

Katja Gutmair, Stella Akouete, Noah Hurmer und Anne Gritto

Weather Frog

- Abschlusspräsentation am 01. März 2021
- Institut: Statistik
- Veranstaltung: Statistisches Praktikum
- Projektpartner: M.Sc. Maximilian Weigert und
M.Sc. Magdalena Mittermeier
- Betreuer: Prof. Dr. Helmut Küchenhoff



Gliederung

1. Einführung

- i. Vorstellen des Projekts
- ii. Datensätze
- iii. Einführung in Clusteranalyse

2. Analyse

- i. Methodik
- ii. Ergebnisse
- iii. Deskriptive Analyse

3. Ausblick

4. Fazit

1. Einführung

i. Vorstellen des Projekts

Vorstellen des Projekts

- Übergeordnete Fragestellung:
Wie verändert sich das Auftreten verschiedener Großwetterlagen (GWL) unter dem Einfluss des Klimawandels?
- Unsere Fragestellung:
Lassen sich Tage anhand von ihren Wettermesswerten sinnvoll clustern?
Wie unterscheiden sich die entstandenen Cluster voneinander?

Vorstellen des Projekts

Definition Großwetterlage

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

Großwetterlagen Beispiele

	Abkürzung	Großwetterlage
1	WA	Westlage, antizyklonal
2	WZ	Westlage, zyklonal
3	WS	Südliche Westlage
4	WW	Winkelförmige Westlage
5	SWA	Südwestlage, antizyklonal
6	SWZ	Südwestlage, zyklonal
...		
29	TRW	Trog Westeuropa
	U	Übergang/Unbestimmt

Vorstellen des Projekts

TODO

Motivation

- untersuchen der Veränderung
- lang anhaltende/gefährliche Wetterlagen herausfinden
- begründung

Ziele des Projekts

Clustereinteilung der beobachteten Wetterdaten

- Anzahl Cluster < Anzahl GWLs
- Berücksichtigung der räumlichen Datenstruktur
- Tage als Beobachtungseinheit
- Ohne Vorinformation der herrschenden GWL

➡ Mit welchem Modell ist dies sinnvoll möglich?

Ziele des Projekts

Vergleich der Cluster

- Verteilung von GWL in den Clustern
- Vergleich der Zusammensetzung der einzelnen Cluster:
Wie scheinen sie sich auffällig zu unterscheiden?

1. Einführung

- i. Vorstellen des Projekts
- ii. Datensätze

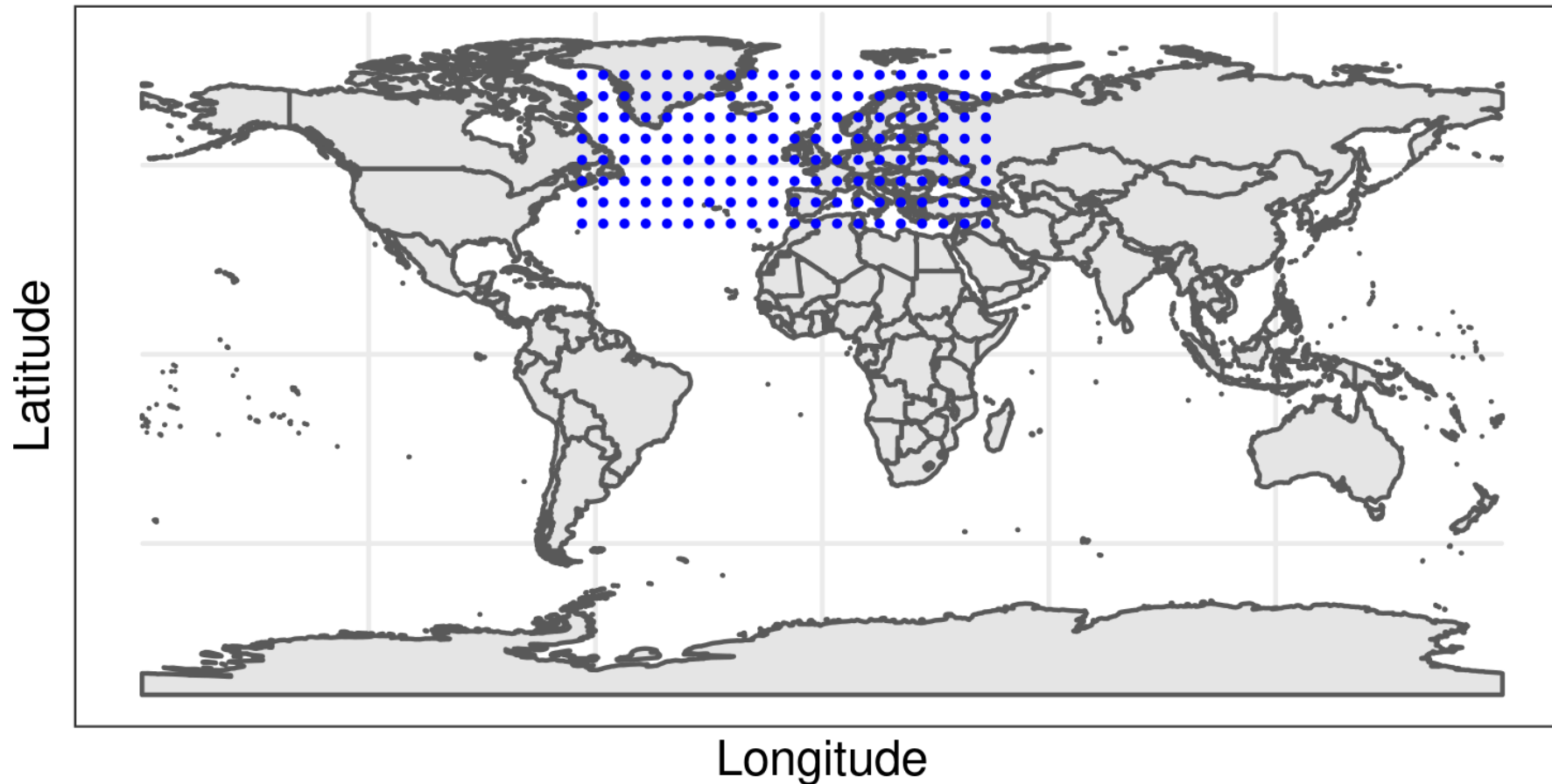
Historischer GWL Datensatz

- Zuteilung einer GWL für jeden Tag
- Für die Jahre 1900 bis 2010

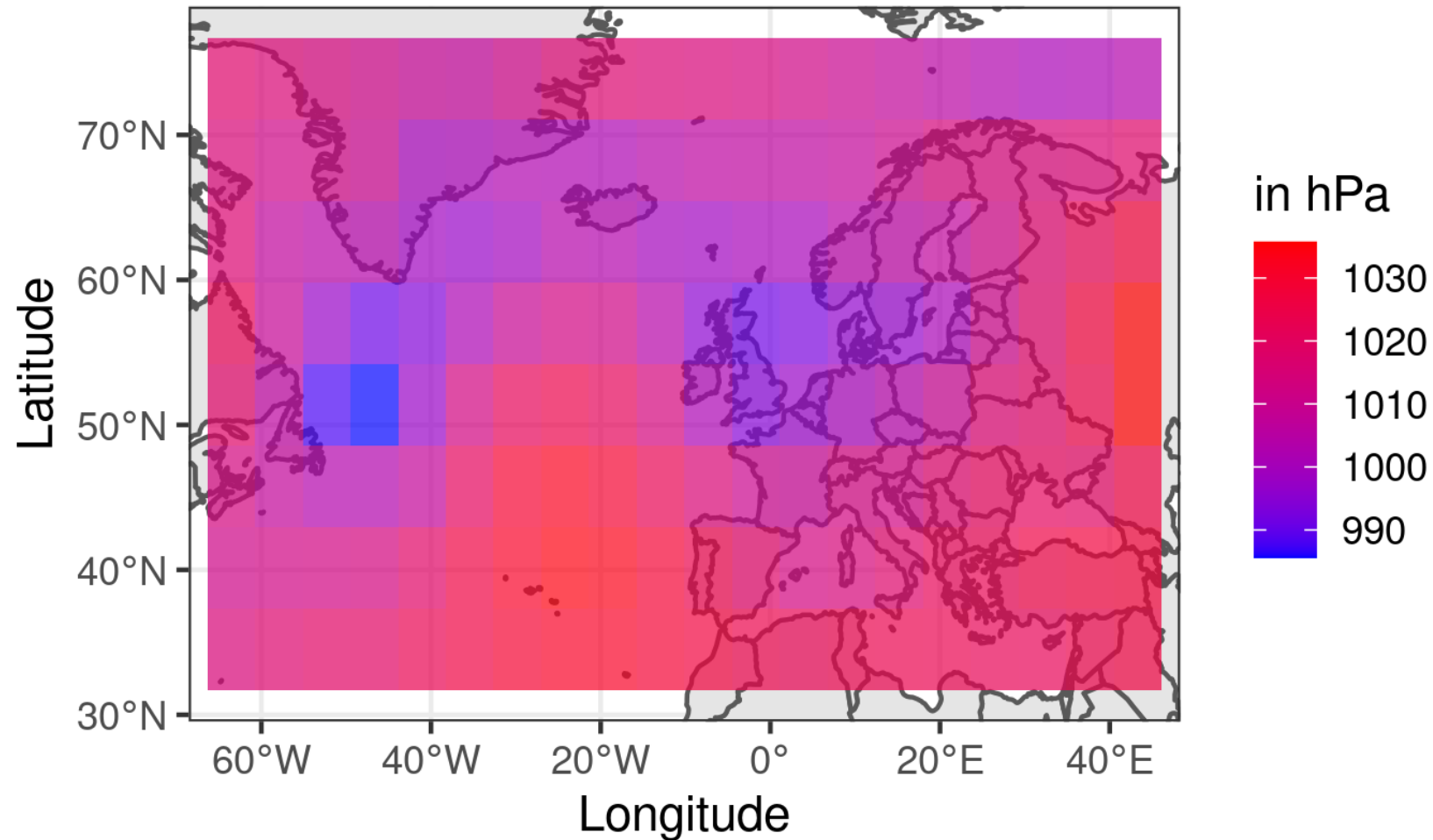
Reanalyse Datensatz

- Pro Tag Messungen an 160 Standorten zu 4 Zeitpunkten
 - Luftdruck in Pa auf Meeresspiegelhöhe (mslp)
 - Geopotential auf 500 hPa in $\frac{m^2}{s^2} = \frac{1}{9.80665} \text{ gpm}$ (geopot)
- Für die Jahre 1900 bis 2010
 - Beschränkung auf eine Klimaperiode: Jahre 1971 bis 2000
- Ohne Information zur herrschenden GWL am Tag
- Standorte im 8x20 Grid über Europa und dem Nordatlantik

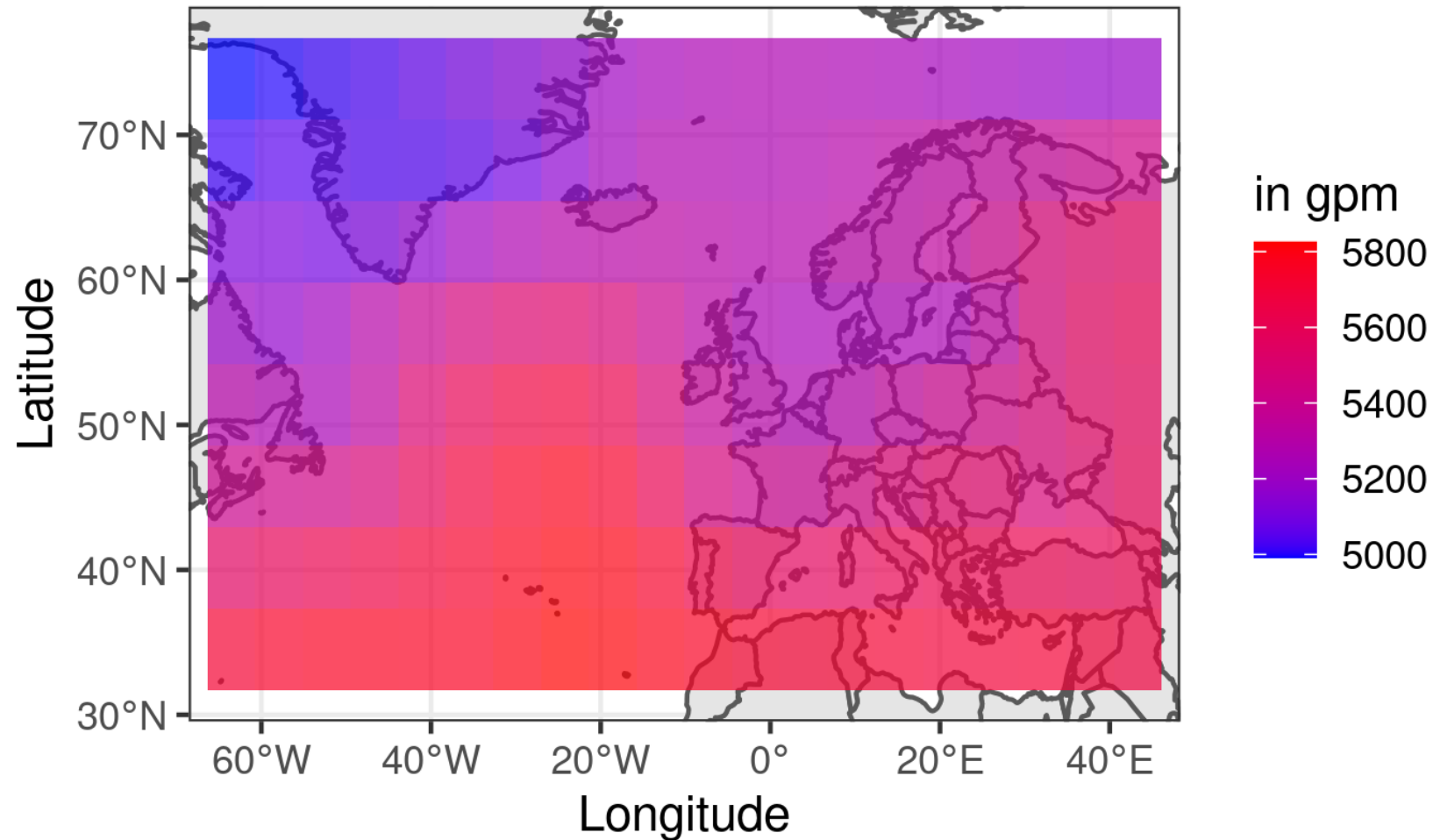
Messpunkte auf einer Weltkarte



Mslp am 01-01-2006 um 0 Uhr



Geopot am 01-01-2006 um 0 Uhr



Auszug aus dem Reanalyse Datensatz

	time	longitude	latitude	mslp	geopotential
1	1900-01-01 00:00:00	-63.56287	73.85311	100428.99	48268.86
2	1900-01-01 00:00:00	-63.56287	68.23695	100553.77	48770.82
3	1900-01-01 00:00:00	-63.56287	62.62077	99920.18	49171.14
4	1900-01-01 00:00:00	-63.56287	57.00457	100049.80	49487.83
...					
640	1900-01-01 18:00:00	43.31280	34.53973	102281.97	55097.32
641	1900-01-02 00:00:00	-63.56287	73.85311	99886.71	47843.04
...					
25946239	2010-12-31 18:00:00	43.31280	40.15595	101758.62	54154.39
25946240	2010-12-31 18:00:00	43.31280	34.53973	101400.51	54491.94

Auszug aus dem Reanalyse Datensatz

	time	longitude	latitude	mslp	geopotential
1	1900-01-01 00:00:00	-63.56287	73.85311	100428.99	48268.86
2	1900-01-01 00:00:00	-63.56287	68.23695	100553.77	48770.82
3	1900-01-01 00:00:00	-63.56287	62.62077	99920.18	49171.14
4	1900-01-01 00:00:00	-63.56287	57.00457	100049.80	49487.83
...					
640	1900-01-01 18:00:00	43.31280	34.53973	102281.97	55097.32
641	1900-01-02 00:00:00	-63.56287	73.85311	99886.71	47843.04
...					
25946239	2010-12-31 18:00:00	43.31280	40.15595	101758.62	54154.39
25946240	2010-12-31 18:00:00	43.31280	34.53973	101400.51	54491.94

Daten pro Tag

Der Tag ist die Beobachtungseinheit

➡ $2 \text{ Parameter} * 4 \text{ Zeitpunkte} * 160 \text{ Messpunkte} = 1280 \text{ Dimensionen}$

➡ 8 Bilder pro Tag

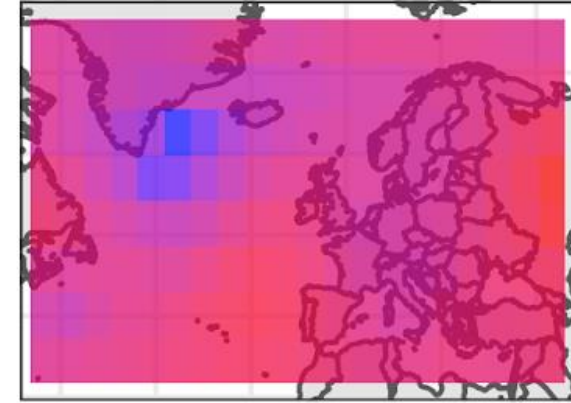
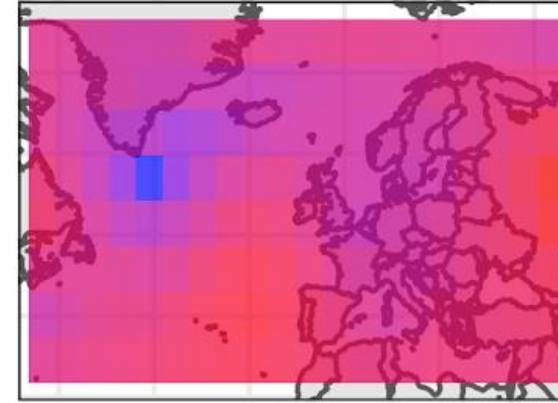
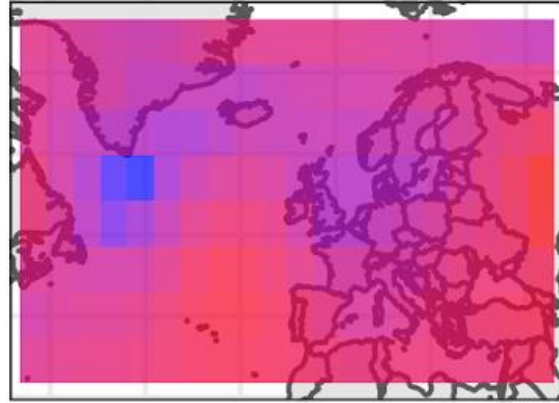
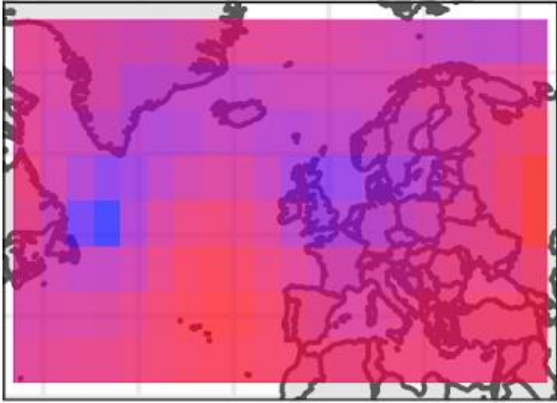
0 Uhr

6 Uhr

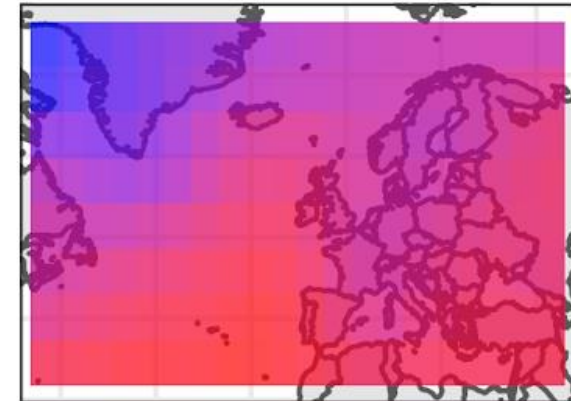
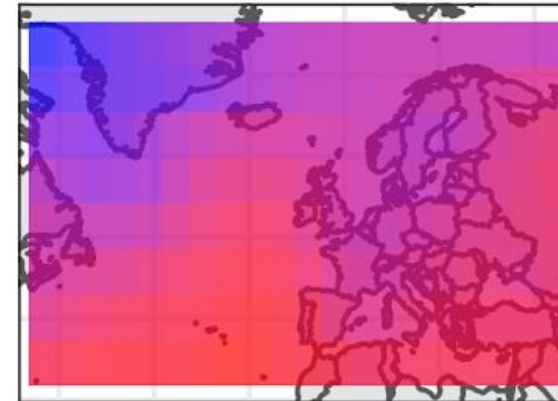
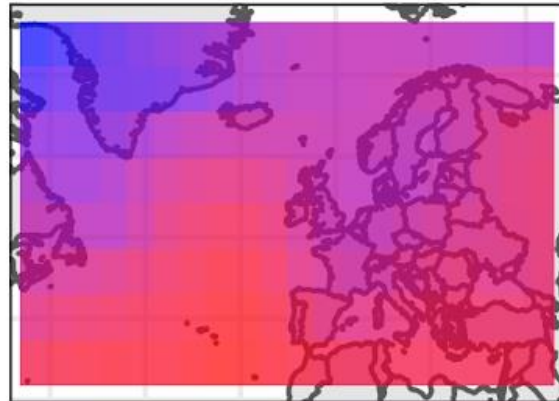
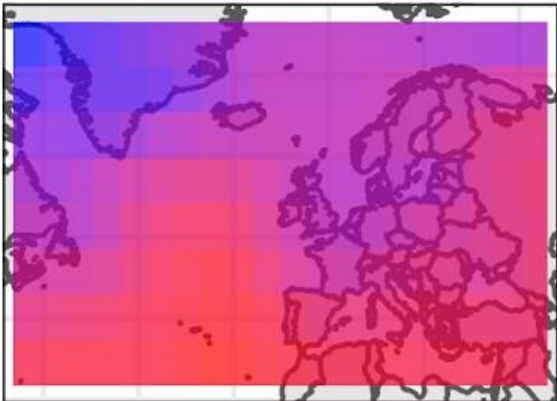
12 Uhr

18 Uhr

Mslp



Geopotential



Daten pro Tag

Der Tag ist die Beobachtungseinheit

➡ 2 Parameter * 4 Zeitpunkte * 160 Messpunkte = 1280 Dimensionen

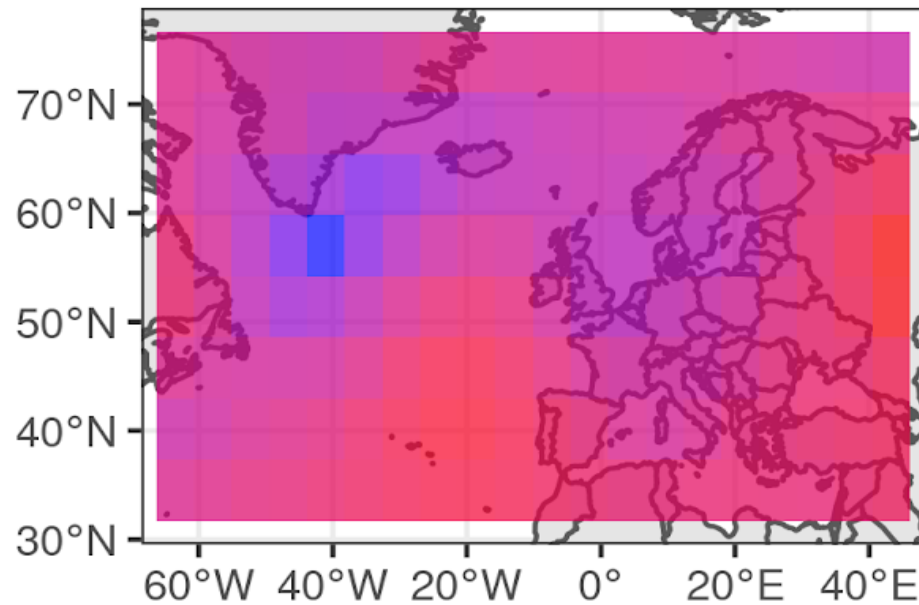
➡ 8 Bilder pro Tag

Reduzierung der Dimensionen

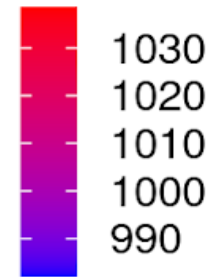
➡ Mittelwert über 4 Messzeiten pro Messpunkt

Mittelwerte am 01.01.2006

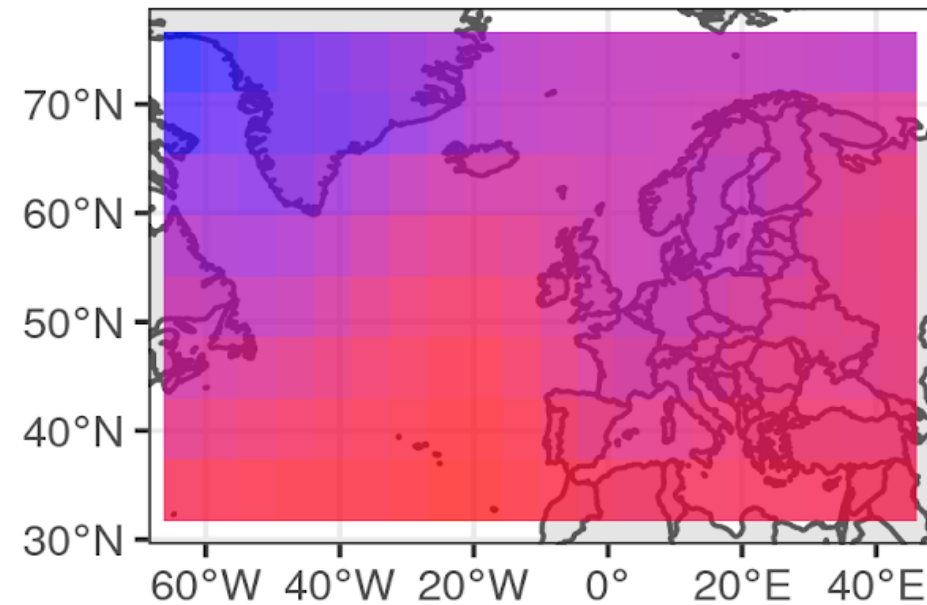
Mslp



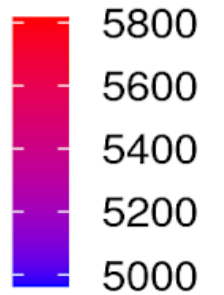
in hPa



Geopot



in gpm



Daten pro Tag

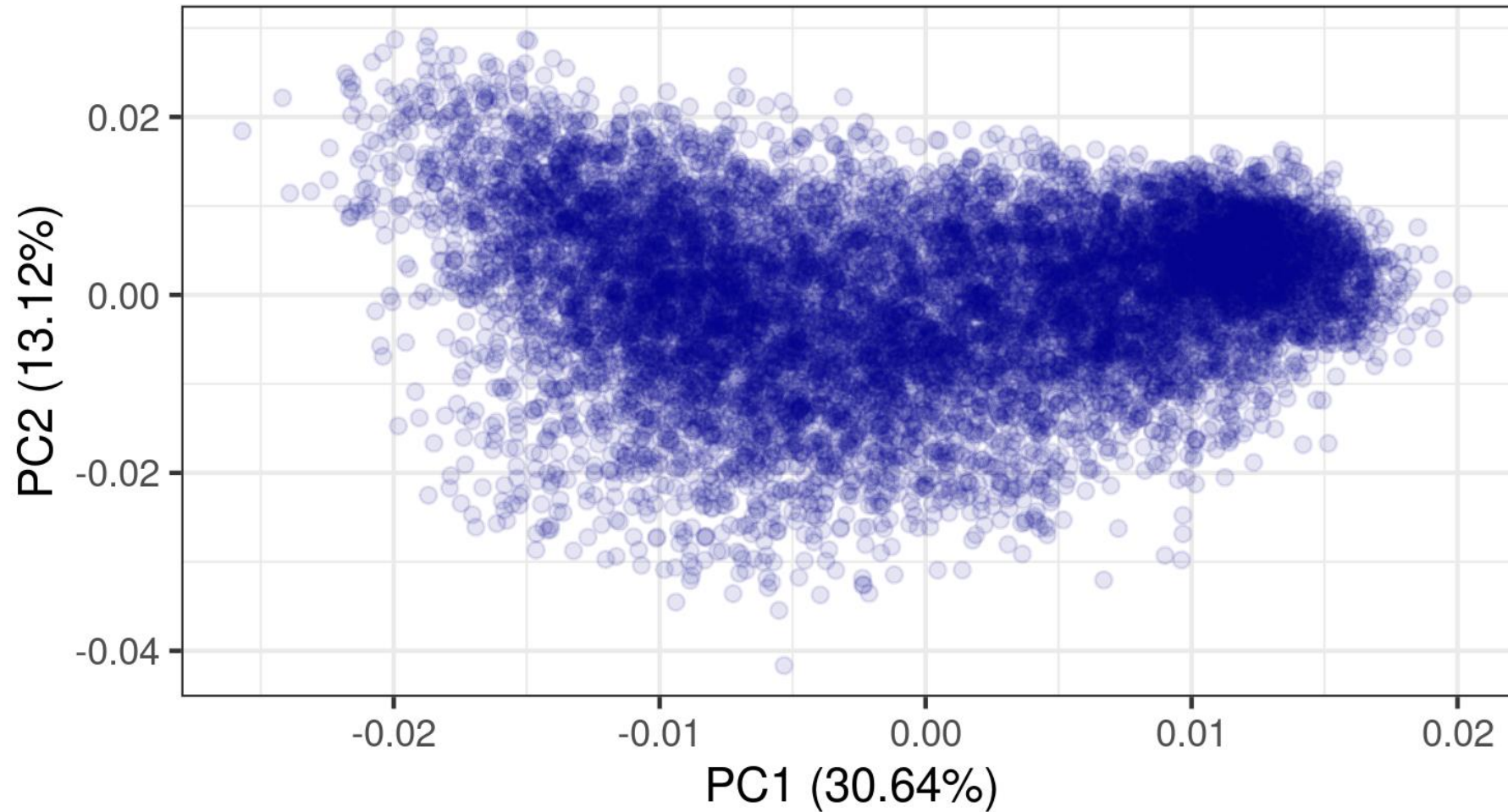
Der Tag ist die Beobachtungseinheit

- ➡ $2 \text{ Parameter} * 4 \text{ Zeitpunkte} * 160 \text{ Messpunkte} = 1280 \text{ Dimensionen}$
- ➡ 8 Bilder pro Tag

Reduzierung der Dimensionen

- ➡ Mittelwert über 4 Messzeiten pro Messpunkt
- ➡ 10958 Tage mit jeweils 320 Dimensionen

Visualisierung der Daten mit PCA



1. Einführung

- i. Vorstellen des Projekts
- ii. Datensätze
- iii. Einführung in Clusteranalyse

Clusteranalyse

- Grundidee: Bildung von möglichst homogenen Gruppen, Cluster untereinander möglichst heterogen
- Betrachten von n Objekten a_1, \dots, a_n mit zugehörigen Merkmalsvektoren x_1, \dots, x_n
 - ➡ Suchen einer Partition C_1, \dots, C_g mit $\bigcup_{k=1}^g C_k = \{a_1, \dots, a_n\}$ wobei $C_k \cap C_l = \emptyset \quad \forall k \neq l$
- Clusteranalyse ist Verfahren des "unsupervised learning"
- Verschiedene Ansätze für Clustering
 - ➡ Optimale Partitionen: Messen der Qualität einer Partition C anhand eines Gütekriteriums

Clusteranalyse

Grundlage für die Clusterbildung ist ein Ähnlichkeits- bzw. Distanzmaß

➡ Manhattan-Metrik, bei der die Distanz d zwischen zwei Punkten a und b definiert ist als

$$d(a, b) = \sum_{i=1}^n |a_i - b_i|$$

2. Analyse

i. Methodik

Bewertungskriterien für Clustering

- Durchschnittliche Silhouettenweite
 - Maßzahl für die Qualität eines Clusterings
 - Unabhängig von der Anzahl der Cluster

Bewertungskriterien für Clustering

- Durchschnittliche Silhouettenweite
 - Gehört das Objekt x zum Cluster A , so ist die Silhouette von x definiert als

$$S(x) = \begin{cases} 0 & \text{Wenn } x \text{ einziges Element von } A, \text{ ist} \\ \frac{\text{dist}(B, x) - \text{dist}(A, x)}{\max\{\text{dist}(A, x), \text{dist}(B, x)\}} & \text{sonst,} \end{cases}$$

wobei $\text{dist}(A, x)$ die Distanz eines Objektes x zum Cluster A und
 $\text{dist}(B, x)$ die Distanz eines Objektes x zum nächstgelegenen Cluster B

Bewertungskriterien für Clustering

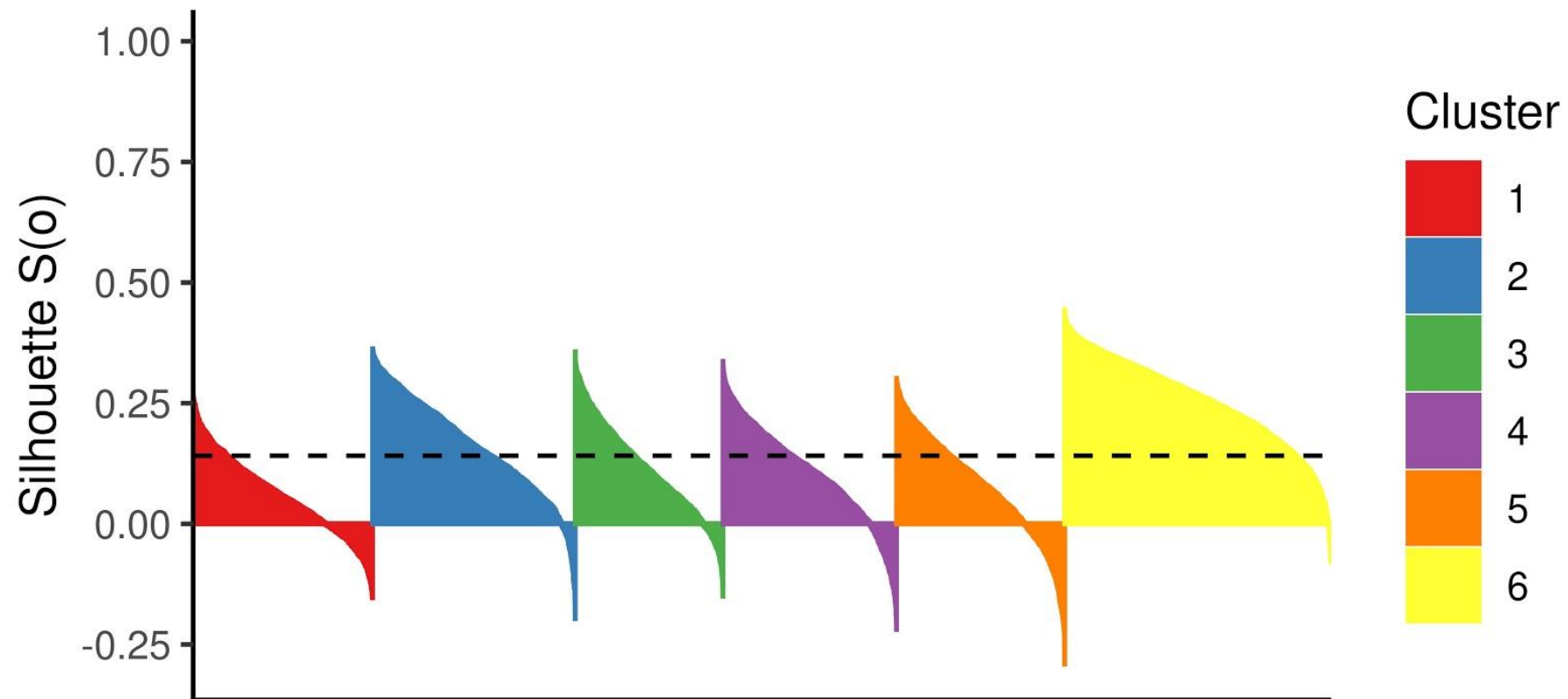
- Durchschnittliche Silhouettenweite
 - Sei C die Anzahl an Cluster, dann ist der Silhouettenkoeffizient definiert durch

$$s_C = \frac{1}{n_C} \sum_{x \in C} S(x)$$

$$\text{wobei } S(x) = \begin{cases} 0 & \text{Wenn } x \text{ einziges Element von } A, \text{ ist} \\ \frac{\text{dist}(B, x) - \text{dist}(A, x)}{\max\{\text{dist}(A, x), \text{dist}(B, x)\}} & \text{sonst,} \end{cases}$$

Silhouettenplot

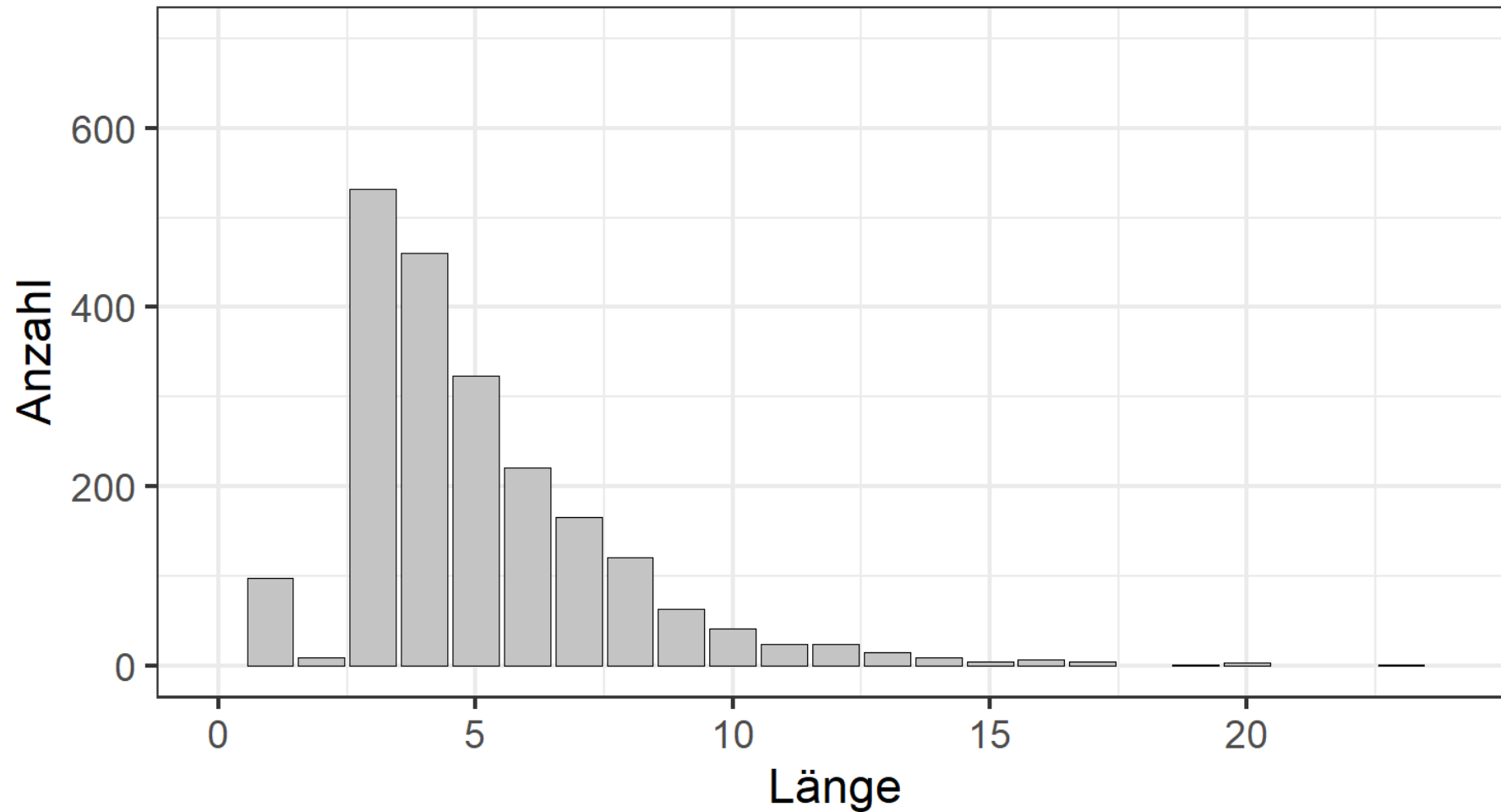
Silhouettenkoeffizient: 0.141



Bewertungskriterien für Clustering

- Durchschnittliche Silhouettenweite
- Verteilung der aufeinanderfolgenden Tage, die im selben Cluster sind (Timeline)

Länge der aufeinanderfolgenden, gleichen GWL



Datensatz Mutation

- Idee: Erstellen eines neuen Datensatzes durch Extrahieren gezielter Information
- Gezielte Informationen
 - Verteilung der Parameter (im Vergleich zu anderen Tagen)
 - Örtliche Lage und Form der „Hoch-“ und „Tiefgebiete“
 - „Bildmuster“ des Tages
 - Veränderung über den Tag

Datensatz Mutation

- Idee: Erstellen eines neuen Datensatzes durch Extrahieren gezielter Information
- Gezielte Informationen
 - Verteilung der Parameter (im Vergleich zu anderen Tagen)
 - Örtliche Lage und Form der „Hoch-“ und „Tiefgebiete“
 - „Bildmuster“ des Tages
 - Veränderung über den Tag
- Erhoffte Wirkung
 - Dimensionen reduzieren
 - Spezifische Gewichtung wichtiger Größen

Vorgehen

- Ausgangslage: Datensatz mit 320 Dimensionen roher Messdaten
- Transformation zu Variablen, die jeweils eine interessierende Größe über alle Standorte zusammengefasst verkörpern
 - Beispiel: Mittelwert des Luftdrucks über alle Standorte am Tag
- ➡ Beobachtungseinheit bleibt der Tag
- ➡ Ziel: Erkennen, welche Tage ähnliche Merkmale aufweisen

Extrahierte Variablen

Variable	Erklärung
Datum	
Minimum/Maximum	Minimaler/Maximaler Wert am Tag
Mittelwert/ Median/Quartile	Mittelwert/Median und Quartile für beide Variablen pro Tag
Intensität	Anzahl der Messpunkte von beiden Variablen pro Tag die über/unter den Quartilen liegen
Differenz am Tag	Summierte Differenzen von 4 Messzeitpunkten am Tag an allen Standorten

Extrahierte Variablen

Variable	Erklärung
Datum	
Minimum/Maximum	Minimaler/Maximaler Wert am Tag
Mittelwert/ Median/Quartile	Mittelwert/Median und Quartile für beide Variablen pro Tag
Intensität	Anzahl der Messpunkte von beiden Variablen pro Tag die über/unter den Quartilen liegen
Differenz am Tag	Summierte Differenzen von 4 Messzeitpunkten am Tag an allen Standorten

} Verteilung der Parameter

Extrahierte Variablen

Variable	Erklärung
Distanz von Maximum und Minimum	Euklidische Distanz
Distanz der beiden Minima und Maxima	Euklidischer Abstand vom Minimum/Maximum der Parameter Geopotential zu Mslp
Spalte vom Minimum/Maximum	In welchem Bereich liegt das Minimum/Maximum? Karte aufgeteilt in 3 Spalten
Zeile vom Minimum/Maximum	In welchem Bereich liegt das Minimum/Maximum? Karte aufgeteilt in 3 Zeilen
Mittelwerte in den Quadranten	Mittelwerte in allen 9 Quadranten von beiden Variablen

Extrahierte Variablen

Variable	Erklärung
Distanz von Maximum und Minimum	Euklidische Distanz
Distanz der beiden Minima und Maxima	Euklidischer Abstand vom Minimum/Maximum der Parameter Geopotential zu Mslp
Spalte vom Minimum/Maximum	In welchem Bereich liegt das Minimum/Maximum? Karte aufgeteilt in 3 Spalten
Zeile vom Minimum/Maximum	In welchem Bereich liegt das Minimum/Maximum? Karte aufgeteilt in 3 Zeilen
Mittelwerte in den Quadranten	Mittelwerte in allen 9 Quadranten von beiden Variablen



Räumliche Ebene

Zusammenhang der räumlichen Ebene und der Verteilung

Skalierung und Gewichtung

- Datensatz wird standardisiert, da die Skalen der einzelnen Variablen unterschiedlich sind

$$x_{neu} = \frac{x - \mu}{\sigma}$$

- Variablen werden zudem gewichtet, unterteilt nach Kategorien
➡ Gewichte einer Kategorie summieren sich auf 1

Skalierung und Gewichtung

Variable	Gewichte	Variable	Gewichte
Datum		Distanz von Maximum und Minimum	$\frac{1}{6}$
Minimum/Maximum	$\frac{1}{3}$	Distanz der beiden Minima und Maxima	$\frac{1}{6}$
Mittelwert/ Median/Quartile	$\frac{1}{3}$ bzw. $\frac{1}{6}$	Spalte vom Minimum/Maximum	$\frac{1}{6}$
Intensität	$\frac{1}{6}$	Zeile vom Minimum/Maximum	$\frac{1}{6}$
Differenz am Tag	$\frac{1}{6}$	Mittelwerte in den Quadranten	$\frac{1}{9}$

Clusteralgorithmus PAM

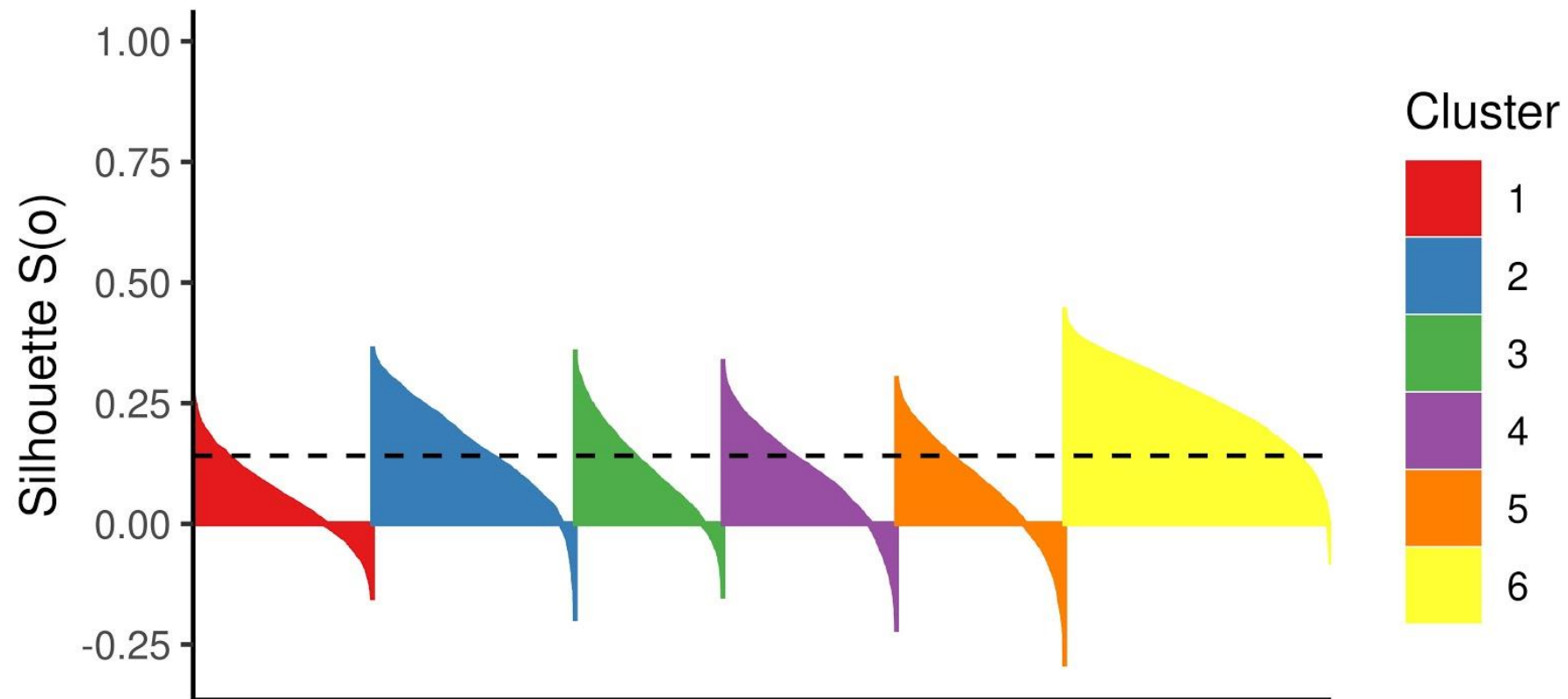
- PAM steht für Partitioning Around Medoids
- Gehört zu den Partitionierenden Verfahren
- Vorgehen: 1. Anzahl k an Cluster festlegen
 2. Wahl von k repräsentativen Objekten (Medoids) aus allen Beobachtungen
 3. Für jeden Medoid m und jeden restlichen Datenpunkt o :
 - i. Entscheiden, ob ein Datenpunkt o einen Medoid m ersetzen soll anhand der Summe S der Distanzen von allen Datenpunkten zu deren jeweiligen Medoid
 - ii. Durchführen für alle Datenpunkte
 - iii. Auswahl der Datenpunkte als Medoids, die die Summe S am stärksten minimieren
 4. Datenpunkte dem Cluster zuteilen, dessen Medoid am nächsten zu o liegt

2. Analyse

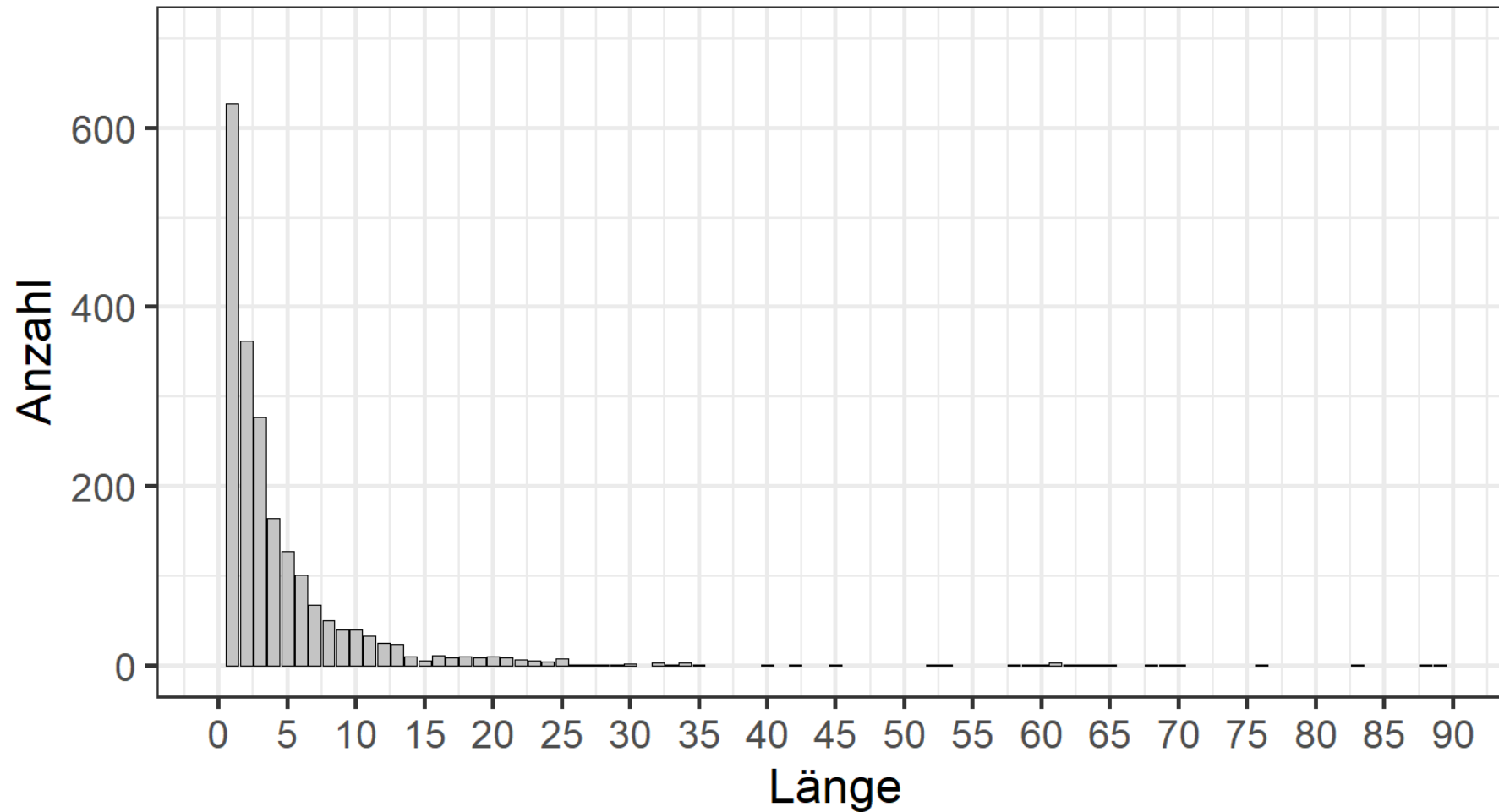
- i. Methodik
- ii. Ergebnisse

Silhouettenplot

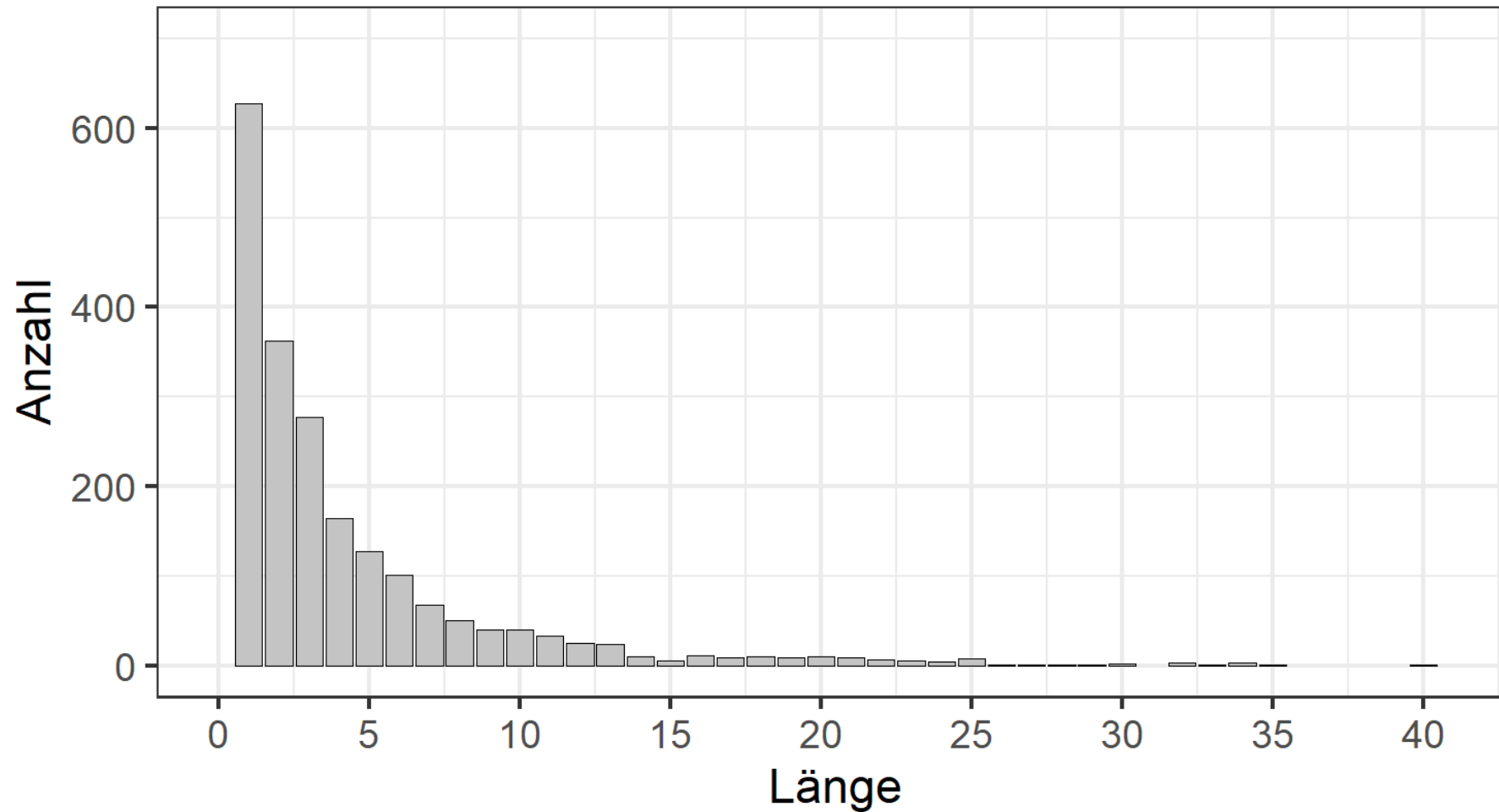
Silhouettenkoeffizient: 0.141



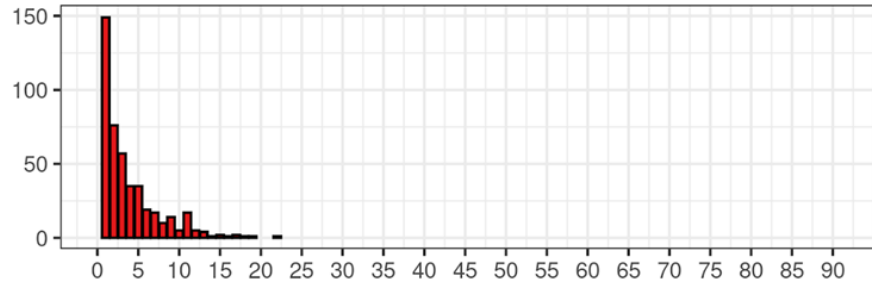
Länge der aufeinanderfolgenden, gleichen Cluster



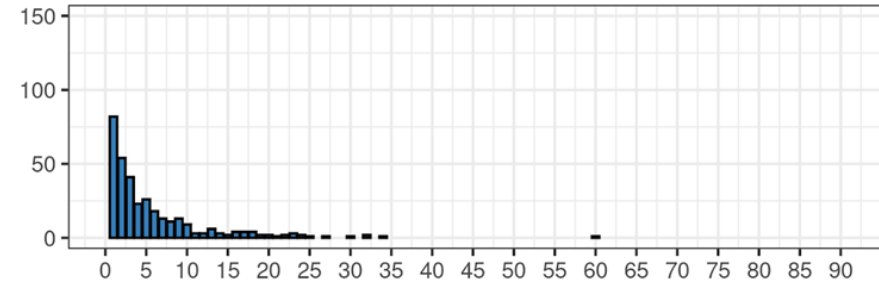
Länge der aufeinanderfolgenden, gleichen Cluster



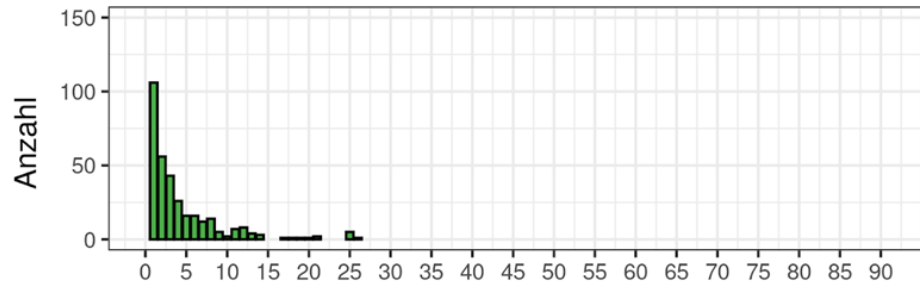
Cluster 1



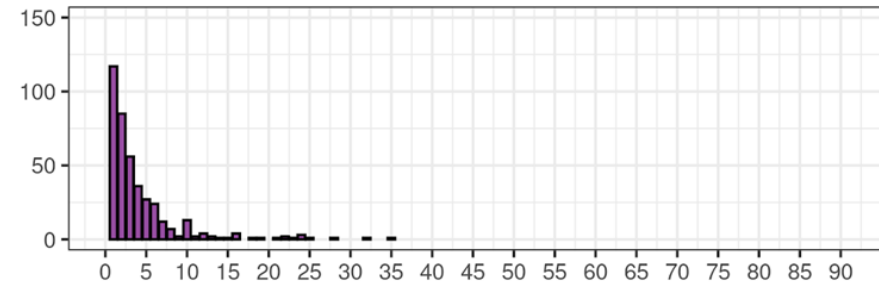
Cluster 2



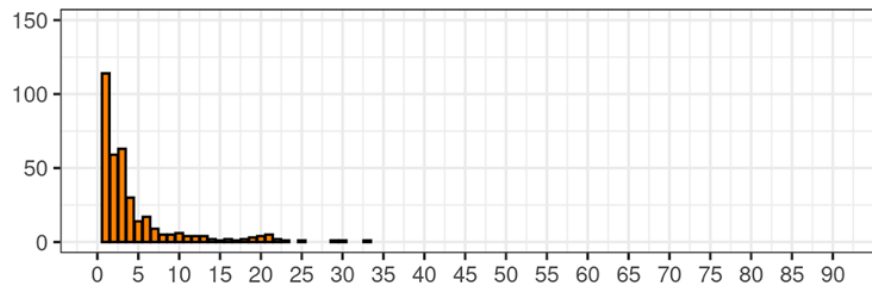
Cluster 3



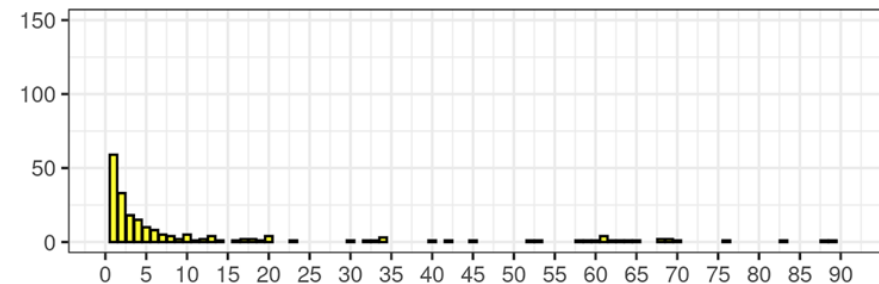
Cluster 4



Cluster 5

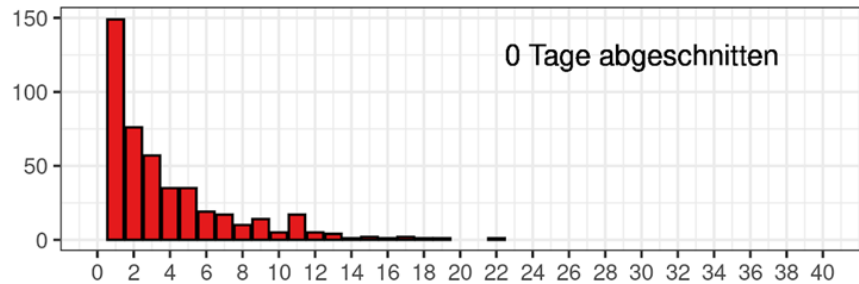


Cluster 6

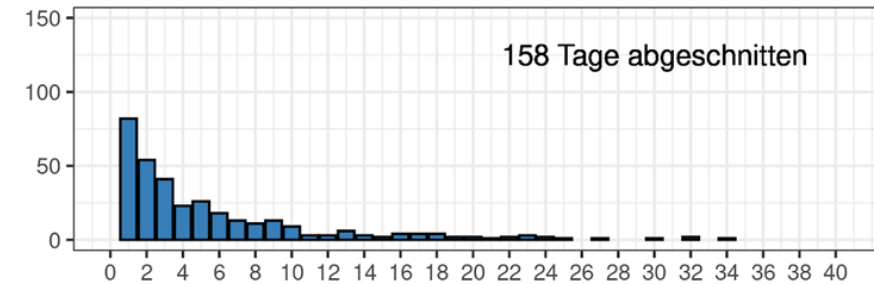


Länge

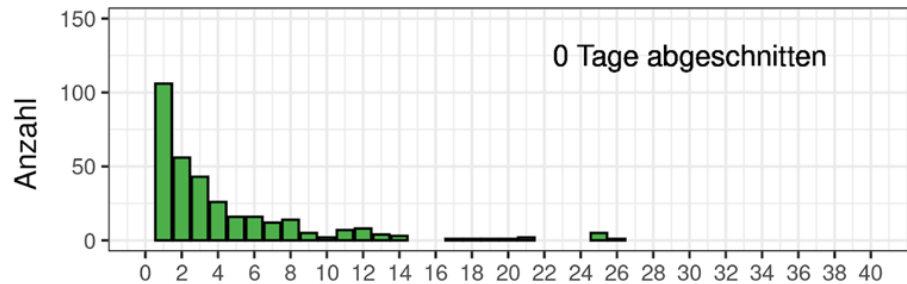
Cluster 1



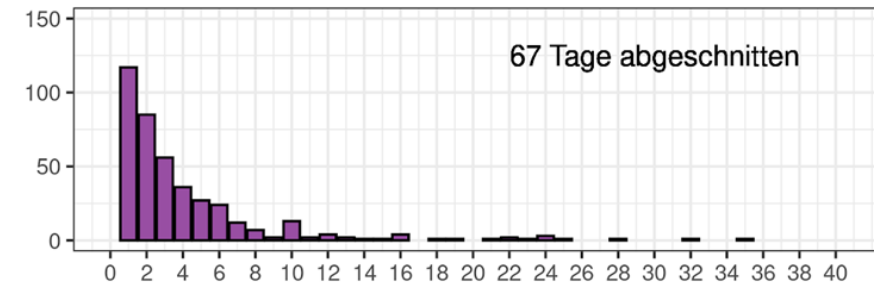
Cluster 2



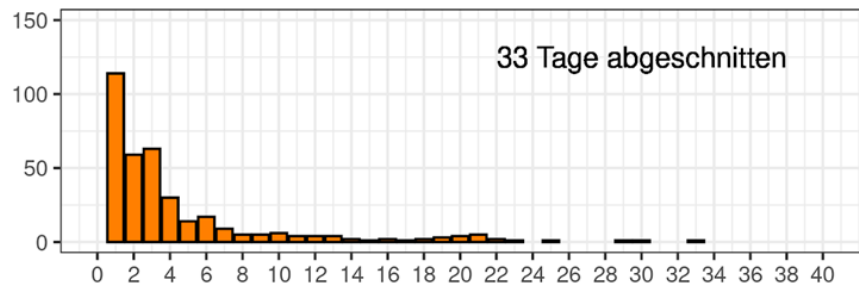
Cluster 3



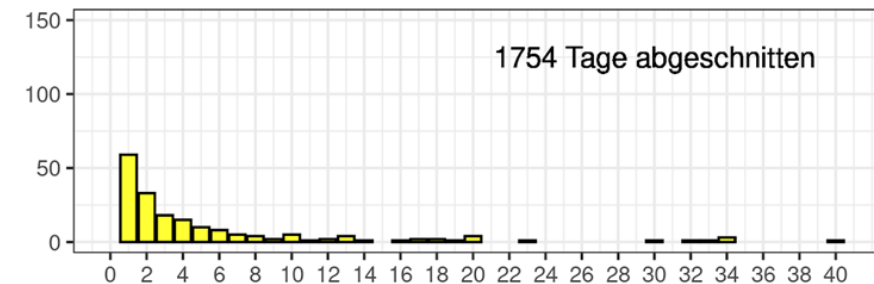
Cluster 4



Cluster 5

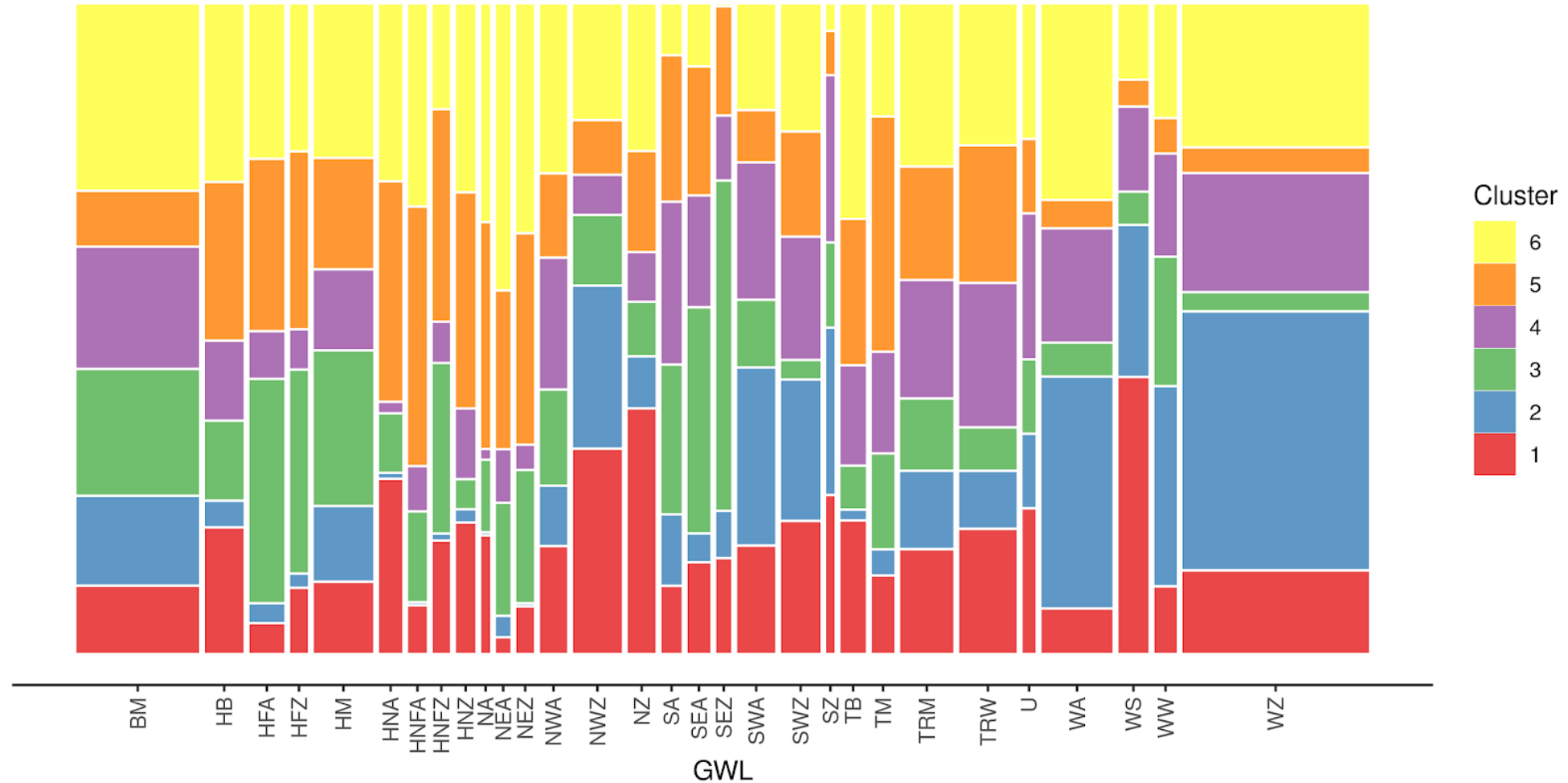


Cluster 6



Länge

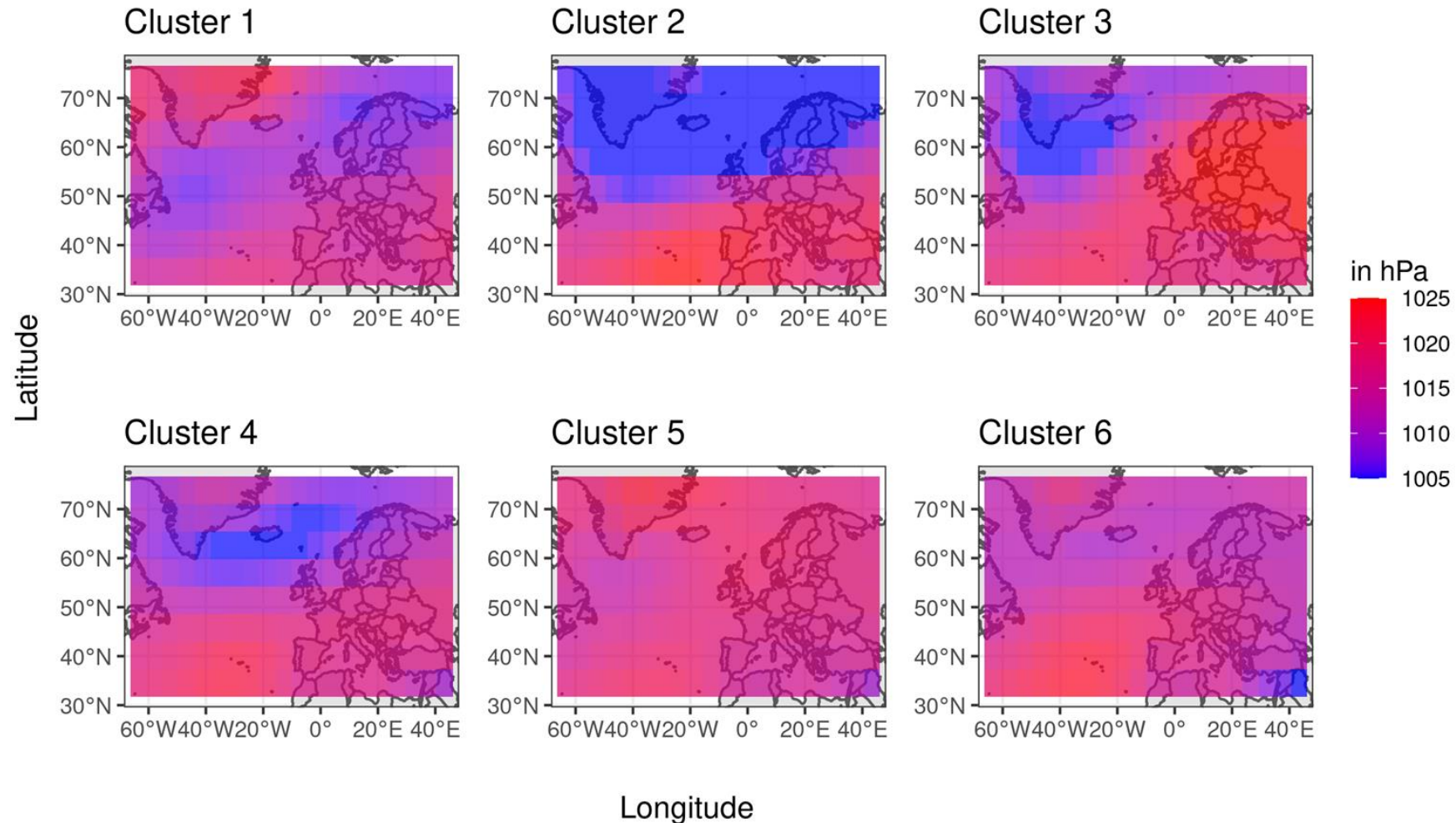
Mosaikplot für Cluster ~ GWL



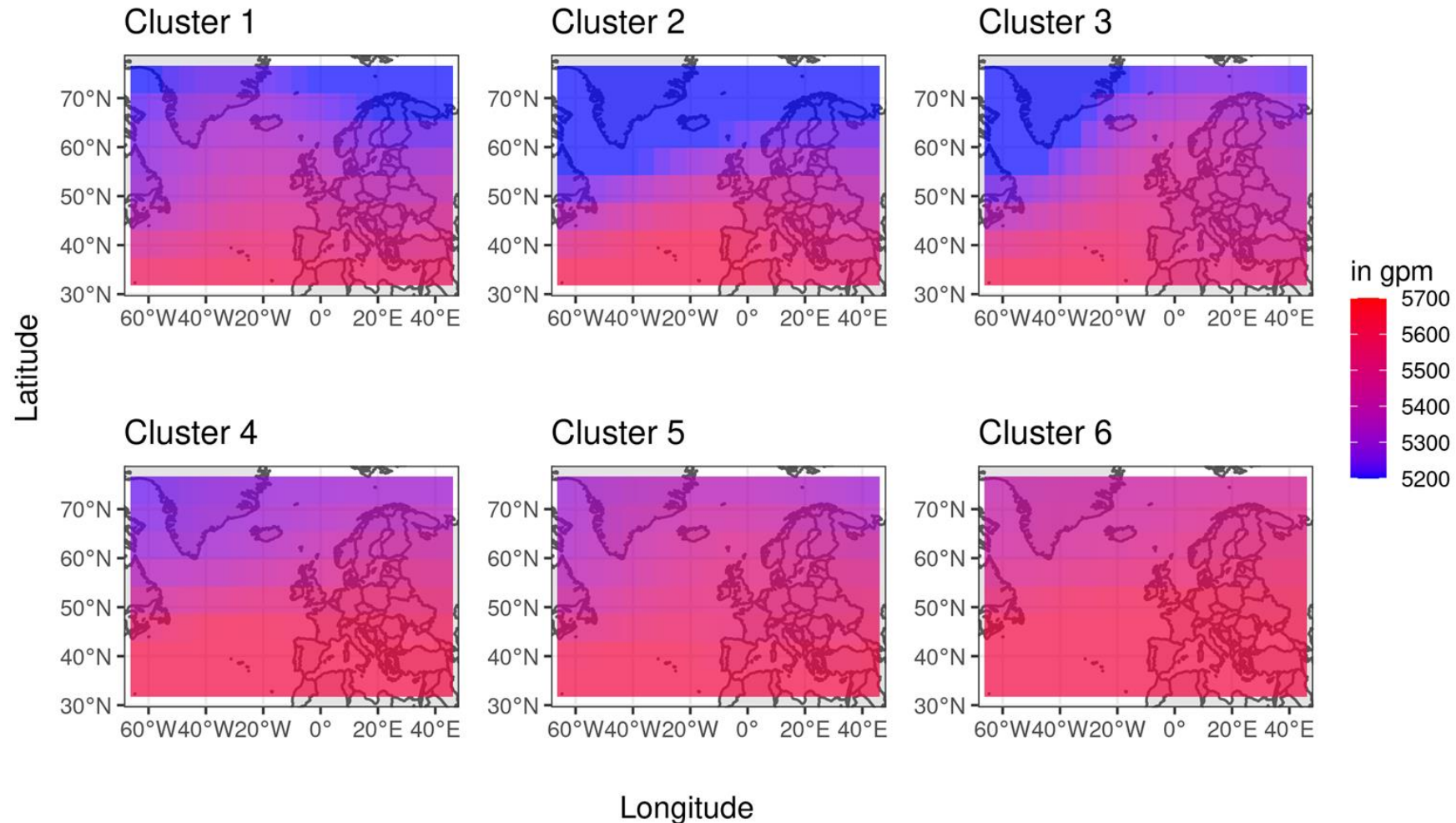
2. Analyse

- i. Methodik
- ii. Ergebnisse
- iii. Deskriptive Analyse der Cluster

Mslp im Mittel über Messpunkte



Geopot im Mittel über Messpunkte



3. Ausblick

Nicht benutzen der zeitlichen struktur -> video statt bilder für zeitliche komponente

Ansatz mit filtern

4. Fazit

Fazit

Quellen

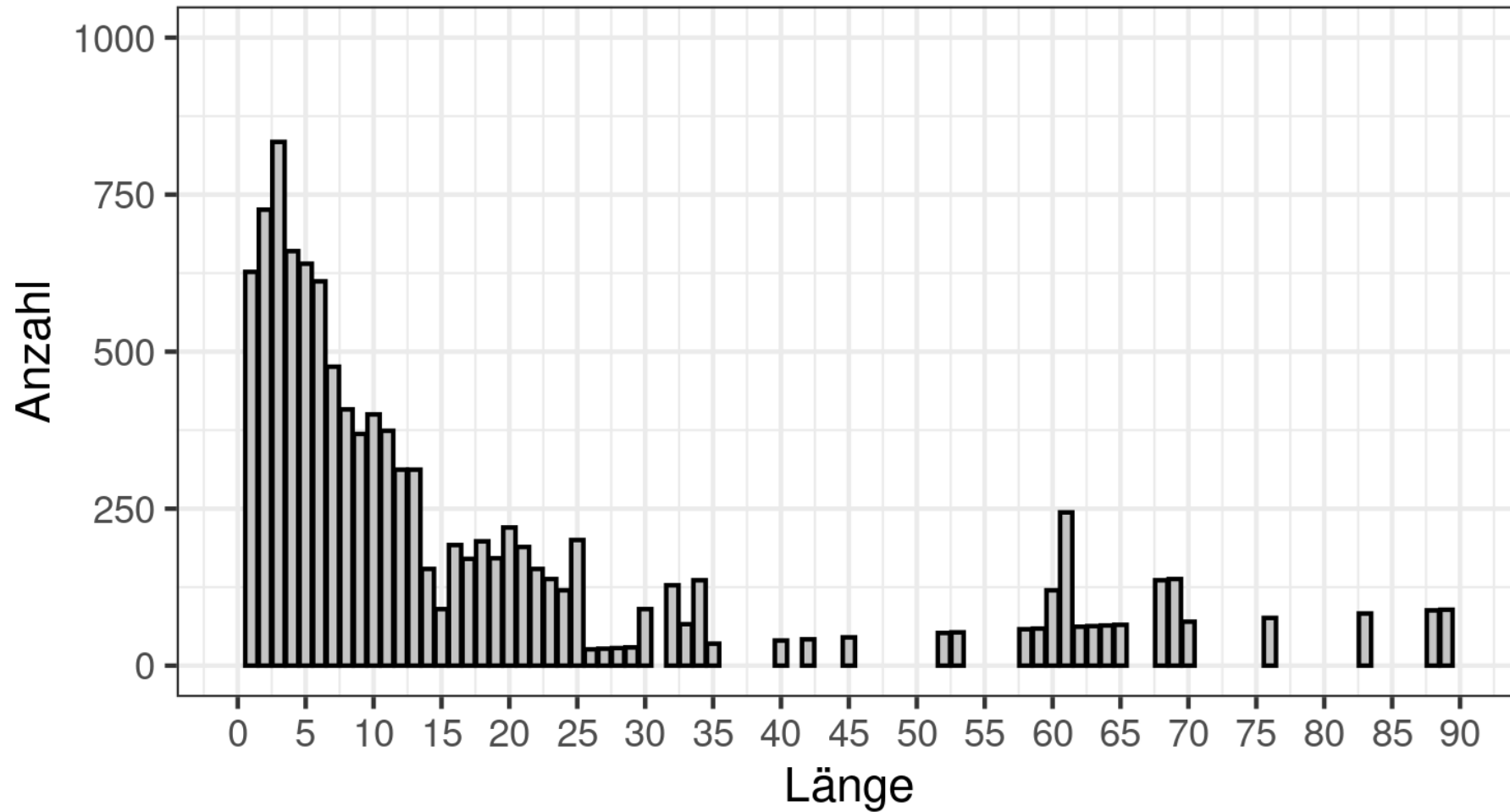
<https://arxiv.org/ftp/arxiv/papers/1302/1302.6602.pdf>

Anhang

Versuchte Algorithmen/Metriken

- Cluster Algorithmen:
 - PAM
 - K-means
 - Fuzzy
 - GMM
 - DBSCAN
- Metriken
 - Euklidisch
 - Manhattan
 - Mahalanobis

Timeline (Länge * Anzahl)



Länge der aufeinanderfolgenden, gleichen GWL

