



## Verslag project 1

Academiejaar 2018-2019

Thomas Bamelis R0640219 & Michiel Jonckheere R0665594

### Inhoudsopgave

<b>1 Clustering</b>	<b>2</b>
1.1 Zonder schalen . . . . .	2
1.2 Met schalen . . . . .	2
1.3 Beschrijving clusters . . . . .	2
<b>2 Bijlage</b>	<b>5</b>
2.1 Clustering . . . . .	5
2.1.1 Zonder schalen . . . . .	5
2.1.2 Met schalen . . . . .	9
2.2 Beschrijving clusters . . . . .	12

# Introductie

In dit verslag wordt nagegaan hoe de oorzaken van overlijden verschillen tussen landen en regio's in de wereld. Er zijn schattingen van het aantal overlijdens beschikbaar voor 183 landen, opgesplitst naar 32 verschillende doodsoorzaken. De landen worden gegroepeerd in 6 groepen volgens geografische ligging en 2 groepen naargelang de globale ontwikkeling van het betreffende land. De gegevens met betrekking tot de doodsoorzaken zijn afkomstig van de Wereldgezondheidsorganisatie [1] en betreffen het jaar 2016, de indeling in groepen is deze volgens de Verenigde Naties [2]. Deze gegevens werden verwerkt en geïnterpreteerd als proporties van de soorten sterfgevallen per land.

## 1 Clustering

Als eerste werd een cluster-analyse uitgevoerd op de gegevens. Eerst bespreken we de gegevens zonder schalen, daarna met.

### 1.1 Zonder schalen

Om een idee te krijgen van hoeveel clusters er best worden genomen, werden het agglomerate nesting algoritme en divisive analysis toegepast. Agglomerate nesting werd gedaan met de volgende dissimilariteiten: group average, nearest neighbour en furthest neighbour, in die volgorde met daarna divisive analysis. Zie figuren 1 op p5 en 2 op p6 in de bijlage 2. Gegeven deze figuren lijkt het meest aannemelijk om 2, 4 en 6 klassen te proberen. De gebruikte clustering algoritmes zijn in volgorde k-means, partitioning around mediods en fuzzy analysis. De clustering ermee voor 2, 4 en 6 klassen werd geëvalueerd via een silhouette plot en een clusplot. Zie figuur 3 op p7. Hieruit blijkt dat partitioning around mediods met 2 clusters het beste presteert met een silhouette coëfficiënt van 0.50 (cluster 1 : 0.69 en cluster 2 : 0.43). Dit is niet bepaald goed en balanceert op het randje van een zwakke structuur.

### 1.2 Met schalen

We trekken hierbij dezelfde conclusies omtrent het aantal klassen, 2, 4, en 6. Zie figuren 4 op p9en 5 op p10. Na dezelfde clustering algoritmes toegepast te hebben (figuur 6 op p11), is de best geobserveerde silhouette coëfficiënt 0.23. Hieruit besluiten we dat clustering met schalen aanzienlijk slechter is dan zonder schalen. We besluiten dus verder te werken met het beste resultaat zonder schalen.

### 1.3 Beschrijving clusters

We bekijken nu nader de twee clusters geselecteerd door pam met twee clusters zonder schalen. We bekijken eerst hoeveel landen uit een bepaalde regio in een bepaalde cluster zitten.

Cluster	Africa	America	Asia	Europe	Oceania
1	45	0	2	0	1
2	9	33	44	40	9

Hieruit kunnen we afleiden dat 45 van de 48 landen in de eerste cluster Afrikaanse landen zijn. Cluster twee bevat bijna alle landen uit America, Asia, Europe en Oceania. Ze bevat ook nog 9 Afrikaanse landen. De eerste cluster neem dus 4/5 van de Afrikaanse landen en op 3 landen na. Het clustering algoritme vindt dus vooral onderscheid tussen Afrikaanse landen tegenover de rest van de wereld qua doodsoorzaken.

Cluster	#N/B	Developed	Developing	Transition
1	1	0	47	0
2	0	36	90	9

Clustering tegenover ontwikkeling toont dat de eerste cluster enkel developing landen selecteert, op Congo na waarvan de ontwikkeling onbepaald is. Het valt echter op dat cluster twee dubbel zoveel developing landen selecteert vergeleken met de eerste cluster, maar ook alle developed en transition landen.



Het is dus niet zo dat de eerste cluster focused op alle developing landen. Als we dit samen leggen met de region table, kunnen we besluiten dat de eerste cluster hoofdzakelijk Afrikaanse developing landen bevat en de tweede cluster “de rest”.

Daarnaast kunnen we de verschillen van tussen de clusters bekijken qua doodsoorzaken. We plotten daarom de marginale gemiddelden van de eerste cluster, afgetrokken met de marginale gemiddelden van de tweede cluster. Zie figuur 7 op p12. Als we de drie hoofdcategoriën van doodsoorzaken bekijken (de verschillende kleuren), blijkt dat communicable, maternal, perinatal and nutritional conditions meer voorkomen bij de eerste cluster dan bij de tweede. We zien ook dat twee noncommunicable diseases aanzienlijk meer voorkomen in de tweede cluster, met nog eens 5 van die doodsoorzaken licht meer voor komen in de tweede cluster. Over de injuries tussen de twee clusters valt niets significant te zeggen. Met dit alles samen kunnen we besluiten dat de meeste Afrikaanse landen die developing zijn een propotioneel opvallend hoger aantal communicable, maternal, perinatal and nutritional conditions bevatten en een propotioneel lager aantal noncommunicable diseases hebben tegenover de rest van de wereld.

## Besluit

TODO

## Referenties

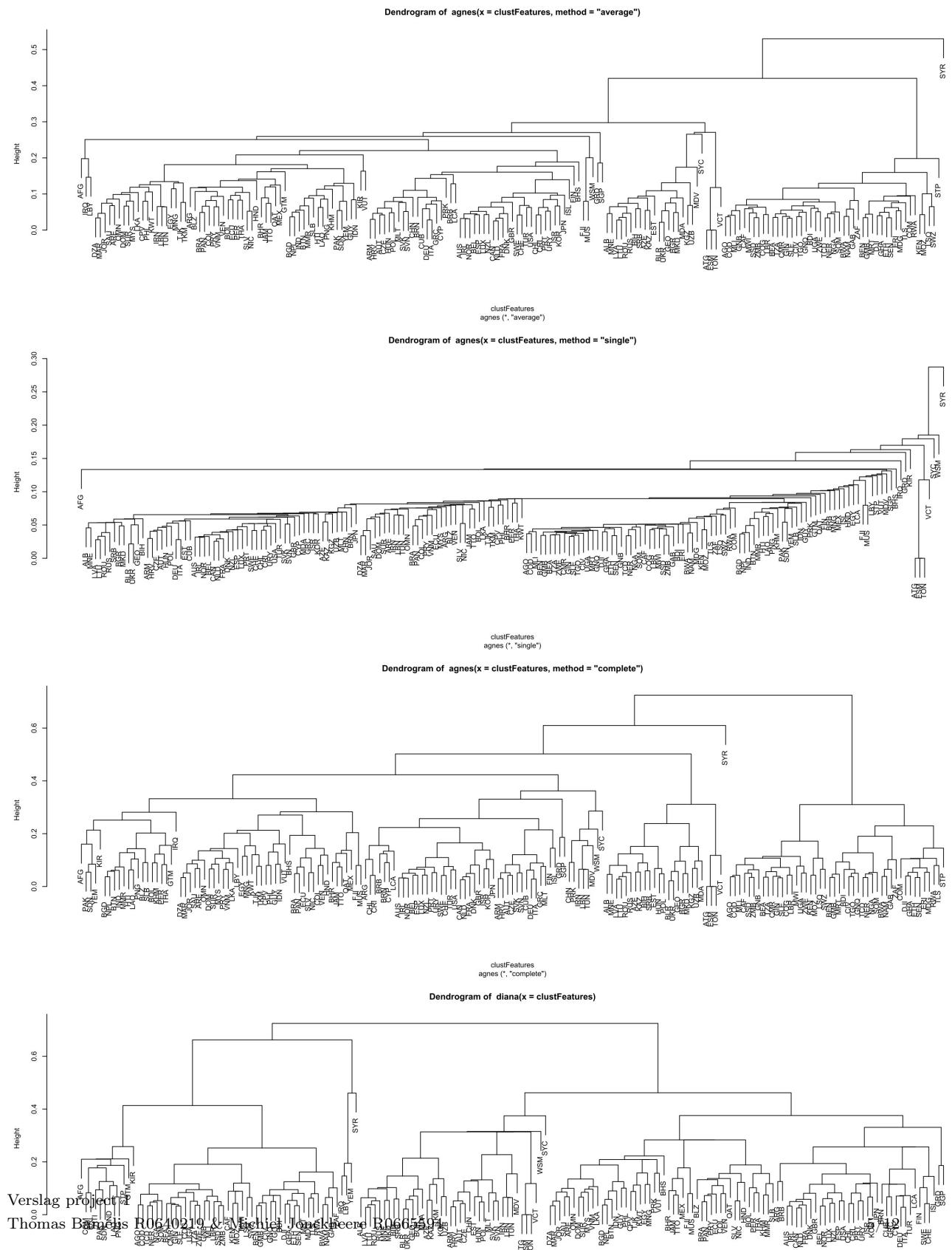
- [1] Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
- [2] Country classification, june 2018. Geneva, United Nations Conference on Trade and Development; 2018.

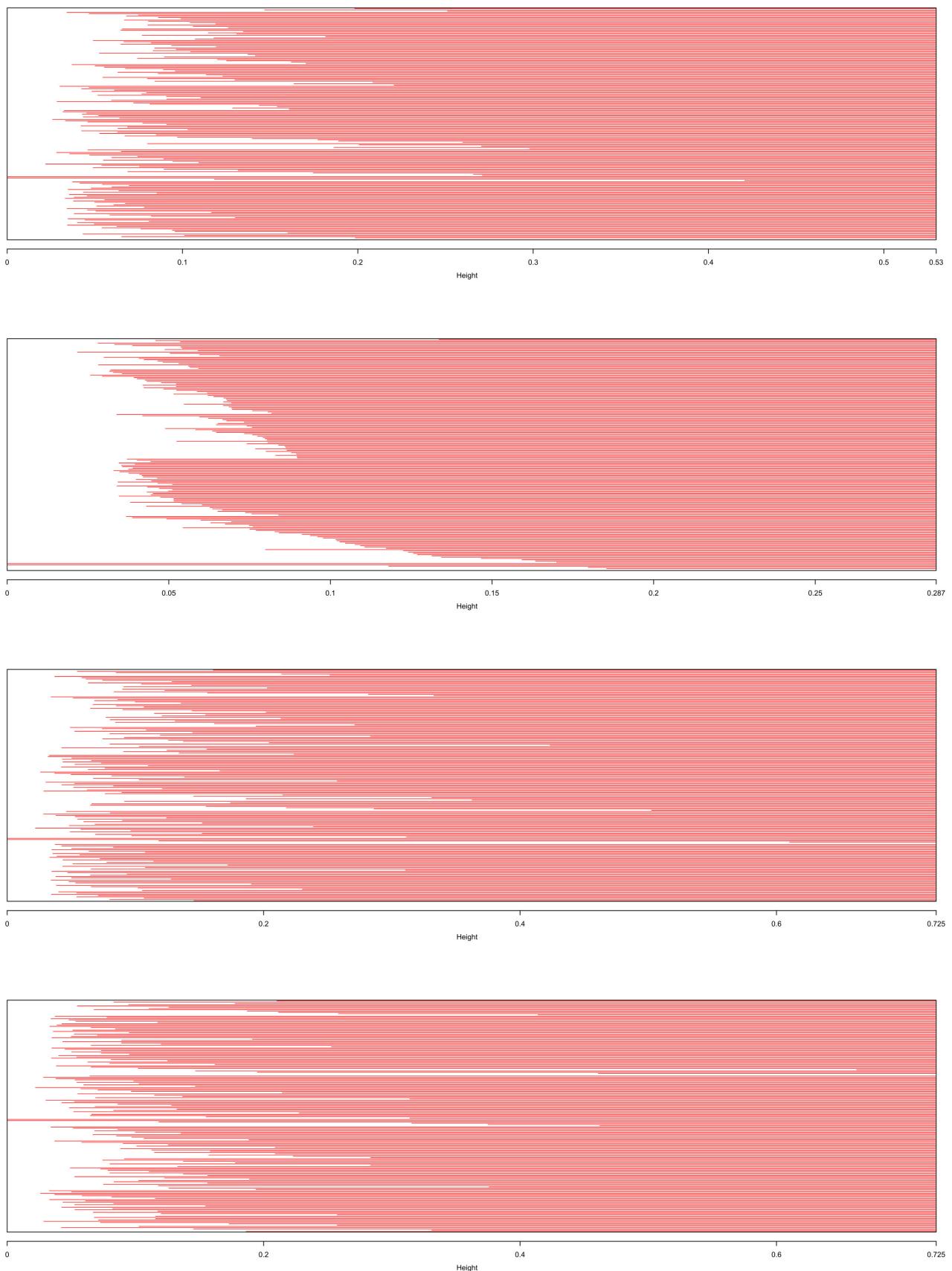


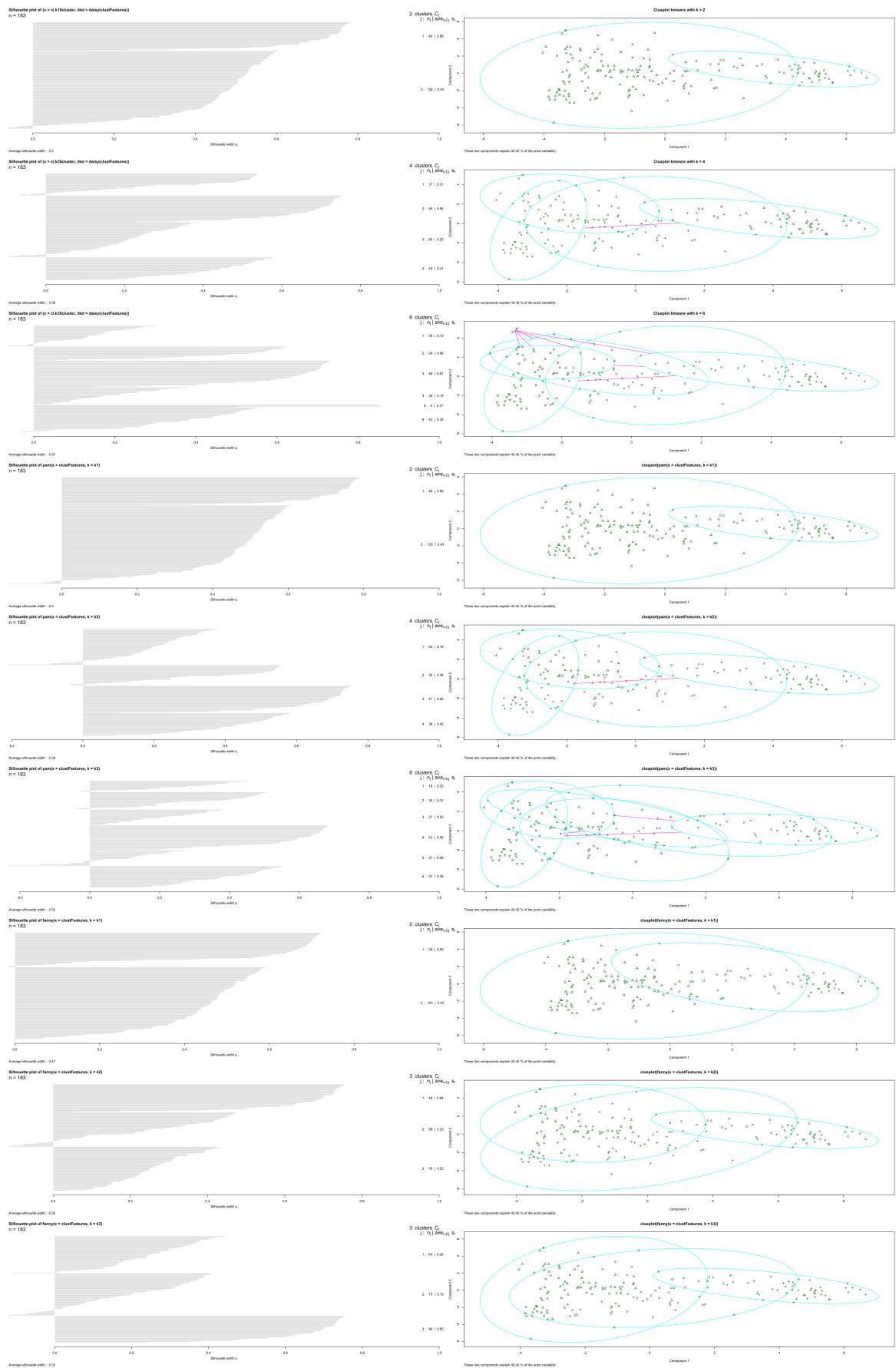
## 2 Bijlage

### 2.1 Clustering

#### 2.1.1 Zonder schalen

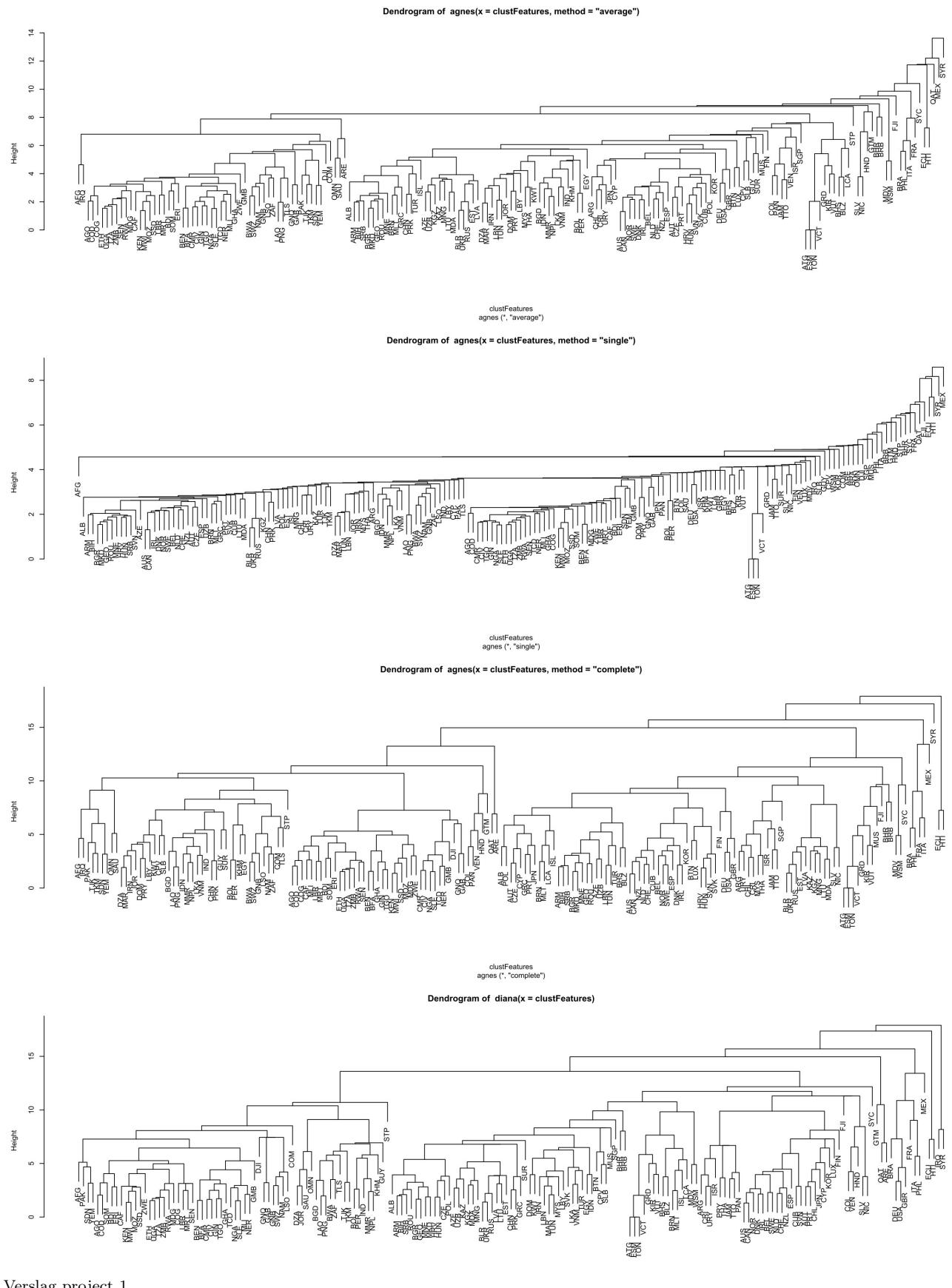




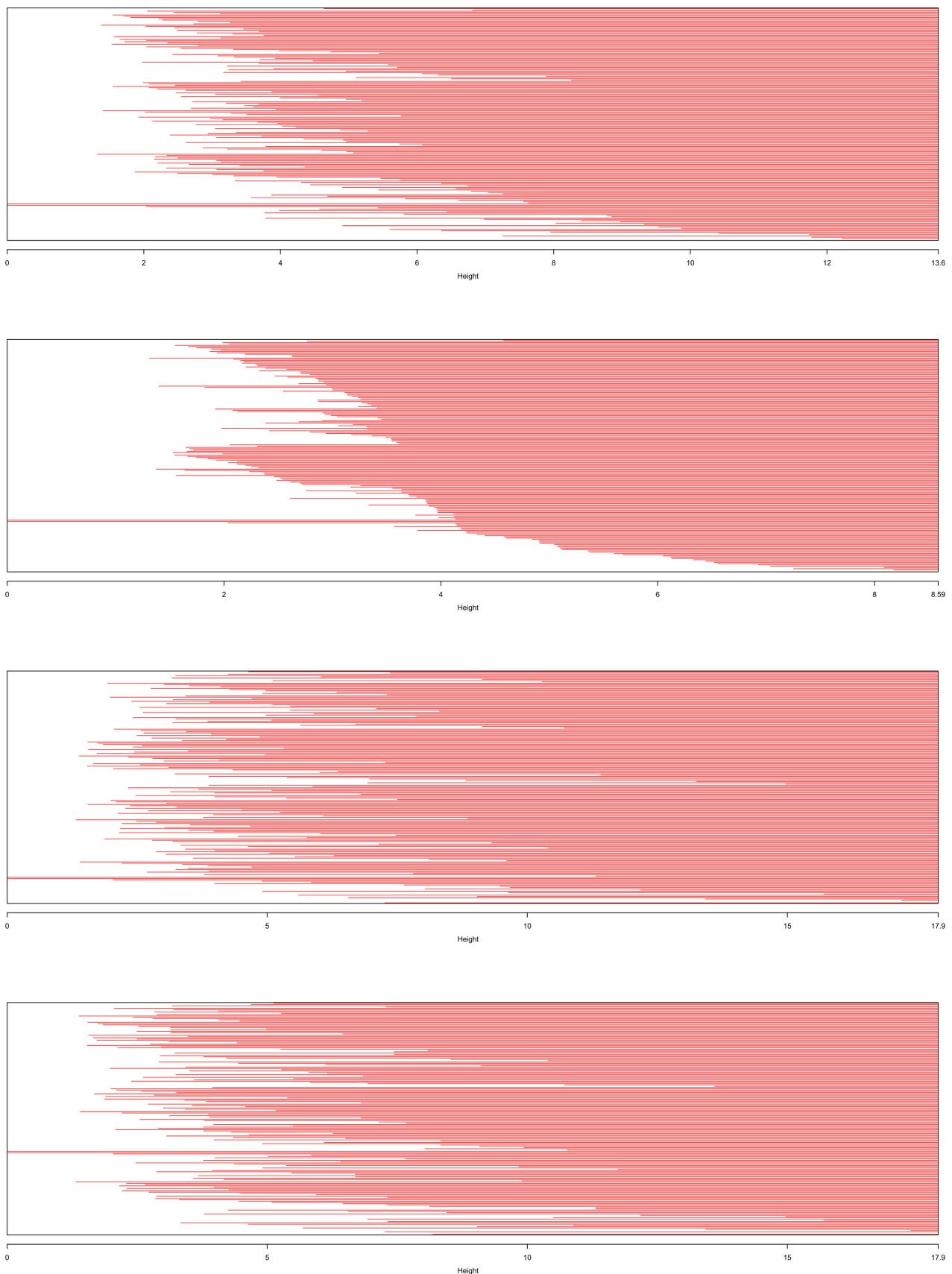


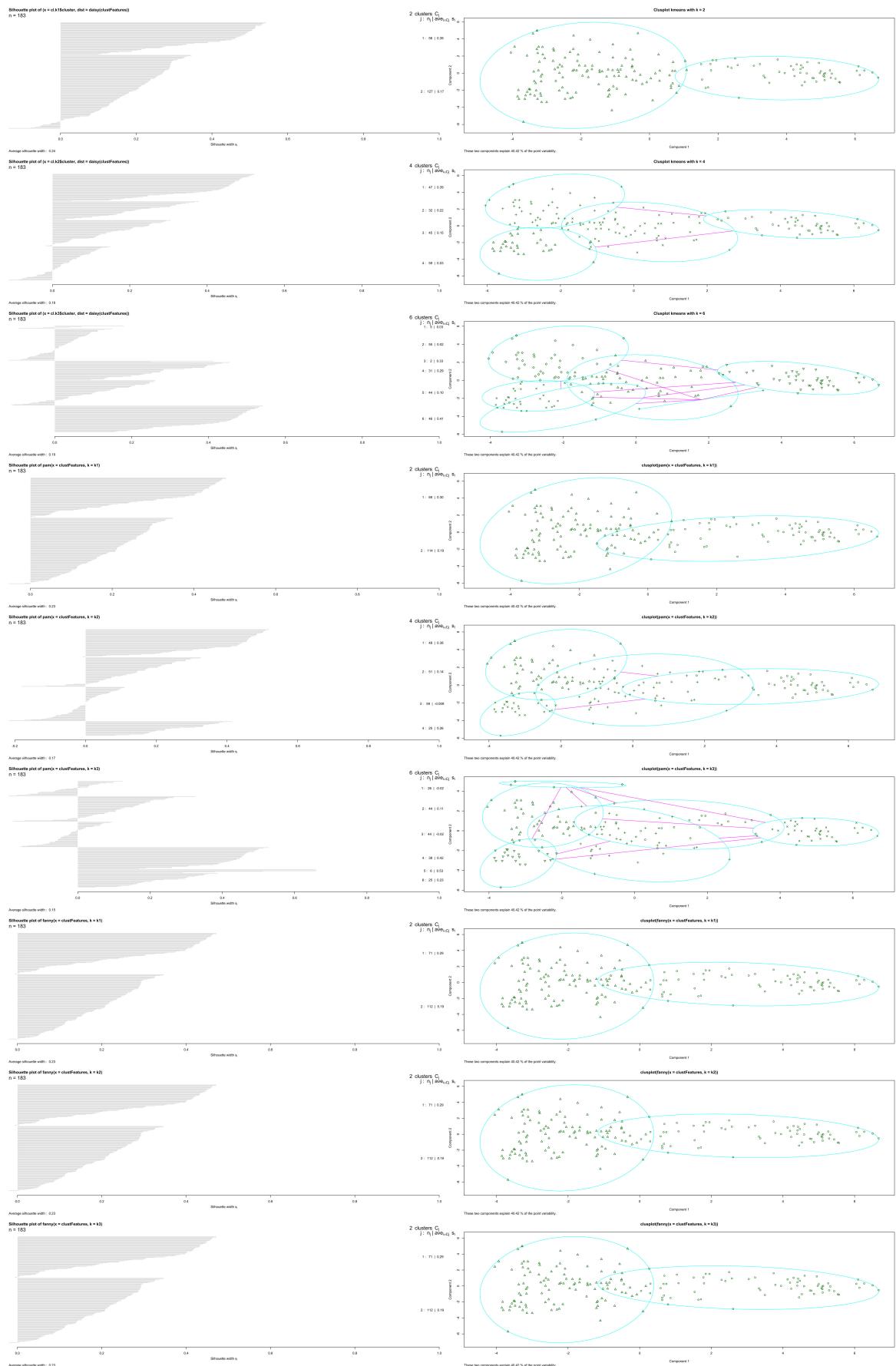


### 2.1.2 Met schalen

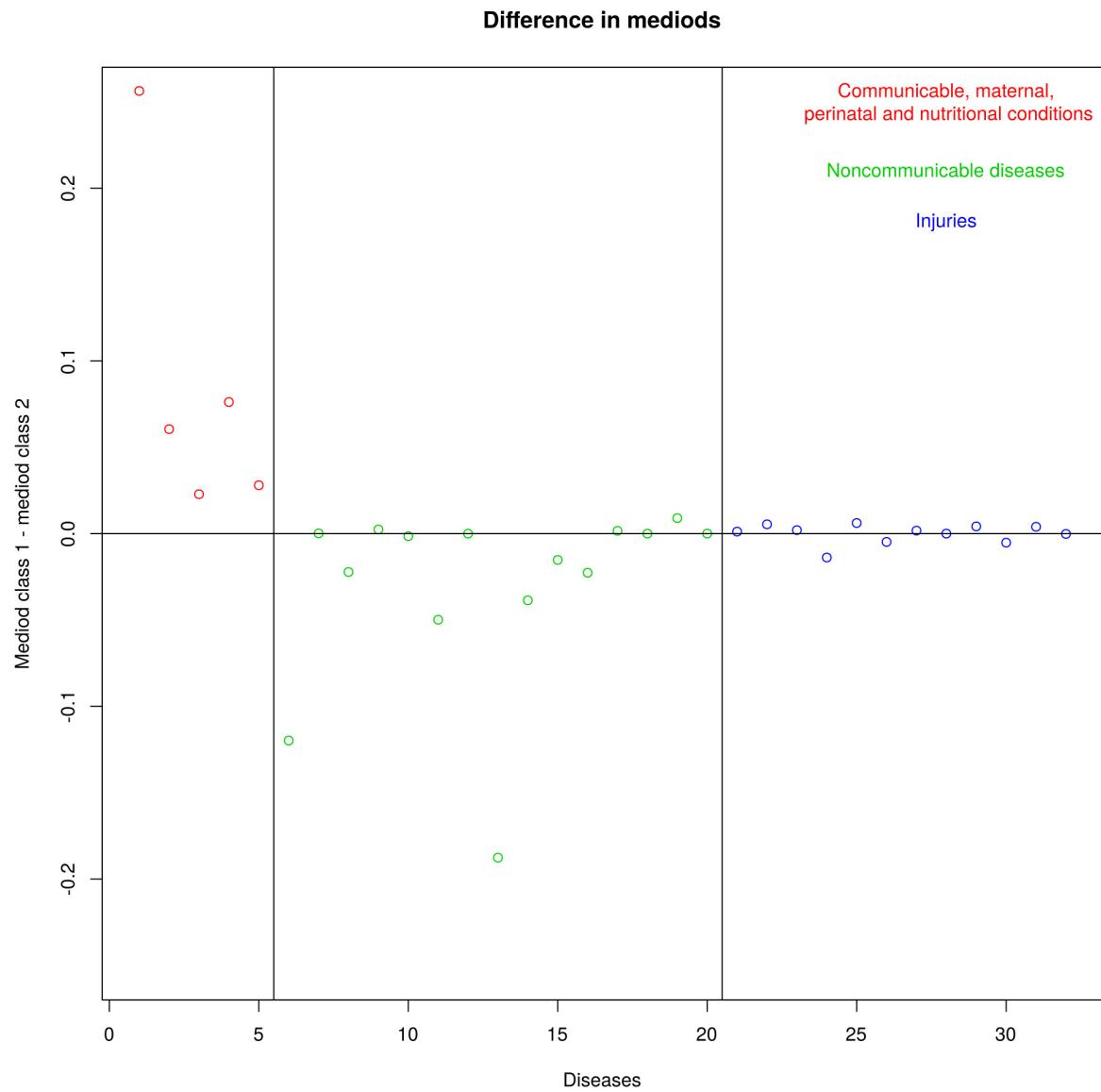


Figuur 4: Hierchical clustering dendograms met schalen





## 2.2 Beschrijving clusters



Figuur 7: De verschillen van de gemiddelden