Data learning

Курс “Машинное обучение”

Лабораторная работа

Исследование применимости алгоритмов k-means и DBSCAN с помощью различных критериев кластеризации

Харитонов Е.А., М16-524

Вариант 3-09

2017

Исходные данные (тестовые)

Данные: Классифицированные объекты, характеризуемые двумя признаками x и y

Размер выборки: 770

Классы: 10 классов

Представление: данные удобно визуализировать в качестве точек на координатной плоскости с координатами (x,y)

p	x	y	k
1	0.11	0.48	1
2	0.106	0.369	1
3	0.029	0.273	1
4	0.215	0.368	1
5	0.145	0.437	1
6	0.154	0.461	1
7	0.068	0.426	1
8	0.148	0.608	1
9	0.149	0.29	1
10	0.177	0.424	1
11	0.086	0.398	1
12	0.09	0.357	1
13	0.131	0.482	1
14	0	0.42	1
15	0.178	0.426	1
16	0.118	0.38	1
17	0.152	0.551	1
18	0.176	0.468	1

Рисунок 1. Пример исходных данных

Визуализация данных

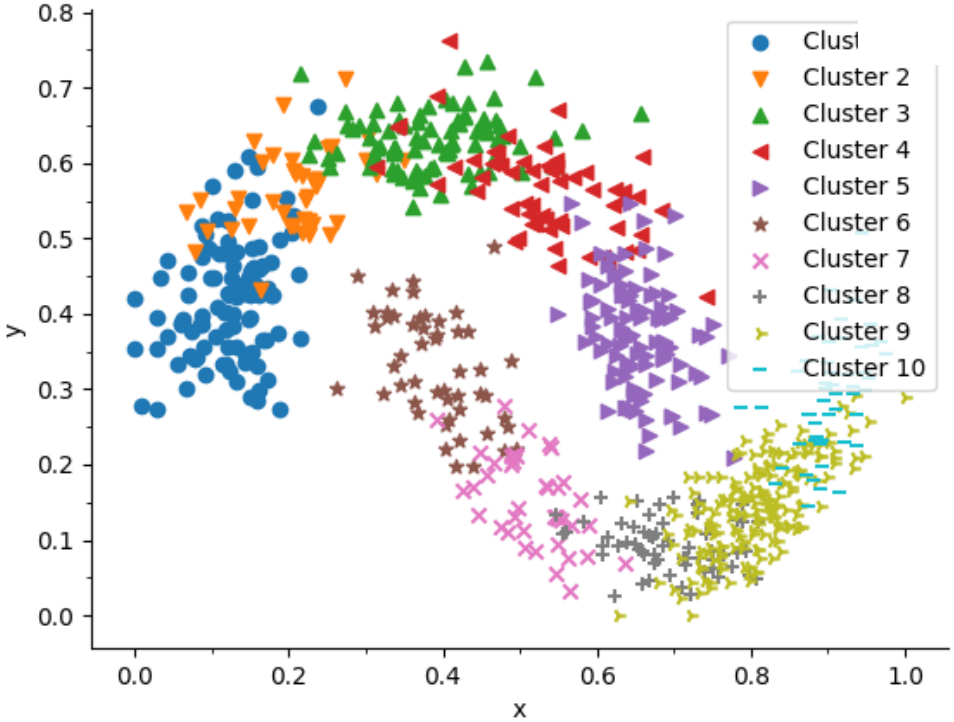


Рисунок 2. Визуализация исходных данных в виде точек на координатной плоскости

Визуализация данных

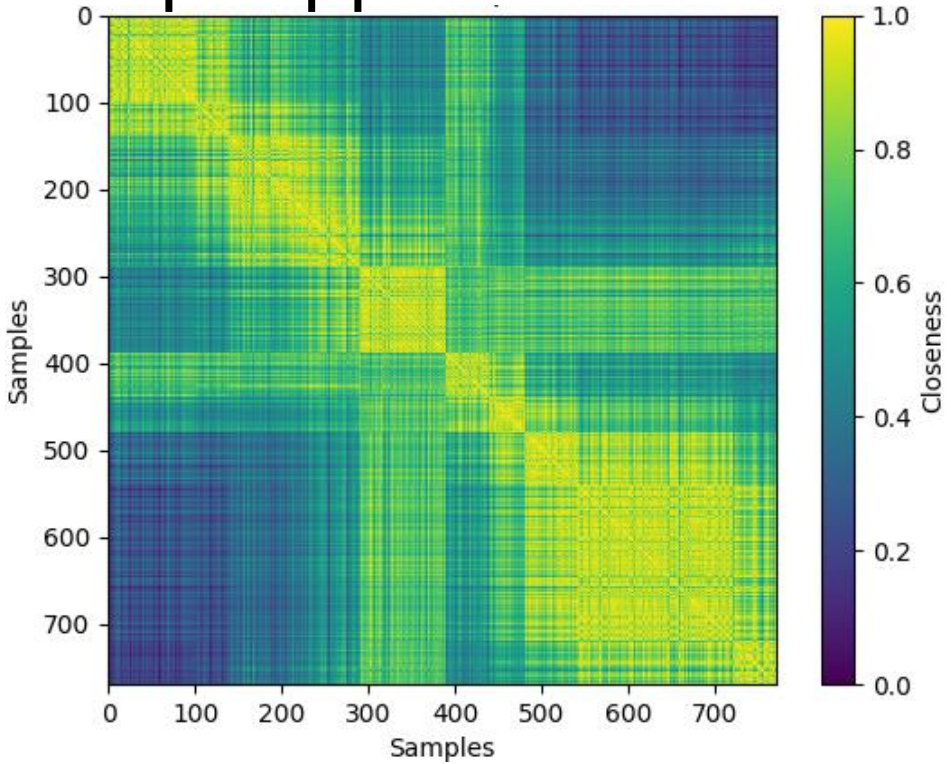
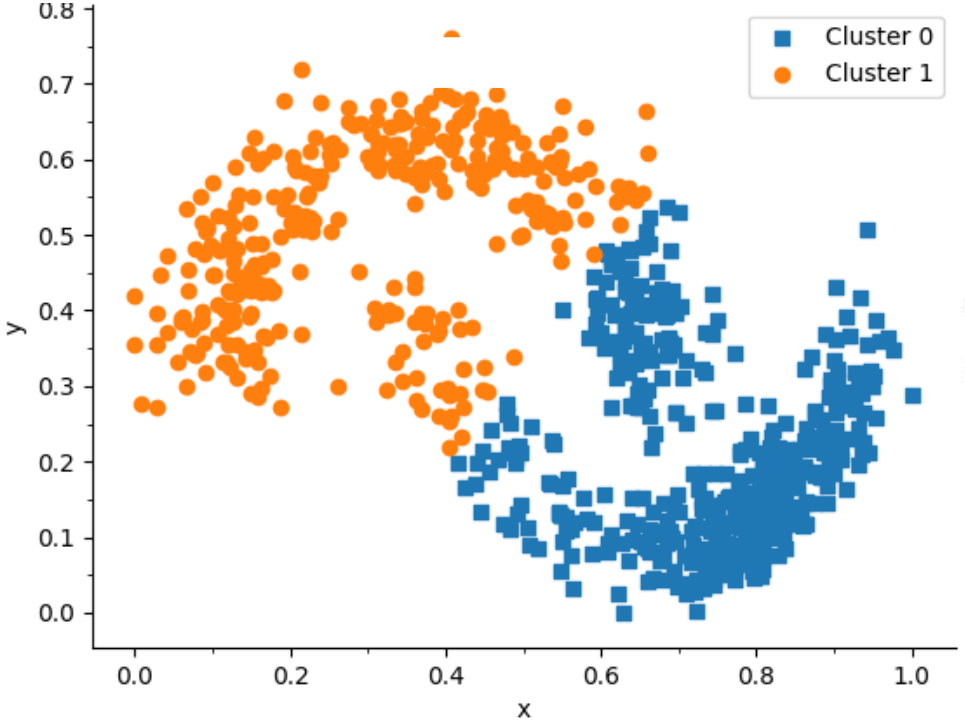


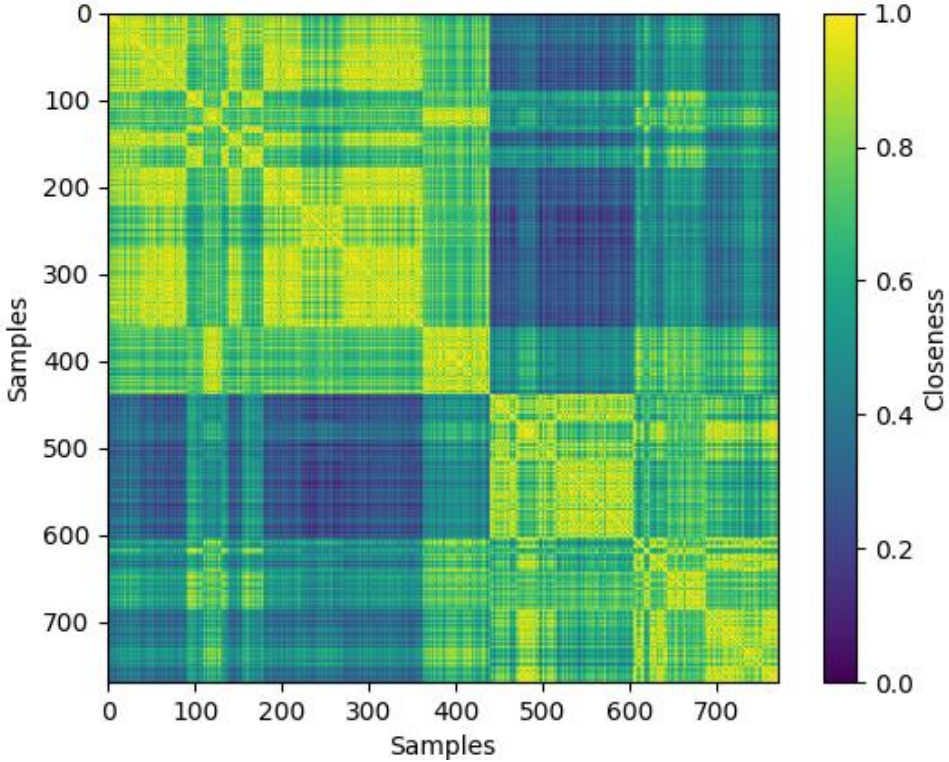
Рисунок 3. Визуализация исходных данных в виде heatmap

Используемые методы

- Алгоритмы:
  - k-means
  - DBSCAN
- Индексы:
  - Суммарное внутрикластерное квадратичное отклонение (S)
  - Silhouette index (SI)
  - Calinski-Harabasz Index (CHI)
  - Adjusted Rand index (ARI)

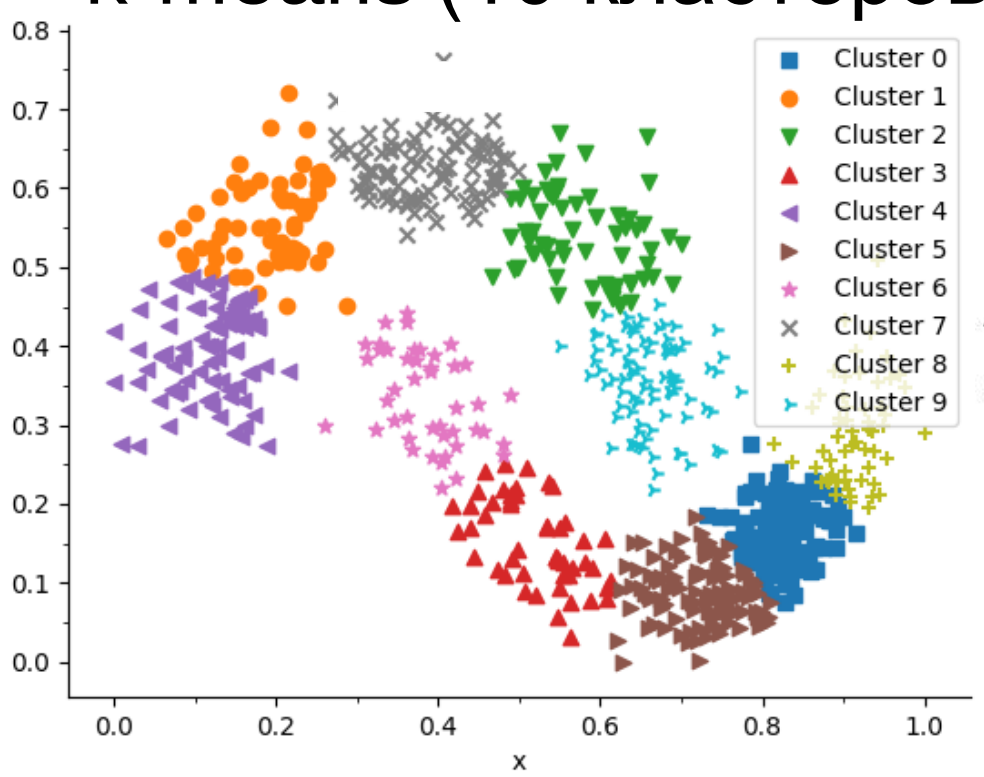
k-means (2 кластера)

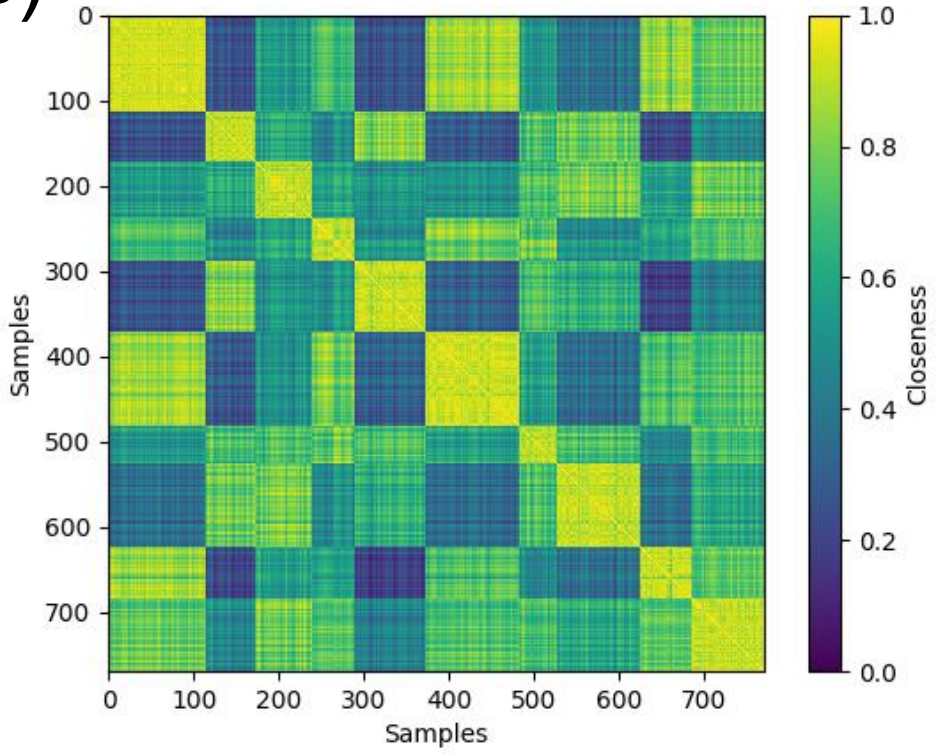




Рисунки 4-5. Кластеризация алгоритмом k-means с K=2. S=26.7, CHI=1488, SI=0.56, ARI(true)=0.24.

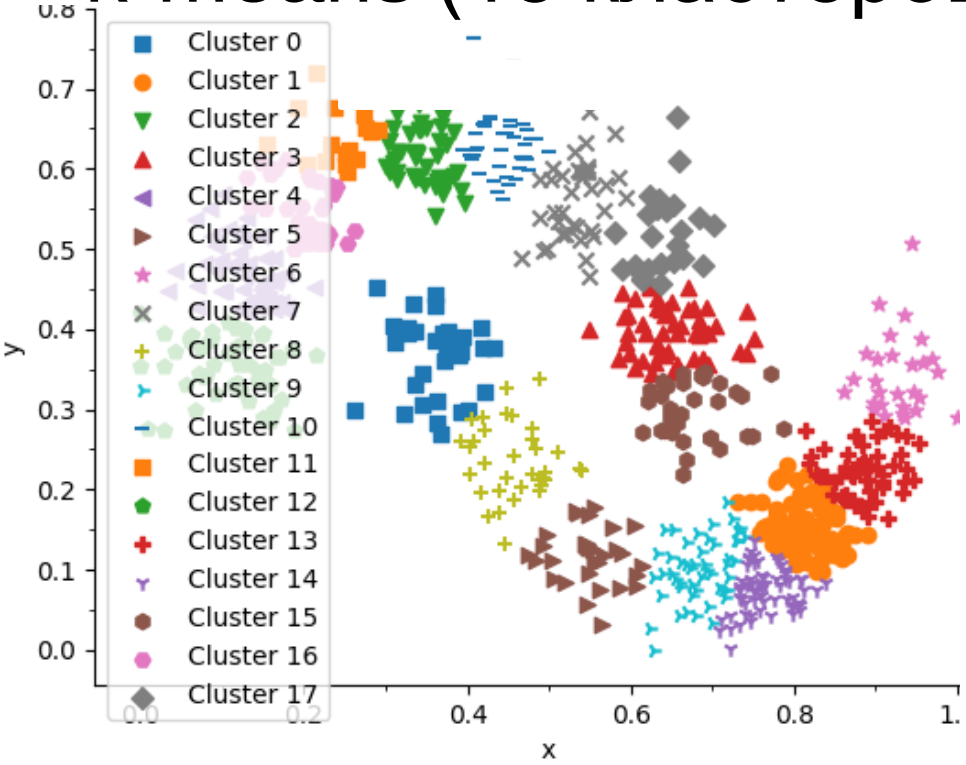
k-means (10 кластеров)

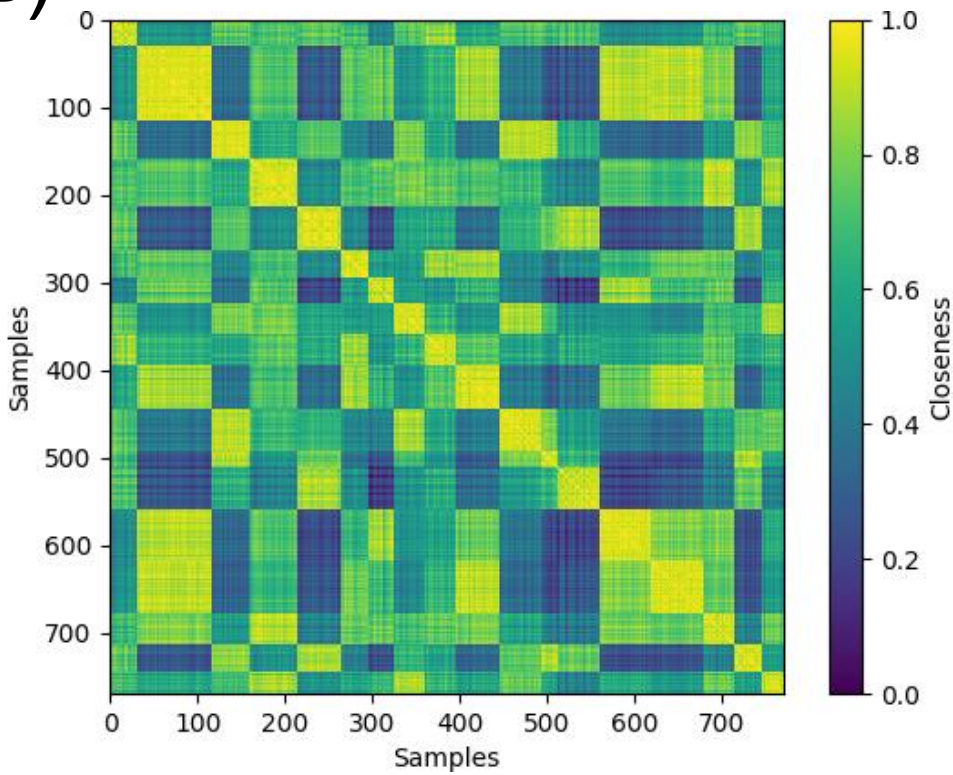




Рисунки 6-7. Кластеризация алгоритмом k-means с K=10. S=3.7, CHI=1695, SI=0.43, ARI(true)=0.59.

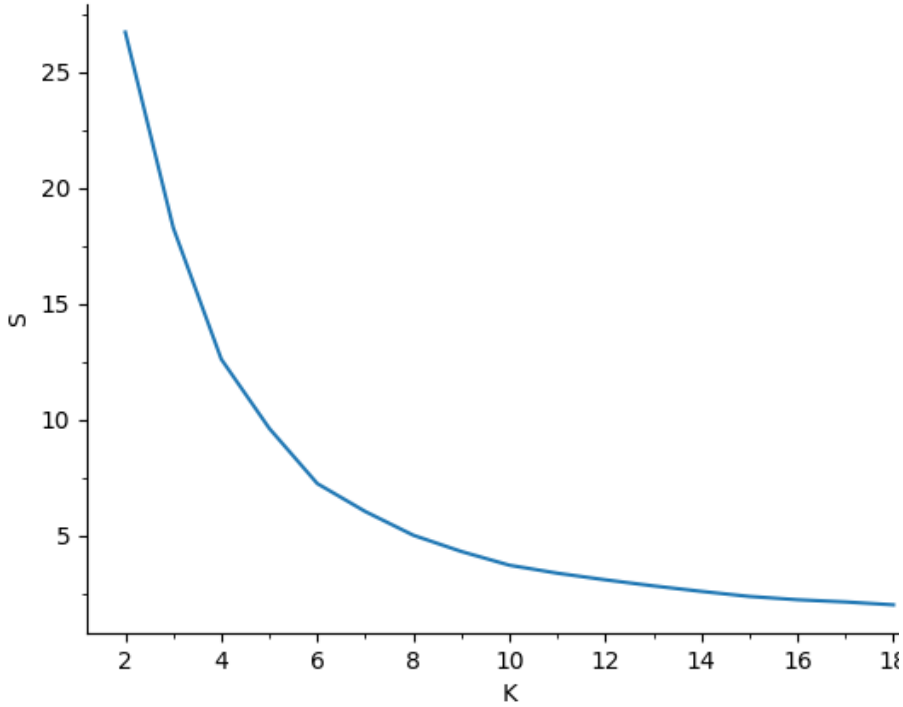
k-means (18 кластеров)

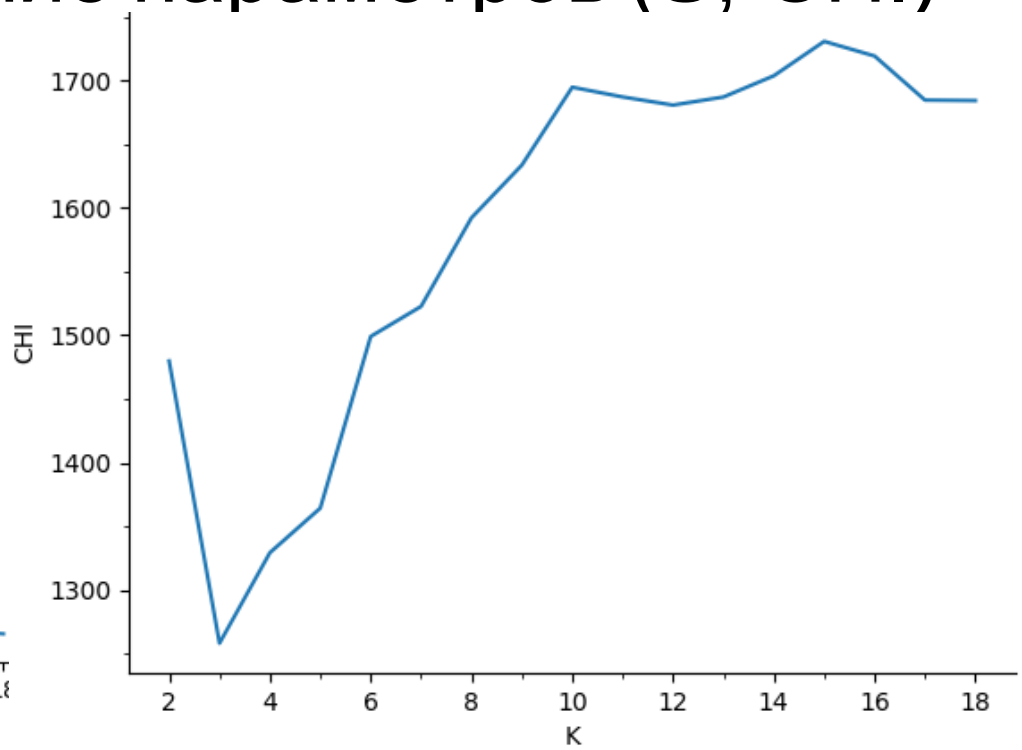




Рисунки 7-8. Кластеризация алгоритмом k-means с K=18. S=2.0, CHI=1672, SI=0.37, ARI(true)=0.42.

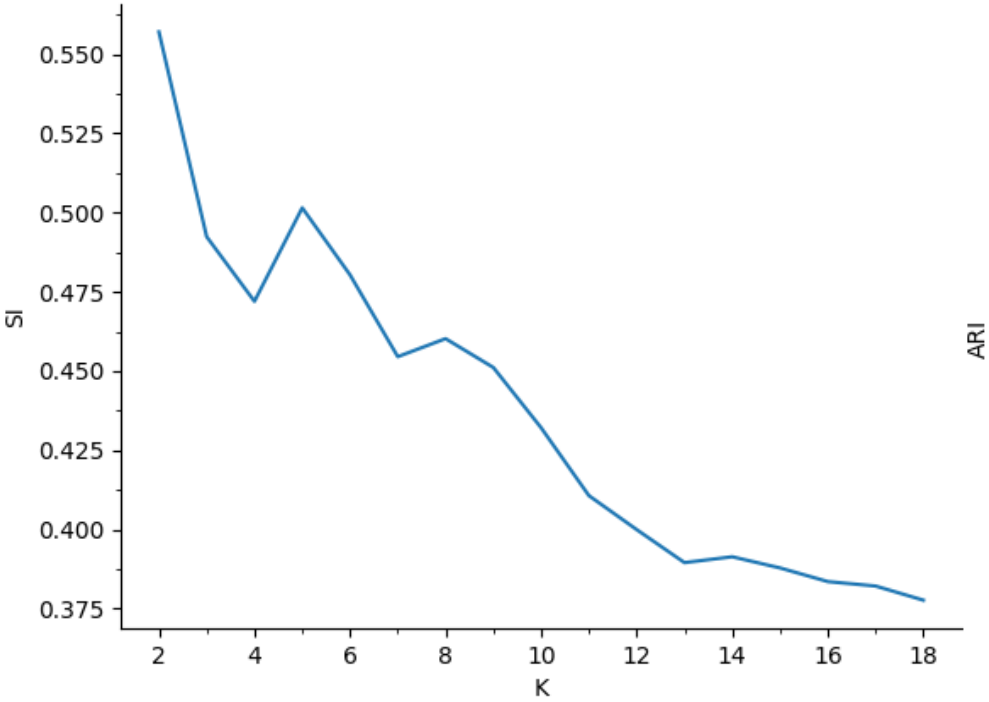
k-means исследование параметров (S, CHI)

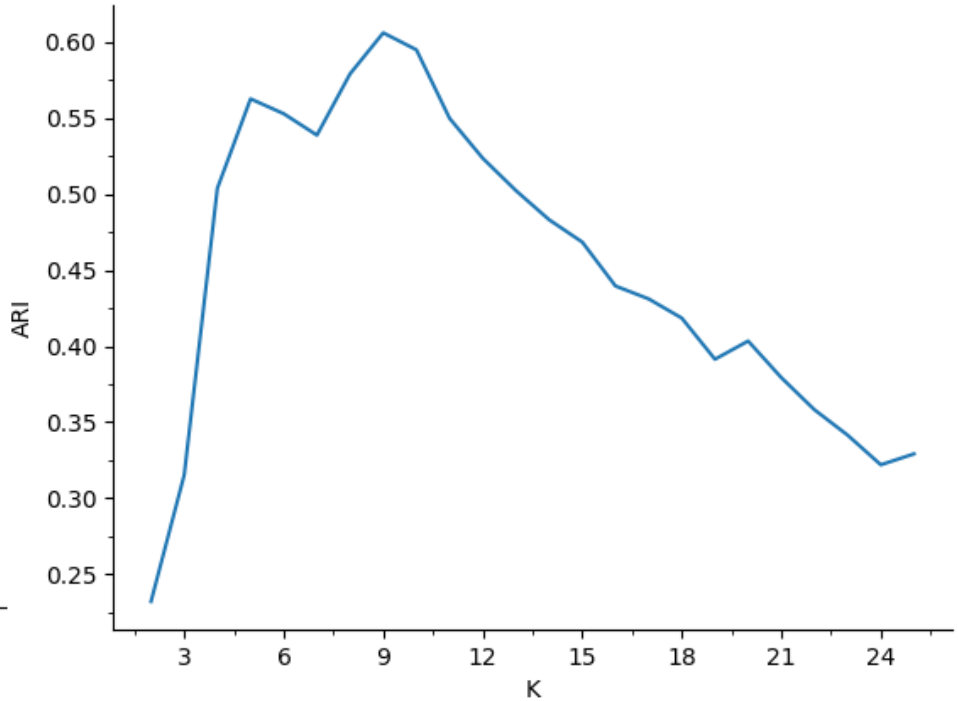




Рисунки 9-10. Зависимость внутрикластерной суммы квадратов отклонений S (слева) и значения индекса Calinski-Harabasz (справа) от количества кластеров.

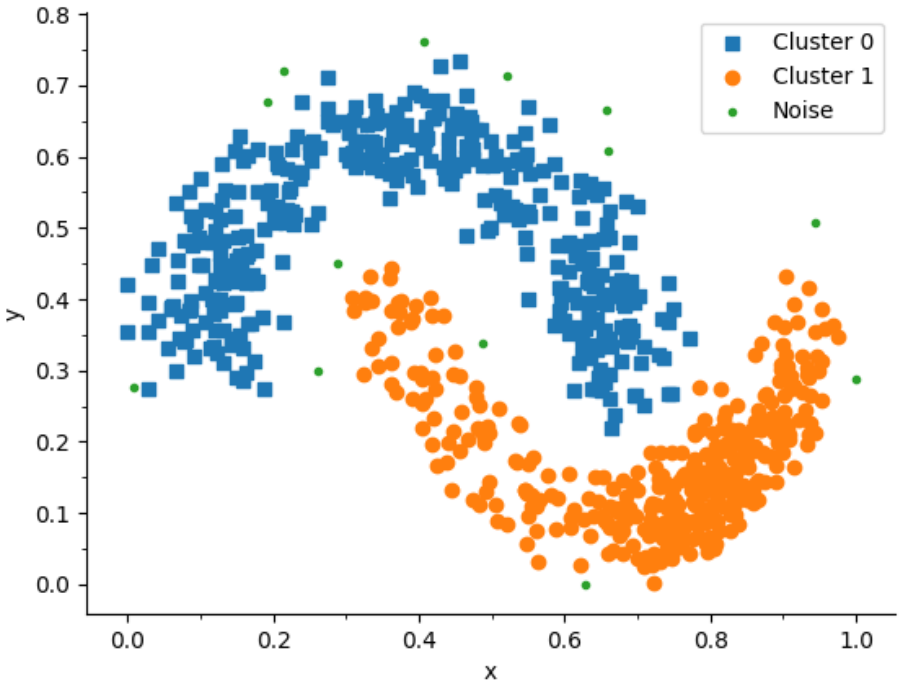
k-means исследование параметров (SI, ARI)

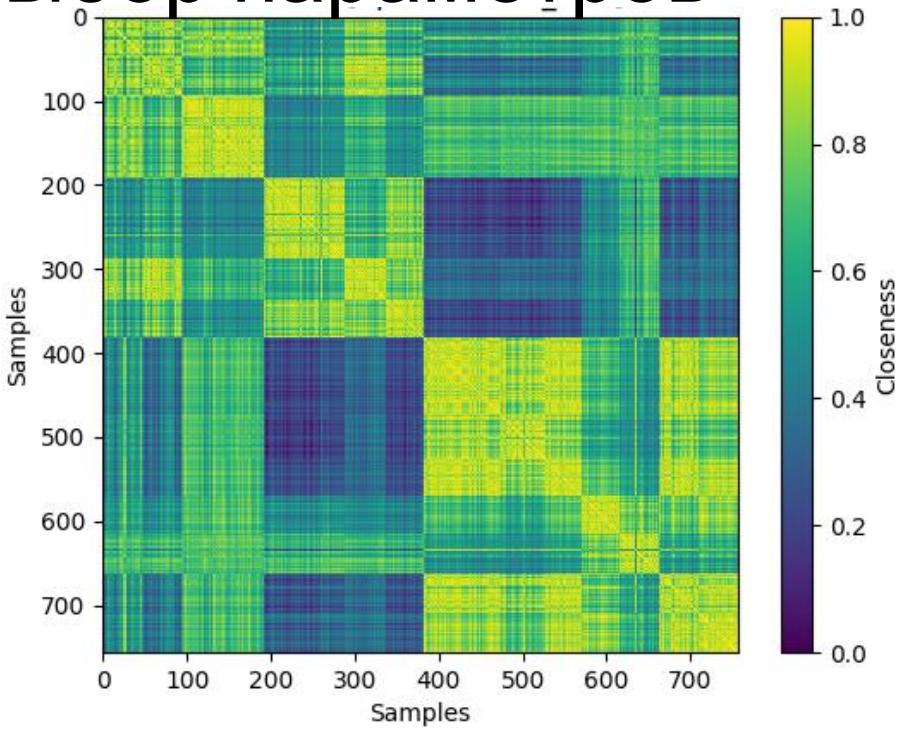




Рисунки 11-12. Зависимость значения индексов Silhouette (слева) и Adjusted Rand Index (по сравнению с реальной кластеризацией) (справа) от количества кластеров.

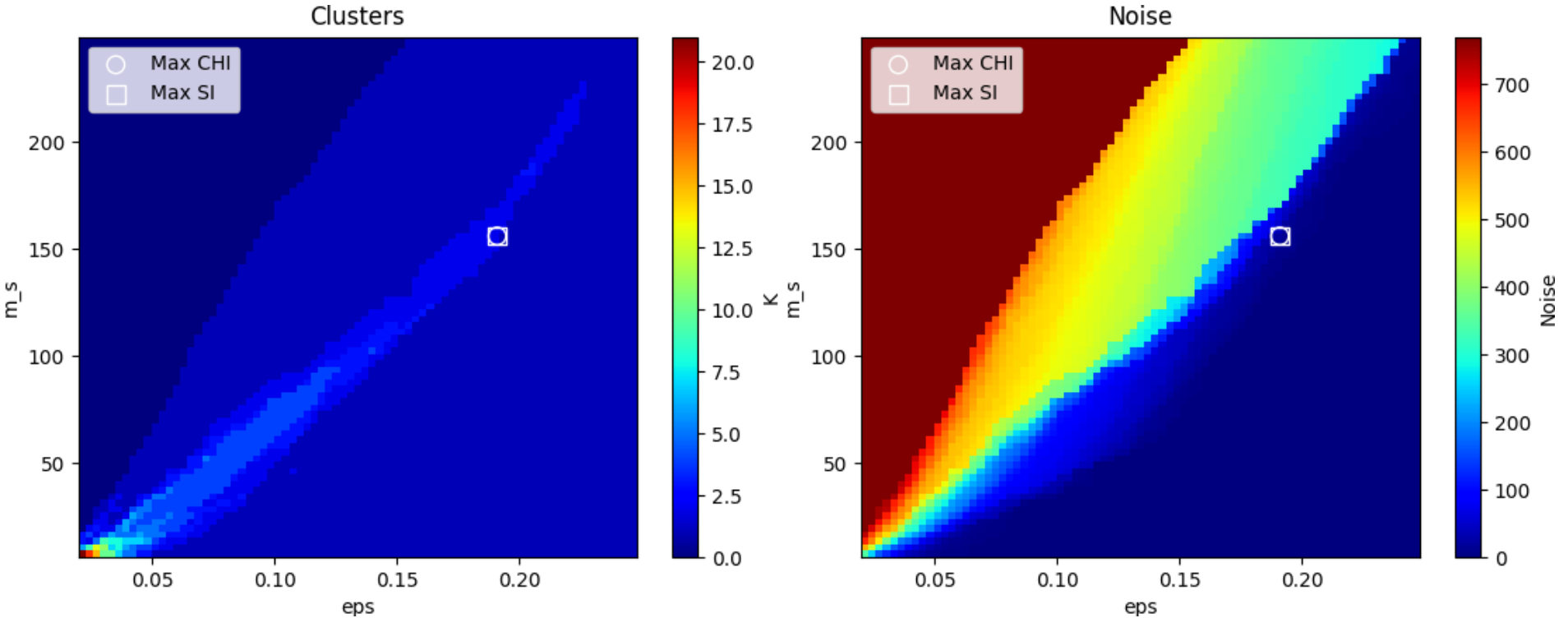
DBSCAN визуальный выбор параметров





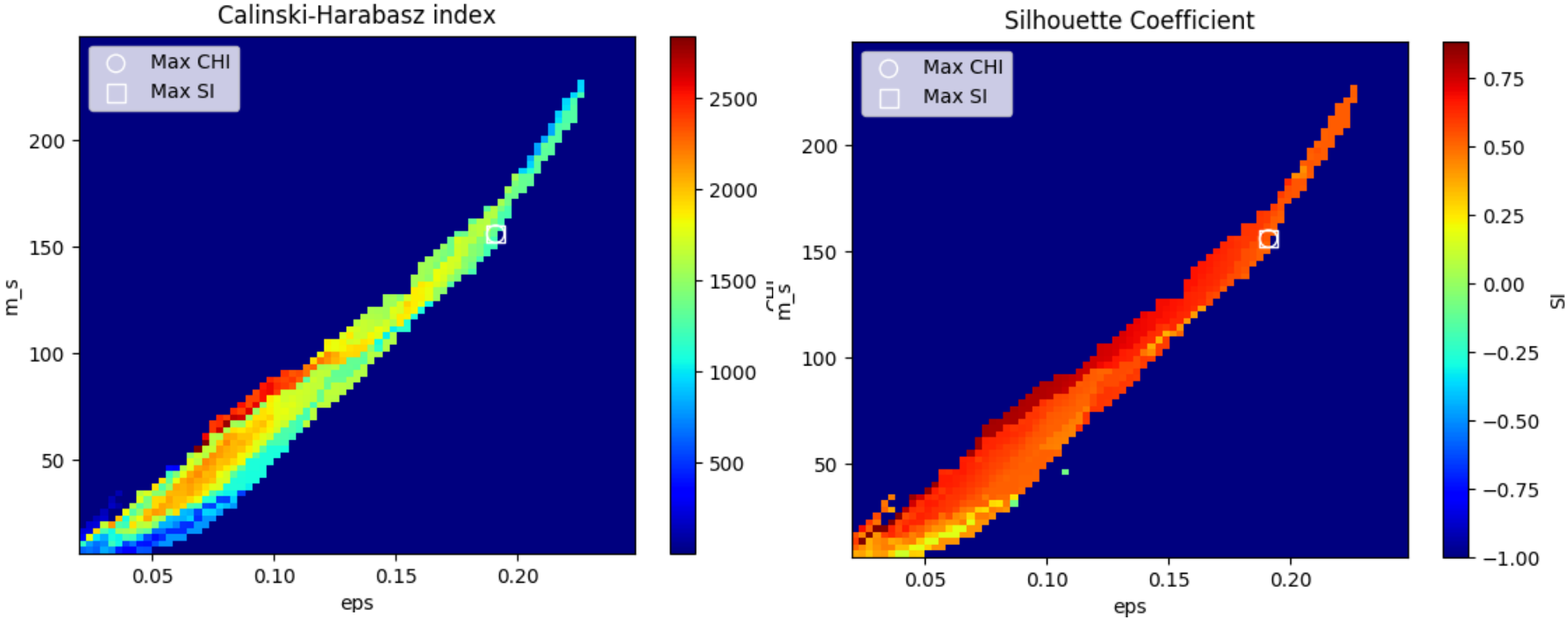
Рисунки 13-14. Кластеризация алгоритмом DBSCAN с параметрами, подобранными визуально. Eps=0.05, m_s=7, K=2, noise=13, CHI=776, SI=0.42, ARI(true)=0.25.

DBSCAN исследование параметров (K, noise)



Рисунки 15-16. Зависимость количества кластеров (слева) и шума (справа) от параметров алгоритма DBSCAN. Лучшие значения индексов при ограничении шума 3%

DBSCAN исследование параметров (CHI, SI)



Рисунки 17-18. Зависимость значения индексов Calinski-Harabasz (слева) и Silhouette (справа) от параметров алгоритма DBSCAN. Лучшие значения индексов при ограничении шума 3%

DBSCAN исследование параметров (ARI)

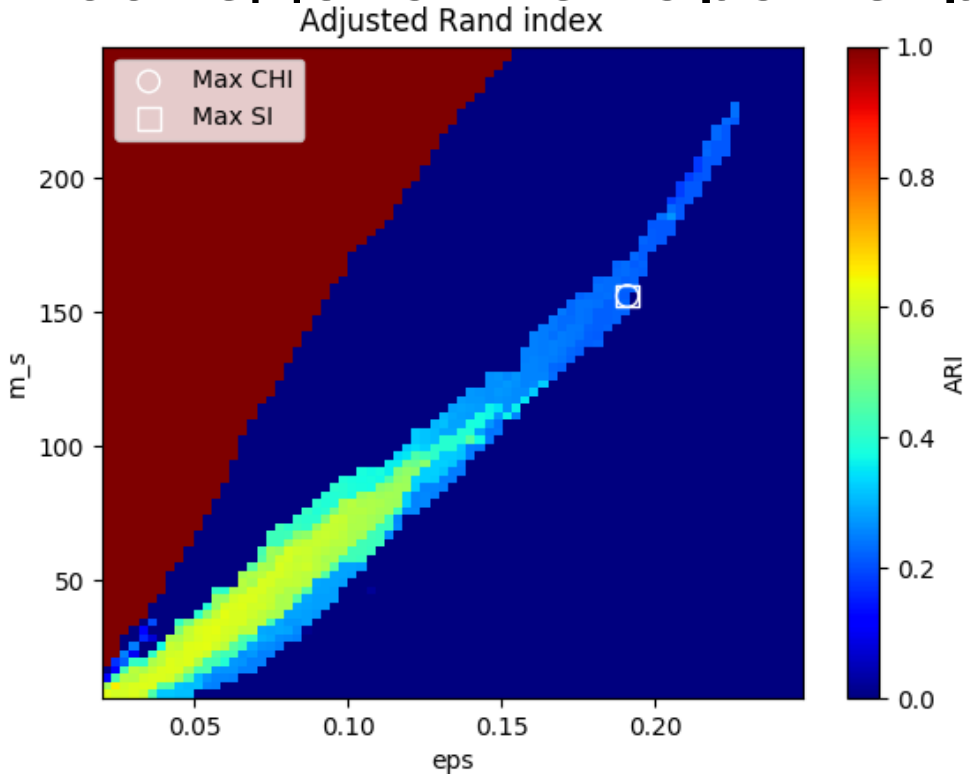
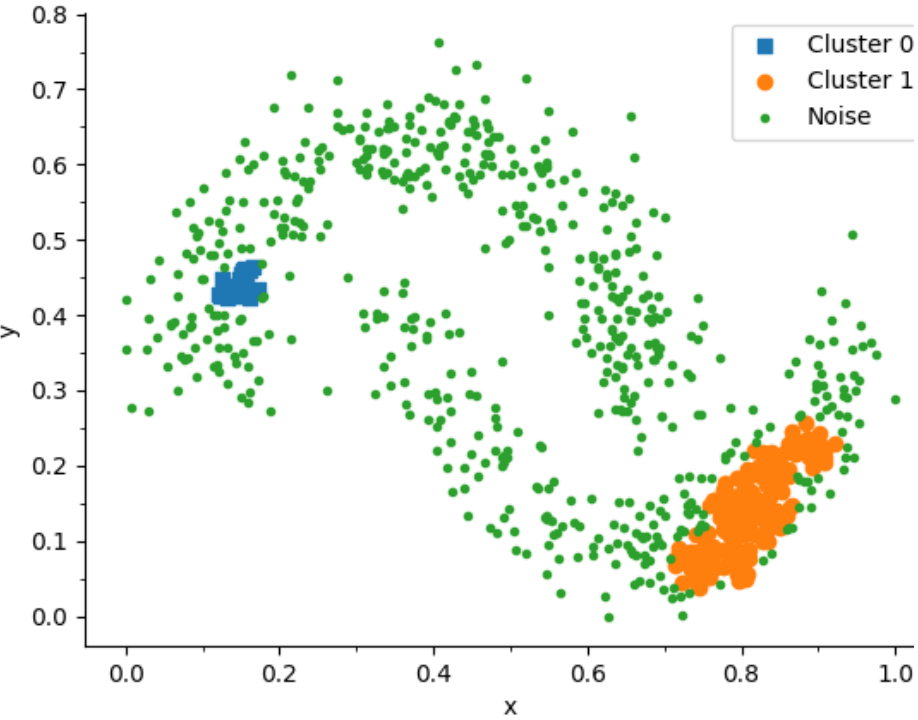
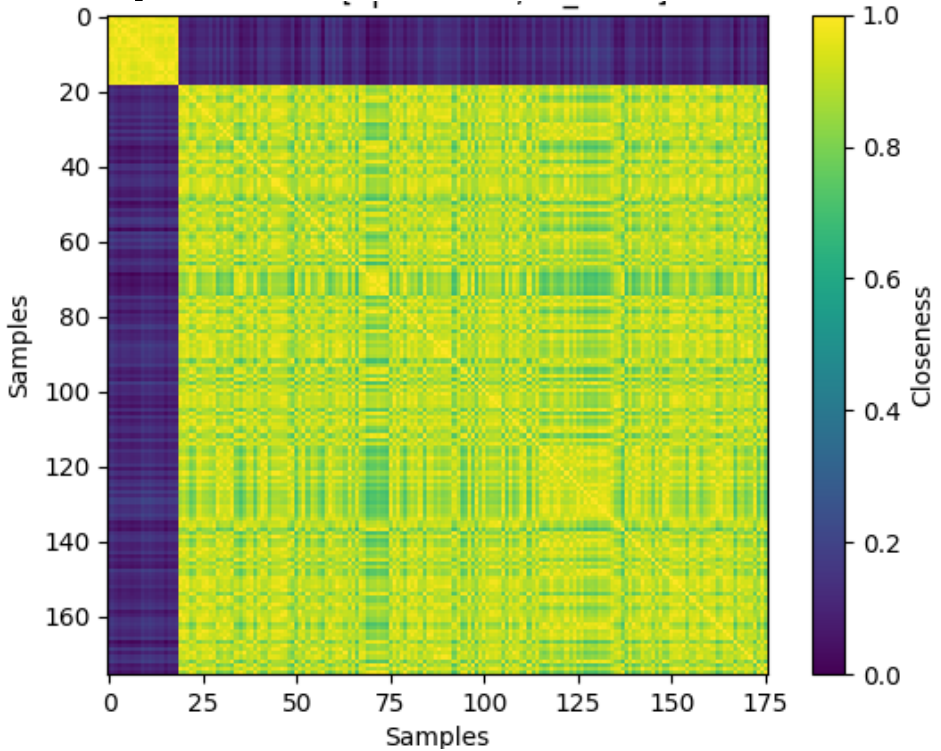


Рисунок 19. Зависимость значения индекса Adjusted Rand Index (по сравнению с реальной кластеризацией) от параметров алгоритма DBSCAN. Лучшие значения индексов при ограничении шума 3%

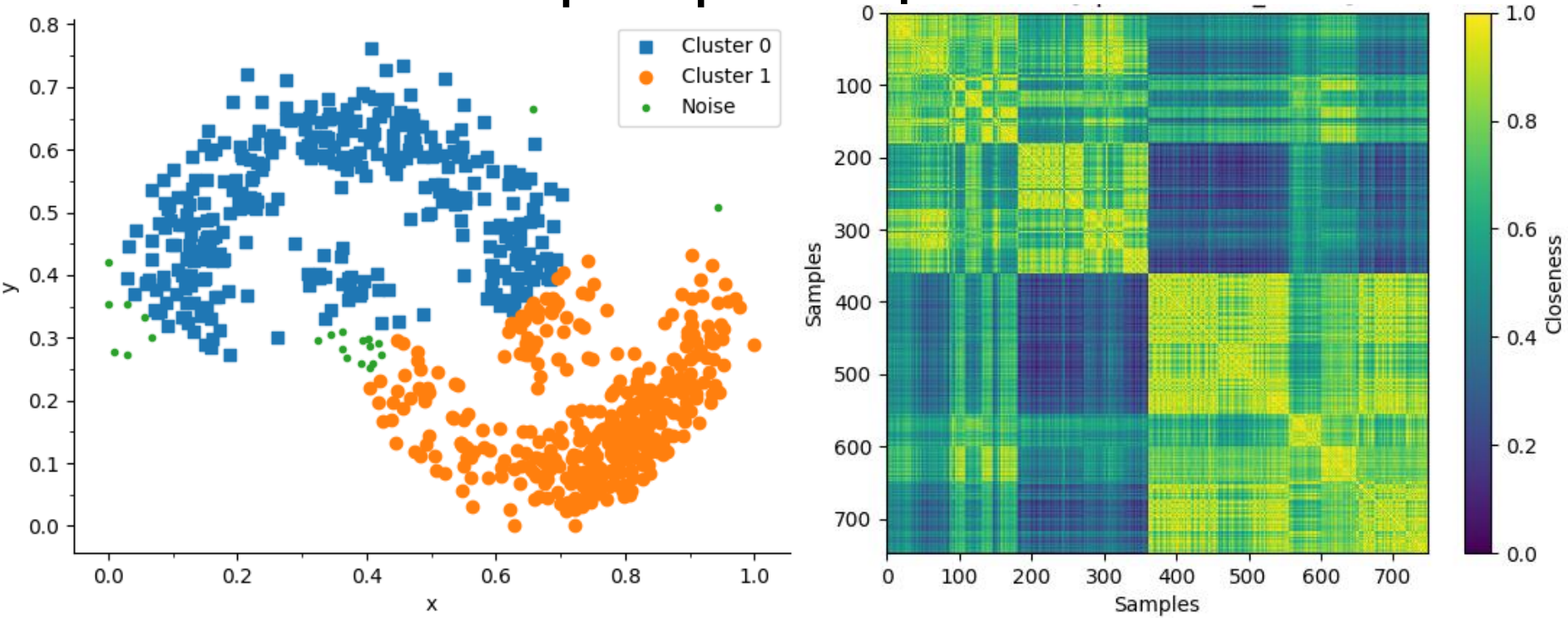
DBSCAN выбор параметров по CHI и SI





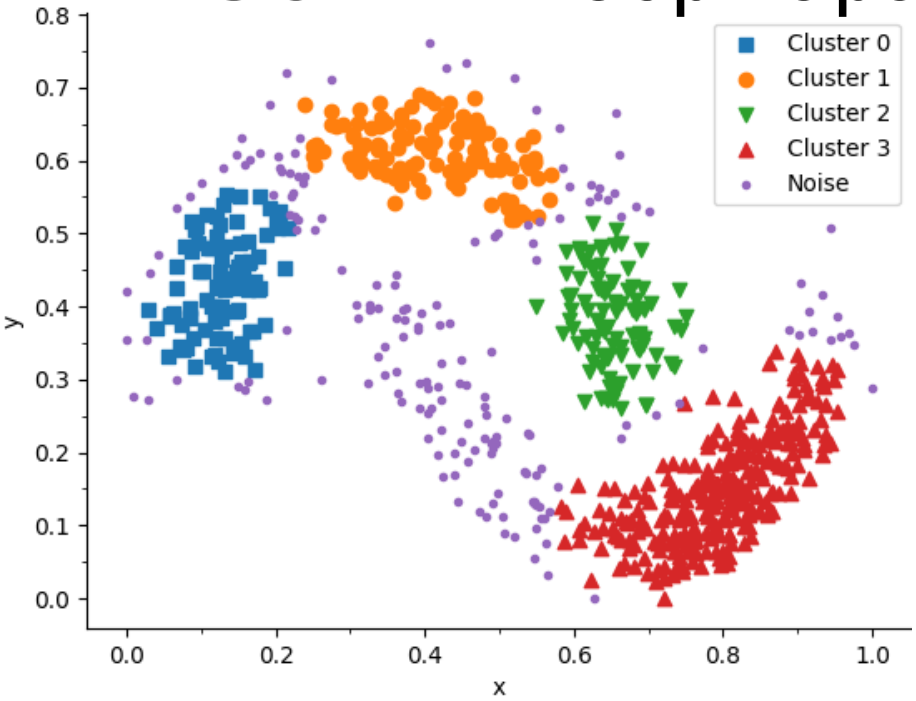
Рисунки 20-21. Кластеризация алгоритмом DBSCAN с параметрами, полученными путем оптимизации индексов CHI и SI Eps=0.023, m_s=12, K=2, noise=594, CHI=1788, SI=0.88, ARI(true)=0.44.

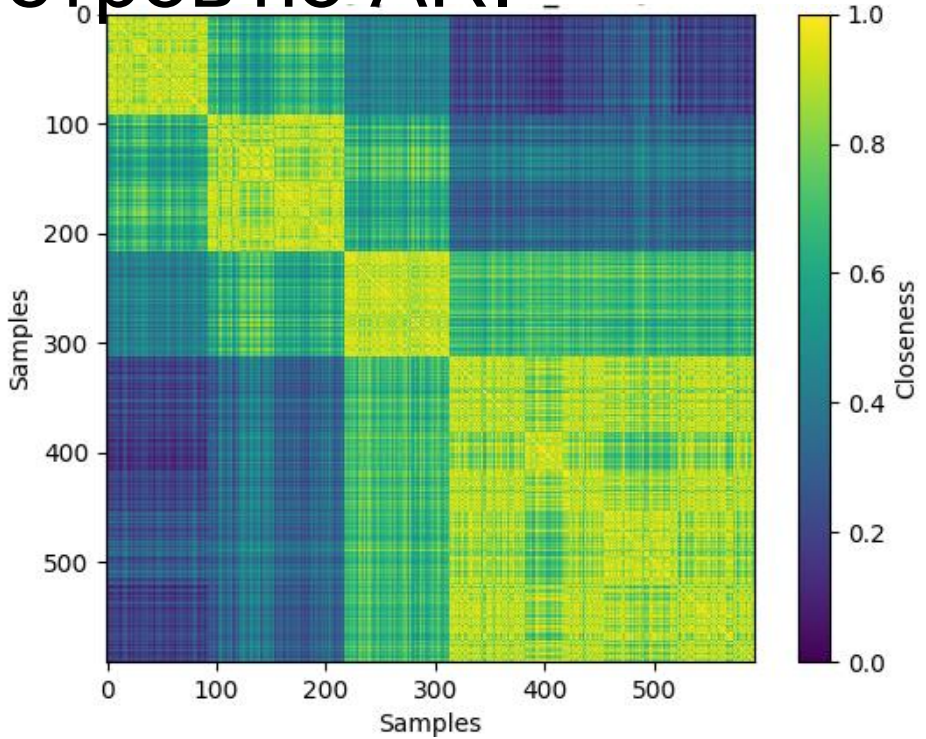
DBSCAN выбор параметров по CHI и SI



Рисунки 22-23. Кластеризация алгоритмом DBSCAN с параметрами, полученными путем оптимизации индексов CHI и SI с ограничением относительного показателя шума (3%). Eps=0.191, m_s=156, K=2, noise=22, CHI=1341, SI=0.54, ARI(true)=0.21.

DBSCAN выбор параметров по ARI

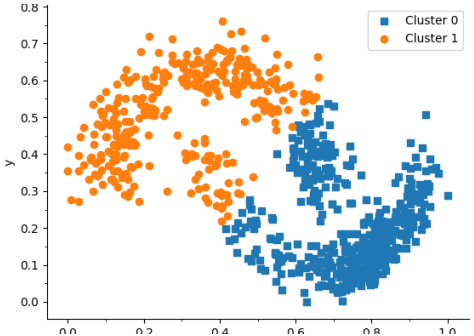


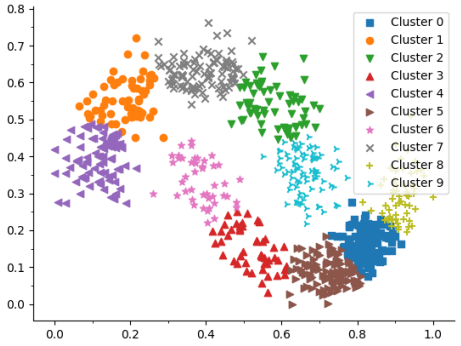


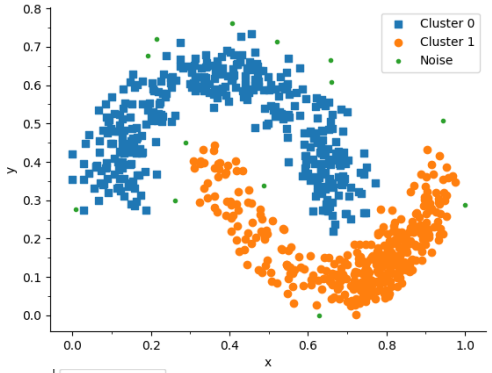
Рисунки 24-25. Кластеризация алгоритмом DBSCAN с параметрами, полученными путем оптимизации ARI по сравнению с реальной кластеризацией Eps=0.06, m_s=27, K=4, noise=178, CHI=1990, SI=0.58, ARI(true)=0.625.

Сравнение k-means и DBSCAN (ARI)

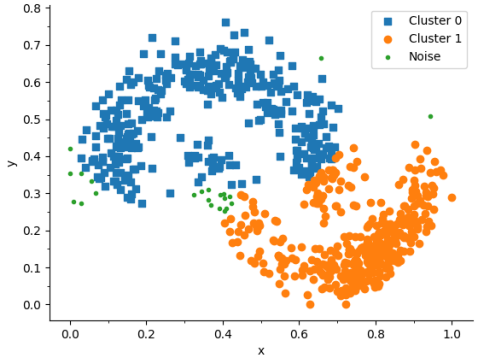
k-means

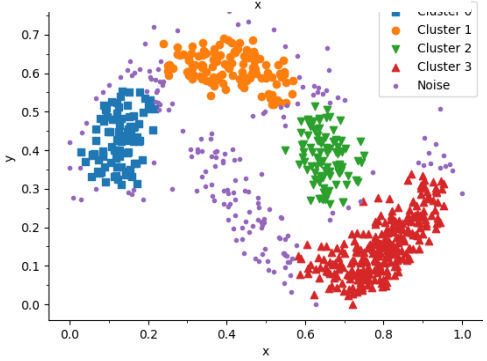


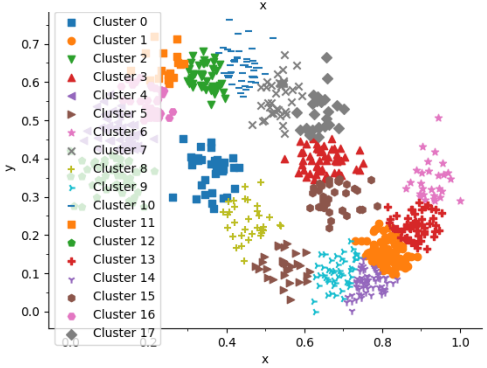




DBSCAN







ARI 0.70 0.51 0.12

Исходные данные (реальные)

Данные: Данные о знаниях пользователей, характеризуемые пятью признаками

Размер выборки: 258

Классы: 4 класса

STG	SCG	STR	LPR	PEG	UNS
0	0	0	0	0	very_low
0,08	0,08	0,1	0,24	0,9	High
0,06	0,06	0,05	0,25	0,33	Low
0,1	0,1	0,15	0,65	0,3	Middle
0,08	0,08	0,08	0,98	0,24	Low
0,09	0,15	0,4	0,1	0,66	Middle
0,1	0,1	0,43	0,29	0,56	Middle
0,15	0,02	0,34	0,4	0,01	very_low
0,2	0,14	0,35	0,72	0,25	Low
0	0	0,5	0,2	0,85	High
0,18	0,18	0,55	0,3	0,81	High
0,06	0,06	0,51	0,41	0,3	Low
0,1	0,1	0,52	0,78	0,34	Middle
0,1	0,1	0,7	0,15	0,9	High
0,2	0,2	0,7	0,3	0,6	Middle
0,12	0,12	0,75	0,35	0,8	High
0,05	0,07	0,7	0,01	0,05	very_low
0,1	0,25	0,1	0,08	0,33	Low
0,15	0,32	0,05	0,27	0,29	Low

Рисунок 26. Пример исходных данных

Визуализация данных

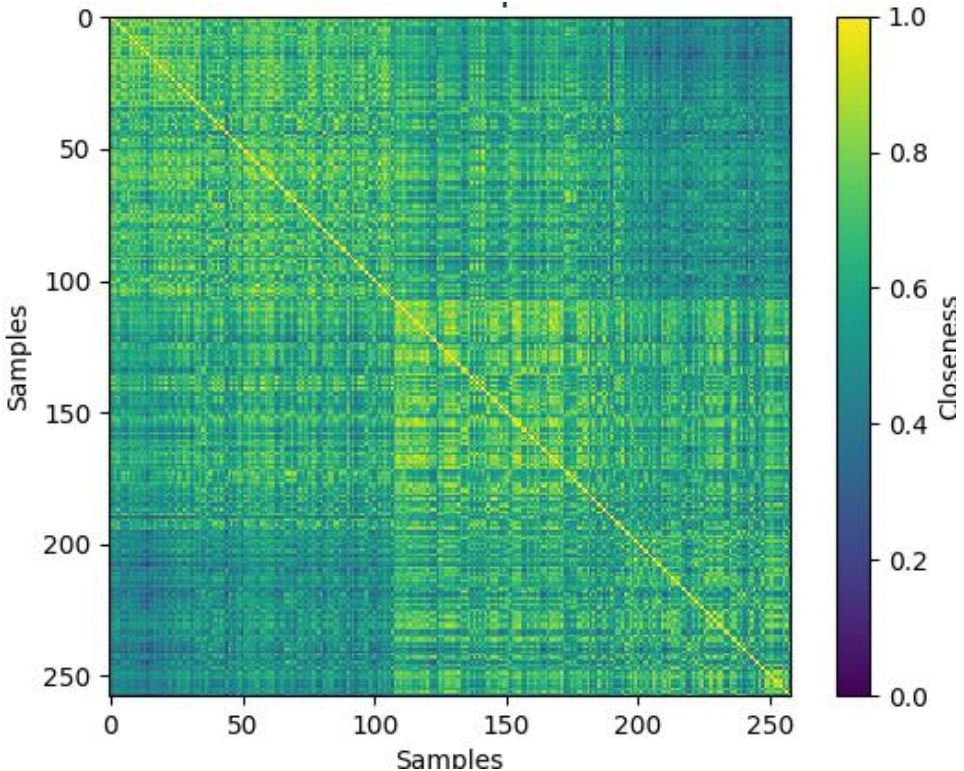


Рисунок 27. Визуализация исходных данных в виде heatmap

k-means (2 кластера)

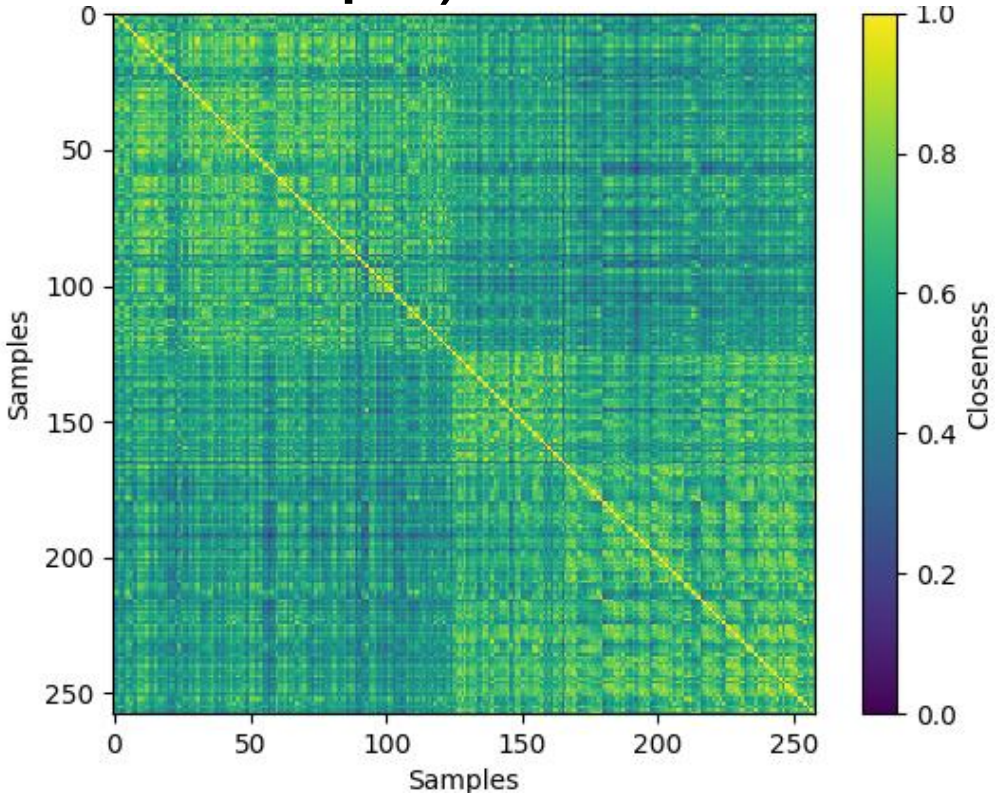


Рисунок 28. Кластеризация алгоритмом k-means с K=2. S=55, CHI=62, SI=0.22, ARI(true)=0.33.

k-means (4 кластера)

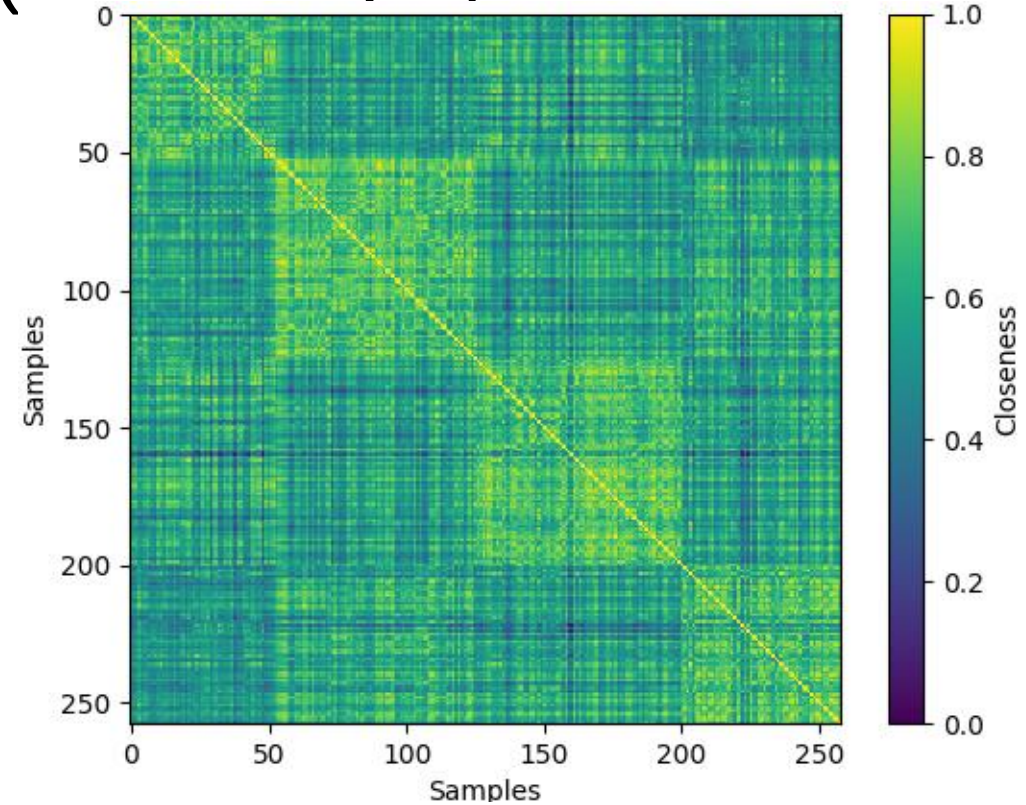


Рисунок 29. Кластеризация алгоритмом k-means с K=4. S=40, CHI=47, SI=0.21, ARI(true)=0.17.

k-means (11 кластеров)

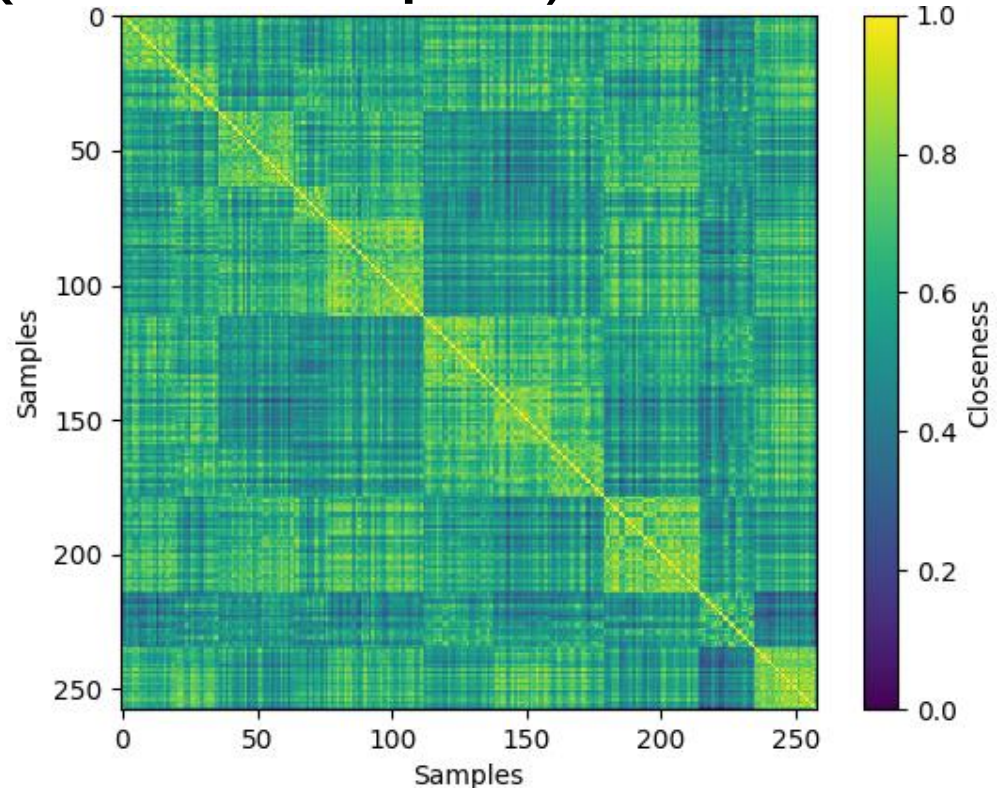
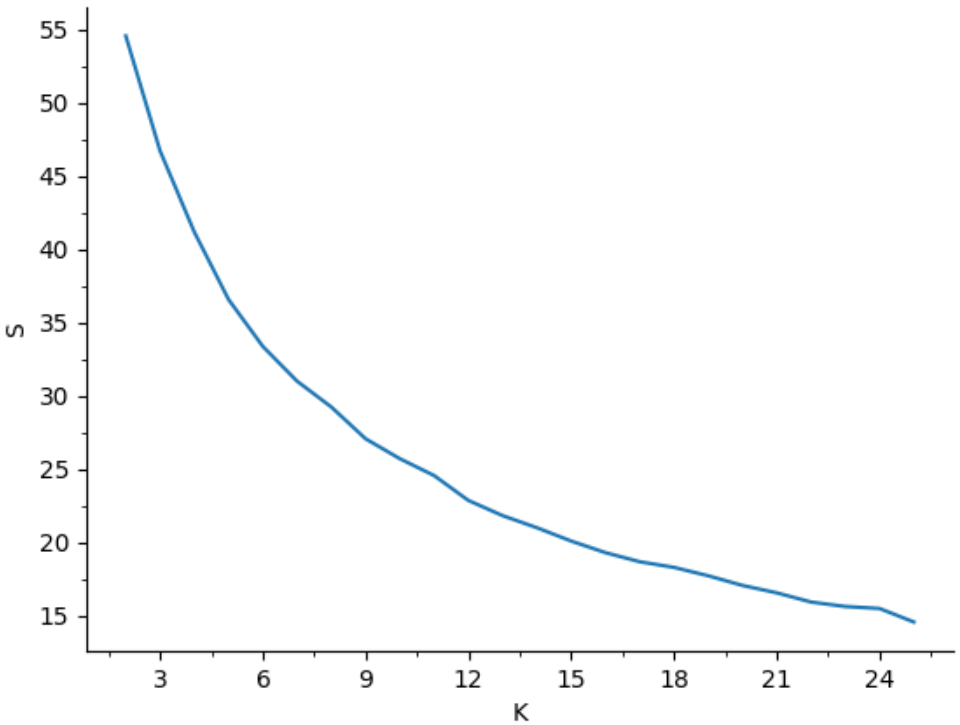
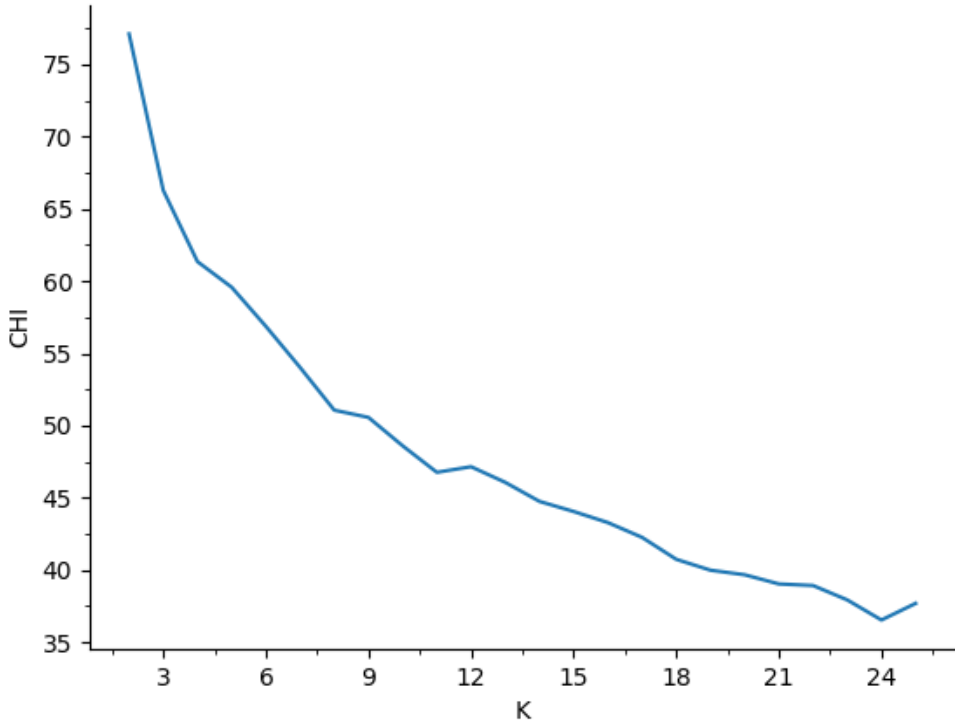


Рисунок 30. Кластеризация алгоритмом k-means с K=11. S=24, CHI=77, SI=0.20, ARI(true)=0.11.

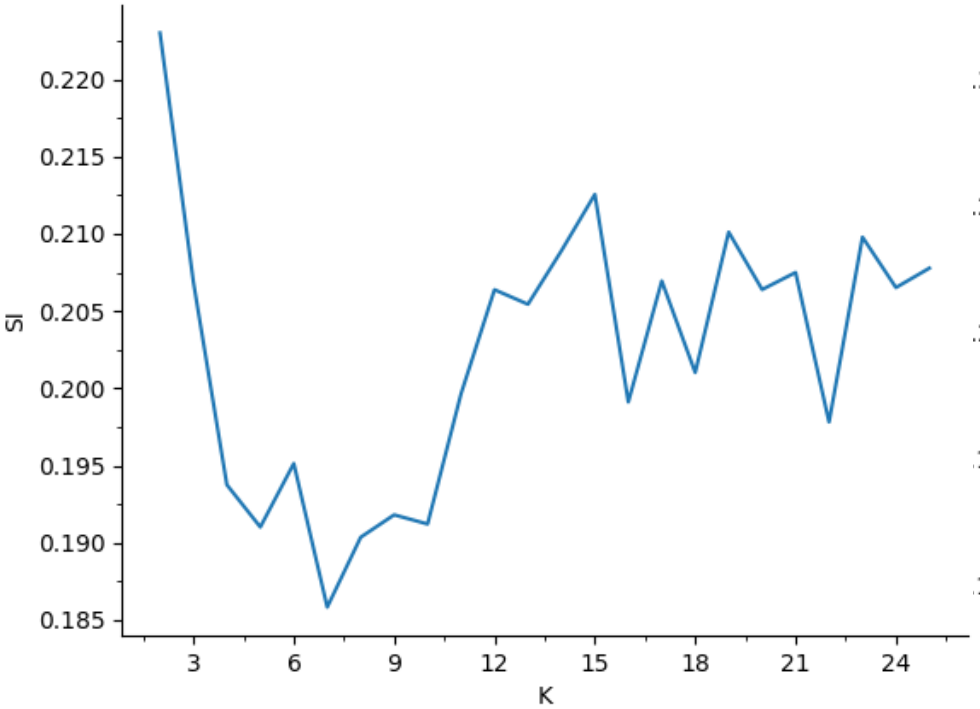
k-means исследование параметров (S, CHI)

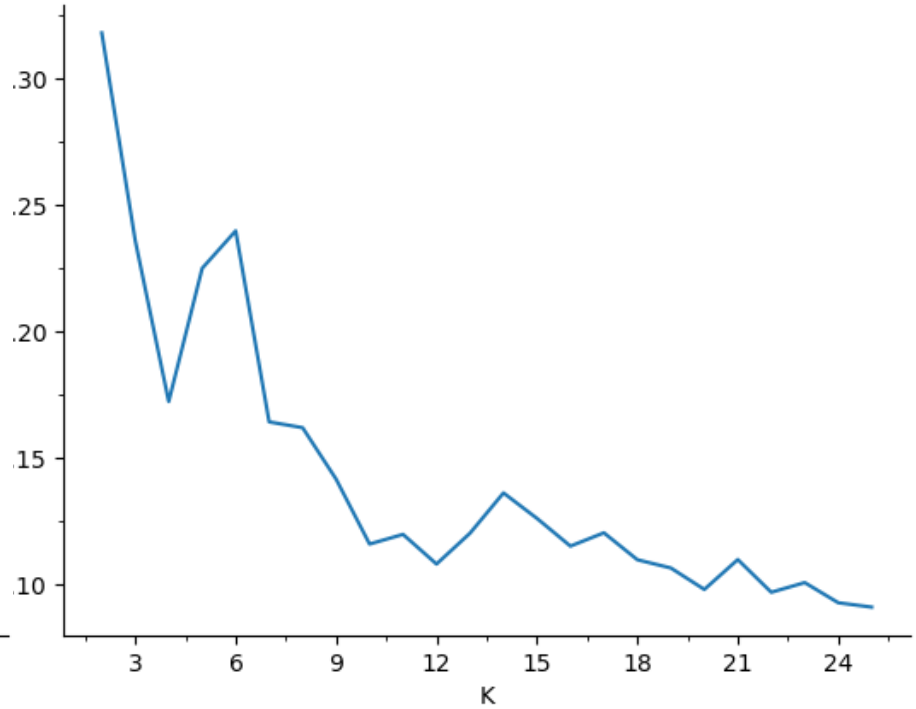




Рисунки 31-32. Зависимость внутрикластерной суммы квадратов отклонений S (слева) и значения индекса Calinski-Harabasz (справа) от количества кластеров.

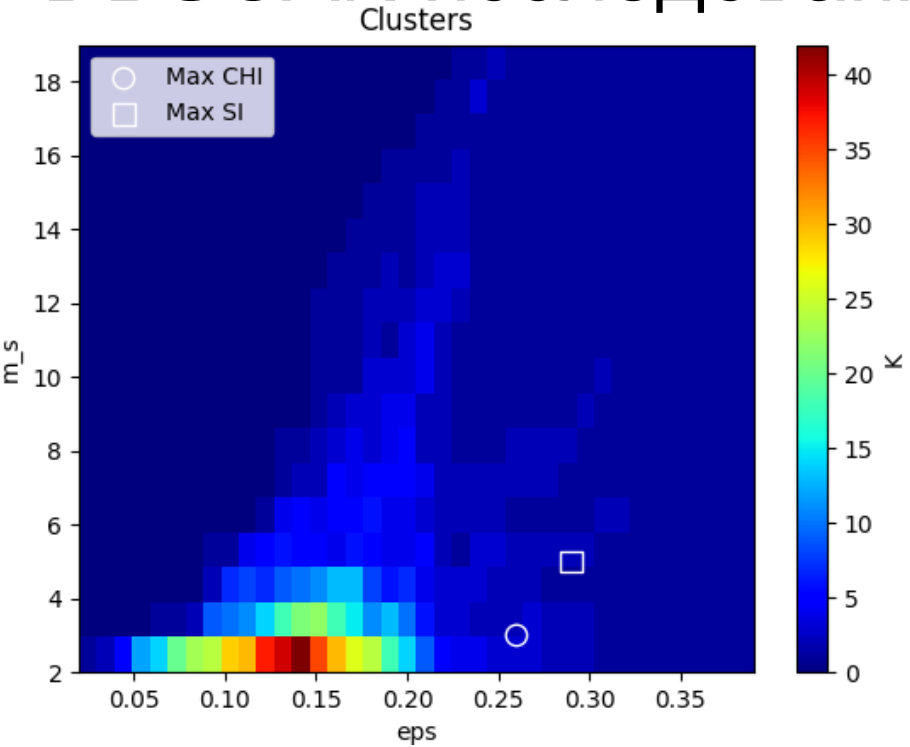
k-means исследование параметров (SI, ARI)

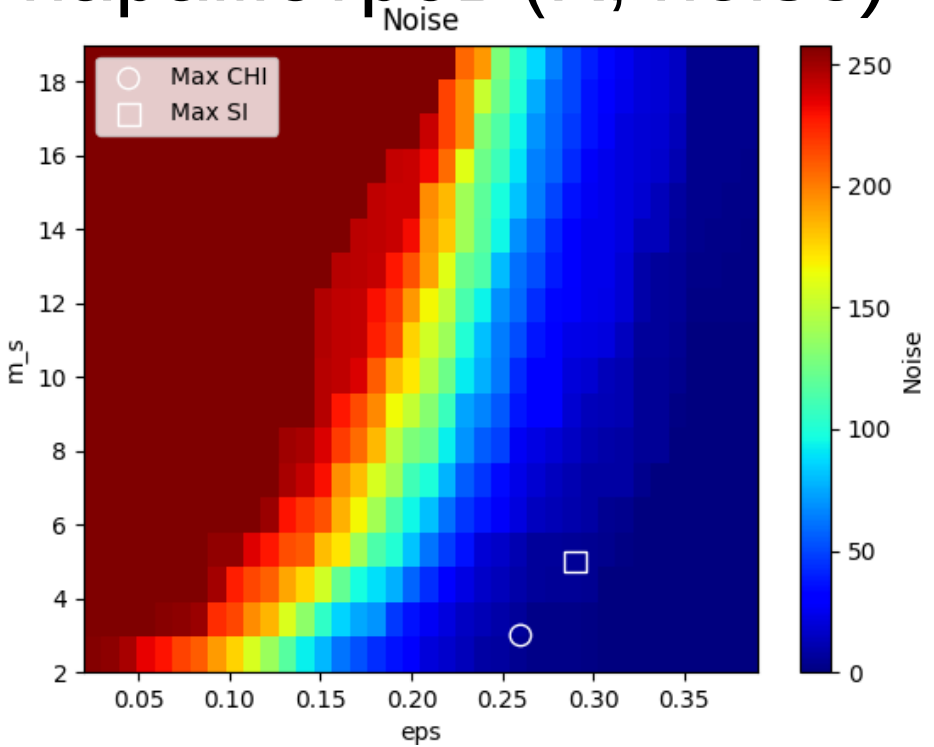




Рисунки 33-34. Зависимость значения индексов Silhouette (слева) и Adjusted Rand Index (по сравнению с реальной кластеризацией) (справа) от количества кластеров.

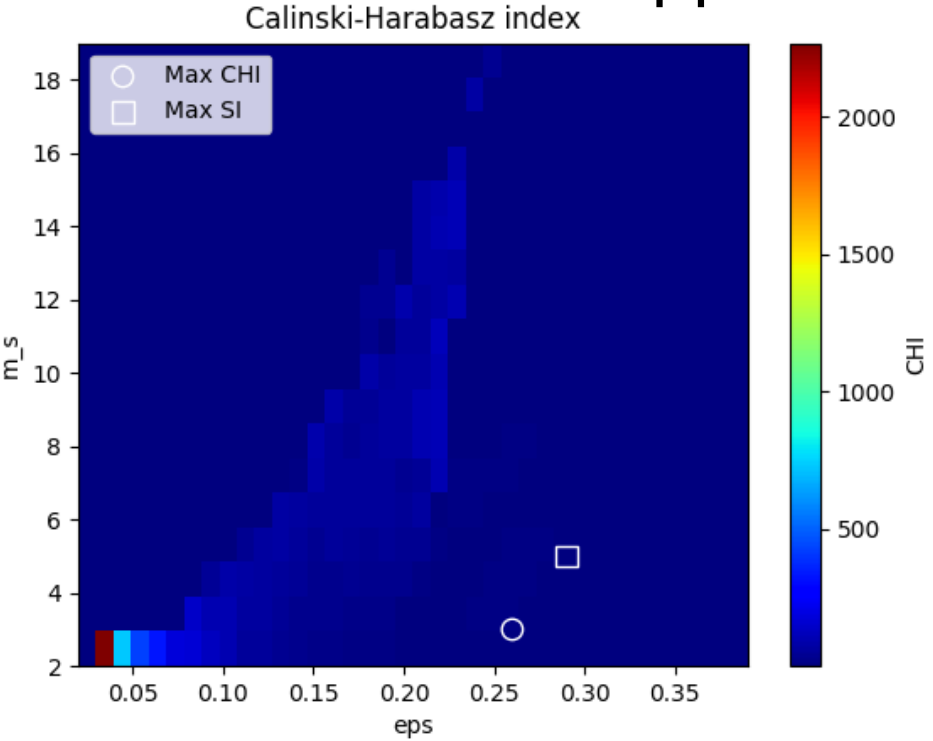
DBSCAN исследование параметров (K, noise)

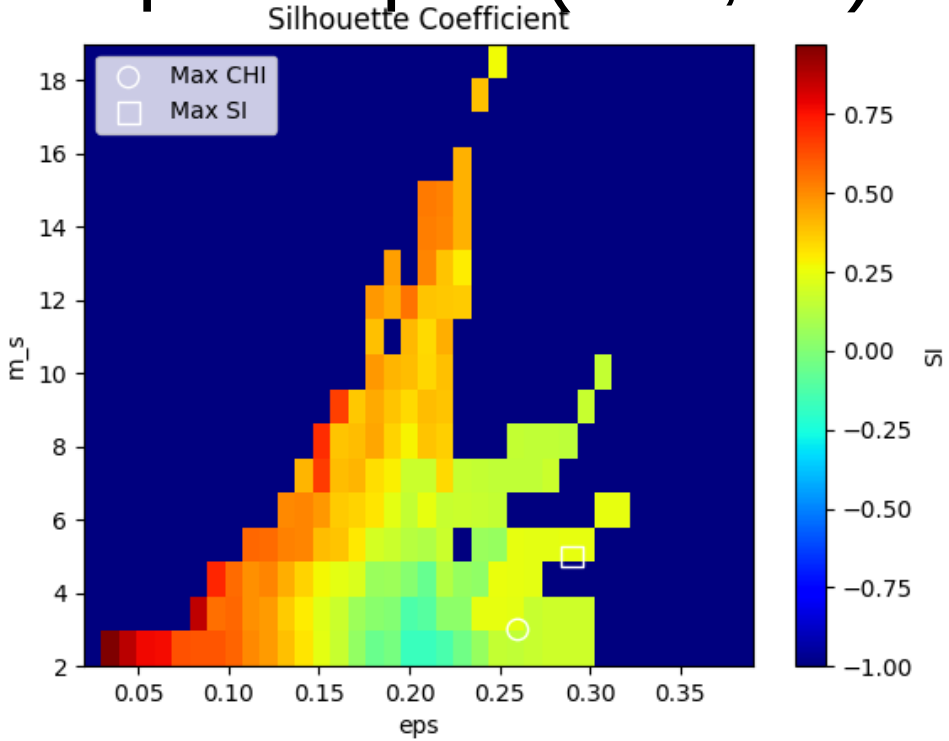




Рисунки 35-36. Зависимость количества кластеров (слева) и шума (справа) от параметров алгоритма DBSCAN. Лучшие значения индексов при ограничении шума 3%

DBSCAN исследование параметров (CHI, SI)





Рисунки 37-38. Зависимость значения индексов Calinski-Harabasz (слева) и Silhouette (справа) от параметров алгоритма DBSCAN. Лучшие значения индексов при ограничении шума 3%

DBSCAN исследование параметров (ARI)

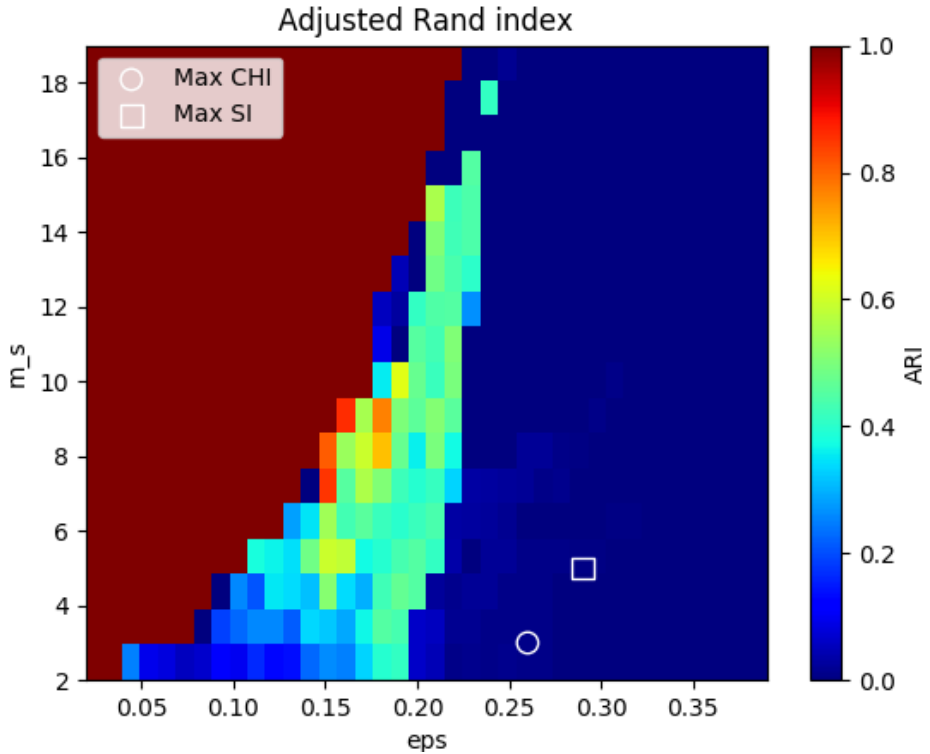


Рисунок 39-40. Зависимость значения индекса Adjusted Rand Index (по сравнению с реальной кластеризацией) от параметров алгоритма DBSCAN. Лучшие значения индексов при ограничении шума 3%

DBSCAN выбор параметров

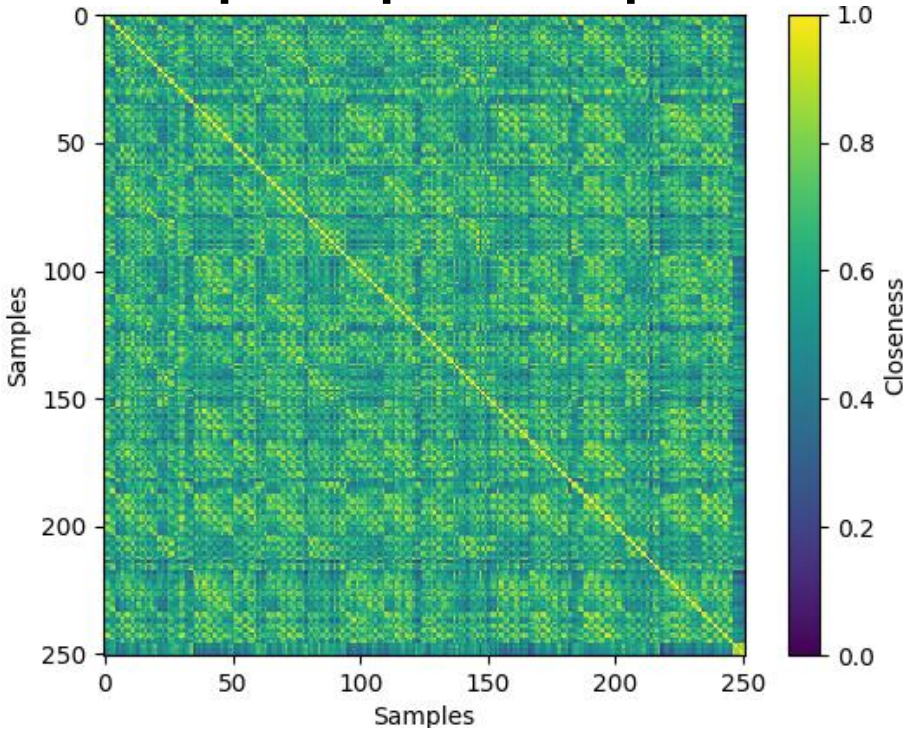
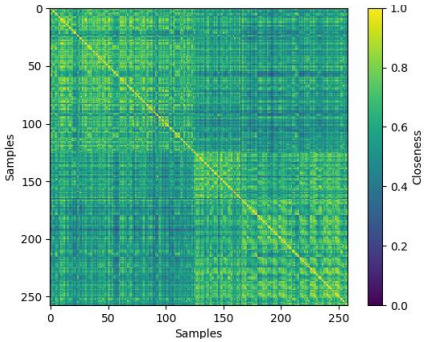
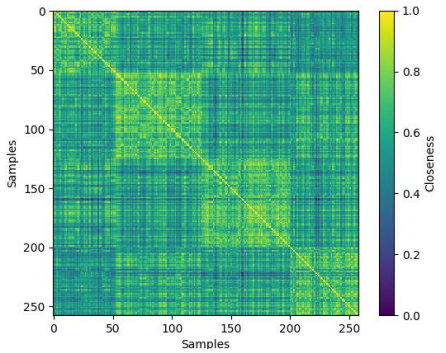


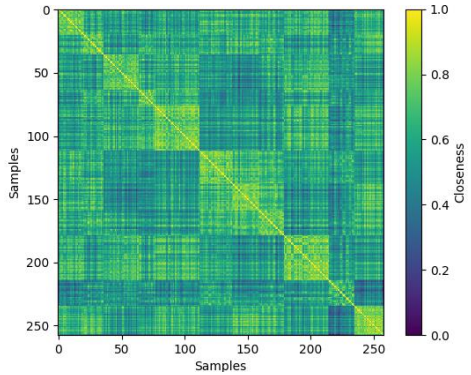
Рисунок 41. Кластеризация алгоритмом DBSCAN с параметрами, полученными путем оптимизации индексов CHI и SI Eps=0.26, m_s=3, K=2, noise=7, CHI=11.4, SI=0.25, ARI(true)=0.006.

Сравнение k-means и DBSCAN (ARI)

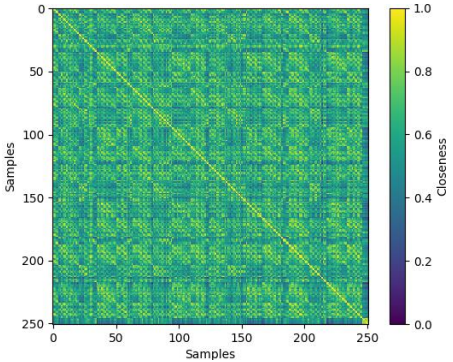
k-means

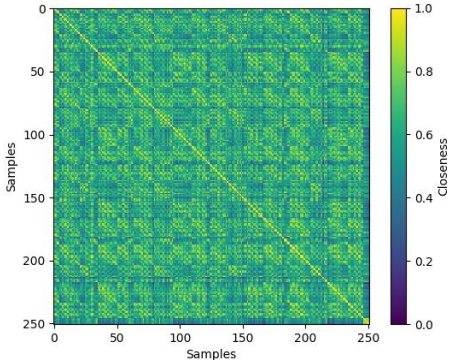


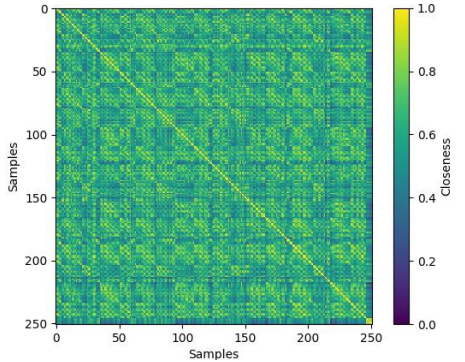




DBSCAN







ARI

0.002

0.012

0.004

Выводы

1. k-means лучше подходит для сферических кластеров, DBSCAN – для ленточных
2. DBSCAN хорошо кластеризовал тестовые данные, потому что они ленточные
3. SSE не показателен для ленточных кластеров
4. CHI и SI тоже дают лучшие значения на непересикающихся сферических кластера, однако не так чувствительны к форме, как SSE
5. Нельзя полагаться на показатели CHI и SI при выборе параметров DBSCAN

Выводы

6. Индексы CHI и SI могут показывать хорошие результаты работы DBSCAN при большем количестве шума, однако таким результатом нельзя доверять
7. С помощью heatmap можно визуализировать результаты кластеризации при любом количестве параметров объекта
8. С помощью ARI можно сравнивать результаты кластеризации в независимости от порядка данных и имен кластеров