

# Cluster Analysis: Identifying Parkinson's Disease Subtypes

Jesse Mu

October 22, 2015

## 1 Preprocessing

### 1.1 Dataset Description

951 subjects, 145 metrics, collected 15-4-2012 from Pablo Martinez Martín. Only 19 features used for clustering and/or interpretation. 50 subjects with missing values of the features to be used in clustering (brought down to 901). It was decided to not impute the data.

### 1.2 Selected Features

Combination of non-motor scale (NMS) symptoms and standard motor symptoms. PIGD was deleted after 2015-07-16 meeting.

Name	Type	Description
nms_d1	byte	cardiovascular
nms_d2	byte	sleep/fatigue
nms_d3	byte	mood/cognition
nms_d4	byte	percep/hallucinations
nms_d5	byte	attention/memory
nms_d6	byte	gastrointestinal
nms_d7	byte	urinary
nms_d8	byte	sexual function
nms_d9	byte	miscellaneous
tremor	float	tremor
bradykin	float	bradykinesia <sup>1</sup>
rigidity	float	rigidity
axial	float	axial <sup>2</sup>

Table 1: Selected Features and Details

---

<sup>1</sup>Impaired ability to adjust the body's position.

<sup>2</sup>Issues affecting the middle of the body.

Name	$\mu$	$\sigma$	min-max
nms_d1	1.73	3.35	0-24
nms_d2	8.75	8.70	0-48
nms_d3	8.68	11.55	0-60
nms_d4	1.64	3.86	0-33
nms_d5	5.42	7.43	0-36
nms_d6	5.53	6.79	0-36
nms_d7	8.08	8.94	0-36
nms_d8	3.52	5.97	0-24
nms_d9	7.13	7.79	0-48
tremor	2.59	2.58	0-12
bradykin	2.40	1.41	0-6
rigidity	2.24	1.36	0-6
axial	3.25	2.68	0-12

Table 2: Descriptive Statistics

## 2 Clustering

$k$ -means clustering with  $k = 4$  was tried. Statistics for determining the optimal number of clusters were used, but were inconclusive: results in Figure 3. This probably indicates that the data is not very well clustered.  $k = 2, 3$  provided models that were too simplistic.  $k = 5$  did not provide any new information, but rather just fragmented existing groups.

Criterion	Optimal $k$
Minimum ASW	2
BIC	18
SSE Scree Plot	Inconclusive
Gap Statistic	4, 13?
Affinity Propagation <sup>3</sup>	8

Table 3: Results of various techniques for determining  $k$

### 2.1 Decision tree

Decision tree for  $k = 4$  created via recursive partitioning is available in Figure 1. More discussion about the decision tree is located in Section 2.2.4.

### 2.2 Interpretation of Clusters

#### 2.2.1 Cluster summaries

Available in Figure 2. Error bar is standard error.

---

<sup>3</sup> $\lambda = 0.98$ ,  $q = 0$ ,  $\text{maxits} = 1000$ ,  $\text{convits} = 100$

### 2.2.2 Interpretation

$k$ -means clustering ( $k = 4$ ) found four clusters. With a brief description, they are:

1. ( $n = 406$ ) Mildly affected in all domains.
2. ( $n = 189$ ) Severely affected in nonmotor domains; mildly affected in motor domains.
3. ( $n = 221$ ) Severely affected in motor domains; mildly affected in nonmotor domains.
4. ( $n = 88$ ) Severely affected in all domains.

These cluster results are identical to the four clusters found in van Rooden et al [2], which was done with a separate dataset using a different modeling method (expectation-maximization), supporting these subtype classifications. Unlike van Rooden, mean disease durations differences do exist between subtypes 1 (mild) and 4 (severe), likely due to further development of the disease, although the differences between 2 and 3 (nonmotor/motor predominated) subtypes are insignificant, suggesting different developmental paths of the disease.

### 2.2.3 Statistical Significance Tests, $k = 4$

For each variable  $i$  and cluster means  $\mu_i^1, \mu_i^2, \mu_i^3, \mu_i^4$ , we use one-way ANOVA for multiple means and reject the null hypothesis that  $\mu_i^1 = \mu_i^2 = \mu_i^3 = \mu_i^4$  with  $p < 0.05$  for every variable except pdonset.

Post-hoc analysis using Tukey’s HSD to examine statistically significant differences between individual means is available in Table 4. For brevity, only statistically insignificant relations are provided; all other relations are significant with  $p < 0.05$ .

### 2.2.4 Feature importance

Features ranked by information gain with respect to cluster are available in Table 5. Also, in the 4-cluster decision tree in Figure 1, features are ranked implicitly by importance in determining clusters. We see, quite naturally, that standard measures of motor symptoms rank very highly (1, 2, 4, 5) in information gain *except* tremor (12). Similarly, bradykinesia (1) is used as the root node of the 4-cluster decision tree, although other motor symptoms are used further down the tree, since immediately successive motor symptom decision nodes would, due to their determination of clusters, be redundant.

The most informative nonmotor symptoms are nms\_d2 (sleep/fatigue) at 2, along with nms\_d3 (mood/cognition). As discussed later in Section 4.1 these features become critical in one-versus-all decision trees for distinguishing various subtypes. The importance of these nonmotor symptoms confirms the longitudinal study by Fereshtehnejad et al. [1] who cites a 3-cluster PD subtype identification based primarily on nonmotor symptoms including cognitive impairment, rapid eye movement sleep disorder (RBD), anxiety, and depression, conditions that align closely with nms\_d2 and nms\_d3 as tested in this dataset. More analysis needs to be done on whether there are parallels between Fereshtehnejad’s 3-cluster longitudinal study and the clusters found in both this investigation and van Rooden.

Variable	Cluster Relation	$p$
age	3-1	0.724
	4-1	0.428
sex	2-1	0.849
	4-1	0.092
	3-2	0.161
	4-2	0.827
	4-3	0.216
pdonset	2-1	0.147
	3-1	0.370
	4-1	0.859
	3-2	0.803
	4-2	0.305
	4-3	0.700
durat_pd	3-1	0.562
cisitot	3-1	0.523
nms_d1	4-3	0.333
nms_d4	4-3	0.557
nms_d5	4-3	0.856
nms_d8	4-3	0.122
nms_d9	2-1	0.730
	4-3	0.074
tremor	4-1	0.360

Table 4: Tukey’s HSD Insignificant Differences

Interestingly, demographic information, including durat\_pd, age, sex, and pdonset, plays almost no role in the determination of these clusters. That the time of onset of PD or sex provides an important negative answer to clinically-relevant questions about the demographic sources of these different subtypes.

### 2.2.5 Correlation Plots

The interplay between specific symptoms in each of the four clusters was examined in Figure 5. There are two points of note. The first is that there is a higher correlation in cluster 4 (severe) between overall severity (cisitot) and bradykinesia and rigidity, illustrated in Figure 3. Second, there is a somewhat higher correlation between bradykinesia, rigidity, and nms\_d6 (gastrointestinal). There is, perhaps somewhat interestingly, a somewhat higher correlation between bradykinesia, rigidity, and nms\_d6 (gastrointestinal) in cluster 4, illustrated in Figure 4. These differences are statistically significant; correlation tests are located in Table 6.

rank	variable	information gain
1	bradykin	0.31574672
2	rigidity	0.29560018
3	nms_d2	0.24218407
4	cisitot	0.22920103
5	axial	0.22780750
6	nms_d3	0.20480570
7	nms_d9	0.15782743
8	nms_d7	0.15290569
9	nms_d5	0.14454931
10	nms_d6	0.14025139
11	nms_d1	0.13212756
12	tremor	0.10937168
13	nms_d4	0.10710526
14	nms_d8	0.10005480
15	durat_pd	0.02876190
16	age	0.02346158
17	sex	0.00000000
18	pdonset	0.00000000

Table 5: Features ranked by information gain

### 3 Nonmotor-predominant subtype analysis

#### 3.1 $k$ -means sub-subdivision on Cluster 2

In an attempt to understand further the properties of the nonmotor-dominated subtypes,  $k$ -means analysis was run again on specifically this subtype to examine any possible patterns.

The same  $k$ -determining tests were run on subtype 2 and are displayed in Table 7.

Boxplots for  $k$ -means run for  $k = 2, 3, 4$  can be seen in Figures ??, ??, and ?. Clusters are ordered by increasing cisitot.

#### 3.2 Interpretation

An interesting set of subtleties occurs when  $k = 2$  and 3. When  $k = 2$ , the two groups are divided by PD severity (see cisitot and especially axial). The specific symptoms of the two groups follow this trend, except nms\_d3 and tremor, which are actually decreasing, and other symptoms like rigidity, nms\_d4, and nms\_d9, which are more indeterminate.

When  $k = 3$ , the symptoms that continue show a non-monotonically increasing trend are nms\_d2, tremor, and rigidity scores, where patients in the 3rd subtype exhibit lower severities. (nms\_d4 and nms\_d9 turn out to be insignificant). In particular, the divergence of nonmotor symptoms nms\_d2 and nms\_d3 and the corresponding motor symptom axial is important in classifying the entire nonmotor subtype, and is discussed in the decision trees for the nonmotor-dominated group in Section 4.1.

---

<sup>4</sup> $\lambda = 0.98$ ,  $q = 0$ ,  $\text{maxits} = 1000$ ,  $\text{convits} = 100$

Cluster	Variables	95% CI	$p$
1	bradykin, cisitot	[-0.0225, 0.171]	0.131
	rigidity, cisitot	[-0.000406, 0.192]	0.0510
	bradykin, nms_d6	[-0.0634, 0.131]	0.493
	rigidity, nms_d6	[-0.101, 0.0932]	0.934
2	bradykin, cisitot	[0.0786, 0.351]	0.00248(**)
	rigidity, cisitot	[-0.215, 0.069]	0.310
	bradykin, nms_d6	[-0.152, 0.133]	0.897
	rigidity, nms_d6	[-0.123, 0.163]	0.781
3	bradykin, cisitot	[0.0995, 0.350]	0.000620(***)
	rigidity, cisitot	[0.0687, 0.322]	0.00298(**)
	bradykin, nms_d6	[0.0350, 0.292]	0.0134(*)
	rigidity, nms_d6	[-0.0846, 0.179]	0.478
4	bradykin, cisitot	[0.454, 0.724]	$3.97 \times 10^{-10}(***)$
	rigidity, cisitot	[0.375, 0.675]	$4.99 \times 10^{-08}(***)$
	bradykin, nms_d6	[0.297, 0.624]	$2.60 \times 10^{-06}(***)$
	rigidity, nms_d6	[0.278, 0.611]	$6.43 \times 10^{-06}(***)$

Table 6: Correlation tests. (\*)  $p < 0.05$ , (\*\*)  $p < 0.01$ , (\*\*\*),  $p < 0.001$

Criterion	Optimal $k$
Minimum ASW	2
BIC	1 (?)
SSE Scree Plot	Inconclusive
Gap Statistic	3
Affinity Propagation <sup>4</sup>	5

Table 7: Results of various techniques for determining  $k$ , applied to subtype 2

## 4 Modeling

One further step of this investigation was to produce accurate, practical models that could be used in a clinical setting to predict the subtype of PD based on previous clustering results. Cluster assignments were treated as labels in a supervised classification problem in an attempt to produce useful models.

### 4.1 One-versus-all decision trees

While the decision tree in Figure 1 is useful, it could be considered overly complicated. Additionally, a model is not necessarily needed to make simpler diagnoses such as classifying a patient as mildly affected (subtype 1) or severely affected (subtype 4). One-versus-all (OVA) decision trees were thus considered, in order to isolate the classification problem and look at possible distinguishing characteristics of individual subtypes. These OVA decision trees for all 4 subtypes are located in Figures 9, 10, 11, and 12. Trees are pruned by selecting

the version of tree with the minimum 10-fold cross-validated error.

#### 4.1.1 1 (mild)

The tree for the mild subtype classifies mainly based on negative responses to nodes asking whether the patient has a relatively severe manifestation of a symptom. The majority of examples are classified by following the bradykinesia  $\leq 2.5$ , which subsequently tests the severity of several nonmotor symptoms. Most of subtype 1 patients that score relatively mildly on these scales are classified this way. There are also small populations of patients who 1) score higher on bradykinesia but lower with axial, tremor, rigidity, and nms\_d2 and 2) score higher in nms\_d2 (sleep) but lower with nms\_d7 (urinary).

#### 4.1.2 2 (nonmotor-predominant)

The decision tree for the nonmotor-predominant subtype is quite simple.

Interestingly, although nms\_d9 (miscellaneous) is not the most important nonmotor symptom, since the information gain is less than nms\_d2 and nms\_d3 and it does not appear very high in the 4-class decision tree, it is used as the root node of this decision tree, classifying over half of the negative examples based on whether the subject has a low severity of miscellaneous symptoms (nms\_d9  $\leq 7.5$ )<sup>5</sup>. This could be an indication that nonmotor-predominant PD patients do indeed have a wide manifestation and variety of nonmotor symptoms. After classifying on nms\_d9, the tree then classifies negative examples as having rigidity  $\leq 3.5$ , an example of how subtype 2 patients have relatively low motor symptoms. Finally, the tree classifies on the nonmotor symptom with the most information gain, nms\_d2, where patients  $\leq 7.5$  are classified as falling into subtype 2.

#### 4.1.3 3 (motor-predominant)

This tree classifies overwhelmingly on severity of bradykinesia, with 476 negative examples when bradykinesia is less than 2.5. The resulting tree is quite complex, but generally, nodes check again for severity of motor symptoms (tremor is the next node) and end up classifying positive examples based on both mildness of nonmotor symptoms and severity of motor symptoms. For example, in the furthest right branch, once nms\_d2 (as we know, an important feature) is established to be relatively mild ( $\leq 12$ ), the test for subtype 3 involves several more nodes verifying the severity of rigidity, tremor, and axial, and the mildness of nms\_d7 (urinary).

#### 4.1.4 4 (severe)

The OVA tree for patients severely affected in all areas is predictable, testing entirely on whether or not symptoms (both motor and nonmotor) are relatively severe. Positive nodes always appear to the right (no) of less-than checks. Interestingly, however, nms\_d4 (percep/hallucinations), previously not of note, is used twice as the root node of a tree and again further down. As the boxplot display in Figure 2 shows, nms\_d4 is perhaps the most

---

<sup>5</sup>Remember 0  $\leq$  nms\_d9  $\leq 48$ .

distinguishing symptom of subtype 4 (severe) against nonmotor-predominant subtype 2 in particular, as subtype 2 has relatively mild percep/hallucination symptoms, in contrast to the comparable levels of severity for other nonmotor symptoms in both groups. This shows that issues with perception and hallucinations generally occur in only the most severe cases of PD, and are relatively rare when a patient exhibits a nonmotor-predominant form of PD.

## 4.2 Different angles of exploration: 2 and 4 vs rest

There are many more interesting questions to be asked when examining the relationship between these clusters. One thing that may be helpful in understanding the relationship between the clusters is exploring different groupings of clusters for decision trees. The tree in Figure 13 is a preliminary example of this kind of exploration. In this tree, the node classifying examples as subtype 4 is localized to the furthest right branch. Predictably, examples in this node have scored relatively higher in rigidity ( $i=3.5$ ). Interestingly, a classification decision that is replicated in the 4 versus all decision tree is the decision to use nms\_d7 (urinary) as a node, where subtype 4 is classified as having relatively high nms\_d7 components ( $i=12$ ). Indeed, as shown in Figure 2, the mean of nms\_d7 severity is especially higher in cluster 4 than cluster 2.

## 5 Conclusion

$k$ -means clustering on this Parkinsons' Disease data set reveals clusters that confirm previous computationally-based findings in the field [2], mainly concerning the identification of four subtypes of Parkinson's disease: mild, nonmotor-predominant, motor-predominant, and severe. The most important nonmotor symptoms in determining these clusters were nms\_d2 (sleep) and nms\_d3 (mood/cognition), which echo findings of Fereshtehnejad's longitudinal study [1]. More work needs to be done ???

Nonmotor symptoms nms\_d2 and nms\_d3 became critical not only in classification trees distinguishing between the various symptoms but in the nonmotor-predominant subgroup itself, where both standard  $k$ -means analysis and decision tree branches show two possible trends in the manifestation of nonmotor-predominant PD:

1. Axial and sleep-severe PD
2. mood/cognition-severe PD

I am, however, not sure if these groups have enough members to warrant a subtype, or whether this could just be chance or noise.

It remains to be seen whether these classification models, especially the one-vs-all decision trees, are useful in clinical practice.

## 5.1 Bayesian Networks

### 5.1.1 On all data

I decided to discretize the data into three uniform-width groups based on the scales of each symptom. In other words, each symptom was discretized into a mild, moderate, and severe



bin. Continuous data was unreliable on my computer, and updating intricately connected nodes like `nms_d2` resulted in slowdowns and crashes on my computer.

Two bayesian network algorithms were tried: the default Bayesian score-search algorithm and the PC conditional independence tests algorithm. I couldn't find the exact name of the Bayesian search implementation, but it was the default method used by GeNIe. GeNIe files will be attached electronically.

I assume these models are to be looked at by Dr. Martín. I have not done too much investigation myself, as I'm not exactly sure what I'm looking for.

### **5.1.2 On nms-dominated data**

I tried to construct Bayesian networks based on the nms-dominated subtype, but the data was too sparse to create a very informative network, even when leaving the information continuous. However, I'm not sure this is necessary. If it is, I can work on this problem more.

## **References**

- [1] Fereshtehnejad et al (June 15, 2015). New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression
- [2] van Rooden et al (2010). The Identification of Parkinson's Disease Subtypes Using Cluster Analysis: A Systematic Review

# UNSCALED Pruned Tree, 904 clusters

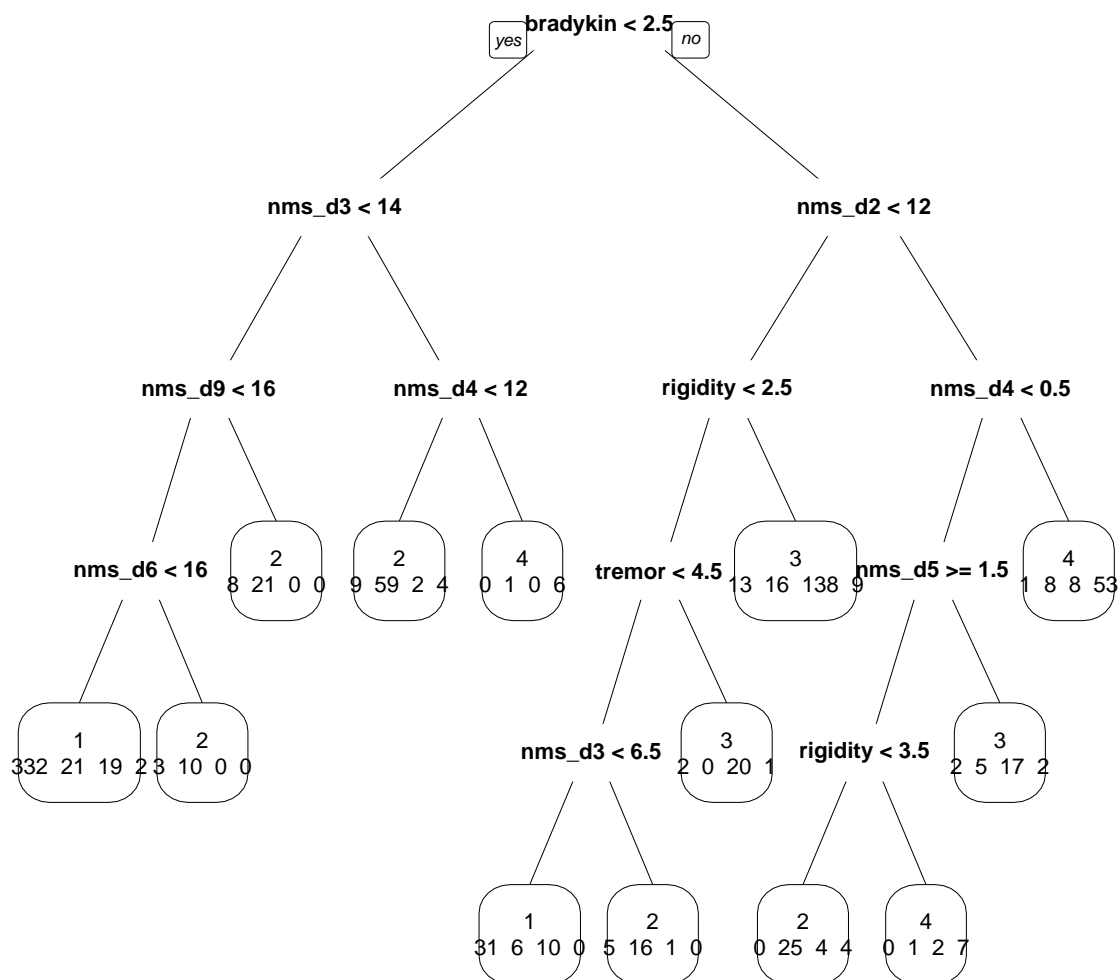


Figure 1: Decision Tree from  $k$ -means clustering, 4 clusters

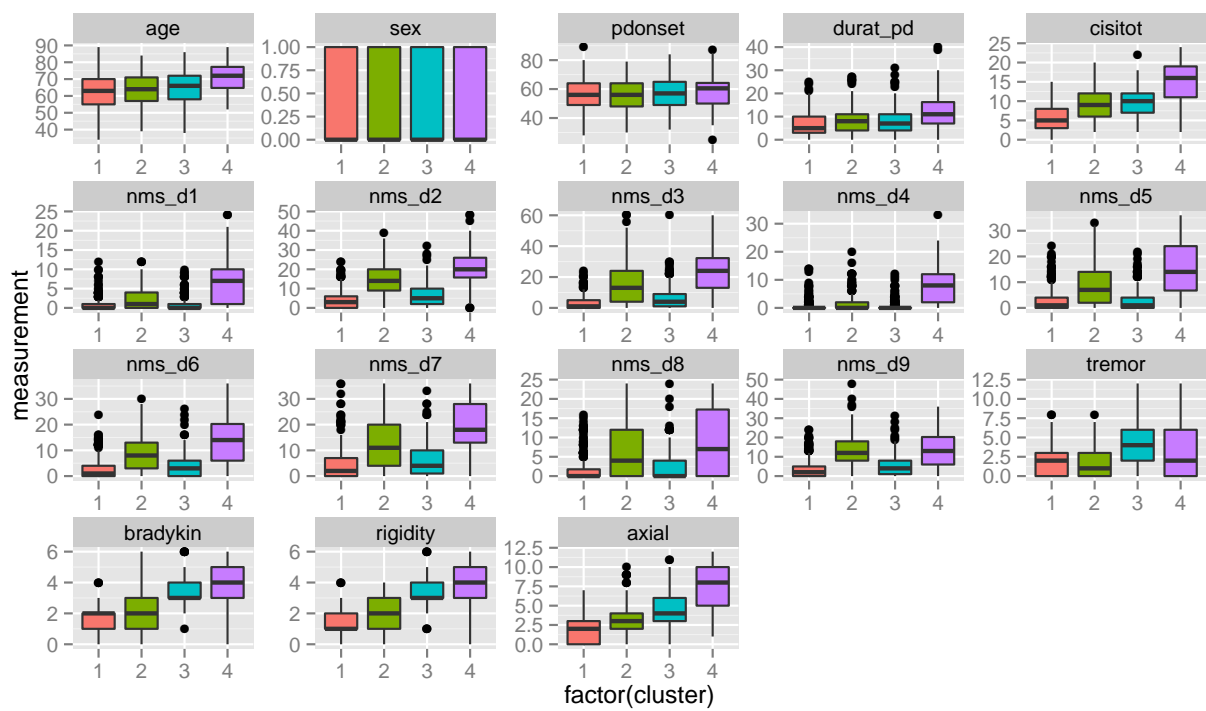
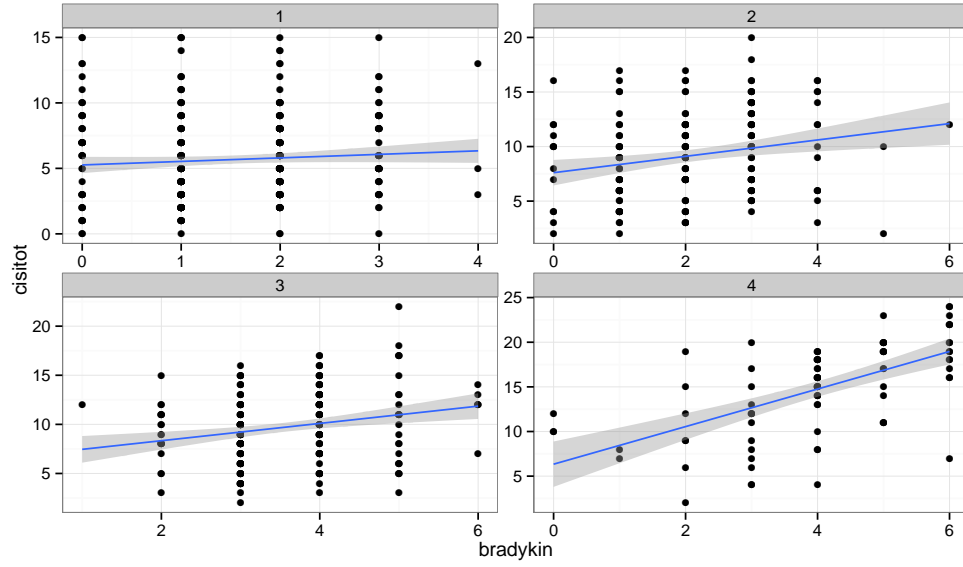
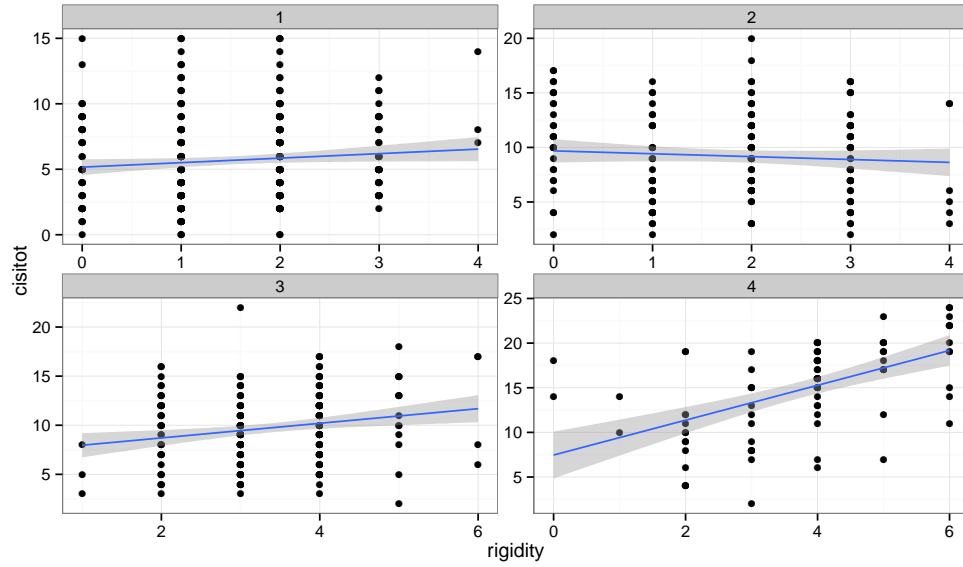


Figure 2: Cluster Summaries,  $k = 4$

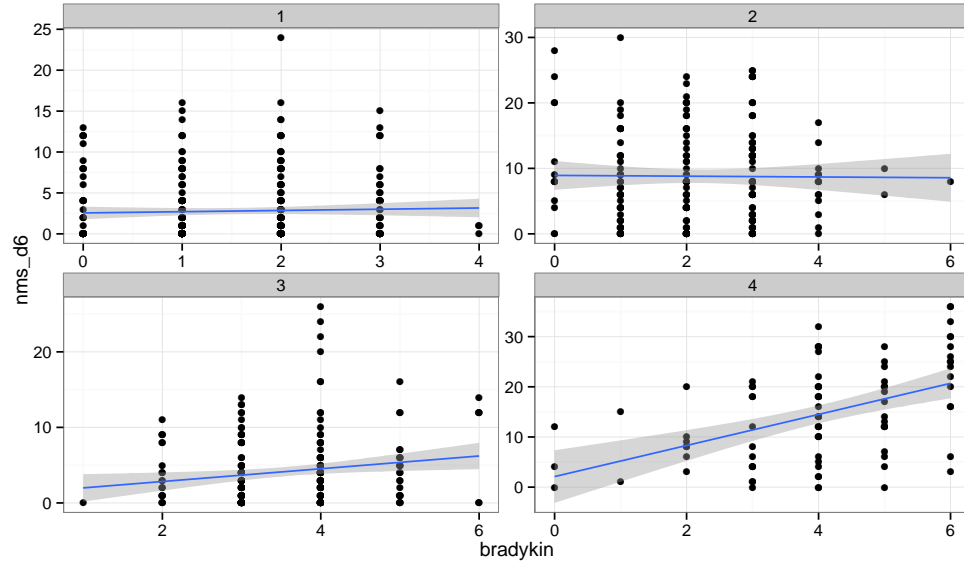


(a)

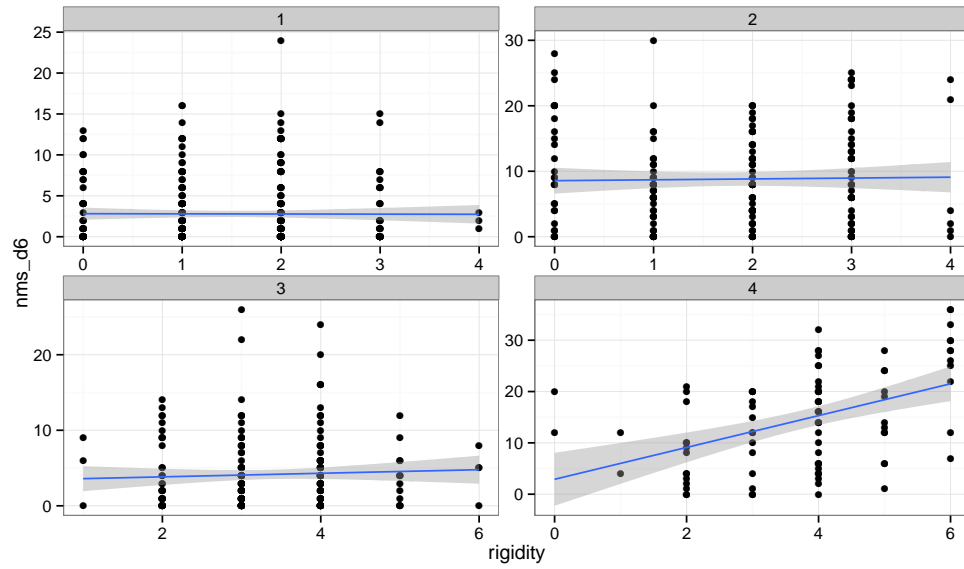


(b)

Figure 3: Relationship between (a) bradykinesia, (b) rigidity and overall severity (cisitot). Shaded band is 95% confidence interval.



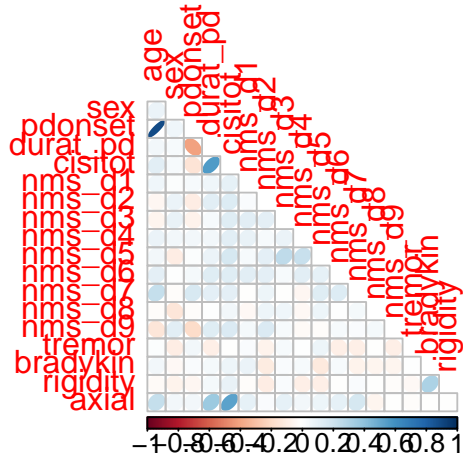
(a)



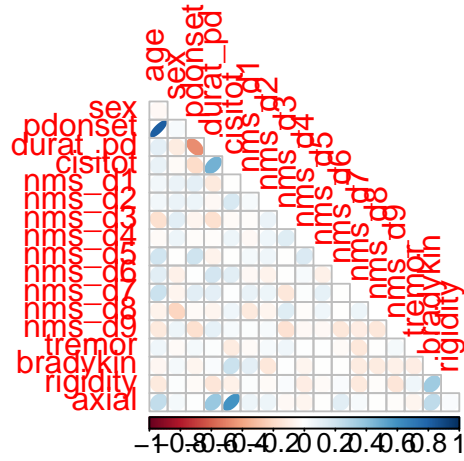
(b)

Figure 4: Relationship between (a) bradykinesia, (b) rigidity and nms\_d6 (gastrointestinal). Shaded band is 95% confidence interval.

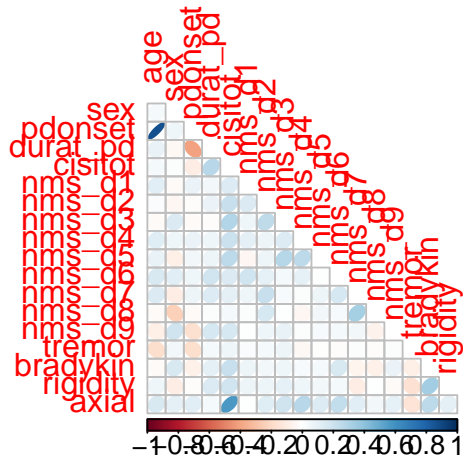
**Correlation 1**



**Correlation 2**



**Correlation 3**



**Correlation 4**

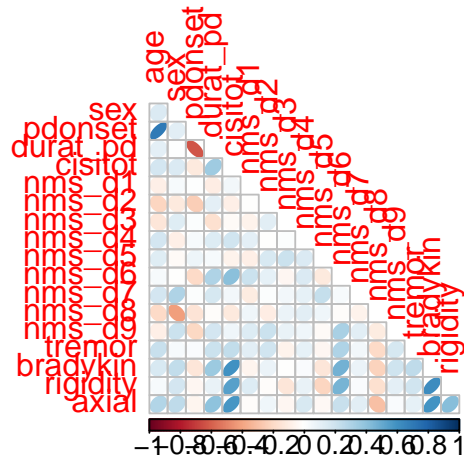


Figure 5: Correlation plots

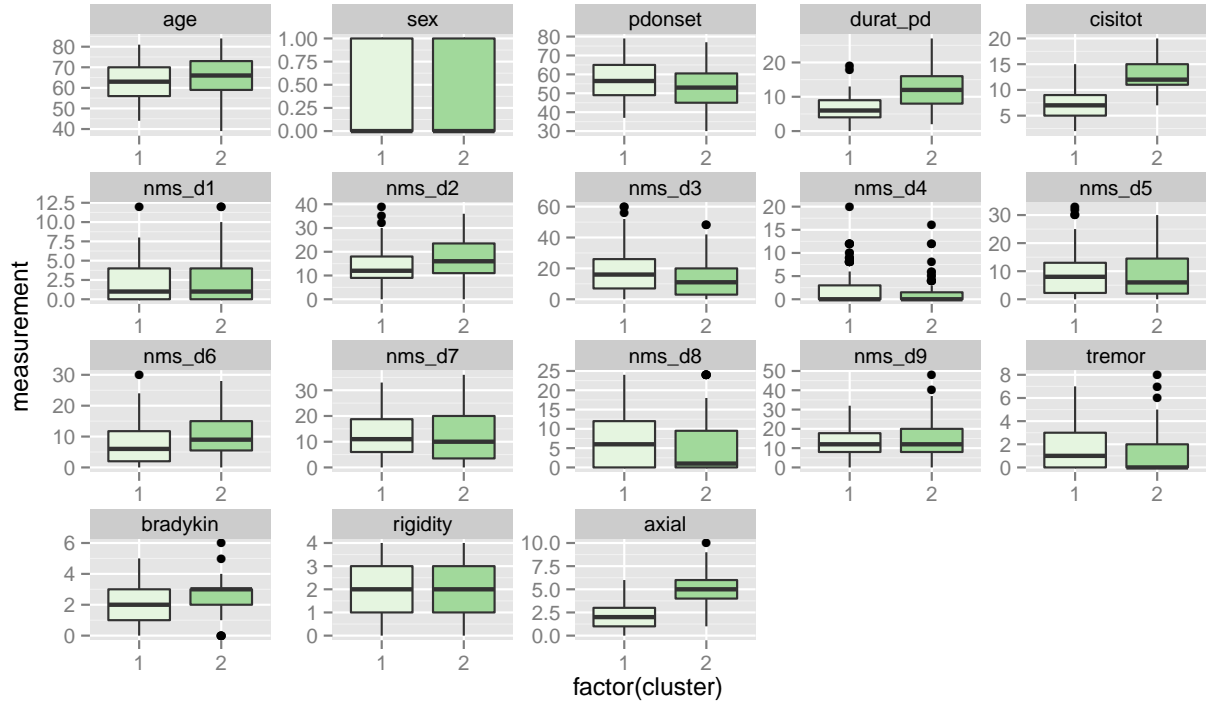


Figure 6: Clustering on nonmotor group:  $k = 2$

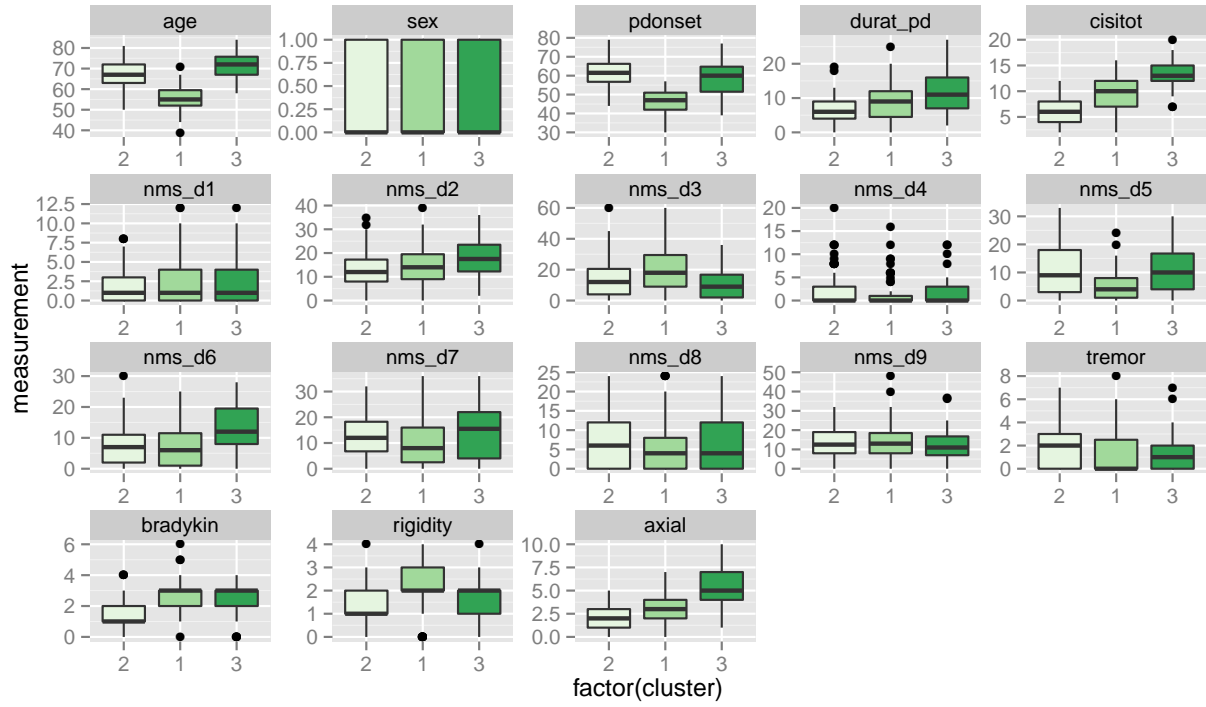


Figure 7: Clustering on nonmotor group:  $k = 3$

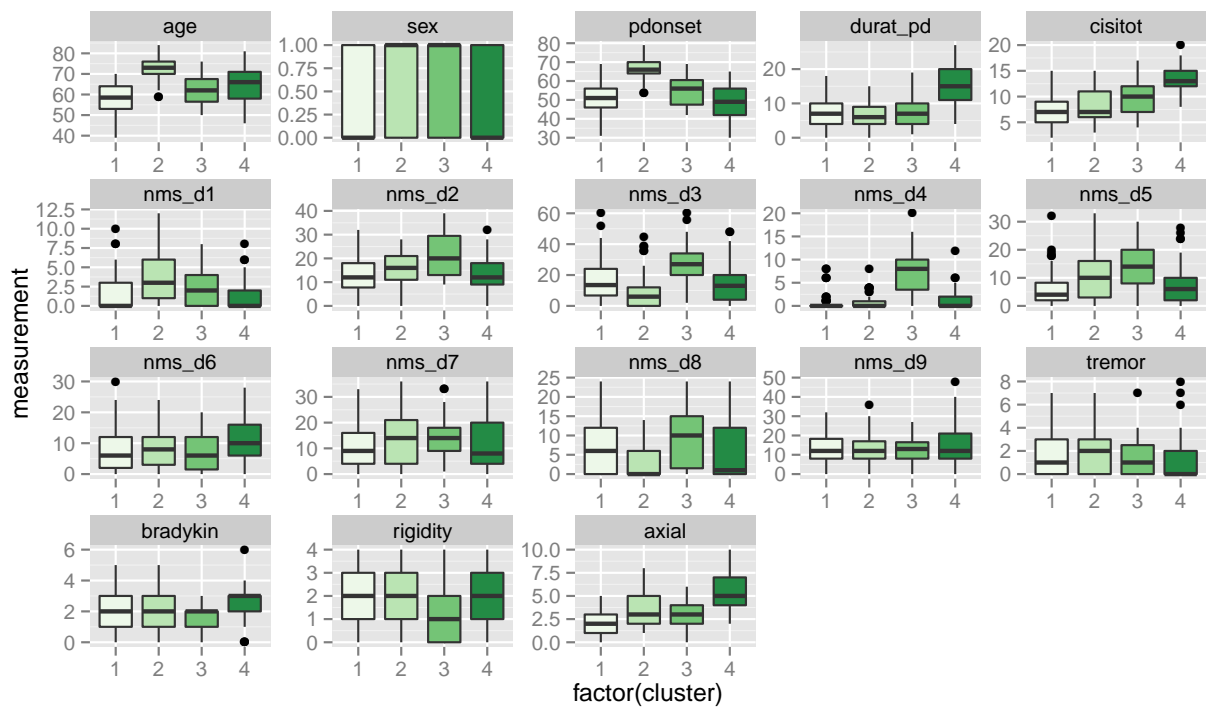


Figure 8: Clustering on nonmotor group:  $k = 4$



## Pruned 1 vs all

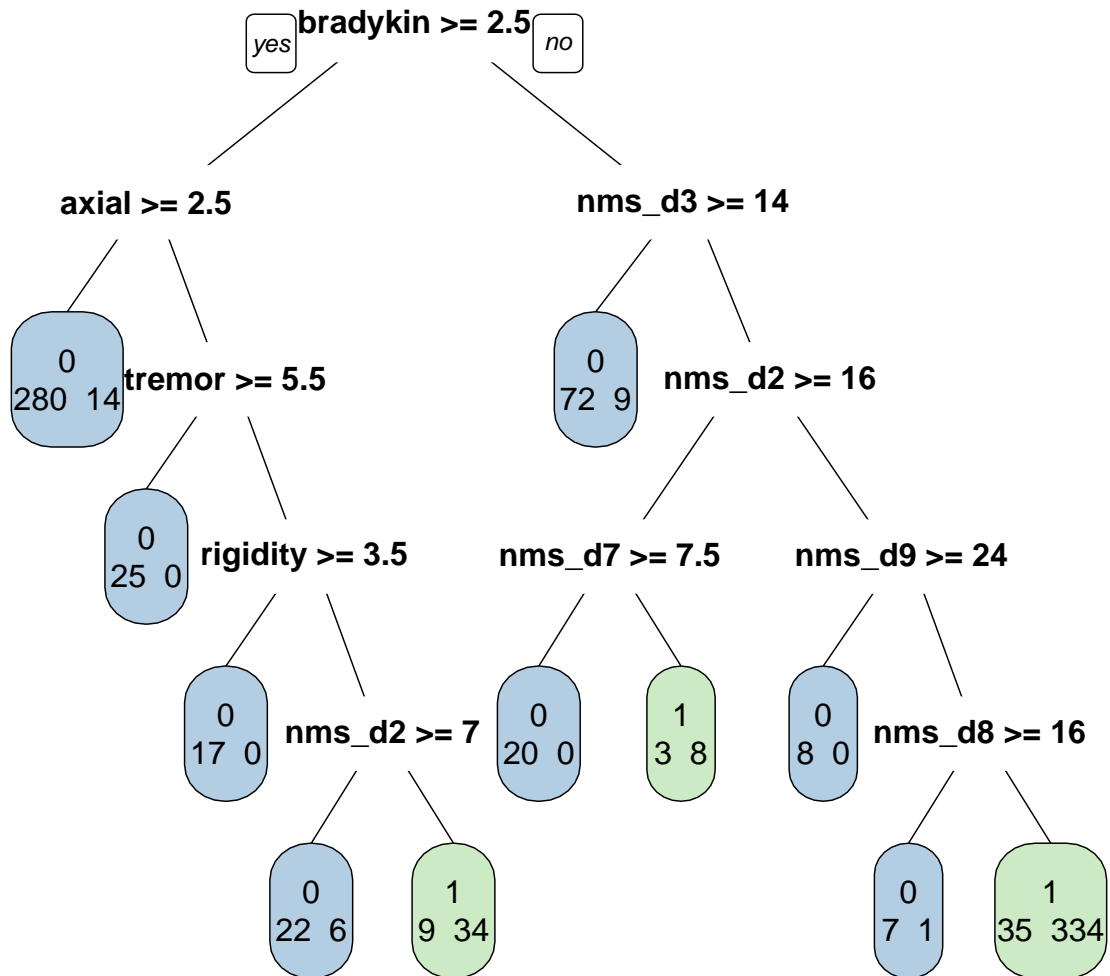


Figure 9: Cluster 1 (mild) vs all

## Pruned 2 vs all

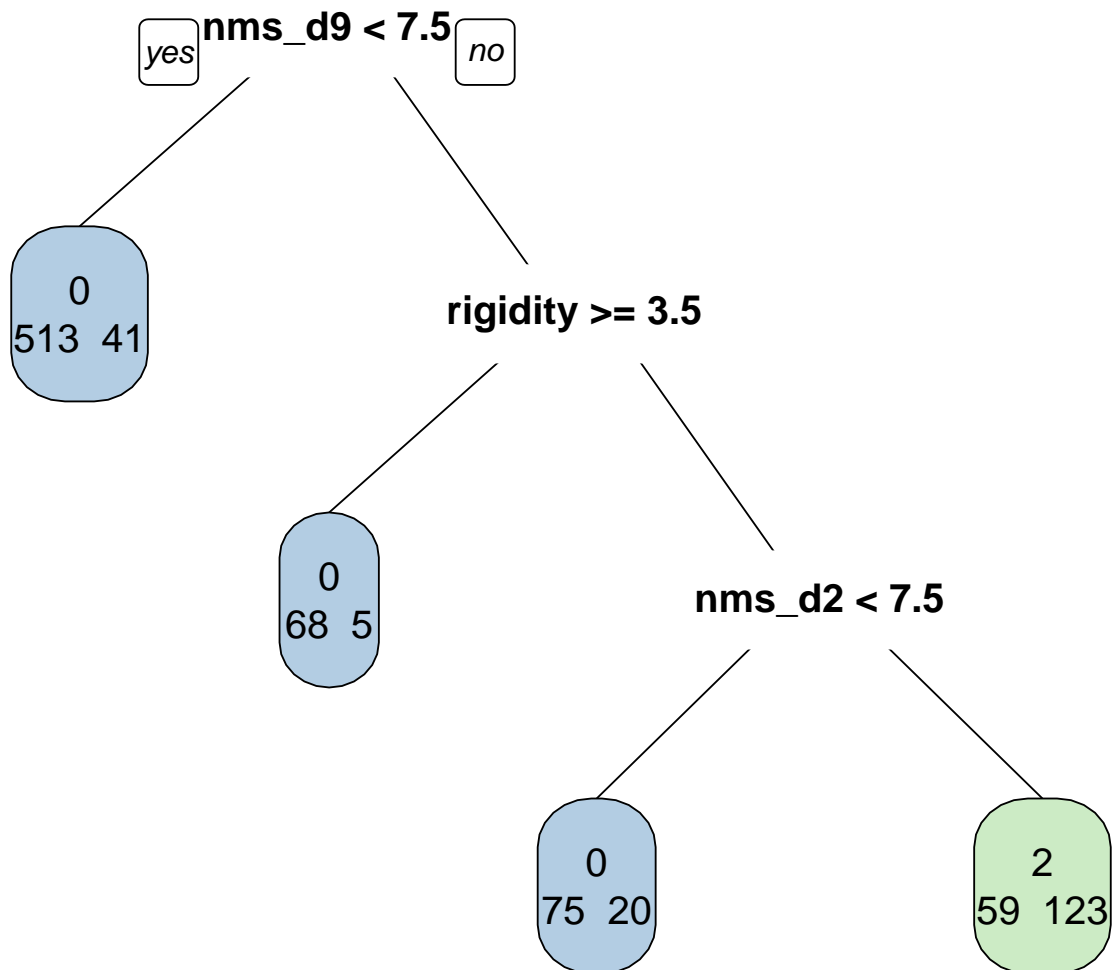


Figure 10: Cluster 2 (nonmotor-dominated) vs all

## Pruned 3 vs all

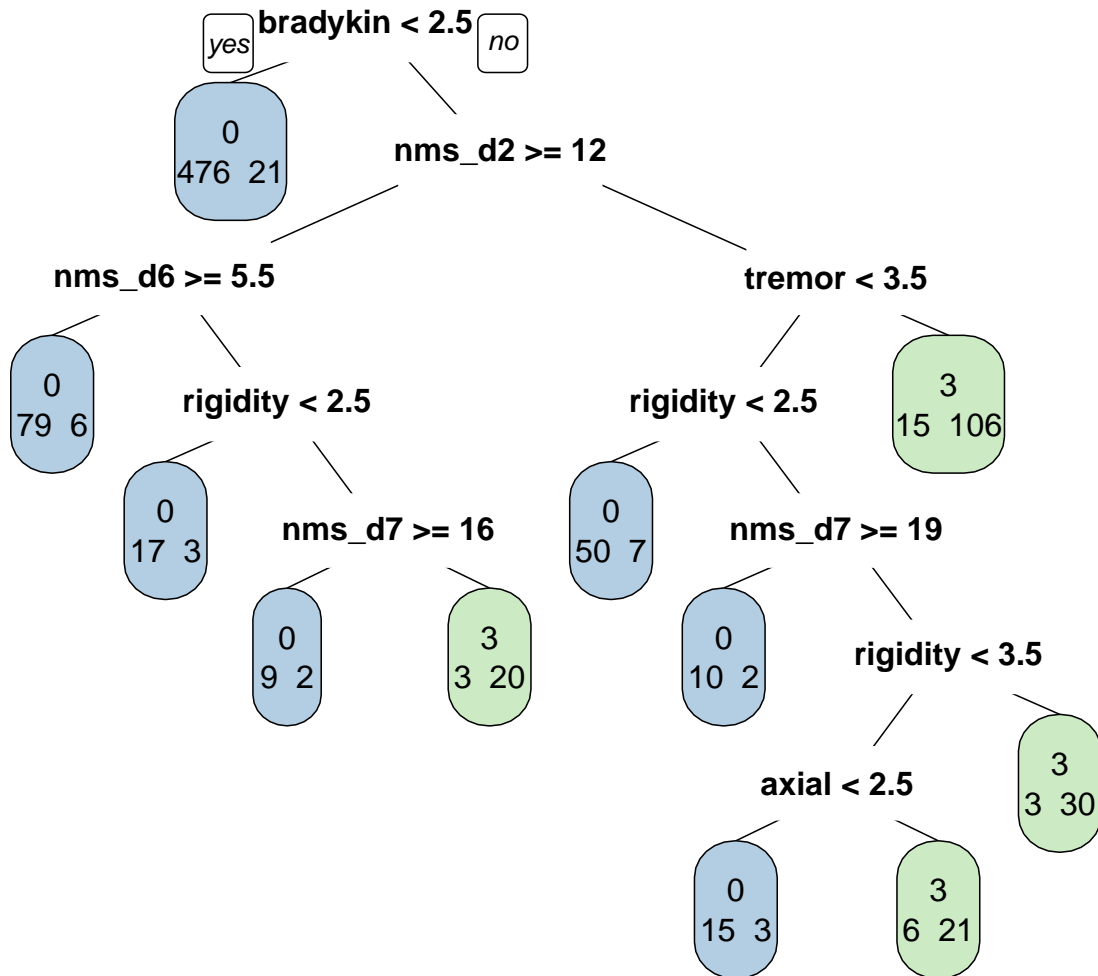


Figure 11: Cluster 3 (motor-dominated) vs all

## Pruned 4 vs all

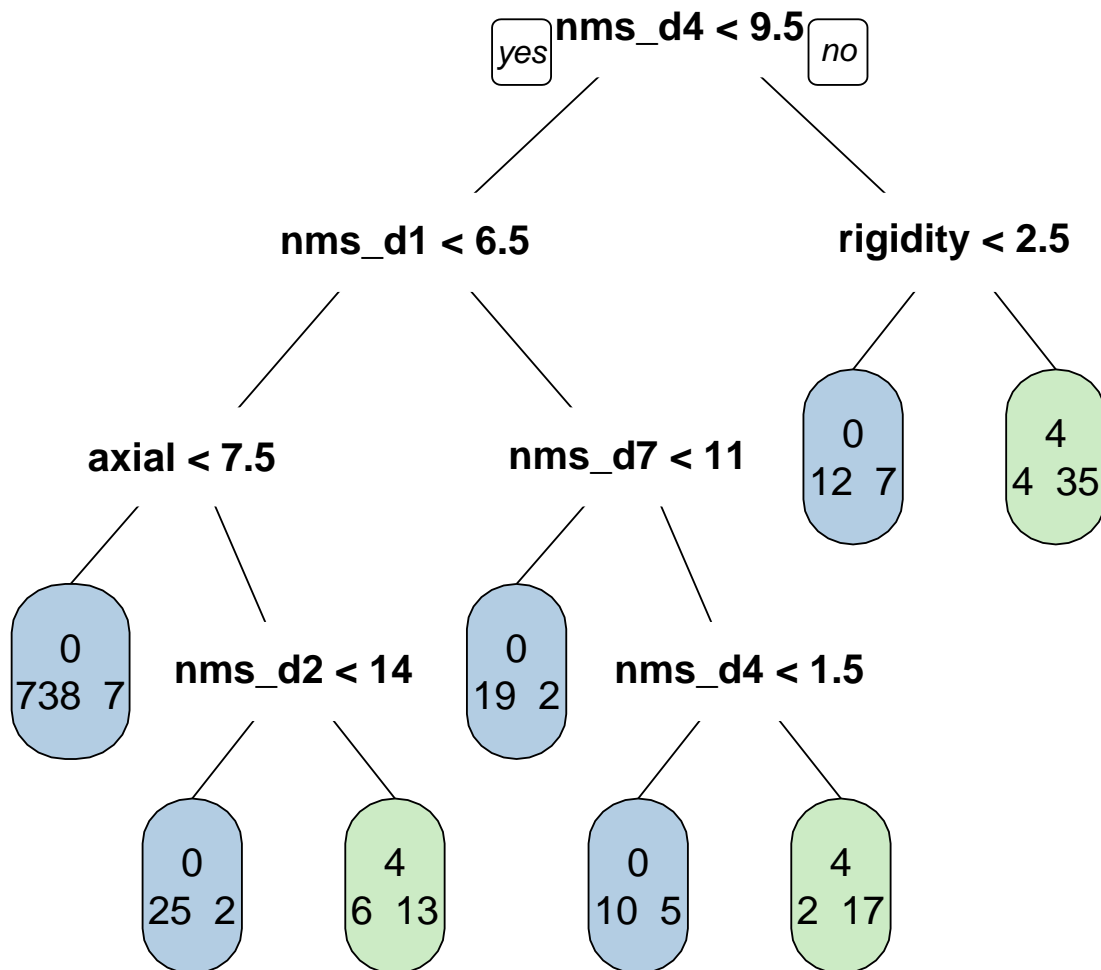


Figure 12: Cluster 4 (severe) vs all

## Pruned 2 and 4 vs rest

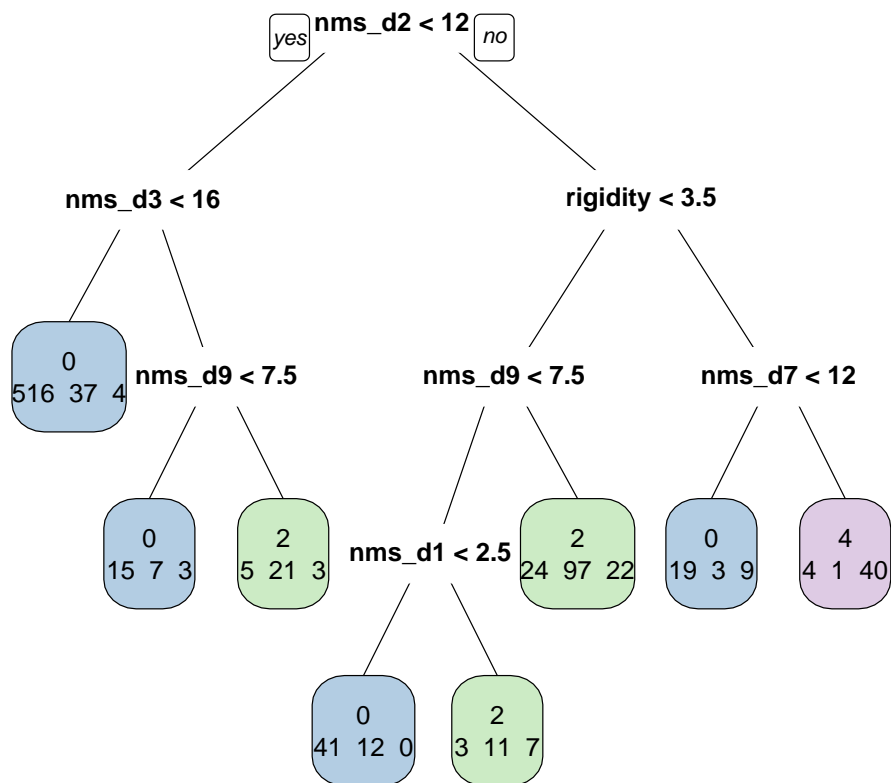


Figure 13: Clusters 2 (nms) and 4 (severe) vs rest (1 and 3)