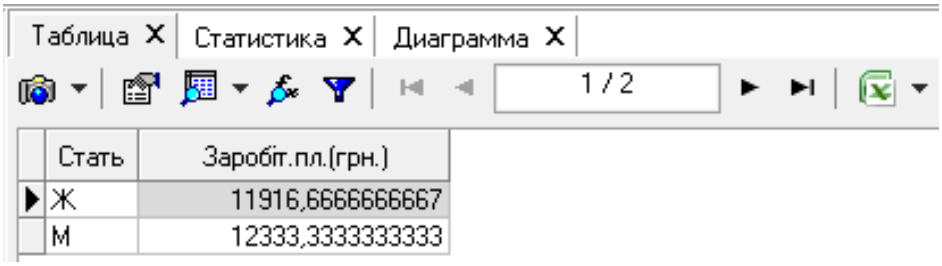


Рис.22. Результат групування.

2.6. Ковзаюче вікно

В задачах прогнозування часто зустрічаються часові ряди із періодичною поведінкою (сезонність). Періодичні закономірності рядів можна виявити за допомогою локальних максимумів автокореляційної функції. Часовий проміжок між двома сусідніми локальними максимумами авто кореляційної функції відповідає одному періоду повтору процесу що описується даним часовим рядом. Таким чином, якщо

змістити часовий ряд праворуч на декілька кроків, що відповідають одному такому періоду, то можна отримати набір ретроспективних даних для побудови моделей прогнозування на основі нейронних мереж та лінійних регресій (Тема 5). Такий метод побудови зсунутих в часі зв'язаних даних називається ковзаючим вікном [1] (вікно – оскільки фіксується графік усього ряду у вікні, ковзаюче – бо це вікно зміщується вперед або назад на визначену кількість кроків). Перетворення ковзаючого вікна має два параметри: "глибина занурення" - кількість "минулих" записів, які потрапили у вікно, і "горизонт прогнозування" - кількість "майбутніх" записів.

Слід відзначити, що для декількох перших (останніх) комірок масиву даних (відносно початку вибірки) при зсуві масиву уперед (назад) будуть формуватися записи, що містять порожні комірки. Метод ковзаючого вікна дозволяє користувачу не використовувати ці порожні записи у результуючій вибірки даних або навпаки використовувати результуючу вибірку даних з неповними записами. Вхідний набір зв'язаних даних має бути належним чином впорядкованим для коректного формування ковзаючого вікна.

Імпортуємо дані що містять часовий ряд з файлу task32.txt та відкриваємо майстер обробки. На першому етапі за допомогою авто кореляційної функції досліджуємо наявність періодичних процесів часового рядка. Слід зауважити що у разі наявності високочастотного шуму у імпортованих даних необхідно провести також попередню фільтрацію або придушення шуму (Тема 1). Початкова функція зображена на Рис.23, відповідна їй автокореляційна функція (Майстер обробки → Автокореляція, Глубина відліку 24) - на Рис.24. Автокореляційна функція виявляє період осциляцій початкової функції з лагом (кроком) 15 (Рис.24).

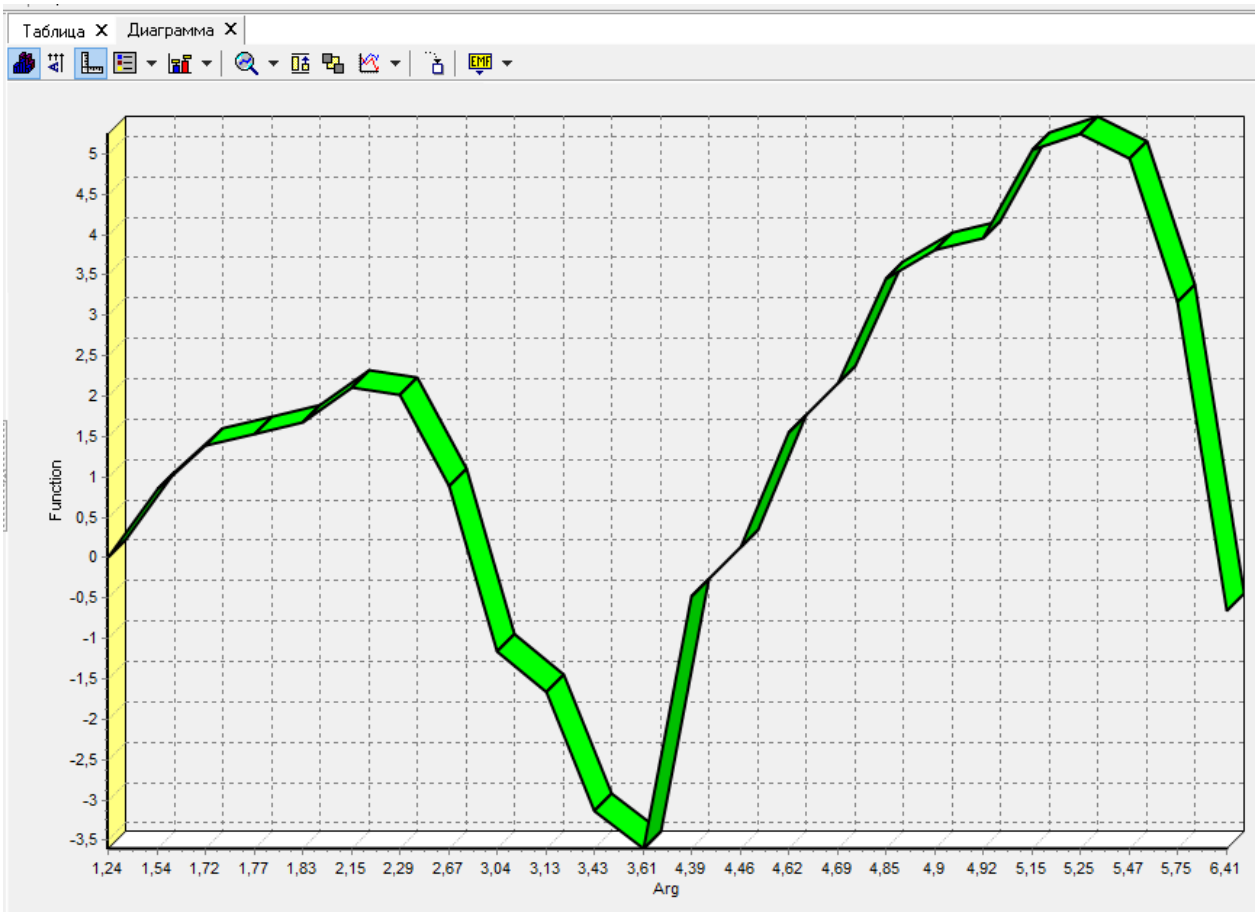


Рис.23. Початкова функція.

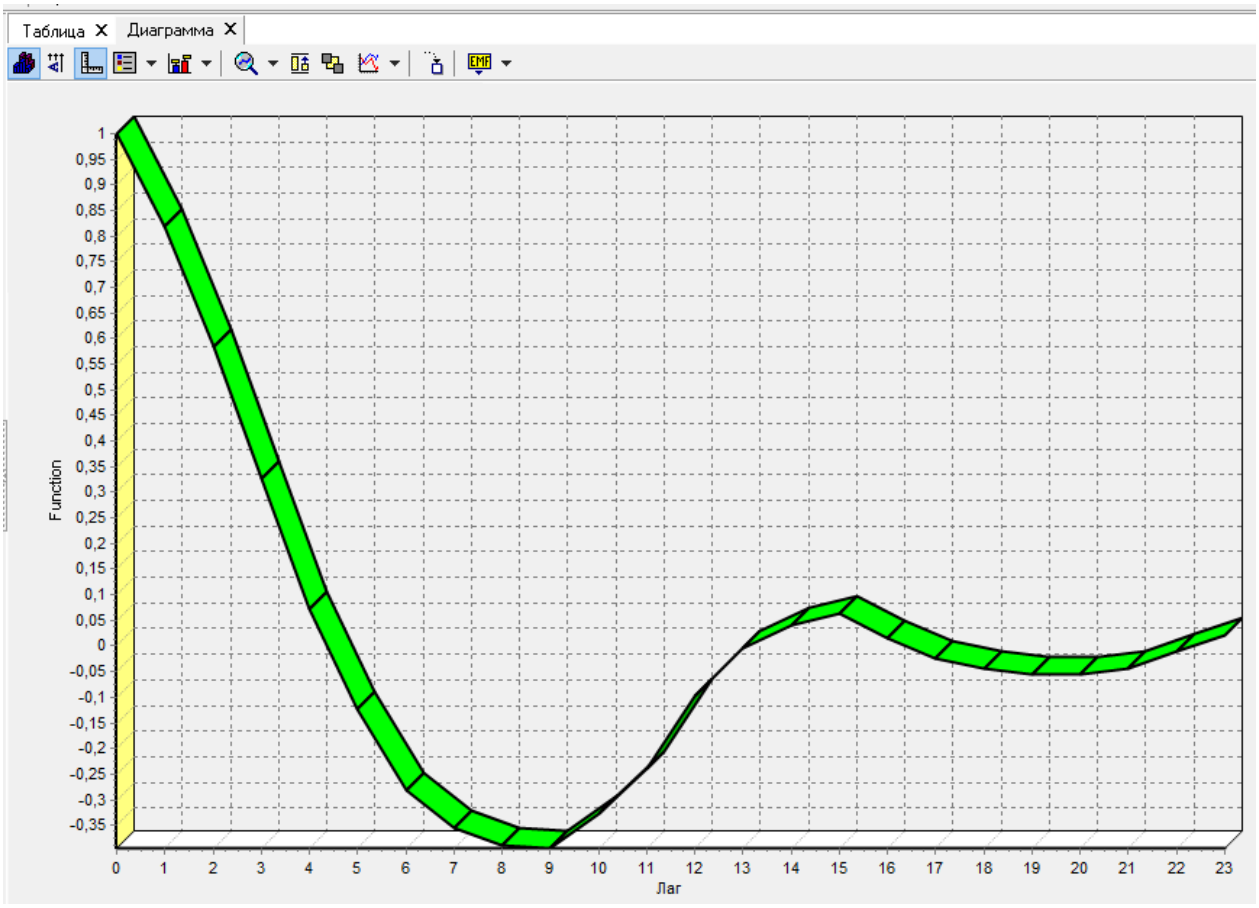


Рис.24. Автокореляційна функція.

У вікні майстра (Рис.9) відкриваємо “Скользящее окно”, вказуючи Arg як “Информационное”, Function - “Используемое”, “Глубина погружения” (зсув графіка вперед) дорівнює 15, а “Горизонт прогнозирования” (зсув графіка назад) – 0, відмічаємо “Оставлять неполные записи”.

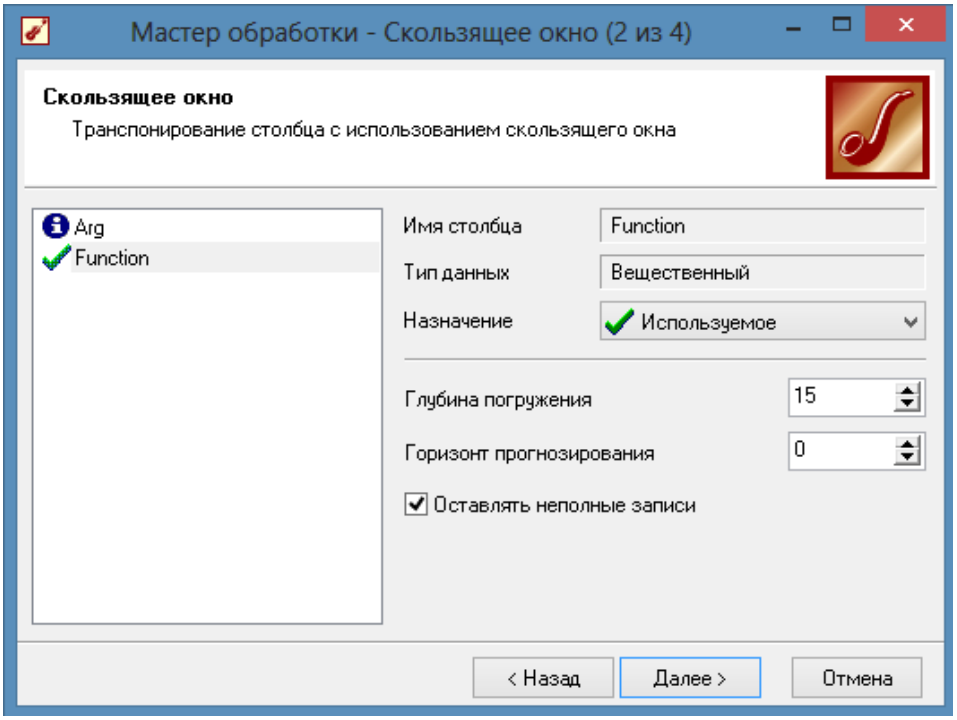


Рис.25. Налаштування ковзаю чого вікна.

Результат зсуву графіка у вигляді діаграми показано на Рис.26.

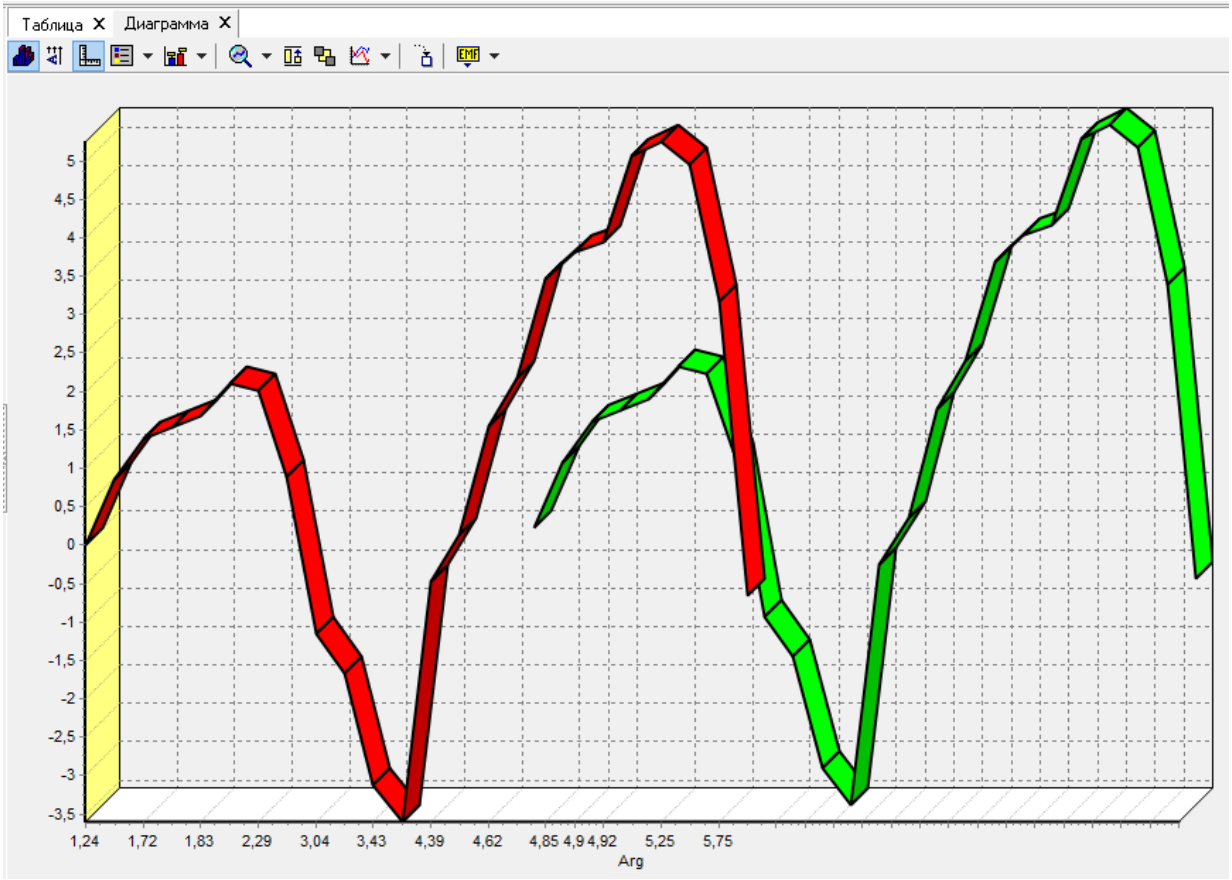


Рис.26. Початкова функція (праворуч) та її зсув на 15 кроків (ліворуч).

2.7. Обчислення неперервних, гладких, кусково-гладких, розривних функцій одного та двох аргументів

При аналізі зв'язаних масивів даних може виникнути необхідність у графічному представленні функції одного чи двох аргументів, гладких, кусково-гладких та розривних функцій. Кожна функція задається набором дозволених символів та математичних операцій та може використовувати задані умови розгалуження. За допомогою умов розгалуження можна будувати кусково-гладкі та розривні функції. Від того, яку функцію треба обчислити, залежить спосіб візуалізації результату. Наприклад, для функції двох аргументів необхідно вибирати багатовимірну діаграму як засіб візуалізації поверхні функції.

1. Неперервні функції одного аргументу $f(x) = (x-8)^2/6 - 8$, $x \in [0,23]$. Після імпорту даних з файлу task33.txt у вікні Майстра обробки обираємо пункт "Калькулятор" (Рис.9). У "Список выражений" записуємо Function1. З випадаючого списку "Функции" (праворуч) обираємо Pow(;) та підставляємо Arg1 (стовбчик імпортованих даних із файлу task34.txt) з випадаючої вкладки "Поля" та ступень 2, підставляємо константи, Рис.27.

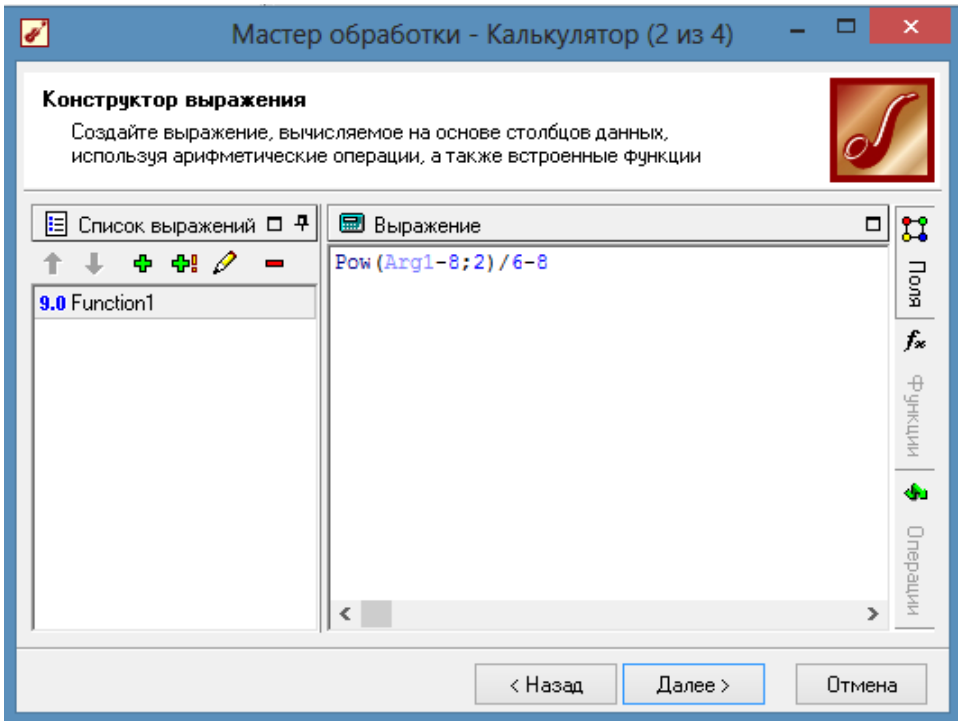


Рис.27. Налаштування конструктора виразу функції.

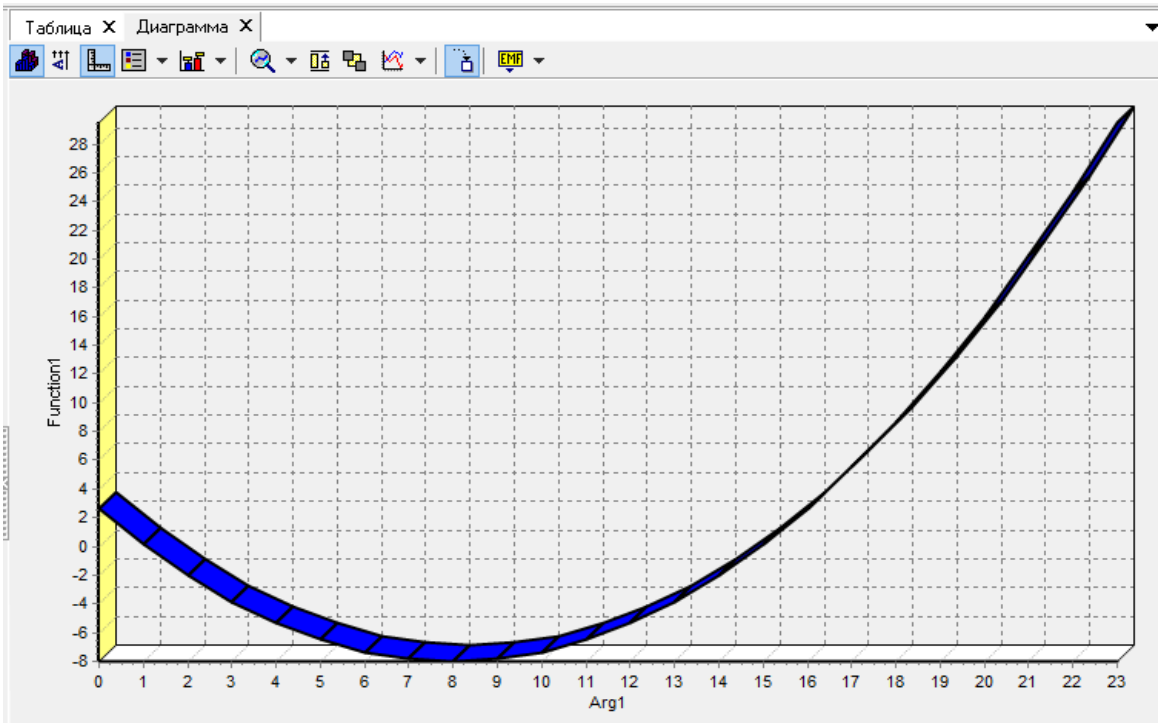


Рис.28. Графік неперервної функції однієї змінної.

2. Неперервна функція двох аргументів $f(x, y) = x^2 + y^2$, $x \in [0,9]$, $y \in [-4,5]$. Аналогічно налаштовуємо “Конструктор виражения” функції (Рис.29).

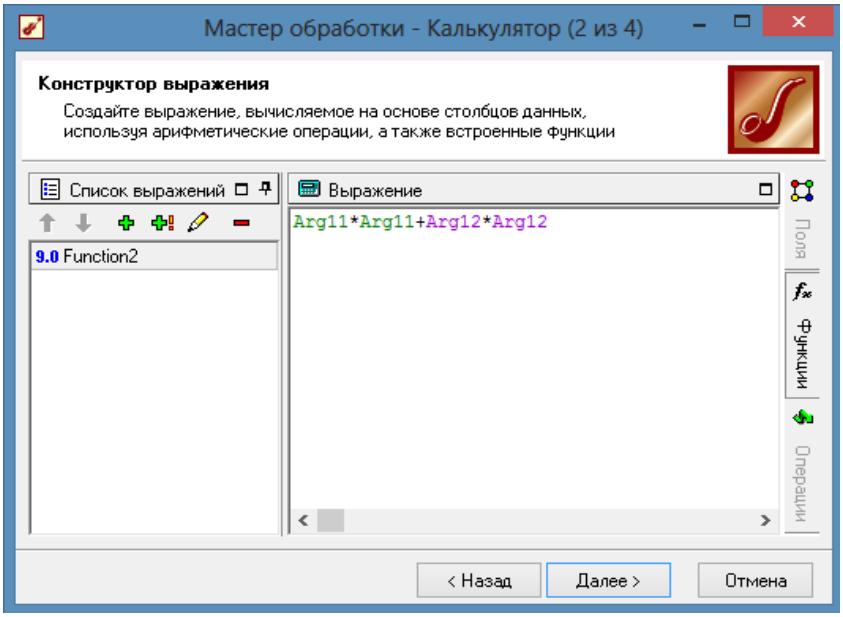


Рис.29. Налаштування конструктора виразу функції.

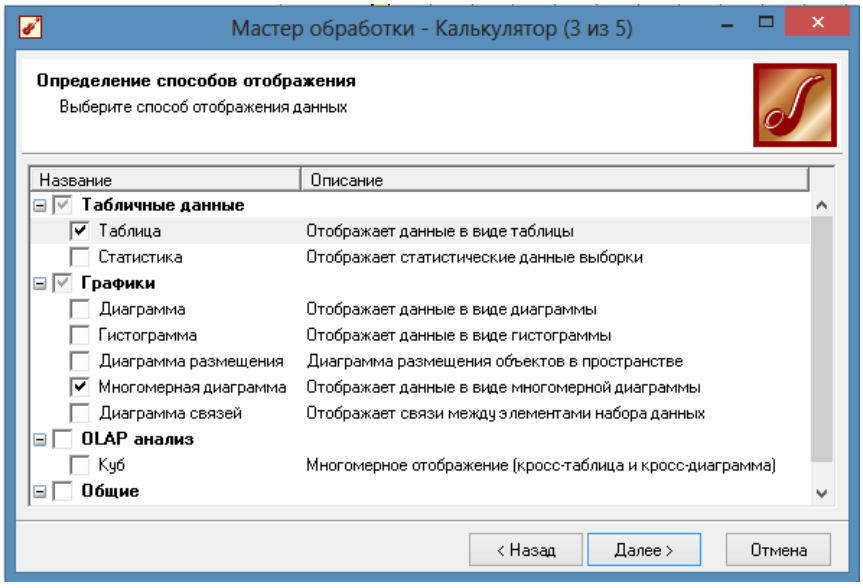


Рис.30. Налаштування засобів візуалізації.

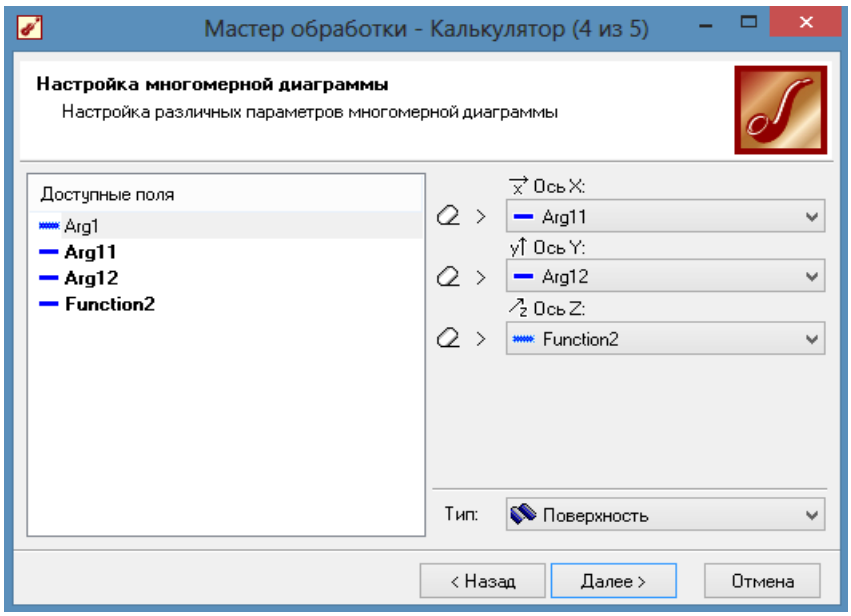


Рис.31. Спосіб відображення – “Многомерная диаграмма” та її налаштування.

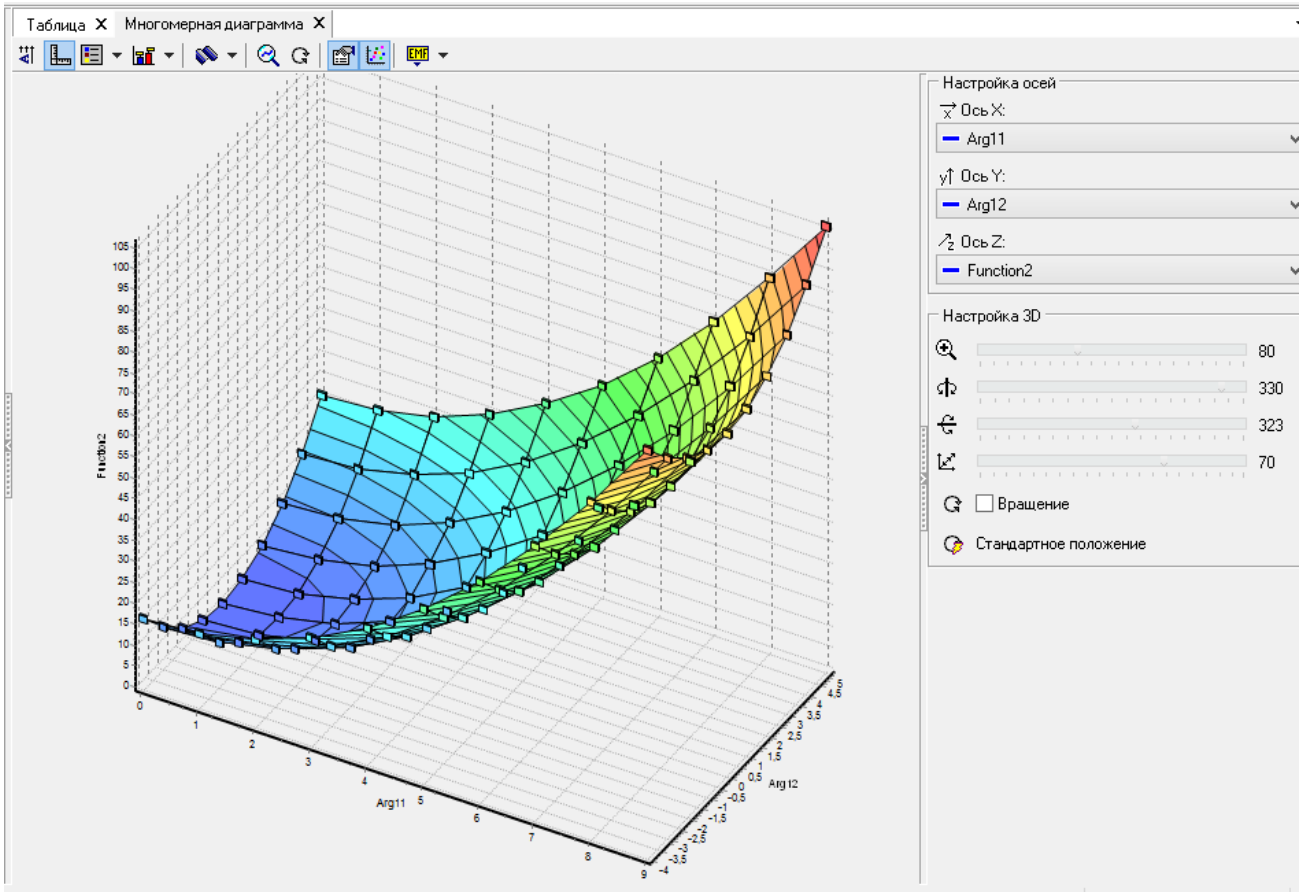


Рис.32. Графік неперевної функції двох аргументів.

3. Кусково-гладка функція одного аргументу $f(x)=\begin{cases}\sqrt{x-2}, & x\geq 2,\\ 0, & x<2.\end{cases}$,

$x\in[0,98]$**.** Налаштовуємо “Конструктор выражения” функції (Рис.33).

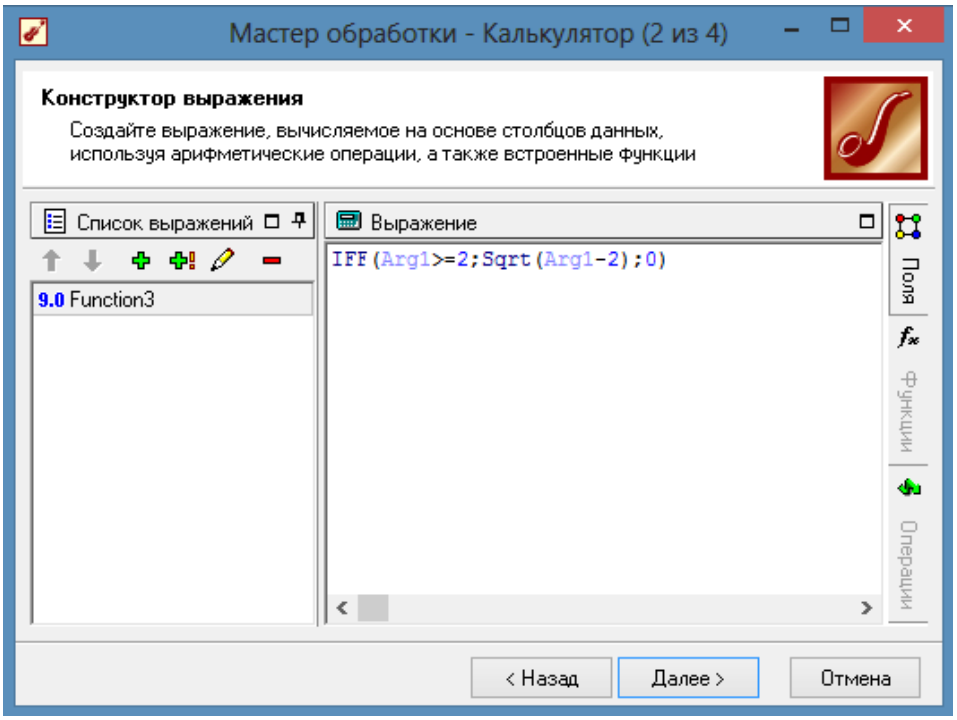


Рис.33. Налаштування конструктора виразу функції.

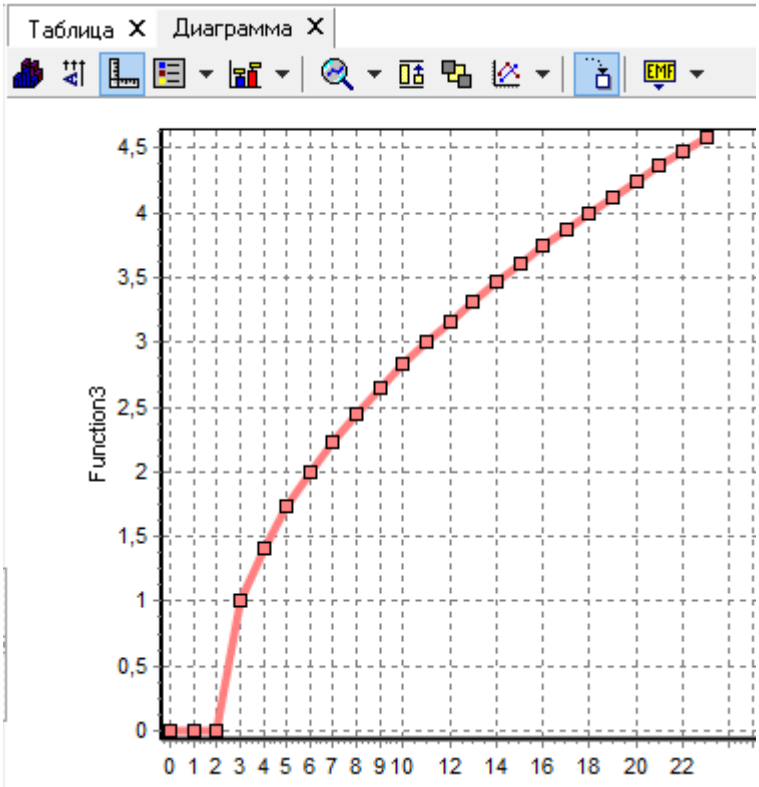


Рис.34. Графік кусково-гладкої функції одного аргументу.

4. Розривна функція одного аргументу $f(x) = \begin{cases} x^2, x <= 80 \\ (x-80)^3, x > 80 \end{cases}$. Налаштовуємо “Конструктор выражения” функції (Рис.35).

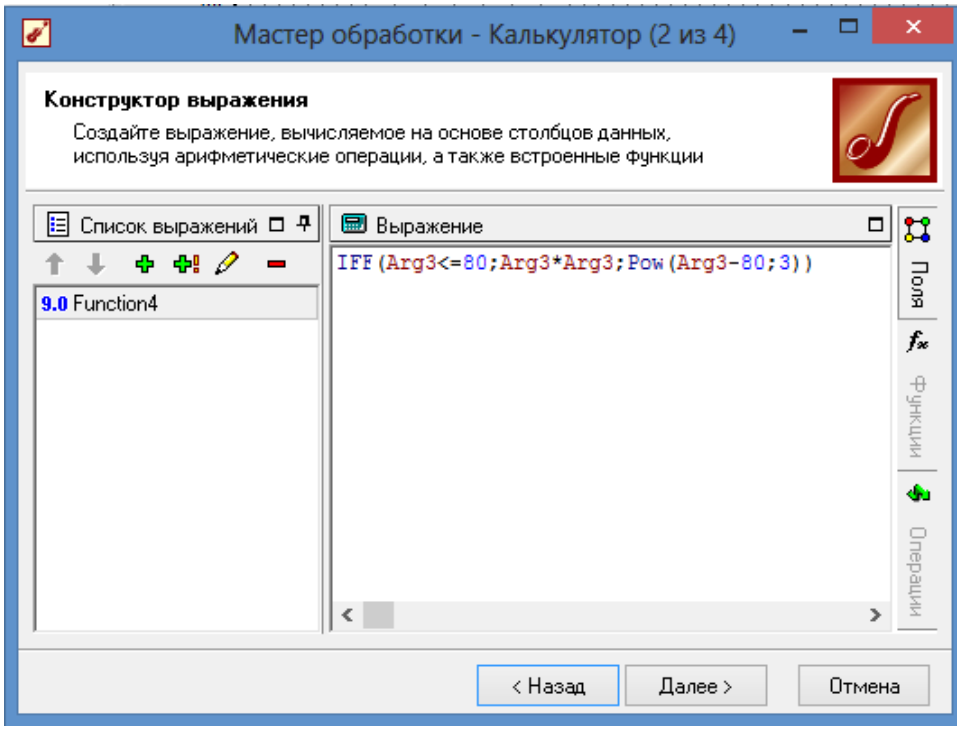


Рис.35. Налаштування конструктора виразу функції.

Графік функції представлений на Рис.36. Редактор автоматично згладжує функцію в точці розриву до неперервної кусково-гладкої функції.

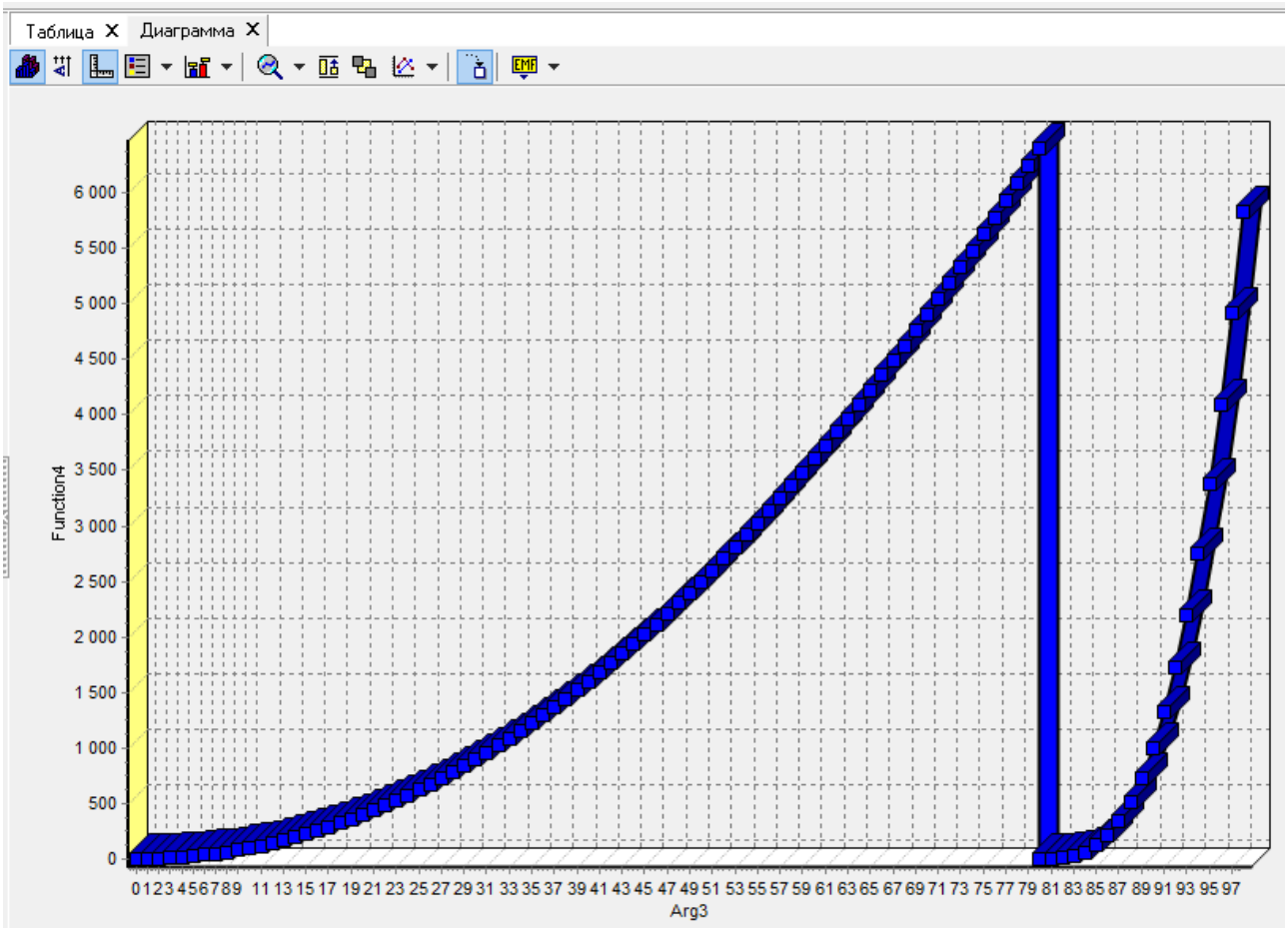


Рис.36. Графік розривної функції (точка Arg3=80) одного аргументу.

5. Розривна функція одного аргументу $f(x)=\dfrac{|x-40|}{x-40}x+6$ з однією точкою розриву першого роду. Налаштовуємо “Конструктор выражения” функції (Рис.37). Графік розривної функції представлений на Рис.38.

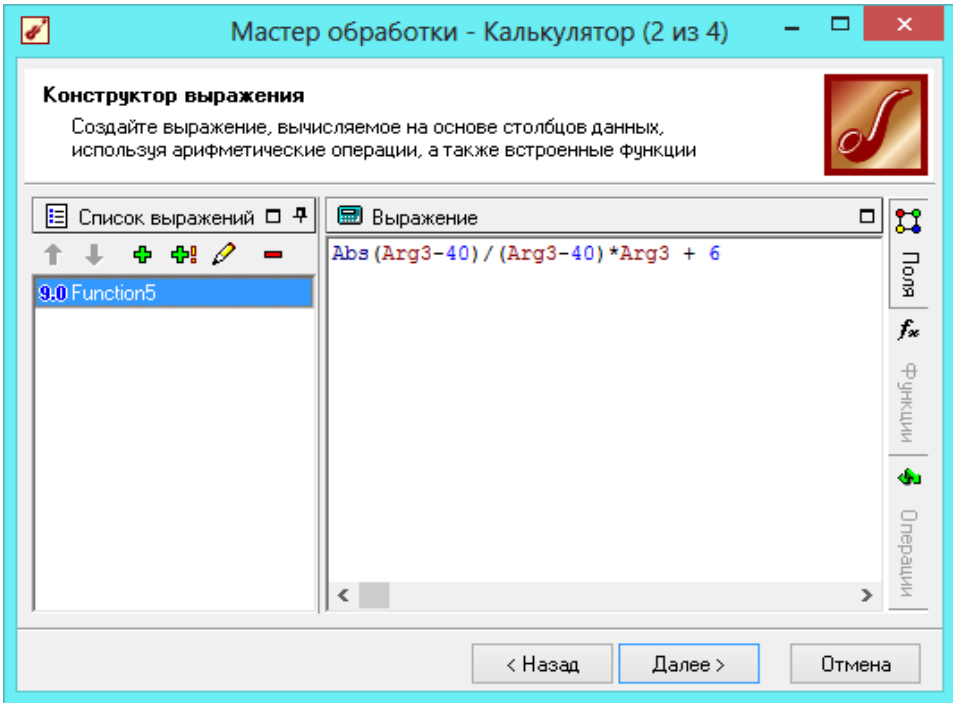


Рис.37. Налаштування конструктора виразу функції.

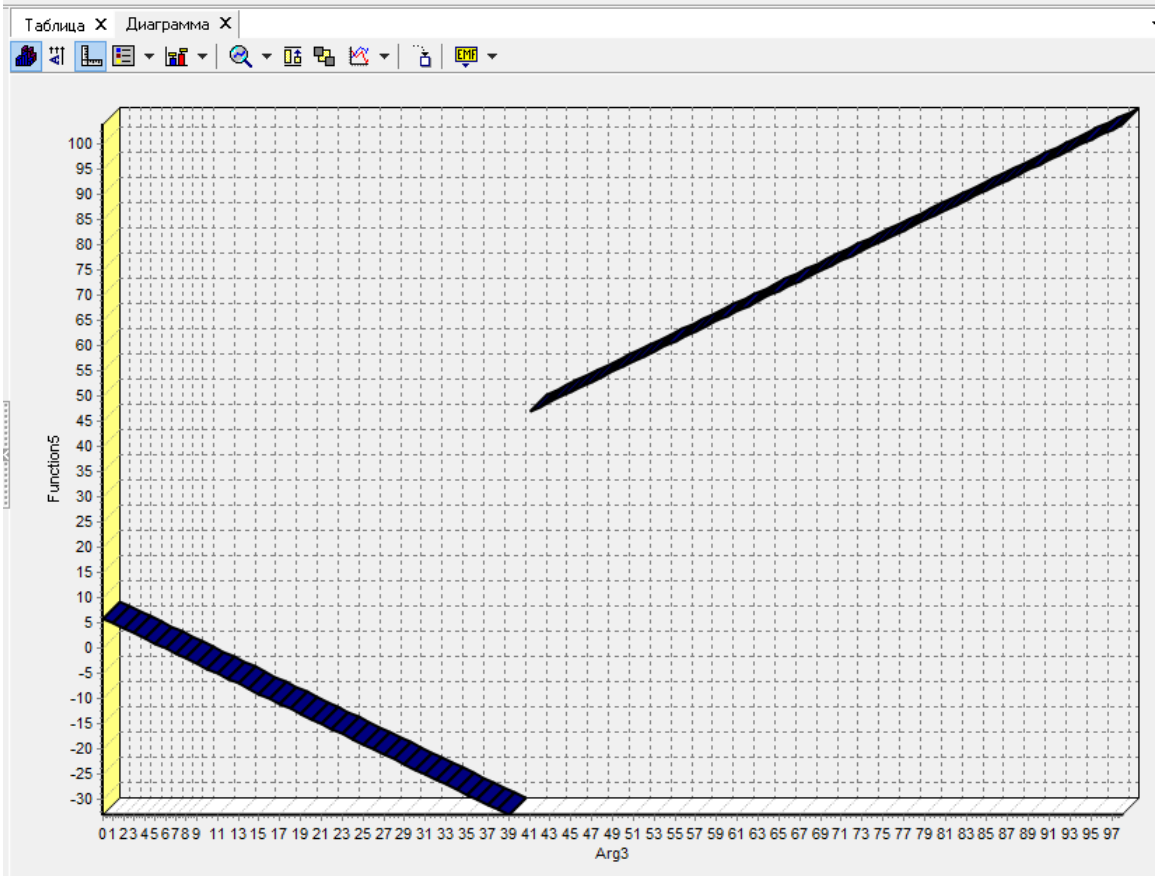


Рис.38. Графік розривної функції (точка Arg3=40) одного аргументу.

Таким чином графічні засоби Deductor Studio дозволяють візуалізувати усі типи функцій одного та двох аргументів.

Завдання для лабораторної роботи 3.

1. Сформувати файл Lab3.xlsx. Зробити таблицю із двох соціально-економічних та демографічних факторів по регіонах України, по роках 2009-2012, наприклад:

Регіон	Рік	Реальна середньомісячна заробітна плата по регіонах (грн.)	Число народжених живими на 1000 нас., обидві статі
АР Крим	2009	3960,3	7,7
Вінницька	2009	3750,6	7,2
…	…	…	…

2. Перевести дані за кожен рік у безрозмірний формат із [0,1]. Зберігти Lab3.txt
3. Імпортувати дані. Розбити їх по групах (зрізи тривимірного куба даних):
- по одному фактору – регіони (по рядках) за 2010, 2012 р.р.;
- регіони - два фактори за 2012 р. (рядок – області, фактори - стовпчики);
- розбиття на дати – регіони по економічному, демографічному факторах по дворіччях (2009-2010, 2011-2012);
- фільтрація даних - відобразити та створити окремі таблиці груп регіонів по роках в 4 групах (рівномірно) по демографічному показнику, по економічному показнику.

4. Побудувати графік функції двох змінних $f(x, y) = xe^{-x} + y^2$, $x \in [0,1]$, $y \in [0,1]$.

Тема 4. Видалення аномалій, згладжування даних та видалення шуму

Задачі: комплексна попередня обробка даних - видалення аномалій, згладжування даних та видалення шуму.

Методи: спектральна обробка даних на основі Фур'є та вейвлет- перетворень.

1. Основні положення спектральної обробки даних

Іноді рядки даних містять аномальні збурення, флуктуації та випадкові шуми. Випадкові збурення заважають виконувати аналіз даних та стають причиною нестійкої роботи алгоритмів обробки даних. В цих умовах складно розв'язувати задачі аналізу та прогнозування, виявляти в даних приховані закономірності, структури, тенденції [6], [7], [16].

Спектральна обробка даних призначена для очищення їх від шумової складової і подальшого згладжування. Згладжування даних необхідно в тому випадку, коли рядки даних виявляються нерівномірними, містять велику кількість дрібних структур, що перешкоджають дослідженню більш значних об'єктів і виявлення прихованих закономірностей. Для цих цілей у Deductor Studio використовуються два схожих методи: **перетворення Фур'є [6] та вейвлет--перетворення [7].** Принцип подібної обробки полягає в тому що функція часового рядку розкладається на базисні функції з певною частотою і амплітудою, після чого зменшується (фільтрується) амплітуда високочастотних складових. Наприклад, для часового рядку добового курсу валюти, це означає фільтрацію інформації щодо добових коливань курсу протягом декількох діб, що схильні до впливу випадкових факторів, і залишити більш стійкі тенденції, наприклад, за декаду. Можна, навпаки, фільтрувати складові з низькою частотою, що дозволить зменшити вплив повільних змін, а залишити тільки швидкі.

Рис. 1. Часовий ряд до спектральної обробки.

Спектральна обробка найбільш часто застосовується для попередньої підготовки даних в задачах прогнозування, оскільки дозволяє зробити часовий ряд

більш гладким. Це дає можливість застосувати більш широкий спектр математичних методів для побудови моделі прогнозування.

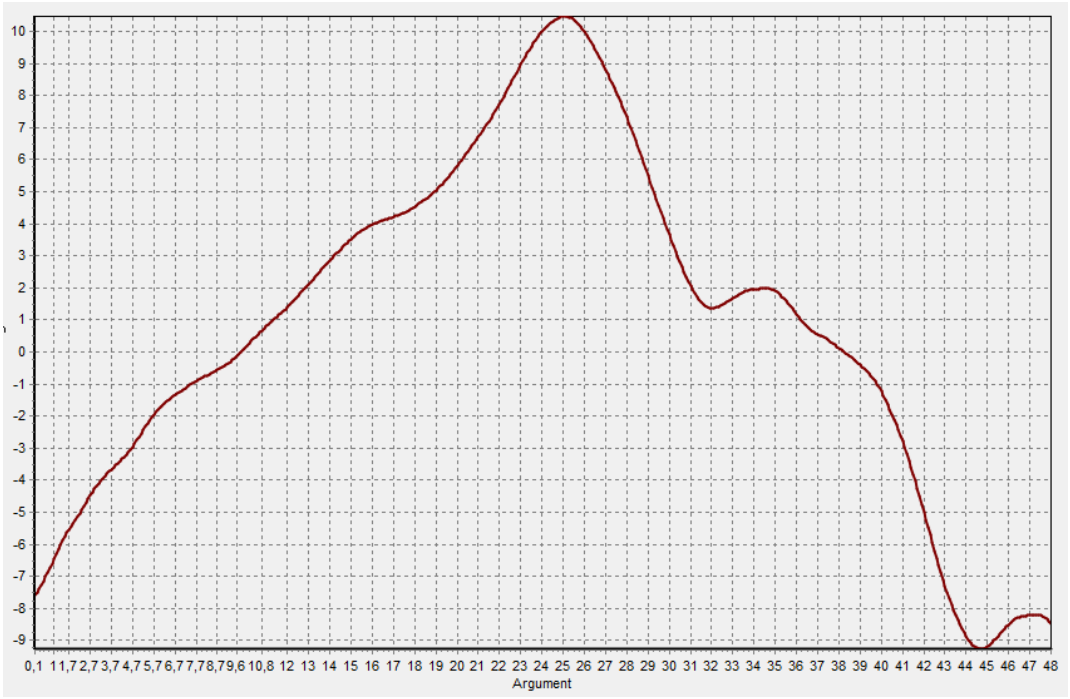


Рис. 2. Часовий ряд після згладжування (спектральної обробки).

Перетворення Фур'є використовується для обробки даних за рахунок перетворення їх частотного спектру. Частотний спектр отримується в результаті розкладання функції часу (часовий ряд) або просторових координат (наприклад, зображення), в базис набору гармонійних функцій. Перетворення Фур'є полягає в тому, що задана дискретна функція надається у вигляді комбінації синусів та косинусів з різною частотою і амплітудою. Кожна синусоїда (або косинусоїда) з певною частотою і амплітудою, отримана в результаті розкладання Фур'є, називається спектральною складовою або гармонікою. Спектральні складові утворюють спектр Фур'є.

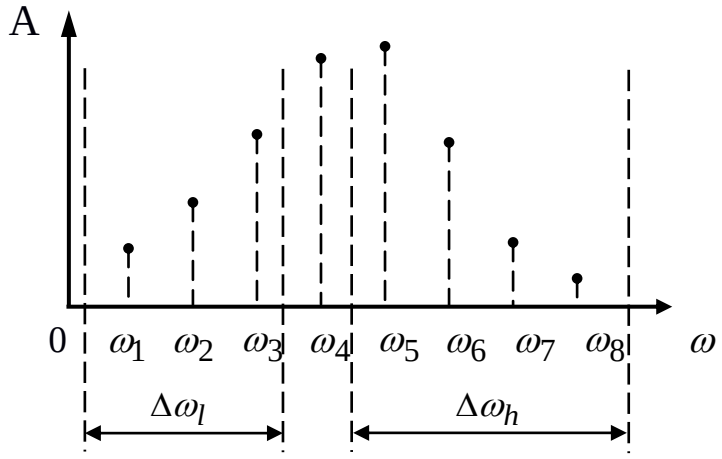


Рис.3. Амплітудно-частотна характеристика (АЧХ) функції. Частотні смуги пропускання для фільтру низьких частот $\Delta\omega_l$ (ФНЧ) та фільтру високих частот $\Delta\omega_h$ (ФВЧ).

Спектр Фур'є функції представляється у вигляді графіка, на якому по горизонтальній осі відкладається кутова частота ω, а по вертикалі – амплітуда спектральних складових A (Рис. 3). Такий графік надає амплітудно-частотну характеристику (АЧХ) функції. АЧХ містить важливу інформацію про характер поведінки функції: швидкі зміни вихідних даних породжують у спектрі складові з високою частотою, а повільні – з низькою. Якщо в АЧХ амплітуда спектральних складових швидко спадає зі зростанням частоти, то функція виявляється більш плавною. А якщо в АЧХ функції є високочастотні компоненти з великою амплітудою, то така функція має високочастотні осцилюючі складові. Останнє, в свою чергу, може свідчити про наявність випадкової складової у функції (наявність шумів).

Спектральна обробки (згладжування) функцій основане на операціях з її спектром Фур'є. Зміна спектру функції за рахунок зменшення (фільтрування з заданим рівнем) амплітуди його високочастотних компонентів приводить до зменшення або зникнення високочастотних осциляцій функції, її згладжування. Алгоритмічно ці операції реалізуються через пряме та зворотне Фур'є - перетворення функції.

Регулюючи параметри фільтрації спектру можна досягти різного ступеня згладжування функції з шумами. Для цього застосовуються спеціальні алгоритми, що здатні редагувати компоненти спектра функції, зменшувати або збільшувати їх. АЧХ є головною характеристикою фільтра. Фільтри низьких частот (ФНЧ) пропускають низькі частоти спектру та послаблюють високі частоти, що знаходяться вище частоти відсікання фільтру (cut off frequency). ФНЧ використовуються для видалення аномальних флуктуацій та шумів функцій. Фільтри високих частот (ФВЧ) навпаки, пропускають високі частоти спектру та послаблюють низькі частоти, що знаходяться нижче частоти відсікання фільтру. Важливе практичне застосування ФВЧ в різних задачах – усунення постійної складової сигналу. ФВЧ використовуються також при обробці аудіосигналів, обробці зображень для виділення контуру зображення та інших.

В Deductor Studio спектральне згладжування функцій з аномаліями та шумами здійснюється на основі заданої користувачем верхньої планки полоси пропускання частот спектру (cut off frequency). Тобто, в аналітичній платформі використовується ФНЧ що відсікає усі складові спектра з частотами що знаходяться вище частоти відсікання. Вибір оптимального значення частоти відсікання можливе за рахунок проведення серії експериментів на тестових прикладах певних класах функцій, з якими працює аналітик при розв'язку прикладної задачі.

Вейвлет—перетворення є узагальненням спектрального аналізу, у тому числі перетворення Фур'є. Слово вейвлет (wavelet, англ.) означає коротку (маленьку) хвильку або хвилі, що ідуть одна за одній. Основна ідея методу полягає у можливості розкладення довільної функції в ряд по базисному набору функцій локальних у часі та по частоті та визначеному у просторі комплекснозначних функцій з обмеженою нормою в $L_2(R)$. Базисний набір функцій утворюється з однієї і тієї ж функції-прототипу (що породжує) за допомогою її зсувів, розтягнень та різномасштабними копіями на осі часу таким чином, що середнє значення кожної функції в області визначення дорівнює нулю (умова допустимос-

ті). Такі базисні функції називаються вейвлетами (приклад на Рис. 4). По локалізації у часовому та частотному представленні вони займають проміжну нішу між гармонійними функціями локалізованими по частоті і функцією Дірака, локалізованою в часі. Локалізація вейвлету у часі та по частоті дозволяє отримати характерні особливості функції в області локалізації при обчислені її коефіцієнтів розкладення в ряд (згорток). Більш широкий масштаб локалізації вейвлету у часі (compact support) означає його звужену локалізацію у частотному діапазоні що проводить до того що більш широка область функції впливає на результат обчислення згортки. Навпаки, чим більш вейвлет локалізований у часі тим менш він локалізований у частотному діапазоні (розмазаний по частоті) тим менш широка область функції впливає на результат обчислення згортки. Вейвлети інваріантні до зсуву на осі часу та до масштабування, задовольняють умовам лінійного перетворення.

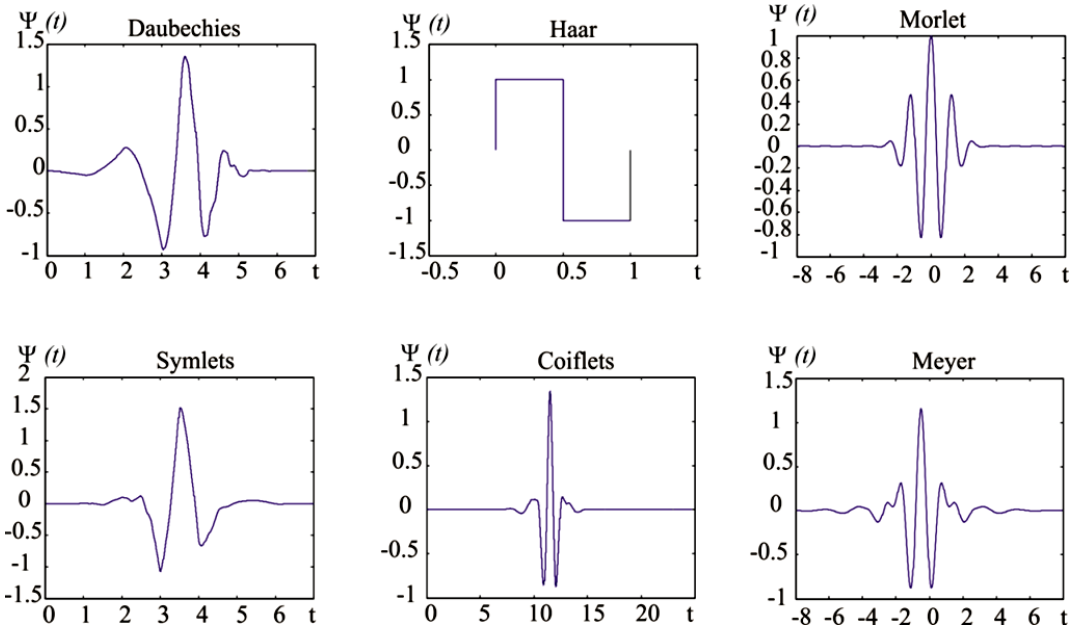


Рис.4. Приклади різних вейвлетів.

На відміну від класичних методів спектрального аналізу (наприклад, Фур'є – перетворення, косинус - перетворення) вейвлет-перетворення покриває фазову площину у вимірах частота – час фігурами однакової площі, але різної форми. Завдяки цій властивості низькочастотна складова функцій краще локалізується вейвлетами у частотній області (домінуючі гармоніки), високочастотна складова – у часовій області (піки функцій, стрибки тощо).

Вейвлет-перетворення дозволяє також проводити згладжування (фільтрацію) розривних та навіть, фрактальних функцій, які не є гладкими та можуть не мати взагалі вейвлетів поняття кратномасштабного аналізу є фундаментальним. Розкладання функції здійснюється на різнотипні складові: регіональної функції тренда, циклічні компоненти з певним періодом повторення, локальні аномалії різного порядку тощо. За рахунок такого представлення з'являється можливість фільтрації незначних фрагментів функції (шуми, аномалії, швидкі хвилі тощо), зберігаючи її основні складові. Або навпаки, акцентувати увагу на дрібних деталях фу-

нкції. Експерименти в фізиці (одним з перших був броунівський рух частинок) виявили переваги вейвлет-перетворення перед традиційними методами спектрального згладжування. Оптимальні значення параметрів вейвлет-перетворення залежать від характеристик вхідної функції і повинні підбиратися у чисельних експериментах. Більш детально ці питання будуть розглянуті нижче.

2. Імпорт даних, візуалізація даних та видалення аномалій.

У текстовому файлі task4.txt задана таблиця ВПОРЯДКОВАНИХ даних: дискретні значення аргументу функції, дискретні значення функції, її значення з аномальними збуреннями, її значення з сильними, середніми та слабкими шумами. Необхідно імпортувати данні, згладити аномальні значення функції та видалити усі типи шумів методами спектральної обробки (Фур'є та вейвлет - перетворення).

У наступному вікні відображається інформація про імпортовані дані з файлу task4.txt. Зліва знаходяться ім'я змінних, праворуч — налаштування атрибутів змінних. Хоча Deductor Studio автоматично знаходить тип даних, рекомендується перевірити коректність імпорту даних. Задаємо кожну змінну як дійсну та неперервну. Імена змінних:

- ✓ Argument — дискретні значення аргументу функції;
- ✓ Function — дискретні значення функції у точках Argument;
- ✓ Anomaly — дискретні значення функції з аномальними збуреннями;
- ✓ Low — дискретні значення функції з малими шумами;
- ✓ Middle — дискретні значення функції з середніми шумами;
- ✓ High — дискретні значення функції з високими шумами.

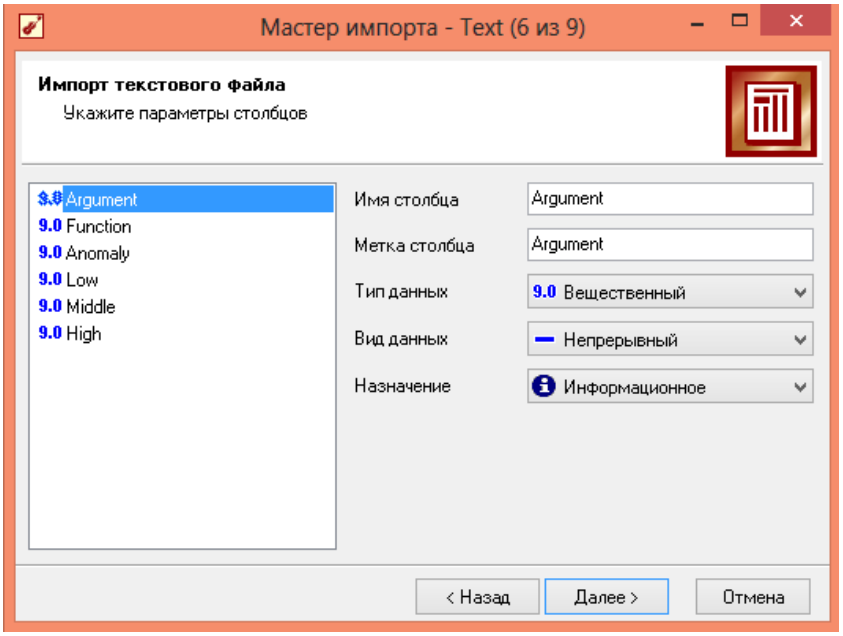


Рис.5. Знайдені значення аргументу та функцій.

У вікні ліворуч, в дереві сценаріїв з'явився новий вузол з імпортованими даними. Результат імпорту відображується на діаграмах (лінії) (Рис.6-10).

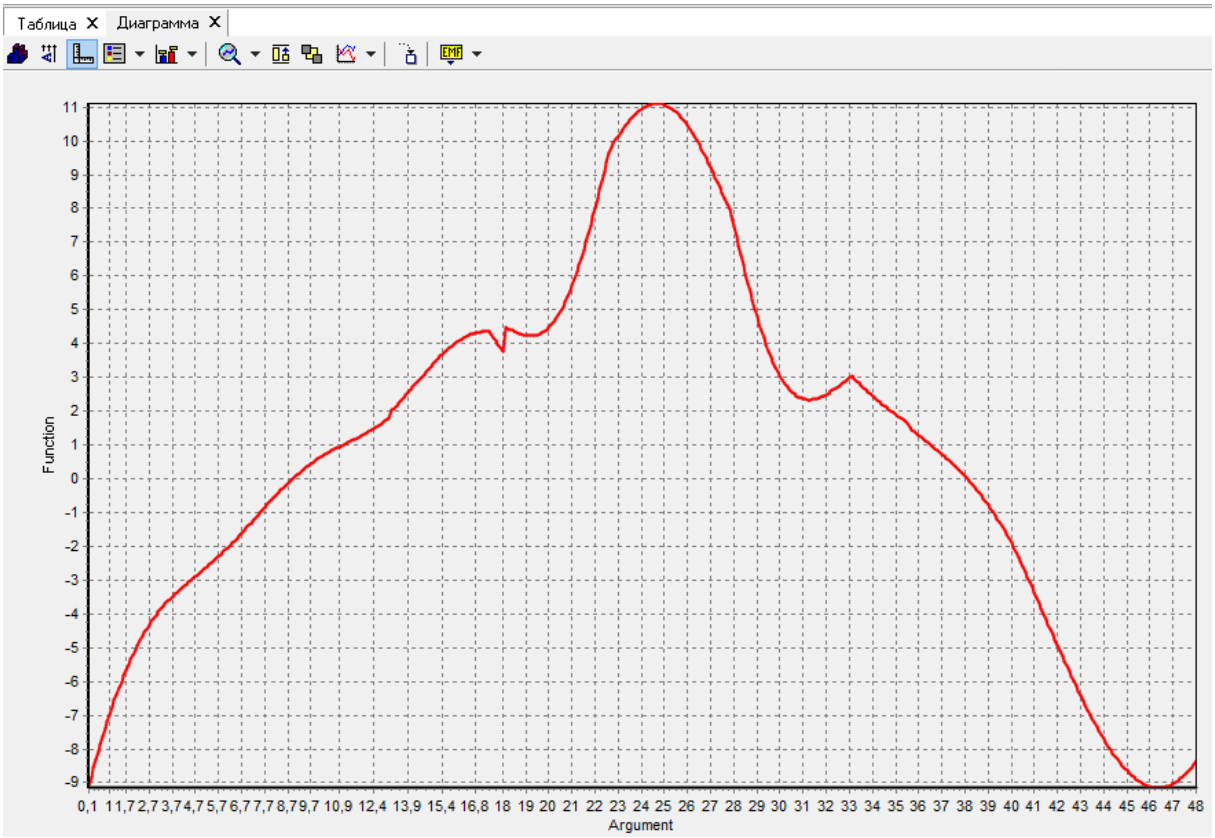


Рис.6. Графік початкової функції.

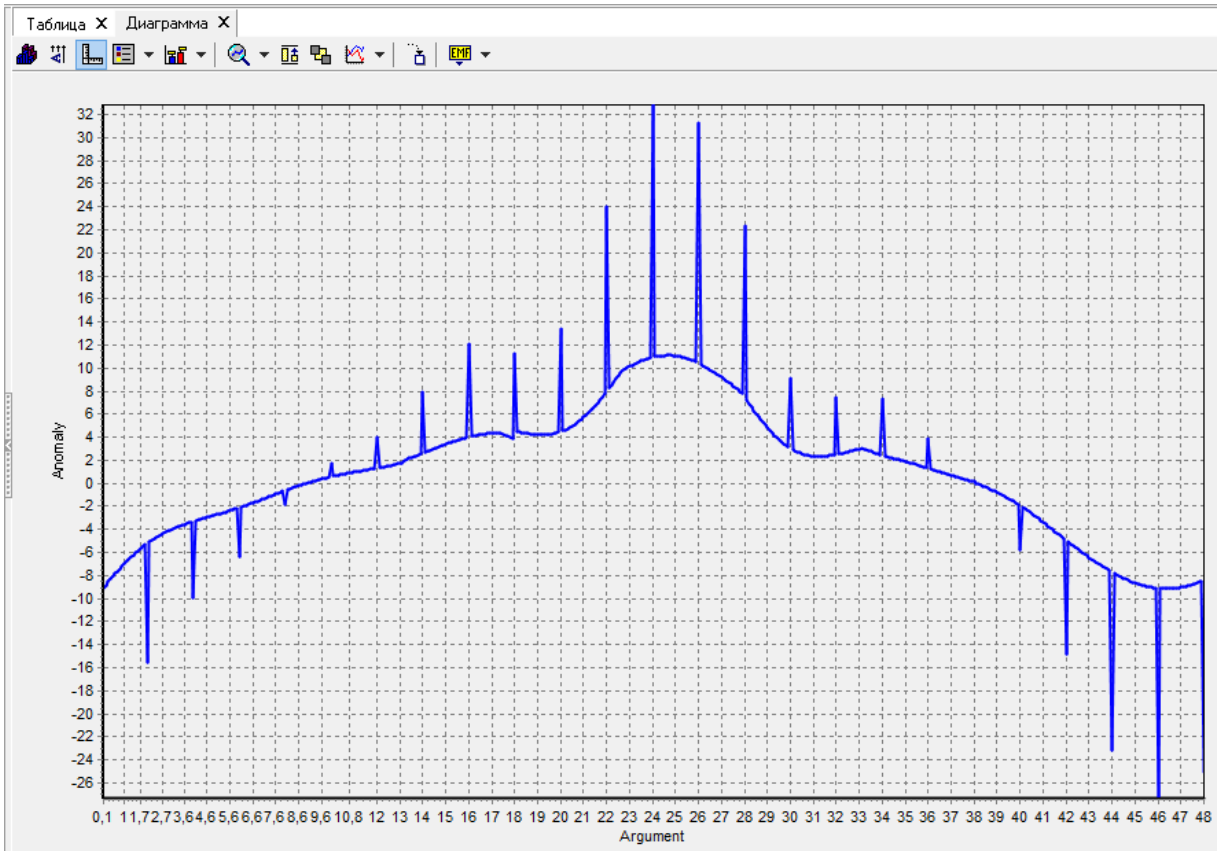


Рис.7. Функція з аномальними значеннями.

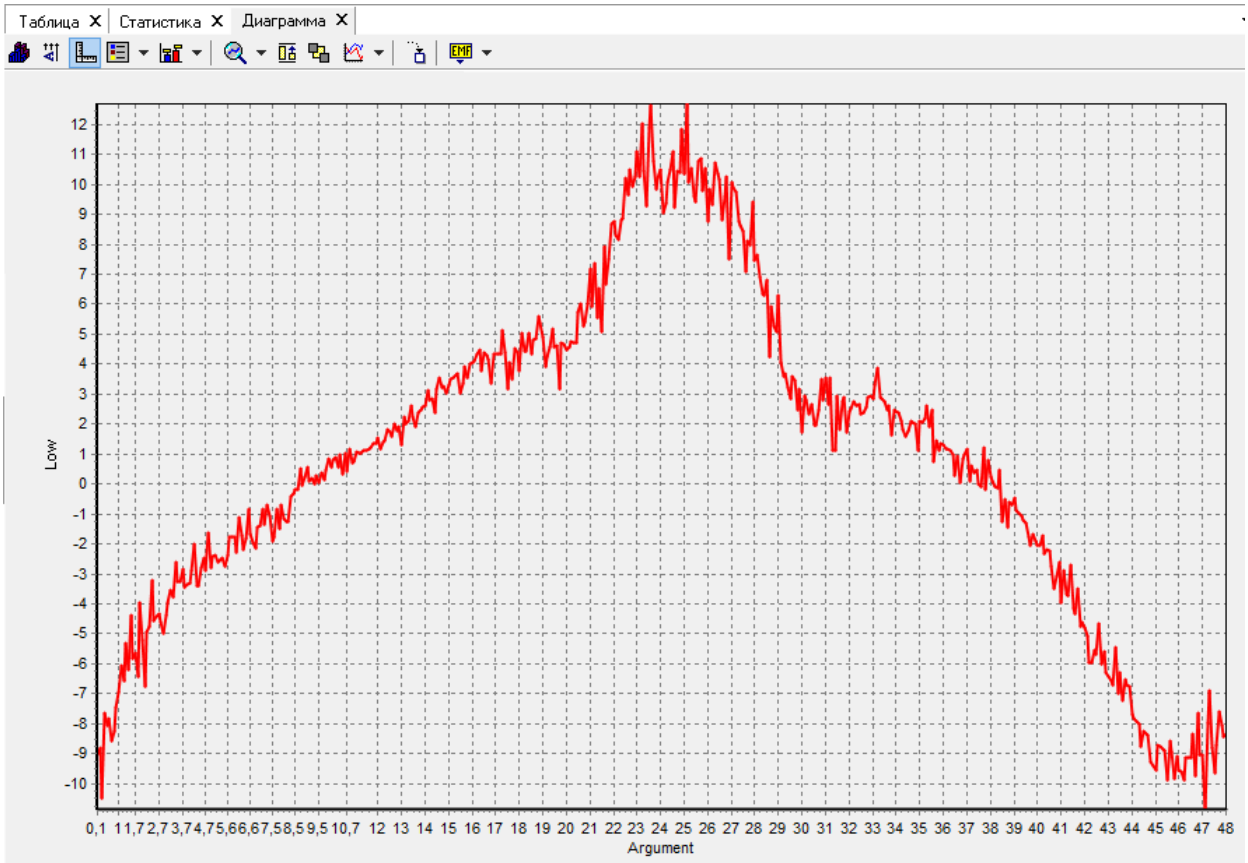


Рис.8. Функція з малими шумами.

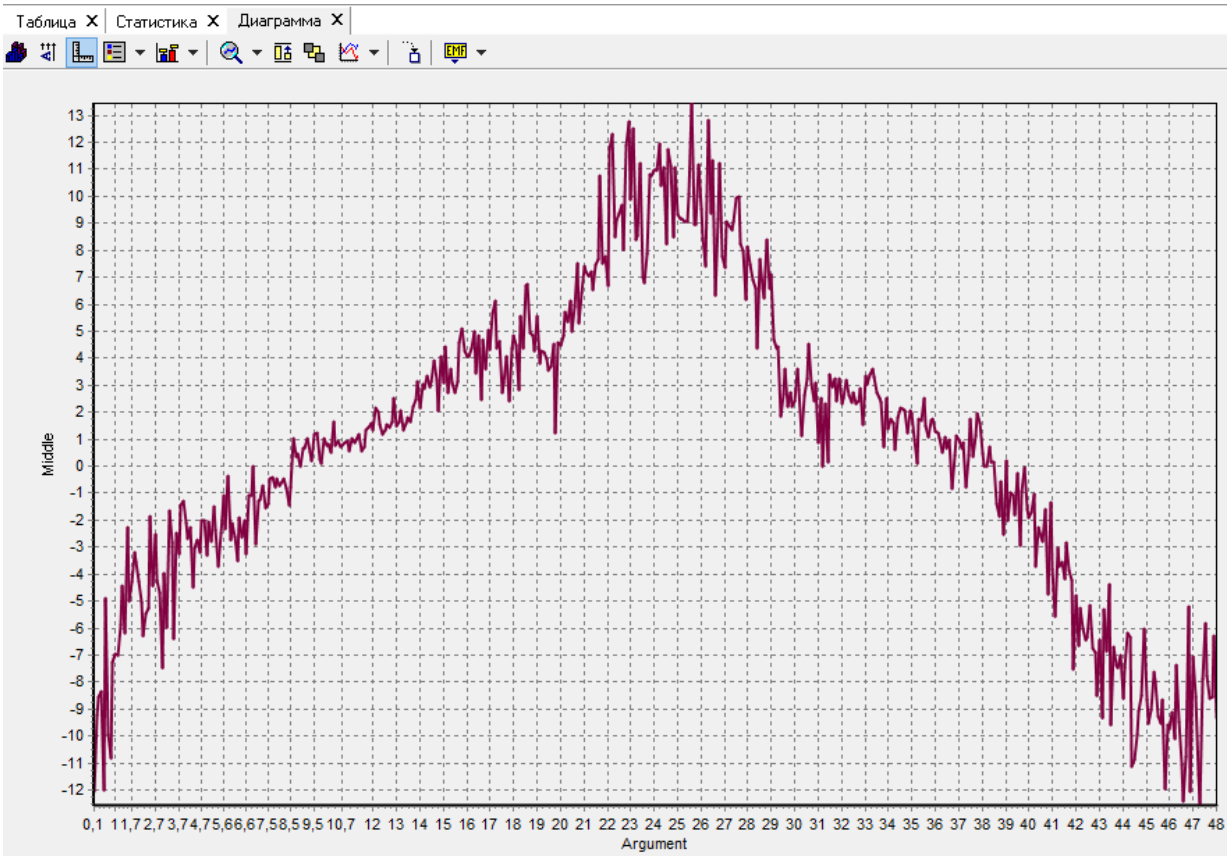


Рис.9. Функція із середніми шумами.

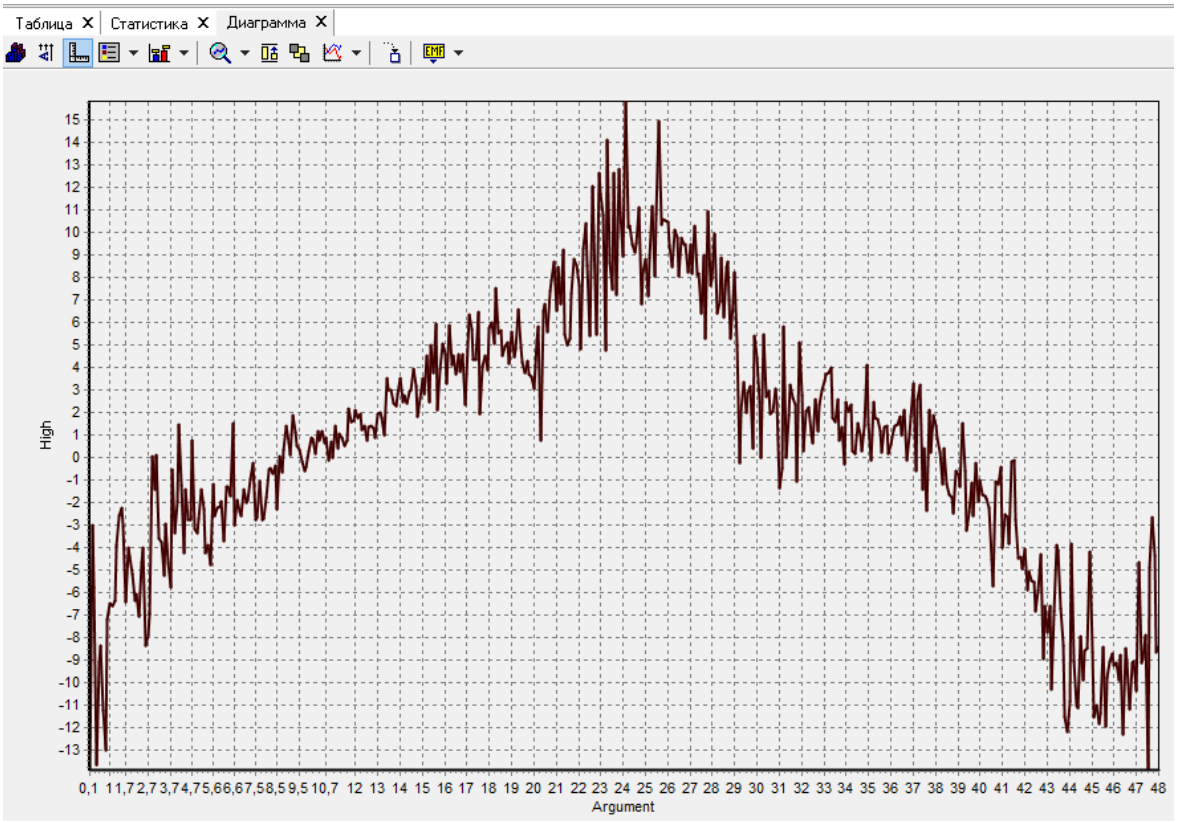


Рис.10. Функція з великими шумами.

Для видалення аномалій методом робастної фільтрації у майстра обробки обираємо “Спектральная обработка” (Рис.11). Далі на другому кроці у вікні для стовпчика Anomaly обираємо тип даних “Используемые”, для решти - “Информационные”. У “Вычитание шума” встановлюємо “Большая” (Рис.12). Результат видалення аномалій у вигляді графіка відновленої функції та графік початкової функції без аномалій показано на Рис.13.

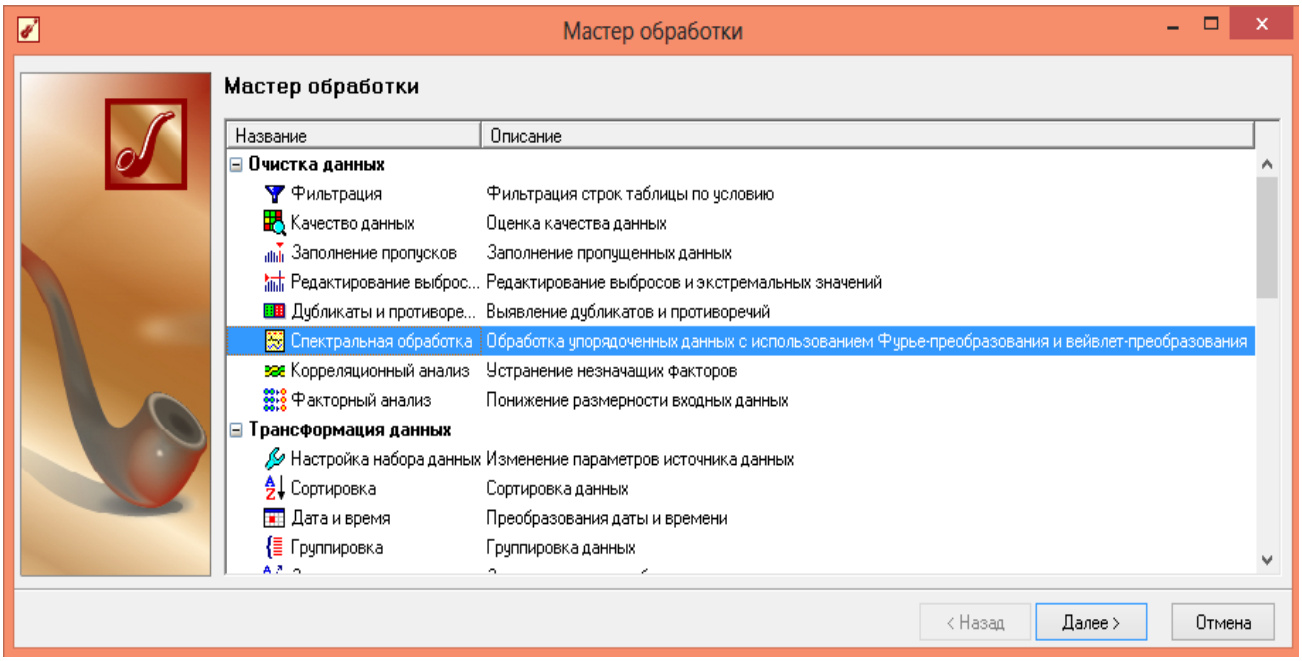


Рис.11. Вікно "Мастер обработки"

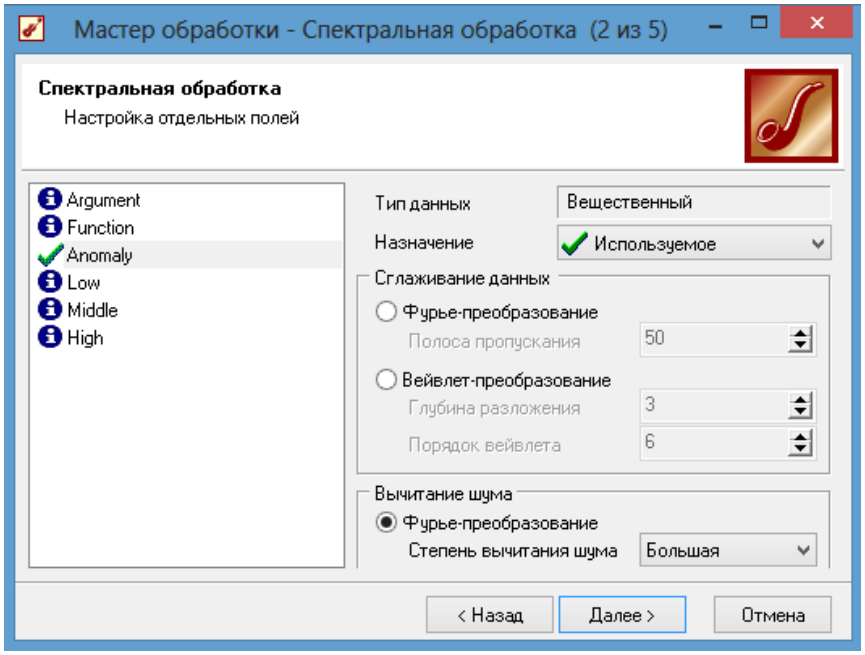


Рис.12. Крок видалення аномальних значень.

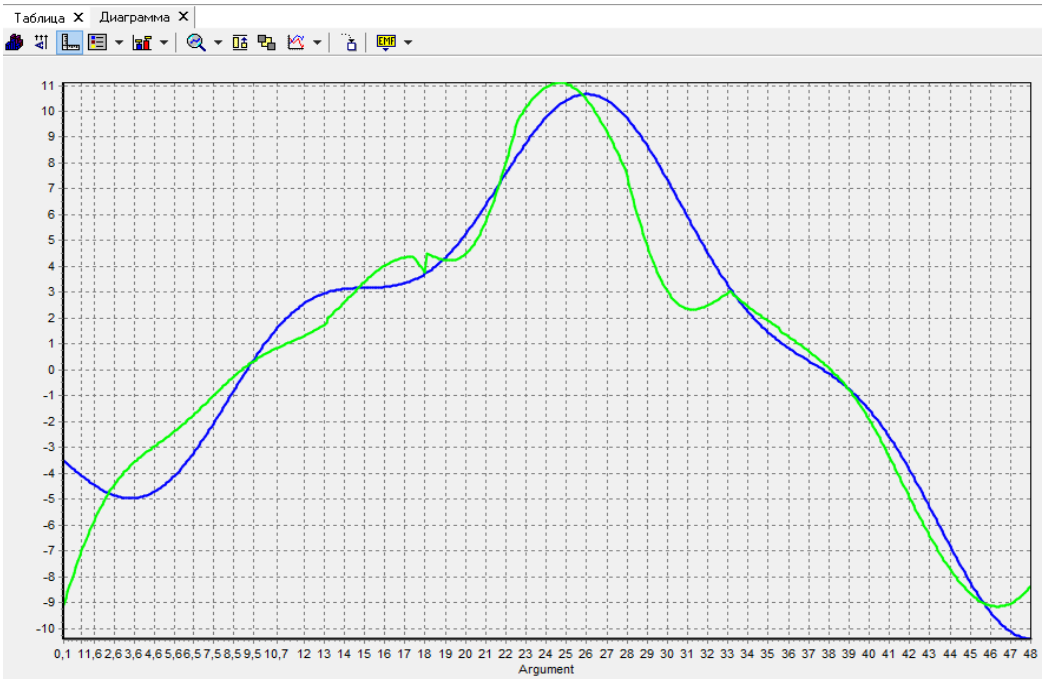


Рис.13. Результат видалення аномалій та початкова функція з Рис.6.

3. Аналіз та порівняння результатів згладжування даних за допомогою вейвлет-перетворення та Фур'є-перетворення.

У цьому пункті для часового ряду з високим ступенем шуму необхідно:

- Провести аналіз роботи вейвлет-перетворення із згладжування функцій із різними параметрами методу. Зробити висновки про вплив параметрів вейвлет-перетворення на точність згладжування функції.
- Провести аналіз роботи Фур'є-перетворення із згладжування функцій із різними параметрами методу. Зробити висновки про вплив параметрів методу Фур'є-перетворення на точність відновлення функції за рахунок фільтрації шуму.
- Порівняти роботу обох методів спектральної фільтрації.

3.1. Метод вейвлет-перетворення

Переходимо на вузол з імпортованими даними з файлу task4.txt, що містить стовпчики з даними на Рис.5 графіки яких зображено на Рис.6 - 10. Обираємо стовпчик High (Рис.10) для згладжування. Будемо проводити візуальний аналіз співставлення кінцевої згладженої функції з початковою (без шуму). Також можна знайти суму квадратів відхилень згладженої функції від початкової на основі даних наданих Deductor Studio у вигляді Таблиць. Перший метод вейвлет-перетворення [1], [7] містить два параметри настройки фільтру шуму:

1. Глибина розкладання (цілочисельний параметр **Level**) змінюється від 1 до 9 та визначає масштаб фільтрації деталей: чим більше ця величина, тим більше "великі" зміни функції будуть відфільтровані. При великих значеннях параметру **Level** (близько 7-9) виконується не тільки фільтрація шуму, але і згладжування функції за рахунок обрізання екстремальних збурень шуму. При використанні великих значень глибини розкладання також може виникнути помилка "Довжина сигналу менш довжини фільтра".

2. Порядок вейвлету (цілочисельний параметр **Order)** змінюється від 1 до 10 та визначає гладкість відновленої функції: чим менше значення параметра **Order**, тим точніше будуть виражені "піки" локальних екстремумів функції, і навпаки – при великих значення параметра Order піки функції будуть більш згладженими.

На Рис. 14 — 20 можна побачити певну залежність якості згладжування функції та видалення шумів від глибини розкладання вейвлет-перетворення. Тут та всюди далі зеленим кольором позначена початкова функція (Рис.6), коричньовим – згладжена функція. На Рис.14 показано результат роботи метода з мінімальною глибиною розкладання **Level**=1, та порядком вейвлету **Order**=6. Зміни **Order** від 1 до 10 у даному випадку дає дуже близький результат. Функція шумів майже не змінила свою форму та потребує подальшого згладжування. Навіть з малими шумами, мінімальна глибина розкладання не виділяє шуми, а тільки трохи їх зменшує.

На Рис.15 представлено результат роботи метода із середнім значенням глибиною розкладання **Level**=4, та мінімальним порядком вейвлету **Order**=1. Отримуємо кусково-постійну функцію як результат фільтрації функції з шумом. Хоча отримана функція краще відтворює початкову функцію, у багатьох випадках такий результат у вигляді не гладкої результуючої функції не є сприятливим. Тому розглянемо подальшу фільтрацію для інших параметрів методу.

На Рис.16 показано результат роботи метода із середнім значенням глибиною розкладання **Level**=4, та середнім порядком вейвлету **Order**=6. Результат видалення шумів дуже близький до початкової функції, а також збережені піки функції. Але вихідна функція не являється гладкою та містить зайві “піки”.

Дуже близький результат, але з більш згладженою вихідною функцією отримано для більшого значення порядку вейвлету **Order**=6, Рис.17. При зменшенні значення глибини розкладання до **Level**=3 отримуємо недостатньо гладке згладжування на виході, присутність високочастотних складових, Рис.18.

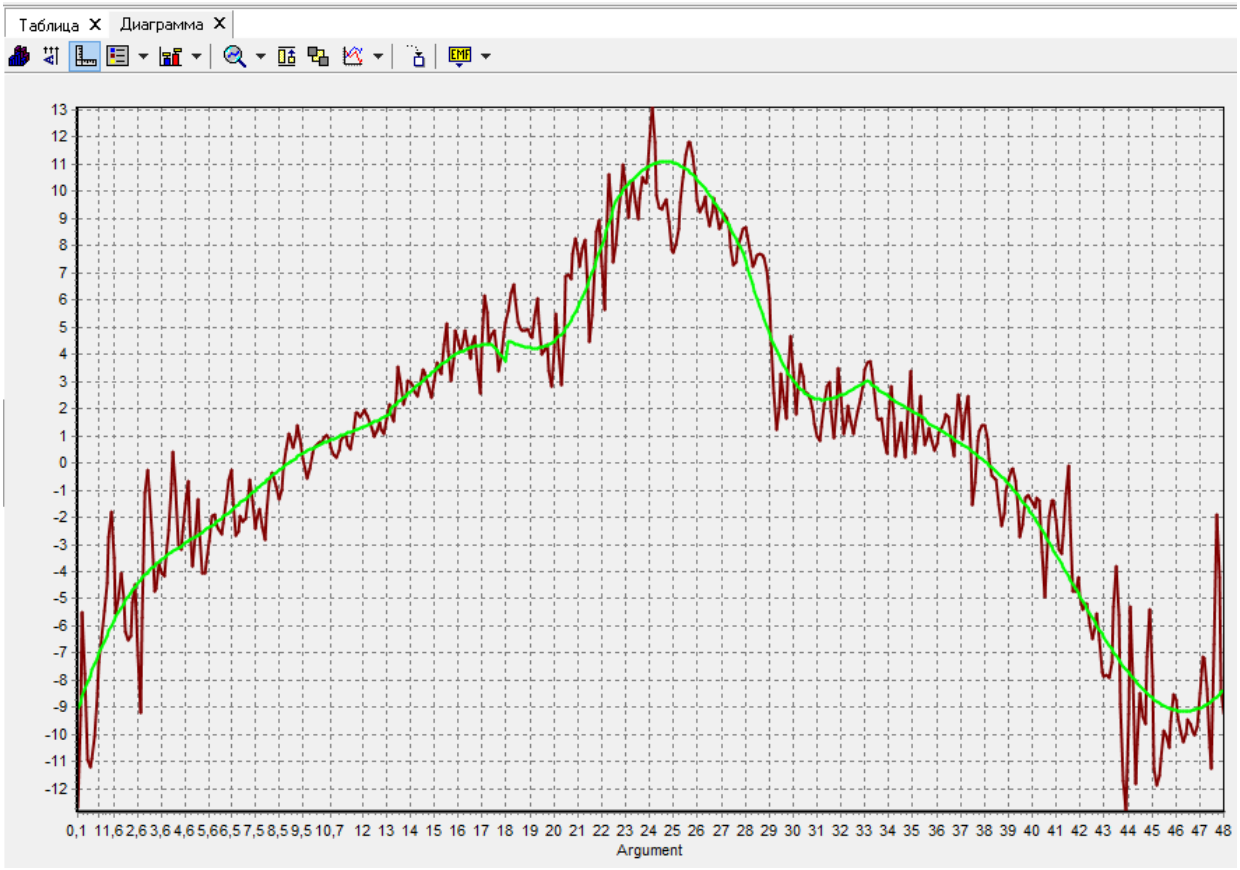


Рис. 14. Невелика фільтрація шуму вейвлет-перетворенням з **Level**=1, **Order**=6 (аналогічні графіки для усього діапазону **Order**=1,…,10).

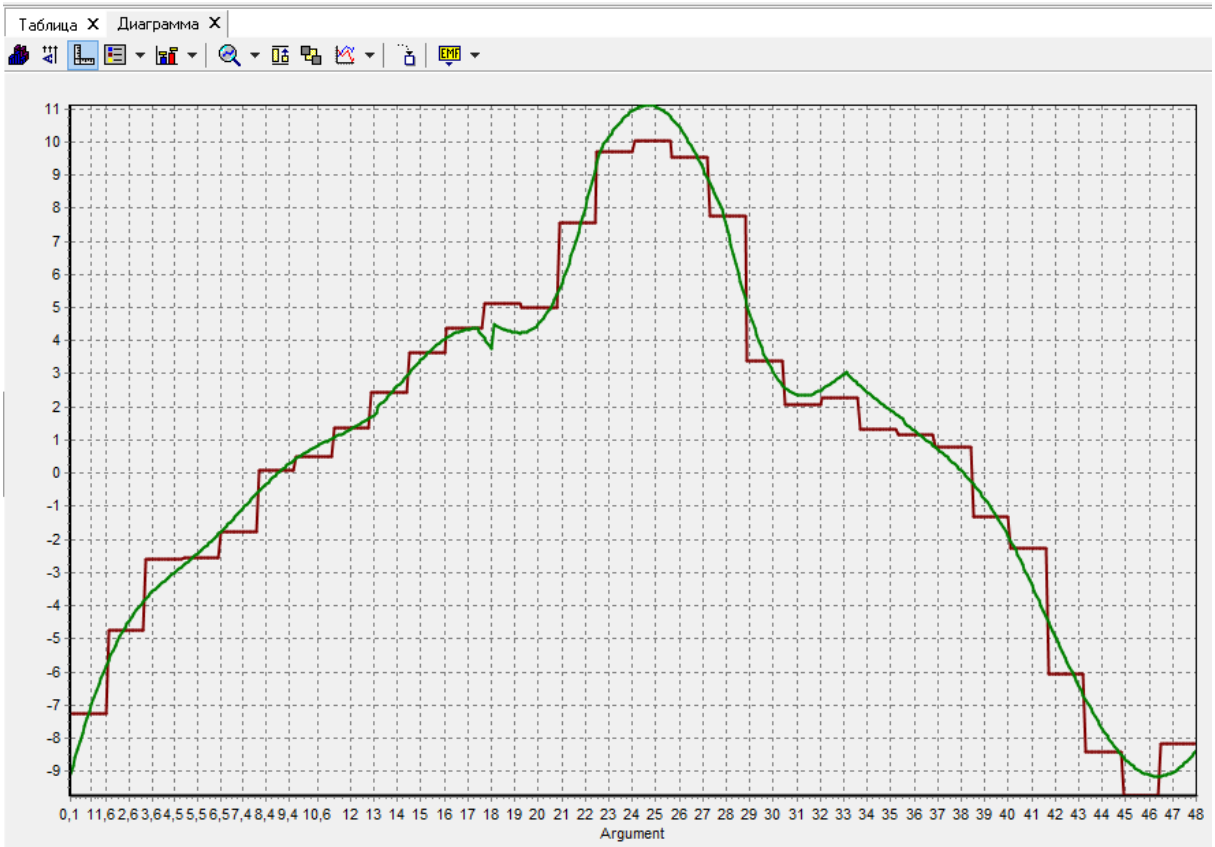


Рис.15. Кусково-постійний графік з фільтрації шуму вейвлет-перетворенням з **Level**=4, **Order**=1.

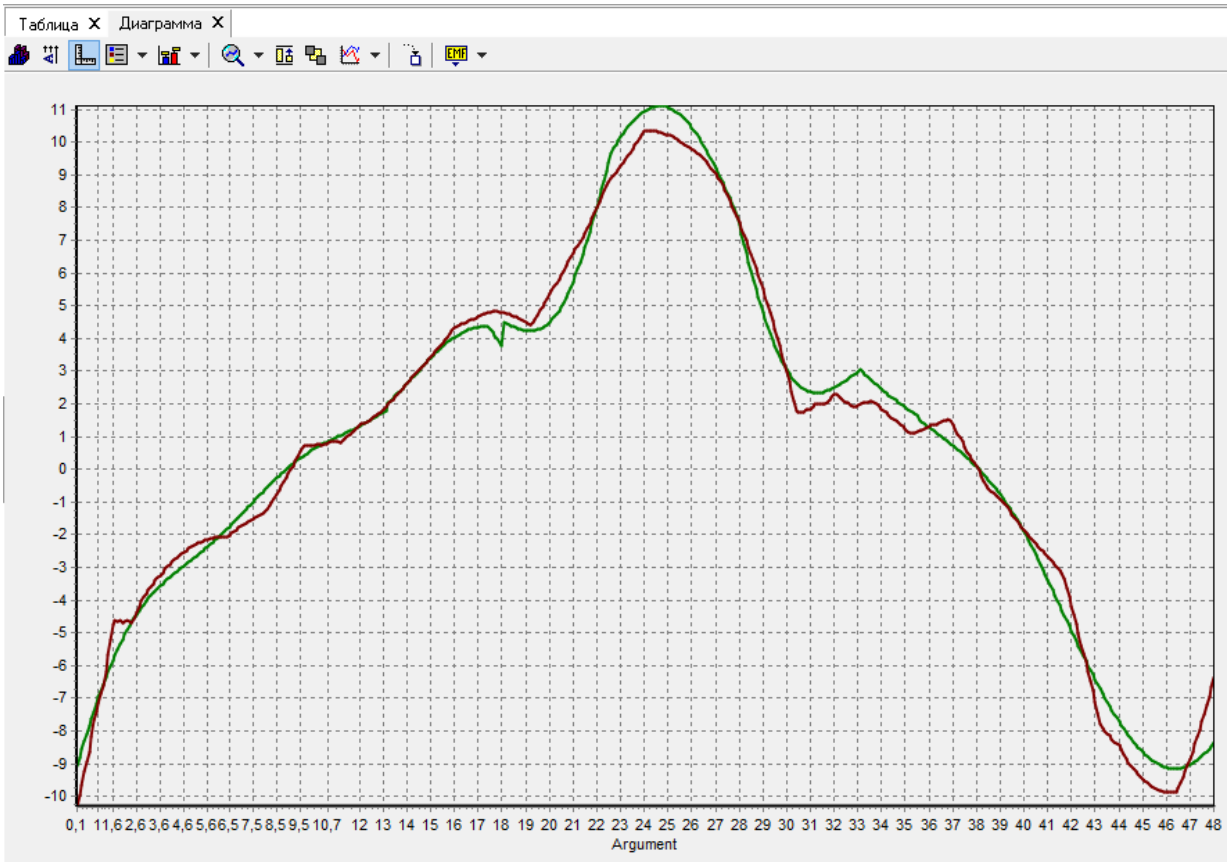


Рис.16. Згладжування шуму вейвлет-перетворенням з **Level**=4, **Order**=3.

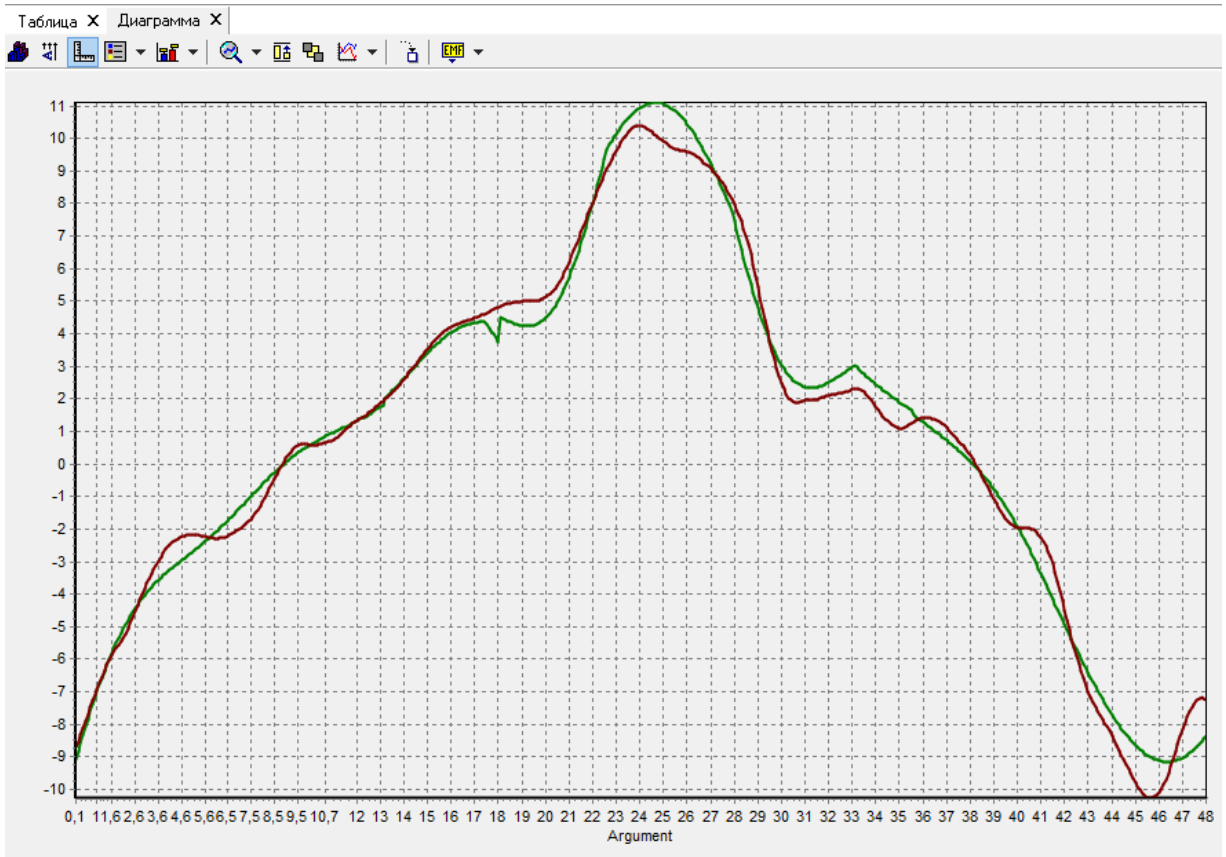


Рис.17. Згладжування шуму вейвлет-перетворенням з **Level**=4, **Order**=6.

Рис.18. Недостатньо гладке згладжування вейвлет-перетворенням з **Level**=3, **Order**=6.

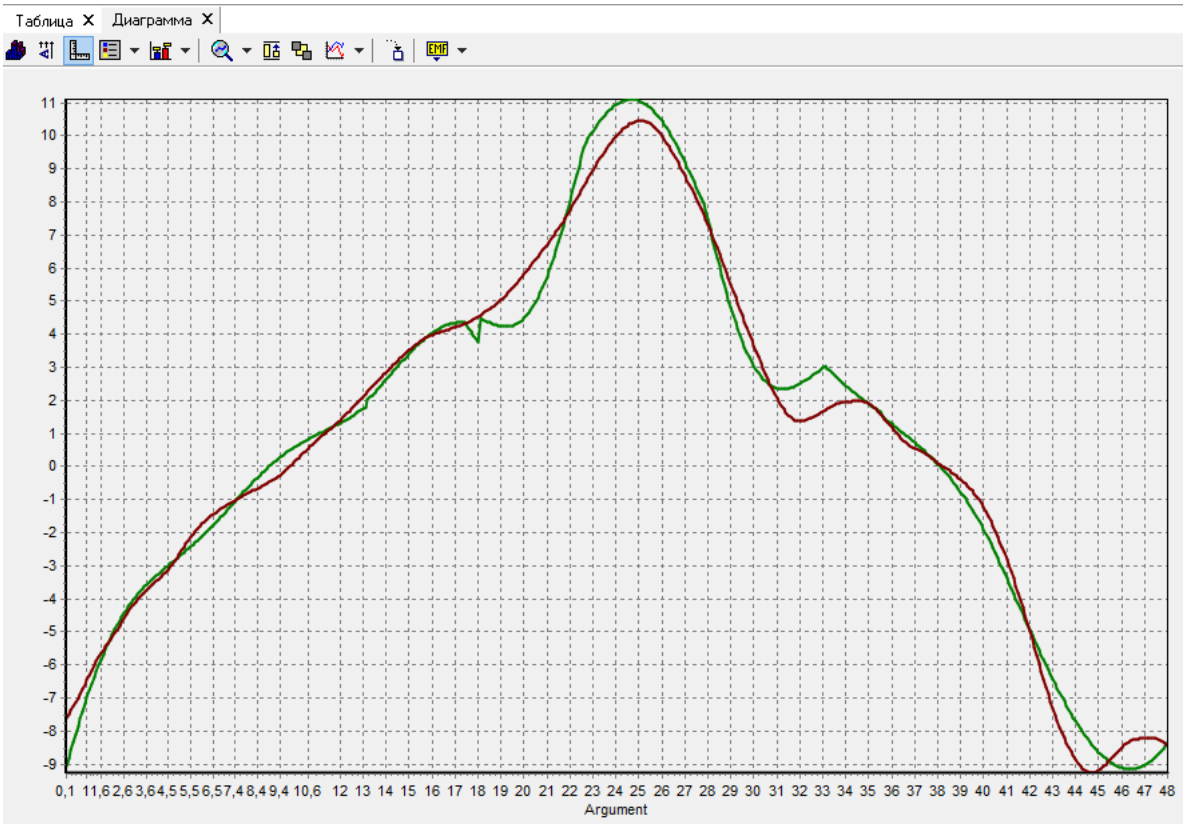


Рис.19. Значне згладжування шуму вейвлет-перетворенням з **Level**=5, **Order**=6.

При збільшенні глибини розкладання до **Level**=5 для фіксованого порядку вейвлету **Order**=6 отримуємо більш згладжену функцію на виході (Рис.19) у

порівнянні з графіком на Рис. 18. У даному випадку відбувається фільтрація негладких локальних екстремумів функції з шумом. При граничних значеннях **Level**=9, **Order**=1 функція на виході вироджується у константу (Рис.20).

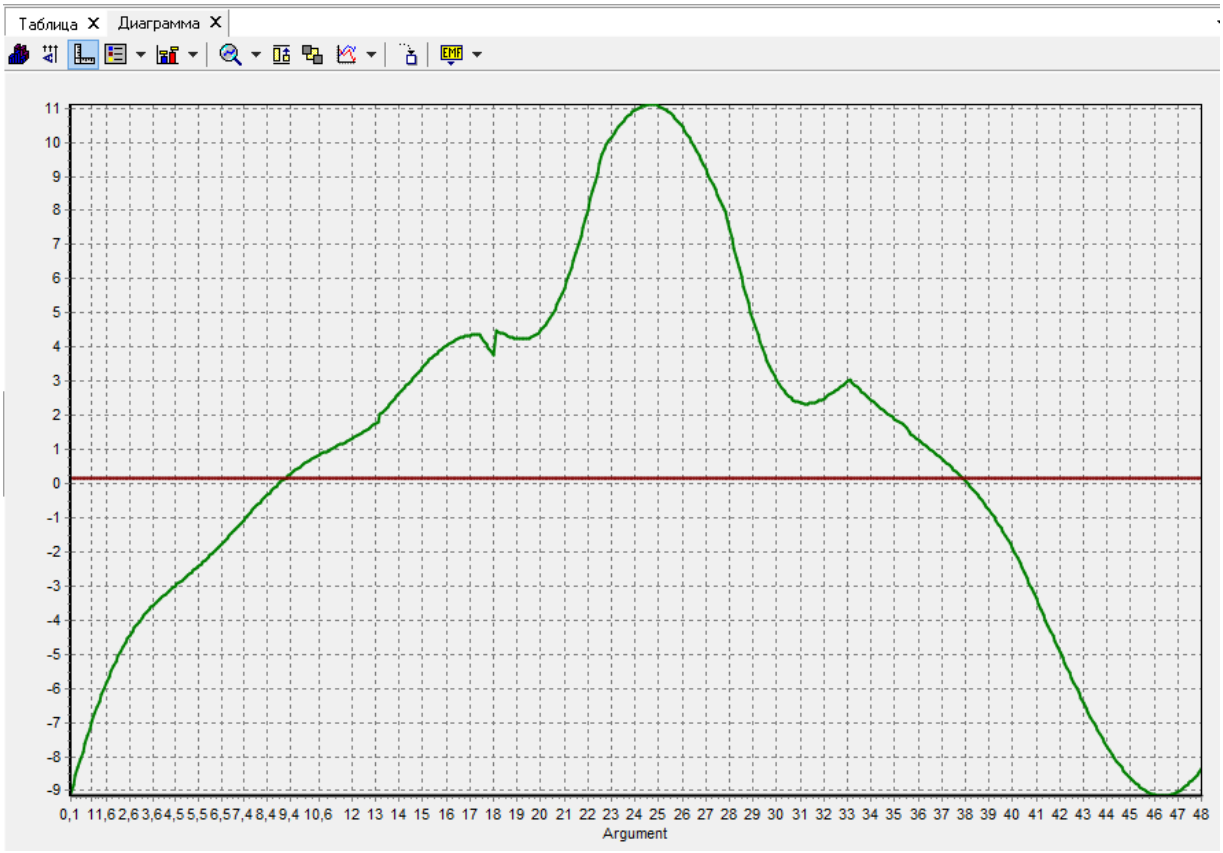


Рис.20. Виродження згладженої функції у константу, з **Level**=9, **Order**=1.

Таким чином, можна стверджувати що для оптимального результату згладжування функції з високочастотним шумом з відносним відхиленням ~ 0,3 – 0,4 від амплітуди початкової функції краще обирати вейвлет-перетворення з параметрами **Level**=4, **Order** від 3 до 8 в залежності від апріорної інформації про наявність гострих локальних екстремумів у початкової функції (наскільки вона гладка). Тобто, при наявності інформації про існування гострих локальних екстремумів треба обирати менші значення **Order**=3 або 4. Чим більш **Order** тим більш згладжену (менш фрактальну) функцію отримуємо на виході. У розглянутому прикладі ми обираємо оптимальні значення **Level**=4, **Order**=6.

3.2. Метод згладжування - Фур'є-перетворення

Другий метод згладжування - Фур'є-перетворення містить один цілочисельний параметр настройки фільтру шуму: величина частотної смуги пропускання (**СП**) від 1 до 100. Це безрозмірний параметр, що прив'язаний до спектру частот вхідної функції. Чим менш значення **СП**, тим менш частотна смуга пропуску гармонік спектру, тим більш згладжується функція. Для розглянутого вище приклада функції з великим шумом (стовпчик High, функція з великими шумами зображена на Рис.10) проведено згладжування для різних значень **СП**. Результати аналізу надані на Рис.21 – 26, де зеленим кольором зображена початкова функція, коричньовим – згладжена Фур'є-перетворенням функція на виході.

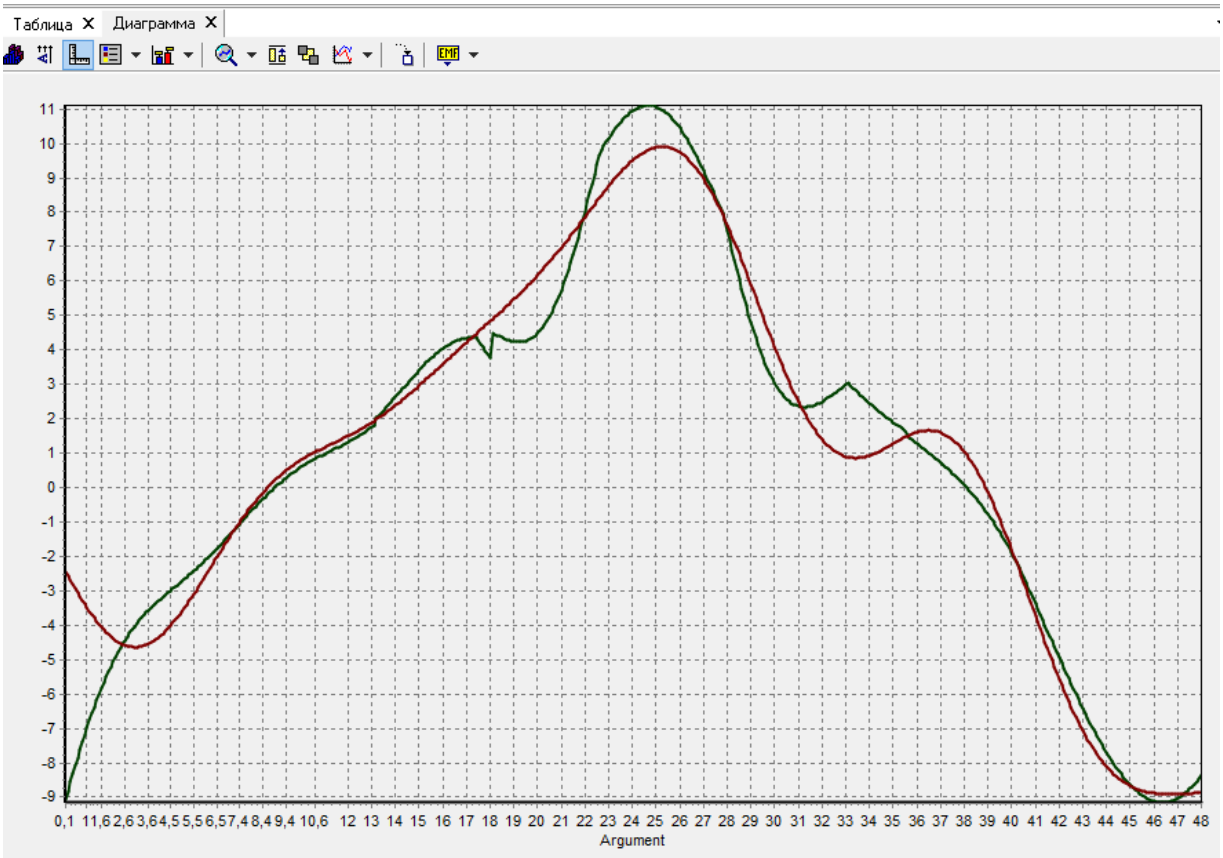


Рис.21. Значне згладжування шуму Фур’є-перетворенням, **СП**=2.

Рис.22. Згладжування шуму Фур’є-перетворенням, **СП**=4.

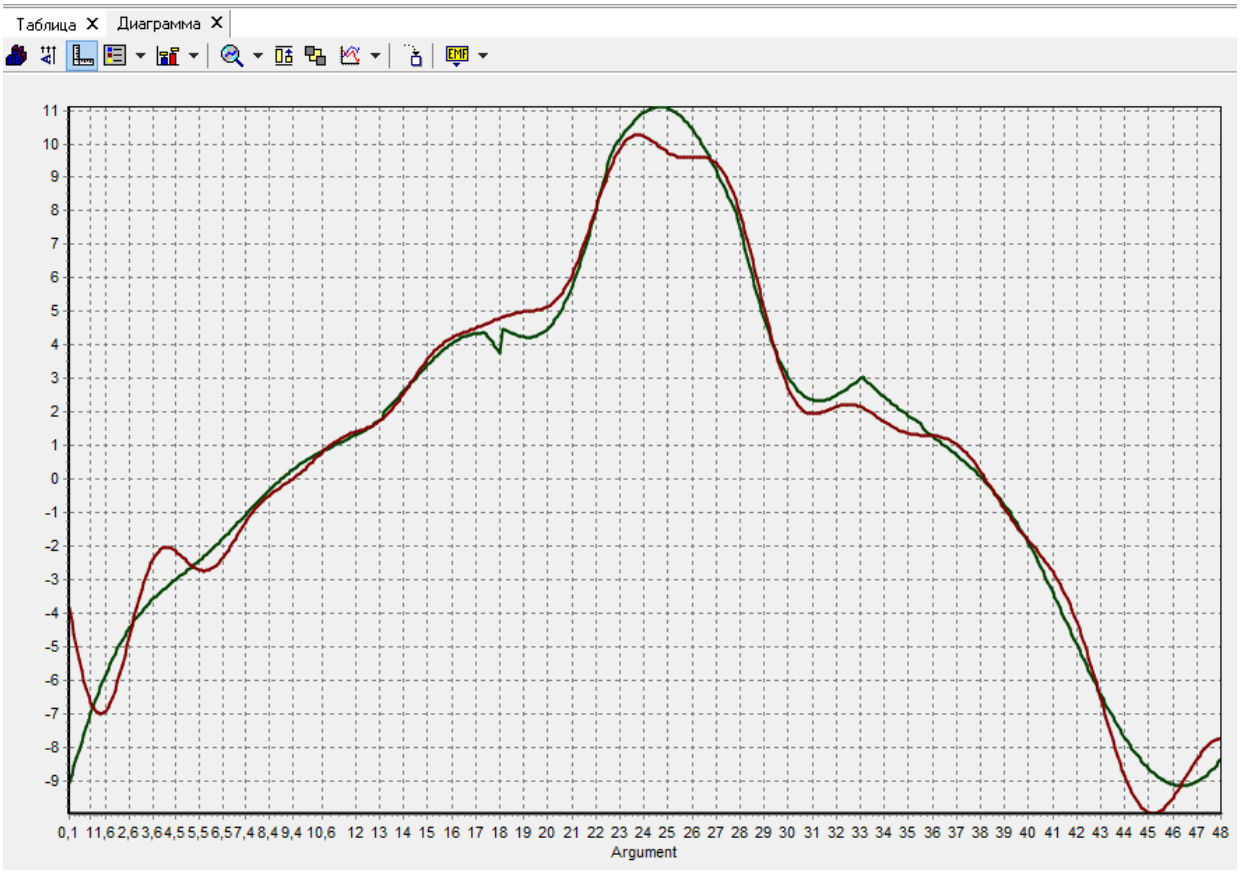


Рис.23. Згладжування шуму Фур’є-перетворенням, **СП**=5.

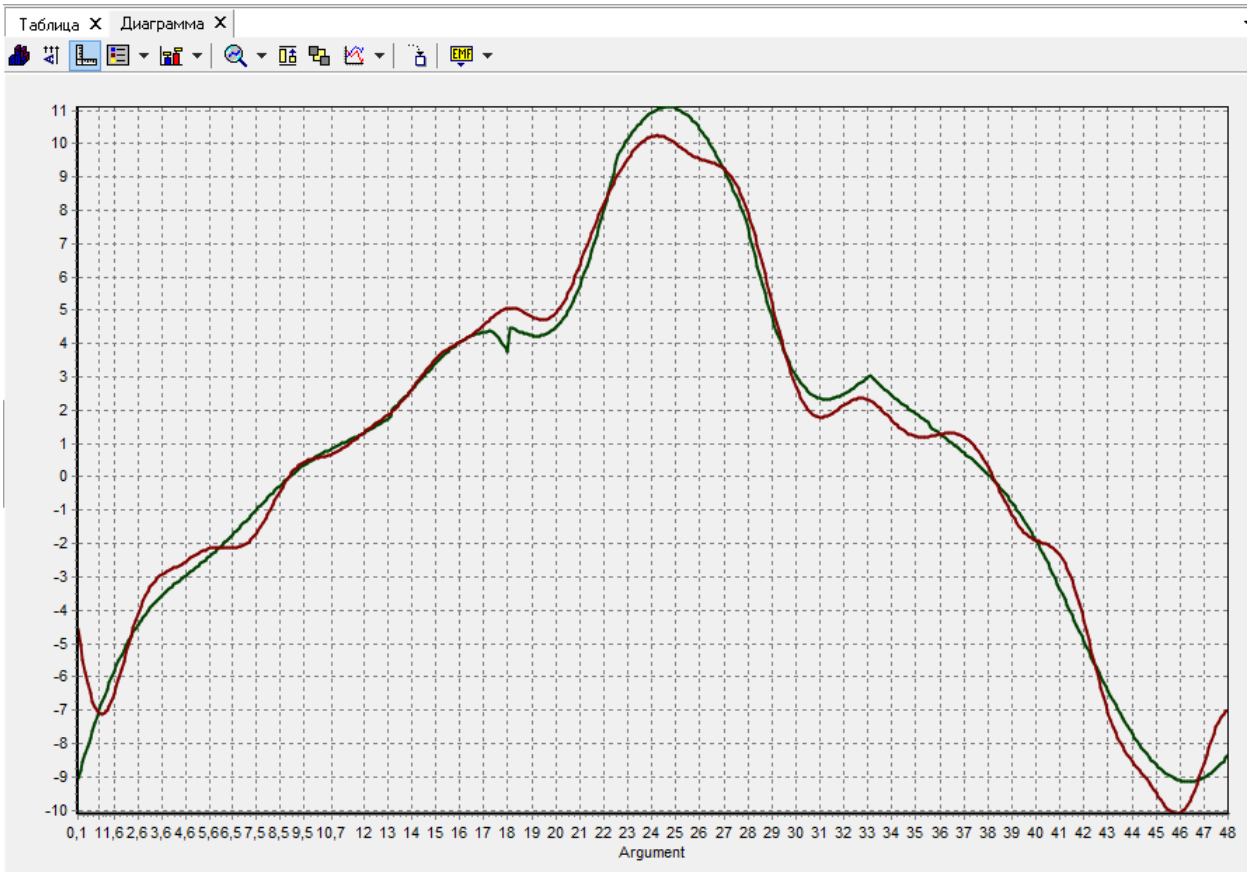


Рис.24. Згладжування шуму Фур’є-перетворенням, **СП**=7.

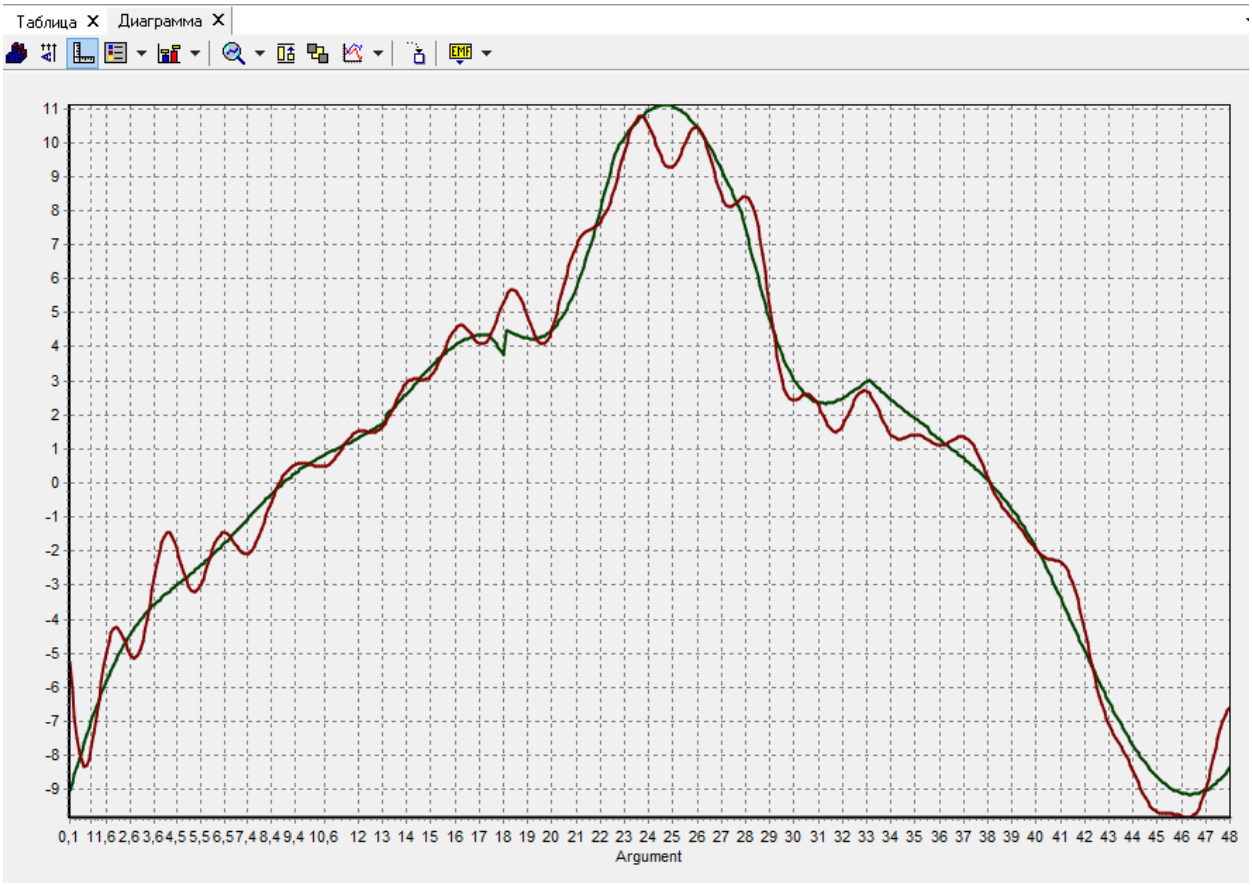


Рис.25. Слабке згладжування шуму Фур’є-перетворенням, **СП**=10.

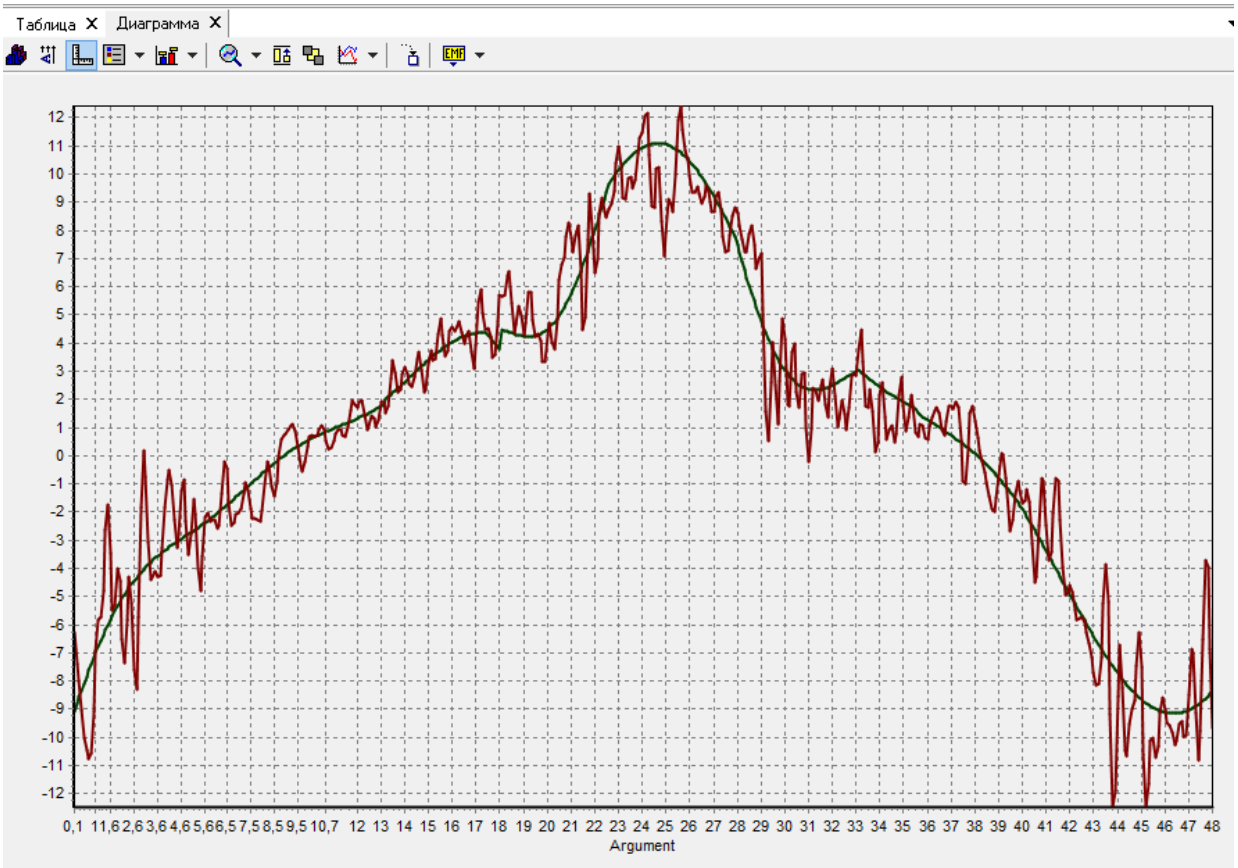


Рис.26. Незначне згладжування шуму Фур’є-перетворенням, **СП**=50.

На Рис.21 – 26 представлені результати згладжування функції з шумом при різних значеннях частотної смуги пропускання **СП**. Чим менш значення **СП**, тим більш частотний діапазон гармонік що підлягають фільтрації. Найбільш значна фільтрація (**СП**=2) показана на Рис.21. За рахунок значної фільтрації згладжена функція на виході навіть не відтворює усі локальні екстремуми початкової функції, що приводить до втрачання важливої інформації. Подальша фільтрація з розширенням смуги пропускання до **СП**=4 приводить до побудови згладженої функції що більш точно апроксимує початкову функцію (Рис.22). Аналогічний результат отримуємо для більших значень **СП**=5, 6, 7 (Рис.23, 24). Слід відзначити що для цих значень смуги пропускання зростає кількість неправдивих локальних екстремумів у згладженої функції, що може привести до помилкових висновків про поведінку початкової функції. При великих значеннях **СП**=10, 50,…,100 (Рис.25, 26) відбувається незначне згладжування функції з шумом що приводе до результуючих функцій з великою кількістю неправдивих локальних екстремумів та невірному результату аналізу даних. Якщо обрати значення **СП** занадто великим, то ступінь згладжування буде недостатньою, а шум буде пригнічений не повністю. Якщо **СП** буде занадто малою, то разом з шумом можуть виявитися пригніченими і зміни, що несуть корисну інформацію. Остаточний вибір **СП** залежить від суб'єктивних критеріїв аналітика при роботі з даними. На даному прикладі можна побачити, що для фільтрації високих шумів можна обирати значення смуги пропускання у діапазоні від 4 до 7. Оптимальним значенням **СП** що дає максимально наближену згладжену функцію до початкової є **СП**=4.

Порівняємо результати роботи обох методів. Для аналізу візьмемо згладжені вейвлет-перетворенням та Фур'є-перетворенням функції, що більш точно апроксимують початкову тестову функцію. Для вейвлет-перетворення **Level**=4, **Order**=3 (Рис.27), для Фур'є-перетворення **СП** = 7 (Рис.28).

Для порівняльного аналізу двох методів проведено також згладжування функцій з малими шумами (стовпчик Low, Рис.8) та середніми шумами (стовпчик Middle, Рис.9). Для вейвлет-перетворення обрано кращі параметри **Level**=4, **Order**=3 (Рис.29, 30), для Фур'є-перетворення **СП** = 7 (Рис.31, 32).

Аналіз графіків на Рис. 27 - 32 говорить про те що при відповідному виборі параметрів фільтрації обидві методи видають приблизно однакові результати – графіки обох згладжених функцій приблизно співпадають. До плюсів вайвлет-перетворення можна віднести більшу точність відтворення первинної функції у випадку коли ця функція не є гладкою та містить декілька точок розриву першої похідної (т.з. фрактальна функція). Так, відтворення гострого локального максимуму згладженою функцією показано на Рис.27, 29, 30 при значенні аргументу Arumentg=33. Фільтрація функцій з шумом методом Фур'є – перетворення не в змозі відтворювати первинні функції даного класу. Але, з іншого боку, результуючі (відфільтровані) методом вейвлет-перетворення функції, на відміну від Фур'є – перетворення, можуть містити неправдиві гострі локальні екстремуми що приводить до невірних інтерпретацій результату та знижує точність аналізу даних. Метод Фур'є-перетворення може ефективно використовуватися для фільтрації шуму та відтворення саме гладких (неперервно-диференційованих) функцій.

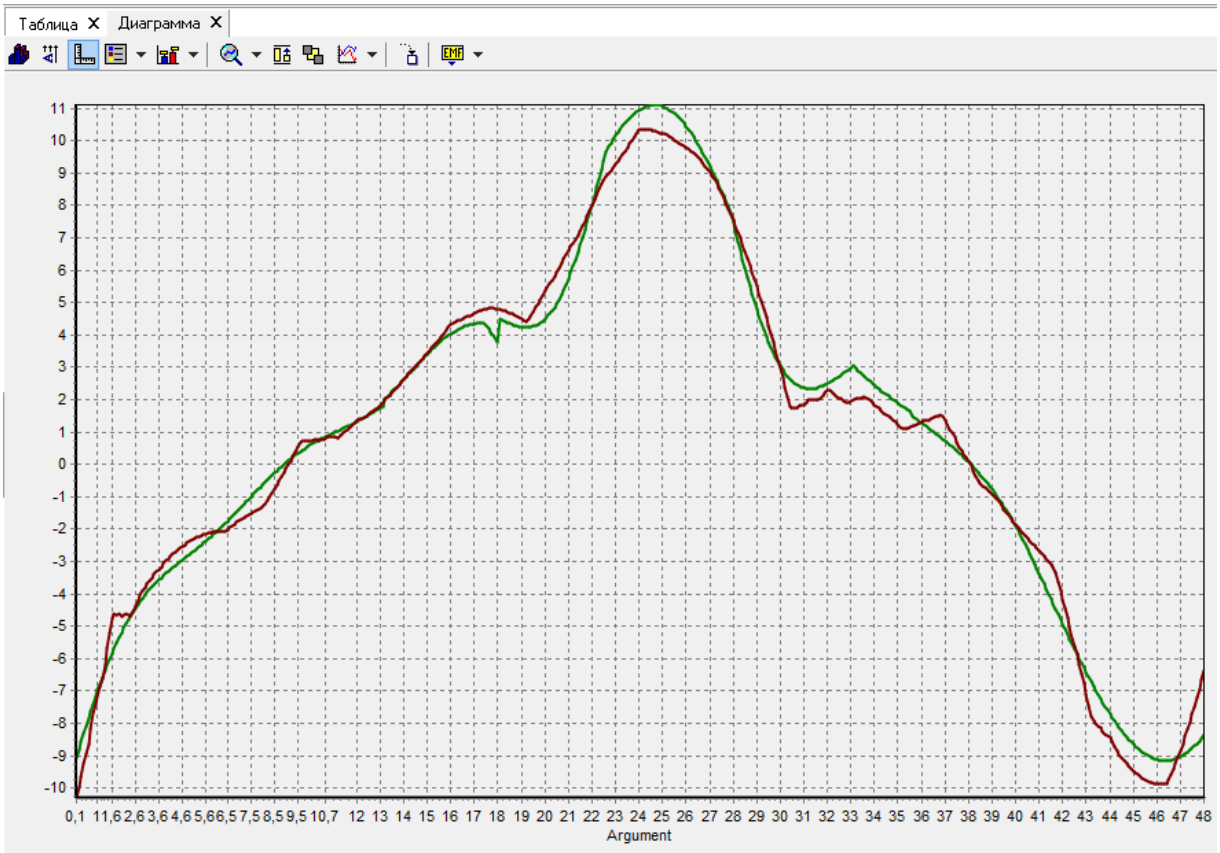


Рис. 27. Згладжування великого шуму вейвлет-перетворенням з **Level**=4, **Order**=3.

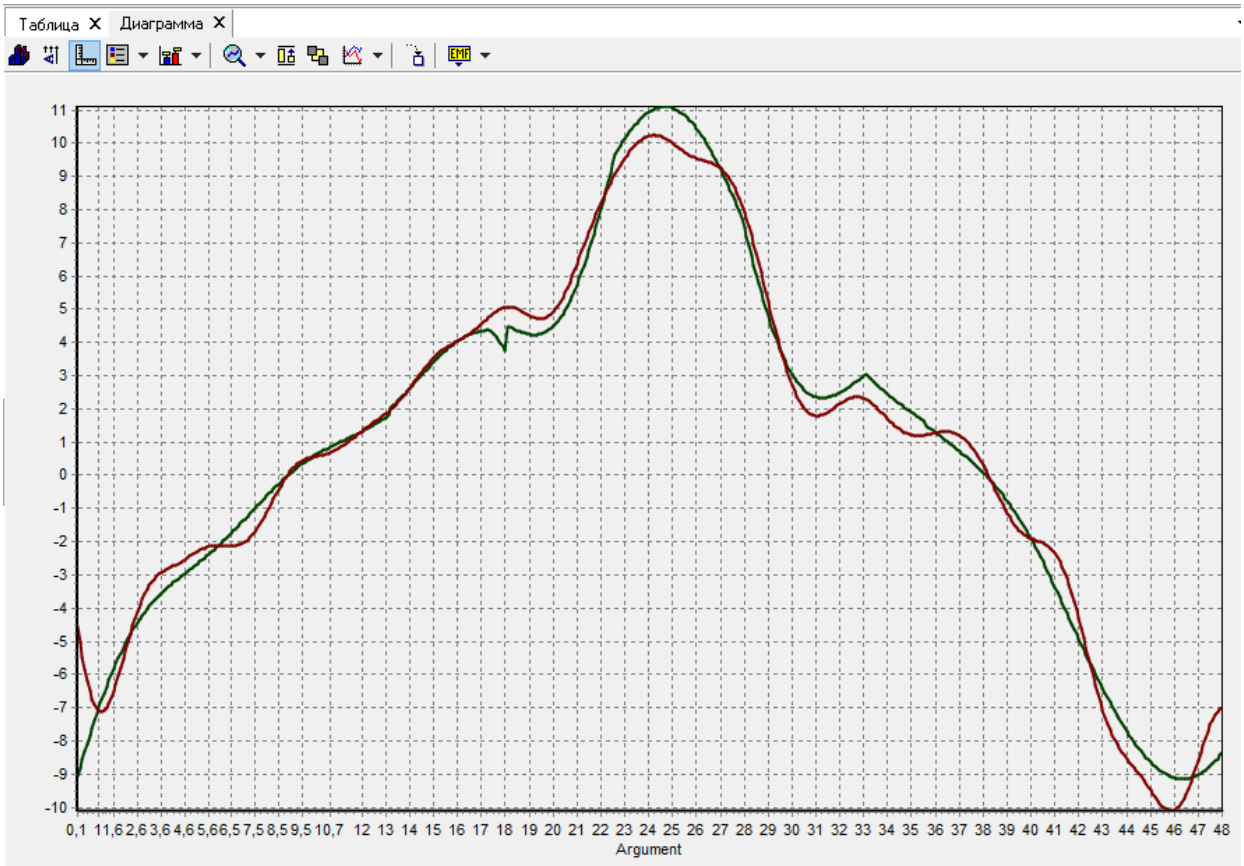


Рис.28. Згладжування великого шуму Фур’є-перетворенням, СП=7.

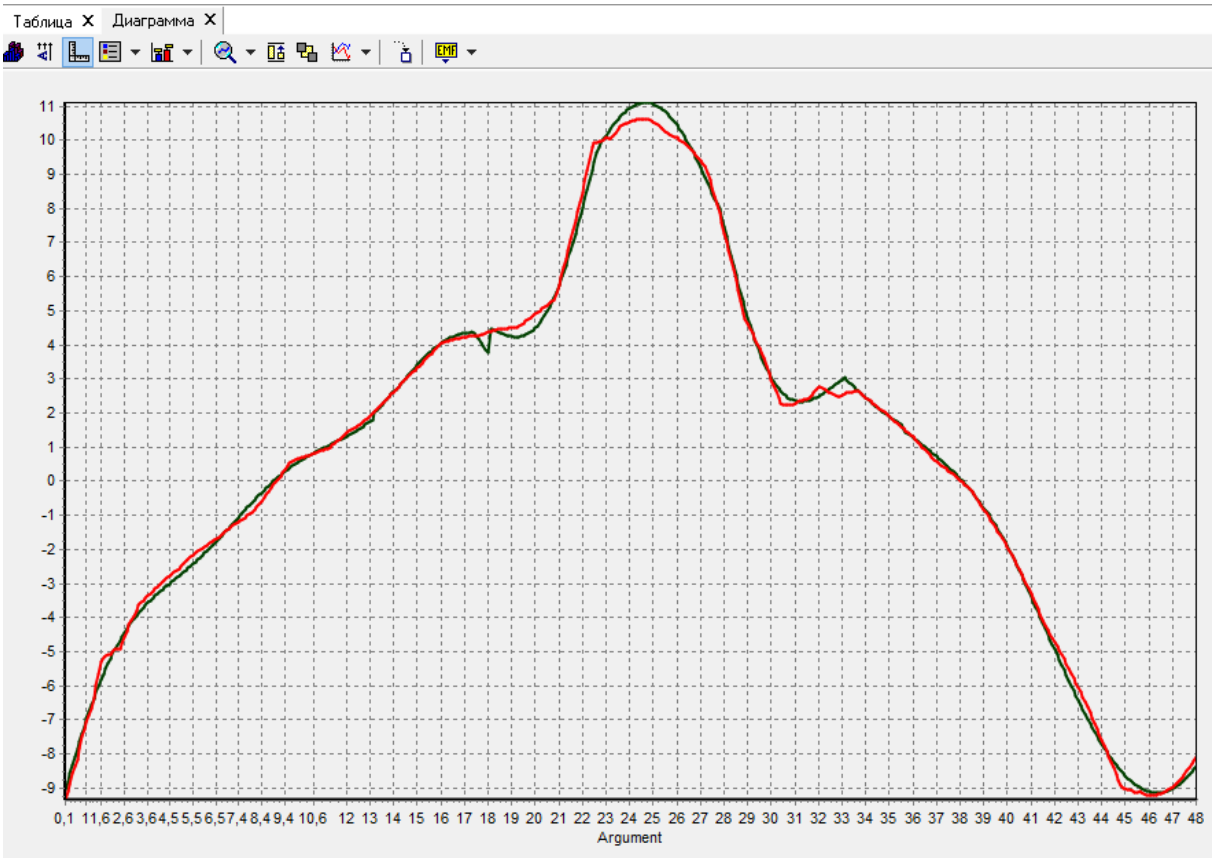


Рис. 29. Згладжування маленького шуму вейвлет-перетворенням з **Level**=4, **Order**=3.

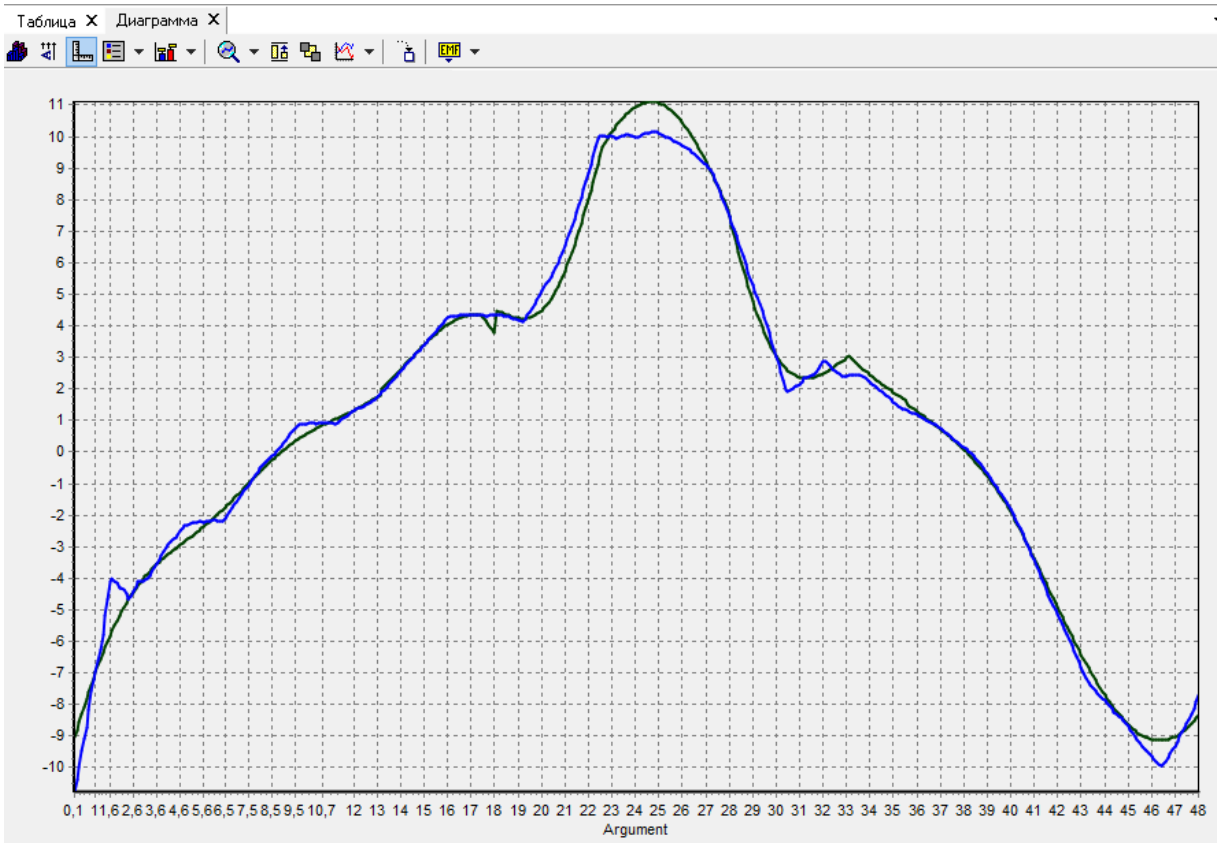


Рис.30. Згладжування середнього шуму вейвлет-перетворенням з **Level**=4, **Order**=3.

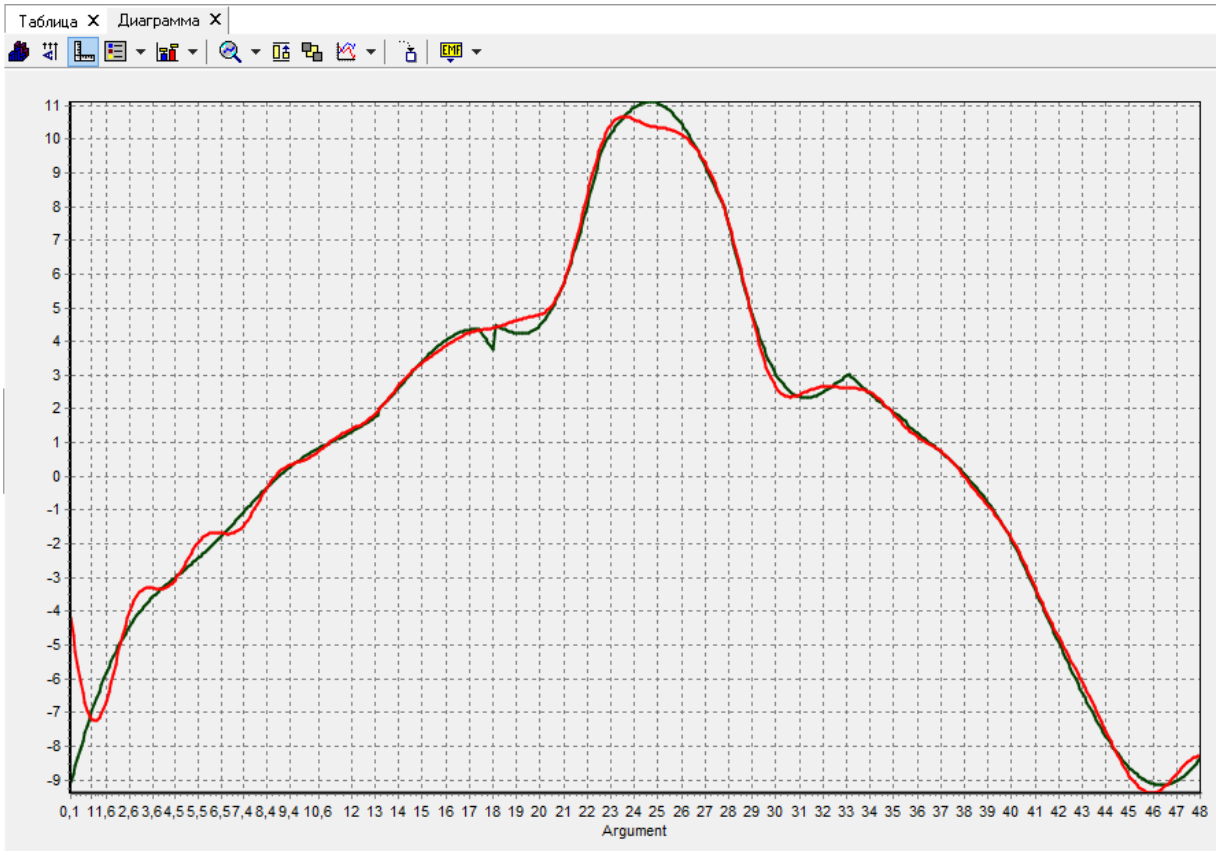


Рис.31. Згладжування маленького шуму Фур’є-перетворенням, СП=7.

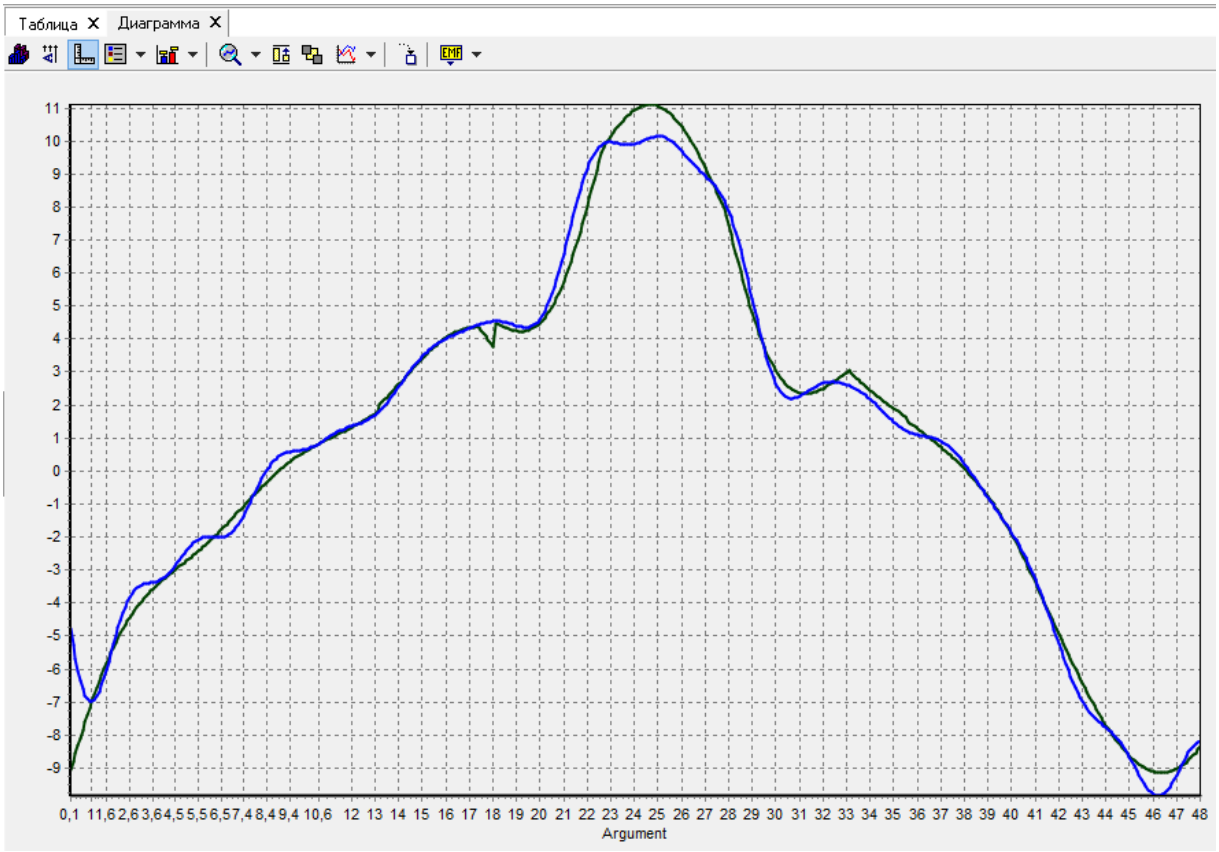


Рис.32. Згладжування середнього шуму Фур’є-перетворенням, СП=7.

Таким чином, для отримання достовірного результату та проведення ефективного аналізу даних доцільно залучати додаткову інформації про тип та властивості вхідної функції та використовувати обидві розглянуті методи. Для методів необхідно обирати параметри згладжування з діапазонів значень, отриманих на тестових функціях.

Завдання для лабораторної роботи 4.

1. Створити файл Lab4.xlsx. Побудувати таблицю даних
2. Перший стовпчик A з назвою N містить цифри 0, 1, 2, ..., 50 - номери точок.
3. Другий стовпчик B з назвою Arg містить формулу - значення ячейки ліворуч поділити на 10, тобто N/10 - 0, 0.1, 0.2, ... - аргумент функції.
4. Третій стовпчик C з назвою Func містить формулу $=1+\mathrm{COS}(\mathrm{Arg})*\mathrm{SIN}(\mathrm{Arg})$, де Arg - значення комірок, розташованих ліворуч.
5. Четвертий стовпчик D з назвою Anomal співпадає з третім стовпчиком крім 7, 12, 22, 32, 42 рядків.

У 7 рядку - формула $=10+\mathrm{COS}(\mathrm{Arg})*\mathrm{SIN}(\mathrm{Arg})$,

у 12 - формула $=-15+\mathrm{COS}(\mathrm{Arg})*\mathrm{SIN}(\mathrm{Arg})$,

у 22 - формула $=20+\mathrm{COS}(\mathrm{Arg})*\mathrm{SIN}(\mathrm{Arg})$,

у 32 - формула $=5+\mathrm{COS}(\mathrm{Arg})*\mathrm{SIN}(\mathrm{Arg})$,

у 42 - формула $=-15+\mathrm{COS}(\mathrm{Arg})*\mathrm{SIN}(\mathrm{Arg})$.

6. П'ятий стовпчик E з назвою Fluct_1 містить формулу 0,1*(СЛЧИС()-0,5)+Func, де Func - значення комірки даного рядка зі стовпчика C.
7. Шостий стовпчик F з назвою Fluct_2 містить формулу 0,3*(СЛЧИС()-0,5)+Func, де Func - значення комірки даного рядка зі стовпчика C.
8. Сьомий стовпчик G з назвою Fluct_3 містить формулу 0,5*(СЛЧИС()-0,5)+Func, де Func - значення комірки даного рядка зі стовпчика C.
9. Зберігаємо файл Lab4.xlsx. Зберігаємо цей файл також як Lab1.txt.
10. Завантажуємо Deductor. Робимо імпорт даних з файлу Lab1.txt. Налаштовуємо візуалізацію даних у тривимірному форматі та демонструємо результат викладачу.
11. Майстром обробки даних видаляємо аномалії зі стовпчику Anomal. Візуалізація результату.
12. Майстром обробки даних згладжуємо аномалії зі стовпчику Anomal за допомогою "Вєйвлет - перетворення". Візуалізація результату.
13. Майстром обробки даних видаляємо шуми зі стовпчиків Fluct_1, Fluct_2, Fluct_3 за допомогою "Видалення шуму" (відповідно до стовпчиків - ступень видалення мала, середня та велика) Візуалізація результату.
14. Зробити цю процедуру також за допомогою "Вєйвлет - перетворення". Візуалізація результату. Порівняння результату.

Тема 5. Задачі прогнозування часових рядів на основі штучних нейронних мереж та лінійних регресій декількох змінних

Задачі: побудувати моделі штучної нейронної мережі та лінійної регресії декількох змінних на основі навчаючих та тестових вибірок; побудувати за допомогою нейромережі та лінійної регресії прогноз часового ряду.

Методи: методи прогнозування на основі моделей штучних нейронних мереж та лінійних регресій декількох змінних.

1. Прогнозування часового ряду за допомогою нейронних мереж

Математичний апарат штучних нейронних мереж часто використовуються для побудови кібернетичних моделей що імітують динаміку складних нелінійних процесів різного типу для яких неможливо записати в аналітичній формі закон балансу та отримати залежності між вхідними та вихідними параметрами, аналітичні розв'язки у явному вигляді. Загальна теоретична модель штучної нейронної мережі надана в [8], [9], [16]. Для побудови моделей нейронних мереж використовується вибірка ретроспективних даних спостережень за динамікою процесу у вигляді часового ряду. На першому етапі вибірка даних розбивається на навчаючу та тестуючу вибірки. Навчаюча вибірка використовується для "навчання" нейронної мережі, тобто визначення набору констант – вагових коефіцієнтів нейронів. Тестуюча вибірка використовується для обчислення величини похибки прогнозу нейронної мережі.

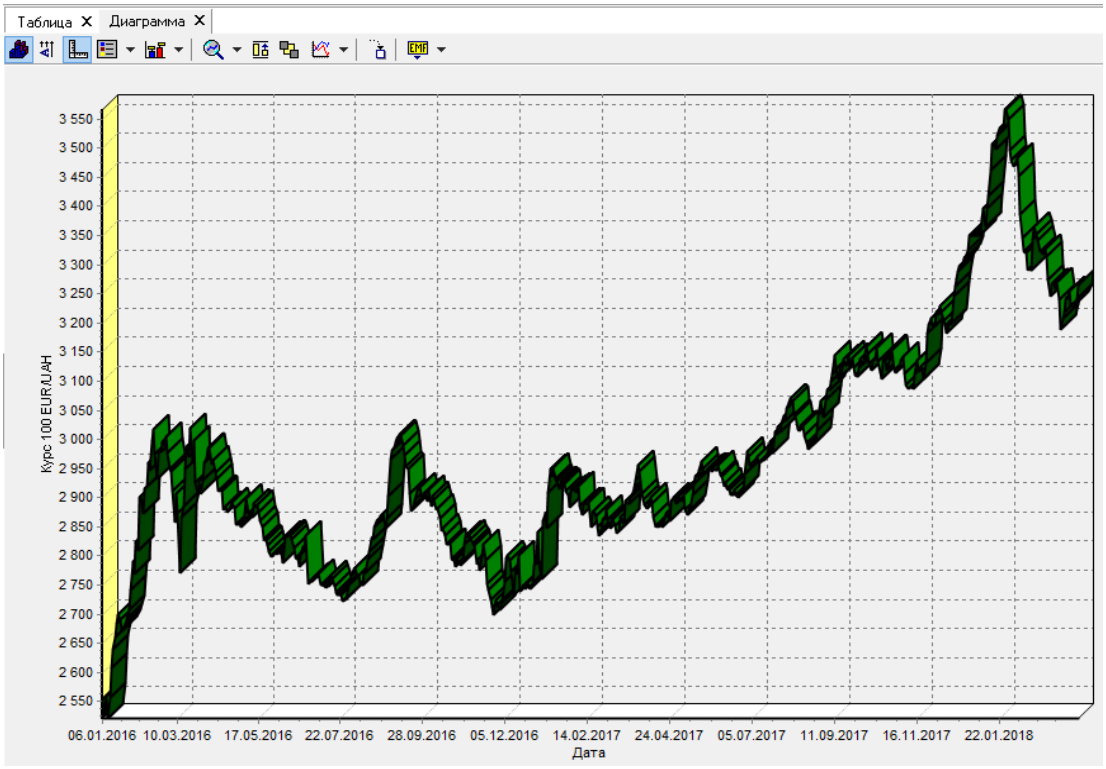


Рис.1. Імпортовані дані офіційного курсу пари 100 Євро/Гривня (EUR/UAH).

Для розв'язку задачі прогнозування часового ряду будемо використовувати дані офіційного курсу НБ України пари 100 EUR/UAH за період 06.01.2016 – 27.03.2018 збережені у вигляді таблиці у файлі task51.txt. Для аналізу якості

прогнозу на 10 днів будемо використовувати дані за період 28.03.2018-11.04.2018 збережені у файлі test51.txt. Після імпорту даних з файлу task51.txt система створює інформаційний вузол у лівому вікні користувача. Імпортовані дані часового ряду показані на Рис.1. Далі необхідно провести автокореляційний аналіз для виявлення сезонних складових часового ряду. Встановлюємо курсор на створеному вузлі та запускаємо майстер обробки . Обираємо “Автокореляция” та встановлюємо атрибут стовпчика “Курс 100 EUR/UAH” “Используемое”, “Количество отсчетов” максимальне 557 (Рис.2). Із отриманого графіку функції автокореляції (Рис.3) випливає що часовий ряд має сезонну компоненту протягом 344 днів та покриває даний період.

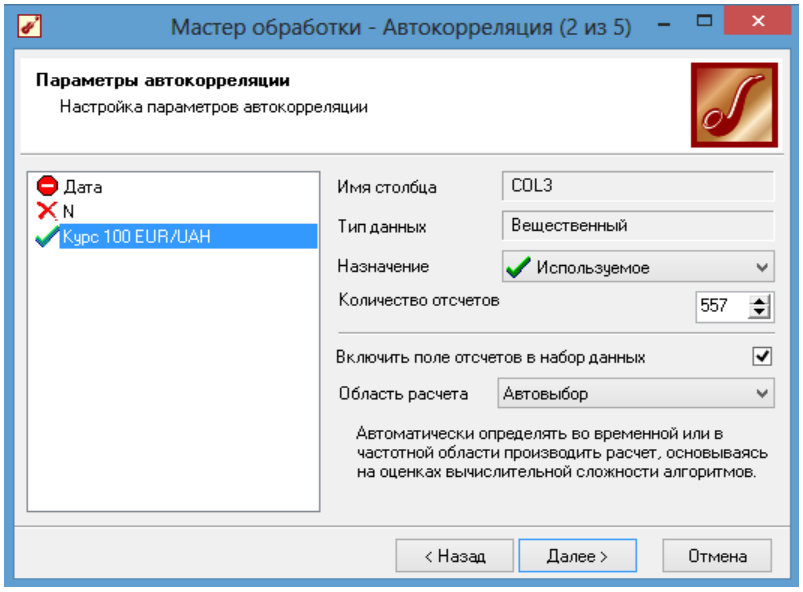


Рис.2. Вікно налаштування автокореляції.

Рис.3. Функція автокореляції часового ряду.

Далі визначаємо гіпотетичну кількість ретроспективних даних що складають набір вхідних даних нейронної мережі – 30 одиниць (орієнтиром виступають періоди локальних коливань ряду на Рис.1, ~ 30 днів). Дане припущення буде перевірятися далі результатами навчання та тестування нейромережі та може бути змінено у разі незадовільних результатів. На основі імпортованих даних необхідно сформувати таблицю вхідних даних методом ковзаючого вікна. Встановлюємо курсор на вузлі імпортованих даних та після запуску майстра обробки обираємо “Скользящее окно”. Налаштування методу із “Глубиною погружения” 30 показано на Рис.4.

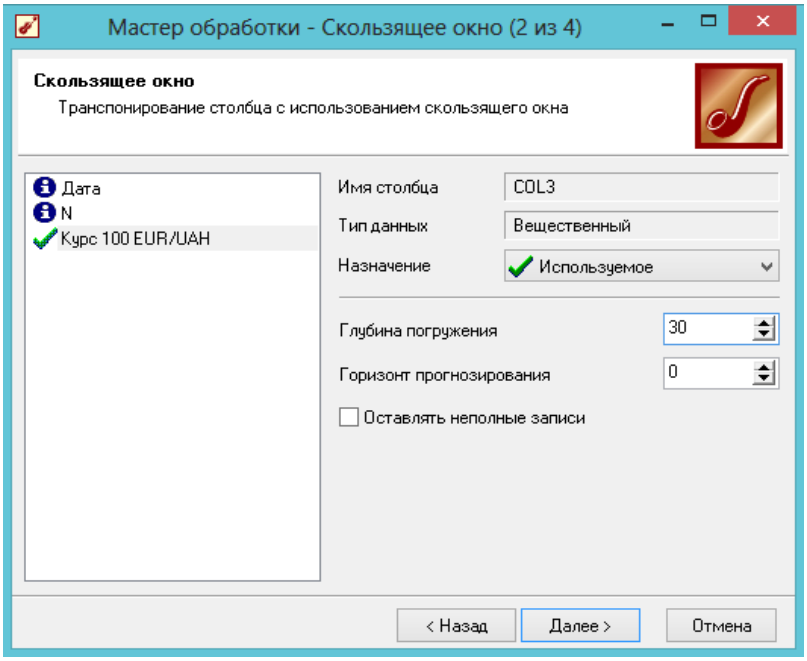


Рис.4. Налаштування ковзаючого вікна.

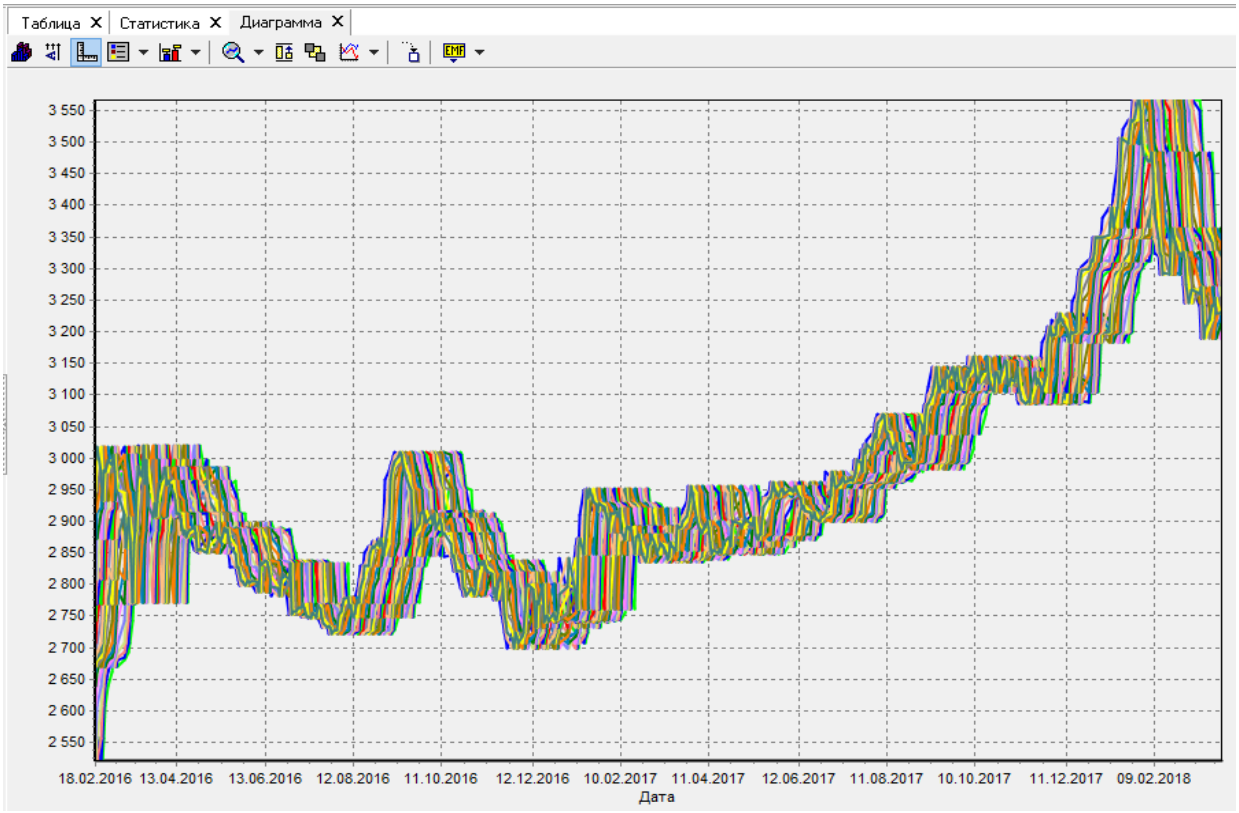


Рис.5. Результат роботи ковзаючого вікна.

В результаті отримаємо таблицю з 33 стовпчиками Дата, N, Курс 100 EUR/UAH - 30, Курс 100 EUR/UAH – 29, …, Курс 100 EUR/UAH. Встановлюємо курсор на вузлі, створеному ковзаючим вікном та після завантаження майстра обробки обираємо “Нейросеть”. Налаштовуємо поля Дата, N – інформаційні, Курс 100 EUR/UAH – вихідне, решта – вхідні (Рис.6).

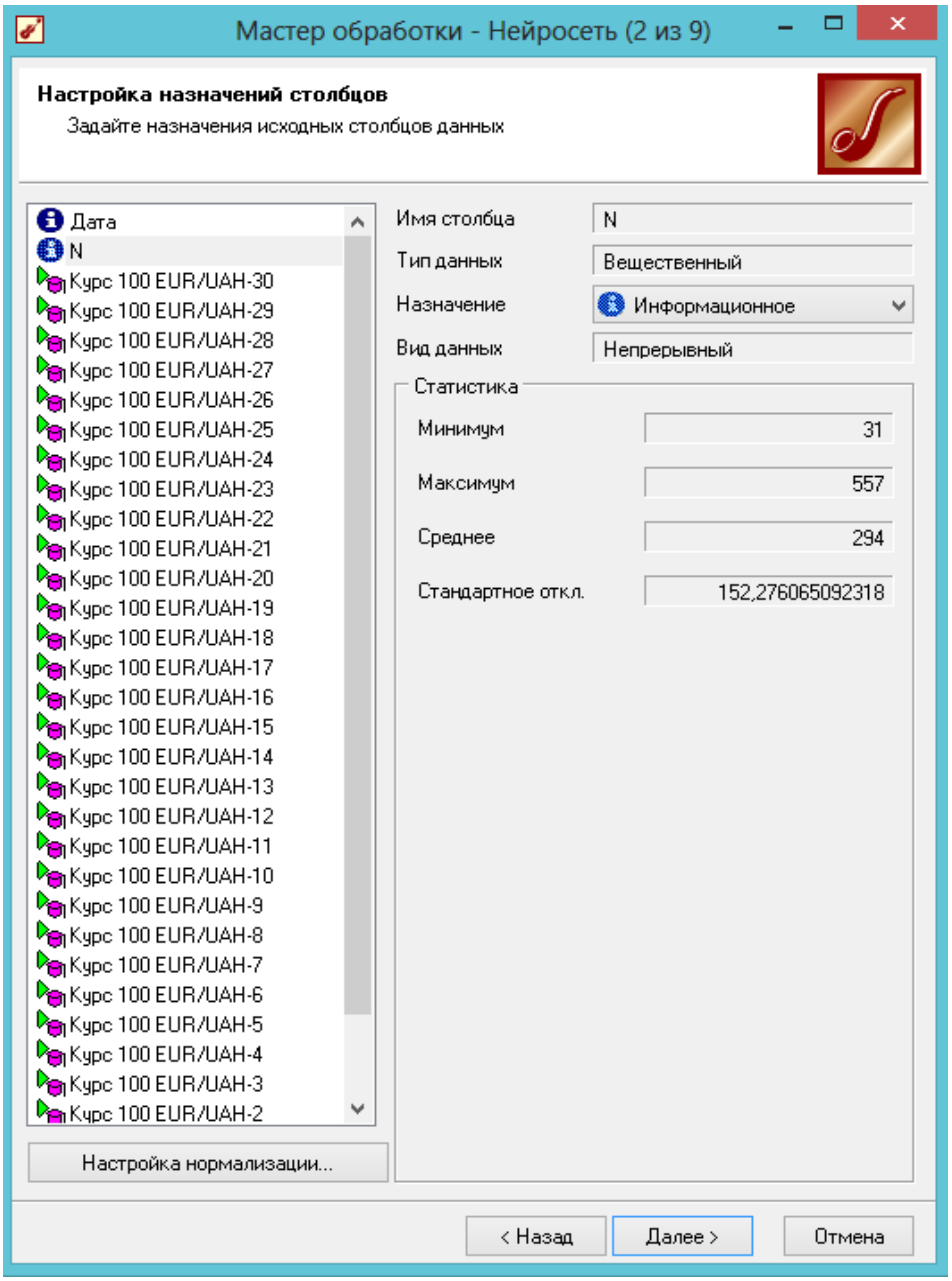


Рис.6. Налаштування вхідних даних нейромережі.

У наступному вікні вказуємо параметри розбиття вибірки даних: розділення вибірки “Случайно”, навчаюча вибірка 90%, тестуюча 10% (Рис.7). На наступному кроці налаштовуємо у першому наближенні таку структуру нейронної мережі: кількість прихованих шарів 3; кількість нейронів у вхідному шарі - 30; у вихідному шарі – 1, у кожному прихованому шарі – 3; функція активації нейронів – “Сигмоида” (Рис.8).

Далі обираємо алгоритм навчання мережі “Resilient propagation” (RPROP) та встановлюємо його параметри “Шаг спуска” - 0,01 та “Шаг подъема” - 1,2 (Рис.6). Алгоритм називається “resilient back propagation” [8], [9], [16] та являє

собою модифікацію методу “back propagation” з підвищеною швидкодією навчання нейромережі. На кожній ітерації алгоритму RPROP вагові коефіцієнти нейронної мережі модифікуються згідно тільки знаку частинних похідних мінімізуючої функції (на відміну від обчислення їх значень у методі “back propagation”) та заданих фіксованих параметрів спуску та підйому на поверхні функції, щоб поліпшити рішення одного прикладу. Для підвищення відсотку розпізнавання даних нейромережою користувач може змінювати значення параметрів методу навчання (Рис.9).

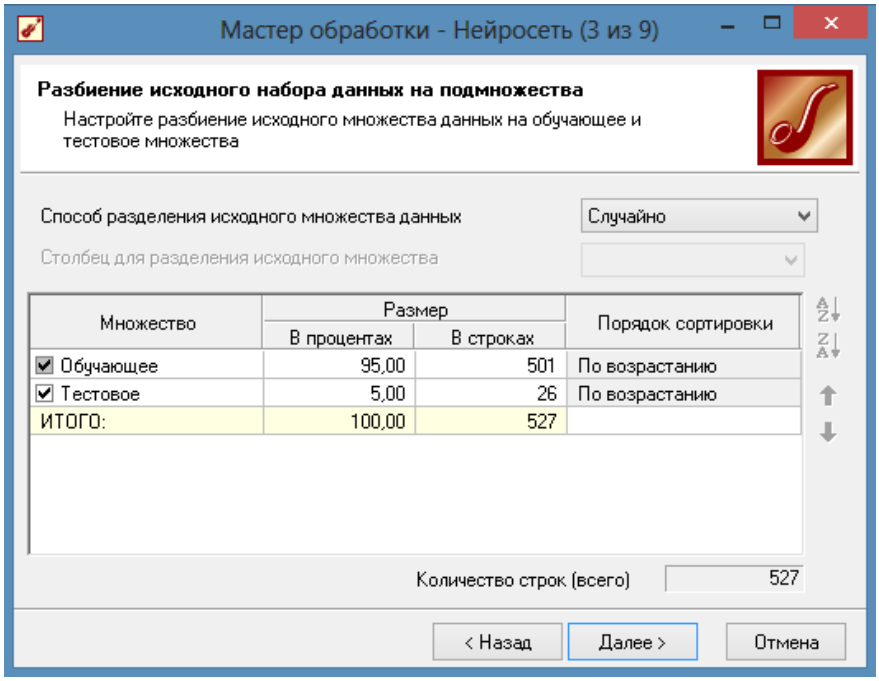


Рис.7. Розбиття вибірки даних.

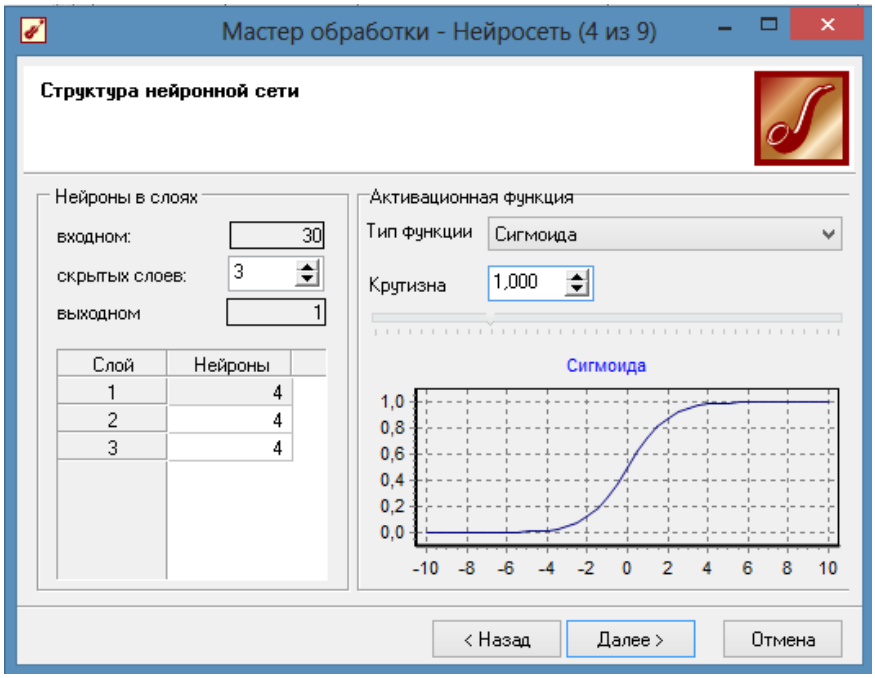


Рис.8. Налаштування структури нейронної мережі [30x4x4x4x1].

Далі налаштовуємо умови зупинки навчання. Приклад можна вважати розпізнаним, якщо відносна помилка відхилення менша за 0,005 (0,5%). Навчання автоматично зупиняється при досягненні епохи (кроків) 20000 (Рис.10).

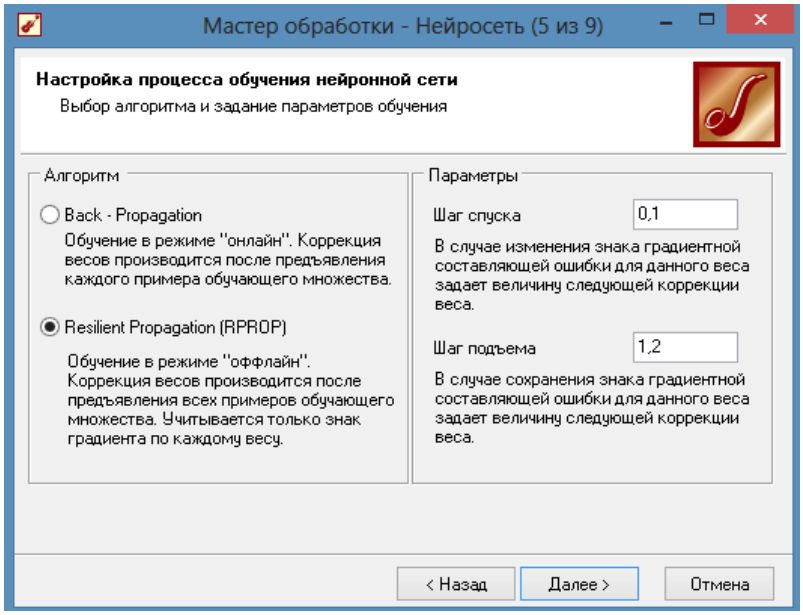


Рис.9. Налаштування параметрів навчання нейронної мережі.

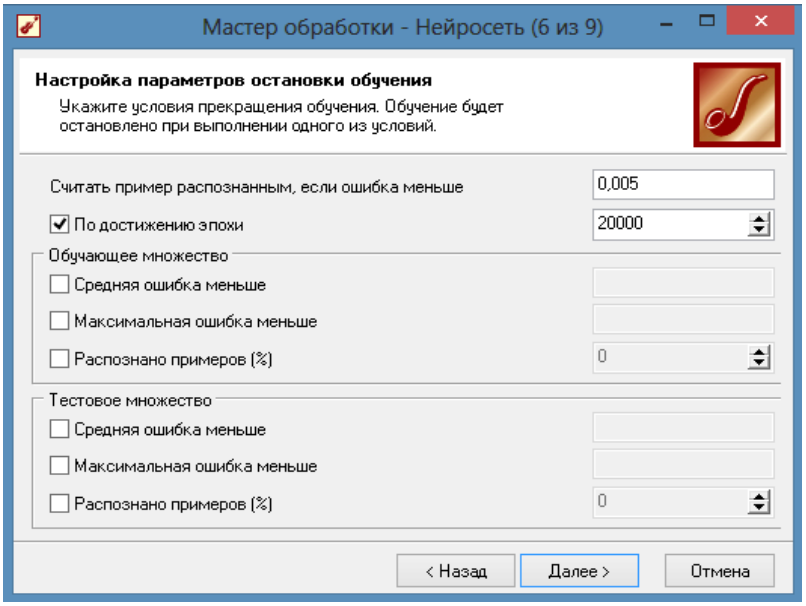


Рис.10. Параметри зупинки навчання.

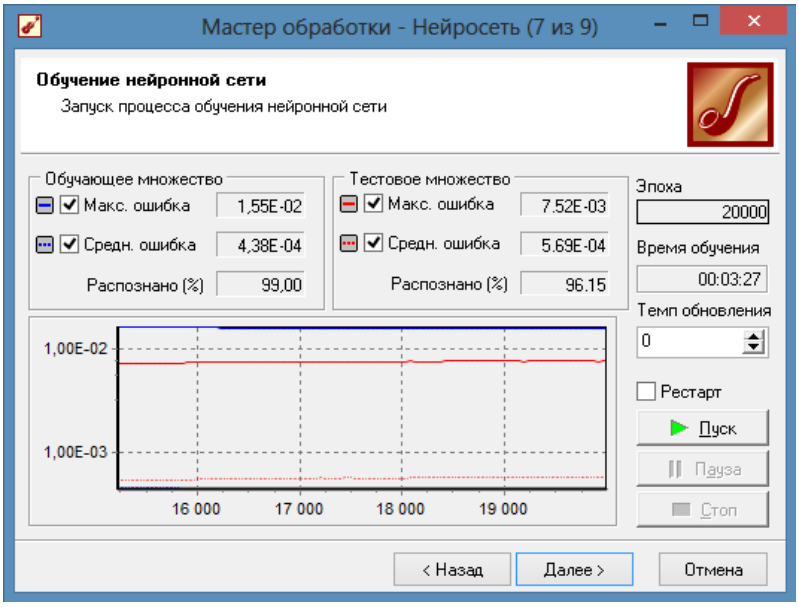


Рис.11. Навчання та тестування мережі з 3 прихованими шарами (по 4 нейрони).

На наступному кроці запускаємо процес та спостерігаємо результат навчання та тестування мережі у графічному форматі (Рис.11). Можна примусово зупинити процес навчання (кнопка Стоп) коли нейромережа розпізнала вже 100% точок. На Рис.11 показано, що за 20000 кроків (“Епоха”) процес зупинено при розпізнанні 99% навчаючої та 96,2% тестової вибірках. Якщо результати роботи алгоритму не є задовільними, на цьому кроці можна повернутися назад, задати нову структуру мережі та скорегувати параметри навчаючої та тестуючої вибірки. Далі обираємо набір способів відображення (Рис.12).

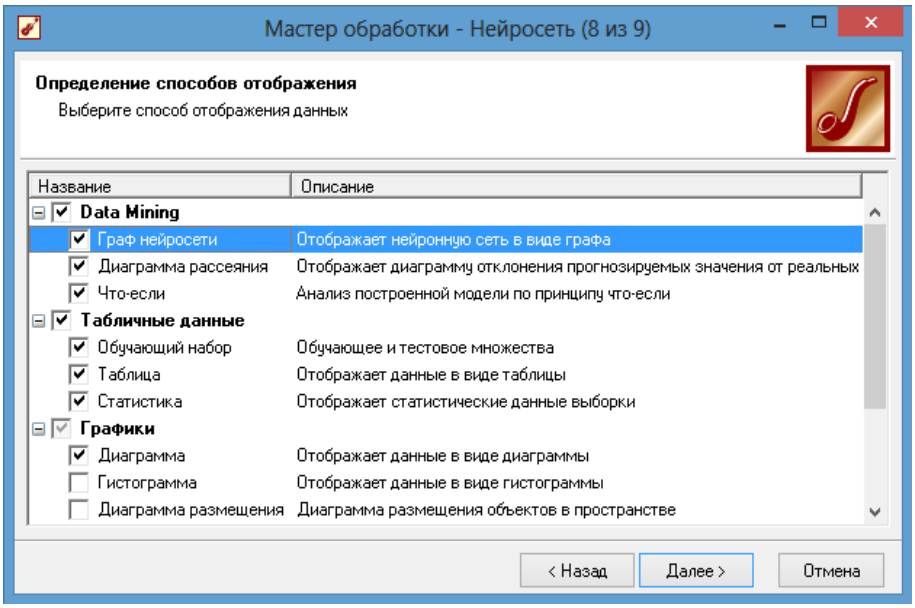


Рис.12. Набір способів відображення результатів.

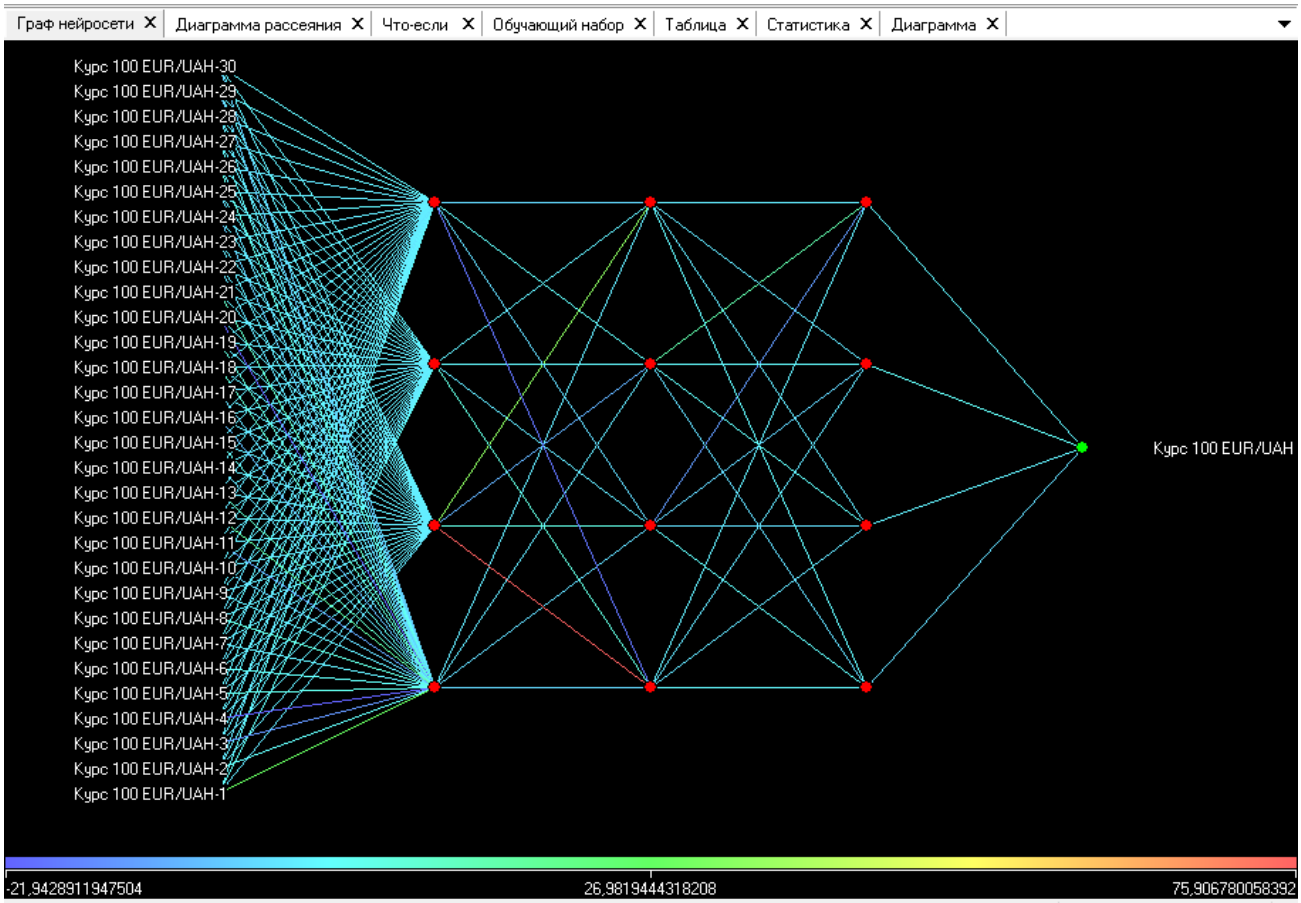


Рис.13. Граф нейронної мережі, модель [30x4x4x4x1].

На Рис.13 показано граф побудованої нейронної мережі з 30 вхідними, 3 прихованими (по 4 нейрони у кожному) та 1 вихідними шарами - [30x4x4x4x1]. Результати навчання представлені на діаграмі розсіювання (Рис.14), яка показує відхилення прогнозованих даних від «еталонних». По вісі абсцисс відкладені значення навчальної вибірки (еталон), по вісі ординат – значення, обчислені за допомогою неромережі на тому ж самому прикладі. Таким чином, дані навчальної вибірки розташовані на еталонній прямій, а дані, що обчислені за допомогою нейромережі в околі прямої лінії. Більш вузький коридор похибки означає наближення обчислених даних до навчаючих і тим більш якісну побудовану модель нейромережі. На Рис.15 показано графіки початкового часового ряду (порівняно з Рис.1) та графік даних на виході нейромережі.

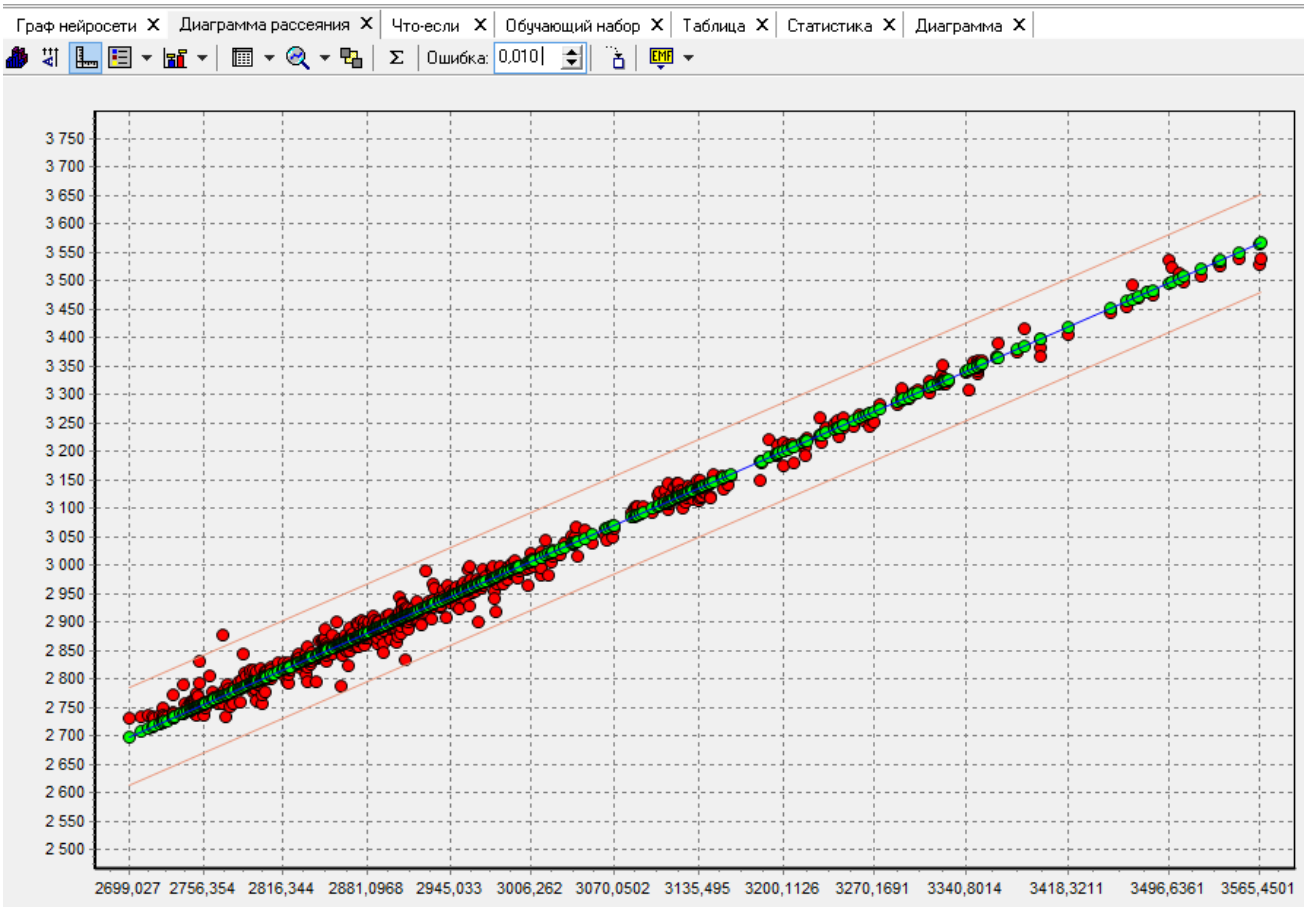


Рис.14. Діаграма розсіювання нейромережі з коридором помилки ±0,01 (1%).

Успішність навчання (99,2% розпізнавання з похибкою 0,5%) та тестування (відповідно, 96,2%) нейронної мережі підтверджує вдалість вибору структури вхідних даних, структури шарів та нейронів мережі, методу та параметрів її навчання. Побудовану модель можна використовувати тепер на наступному етапі для прогнозування часового ряду. Встановлюємо курсор на вузол, створений нейронмережею у лівому вікні користувача та після запуску майстру обробки обираємо прогнозування. У вікні налаштувань (Рис.16) програма автоматично встановлює типи усіх даних та оформлює стовпчики, що будуть використовуватися на кожному наступному кроці алгоритму (із зсувом на один крок уперед). У нижньому вікні вказуємо горизонт прогнозу 10 днів та запускаємо процес із налаштуваннями параметрів системи за замовченням на наступних кроках.

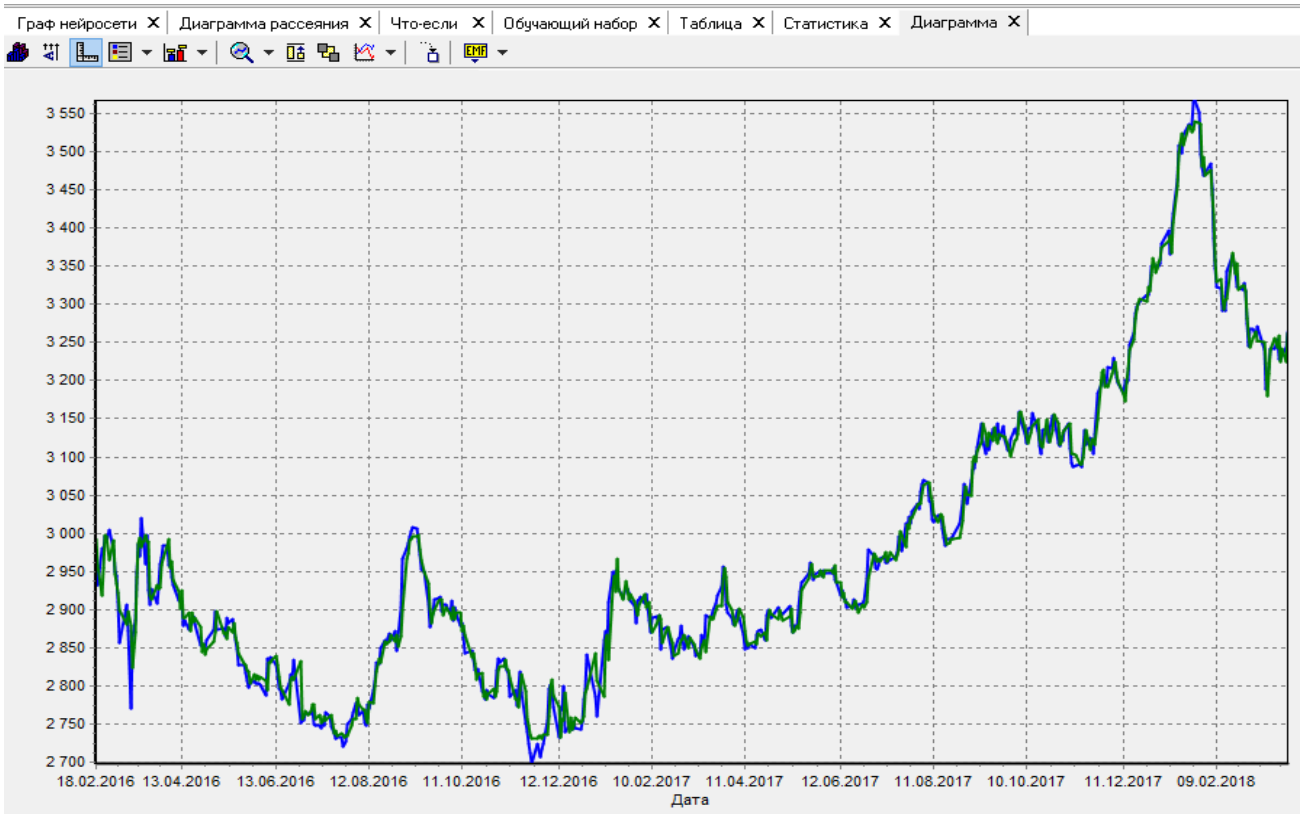


Рис.15. Графіки початкового часового ряду та вихідних значень нейромережі.

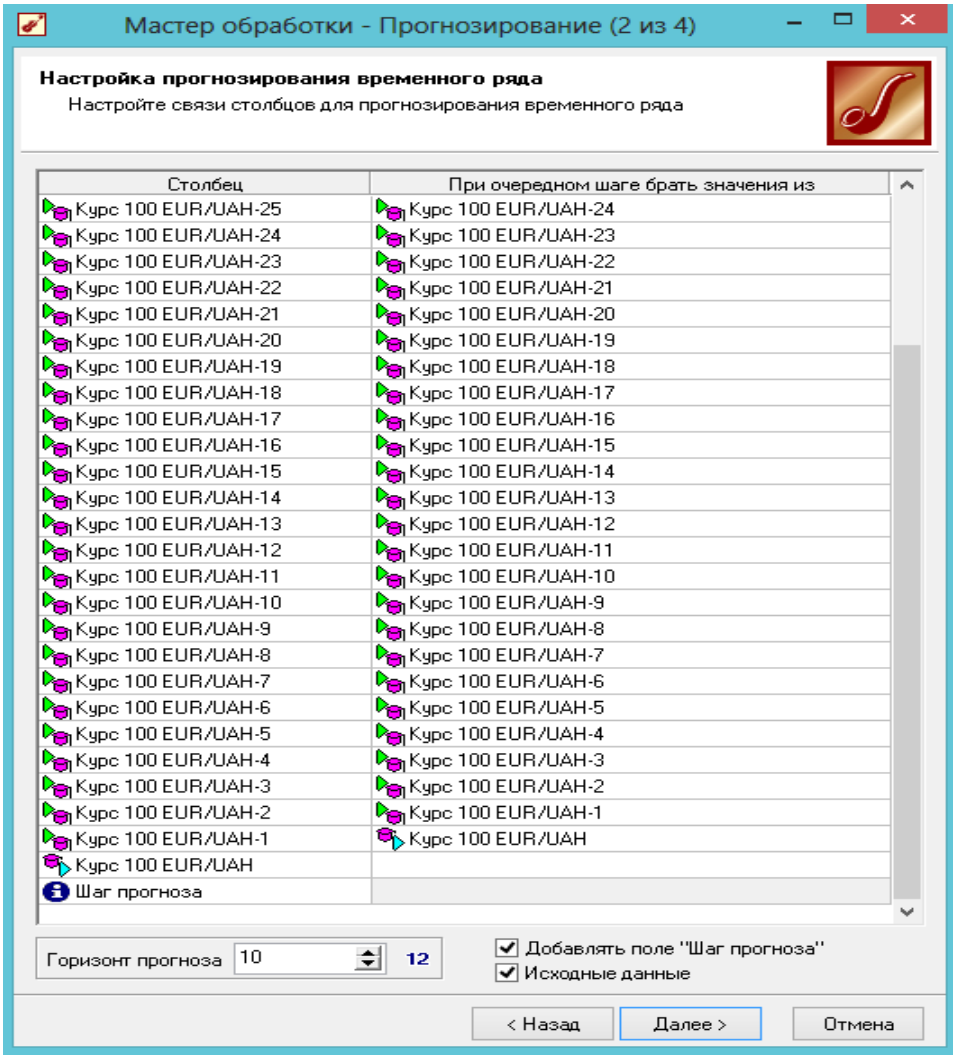


Рис.16. Налаштування вікна «Прогнозирование”.

Результати прогнозування надаються у графічному (Рис.17) та табличному (Таблиця 1) вигляді. Табличні дані можна зберегти у текстовому файлі за допомогою майстру експорту . На графіку на Рис.17 прогноз на 10 днів показаний темною лінією у кінці. Для експортованих у текстовий файл export.txt даних в Excel побудовані графіки на Рис.22. З Таблиці 1 слідує, що максимальна відносна похибка прогнозування побудованої моделі штучної нейронної мережі не перевищує 1%, що є сприятливим у багатьох практичних задач.

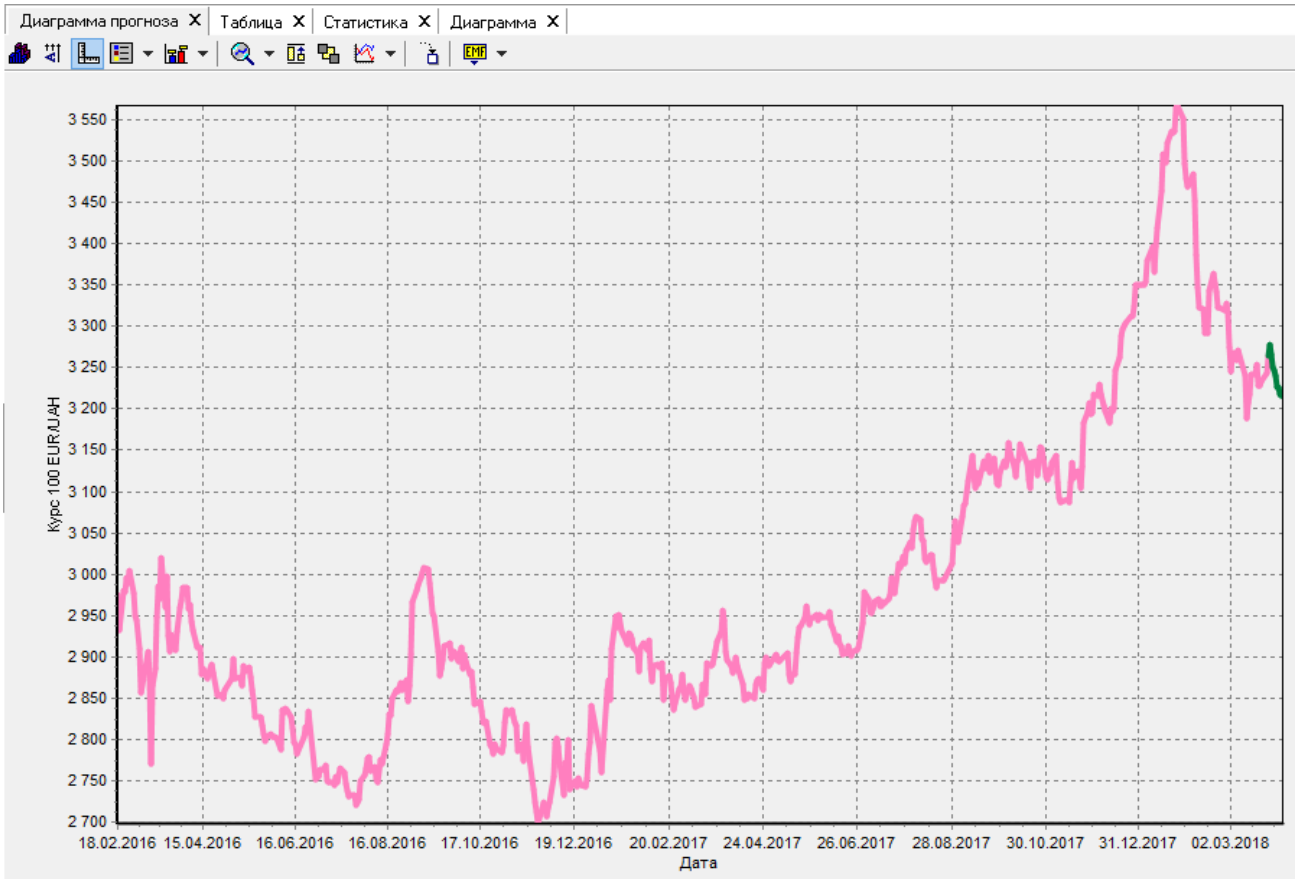


Рис. 17. Результуюча діаграма з прогностичними даними.

2. Прогнозування часового ряду за допомогою лінійної регресії

Для розв’язання задачі прогнозування часового ряду розглянемо ще одну модель – лінійну регресію як функцію багатьох змінних [2]. З самої назви моделі випливає її основна структурна властивість. На відміну від нелінійної моделі штучної нейронної мережі у лінійній регресії вихідний параметр лінійно залежить від кожного вхідного параметра. Для обчислення значень вільної константи та коефіцієнтів при вхідних параметрах регресії використовується метод найменьших квадратів, що використовує набір ретроспективних даних. “Навчання” лінійної регресії здійснюється на основі мінімізації квадратів відхилень обчислених значень регресії (на вхідних ретроспективних даних) від фактичних ретроспективних даних вихідного параметру.

Розглянемо задачу прогнозування часового ряду Курс 100 EUR/UAH з попереднього параграфу. Використовуємо так само набір із 30 вхідних параметрів Курс 100 EUR/UAH - 30, Курс 100 EUR/UAH – 29, …, Курс 100 EUR/UAH-1, таблиця з якими вже підготовлена на етапі використання ковзаючого вікна, та одним вихідним параметром Курс 100 EUR/UAH, тобто схему [30x1].

Встановлюємо курсор на вузол з результатами ковзаю чого вікна та після запуску майстра обробки обираємо “Линейная регрессия”. Далі налаштовуємо вхідні параметри так же само як і для нейронної мережі (Рис.6), розбиваємо вибірку з ретроспективними даними на навчаючу та тестову (Рис.7) та запускаємо навчання лінійної регресії (Рис.18). В результаті навчання розпізнано 100% навчаючої та тестової вибірок. Максимальна похибка навчання 0,012 (1,2%). Застосовуємо налаштування засобів візуалізації результатів так саме як на Рис.12. На Рис.19 показано діаграму розсіювання лінійної регресії з коридором помилки ±0,01 (1%), на Рис.20 - графіки початкового часового ряду (із Рис.1) та вихідних значень лінійної регресії.

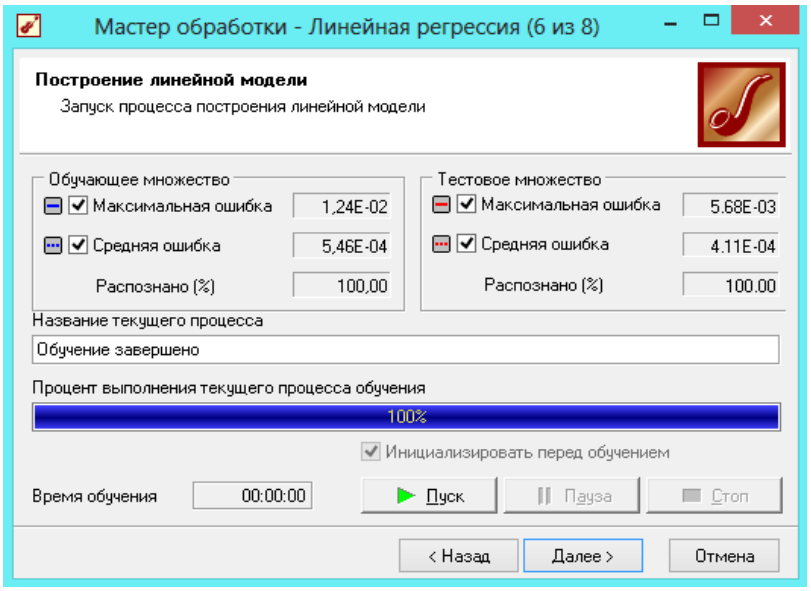


Рис. 18. Навчання лінійної регресії.

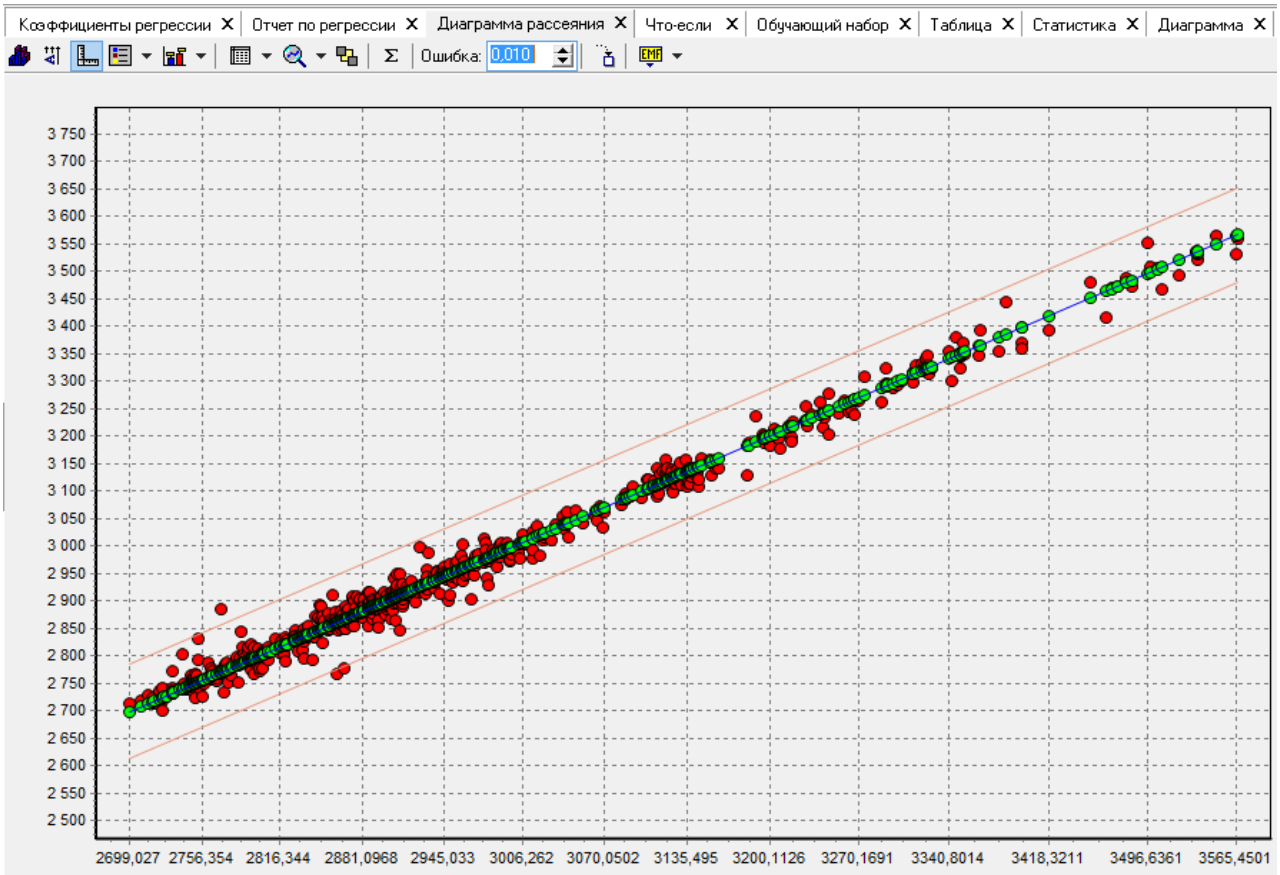


Рис.19. Діаграма розсіювання лінійної регресії з коридором помилки ±0,01 (1%).

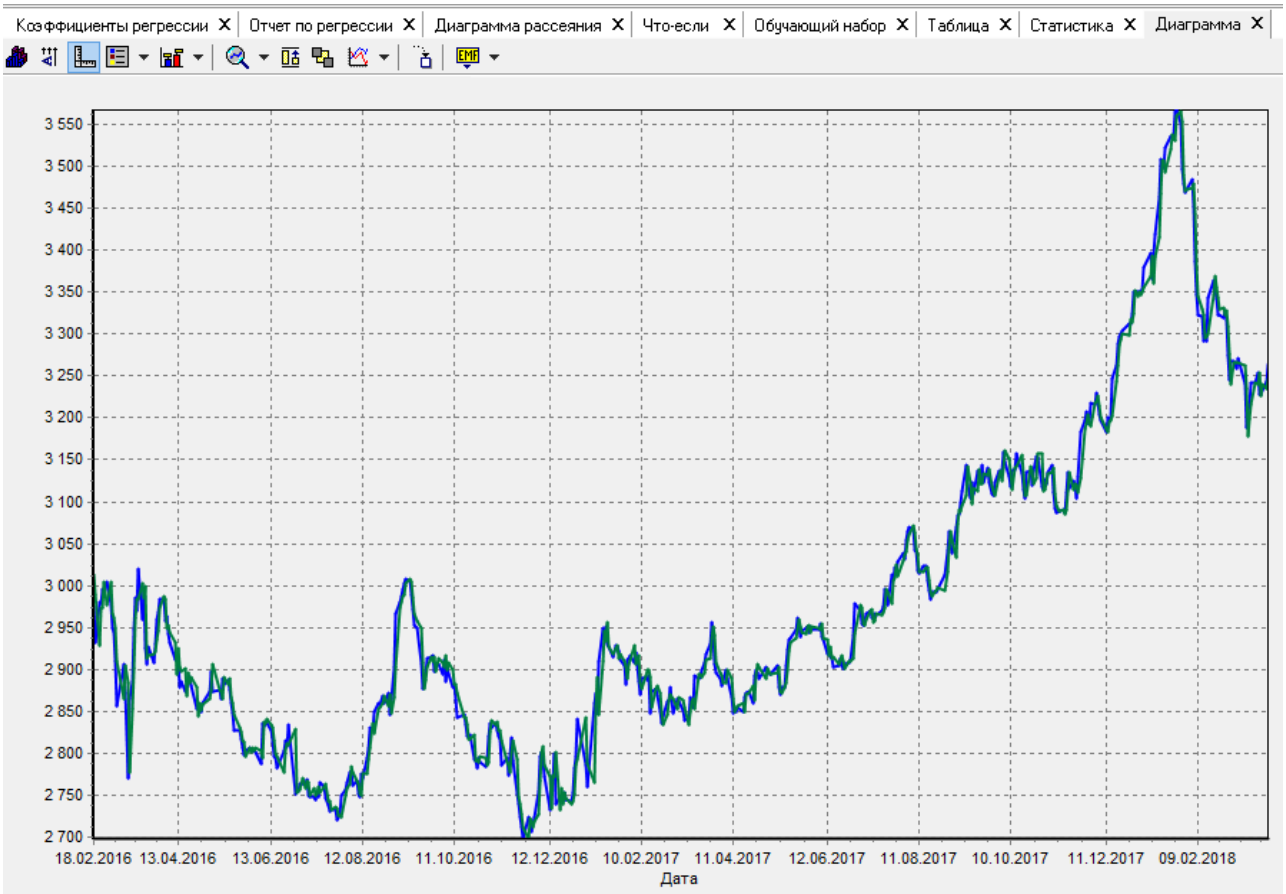


Рис.20. Графіки початкового часового ряду та вихідних значень лінійної регресії.

Із отриманих результатів слідує, що процес навчання закінчився успішно і побудовану модель лінійної регресії можна використовувати для прогнозування.

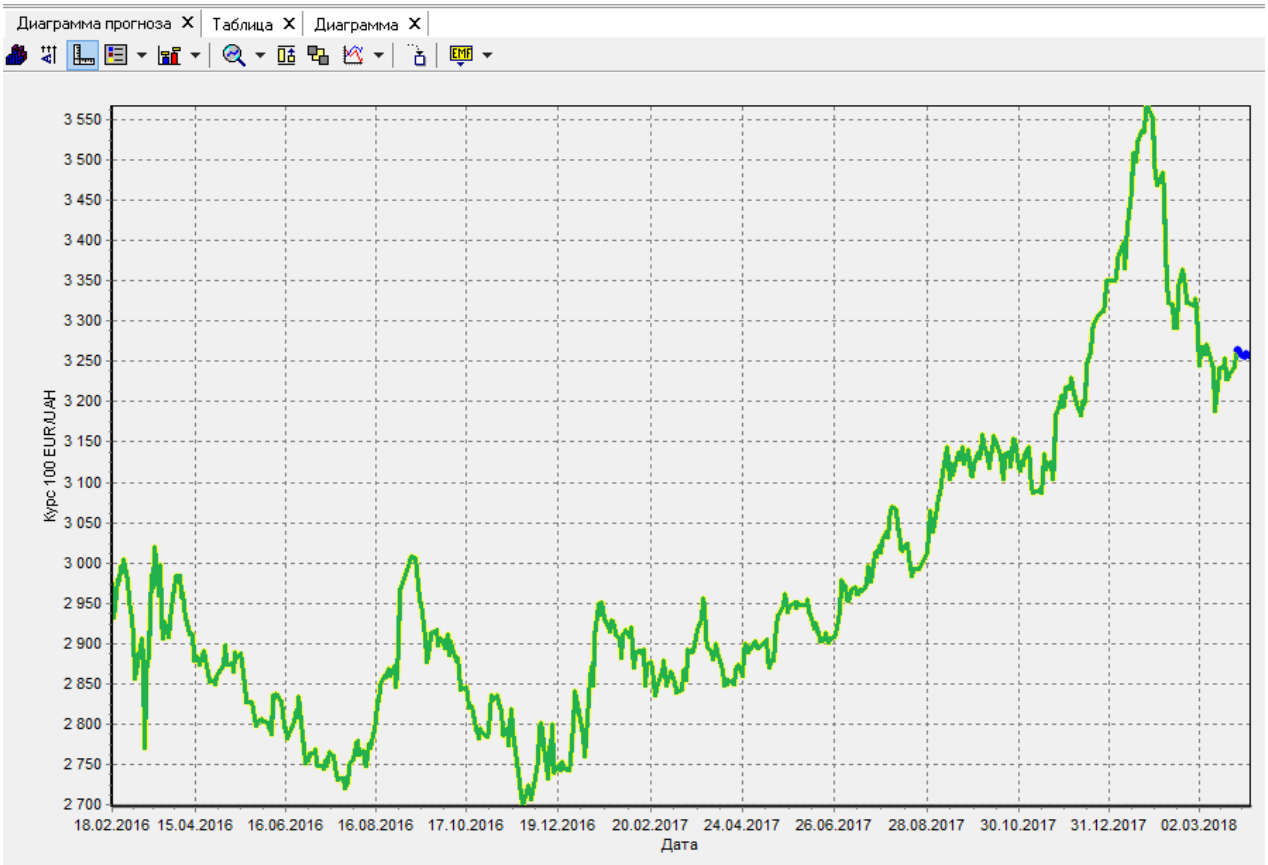


Рис.21. Результуюча діаграма з прогностичними даними.

Встановлюємо далі курсор на вузлі, створеному лінійною регресією та після завантаження майстра обробки, обираємо “Прогнозирование”. Налаштування стовбців даних для прогнозування здійснюється автоматично так саме як на Рис.16. Встановлюємо у нижньому вікні “Горизонт прогноза” 10 (днів). Далі у наступному вікні обираємо засоби візуалізації Таблиця та Діаграма та відмічаємо вихідний стовпчик для графіку. Результат прогнозу показані на Рис.21 у кінці графіку товстою лінією. Після експорту таблиці у текстовий файл, створюємо в Excel таблицю з даними (Таблиця 1) та відповідними графіками (Рис.22).

Прогноз лінійної регресії має більшу максимальну відносну похибку ніж прогноз нейромережі – до 1,9% (Таблиця 1). Крім того, із графіків прогнозу (Рис.22) слідує що прогноз нейронної мережі більш точно показує тенденції що до поведінки часового ряду у майбутньому. Тобто, нелінійна модель штучної нейронної мережі надає більш точний прогноз тенденцій часового ряду з мінімальною відносною похибкою прогностичних даних.

Таблиця 1. Порівняння фактичних та прогностичних даних за 10 днів.

Дата	Факт	Нейро-мережа	Похибка (%)	Лінійна регресія	Похибка (%)
28.03.2018	3255,7	3278,67	0,7	3263,897	0,3
29.03.2018	3273,7	3265,96	0,2	3261,29	0,4
30.03.2018	3270,4	3251,09	0,6	3258,774	0,4
02.04.2018	3241,3	3247,76	0,2	3255,426	0,4
03.04.2018	3223,3	3240,2	0,5	3255,502	1,0
04.04.2018	3240,6	3226,15	0,4	3249,132	0,3
05.04.2018	3224,1	3226,8	0,1	3251,446	0,8
06.04.2018	3195,8	3217,87	0,7	3246,201	1,6
10.04.2018	3184,1	3215,95	1,0	3243,907	1,9
11.04.2018	3210,8	3218,82	0,2	3241,838	1,0

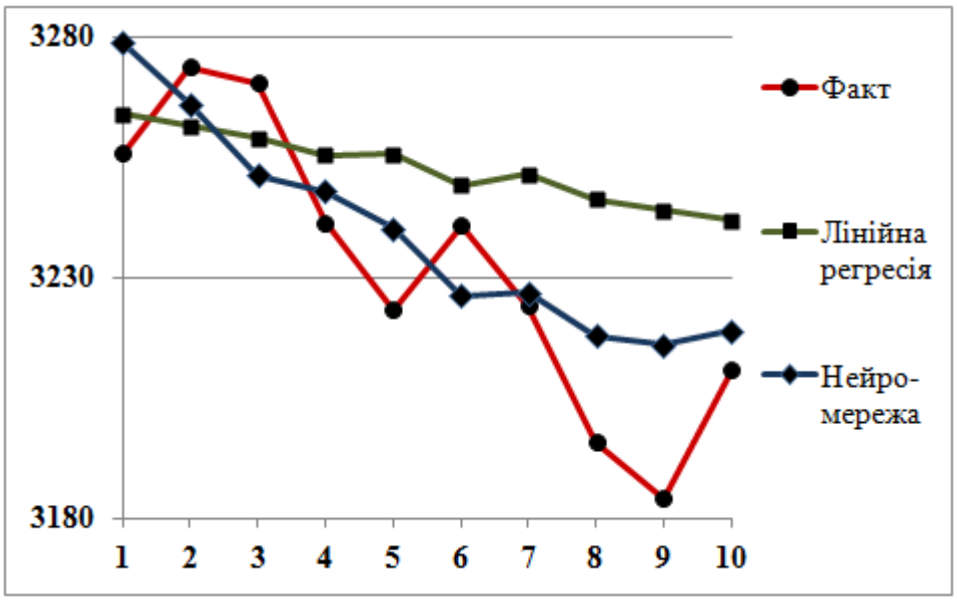


Рис.22. Порівняння прогнозу із фактичними даними для двох моделей.

Таким чином середовище Deductor Studio дозволяє будувати модель нейронної мережі довільної структури для ефективного розв'язання задач прогнозування часових рядів для об'єктів з декількома вхідними та вихідними параметрами. Дана модель показує кращій результат на тестових прикладах у порівнянні з моделлю лінійної регресії багатьох змінних. Для побудови моделі нейронної мережі немає необхідності в строгій математичній специфікації моделі, що є особливо цінним фактором при аналізі слабко структурованих та слабко формалізованих процесів та об'єктів. Це означає, що завдяки розвиненому та зручному програмному середовищу розробки аналітик при побудові моделі нейронної мережі орієнтується більш на визначення структури мережі, методу та параметрів її навчання, ніж на побудову математичної моделі.

Завдання для лабораторної роботи 5.

1. Створити файл Lab5.xlsx із трендом офіційного курсу НБ України пари 100 USD/UAH (доларів США/Гривні) за період 06.01.2010 – 27.03.2012. Зберегти файл як Lab5.txt. Для аналізу якості прогнозу на 10 днів зберегти дані за період 28.03.2012-11.04.2012 у файлі Labtest5.txt.
2. Імпортувати файл Lab5.txt, завантажити майстер обробки та обрати метод автокореляції.
3. Проаналізувати сезонність та наявність коливань тренда. Виявити глибину занурення.
4. Методом ковзаю чого вікна побудувати серію трендів на визначену глибину занурення.
5. Підібрати методом перебору кращу структуру нейронної мережі що відповідає мінімальній похибці навчання нейро-мережі на діаграмі розсіювання.
6. Зробити прогноз тренду нейро-мережею на 10 днів.
7. Навчити та зробити прогноз тренду лінійною регресією на 10 днів. Порівняти результати обох прогнозів.

Тема 6. Задачі підтримки прийняття рішень на основі штучної нейронної мережі, лінійної регресії декількох змінних та дерева рішень

Задачі: побудувати моделі штучної нейронної мережі, лінійної регресії декількох змінних та дерева рішень на основі навчаючих та тестових вибірок; отримати за допомогою нейромережі, лінійної регресії та дерева рішень розв'язки прикладних завдань з підтримки прийняття рішень на тестових даних.

Методи: методи підтримки прийняття рішень на основі моделей штучної нейронної мережі, лінійної регресії декількох змінних та дерев рішень.

1. Підтримка прийняття рішень за допомогою нейронних мереж

Розглянемо приклад використання моделі штучної нейронної мережі у задачі прийняття рішень з іпотечного кредитування 1-кімнатної квартири фізичних осіб. Після імпорту даних з файлу task6.txt (Рис.1) відкриваємо майстер обробки та обираємо "Нейросеть". Далі налаштовуємо тип даних стовпців (Рис.2), проводимо їх нормалізацію (Рис.3) та розбиваємо дані на навчаючу та тестову вибірки (Рис.4). Для навчання нейромережі обираємо метод "back propagation" [8], [9], [16] із параметрами, що вказані за замовчуванням.

Таблица

1 / 50

№	Вік>30	Прибуток (т.грн.)	Будинок здано	Міська зона	Метро поруч	Рішення
1	так	14	так	1	так	видати
2	ні	8	так	1	так	видати
3	ні	7	ні	1	ні	відмовити
4	так	9	ні	2	ні	відмовити
5	так	11	ні	1	так	видати
6	так	6	ні	1	ні	відмовити
7	ні	8	ні	2	ні	відмовити
8	так	6	так	1	так	видати
9	так	14	так	1	так	видати
10	ні	9	ні	2	ні	відмовити
11	так	7	так	1	так	видати
12	ні	7	так	2	ні	відмовити
13	так	25	так	1	так	видати
14	так	9	ні	2	ні	відмовити
15	так	15	ні	1	так	видати
16	так	8	ні	1	ні	відмовити
17	так	20	ні	2	так	видати
18	ні	8	ні	2	ні	відмовити
19	так	9	так	1	так	видати
20	ні	7	так	2	ні	відмовити
21	так	25	так	1	так	видати
22	ні	18	так	1	так	видати
23	так	8	ні	2	ні	відмовити
24	так	9	ні	1	ні	відмовити
25	ні	11	ні	1	ні	відмовити

Рис.1. Фрагмент таблиці імпортованих даних.

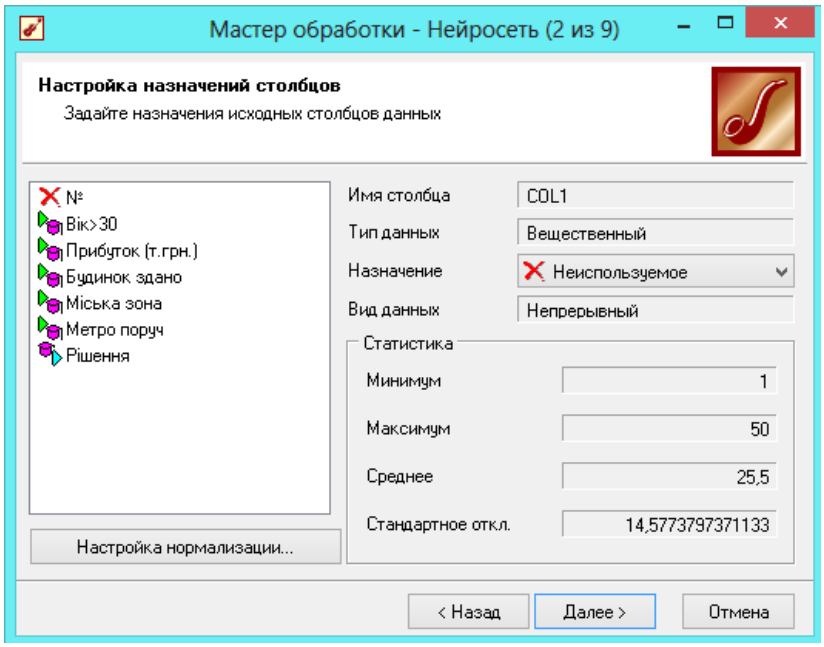


Рис.2. Налаштування вхідних та вихідного параметрів нейромережі.

При налаштуванні вхідних параметрів різного типу (дійсний, строковий, та ін.) можна виконати процедуру “Настройка нормализации данных” (знизу у вікні на Рис.2). Нормалізатор (Рис.3) у автоматичному режимі аналізує значення, що приймають вхідні змінні. Наприклад, данні стовпчику “Вік>30” містять тільки значення “так” чи “ні”. Відповідно їм, в моделі нейронної мережі будуть використовуватися цілочисельні значення -1 та 1. Так саме буде виконана автоматична нормалізація строкових даних інших стовпчиків “Будинок здано” та “Метро поруч” які теж приймають тільки два значення “так” чи “ні” (Рис.3). Користувач може змінити параметри нормалізації згідно влісних міркувань у відповідному вікні. Ми залишаємо налаштування нормалізації без змін (за замовчанням). На Рис.4 – 7 показано налаштування параметрів нейронної мережі та результат навчання у вигляді графіка відносної похибки розпізнавання нейромережею даних навчаючої та тестової вибірок. Розпізнавання складає 100%.

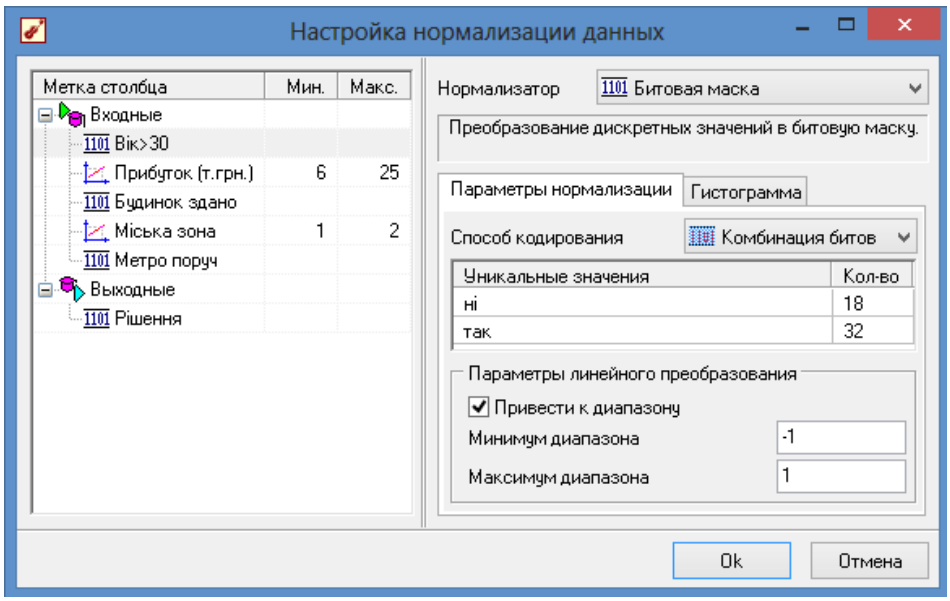


Рис.3. Вікно нормалізації даних.

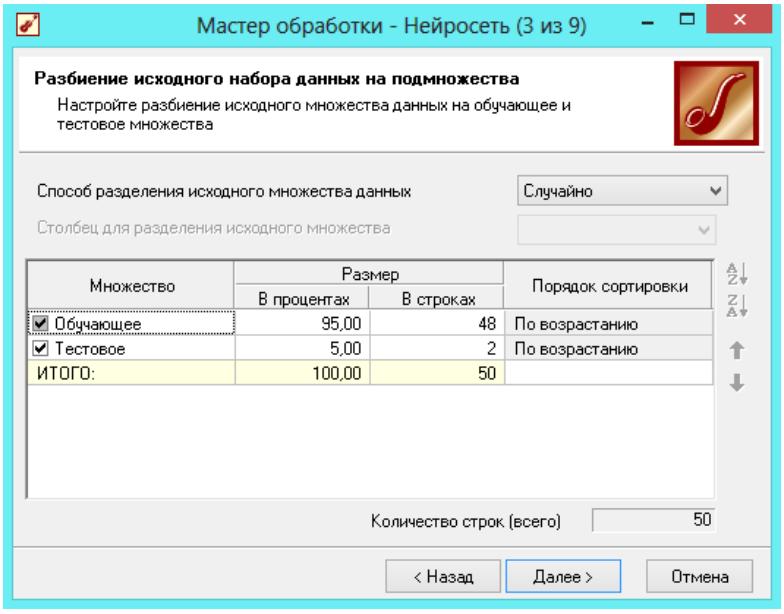


Рис.4. Розбиття даних на навчаючу та тестову вибірки.

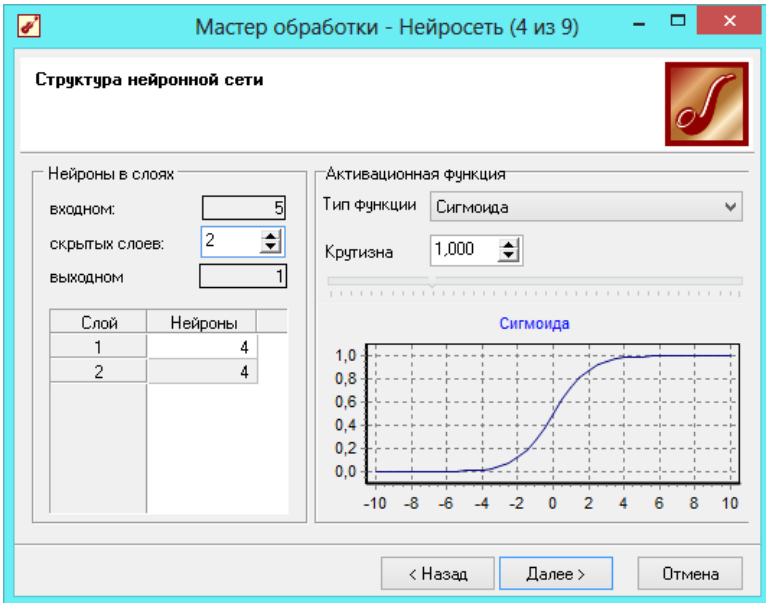


Рис.5. Налаштування нейромережі з 2-ма прихованими шарами по 4 нейрони.

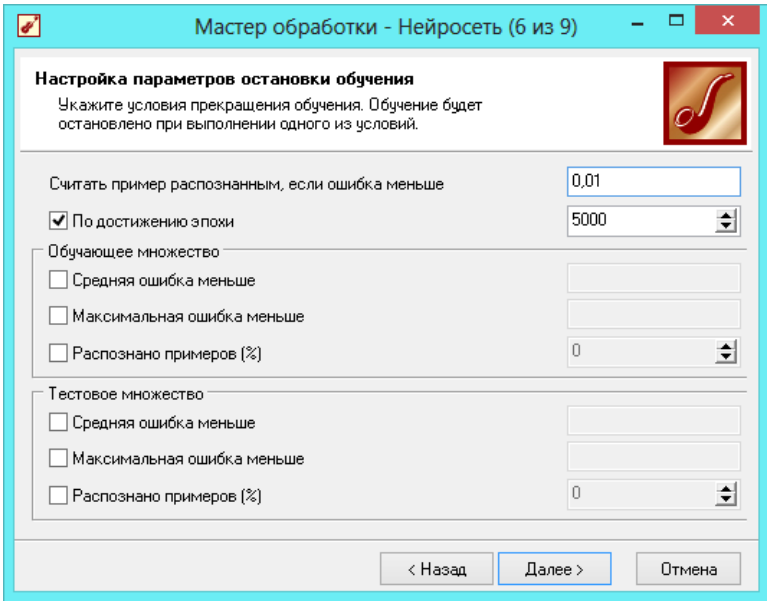


Рис.6. Параметри методу “back propagation” з похибкою 0,01 (1%) та 5000 ітерацій

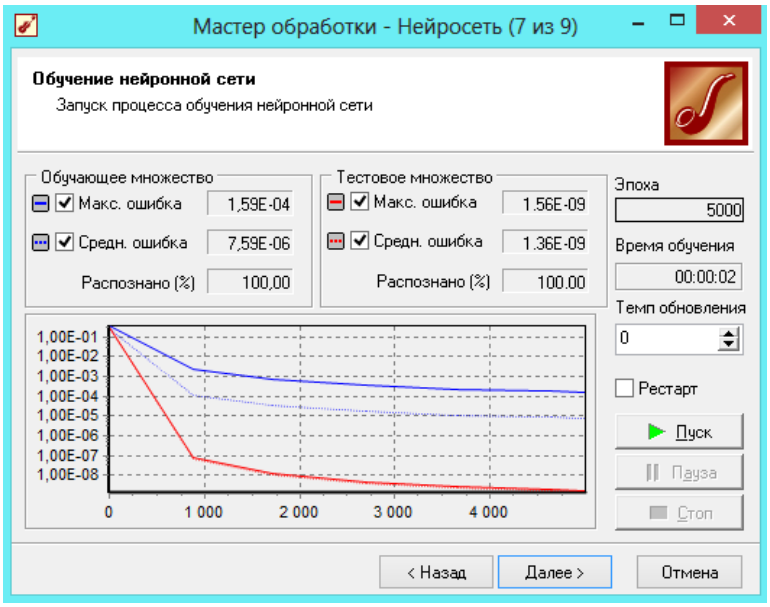


Рис.7. Результати навчання нейромережі із 100% розпізнаванням даних навчаючої та тестової вибірок та відносною похибкою не більш 0,0002 (0,02%).

На Рис.8 показано структуру побудованої нейронної мережі з одним вхідним (5 нейронів), двома прихованими (по 4 нейрони на кожному) та одним вихідним (1 нейрон) шарами.

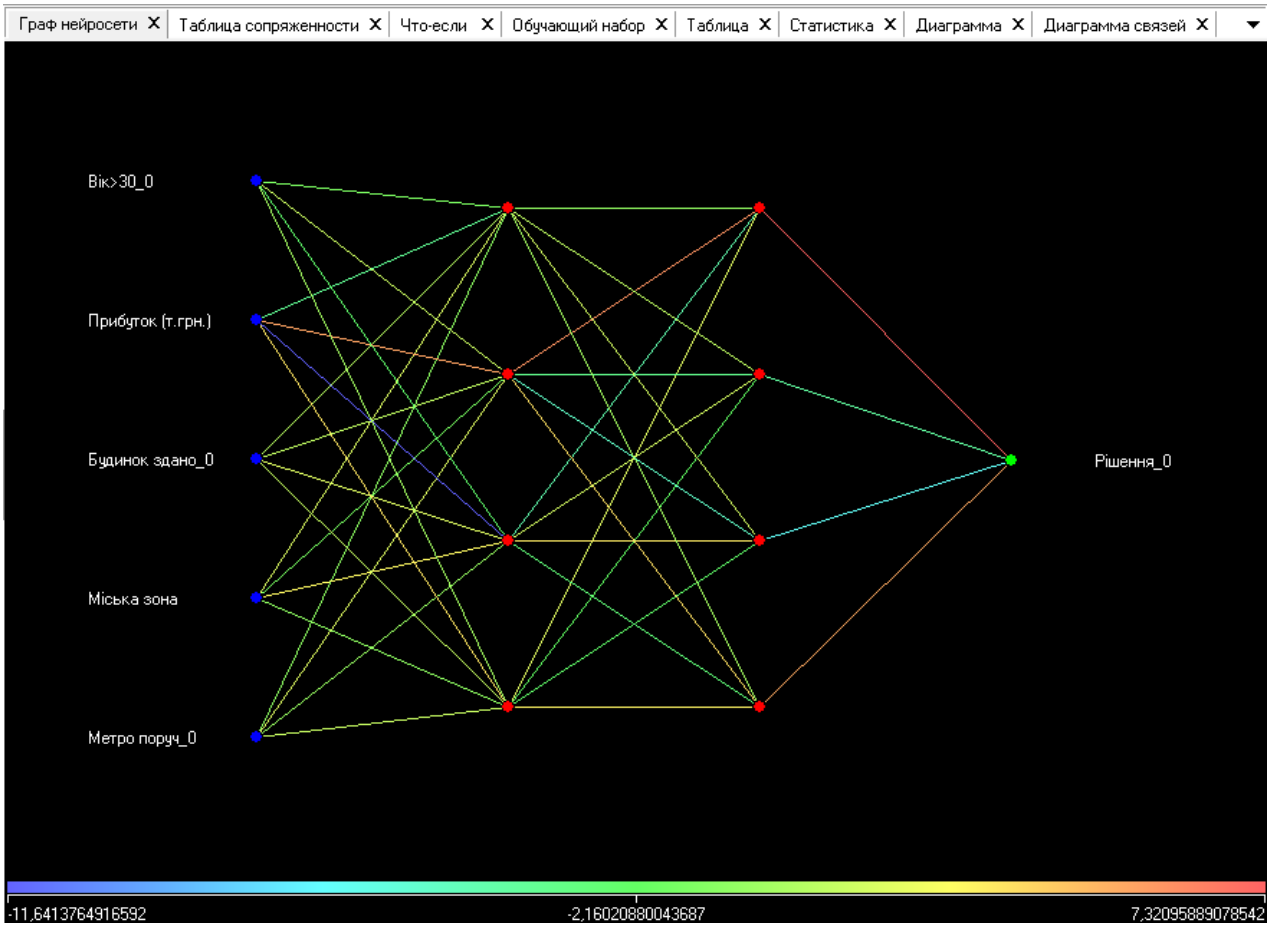


Рис.8. Структура побудованої нейромережі з 1 вхідним (5 нейронів), 2 прихованими (по 4 нейрони на кожному) та 1 вихідним (1 нейрон) шарами.

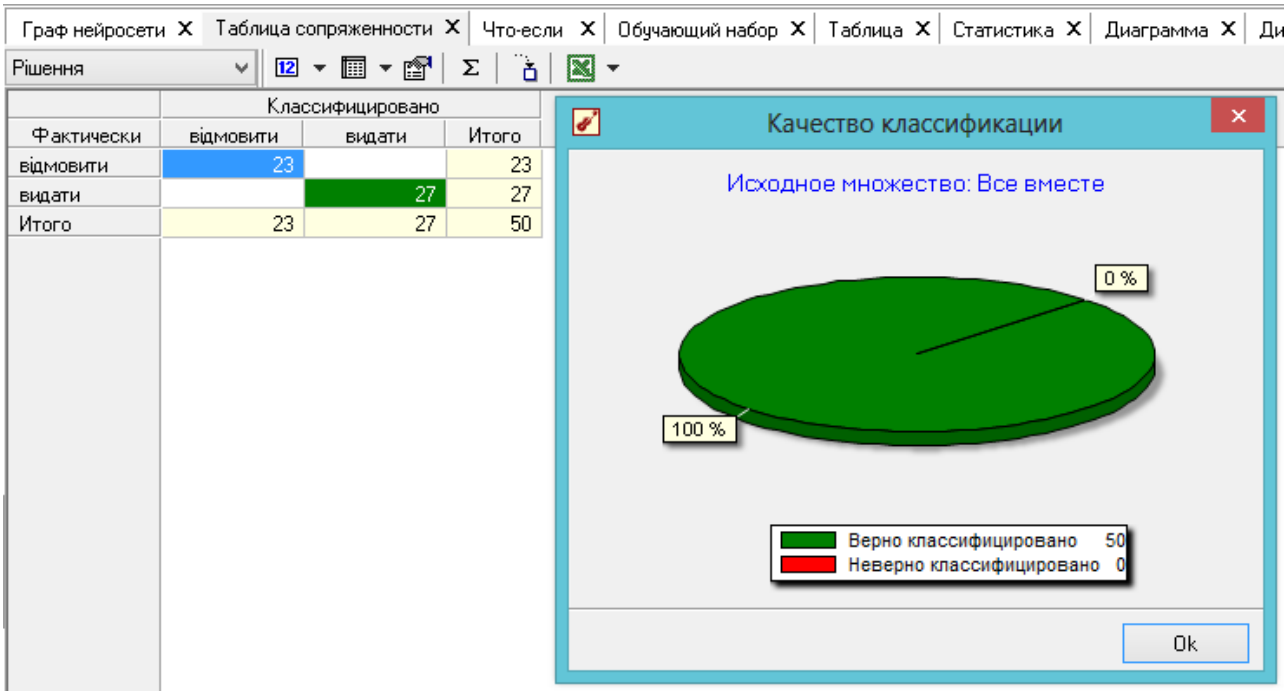
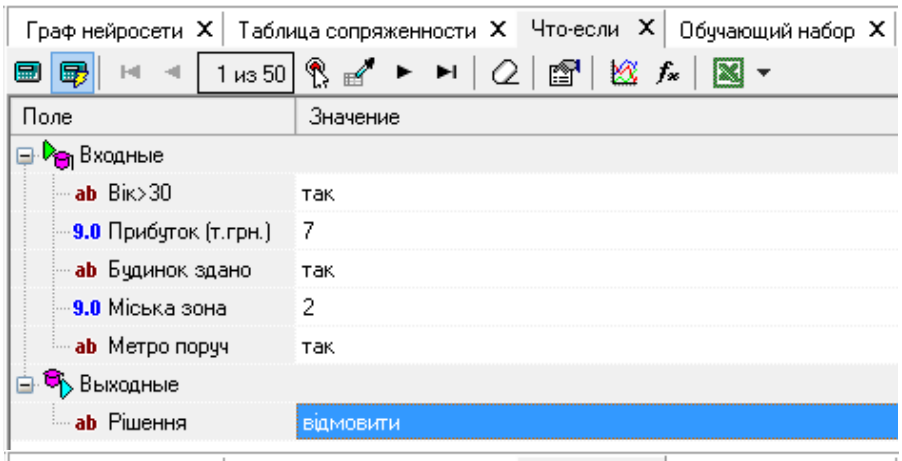


Рис.9. Перевірка якості побудованої моделі на основі таблиці спряженості.

Ефективність вибору структури побудованої моделі нейромережі підтверджена результатами навчання із 100% розпізнаванням даних навчаючої та тестової вибірок, відносною похибкою не більш 0,0002 (0,02%) (Рис.7) та 100% (50 записів) вірною класифікацією даних (таблиця спряженості на Рис.9). Підтримка прийняття рішень на основі даної моделі нейронної мережі здійснюється за допомогою функції “Что-если” (Рис.10а, 10б).

а)



б)

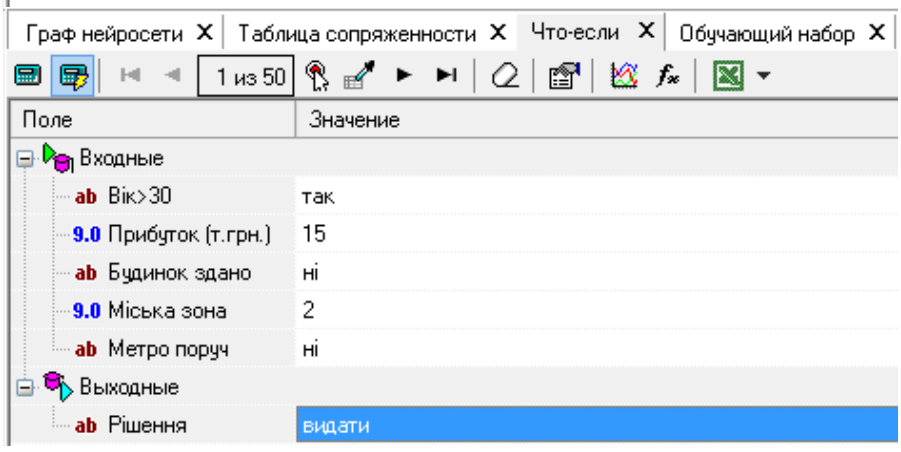


Рис.10. Приклади надання рішень про видачу кредиту нейромережею.

В полях вхідних параметрів необхідно ввести фактичні данні відповідного формату та форми про фізичну особу та об’єкт нерухомості та отримати рекомендоване нейромережею рішення з видачі кредиту. Приклади таких розв’язків показано на Рис.10а, 10б. Наведені приклади демонструють коректність роботи моделі нейронної мережі.

2. Прийняття рішень за допомогою лінійної регресії декількох змінних

Лінійна регресія так саме як і штучна нейронна мережа відноситься до класу кібернетичних моделей, коли не відомий точно характер залежності між вхідними та вихідними параметрами об’єкта [2], [16]. Найбільш розповсюджена форма регресії - лінійна. Так же саме як і нейронна мережа лінійна регресія багатьох змінних може використовуватися для підтримки прийняття рішень в складних системах.

Розглянемо задачу з п.1 про іпотечне кредитування 1-кімнатної квартири фізичних осіб. Встановлюємо курсор на вузлі імпортованих даних з файлу task6.txt, відкриваємо майстер обробки та обираємо “Линейная регрессия”. Далі налаштовуємо тип даних стовпців так саме як на Рис.2, проводимо їх нормалізацію (Рис.3) та розбиваємо дані на навчаючу та тестову вибірки (Рис.4). Після запуску отримуємо результат навчання лінійної регресії на Рис.11. Результати розпізнавання даних у лінійній регресії (83,3%) гірші ніж у нейромережі (100%, Рис.7). Покращити результат навчання лінійної регресії неможливо, оскільки її структура жорстко фіксована.

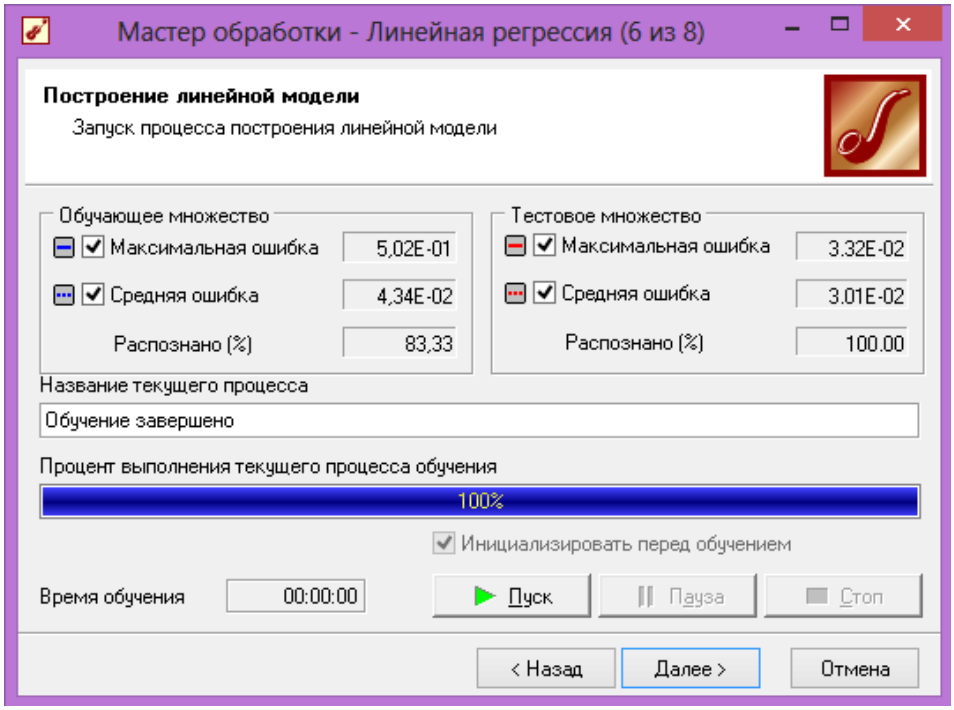


Рис.11. Запуск процесу навчання лінійної регресії.

Налаштування засобів відображення результатів показано на Рис.12. Результати навчання побудованої моделі лінійної регресії, що показані у

таблиці спряженості на Рис.13, гірші у порівнянні із результатами навчання нейромережі (Рис.31) – розпізнано 47 правил із 50 (94%). Такий результат, безумовно, впливає на якість рішень, що пропонує модель. На Рис.14а, 14б представлені результати генерації рішень моделі на тих самих тестових завданнях що на Рис.10. Рішення лінійної регресії у даному випадку повністю протилежні рішенням нейронної мережі. При тому, що рішення нейронної мережі цілком відповідають істинності тестового завдання.

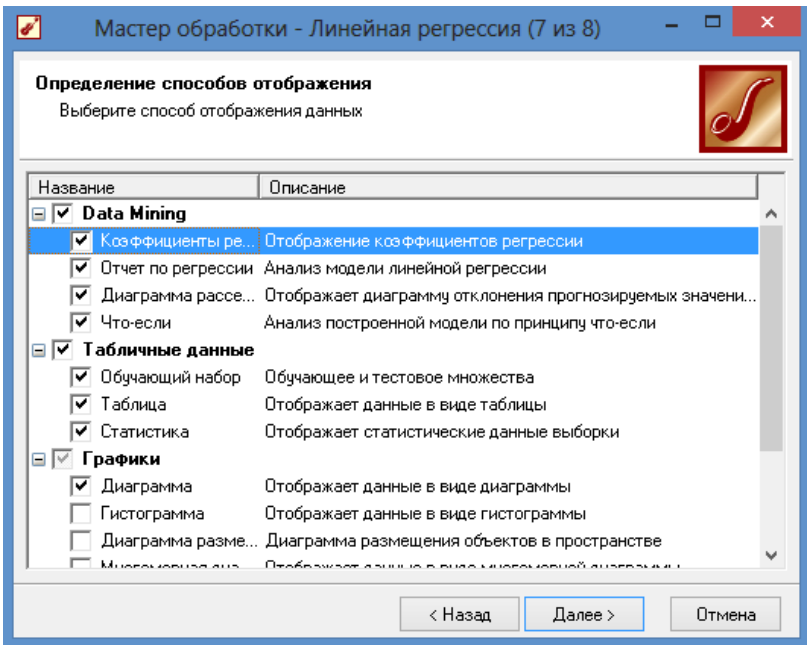


Рис.12. Налаштування засобів відображення лінійної регресії.

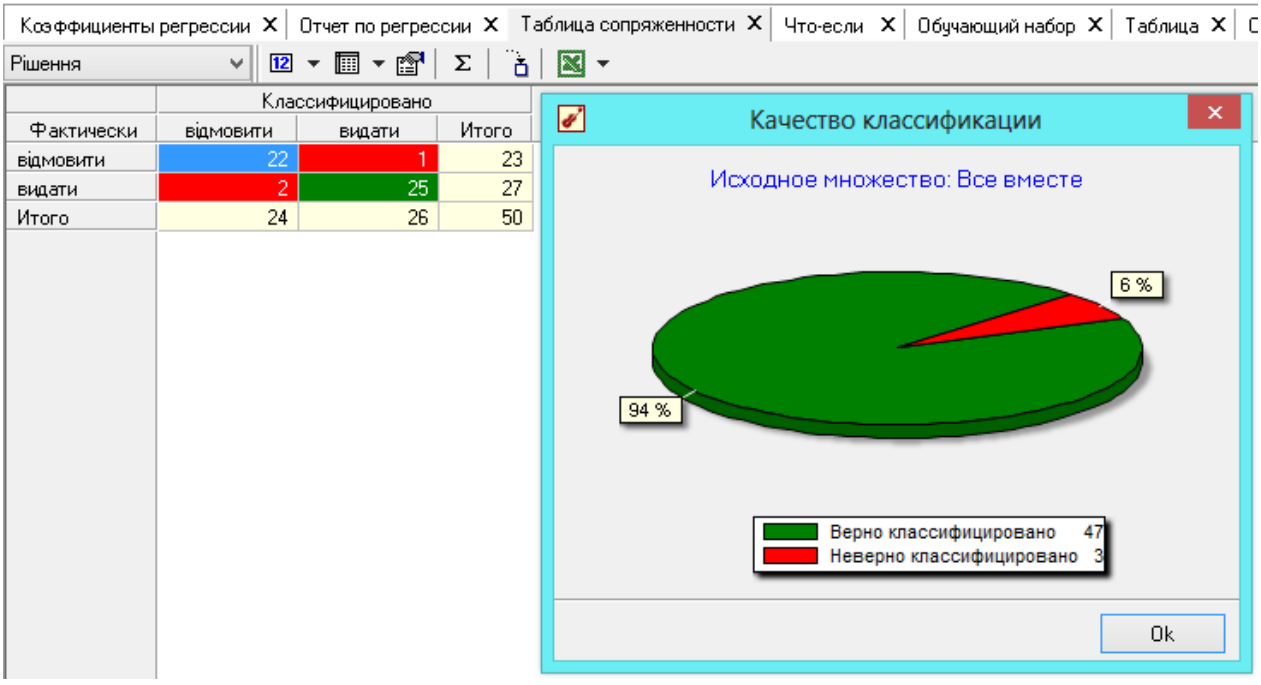
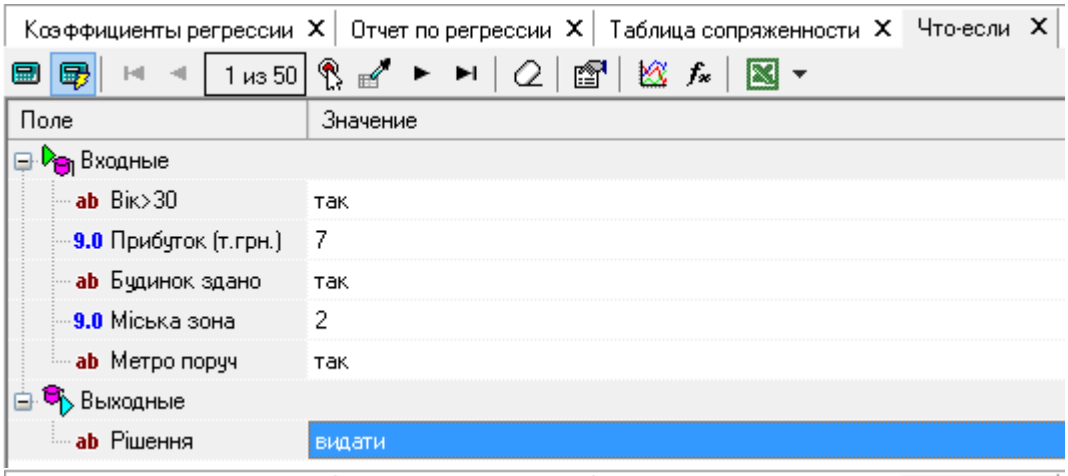


Рис.13. Перевірка якості навчання моделі на основі таблиці спряженості.

а)



б)

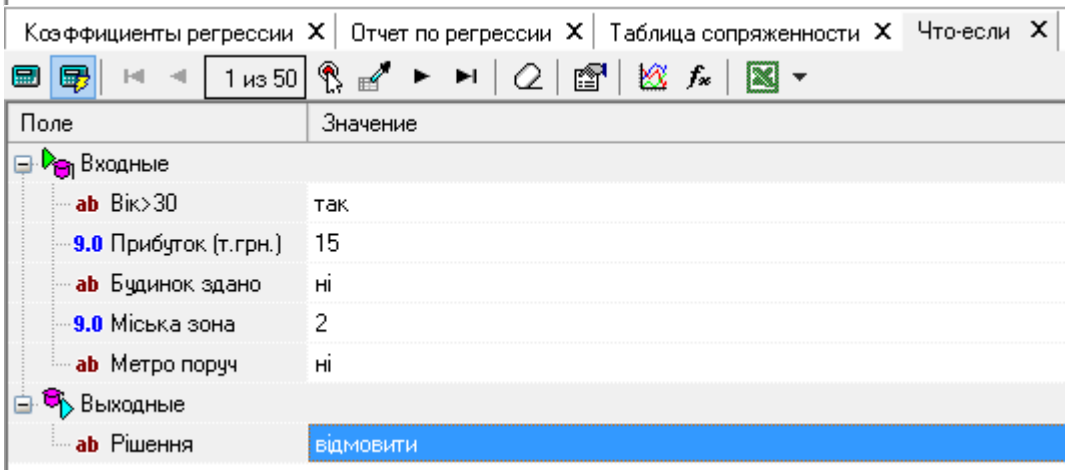


Рис.14. Приклади надання рішень про видачу кредиту лінійною регресією.

3. Підтримка прийняття рішень за допомогою дерев рішень

Дерева рішень представляють собою продукційні моделі представлення знань [10] та разом із нейронними мережами та лінійними регресіями можуть застосовуватися у системах підтримки прийняття рішень. Даний тип моделі представлення знань використовується в задачах діагностики, класифікації, управлінні, тощо. Дерево рішень являє собою узгоджену множину правил продукцій, побудованих від фактів до мети. В даному випадку ми використовуємо прямий метод виведення висновків продукційної системи (від “фактів до мети”). Алгоритм побудови продукційної системи у Deductor Studio ґрунтується на основі навчання правил продукцій на ретроспективних даних вибірки вхідних параметрів, яка розбивається на навчаючу та тестову вибірки (так саме як і у попередніх моделях підтримки прийняття рішень). Після створення дерева рішень можна дізнатися ранг значущості кожного фактора – вхідного параметра що знаходиться в умові ядра кожної продукції (найбільш значущі фактори знаходяться на верхніх рівнях дерева рішень).

Розглянемо задачу з п.1 про іпотечне кредитування 1-кімнатної квартири фізичних осіб. Встановлюємо курсор на вузлі імпортованих даних з файлу task6.txt, завантажуємо майстер обробки та обираємо “Дерево решений”. Налаштування вхідних та вихідних параметрів таке саме як на Рис.2. Проводимо їх нормалізацію (Рис.2) та розбиваємо дані на навчаючу та тестову вибірки (Рис.3). На наступному кроці налаштовуємо параметри навчання дерева рішень -

мінімальну кількість прикладів, при якому буде створено новий вузол (встановлюємо 2), та параметр відсічу вузлів дерева - рівень довіри, що регулює компактність дерева: чим меньш його значення тим меньш розмір дерева, (встановлюємо рівень довіри 40%) (Рис.15). Після запуску метода отримуємо результат навчання дерева рішень (Рис.16). Розпізнано 95% навчаючої та 100% тестуючої вибірок.

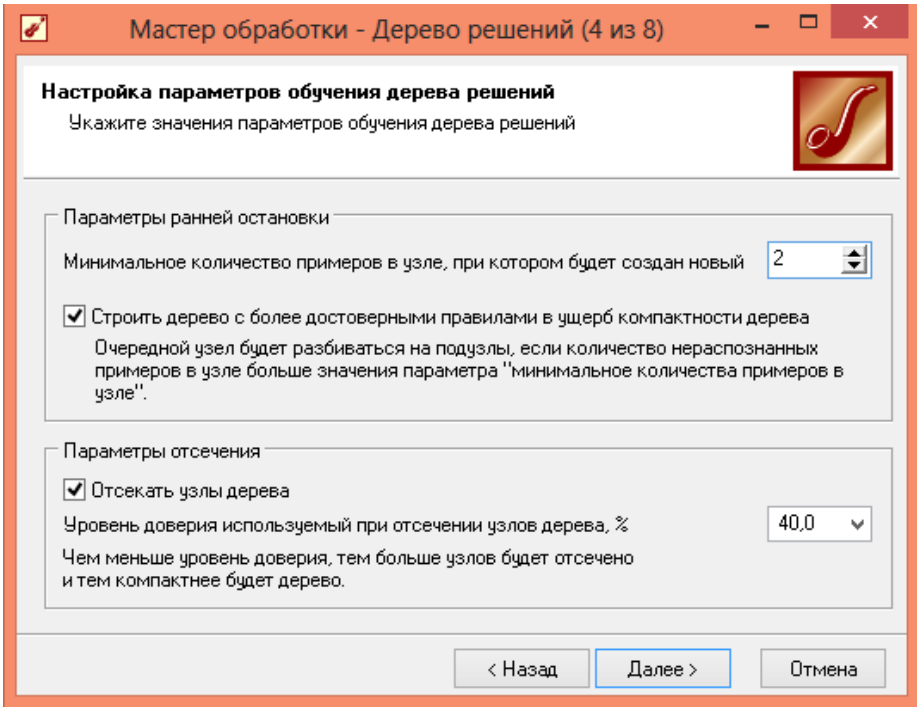


Рис.15. Налаштування параметрів навчання дерева рішень.

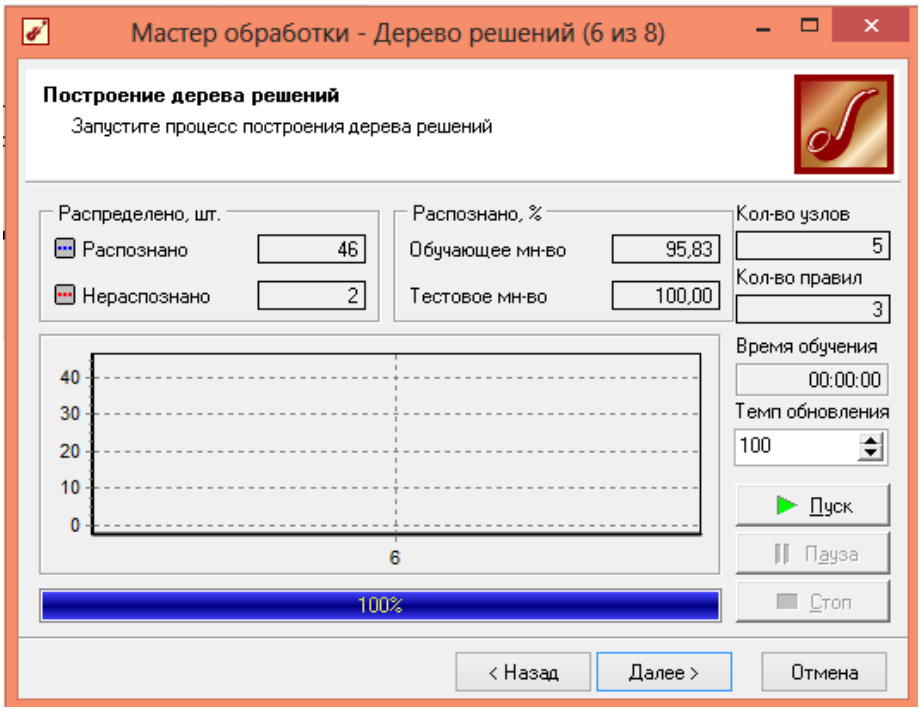


Рис.16. Запуск та результат навчання дерева рішень.

Після налаштування вікна засобів візуалізації (Рис.17) перш за все аналізуємо дерево рішень (Рис.18). Воно побудовано у формі дерева правил

продукцій вигляду “ЯКЩО “А” ТО “В” ІНАКШЕ “С””. Більш важливі фактори (вхідні параметри) розташовані на верхніх вузлахи дерева (“Метро поруч”), меньш вагомі - на нижчому рівні (“Прибуток (т.грн.)”>=13), а не впливові фактори взагалі не присутні в дереві рішень (“Вік>30”, “Будинок здано”, “Міська зона”). Дана візуалізація дає можливість побачити усі наявні правила на підтримку та достовірність, по кожному фактору окремо.

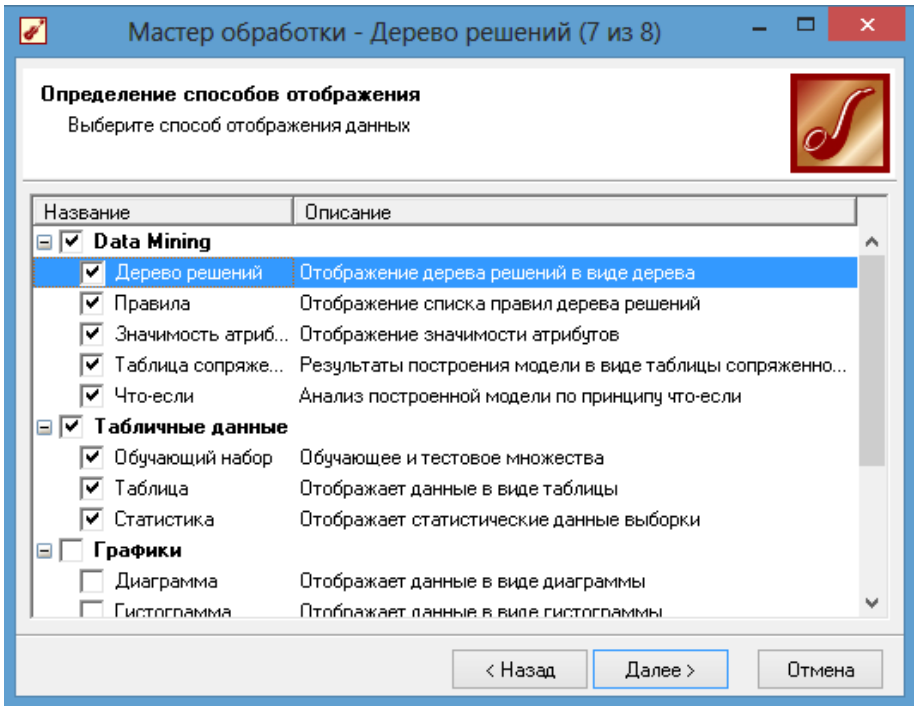


Рис.17. Налаштування засобів візуалізації дерева рішень.

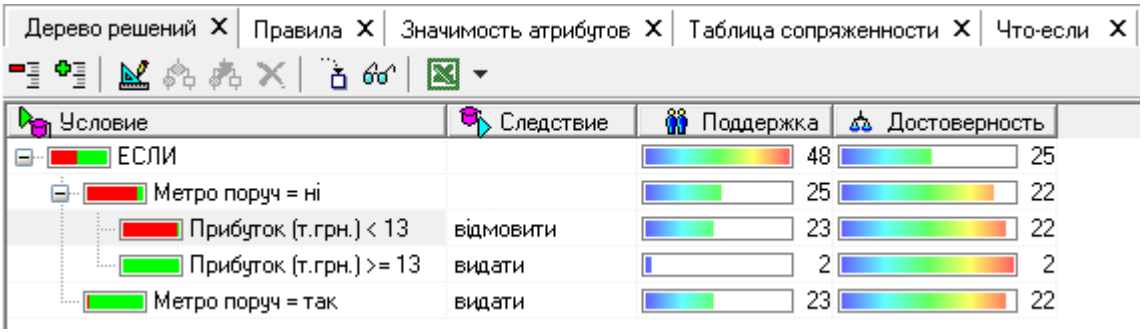


Рис.18. Структура побудованого дерева рішень.

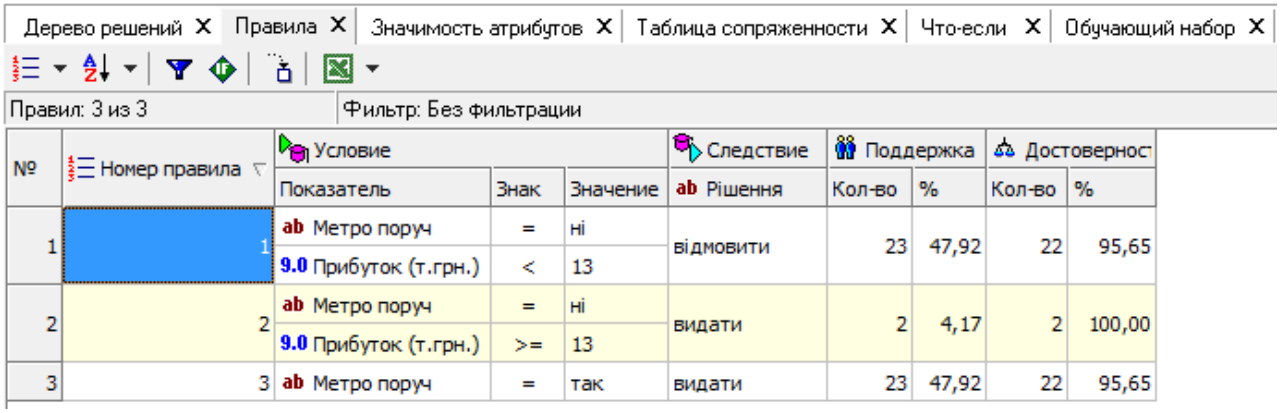


Рис.19. Набір правил дерева рішень.

Набір правил дерева рішень показано на Рис.19. Тут ми спостерігаємо аналогічний результат що зображений на Рис.18. Кількісну оцінку значущісті вхідних параметрів можна дізнатися у вікні “Значимость атрибутов” (Рис.20).

Кількість невірно розпізнаних прикладів із вказівкою які саме приклади були віднесені до якого класу помилково можна побачити у “Таблице сопряженности” (Рис.21). По діагоналі таблиці розташовані приклади, які були правильно розпізнані, в інших комірках - ті, які були віднесені до іншого класу. В даному випадку результат розпізнавання можна вважати задовільним, оскільки дерево правильно класифікувало 48 прикладів (96%) та відкинуло тільки 2 приклади (4%).

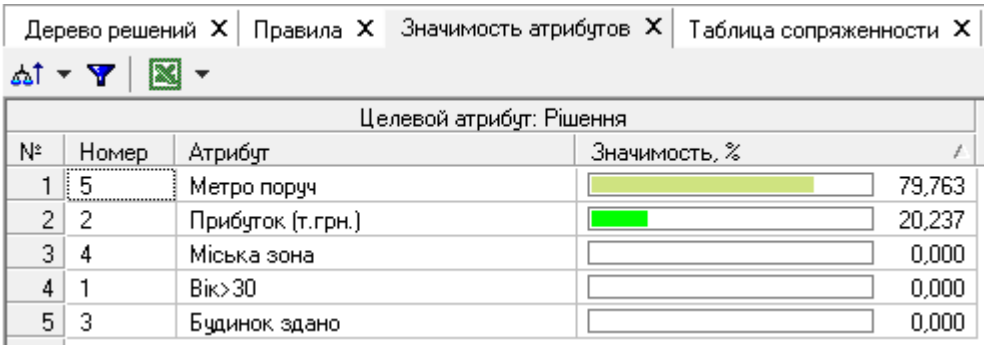


Рис.20. Значущість факторів побудованого дерева рішень.

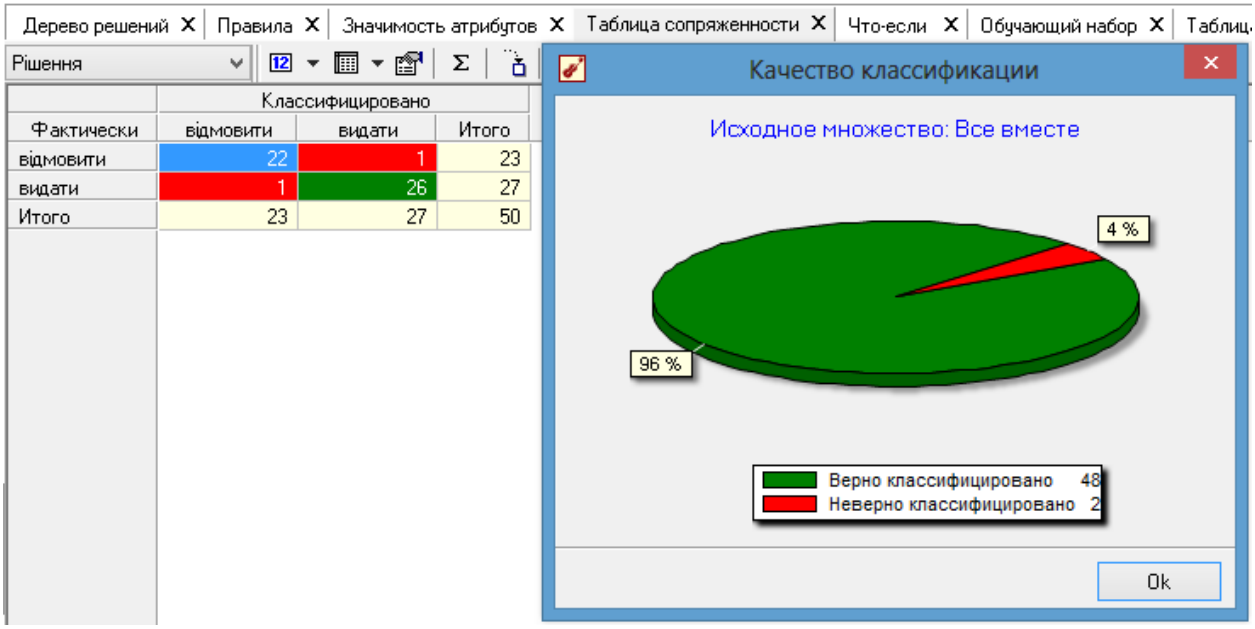
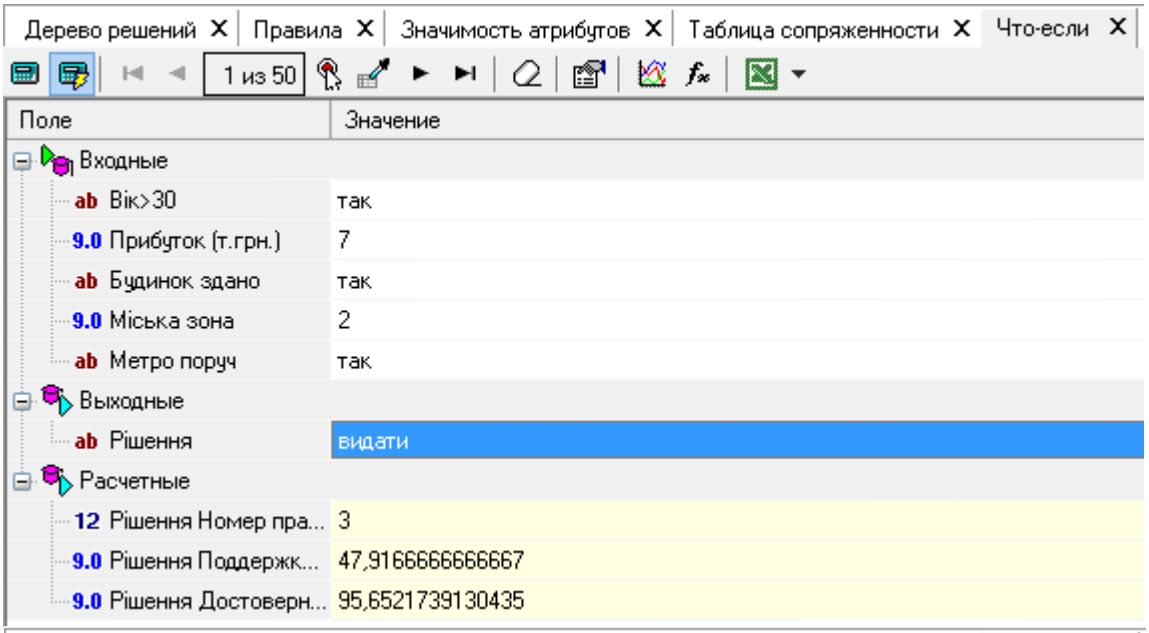


Рис.21. Таблиця спряженості правил дерева рішень.

Перевірка роботи системи на тестовому прикладі, що використовувався у попередніх моделях (Рис. 10а, 10б, 14а, 14б) показано за допомогою візуалізатора “Что-если” на Рис.22а, 22б. Результати роботи дерева рішень з підтримки прийняття рішень по видачі кредиту у першому випадку співпадають з результатом нейронної мережі (Рис.10а та 22а) та не співпадають з лінійною регресією (Рис.14а), у другому випадку співпадають з результатами лінійної регресії (Рис.14б та 22б) та не співпадають з нейромережею (Рис. 10б).

а)



б)

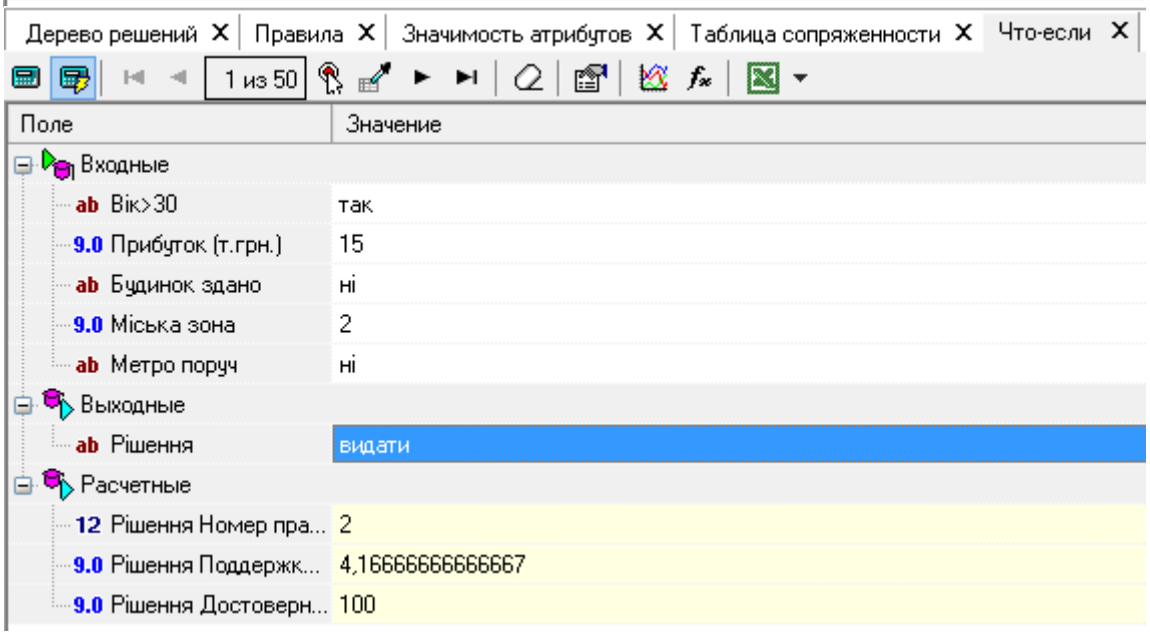


Рис.22. Висновки дерева рішень на тестових прикладах.

Розбіжність результатів наданих нейромережею та лінійною регресією з результатами дерева рішень може бути результатом як нерепрезентативності вибірки ретроспективних даних (недостатньо великий об’єм вибірки) так і не коректністю вибору вхідних факторів у моделі прийняття рішень з видачі кредиту. Дійсно, якщо дерево рішень у 96% використовує тільки 2 фактора та відкидає 3 фактори із 5 у побудові правил, то такий результат не може бути задовільним для аналітика. Тут для побудови більш коректної та реалістичної моделі підтримки прийняття рішень необхідно проаналізувати існування інших факторів що впливають на рішення з видачі кредиту та, по можливості, збільшити кількість ретроспективних даних.

Загалом, проведене у даному розділі дослідження властивостей різних моделей на тестових прикладах показало що для створення коректної та достатньо реалістичної системи підтримки прийняття рішень доцільно спочатку зробити аналіз якості вибірки ретроспективних даних та значущості кожного

фактору моделі прийняття рішень за допомогою дерева рішень і потім, при задовільних результатах такого аналізу, будувати та використовувати модель штучної нейронної мережі з оптимізованої структурою прихованих шарів (за результатами навчання). Модель дерева рішень надає більш детальний аналіз якості вибірки даних та набору факторів але, на відміну від нейронної мережі, не надає аналітику можливості обрати та оптимізувати структуру моделі для навчання та тестування при мінімальній похибці. Модель нейронної мережі виявляє собою найбільш гнучкий та точний теоретичний інструмент для навчання та тестування на ретроспективних даних та підтримки прийняття рішень користувача у складних системах.

Завдання для лабораторної роботи 6.

1. Створити файл Lab6.xlsx. Створити таблицю із кредитною історією клієнтів автосалону (у довільній формі, 5 стовпчиків 50 рядків). Зберегти файл як Lab6.txt.
2. Імпортувати файл Lab6.txt у Deductor studio.
3. У налаштуваннях майстру обробки призначати кожного разу особисті дані клієнтів автосалону як вхідні параметри моделі, рішення про видачу кредиту - як вихідний параметр.
4. Побудувати для імпортованих даних моделі систем підтримки прийняття рішень методами штучної нейронної мережі, лінійної регресії декількох змінних та дерева рішень.
5. Порівняти висновки (рекомендації) побудованих моделей за допомогою засобу візуалізації “Что-если” на однакових тестових даних.

Тема 7. Багатовимірний кластерний аналіз на основі методу К-середніх (G-середніх), самоорганізуючих карт Кохонена та ЕМ-кластеризації

Задачі: провести кластерний аналіз групи об'єктів у багатовимірному просторі.

Методи: методи багатовимірної кластеризації К-середніх (G-середніх), ЕМ-кластеризації та модель самоорганізуючих карт Кохонена,.

1. Багатовимірна кластеризація методами К-середніх та G-середніх

Основною задачею багатовимірного кластерного аналізу є розбиття множини об'єктів у багатовимірному просторі на групи (кластери) що містять схожі за певними критеріями об'єкти. При цьому елементи із різних груп (кластерів) мають відрізнятися між собою. Існує багато різних алгоритмів багатовимірної кластерізації, але найбільш популярним та поширеним серед усіх є метод К-середніх (k-means) що реалізований у Deductor studio [1], [11], [16]. В основі роботи алгоритму К-середніх покладено наступний ітераційний процес:

1. Обираємо число k та визначаємо k точок - центрів ваги груп (кластерів) у багатовимірному просторі атрибутів (факторів) елементів.
2. Кожен з елементів вибірки буде наближений до одного із центрів ваги. Вважаємо що не всі елементи наближаються до єдиного центру, тобто число кластерів більш одного.
3. На основі даних про елементи кожної групи обчислюємо значення центрів кластерів, які становляться новими центрами ваги. Центр кластера обчислюється як середнє значення векторів - об'єктів, що входять до даної групи (звідси назва методу).
4. Для нових значень центрів ваги утворюються нові групи елементів.
5. Якщо нова група елементів не відрізняється від попередньої із заданою похибкою, повторюємо п.3-5. Ітераційний процес може зупинитися також за рахунок досягнення граничної кількості ітерацій що задається користувачем.

Розглянемо роботу методу G-середніх (аналогу К-середніх) для задачі кластеризації регіонів України за 5 демографічними та економічними показниками (далі – фактори від 1 до 5) за економічно однорідний період з 2001 – 2004 р. (за даними Держкомстату України). Після імпорту даних із файлу task7.txt отримуємо таблицю із 108 рядками, фрагмент якої представлений на Рис.1. Запускаємо майстер обробки та обираємо метод "Кластеризация". Далі налаштовуємо стовпчики "Регіон" та "Рік" як інформаційні, решту стовпчиків як вхідні (Рис.2). Знизу ліворуч у вікні натискуємо кнопку налаштування нормалізації даних. Для кожного вхідного стовпчика встановлюємо лінійну нормалізацію та обираємо діапазон значень мінімум 0, максимум 1 (Рис.3). Рівень значущості усіх параметрів залишаємо однаковий 100%. Ця процедура необхідна для коректної роботи алгоритму кластеризації у одиничному п'ятивимірному кубі та пов'язана із переведенням усіх вхідних даних у безрозмірний вигляд із значеннями із діапазону [0,1]. Далі встановлюємо з клавіатури 100% навчаючої вибірки (Рис.4) та обираємо метод G-means з автоматичним вибором числа кластерів (Рис.5).

Таблица | Статистика | Диаграмма

1 / 108

Регіон	Рік	Число народжених живими на 1000 нас., обидві статі	Обсяг реалізованої промислової продукції (робіт, послуг) по регіонах грн. на одну особу	Обсяг реалізованих послуг у розра-хунку на одну особу, грн.	Кількість малих підпри-ємств на 10 000 населення	Обсяг доходів місцевих бюджетів на душу населення, грн.
АР Крим	2001	7,2	1,8	390	54	445,2
Вінницька	2001	8,1	2,1	83	30	155,25
Волинська	2001	10,9	1,6	159	35	148,42
Дніпропетровська	2001	7,1	11,4	282	43	343,07
Донецька	2001	6,1	12,3	157	51	318,83
Житомирська	2001	8,3	1,7	106	41	142,87
Закарпатська	2001	10,7	0,9	95	51	141,23
Запорізька	2001	7,1	10,6	235	45	335,26
Івано-Франківська	2001	9,5	2	138	46	157,95
Київська	2001	7,4	3,1	118	36	205,33
Кіровоградська	2001	7,6	1,4	119	44	172,51
Луганська	2001	6,1	8	103	36	211,53
Львівська	2001	8,7	2,1	265	53	203,32
Миколаївська	2001	7,8	4,1	159	57	248,32
Одеська	2001	8,2	2,4	353	43	303,88
Полтавська	2001	7	7,2	142	40	255,76
Рівненська	2001	11,2	2,4	132	34	148,34
Сумська	2001	6,8	3,3	130	41	216,62
Тернопільська	2001	8,8	1	102	30	117,84
Харківська	2001	6,7	3,8	260	52	273,63
Херсонська	2001	8,1	1,7	185	49	169,51
Хмельницька	2001	8,3	1,9	131	34	148,3
Черкаська	2001	7,1	2,5	125	36	186,48
Чернівецька	2001	9,7	0,9	162	39	136,02
Чернігівська	2001	6,7	2,8	109	28	176,58
Київ	2001	7,3	4,2	1030	130	1597,34
Севастополь	2001	7,2	1,7	319	55	323,64
АР Крим	2002	8	1,66	456	61	420,39
Вінницька	2002	8,6	2,25	114	37	212,09
Волинська	2002	11,1	1,79	179	39	197,15
Дніпропетровська	2002	7,7	11,94	269	48	445,48

Рис.1. Фрагмент таблиці із імпортованими даними.

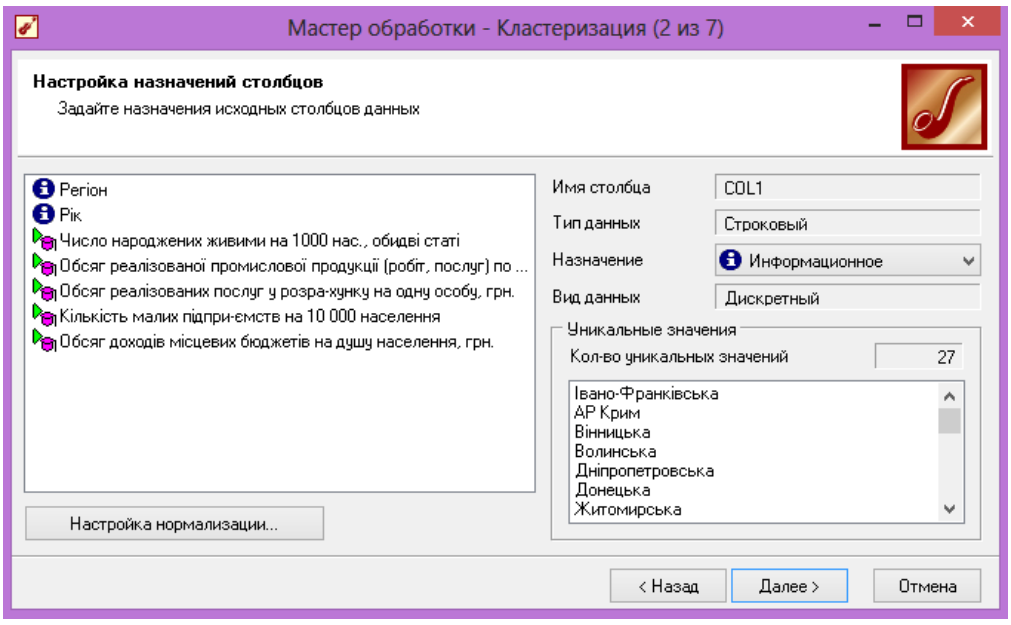


Рис.2. Налаштування стовпчиків вхідних даних.

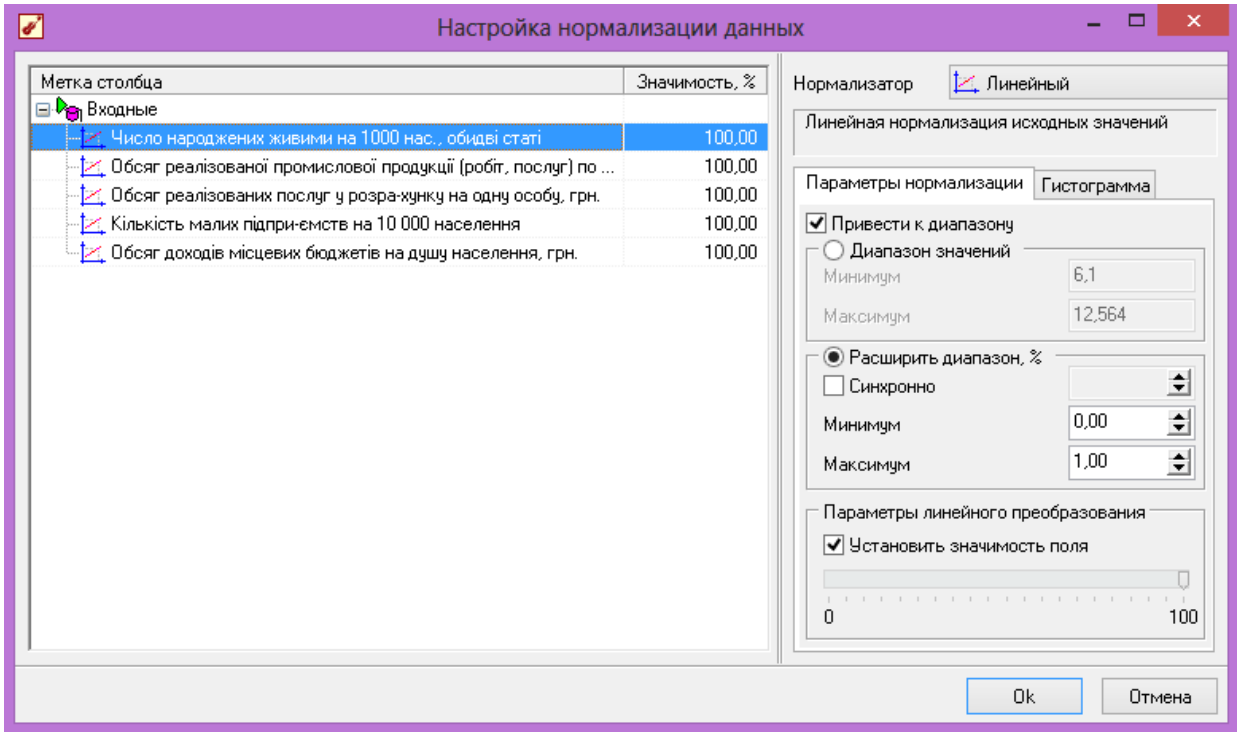


Рис.3. Нормалізація вхідних даних.

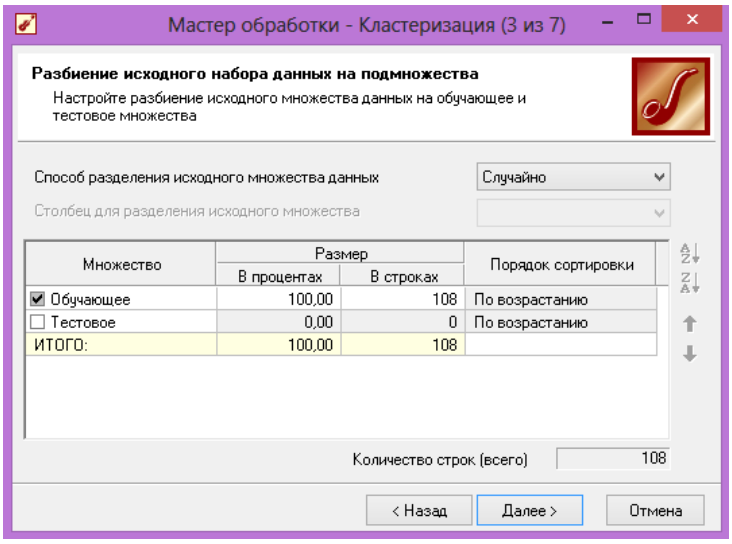


Рис.4. Налаштування навчаючої вибірки.

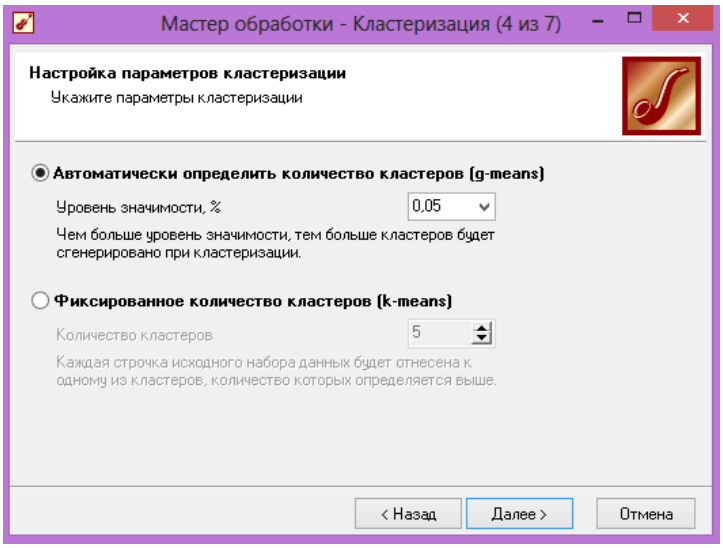


Рис.5. Метод G-means - автоматичний вибір числа кластерів (рівень значущості 0,05).

Далі запускаємо метод кластеризації (Рис.6) та налаштовуємо засоби візуалізації результатів (Рис.7). Слід підкреслити важливість вибору такого засобу візуалізації як Куб (OLAP-аналіз). На Рис.8, 9 показано налаштування полів та вимірів Кубу. Оскільки ми розв'язуємо задачу кластеризації регіонів країни по роках, встановлюємо рядки Регіони та Роки, стовпчики – Номер кластера (Рис.9). Результати кластеризації – кількість автоматично визначених кластерів (6, від 0 до 5) та зв'язки між ними показані на Рис.10. На Рис.11 показано таблицю парного порівняння кластерів з відсотками спільних елементів.

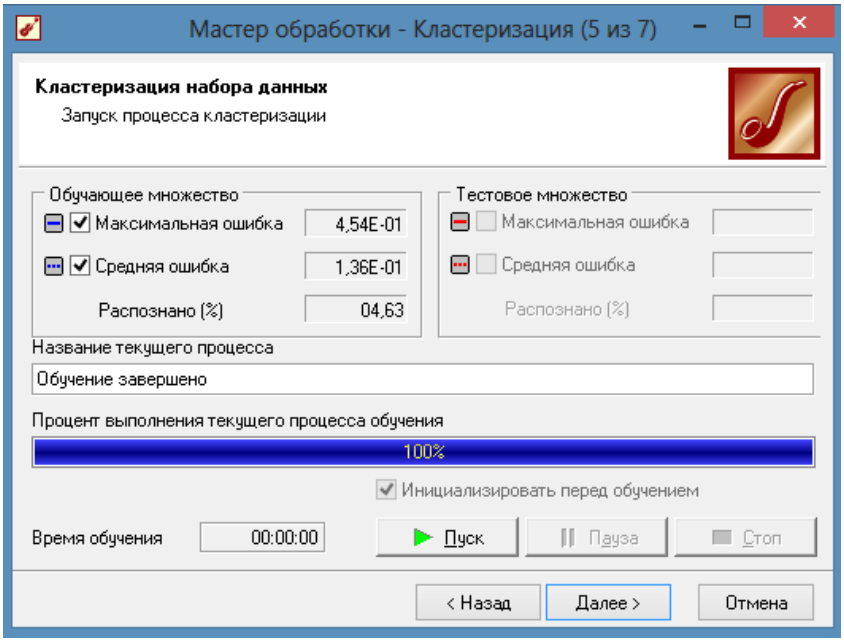


Рис.6. Запуск програми кластеризації.

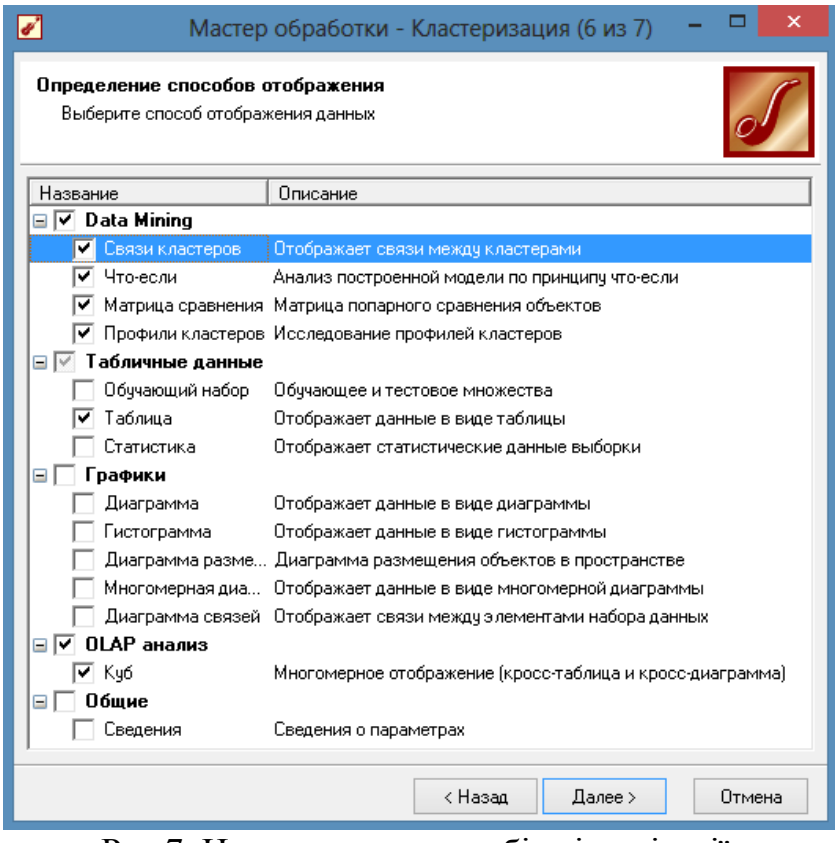


Рис.7. Налаштування засобів візуалізації.

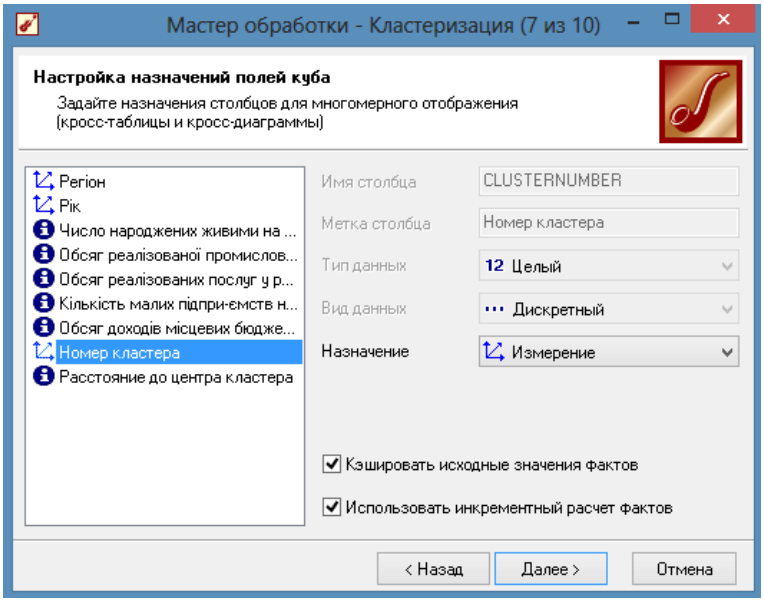


Рис.8. Налаштування полів кубу (OLAP - аналіз).

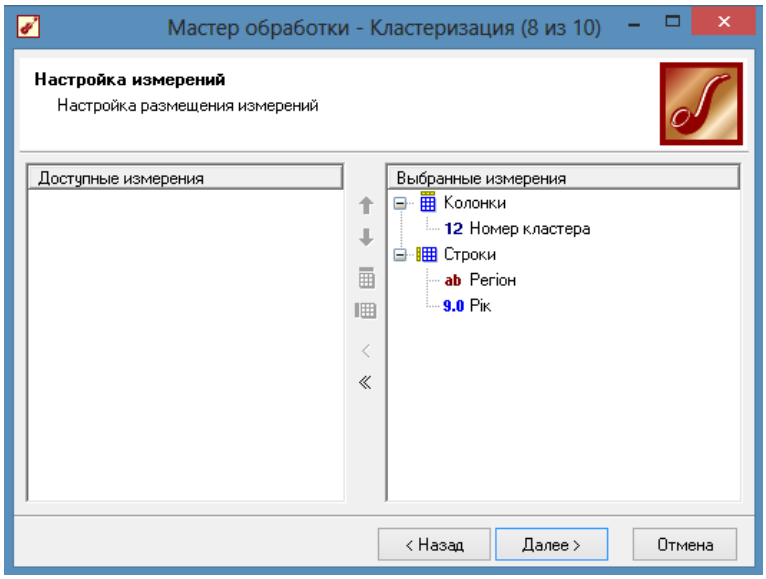


Рис.9. Налаштування вимірів кубу.

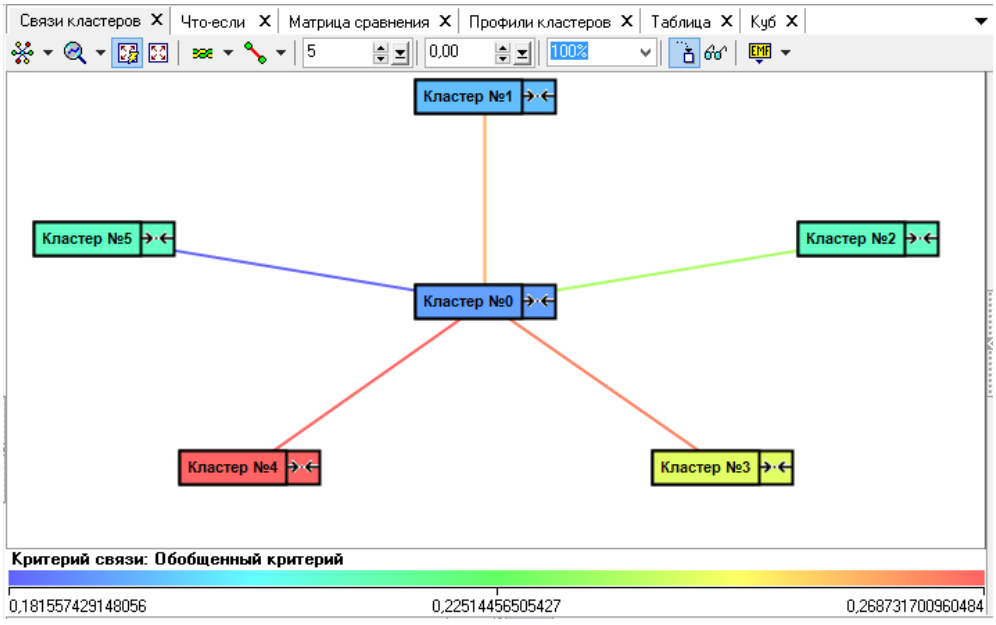
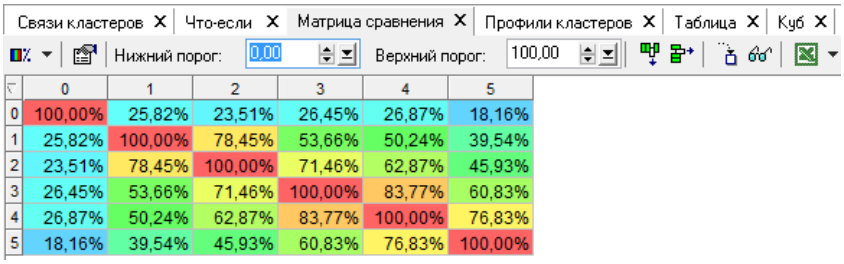


Рис.10. Зв’язки 6 побудованих автоматично кластерів (тип - зірка).



	0	1	2	3	4	5
0	100,00%	25,82%	23,51%	26,45%	26,87%	18,16%
1	25,82%	100,00%	78,45%	53,66%	50,24%	39,54%
2	23,51%	78,45%	100,00%	71,46%	62,87%	45,93%
3	26,45%	53,66%	71,46%	100,00%	83,77%	60,83%
4	26,87%	50,24%	62,87%	83,77%	100,00%	76,83%
5	18,16%	39,54%	45,93%	60,83%	76,83%	100,00%

Рис.11. Матриця порівняння спільних елементів у кластерах (по парам).

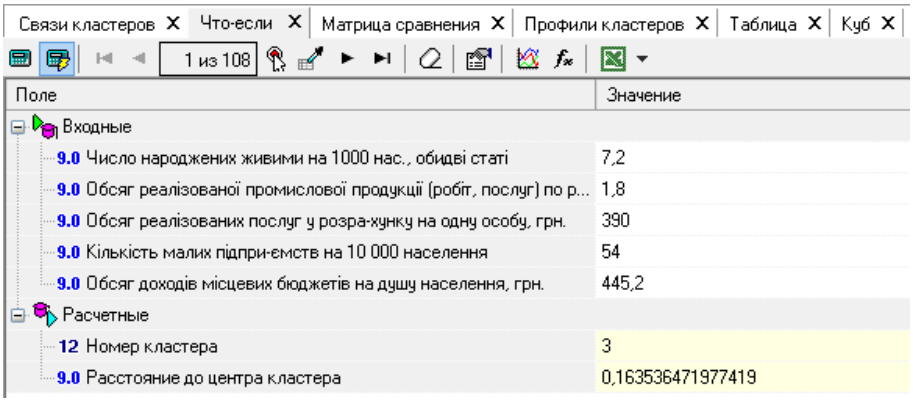


Рис.12. Засіб візуалізації “Что-если”.

За допомогою засобу візуалізації “Что-если” можна ввести набір вхідних даних та отримати номер кластеру до якого відноситься регіон з даним набором показників. Важливу інформацією про якість кластеризації надає візуалізатор “Профили кластеров” (Рис.13). Значущість дорівнює $(1-\alpha)100\%$ (де α - ймовірність нульової гіпотези). Для неперервних даних використовується t-критерій Ст'юдента, а для дискретних – критерій χ^2. Загальна значущість поля визначається за F- критерієм Фішера. Результати кластеризації найкраще представлені в OLAP-кубі, що показує розподіл по кластерах усіх регіонів по роках (Рис.14).

Рис.13. Показники якості кластеризації – профілі кластерів.

Связи кластеров | Что-если | Матрица сравнения | Профили кластеров | Таблица | Куб

Номер кластера

Регіон	Рік	0	1	2	3	4	5	Итого:
АР Крим	2001				1			1
	2002				1			1
	2003					1		1
	2004					1		1
	Итого:				2	2		4
Вінницька	2001				1			1
	2002					1		1
	2003					1		1
	2004					1		1
	Итого:				1	3		4
Волинська	2001						1	1
	2002						1	1
	2003						1	1
	2004						1	1
	Итого:						4	4
Дніпропетровська	2001			1				1
	2002			1				1
	2003		1					1
	2004		1					1
	Итого:		2	2				4
Донецька	2001			1				1
	2002			1				1
	2003		1					1
	2004		1					1
	Итого:		2	2				4

Рис.14. OLAP-куб із результатами кластеризації.

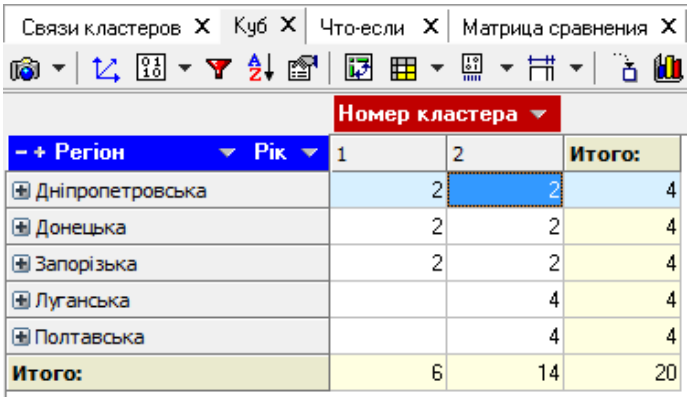
У кубі можна показати дані що відносяться до кожного окремого кластеру. Із матриці порівнянь на Рис.11 слідує, що можна умовно об’єднати перший та другий кластери, а також третій та четвертий. Дійсно, на Рис.15 показано нульовий кластер (у вікні “Номер кластера” ставимо прапорець тільки кластер 0) що містить тільки місто Київ за увесь період 2001-2004 р.р. На Рис.16 показано перший та другий кластери. Вони містять цілком Дніпропетровську, Донецьку, Запорізьку, Луганську та Полтавську області за увесь період 2001 – 2004 р.р. Третій та четвертий кластери містять цілком АР Крим, Вінницьку, Житомирську, Київську (без міста Киів), Кіровоградську, Львівську, Миколаївську, Одеську, Сумську, Тернопільську, Харківську, Херсонську, Хмельницьку, Черкаську, Чернігівську області та місто Севастополь за 2001-2004 р.р., майже цілком Івано-Франківську та Чернівецьку області за 2001-2003 р.р. (Рис.17). П’ятий кластер містить цілком Волинську, Закарпатську, Рівненську області за 2001-2004 р.р. та фрагментарно Івано-Франківську та Чернівецьку області тільки за 2004 р. (Рис.18).

Связи кластеров | Куб | Что-если | Матрица сравнения

Номер кластера

Регіон	Рік	0	Итого:
Київ	2001	1	1
	2002	1	1
	2003	1	1
	2004	1	1
	Итого:	4	4
Итого:		4	4

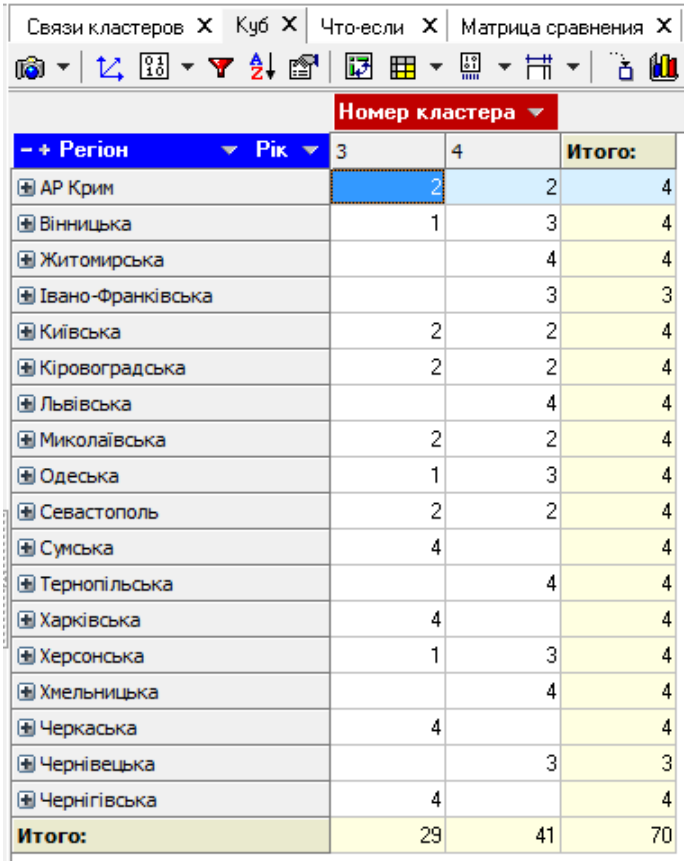
Рис.15. Нульовий кластер.



Связи кластеров | Куб | Что-если | Матрица сравнения

Регіон / Рік	Номер кластера 1	2	Итого:
Дніпропетровська	2	2	4
Донецька	2	2	4
Запорізька	2	2	4
Луганська		4	4
Полтавська		4	4
Итого:	6	14	20

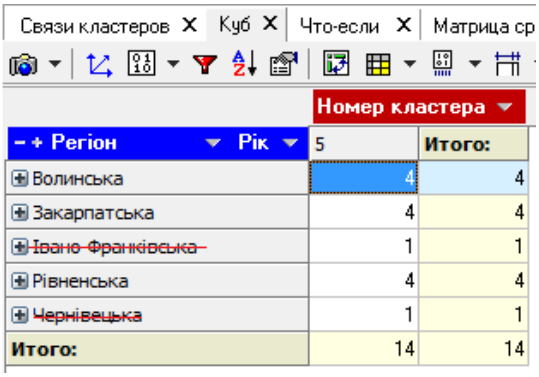
Рис.16. Перший та другий кластери.



Связи кластеров | Куб | Что-если | Матрица сравнения

Регіон / Рік	Номер кластера 3	4	Итого:
АР Крим	2	2	4
Вінницька	1	3	4
Житомирська		4	4
Івано-Франківська		3	3
Київська	2	2	4
Кіровоградська	2	2	4
Львівська		4	4
Миколаївська	2	2	4
Одеська	1	3	4
Севастополь	2	2	4
Сумська	4		4
Тернопільська		4	4
Харківська	4		4
Херсонська	1	3	4
Хмельницька		4	4
Черкаська	4		4
Чернівецька		3	3
Чернігівська	4		4
Итого:	29	41	70

Рис.17. Третій та четвертий кластери.



Связи кластеров | Куб | Что-если | Матрица ср

Регіон / Рік	Номер кластера 5	Итого:
Волинська	4	4
Закарпатська	4	4
Івано-Франківська	1	1
Рівненська	4	4
Чернівецька	1	1
Итого:	14	14

Рис.18. П’ятий кластер.

Таким чином, за результатами роботи алгоритму G-середніх можна зробити висновок про доцільність розбиття вибірки за змістом (тобто, проведення класифікації регіонів) на чотири кластери кожен з яких містить цілком (або переважну бі-

льшість) даних за період по кожному регіону. Івано-Франківьку та Чернівецьку області можна викреслити із п’ятого кластеру та залишити тільки у об’єднаному третьому та четвертому кластері (Рис.18).

2. Багатовимірний кластерний аналіз на основі карт Кохонена

Самоорганізуюча карта Кохонена (Kohonen, Self-Organizing Map (SOM)) – це нейронна мережа спеціального вигляду (завжди тільки з одним прихованим шаром), що виконує задачу багатовимірної кластеризації та візуалізації багатовимірного набору даних [1], [12], [16]. SOM відображує багатовимірні об’єкти у вигляді набору компактних і зручних для інтерпретації двовимірних карт. Пошук закономірностей у великих масивах даних за допомогою навчання нейронної мережі на ретроспективних даних дозволяє проводити розвідувальний аналіз багатовимірних вибірок даних який відрізняється від класичних статистичних процедур аналізу даних (в яких, як правило, перевіряється деякий набір статистичних гіпотез). Головна ідея методу полягає у тому що створюється проекція багатовимірного простору даних на набір двовимірних площин (карт), на яких впорядковані певним чином елементи (нейрони) розфарбовані за шкалою окремого фіксованого атрибуту даних для кожної карти. Карта Кохонена відображується за допомогою комірок шестикутної (частіше) або прямокутної форми (Рис.19).

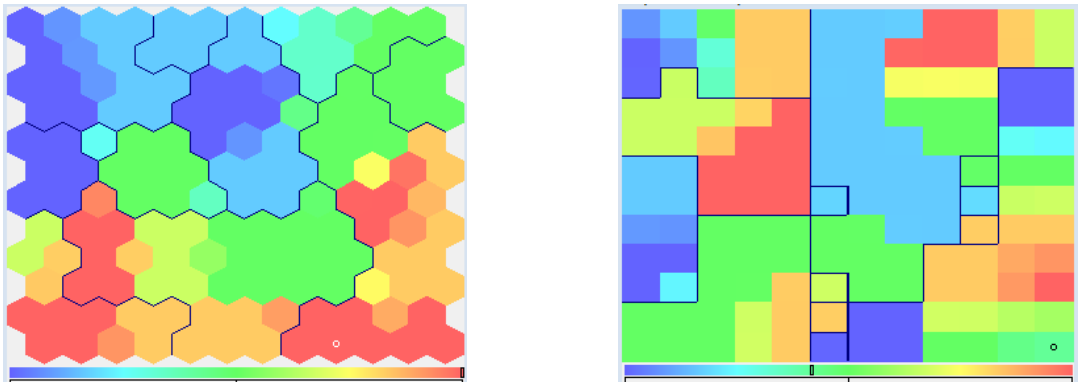


Рис.19. Структура комірок Карти Кохонена.

Розглянемо варіант карт Кохонена що реалізований в Deductor studio. При реалізації алгоритму SOM спочатку задається шестикутна конфігурація сітки, кількість нейронів на кожній карті (вузлів або центрів класів) та початкові значення вагових коефіцієнтів нейронів. Кількість нейронів дорівнює кількості елементів у вхідній виборці даних. Навчання нейронної мережі відбувається циклічно та складається з послідовності корекцій векторів, що представляють собою нейрони. У циклі перебираються усі карти, що відповідають окремому атрибуту вхідних даних. На кожному кроку цикла:

- обирається елемент із початкової вибірки даних (характеризується вектором атрибутів);
- для обраного елементу знаходиться найкращий вузол на даній карті вектор вагових коефіцієнтів якого менш всього відрізняється від вектору даного елементу в Евклідовій метриці (best matching unit, BMU);
- визначається кількість сусідів із вхідної вибірки для знайденого вузла та здійснюється навчання нейромережі – зміна набору вагових коефіцієнтів вузла та йо-

го сусідів для наближення їх значень до компонент вектору обраного елементу вхідної вибірки;
- обчислюється похибка даної карти та порівнюється із мінімальною із попередніх циклів.

При використанні цього алгоритму вектори, що схожі в вихідному просторі, розташовуються поруч і на окремій двовимірній карті. Для модифікації їх вагових коефіцієнтів $r_i(t_j)$ використовується формула:

$$r_i(t_j+1) = r_i(t_j) + h_{ci}(t_j+1)\big(x(t_j+1) - r_i(t_j)\big),$$

де t_j - номер епохи навчання (дискретний час), $x(t_j)$ - вектор елементу навчальної вибірки на ітерації t_j, $h_{ci}(t_j)$ - функція сусідства нейронів. В Deductor Studio можна використовувати одну з двох функцій $h_{ci}(t_j)$ - ступінчату або функцію Гауса. Функція Гауса має вигляд:

$$h_{ci}(t_j) = \alpha(t_j)\exp\left(-\frac{\sum_{\mathrm{i}}(\mathrm{y_{ki}}(t_j) - \mathrm{y_{mi}}(t_j))^2}{2\sigma^2(t_j)}\right)$$

де $\mathrm{y_k}(t_j)$, $\mathrm{y_k}(t_j)$ - вектори-координати двох сусідніх вузлів мережі, $\sigma(t_j)$ - коефіцієнт, що зменшує кількість сусідів з кожною ітерацією, $\sigma(t_j) \underset{t_j\to\infty}{\to} 0$, $\alpha(t_j) \in (0,1)$ - функція швидкості навчання. Найбільш поширена форма використання функції швидкості навчання $\alpha(t_j) = \frac{\mathrm{C}}{K + t_j}$, де C, K - додатні константи, $\alpha(t_j) \underset{t_j\to\infty}{\to} 0$. Завдяки використанню $\alpha(t_j)$ усі елементи $x(t_j)$ навчальної вибірки приблизно однаково впливають на кінцевий результат навчання.

Процес навчання нейронної мережі складається з двох етапів. Спочатку обираємо достатньо великі значення $C > 0$ (швидкості навчання) та радіусу навчання. Останній визначає кількість нейронів (вузлів) що задіяні у навчанні на поточній ітерації та зменьшується з кожною ітерацією аж до одного нейрона (вузла). Ці налаштування дозволяють на початку процесу розташувати вектора нейронів відповідно до розподілу екземплярів вибірки. Після цього на кожній проводиться обчислення вагових коефіцієнтів $r_i(t_j)$ разом із зменшенням значень функції швидкості навчання та радіусу навчання. Параметри навчання карт Коханена задаються у відповідному вікні налаштування методу.

Розглянемо роботу карт Коханена для розв'язування задачі кластеризації регіонів України за 5 демографічними та економічними показниками що розглянута у п.1. Встановлюємо курсор на вузол з імпортованими даними із файлу task7.txt, запускаємо майстер обробки та обираємо метод "Карта Кохонена". На наступних кроках налаштовуємо стовпчики вхідних даних так саме як показано на Рис.2 - 4. Далі налаштовуємо параметри методу (Рис.20, 21, 22).

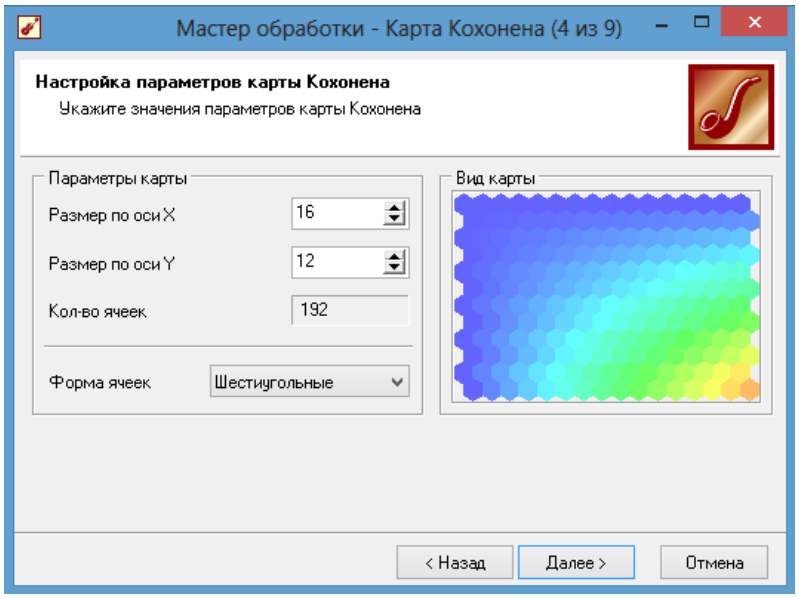


Рис.20. Налаштування параметрів карти Кохонена.

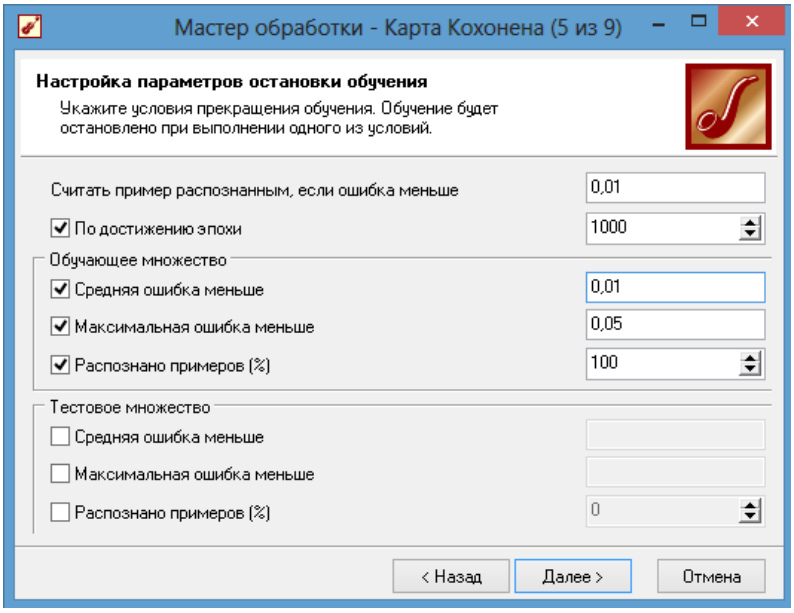


Рис.21. Налаштування параметрів зупинки навчання карти Кохонена.

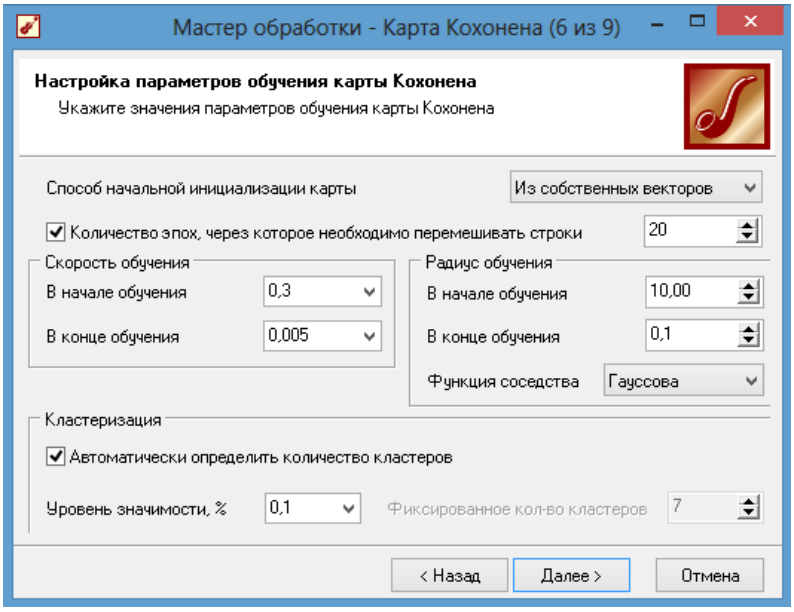


Рис.22. Налаштування параметрів навчання карти Кохонена.

У параметрах навчання вказуємо кількість епох 1000, середня та максимальна відносна похибка 0,01 (1%) та 0,05 (5%) (Рис.21), спосіб початкової ініціалізації “Из собственных векторов” (можна експериментувати також із “Случайними значениями” або “Из обучающего множества”), тип функції сусідства “Гауссова” із константами 0,3 (початкове) та 0,05 (кінцеве значення), радіусом навчання 10 (початкове) та 0,1 (кінцеве значення), автоматичне визначення числа кластерів з рівнем значущості 0,1 (Рис.22). Запускаємо процес навчання (Рис.23). Після налаштування засобів візуалізації (так саме як на Рис.7 - 9) отримуємо на Рис.24 результати у вигляді зв’язків побудованих 3 кластерів (від 0 до 2).

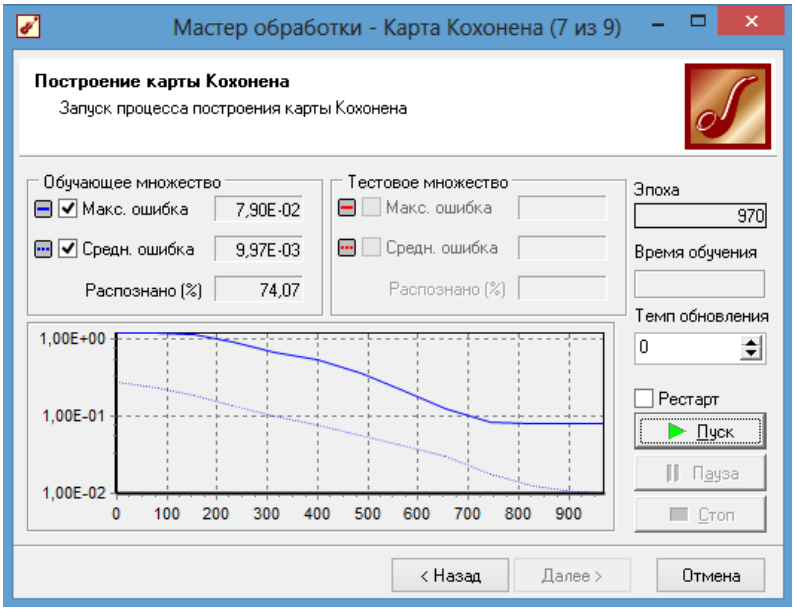


Рис.23. Запуск процесу навчання карт Кохонена.

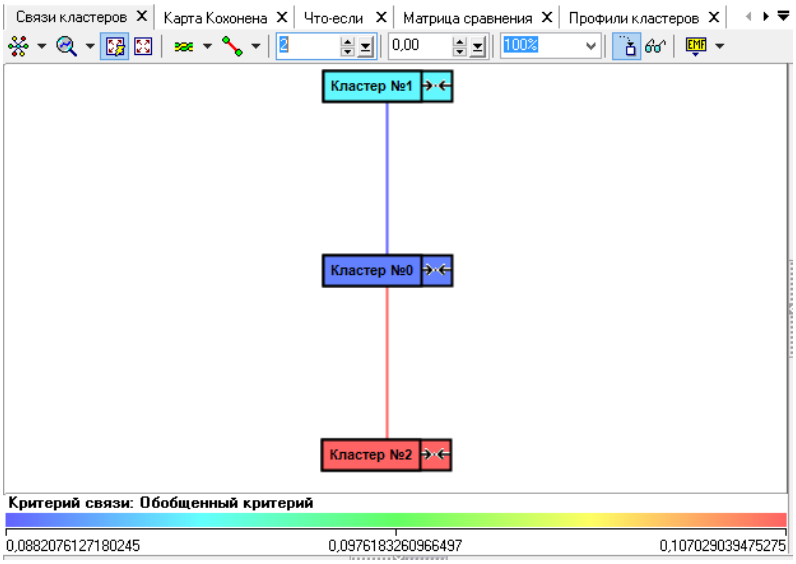


Рис.24. Зв’язки кластерів.

На Рис. 25 зображені карти Кохонена для побудованої кластеризації регіонів із трьома кластерами. Слід відзначити що на кожній карті візуально чітко виділено по одному кластеру (розфарбованому окремо згідно кольоровій шкалі окремої ознаки). На першій карті “Число народженими живими на 1000 нас.” візуально виокремлено другий кластер, на другій карті “Обсяг реалізованої промислової продукції…” виокремлено перший кластер, на решті карт виокре-

млено нульовий кластер. На карті “Кластери” трьома кольорами показано розташування вузлів (нейронів мережі) по трьом кластерам. Відсоток елементів вибірки, що потрапили до суміжних кластерів показано у матриці порівняння спільних елементів у кластерах (по парам) на Рис.26. Можна зробити висновок, що границя між першим та другим кластерами досить не чітка (53% спільних елементів). Візуалізатор “Что-если” на Рис.27 дозволяє встановлювати за набором вхідних параметрів для даної моделі номер кластеру регіону.

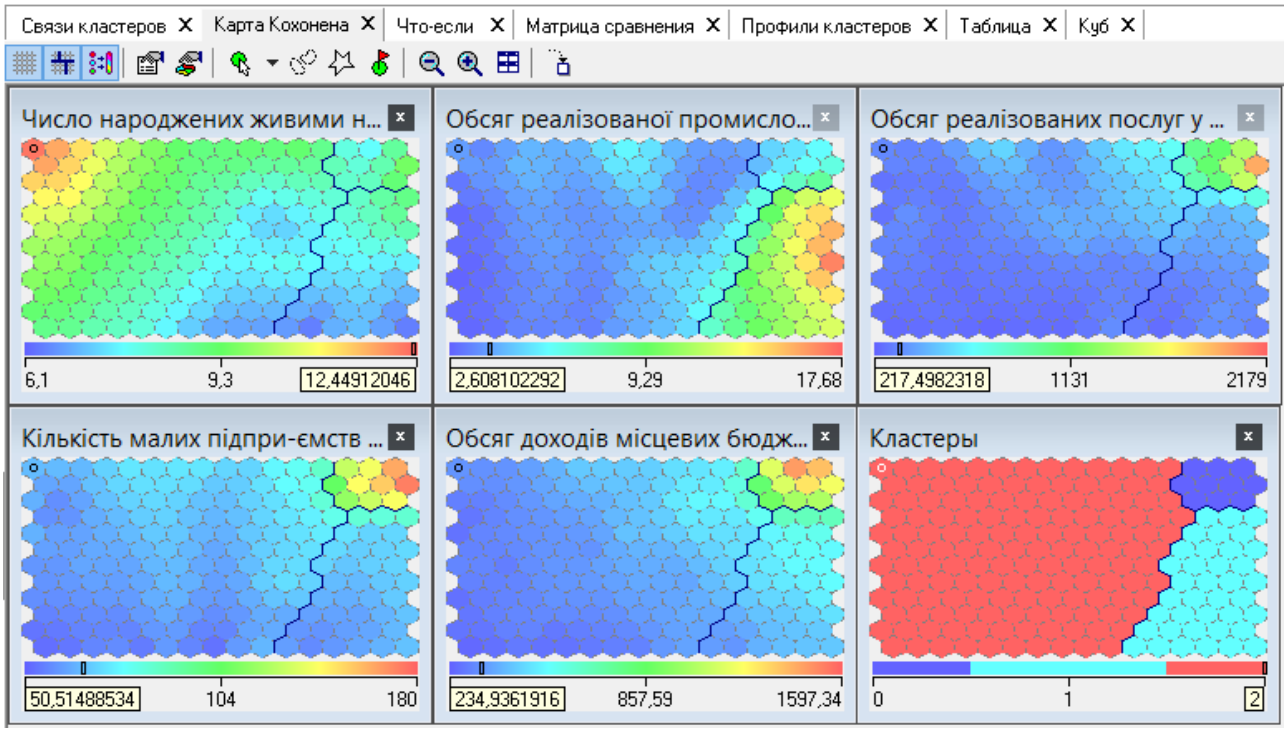


Рис.25. Побудовані карти Кохонена.

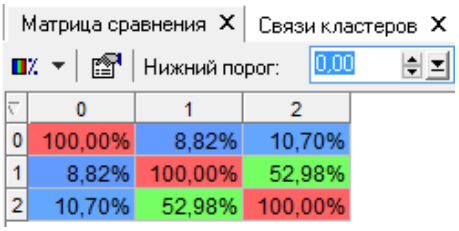


Рис.26. Матриця порівняння спільних елементів у кластерах (по парам).

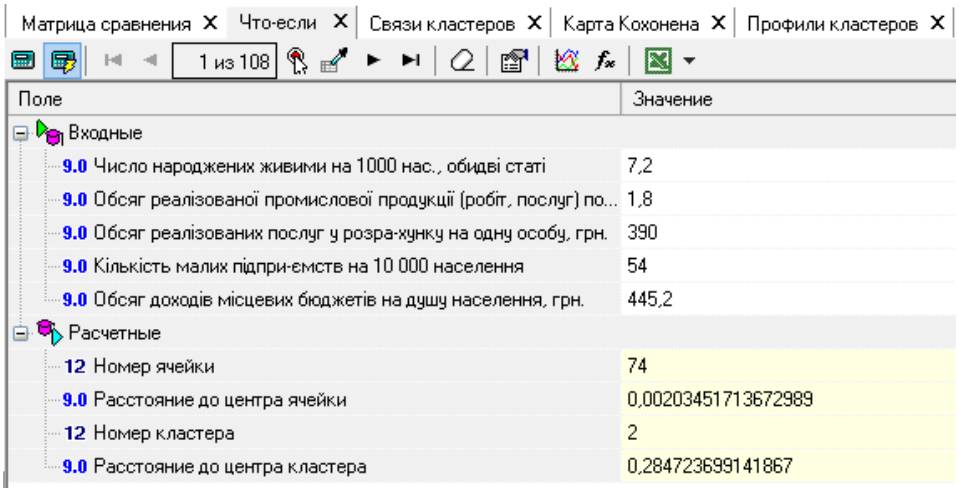


Рис.27. Засіб візуалізації “Что-если”.

Статистичні показники кожного кластеру по кожному фактору показані у вікні «Профили кластеров” на Рис.28. Слід відзначити що перший та четвертий фактори є значущими для другого кластеру (99-100%), четвертий та п’ятий фактори – для першого кластеру (100%), перший другий та третій фактори – для нульового кластеру (100%). Показано розподіл елементів вибірки по кластерах: другий кластер містить 77,8% вибірки, перший кластер – 18,5%, нульовий кластер – 3,7%.

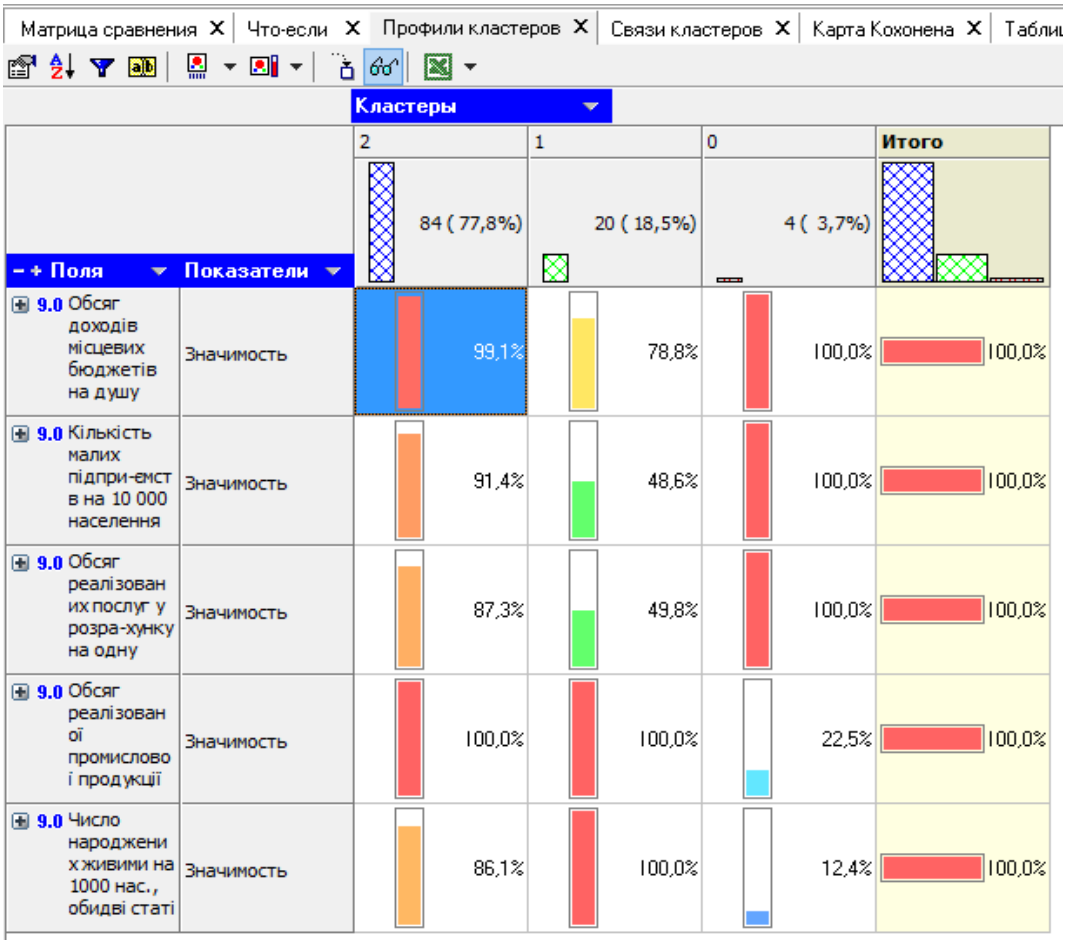


Рис.28. Профілі кластерів.

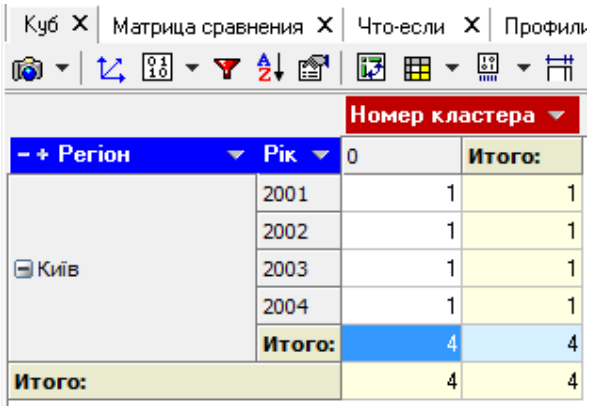


Рис.29. Нульовий кластер.

Зміст кожного з трьох кластерів показано на Рис. 29 – 31. Нульовий та перший кластери повністю співпадають з нульовим та об’єднаними першим-другим кластерами методу G-середніх (Рис.15, 16, відповідно). Але, на відміну від попере-

днього методу карти Кохонена поміщають решту регіонів у один кластер (другий кластер на Рис.31). З точки зору історії соціально-економічного розвитку країни виокремлення декількох західних регіонів у окремий кластер є більш доцільним, так як це зроблено методом G-середніх у п'ятому кластері (Рис.18). Тому, кластеризація методом G-середніх виявляється більш реалістичною. З іншого боку, карти Кохонена показують наглядно (Рис.25, матриця порівняння Рис.26) що межа між кластерами досить розмита і виявити тонку різницю між регіонами що входять до складу побудованих методом G-середніх об'єднаних третього з четвертим та відокремленого п'ятого кластерів дуже складно.

Куб | Матрица сравнения | Что-если | Профиль

	Номер кластера	
Регіон / Рік	1	Итого:
Дніпропетровська	4	4
Донецька	4	4
Запорізька	4	4
Луганська	4	4
Полтавська	4	4
Итого:	20	20

Рис.30. Перший кластер.

Куб | Матрица сравнения | Что-если | Профиль

	Номер кластера	
Регіон / Рік	2	Итого:
АР Крим	4	4
Вінницька	4	4
Волинська	4	4
Житомирська	4	4
Закарпатська	4	4
Івано-Франківська	4	4
Київська	4	4
Кіровоградська	4	4
Львівська	4	4
Миколаївська	4	4
Одеська	4	4
Рівненська	4	4
Севастополь	4	4
Сумська	4	4
Тернопільська	4	4
Харківська	4	4
Херсонська	4	4
Хмельницька	4	4
Черкаська	4	4
Чернівецька	4	4
Чернігівська	4	4
Итого:	84	84

Рис.31. Другий кластер.

Слід відзначити також, що спроби покращити результати побудованої в автоматичному режимі кластеризації регіонів методом карт Кохонена за рахунок заданих в ручному режимі кількості кластерів та інших налаштувань методу не привели до бажаних результатів. Аналітик завжди має проводити дослідження різними методами для порівняння результатів їх роботи та отримання нових прихованих закономірностей.

3. Метод багатовимірної ЕМ-кластеризації

Пакет Deductor Studio пропонує аналітикам ще один метод ЕМ-кластеризації (Expectation-Maximization algorithm). Цей алгоритм рекомендовано використовувати [1] якщо результати роботи двох попередніх методів не задовольнили аналітика та експертів. По суті алгоритм ЕМ-кластеризації означає ітераційний процес максимізації математичного очікування логарифма функції правдоподобія для даних вибірки [13], [16].

Розглянемо роботу даного алгоритму для розв’язування задачі кластеризації регіонів України із п.1. Встановлюємо курсор на вузлі імпортованих даних із файлу task7.txt, запускаємо майстер обробки та обираємо у вікні “ЕМ-кластеризация”. Налаштовуємо параметри вхідних стовпчиків так саме як на Рис.2 – 4 та обираємо на наступному кроці автоматичне визначення кількості кластерів з вказівкою максимальна кількість 6 та мінімальна кількість елементів в кластері 4 (враховуючи досвід роботи із іншими методами кластеризації у попередніх пунктах), нижній поріг правдоподобія 0,1 (Рис.32).

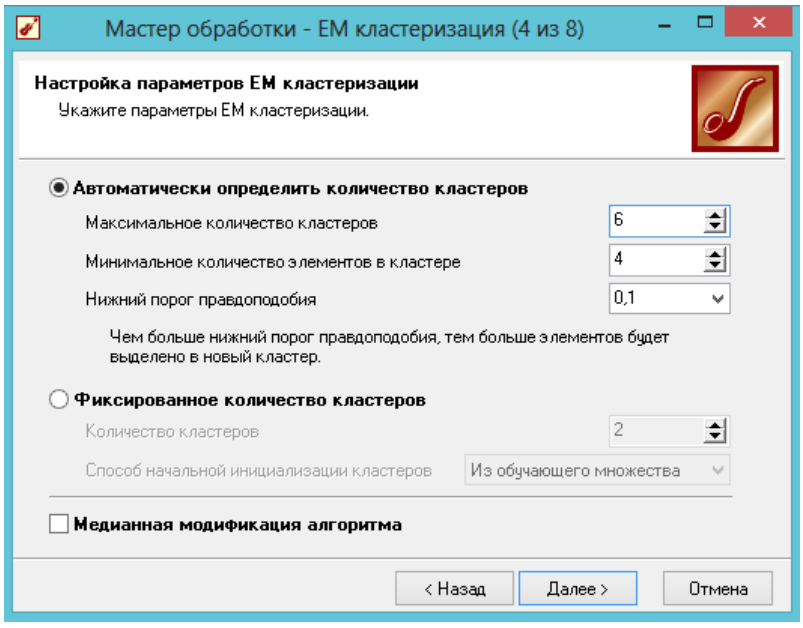


Рис.32 Налаштування параметрів ЕМ-кластеризації.

Далі встановлюємо параметри зупинки ітераційного процесу навчання методу (обчислення максимуму логарифма функції правдоподобія): рівень точності моделі 10^{-5}, максимальна кількість ітерацій 10000 (Рис.33) та запускаємо процес навчання (Рис.34). Після налаштування засобів візуалізації (так саме як на Рис.7 - 9) отримуємо на Рис.35 результати у вигляді зв’язків побудованих 6 кластерів (від 0 до 5).

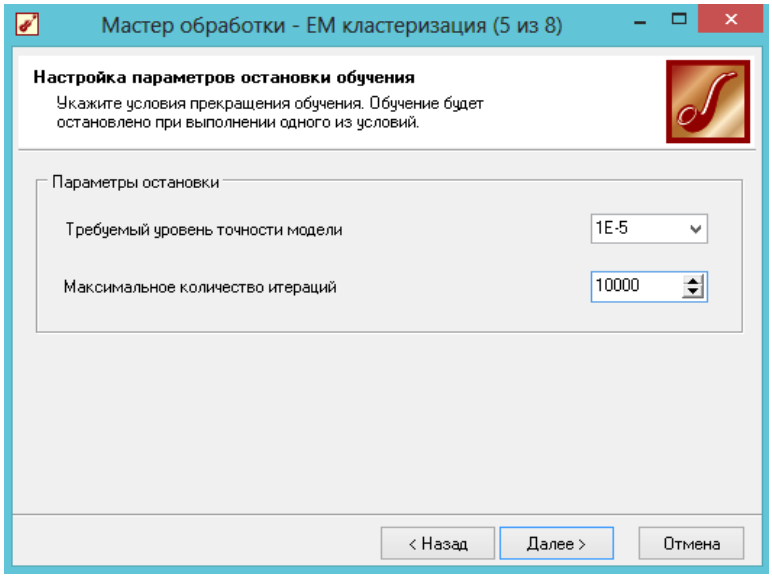


Рис.33. Налаштування параметрів зупинки навчання EM-кластеризації.

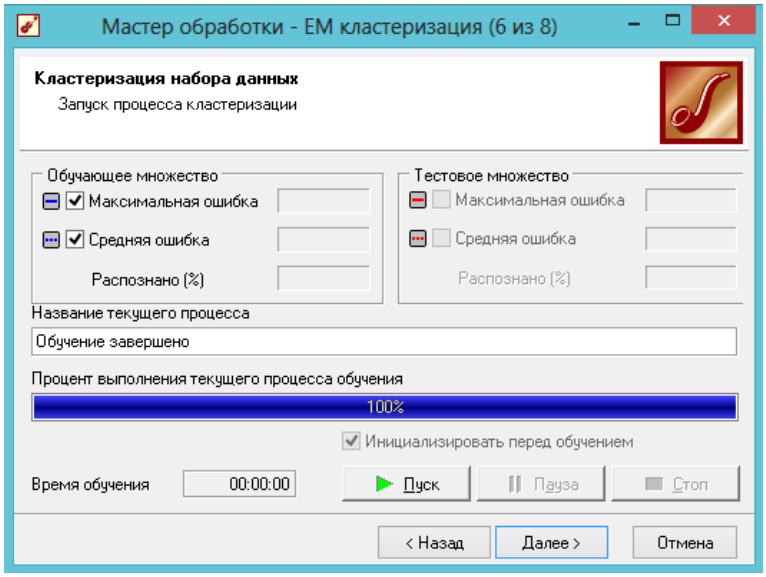


Рис.34. Запуск методу EM-кластеризації.

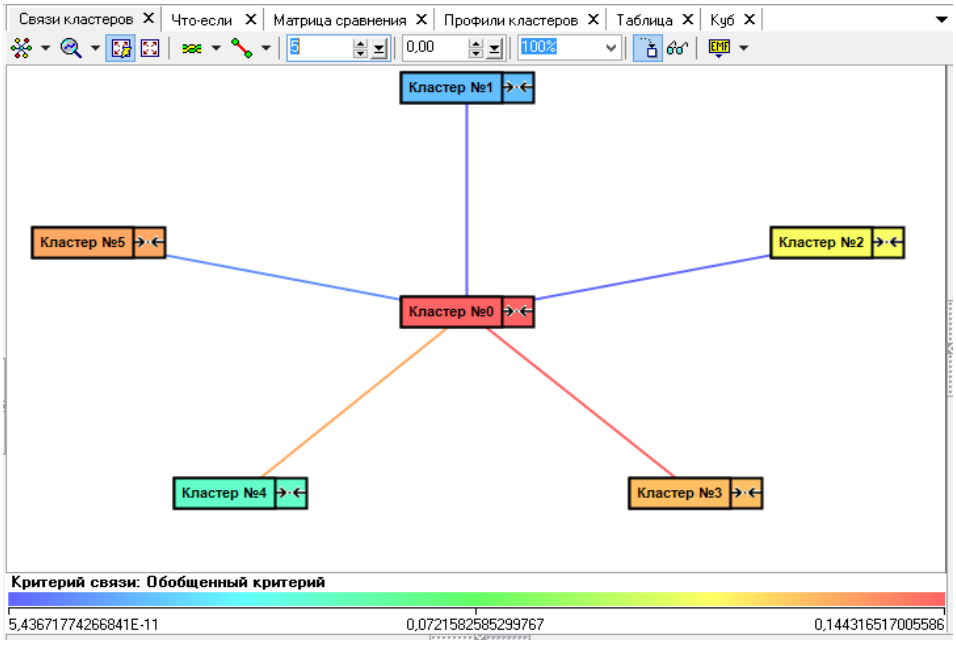
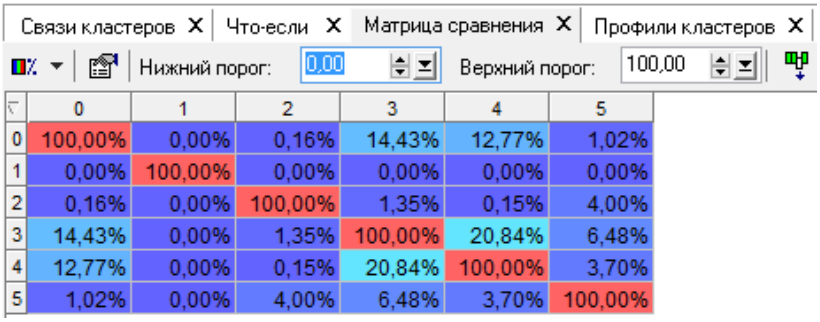


Рис.35. Зв’язки кластерів.

Із таблиці порівняння спільних елементів у кластерах (по парам кластерів) (Рис.36) слід що більшість спільних елементів знаходяться у нульовому з третім та четвертим кластерах (>10%), та у третьому з п’ятим (>6%). Це говорить про те що, за необхідністю, у разі отримання додаткової експертної інформації про об’єкти спостережень (наприклад, історія розвитку регіонів, особливості географічного розташування, спільні кордони регіонів, тощо) аналітик може розглядати ці кластери як кандидати на об’єднання їх в один кластер.



	0	1	2	3	4	5
0	100,00%	0,00%	0,16%	14,43%	12,77%	1,02%
1	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%
2	0,16%	0,00%	100,00%	1,35%	0,15%	4,00%
3	14,43%	0,00%	1,35%	100,00%	20,84%	6,48%
4	12,77%	0,00%	0,15%	20,84%	100,00%	3,70%
5	1,02%	0,00%	4,00%	6,48%	3,70%	100,00%

Рис.36. Матриця порівняння спільних елементів у кластерах (по парам).

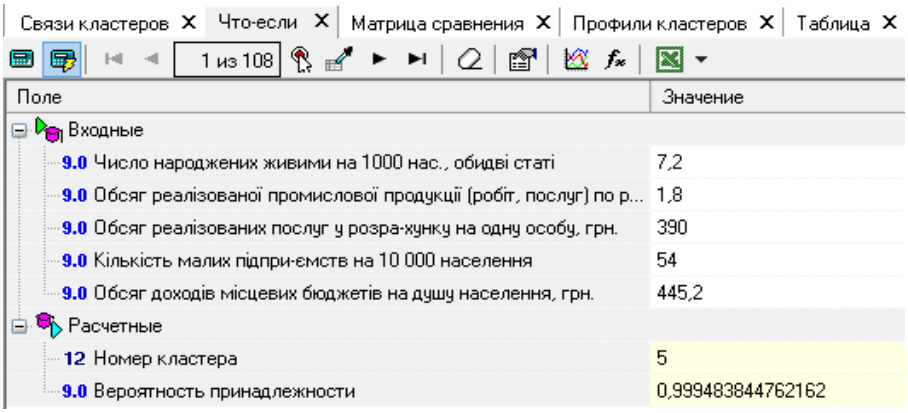


Рис.37. Засіб візуалізації “Что-если”.

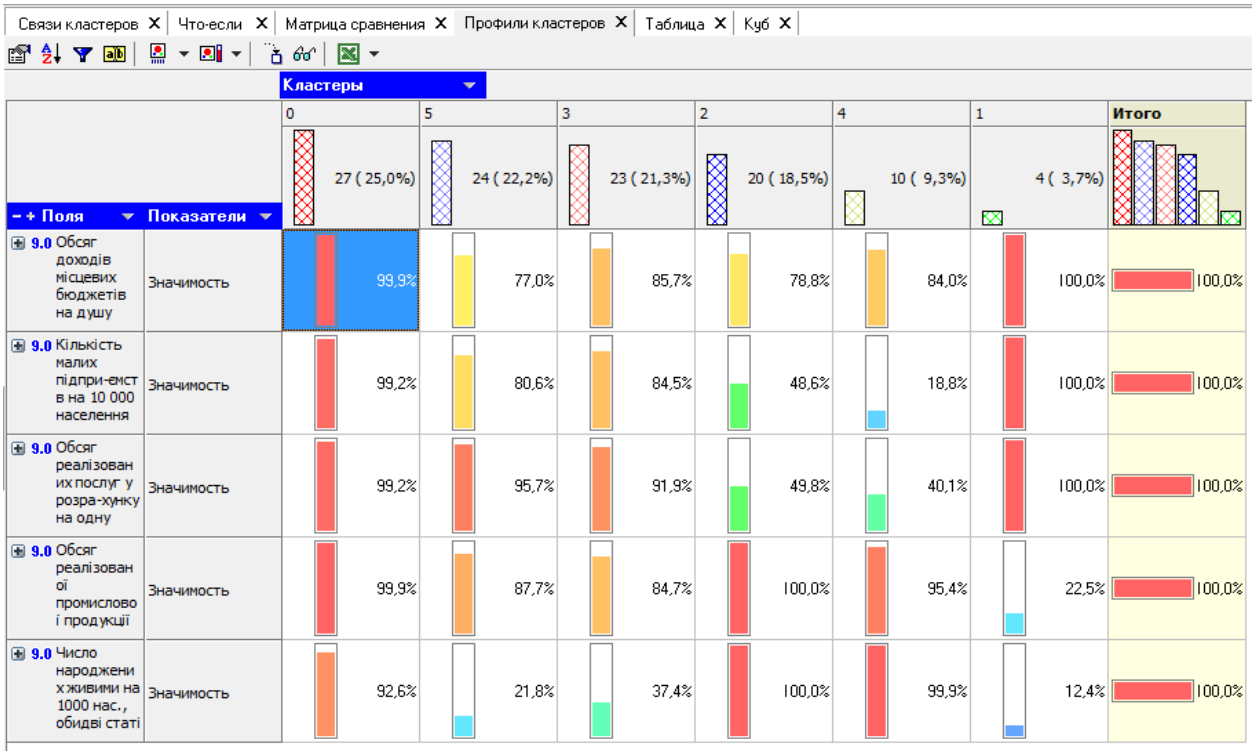


Рис.38. Профілі кластерів.

За допомогою візуалізатору “Что-если” на Рис.37 можна встановлювати номер кластеру регіону за новим набором вхідних параметрів для даної моделі. Кількісні характеристики побудованої моделі кластеризації показані у таблиці профілів кластерів на Рис.38. Можна зробити висновки про значущість кожного із 5 факторів для кожного з 6 кластерів.

Зміст кожного кластера побудованої моделі показано на Рис.39 – 41. Перший кластер (Рис.39) співпадає з нульовим кластером методу G-середніх (Рис.15) та нульовим кластером карт Кохонена (Рис.29). Другий кластер (Рис.40) співпадає з об’єднаним першим-другим кластером методу G-середніх (Рис.16) та першим кластером карт Кохонена (Рис.30). Решта кластерів - нульовий, третій, четвертий та п’ятий разом (Рис.41) співпадають із об’єднанням третім, четвертим та п’ятим кластерами методу G-середніх (Рис17, 18) та з другим кластером карт Кохонена (Рис.31). Можна зробити висновок, що результати роботи ЕМ-кластеризації узгоджуються із результатами двох попердніх методів кластеризації у випадку якщо (згідно матриці порівняння) об’єднати нульовий, третій, четвертий та п’ятий кластери в один.

Таким чином, усі три методи дозволяють ефективно проводити кластеризацію даних у багатовимірному просторі елементів вибірки в режимі автоматичного вибору числа кластерів. Отримані різними методами результати кластеризації узгоджуються між собою та за умови комплексного застосування дозволяють аналітику проводити класифікацію елементів вибірки з максимальним правдоподобієм та мінімальною похибкою.

Куб | Связи кластеров | Что-если | Матрица сравнения

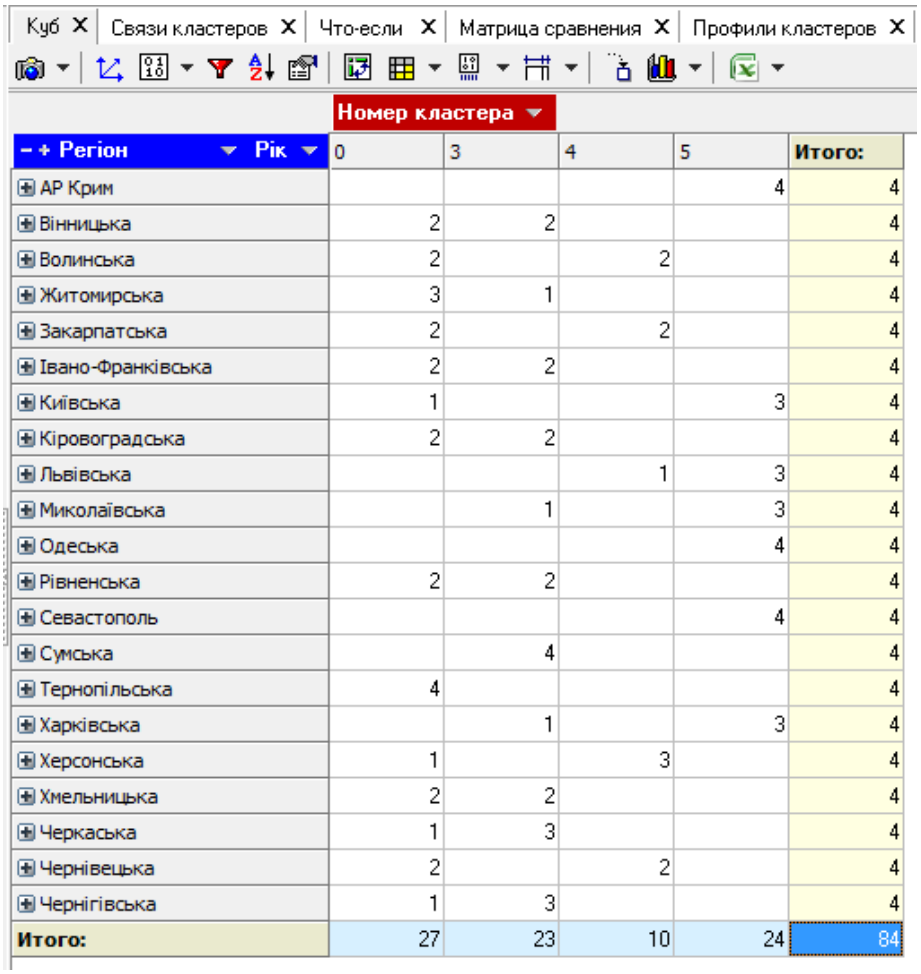
		Номер кластера	
Регіон	Рік	1	Итого:
Київ	2001	1	1
	2002	1	1
	2003	1	1
	2004	1	1
	Итого:	4	4
Итого:		4	4

Рис.39. Перший кластер.

Куб | Связи кластеров | Что-если | Матрица сравнения

		Номер кластера	
Регіон	Рік	2	Итого:
Дніпропетровська		4	4
Донецька		4	4
Запорізька		4	4
Луганська		4	4
Полтавська		4	4
Итого:		20	20

Рис.40. Другий кластер.



Куб | Связи кластеров | Что-если | Матрица сравнения | Профили кластеров

Регіон / Рік	Номер кластера 0	3	4	5	Итого:
АР Крим				4	4
Вінницька	2	2			4
Волинська	2		2		4
Житомирська	3	1			4
Закарпатська	2		2		4
Івано-Франківська	2	2			4
Київська	1			3	4
Кіровоградська	2	2			4
Львівська			1	3	4
Миколаївська		1		3	4
Одеська				4	4
Рівненська	2	2			4
Севастополь				4	4
Сумська		4			4
Тернопільська	4				4
Харківська		1		3	4
Херсонська	1		3		4
Хмельницька	2	2			4
Черкаська	1	3			4
Чернівецька	2		2		4
Чернігівська	1	3			4
Итого:	27	23	10	24	84

Рис.41. Нульовий, третій, четвертий та п’ятий кластери.

Завдання для лабораторної роботи 7.

1. Створити файл Lab7.xlsx. Створити таблицю з 2 демографічними та 3 соціально-економічними показниками для 25 областей України за 2009-2012 р.р. (7 стовпчиків 26 рядків). Назви стовпчиків: Область, Рік, Демографічний 1, Демографічний 2, Соц.-ек.1, Соц.-ек.2, Соц.-ек.3.
2. Перевести дані у безрозмірний формат за кожен рік окремо (нормалізація даних у відрізок [0,1]).
3. Зберегти файл Lab7.xlsx та як Lab7.txt.
2. Імпортувати файл Lab7.txt у Deductor studio.
2. Виконати кластеризацію даних методами G-середніх, самоорганізуючих карт Кохонена та ЕМ-кластеризації.
3. Порівняти побудовані різними методами різні моделі кластерів за допомогою засобів візуалізації.

Тема 8. Пошук прихованих закономірностей у транзакціях на основі кластеризації транзакцій та асоціативних правил

Задача: Пошук прихованих закономірностей у транзакціях.
Методи: Кластеризація транзакцій. Асоціативні правила.

1. Кластеризація транзакцій

Транзакція – це сукупність подій що відбуваються майже одночасно у фіксований проміжок часу. Формально, транзакція задається за допомогою записів із двох полів в вибірці даних. Перше поле – це ідентифікатор (ID) транзакції (не число), друге поле – значення транзакції (не число). Наприклад:

№101 Чай, цукор, печиво

де "№101" – ідентифікатор чека покупки, "Чай, цукор, печиво" – товар у чеку - це елементи зліченої множини. Множина усіх транзакцій утворює таблицю із двох стовпчиків. Перший стовпчик містить ID транзакцій, які можуть повторюватись у стовпчику декілька разів, оскільки одна транзакція може відповідати декільком елементам вибірки (наприклад, декілька товарів (елементів) із одного чека). Також різні ID можуть відповідати одному й тому ж самому елементу (наприклад, різні чеки містять один товар). У Deductor Studio обидві стовпчики вибірки повинні містити елементи зліченої множини не числового типу [1].

Для первинного аналізу множини транзакцій використовується кластеризація транзакцій, що реалізована на основі алгоритму CLOPE (Clastering with sLOPE) [1], [14], [16]. Основною ідею неієрархічного ітеративного методу CLOPE є використання глобального критерію оптимізації на основі пошуку максимуму функції вартості спеціального вигляду, яка підвищує близькість транзакцій в кластерах за рахунок збільшення параметру кластерної гістограми [14], [16].

Розглянемо приклад роботи алгоритму для 5 соціально-економічних та демографічних показників по областях України за період з 2001 – 2004 р.р. (див. Тема 7). Для підготовки файлу з транзакціями, кожен і-елемент j-стовпчика x_{ij} (за кожен рік спостережень окремо) спочатку приводиться до безрозмірного цілечисельного вигляду із значеннями з діапазону $\hat{x}_{ij} = 0,...,8$ за формулою:

$$\hat{x}_{ij} = round\left(\frac{(x_{ij} - x_{min\ j})}{(x_{max\ j} - x_{min\ j})}8\right),$$

де *round*(x) - функція округлення дійсного числа x до найближчого цілого, $x_{max\ j}$ та $x_{min\ j}$ - максимальне та мінімальне значення j-стовпчика за один рік. Кожному із 9 значень безрозмірного параметра $\hat{x}_{ij}$ ставиться у відповідність терм лінгвістичної змінної (Таблиця 1). Так, для даних із 5 соціально-економічних та демографічних показників по областях України за період з 2001 – 2004 р.р. (див. Тема 7) встановлюємо ID кожної транзакції – назва області та рік. Наприклад, Київська 01 відповідає Київській області за 2001р. Розташовуємо в один стовпчик області за 2001 р., під ним області за 2002 р., 2003 р., 2004 р.

Таблиця 1.

Числове значення	Терм лінгвістичної змінної
0	дуже низьке значення
1	низьке значення
2	не низьке значення
3	нижче середнього значення
4	середнє значення
5	вище середнього значення
6	не високе значення
7	високе значення
8	дуже високе значення

У елементах транзакцій використовуємо назви соціально-економічних та демографічних показників як лінгвістичні змінні та для їх термів використовуємо наступні скорочення:

- “Число народжених живими на 1000 нас., обидві статі” = {“Дуже низька народж.”,…, “Дуже висока народж.”};
- “Обсяг реалізованої промислової продукції (робіт, послуг) по регіонах грн. на одну особу”={“Дуже низьк. обсяг реал.прод.”,…, “Дуже висок. обсяг реал. прод.”};
- “Обсяг реалізованих послуг у розрахунку на одну особу, грн.”= {“Дуже низьк. обсяг реал. послуг”,…, “Дуже висок. обсяг реал. послуг”};
- «Кількість малих підприємств на 10000 населення”= {“Дуже низька кільк. м.п.”,…, “Дуже висок. кільк. м.п.”};
- “Обсяг доходів місцевих бюджетів на душу населення, грн.”= {“Дуже низьк. обсяг дох. бдж.”,…, “Дуже висок. обсяг дох. бдж.”}

Дані у форматі двох стовпчиків заносимо у текстовий файл task8.txt. Після імпорту даних із файлу отримуємо таблицю із 500 рядків (25 областей х 4 роки х 5 показників), фрагмент якої зображений на Рис.1. Після встановлення курсору на вузлі імпортованого файлу завантажуємо майстер обробки та обираємо у вікні метод “Кластеризация транзакций”. Далі програма розпізнає два стовпчики – перший має призначення ID та назву «Область”, другий – призначення “Елемент” та назву “Транзакція” (Рис.2). Обидві стовпчики мають тип даних строковий, вид даних – дискретний.

Далі налаштовуємо параметри кластеризації транзакцій (Рис.3) - максимальна кількість кластерів 100, максимальна кількість ітерацій 10 та запускаємо метод на виконання (Рис.4). Програма успішно побудувала 62 кластери за 3 ітерації досягнувши максимуму функції вартості. Налаштовуємо засоби візуалізації (Рис.5) та поля і виміри OLAP-куба (Рис.6, 7).

Результати кластеризації показані на Рис.8 – 10. Матриця парних порівнянь кластерів (Рис.8) з порогом відсічі 40% спільних елементів демонструє якість проведеної кластеризації. За допомогою даної матриці можна виокремлювати групи кластерів з великою кількістю спільних елементів для подальшого об’єднання у разі необхідності (при наявності додаткової інформації про іс-

нуючі зв’язки між кластерами) та, навпаки, підкреслити ізольовані, чітко відокремлені від решти кластери для подальшого вивчення.

Таблица | Статистика

1 / 500

Область	Транзакція
АР Крим 01	Не низька народж.
Вінницька 01	Нижче серед. народж.
Волинська 01	Дуже висока народж.
Дніпропетровська 01	Не низька народж.
Донецька 01	Дуже низька народж.
Житомирська 01	Нижче серед. народж.
Закарпатська 01	Висока народж.
Запорізька 01	Не низька народж.
Івано-Франківська 01	Вище серед. народж.
Київська 01	Не низька народж.
Кіровоградська 01	Не низька народж.
Луганська 01	Дуже низька народж.
Львівська 01	Серед. народж.
Миколаївська 01	Нижче серед. народж.
Одеська 01	Нижче серед. народж.
Полтавська 01	Низька народж.
Рівненська 01	Дуже висока народж.
Сумська 01	Низька народж.
Тернопільська 01	Серед. народж.
Харківська 01	Низька народж.
Херсонська 01	Нижче серед. народж.
Хмельницька 01	Нижче серед. народж.
Черкаська 01	Не низька народж.
Чернівецька 01	Не висока народж.
Чернігівська 01	Низька народж.
АР Крим 02	Не низька народж.
Вінницька 02	Нижче серед. народж.
Волинська 02	Висока народж.
Дніпропетровська 02	Не низька народж.

Рис.1. Фрагмент таблиці імпортованого файлу із транзакціями.

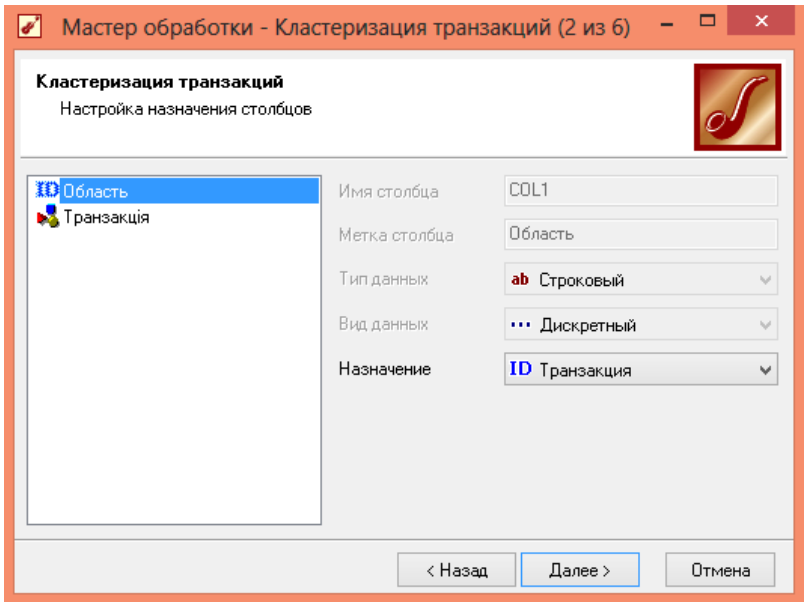


Рис.2. Налаштування стовпчиків транзакцій.

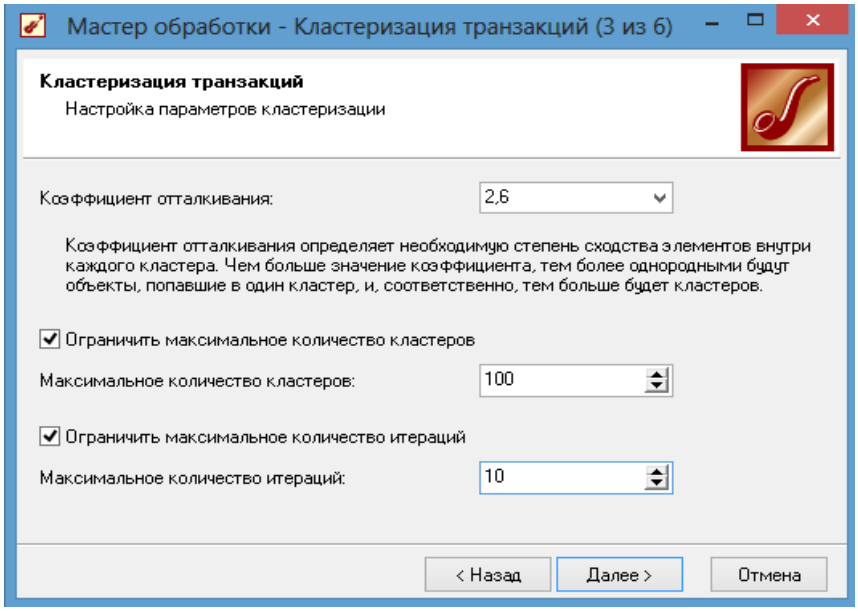


Рис.3. Налаштування параметрів кластеризації транзакцій.

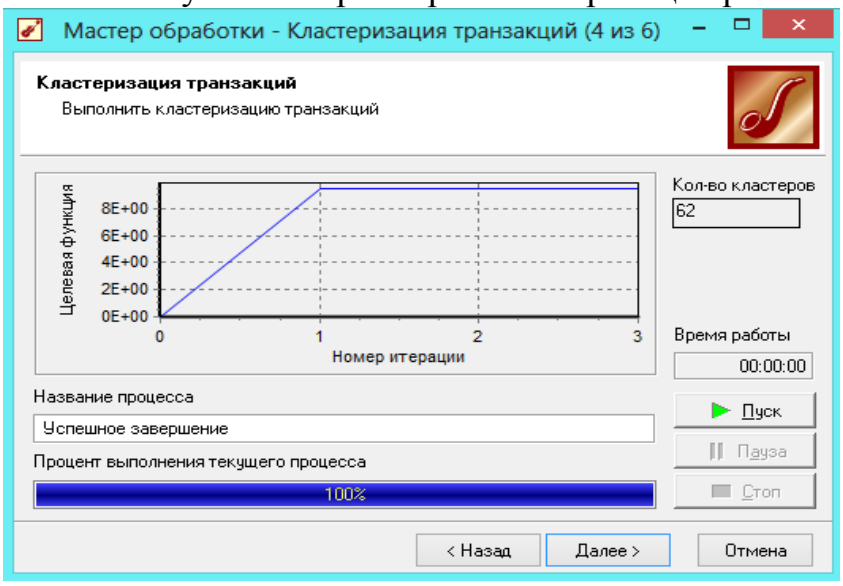


Рис.4. Запуск кластеризації транзакцій на виконання.

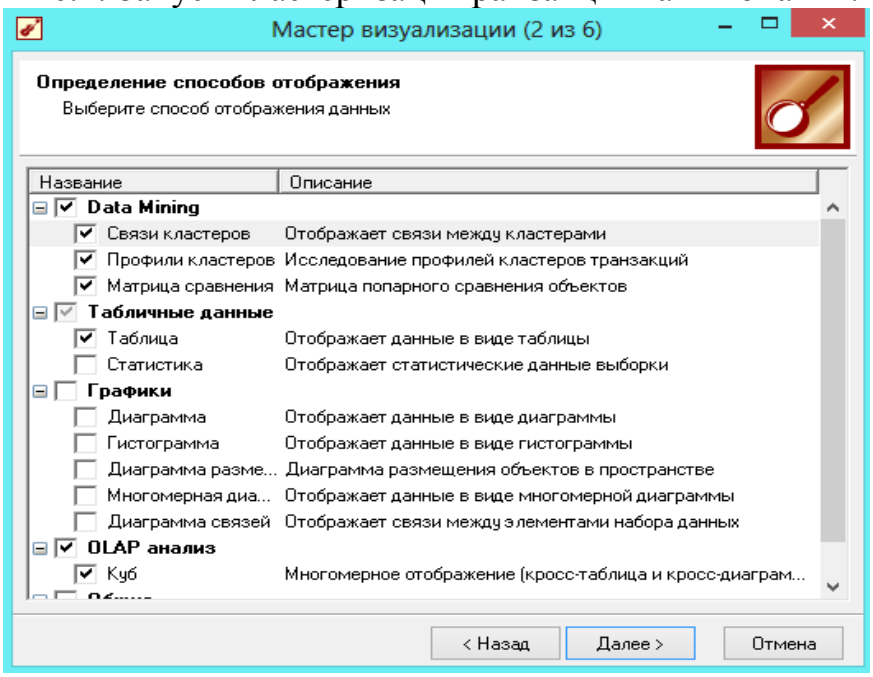


Рис.5. Налаштування засобів візуалізації.

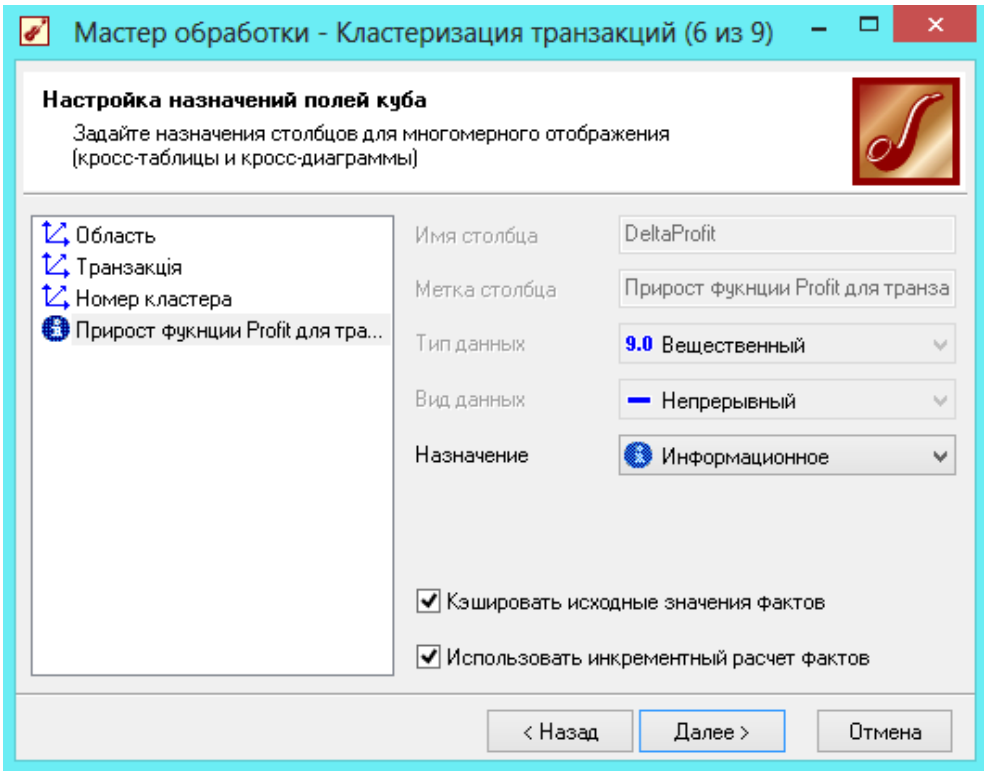


Рис.6. Налаштування полів OLAP-куба.

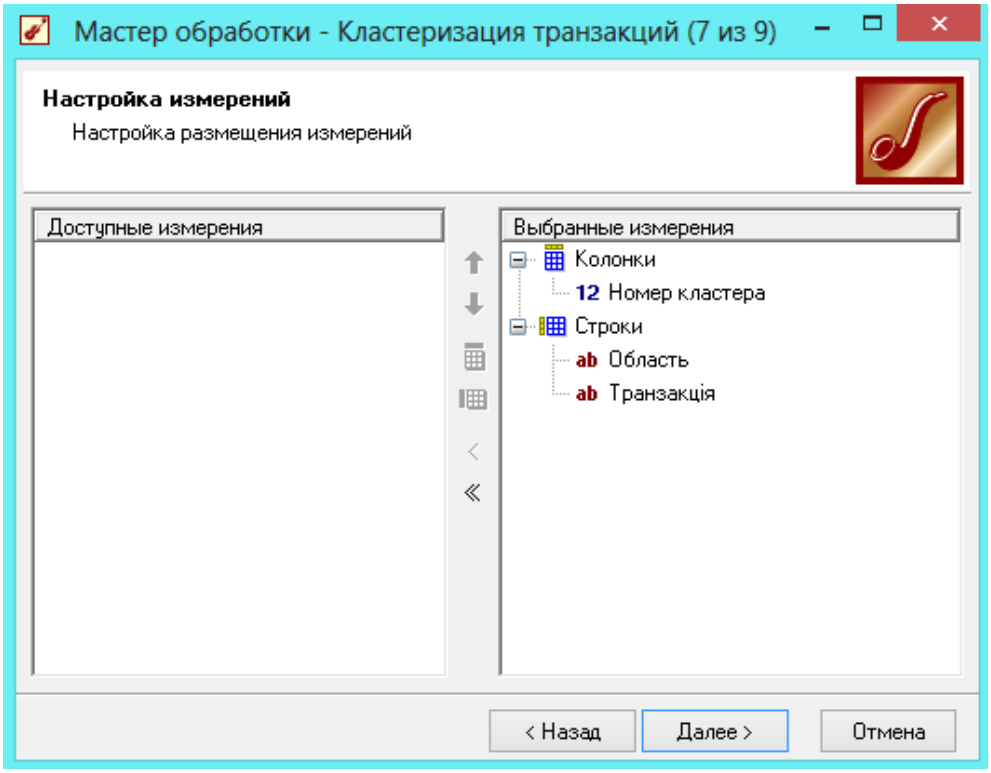


Рис.7. Налаштування вимірів OLAP-куба.

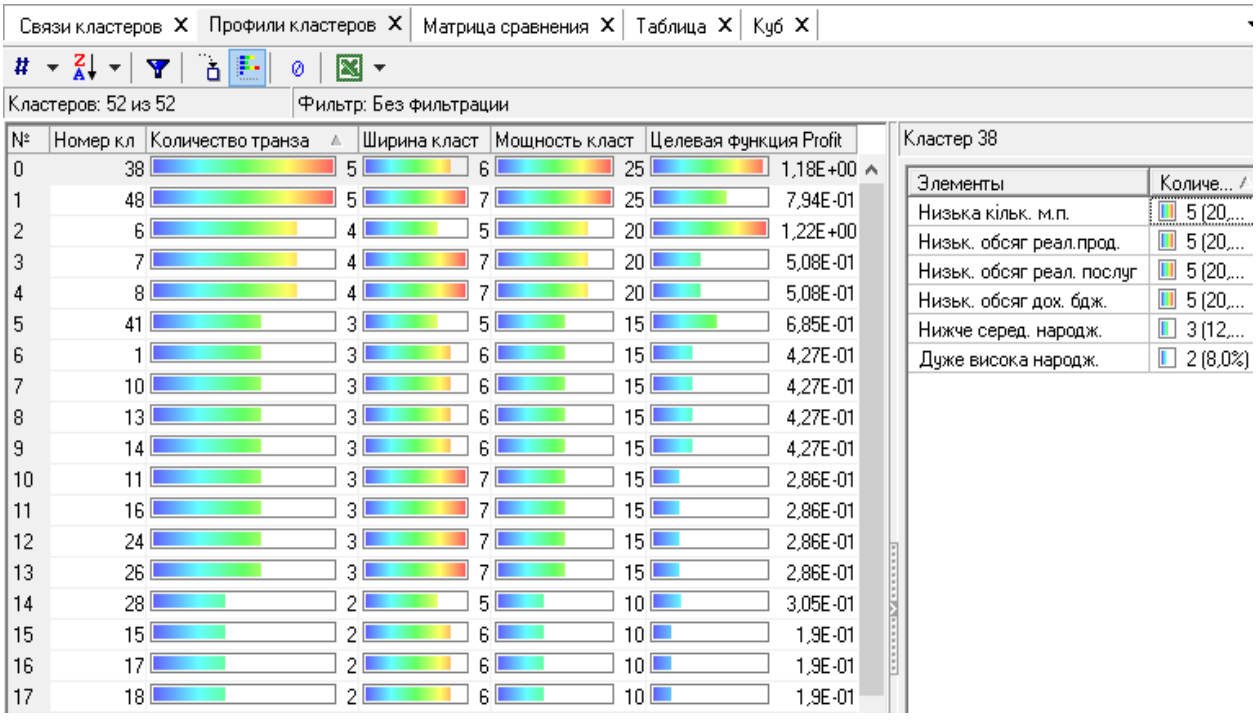
Характеристики кластерів представлені в таблиці “Профили кластеров” (Рис.9). У третьому стовпчику показано скільки транзакцій містить кожен кластер. Два самих великих кластери містять по 5 транзакцій, три кластери – по 4, дев’ять кластерів – по 3 і решта по 2 і менш. Табличка праворуч показує що у 38 кластері (№ 0) у усіх 5 транзакціях співпадають 4 елементи, а у трьох і інших двох транзакціях співпадають по п’ять елементів.

Связи кластеров | Профили кластеров | Матрица сравнения | Таблица | Куб

Нижний порог: 40,00 Верхний порог: 100,00

	0	1	2	3	4	5	6	7	8	9	10	11
0	100,00%	60,00%					40,00%		50,00%			
1	60,00%	100,00%					60,00%		40,00%			
2			100,00%	60,00%		40,00%						
3			60,00%	100,00%	40,00%	60,00%						
4				40,00%	100,00%	60,00%						
5			40,00%	60,00%	60,00%	100,00%	40,00%					
6	40,00%	60,00%				40,00%	100,00%		40,00%			
7								100,00%	45,00%			
8	50,00%	40,00%					40,00%	45,00%	100,00%			
9										100,00%	53,33%	
10										53,33%	100,00%	
11												100,00%
12												53,33%
13	60,00%	53,33%				40,00%	60,00%		56,67%			
14	40,00%	66,67%					60,00%		53,33%			
15		40,00%					40,00%	45,00%				
16	46,67%							45,00%				
17										40,00%	63,33%	
18				40,00%						40,00%	43,33%	
19										40,00%		
20												
21	40,00%		40,00%	40,00%					40,00%			
22												
23												
24						40,00%						

Рис.8. Фрагмент матриці парних порівнянь кластерів із нижнім порогом 40%.

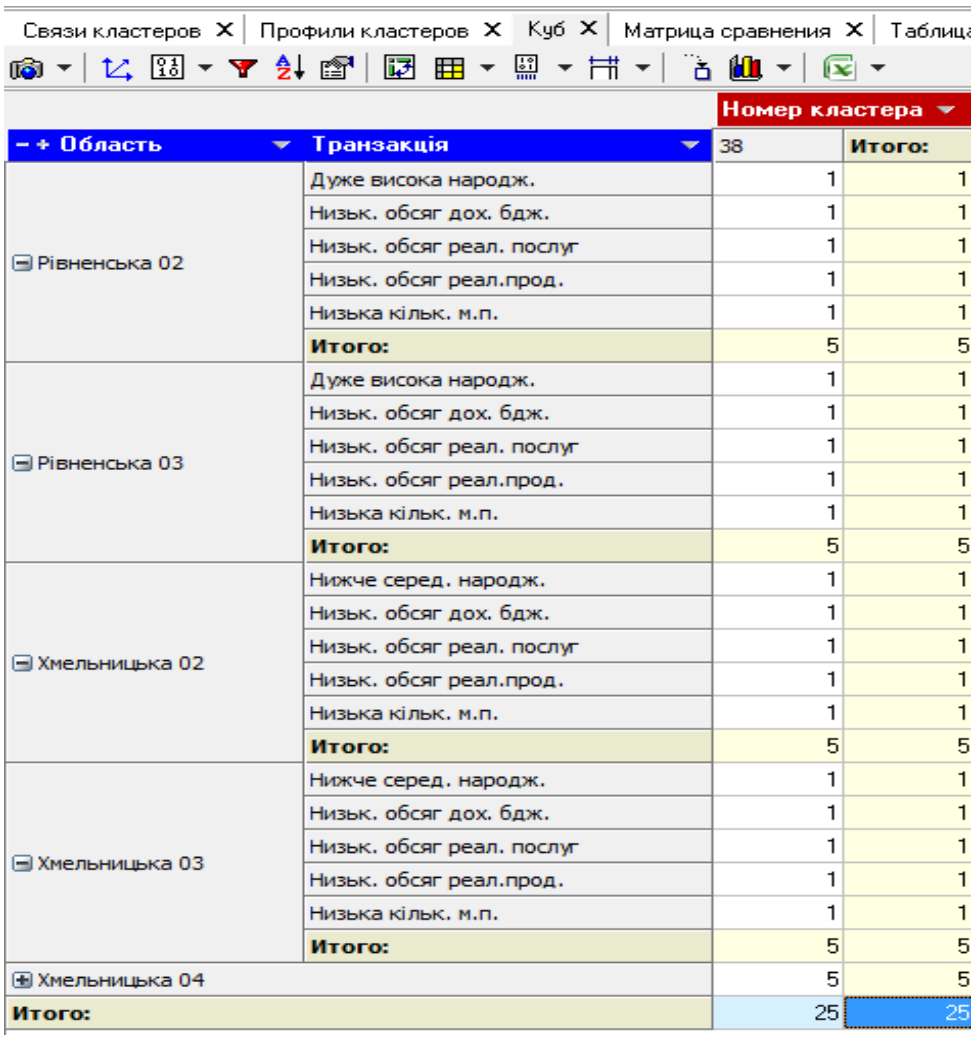


№	Номер кл	Количество транза	Ширина класт	Мощность класт	Целевая функция Profit
0	38	5	6	25	1,18E+00
1	48	5	7	25	7,94E-01
2	6	4	5	20	1,22E+00
3	7	4	7	20	5,08E-01
4	8	4	7	20	5,08E-01
5	41	3	5	15	6,85E-01
6	1	3	6	15	4,27E-01
7	10	3	6	15	4,27E-01
8	13	3	6	15	4,27E-01
9	14	3	6	15	4,27E-01
10	11	3	7	15	2,86E-01
11	16	3	7	15	2,86E-01
12	24	3	7	15	2,86E-01
13	26	3	7	15	2,86E-01
14	28	2	5	10	3,05E-01
15	15	2	6	10	1,9E-01
16	17	2	6	10	1,9E-01
17	18	2	6	10	1,9E-01

Рис. 9. Фрагмент таблиці профілів кластерів транзакцій.

Ширина 38-го кластера 6 означає що 6 різних елементів використовується у його 5 транзакціях (список елементів та скільки разів вони використовуються надано у таблиці праворуч). Потужність 38-го кластера 25 визначає суму кількості входжень кожного елементу в кластер. Це значення можна обчислити також із таблиці праворуч: 5+5+5+5+3+2=25. Останній стовпчик таблиці містить значення цільової функції Profit (або функції вартості) що підлягає максимізації під час роботи алгоритму. Данні в таблиці можна відсортувати також за цим параметром

(для цього треба натиснути двічі мишкою на назві самого стовпчика). Слід відзначити, що максимальні значення функції приходяться на перші 15 кластерів. Далі, за допомогою куба можна проаналізувати зміст кожного окремого кластера. Наприклад, транзакції що входять до складу 38 кластеру показані на Рис. 10. Ці результати виявляють приховані закономірності про схожість соціально-економічних та демографічних процесів за обраними 5 показниками у Рівенський обл. у 2002-2003 р.р., Хмельницькій обл. у 2003-2005 р.р.



Область	Транзакція	Номер кластера 38	Итого:
Рівненська 02	Дуже висока народж.	1	1
	Низьк. обсяг дох. бдж.	1	1
	Низьк. обсяг реал. послуг	1	1
	Низьк. обсяг реал.прод.	1	1
	Низька кільк. м.п.	1	1
	Итого:	5	5
Рівненська 03	Дуже висока народж.	1	1
	Низьк. обсяг дох. бдж.	1	1
	Низьк. обсяг реал. послуг	1	1
	Низьк. обсяг реал.прод.	1	1
	Низька кільк. м.п.	1	1
	Итого:	5	5
Хмельницька 02	Нижче серед. народж.	1	1
	Низьк. обсяг дох. бдж.	1	1
	Низьк. обсяг реал. послуг	1	1
	Низьк. обсяг реал.прод.	1	1
	Низька кільк. м.п.	1	1
	Итого:	5	5
Хмельницька 03	Нижче серед. народж.	1	1
	Низьк. обсяг дох. бдж.	1	1
	Низьк. обсяг реал. послуг	1	1
	Низьк. обсяг реал.прод.	1	1
	Низька кільк. м.п.	1	1
	Итого:	5	5
Хмельницька 04		5	5
Итого:		25	25

Рис.10. Зріз куба для 38 кластера.

З іншого боку, побудова кластерів транзакцій із великим значенням коефіцієнта відштовхування 2,6 (задано на Рис.3) унеможливлює виявлення схожих процесів для більшості транзакцій, оскільки їх переважна більшість потрапила у малопотужні кластери із кількістю транзакцій від 1 до 2. Тому, у даному випадку доцільно розглядати проведену кластеризацію як перший крок аналізу.

На наступному кроці повторимо усі етапи алгоритму, що представлені на Рис.2 – 7 тільки із зменшеним значенням коефіцієнта відштовхування 1,4 (обираємо число із випадаючого списку). Профілі 13 побудованих кластерів представлені на Рис.11. Кількість кластерів зменшилась у 4 рази у порівнянні із попередньою кластеризацією. Хоча ширина кластерів значно збільшилась, матриця парних порівнянь кластерів із нижнім порогом відсічі 40% (Рис.12) показує кращій результат у порівнянні із таблицею на Рис.8. Це означає що зменшення коефіцієнта відштовхування привело до покращення результатів кластнризації у даному прикладі.

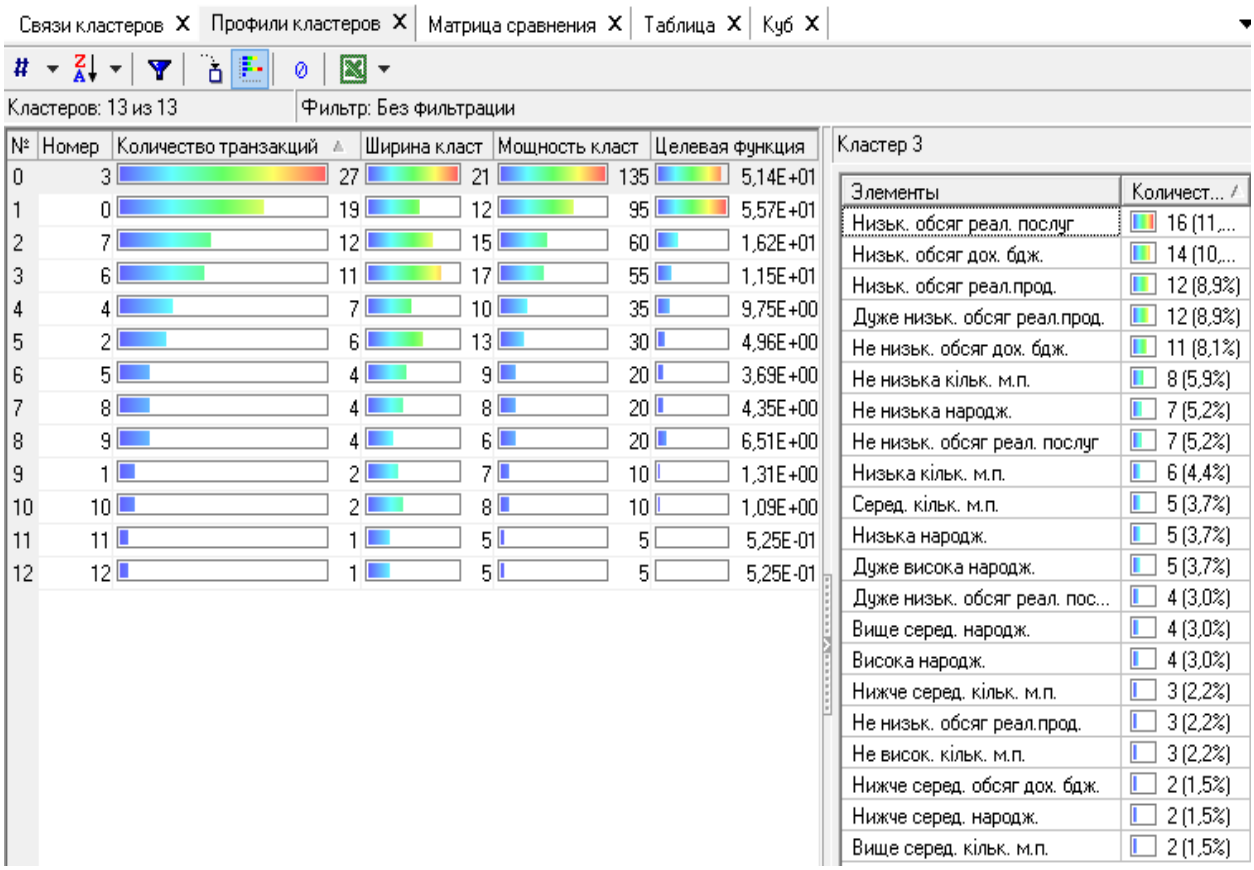


Рис. 11. Таблиця профілів кластерів транзакцій.

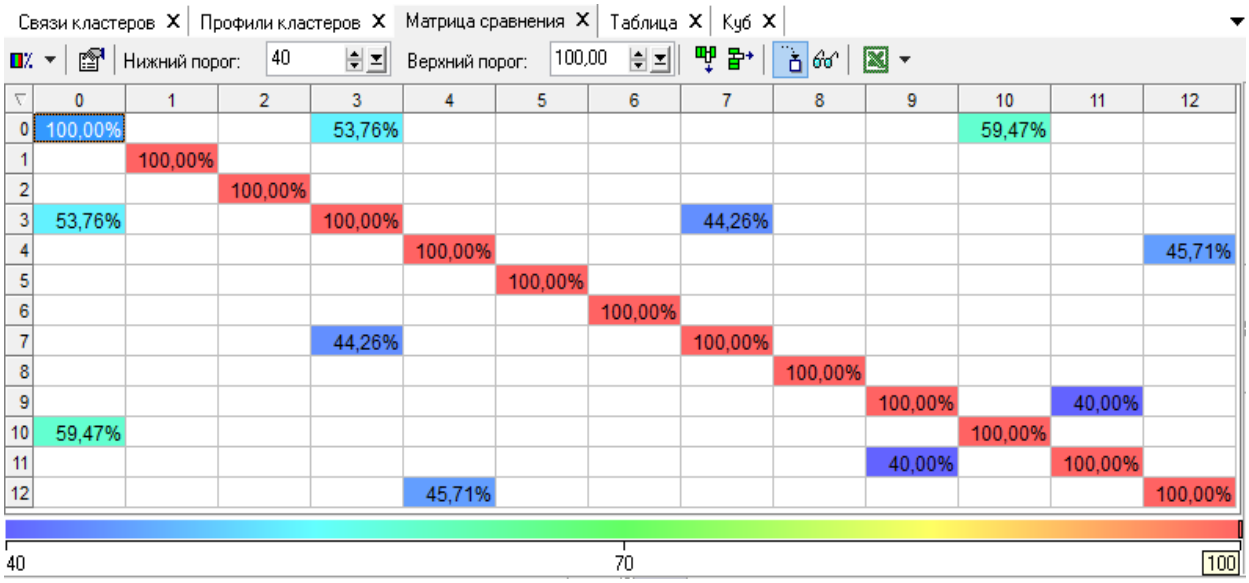


Рис.12. Матриця парних порівнянь кластерів із нижнім порогом 40%.

Зменшення числа кластерів також дозволяє зробити висновки про існування спільних або близьких властивостей у більшій групі транзакцій, у даному випадку областей України в різні роки. На Рис.13 показано зріз OLAP-кубу для найбільшого 3-го кластера. Можна зробити висновки про схожість процесів у Волинській, Закарпатській, Кіровоградській, Черкаській областях за увесь період 2001-2004 р.р. та схожість з іншими областями за певні роки. На Рис.14 показані зв'язки між кластерами (колір відповідає потужності зв'язку). Схожі кластери розфарбовані близькими кольорами (згідно кольорової шкали знизу вікна).

Связи кластеров | Профили кластеров | Матрица сравнения | Куб | Таблица

Область / Транзакція	Номер кластера: 3	Итого:
Волинська 01	5	5
Волинська 02	5	5
Волинська 03	5	5
Волинська 04	5	5
Закарпатська 01	5	5
Закарпатська 02	5	5
Закарпатська 03	5	5
Закарпатська 04	5	5
Івано-Франківська 01	5	5
Київська 01	5	5
Київська 03	5	5
Кіровоградська 01	5	5
Кіровоградська 02	5	5
Кіровоградська 03	5	5
Кіровоградська 04	5	5
Рівненська 04	5	5
Сумська 01	5	5
Херсонська 02	5	5
Херсонська 04	5	5
Черкаська 01	5	5
Черкаська 02	5	5
Черкаська 03	5	5
Черкаська 04	5	5
Чернівецька 02	5	5
Чернівецька 03	5	5
Чернівецька 04	5	5
Чернігівська 02	5	5

Рис.13. Зміст найбільшого 3-го кластера.

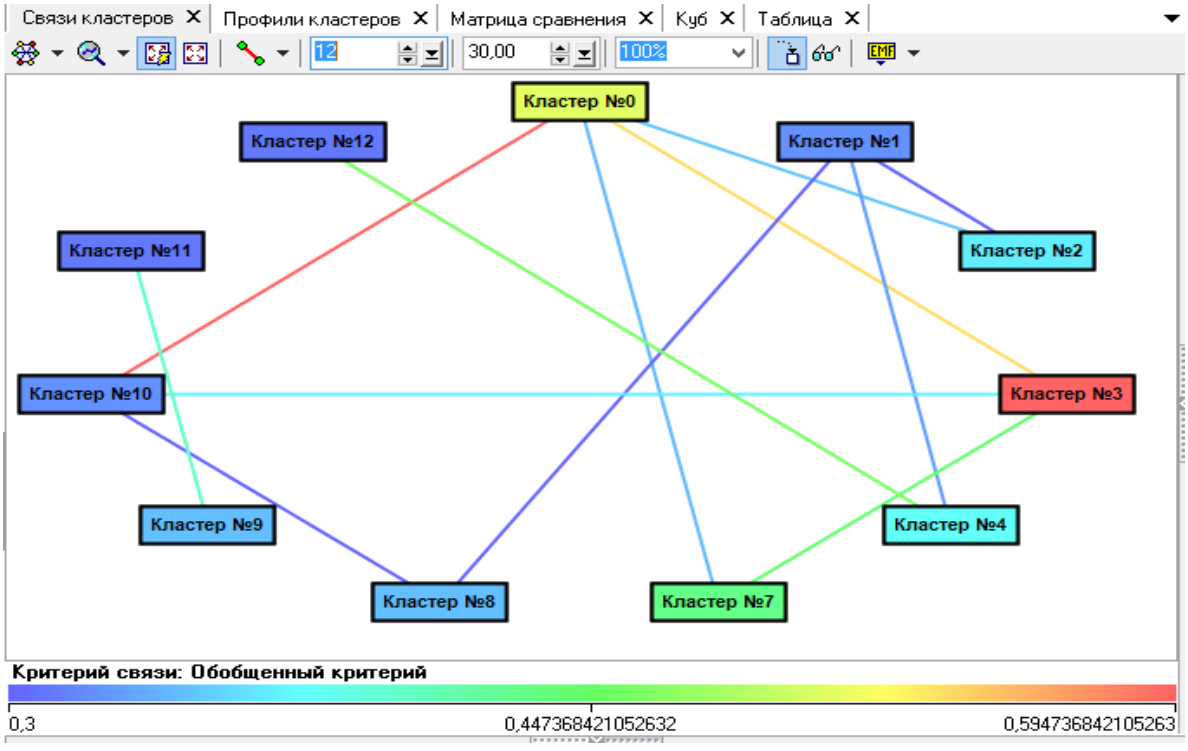


Рис.14. Зв’язки між кластерами (схема “Сеть”).

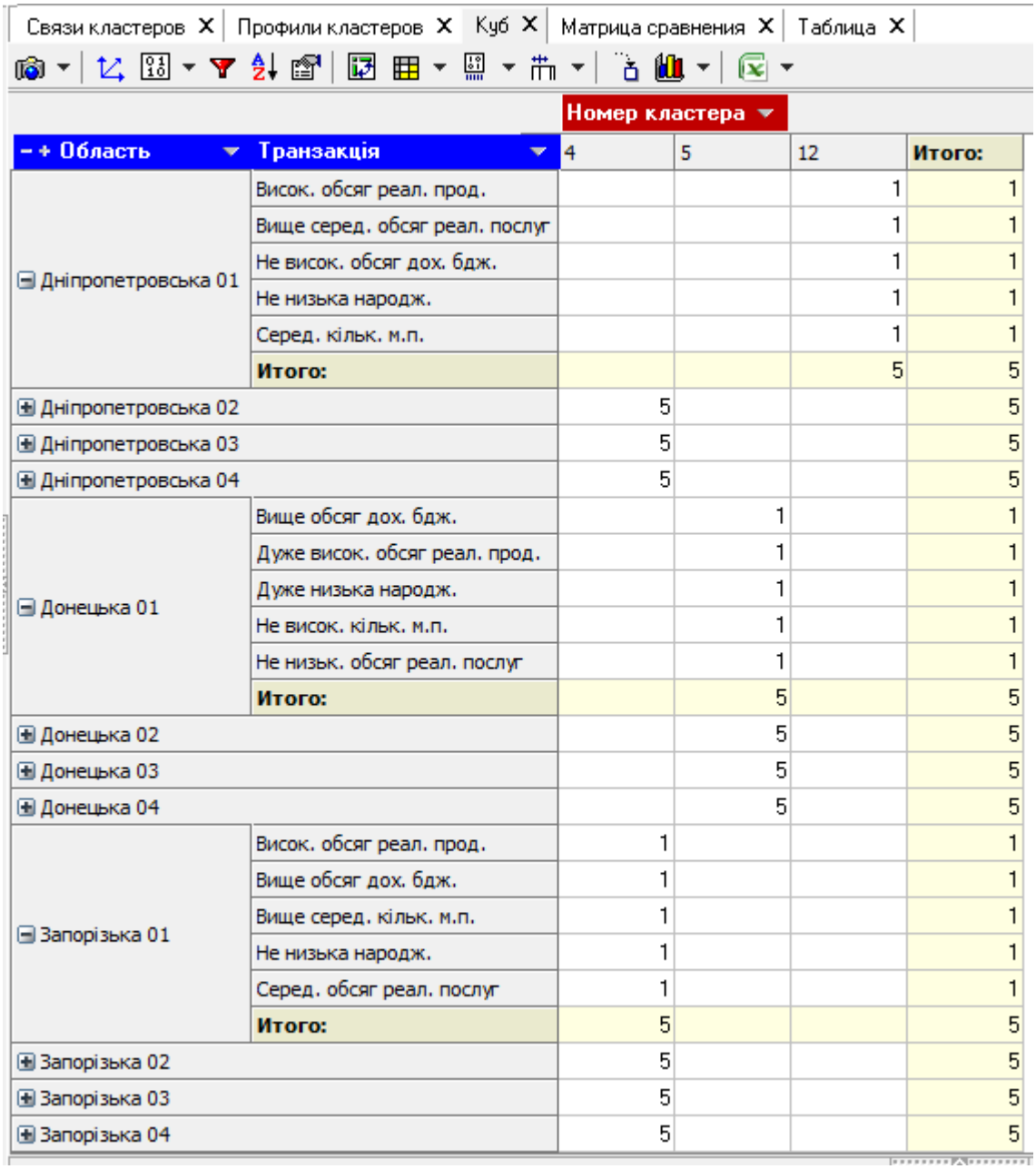
Згідно порівняльній матриці на Рис.12 та графу зв'язків між кластерами на Рис.14 можна проаналізувати окремо об'єднання (групи) кластерів: 0-3-7-10, 1-2, 4-5-12, 6-8, 9-11. Так, наприклад, дуже показовими є групи 1-2, 4-5-12 зміст яких показаний на Рис.15, 16. Якщо навіть елементи різних транзакцій не співпадають, аналітик може виділити групи областей із близькими значеннями транзакцій (див. Таблицю 1). Наприклад, АР Крим та Одеська обл. (Рис.15), Дніпропетровська, Донецька та Запорізька обл. (Рис.16).

Связи кластеров | Профили кластеров | Куб | Матрица сравнения

		Номер кластера		
Область	Транзакція	1	2	Итого:
АР Крим 01	Висок. кільк. м.п.	1		1
	Дуже висок. обсяг дох. бдж.	1		1
	Дуже висок. обсяг реал. послуг	1		1
	Не низька народж.	1		1
	Низьк. обсяг реал.прод.	1		1
	Итого:	5		5
АР Крим 02		5		5
АР Крим 03			5	5
АР Крим 04			5	5
Одеська 01	Висок. обсяг реал. послуг		1	1
	Вище обсяг дох. бдж.		1	1
	Нижче серед. народж.		1	1
	Низьк. обсяг реал.прод.		1	1
	Серед. кільк. м.п.		1	1
	Итого:		5	5
Одеська 02			5	5
Одеська 03			5	5
Одеська 04			5	5
Итого:		10	30	40

Рис.15. Зріз OLAP-куба по кластерам групи 1-2.

Практична значущість методу кластеризації транзакцій для даного прикладу полягає у виявленні схожих (або принципово різних) властивостей та близьких (або не схожих) процесах що мають місце в різних суб'єктах аналізу (у даному випадку – областей України в різні роки). Виявлення прихованих закономірностей дозволяє зробити висновки про ефективність керування процесами у суб'єктах складної системи та звернути увагу на зовнішні неконтрольовані фактори що впливають негативно (позитивно) на зміну властивостей суб'єктів аналізу.



		Номер кластера			
Область	Транзакція	4	5	12	Итого:
Дніпропетровська 01	Висок. обсяг реал. прод.			1	1
	Вище серед. обсяг реал. послуг			1	1
	Не висок. обсяг дох. бдж.			1	1
	Не низька народж.			1	1
	Серед. кільк. м.п.			1	1
	Итого:			5	5
Дніпропетровська 02		5			5
Дніпропетровська 03		5			5
Дніпропетровська 04		5			5
Донецька 01	Вище обсяг дох. бдж.		1		1
	Дуже висок. обсяг реал. прод.		1		1
	Дуже низька народж.		1		1
	Не висок. кільк. м.п.		1		1
	Не низьк. обсяг реал. послуг		1		1
	Итого:		5		5
Донецька 02			5		5
Донецька 03			5		5
Донецька 04			5		5
Запорізька 01	Висок. обсяг реал. прод.	1			1
	Вище обсяг дох. бдж.	1			1
	Вище серед. кільк. м.п.	1			1
	Не низька народж.	1			1
	Серед. обсяг реал. послуг	1			1
	Итого:	5			5
Запорізька 02		5			5
Запорізька 03		5			5
Запорізька 04		5			5

Рис.16. Зріз OLAP-куба по кластерам групи 4-5-12.

2. Побудова асоціативних правил для транзакцій

Асоціативні правила представляють собою правила продукцій стандартного вигляду “Якщо A То B”, де A и B - вирази логічного типу [10] що побудовані на основі елементів транзакцій (наприклад, термів лінгвістичних змінних) [15], [16]. Головна ідея побудови асоціативних правил – це виявлення прихованих логічних зв’язків між елементами транзакцій. Якщо розглядати приклад із попереднього пункту, то асоціативне правило може мати вигляд:

Якщо “Дуже низька народж.” Та “Низьк.обсяг реал.прод.” ТО “Дуже низька кільк. м.п.”

Основною характеристикою набору елементів (навіть якщо він складається з одного елементу) в транзакціях є “**підтримка набору елементів**” S (*support*). S визначається як відсоток транзакцій із усієї вибірки що одночасно містять усі елементи даного набору.

Наприклад, якщо із усіх 25 областей за усі 4 роки (100 транзакцій) тільки три транзакції Тернопільська 01, Тернопільська 02 та Сумська 02 містять одночасно “Дуже низька народж.”, “Низьк. обсяг реал.прод.”, то підтримка даного набору елементів у вибірці складає s= 3%. Підтримку також називають забезпеченням набору. Асоціативні правила будуються тільки при наявності у вибірці транзакцій наборів мінімум з двох елементів ($S>0$).

Асоціативне правило “Якщо A То B” сформоване за допомогою вибірки транзакцій має **підтримку** S (*support*), якщо S% транзакцій містять елементи із виразу A та виразу B одночасно [1], [15], [16].

Наприклад, якщо із усіх 25 областей за усі 4 роки (100 транзакцій) тільки Тернопільська 01 та Сумська 02 містять одночасно “Дуже низька народж.”, “Низьк. обсяг реал.прод.”, “Дуже низька кільк. м.п.” то підтримка наведеного вище асоціативного правила $S = 2\%$.

Вірогідність правила - це міра його істинності. Правило “Якщо A То B” істинне з вірогідністю c (*confidence*), якщо c% транзакцій із множини, що містить групу елементів виразу A, містять й групу елементів виразу B [1], [15], [16].

Наприклад, якщо 20% транзакцій містять набір елементів умови та наслідку правила “Дуже низька народж.”, “Низьк. обсяг реал. прод.”, “Дуже низька кільк. м.п.” (підтримка асоціативного правила $S = 20\%$) та 40% транзакцій вибірки містять набір елементів умови правила “Дуже низька народж.”, “Низьк. обсяг реал. прод.” (підтримка набору елементів умови правила $S_A = 40\%$), то вірогідність наведеного вище асоціативного правила $C = (S/S_A)100\% = 50\%$.

Вочевидь, якщо підтримка та вірогідність асоціативного правила досить високі то можна очікувати що у майбутній транзакції за наявності тієї ж самої умови буде присутній такий саме наслідок. Крім об’єктивних характеристик наборів елементів та асоціативних правил існує ще суб’єктивні характеристики, що основані на даних об’єктивних характеристиках.

Ліфт (від interest lift – підвищення зацікавленості) L асоціативного правила обчислюється як відношення вірогідності правила C до підтримки набору елементів із наслідку правила S_B: $L = C/S_B = S/(S_A S_B)$ [15], [16]. Чим більш значення ліфта тим частіше наслідок визначається умовою правила, ніж у випадках коли умова відсутня. Тобто, якщо $L > 1$, то умова з’являється більш часто у транзакціях що містить наслідок ніж у транзакціях без нього. Ліфт є важливою, але другорядною характеристикою правил, після підтримки та вірогідності.

Наприклад, якщо 20% транзакцій містять набір елементів умови та наслідку правила “Дуже низька народж.”, “Низьк. обсяг реал. прод.”, “Дуже низька кільк. м.п.” (підтримка асоціативного правила $S = 20\%$), 40% транзакцій вибірки містять набір елементів умови правила “Дуже низька народж.”, “Низьк. обсяг реал. прод.” (підтримка набору елементів умови правила $S_A = 40\%$) та 50% транзакцій вибірки містять набір елементів наслідку правила “Дуже низька кільк. м.п.” (підтримка набору елементів наслідку правила $S_B = 50\%$) то ліфт наведеного вище асоціативного правила $L = S/(S_A S_B) = 0{,}01$, що означає досить слабкий зв’язок між умовою правила та його наслідком. **Впевненість** (conviction) асоціативного правила визначається як відношення очікуваної частоти появи умови A без наслідку B у транзакціях

(тобто, частота того що правило “Якщо A То B” не коректне) до частоти не вірогідності самого правила “Якщо A То B ” [15], [16] та обчислюється як $\mathrm{Conv} = \frac{1 - S_B/100}{1 - C/100}$. Чим більш цей показник тим вище значущість правила.

Алгоритми пошуку асоціативних правил призначені для знаходження всіх правил виду “Якщо A То B”, причому підтримка і достовірність цих правил повинні знаходитися в рамках деяких наперед заданих меж, які називаються відповідно мінімальною і максимальною підтримкою і мінімальної і максимальною достовірністю.

Для побудови моделі асоціативних правил для розглянутого у попередньому пункті прикладу встановлюємо курсор на вузлі імпортованої із файлу task8.txt вибірки транзакцій, та після запуску майстру обробки обираємо асоціативні правила. Налаштування стовпчиків транзакцій таке саме як на Рис.2. На наступному кроці налаштовуємо параметри побудови асоціативних правил: мінімальну і максимальну підтримку, мінімальну і максимальну вірогідність та максимальну потужність множини (Рис.17).

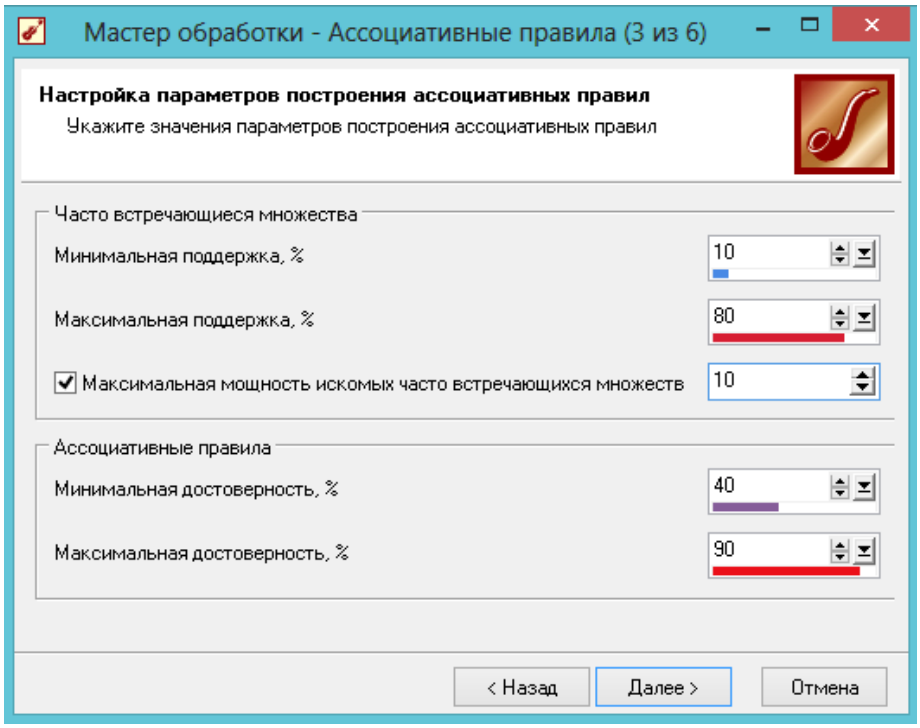


Рис.17. Налаштування параметрів побудови асоціативних правил.

Алгоритми побудови асоціативних правил використовують тільки множини елементів транзакцій що мають підтримку із вказаного користувачем діапазону мінімальної та максимальної підтримки та залишають в базі асоціативні правила, що мають вірогідність із вказаного діапазону значень від мінімальної до максимальної. Тобто межі діапазонів підтримки і вірогідності обмежують кількість використаних множин елементів та, відповідно, побудованих правил. Якщо підтримка множини елементів має велике значення, то алгоритм знаходить добре відомі та очевидні правила, що знецінює результати аналіз. У той же час, низьке значення підтримки множини елементів приводить до побудови занадто великої кількості асоціативних правил, що з одного боку, вимагає суттє-

вого збільшення обчислювальних ресурсів, з іншого - робить аналіз транзакцій набагато менш зручним та зрозумілим для аналітика. Але добре відомий факт, що більшість корисних правил, що виявляють нові закономірності, знаходиться саме при низькому значенні мінімального порогу підтримки множин елементів транзакцій. З іншого боку, занадто низьке значення мінімального порогу підтримки веде до побудови не значущих асоціативних правил. Таким чином, необхідно знайти оптимальні значення меж діапазонів підтримки і вірогідності, які, з одного боку, забезпечують виявлення нових закономірностей на основі побудови нових корисних правил, з іншого боку, забезпечують значущість асоціативних правил. Вибір оптимальних значень параметрів методи можливий за рахунок серії випробувань та аналізу результатів побудови асоціативних правил.

Спочатку встановимо максимальні значення меж для діапазонів підтримки 10% та 80%, та вірогідності 40%, 90% (Рис.17). Запускаємо алгоритм на виконання та отримуємо 19 правил (Рис.18). Отримуємо також гістограму потужності множин елементів, що часто зустрічаються в вибірці.

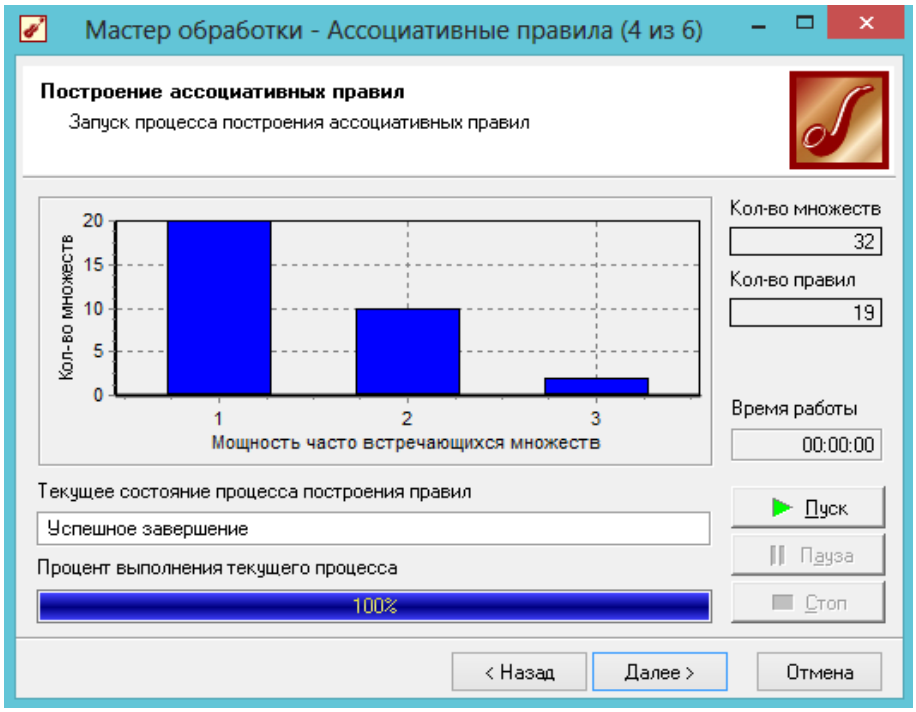


Рис.18. Запуск процесу побудови асоціативних правил.

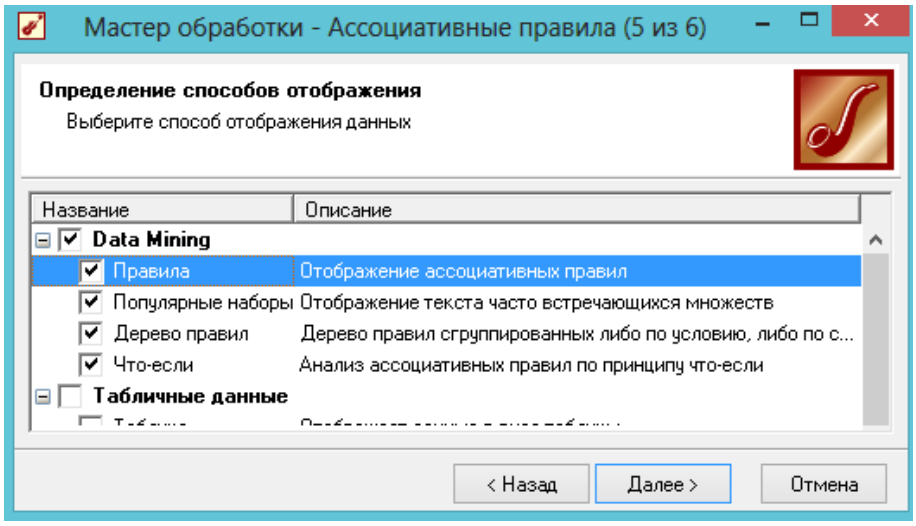


Рис.19. Налаштування засобів візуалізації.

Після налаштувань засобів візуалізації (Рис.19) отримуємо результат у вигляді таблиці асоціативних правил. Обираємо “Порядок сортировки” – по “Поддержке” та по спаданню значень (кнопки меню  , Рис.20). Побудовані асоціативні правила, що мають одночасно підтримку S>10%, достовірність C>40% та ліфт L>1 містять достатньо очевидні та не корисні знання про зв’язки низьких значень соціально - економічних та демографічних показників. Наприклад, якщо в регіоні відносно низький обсяг реалізованої продукції, то отримуємо відносно низький обсяг доходів у місцевий бюджет.

Правила | Популярные наборы | Дерево правил | Что-если

Правил: 19 из 19 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддерж		Достоверность	Лифт
				Кол-во	%		
1	9	Низьк. обсяг реал.прод.	Низьк. обсяг дох. бдж.	25	25,00	54,35	1,553
2	8	Низьк. обсяг дох. бдж.	Низьк. обсяг реал.прод.	25	25,00	71,43	1,553
3	12	Низьк. обсяг реал.прод.	Низьк. обсяг реал. послуг	19	19,00	41,30	1,180
4	11	Низьк. обсяг реал. послуг	Низьк. обсяг реал.прод.	19	19,00	54,29	1,180
5	7	Низьк. обсяг реал. послуг	Низьк. обсяг дох. бдж.	16	16,00	45,71	1,306
6	6	Низьк. обсяг дох. бдж.	Низьк. обсяг реал. послуг	16	16,00	45,71	1,306
7	13	Низька кільк. м.п.	Низьк. обсяг реал.прод.	15	15,00	75,00	1,630
8	5	Нижче серед. народж.	Низьк. обсяг реал.прод.	15	15,00	68,18	1,482
9	2	Дуже низьк. обсяг реал. послуг	Низьк. обсяг реал.прод.	13	13,00	56,52	1,229
10	19	Низьк. обсяг реал. послуг Низьк. обсяг реал.прод.	Низьк. обсяг дох. бдж.	12	12,00	63,16	1,805
11	18	Низьк. обсяг дох. бдж. Низьк. обсяг реал.прод.	Низьк. обсяг реал. послуг	12	12,00	48,00	1,371
12	17	Низьк. обсяг дох. бдж. Низьк. обсяг реал. послуг	Низьк. обсяг реал.прод.	12	12,00	75,00	1,630
13	4	Нижче серед. народж.	Низьк. обсяг дох. бдж.	12	12,00	54,55	1,558
14	1	Дуже низьк. обсяг реал. послуг	Низьк. обсяг дох. бдж.	12	12,00	52,17	1,491
15	16	Нижче серед. народж.	Низьк. обсяг дох. бдж. Низьк. обсяг реал.прод.	11	11,00	50,00	2,000
16	15	Низьк. обсяг дох. бдж. Низьк. обсяг реал.прод.	Нижче серед. народж.	11	11,00	44,00	2,000
17	14	Нижче серед. народж. Низьк. обсяг реал.прод.	Низьк. обсяг дох. бдж.	11	11,00	73,33	2,095
18	10	Низька кільк. м.п.	Низьк. обсяг дох. бдж.	11	11,00	55,00	1,571
19	3	Не низьк. обсяг дох. бдж.	Низьк. обсяг реал. послуг	11	11,00	68,75	1,964

Рис.20. Набір правил асоціацій.

На Рис.21 показані популярні набори елементів, які найбільш часто зустрічаються в транзакціях одночасно. Для аналізу цих наборів необхідно зробити сортування таблиці (спадаюче) по підтримці (кнопки меню ). Переважна більшість наборів містить елементи з низькими значеннями соціально-економічних та демографічних показників, що пояснює закономірності формування бази асоціативних правил.

Правила | Популярные наборы | Дерево правил | Что-если

Множеств: 32 из 32 | Фильтр: Без фильтрации

№	Номер множества	Элементы	Поддержка		Мощность
			Кол-во	%	
1	16	Низьк. обсяг реал.прод.	46	46,00	1
2	15	Низьк. обсяг реал. послуг	35	35,00	1
3	14	Низьк. обсяг дох. бдж.	35	35,00	1
4	27	Низьк. обсяг дох. бдж. Низьк. обсяг реал.прод.	25	25,00	2
5	2	Дуже низьк. обсяг реал. послуг	23	23,00	1
6	12	Нижче серед. народж.	22	22,00	1
7	17	Низька кільк. м.п.	20	20,00	1
8	29	Низьк. обсяг реал. послуг Низьк. обсяг реал.прод.	19	19,00	2
9	10	Не низька народж.	18	18,00	1
10	3	Дуже низьк. обсяг реал.прод.	18	18,00	1
11	19	Серед. кільк. м.п.	17	17,00	1
12	9	Не низька кільк. м.п.	17	17,00	1
13	26	Низьк. обсяг дох. бдж. Низьк. обсяг реал. послуг	16	16,00	2
14	8	Не низьк. обсяг реал.прод.	16	16,00	1
15	7	Не низьк. обсяг реал. послуг	16	16,00	1
16	6	Не низьк. обсяг дох. бдж.	16	16,00	1
17	30	Низьк. обсяг реал.прод. Низька кільк. м.п.	15	15,00	2
18	25	Нижче серед. народж. Низьк. обсяг реал.прод.	15	15,00	2
19	18	Низька народж.	14	14,00	1

Рис.21. Популярні набори елементів транзакцій.

Правила | Популярные наборы | Дерево правил | Что-если

- Ассоциативные правила (по следствию)
  - Низьк. обсяг дох. бдж. (35,00%; 35)
  - Низьк. обсяг реал.прод. (46,00%; 46)
    - Дуже низьк. обсяг реал. послуг (13,00%; 13)
    - Нижче серед. народж. (15,00%; 15)
    - Низьк. обсяг дох. бдж. (25,00%; 25)
    - Низьк. обсяг реал. послуг (19,00%; 19)
    - Низька кільк. м.п. (15,00%; 15)
    - Низьк. обсяг дох. бдж. И Низьк. обсяг реал. послуг (12,00%; 1
  - Низьк. обсяг реал. послуг (35,00%; 35)
  - Нижче серед. народж. (22,00%; 22)
  - Низьк. обсяг дох. бдж. И Низьк. обсяг реал.прод. (25,00%; 25)

Количество правил: 6; Следствие: Низьк. обсяг реал.прод.

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
Дуже низьк. обсяг реал. послуг	13	13,00	56,50	1,229
Нижче серед. народж.	15	15,00	68,20	1,482
Низьк. обсяг дох. бдж.	25	25,00	71,40	1,553
Низьк. обсяг реал. послуг	19	19,00	54,30	1,18
Низька кільк. м.п.	15	15,00	75,00	1,63
Низьк. обсяг дох. бдж. И Низь...	12	12,00	75,00	1,63

Рис.22. Дерева асоціативних правил (за наслідками).

Побудовані 19 асоціативних правил представлені у дереві правил за наслідками. Тобто усі правила розбиті на 5 груп за однаковим наслідком (висновком). На Рис.22 показано детально інформацію про 6 правил що мають єдиний наслідок “Низьк. Обсяг реал. прод”. Усі 6 правил мають підтримку від 12% до 25 % та достовірність від 54% до 75%. Але, як було вже відзначено вище даний набір правил не виявляє нових закономірностей у вибірці транзакцій.

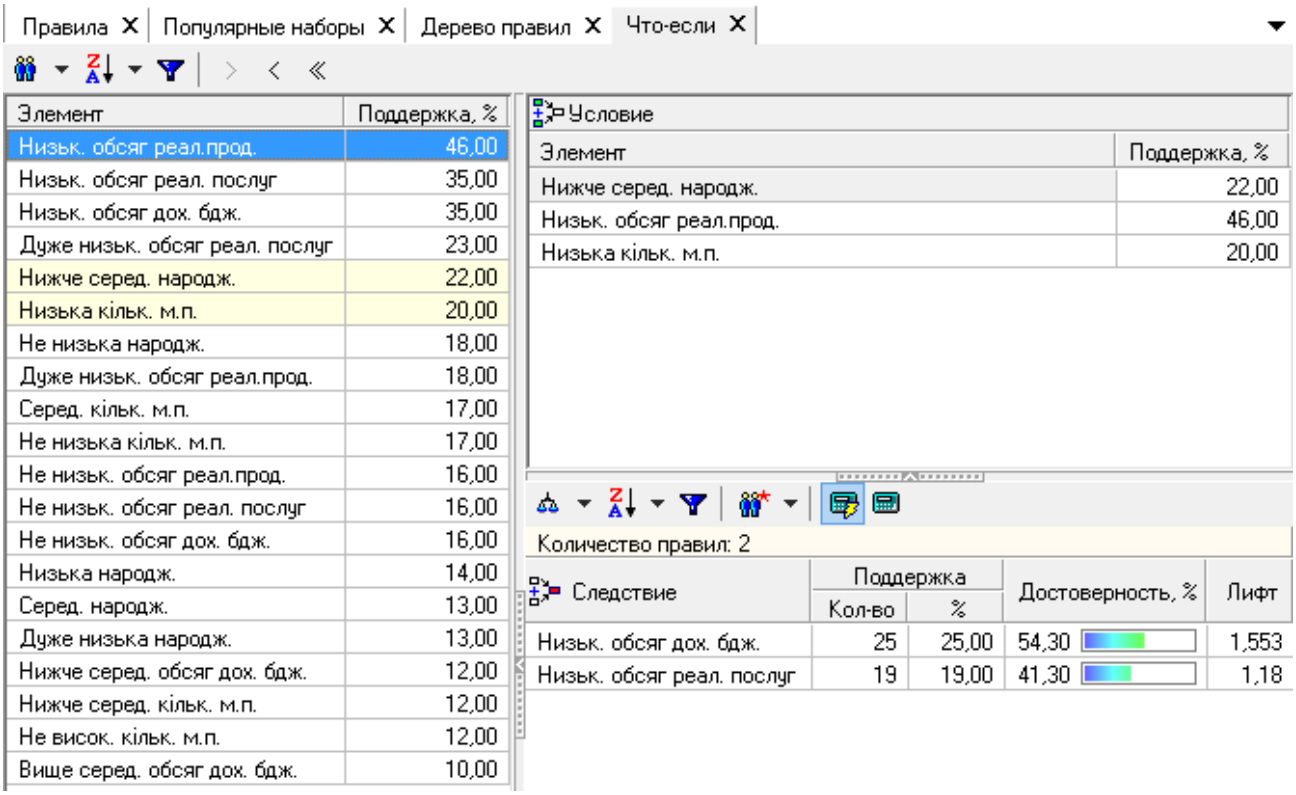


Рис.23. Візуалізатор “Что-если”.

За допомогою візуалізатора “Что-если” (Рис.23) можна сформувати умову правила (вікно праворуч, зверху) обираючи мишкою елементи із списку (вікно ліворуч) орієнтуючись на показник підтримки кожного елементу та отримати наслідок (заключення) асоціативного правила натиснувши кнопку . Знову (Рис.23) ми отримуємо правило із відносно низькими значеннями соціально-економічних та демографічних показників. На наступному етапі зменшуємо мінімальну та максимальну підтримку множин елементів та запускаємо процес побудови правил (Рис.24).

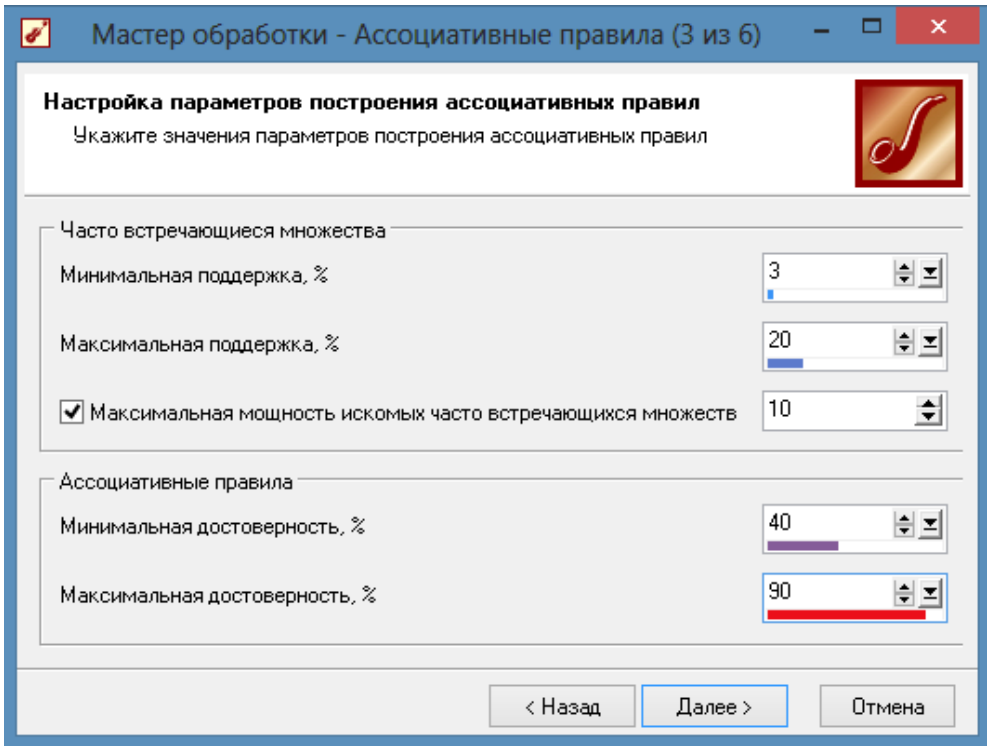


Рис.24. Налаштування параметрів побудови асоціативних правил.

Отримуємо 128 правил, що у 7 разів більше (!) ніж у попередньому випадку (Рис. 25). Фрагмент таблиці з правилами представлений на Рис.26.

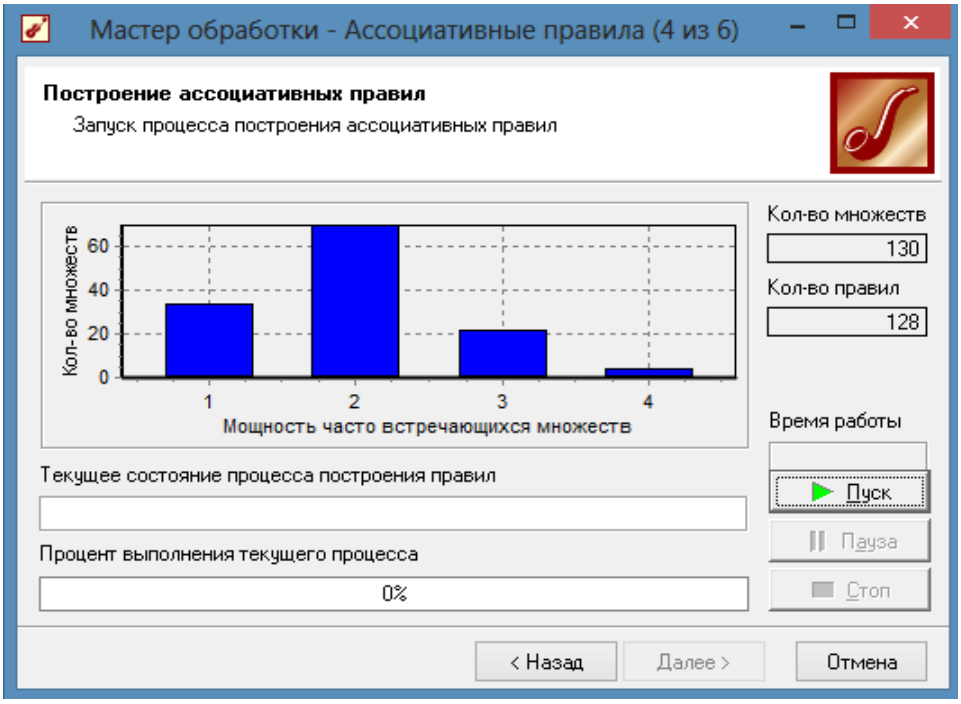


Рис.25. Запуск процесу побудови асоціативних правил.

Правила | Популярные наборы | Дерево правил | Что-если

Правил: 128 из 128 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	12	Висок. обсяг реал. прод.	Не низька народж.	7	7,00	87,50	4,861
2	15	Серед. обсяг реал. послуг	Висок. обсяг реал. прод.	6	6,00	75,00	9,375
3	14	Висок. обсяг реал. прод.	Серед. обсяг реал. послуг	6	6,00	75,00	9,375
4	65	Серед. обсяг реал. послуг	Висок. обсяг реал. прод. Не низька народж.	5	5,00	62,50	8,929
5	64	Висок. обсяг реал. прод.	Не низька народж. Серед. обсяг реал. послуг	5	5,00	62,50	12,500
6	63	Висок. обсяг реал. прод. Серед. обсяг реал. послуг	Не низька народж.	5	5,00	83,33	4,630
7	62	Висок. обсяг реал. прод. Не низька народж.	Серед. обсяг реал. послуг	5	5,00	71,43	8,929
8	39	Серед. обсяг реал. послуг	Не низька народж.	5	5,00	62,50	3,472
9	37	Нижче серед. обсяг дох. бдж.	Не низьк. обсяг реал.прод.	5	5,00	41,67	2,604
10	22	Вище серед.обсяг реал. прод.	Дуже низька народж.	5	5,00	83,33	6,410
11	90	Дуже низьк. обсяг дох. бдж.	Дуже низьк. обсяг реал.прод. Серед. народж.	4	4,00	80,00	20,000
12	89	Дуже низьк. обсяг дох. бдж. Дуже низьк. обсяг реал.прод.	Серед. народж.	4	4,00	80,00	6,154
13	58	Серед. обсяг реал. послуг	Висок. обсяг реал. прод. Дуже висок. обсяг дох. бдж.	4	4,00	50,00	12,500
14	57	Дуже висок. обсяг дох. бдж.	Висок. обсяг реал. прод. Серед. обсяг реал. послуг	4	4,00	80,00	13,333
15	56	Висок. обсяг реал. прод.	Дуже висок. обсяг дох. бдж. Серед. обсяг реал. послуг	4	4,00	50,00	12,500

Рис. 26. Фрагмент таблиці з набором правил асоціацій.

На відміну від набору асоціативних правил отриманих при великих значеннях мінімального та максимального порогів підтримки у даному випадку ми отримуємо набір принципово інших правил із відносними значеннями соціально-економічних та демографічних показників на середньому та високому рівнях. Наприклад, при “Висок. обсяг реал. прод.” Отримуємо “Не низька народж.”, що говорить про позитивний вплив економічних показників на демографічні в регіонах. На Рис. 27 зображено фрагмент таблиці з популярними наборами елементів в транзакціях (відсортованих за значенням підтримки). На відміну від набору даних отриманих у попередньому випадку, тут з’являються елементи з відносними значеннями соціально-економічних та демографічних показників на середньому та високому рівнях.

Правила | Популярные наборы | Дерево правил | Что-если

Множеств: 130 из 130 | Фильтр: Без фильтрации

№	Номер множества	Элементы	Поддерж		Мощность
			Кол-во	%	
1	29	Низька кільк. м.п.	20	20,00	1
2	26	Не низька народж.	18	18,00	1
3	16	Дуже низьк. обсяг реал.прод.	18	18,00	1
4	31	Серед. кільк. м.п.	17	17,00	1
5	25	Не низька кільк. м.п.	17	17,00	1
6	24	Не низьк. обсяг реал.прод.	16	16,00	1
7	23	Не низьк. обсяг реал. послуг	16	16,00	1
8	22	Не низьк. обсяг дох. бдж.	16	16,00	1
9	30	Низька народж.	14	14,00	1
10	32	Серед. народж.	13	13,00	1
11	18	Дуже низька народж.	13	13,00	1
12	28	Нижче серед. обсяг дох. бдж.	12	12,00	1
13	27	Нижче серед. кільк. м.п.	12	12,00	1
14	19	Не висок. кільк. м.п.	12	12,00	1
15	7	Вище серед. обсяг дох. бдж.	10	10,00	1
16	34	Серед. обсяг реал. послуг	8	8,00	1
17	14	Дуже висока народж.	8	8,00	1
18	5	Вище серед. кільк. м.п.	8	8,00	1
19	3	Висок. обсяг реал. прод.	8	8,00	1
20	42	Висок. обсяг реал. прод. Не низька народж.	7	7,00	2
21	33	Серед. обсяг дох. бдж.	7	7,00	1
22	8	Вище серед. обсяг реал. послуг	7	7,00	1
23	6	Вище серед. народж.	7	7,00	1

Рис. 27. Фрагмент списку популярних наборів.

Побудовані 128 асоціативних правил представлені у дереві правил за наслідками. Фрагмент відсортованої за кількістю умов в правилах таблиці дерева правил представлений на Рис. 28.

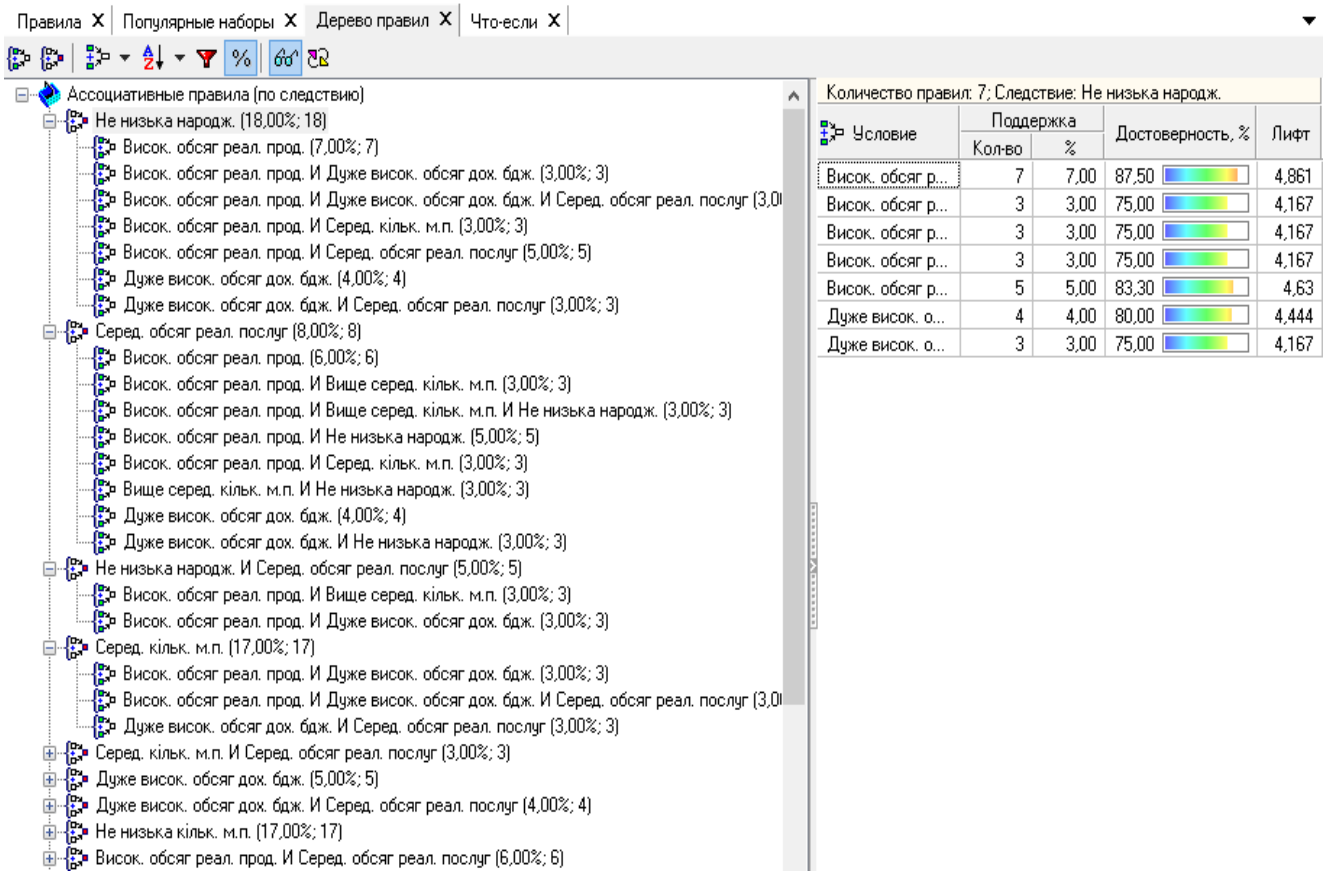


Рис. 28. Фрагмент дерева правил.

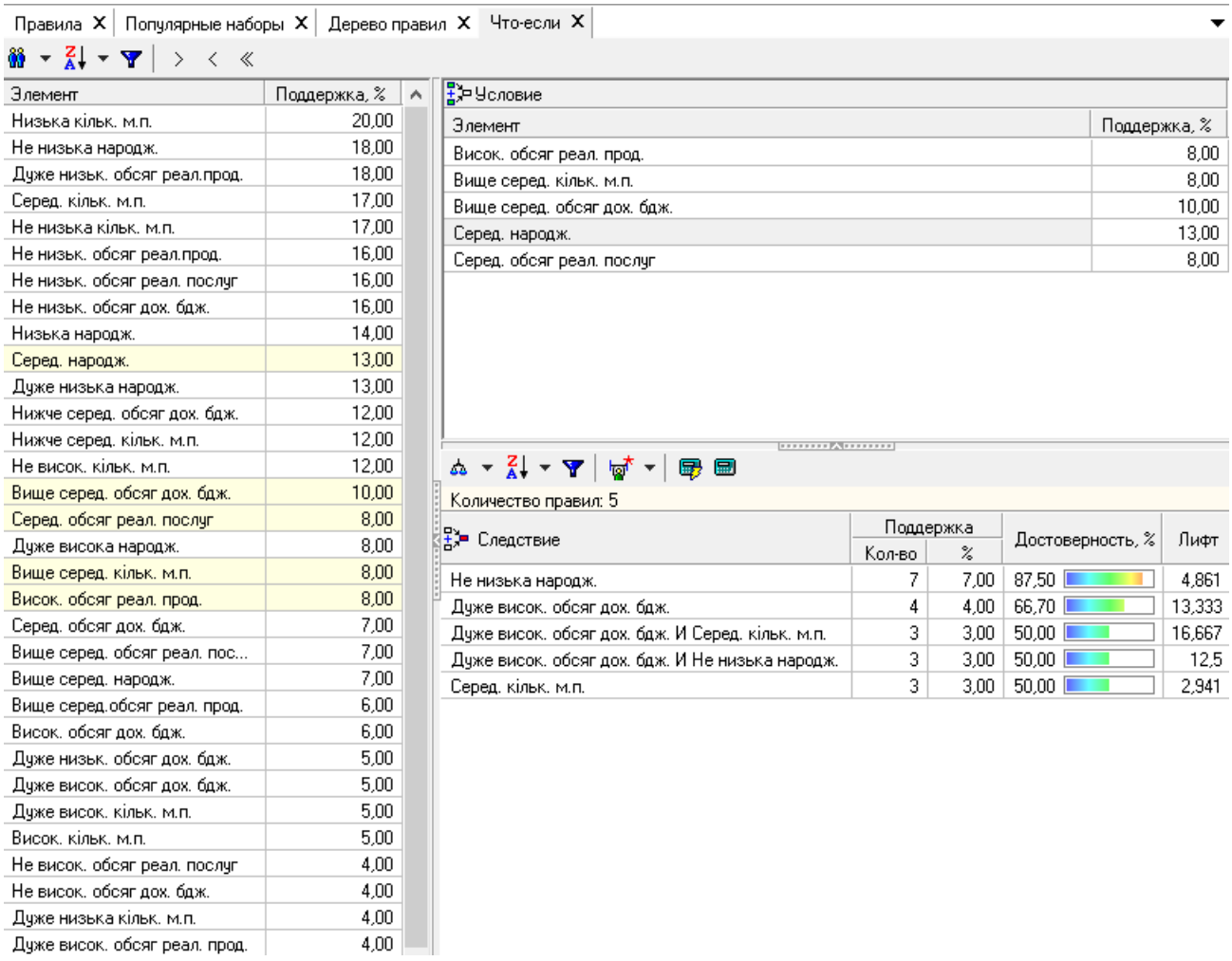


Рис.29. Візуалізатор “Что-если”.

Наприклад, наслідок “Не низька народж.” знаходиться у 7 правилах (вікно праворуч). Підтримка правил досить низька: від S=3% до S=7%. У той же час їх достовірність досить висока: від C=75% до C=87,5% та правила мають високий ліфт від L=4,167% до 4,861%. Отримані із цих 7 правил умови “Не низької народжуваності” в регіонах (Рис.28):

- “Висок. обсяг реал. прод.” ;
- “Висок. обсяг реал. прод.” ТА “Дуже висок. обсяг дох. бдж.”;
- “Висок. обсяг реал. прод.” ТА “Дуже висок. обсяг дох. бдж.” ТА “Серед. обсяг реал. послуг”;
- “Висок. обсяг реал. прод.” ТА “Серед. кільк. м.п.”;
- “Висок. обсяг реал. прод.” ТА “Серед. обсяг реал. послуг”;
- “Дуже висок. обсяг дох. бдж.”;
- “Дуже висок. обсяг дох. бдж.” ТА “Серед. обсяг реал. послуг”.

Отриманні в наведених правилах висновки (знання) достатньо корисні та не очевидні, оскільки в різних країнах існують приклади коли рівень народжуваності з часом знижується на фоні зростання соціально-економічних показників, що приводить в решті до відносно низького або дуже низького рівня народжуваності для відносно середніх та високих показників соціально - економічного розвитку. Слід відзначити, що побудовані асоціативні правила були виявлені у групах транзакцій з низьким рівнем підтримки (3-7%), що відносяться до невеликої групи областей України у 2001-2004 р.р. із відносно високими значеннями соціально-економічних показників. Дані результати узгоджуються із побудованими у попередньому пункті кластерами транзакцій (Рис.15, 16). Таким чином, кластеризація транзакцій та асоціативні правила дозволяють провести детальний аналіз із встановлення причинно-наслідних зв’язків між транзакціями для виявлення нових закономірностей та генерування практично значущих висновків.

Завдання для лабораторної роботи 8.

1. Створити файл Lab8.xlsx. Створити таблицю із 2 демографічними та 3 соціально-економічними показниками для 25 областей України за 2009-2012 р.р. (7 стовпчиків 26 рядків). Назви стовпчиків: Область, Рік, Демографічний 1, Демографічний 2, Соц.-ек.1, Соц.-ек.2, Соц.-ек.3.
2. Перевести дані у безрозмірний формат за кожен рік окремо (нормалізація даних у відрізок [0,1]).
3. Кожен з числових показників замінити на один із термів лінгвістичної змінної із 9 значень (від дуже низького до дуже великого).
4. На новому листі сформувати нову таблицю із транзакціями у два стовпчики. Перший стовпчик з назвою ІД містить елементи із назвою регіону та року, наприклад Тернопільска_09. Другий стовпчик з назвою Транзакція містить зна-

чення термів лінгвістичних змінних по кожному демографічному або соціально-економічному параметру (наприклад, Рис.1).

5. Зберегти файл як Lab8.xlsx та як Lab8.txt.

6. Імпортувати файл Lab8.txt у Deductor studio.

7. Виконати кластеризацію транзакцій та побудувати асоціативні правила для різних параметрів діапазонів підтримки множин елементів транзакцій.

8. Проаналізувати отримані висновки моделі на предмет формулювання нових знань та правил про зв'язок демографічних та соціально-економічних показників країни.

Список літератури

1. Deductor. Руководство аналитика. Версия 5.3. Компания BaseGroup Labs, 2013.
2. Фельдман Л.П., Петренко А.І., Дмитрієва О.А. Чисельні методи в інформатиці. – Київ: BHV, 2004. – 420 с.
3. Jae-On Kim, Charles W. Mueller. Factor Analysis: Statistical Methods and Practical Issues. Sage Publications, Newbury Park, 1978, 82p.
4. Айвазян С.А., Бухштабер В.М., Енюков Е.С., Мешалкин Л.Д. Прикладная статитстика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
5. Шаховська Н.Б., Пасічник В.В. Сховища та простори даних. - Львів, Львівська політехніка, 2009. - 244с.
6. Лупенко С.А. Теоретичні основи моделювання та опрацювання циклічних сигналів в інформаційних системах. – Львів: Магнолія 2006, 2016 – 344с.
7. Капшій О.В., Коваль О.І., Русин Б.П. Вейвлет - перетворення у компресії та попередній обробці зображень. – Львів: Сполом, 2008. – 206с.
8. Новотарський М.А., Нестеренко Б.Б. Штучні нейронні мережі: обчислення. – К.: Інститут Математики НАН України, 2004. – 408с.
9. P.J. Braspenning, F.Thuijsman, A.J.M.M. Weijters. Artificial Neural Networks. Springer, 1995. – 295.
10. Литвин В.В. Бази знань інтелектуальних систем підтримки прийняття рішень. – Львів: Видавництво Львівської політехніки, 2011. – 240с.
11. Wu J. Advances in K-means Clustering: A Data Mining Thinking. Berlin: Springer, 2012. – 180 p.
12. Kohonen T., Honkela T. Kohonen network. Scholarpedia, 2007.
13. Hastie N., Tibshirani R., Friedman J. 8.5 The EM algorithm. In book: The Elements of Statistical Learning. NY: Springer, 2001, pp. 236 – 243.
14. Y. Yang, X. Guan J. You. CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data. http://www.inf.ufrgs.br/~alvares/CMP259DCBD/clope.pdf
15. Tan P.N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. Pearson Education Limited, 2018. – 864 p.
16. https://en.m.wikipedia.org/wiki

Навчальне видання

АКІМЕНКО Віталій Володимирович

Прикладні задачі інтелектуального аналізу даних (DATA MINING)

Київський національний університет імені Тараса Шевченка

Підписано до друку 01.06.2018.
Папір офсетний Формат 60x84 1/16.
Умов. друк. арк. 9,5 Обл. вид. арк. 9,1
Наклад 200 прим.

ІВЦ КНУ імені Тараса Шевченка