Rezultati za sve značajke bez kombinacije značajki koji jako koreliraju zbog fizikalne sličnosti

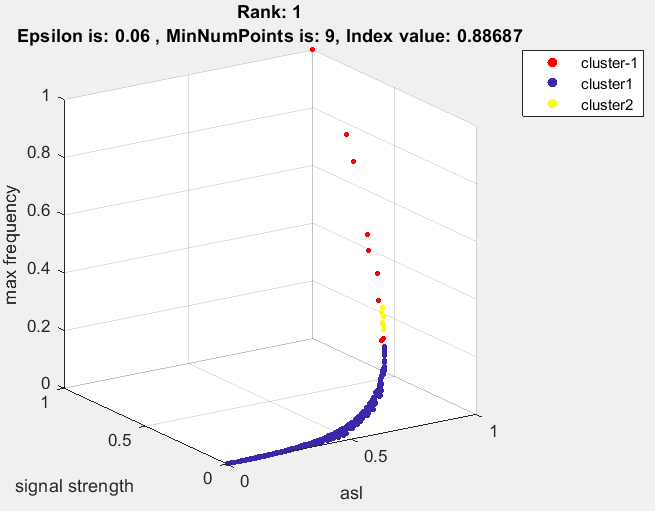
Ispod je prikazan koreleogram svih značajki za korišteni skup podataka.

Prva grupa značajki koje jako koreliraju su PEAK AMP, RMS, ABSOULUTE ENERGY i SIGNAL STRENGTH.

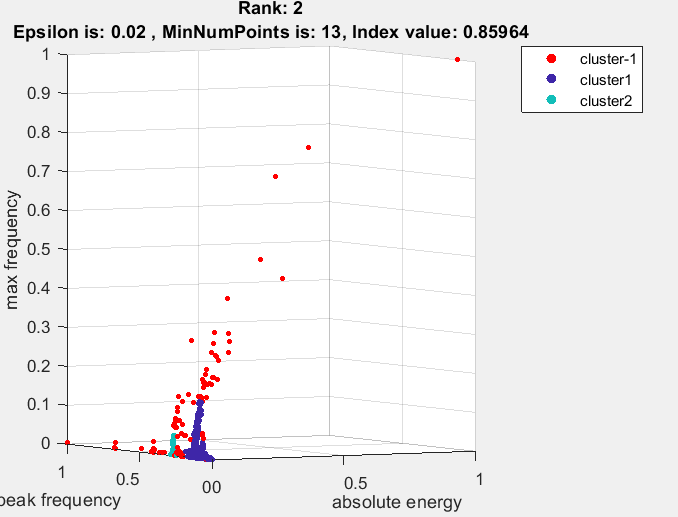
Druga grupa značajki koje jako koreliraju su WEIGHTED PEAK FREQUENCY i PEAK FREQUENCY.

Selekcija značajki provedena je kroz vrijednosti:

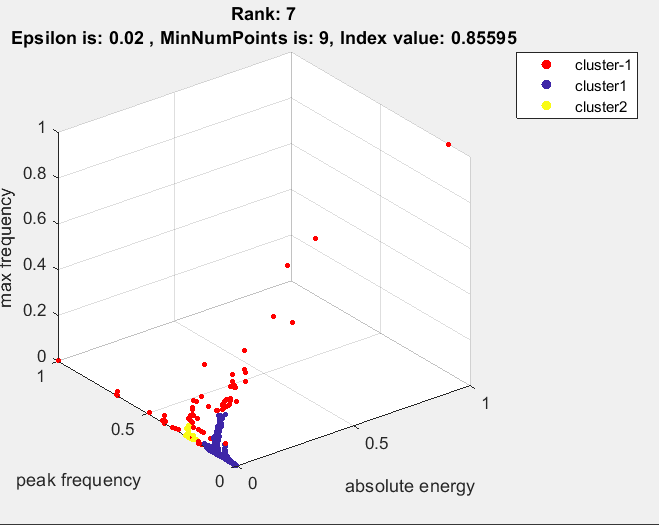
1. Epsilon – 0.02 do 0.06 s korakom 0.01
2. NMinPoints – 9 do 20 s korakom 1



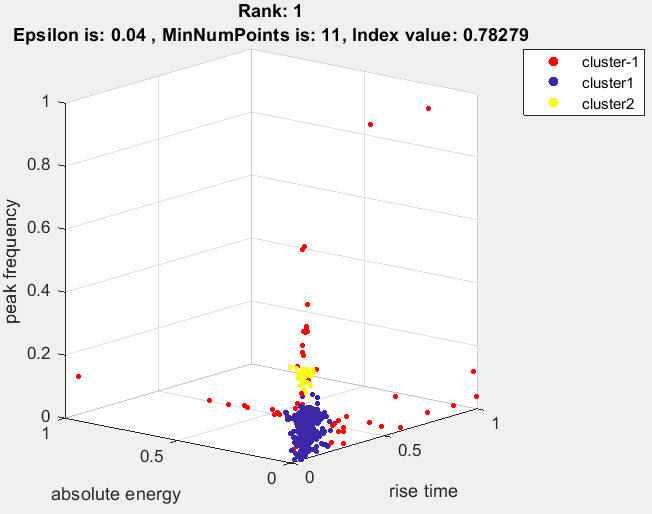
Zaključak da je potrebno koristiti veći broj NMinPoints jer kod male razlike u broju outliera indeks više preferira manji broj zgusnutih točaka.



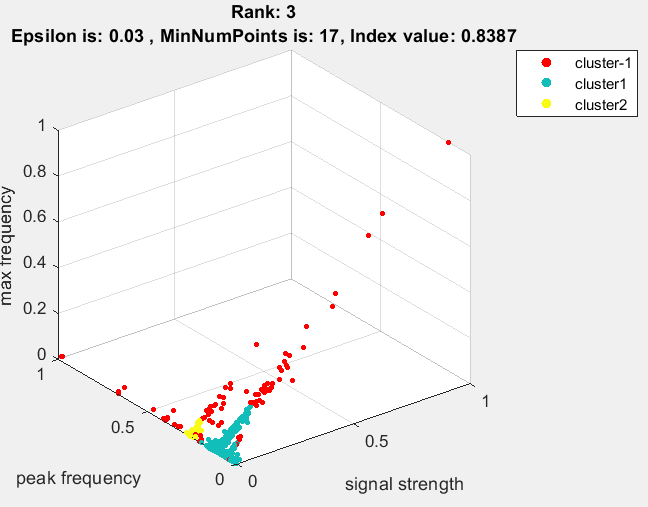
Ovdje se barem izabire veći broj točaka u svakom clusteru.



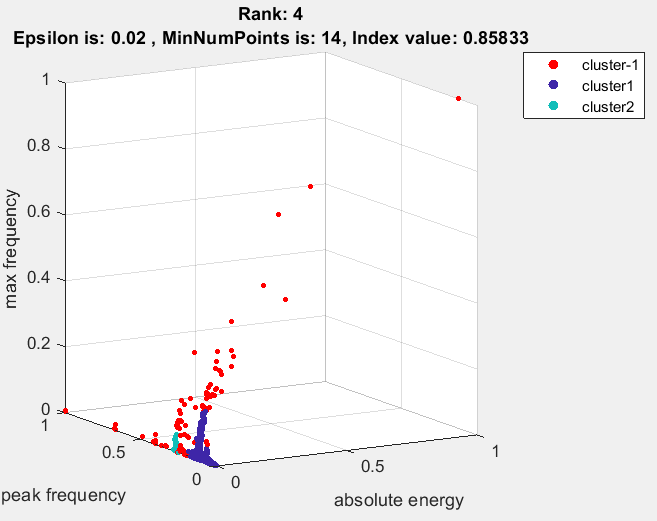
Za slihouette dobivamo dosta dobre rezultate.



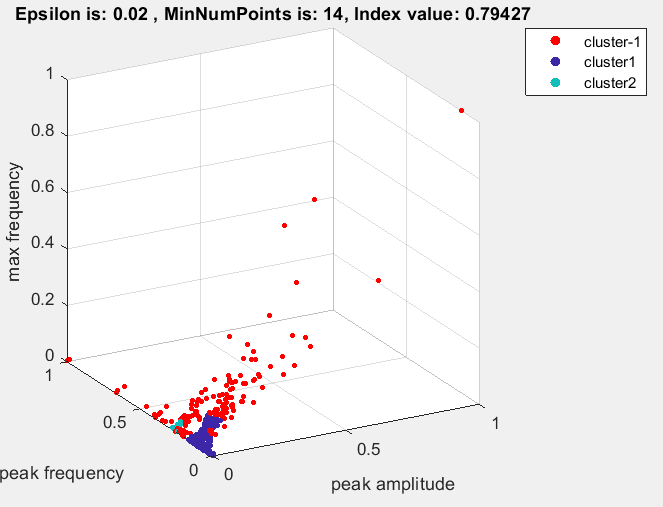
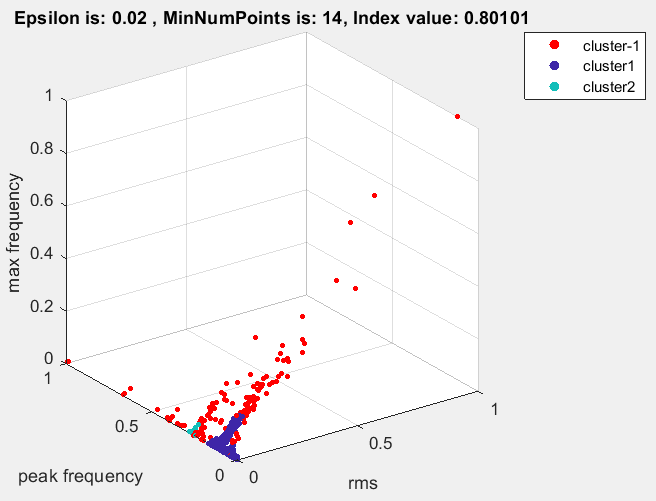
Promjenom algoritma selekcije značajki da za svaki podskup značajki prvo izračuna optimalni cluster rezultat pa onda prikazuje 10 najboljih subsetova.

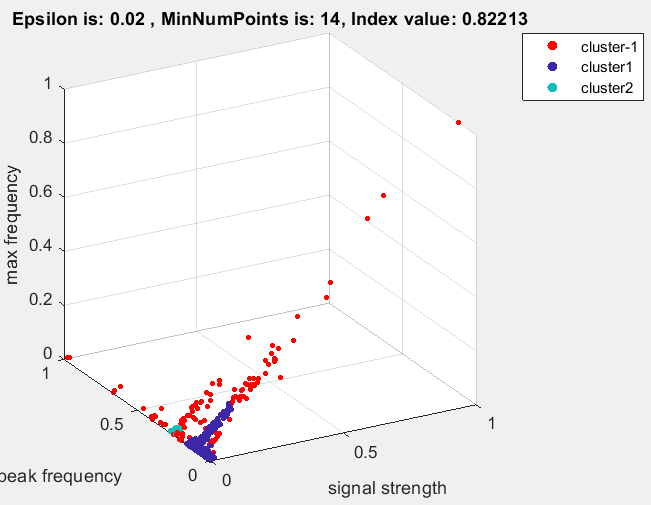
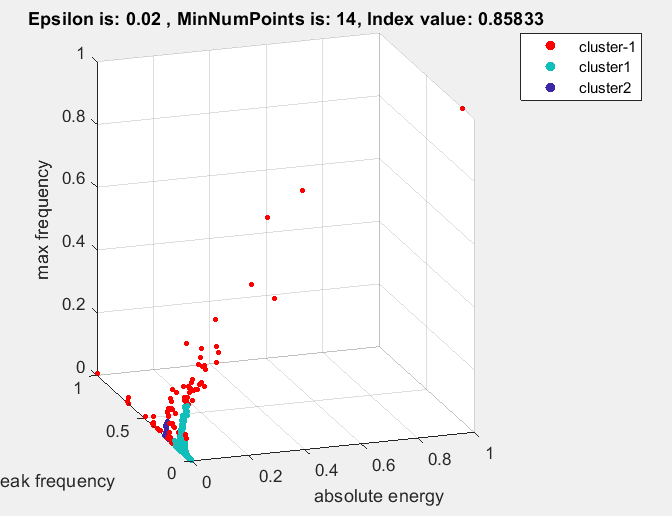


Vidimo kako abs energy zamjenjuje signal strength jer su dosta korelirani, možemo za smanjiti feature selection samo raditi s jednim od koreliranih značajki.



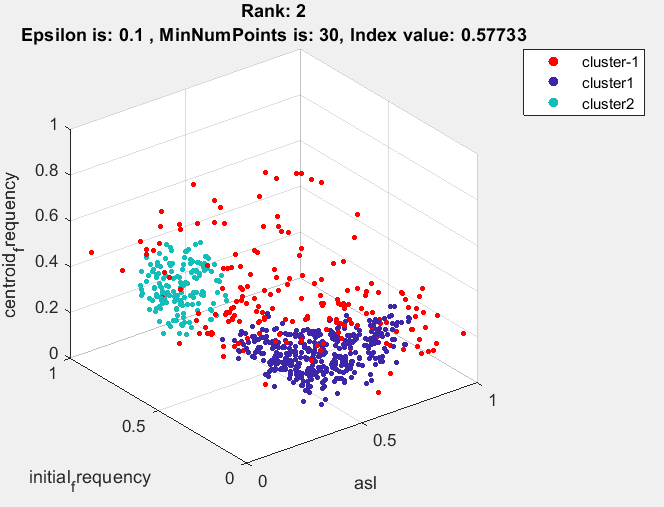
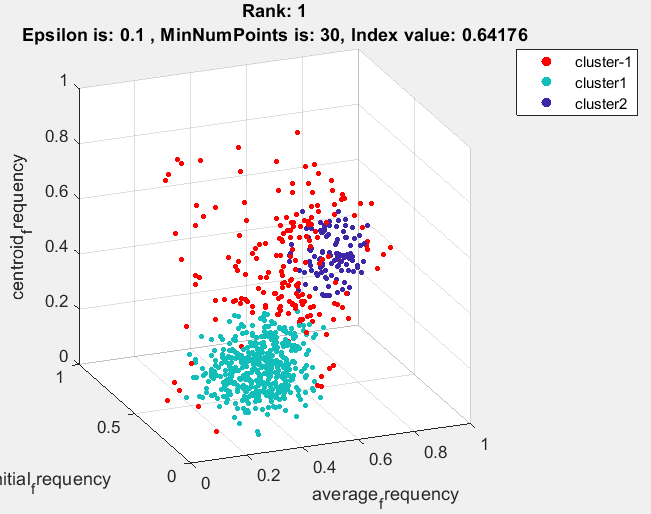
Za 4 korelirane značajke PEAK AMP, RMS, ABSOULUTE ENERGY i SIGNAL STRENGTH vidimo kako zamjenom jedne korelirane značajke s drugom u podskupu značajki čini slične skupove podataka te rezultira sličnim clusteringom.

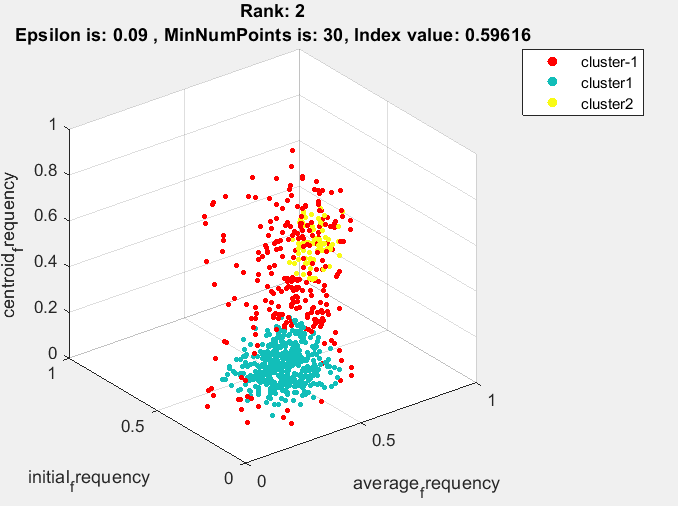
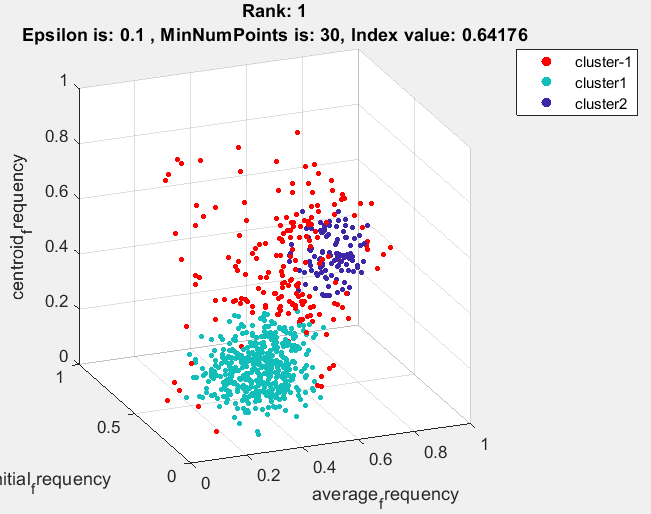
 

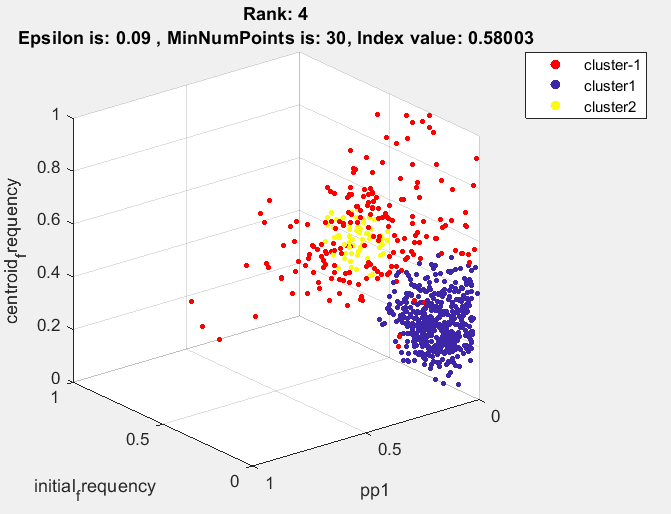
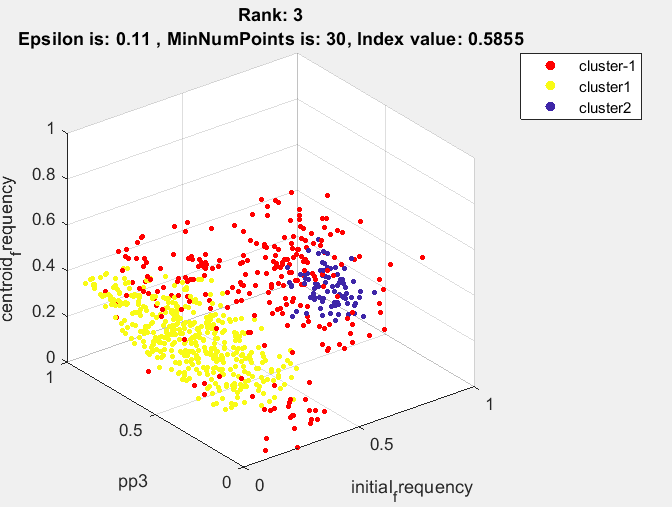
Želimo odrediti optimalne granice epsilona. Znamo da što je manji epsilon biti će potrebno smanjiti NminPoints za kvalitetan rezultat.

Za epsilon od 0.1, 0.15 i 0.2 uz NminPoints = 30 algoritam uvijek izabire epsilon = 0.1 te sam clustering nije optimalan gdje je najveća vrijednost indeksa 0.64176



Za NminPoints = 30 i epsilon 0.09 , 0.1, 0.11 dobivamo podvojene rezultate za najbolji epsilon gdje je najbolji rezultat za epsilon 0.1.





Još jedan koristan tip vizualizacije:

Obzirom da vrijeme ne uzimamo kao zasebnu značajku kod svrstavanja AE značajki u klastere, ne znamo:

* u koje vrijeme se formirao pojedini klaster
* koliko je točaka u nekom vremenskom intervalu detektirano da pripada pojedinom klasteru

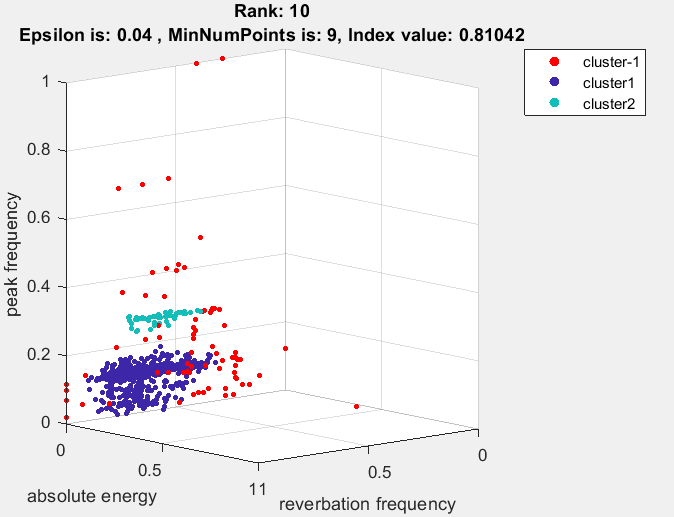
Ideja bi bila, **nakon** klasteriranja, vizualizirati na x-y grafovima u ovisnosti o vremenu:

* kumulativni broj AE emisija pridijeljen svakom klasteru (+ odvojeno neklasterirane točke)
* broj AE emisija pridijeljenih pojedinom klasteru po jednom satu

To pomaže npr. da se vidi npr. u kojoj fazi suše se javlja koji klaster i koliko je značajan.

To bi nam bilo korisno i za ugradbenu implementaciju.

Dobivamo zanimljivi rezultat za nevedeni skup značajki : PEAK FREQ, ABS ENERGY, REVB FREQ



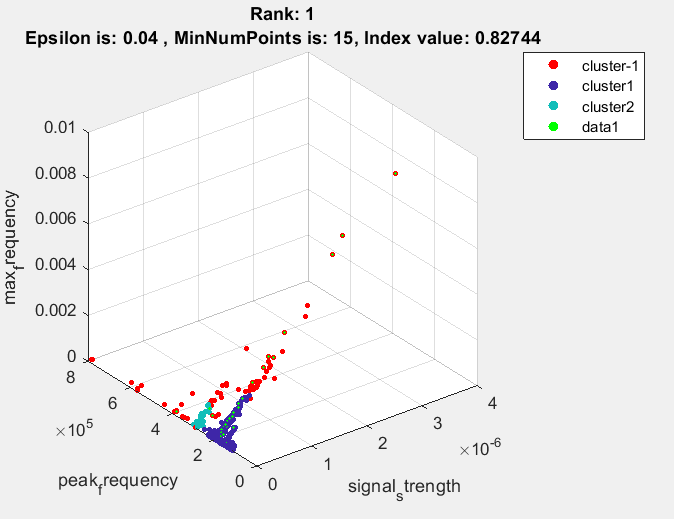
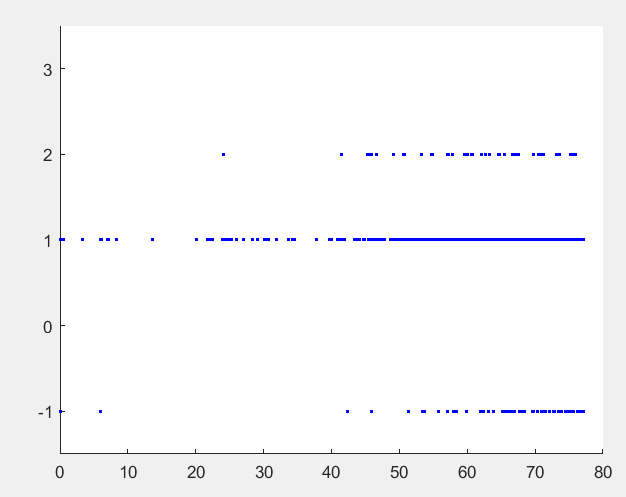
**Fiksiranje Nmin = 15 te određivanje max\_eps kao koljeno krivulje reachability plota:**

1. result: (signal\_strength, peak\_frequency, max\_frequency), Epsilon: 0.04, MinNumPoints: 15, DBCV index value: 0.82744

cluster -1had 9.322% single peaks

cluster 1had 89.8305% single peaks

cluster 2had 0.84746% single peaks

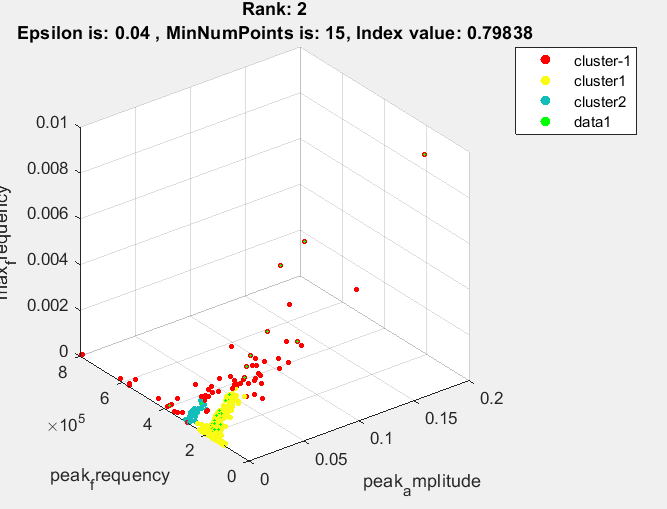
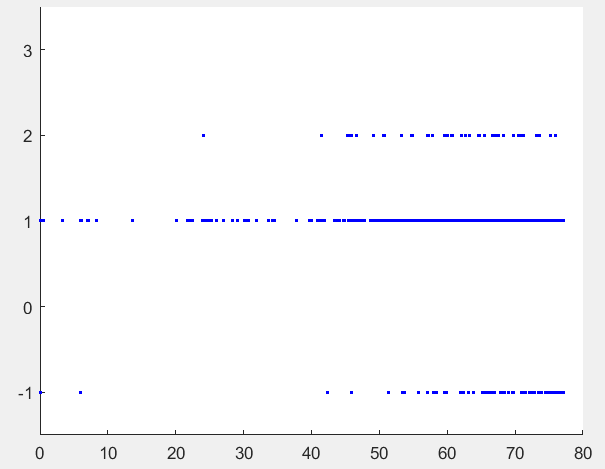
 

2. result: (peak\_amplitude, peak\_frequency, max\_frequency), Epsilon: 0.04, MinNumPoints: 15, DBCV index value: 0.79838

cluster -1had 8.4746% single peaks

cluster 1had 89.8305% single peaks

cluster 2had 1.6949% single peaks

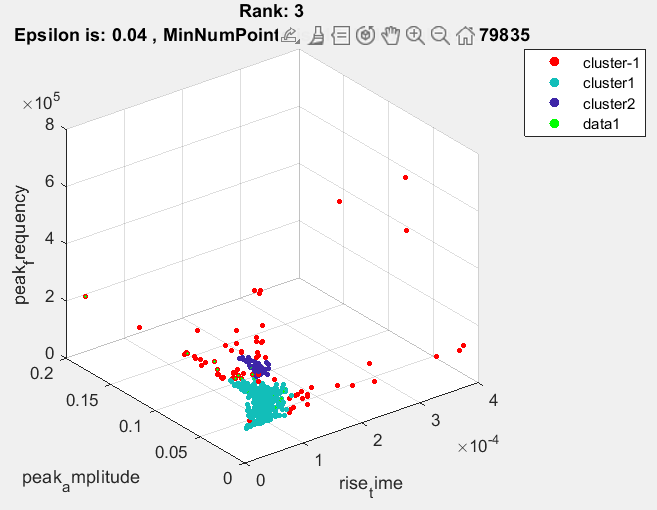
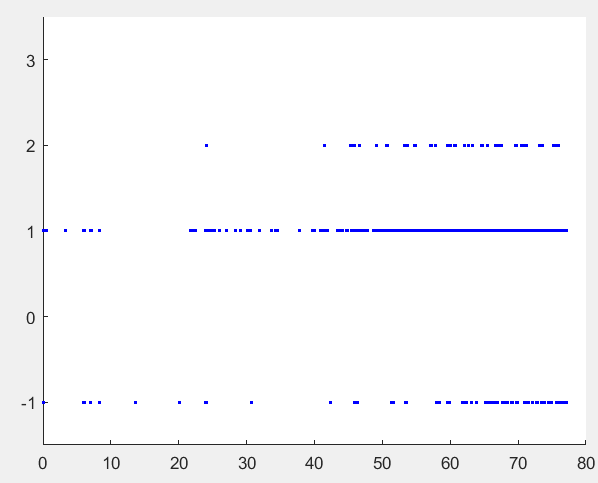
 

3. result: (rise\_time, peak\_amplitude, peak\_frequency), Epsilon: 0.04, MinNumPoints: 15, DBCV index value: 0.79835

cluster -1had 8.4746% single peaks

cluster 1had 90.678% single peaks

cluster 2had 0.84746% single peaks

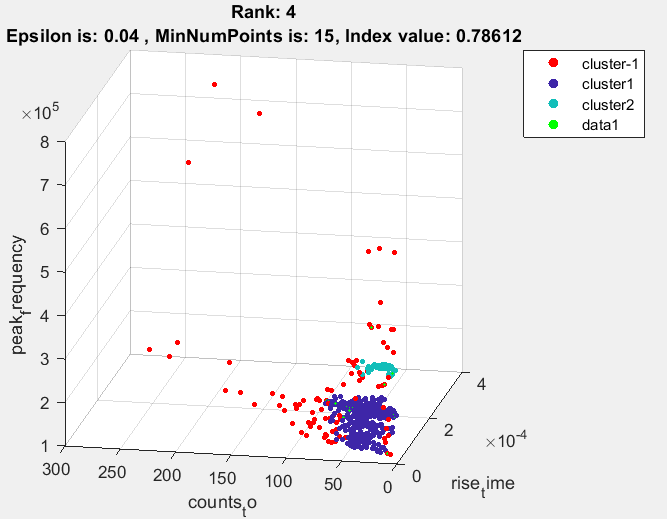
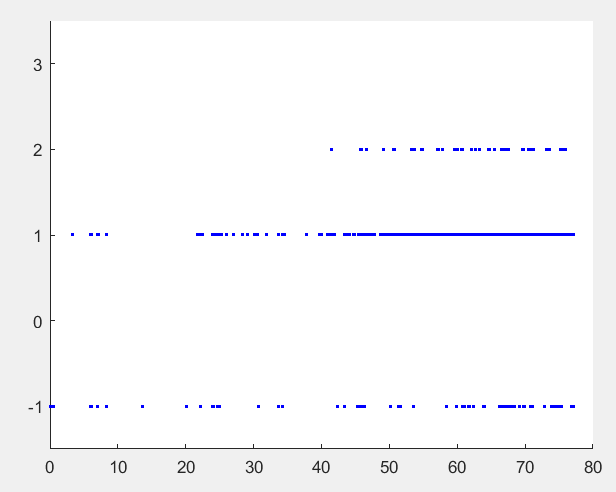
 

4. result: (rise\_time, counts\_to, peak\_frequency), Epsilon: 0.04, MinNumPoints: 15, DBCV index value: 0.78612

cluster -1had 4.2373% single peaks

cluster 1had 94.9153% single peaks

cluster 2had 0.84746% single peaks

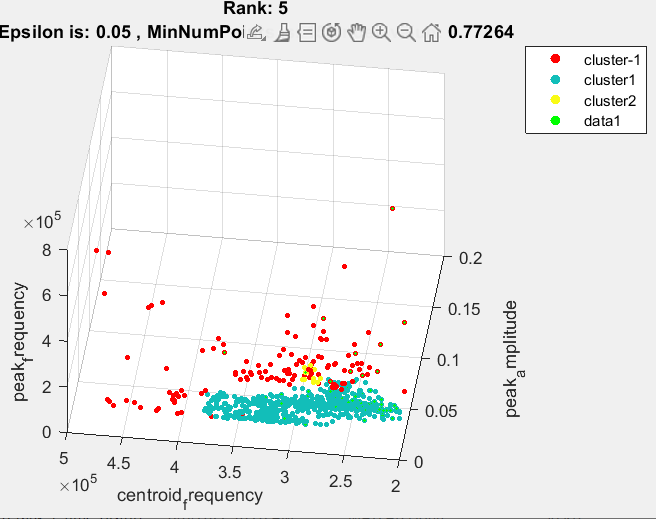
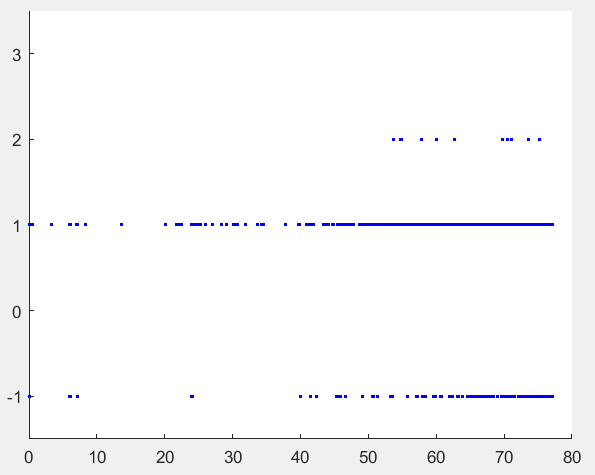
 

5. result: (peak\_amplitude, centroid\_frequency, peak\_frequency), Epsilon: 0.05, MinNumPoints: 15, DBCV index value: 0.77264

cluster -1had 8.4746% single peaks

cluster 1had 90.678% single peaks

cluster 2had 0.84746% single peaks

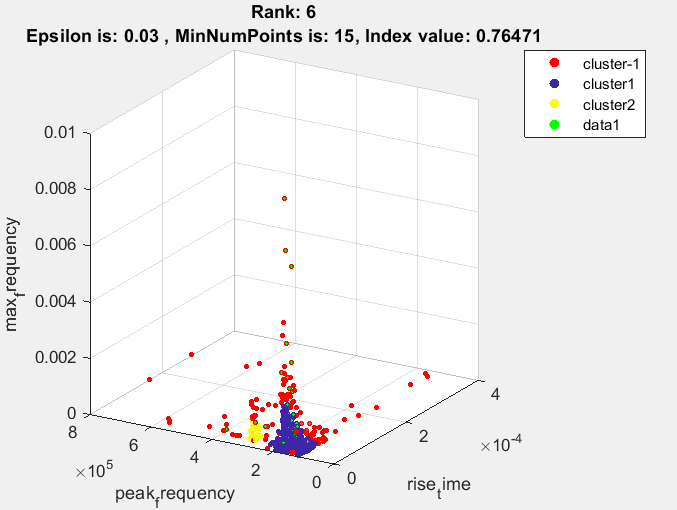
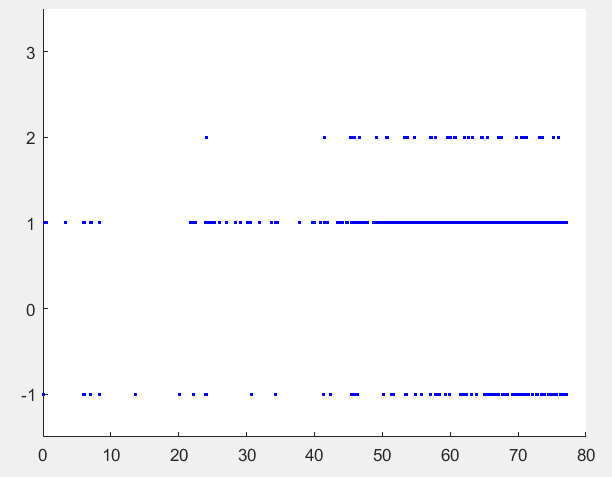
 

6. result: (rise\_time, peak\_frequency, max\_frequency), Epsilon: 0.03, MinNumPoints: 15, DBCV index value: 0.76471

cluster -1had 13.5593% single peaks

cluster 1had 86.4407% single peaks

cluster 2had 0% single peaks

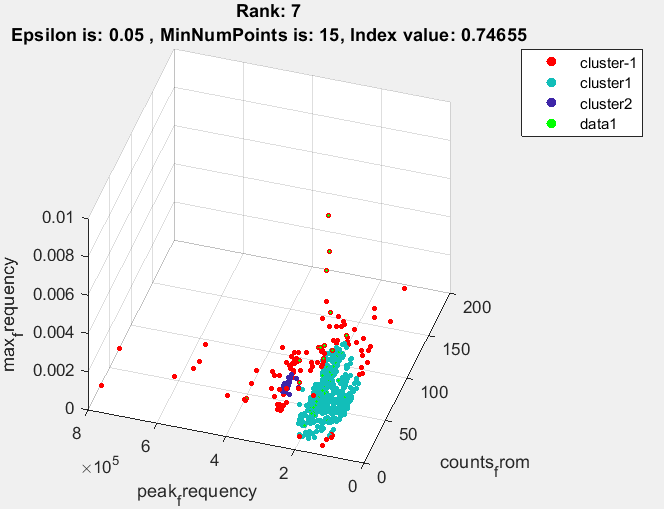
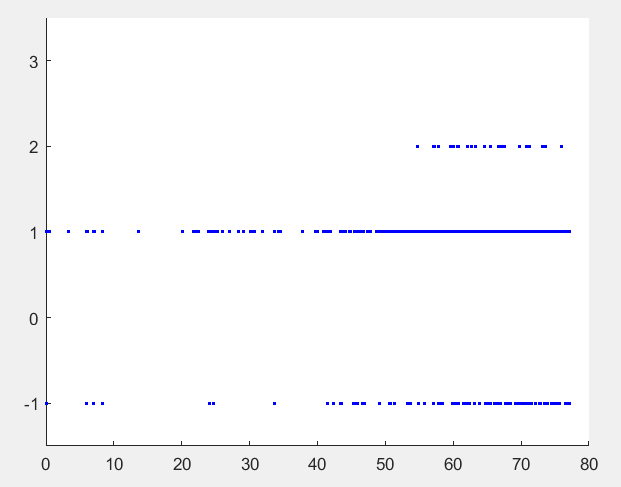
 

7. result: (counts\_from, peak\_frequency, max\_frequency), Epsilon: 0.05, MinNumPoints: 15, DBCV index value: 0.74655

cluster -1had 11.0169% single peaks

cluster 1had 88.9831% single peaks

cluster 2had 0% single peaks

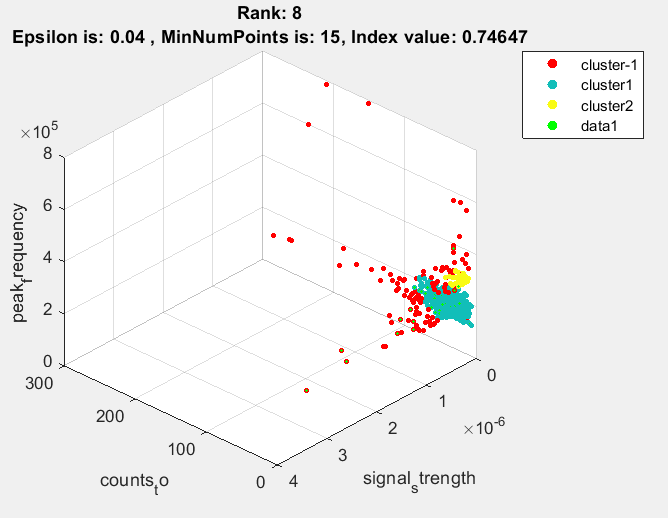
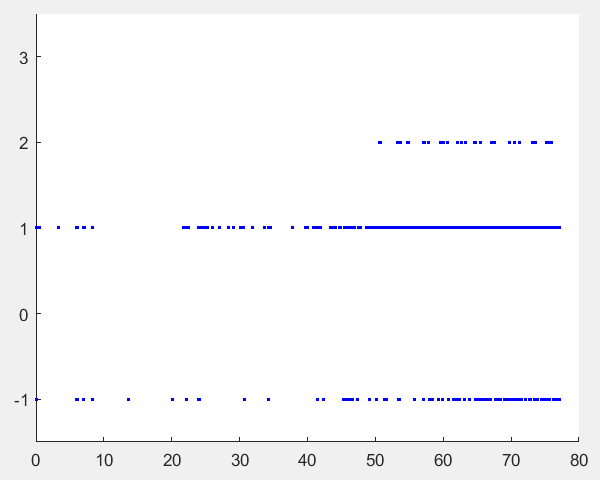
 

8. result: (counts\_to, signal\_strength, peak\_frequency), Epsilon: 0.04, MinNumPoints: 15, DBCV index value: 0.74647

cluster -1had 10.1695% single peaks

cluster 1had 88.9831% single peaks

cluster 2had 0.84746% single peaks

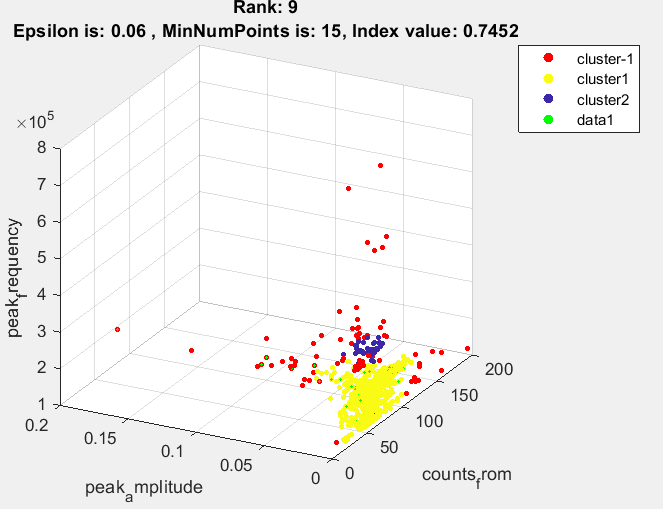
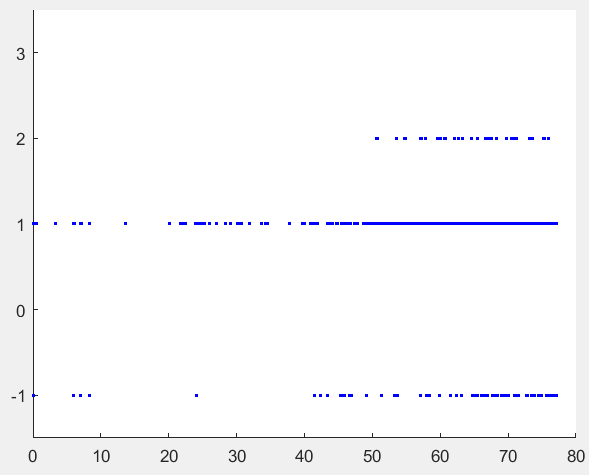
 

9. result: (counts\_from, peak\_amplitude, peak\_frequency), Epsilon: 0.06, MinNumPoints: 15, DBCV index value: 0.7452

cluster -1had 5.9322% single peaks

cluster 1had 92.3729% single peaks

cluster 2had 1.6949% single peaks

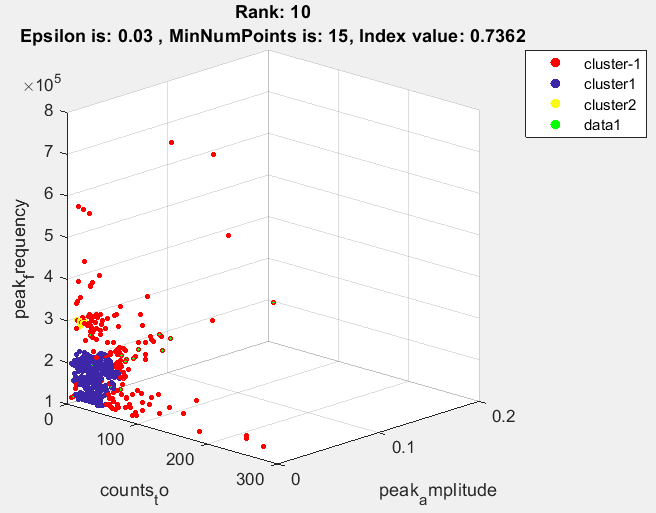
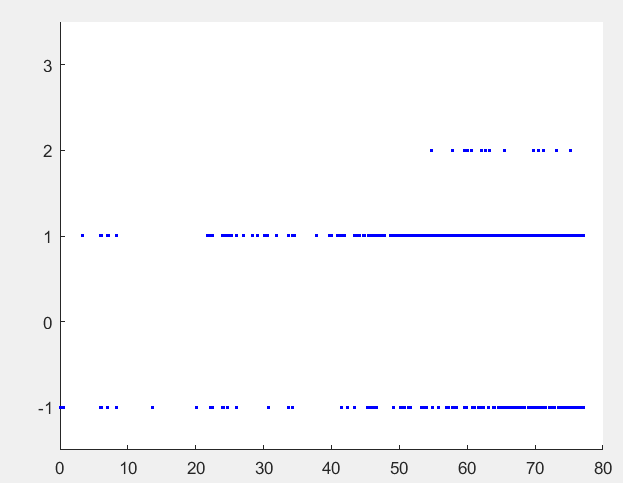
 

10. result: (counts\_to, peak\_amplitude, peak\_frequency), Epsilon: 0.03, MinNumPoints: 15, DBCV index value: 0.7362

cluster -1had 12.7119% single peaks

cluster 1had 87.2881% single peaks

cluster 2had 0% single peaks

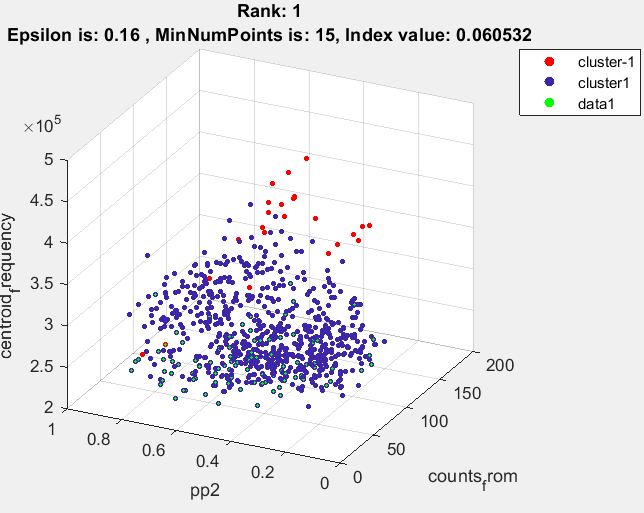
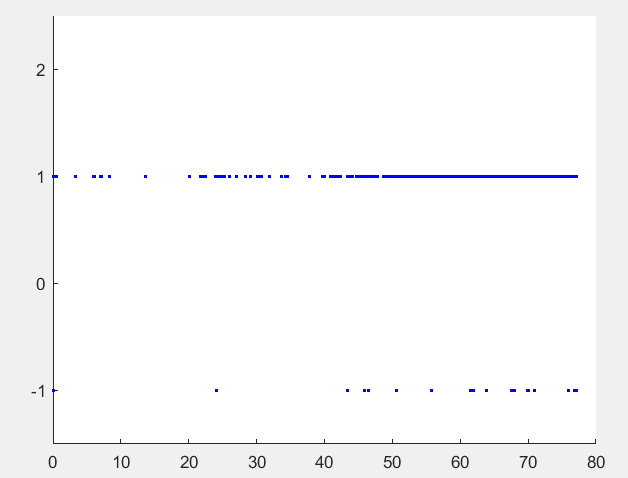
 

Najbolji rezultati za jedan pronađeni cluster:

Single cluster: 1. result: (counts\_from, pp2, centroid\_frequency), Epsilon: 0.16, MinNumPoints: 15, DBCV index value: 0.060532

cluster -1 had 0.84746% single peaks

cluster 1 had 99.1525% single peaks

Svi rezultati imaju vremensku distribuciju tijekom čitavog eksperimenta.

**Nmin od 9 do 14 te određivanje max\_eps kao koljeno normalizirane krivulje reachability plota:**

1. result: (signal\_strength, peak\_frequency, max\_frequency), Estimated Epsilon: 0.036924, Picked Epsilon: 0.026924, MinNumPoints: 14, DBCV index value: 0.84842

cluster -1 had 11.8644% single peaks

cluster 1 had 88.1356% single peaks

cluster 2 had 0% single peaks

2. result: (peak\_amplitude, peak\_frequency, max\_frequency), Estimated Epsilon: 0.039535, Picked Epsilon: 0.019535, MinNumPoints: 9, DBCV index value: 0.80914

cluster -1 had 14.4068% single peaks

cluster 1 had 85.5932% single peaks

cluster 2 had 0% single peaks

3. result: (rise\_time, peak\_amplitude, peak\_frequency), Estimated Epsilon: 0.050828, Picked Epsilon: 0.030828, MinNumPoints: 9, DBCV index value: 0.8055

cluster -1 had 9.322% single peaks

cluster 1 had 89.8305% single peaks

cluster 2 had 0.84746% single peaks

4. result: (rise\_time, counts\_to, peak\_frequency), Estimated Epsilon: 0.049931, Picked Epsilon: 0.039931, MinNumPoints: 13, DBCV index value: 0.78855

cluster -1 had 4.2373% single peaks

cluster 1 had 94.9153% single peaks

cluster 2 had 0.84746% single peaks

5. result: (counts\_from, peak\_amplitude, peak\_frequency), Estimated Epsilon: 0.064638, Picked Epsilon: 0.054638, MinNumPoints: 10, DBCV index value: 0.78577

cluster -1 had 5.0847% single peaks

cluster 1 had 93.2203% single peaks

cluster 2 had 1.6949% single peaks

6. result: (peak\_amplitude, pp1, peak\_frequency), Estimated Epsilon: 0.04559, Picked Epsilon: 0.03559, MinNumPoints: 12, DBCV index value: 0.7816

cluster -1 had 7.6271% single peaks

cluster 1 had 91.5254% single peaks

cluster 2 had 0.84746% single peaks

7. result: (pp1, peak\_frequency, max\_frequency), Estimated Epsilon: 0.041806, Picked Epsilon: 0.031806, MinNumPoints: 11, DBCV index value: 0.77573

cluster -1 had 9.322% single peaks

cluster 1 had 89.8305% single peaks

cluster 2 had 0.84746% single peaks

8. result: (counts\_to, peak\_amplitude, peak\_frequency), Estimated Epsilon: 0.055068, Picked Epsilon: 0.035068, MinNumPoints: 13, DBCV index value: 0.77353

cluster -1 had 9.322% single peaks

cluster 1 had 89.8305% single peaks

cluster 2 had 0.84746% single peaks

9. result: (rise\_time, peak\_frequency, max\_frequency), Estimated Epsilon: 0.046756, Picked Epsilon: 0.036756, MinNumPoints: 11, DBCV index value: 0.76882

cluster -1 had 9.322% single peaks

cluster 1 had 89.8305% single peaks

cluster 2 had 0.84746% single peaks

10. result: (rise\_time, peak\_amplitude, centroid\_frequency), Estimated Epsilon: 0.055749, Picked Epsilon: 0.055749, MinNumPoints: 9, DBCV index value: 0.7587

cluster -1 had 4.2373% single peaks

cluster 1 had 95.7627% single peaks

cluster 2 had 0% single peaks

**Želimo fiksni Nmin ako za drugi Nmin i sve moguće kombinacije epsilon isti feature set ima slične rezultate dbcv indeksa. Iz matrice dbcv indeksa za naveden kombinacije eps i Nmin parametara te određene podksupove značajki vidimo baš za najbolje izabarne feature setove.**

**Npr.** (signal\_strength, peak\_frequency, max\_frequency) (52. element)

|  |  |  |  |
| --- | --- | --- | --- |
| **Nmin/epsilon** | **Procjena** | **Procjena -0.01** | **Procjena – 0.02** |
| **9** | 0.0025 | 0.7316 | **0.7673** |
| **10** | 0.7664 | **0.8356** | 0.5617 |
| **11** | 0.8247 | **0.8369** | 0.7705 |
| **12** | 0.8260 | **0.8394** | 0.7579 |
| **13** | **0.8272** | 0.8126 | 0.7819 |
| **14** | 0.8274 | **0.8484** | 0.7858 |

**Npr.** (peak\_amplitude, peak\_frequency, max\_frequency) (79. element)

|  |  |  |  |
| --- | --- | --- | --- |
| **Nmin/epsilon** | **Procjena** | **Procjena -0.01** | **Procjena – 0.02** |
| **9** | 0.8056 | 0.7901 | **0.8091** |
| **10** | 0.0022 | 0.7913 | **0.8042** |
| **11** | 0.0027 | 0.7936 | **0.8029** |
| **12** | 0.0028 | **0.7948** | 0.7793 |
| **13** | 0.0028 | **0.8019** | 0.7865 |
| **14** | 0.0028 | 0.3806 | **0.7843** |

**Single cluster rezultati: datoteke: DBCV\_single\_clust\_features\_selected\_for\_Nmin\_9\_14\_Eps\_0\_0058583\_0\_12096\_rank\_1-10**

1. result: (counts\_from, pp2, centroid\_frequency), Estimated Epsilon: 0.11784, Picked Epsilon: 0.10784, MinNumPoints: 13, DBCV\_single\_clust index value: 0.024425

cluster -1 had 6.7797% single peaks

cluster 1 had 93.2203% single peaks

2. result: (counts\_from, pp3, centroid\_frequency), Estimated Epsilon: 0.10186, Picked Epsilon: 0.10186, MinNumPoints: 13, DBCV\_single\_clust index value: 0.02275

cluster -1 had 7.6271% single peaks

cluster 1 had 92.3729% single peaks

3. result: (pp1, pp4, peak\_frequency), Estimated Epsilon: 0.084742, Picked Epsilon: 0.084742, MinNumPoints: 11, DBCV\_single\_clust index value: 0.018438

cluster -1 had 1.6949% single peaks

cluster 1 had 98.3051% single peaks

4. result: (counts\_to, pp2, centroid\_frequency), Estimated Epsilon: 0.080206, Picked Epsilon: 0.080206, MinNumPoints: 9, DBCV\_single\_clust index value: 0.018313

cluster -1 had 15.2542% single peaks

cluster 1 had 84.7458% single peaks

5. result: (pp2, centroid\_frequency, weighted\_peak\_frequency), Estimated Epsilon: 0.081615, Picked Epsilon: 0.081615, MinNumPoints: 12, DBCV\_single\_clust index value: 0.018061

cluster -1 had 4.2373% single peaks

cluster 1 had 95.7627% single peaks

6. result: (pp2, pp4, peak\_frequency), Estimated Epsilon: 0.081793, Picked Epsilon: 0.071793, MinNumPoints: 13, DBCV\_single\_clust index value: 0.017765

cluster -1 had 1.6949% single peaks

cluster 1 had 98.3051% single peaks

7. result: (pp3, pp4, centroid\_frequency), Estimated Epsilon: 0.079862, Picked Epsilon: 0.079862, MinNumPoints: 14, DBCV\_single\_clust index value: 0.017746

cluster -1 had 3.3898% single peaks

cluster 1 had 96.6102% single peaks

8. result: (pp2, pp4, centroid\_frequency), Estimated Epsilon: 0.084647, Picked Epsilon: 0.084647, MinNumPoints: 14, DBCV\_single\_clust index value: 0.01529

cluster -1 had 1.6949% single peaks

cluster 1 had 98.3051% single peaks

9. result: (counts\_to, pp1, pp3), Estimated Epsilon: 0.085511, Picked Epsilon: 0.085511, MinNumPoints: 14, DBCV\_single\_clust index value: 0.014828

cluster -1 had 3.3898% single peaks

cluster 1 had 96.6102% single peaks

10. result: (counts\_from, pp1, pp4), Estimated Epsilon: 0.086774, Picked Epsilon: 0.086774, MinNumPoints: 12, DBCV\_single\_clust index value: 0.01467

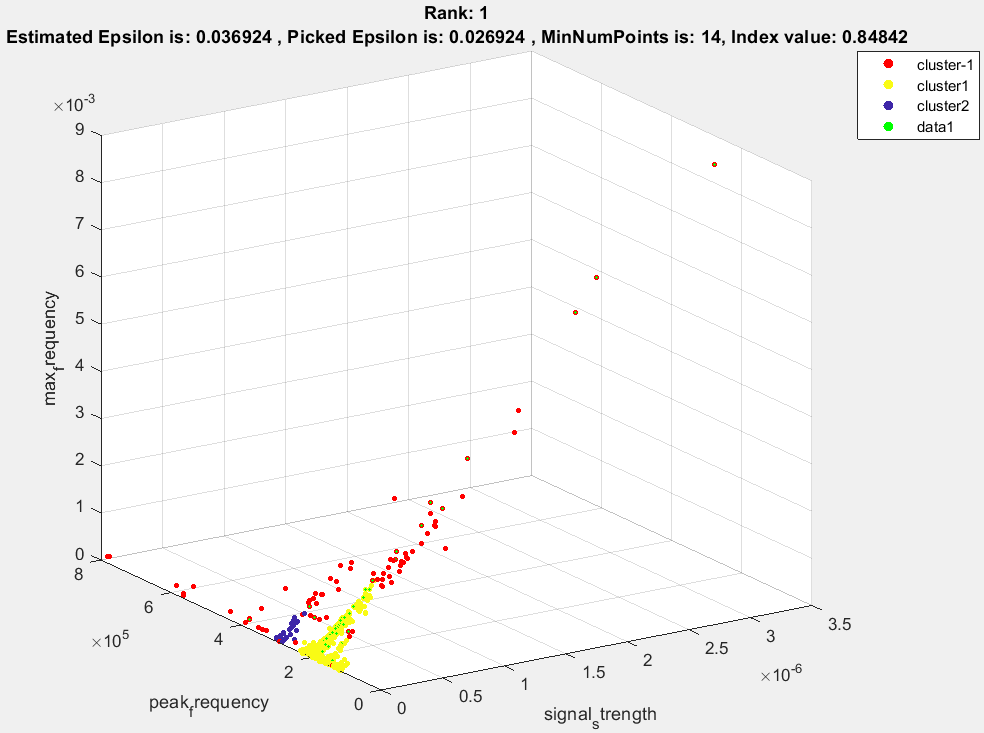
cluster -1 had 2.5424% single peaks

cluster 1 had 97.4576% single peaks

**Provođenje voting sheme na fiksnom Nmin=14 i epsilon od koljena 3\*0.01 prema gore i dolje:**

**1. result:** (signal\_strength, peak\_frequency, max\_frequency),

**Estimated Epsilon: 0.036924, Picked Epsilon: 0.026924**, MinNumPoints: 14, DBCV index value: 0.84842



**2. result:** (rise\_time, peak\_amplitude, peak\_frequency),

**Estimated Epsilon: 0.057864, Picked Epsilon: 0.037864**, MinNumPoints: 14, DBCV index value: 0.79715

**3. result:** (peak\_amplitude, peak\_frequency, max\_frequency),

**Estimated Epsilon: 0.051468, Picked Epsilon: 0.031468,** MinNumPoints: 14, DBCV index value: 0.78427

**4. result:** (peak\_amplitude, pp1, peak\_frequency),

**Estimated Epsilon: 0.04284, Picked Epsilon: 0.03284**, MinNumPoints: 14, DBCV index value: 0.7641

**5. result:** (rise\_time, peak\_frequency, max\_frequency),

**Estimated Epsilon: 0.047773, Picked Epsilon: 0.027773**, MinNumPoints: 14, DBCV index value: 0.76224

**6. result:** (signal\_strength, pp1, peak\_frequency),

**Estimated Epsilon: 0.04675, Picked Epsilon: 0.03675**, MinNumPoints: 14, DBCV index value: 0.75568

**7. result:** (rise\_time, counts\_to, peak\_frequency),

**Estimated Epsilon: 0.053326, Picked Epsilon: 0.033326**, MinNumPoints: 14, DBCV index value: 0.75365

**8. result:** (signal\_strength, centroid\_frequency, peak\_frequency),

**Estimated Epsilon: 0.066667, Picked Epsilon: 0.056667**, MinNumPoints: 14, DBCV index value: 0.75346

**9. result:** (counts\_to, signal\_strength, peak\_frequency),

**Estimated Epsilon: 0.060104, Picked Epsilon: 0.040104**, MinNumPoints: 14, DBCV index value: 0.75247

**10. result:** (counts\_to, peak\_frequency, max\_frequency),

**Estimated Epsilon: 0.059598, Picked Epsilon: 0.039598**, MinNumPoints: 14, DBCV index value: 0.75201

Iz navedenog rezultata vidimo da preferirani epsilon je uvijek niži od prvobitno procijenjenog epsilon pomoću koljena krivulje reachability plota. Isto tako Preferirani epsilon je niži za 0.01 ili 0.02. Time za buduće računanje optimlano je ograničiti epsilon od prvobitno procijenjenog epsilon do za 0.02 niži.

**Kod single cluster rezultat vidi se da se uvijek preferira što viša vrijednost epsilona od procijenjene. Što znači da je procijenjeni epsilon taman na granici da clusteriranjem ovaja na više od jedan cluster.**

**1. result:** (counts\_from, pp1, centroid\_frequency),

**Estimated Epsilon: 0.094146, Picked Epsilon: 0.12415**, MinNumPoints: 14, DBCV\_single\_clust index value: 0.067152

**2. result:** (counts\_from, pp2, centroid\_frequency),

**Estimated Epsilon: 0.12096, Picked Epsilon: 0.15096**, MinNumPoints: 14, DBCV\_single\_clust index value: 0.060421

**3. result:** (pp1, pp3, pp4),

**Estimated Epsilon: 0.10619, Picked Epsilon: 0.13619**, MinNumPoints: 14, DBCV\_single\_clust index value: 0.032587

**4. result:** (counts\_from, pp1, pp2),

**Estimated Epsilon: 0.091842, Picked Epsilon: 0.12184**, MinNumPoints: 14, DBCV\_single\_clust index value: 0.027771

**5. result:** (counts\_from, pp3, centroid\_frequency),

**Estimated Epsilon: 0.10525, Picked Epsilon: 0.13525,** MinNumPoints: 14, DBCV\_single\_clust index value: 0.022269

**6. result:** (counts\_to, pp1, peak\_frequency),

**Estimated Epsilon: 0.088149, Picked Epsilon: 0.11815,** MinNumPoints: 14, DBCV\_single\_clust index value: 0.021379

**7. result:** (counts\_from, pp4, centroid\_frequency),

**Estimated Epsilon: 0.073603, Picked Epsilon: 0.1036,** MinNumPoints: 14, DBCV\_single\_clust index value: 0.021028

**8. result:** (counts\_to, pp3, centroid\_frequency),

**Estimated Epsilon: 0.099178, Picked Epsilon: 0.12918,** MinNumPoints: 14, DBCV\_single\_clust index value: 0.020663

**9. result:** (counts\_from, pp1, pp4),

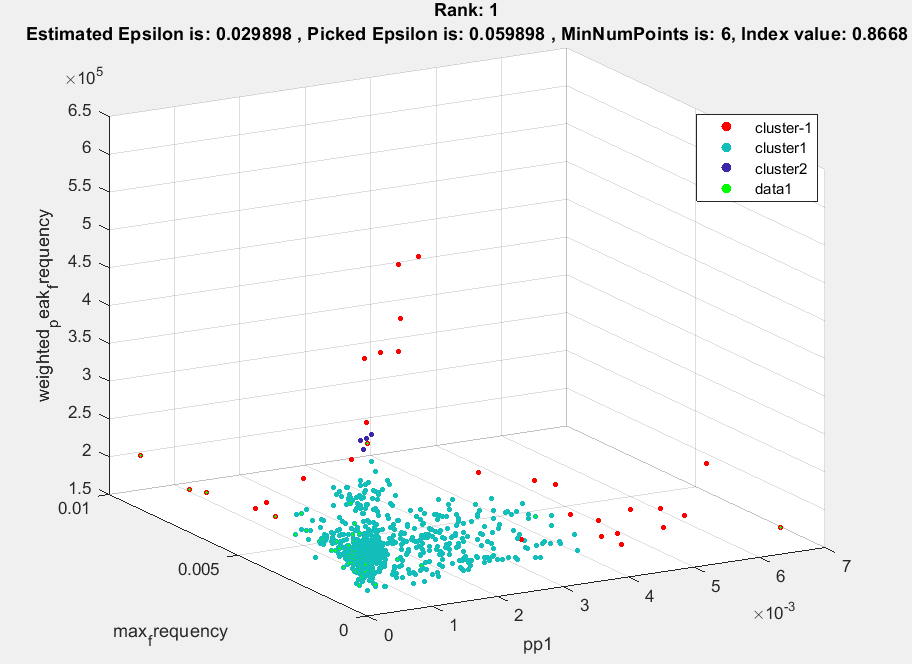
**Estimated Epsilon: 0.10086, Picked Epsilon: 0.13086,** MinNumPoints: 14, DBCV\_single\_clust index value: 0.020317

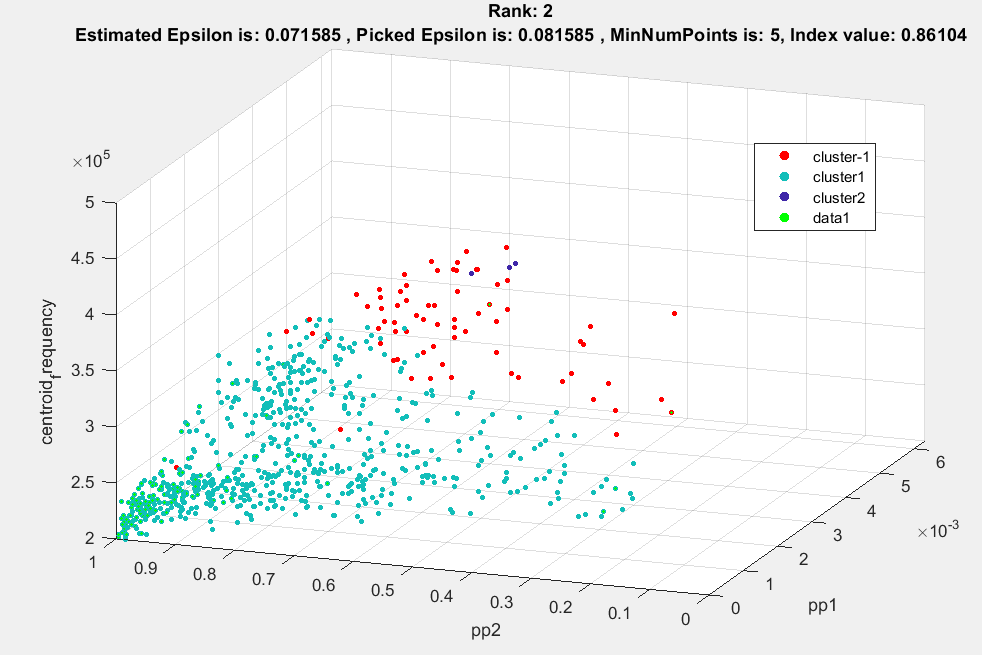
**10. result:** (pp1, pp2, pp4),

**Estimated Epsilon: 0.10824, Picked Epsilon: 0.12824,** MinNumPoints: 14, DBCV\_single\_clust index value: 0.020155

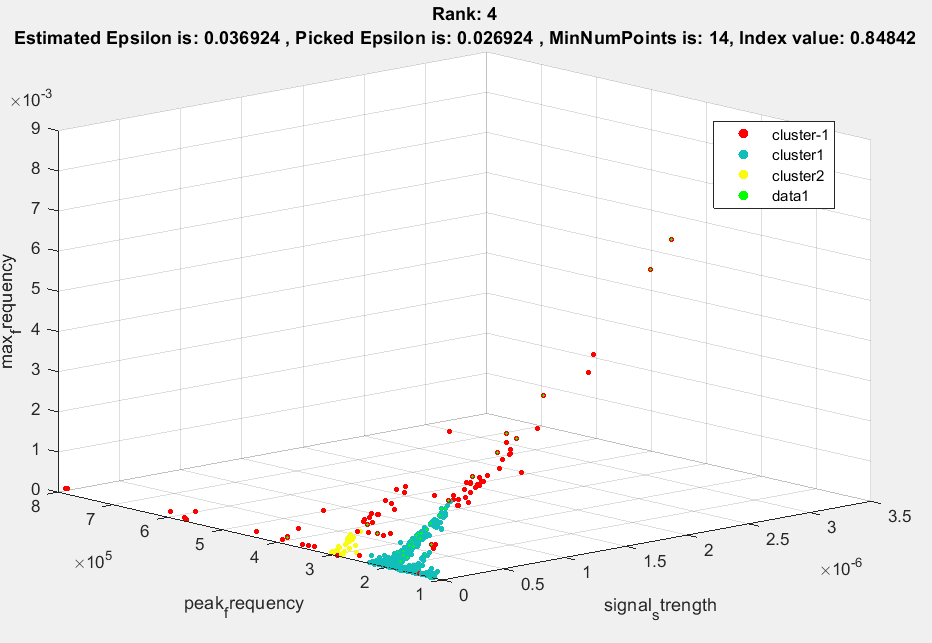
Provođenje voting sheme na premalim Nmin će DBCV indeks izabrati jedan cluster s većinom točaka te drugi cluster s jako malim brojem točaka koliko mu Nmin dopušta. Time će density sparsness cluster biti jako mali te će vrijednost indeksa biti velika.

**Provjeri!**

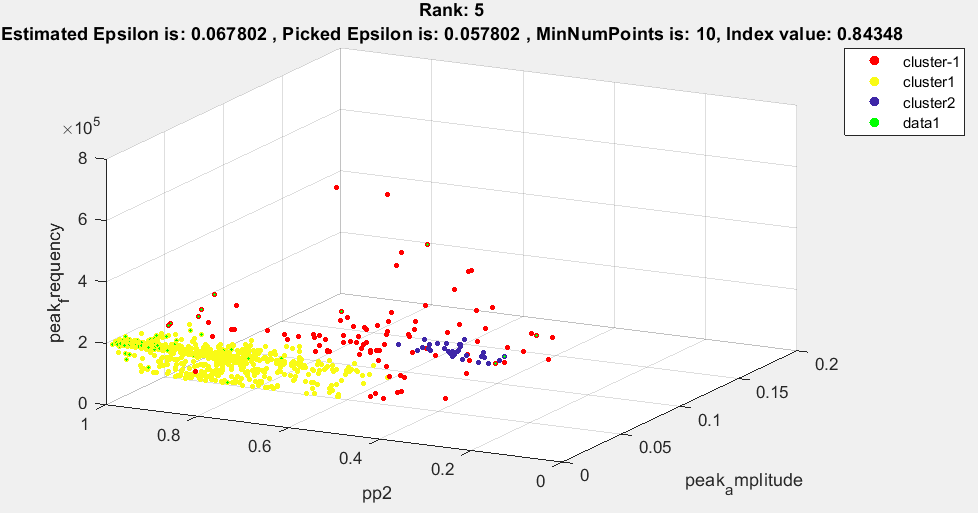




**Tek na 4. mjestu dobivamo pravi najbolji subset:**



**Ostali rezultati prate prijašnje opisanu mehaniku velikog clustera i jako malog do Nmin = 10.**



**Navedeno se može izbjeći ograničavanjem Nmin na vrijednosti veće od 10 i postavljanje large\_perc parametra u gradient clustering algoritmu na manje vrijednosti od 1 (100%). Navedeni parametar određuje maksimalnu veličinu clustera u odnosu na čitavi skup podataka. U slučaju da je 0.8 to znači da će cluster biti identificiran kao cluster samo ako sadrži manje od 80% točaka skupa podataka.**

**Najbolji subseti i clusteri povećavanjem granice Nmin su slijedeći:**

1. **Za Nmin = [11,20]**

**Problem gustoće:**

Povećanjem gustoće za malo može dovesti do spajanja dva veća clustera ako su im krajnje točke dovoljno blizu. Želimo smanjiti ili povećati broj točaka unutar određenog klustera mijenjanjem parametara, no time gustoća utječe na sve clustere

Veliki dataset:

