

HDBSCAN

With Log-Transformed Data

PMS

23 May, 2023

Introduction

The following analysis is an implementation of the [HDBSCAN](#) algorithm on individual, log-scaled proximity measures (no auxiliary variables included).

Please note that for the following plots, HDBSCAN considers cluster 1 to be “noise points” and are thus not part of an actual cluster.

Assumptions of the Algorithm

[This](#) fast implementation of HDBSCAN (Campello et al., 2013) computes the hierarchical cluster tree representing density estimates along with the stability-based flat cluster extraction. HDBSCAN essentially computes the hierarchy of all DBSCAN* clusterings, and then uses a stability-based extraction method to find optimal cuts in the hierarchy, thus producing a flat solution.

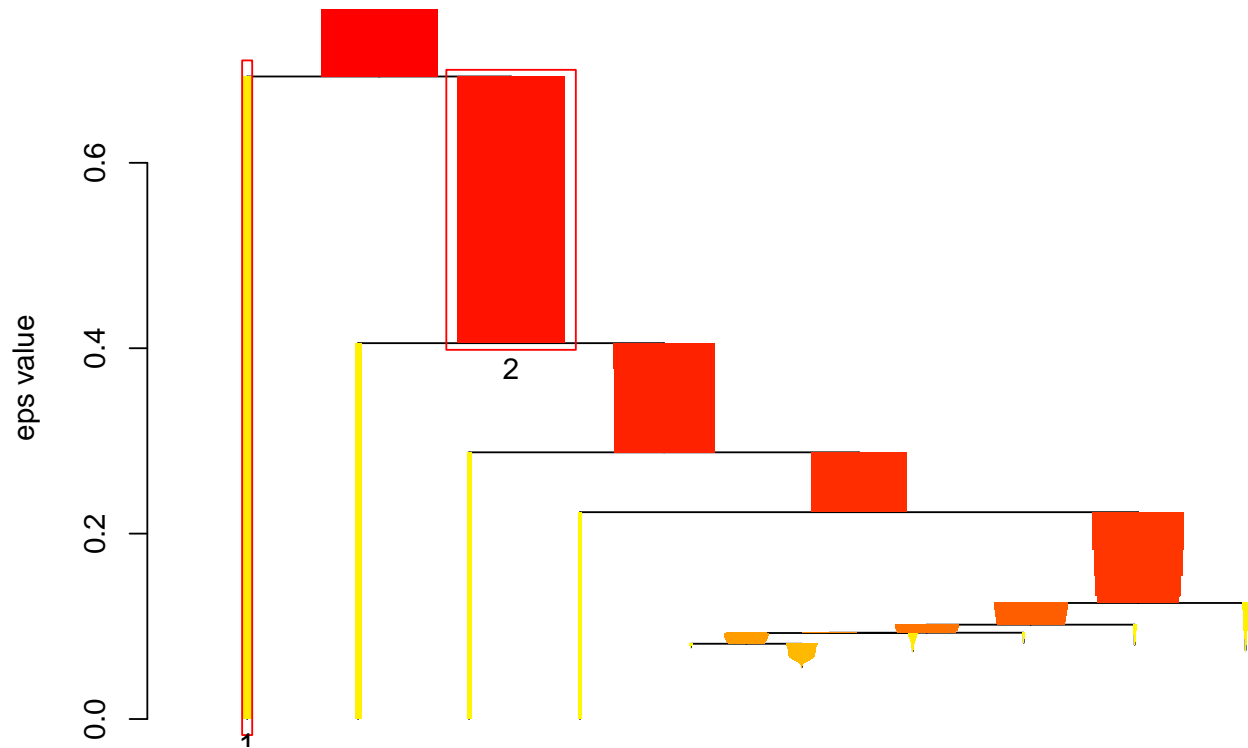
HDBSCAN performs the following steps:

- Compute mutual reachability distance mrd between points (based on distances and core distances).
 - Use mdr as a distance measure to construct a minimum spanning tree.
 - Prune the tree using stability.
 - Extract the clusters.
-

Amenities

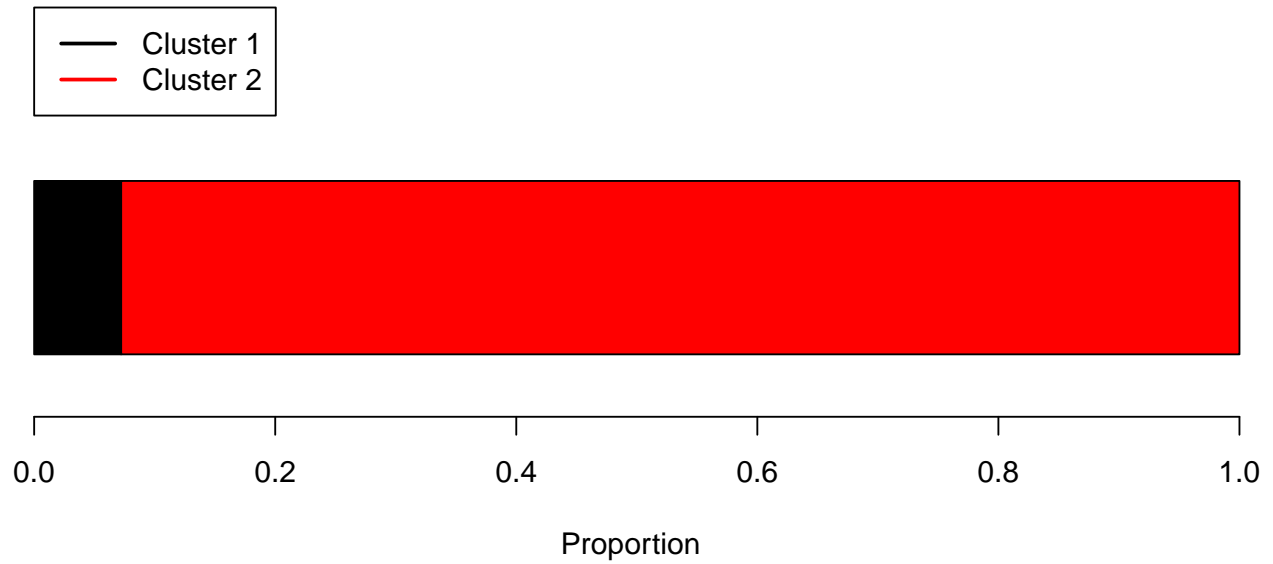
Employment

HDBSCAN*



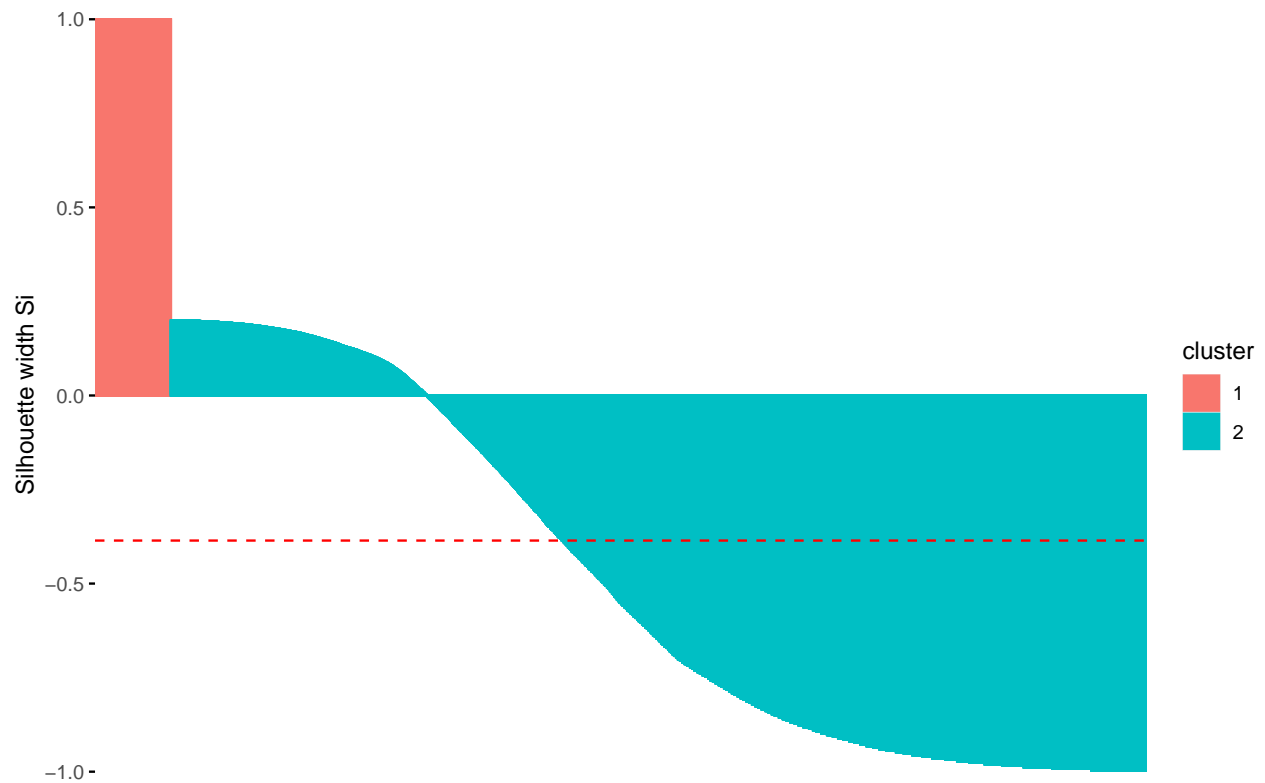
```
## [1] "Silhouette coefficient: 0.669838211728378"  
## [1] "Xie Beni coefficient: 27759.4905301196"  
## [1] "Davies Bouldin coefficient: 0.42742402852972"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.16035  
##   cluster  size ave.sil.width  
## 1         1  1521          1.00  
## 2         2 19546         -0.49
```

Clusters silhouette plot
Average silhouette width: -0.39



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
## Cluster 1 Cluster 2
##      1787      22696
##
##
##
## DB Population:
## Cluster 1 Cluster 2
##      67.9      71.8
##
##
##
## CSD Population:
## Cluster 1 Cluster 2
## 230600.9 241133.7
##
##
##
## CMA Type:
## Cluster 1 Cluster 2
##      813      9814
## B      732      9580
## D      174      2486
## K       68       816
```

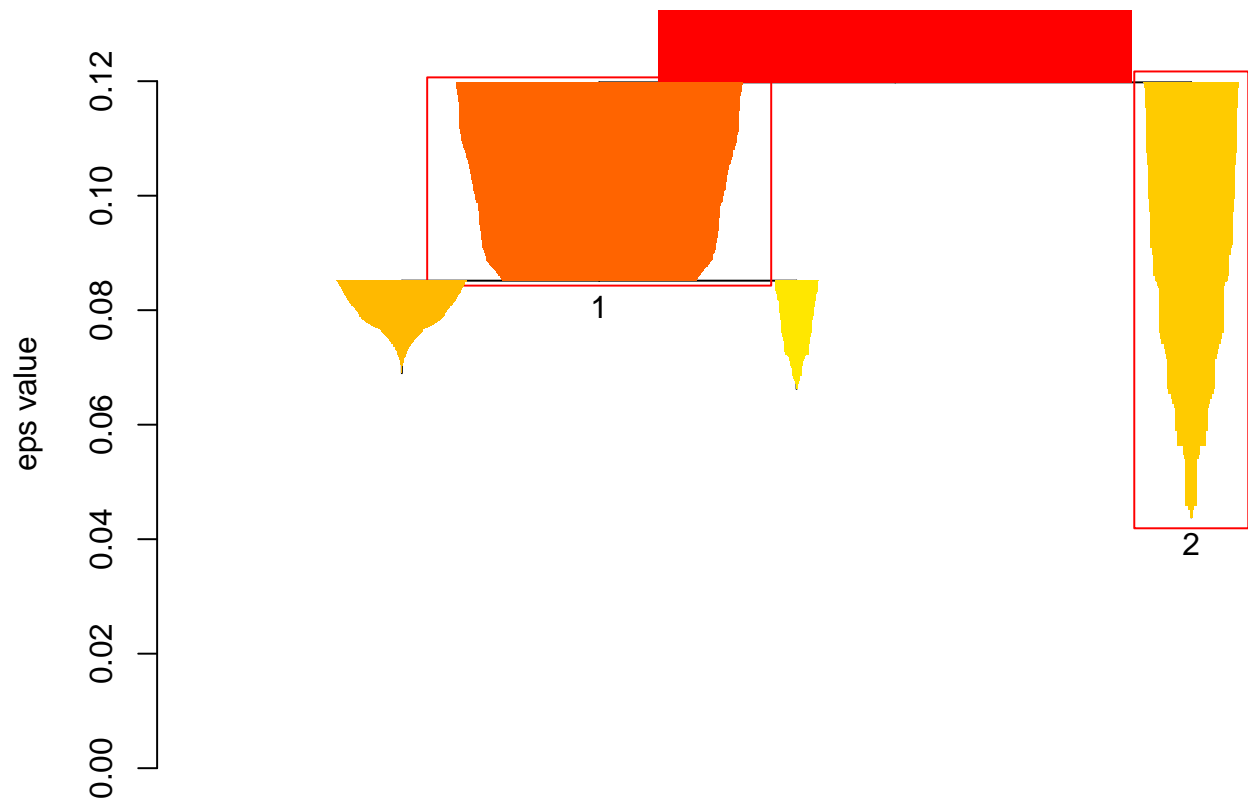
```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2
##      0.225      0.226
##
##
##
## Provinces:
##              Cluster 1 Cluster 2
## Alberta              46       717
## BritishColumbia      90      1109
## NewBrunswick         16       191
## NorthwestTerritories  0        16
## NovaScotia           55       724
## Ontario              275      3583
## Quebec               105      1273
## Saskatchewan          6       115
## NA's                 1194     14968
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2
## 0      1629     20489
## 1       121     1694
## 2        19      248
## F        18      265
##
##
##
## PMS_prox_idx_emp :
## Cluster 1 Cluster 2
##      0.02616     0.0256

```

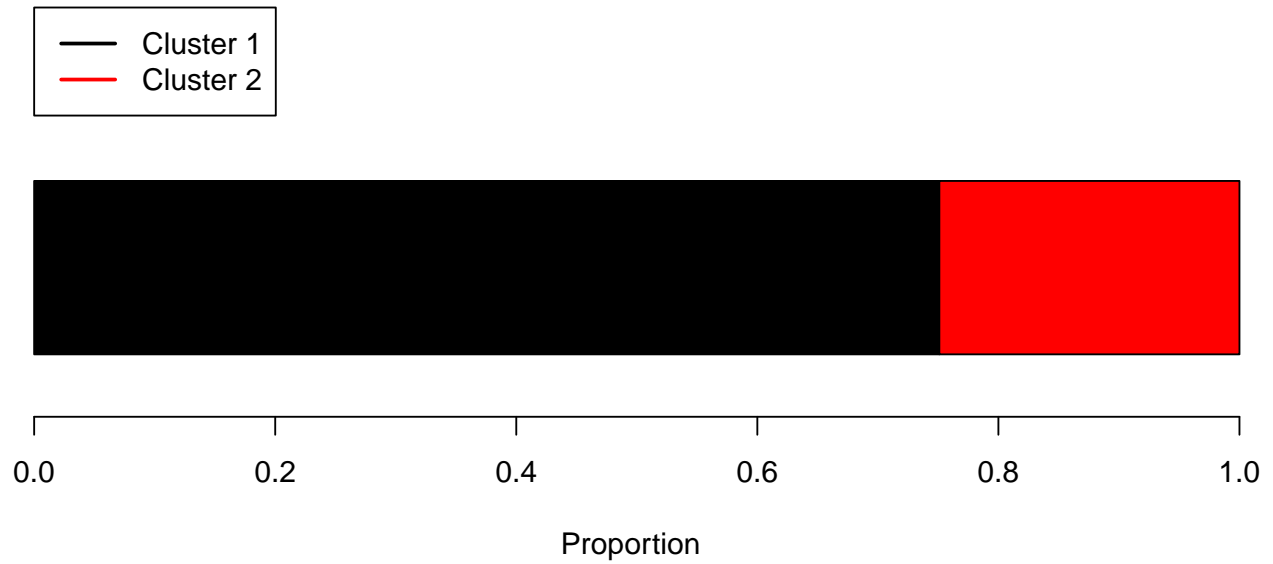
Pharmacy

HDBSCAN*

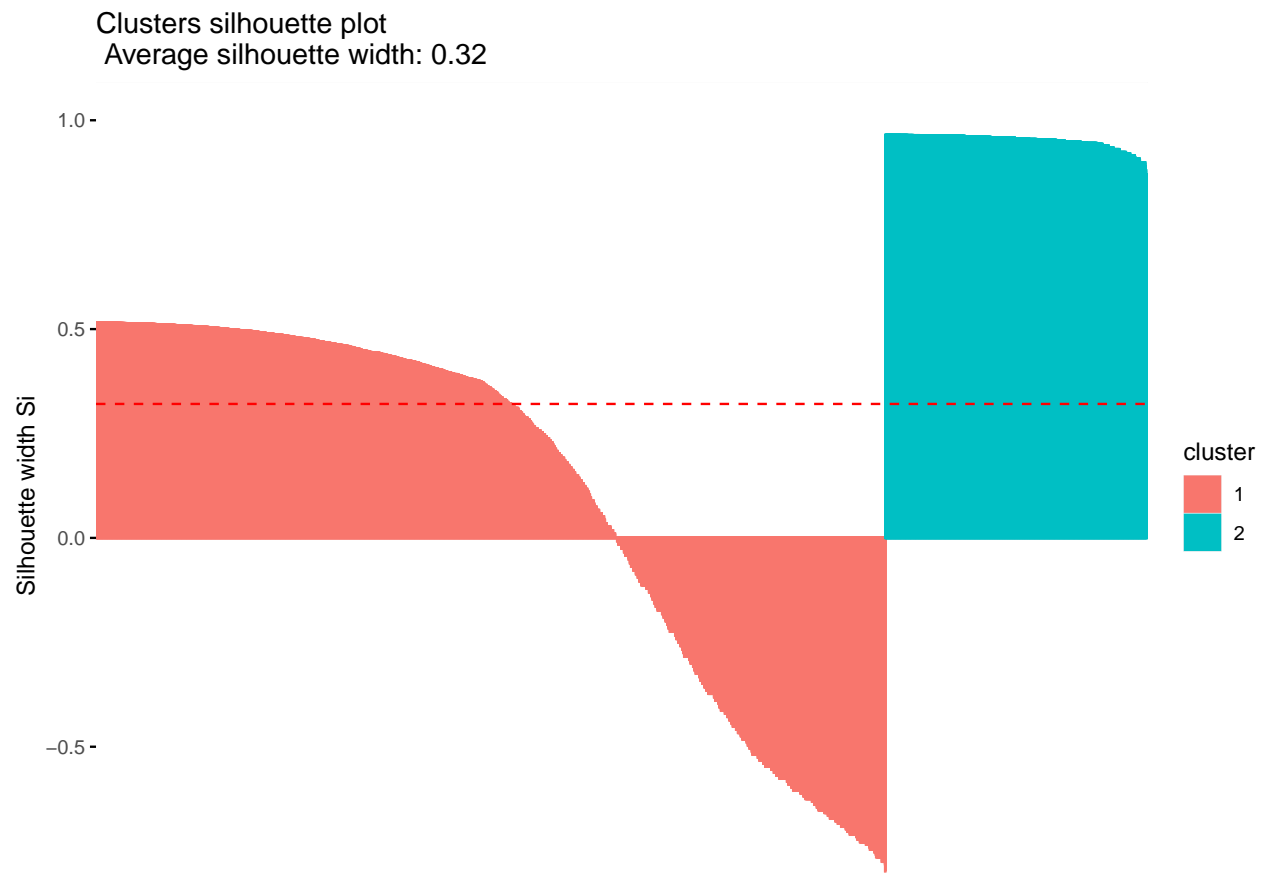


```
## [1] "Silhouette coefficient: 0.443449260743095"  
## [1] "Xie Beni coefficient: Inf"  
## [1] "Davies Bouldin coefficient: 1.14787619024023"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.01155  
##   cluster size ave.sil.width  
## 1       1 5366         0.11  
## 2       2 1779         0.95
```



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##   Cluster 1 Cluster 2
##      18398      6085
##
##
##
## DB Population:
##   Cluster 1 Cluster 2
##       71.3       72
##
##
##
## CSD Population:
##   Cluster 1 Cluster 2
##  235983.9  253622.2
##
##
##
## CMA Type:
##   Cluster 1 Cluster 2
##       8024      2603
## B       7703      2609
## D       2001       659
## K        670       214
```

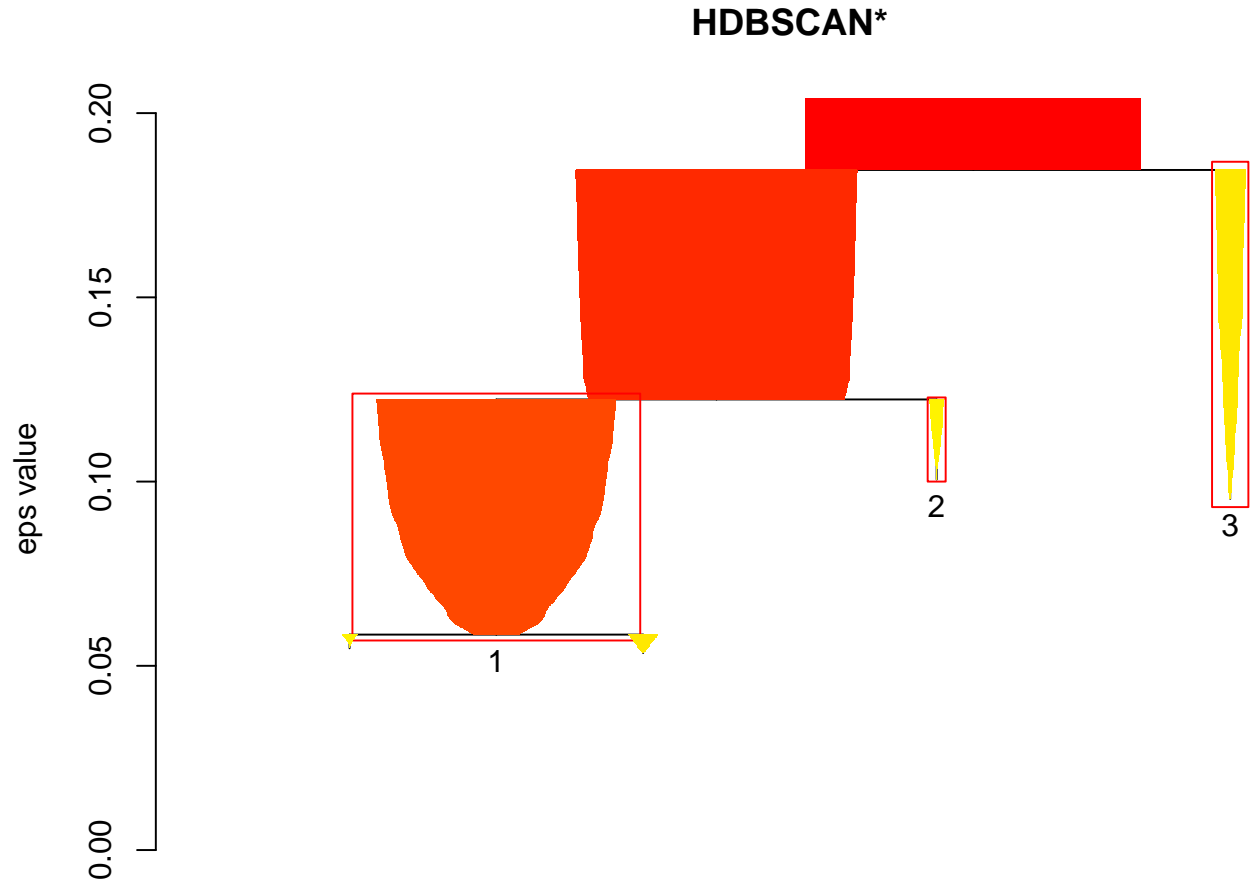


```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2
##      0.227      0.223
##
##
##
## Provinces:
##              Cluster 1 Cluster 2
## Alberta              570      193
## BritishColumbia      900      299
## NewBrunswick          154       53
## NorthwestTerritories   12        4
## NovaScotia            581      198
## Ontario              2878     980
## Quebec               1064     314
## Saskatchewan           98       23
## NA's                 12141    4021
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2
## 0      16596     5522
## 1       1370      445
## 2        208       59
## F        224       59
##
##
##
## PMS_prox_idx_pharma :
## Cluster 1 Cluster 2
##      0.04493    0.04319

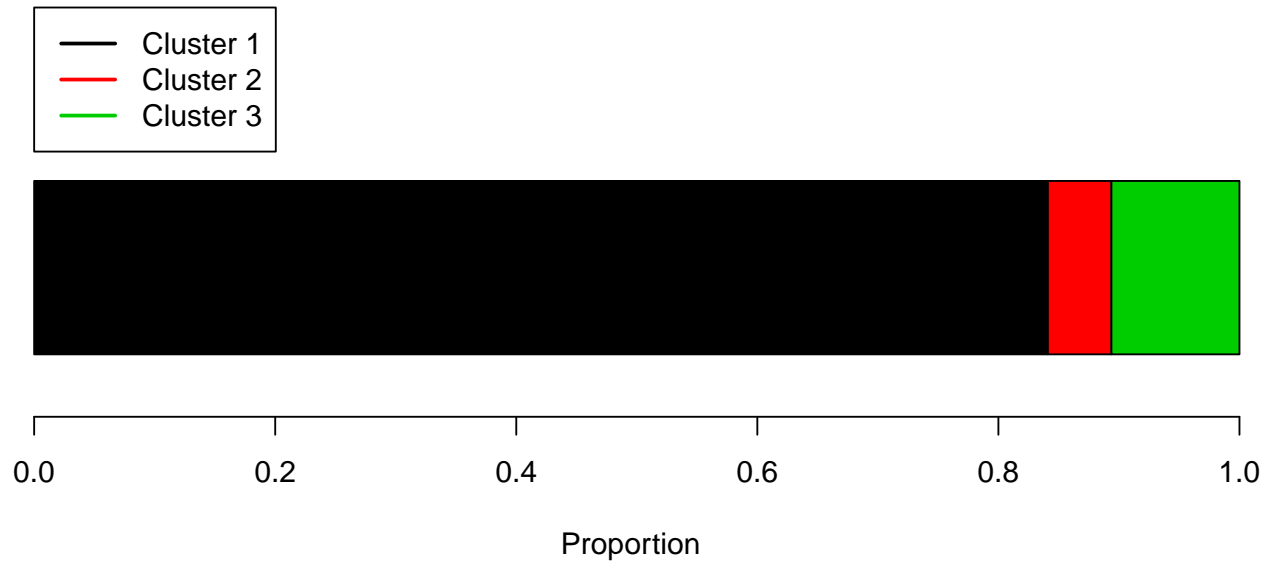
```

Childcare



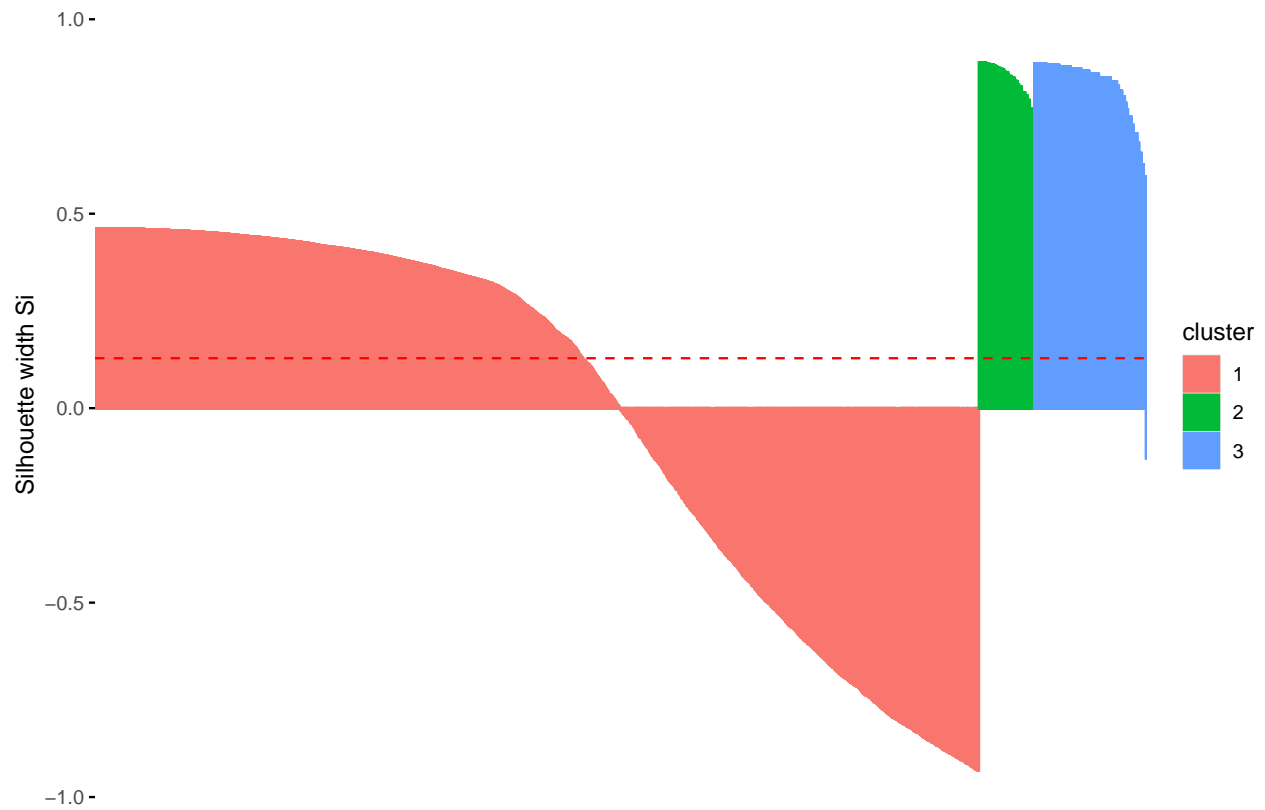
```
## [1] "Silhouette coefficient: 0.400812993352801"  
## [1] "Xie Beni coefficient: Inf"  
## [1] "Davies Bouldin coefficient: 6.14542021462162"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.01205  
## [1] 0.101  
##   cluster size ave.sil.width  
## 1      1 8677      -0.01  
## 2      2  544       0.85  
## 3      3 1096       0.83
```

Clusters silhouette plot Average silhouette width: 0.13



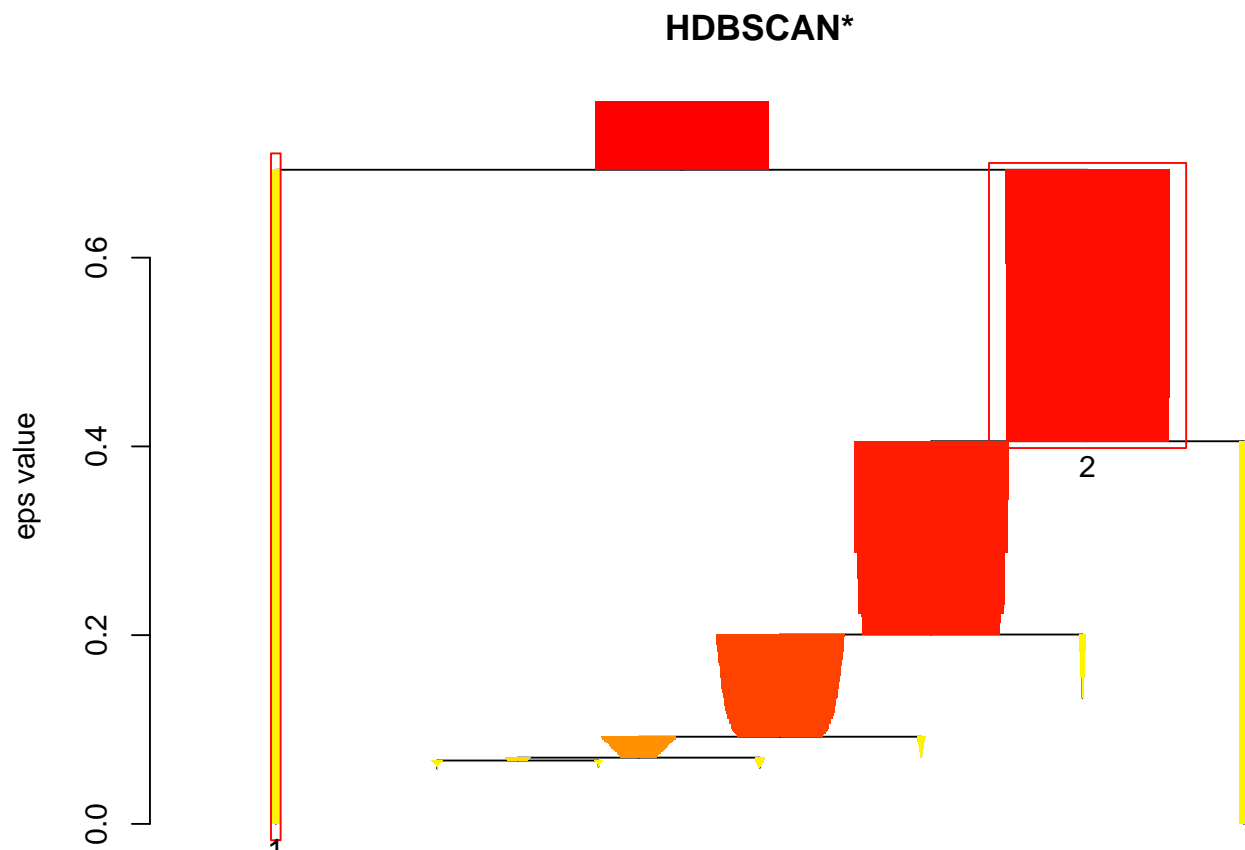
```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##   Cluster 1 Cluster 2 Cluster 3
##      20578      1314      2591
##
##
##
## DB Population:
##   Cluster 1 Cluster 2 Cluster 3
##       71.6      68.7      71.8
##
##
##
## CSD Population:
##   Cluster 1 Cluster 2 Cluster 3
##  238226.5  260338.3  247224.8
##
##
##
## CMA Type:
##   Cluster 1 Cluster 2 Cluster 3
##      8932      565      1130
## B      8645      564      1103
## D      2247      139       274
## K       754       46       84
```

```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2 Cluster 3
##      0.227      0.23      0.221
##
##
##
## Provinces:
##              Cluster 1 Cluster 2 Cluster 3
## Alberta              646          39          78
## BritishColumbia      1022          55         122
## NewBrunswick          181           9          17
## NorthwestTerritories   15           0           1
## NovaScotia            655          39          85
## Ontario              3237         202         419
## Quebec               1160          72         146
## Saskatchewan          103           6          12
## NA's                 13559         892        1711
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2 Cluster 3
## 0      18610      1181      2327
## 1       1513       100       202
## 2        220        14        33
## F        235        19        29
##
##
##
## PMS_prox_idx_childcare :
## Cluster 1 Cluster 2 Cluster 3
##      0.07613  0.08299  0.07877

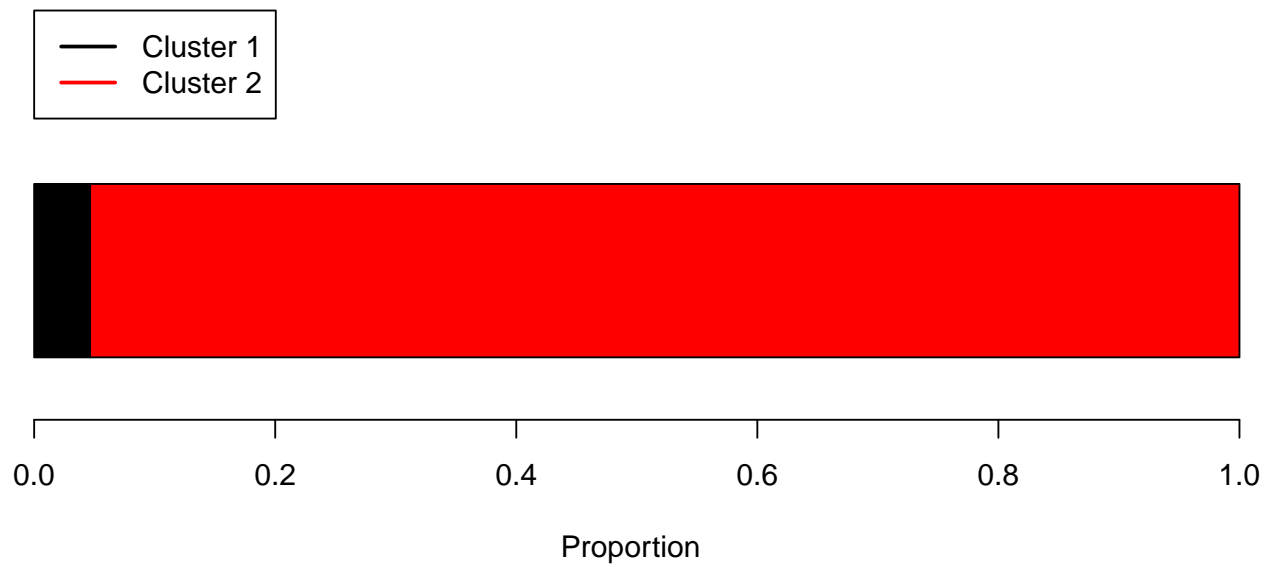
```

Health



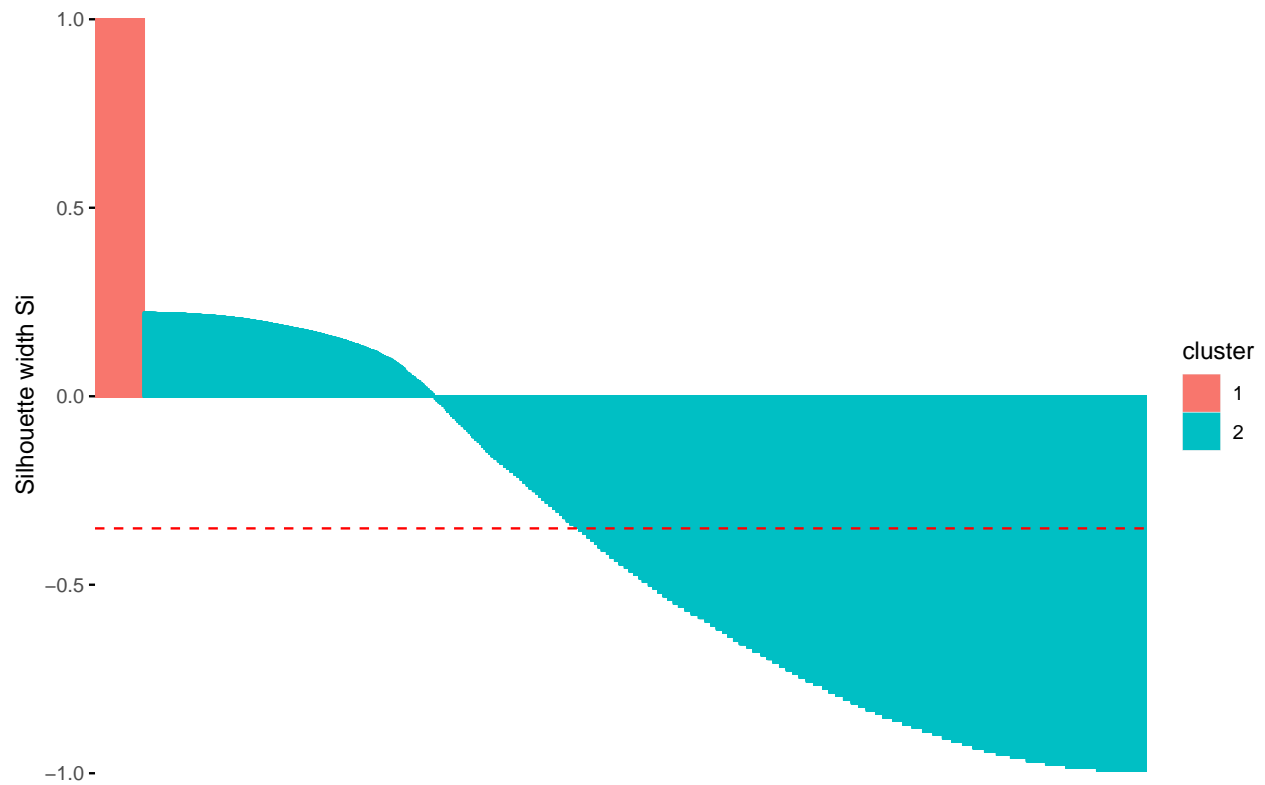
```
## [1] "Silhouette coefficient: 0.71150342405053"  
## [1] "Xie Beni coefficient: 119651.993317046"  
## [1] "Davies Bouldin coefficient: 0.380298291113713"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.06835  
##   cluster size ave.sil.width  
## 1         1   687         1.00  
## 2         2 14109        -0.42
```

Clusters silhouette plot
Average silhouette width: -0.35



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
## Cluster 1 Cluster 2
##      1142      23341
##
##
##
## DB Population:
## Cluster 1 Cluster 2
##      68.9      71.6
##
##
##
## CSD Population:
## Cluster 1 Cluster 2
## 217834.9 241467.9
##
##
##
## CMA Type:
## Cluster 1 Cluster 2
##      509      10118
## B      450      9862
## D      137      2523
## K       46       838
```



```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2
##      0.235      0.226
##
##
##
## Provinces:
##              Cluster 1 Cluster 2
## Alberta              32      731
## BritishColumbia      62     1137
## NewBrunswick          9      198
## NorthwestTerritories  1       15
## NovaScotia           28      751
## Ontario              169     3689
## Quebec               55     1323
## Saskatchewan          8      113
## NA's                 778    15384
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2
## 0      1046    21072
## 1        73    1742
## 2         8     259
## F       15     268
##
##
##
## PMS_prox_idx_health :
## Cluster 1 Cluster 2
##      0.01159   0.01382

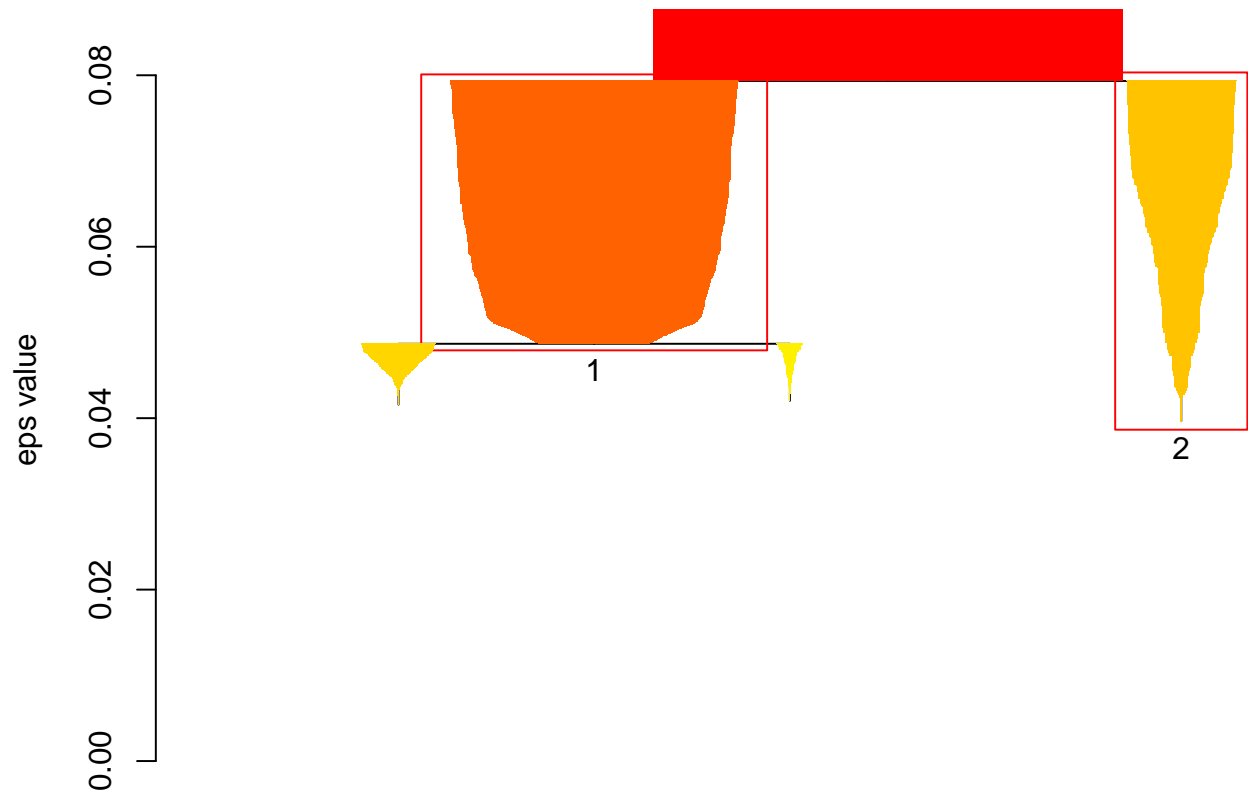
```

Grocery

```
## [1] "NO CLUSTERS DETECTED"
```

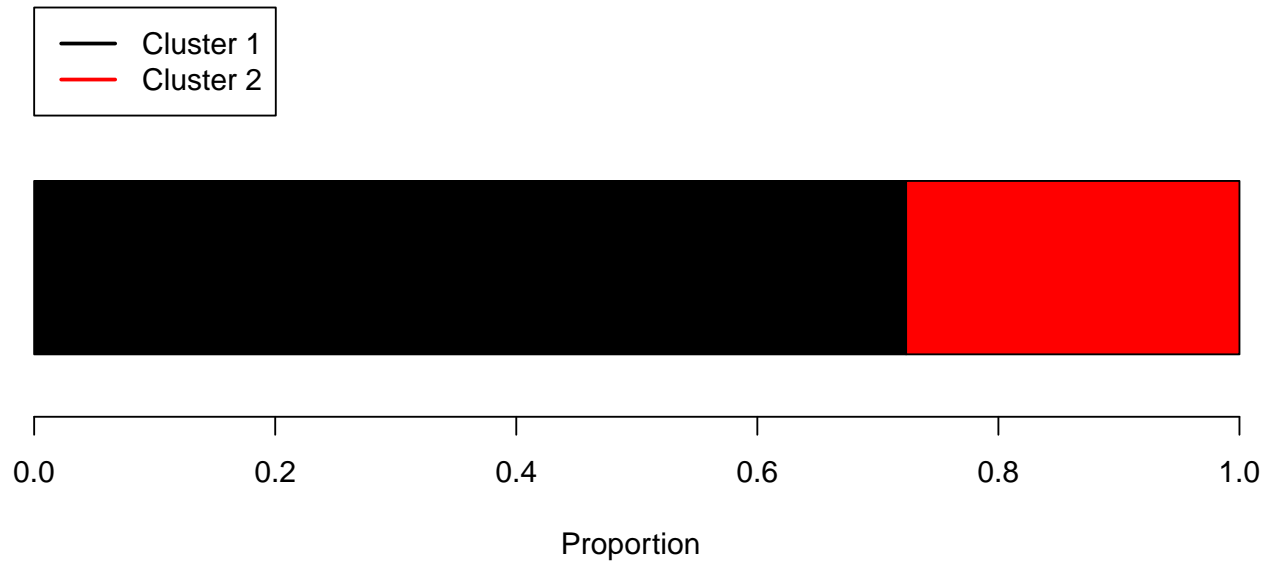
Primary Education

HDBSCAN*



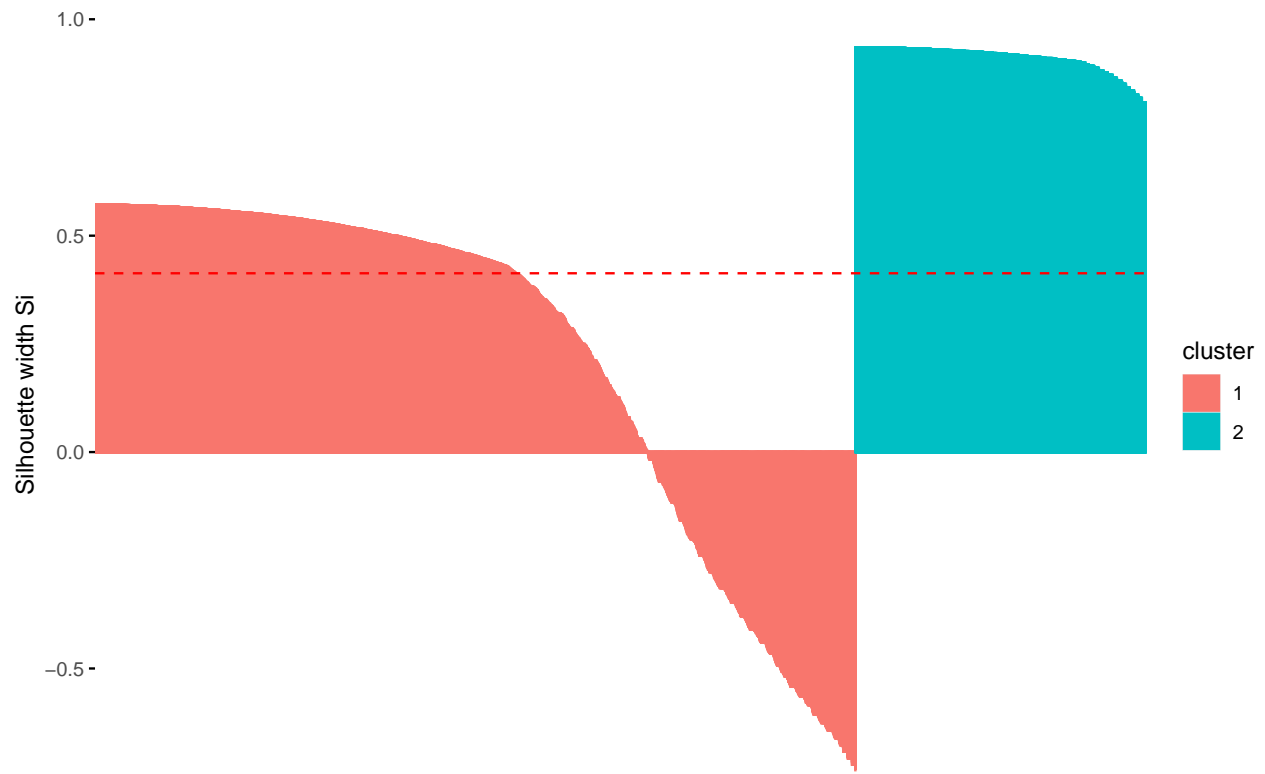
```
## [1] "Silhouette coefficient: 0.491369087890728"  
## [1] "Xie Beni coefficient: 855312.566930832"  
## [1] "Davies Bouldin coefficient: 0.886665004577807"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.04895  
##   cluster size ave.sil.width  
## 1       1 6896         0.22  
## 2       2 2636         0.91
```

Clusters silhouette plot
Average silhouette width: 0.41



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##   Cluster 1 Cluster 2
##      17684      6799
##
##
##
## DB Population:
##   Cluster 1 Cluster 2
##        71      72.6
##
##
##
## CSD Population:
##   Cluster 1 Cluster 2
##  238514.4  245177.7
##
##
##
## CMA Type:
##   Cluster 1 Cluster 2
##      7712      2915
## B      7465      2847
## D      1880       780
## K       627       257
```

```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2
##      0.226      0.227
##
##
##
## Provinces:
##              Cluster 1 Cluster 2
## Alberta              533      230
## BritishColumbia      866      333
## NewBrunswick          143       64
## NorthwestTerritories   12        4
## NovaScotia            582      197
## Ontario              2806     1052
## Quebec                981      397
## Saskatchewan           90       31
## NA's                 11671     4491
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2
## 0      15995      6123
## 1       1308       507
## 2        182        85
## F        199        84
##
##
##
## PMS_prox_idx_educpri :
## Cluster 1 Cluster 2
##      0.11698      0.11856

```

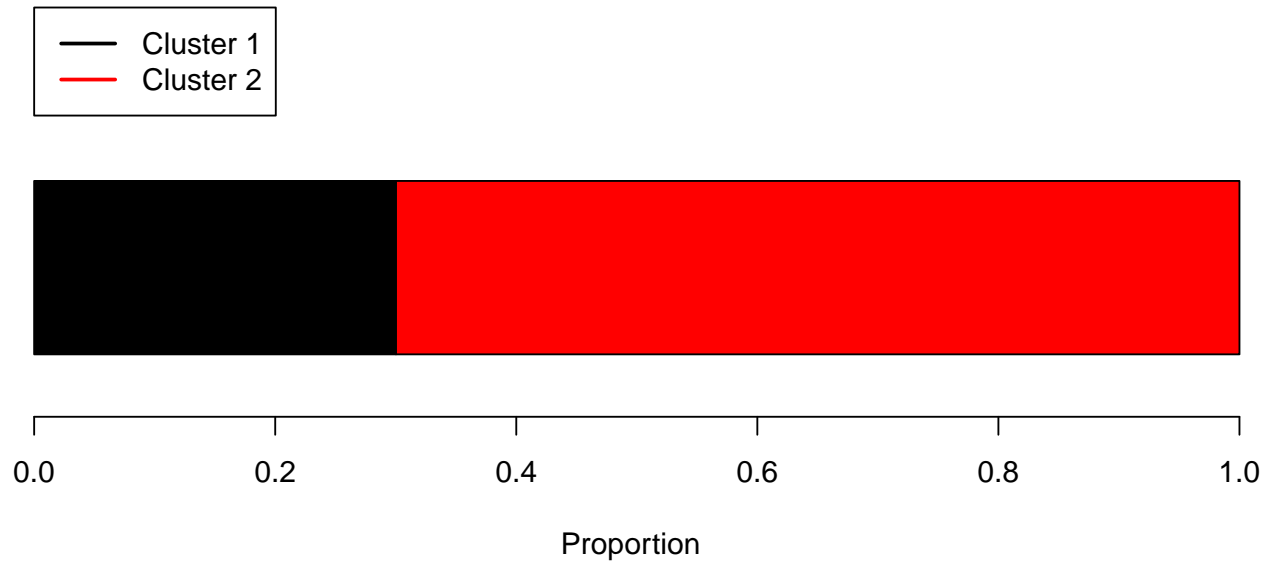
Secondary Education

HDBSCAN*



```
## [1] "Silhouette coefficient: 0.502911585364128"  
## [1] "Xie Beni coefficient: 121053.51347719"  
## [1] "Davies Bouldin coefficient: 0.688018349174813"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.0655  
##   cluster size ave.sil.width  
## 1      1 1255      0.75  
## 2      2 2925      0.76
```


Clusters silhouette plot
Average silhouette width: 0.76



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
## Cluster 1 Cluster 2
##      7345      17138
##
##
##
## DB Population:
## Cluster 1 Cluster 2
##      70.5      71.9
##
##
##
## CSD Population:
## Cluster 1 Cluster 2
## 245038.3 238361.7
##
##
##
## CMA Type:
## Cluster 1 Cluster 2
##      3210      7417
## B      3102      7210
## D       756      1904
## K       277       607
```

```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2
##      0.226      0.226
##
##
##
## Provinces:
##              Cluster 1 Cluster 2
## Alberta              224      539
## BritishColumbia      349      850
## NewBrunswick          58      149
## NorthwestTerritories    6       10
## NovaScotia            242      537
## Ontario              1150     2708
## Quebec                430      948
## Saskatchewan          53       68
## NA's                  4833     11329
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2
## 0      6581      15537
## 1       604      1211
## 2        81      186
## F       79       204
##
##
##
## PMS_prox_idx_educsec :
## Cluster 1 Cluster 2
##      0.10348    0.10371

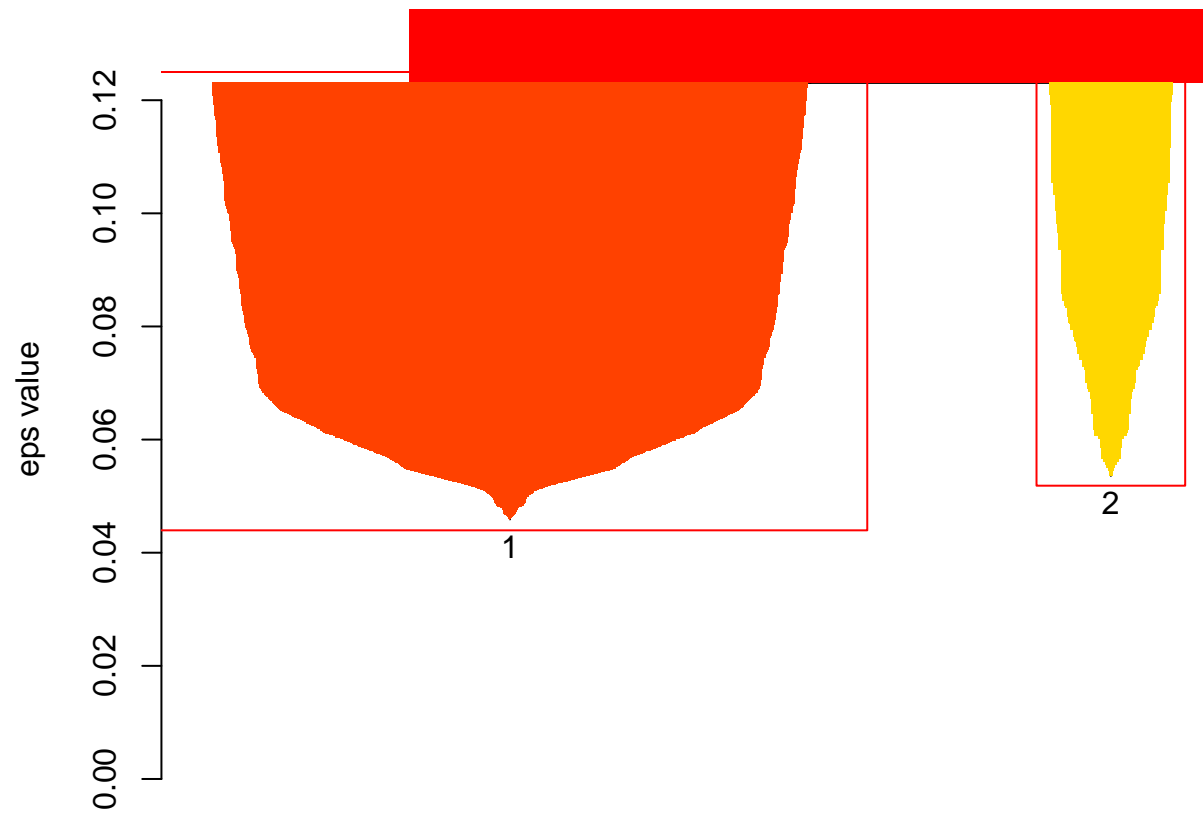
```

Libraries

```
## [1] "NO CLUSTERS DETECTED"
```

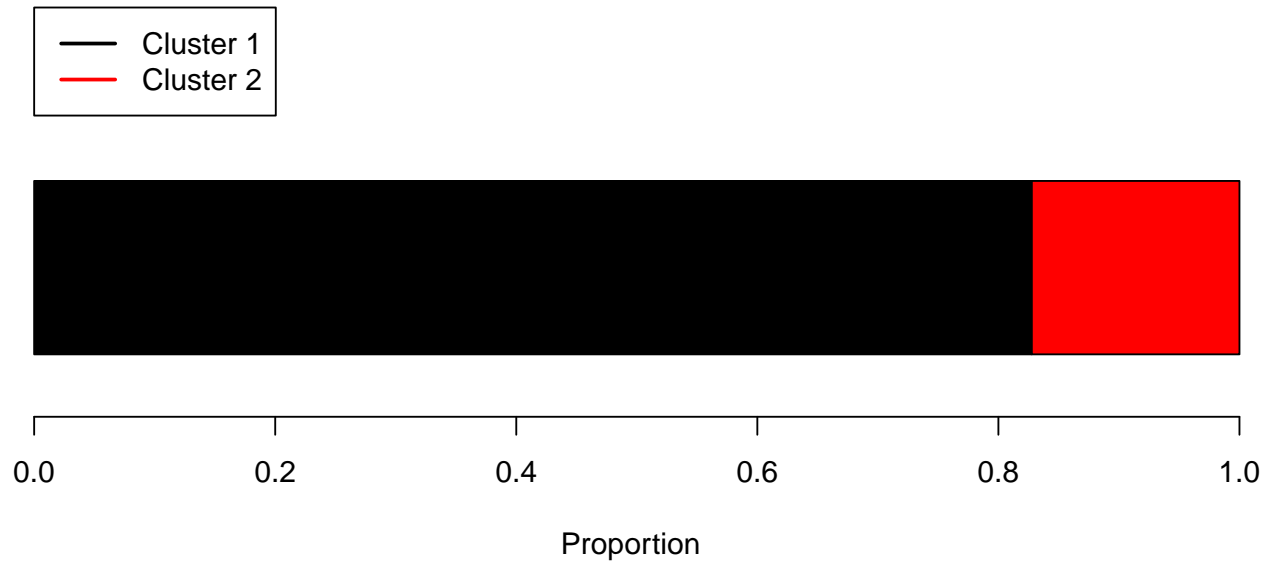
Parks

HDBSCAN*



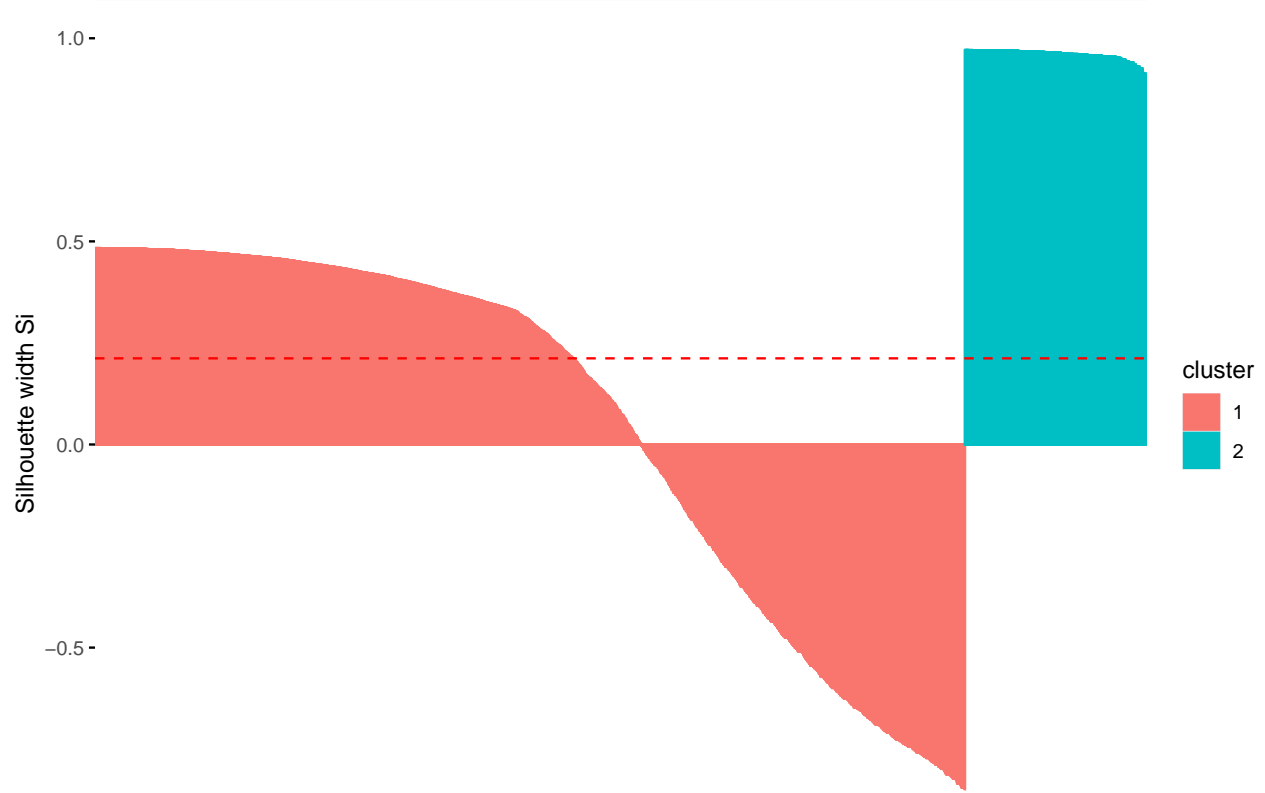
```
## [1] "Silhouette coefficient: 0.343965286339036"  
## [1] "Xie Beni coefficient: Inf"  
## [1] "Davies Bouldin coefficient: 4.9823162347747"
```

Proportion of DBs in each cluster



```
## [1] "Segment cutoff values:"  
## [1] 0.01935  
##   cluster size ave.sil.width  
## 1      1 8696      0.06  
## 2      2 1810      0.96
```

Clusters silhouette plot
Average silhouette width: 0.21



```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##   Cluster 1 Cluster 2
##      20277      4206
##
##
##
## DB Population:
##   Cluster 1 Cluster 2
##       71.3      72.1
##
##
##
## CSD Population:
##   Cluster 1 Cluster 2
##  241671.3  234067.8
##
##
##
## CMA Type:
##   Cluster 1 Cluster 2
##      8769      1858
## B      8533      1779
## D      2226       434
## K       749       135
```

```

##
##
##
## Index of Remoteness:
## Cluster 1 Cluster 2
##      0.226      0.228
##
##
##
## Provinces:
##              Cluster 1 Cluster 2
## Alberta              610      153
## BritishColumbia      993      206
## NewBrunswick          166       41
## NorthwestTerritories   13        3
## NovaScotia            639      140
## Ontario              3178      680
## Quebec                1153      225
## Saskatchewan           102       19
## NA's                  13423     2739
##
##
##
## Amenity dense:
## Cluster 1 Cluster 2
## 0      18312      3806
## 1       1511       304
## 2        214        53
## F        240        43
##
##
##
## PMS_prox_idx_parks :
## Cluster 1 Cluster 2
##      0.06836      0.06987

```

Transit

```
## [1] "NO CLUSTERS DETECTED"
```

Conclusion

text

```
## list(PMS_prox_idx_emp = 0.16035, PMS_prox_idx_pharma = 0.01155,  
##      PMS_prox_idx_childcare = c(0.01205, 0.101), PMS_prox_idx_health = 0.06835,  
##      NA, PMS_prox_idx_educpri = 0.04895, PMS_prox_idx_educsec = 0.0655,  
##      NA, PMS_prox_idx_parks = 0.01935, NA)
```