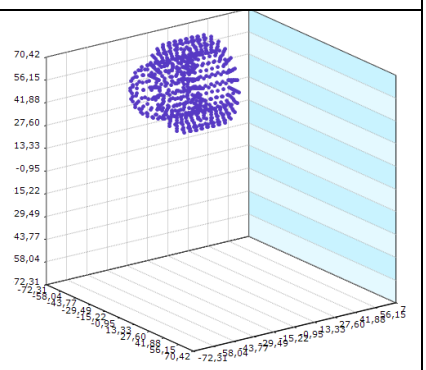
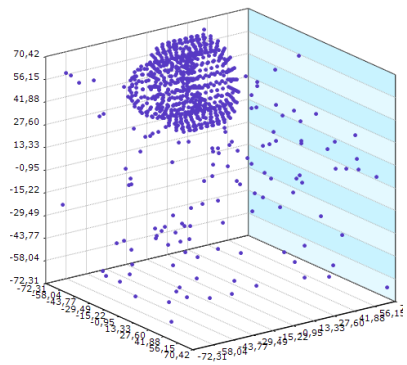
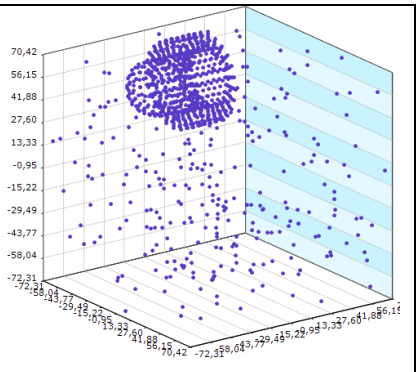
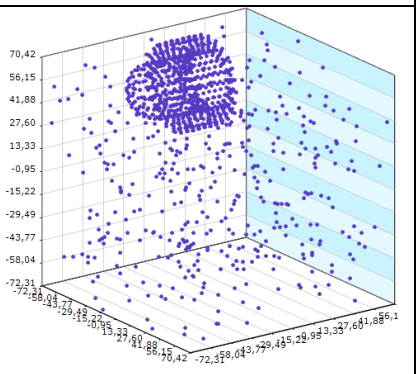
Додаток А

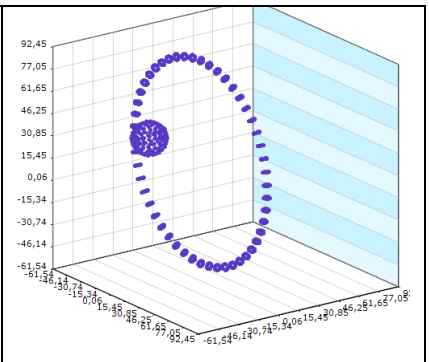
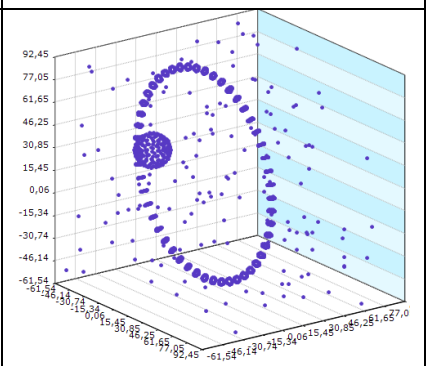
Таблиця А.1 – Експериментальні та реальні дані.

Кількість точок	% шуму	Характеристики вибірки	Просторове зображення
722	0	AveExpectation= 2,10 maxExpectation= 7,08 minExpectation= -0,77 AveStdDeviation= 1600092,39 maxStdDeviation= 2387575,17 minStdDeviation= 321328,11 AveRange= 59,49 maxRange= 72,31 minRange= 52,44 AveCorrelation= 3,03 maxCorrelation= 9,02 minCorrelation= -1,58	
	20	AveExpectation= 0,19 maxExpectation= 0,58 minExpectation= 0,00 AveStdDeviation= 3325480,53 maxStdDeviation= 4479679,53 minStdDeviation= 1069160,12 AveRange= 140,25 maxRange= 141,31 minRange= 138,42 AveCorrelation= 0,67 maxCorrelation= 1,98 minCorrelation= -0,35	

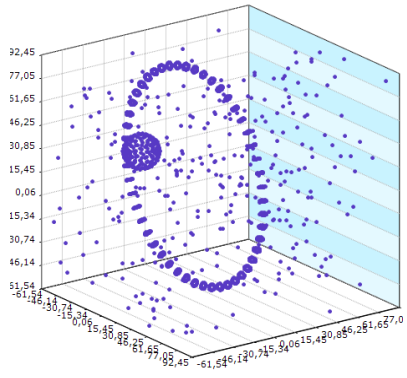
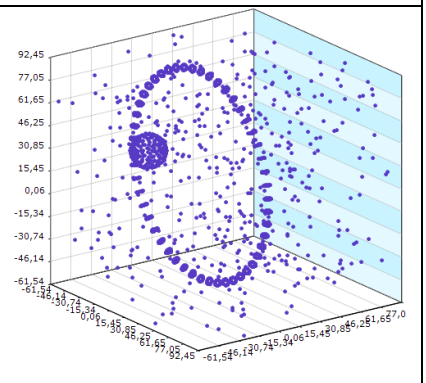
Продовження таблиці А.1.

	40	AveExpectation= -0,05 maxExpectation= 0,53 minExpectation= -0,68 AveStdDeviation= 4625156,99 maxStdDeviation= 6082670,76 minStdDeviation= 1798336,22 AveRange= 141,58 maxRange= 142,42 minRange= 141,00 AveCorrelation= 0,31 maxCorrelation= 0,91 minCorrelation= -0,15	
	60	AveExpectation= -0,05 maxExpectation= 0,49 minExpectation= -0,65 AveStdDeviation= 5941195,32 maxStdDeviation= 7593426,39 minStdDeviation= 2672303,41 AveRange= 141,25 maxRange= 142,42 minRange= 140,31 AveCorrelation= 0,23 maxCorrelation= 0,67 minCorrelation= -0,09	

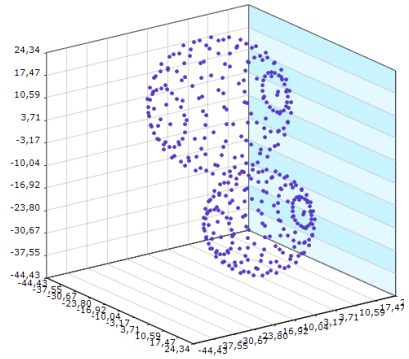
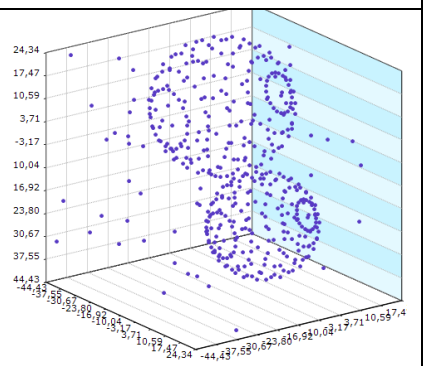
Продовження таблиці А.1.

782	0	AveExpectation= -0,39 maxExpectation= 0,00 minExpectation= -0,67 AveStdDeviation= 4241167,92 maxStdDeviation= 6547451,32 minStdDeviation= 82153,88 AveRange= 103,39 maxRange= 144,09 minRange= 22,27 AveCorrelation= 0,01 maxCorrelation= 0,02 minCorrelation= 0,00	
	20	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 6901696,96 maxStdDeviation= 10357365,41 minStdDeviation= 771019,95 AveRange= 152,00 maxRange= 152,54 minRange= 151,00 AveCorrelation= -0,02 maxCorrelation= 0,01 minCorrelation= -0,06	

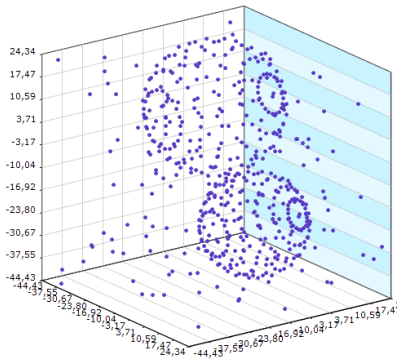
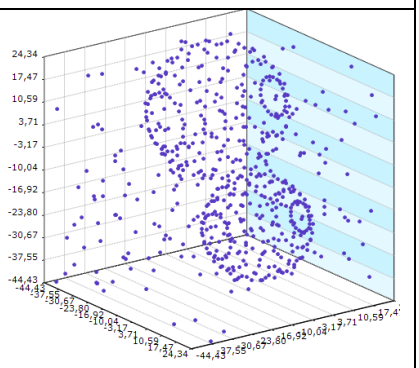
Продовження таблиці А.1.

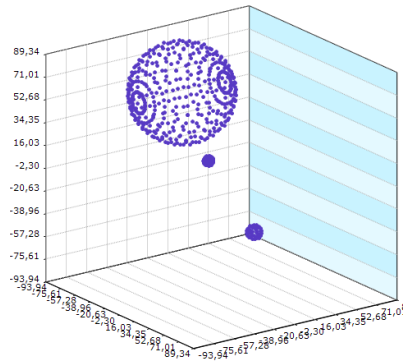
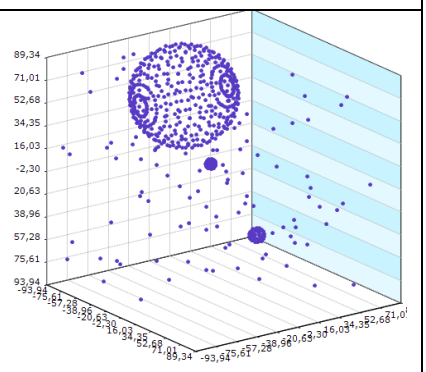
	40	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 9418044,41 maxStdDeviation= 13890736,90 minStdDeviation= 1521326,92 AveRange= 152,33 maxRange= 153,45 minRange= 151,00 AveCorrelation= -0,03 maxCorrelation= 0,00 minCorrelation= -0,06	
	60	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 11845187,01 maxStdDeviation= 17273110,25 minStdDeviation= 2607450,88 AveRange= 152,33 maxRange= 153,45 minRange= 151,54 AveCorrelation= -0,05 maxCorrelation= 0,00 minCorrelation= -0,10	

Продовження таблиці А.1.

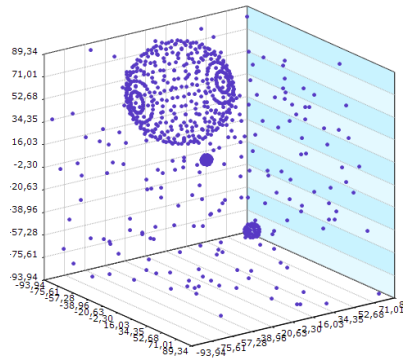
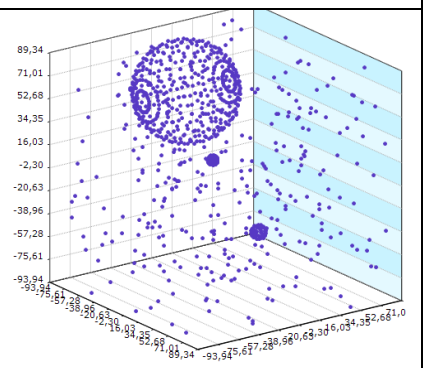
382	0	AveExpectation= 0,10 maxExpectation= 0,70 minExpectation= -0,31 AveStdDeviation= 766331,19 maxStdDeviation= 1405571,80 minStdDeviation= 77981,37 AveRange= 52,79 maxRange= 65,17 minRange= 39,69 AveCorrelation= -0,18 maxCorrelation= 0,10 minCorrelation= -0,69	
	20	AveExpectation= -0,61 maxExpectation= 0,00 minExpectation= -1,50 AveStdDeviation= 1144620,84 maxStdDeviation= 1940495,72 minStdDeviation= 225749,65 AveRange= 66,26 maxRange= 68,34 minRange= 63,43 AveCorrelation= -0,26 maxCorrelation= 0,00 minCorrelation= -0,62	

Продовження таблиці А.1.

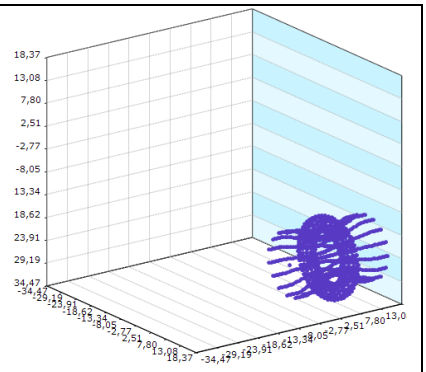
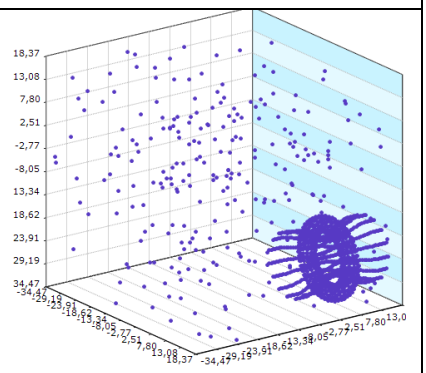
	40	AveExpectation= -0,10 maxExpectation= 0,00 minExpectation= -0,24 AveStdDeviation= 1496430,74 maxStdDeviation= 2447026,02 minStdDeviation= 336807,99 AveRange= 67,59 maxRange= 68,34 minRange= 67,00 AveCorrelation= -0,25 maxCorrelation= 0,00 minCorrelation= -0,55	
	60	AveExpectation= -0,10 maxExpectation= 0,00 minExpectation= -0,24 AveStdDeviation= 1821133,68 maxStdDeviation= 2894346,77 minStdDeviation= 442973,54 AveRange= 67,59 maxRange= 68,34 minRange= 67,00 AveCorrelation= -0,26 maxCorrelation= 0,00 minCorrelation= -0,48	

556	0	AveExpectation= -10,72 maxExpectation= 0,24 minExpectation= -32,11 AveStdDeviation= 2464777,25 maxStdDeviation= 4209462,00 minStdDeviation= 500369,72 AveRange= 114,75 maxRange= 134,30 minRange= 79,37 AveCorrelation= 0,51 maxCorrelation= 1,54 minCorrelation= 0,00	
	20	AveExpectation= -0,14 maxExpectation= 0,17 minExpectation= -0,38 AveStdDeviation= 3973166,20 maxStdDeviation= 6541042,98 minStdDeviation= 948258,37 AveRange= 178,09 maxRange= 180,94 minRange= 175,00 AveCorrelation= 0,25 maxCorrelation= 0,76 minCorrelation= -0,06	

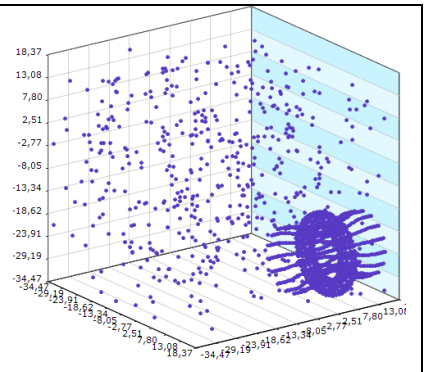
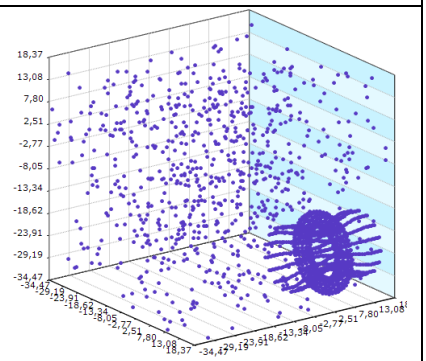
Продовження таблиці А.1.

	40	AveExpectation= -0,02 maxExpectation= 0,09 minExpectation= -0,17 AveStdDeviation= 5522130,71 maxStdDeviation= 8696946,19 minStdDeviation= 1748233,48 AveRange= 180,09 maxRange= 180,34 minRange= 179,94 AveCorrelation= 0,17 maxCorrelation= 0,50 minCorrelation= -0,01	
	60	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 7203610,91 maxStdDeviation= 10850564,28 minStdDeviation= 2779506,90 AveRange= 181,42 maxRange= 182,34 minRange= 180,94 AveCorrelation= 0,09 maxCorrelation= 0,27 minCorrelation= -0,02	

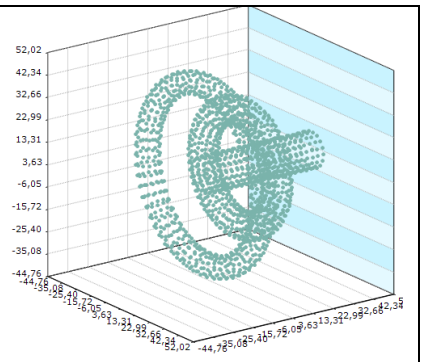
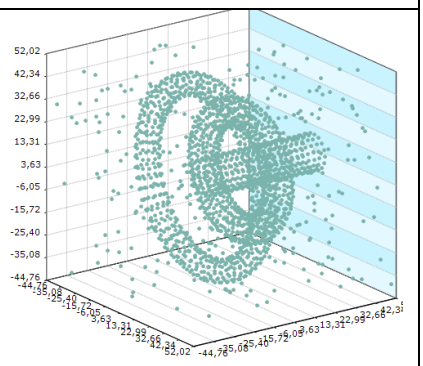
Продовження таблиці А.1.

1292	0	AveExpectation= 0,91 maxExpectation= 2,81 minExpectation= -0,10 AveStdDeviation= 4212128,29 maxStdDeviation= 10029279,67 minStdDeviation= 349007,62 AveRange= 24,43 maxRange= 34,47 minRange= 18,81 AveCorrelation= 4,49 maxCorrelation= 9,42 minCorrelation= -2,38	
	20	AveExpectation= -0,06 maxExpectation= 0,00 minExpectation= -0,17 AveStdDeviation= 5999719,90 maxStdDeviation= 13782016,65 minStdDeviation= 917930,93 AveRange= 51,61 maxRange= 52,37 minRange= 51,00 AveCorrelation= 0,69 maxCorrelation= 1,32 minCorrelation= -0,24	

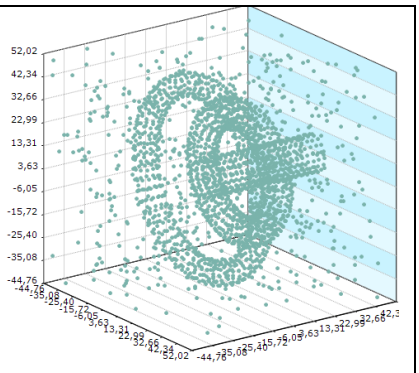
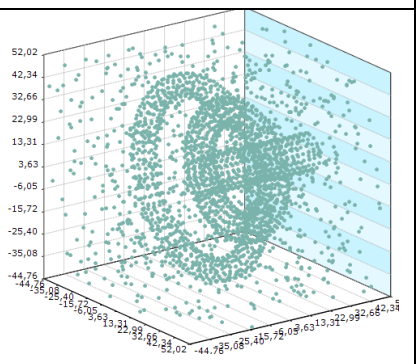
Продовження таблиці А.1.

	40	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 7457980,14 maxStdDeviation= 16997770,61 minStdDeviation= 1210568,25 AveRange= 51,61 maxRange= 52,37 minRange= 51,00 AveCorrelation= 0,26 maxCorrelation= 0,51 minCorrelation= -0,06	
	60	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 8915388,80 maxStdDeviation= 20191761,83 minStdDeviation= 1526109,33 AveRange= 51,61 maxRange= 52,37 minRange= 51,00 AveCorrelation= 0,10 maxCorrelation= 0,21 minCorrelation= -0,01	

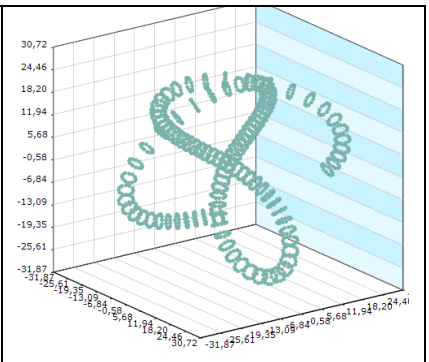
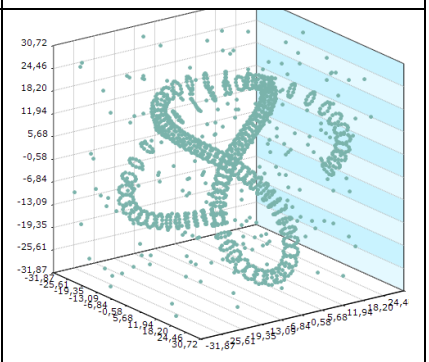
Продовження таблиці А.1.

1751	0	AveExpectation= -0,03 maxExpectation= 0,00 minExpectation= -0,04 AveStdDeviation= 8698787,61 maxStdDeviation= 13130396,47 minStdDeviation= 1260447,72 AveRange= 76,50 maxRange= 86,39 minRange= 56,84 AveCorrelation= -0,02 maxCorrelation= 0,04 minCorrelation= -0,10	
	20	AveExpectation= 0,05 maxExpectation= 0,16 minExpectation= 0,00 AveStdDeviation= 13304925,36 maxStdDeviation= 19083532,59 minStdDeviation= 3392606,23 AveRange= 95,59 maxRange= 96,02 minRange= 95,00 AveCorrelation= -0,03 maxCorrelation= 0,00 minCorrelation= -0,08	

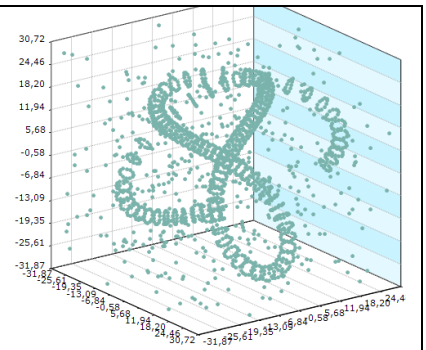
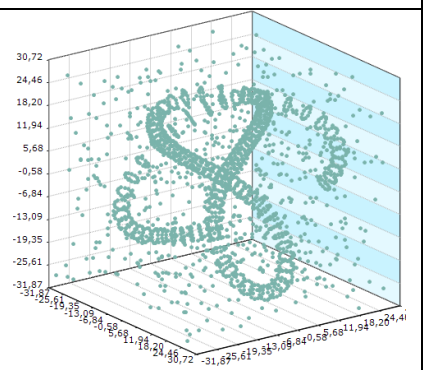
Продовження таблиці А.1.

	40	AveExpectation= -0,05 maxExpectation= 0,00 minExpectation= -0,16 AveStdDeviation= 16912056,08 maxStdDeviation= 24188415,21 minStdDeviation= 4333312,74 AveRange= 95,59 maxRange= 96,02 minRange= 95,00 AveCorrelation= -0,02 maxCorrelation= 0,00 minCorrelation= -0,06	
	60	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 20458703,13 maxStdDeviation= 29141035,22 minStdDeviation= 5281287,93 AveRange= 95,59 maxRange= 96,02 minRange= 95,00 AveCorrelation= -0,02 maxCorrelation= 0,00 minCorrelation= -0,05	

Продовження таблиці А.1.

1294	0	AveExpectation= 0,00 maxExpectation= 0,01 minExpectation= 0,00 AveStdDeviation= 14292162,05 maxStdDeviation= 17868392,54 minStdDeviation= 9762846,59 AveRange= 52,45 maxRange= 60,01 minRange= 41,37 AveCorrelation= 0,00 maxCorrelation= 0,01 minCorrelation= 0,00	
	20	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 18999204,54 maxStdDeviation= 23385494,06 minStdDeviation= 13156549,88 AveRange= 60,86 maxRange= 61,72 minRange= 60,00 AveCorrelation= 0,00 maxCorrelation= 0,00 minCorrelation= 0,00	

Кінець таблиці А.1.

	40	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 23367447,26 maxStdDeviation= 28523867,45 minStdDeviation= 16244155,10 AveRange= 60,86 maxRange= 61,72 minRange= 60,00 AveCorrelation= 0,00 maxCorrelation= 0,00 minCorrelation= 0,00	
	60	AveExpectation= 0,00 maxExpectation= 0,00 minExpectation= 0,00 AveStdDeviation= 27727745,75 maxStdDeviation= 33648341,17 minStdDeviation= 19328719,57 AveRange= 60,86 maxRange= 61,72 minRange= 60,00 AveCorrelation= 0,00 maxCorrelation= 0,00 minCorrelation= 0,00	

Додаток Б

Слайди презентації

Актуальность исследований

- **Необходимость синхронизировать и организовать на единой критериальной основе выбор метода кластеризации на основании данных анализируемой выборки**
- **Потребность унифицировать алгоритмы кластеризации и за счет этого сократить время на выбор метода**
- **Необходимость обеспечения пользователей качественным решением задачи анализа при различных исследуемых данных**
- **Постоянное увеличение объема поступающей информации и разнородность этой информации требует развития технологий анализа этих данных**
- **Необходимость создания универсального метода кластеризации. Отдельные методы хорошо работают на определенных данных, но не являются универсальными**
- **Необходимость анализа сложных наборов данных, с пересекающимися и накладывающимися классами**

Цель работы

Разработка универсальной математической модели зависимости выбора комбинации алгоритмов на различных этапах иерархического алгоритма Хамелеон от исходных характеристик анализируемого набора данных с целью улучшения качества кластеризации.

2

Постановка задачи исследования

Представим постановку задачи как разработку такой математической модели выбора алгоритмов A_r для разных этапов алгоритма кластеризации Хамелеон для решения R задачи выбора оптимальной комбинации алгоритмов $A = \sum_{r=1}^{r} A_r$, состоящей из множества $A_r, r = \overline{1, r}$. Выбор каждой комбинации сводится к выбору такого подмножества, решения задачи, которое удовлетворяло бы заданному критерию качества и статистическим характеристикам объекта. Представим модель следующим образом:

$$\{D, H_D\} \xrightarrow{A_r \in A} \{E^r\}$$

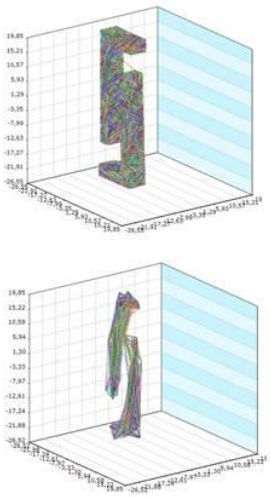
где A – множество алгоритмов; D – исследуемая выборка данных; H_D – множество характеристик исследуемой выборки данных; E – множество показателей качества решаемой задачи по выбору алгоритмов для разных этапов алгоритма кластеризации Хамелеон, оптимального для данной выборки.

3

Модификация этапа огрубления графа

В процессе стадии огрубления строится последовательность меньших графов, каждый с меньшим количеством узлов.

- Случайное паросочетание (RM).
- Паросочетание из тяжелых ребер (HEM).
- Модифицированный алгоритм паросочетания тяжелых ребер (Modified Heavy Edge Matching - HEM*).
- Паросочетание из наиболее тяжелых ребер (heaviest-edge matching).
- Модифицированное паросочетание из наиболее тяжелых ребер HEM*+.
- Паросочетание легких ребер (LEM).
- Паросочетание из тяжелых клик (HCM).
- Сочетание тяжелых треугольников. Heavy-triangle matching (HTM).
- Сочетание тяжелых схем (Heaviest Schema Matching (HSM)).
- Сочетание гиперребер (Hyperedge Coarsening HEC).
- Видоизмененое сочетание гиперребер (Modified Hyperedge Coarsening MHEC).
- Сочетание лучшего (первого) выбора (First Choice Coarsening FCC).



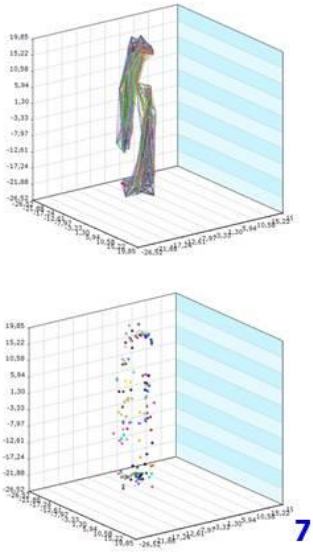
6

Модификация этапа начального разделения графа в рамках алгоритма Хамелеон (Initial partitioning)

Для решения задачи разбиения графа можно рекурсивно применить метод бинарного деления, при котором на первой итерации граф разделяется на две равные части, далее на втором шаге каждая из полученных частей также разбивается на две части и т.д.

В случае, когда требуемое количество разбиений *n* не является степенью двойки, каждое деление пополам необходимо осуществлять в соответствующем соотношении.

- Покоординатное разбиение (Coordinate NestedDissection (CND)).
- Деление сети с использованием кривых, заполняющих пространство (Space-flling Curv Techniques (SFCT)).
- Рекурсивное разделение графа (Recursive Graph Bisection (RGB)).
- Алгоритм возрастающего графа (Graph Growth Partitioning (GGP)).
- Алгоритм возрастающего графа с учетом выгод (Greedy Graph Growing Algorithm (GGGP)).
- Уровневое ячеечное разбиение(Levelized Nested Dissection (LND)).
- Seed-Growth bisection (SGB).



7

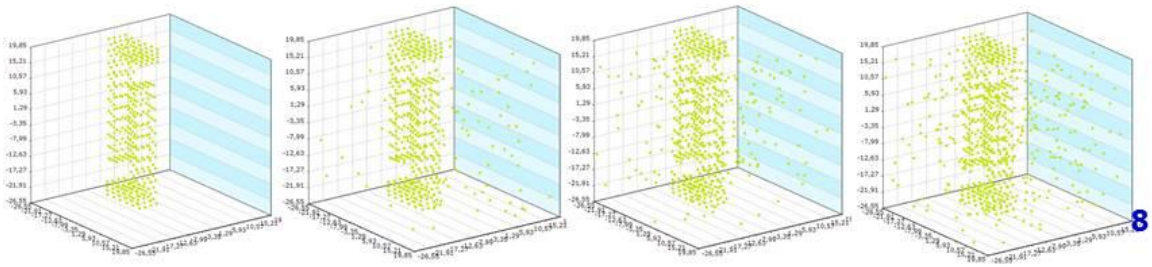
Описание реальных и экспериментальных выборок

- Реальные выборки данных были получены с профильных ресурсов. Выбрано 46 выборок для анализа.

- С ресурсов так же получены экспериментальные выборки используемые для проверки кластеризации. Выбрано 35 выборок с различными характеристиками и количеством объектов в выборке.

- Предложен метод генерации трехмерных экспериментальных выборок на основе экспортированных данных построенной модели в системе 3Ds Max. Сгенерированно 33 выборки. Каждая из выборок анализировалась в 4 вариантах:

  - без добавления шума;
  - с добавлением 20% шума;
  - с добавлением 40% шума;
  - с добавлением 60% шума.



8

Математическая модель зависимости выбора параметра k при построении k-nn графа от исходных характеристик выборки

В рамках работы было проведено 3 эксперимента для выбора управляемых параметров:

- В первом эксперименте анализировались такие характеристики как количество объектов в выборке, минимальные и максимальные значения матожидания, дисперсии и разброса. Зависимости между данными параметрами и значением k не выявлено;
- Во втором эксперименте в качестве управляемого параметра были выбраны длинна наибольшего остовного ребра полносвязного графа и среднее значение длинны всех остальных ребер остова. Данные характеристики показывают зависимость, но использование данного подхода не является целесообразным в связи с трудоемкостью построения остова полносвязного графа;
- В третьем эксперименте в качестве характеристики использовались количество компонент связности и максимальное расстояние между компонентами связности и количество элементов в компоненте связности. Вторая характеристика вычисляется следующим образом:

$$SetDist = \max\left(\frac{dist(avComponent_i, avComponent_j)}{\max\left(\frac{\max\ ComponentOstovEdge_{ij}}{ComponentVertexNum_{ij}}\right)}\right)$$

где *avComponent* - центроид компоненты связности, *ComponentOstovEdge* - ребро соединяющее вершны принадлежащие одной компоненте, *ComponentVertexNum* – количество вершин в компоненте.
Данные характеристики нетрудоемки в расчете и существует зависимость между ними и значением k.

9

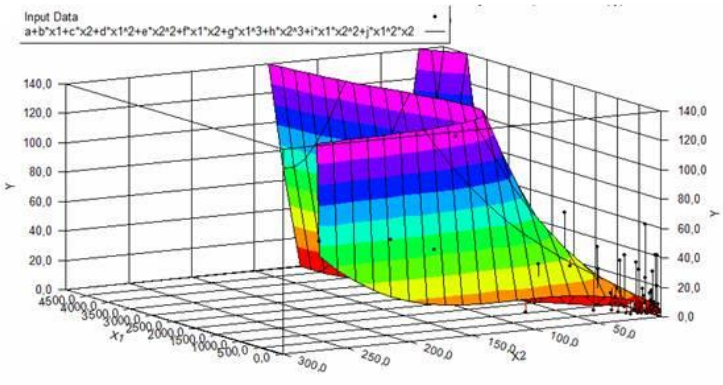
Математическая модель зависимости выбора параметра k для асимметричного графа

В результате исследования была получена следующая мат модель:

$$k = a + b \cdot x_1 + c \cdot x_2 + d \cdot x_1^2 + e \cdot x_2^2 + f \cdot x_1 \cdot x_2 + g \cdot x_1^3 + h \cdot x_2^3 + i \cdot x_1 \cdot x_2^2 + j \cdot x_1^2 \cdot x_2$$

где x_1- коэффициент расстояния, x_2 – количество компонент связности.

a	4,963024
b	2,33E-02
c	0,42939
d	-4,45E-05
e	-3,86E-03
f	4,18E-04
g	1,05E-08
h	1,14E-05
i	1,19E-05
j	-4,73E-07



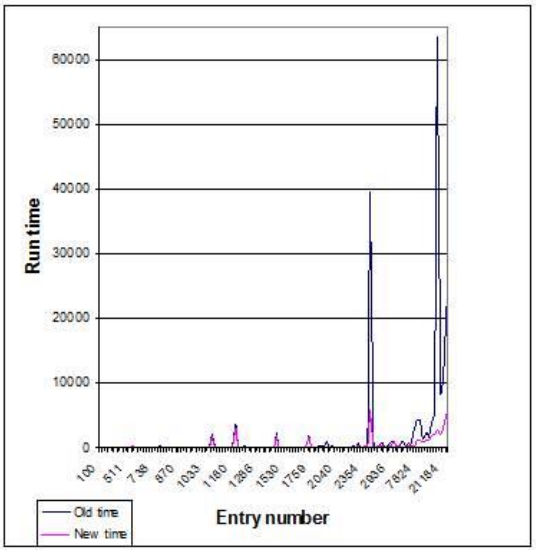
Стандартная ошибка оценки 11,2986020522291.

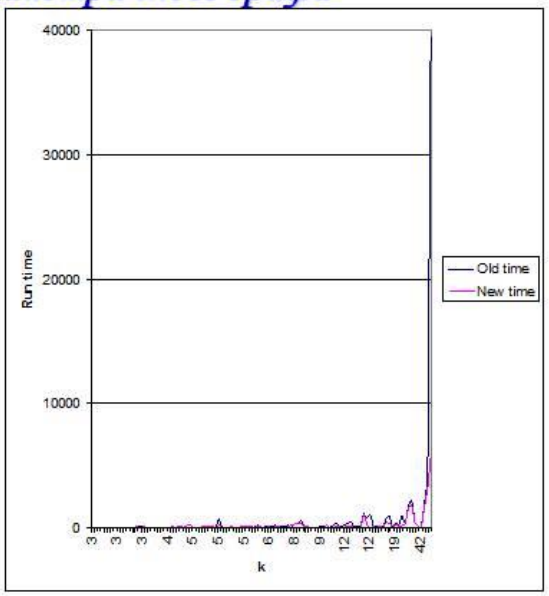
Коэффициент множественной детерминации 0,6452864929.

Статистика Дублина-Ватсона 1,24157318003058.

10

Применение математической модели зависимости выбора параметра k для асимметричного графа





Применение данной модели улучшили время выполнения этапа построения графа в 62.45% случаев. В 37.55% случаев время выполнения ухудшилось. Отрицательный результат применения модели получен в 7.71% случаев. В среднем время выполнения улучшилось на 161%.

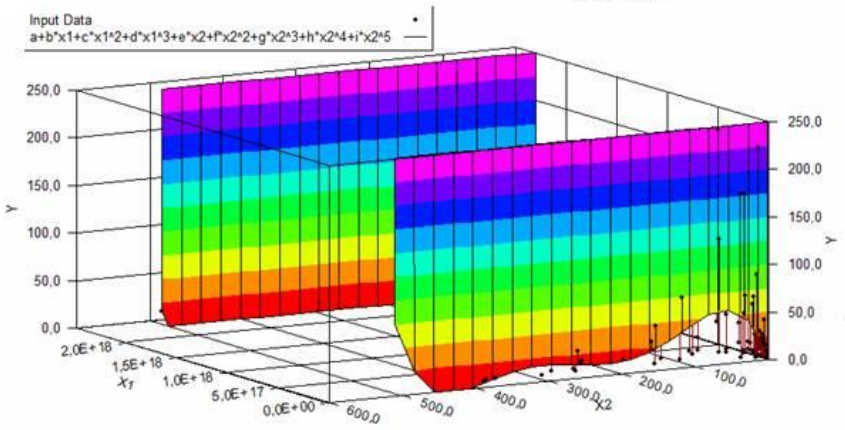
11

Математическая модель зависимости выбора параметра k для симметричного графа

В результате исследования была получена следующая мат модель:

$$k = a + b \cdot x_1 + c \cdot x_1^2 + d \cdot x_1^3 + e \cdot x_2 + f \cdot x_2^2 + g \cdot x_2^3 + h \cdot x_2^4 + i \cdot x_2^5$$

где x_1- коэффициент расстояния, x_2 – количество компонент связности.



a	-0,547360564
b	-7,46E-14
c	1,51E-29
d	-6,56E-48
e	2,323285358
f	-3,09E-02
g	1,55E-04
h	-3,34E-07
i	2,61E-10

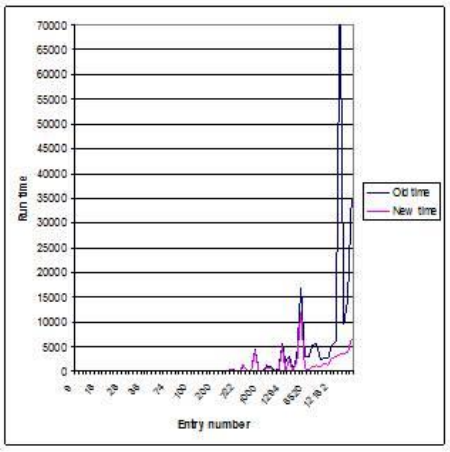
Стандартная ошибка оценки 42,8805641130193.

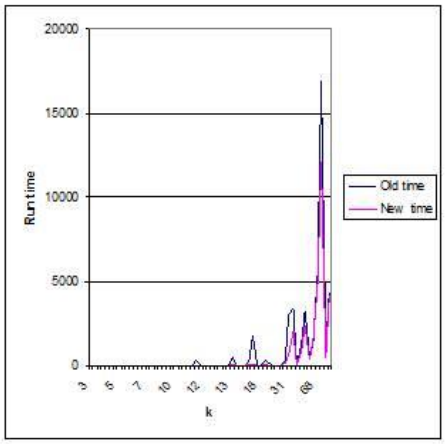
Коэффициент множественной детерминации 0,15118817.

Статистика Дублина-Ватсона 1,26055939255469.

12

Применение математической модели зависимости выбора параметра k для симметричного графа

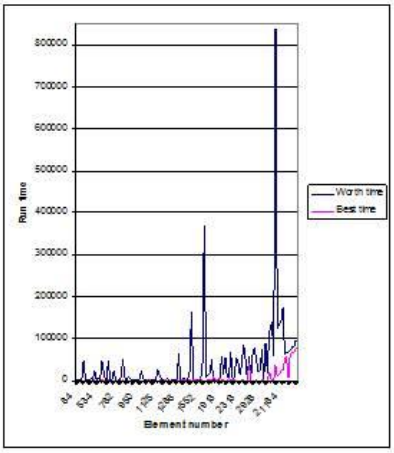




Применение данной модели улучшили время выполнения этапа построения графа в 69.23% случаев. В 20.51% случаев время выполнения ухудшилось. Отрицательный результат применения модели получен в 5.12% случаев. В среднем время выполнения улучшилось на 169%

13

Выводы по экспериментальным результатам работы модифицированного алгоритма Хамелеон



- Для сложных выборок иногда есть смысл использовать алгоритм KL, для остальных в большинстве случаев целесообразно использовать более быстрый алгоритм FM и более простые граничные алгоритмы KL и FM.
- Для выборок со сложной структурой данных качество разделения наиболее точно оценивается с помощью индекса Данна.
- Применение алгоритма построения симметричного графа нецелесообразно из-за низкой скорости работы.
- Качество работы алгоритмов на этапе восстановления и улучшения графа падает с увеличением количества классов.

14

ВЫВОДЫ

- Впервые построена **модель зависимости качества кластеризации сложных линейнонеразделимых защумленных данных различного объема** от характеристик данных и алгоритмов динамической кластеризации.
- **Дальнейшее развитие получил метод Хамелеон**, который в отличие от существующего использует разные алгоритмы на разных этапах кластеризации, что позволяет учитывать разницу в данных и использовать методы, которые работают лучше на конкретных данных.
- **Усовершенствован критерий сравнения алгоритмов и анализа исходных данных** применяемых в рамках алгоритма Хамелеон
- **Усовершенствован метод построения графа**, который позволяет ускорить процесс построения графа посредством выбора оптимального k, основываясь на характеристиках анализируемых данных. **Созданы модели выбора k для алгоритма k-ближайшах соседей используемого в алгоритме Хамелеон**, основанная на характеристиках данных.

16