МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ

по лабораторной работе №6

по дисциплине «Машинное обучение»

ТЕМА: Кластеризация (DBSCAN, OPTICS)

Студент гр. 6307 _____ Мишанов А. А.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы: ознакомиться с методами кластеризации (DBSCAN, OPTICS) библиотеки sklearn.

Ход работы

1. Датасет скачан и загружен в датафрейм.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	40.900749	0.818182	95.40	0.00	95.40	0.000000	0.166667	0.000000	0.083333	0.00	0.0	2.0	1000.0	201.802084	139.509787	0.000000	12.0
1	3202.467416	0.909091	0.00	0.00	0.00	6442.945483	0.000000	0.000000	0.000000	0.25	4.0	0.0	7000.0	4103.032597	1072.340217	0.222222	12.0
2	2495.148862	1.000000	773.17	773.17	0.00	0.000000	1.000000	1.000000	0.000000	0.00	0.0	12.0	7500.0	622.066742	627.284787	0.000000	12.0
3	817.714335	1.000000	16.00	16.00	0.00	0.000000	0.083333	0.083333	0.000000	0.00	0.0	1.0	1200.0	678.334763	244.791237	0.000000	12.0
4	1809.828751	1.000000	1333.28	0.00	1333.28	0.000000	0.666667	0.000000	0.583333	0.00	0.0	8.0	1800.0	1400.057770	2407.246035	0.000000	12.0

DBSCAN

1. Кластеризация методом K-средних.

```
k_means = KMeans(init='k-means++', n_clusters=3, n_init=15)
k_means.fit(no_labeled_data)
```

```
KMeans(n_clusters=3, n_init=15)
```

2. Стандартизация данных.

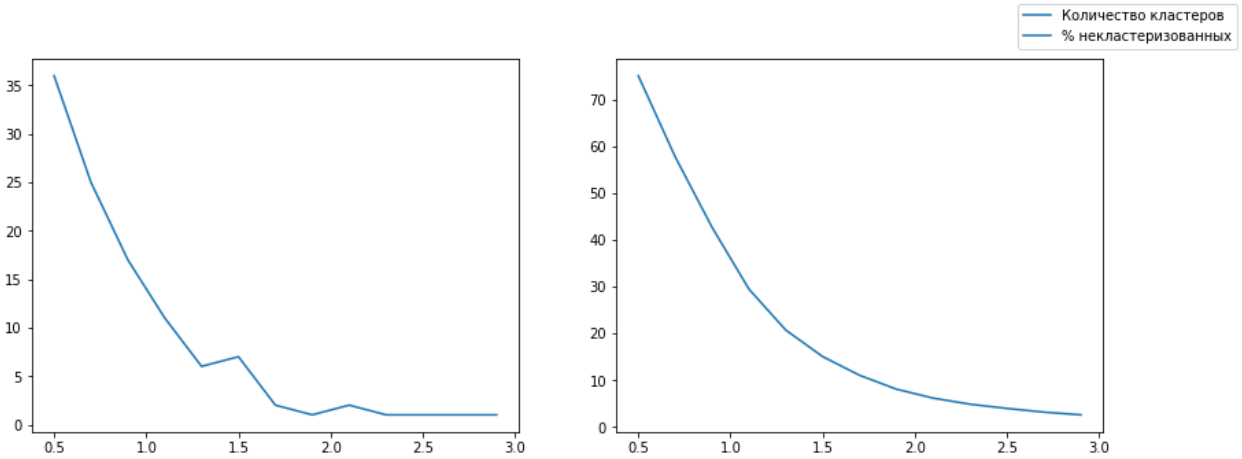
```
data = np.array(data, dtype='float')
min_max_scaler = StandardScaler()
scaled_data = min_max_scaler.fit_transform(data)
```

3. Кластеризация методом DBSCAN при параметрах по умолчанию. Получены метки кластеров, количество кластеров и оценочная часть наблюдений, которые не удалось кластеризировать.

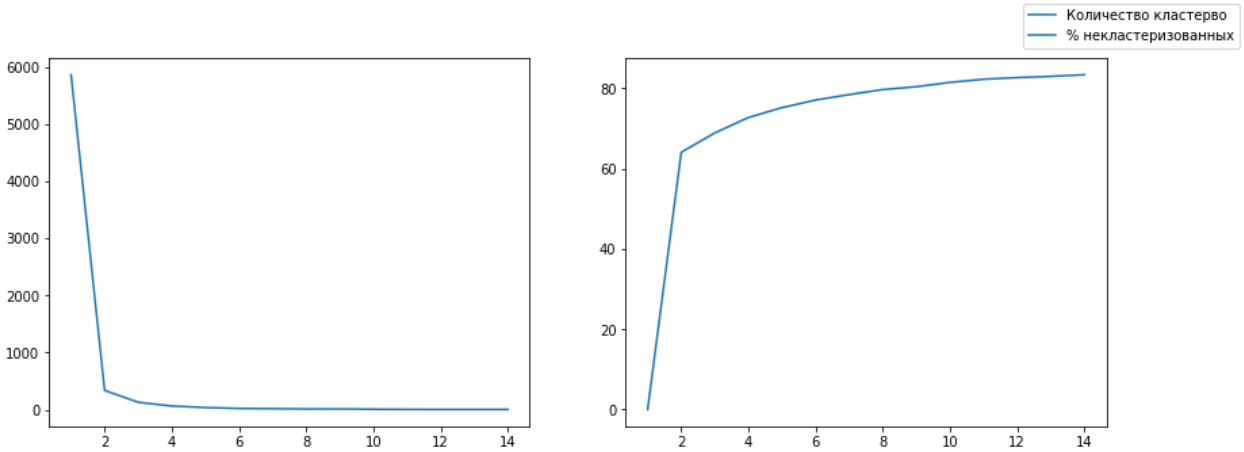
```
clustering = DBSCAN().fit(scaled_data)
print(set(clustering.labels_))
print(len(set(clustering.labels_)) - 1)
print(list(clustering.labels_).count(-1) / len(list(clustering.labels_)))
```

```
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, -1}
36
0.7512737378415933
```

4. Построен график зависимости количества кластеров от максимальной дистанции между наблюдениями и график зависимости процента некластеризованных данных от максимальной дистанции между наблюдениями.



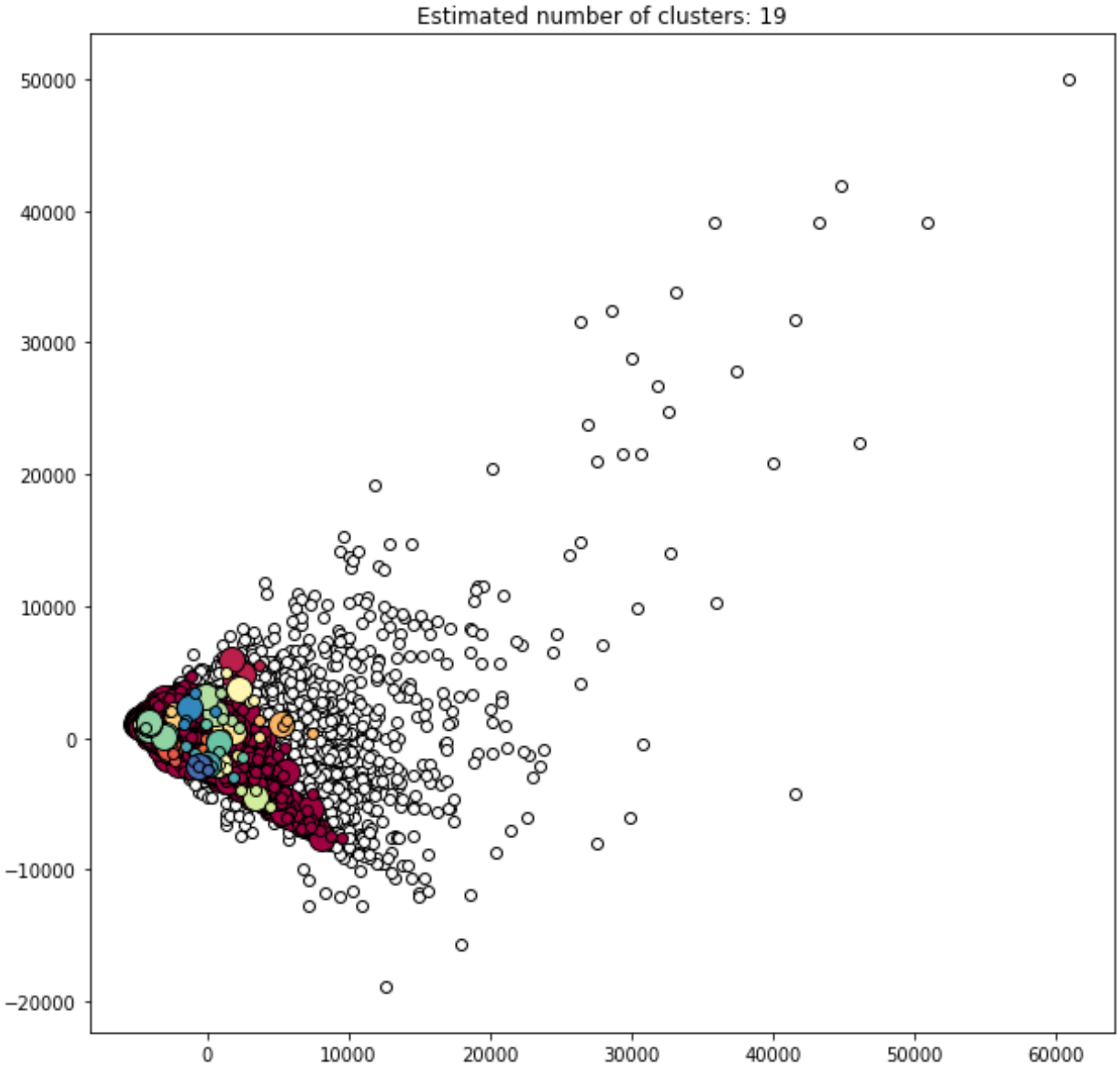
5. Построен график зависимости количества кластеров от минимального значения количества точек и график зависимости процента некластеризованных данных от минимального значения точек



6. Определены значения параметров, при котором количество кластеров получается от 5 до 7, и процент не кластеризированных не превышает 12.

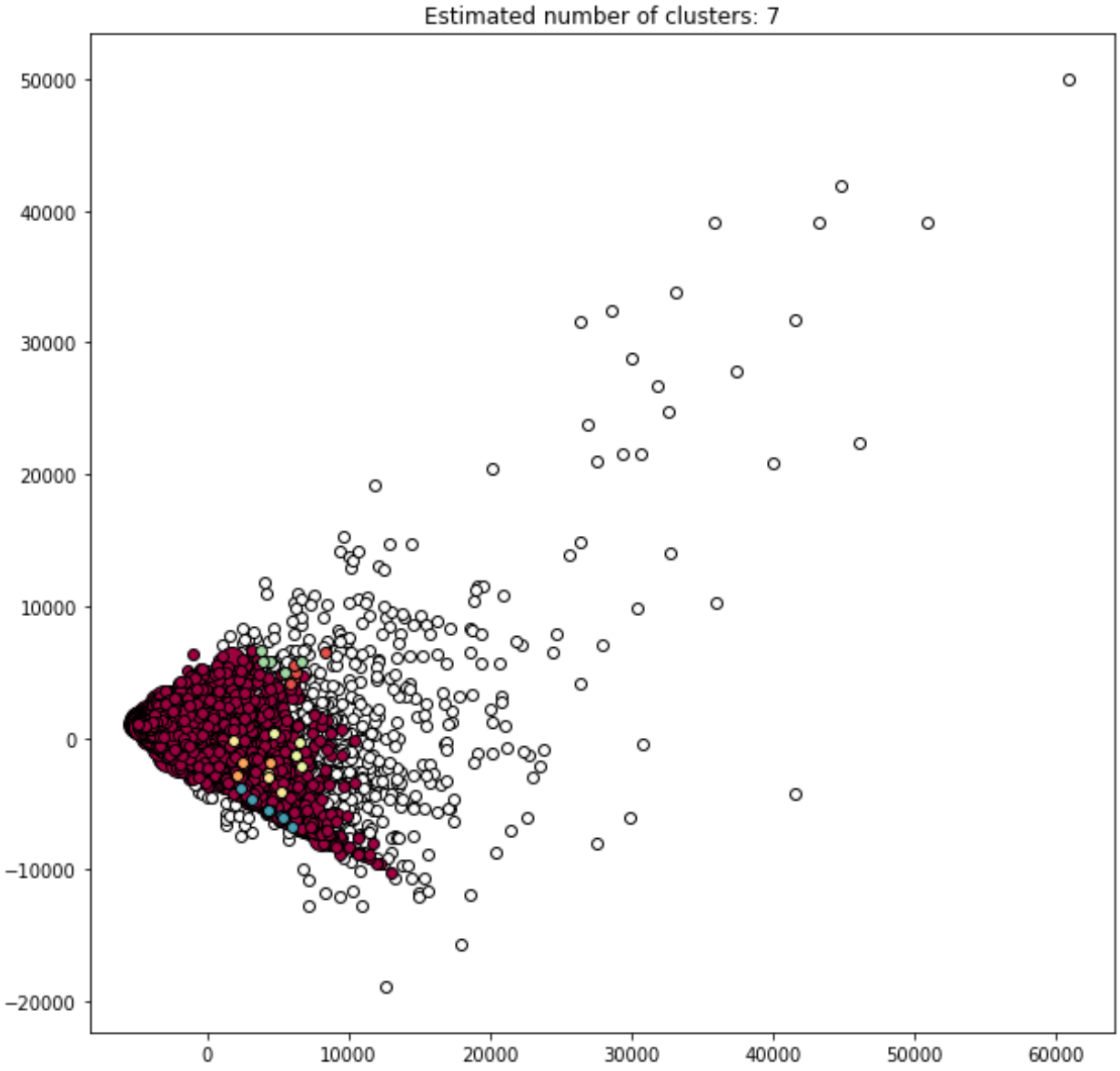
```
(3, 1.9999999999999996) [6, 0.06287633163501621]
(3, 2.5999999999999996) [5, 0.030917091245947197]
(3, 2.6999999999999993) [5, 0.027095877721167207]
(3, 2.8999999999999995) [5, 0.0222325150532654]
(4, 1.6999999999999997) [5, 0.1024779990736452]
```

7. Понижение размерности до 2 с помощью метода главных компонент. Визуализация кластеризации.

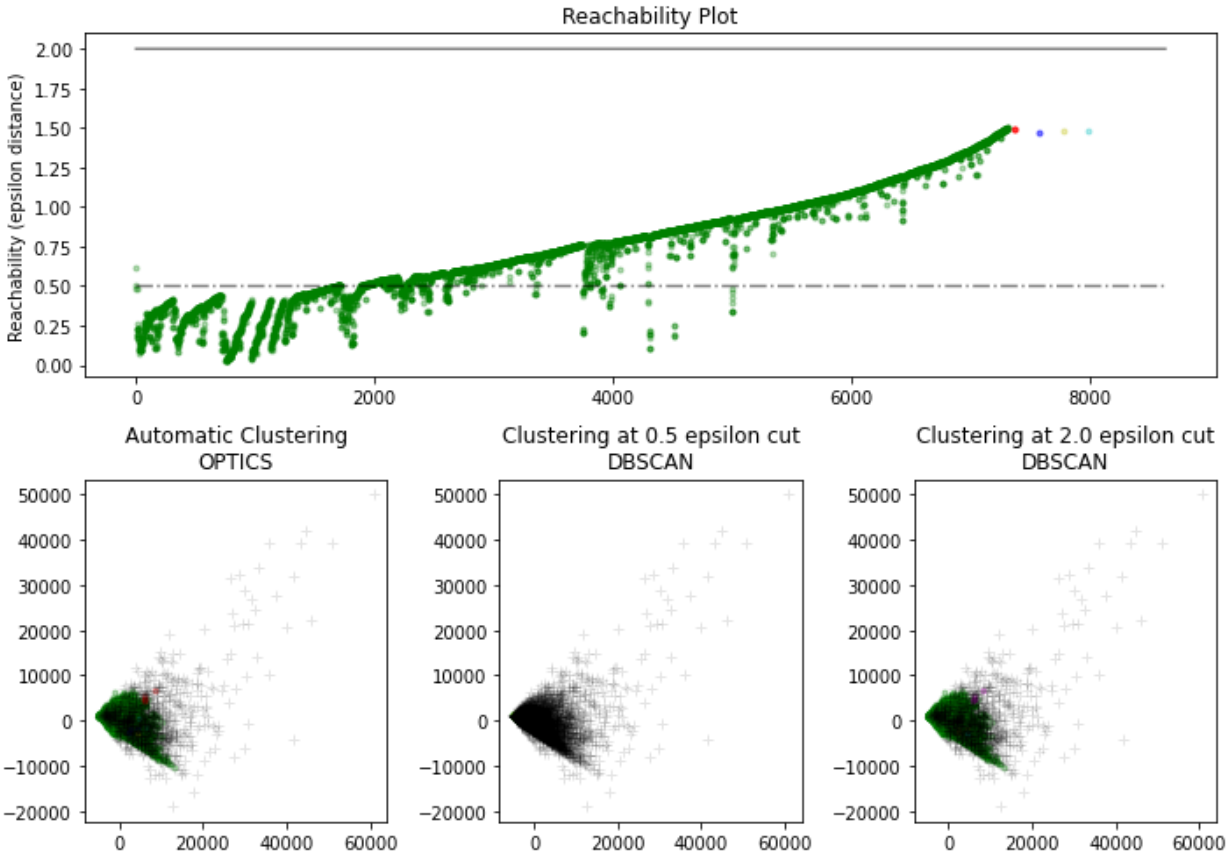


OPTICS

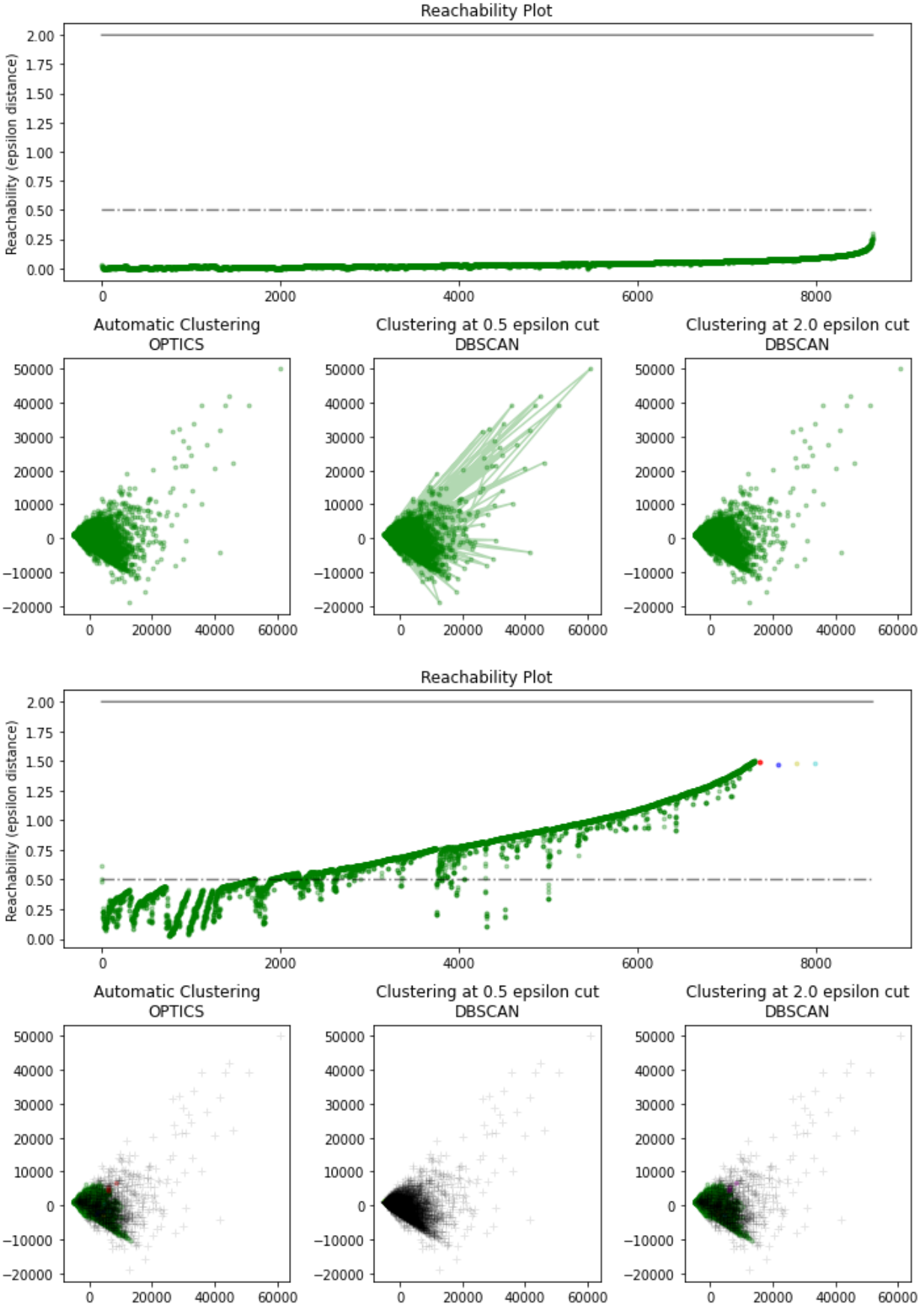
1. При тех же параметрах, как и у DBSCAN, и cluster_method=’dbscan’ кластеризация получилась близкой к кластеризации DBSCAN.

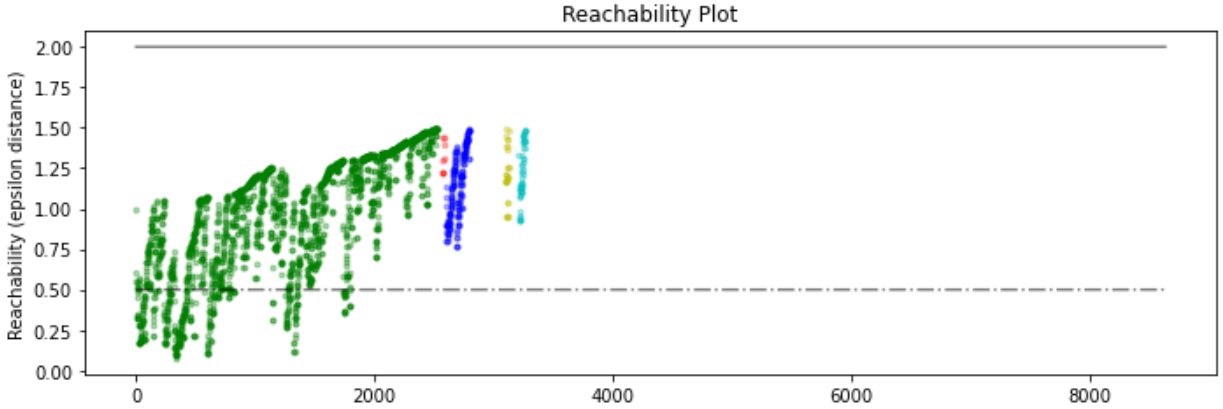


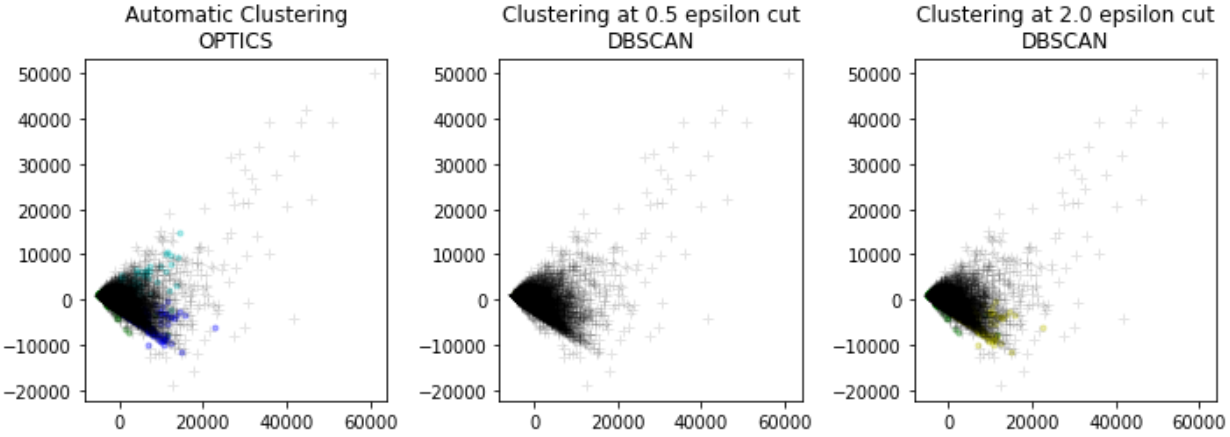
2. Построен график достижимости.

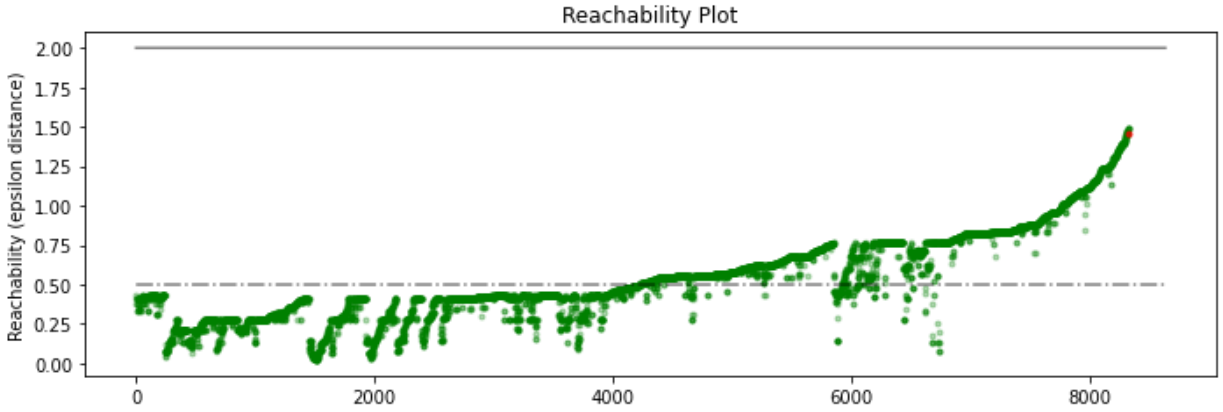


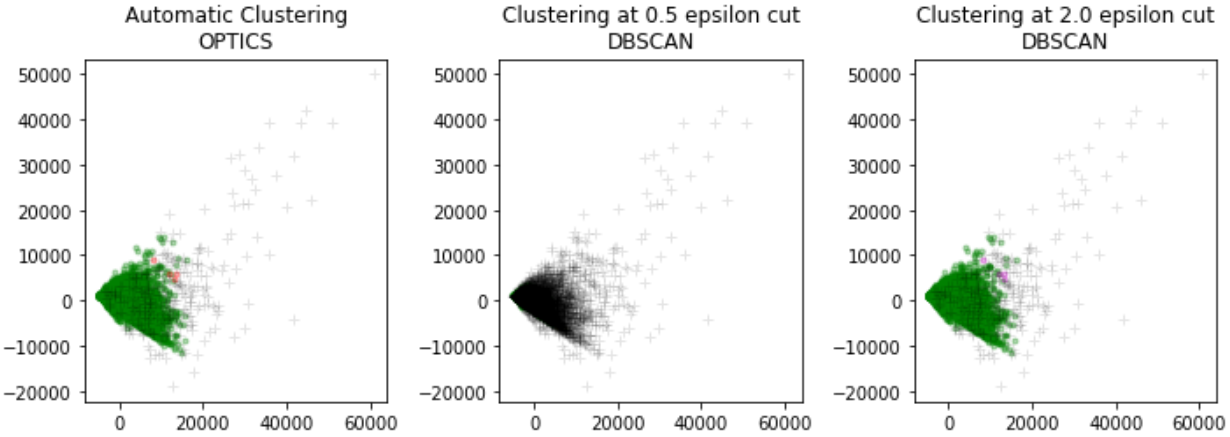
3. Исследована работа метода OPTICS с использованием различных метрик, а также построены графики достижимости. Исследуемые метрики: cosine, euclidean, canberra, chebyshev, cityblock.

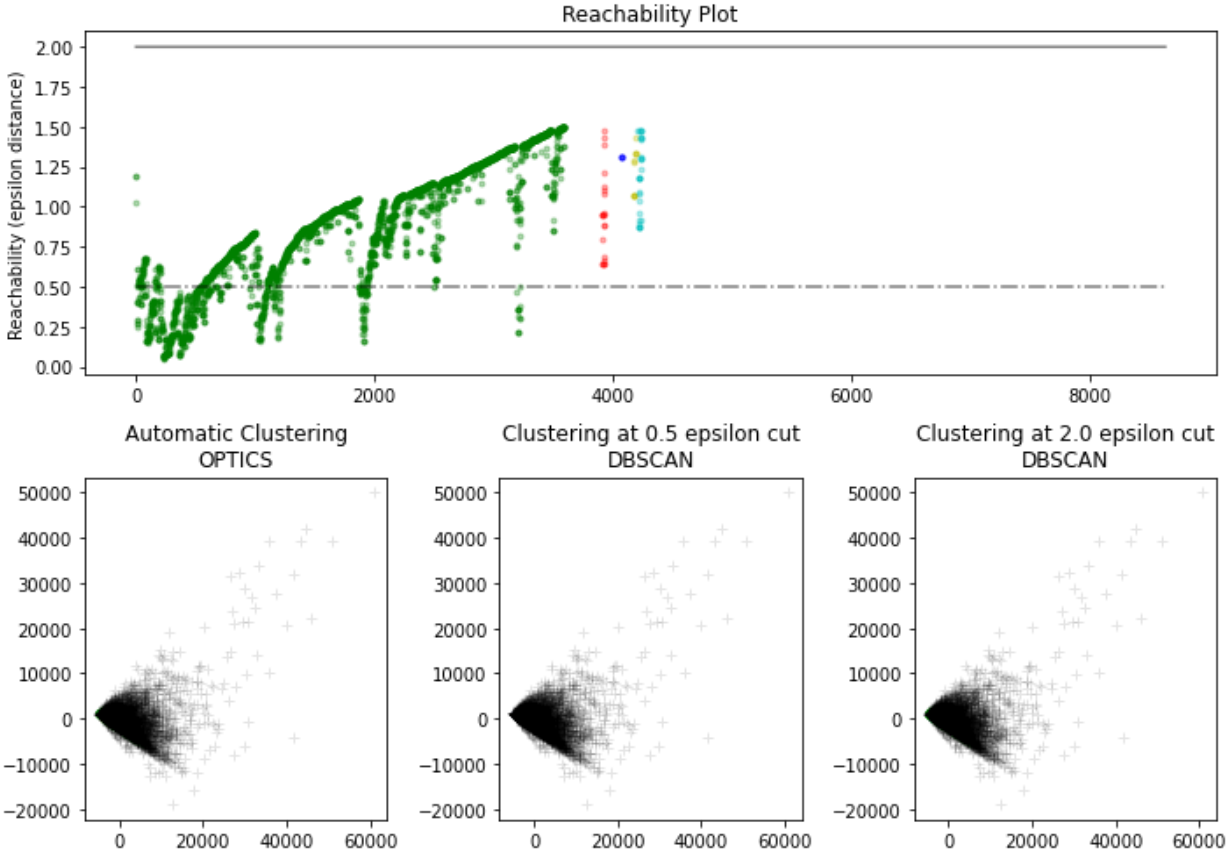

Reachability Plot
Reachability (epsilon distance)
Automatic Clustering
OPTICS
Clustering at 0.5 epsilon cut
DBSCAN
Clustering at 2.0 epsilon cut
DBSCAN
Reachability Plot
Reachability (epsilon distance)
Automatic Clustering
OPTICS
Clustering at 0.5 epsilon cut
DBSCAN
Clustering at 2.0 epsilon cut
DBSCAN


Reachability Plot
Reachability (epsilon distance)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
8000


Automatic Clustering
OPTICS
Clustering at 0.5 epsilon cut
DBSCAN
Clustering at 2.0 epsilon cut
DBSCAN
50000
40000
30000
20000
10000
0
−10000
−20000
0
20000
40000
60000


Reachability Plot
Reachability (epsilon distance)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
8000


Automatic Clustering
OPTICS
Clustering at 0.5 epsilon cut
DBSCAN
Clustering at 2.0 epsilon cut
DBSCAN
50000
40000
30000
20000
10000
0
−10000
−20000
0
20000
40000
60000


Reachability Plot
Reachability (epsilon distance)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
8000
Automatic Clustering
OPTICS
Clustering at 0.5 epsilon cut
DBSCAN
Clustering at 2.0 epsilon cut
DBSCAN
50000
40000
30000
20000
10000
0
−10000
−20000
0
20000
40000
60000