**МИНОБРНАУКИ РОССИИ**
**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**
**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**
**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**
**Кафедра математического обеспечения и применения ЭВМ**

# ОТЧЕТ

**по лабораторной работе №7**

**по дисциплине «Статистические методы обработки экспериментальных данных»**

**Тема: Кластерный анализ. Метод поиска сгущений.**

Студентка гр. 8382 _____ Звегинцева Е.Н.

Студент гр. 8382 _____ Мирончик П.Д.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург

2022

**Цель работы.**

Освоение основных понятий и некоторых методов кластерного анализа, в частности, метода поиска сгущений.

**Основные теоретические положения.**

Нормировка, т.е. стандартизация, переменных применяется для того, чтобы характеристики имели один масштаб, в следствии чего было возможно корректное разбиение на кластеры. В данной работе для нормировки была использована формула

$$z = \frac{x - \bar{x}}{\sigma_x}$$

Где $\sigma_x$ – стандартное отклонение переменной, а $\bar{x}$ – ее среднее значение.

Метод поиска сгущений является еще одним итеративным методом кластерного анализа. Основная идея метода заключается в построении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков в поисках локальных сгущений объектов.

Алгоритм можно описать так:

1. В качестве начального центра кластера выбирается наблюдение из выборки еще не распределенных наблюдений, на расстоянии R от которого находится наибольшее число других наблюдений;
2. Формируется кластер относительно текущего центра;
3. В качестве нового центра выбирается среднее геометрическое место точек, входящих в кластер

$$\left(v_{c_k}, E_{c_k}\right) = \left(\frac{1}{N_k}\sum_{i=1}^{N_k} v_{k,i}\,, \frac{1}{N_k}\sum_{i=1}^{N_k} E_{k,i}\right)$$

4. Повторяем пункты 2 и 3, пока состав нового кластера не совпадет с составом кластера, полученного на предыдущей итерации.
5. Добавляем полученные данные (кластер и его центр) и запоминаем выборку нераспределенных элементов.

Для оценки устойчивости полученного разбиения целесообразно повторить процесс кластеризации несколько раз для различных значений радиуса сферы, изменяя каждый раз радиус на небольшую величину.

Существуют различные способы выбора начального радиуса сферы. В частности, если обозначить через $d_{ij}$ расстояние между $i$-м и $j$-м объектами, то в качестве нижней границы значения радиуса сферы можно выбрать минимальное из таких расстояний, а в качестве верхней границы - максимальное:

$$R_{min} = \min_{i,j} d_{ij};$$

$$R_{max} = \max_{i,j} d_{ij}.$$

Тогда, если начинать работу алгоритма с

$$R = R_{min} + \delta;\ \delta > 0$$

и при каждом его повторении увеличивать значение $\delta$ на некоторую величину, то в конечном итоге можно найти значения радиусов, которые приводят к устойчивому разбиению на кластеры.

Наилучшим разбиением считается такое, при котором достигается экстремальное (минимальное или максимальное) значение выбранного функционала качества.

В качестве таких функционалов могут быть использованы:

- Сумма по всем кластерам квадратов расстояний элементов кластеров до центров соответствующих кластеров:

$$F_1 = \sum_k \sum_{i=1}^{N_k} ||x_{k,i} - c_k||^2 = \sum_k \sum_{i=1}^{N_k} \left((\nu_{k,i} - \nu_{c_k})^2 + (E_{k,i} - E_{c_k})^2\right)$$

- Сумма по всем кластерам внутрикластерных расстояний между элементами кластеров:

$$F_2 = \sum_k \sum_{i=1}^{N_k} \sum_{j=i+1}^{N_k} ||x_{k,i} - x_{k,j}||^2 = \sum_k \sum_{i=1}^{N_k} \sum_{j=i+1}^{N_k} \left((\nu_{k,i} - \nu_{k,j})^2 + (E_{k,i} - E_{k,j})^2\right)$$

- Сумма по всем кластерам внутрикластерных дисперсий (относительно центров кластеров):

$$F_3 = \sum_k \sum_{i=1}^{N_k} \sigma_{k,i}^2 = \sum_k \frac{1}{N_k} \sum_{i=1}^{N_k} (x_{k,i} - c_k)^2 = \sum_k \frac{1}{N_k} \sum_{i=1}^{N_k} \left( (\nu_{k,i} - \nu_{c_k})^2 + (E_{k,i} - E_{c_k})^2 \right)$$

Оптимальным следует считать разбиение, при котором сумма внутрикластерных (внутригрупповых) дисперсий будет минимальной.

**Постановка задачи.**

Дано конечное множество из объектов, представленных двумя признаками (в качестве этого множества принимаем исходную двумерную выборку, сформированную ранее в лабораторной работе №4). Выполнить разбиение исходного множества объектов на конечное число подмножеств (кластеров) с использованием метода поиска сгущений. Полученные результаты содержательно проинтерпретировать.

**Выполнение работы.**

Для выполнения данной работы была использована выборка, сформированная в первой лабораторной работе.

Нам нужно корректно реализовать методы кластерного анализа, для чего мы нормализуем нашу выборку по формуле:

$$z = \frac{x - \bar{x}}{\sigma_x}$$

Где $\sigma_x$ – стандартное отклонение переменной, а $\bar{x}$ – ее среднее значение.

Нормализированное множество представлено на рис.1 в таблице, а также на рис.2 на диаграмме рассеяния.

| | 0 | 1 |
|---|---|---|
| **0** | -0.893663 | -1.027792 |
| **1** | 1.325225 | 1.241278 |
| **2** | -1.333615 | -1.477776 |
| **3** | -1.716182 | -0.826735 |
| **4** | -0.147657 | 0.049298 |
| **...** | ... | ... |
| **109** | 1.000043 | 0.781720 |
| **110** | 0.368808 | 0.714701 |
| **111** | 0.789631 | 1.222130 |
| **112** | 1.765177 | 1.806152 |
| **113** | 0.158396 | -0.314519 |

114 rows × 2 columns

Рисунок 1 – таблица, сгенерированная программой

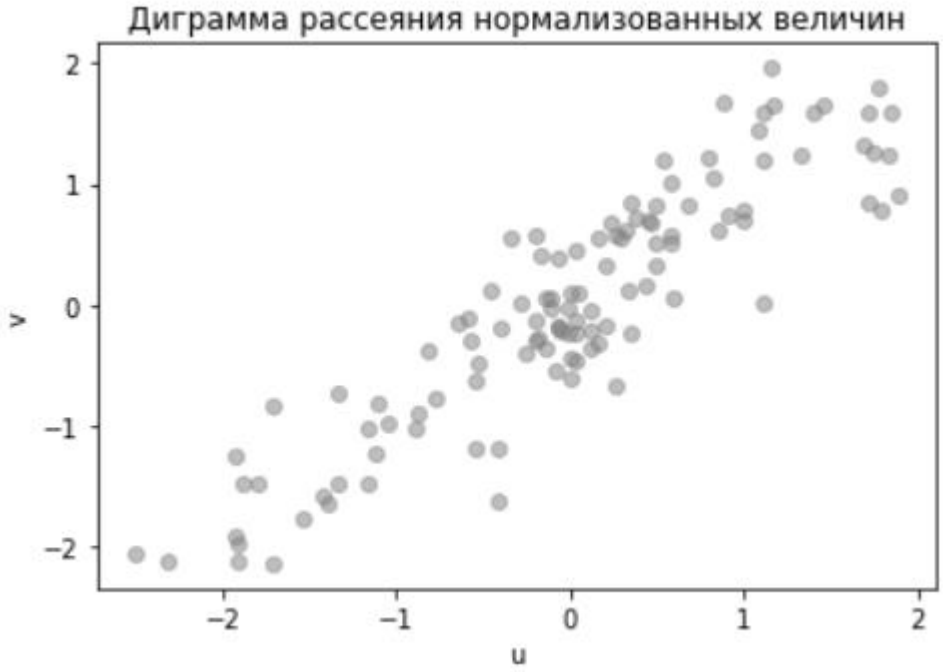



Рисунок 2 – диаграмма рассеяния нормализованных величин

Реализуем алгоритм поиска сгущений. Отобразим полученные кластеры, выделим каждый кластер разным цветом, отметим центроиды.

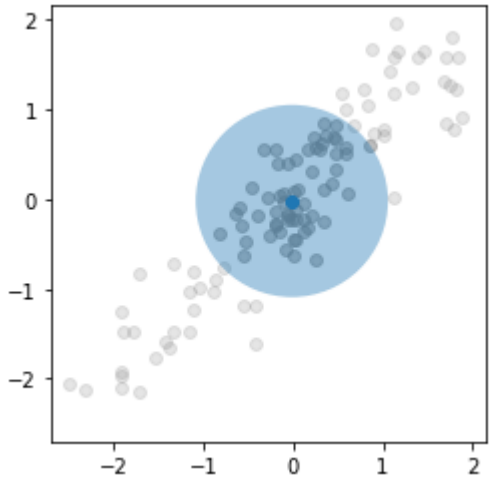
Определим нижнюю и верхнюю границы радиуса сферы:

$$R_{min} = mind_{ij} = 0.0213905953681577777;$$

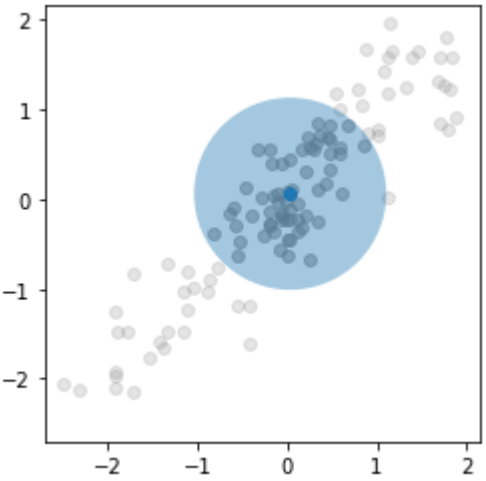
$$R_{max} = maxd_{ij} = 5.751746303724612.$$

Сам алгоритм поиска сгущений описан в основных теоретических положениях. Визуально представим его на примере радиуса, входящего в промежуток, $R$=1.0719558085668415 (при дальнейшей проверке устойчивости соединений, мы опустим промежуточные шаги, для удобства прочтения отчета).

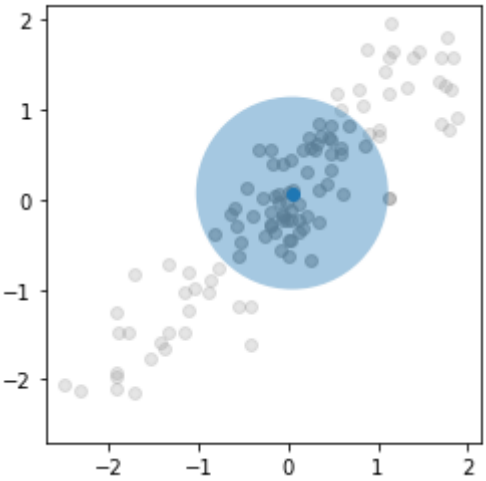
Формирование 1го кластера



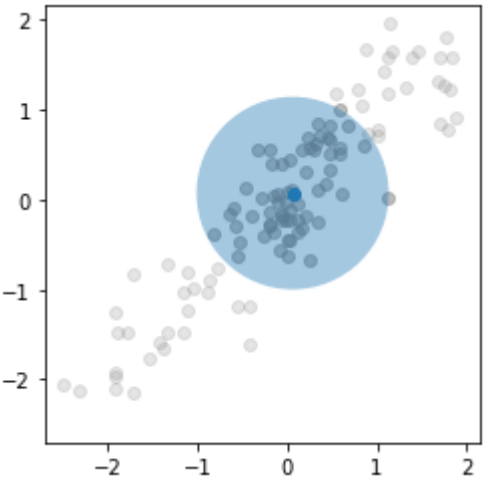

*1.1*



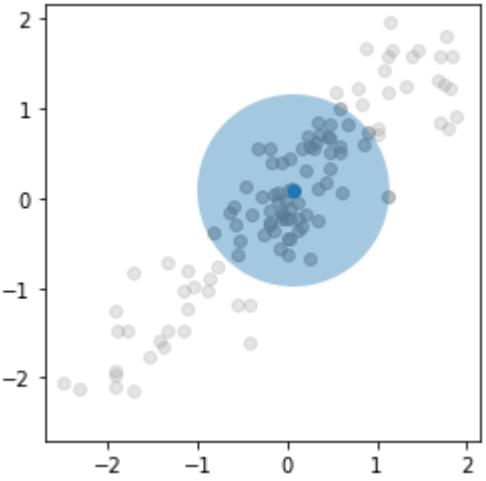

*1.2*



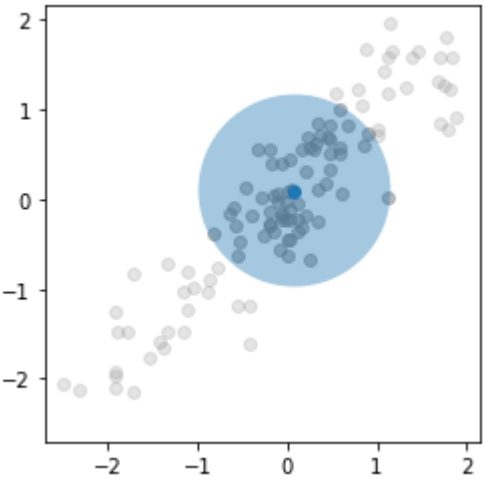

*1.3*




*1.4*

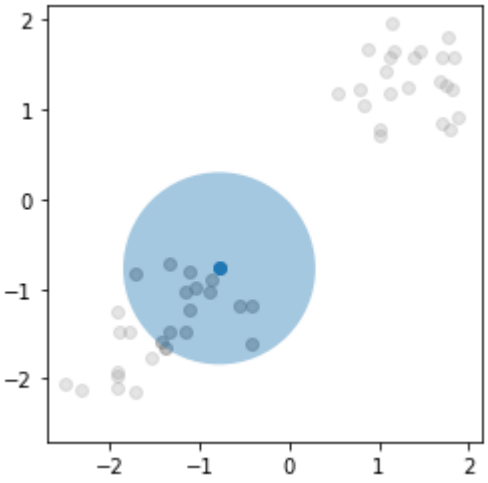



*1.5*

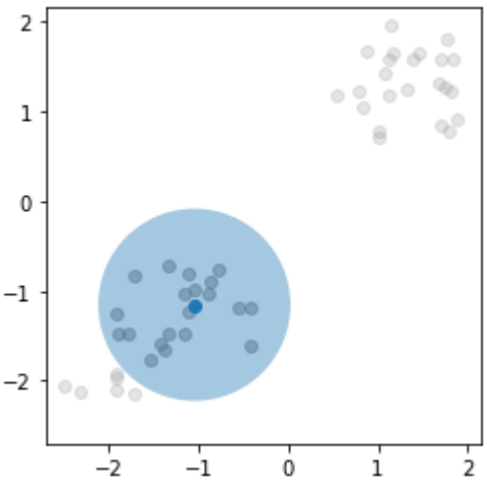



*1.6*

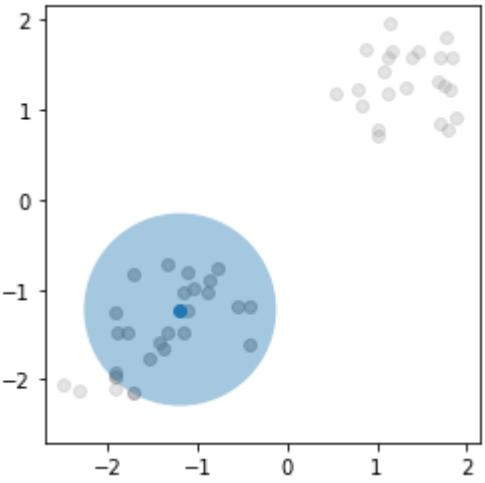
Формирование 2 кластера



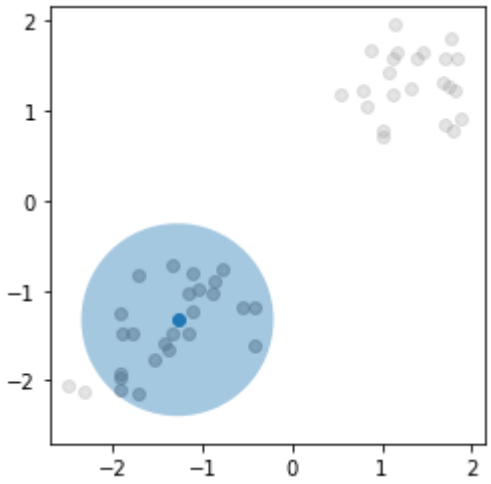

*2.1*



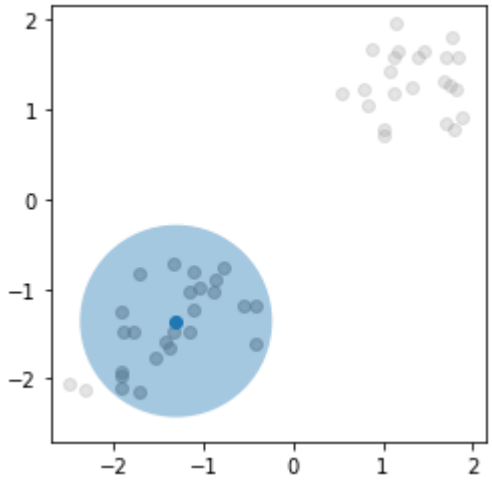

*2.2*



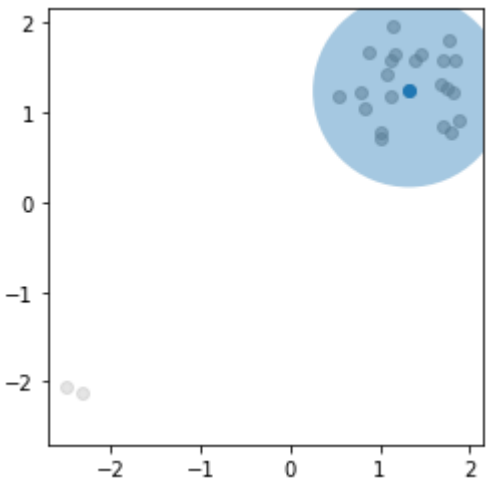
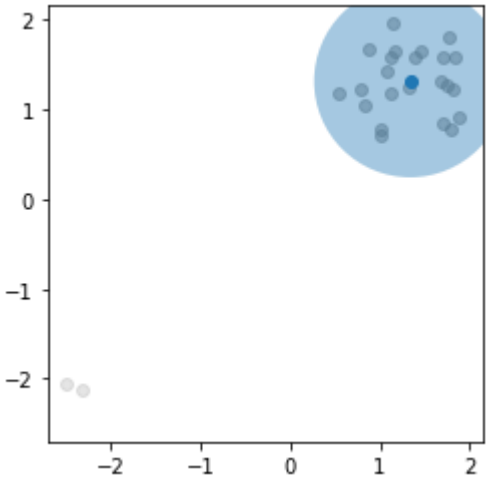

*2.3*

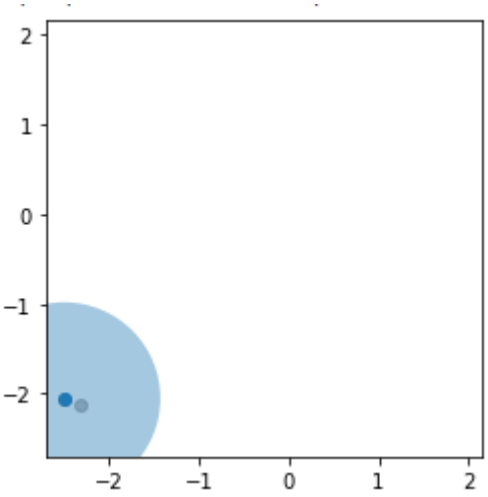
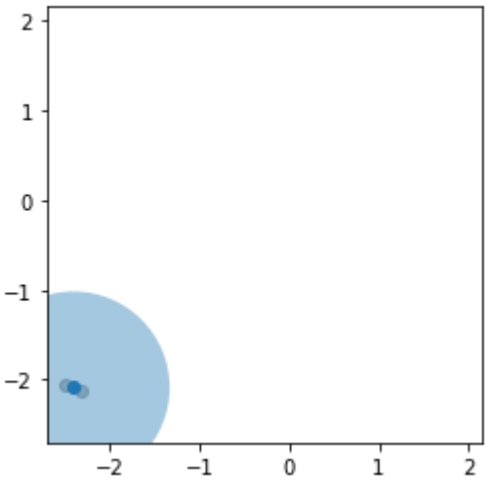



*2.4*




*2.5*

| Формирование 3го кластера |
| --- |
|  *3.1*  *3.2* |

| Формирование 4го кластера |
| --- |
|  *4.1*  *4.2* |

Результат кластеризации представлен на рис. 3.

```
Число кластеров: 4
F1: 38.32866626439944, F2: 1921.4228010455267, F3: 1.0352791085740767
```

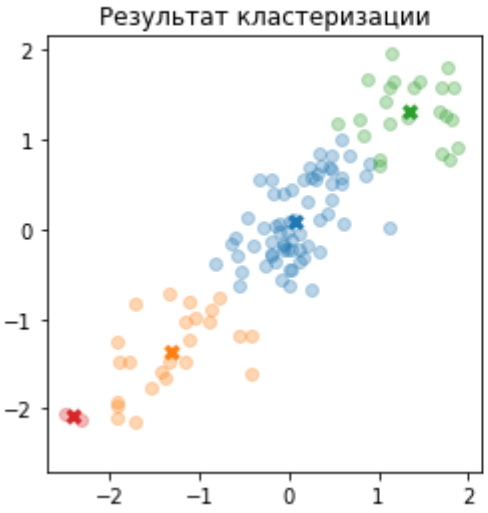



Рисунок 3 – Результат кластеризации

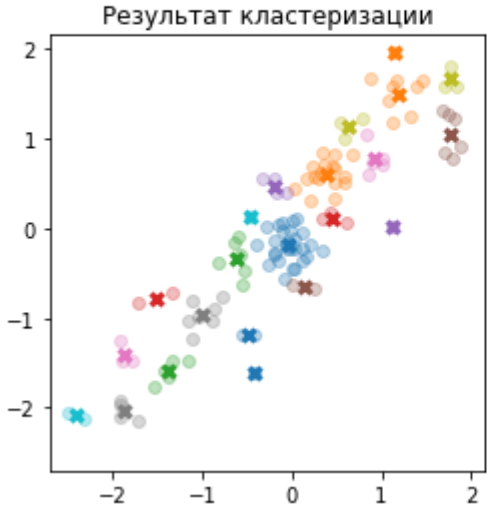
Таблица 1

| Номер кластера | Центр кластера | Количество элементов в кластере |
|---|---|---|
| 1 | (0.08217701526223979, 0.09878930409319236) | 65 |
| 2 | (-1.3057191761513063, -1.3567031373700817) | 24 |
| 3 | (1.339363393379832, 1.3180793373874455) | 23 |
| 4 | (-2,1754394392808516 ; -2,465932878978102) | 2 |

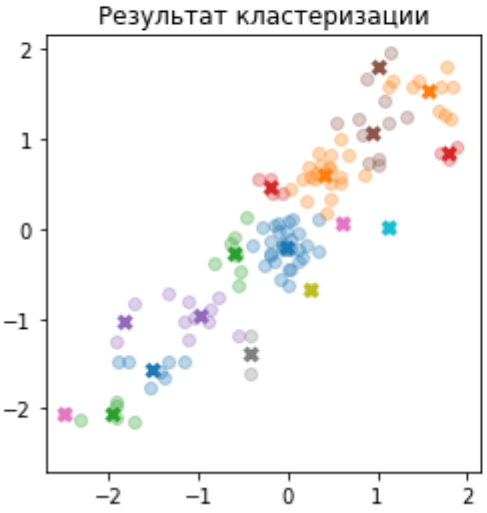
Далее мы выбираем величину, на которую будем изменять радиус окружности, чтобы проверить устойчивость разбиения. Ближайшие к границам значения в дальнейших примерах мы опустим, так как при них создавались либо слишком большие кластеры, куда входила почти вся выборка, либо слишком маленькие, куда входило 1-3 элемента, что является нецелесообразной кластеризацией.

Ниже приведены вычисления разбиения на кластеры с постепенным увеличением радиуса. Для каждого из которых приведено 3 функционала качества.

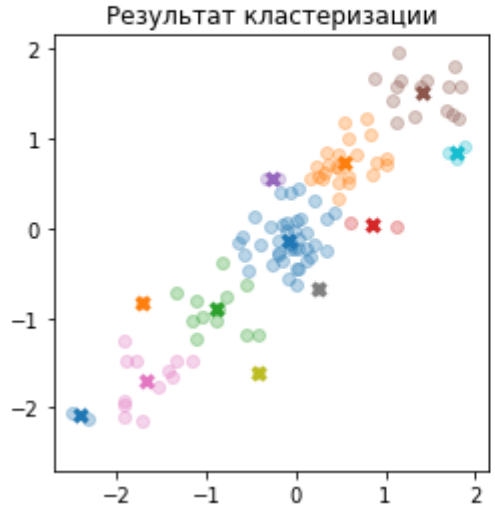
```
Выбранный радиус кластера: 0.4034143092585881
Число кластеров: 22
F1: 4.450114335400133, F2: 190.34063733373287, F3: 0.521947284557638
```



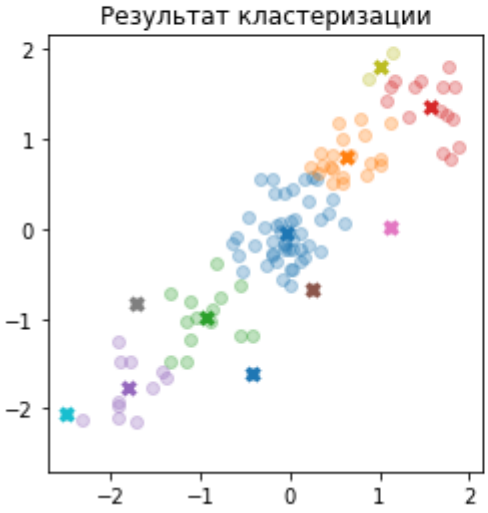

```
Выбранный радиус кластера: 0.4989202377311957
Число кластеров: 17
F1: 7.508712276391269, F2: 279.0524174132914, F3: 0.7577992928985261
```



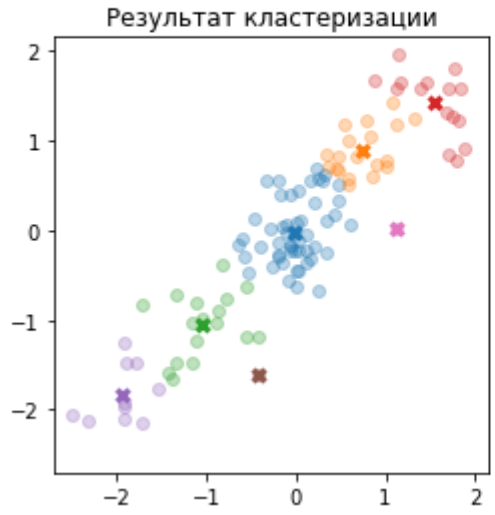
```
Выбранный радиус кластера: 0.5944261662038033
Число кластеров: 12
F1: 13.13241617587502, F2: 580.8517602492781, F3: 0.7428716049259078
```



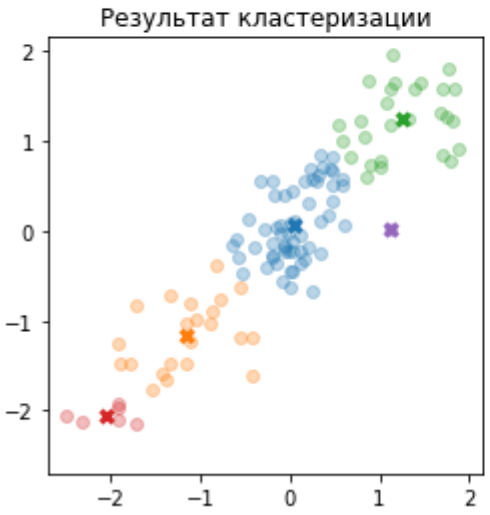

```
Выбранный радиус кластера: 0.6899320946764109
Число кластеров: 11
F1: 17.14281306632721, F2: 798.1919785163991, F3: 0.8115777773607646
```



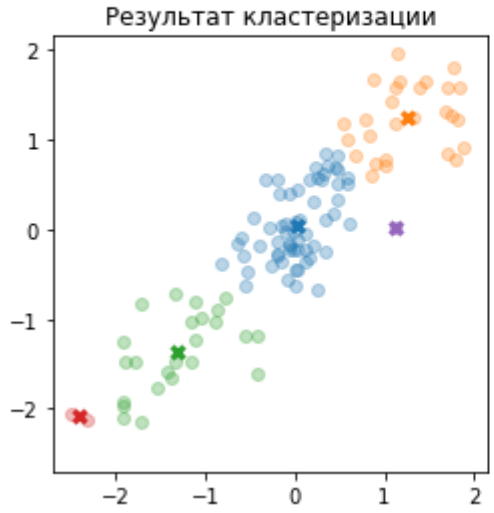
```
Выбранный радиус кластера: 0.7854380231490186
Число кластеров: 7
F1: 22.036308983731626, F2: 993.5196194508609, F3: 0.9591392726973078
```



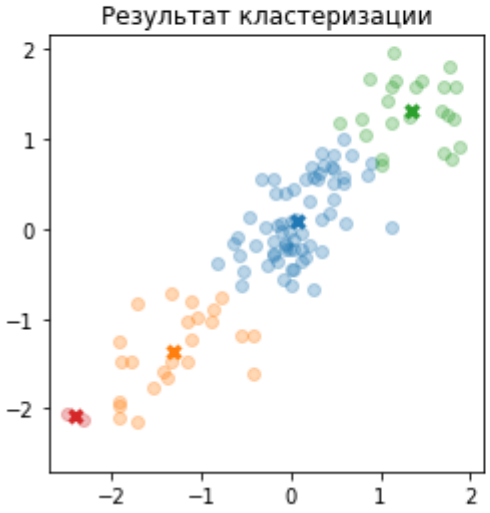

```
Выбранный радиус кластера: 0.8809439516216262
Число кластеров: 5
F1: 31.7209342889367, F2: 1505.207603222468, F3: 0.9999038935764566
```



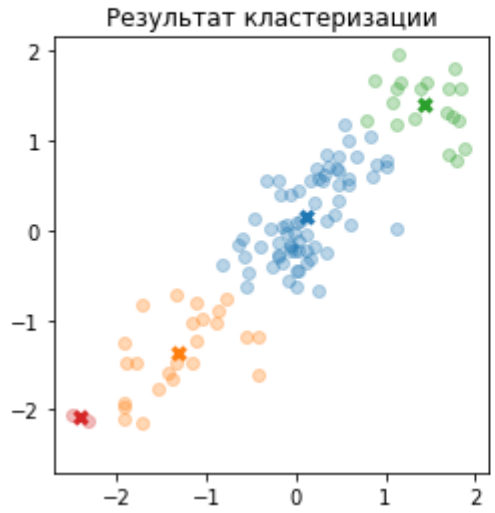
```
Выбранный радиус кластера: 0.9764498800942338
Число кластеров: 5
F1: 35.28609636933005, F2: 1666.604660038646, F3: 1.0179869824822763
```



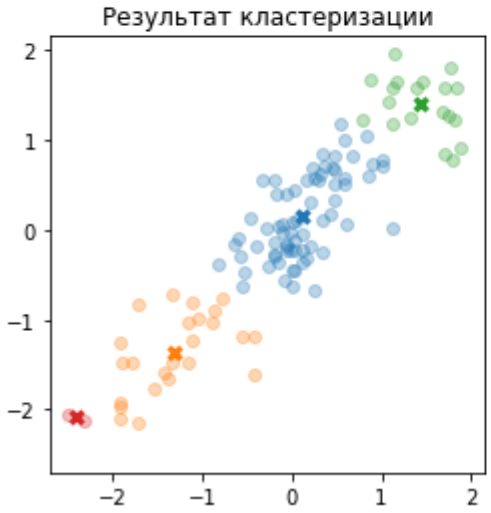

```
Выбранный радиус кластера: 1.0719558085668415
Число кластеров: 4
F1: 38.32866626439944, F2: 1921.4228010455267, F3: 1.0352791085740767
```



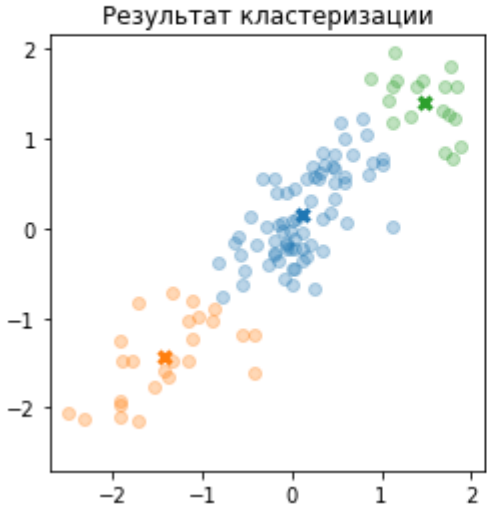
```
Выбранный радиус кластера: 1.167461737039449
Число кластеров: 4
F1: 41.2201386300698, F2: 2166.731069633311, F3: 1.0316668365470598
```




```
Выбранный радиус кластера: 1.2629676655120567
Число кластеров: 4
F1: 41.2201386300698, F2: 2166.731069633311, F3: 1.0316668365470598
```

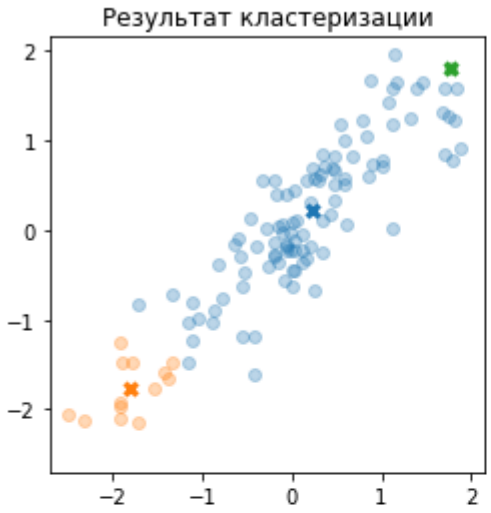


```
Выбранный радиус кластера: 1.3584735939846644
Число кластеров: 3
F1: 46.386869051767135, F2: 2375.949888724397, F3: 1.1217040163657144
```

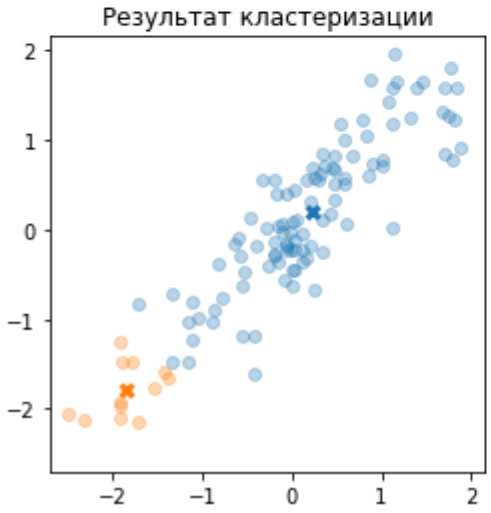



Далее промежуток радиусов, которые разделяют выборку на 3 кластера, вплоть до радиуса *R= 2.31353287871074*

```
Выбранный радиус кластера: 2.2180269502381327
Число кластеров: 3
F1: 126.98483987339817, F2: 6659.845847464388, F3: 1.4377015706821323
```




```
Выбранный радиус кластера: 2.31353287871074
Число кластеров: 2
F1: 136.8419312977539, F2: 7114.7850526322, F3: 1.5009579110496676
```




Оценив приведенные результаты можно сделать вывод, что малые радиусы окружности порождают неустойчивые разбиения. Можно сделать предположение, что это вызвано очень малым расстоянием наблюдений, близких к выборочным средним. При всем перечисленном, значения функционалов качества до $R=0.6899320946764109$ меньше, чем было получено при использовании метода k-means. Можно увидеть, что при увеличении радиуса увеличиваются значения функционалов качества, но при этом мы наблюдаем более устойчивые разбиения.

**Выводы.**

Таким образом, были освоены основные понятия кластерного анализа и реализован метод поиска сгущений. Для применения алгоритма были применены различные радиусы окружностей для проверки соединений на устойчивость. В следствии чего, было сделано предположение о том, что при увеличении радиуса, увеличивается и устойчивость соединения. Но в связи с не таким плотным расположением по краям и сгущением около выборочных средних, значение радиуса сильно влияет на состав кластеров. Функционалы качества показывают нам, что качество разбиения значительно снижается от увеличения радиуса, не смотря на устойчивость соединений.

При сравнении реализованных методов кластеризации, можно сказать о том, что при больших значениях радиуса окружности, метод поиска сгущений дает худшие функционалы качества, чем при методе k-means. Также этот метод является более медленным, чем k-means. Преимущество k-means в том, что можно оценить число кластеров. Лишь при малых значениях радиуса окружности можно наблюдать меньшие показатели функционалов качества, чем было приведено в предыдущей работе при использовании.