



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
2 Bijlage	2
2.1 Clustering	2
2.1.1 Zonder schalen	2
2.1.2 Met schalen	2

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].

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 dissimiliteden: group average, nearest neighbour en furthest neighbour, in die volgorde met daarna divisive analysis. Zie figuren 1 en 2 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. 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 en 5 Na dezelfde clustering algoritmes toegepast te hebben (figuur 6), 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

Besluit

TODO

Referenties

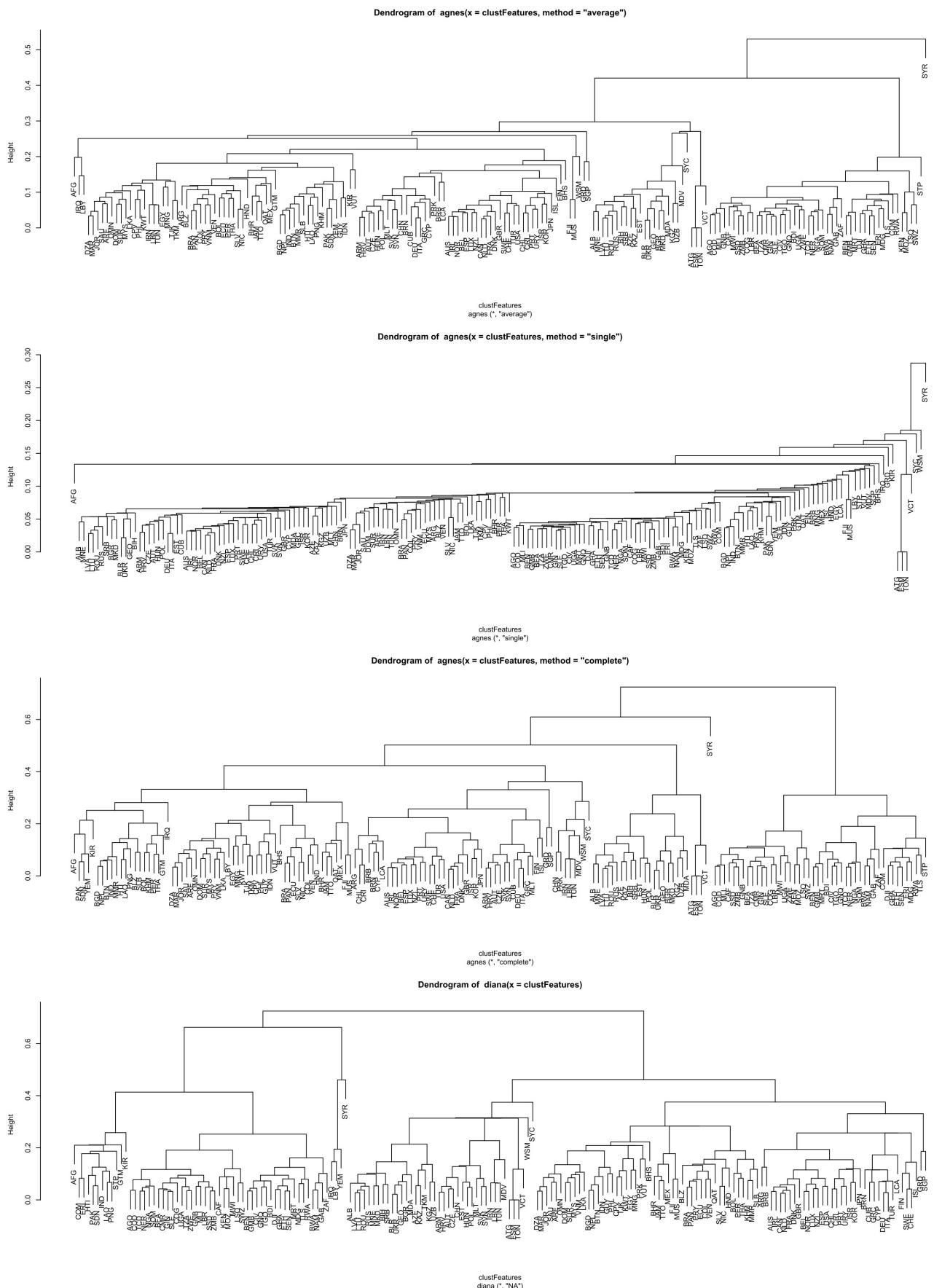
- [1] Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
- [2] Country classification, june 2018. Geneva, United Nations Conference on Trade and Development; 2018.

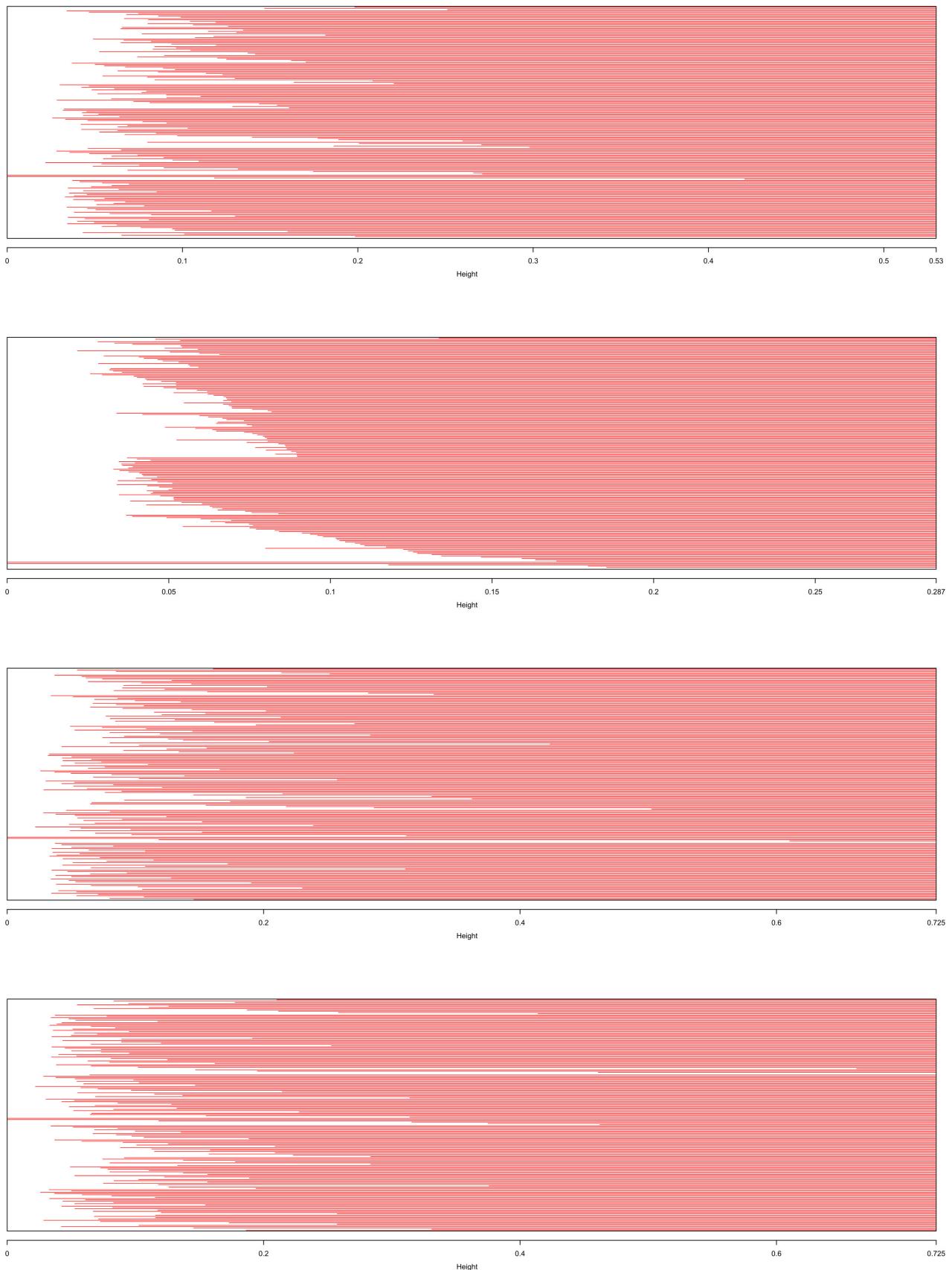
2 Bijlage

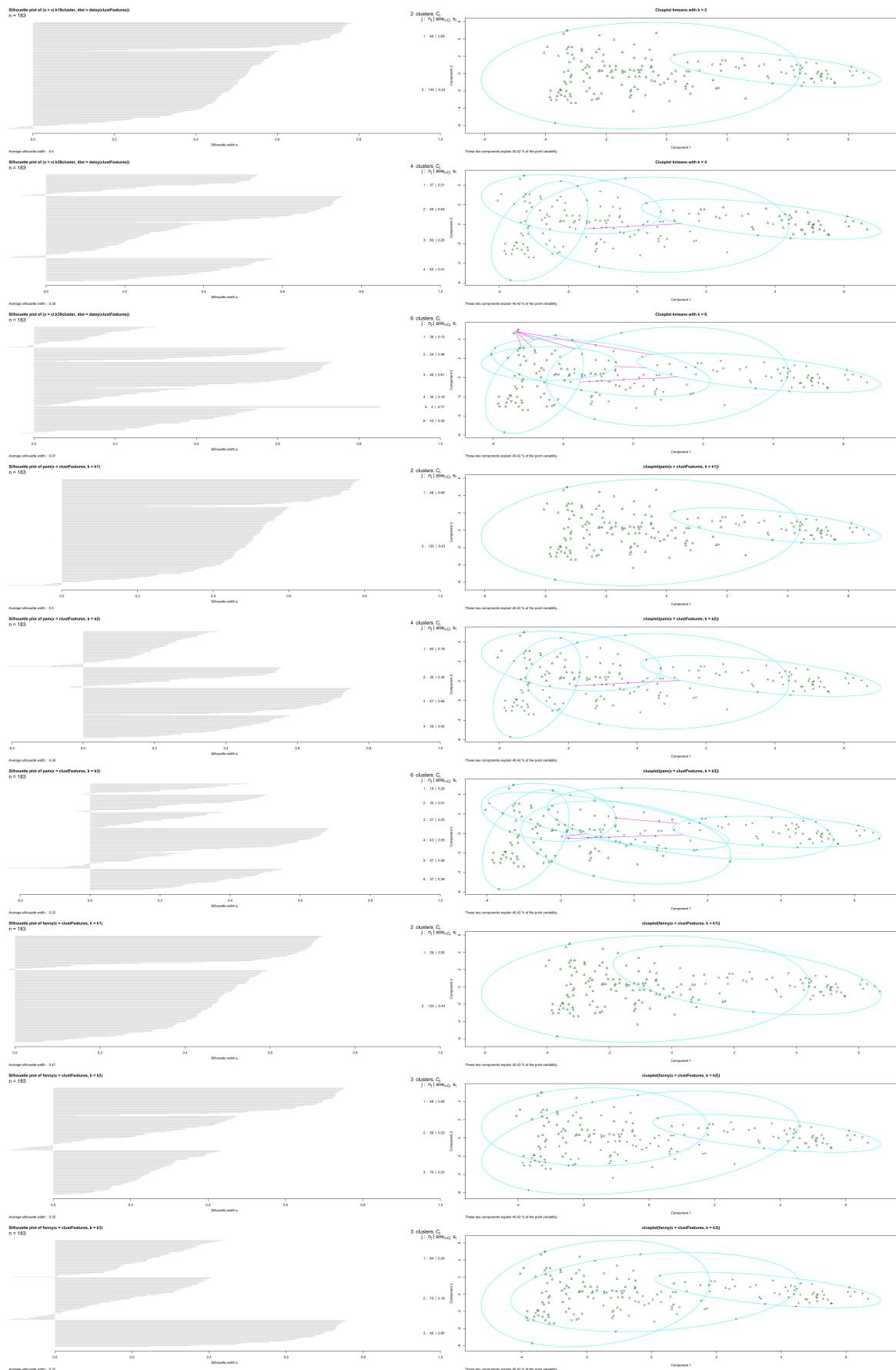
2.1 Clustering

2.1.1 Zonder schalen

2.1.2 Met schalen







Figuur 3: Clustering evaluatie

