

Katja Gutmair, Stella Akouete, Noah Hurmer und Anne Gritto

# Weather Frog

- Zwischenpräsentation am 21.12.2020
- Institut: Statistik
- Veranstaltung: Statistisches Praktikum
- Projektpartner: Maximilian Weigert und Magdalena Mittermeier
- Betreuer: Dr. Helmut Küchenhoff



# Gliederung

1. Vorstellen des Projekts
2. Einführung in Clustern
3. Ziele
4. Probleme und Ansätze
5. Konzept der Methodik

# Vorstellen des Projekts I

- Übergeordnete Fragestellung:
  - Wie verändert sich das Auftreten verschiedener Großwetterlagen (GWL) unter dem Einfluss des Klimawandels?
- Unsere Fragestellung:
  - Wie lassen sich Tage anhand von ihren Wettermesswerten clustern, um diese GWL-übergreifend in Gruppen einzuteilen?
  - Wie unterscheiden sich die Gruppen voneinander?

# Vorstellen des Projekts II

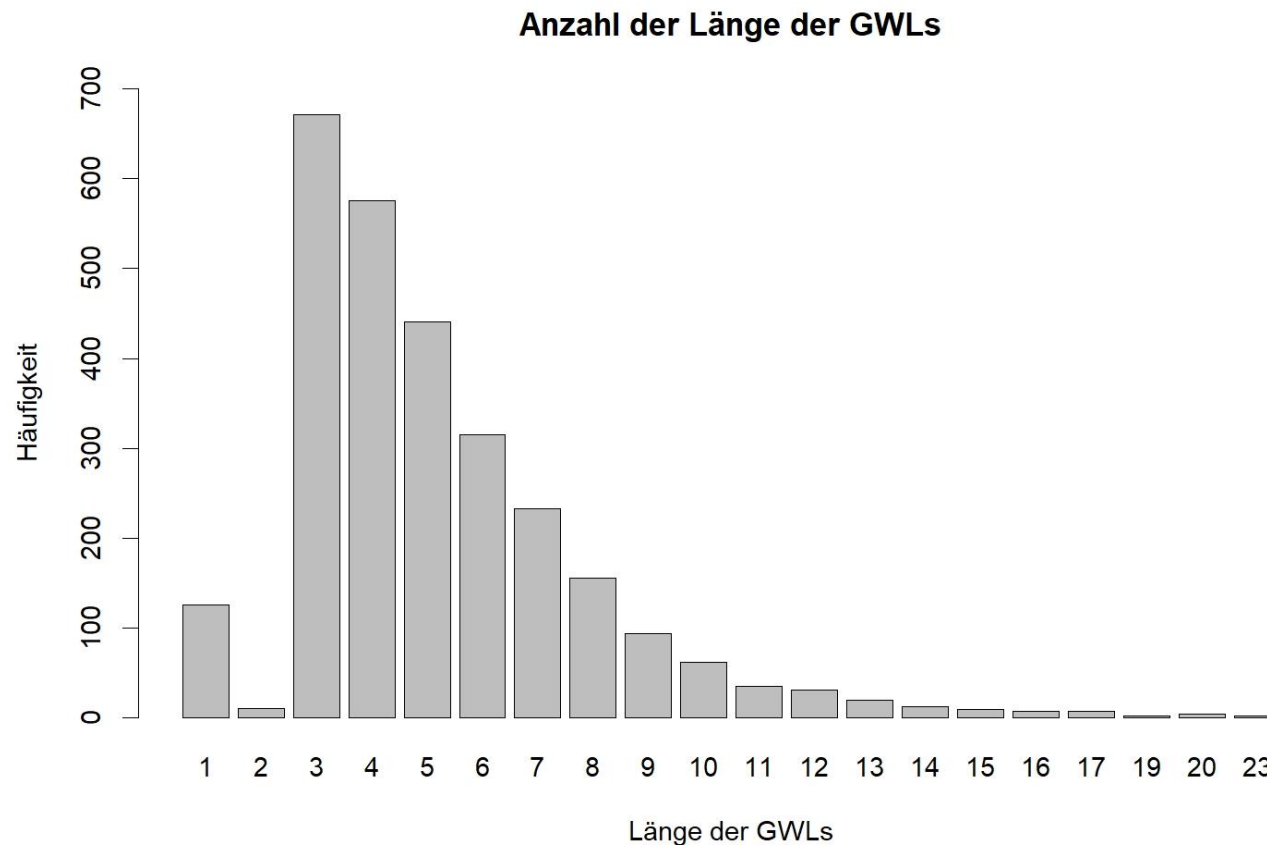
## Definition Großwetterlage

- Atmosphärischer Zustand, definiert durch Strömungsanordnungen
- Definiert über ganz Europa
- Dauer:  $\geq 3$  Tage
- Kategorisierung nach dem Katalog von Hess & Brezowsky
- 29 GWL nach Hess & Brezowsky

# Großwetterlagen Beispiele

1	<b>Wa</b>	Westlage, Mitteleuropa überwiegend antizyklonal
2	<b>Wz</b>	Westlage, Mitteleuropa überwiegend zyklonal
3	<b>WS</b>	Südliche Westlage
4	<b>WW</b>	Winkelförmige Westlage
5	<b>SWa</b>	Südwestlage, Mitteleuropa überwiegend antizyklonal
6	<b>SWz</b>	Südwestlage, Mitteleuropa überwiegend zyklonal
7	<b>NWa</b>	Nordwestlage, Mitteleuropa überwiegend antizyklonal
8	<b>NWz</b>	Nordwestlage, Mitteleuropa überwiegend zyklonal
...		
29	<b>TrW</b>	Trog Westeuropa
30	<b>Ü</b>	Übergangslage / Unbestimmt

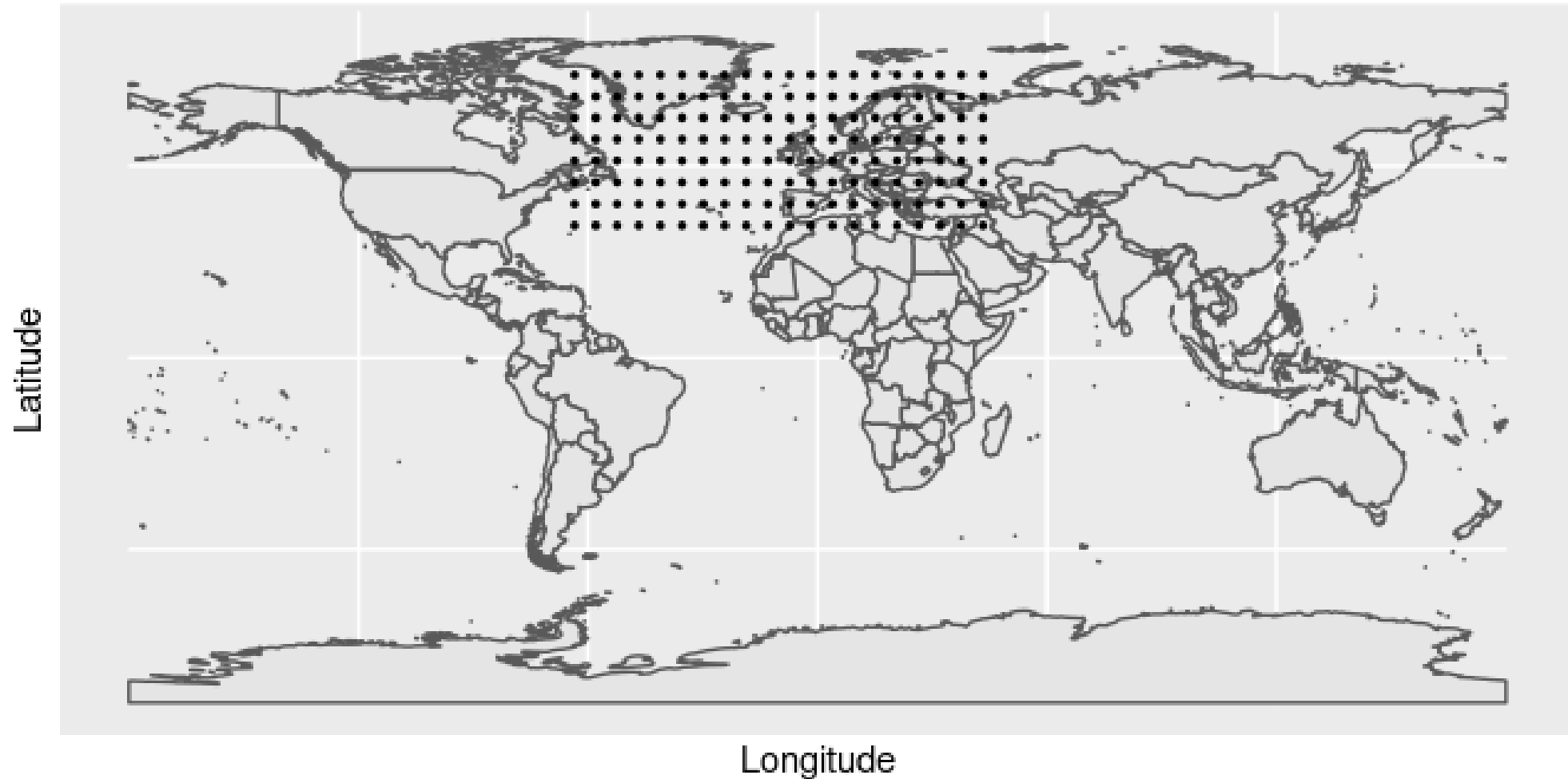
# Vorstellen des Projekts II



# Vorstellen des Projekts III

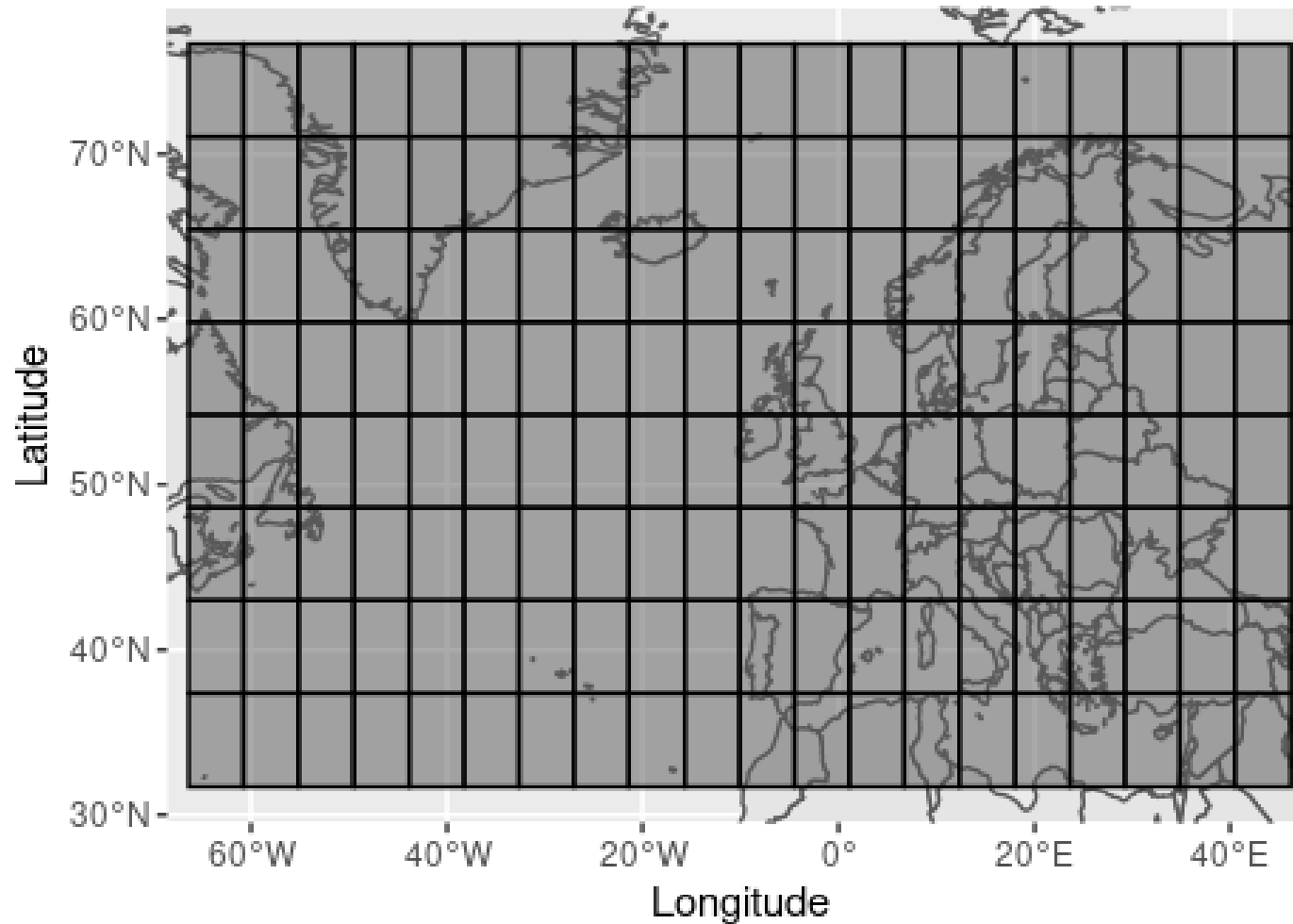
- Reanalyse Datensatz
  - Pro Tag Messungen an 160 Standorten zu 4 Zeitpunkten
    - Luftdruck in Pa auf Meeresspiegelhöhe (mslp)
    - Geopotential auf 500 hPa in  $\frac{m^2}{s^2}$  (geopot)
  - Für die Jahre 1900 bis 2010
  - Ohne Information zur herrschenden GWL am Tag
  - Standorte im 8x20 Grid über Europa und dem Nordatlantik

## Messpunkte auf einer Weltkarte





## Messpunkte



# Auszug aus dem Reanalyse Datensatz

	time	longitude	latitude	mslp	geopotential
1	1900-01-01 00:00:00	-63.562874	73.85311	100428.99	48268.86
2	1900-01-01 00:00:00	-63.562874	68.23695	100553.77	48770.82
3	1900-01-01 00:00:00	-63.562874	62.62077	99920.18	49171.14
4	1900-01-01 00:00:00	-63.562874	57.00457	100049.80	49487.83
. . .					
640	1900-01-01 18:00:00	43.312801	34.53973	102281.97	55097.32
641	1900-01-02 00:00:00	-63.562874	73.85311	99886.71	47843.04
. . .					
25946239	2010-12-31 18:00:00	43.312801	40.15595	101758.62	54154.39
25946240	2010-12-31 18:00:00	43.312801	34.53973	101400.51	54491.94

# Einführung in Clusteranalyse

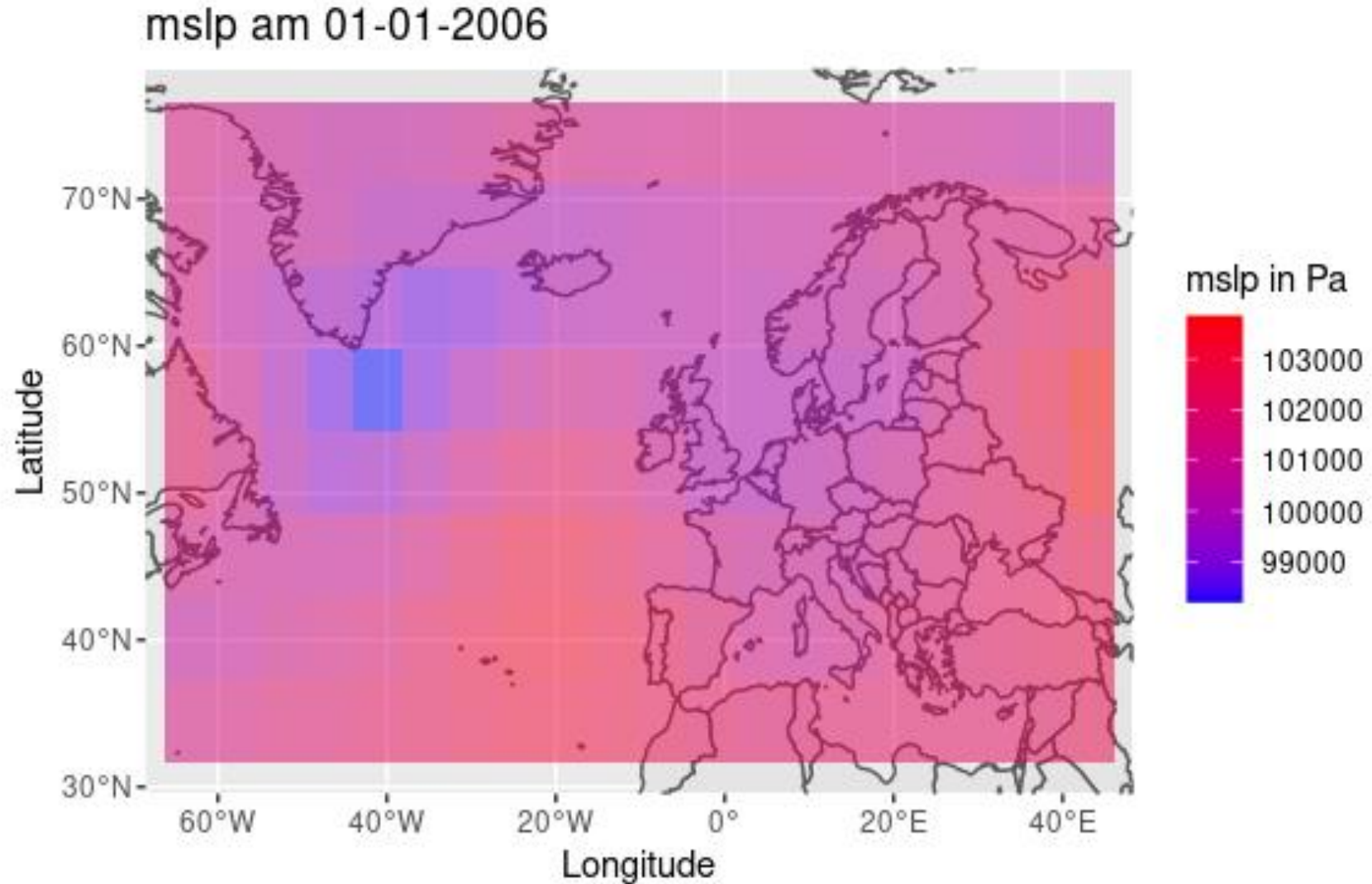
- Grundidee: Bildung von möglichst homogenen Gruppen, Cluster untereinander möglichst heterogen
- Clusteranalyse ist Verfahren des "unsupervised learning"
- Verschiedene Distanzmetriken
- Verschiedene Ansätze für Cluster
  - Optimale Partitionen
  - Dichtebasierte Verfahren
  - Und andere

# Ziele des Projekts

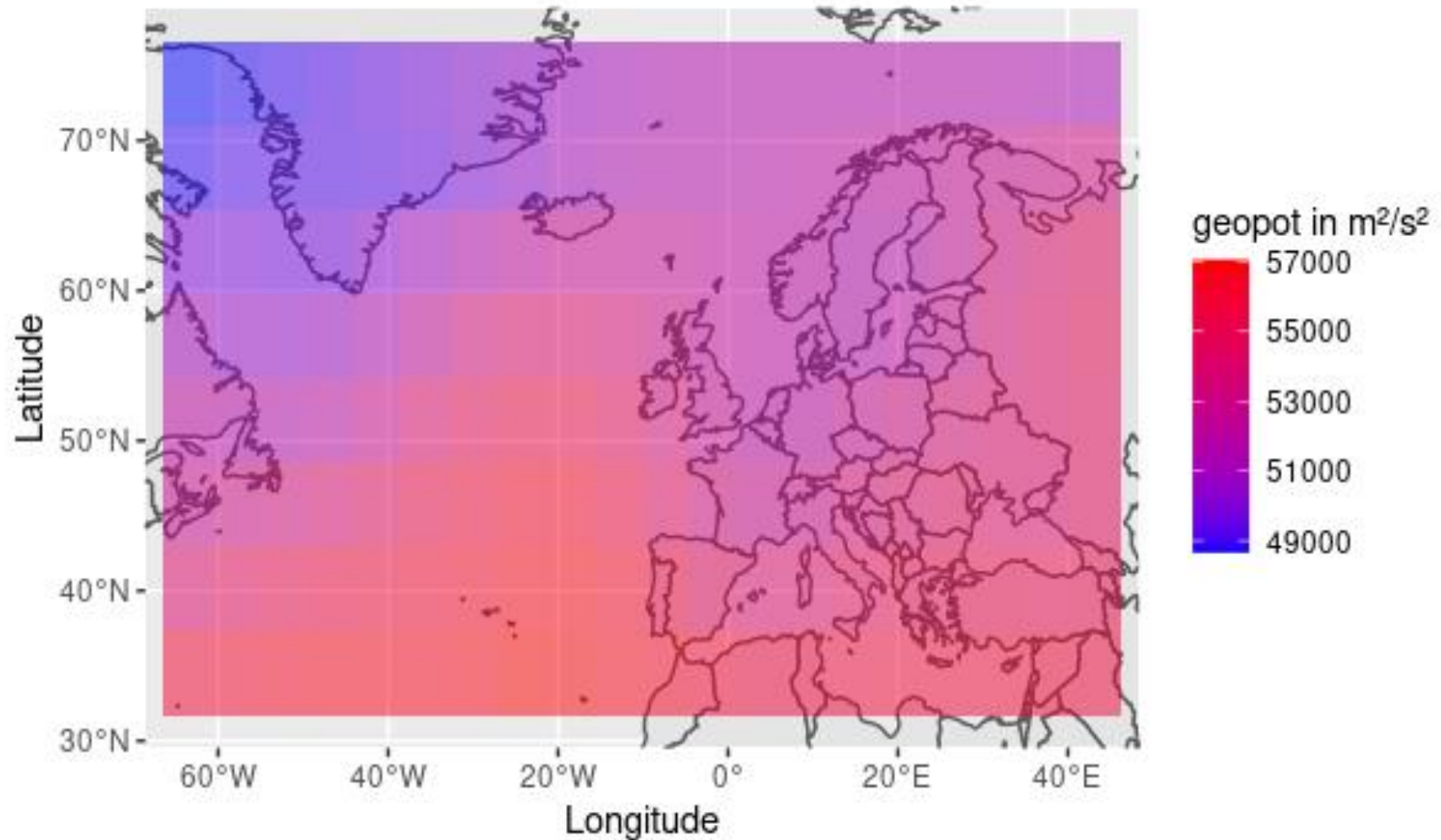
- Clustereinteilung der beobachteten Wetterdaten
  - Ein GWL soll sich in einem Cluster befinden
  - Anzahl Cluster < Anzahl GWLs
  - Berücksichtigung der räumlichen und zeitlichen Datenstruktur
- Vergleich der Cluster
  - Verteilung der GWLs in den Clustern
  - Vergleich der Zusammensetzung der einzelnen Cluster: max./min Luftdruck/Geopotential, Quantile, Ermittlung von Ausreißern, Stabilitätsprüfung?

# Probleme und Ansätze I

- Größe des Datensatzes
  - Erstmal Reduzierung auf 5 Jahre (2006-2010)
  - Grundsätzlich auf „Klimaperiode“ 1981-2010
- Anzahl der Dimensionen
  - Tagesdurchschnitt der 4 Messungen



## Geopotential am 01-01-2006

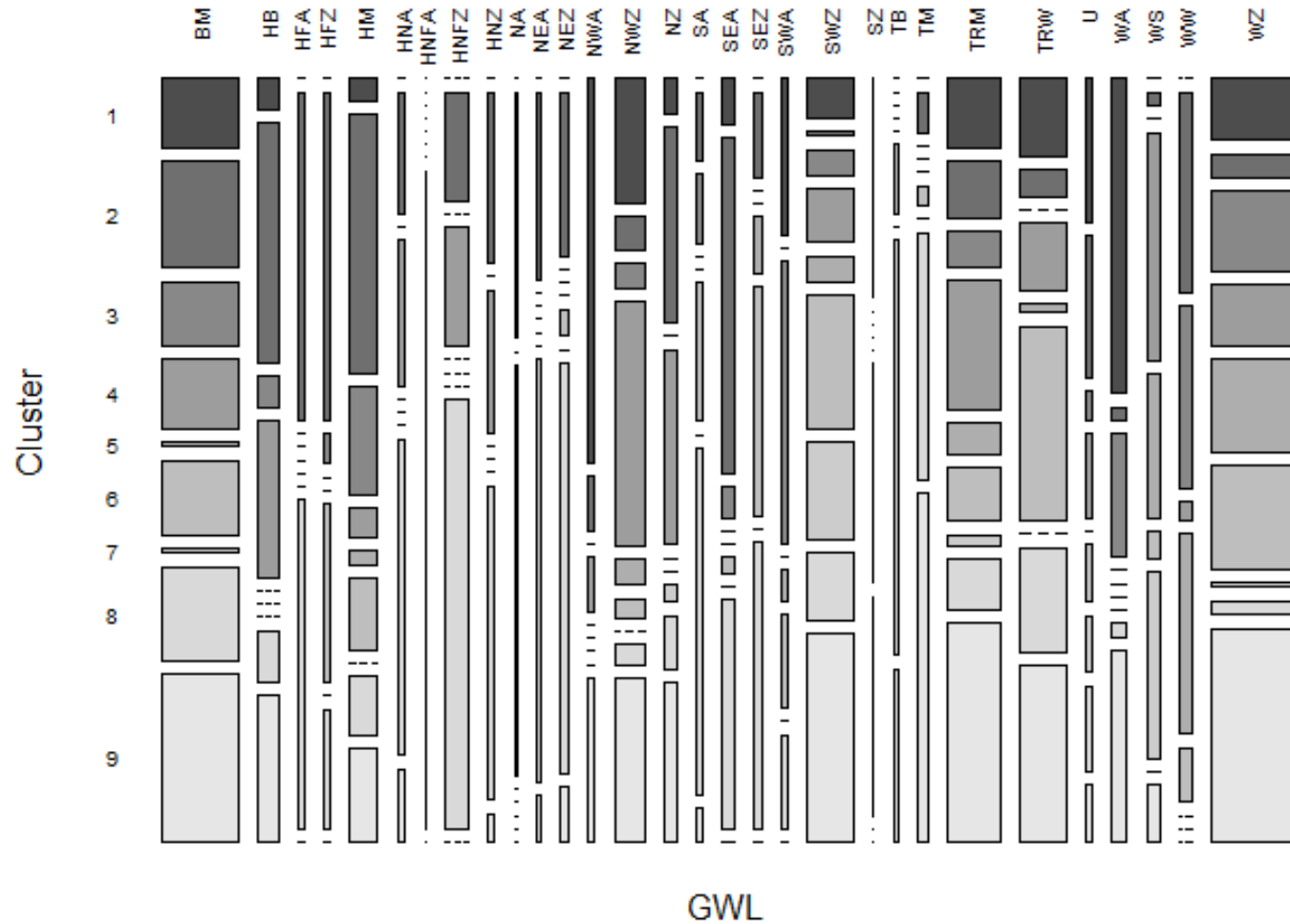


# Cluster I

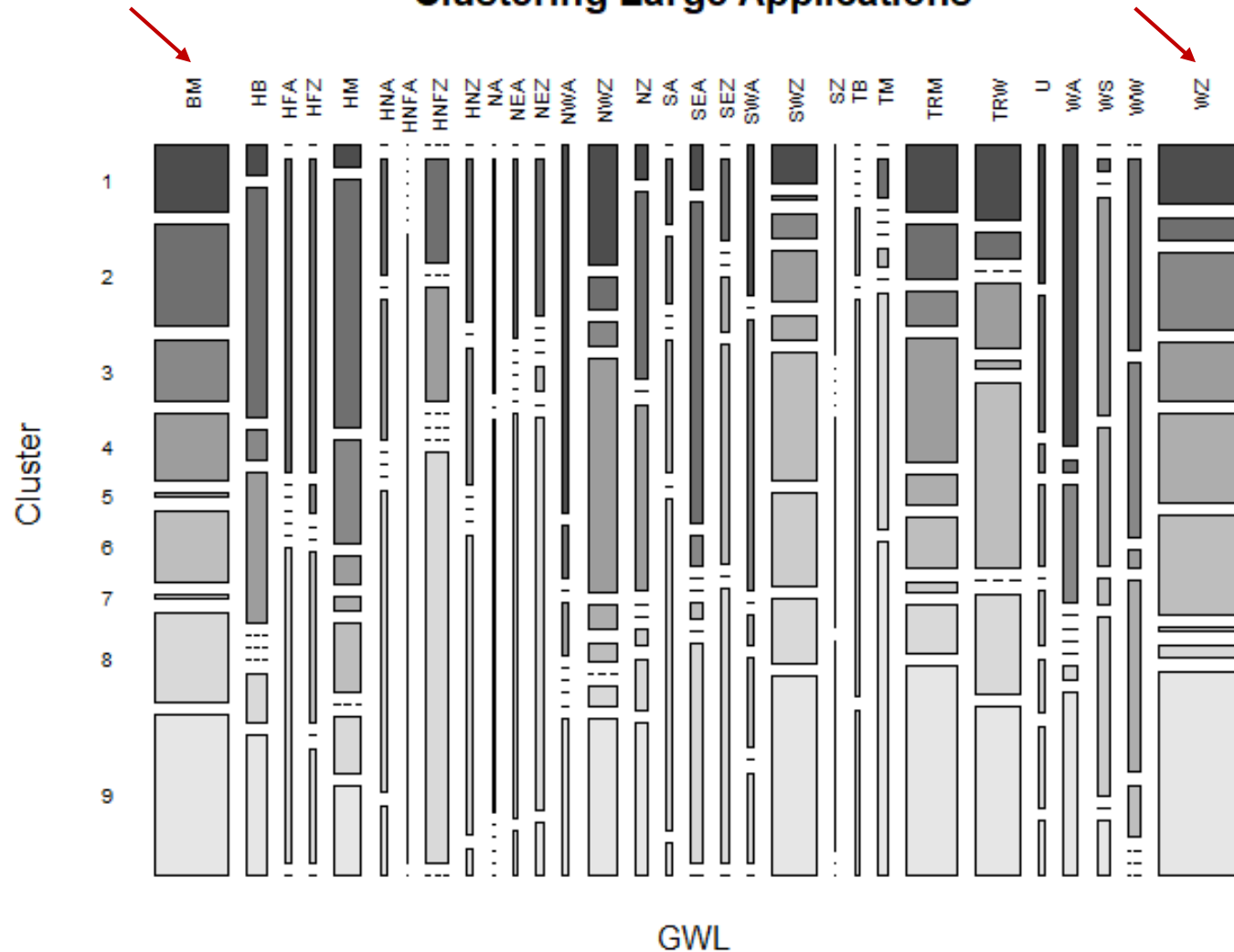
- Viele Beobachtungen und clustern mit hohen (320) Dimensionen
  - Algorithmus Clustering Large Applications (CLARA)
  - Euklidische Distanzmetrik
- Methodik
  - Stichprobe aus Datensatz ziehen und in  $k$  Cluster einteilen
  - Die restlichen Objekte den Clustern zuteilen, die am nächsten liegen
  - $N$  Wiederholungen und beste Variante auswählen



## Clustering Large Applications

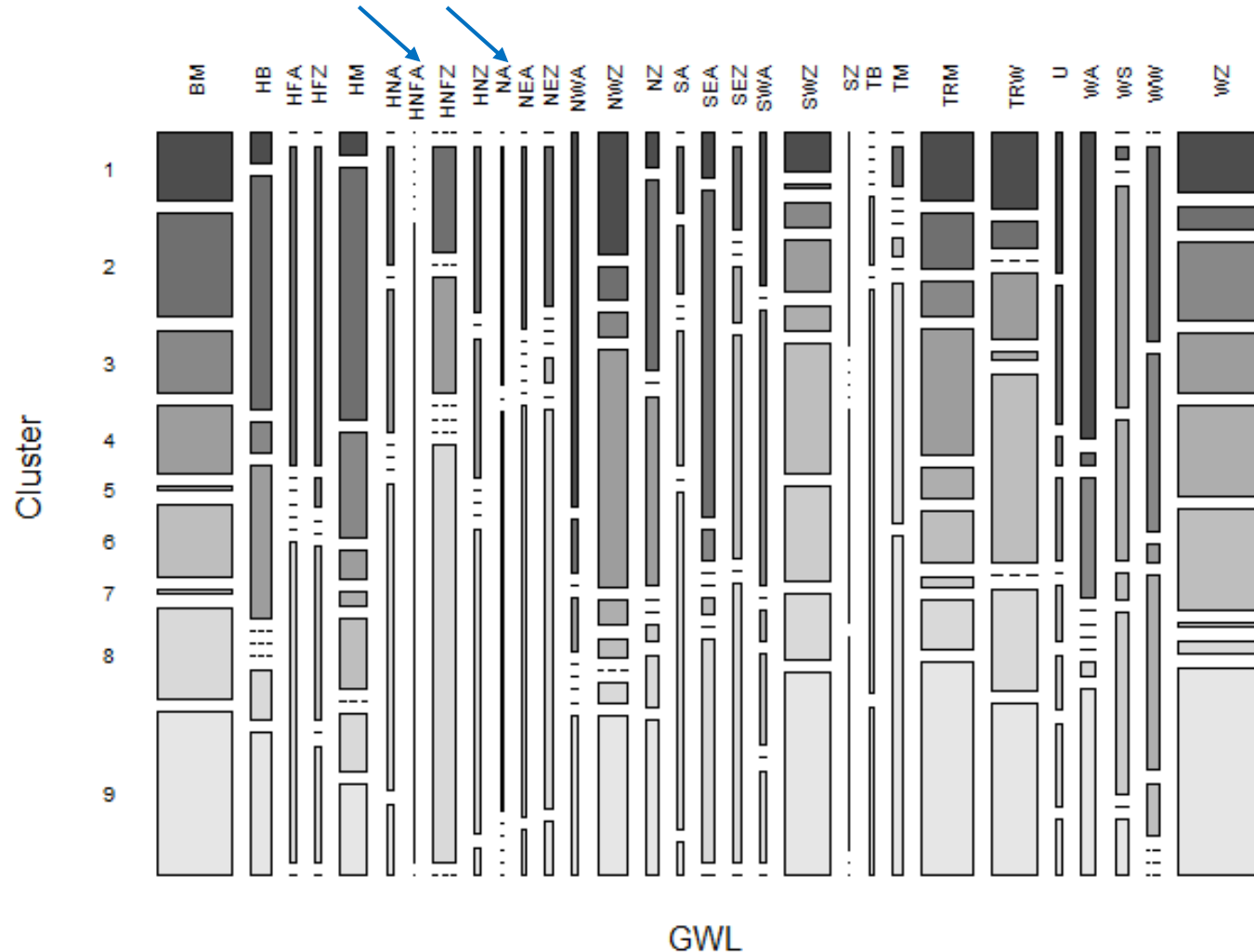


## Clustering Large Applications



➔ GWL, die häufig auftreten wie z.B. **BM oder WZ** sind in allen Clustern vertreten

## Clustering Large Applications



- ➔ GWL, die häufig auftreten wie z.B. BM oder WZ sind in allen Clustern vertreten
- ➔ GWL, die seltener auftreten wie z.B. NA oder HNFA lassen sich in ein bzw. zwei Cluster zuordnen

# Probleme und Ansätze II

- Korrelation zwischen Variablen
  - Luftdruck und Geopotential
  - Standort

➡ Mahalanobisdistanz

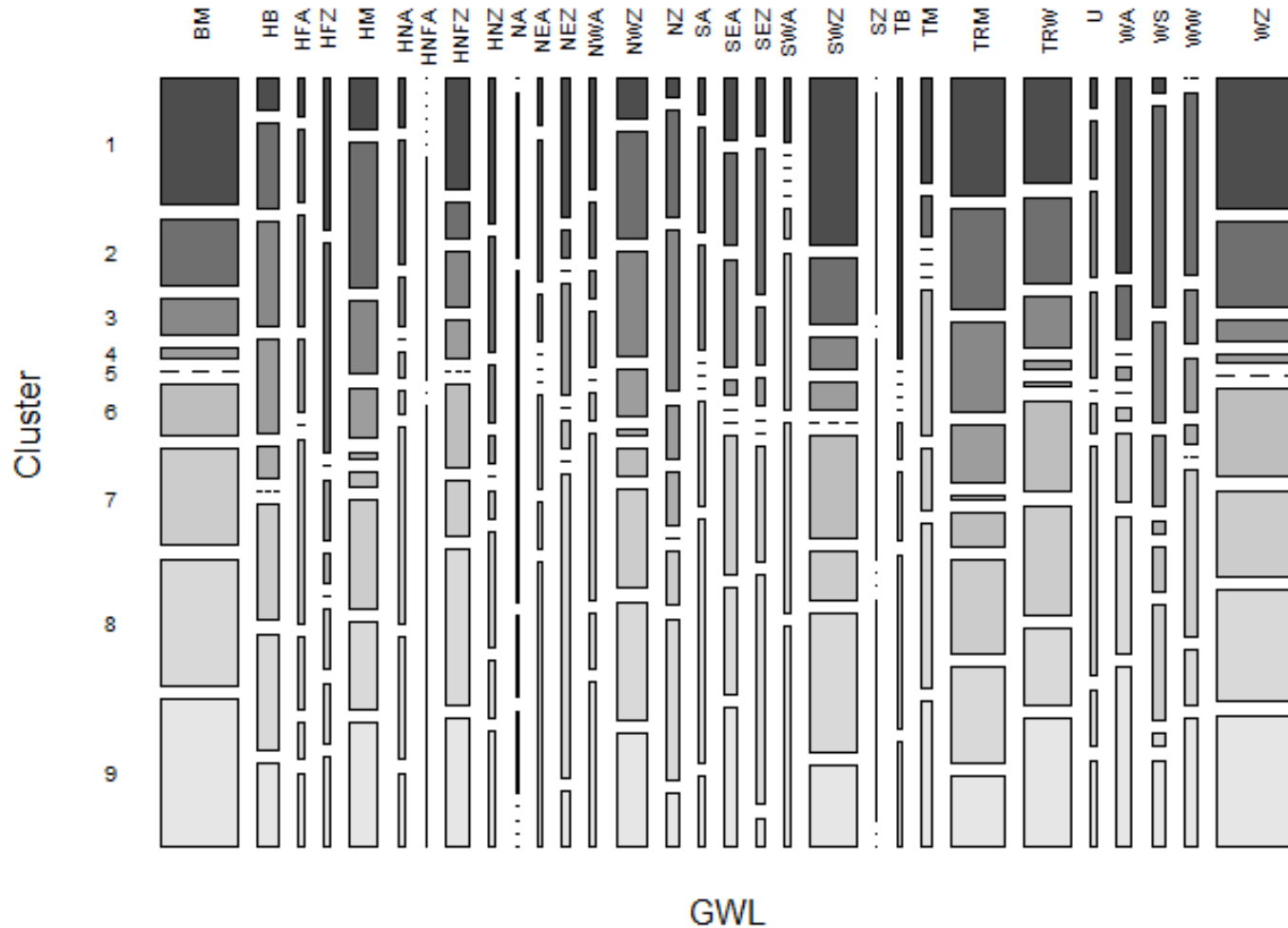
# Mahalanobisdistanz

- Distanz zwischen zwei Punkten im multivariaten Raum
- Geeignet für korrelierte Daten
- $MD(x, y) = \sqrt{(y - x)^T \cdot C^{-1} \cdot (y - x)}$

mit C als Kovarianzmatrix

# Cluster II - K-Means-Algorithmus

- Gehört zu den Partitionierenden Verfahren
- Varianzkriterium: Minimieren der Gesamtsumme der quadrierten Abweichungen
- Vorgehen: 1. Vorgeben einer Anfangspartition
  2. Berechnen der jeweiligen Gruppenschwerpunkte
  3. Verschieben der Elemente in die nächstgelegene Gruppe
  4. Wiederholen der Schritte 2 und 3 bis kein Element mehr die Gruppe wechseln muss

**Cluster mit k-Means und Mahalanobis**

➔ Ähnliche Aufteilung wie bei CLARA

# Probleme und Ansätze III

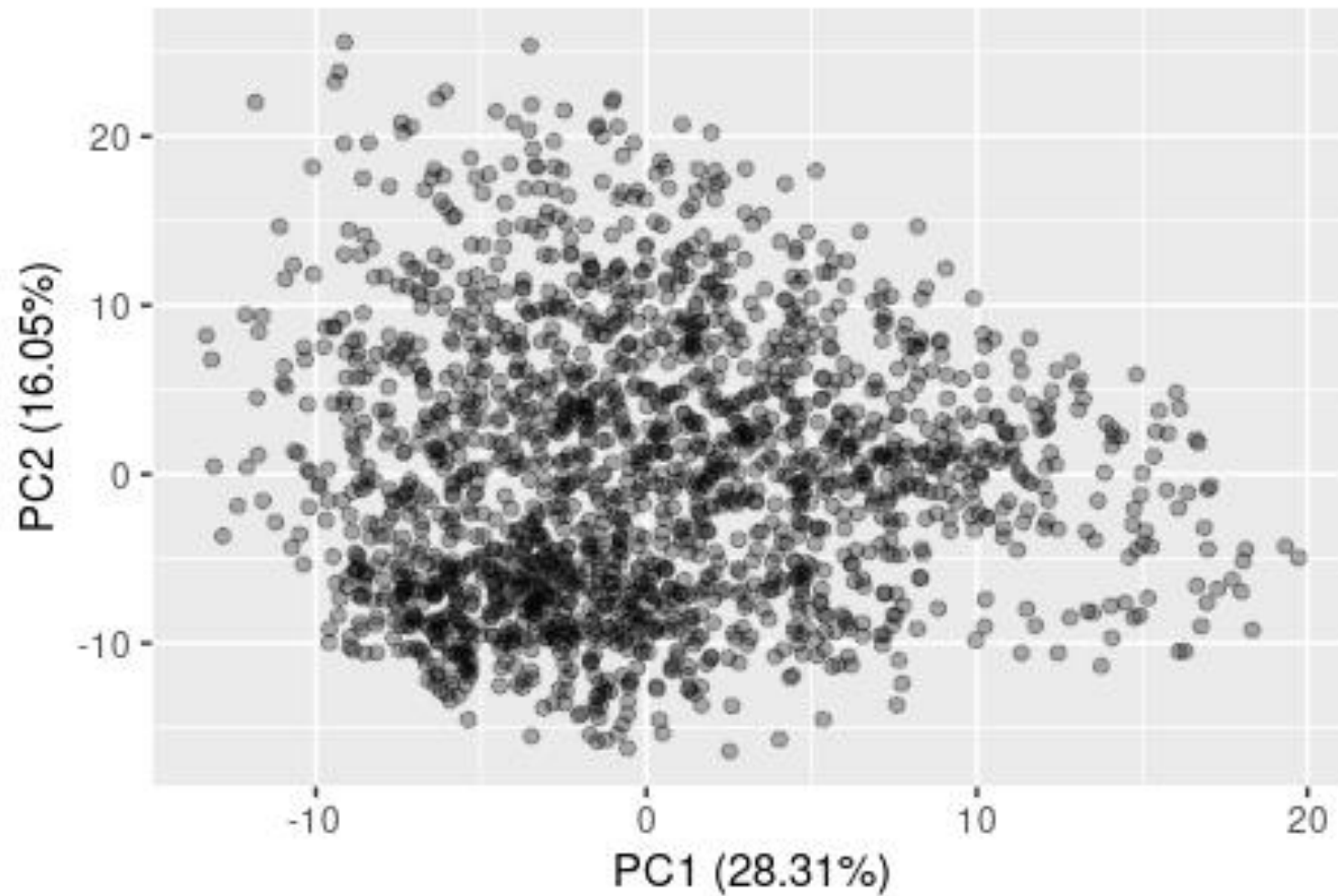
- Anzahl Dimensionen

## → Principle Component Analysis

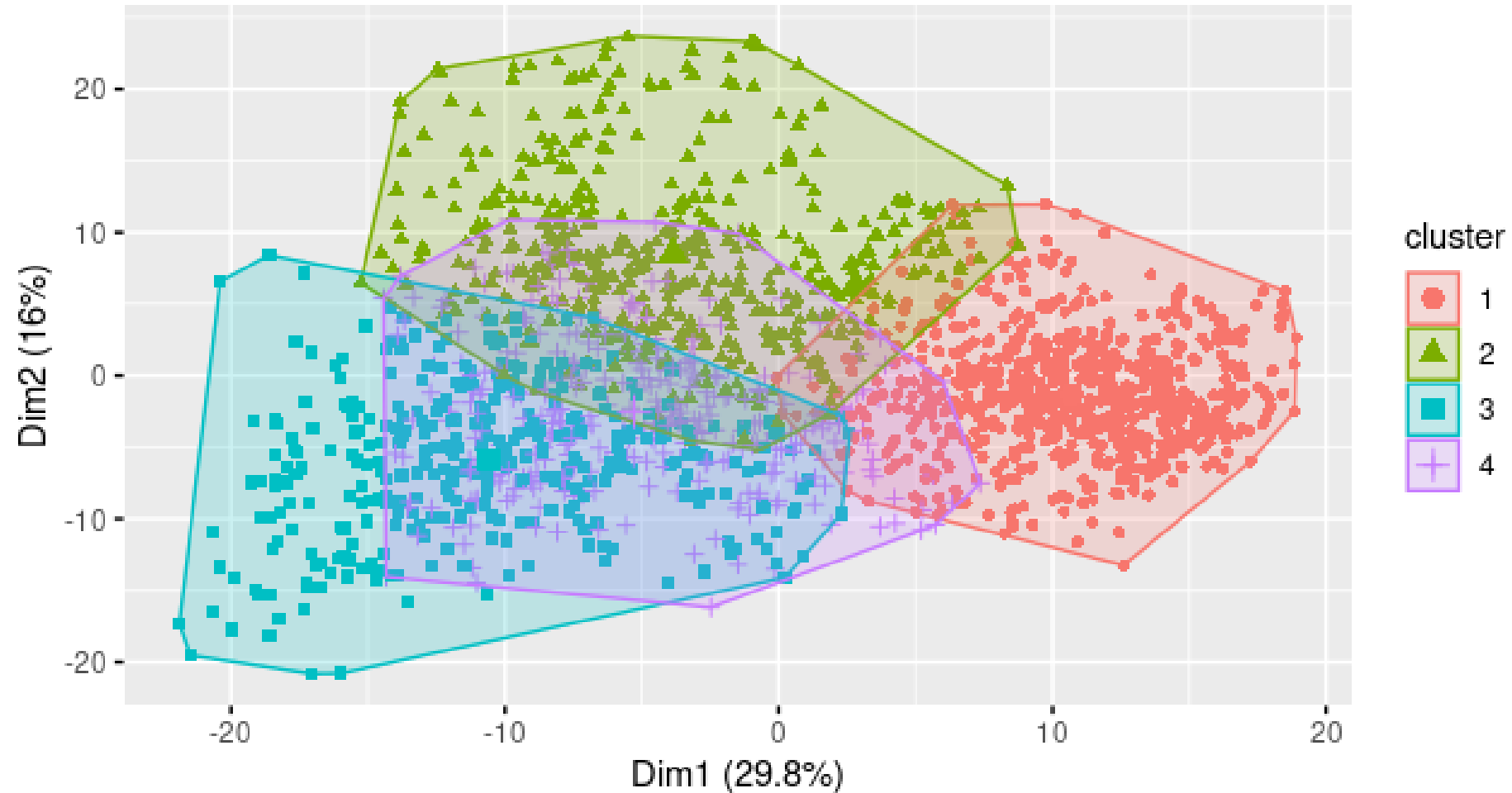
- Aus Eigenvektoren der Kovarianzmatrix
- Erklären der meisten Varianz mit weniger Dimensionen
- Hier 85% der Varianz mit 10 Dimensionen erklärt



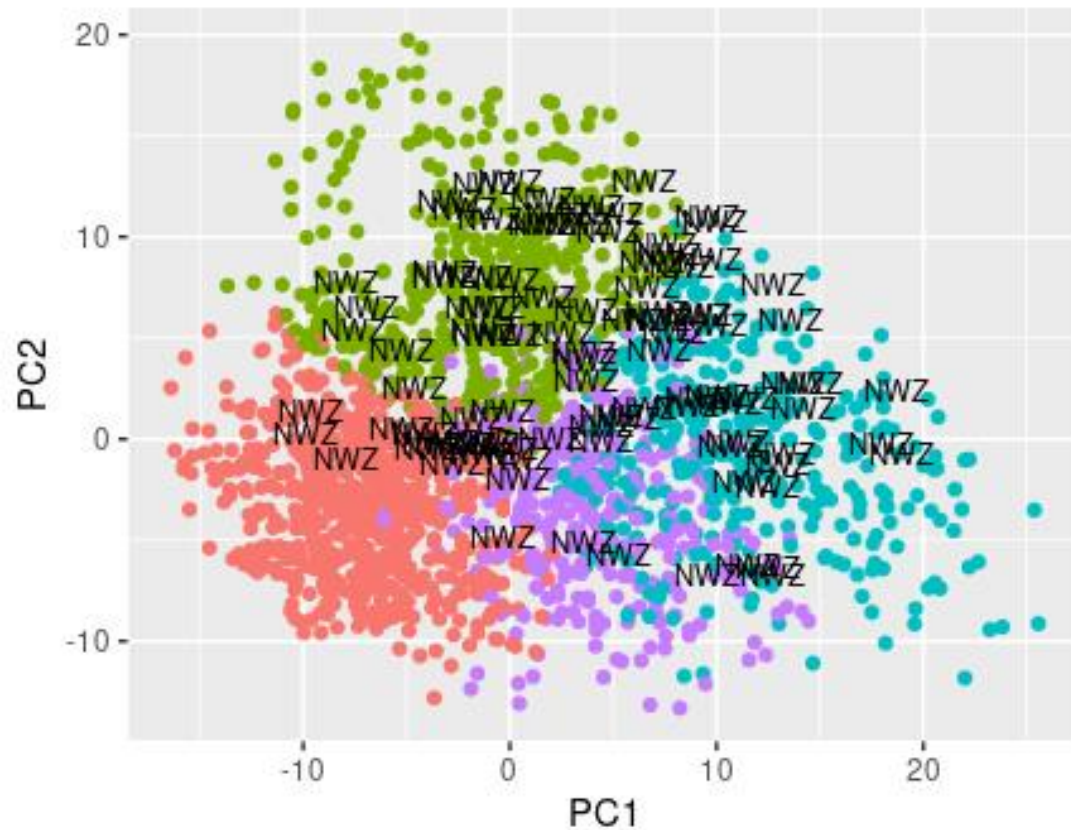
Ersten zwei PC (skaliert)



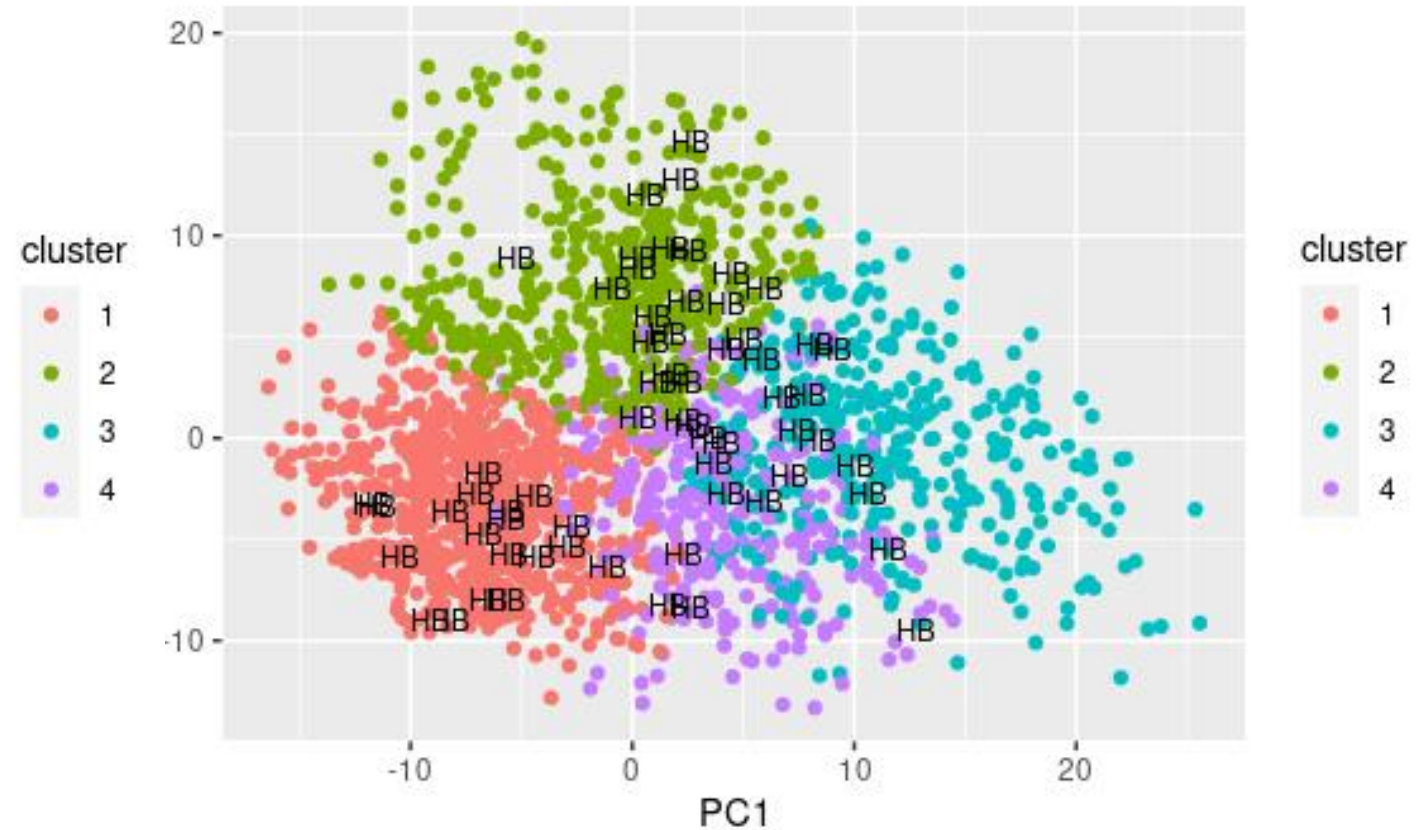
Cluster plot



kmeans 10pc



kmeans 10pc

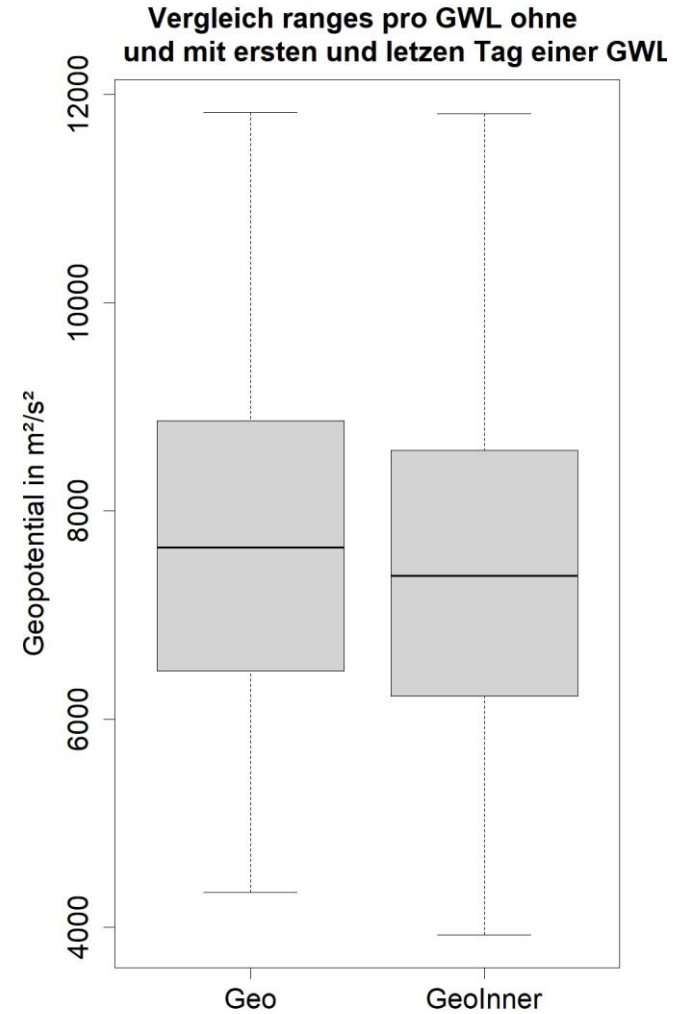
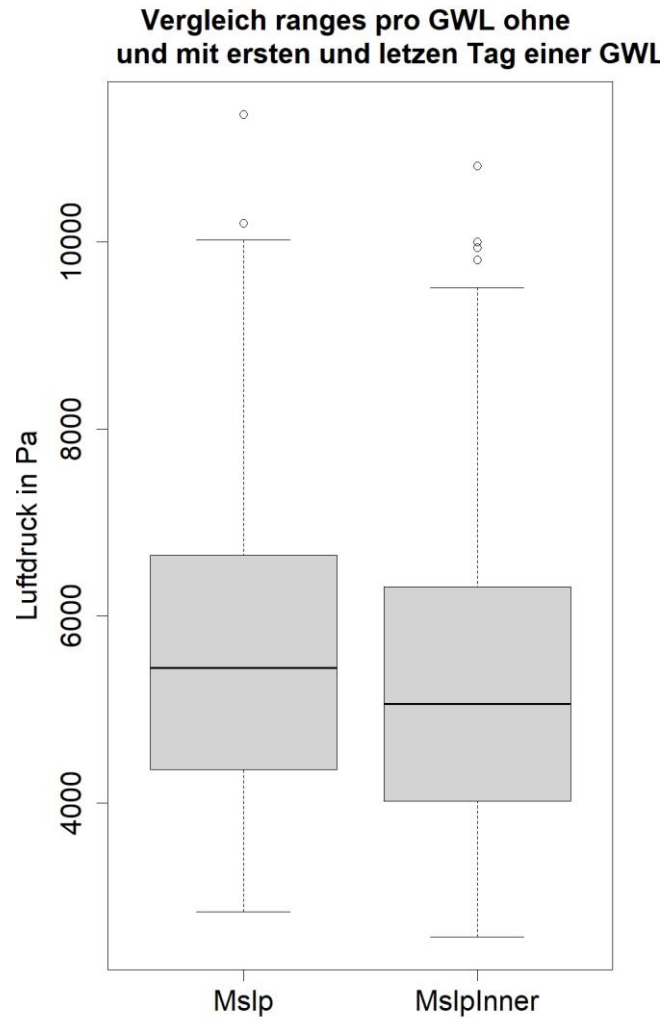


## Mosaikplot GWL zu Cluster durch PCA&kmeans



# Probleme und Ansätze IV

- Örtliche Komponente sehr wichtig
    - Art von Pattern Recognition?
  - Definieren einer GWL normalerweise anhand eines kleineren geographischen Ausschnittes
    - Gewichtung von Europa?
  - Unterschiede der ersten und letzten Tage einer GWL
  - Saisonale Unterschiede in GWL
    - Saisonbereinigung?
- *Datensatz herunterbrechen auf diskrete Variablen*

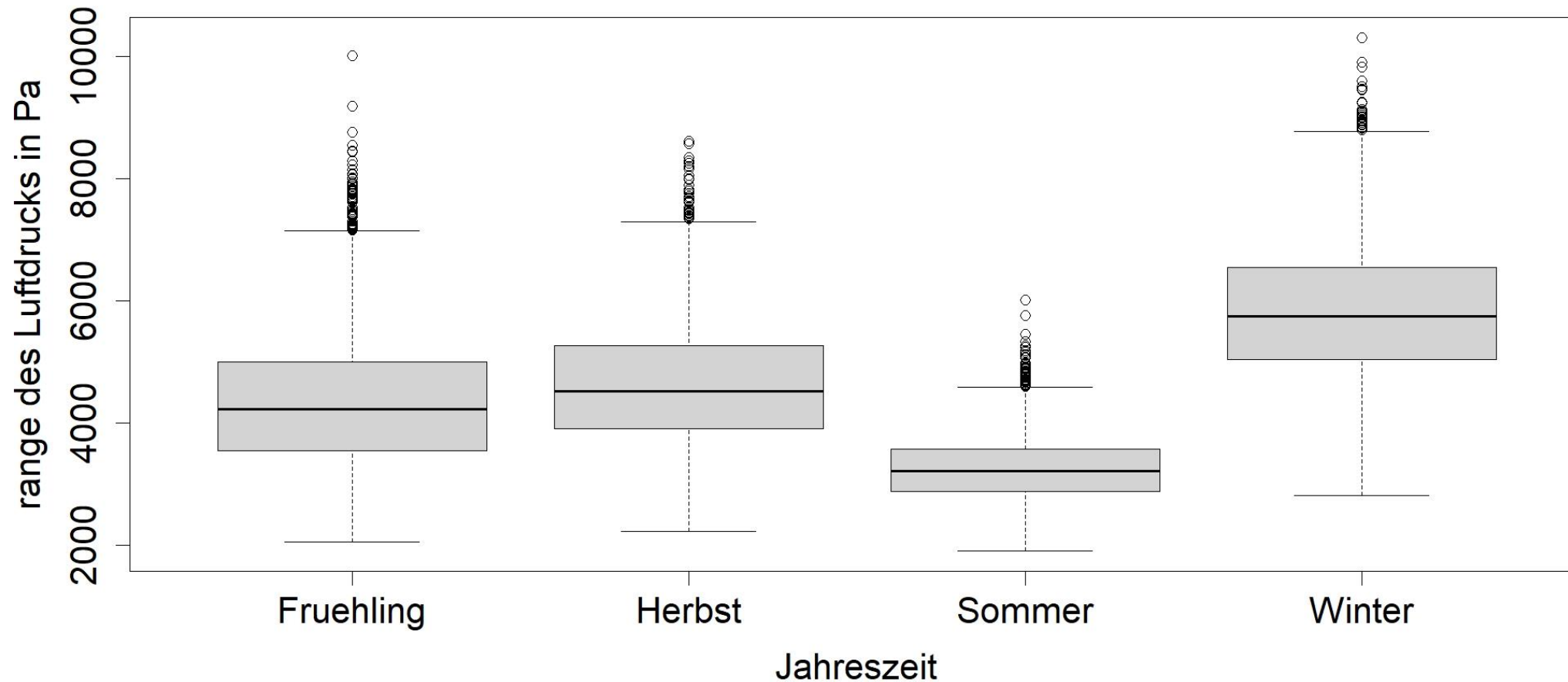


# Probleme und Ansätze IV

- Örtliche Komponente sehr wichtig
    - Art von Pattern Recognition?
  - Definieren einer GWL normalerweise anhand eines kleineren geographischen Ausschnittes
    - Gewichtung von Europa?
  - Unterschiede der ersten und letzten Tage einer GWL
  - Saisonale Unterschiede in GWL
    - Saisonbereinigung?
- *Datensatz herunterbrechen auf diskrete Variablen*



### range des Luftdrucks in Abhängigkeit der Jahreszeiten





# Probleme und Ansätze IV

- Örtliche Komponente sehr wichtig
    - Art von Pattern Recognition?
  - Definieren einer GWL normalerweise anhand eines kleineren geographischen Ausschnittes
    - Gewichtung von Europa?
  - Unterschiede der ersten und letzten Tage einer GWL
  - Saisonale Unterschiede in GWL
    - Saisonbereinigung?
- *Datensatz herunterbrechen auf diskrete Variablen*

# Methodik

- Erstellen eines neuen Datensatzes
  - Extrahieren von neuen Variablen, zum Teil auf diskreter Ebene
- Vorteile dieses Vorgehens
  - Reduzierung der Dimensionen
  - Besseres Einbeziehen der örtlichen Komponente
  - Einbringen von anderen möglichen Variablen

# Extrahierte Variablen

Variable	Erklärung	Metrik
Zeitpunkt	Evtl. für Saisonbereinigung	Kategorial
Minimum/Maximum	Minimaler/Maximaler Wert am Tag	Numerisch (evtl kategorial)
Quadrant vom Minimum/Maximum	In welchem Bereich befindet sich das Tief/Hoch? Karte aufgeteilt in X Felder • Europa feiner Unterteilt?	Kategorial oder geographischer Abstand der Mittelpunkte der Quadranten
Range der Parameter	Abstände der Maxima und Minima	Numerisch oder kategorial
Abstand Hoch-tief	Geographischer Abstand zwischen Maximalem und Minimalem Wert	Numerisch oder Kategorial

# Cluster III - Filtern pro Tag

- Tagesmesswerte besser in “Gebiete” unterteilen
  - Örtliche Komponente besser einbringen
  - Typische Merkmale der GWLs extrapolieren

→ Tage filtern durch Spatial Clustering

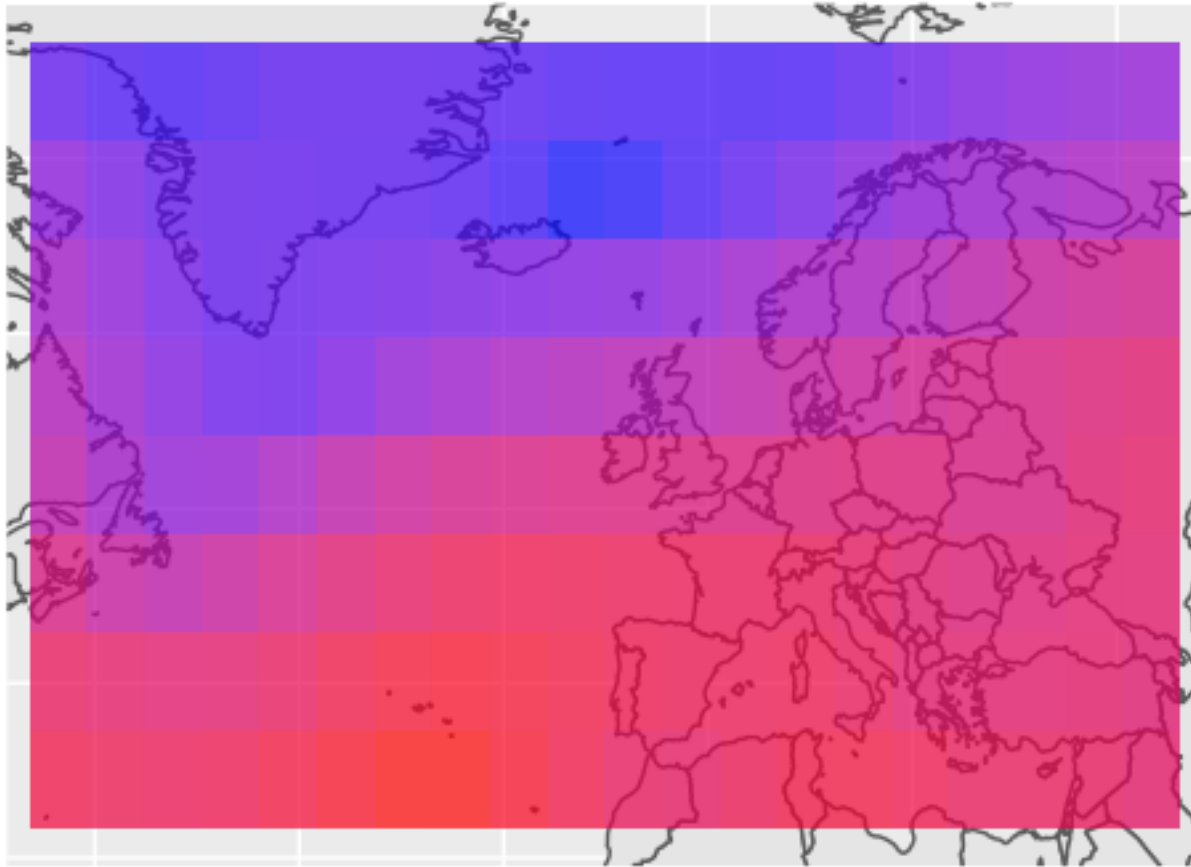
# DBSCAN

- Dichtebasierte räumliche Clusteranalyse mit Rauschen
- Zusammenhängende Gebiete ähnlicher Messwerte
  - z.B. „Hoch“- und „Tiefdruckgebiet“
  - Diskrete Clusterzugehörigkeit statt stetigen Messwerten

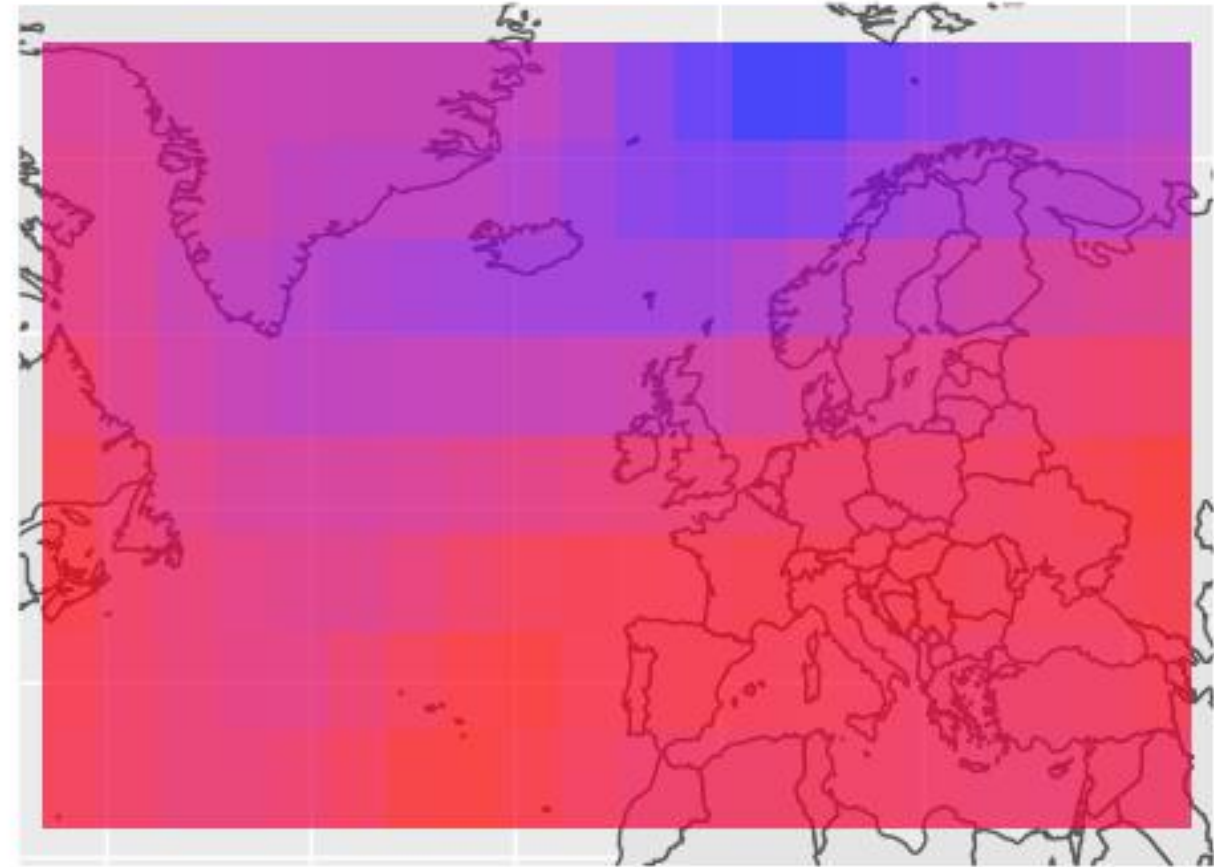
# Cluster III - Filtern pro Tag

- Folgendes als Beispiel anhand von dem 12.12.2006

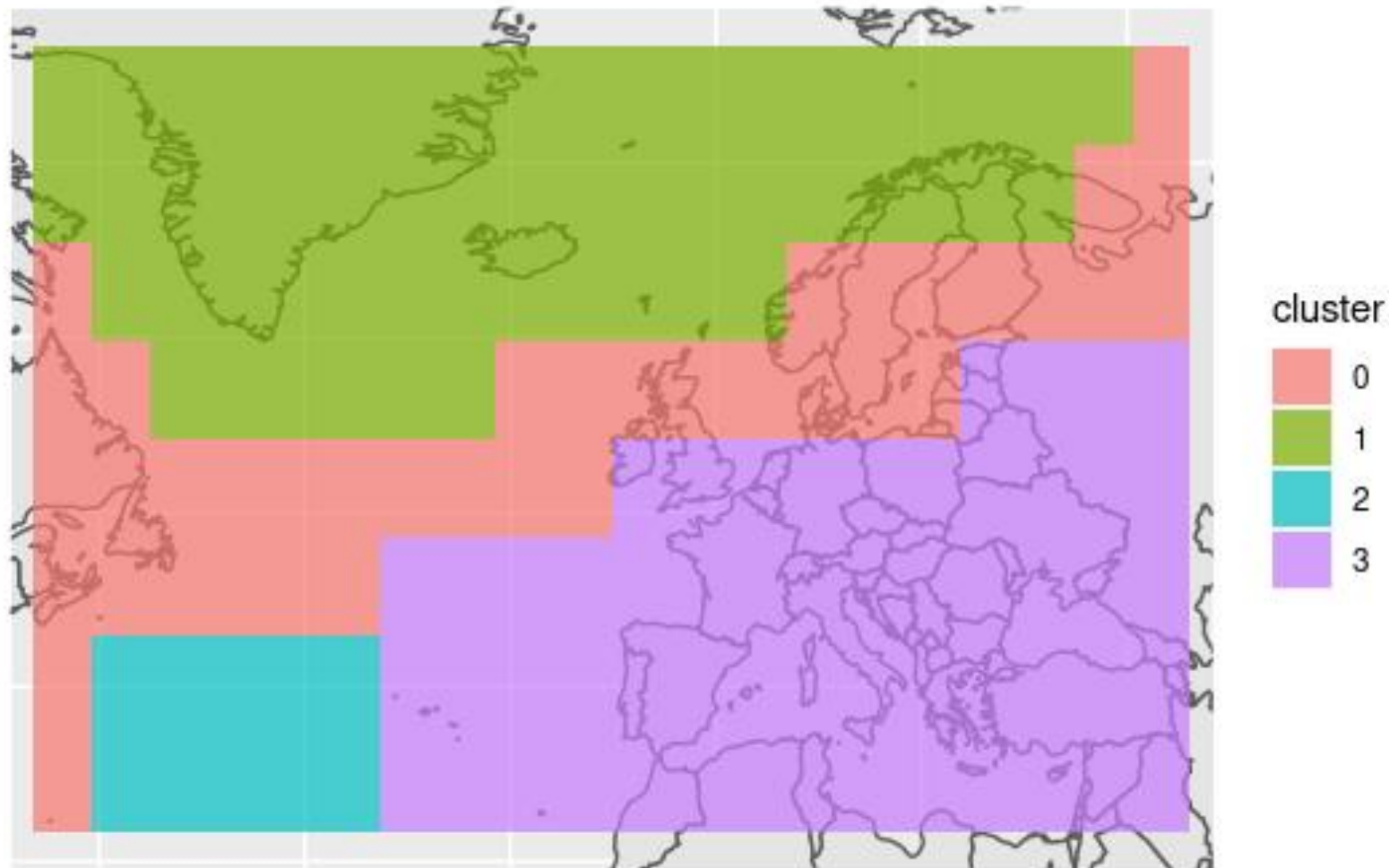
geopot am 2006-12-12



mslp am 2006-12-12

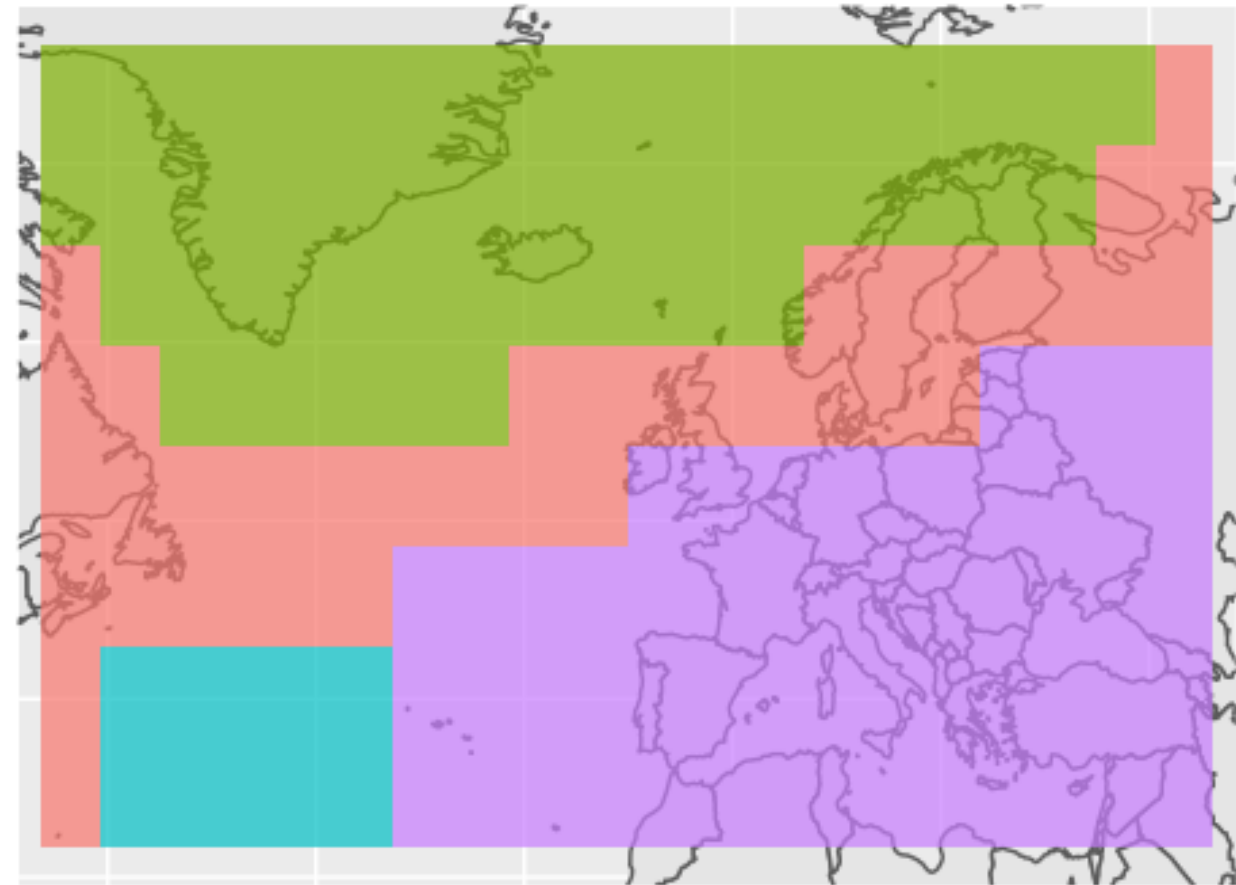
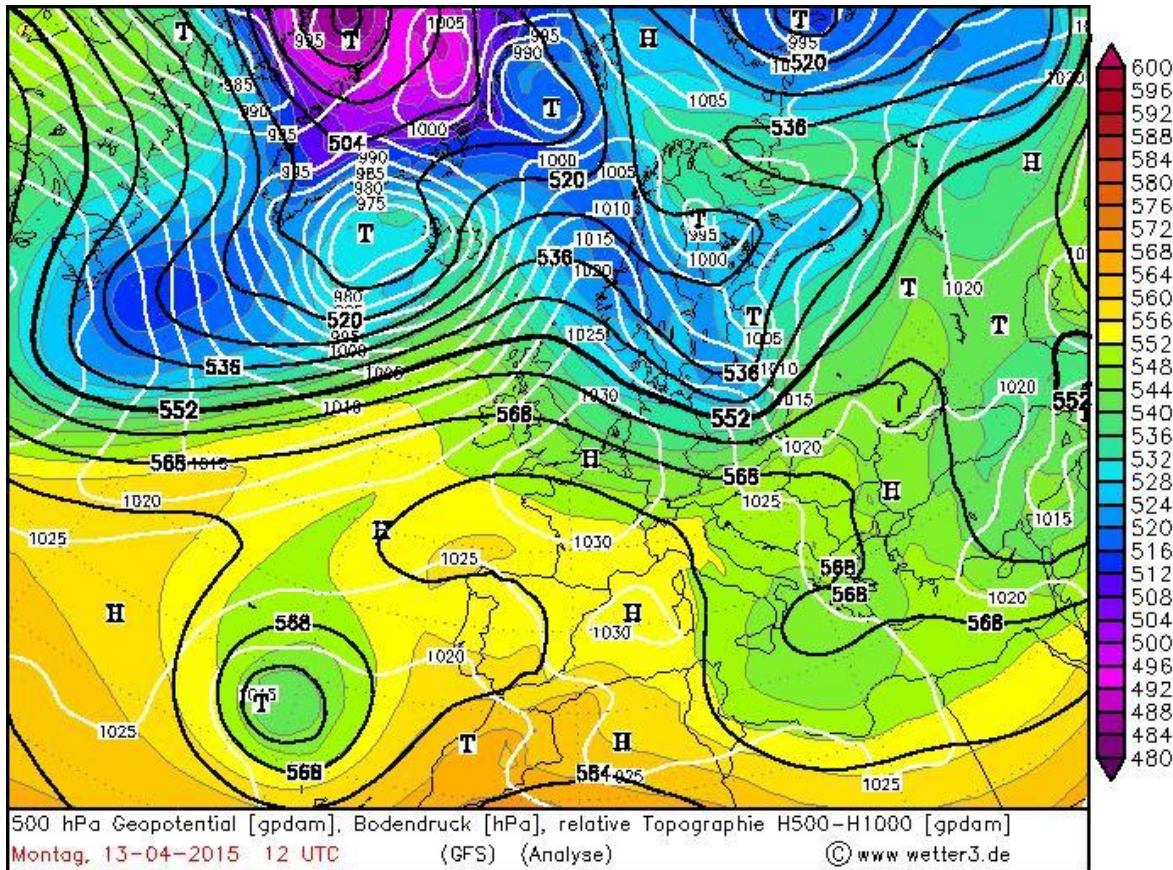


dbscan am 2006-12-12





## GWL ist WA (Westeuropa antizyklonal)



<http://www.schulbiologiezentrum.info/Wetter%20Materialien/Gro%DFwetterlagen%20Material.pdf> - 20.12.2020 2:20Uhr

# Cluster III - Filtern pro Tag

- Variablen extrahieren
  - Definieren eines „max“ und eines „min“ Gebietes

Parameter	Variable	Erklärung	Metrik
Gesamtcluster	Anzahl Cluster		kategorial
Für Max und Min Cluster	Größe des Clusters	Anzahl Punkte im Cluster	numerisch
Für Max und Min Cluster	Räumliche Lage	x Punkte des Clusters liegen in Quadrant y	numerisch

# Probleme

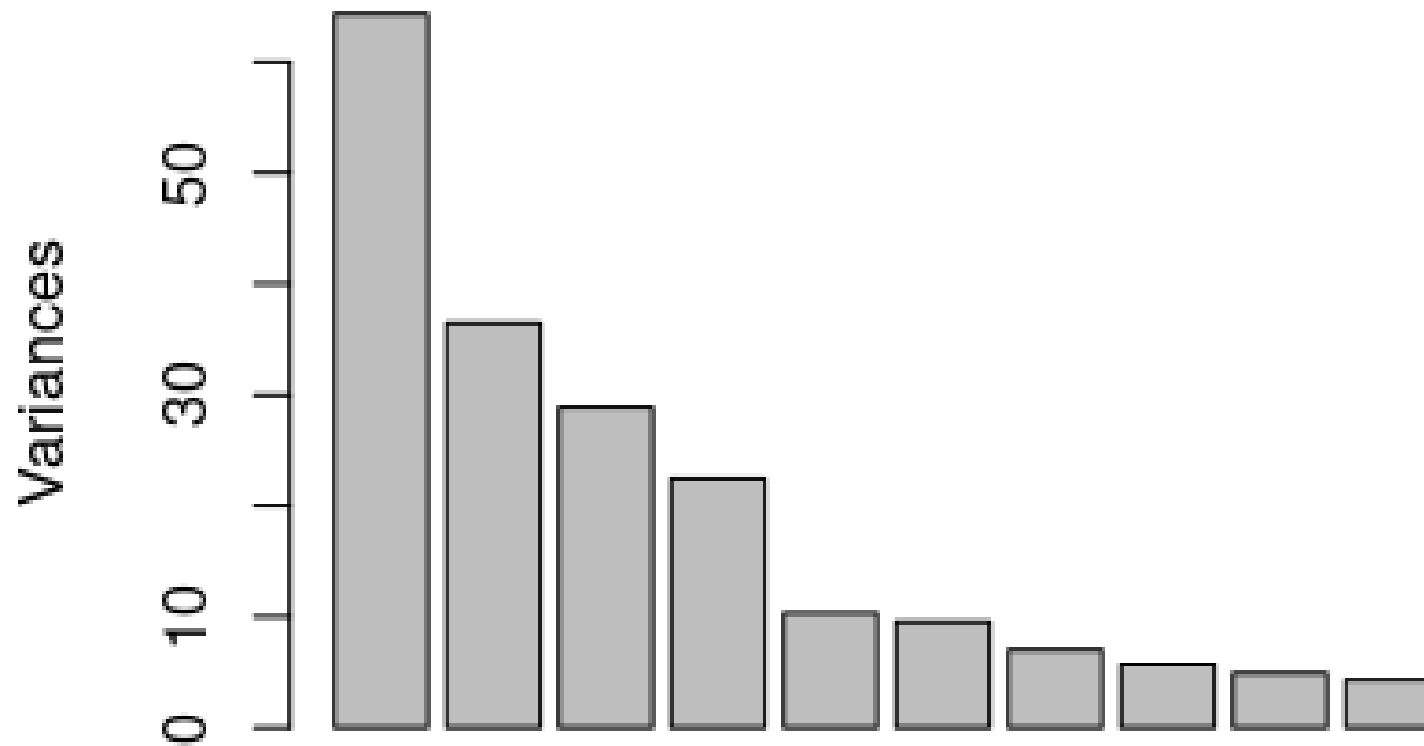
1. Viele Dimensionen (1280 Dimensionen über ca 40.000 Beobachtungen)
2. Große Auswahl an Clusteralgorithmen und Distanzmetriken
3. Wichtigkeit der örtlichen Komponente
4. GWL werden auch anhand von Variablen definiert, die uns nicht zur Verfügung stehen (z.B. Strömungsrichtung)
5. Variablen außerhalb der erhobenen Daten sind auch von Interesse (z.B. Saison, Gewichtung von Europa)

# Anhang

# Auszug aus dem GWL Datensatz

	JAHRMONAT	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1201	200001	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	SWA	SWA	SWA	SWA
1202	200002	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ	WZ
1203	200003	WZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ	NWZ
1204	200004	TM	TRW	TRW	TRW	TRW	HB	HB	HB	HB	HB	TB	TB	TB	SWZ
1205	200005	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HNFA	HM	HM
1206	200006	SWA	SWA	SWA	SWA	SWA	BM	BM	BM	BM	BM	BM	WA	WA	WA
1207	200007	SWZ	SWZ	SWZ	SWZ	SWZ	SWZ	SWZ	TRM	TRM	TRM	TRM	TRM	TRM	TRM
1208	200008	WZ	WZ	WZ	WZ	BM	BM	BM	BM	BM	BM	BM	BM	BM	BM
1209	200009	BM	TRM	TRM	TRM	TRM	WA	WA	WA	WA	WA	WA	WA	WA	WA
1210	200010	TRW	TRW	BM	BM	BM	BM	BM	BM	TB	TB	TB	TB	TB	SA
1211	200011	WZ	WZ	WZ	WZ	WZ	TB	TB	TB	TB	TB	TB	TB	TB	TRW

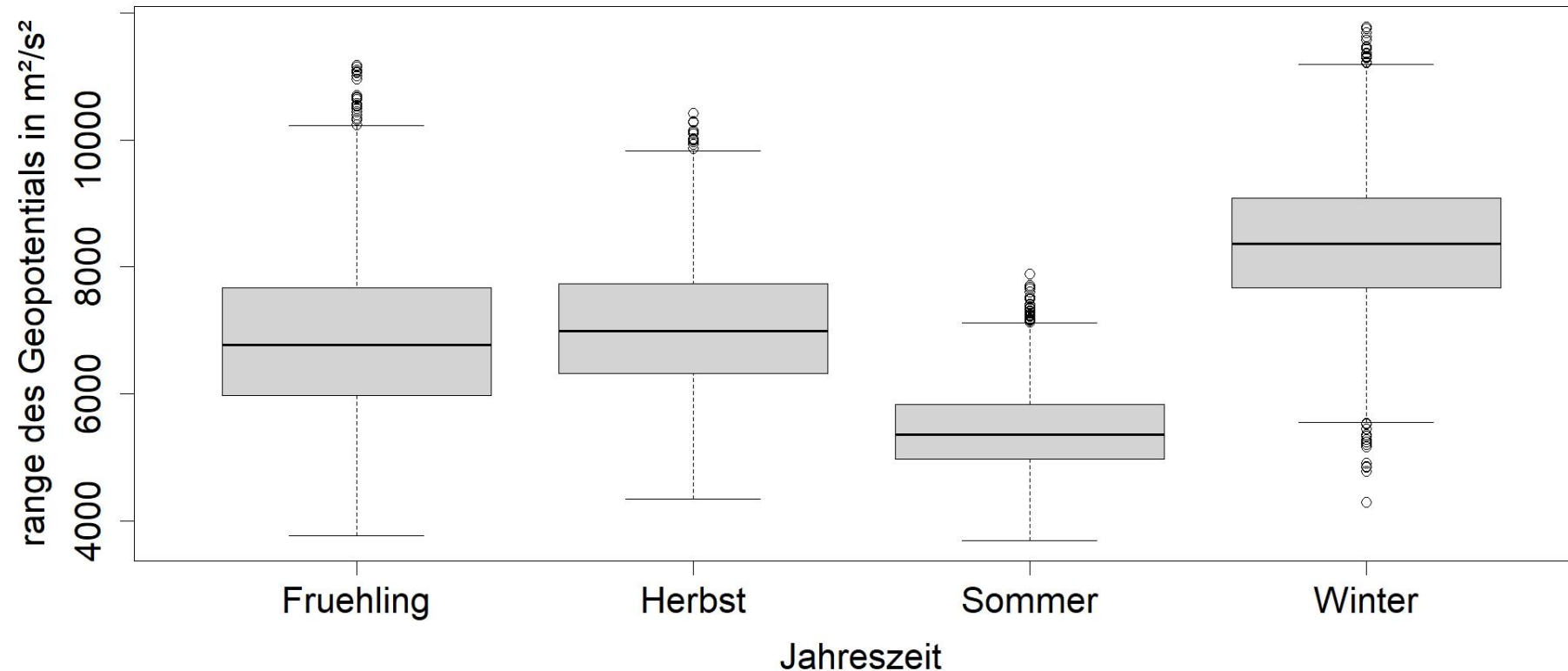
## Screepplot PCA

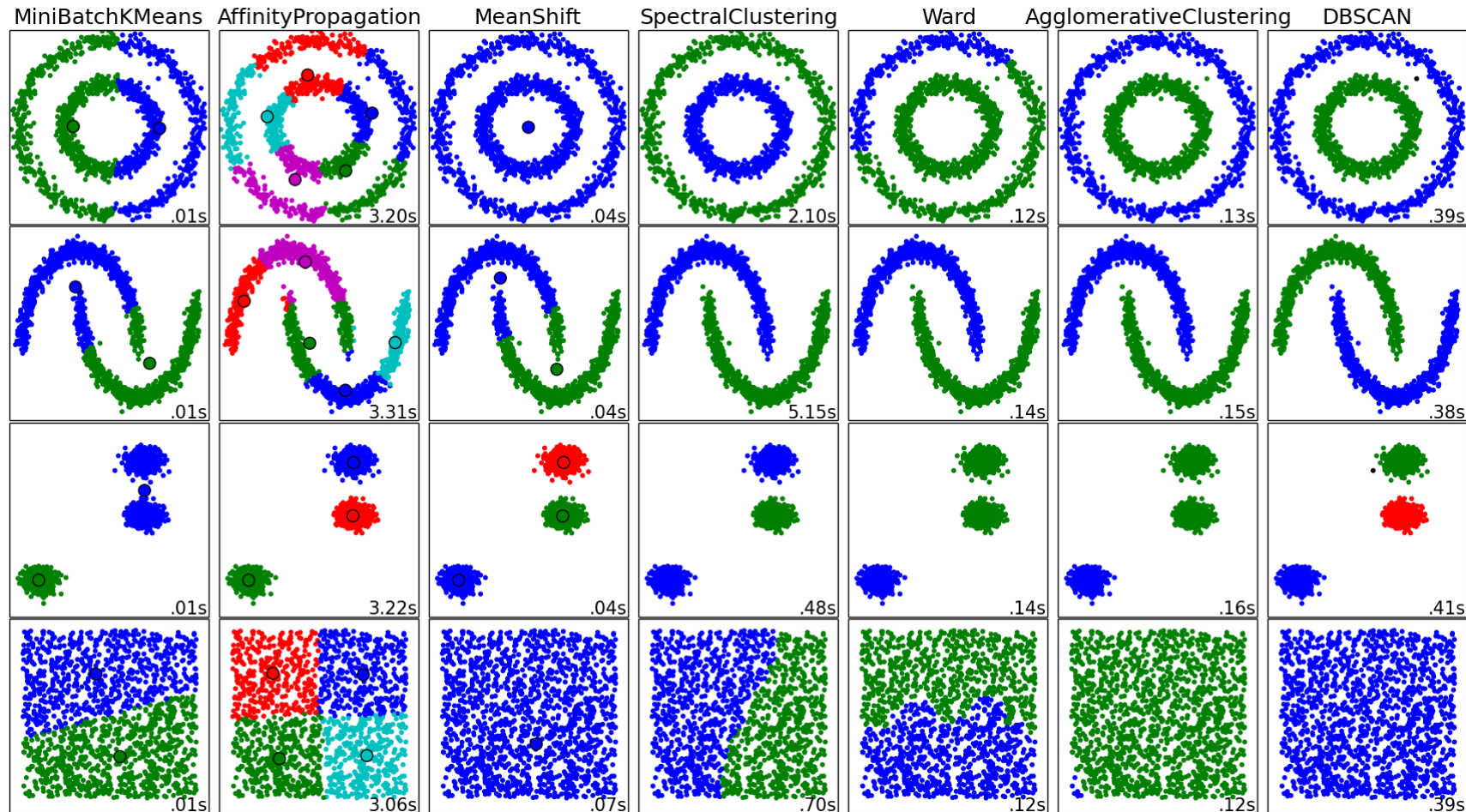




# Probleme und Ansätze IV

range des Geopotentials in Abhängigkeit der Jahreszeiten

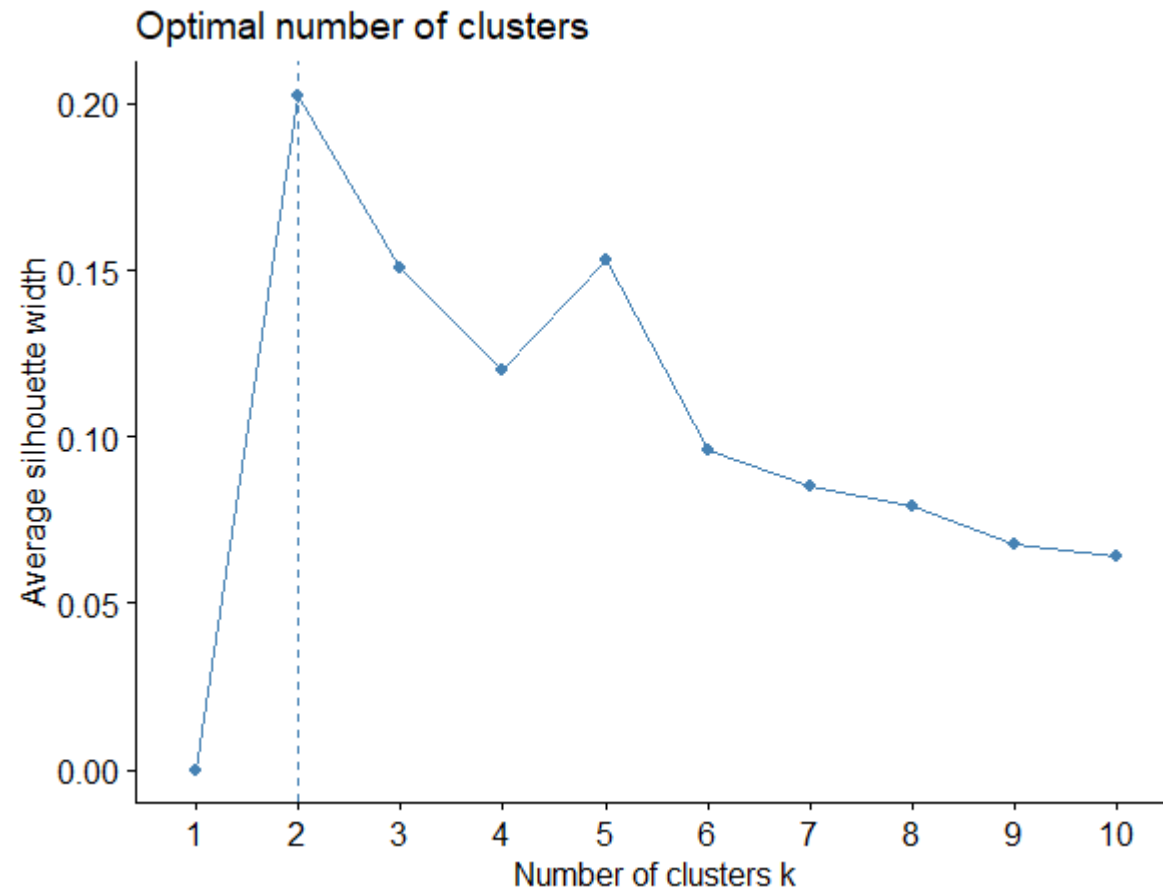




[https://scikit-learn.org/0.15/auto\\_examples/cluster/plot\\_cluster\\_comparison.html](https://scikit-learn.org/0.15/auto_examples/cluster/plot_cluster_comparison.html), 20.12.2020 01:30

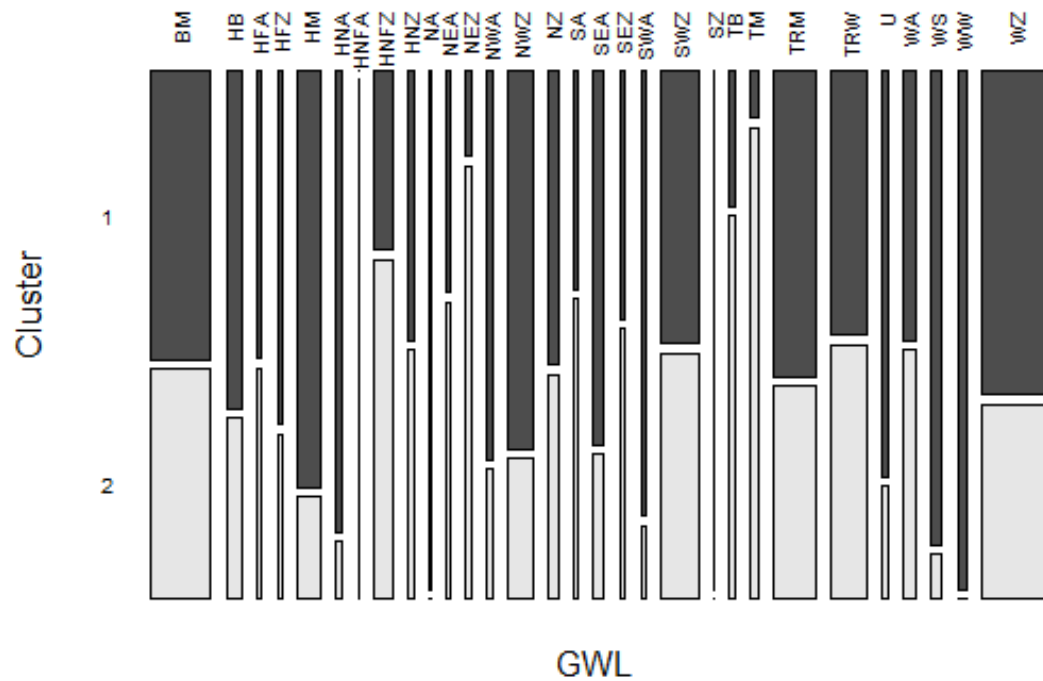


# CLARA

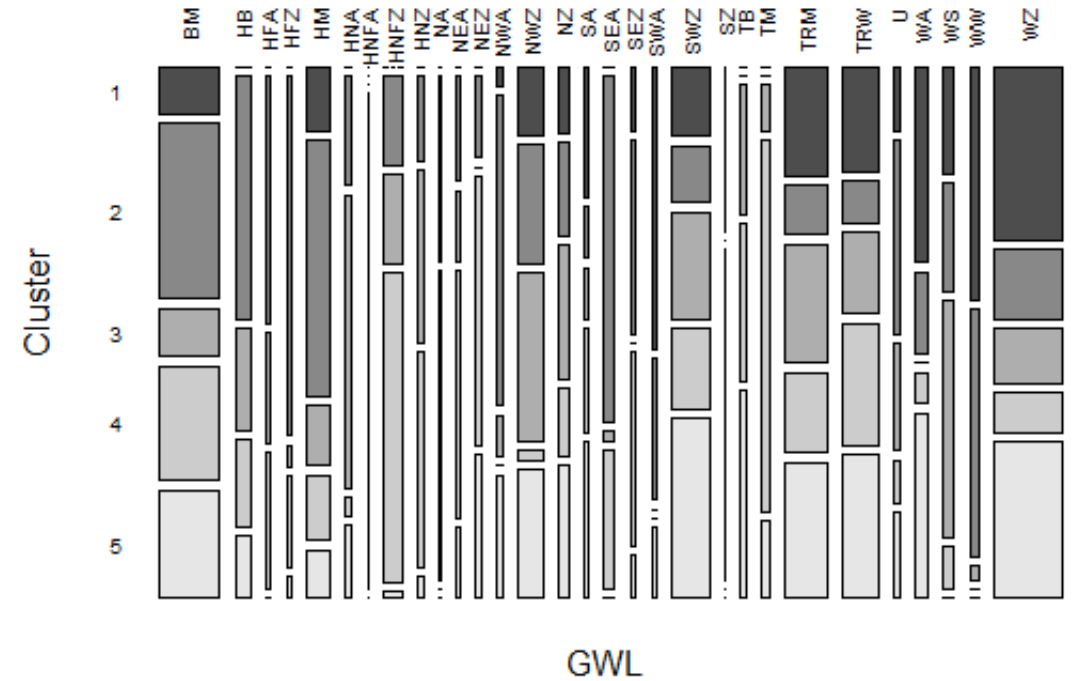


# CLARA

Clustering Large Applications



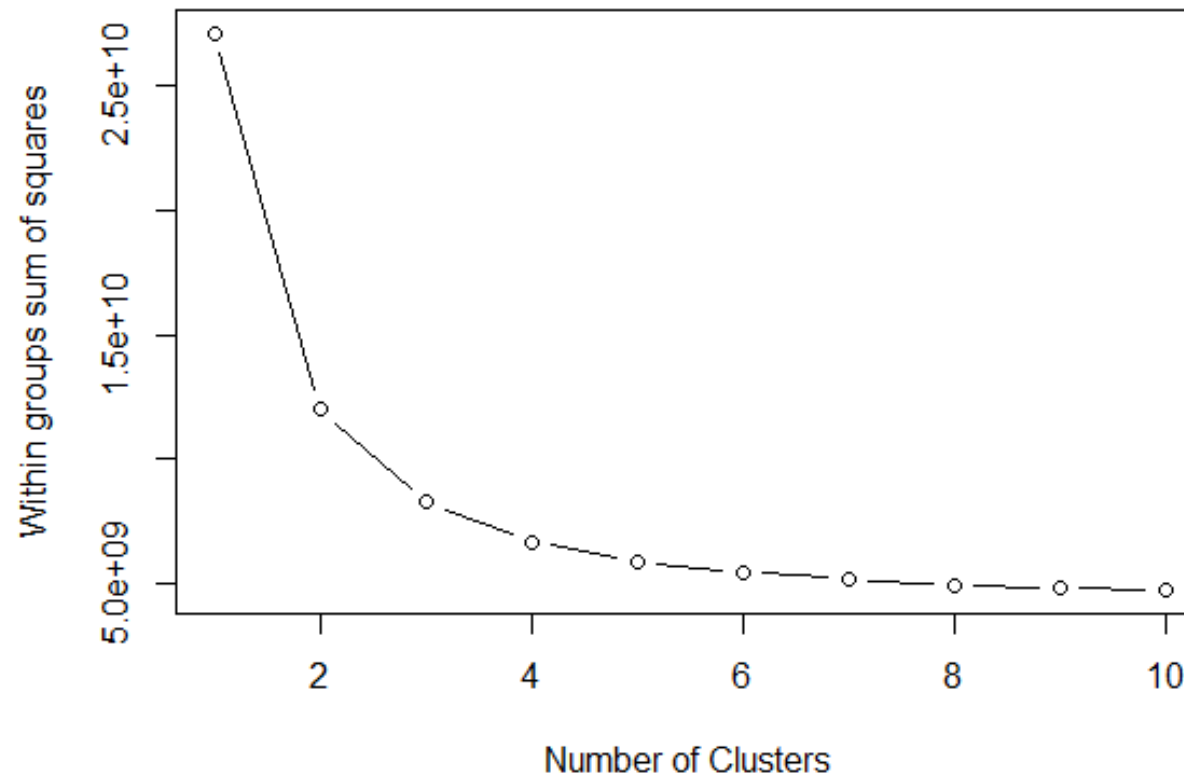
Clustering Large Applications



# CLARA

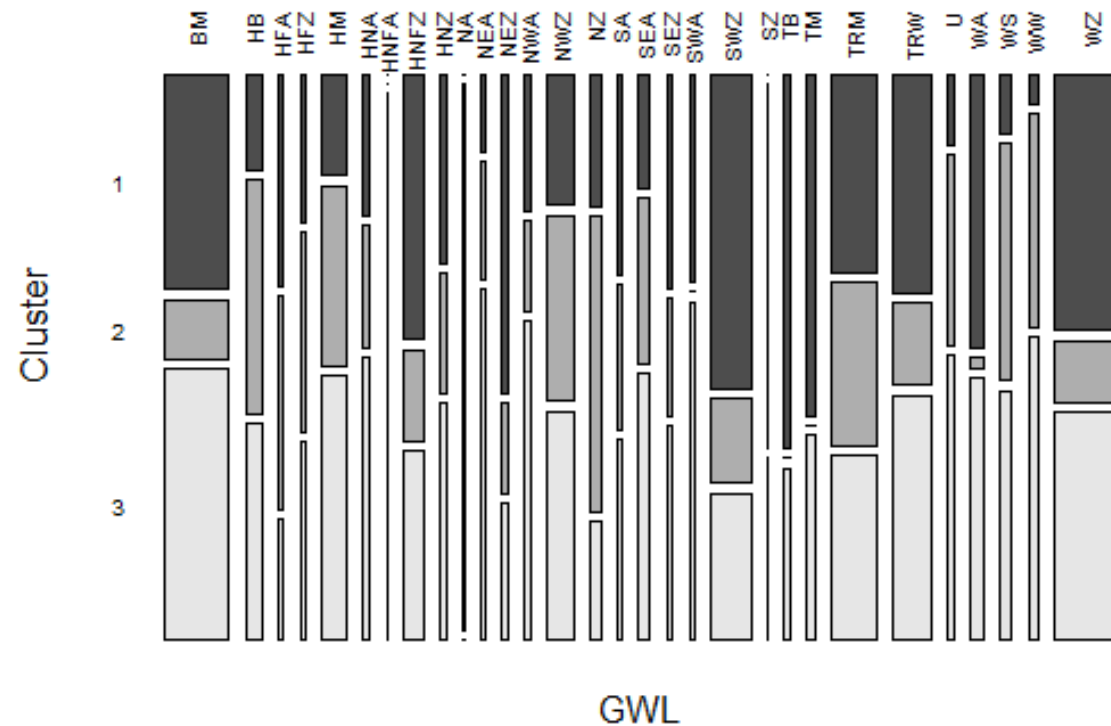


# Mahalanobis mit k-means



# Mahalanobis mit k-means

Cluster mit k-Means und Mahalanobis



# Mahalanobis mit k-means

