Міністерство освіти і науки України

Національний авіаційний університет

ННІІДС

Кафедра прикладної математики

Курсова робота   
 з дисципліни «Аналіз даних»

На тему: «Дослідження вартості курсів з англійської мови »

**Виконав:** студент ІІДС-351

Подлєсний Максим

**Перевірив:**завідувач кафедри

Приставка Пилип Олександрович

Київ 2017

#### Зміст

[Постановка задачі 3](#_Toc482920265)

[Вступ 5](#_Toc482920266)

[Тестування автоматизованої системи 7](#_Toc482920267)

[Двовимірний статистичний аналіз даних 7](#_Toc482920268)

[Багатовимірний статистичний аналіз даних 11](#_Toc482920269)

[Список використаної літератури 17](#_Toc482920275)

# Постановка задачі

Нехай задано Ω5,*N*= {*Xk,l*, *k*=1…5*, l*=1…30}, де:

*  - час доби
* - вік людини
* - час розв'язання арифметичної задачі
* - час розв'язання географічної задачі
* - час розв'язання логічної задачі

Задача полягає в перевірці залежності між даними вибірками на основі проведення багатовимірного аналізу обробки статистичних даних над зібраними даними.

Представлені багатовимірні дані зібрані самостійно.

**Об’єкт дослідження:** багатовимірні дані.

**Предмет дослідження**: залежність між часом доби, віком людини та часом виконання трьох різних задач(арифметична, на знання географії та логічна)

**Мета роботи**: розробити автоматизовану систему для візуального представлення та обробки великих обсягів статистичних даних, дослідити як залежать витрачені кошти на навчання англійської мови від тривалості процесу вивчення людиною.

**Програмне середовище:** що використовувалося в процесі виконання курсової роботи Microsoft Visual Studio Comunity 2015.

**Мова програмування :** C#.

# Вступ

Ця робота ставить ціль дослідити залежність людської кмітливості від різних факторів. Для цього розроблено тест, що складається з 3-х питань(арифметика, географія, логіка). Засікається час, за який людина відповіла на питання.

По-перше, у даних у цій роботі є інформація про час доби, коли людина пройшла даний тест(наприклад днем, ввечері тощо)

Другим фактором є вік людини, який людина ввела на початку тесту.

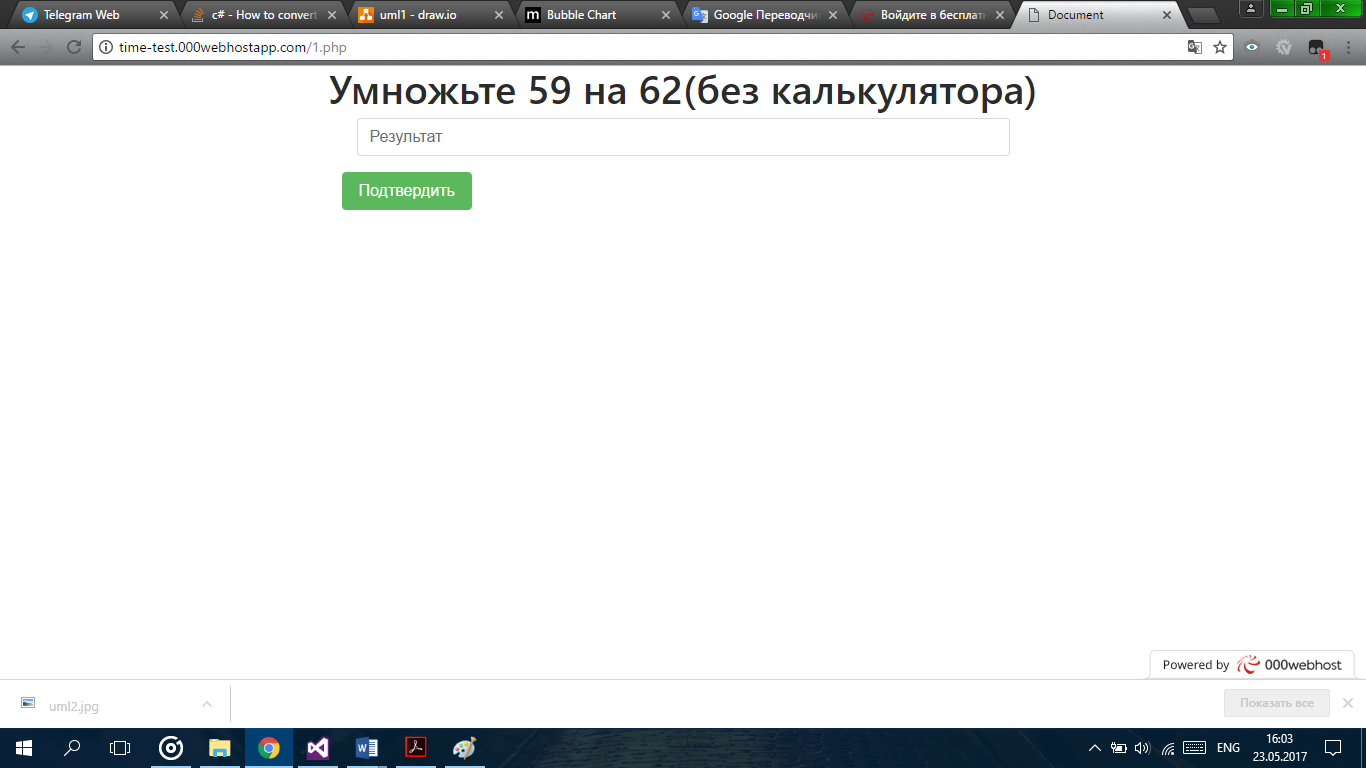
Даний тест був зроблений у формі сайту на мові РНР та викладений у мережу Інтернет. Всього було зроблено 30 спостережень

Рис.1 Сайт з тестом

**Практична частина**

Курсова робота основана на створенні програмного забезпечення для проведення статистичного аналізу, котре реалізоване на основі програм декількох лабораторних робіт.   
Дана програма реалізована на мові програмування високого рівня С#, в середовищі Microsoft Visual Studio з використанням додатку Windows Forms.

Обробка статистичних даних починається з первинного статистичного аналізу: формування варіаційного ряду, обчислення статистичних характеристик параметрів генеральної сукупності, побудова гістограми та емпіричної функції розподілу, відтворення розподілів та перевірка збігу емпіричної і теоретичної функцій, за необхідності перетворення даних.

Під час аналізу двовимірних масивів спостережень додається ще ряд питань, які не розглядаються при одновимірному: перевірку нормальності розподілу двовимірної випадкової величини, встановлення наявності стохастичного зв’язку та перевірка його значущості, за наявності зв'язку проведення регресійного аналізу, а саме відтворення ліній регресії (лінійної, параболічної та квазі-лінійної), обрахунок коефіцієнтів регресії, побудова довірчих інтервалів на лінію регресії, побудова толерантних меж, меж нового спостереження та перевірка адекватності відтвореної лінії регресії.

При обробці багатовимірних спостережень обраховується: вектор середніх, дисперсійно-коваріаційна та кореляційна матриці.

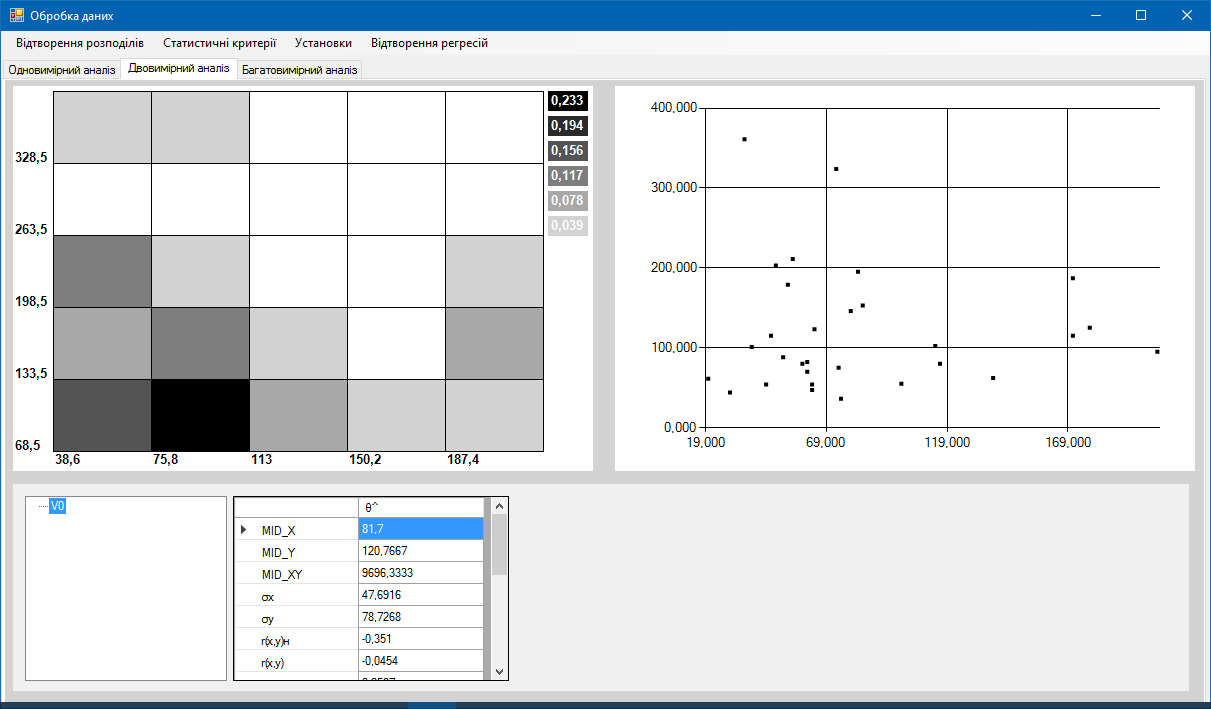
При наявності нормального багатовимірного розподілу виконується регресійний аналіз (відтворення моделі регресії та перевірка адекватності моделі регресії, на основі візуального аналізу діагностичної діаграми). Також в розробленому програмному забезпеченні присутня візуалізація даних: матриця діаграм розкиду, бульбашкова діаграма та паралельні координати.

## Тестування автоматизованої системи

### Двовимірний статистичний аналіз даних

Двохвимірний статистичного аналізу складається з етапів:

* побудови варіаційного ряду, гістограмної оцінки;
* встановлення наявності стохастичного зв’язку між складовими двовимірного випадкового вектора, якщо зв'язок існує, то виникає необхідність ідентифікації та відтворення лінії регресії.

  
.Рис2. Гістограма відносних частот і кореляційне (X3 ) та оцінки (X5 ).

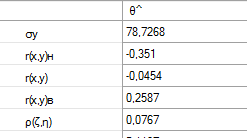
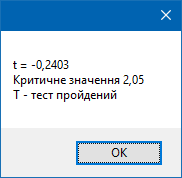
З візуального аналізу кореляційного поля і кореляційного аналізу можемо відтворювати лінійну регресію і проводити регресійний аналіз  
  


Рис.4. Кореляційний аналіз для (X3 ) та (X5  ).

  
  
Рис.5. Перевірка значущості

З візуального аналізу кореляційного поля і кореляційного аналізу можемо відтворювати лінійну регресію і проводити регресійний аналіз.

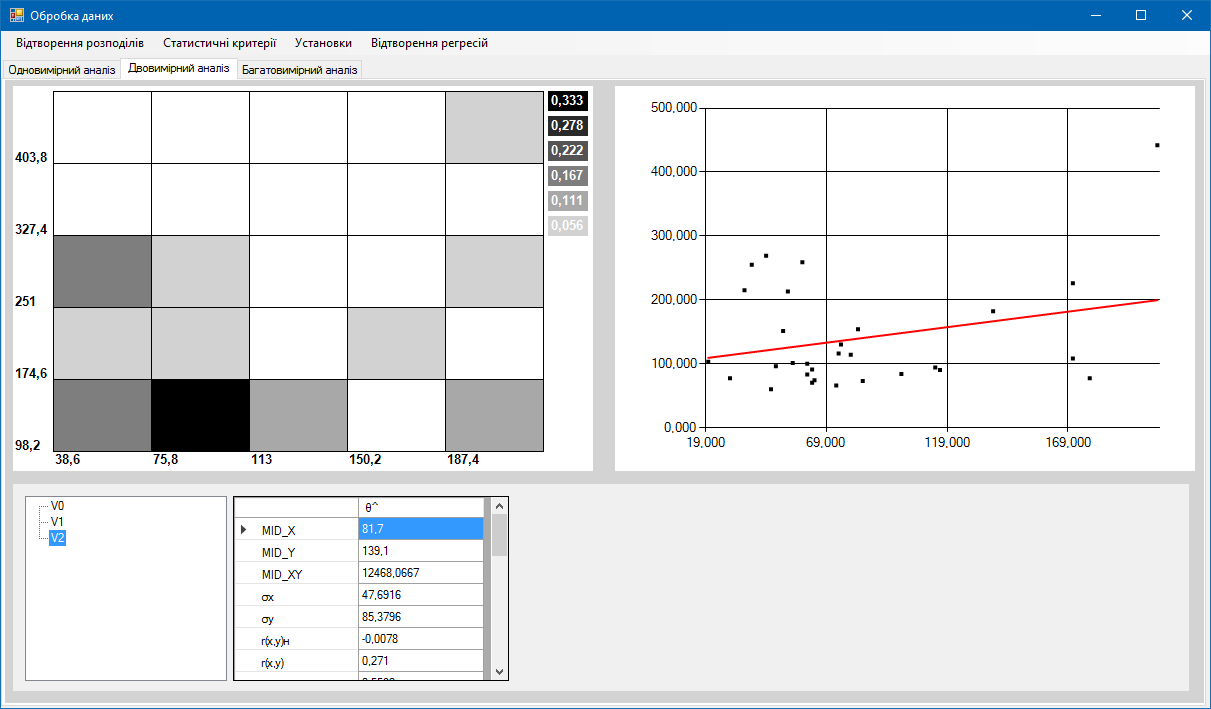


Рис.6. Відтворення лінії регресії для (X3 ) та (X4  ).

Отже, програма показує залежність між часом розв’язання математичної та географічної задачі.

### Багатовимірний статистичний аналіз даних

При обробці багатовимірних спостережень обраховується: вектор середніх, дисперсійно-коваріаційна та кореляційна матриці.

Дисперсійно-коваріаційна матриця відображає ступінь взаємозалежності двох величин.

Коефіцієнти кореляції відображають тісноту зв’язку між показниками.

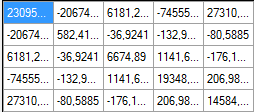


Рис.7.Дисперсійно-коваріаційна матриця

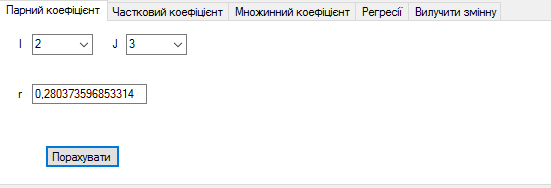


Рис8.Парний коефіцієнт кореляції між часом математичної та географічної задачі

Частковий коефіцієнт кореляції відображає зв'язок між 2-ма змінними без урахування впливу групи інших змінних

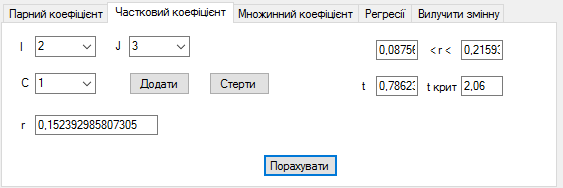


Рис9. Частковий коефіцієнт кореляції між часом математичної та географічної задачі без урахування віку та часу доби

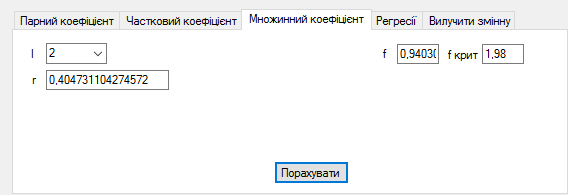
Множинний коефіцієнт кореляції пояснює вплив усіх ознак на і-ту

Рис9. Множинний коефіцієнт кореляції часу математичної задачі

Відтворимо регресію з часом доби

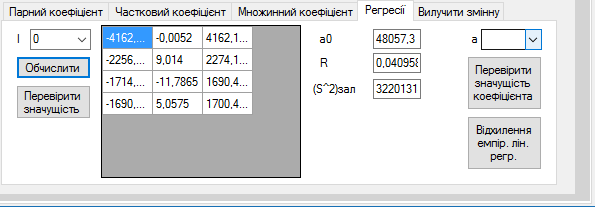


Рис11. Коефіцієнти регресії

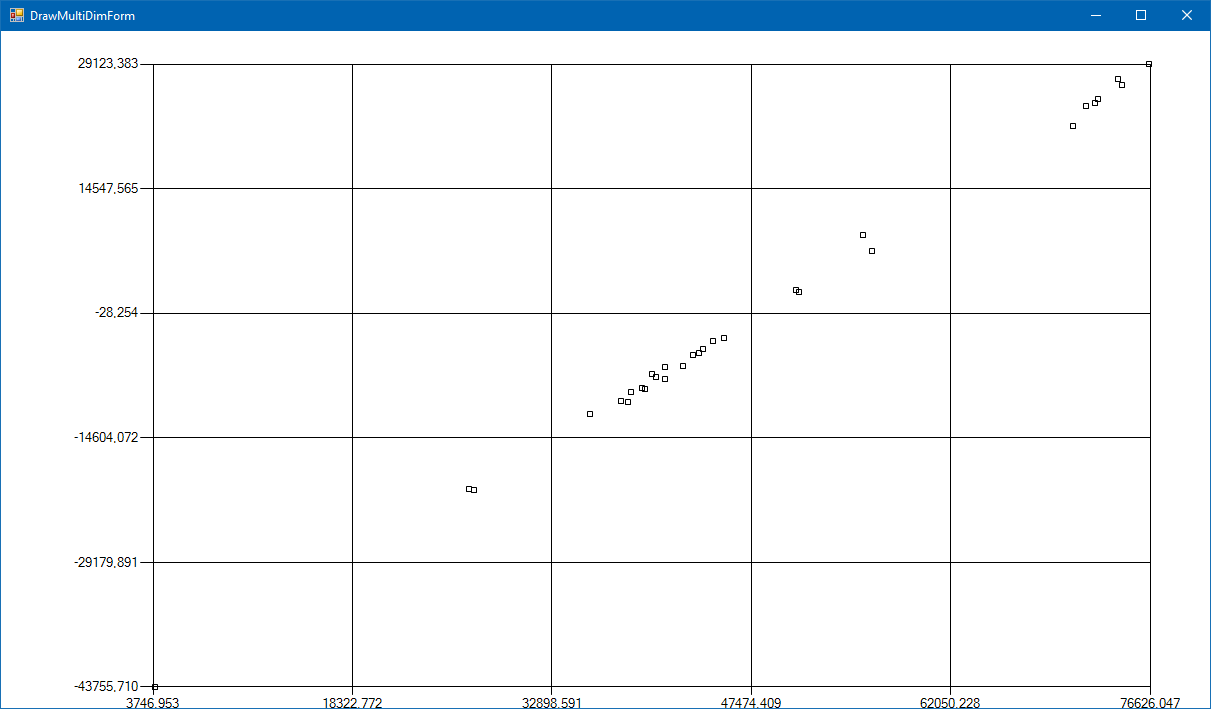


Рис.12.Діагностична діаграма.

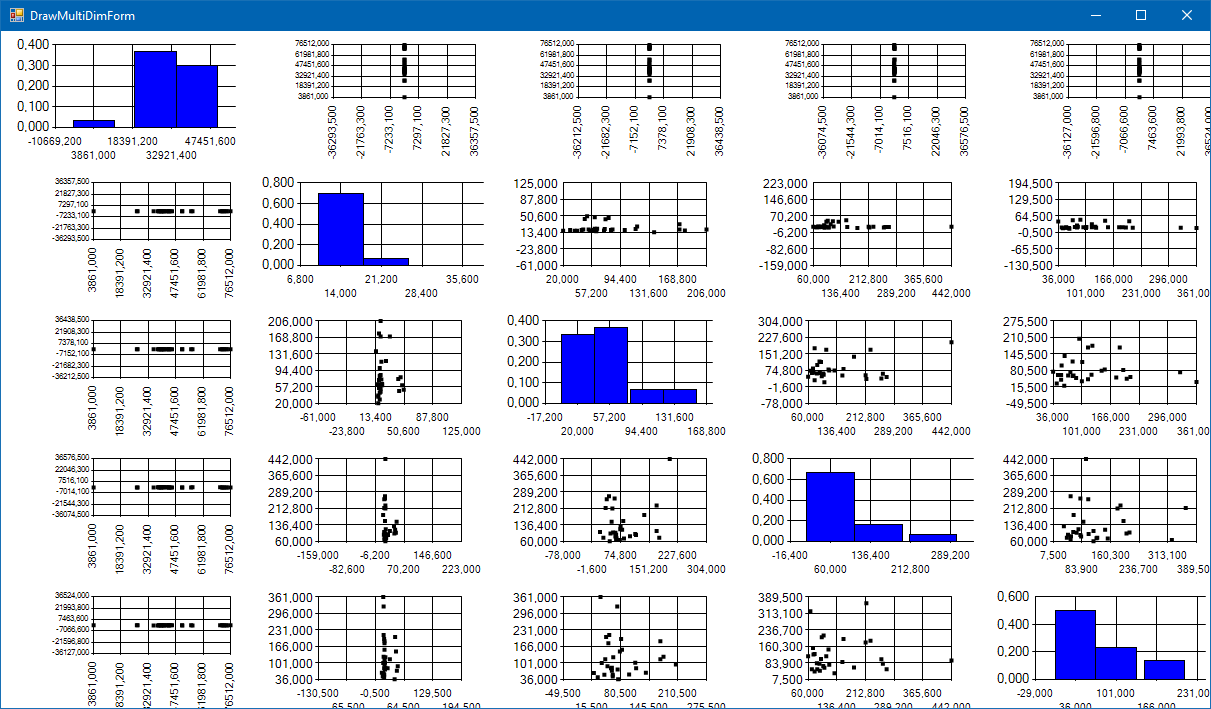
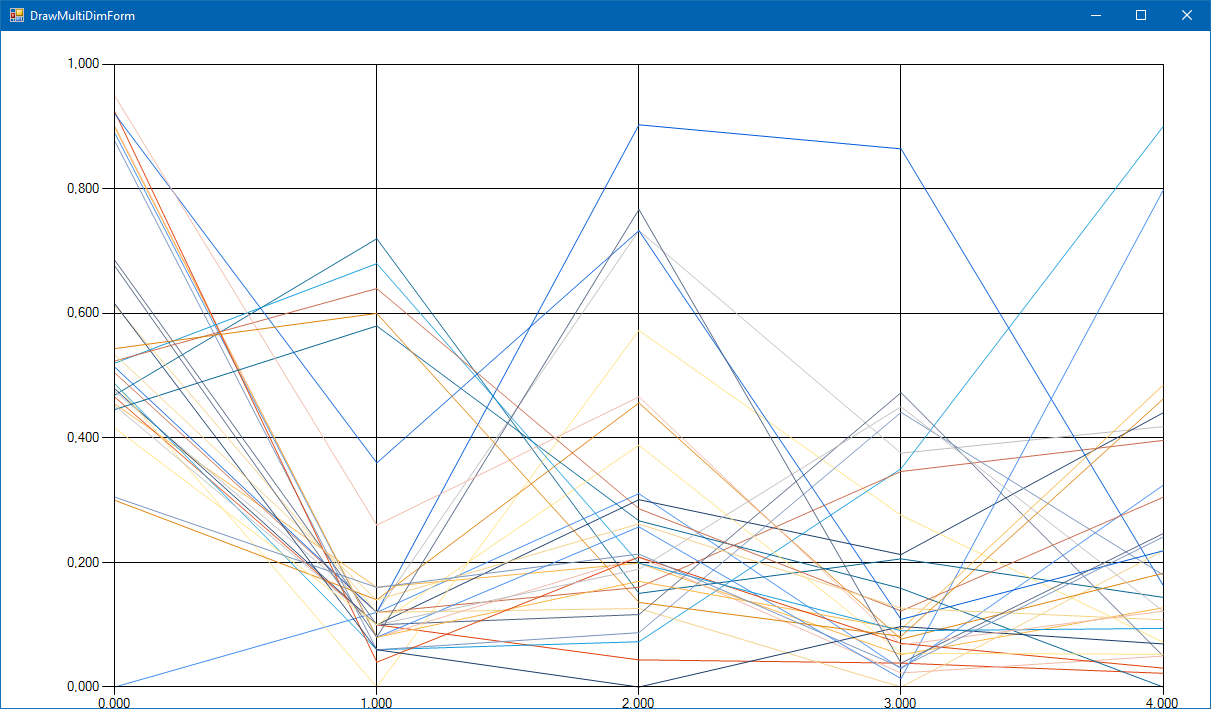


Рис.13.Матриця діагонального розкиду

Рис.14.Паралельні координати

### 

### Бульбашкова діаграма

**Висновки**

В даній курсовій роботі було проведено аналіз технологій обробки багатовимірних даних. Розроблене програмне забезпечення для візуального представлення та обробки великих обсягів статистичних даних.

Проведені дослідження яка залежність між витраченими коштами на навчання та тривалістю вивчення мови людиною.

А саме первинний статистичний аналіз одновимірних даних: формування варіаційного ряду, обчислення статистичних характеристик, вилучення аномальних значень, побудова гістограми та емпіричної функції розподілу. Перевірка нормальності розподілу двовимірної випадкової величини, встановлення наявності стохастичного зв’язку та багатовимірний статистичний аналіз:

* побудова дисперсійно-коваріаційної матриці, кореляційної матриці, діагностичної діаграми;
* відтворення багатовимірної лінійної регресії;

візуалізація багатовимірних даних за допомогою теплової карти, матриці діаграм розкиду, бульбашкової діаграми та паралельних координат.

Після проведення автоматизованої обробки зібраних статистичних даних було прийнято висновок, що існує зв'язок між часом доби, віком, та часом розв'язання задач

# Список використаної літератури

1. П.О. Приставка, О.М.Мацуга АНАЛІЗ ДАНИХ. Електронний посібник для студентів спеціальності «Прикладна математика».-Київ, 2010.-219с.