

Lösung zu Kapitel 12: Beispiel 1

Wir haben hierarchische Verfahren anhand des Datensatzes mit den Demografiekennzahlen (Datenfile [demographie.csv](#)) besprochen.

- Führen Sie eine hierarchische Clusteranalyse mit den Variablen [Fertilityrate](#) und [Annualnetmigrationrate](#) ohne Standardisierung durch.
- Führen Sie dieselbe Analyse nur mit [Annualnetmigrationrate](#) aus.
- Vergleichen Sie die beiden Ergebnisse.

Nach der Datenaufbereitung (Weglassen von Beobachtungen mit fehlenden Werten) rufen wir ein hierarchisches Clusterverfahren auf. Ein Dendrogramm gibt den besten Überblick über den Clusterbildungsprozess.

R

```
> library("cluster")
> demog <- read.csv2("demographie.csv")
> attach(demog)
> demog1 <- na.omit(demog[1:4])
> detach(demog)
> clust_2var <- agnes(demog1[3:4], stand = FALSE)
> pltree(clust_2var, main = "Average-Linkage mit zwei Variablen",
+       labels = demog1$Code)
```

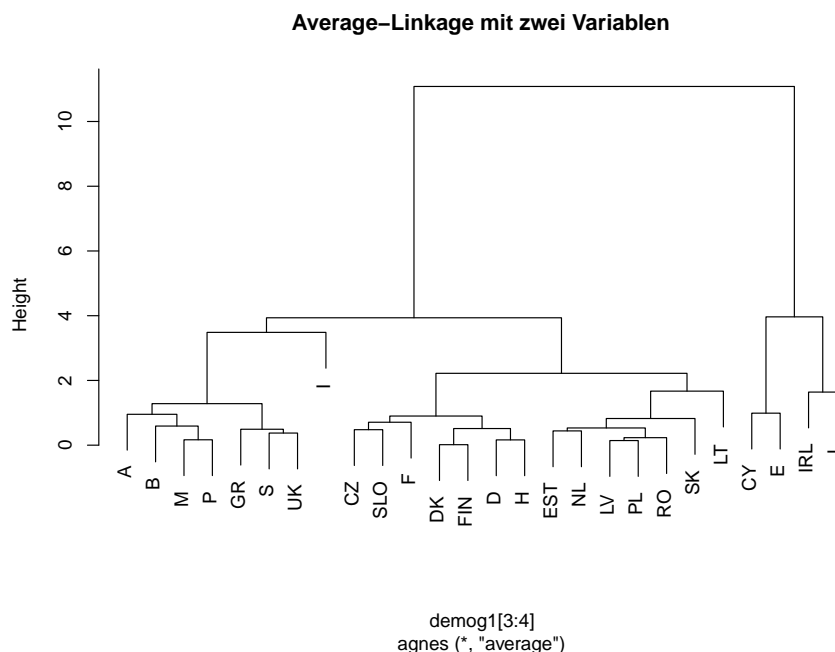


Abbildung 1: Dendrogramm für Clustern mit zwei Variablen

Im Vergleich dazu die Clusterbildung, wenn nur die Variable [Annualnetmigrationrate](#) eingesetzt wird:

R

```
> clust_1var <- agnes(demog1[4], stand = FALSE)
> pltree(clust_1var, main = "Average-Linkage mit einer Variablen",
+       labels = demog1$Code)
```

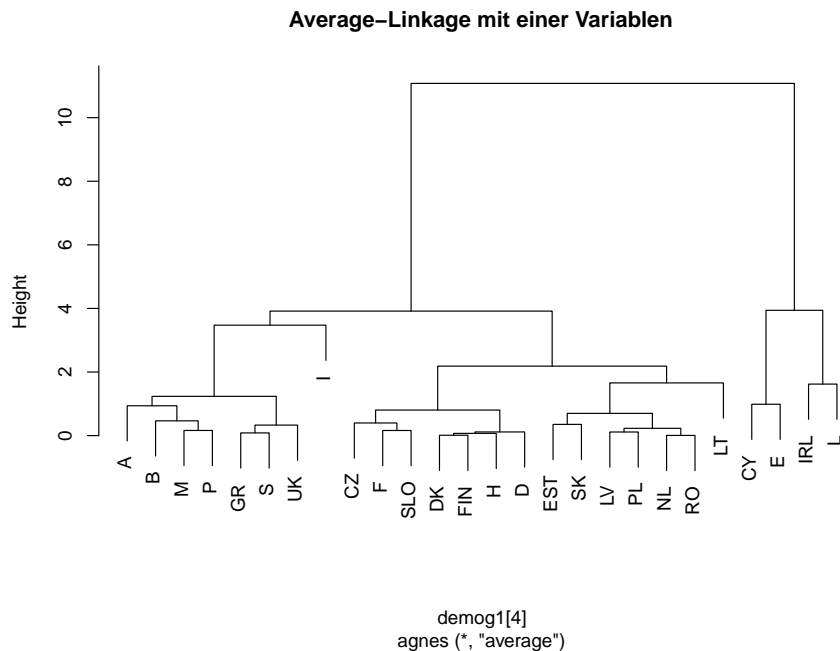


Abbildung 2: Dendrogramm für Clustern mit einer Variablen

Die Dendrogramme sind sehr ähnlich. Einen Vergleich der Clusterbildung kann etwa dadurch anstellen, dass man für eine gegebene Clusterzahl untersucht, wo die Beobachtungen beim einen und beim anderen Verfahren liegen. Dazu speichern wir die Clusterzugehörigkeit der beiden Varianten (in den Variablen `clust_1var_c` und `clust_2var_c`) und zwar bei vier Clustern, die sich nach dem Dendrogramm als gute Clusteranzahl erweisen. Danach erstellen wir eine Kreuztabelle der beiden Clusterzugehörigkeiten.

R

```
> clust_2var_c <- cutree(clust_2var, 4)
> clust_1var_c <- cutree(clust_2var, 4)
> table(clust_1var_c, clust_2var_c)
```

```
      clust_2var_c
clust_1var_c  1  2  3  4
1      8  0  0  0
2      0  2  0  0
3      0  0 14  0
4      0  0  0  2
```

Die Kreuztabelle zeigt, dass die Beobachtungen bei beiden Varianten jeweils in dieselben Cluster eingeteilt wurden.

Die Erklärung dafür ist, dass von den zwei Variablen **Annualnetmigrationrate** eine weit höhere Streuung ($s = 4.6726$) aufweist als **Fertilityrate** ($s = 0.2376$). Das bedeutet bei einer Berechnung

einer Distanzmatrix ohne Standardisierung, dass die Variable mit stärkerer Streuung die Werte dieser Matrix stärker beeinflusst. Sind die Werte – so wie hier – stark unterschiedlich, kann eine Variable für die Clusterbildung nahezu keine Bedeutung hat.