Cluster Analysis: Identifying Parkinson's Disease Subtypes

Jesse Mu

Wednesday, June 10

1 Preprocessing

1.1 Dataset Description

951 subjects, 145 metrics, collected 15-4-2012. From Pablo Martinez Martín. 170 subjects with missing values (brought down to 781); these were removed automatically, even if the missing values were not included in the selected features below. This will need to be changed later on, by keeping those removed that still have all selected features and perhaps with some compensation for missing values.

1.2 Selected Features

Combination of non-motor scale (NMS) symptoms and standard motor symptoms.

Name	Type	Format	Description
nms_d1	byte	%8.0g	cardiovascular
nms_d2	byte	%8.0g	sleep/fatigue
nms_d3	byte	%8.0g	mood/cognition
nms_d4	byte	%8.0g	percep/hallucinations
nms_d5	byte	%8.0g	attention/memory
nms_d6	byte	%8.0g	gastrointestinal
nms_d7	byte	%8.0g	urinary
nms_d8	byte	%8.0g	sexual function
nms_d9	byte	%8.0g	miscellaneous
tremor	float	%9.0g	tremor
bradykin	float	%9.0g	bradykinesia[1]
rigidity	float	%9.0g	rigidity
axial	float	%9.0g	axial[2]
pigd	float	%9.0g	postural instability and gait difficulty

Table 1: Selected Features and Details

Name	μ	σ	min-max
nms_d1	1.76	3.32	0-24
nms_d2	8.71	8.76	0-48
nms_d3	8.70	11.83	0-60
nms_d4	1.65	3.94	0-33
nms_d5	5.22	7.44	0-36
nms_d6	5.67	6.92	0-36
nms_d7	8.02	9.09	0-36
nms_d8	3.57	5.97	0-24
nms_d9	6.99	7.74	0-48
tremor	2.59	2.63	0-12
bradykin	2.49	1.39	0-6
rigidity	2.34	1.36	0-6
axial	3.28	2.75	0-12
pigd	3.36	2.77	0-12

Table 2: Descriptive Statistics

1.3 Dimensionality Reduction: PCA

May not be useful? If we're trying to identify *clinically* relevant features, merging them may not be a good idea.

Figure 1 shows scree test elbow occurs around 2 or 3. Also, eigenvalues 1 and 2 > 1, while 3 is around .9

2 k-means

2.1 Identifying optimal number of clusters

2.1.1 WSS Error Scree Test

Figure 2 shows no optimal elbow in scree test! Maybe 2-3?

2.1.2 Gap Statistic

Optimal cluster is the local maximum of the gap statistic, but it appears to be consistently increasing in Figure 3.

2.1.3 Average Silhouette Width

Figure 4 shows average silhouette width as being consistently under 0.25 for all clusters, implying the data is not well structured.

[1] Impaired ability to adjust the body's position.

[2] Issues affecting the middle of the body.

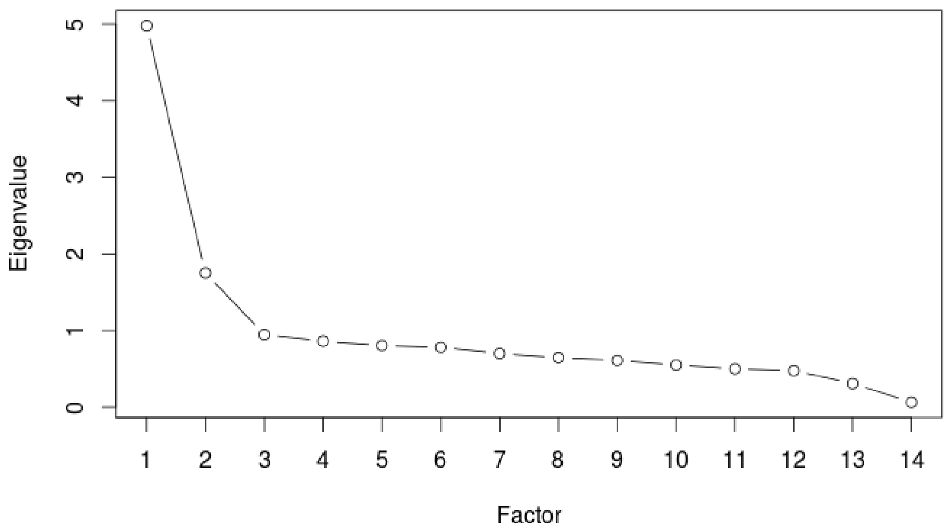

Figure 1: Scree test: eigenvalues by factor

2.2 Cluster statistics

k	n	Within SS	sum(Within SS)
2	201/580	4248.585/4132.434	8381.019
3	420/231/130	2618.368/1973.82/3076.542	7668.73
4	61/372/145/203	1481.25/1845.389/2147.988/1609.555	7084.183

Table 3: Cluster statistics

2.3 Centers

Omitted; too much information.

2.4 Decision tree classifier based on clusters

k	CP[3]	CV Xerror[4]	Root Feature	Root Error	Figure
2	0.0348	0.134	axial ≥ 0.44	0.257	Figure 5
3	0.0100	0.194	bradykin < 0.0041	0.462	Figure 6
4	0.0100	0.248	bradykin < 0.0041	0.523	Figure 7

Table 4: k-kmeans decision trees statistics

[3]Complexity Parameter

[4]10-fold cross validation

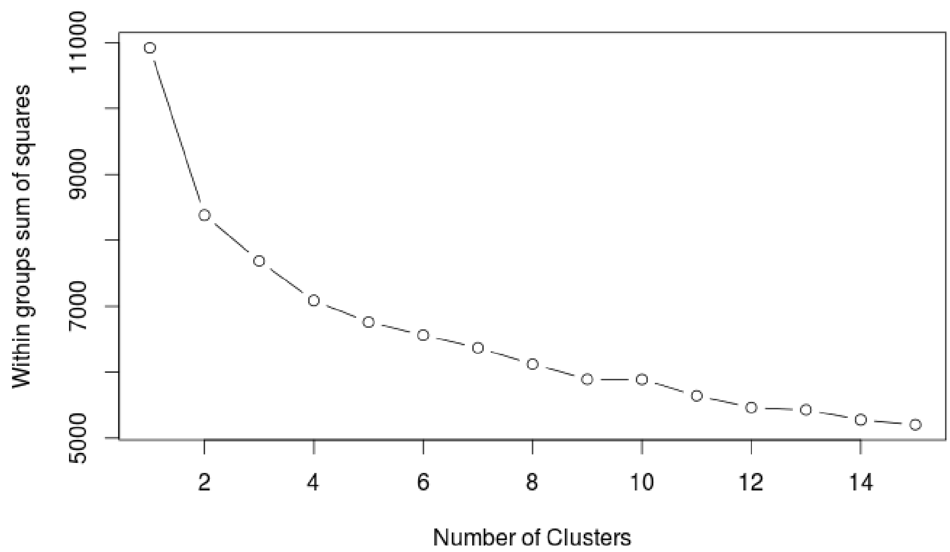

Figure 2: Scree test: WSS error by cluster size

2.5 Summary Statistics based on Clusters

```
k = 2, cluster 1
var,mean,sd,min,max
age,68.38,9.4,42,89
sex,0.43,0.5,0,1
pdonset,57.65,10.81,32,87
durat_pd,10.73,7.14,0,40
cisitot,13.07,4.55,2,24
k = 2, cluster 2
var,mean,sd,min,max
age,63.43,9.55,37,89
sex,0.36,0.48,0,1
pdonset,56.44,10.62,28,89
durat_pd,6.99,4.94,0,28
cisitot,6.64,3.42,0,16

k = 3, cluster 1
var,mean,sd,min,max
age,62.69,9.43,37,89
sex,0.39,0.49,0,1
pdonset,56.04,10.41,28,89
durat_pd,6.65,4.66,0,26
cisitot,5.76,3.15,0,15
```

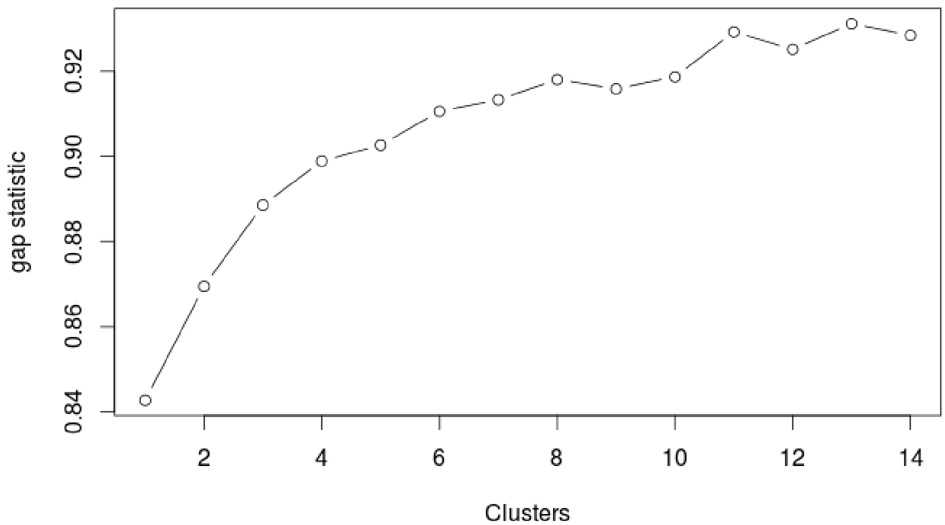

Figure 3: Gap statistic by cluster size

```
k = 3, cluster 2
var,mean,sd,min,max
age,66.17,9.74,38,86
sex,0.32,0.47,0,1
pdonset,57.78,11.19,32,84
durat_pd,8.39,5.79,0,31
cisitot,9.95,3.5,2,22
k = 3, cluster 3
var,mean,sd,min,max
age,68.62,9.2,50,89
sex,0.46,0.5,0,1
pdonset,57.24,10.46,35,87
durat_pd,11.38,7.52,0,40
cisitot,13.52,5.01,2,24

k = 4, cluster 1
var,mean,sd,min,max
age,71.9,8.1,54,89
sex,0.46,0.5,0,1
pdonset,58.28,10.27,35,87
durat_pd,13.62,7.76,0,40
cisitot,16.72,4.14,4,24
k = 4, cluster 2
var,mean,sd,min,max
```

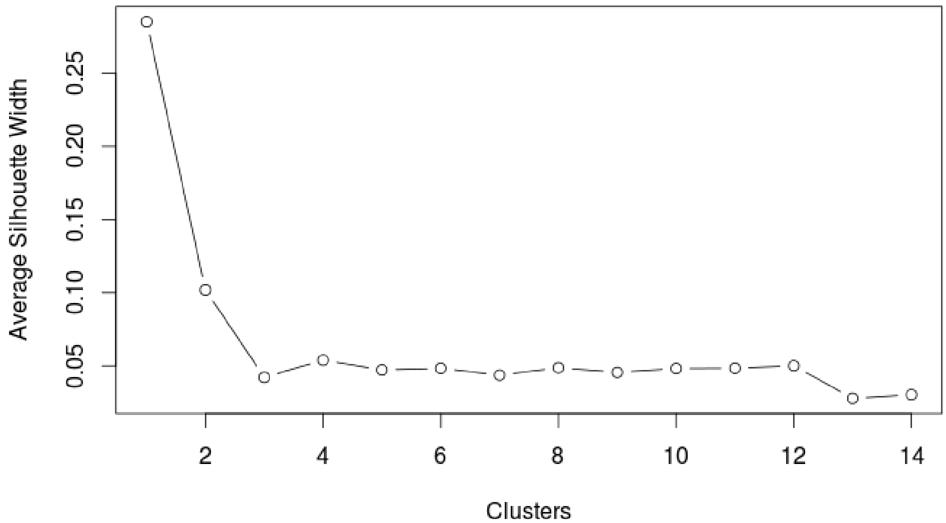

Figure 4: Average silhouette width by cluster size

```
age,62.65,9.63,37,89
sex,0.36,0.48,0,1
pdonset,56.13,10.47,28,89
durat_pd,6.52,4.66,0,25
cisitot,5.58,3.12,0,15
k = 4, cluster 3
var,mean,sd,min,max
age,64.79,9.3,44,86
sex,0.47,0.5,0,1
pdonset,56.2,10.71,35,85
durat_pd,8.59,6.16,0,27
cisitot,9.23,3.88,2,19
k = 4, cluster 4
var,mean,sd,min,max
age,66.27,9.47,40,86
sex,0.33,0.47,0,1
pdonset,57.84,11.06,32,84
durat_pd,8.43,5.64,0,31
cisitot,10.06,3.49,3,22
```

3 Biclustering

Used BCBimax clustering algorithm. Clusters seem quite sparse.

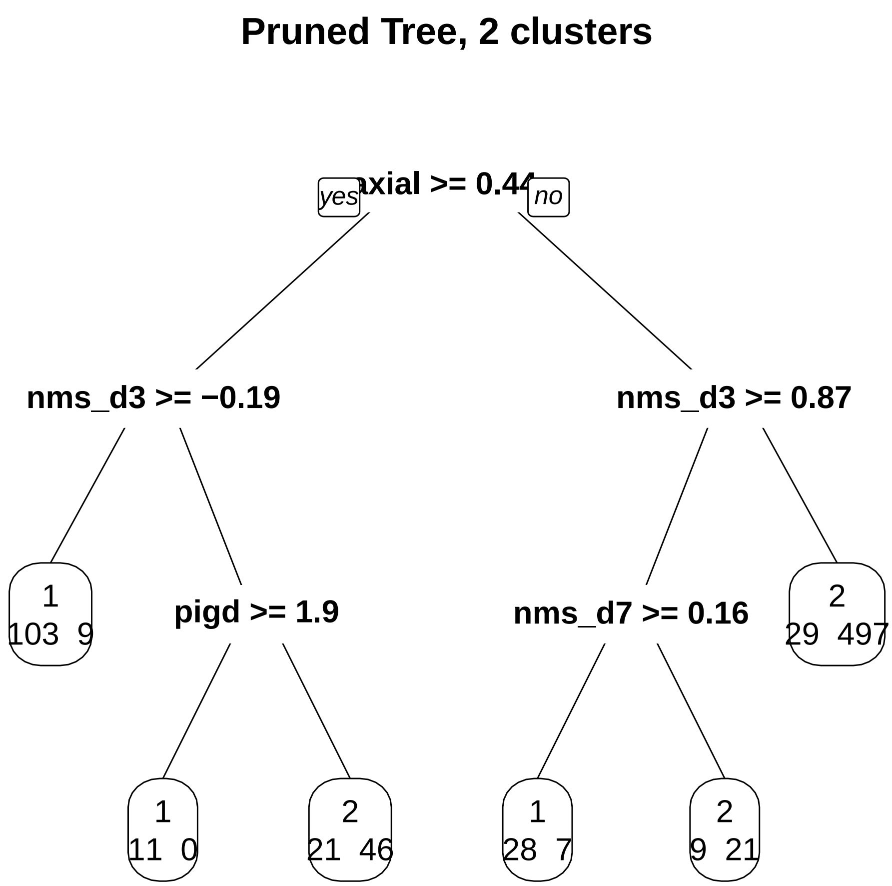

Figure 5: Decision Tree from k-means clustering, 2 clusters

4 Subspace clustering

5 Bayesian Networks

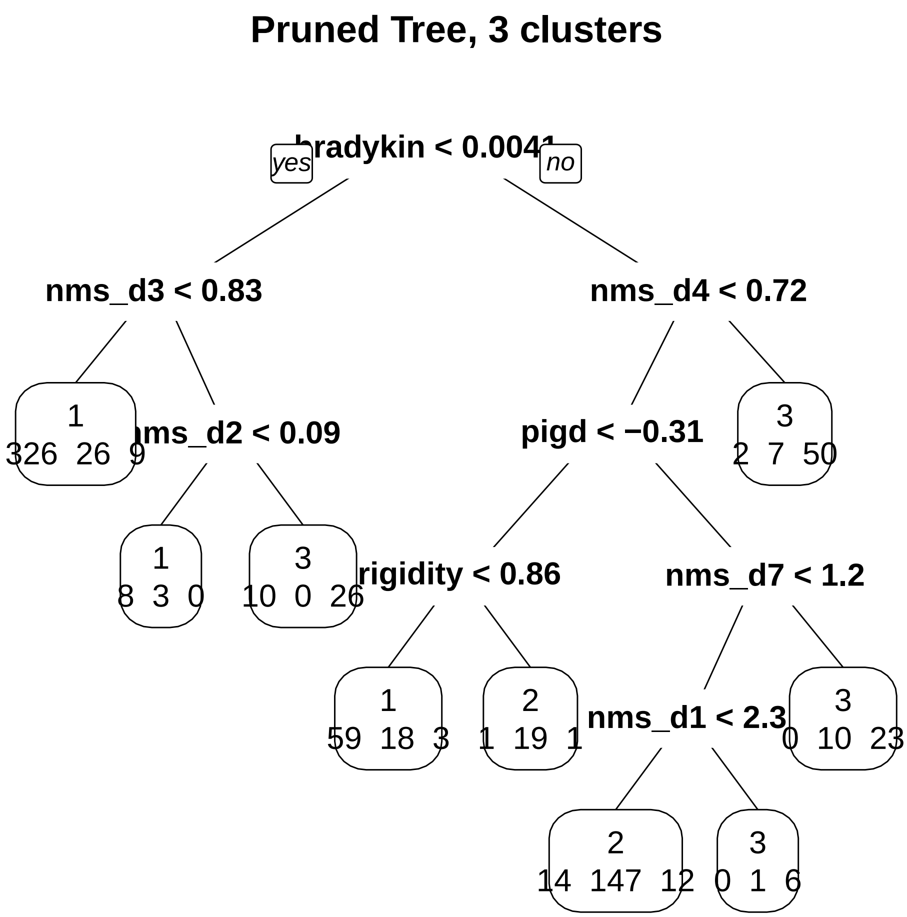

Figure 6: Decision Tree from k-means clustering, 3 clusters

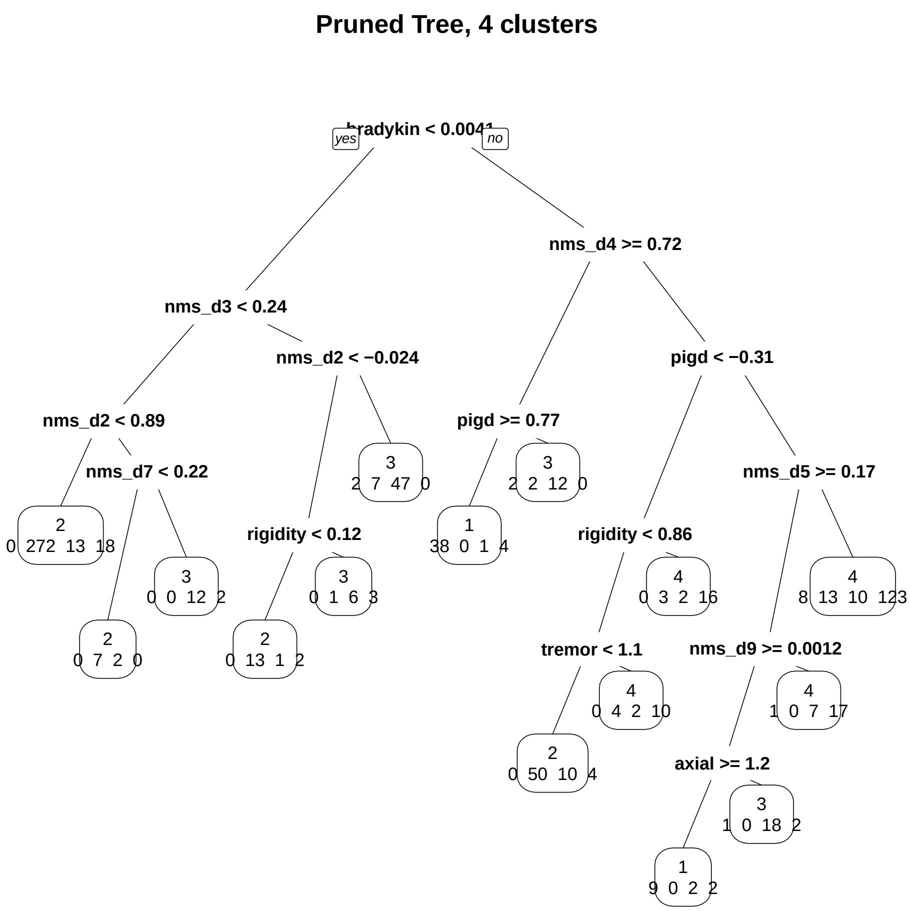

Figure 7: Decision Tree from k-means clustering, 4 clusters

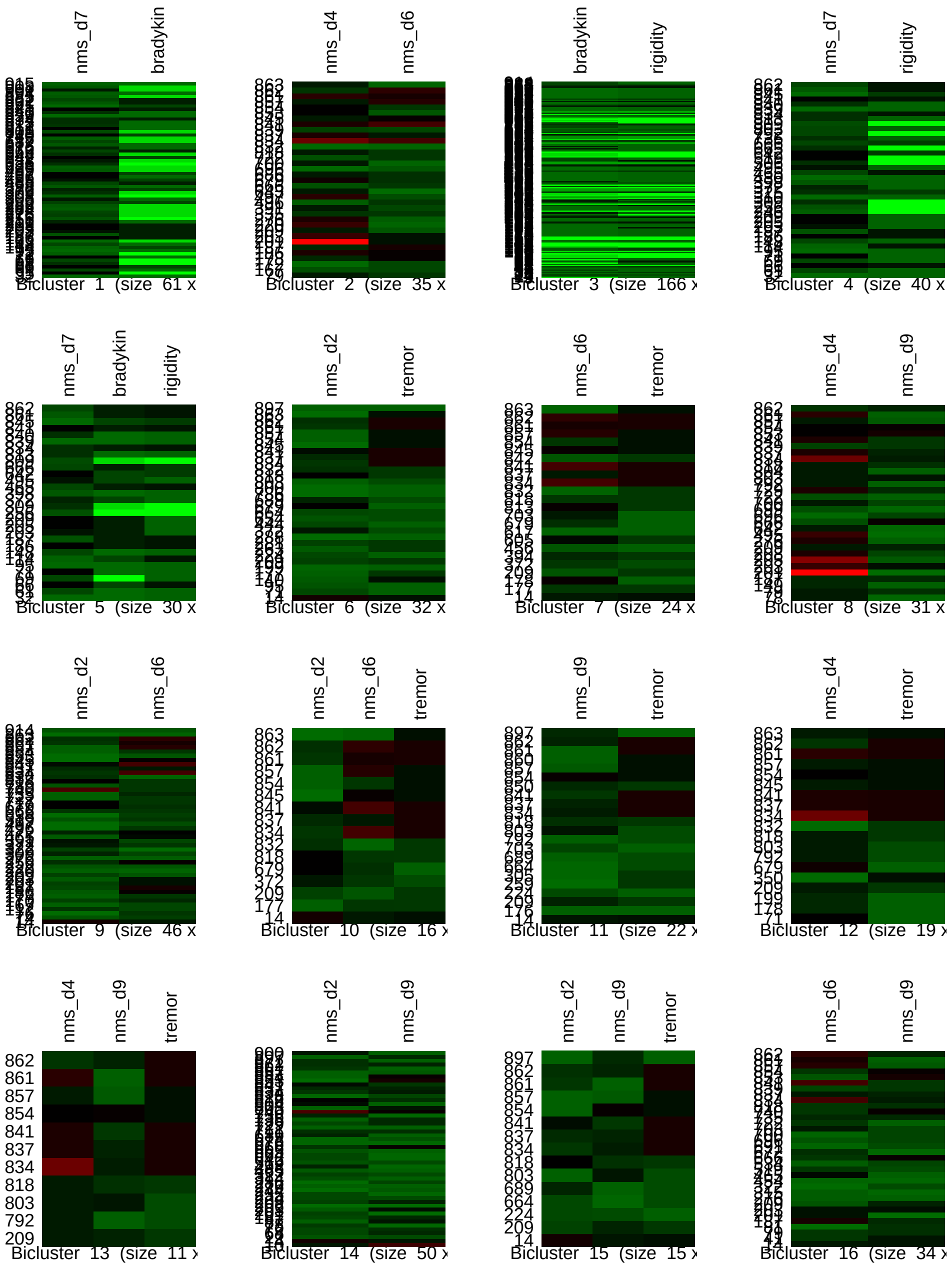

Figure 8: Biclustering $N = 16$

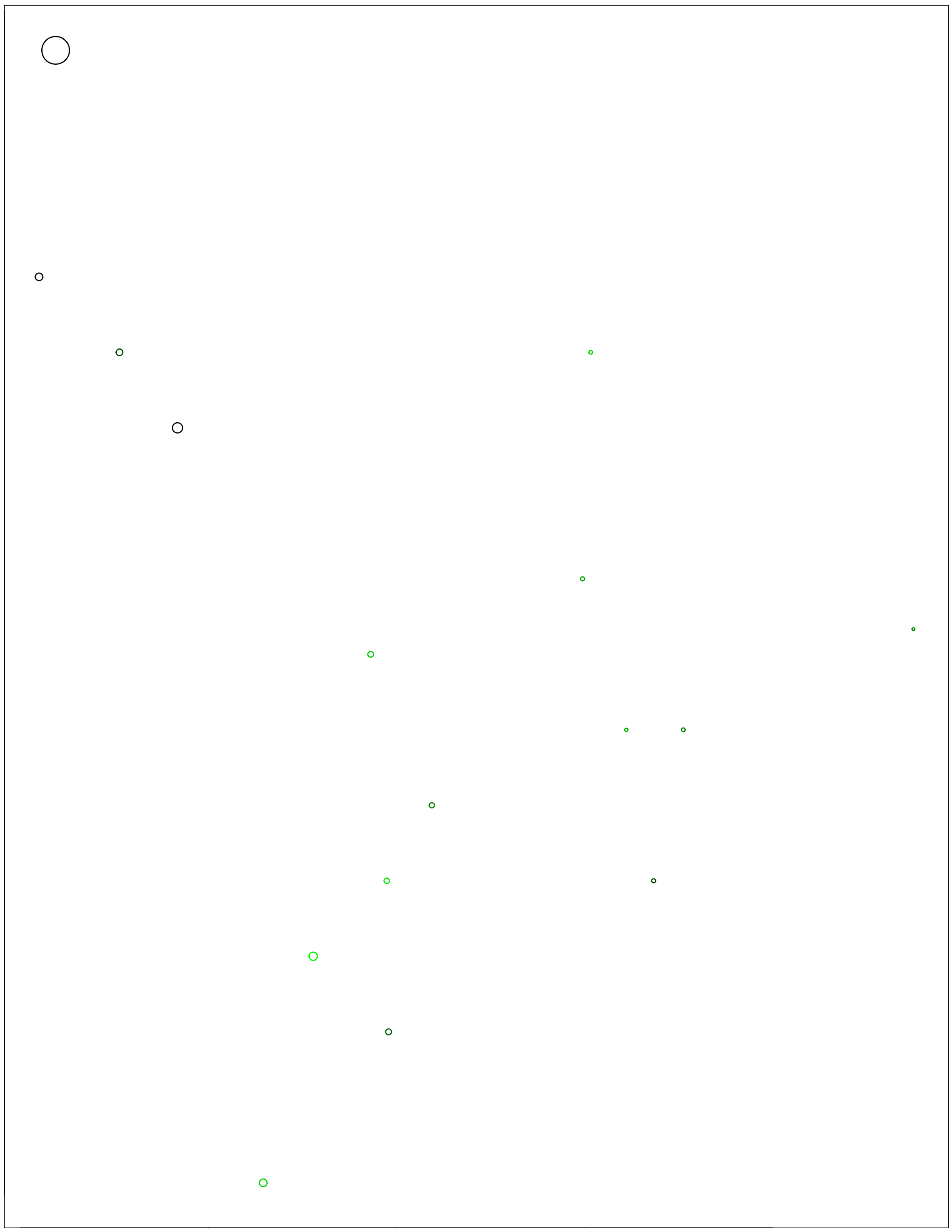

Figure 9: Bubbleplot $N = 16$