Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Prof. Dr. Rainer Stollhoff

Unsupervised Learning
Dimensionsreduktion

Gliederung

- Dimensionsreduktion
  - Motivation
  - PCA: Principal Component Analysis / Hauptkomponentenanalyse
  - Manifold Learning
    - Multidimensionale Skalierung
    - t-SNE: t-distributed Stoachtic Neighbourhood Embedding / t-verteilte stochastische Nachbarschaftseinbettung

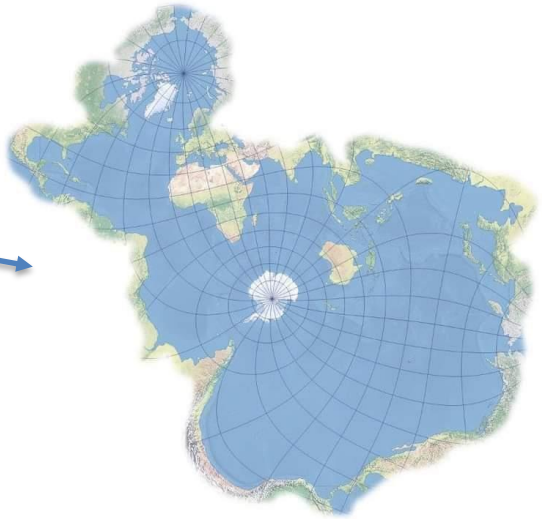
- Clustering
  - Motivation
  - k-Means-Clustering
  - Hierarchisches Clustering

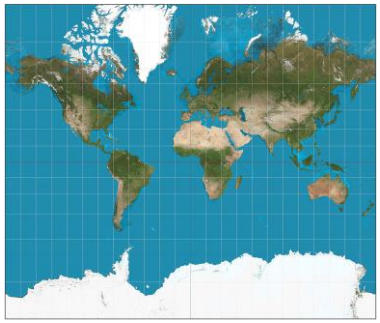
Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

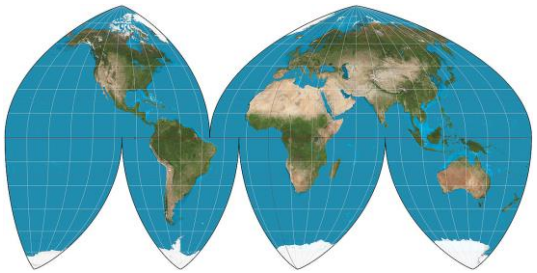
Prof. Dr. Rainer Stollhoff

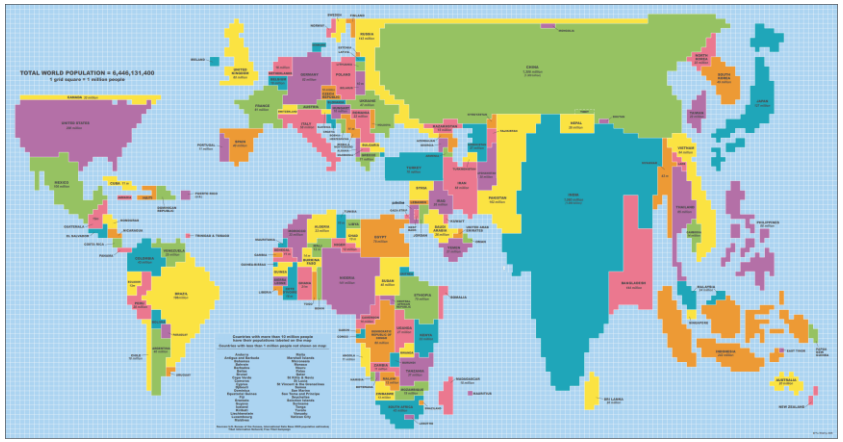
Unsupervised Learning
Dimensionsreduktion

Unsupervised Learning
- Metrische Projektionen





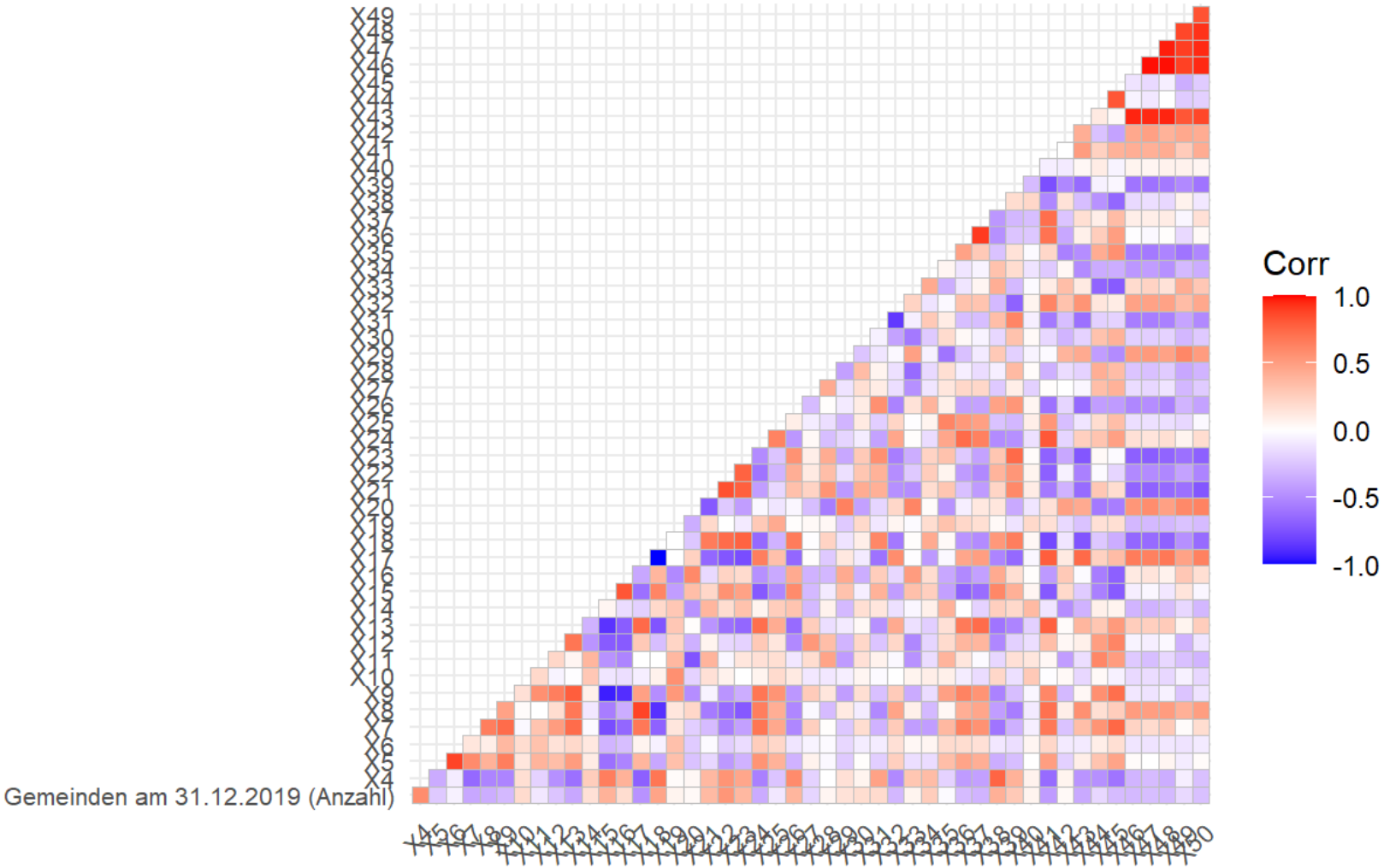




Strukturdaten Bundestagswahl 2021

```
##  [1] "Land"
##  [2] "Wahlkreis-Name"
##  [3] "Gemeinden am 31.12.2019 (Anzahl)"
##  [4] "Fläche am 31.12.2019 (km²)"
##  [5] "Bevölkerung am 31.12.2019 - Insgesamt (in 1000)"
##  [6] "Bevölkerung am 31.12.2019 - Deutsche (in 1000)"
##  [7] "Bevölkerung am 31.12.2019 - Ausländer/-innen (%)"
##  [8] "Bevölkerungsdichte am 31.12.2019 (EW je km²)"
##  [9] "Zu- (+) bzw. Abnahme (-) der Bevölkerung 2019 - Geburtensaldo (je 1000 EW)"
## [10] "Zu- (+) bzw. Abnahme (-) der Bevölkerung 2019 - Wanderungssaldo (je 1000 EW)"
## [11] "Alter von ... bis ... Jahren am 31.12.2019 - unter 18 (%)"
## [12] "Alter von ... bis ... Jahren am 31.12.2019 - 18-24 (%)"
## [13] "Alter von ... bis ... Jahren am 31.12.2019 - 25-34 (%)"
## [14] "Alter von ... bis ... Jahren am 31.12.2019 - 35-59 (%)"
## [15] "Alter von ... bis ... Jahren am 31.12.2019 - 60-74 (%)"
## [16] "Alter von ... bis ... Jahren am 31.12.2019 - 75 und mehr (%)"
## [17] "Bodenfläche nach Art der tatsächlichen Nutzung am 31.12.2019 - Siedlung und Verkehr (%)"
## [18] "Bodenfläche nach Art der tatsächlichen Nutzung am 31.12.2019 - Vegetation und Gewässer (%)"
## [19] "Fertiggestellte Wohnungen 2019 (je 1000 EW)"
## [20] "Bestand an Wohnungen am 31.12.2019 - insgesamt (je 1000 EW)"
## [21] "Wohnfläche am 31.12.2019 (je Wohnung)"
## [22] "Wohnfläche am 31.12.2019 (je EW)"
## [23] "PKW-Bestand am 01.01.2020 - PKW insgesamt (je 1000 EW)"
## [24] "PKW-Bestand am 01.01.2020 - PKW mit Elektro- oder Hybrid-Antrieb (%)"
## [25] "Unternehmensregister 2018 - Unternehmen insgesamt (je 1000 EW)"
## [26] "Unternehmensregister 2018 - Handwerksunternehmen (je 1000 EW)"
## [27] "Schulabgänger/-innen beruflicher Schulen 2019"
## [28] "Schulabgänger/-innen allgemeinbildender Schulen 2019 - insgesamt ohne Externe (je 1000 EW)"
## [29] "Schulabgänger/-innen allgemeinbildender Schulen 2019 - ohne Hauptschulabschluss (%)"
## [30] "Schulabgänger/-innen allgemeinbildender Schulen 2019 - mit Hauptschulabschluss (%)"
## [31] "Schulabgänger/-innen allgemeinbildender Schulen 2019 - mit mittlerem Schulabschluss (%)"
## [32] "Schulabgänger/-innen allgemeinblldender Schulen 2019 - mit allgemeiner und Fachhochschulreife (%)"
## [33] "Kindertagesbetreuung am 01.03.2020 - Betreute Kinder unter 3 Jahre (Betreuungsquote)"
## [34] "Kindertagesbetreuung am 01.03.2020 - Betreute Kinder 3 bis unter 6 Jahre (Betreuungsquote)"
## [35] "Verfügbares Einkommen der privaten Haushalte 2018 (EUR je EW)"
## [36] "Bruttoinlandsprodukt 2018 (EUR je EW)"
## [37] "Sozialversicherungspflichtig Beschäftigte am 30.06.2020 - insgesamt (je 1000 EW)"
## [38] "Sozialversicherungspflichtig Beschäftigte am 30.06.2020 - Land- und Forstwirtschaft, Fischerei (%)"
## [39] "Sozialversicherungspflichtig Beschäftigte am 30.06.2020 - Produzierendes Gewerbe (%)"
## [40] "Sozialversicherungspflichtig Beschäftigte am 30.06.2020 - Handel, Gastgewerbe, Verkehr (%)"
## [41] "Sozialversicherungspflichtig Beschäftigte am 30.06.2020 - Öffentliche und private Dienstleister (%)"
## [42] "Sozialversicherungspflichtig Beschäftigte am 30.06.2020 - Übrige Dienstleister und \"\"ohne Angabe
\"\" (%)"
## [43] "Empfänger/-innen von Leistungen nach SGB II  Oktober 2020 -  insgesamt (je 1000 EW)"
## [44] "Empfänger/-innen von Leistungen nach SGB II  Oktober 2020 -  nicht erwerbsfähige Hilfebedürftige (%)"
## [45] "Empfänger/-innen von Leistungen nach SGB II  Oktober 2020 -  Ausländer/-innen (%)"
## [46] "Arbeitslosenquote Februar 2021 - insgesamt"
## [47] "Arbeitslosenquote Februar 2021 - Männer"
## [48] "Arbeitslosenquote Februar 2021 - Frauen"
## [49] "Arbeitslosenquote Februar 2021 - 15 bis 24 Jahre"
## [50] "Arbeitslosenquote Februar 2021 - 55 bis 64 Jahre"
```

Kombinationen mit viel Redundanz



Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Prof. Dr. Rainer Stollhoff

Unsupervised Learning

Dimensionsreduktion

Hauptkomponentenanalyse - PCA

Hauptkomponentenanalyse - Idee

- Suche nach (Linear-)Kombinationen von Variablen, die
  - eine große Variabilität zwischen den Beobachtungen aufweisen und
  - gegenseitig unkorreliert sind

PCA - iterative Konstruktion

Ableitung - iterative Konstruktion

- Hauptkomponenten sind lineare Kombinationen von Merkmalen

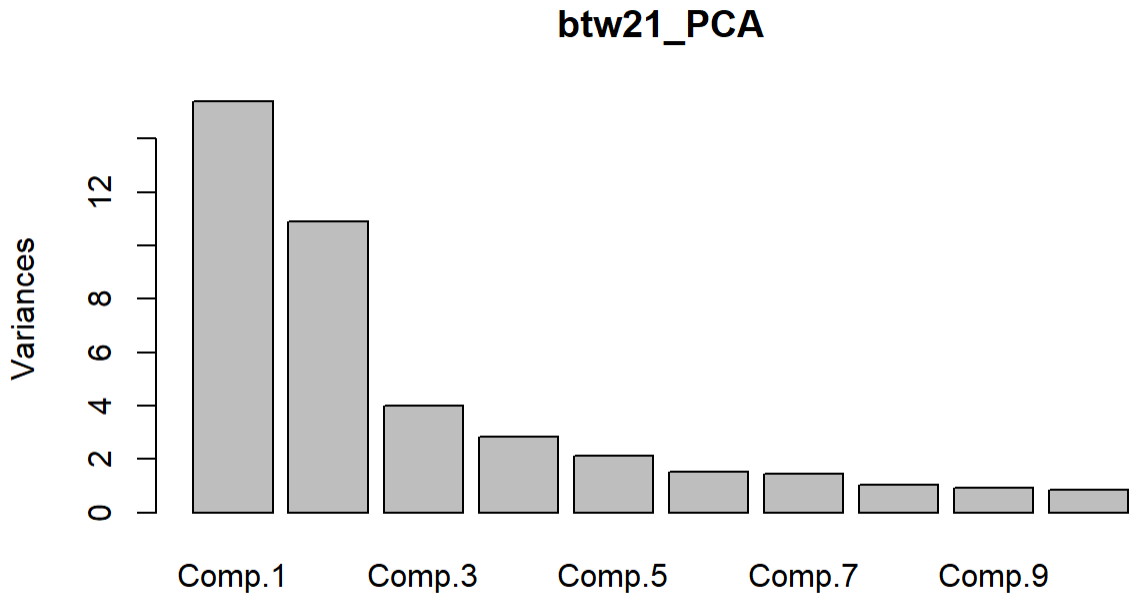
$$PCA_i = \beta_{i,1} X_1 + \cdots + \beta_{i,m} X_m$$

- Die Hauptkomponenten iterativ konstruieren
- Wählen Sie die Richtung im Datenraum mit der größten Varianz als erste Hauptkomponente

$$\begin{aligned} \sigma^2(PCA_1) &= \sigma^2\left(\beta_{1,1} X_1 + \cdots + \beta_{1,m} X_m\right) \\ &= \beta_{1,1}^2 \sigma^2(X_1) + \cdots + \beta_{1,m} \sigma^2(X_m) \end{aligned}$$

- Wählen Sie die zweite Hauptkomponente
  - durch die Mitte der Daten hindurch
  - orthogonal oder unkorreliert zur ersten Hauptkomponente
  - mit der größten Varianz
- Alle anderen Hauptkomponenten auswählen
  - durch die Mitte der Daten hindurch
  - orthogonal oder unkorreliert zu den vorherigen Hauptkomponenten
  - mit der größten Varianz

PCA - Varianzzerlegung

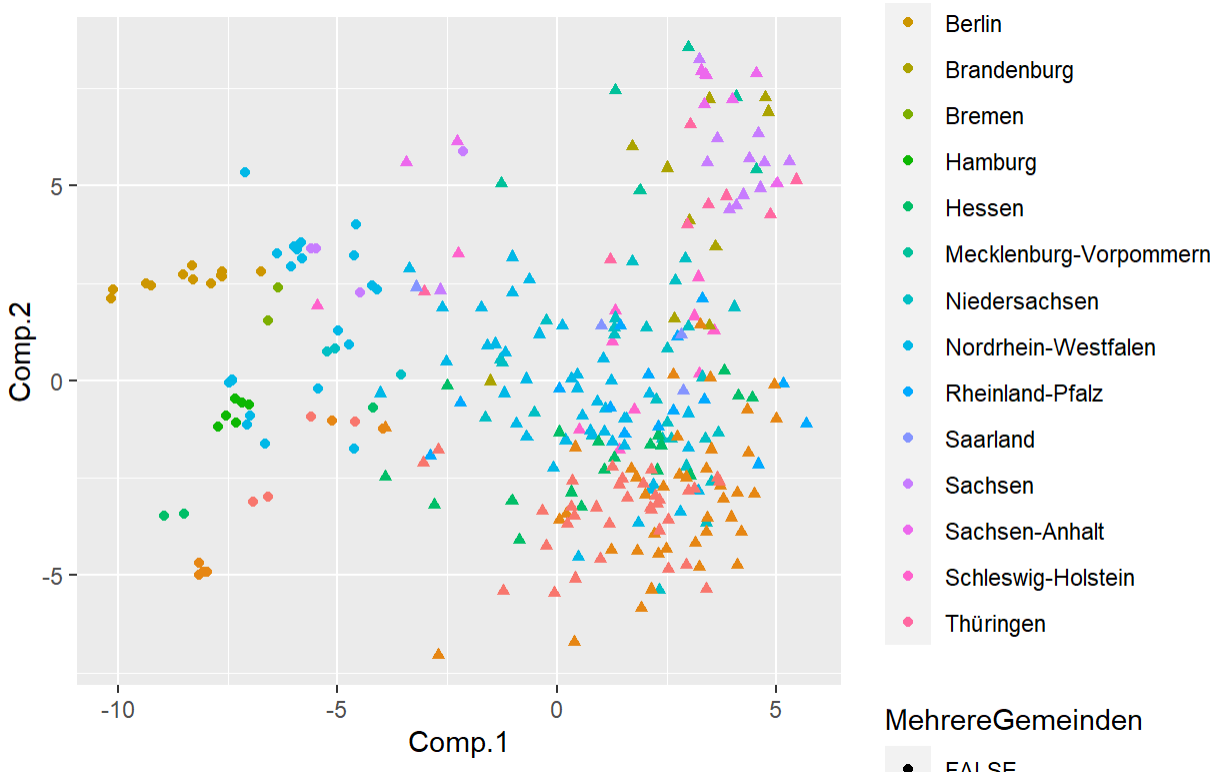


PCA - Komponenten

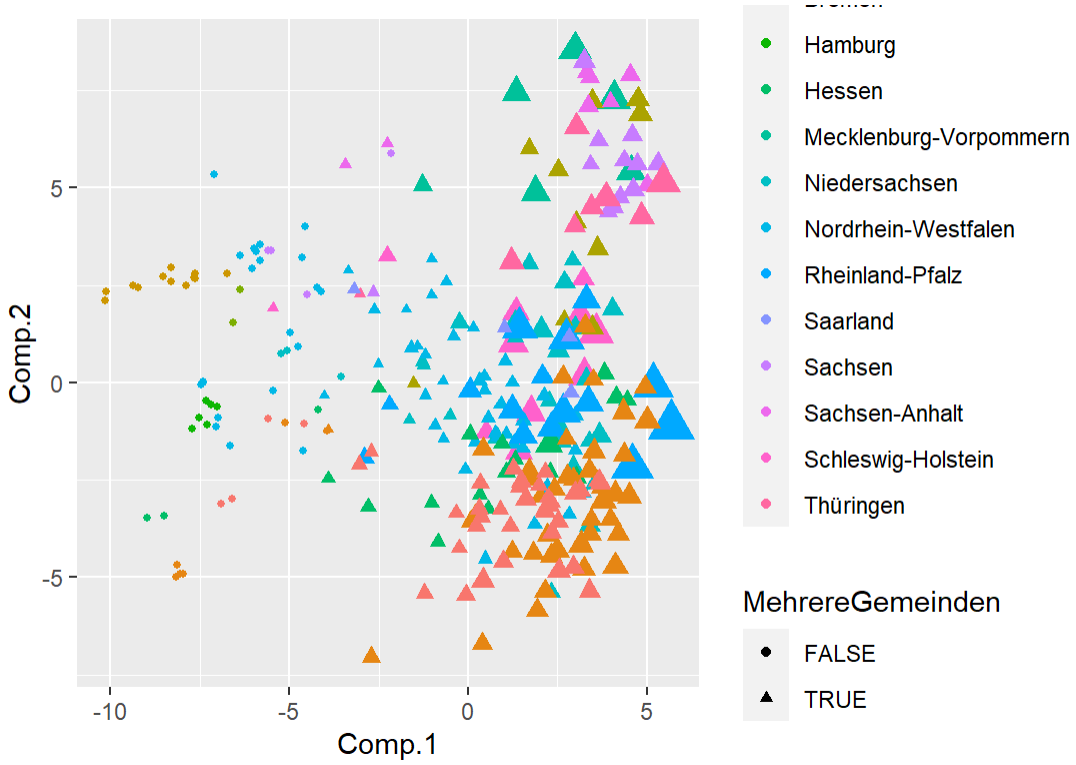
```
btw21_PCA$loadings[,1:2] |> round(1)
```

```
##                                  Comp.1 Comp.2
## Gemeinden am 31.12.2019 (Anzahl)    0.1    0.0
## X4                                  0.2    0.1
## X5                                 -0.1   -0.1
## X6                                  0.0   -0.1
## X7                                 -0.2   -0.1
## X8                                 -0.2    0.0
## X9                                 -0.2   -0.2
## X10                                 0.0    0.0
## X11                                 0.0   -0.2
## X12                                -0.1   -0.2
## X13                                -0.2   -0.1
## X14                                 0.1   -0.1
## X15                                 0.2    0.2
## X16                                 0.1    0.2
## X17                                -0.2    0.0
## X18                                 0.2    0.0
## X19                                 0.0   -0.1
## X20                                 0.0    0.2
## X21                                 0.2   -0.2
## X22                                 0.2    0.0
## X23                                 0.2   -0.1
## X24                                -0.2   -0.1
## X25                                -0.1   -0.1
## X26                                 0.2    0.0
## X27                                 0.0   -0.1
## X28                                 0.1   -0.1
## X29                                 0.0    0.2
## X30                                 0.0   -0.1
## X31                                 0.2    0.0
## X32                                -0.2    0.0
## X33                                 0.0    0.2
## X34                                 0.1    0.0
## X35                                 0.0   -0.2
## X36                                -0.1   -0.1
## X37                                -0.2   -0.1
## X38                                 0.2    0.1
## X39                                 0.2   -0.1
## X40                                 0.0    0.0
## X41                                -0.2    0.0
## X42                                 0.0    0.2
## X43                                -0.2    0.2
## X44                                -0.1   -0.2
## X45                                -0.1   -0.2
## X46                                -0.2    0.2
## X47                                -0.2    0.2
## X48                                -0.2    0.2
## X49                                -0.1    0.2
## X50                                -0.2    0.2
```

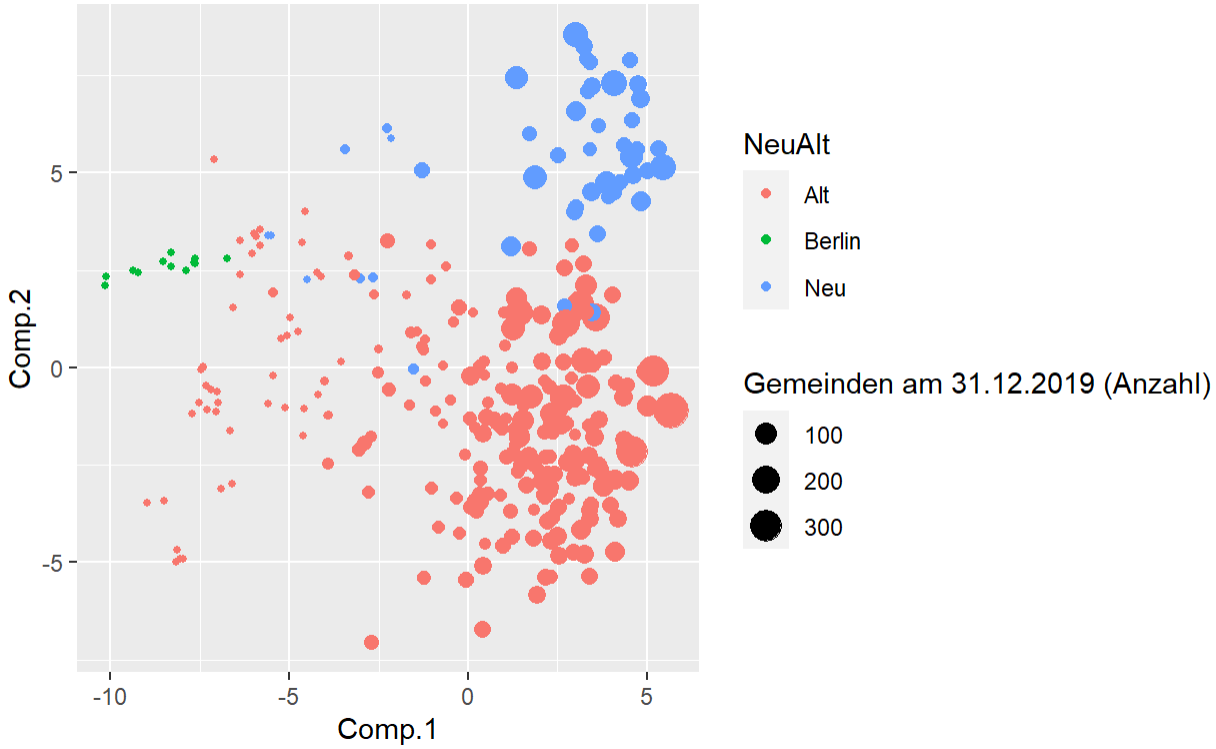
PCA - Visualisierung



PCA - Visualisierung



PCA - Visualisierung

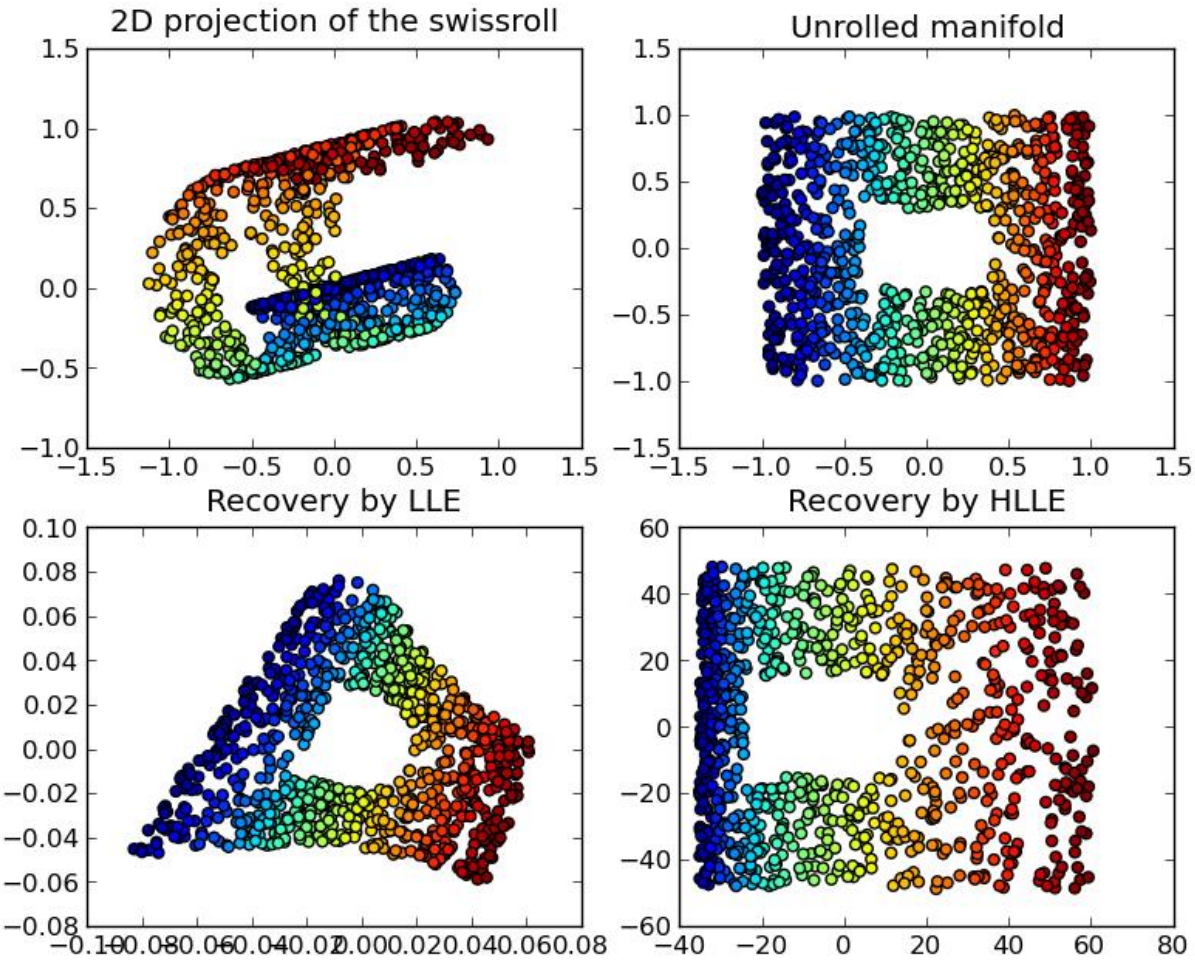


Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Prof. Dr. Rainer Stollhoff

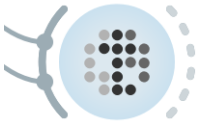
Unsupervised Learning
Dimensionsreduktion
Manifold Learning

Motivation - Nicht-lineare Datenstruktur



Swissroll manifold recovery by the LLE and Hessian LLE algorithms, Olivier Grisel

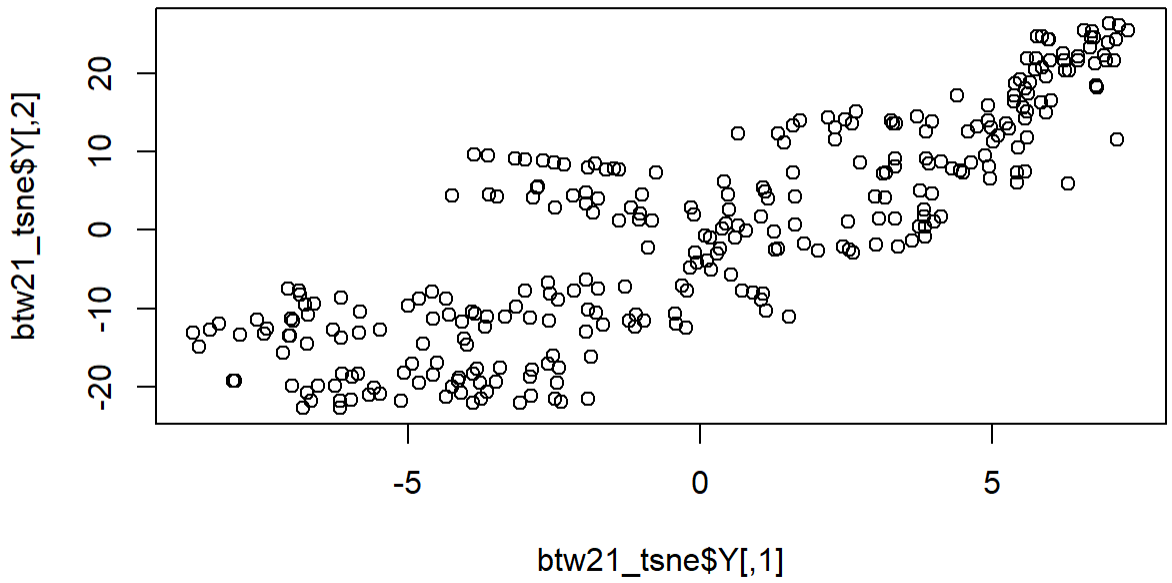
t-verteilte stochastische Nachbarschaftseinbettung



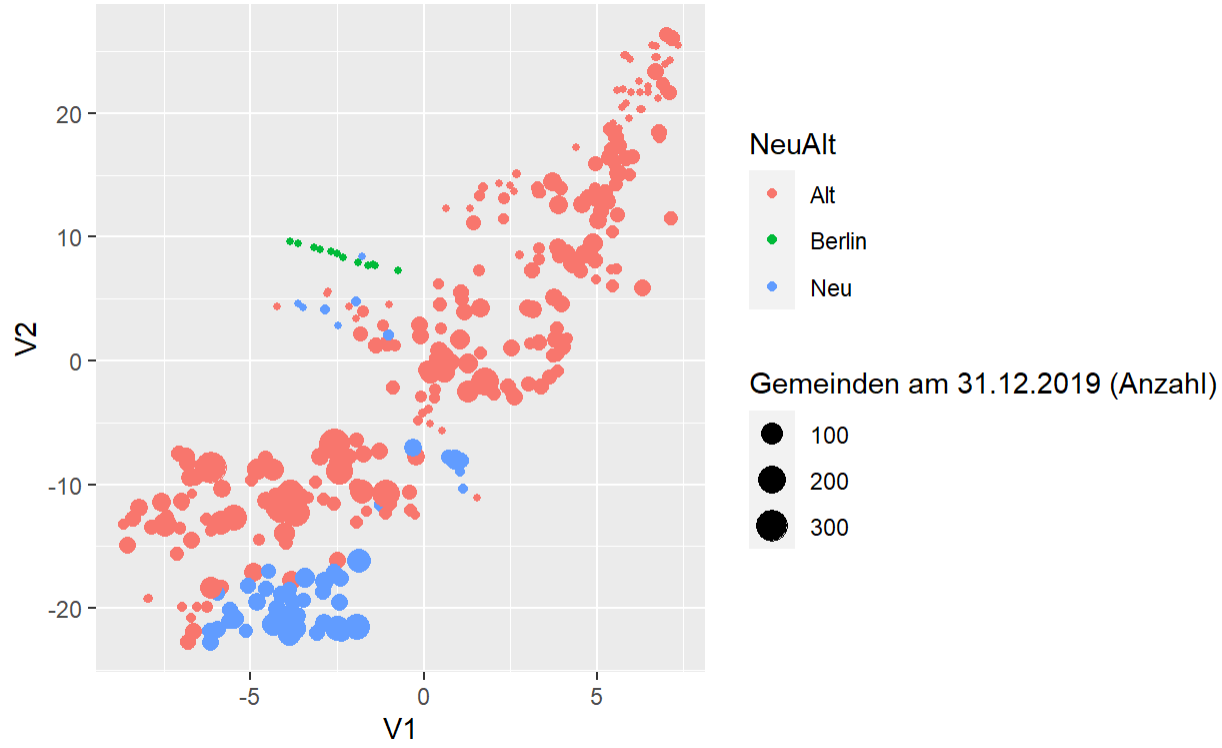
t-SNE

- **Stochastische Nachbarn** für die Eingabedaten definieren
  - Für jede Beobachtung i werden die "Übereinstimmungs"-Wahrscheinlichkeiten $p_{j|i}$ für alle anderen Beobachtungen j unter Verwendung einer normalisierten Gaußschen Dichte mit Mittelpunkt x_i berechnet
  - Beobachtungen j mit hohen "Übereinstimmungswahrscheinlichkeiten" $p_{j|i}$ sind Nachbarn von i
- Neue zweidimensionale **Einbettungen** finden
  - Definieren Sie eine neue „Übereinstimmungs-" Wahrscheinlichkeit $q_{j|i}$ für alle anderen Beobachtungen j unter Verwendung der Dichte einer **t-Verteilung** mit Mittelpunkt x_i
  - Passen Sie die Werte für y_i so an, dass sich $p_{j|i}$ und $q_{j|i}$ so wenig wie möglich unterscheiden und die "Nachbarschaften" stabil bleiben.

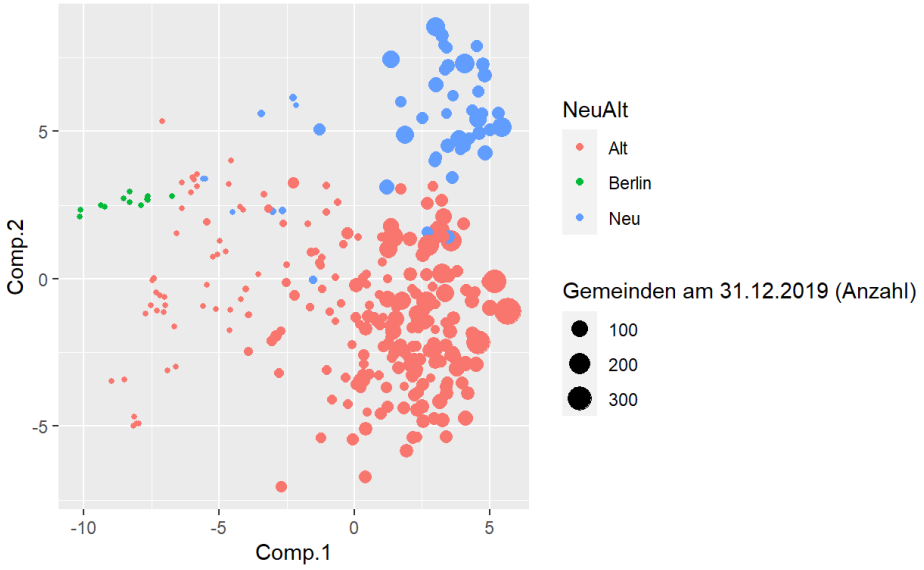
t-SNE-Visualisierung

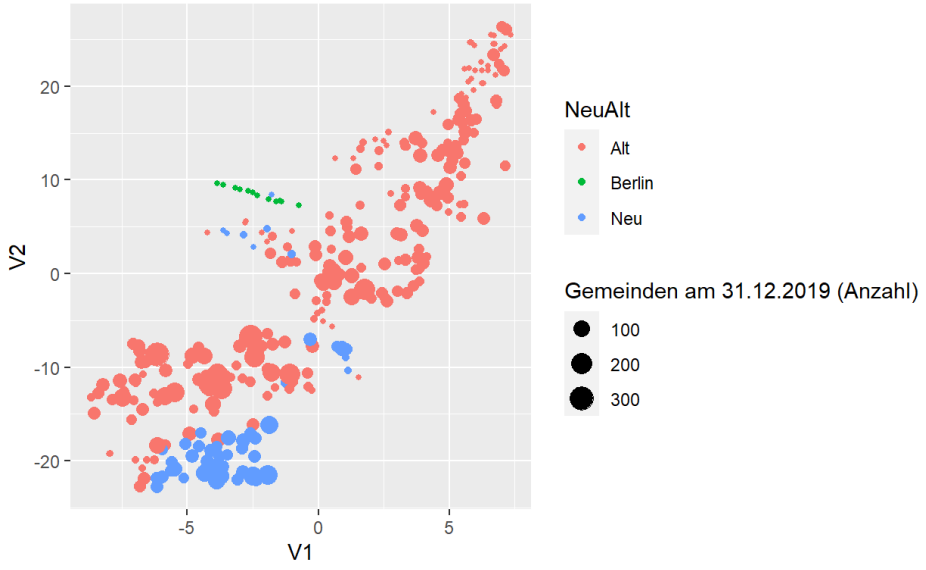


t-SNE-Visualisierung



Vergleich der Methoden





Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Prof. Dr. Rainer Stollhoff

Unsupervised Learning
Clustering

Gliederung

- Motivation
- k-Means-Clustering
- Hierarchisches Clustering

Unsupervised Learning - Diskrete Projektionen

https://pixabay.com/vectors/european-union-flags-stars-eu-1328256/ CC

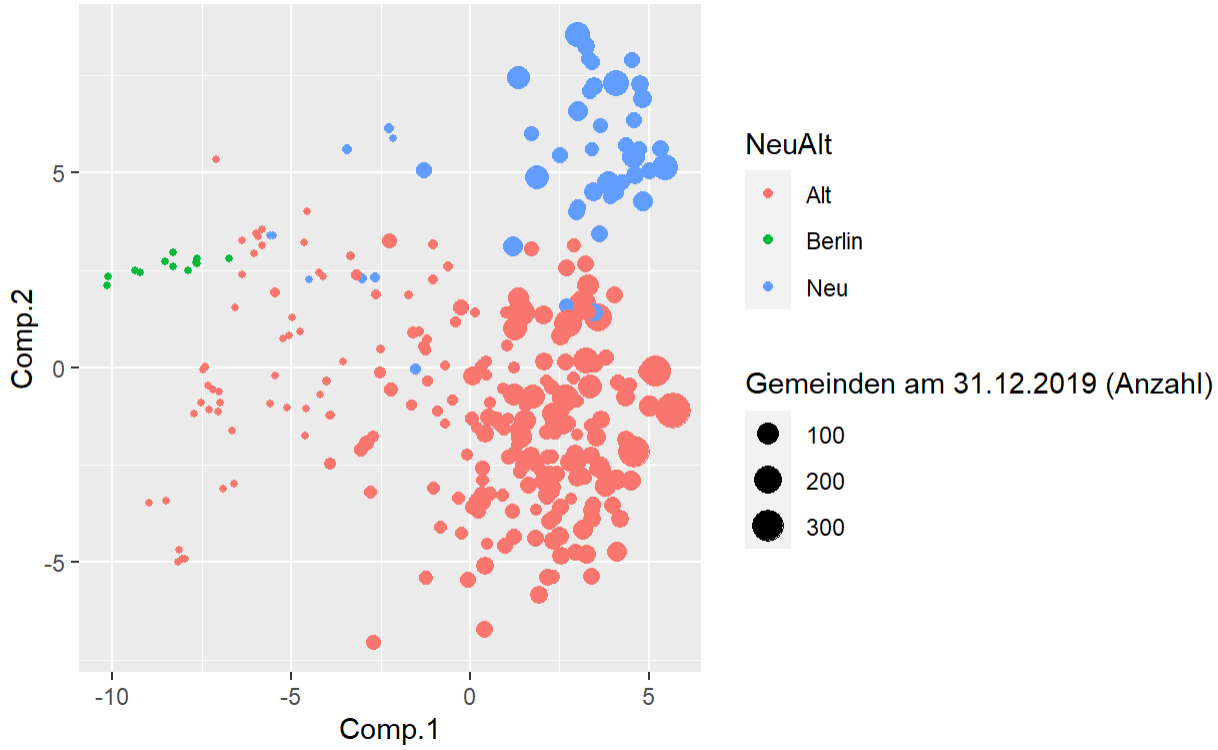
https://www.flickr.com/photos/donkeyhotey/19920864433 CC BY 2.0

Länder
Angola
Benin
Burkina Faso
Kamerun
Republik Congo
Demokratische Republik Kongo
Indonesien
Gambia
Ghana
Guinea-Bissau
Malaysia
Nigeria
Senegal
Sudan
Togo
Uganda
Burundi
Zentralafrikanische Republik
Elfenbeinküste
Äquatorial Guinea
Gabun
Guinea
Liberia
Mali
Ruanda
Sierra Leone
Tansania
Uganda

WWF, 2005

Gibt es einzelne Gruppen von Wahlkreisen?

z. B. urban vs. ländlich, neu vs. alt? ...



Idee des Clustering

- Gruppieren Sie Beobachtungen nach ihrer Ähnlichkeit
- Hierarchisches Clustering:
  - Berechnen von paarweisen Abständen zwischen Beobachtungen
  - Kombinieren Sie mehr und mehr Datenpunkte zu immer größeren Clustern auf der Grundlage ihrer Abstände
- K-Means Clustering:
  - Bilden Sie k zufällige Cluster und tauschen Sie so lange Datenpunkte zwischen den Clustern aus, bis sie sich intern so ähnlich wie möglich sind.
- Ähnlichkeitsmaße
  - Euklidischer Abstand
  - (Pearson) Korrelationskoeffizient

Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Prof. Dr. Rainer Stollhoff

Unsupervised Learning

Clustering

k-Means-Clustering

K-Means Clustering:

- Die Idee:
  Bilden Sie k zufällige Cluster und tauschen Sie so lange Datenpunkte zwischen den Clustern aus, bis sie sich intern so ähnlich wie möglich sind.
- Silhouette Score
  - Für jede Beobachtung
    - a(i) = Durchschnittlicher gleicher Clusterabstand
    - b(i) = Minimum über den durchschnittlichen anderen Clusterabstand
    - $s(i) \sim \frac{b(i)-a(i)}{\max(b(i),a(i))}$
  - Silhouette Score = Durchschnitt s(i)

```
## K-means clustering with 3 clusters of sizes 30, 166, 103
##
## Cluster means:
##   Gemeinden am 31.12.2019 (Anzahl)       X4       X5       X6        X7
## 1                         6.166667  266.410 320.6437 255.0167 19.973333
## 2                        50.487952 1593.027 265.2060 240.2982  9.177711
## 3                        21.961165  826.601 286.6282 244.9379 14.241748
##          X8         X9      X10      X11      X12      X13      X14      X15
## 1 2684.0733  1.4500000 3.306667 16.41000 8.193333 16.35333 35.06667 14.00667
## 2  380.3392 -3.7234940 4.139157 16.23735 6.947590 11.18072 34.87169 18.48313
## 3 1291.3233 -0.8281553 3.615534 16.62524 8.078641 13.57767 34.48252 16.21262
##         X16      X17      X18      X19      X20      X21      X22      X23
## 1  9.953333 52.50333 47.49667 4.050000 521.5033 78.02333 40.56000 514.8467
## 2 12.281325 18.85241 81.14759 3.146386 517.3066 92.42711 47.40663 600.8958
## 3 11.017476 32.10097 67.89903 3.765049 502.6019 89.49126 44.64951 558.3699
##        X24      X25      X26      X27      X28      X29      X30      X31
## 1 2.526667 51.63667 5.233333 3.186667 8.923333 5.786667 15.36333 36.18667
## 2 1.130723 38.91928 7.324096 2.787952 9.527108 7.209639 15.72048 43.87169
## 3 1.554369 43.09903 6.414563 3.415534 9.794175 6.458252 16.67767 41.83398
##        X32      X33      X34      X35      X36      X37       X38      X39
## 1 42.65667 37.90333 91.62000 25633.73 75086.20 591.2533 0.1266667 19.85667
## 2 33.20602 37.25241 92.75482 22033.11 30698.57 340.6693 1.3084337 30.73855
## 3 35.00777 32.15243 92.67087 23148.45 43015.26 426.0621 0.4864078 29.73204
##        X40      X41      X42      X43      X44      X45      X46      X47
## 1 20.98333 34.26000 24.77333 78.4200 27.29333 43.37000 6.786667 7.176667
## 2 23.13614 14.71265 30.10843 63.1006 26.04398 31.38916 6.302410 6.788554
## 3 21.55631 20.46602 27.76408 70.4835 27.47767 40.15534 6.408738 6.856311
##        X48      X49      X50
## 1 6.366667 5.416667 7.616667
## 2 5.756627 6.079518 6.704819
## 3 5.901942 5.765049 6.662136
##
## Clustering vector:
##   [1] 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 3 2 2 2 3 3 2 2 2 2 2
##  [38] 2 3 2 3 3 3 3 2 2 2 3 2 2 1 1 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 1 1 1 1 2 2 2 2 3 2 2 3 3 1 1 3 2 3 2
## [112] 2 2 2 2 2 2 3 3 3 2 2 2 2 2 3 2 2 3 2 3 3 2 3 2 2 2 2 2 2 2 3 3 2 2 2 2 3
## [149] 3 3 2 2 2 2 2 2 2 2 3 3 2 2 2 2 2 2 2 3 2 2 3 2 2 2 2 3 2 2 1 2 3 1 1 3 3
## [186] 3 2 2 2 2 2 2 3 3 2 2 2 2 2 2 3 3 2 2 2 2 2 3 3 2 1 1 1 1 1 1 2
## [223] 2 3 3 3 2 3 3 3 2 3 3 2 3 3 3 3 2 2 2 1 2 1 3 2 2 2 3 3 3 2 3 3 3 3 3 1 1
## [260] 1 3 3 2 2 3 3 3 3 3 3 1 3 3 3 1 2 2 2 2 2 3 2 2 2 3 3 3 2 2 3 2 3 2 3 3 3 3 3 3
## [297] 2 2 3
##
## Within cluster sum of squares by cluster:
## [1] 5875327975 3456694074 3487624972
##  (between_SS / total_SS =  80.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

k=3

```
table(btw21_kM$cluster, btw21$Land)
```

```
##
##     Baden-Württemberg Bayern Berlin Brandenburg Bremen Hamburg Hessen
##   1                 7      8      0           0      0       6      5
##   2                31     38      0          10      0       0     16
##   3                 0      0     12           0      2       0      1
##
##     Mecklenburg-Vorpommern Niedersachsen Nordrhein-Westfalen Rheinland-Pfalz
##   1                      0             1                   7               1
##   2                      5            27                  33              13
##   3                      1             2                  24               1
##
##     Saarland Sachsen Sachsen-Anhalt Schleswig-Holstein Thüringen
##   1        0       0              0                  0         0
##   2        3      11              7                  9         7
##   3        1       5              2                  2         1
```

```
table(btw21_kM$cluster, btw21$MehrereGemeinden)
```

```
##
##     FALSE TRUE
##   1    27    8
##   2     0  210
##   3    34   20
```

```
table(btw21_kM$cluster, btw21$NeuAlt)
```

```
##
##     Alt Berlin Neu
##   1  35      0   0
##   2 170      0  40
##   3  33     12   9
```

k=6

```
set.seed(142)
btw21_kM <- select_if(btw21,is.numeric) |>
  mutate(across(everything(),~./sd(.))) |>
  kmeans(centers = 6)
table(btw21_kM$cluster, btw21$MehrereGemeinden)
```

```
##
##      FALSE TRUE
##    1    27    5
##    2     0   49
##    3    34   12
##    4     0   55
##    5     0   78
##    6     0   39
```

```
table(btw21_kM$cluster, btw21$NeuAlt)
```

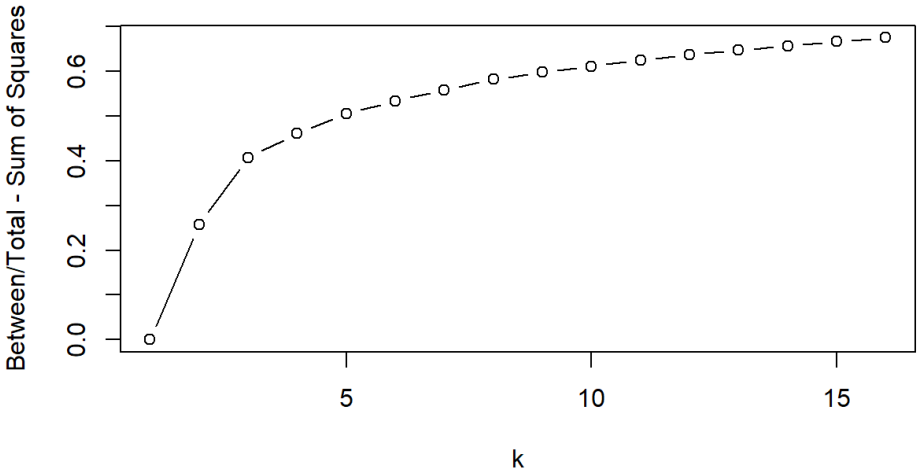
```
##
##      Alt Berlin Neu
##    1  32      0   0
##    2  49      0   0
##    3  26     12   8
##    4  55      0   0
##    5  76      0   2
##    6   0      0  39
```

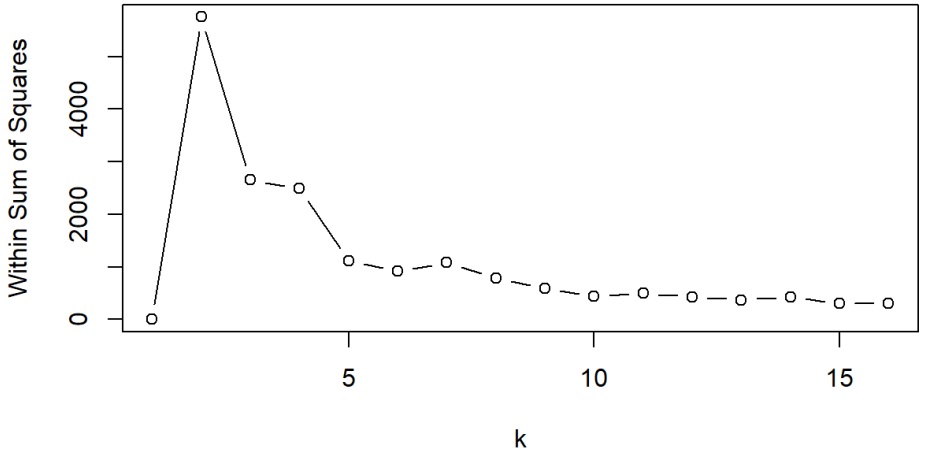
k=16

```
table(btw21_kM16$cluster, btw21$Land)
```

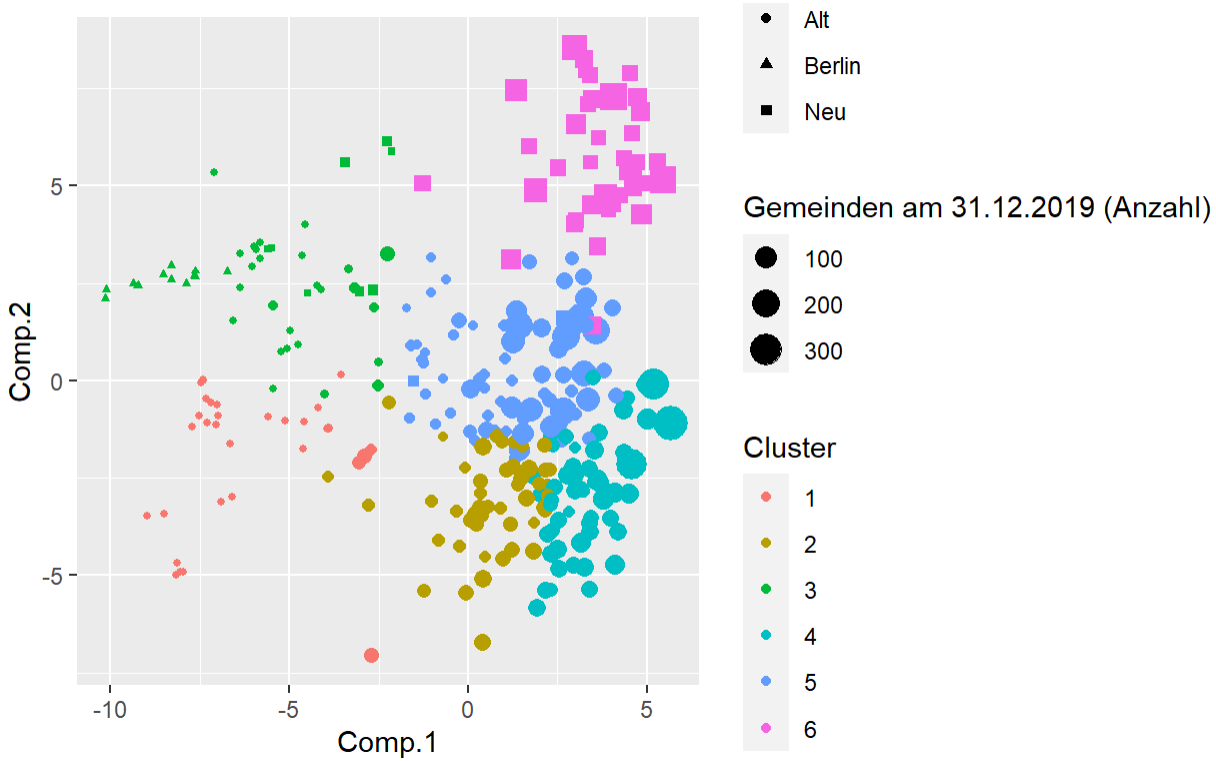
```
##
##    Baden-Württemberg Bayern Berlin Brandenburg Bremen Hamburg Hessen
## 1                  2      5      0           0      0       0      2
## 2                  3      5      0           0      0       0      8
## 3                  4      3      0           0      0       0      3
## 4                  2     15      0           0      0       0      0
## 5                  0      1      0           3      0       0      0
## 6                  0     10      0           0      0       0      3
## 7                  0      0      0           0      0       0      1
## 8                  0      0     12           0      0       6      0
## 9                  7      3      0           0      0       0      0
## 10                 0      0      0           0      0       0      0
## 11                20      4      0           0      0       0      3
## 12                 0      0      0           7      0       0      0
## 13                 0      0      0           0      0       0      2
## 14                 0      0      0           0      0       0      0
## 15                 0      0      0           0      2       0      0
## 16                 0      0      0           0      0       0      0
##
##    Mecklenburg-Vorpommern Niedersachsen Nordrhein-Westfalen Rheinland-Pfalz
## 1                       0             0                   2               0
## 2                       0             0                   0               0
## 3                       0             4                   5               1
## 4                       0             4                   0               0
## 5                       0             4                   7               3
## 6                       0             0                   0               7
## 7                       0             5                  18               2
## 8                       0             0                   3               0
## 9                       0             1                   1               0
## 10                      1             0                   0               0
## 11                      0             0                   9               0
## 12                      5             0                   0               0
## 13                      0             0                   5               0
## 14                      0            12                   0               2
## 15                      0             0                  11               0
## 16                      0             0                   3               0
##
##    Saarland Sachsen Sachsen-Anhalt Schleswig-Holstein Thüringen
## 1         0       0              0                  0         0
## 2         0       0              0                  0         0
## 3         0       0              0                  1         0
## 4         0       0              0                  0         0
## 5         0       0              0                  3         0
## 6         2       0              0                  0         0
## 7         1       0              0                  0         0
## 8         0       0              0                  0         0
## 9         0       0              0                  0         0
## 10        0       5              2                  1         1
## 11        0       0              0                  0         0
## 12        0      11              7                  0         7
## 13        0       0              0                  0         0
## 14        0       0              0                  6         0
## 15        1       0              0                  0         0
## 16        0       0              0                  0         0
```

Welches k?


Between/Total - Sum of Squares
0.0
0.2
0.4
0.6
5
10
15
k


Within Sum of Squares
0
2000
4000
5
10
15
k

Kombination PCA und k-means

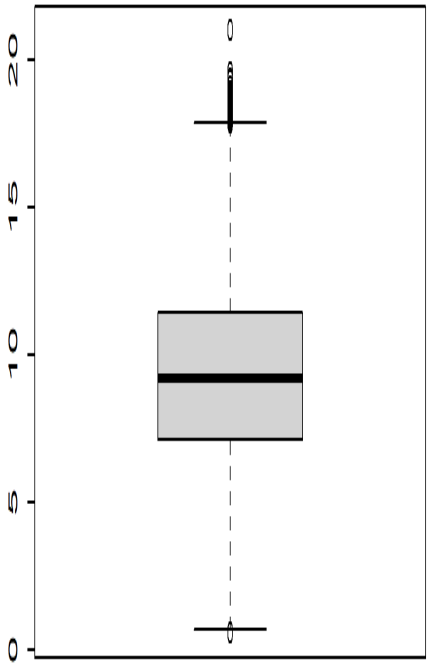


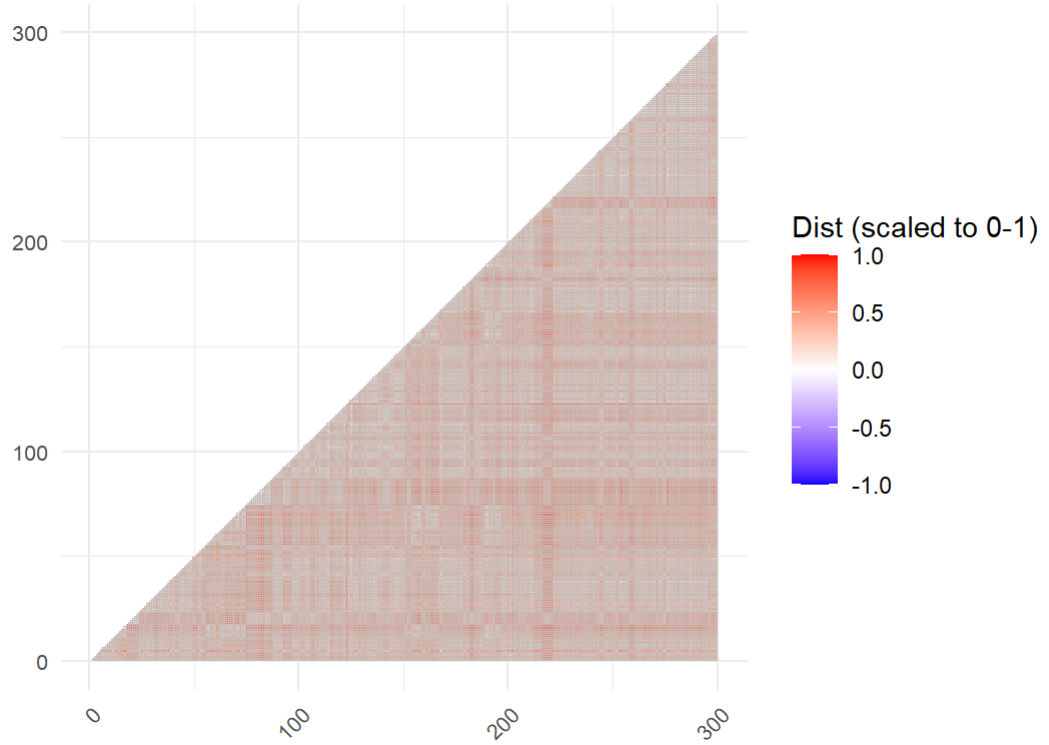
Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Prof. Dr. Rainer Stollhoff

Unsupervised Learning
Clustering
Hierarchisches Clustering

Entfernungen



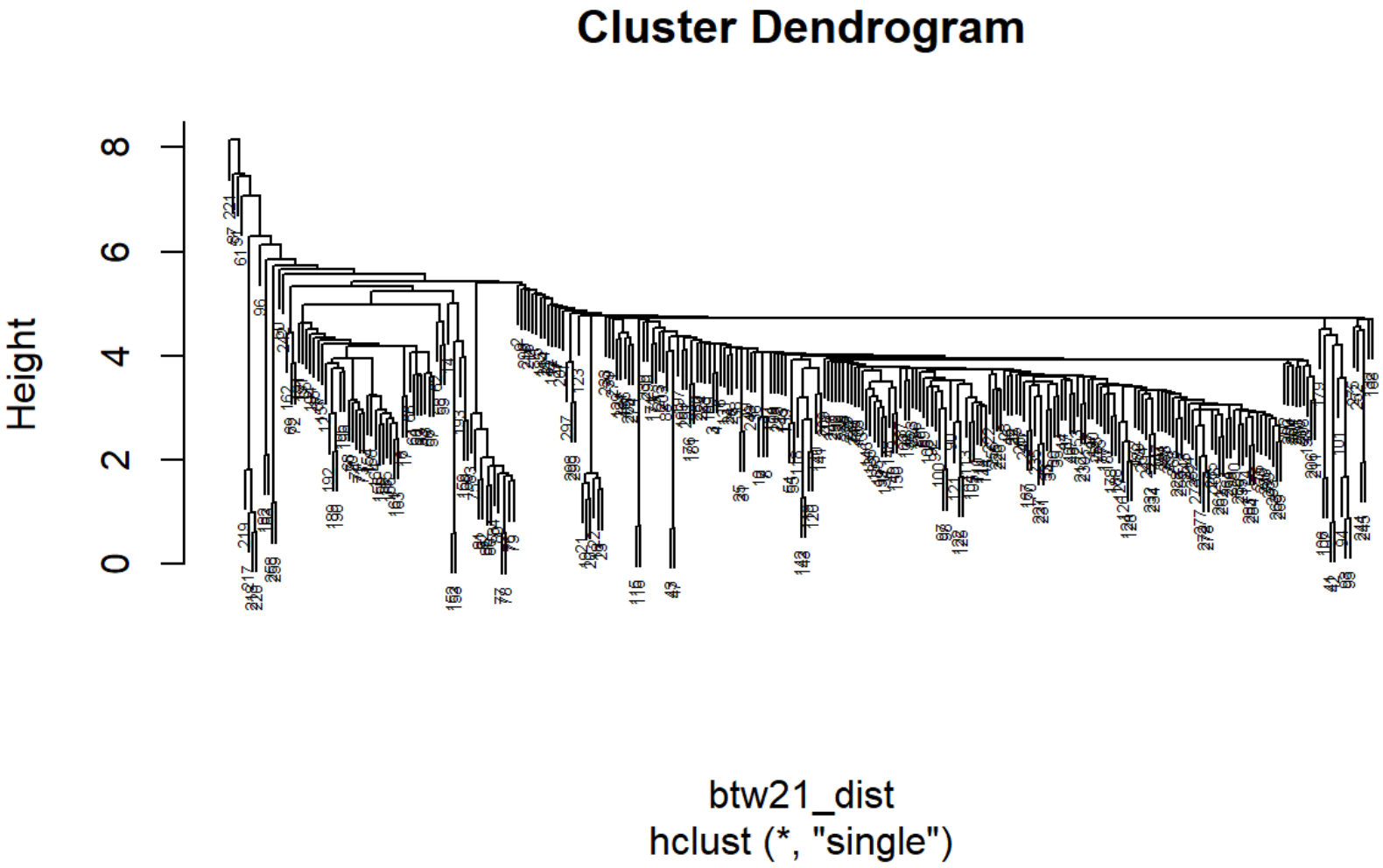


Hierarchisches Clustering

- Idee:
  Berechnen Sie, wie die Kombination von Beobachtungen die Abstände verändern würde (Linkage/Verknüpfung)
- Linkage-Strategien:
  - Single Linkage:
    Minimaler Abstand zwischen Elementen von Clustern
  - Average Linkage:
    Durchschnittlicher Abstand zwischen Elementen von Clustern
  - Complete Linkage:
    Maximaler Abstand zwischen Elementen von Clustern
  - Ward:
    Minimierung der Varianz innerhalb von Clustern

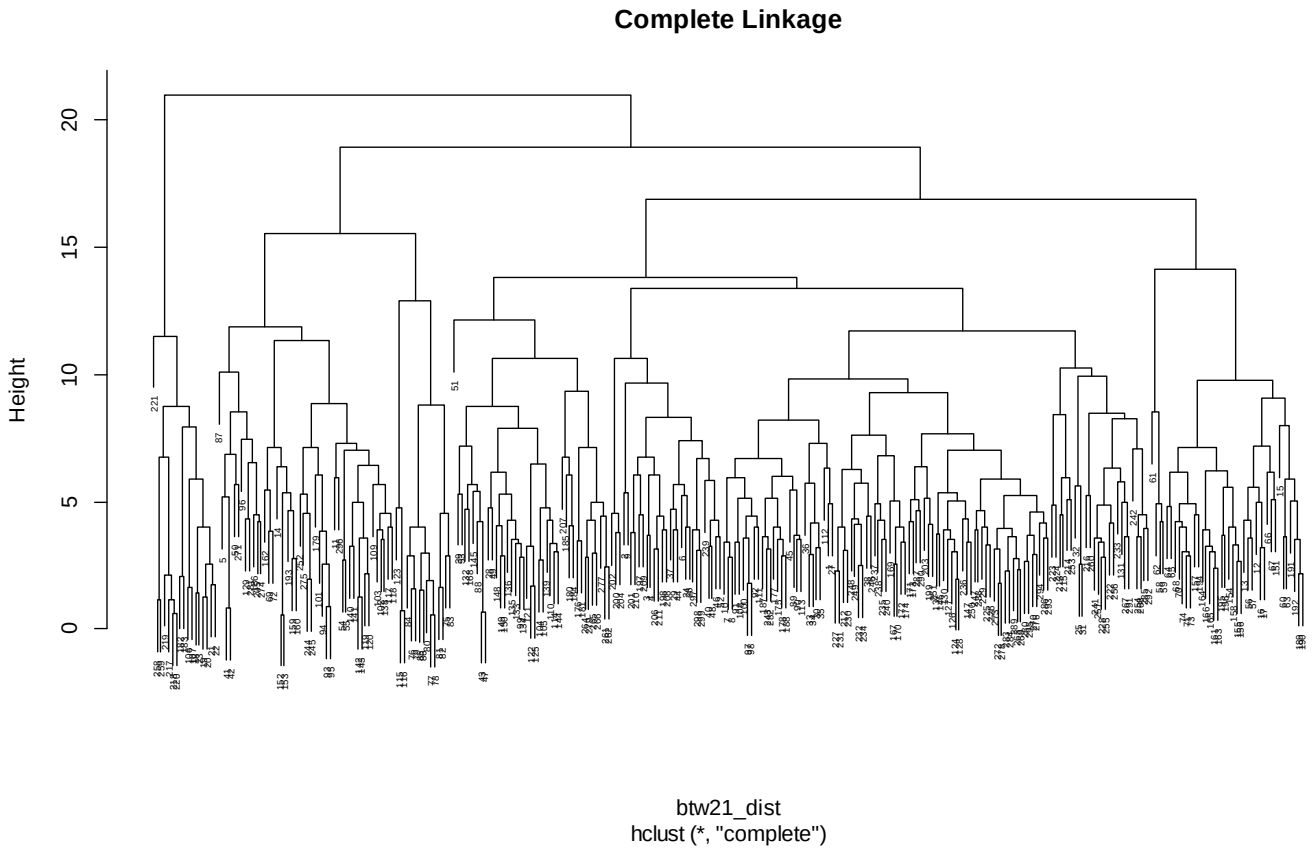
Single Linkage Clustering

Neigt dazu, "lange, dünne, gestreckte, kettenartige" Büschel zu bilden

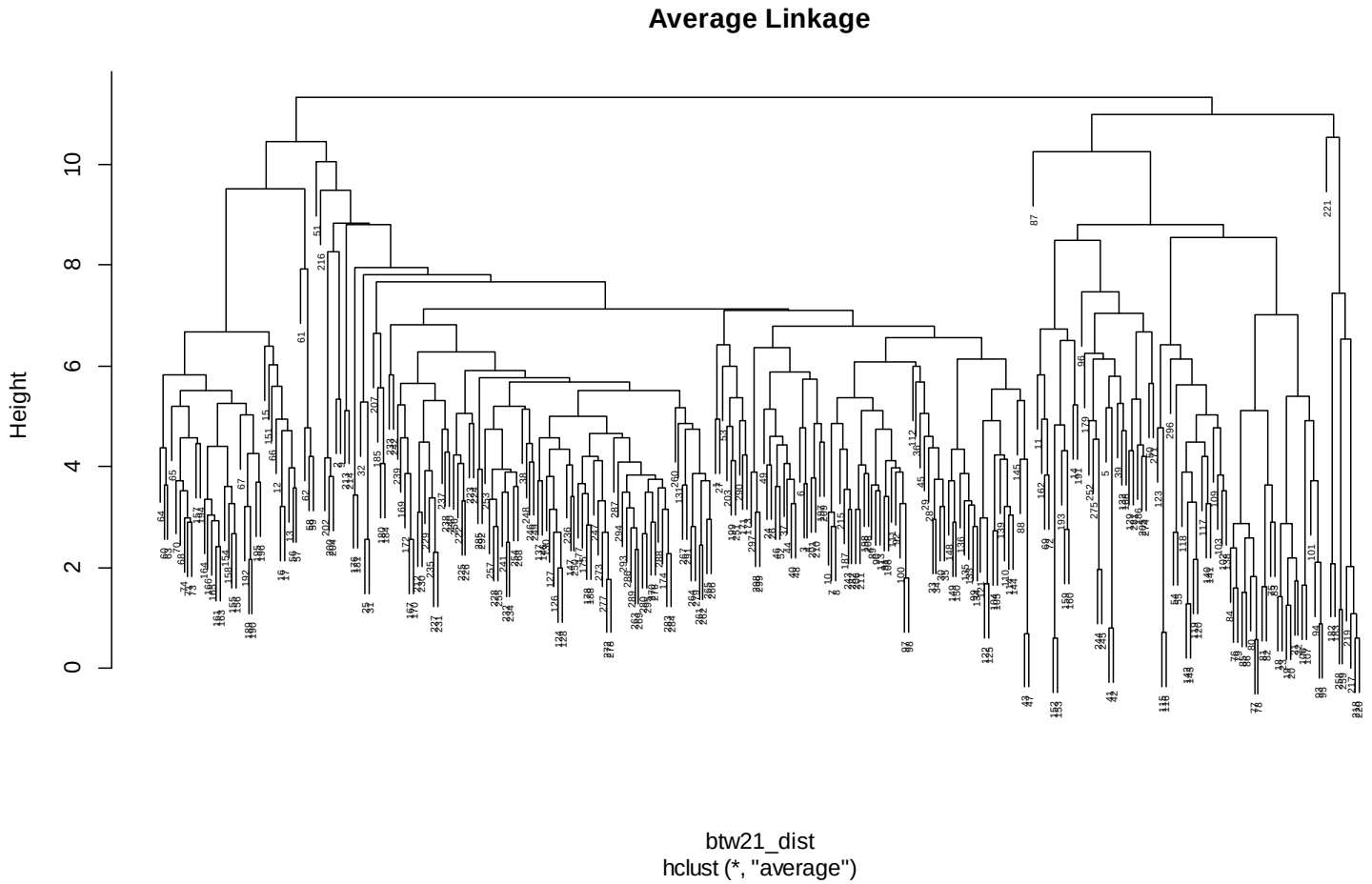


Complete Linkage Clustering

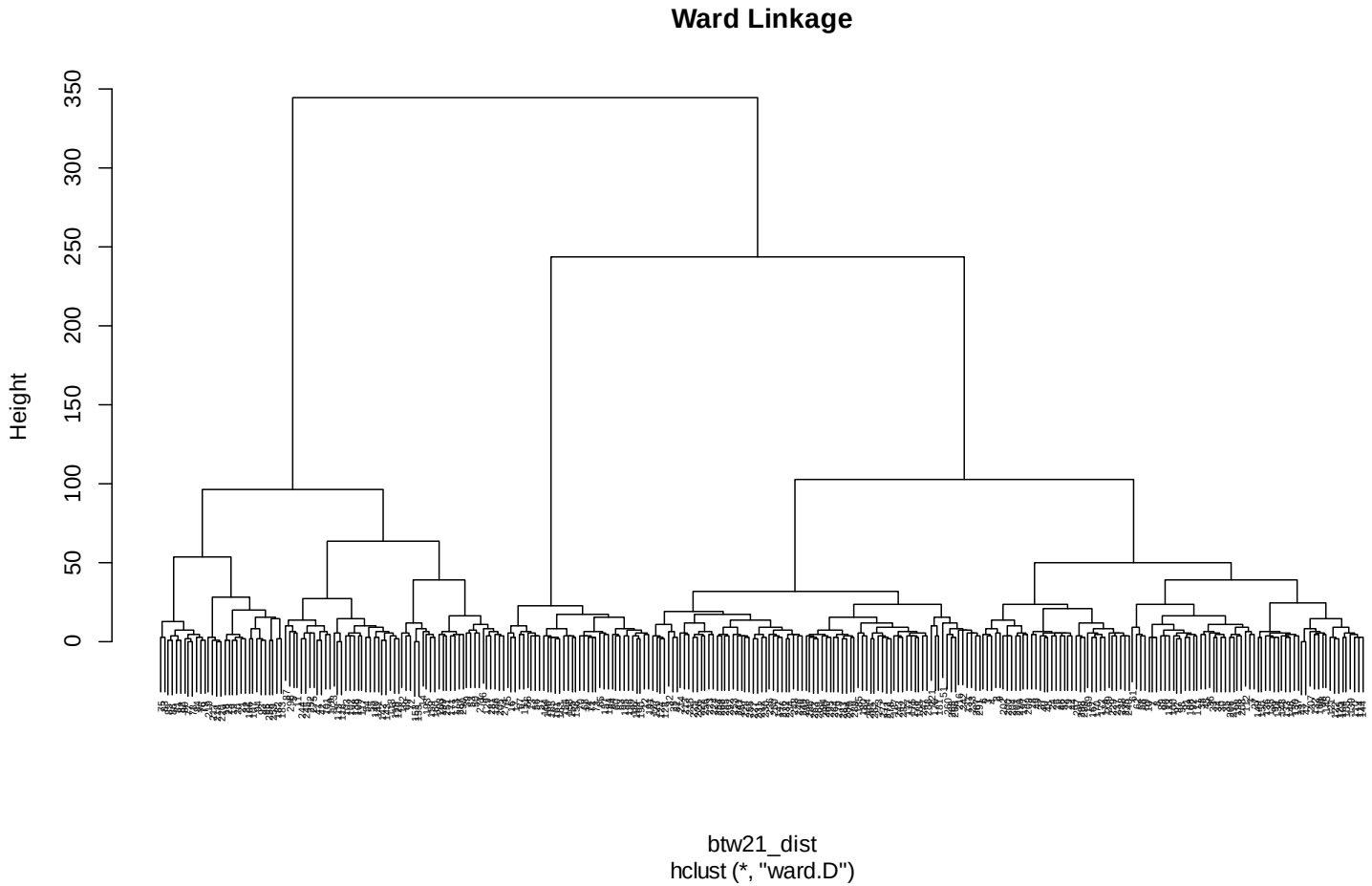
Neigt zur Bildung von "kompakten, gleich großen" Clustern



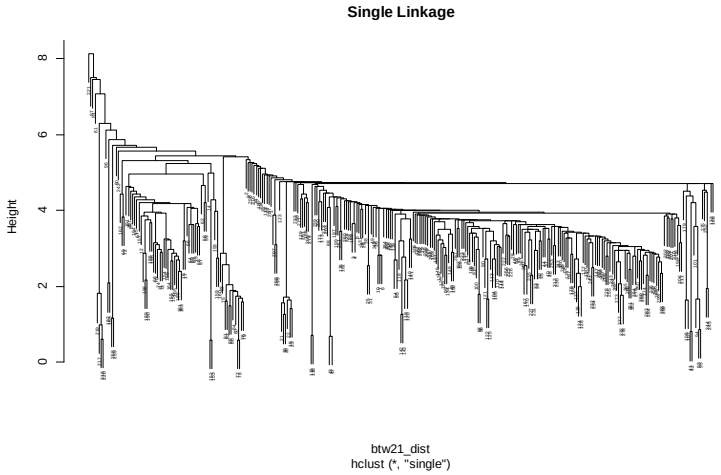
Average Linkage Clustering

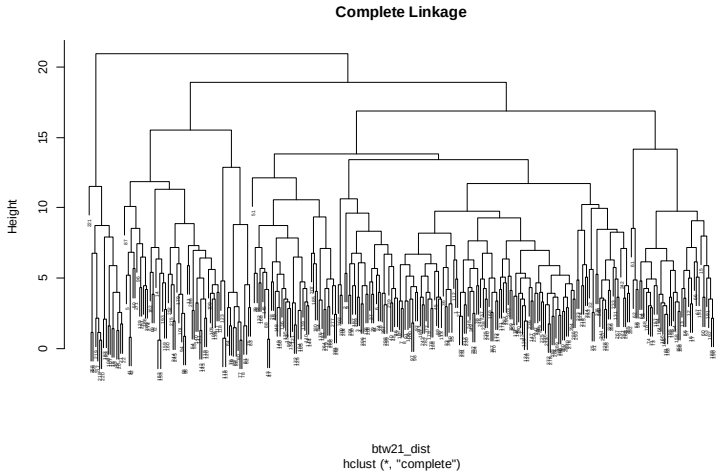


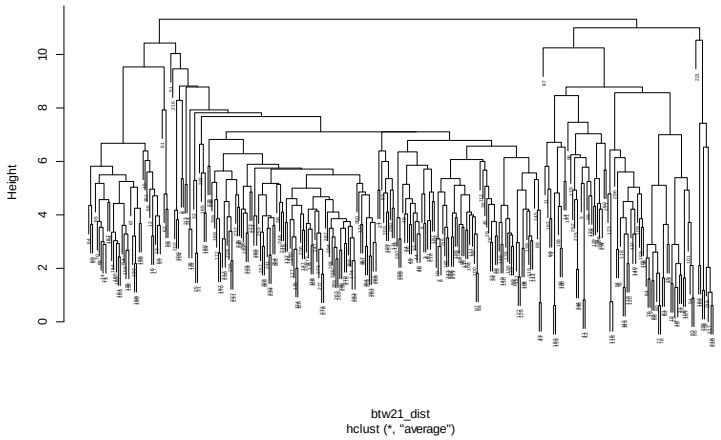
Ward Clustering

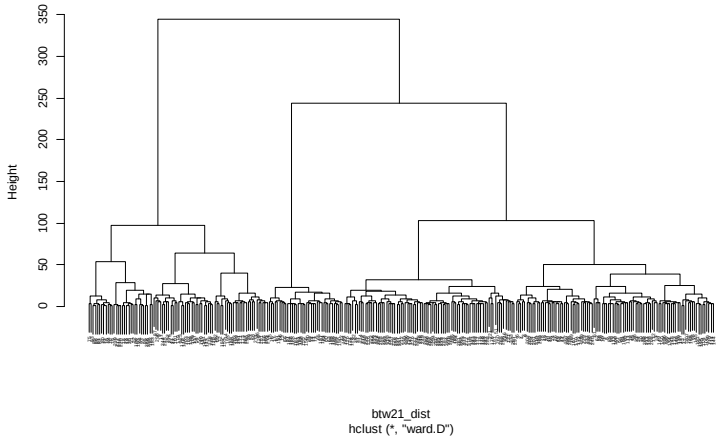


Vergleich


Single Linkage
Height
0
2
4
6
8
btw21_dist
hclust (*, "single")


Complete Linkage
Height
0
5
10
15
20
btw21_dist
hclust (*, "complete")


Average Linkage
Height
0
2
4
6
8
10
btw21_dist
hclust (*, "average")


Ward Linkage
Height
0
50
100
150
200
250
300
350
btw21_dist
hclust (*, "ward.D")

Visualisierung und Vergleich