



Verslag project 1

Academiejaar 2018-2019

Thomas Bamelis R0640219 & Michiel Jonckheere R0665594

Inhoudsopgave

1 Clustering	2
1.1 Zonder schalen	2
1.2 Met schalen	2
1.3 Beschrijving clusters	2
2 Principaalcomponentenanalyse	3
2.1 PC1	3
2.2 PC2	4
2.3 PC3	4
2.4 PC4	4
3 Multivariate Normaliteit	5
4 Classificatie	5
4.1 Regio	5
4.2 Ontwikkeling	5
5 Bijlage	7
5.1 Clustering	7
5.1.1 Zonder schalen	7
5.1.2 Met schalen	11
5.2 Beschrijving clusters	14
5.3 Principaalcomponentenanalyse	14

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 p7 en 2 op p8 in de bijlage 5. 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 silhouet plot en een clusplot. Zie figuur 3 op p9. Hieruit blijkt dat partitioning around mediods met 2 clusters het beste presteert met een silhouet 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 p11 en 5 op p12. Na dezelfde clustering algoritmes toegepast te hebben (figuur 6 op p13), is de best geobserveerde silhouet 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 gemiddelen van de eerste cluster, afgetrokken met de marginale gemiddelen van de tweede cluster. Zie figuur 7 op p14. 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 voorkomen 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 proportioneel opvallend hoger aantal communicable, maternal, perinatal and nutritional conditions bevatten en een proportioneel lager aantal noncommunicable diseases hebben tegenover de rest van de wereld.

2 Principaalcomponentenanalyse

Voor de principaalcomponentenanalyse wordt enkel gekeken naar de niet geschaalde gegevens. Dit zagen we al in de vorige sectie dat de geschaalde gegevens voor minder goede resultaten zorgden. Om voor de principaalcomponenten toch zeker te zijn dat de niet geschaalde gegevens hiervoor ook beter zijn, werd toch eens een vergelijking gedaan. In tabel 1 is deze vergelijking te zien. De eerste zes principaalcomponenten van de originele data verklaren ongeveer 95% van de variantie van de data. Bij de geschaalde gegevens verklaren de eerste zes slechts 60% van die variantie. Om aan 95% te komen zijn er bij de geschaalde data 23 principaalcomponenten nodig.

	PC1	PC2	PC3	PC4	PC5	PC6
Originiele data	68.9%	81.4%	86.4%	90.4%	93.1%	95.2%
Geschaalde data	29.7%	40.4%	46.5%	52.3%	56.5%	60.4%

Tabel 1: De cumulatieve proportie van de variantie van de eerste zes principaalcomponenten van zowel de originele als de geschaalde data.

We hebben besloten om hier verder te gaan met de eerste vier principaalcomponenten aangezien ze 90% van de variantie van de data beschrijven.

2.1 PC1

Voor de variabelen *Malignant Neoplasms* en *Cardiovascular Diseases* vinden we hoge positieve waarden, respectievelijk 0.33 en 0.66. De grootste negatieve waarde voor de eerste principaalcomponent is -0.62 van de variabele *Infections and Parasitic Diseases*. Dit betekent dus bij een hogere PC1 dat er meer doodsoorzaken zijn van *Malignant Neoplasms* en *Cardiovascular Diseases* en minder van *Infections and Parasitic Diseases*.

Als we dit bekijken voor de landen dan vinden we zeven landen met een lagere PC1 dan de rest. Deze landen zijn *Angola*, *Central African Republic*, *Kenya*, *Lesotho*, *Mozambique*, *South Sudan* en *Zambia*. Wat hier opvalt, is dat deze landen allemaal in Afrika liggen. Als we dan kijken naar de landen met een opvallend hogere PC1 waarde, dan zien we dat merendeel in Europa gelegen is. Om in het algemeen te kijken naar alle landen vergelijken we de landen met een positieve PC1 met de landen met een negatieve. Hieruit vinden we dat er 66% van de negatieve PC1 waarden uit Afrika komen en 93% van de positieve waarden komen uit de andere regio's verschillend van Afrika. We kunnen dus uit PC1 Afrika gaan onderscheiden van de andere regio's.

De ontwikkeling afleiden uit PC1 is deels mogelijk. Alle landen die ontwikkeld zijn of in een overgang zitten naar ontwikkeld, hebben allemaal een positieve PC1. Voor de ontwikkelingslanden heeft 50% een positieve en 50% een negatieve PC1. Als we dit dan koppelen aan de regio's kunnen we de ontwikkeling gaan afleiden uit de regio waar de negatieve PC1 waarden voorkomen. Dit betekent dus dat een land die in Afrika gelegen is en een negatieve PC1 waarde heeft, hoogstwaarschijnlijk een ontwikkelingsland zal zijn. Landen die in een andere regio dan Afrika liggen en een positieve PC1 hebben, zijn ofwel

ontwikkelde landen ofwel overgangslanden. We merken ook op dat de landen in Azië bijna allemaal ontwikkelingslanden zijn.

2.2 PC2

De meest positieve waarden voor PC2 zijn voor de variabelen *Cardiovascular Diseases* en *Infectious and Parasitic Diseases*, met de respectievelijke waarden 0.66 en 0.31. De meest negatieve waarde -0.58 hoort bij de variabele *Malignant Neoplasms*. Dit betekent bij een hogere PC2 waarde dat er minder doden door *Malignant Neoplasms* zijn en juist meer door *Cardiovascular Diseases* en *Infectious and Parasitic Diseases*.

Het grootste deel van de landen die een positieve PC2 waarde hebben, zijn landen die in Azië en Afrika liggen. Daar zijn er dus meer doden door *Cardiovascular Diseases* en *Infectious and Parasitic Diseases*. Over de negatieve PC2 waarden zijn er niet echt regio's uit af te leiden. Voornamelijk Amerikaanse en Europese landen hebben het grootste aandeel bij de negatieve waarden en minder opvallend bij de positieve. Wat er ook duidelijk te zien is, is dat er maar een paar Afrikaanse landen een negatieve waarde hebben. Om betere conclusies te trekken, zouden we naar de extremere positieve en negatieve waarden kunnen kijken, maar daarbij valt er nog altijd niets beters te concluderen.

Voor de ontwikkeling van de landen zien we dat bij een positieve PC2 het grootste deel bij de ontwikkelingslanden zit. Ontwikkelde landen hebben bijna allemaal een negatieve PC2, maar ook 33% van de ontwikkelingslanden hebben een negatieve PC2.

We besluiten hieruit dat de positieve PC2 waarden hoogstwaarschijnlijk landen die in Azië of Afrika liggen en ontwikkelingslanden zijn. Bij een negatieve PC2 waarde is de kans het grootst dat het een Amerikaans ontwikkelingsland of een Europees ontwikkeld land is. Als het een Aziatisch land is, dan is dat land een ontwikkelingsland bij een negatieve PC2 waarde.

2.3 PC3

Bij PC3 waarden vinden we voor volgende drie verandelen de meest in het oog springende waarden: *Malignant Neoplasms*, *Infectious and Parasitic Diseases* en *Diabetes Mellitus*. De waarden zijn respectievelijk -0.47, -0.51 en 0.53. Deze principaalcomponent kijkt dus naar het verschil tussen de eerste twee en de laatst genoemde variabelen.

Als we naar de regio's kijken van de landen, zien we dat de lagere PC3 waarden voornamelijk uit de regio's Europa en Afrika komen. De opvallend positieve waarden komen uit de regio's Amerika, Azië en Oceanië. Voor de ontwikkeling hebben alle ontwikkelde landen een negatieve PC3 waarde, voor de andere waarden is dit wat meer verdeeld en komen extrems in beide gevallen voor.

2.4 PC4

De vierde principaalcomponent vergelijkt de variabele *Diabetes Mellitus* met *Collective Violence and Legal Intervention* en *Neurological Conditions*. De PC4 waarden van deze drie variabelen zijn respectievelijk 0.76, -0.34 en -0.23.

Bij het kijken naar de regio's en ontwikkeling zien we dat deze over het algemeen rond 0 hangen, al zijn er een paar ontwikkelingslanden die toch relatief grote PC4 waarden hebben. Deze landen zijn twee landen uit Oceanië en drie Amerikaanse, nl. Micronesia en Tonga (eilanden van Oceanië) en Antigua en Barbuda, Grenada en Saint Vincent and the Grenadines (allemaal eilanden in de buurt van de Caribische Zee). Op deze eilanden is de PC4 het hoogst en zijn er daar meer doden door *Diabetes Mellitus* en minder door *Collective Violence and Legal Intervention* en *Neurological Conditions*.

In figuur 8 staan de plots van de PC1 waarden t.o.v. de PC2, PC3 en PC4 waarden. Hieruit zijn de meest getrokken conclusies ook zichtbaar adhv de kleurtjes en figuurtjes gebruikt voor de verschillende regio's en ontwikkelingen.

3 Multivariate Normaliteit

Om de multivariate normaliteit van de data na te gaan moeten we eerst en vooral kijken of alle variabelen wel univariaat normaal verdeeld zijn. De univariate testen tonen aan dat alle variabelen niet univariaat normaal verdeeld zijn. Hierdoor kan de data ook geen multivariate normaliteit hebben. Ook na het toepassen van de *logit* transformatie was geen enkele variabele normaal verdeeld.

We kunnen hieruit dus besluiten dat de gegevens niet multivariaat normaal verdeeld zijn en we in de volgende sectie over classificatie geen methodes moeten toepassen waarvoor deze eigenschap vereist is.

4 Classificatie

Bij de classificatie gaan we na in hoeverre het mogelijk is om de regio en ontwikkeling van een land te identificeren a.d.h.v. de doodsoorzaken. Er wordt gebruik gemaakt van de lineaire discriminantmethode en *k-nearest neighbours* methode. Het was niet mogelijk van de kwadratische discriminantmethode toe te passen op de data omdat sommige groepen te klein zijn en er niet genoeg data over is. KLOPT DIT???????

4.1 Regio

Om classificatie toe te passen op de regio van landen, zien we dat bij elke methode de *actual error rate* vrij groot is (tussen de 25% en 30%). Dit betekent dat het moeilijk is om effectief de regio te bepalen van een land als de doodsoorzaken gegeven zijn. 46 van de 183 landen worden verkeerd ingedeeld bij de lineaire discriminantmethode. Als we dan kijken naar de *k-nearest neighbours* methode, dan zien we dat deze *AER* iets groter is dan bij de LDA. Er werd gekeken naar $k = \{2, 3, 4, 5\}$. De beste methode voor het classificeren volgens de regio's is de lineaire discriminantmethode.

Het is opmerkelijk dat alle Europese landen juist ingedeeld worden als een Europees land. Er zijn wel relatief veel landen uit andere regio's die ook als Europees land worden ingedeeld.

4.2 Ontwikkeling

Om de ontwikkeling van de landen na te gaan met behulp van de doodsoorzaken, verkregen we betere error ratio's. Deze lagen allemaal dicht bij elkaar rond de 12%. De beste *AER* verkregen we bij de *8-nearest neighbours* methode. Hierbij werden 21 van de 183 landen verkeerd ingedeeld volgens hun ontwikkeling. Deze landen komen voornamelijk uit Europa, Azië en Amerika. De meeste landen die verkeerd ingedeeld worden zijn ontwikkelingslanden of overgangslanden en worden als ontwikkeld land geclassificeerd. Ontwikkelde landen die verkeerd zijn ingedeeld worden als een overgangsland ingedeeld.

In de data zit er één speciaal land (Congo) waarvan de ontwikkeling niet gekend is. Dit land wordt ingedeeld als een ontwikkelingsland. Aangezien Congo in Afrika ligt en we uit de gegeven data konden opmerken dat alle landen uit Afrika ontwikkelingslanden zijn, kunnen we hier wel concluderen dat Congo een ontwikkelingsland is.

Besluit

Referenties

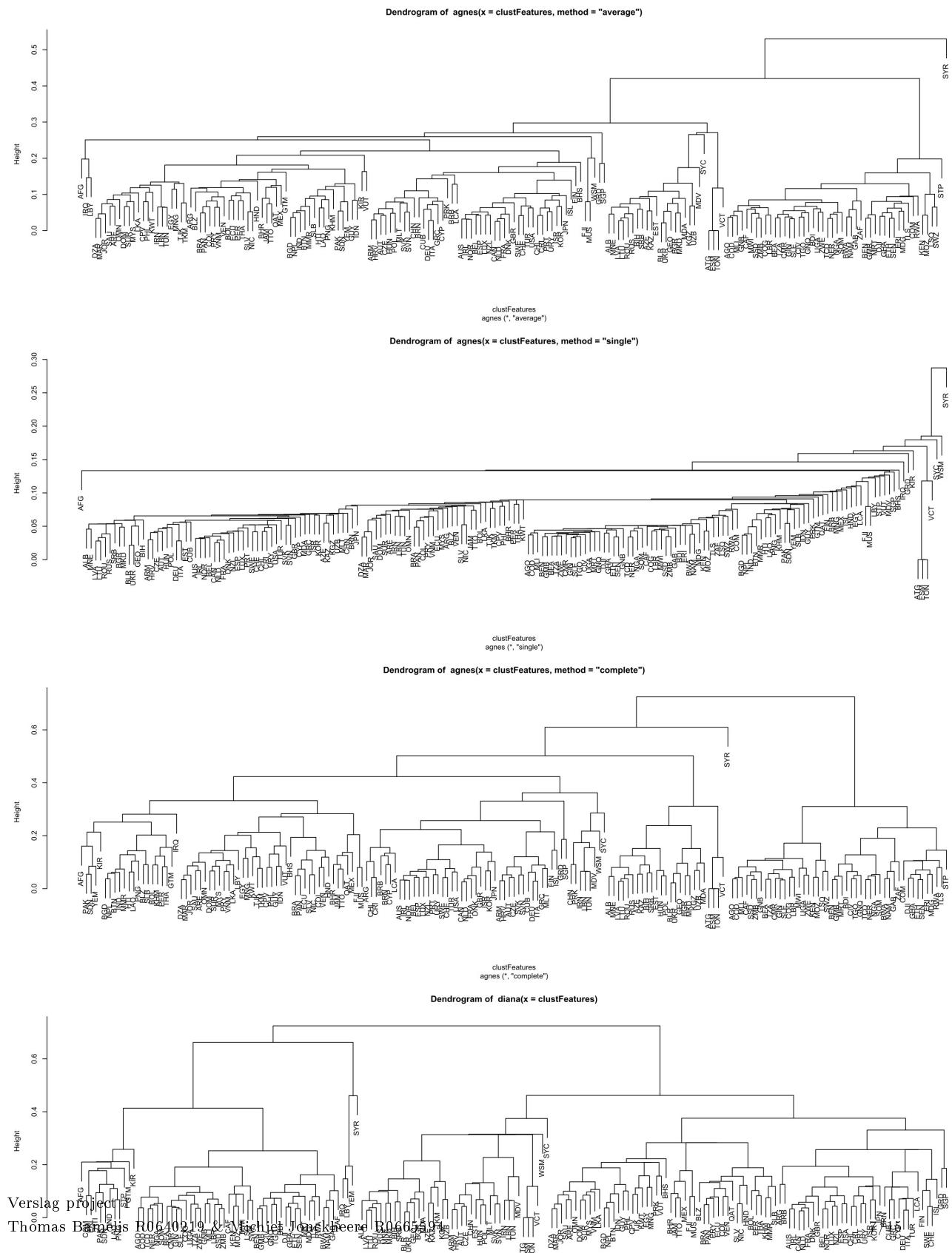
- [1] Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
- [2] Country classification, june 2018. Geneva, United Nations Conference on Trade and Development; 2018.

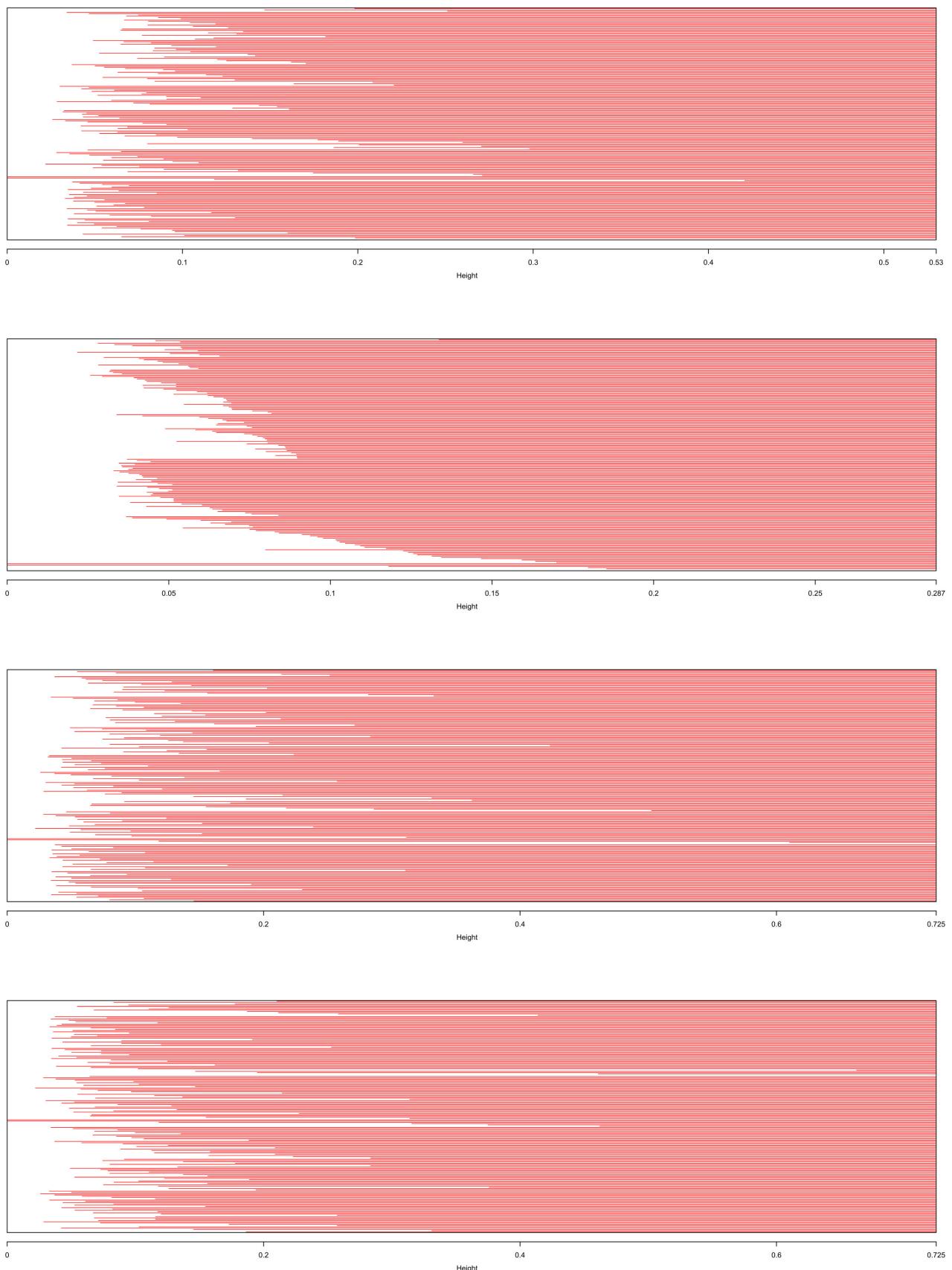


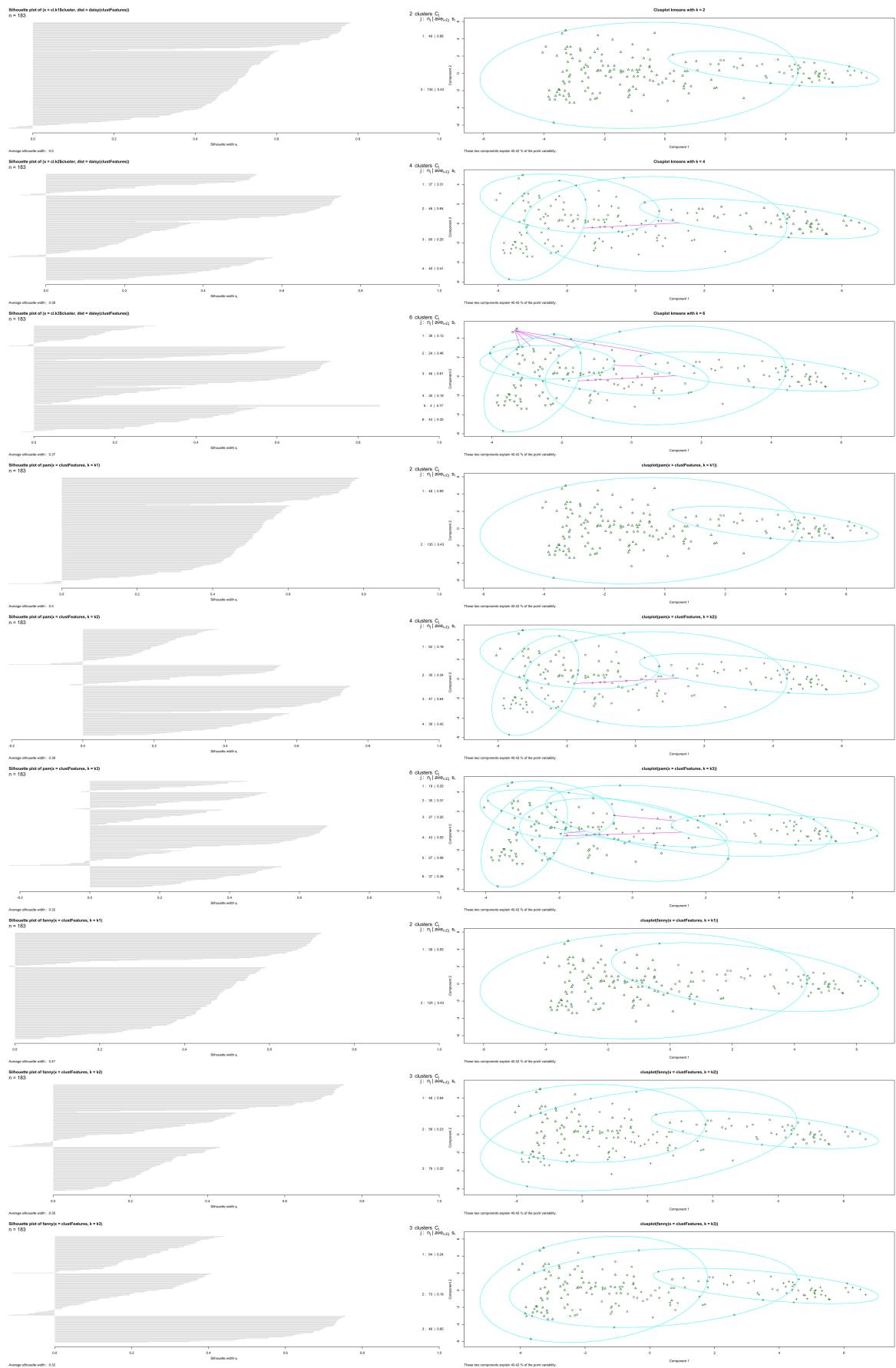
5 Bijlage

5.1 Clustering

5.1.1 Zonder schalen

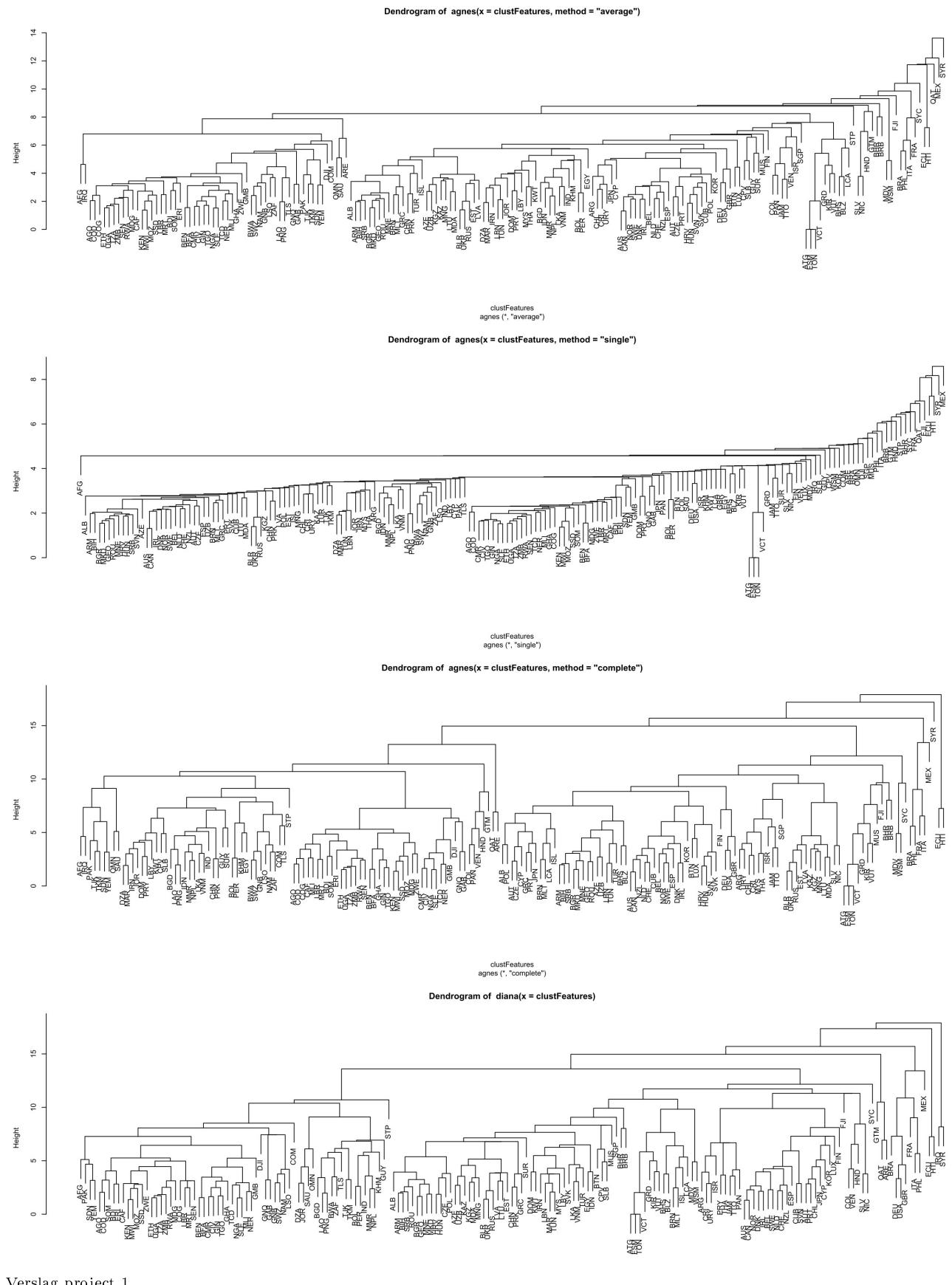




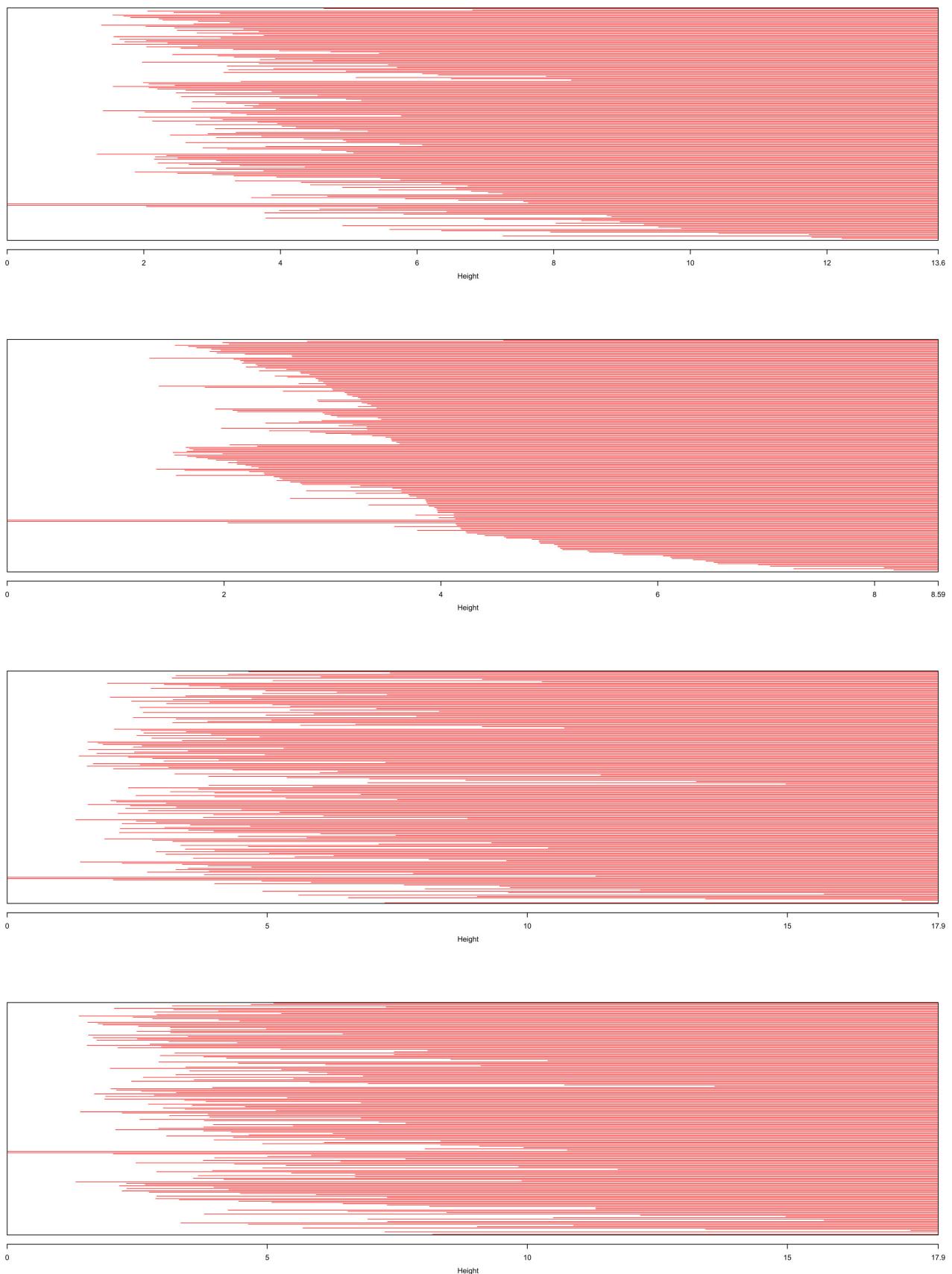


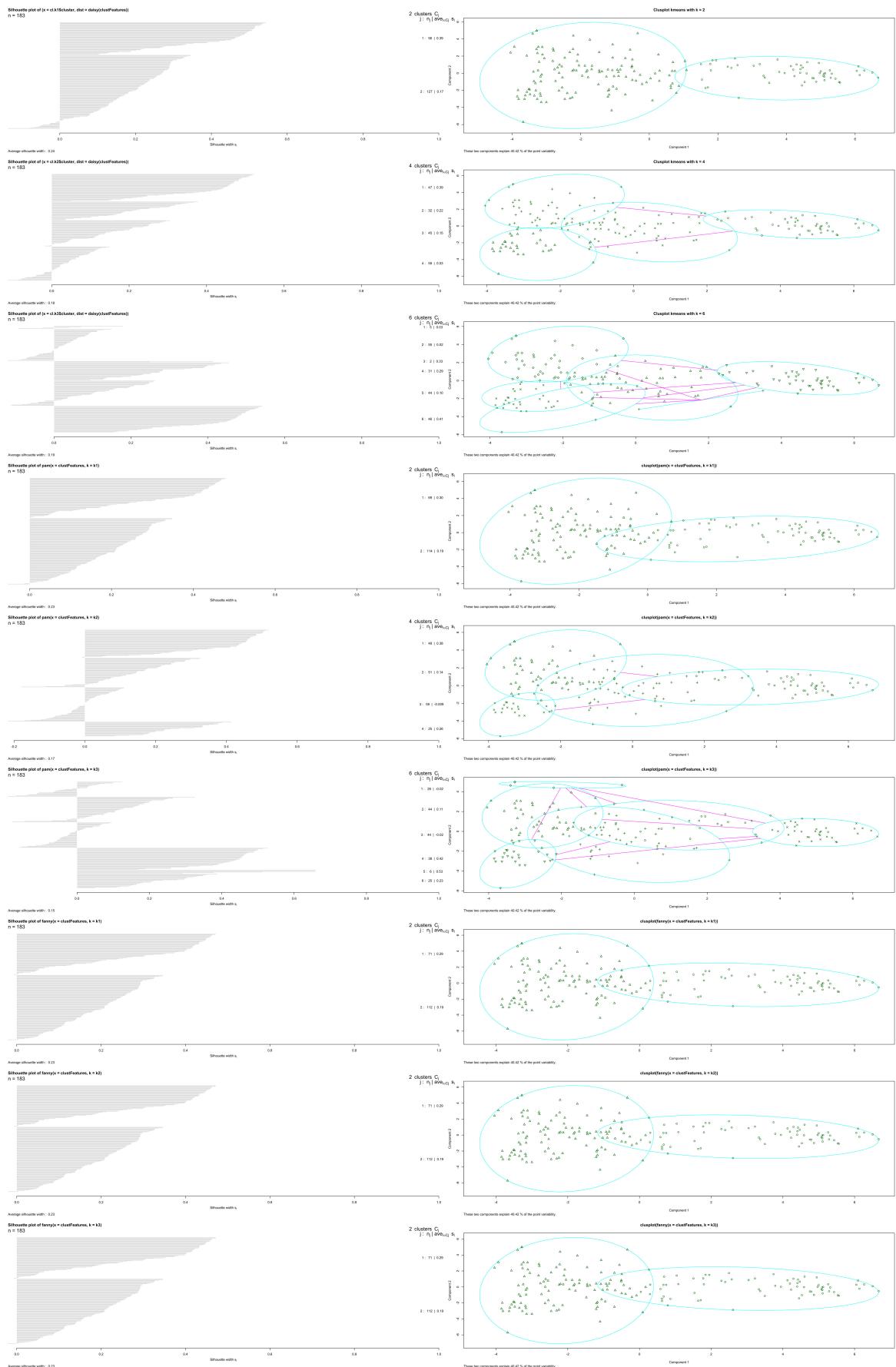


5.1.2 Met schalen

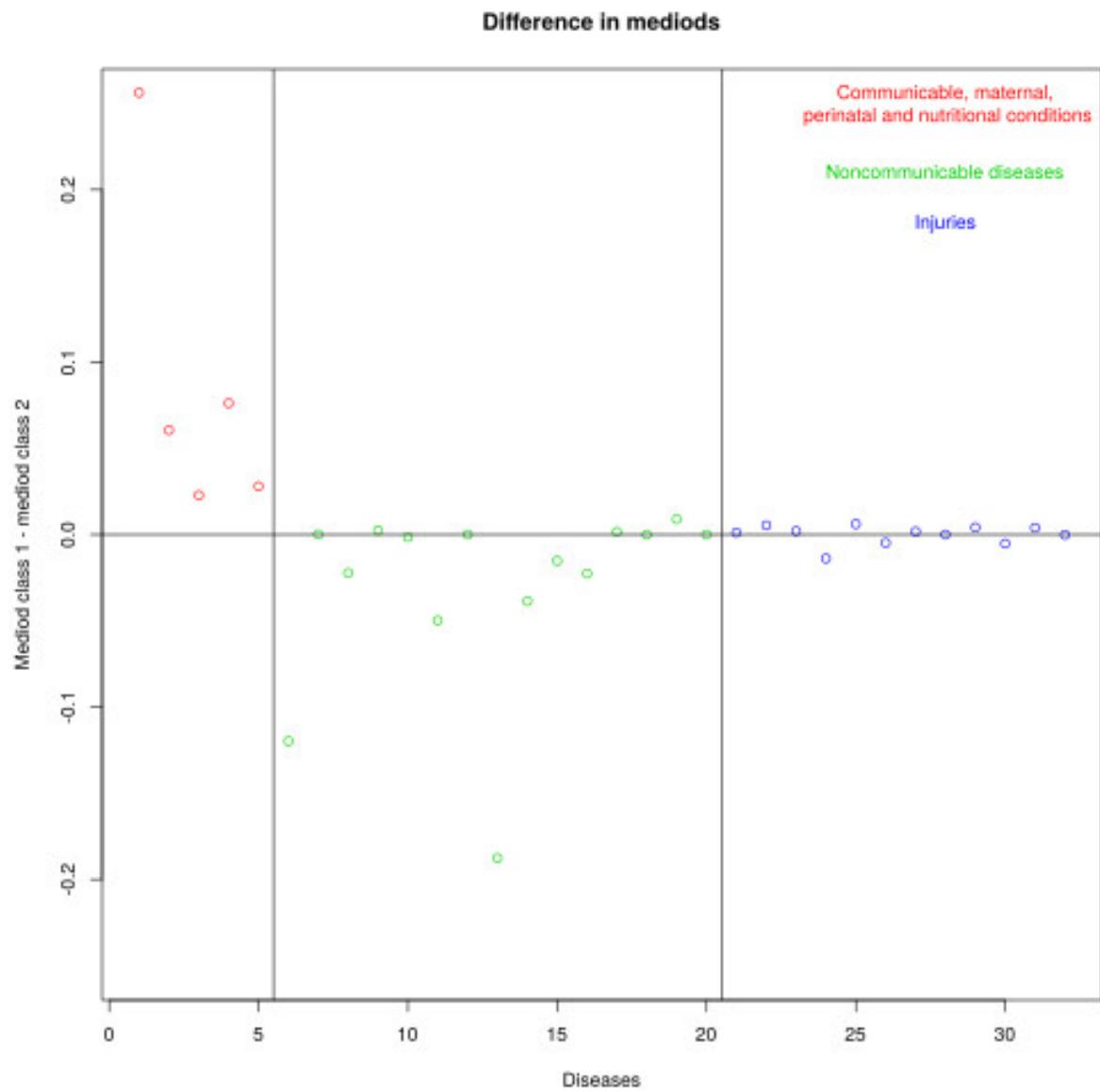


Figuur 4: Hierarchical clustering dendograms met schalen



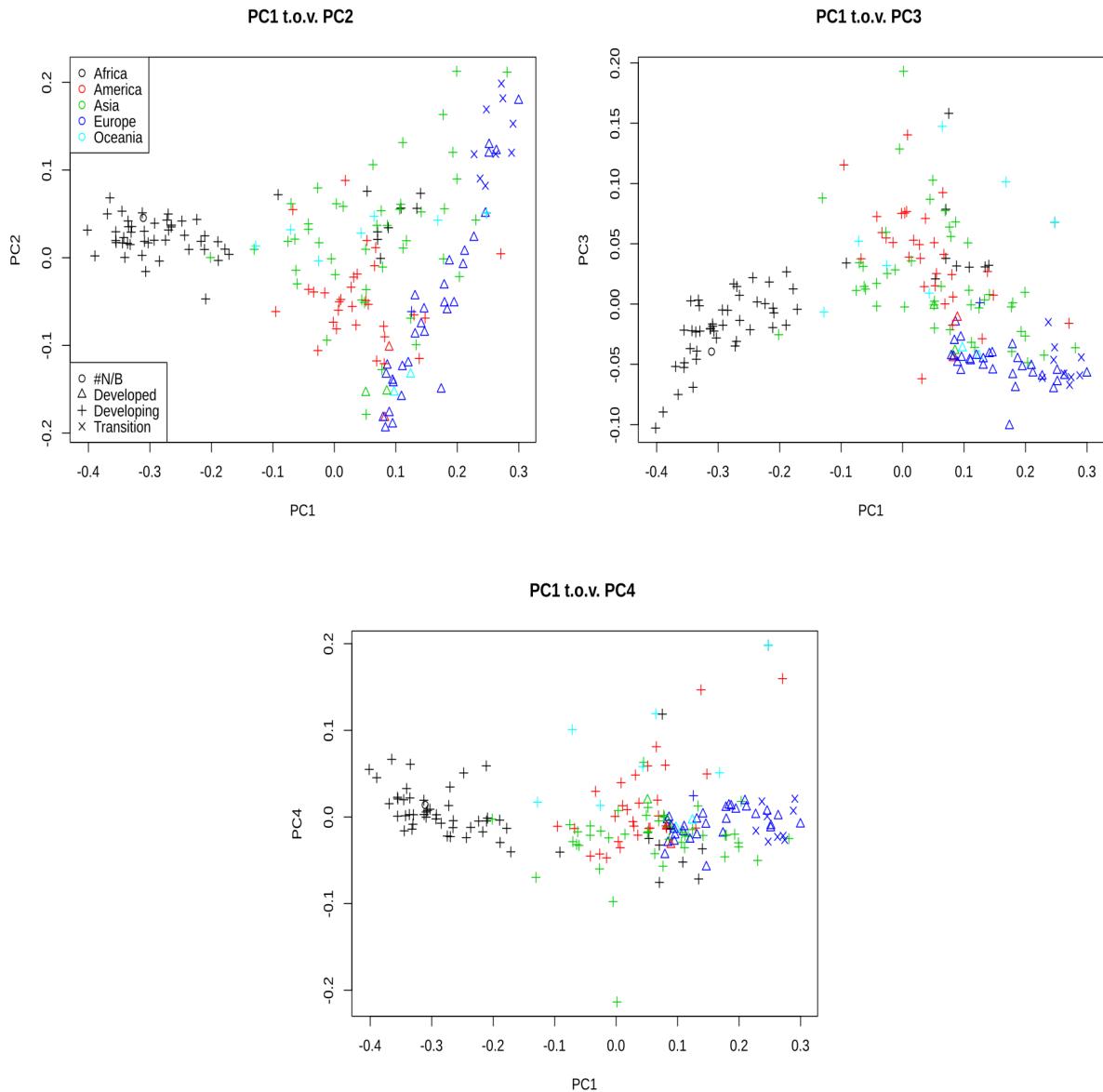


5.2 Beschrijving clusters



Figuur 7: De verschillen van de gemiddelden

5.3 Principaalcomponentenanalyse



Figuur 8: De eerste principaalcomponent vergeleken met de andere drie belangrijkste principaalcomponenten