## MSc. Research Methods – Statistikteil Lösungstext

## - Übung 8.1: Clusteranalyse -

## Methoden

In den USA wurden die Kriminalitätsraten für 7 Verbrechenstypen (Murder, Rape, Robbery, Assault, Burglary, Theft, Vehicle) pro 100000 Einwohner in den 50 Bundesstaaten ermittelt. Zur Klärung der Frage, welche Bundesstaaten sich in ihrer Kriminalitätsstruktur ähneln, wurden eine divisive und eine agglomerative Clusteranalyse durchgeführt.

Als divisives Verfahren wurde k-means Clustering nach vorheriger z-Transformation (Mittelwert = 0; Standardabweichung = 1) der Verbrechensraten durchgeführt. Die beste Lösung mit maximal 6 Gruppen wurde nach dem SSI-Kriterium bestimmt. Als agglomeratives Verfahren fand Ward's Minimum Variance-Clustering statt. Hier wurde die optimale Clusterzahl mit dem Silhouette-Plot ermittelt.

Anschliessend wurde die Gruppen der optimalen Clusterlösung im Falle von k-means-Clustering mittels Varianzanalysen bezüglich Unterschieden in den Raten der sieben Verbrechenstypen verglichen. Signifikante Unterschiede wurden mittels Tukey's post-hoc Test im Package 'multcomp' ermittelt und in Boxplots visualisiert. Da die visuelle Inspektion der Boxplots keine gravierenden Abweichungen von Normalverteilung und Varianzhomogenität der Residuen erbrachte, wurde auf eine Transformation der Daten verzichtet.

## **Ergebnisse**

Die agglomerative Clusteranalyse mit Wards Minimum Variance Clustering erbrachte 2 Cluster als optimale Lösung (Abb. 1).

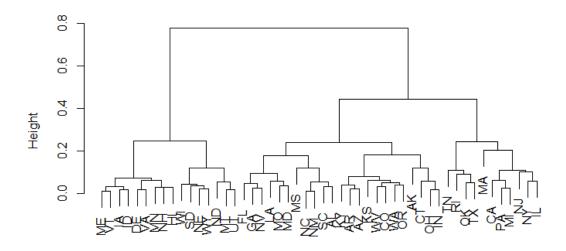


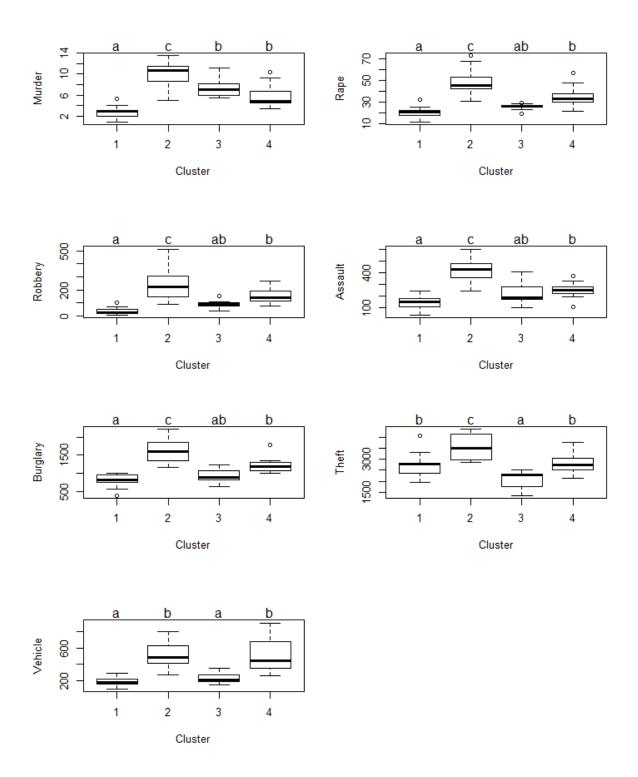
Abb. 1: Agglomerative Clusterung der Verbrechensstrukturen in den 50 Bundesstaaten der USA basierend auf Inzidenzraten von 7 Verbrechenskategorien mittels Ward's Minium Variance Clusterung. Die Kürzel stehen für die Bundesstaaten.

Die k-mean-Clusteranalyse (für 2 bis 6 Cluster) erbrachte 4 Cluster als optimale Lösung gemäss SSI (Tab. 1).

**Tab. 1:** Zugehörigkeit der 50 Bundesstaaten zu den 4 mittels *k*-means-Clusterung unterschiedenen Gruppen.

Cluster	Anzahl Staaten	Staaten	Kurzcharakterisierung
1	13	IA, ID, ME, MN, MT, ND, NE, NH, SD, UT, VT, WI, WY	Niedrige Kriminalität ausser bei Diebstahl
2	17	AK, AZ, CA, CO, FL, GA, IL, LA, MD, MI, NM, NV, NY, OR, SC, TX, WA	Hohe Kriminalität
3	9	AL, AR, IN, KY, MS, NC, PA, VA, WV	Niedrige Kriminallität ausser bei Mord
4	11	CT, DE, HI, KS, NJ, MO, OH, OK, TN	Mittlere Kriminalität

Die vier k-means-Cluster wurden im Folgenden auf Unterschiede bei den 7 Verbrechenskategorien getestet. Es bestanden durchweg signifikante Unterschiede (ANOVA; p < 0.001; Abb. 2). Cluster 1 zeichnet sich für sechs der sieben Kriminalitätsarten durch niedrigste Werte aus, nur bei Diebstahl sind sie mittelhoch. Cluster 3 hat ebenfalls besonders niedrige Kriminalitätsraten in sechs Kategorien, nur bei Morden liegt dieses Cluster im Mittelfeld. Cluster 4 liegt bei allen Kriminalitätsarten im Mittelfeld und Cluster 2 hat in allen sieben Kategorien besonders hohe Raten.



**Abb. 2:** Vergleich der vier mittels k-means-Clustering abgeleiteten Gruppen von Bundesstaaten bezüglich ihrer Kriminalitätsraten (pro 100000 Einwohner und Jahr) in sieben Kategorien von Vergehen. Die Kleinbuchstaben bezeichnen homogene Gruppen gemäss Tukey's post-hoc-Test ( $\alpha$  = 0.05).