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

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 Wetterzustand
- Definiert über ganz Europa
- Dauer: $\geq$ 3 Tage
- Kategorisierung nach dem Katalog von Hess und Brezowsky
- 29 GWL nach Hess und 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

Quelle: http://www.sklima.de/wetterlagen_uebersicht.php

Ziele des Projekts

Clustereinteilung der Tage anhand beobachteter 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

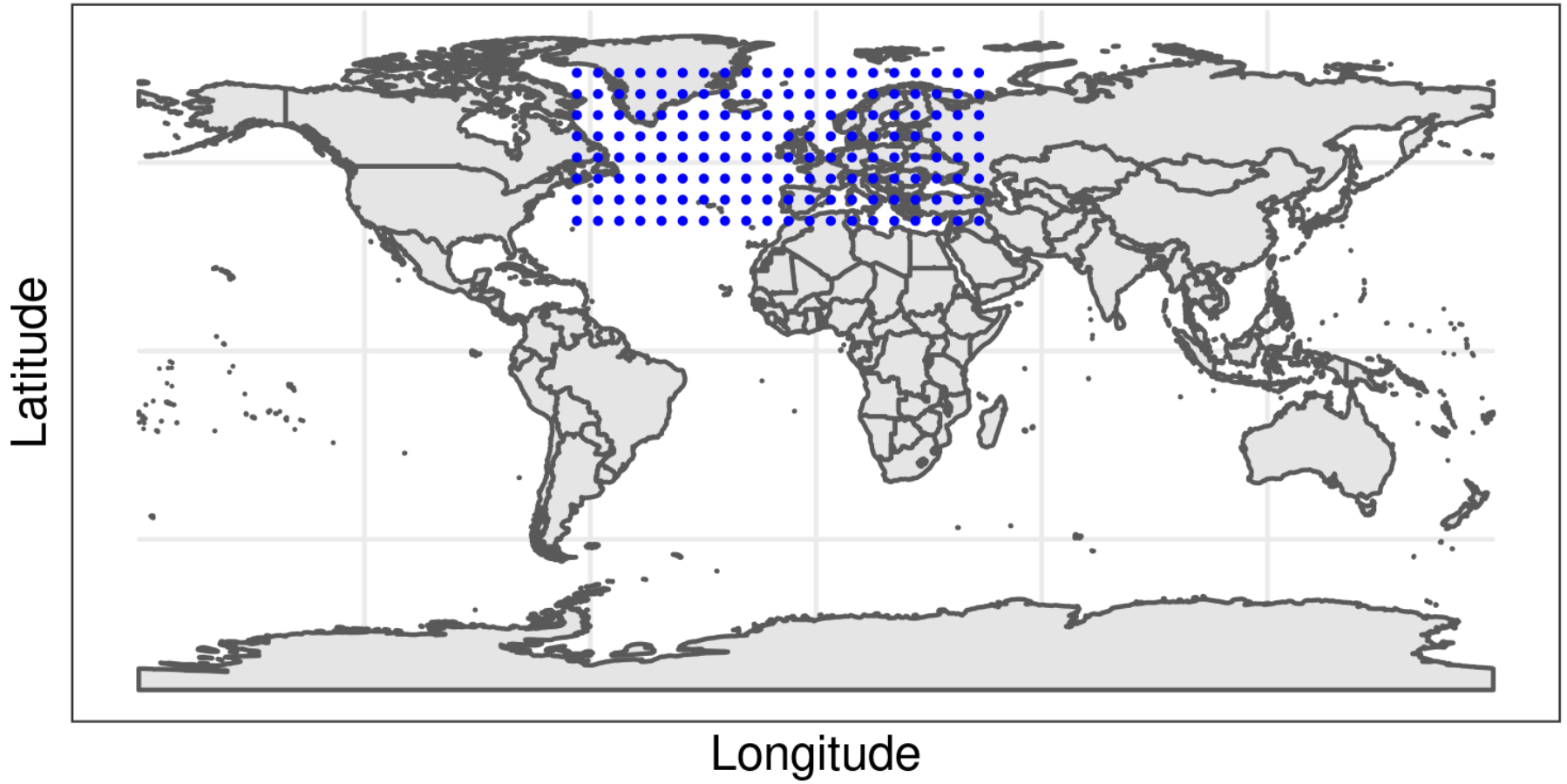
- Verteilung von GWL in den Clustern
- Vergleich der Zusammensetzung der einzelnen Cluster:
  Wie scheinen sie sich auffällig zu unterscheiden?

1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze

Reanalyse Datensatz

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

Messpunkte auf einer Weltkarte



Auszug aus dem Reanalyse Datensatz

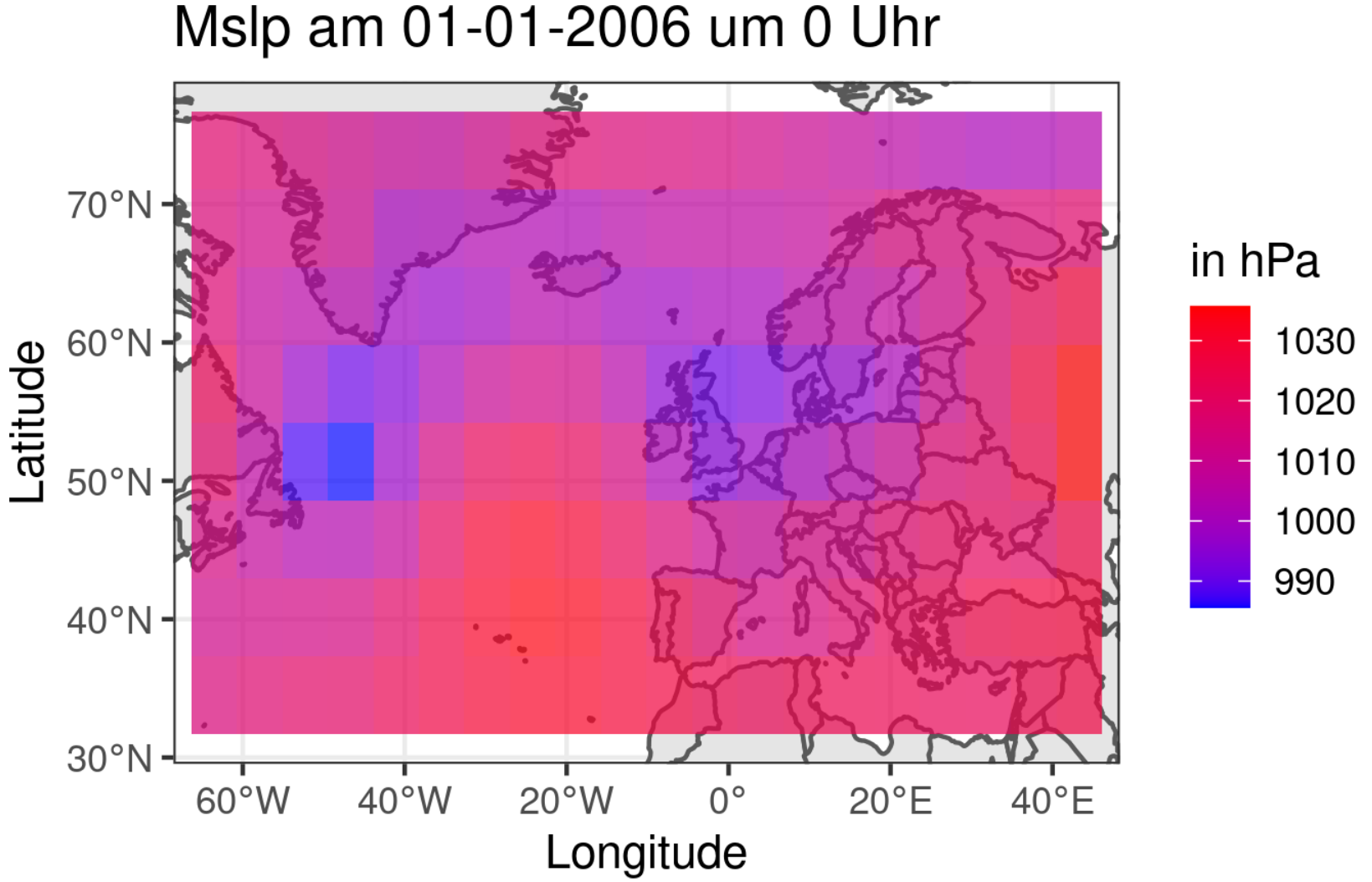
	time	longitude	latitude	mslp	geopotential
1	1900-01-01 00:00:00	-63.56287	73.85311	100428.99	48268.86
2	1900-01-01 00:00:00	-63.56287	68.23695	100553.77	48770.82
3	1900-01-01 00:00:00	-63.56287	62.62077	99920.18	49171.14
4	1900-01-01 00:00:00	-63.56287	57.00457	100049.80	49487.83
640	1900-01-01 18:00:00	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

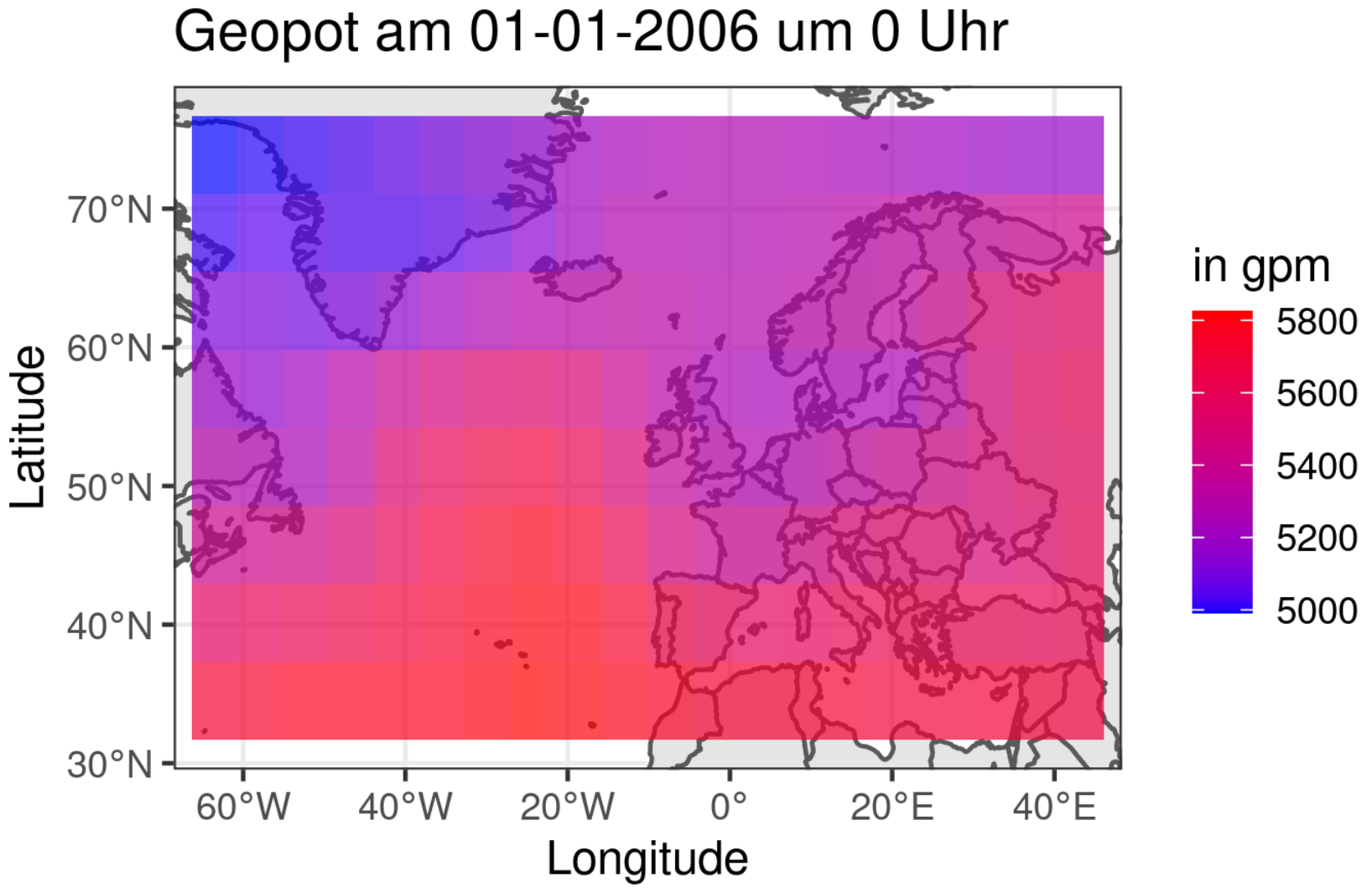
Auszug aus dem Reanalyse Datensatz

	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

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

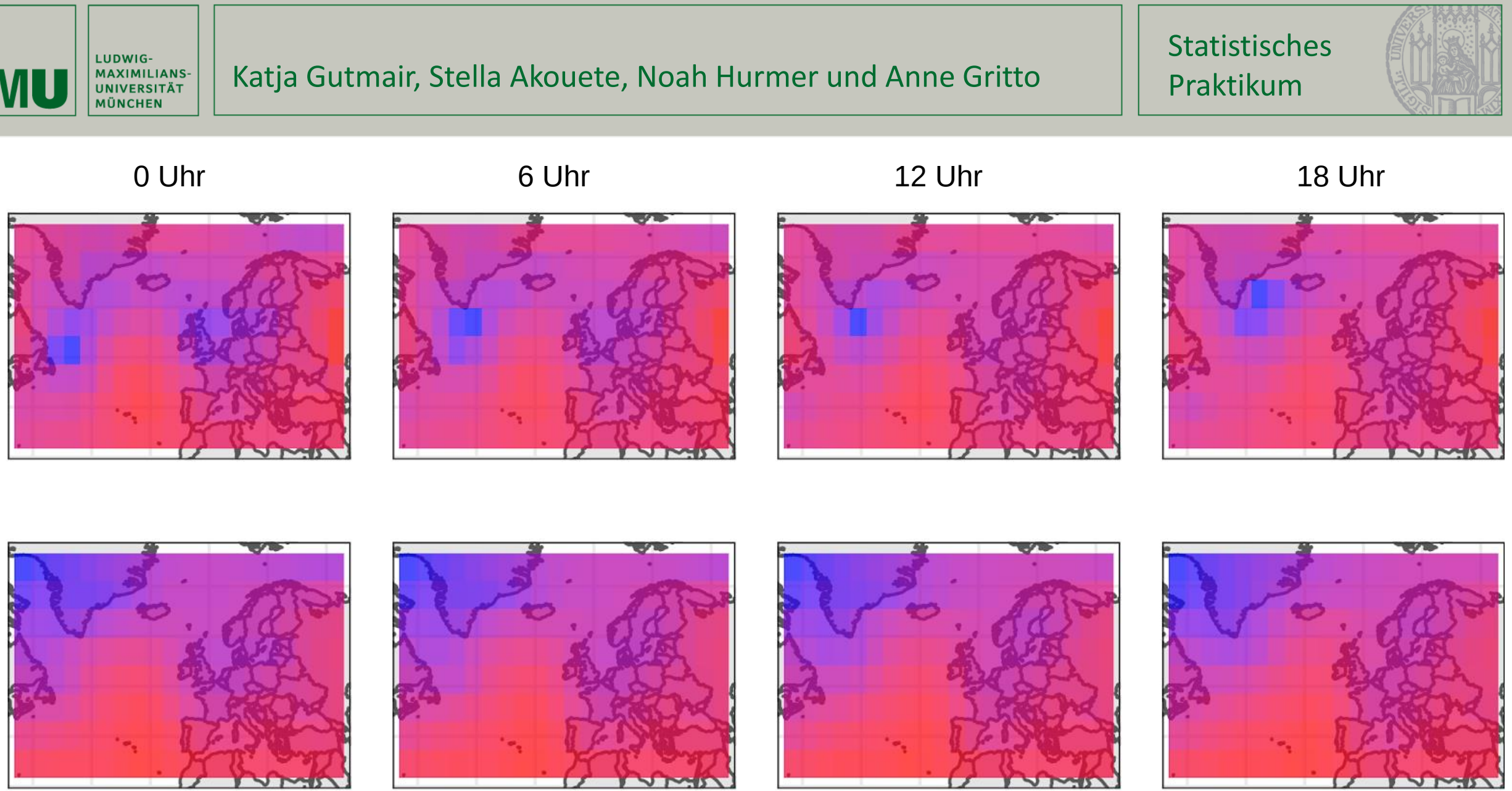

Mslp am 01-01-2006 um 0 Uhr
Latitude
70°N
60°N
50°N
40°N
30°N
60°W
40°W
20°W
0°
20°E
40°E
Longitude
in hPa
1030
1020
1010
1000
990


Geopot am 01-01-2006 um 0 Uhr
Latitude
70°N
60°N
50°N
40°N
30°N
60°W
40°W
20°W
0°
20°E
40°E
Longitude
in gpm
5800
5600
5400
5200
5000

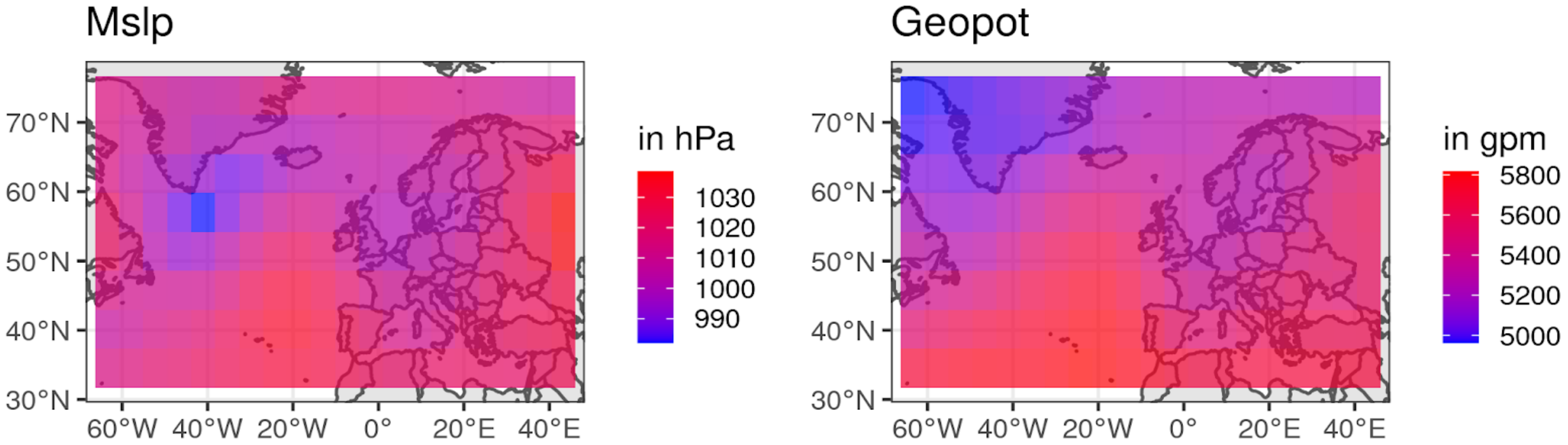
Daten pro Tag

Der Tag ist die Beobachtungseinheit

- 2 Parameter * 4 Zeitpunkte * 160 Messpunkte = 1280 Dimensionen
- 8 Bilder pro Tag


0 Uhr
6 Uhr
12 Uhr
18 Uhr
Mslp
Geopotential

Mittelwerte am 01.01.2006



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
- 10958 Tage mit jeweils 320 Dimensionen

1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing

Datensatz Mutation

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

Datensatz Mutation

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

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

Extrahierte Variablen

Variable	Erklärung
Minimum/Maximum	Minimaler/Maximaler Wert am Tag
Mittelwert	Mittelwert für beide Variablen pro Tag
Median/Quartile	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

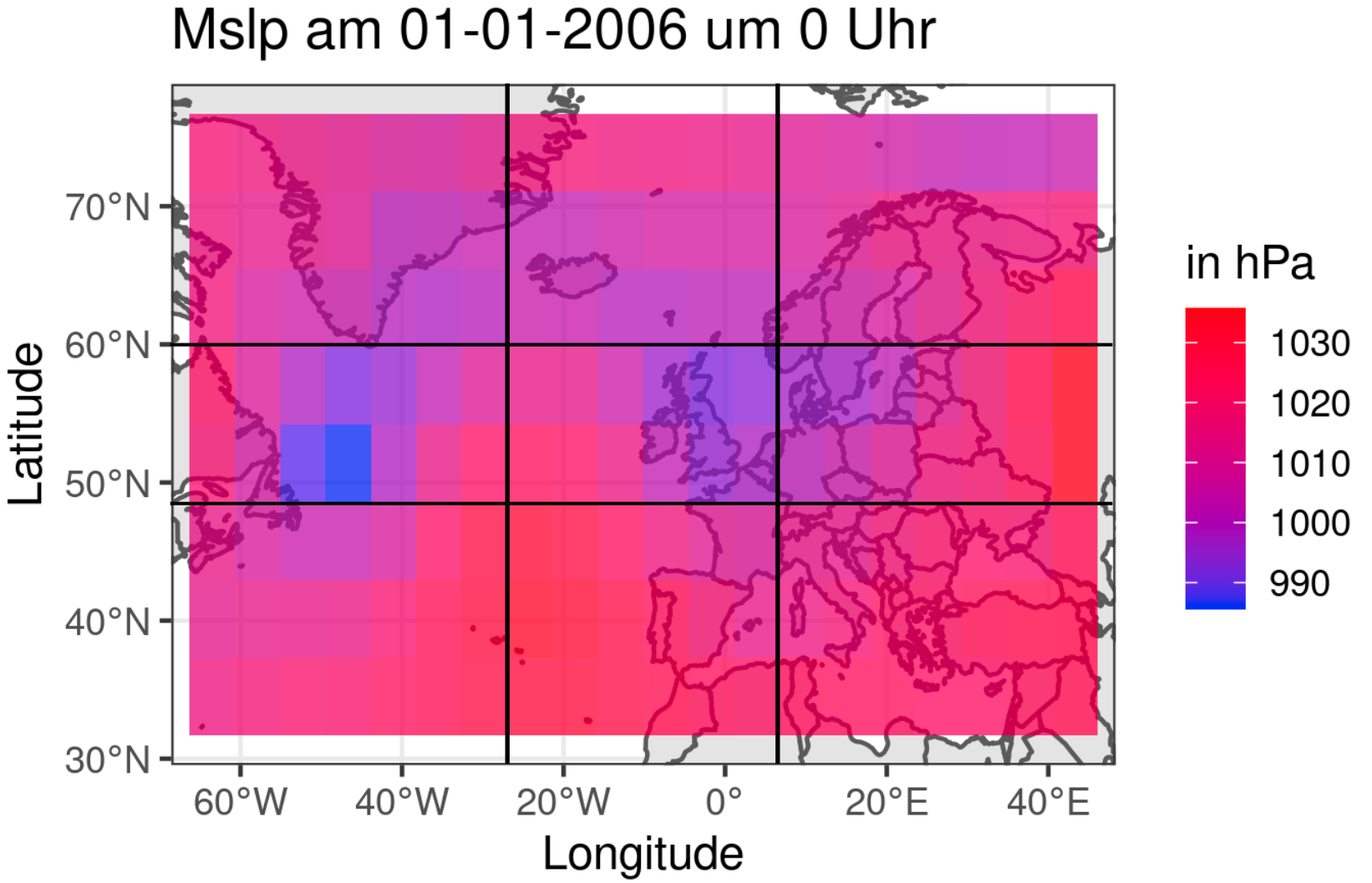
Extrahierte Variablen

Variable	Erklärung
Minimum/Maximum	Minimaler/Maximaler Wert am Tag
Mittelwert	Mittelwert für beide Variablen pro Tag
Median/Quartile	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

} Verteilungsvariablen

Extrahierte Variablen

Variable	Erklärung	
Distanz zwischen Extrema	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	} Lage der Extrema
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	


Mslp am 01-01-2006 um 0 Uhr
Latitude
70°N
60°N
50°N
40°N
30°N
60°W
40°W
20°W
0°
20°E
40°E
Longitude
in hPa
1030
1020
1010
1000
990

Extrahierte Variablen

Variable	Erklärung
Distanz zwischen Extrema	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 Variablen

Skalierung und Gewichtung

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

$$x_{i,neu} = \frac{x_i - \mu_i}{\sigma_i} \quad \text{mit } i = 1, \dots, 48$$

- Variablen werden zudem gewichtet
  - Aufgeteilt in Kategorien, die jeweils in Summe gleich gewichtet sind

Skalierung und Gewichtung

Variablen	Gewichte
Minimum, Maximum, Mittelwert	$\frac{1}{3}$
Median, Quartile, Intensität und Differenz am Tag	$\frac{1}{6}$
Euklidische Distanzen, Spalten und Zeilen vom Minimum/Maximum	$\frac{1}{6}$
Mittelwert in den Quadranten	$\frac{1}{9}$

1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Clusterverfahrens

Clusteranalyse

- Verfahren des ”unsupervised learning” (kein Target)
- Grundidee: Bildung von möglichst homogenen Gruppen, Cluster untereinander möglichst heterogen
- Betrachten von n Objekten $a_1, \dots, a_n$ mit zugehörigen Merkmalsvektoren $x_1, \dots, x_p$

  Suchen einer Partition $C_1, \dots, C_k$ mit $\bigcup_{i=1}^{k} C_i = \{a_1, \dots, a_{ni}\}$ wobei $C_i \cap C_j = \emptyset \quad \forall\, i \neq j$

- Verschiedene Ansätze für Clustering
- Distanz zwischen Objekten durch Ähnlichkeits- bzw. Distanzmaß

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

Distanzmaß

- Manhattan-Metrik
  - die Distanz d zwischen zwei Objekten a und b definiert ist als

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

wobei a $= (a_1, \ldots, a_p)$, b $= (b_1, \ldots, b_p)$

1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Algorithmus
    iii. Ergebnisse

Bewertungskriterien für Clustering

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

Bewertungskriterien für Clustering

- Silhouettenkoeffizient
  - Maßzahl für die Qualität eines Clusterings
  - Unabhängig von der Anzahl der Cluster
  - Gehört das Objekt o zum Cluster A, so ist die Silhouette von o definiert als

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

wobei $dist(A,\ o)$ die durschnittliche Distanz eines Objektes o zu anderen Punkten des Clusters A
$dist(B,\ o)$ die Distanz eines Objektes o zum nächstgelegensten Objekt des Clusters B

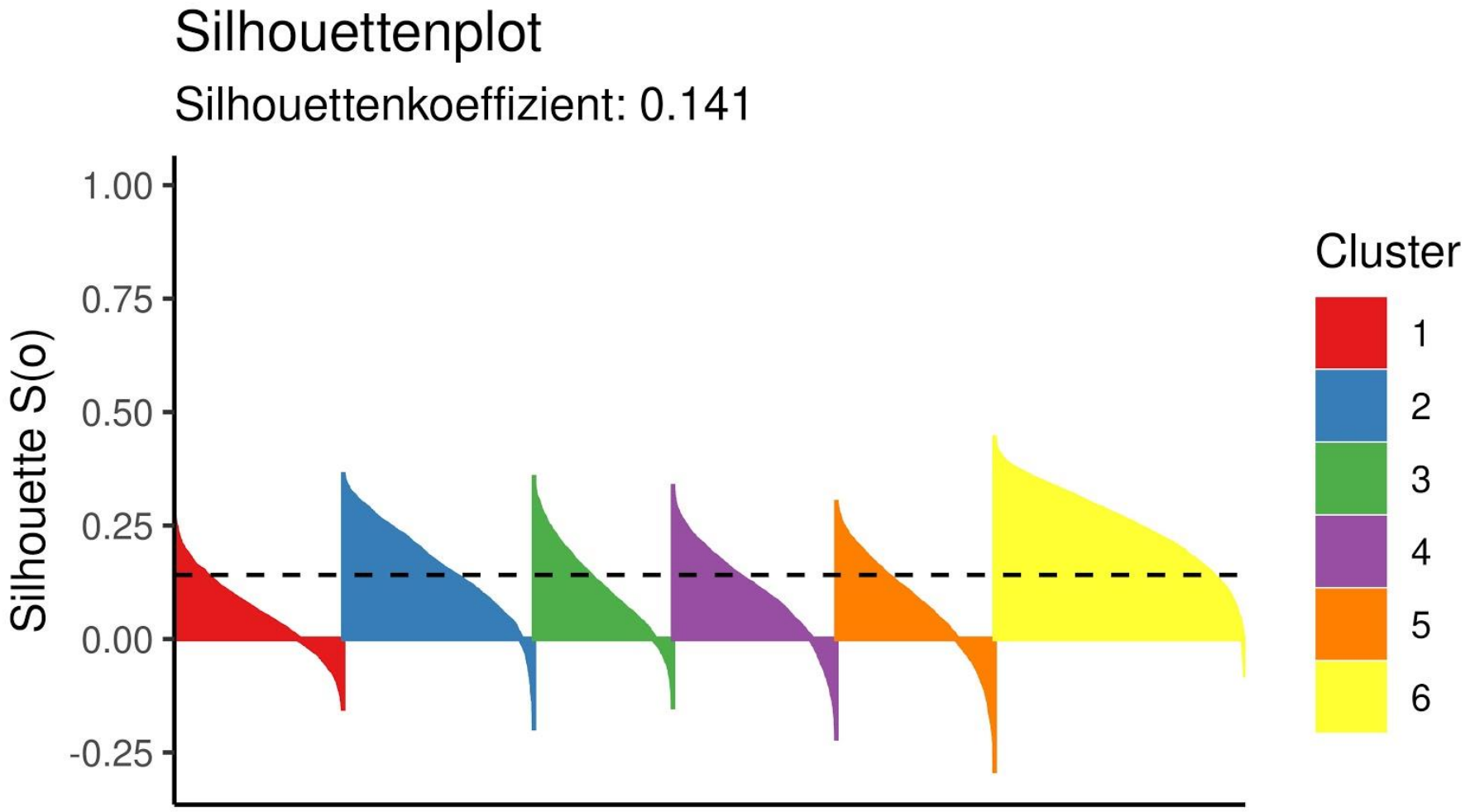
Bewertungskriterien für Clustering

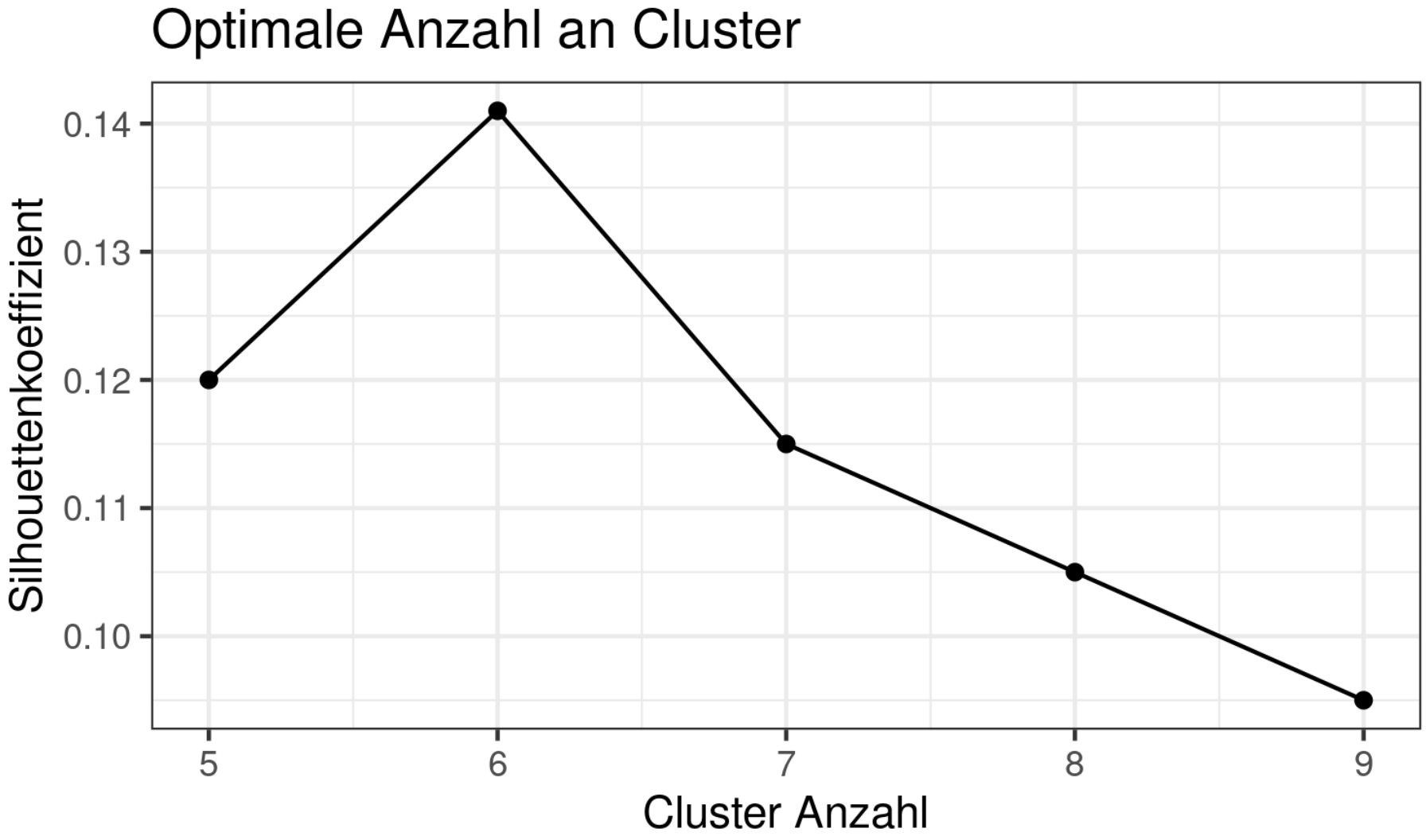
- Silhouettenkoeffizient
  - Sei k die Anzahl an Cluster, dann ist der Silhouettenkoeffizient definiert durch

$$s = \frac{1}{n} \sum_{o \in N} S(o)$$

wobei

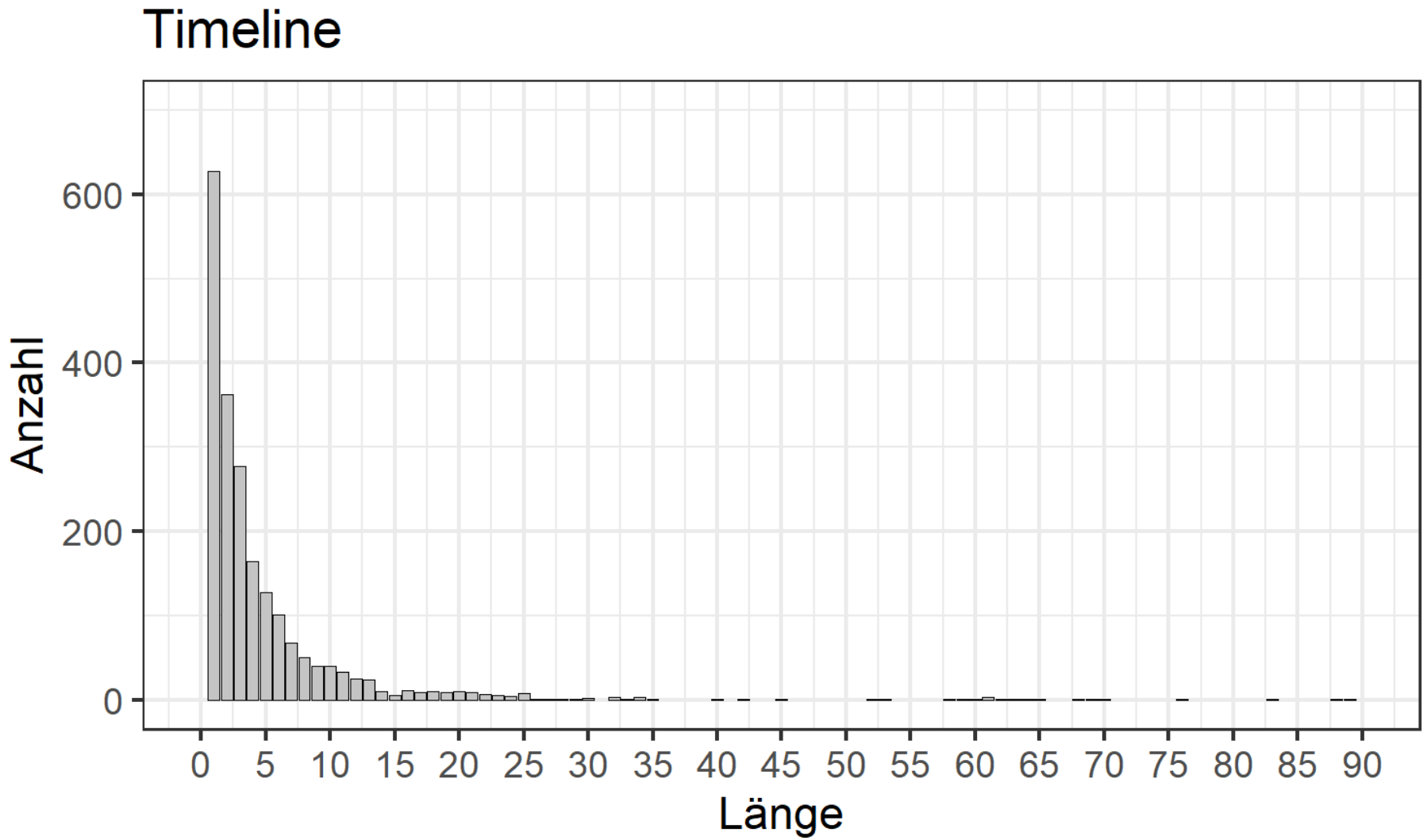
$$S(o) = \begin{cases} 0 & \text{Wenn } x \text{ einziges Element von } A, \text{ ist} \\ \frac{dist(B,o) - dist(A,o)}{\max\{dist(A,o), dist(B,o)\}} & \text{sonst,} \end{cases}$$


Silhouettenplot
Silhouettenkoeffizient: 0.141
Silhouette S(o)
1.00
0.75
0.50
0.25
0.00
-0.25
Cluster
1
2
3
4
5
6


Optimale Anzahl an Cluster
Silhouettenkoeffizient
0.14
0.13
0.12
0.11
0.10
5
6
7
8
9
Cluster Anzahl

Bewertungskriterien für Clustering

- Timeline
  - Häufigkeiten bestimmter Längen an aufeinanderfolgenden Tagen im selben Cluster
  - Erwünscht:
    - Längen ab 3 Tagen
    - Nach oben limitiert

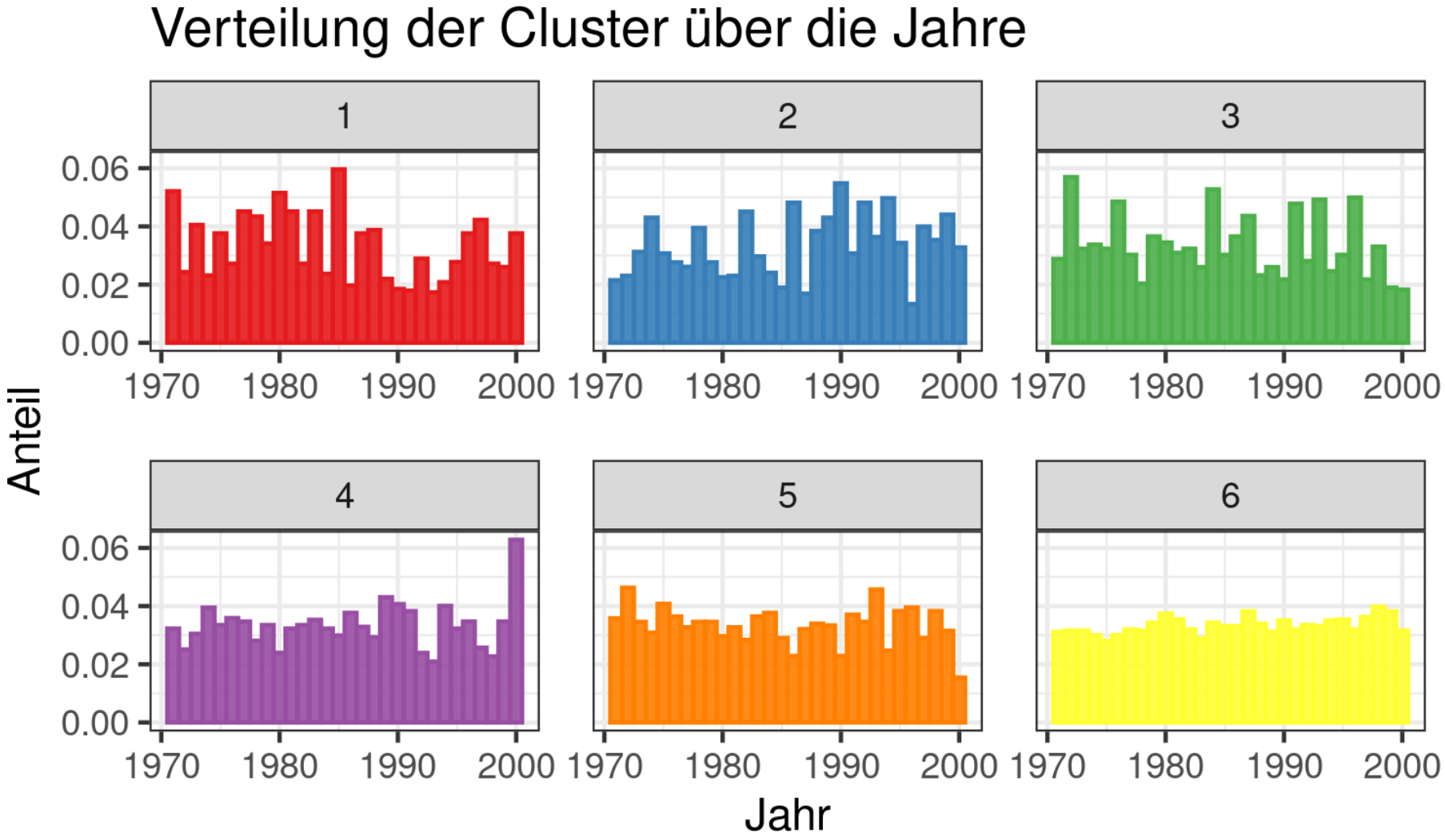

Timeline
Anzahl
600
400
200
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
Länge

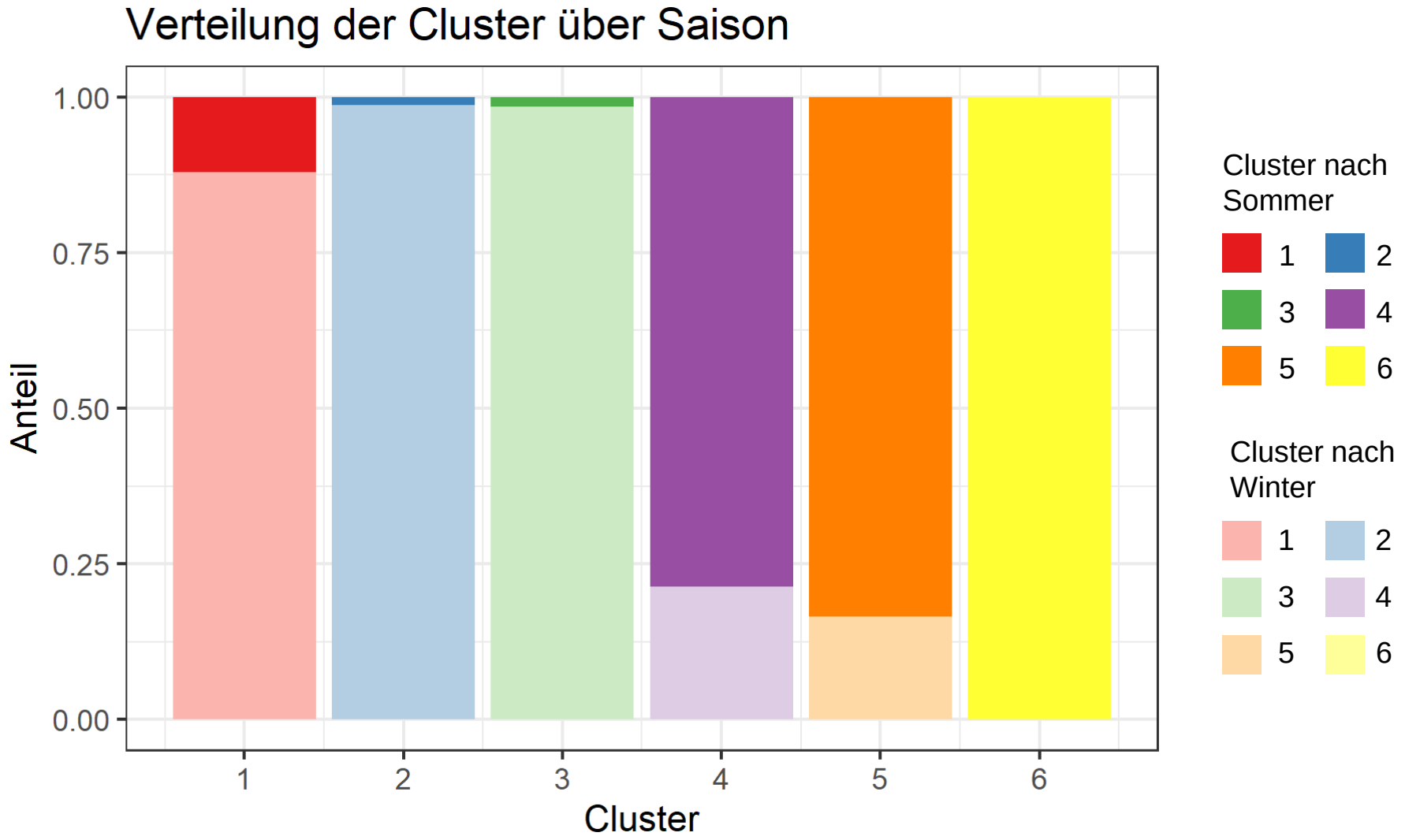
Beispiele

	date	cluster
1	1971-04-26	1
2	1971-04-27	1
3	1971-04-28	1
4	1971-04-29	4
5	1971-04-30	1
6	1971-05-01	4
7	1971-05-02	4
8	1971-05-03	4

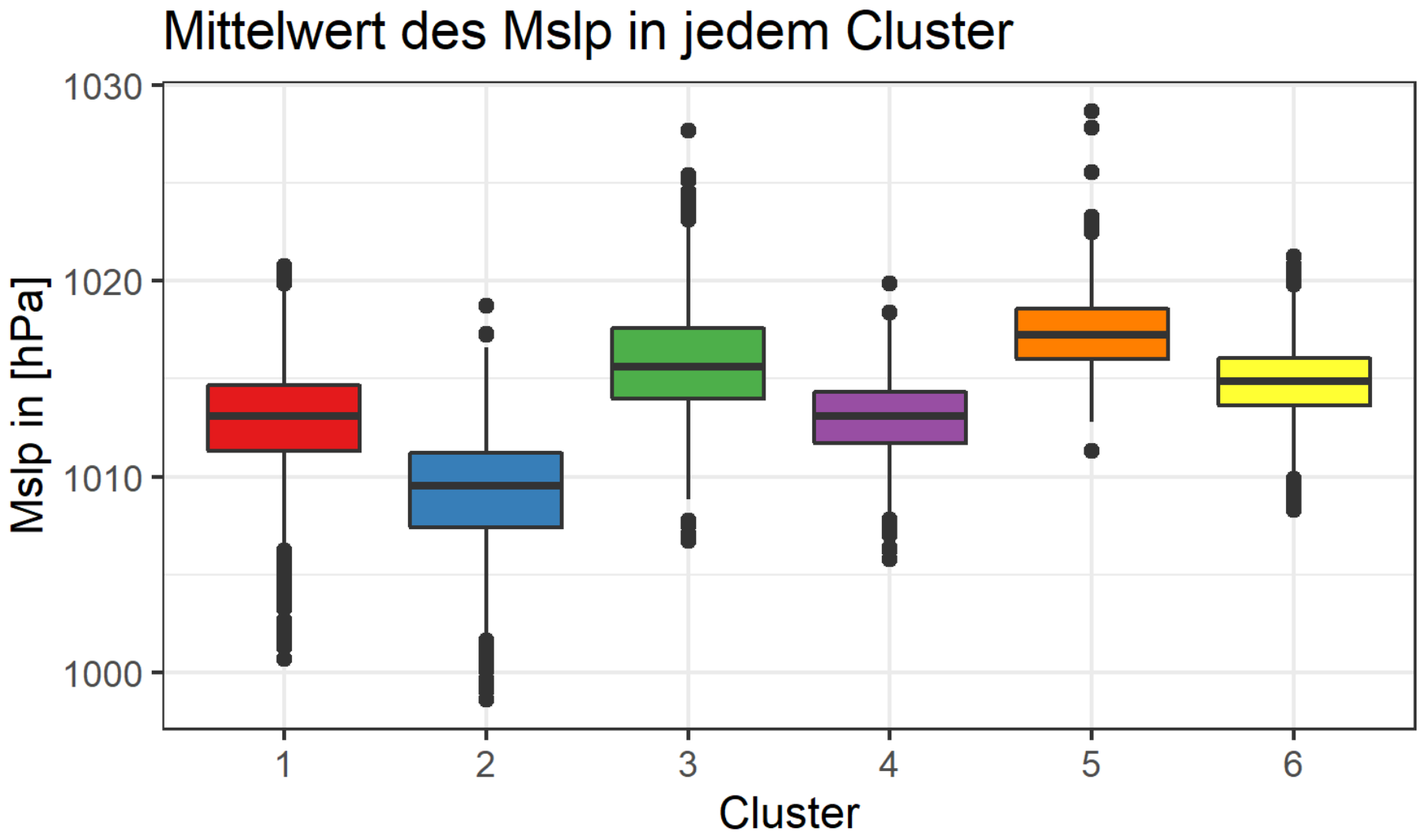
⇨ Übergang zwischen Clustern oft nicht sauber

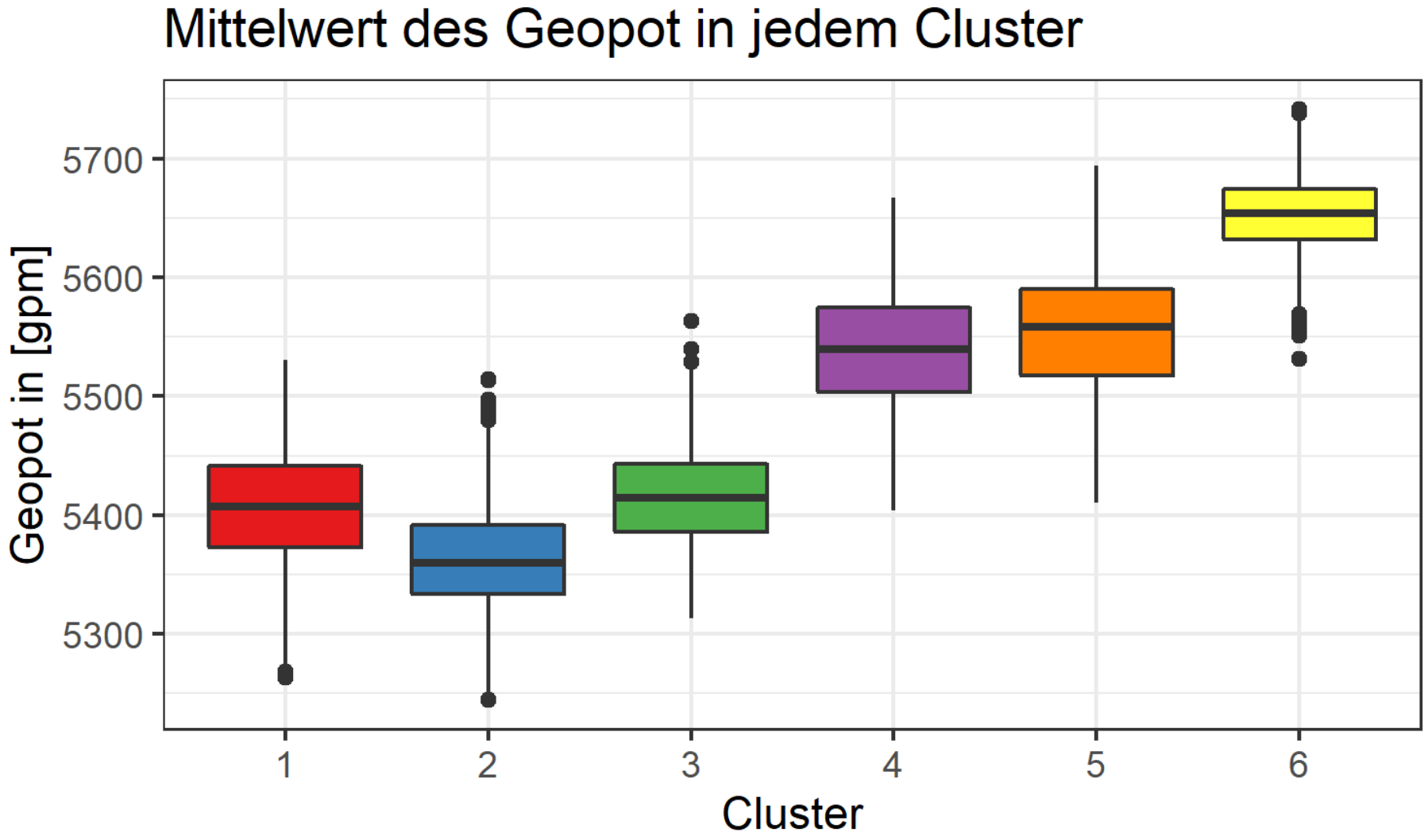
1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Algorithmus
    iii. Ergebnisse
3. Deskriptive Analyse
    i. Verteilung über die Zeit

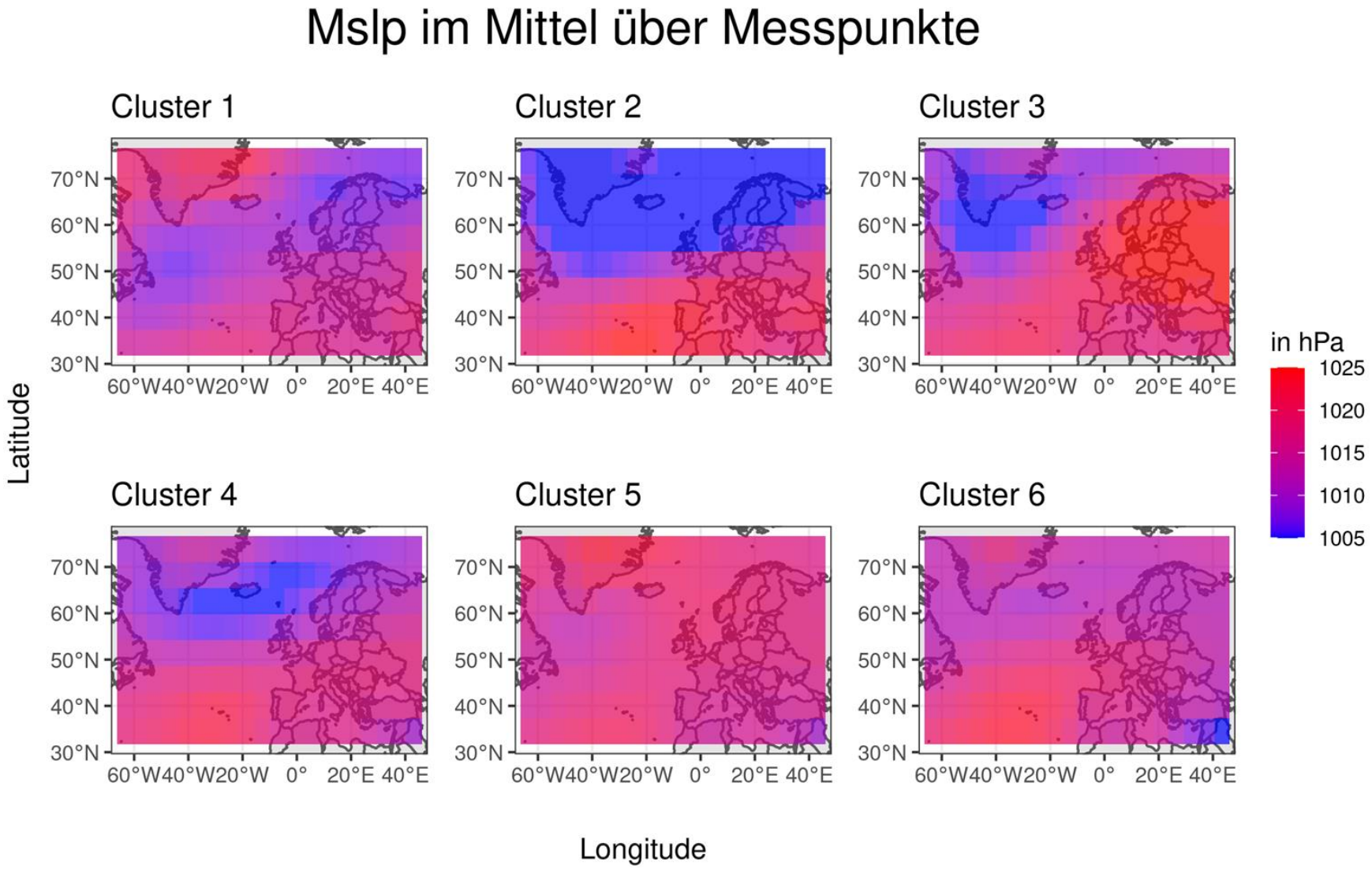

Verteilung der Cluster über die Jahre
1
2
3
4
5
6
0.06
0.04
0.02
0.00
1970
1980
1990
2000
Anteil
Jahr


Verteilung der Cluster über Saison
Anteil
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
Cluster
Cluster nach Sommer
1
2
3
4
5
6
Cluster nach Winter
1
2
3
4
5
6

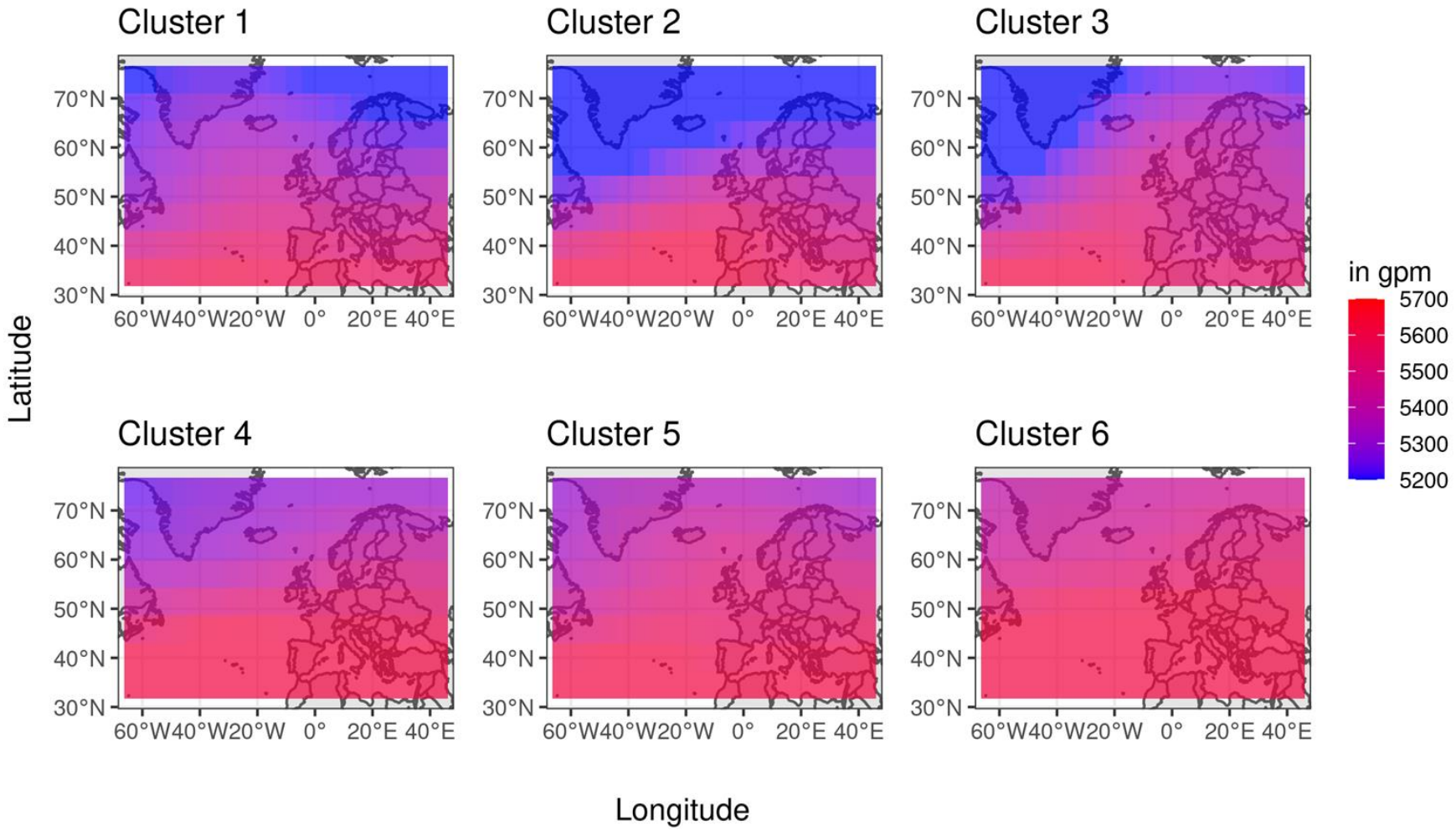
1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Algorithmus
    iii. Ergebnisse
3. Deskriptive Analyse
    i. Verteilung über die Zeit
    ii. Unterschiede und Ähnlichkeiten in den Clustern


Mittelwert des Mslp in jedem Cluster
Mslp in [hPa]
1030
1020
1010
1000
1
2
3
4
5
6
Cluster


Mittelwert des Geopot in jedem Cluster
Geopot in [gpm]
5700
5600
5500
5400
5300
1
2
3
4
5
6
Cluster

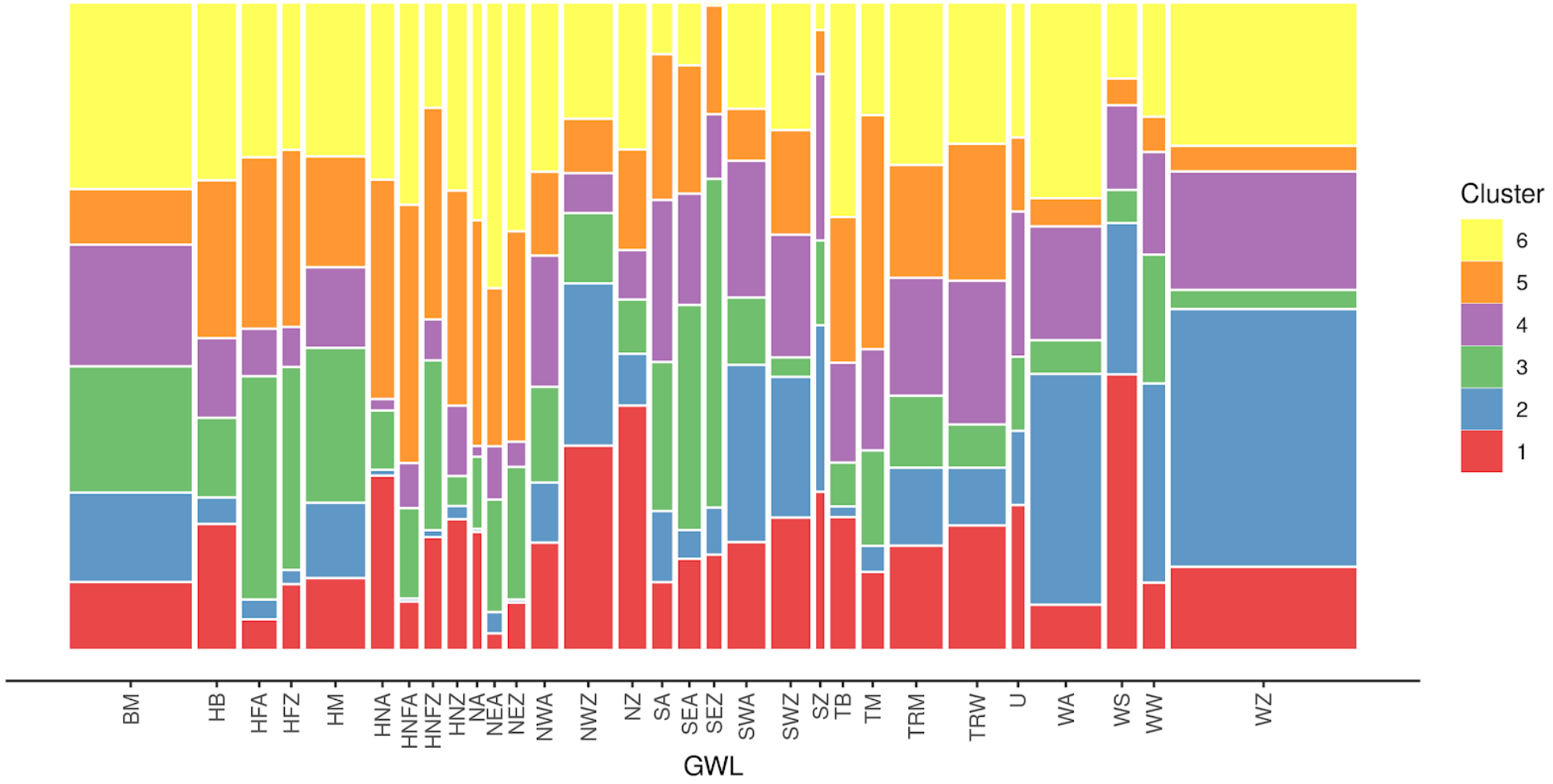

Mslp im Mittel über Messpunkte
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
70°N
60°N
50°N
40°N
30°N
60°W 40°W 20°W 0° 20°E 40°E
Latitude
Longitude
in hPa
1025
1020
1015
1010
1005

Geopot im Mittel über Messpunkte



1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Algorithmus
    iii. Ergebnisse
3. Deskriptive Analyse
    i. Verteilung über die Zeit
    ii. Unterschiede und Ähnlichkeiten in den Clustern
    iii. Vergleich zur gegebenen GWL-Einteilung

Mosaikplot für Cluster ~ GWL



Beispiele

	date	cluster	gwl
1	1971-04-22	5	SA
2	1971-04-23	5	SA
3	1971-04-24	5	SA
4	1971-04-25	1	HNZ
5	1971-04-26	1	HNZ
6	1971-04-27	1	HNZ

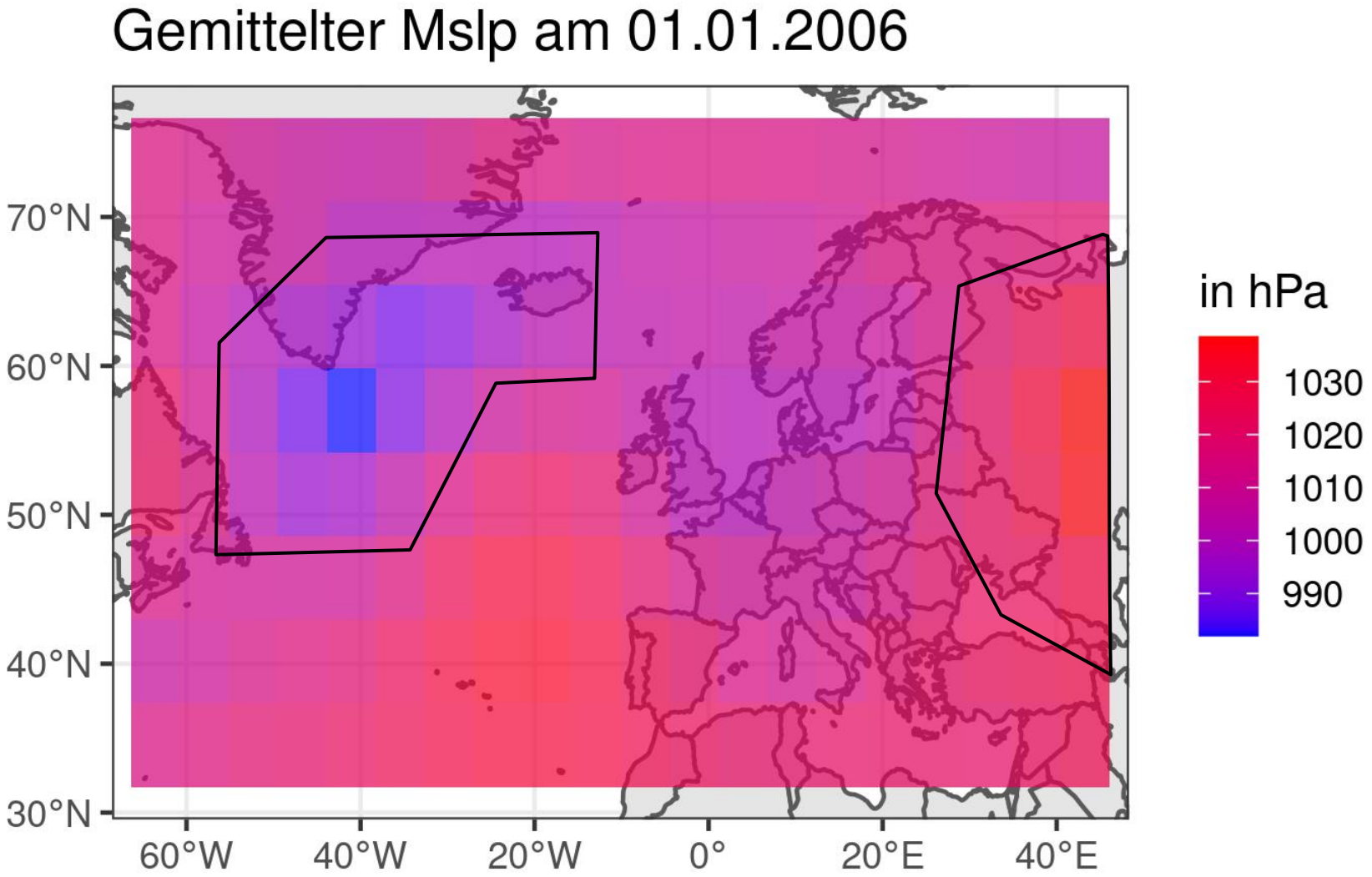
⇨ Zum Teil wechseln Cluster passend mit den GWL am Tag

Gliederung

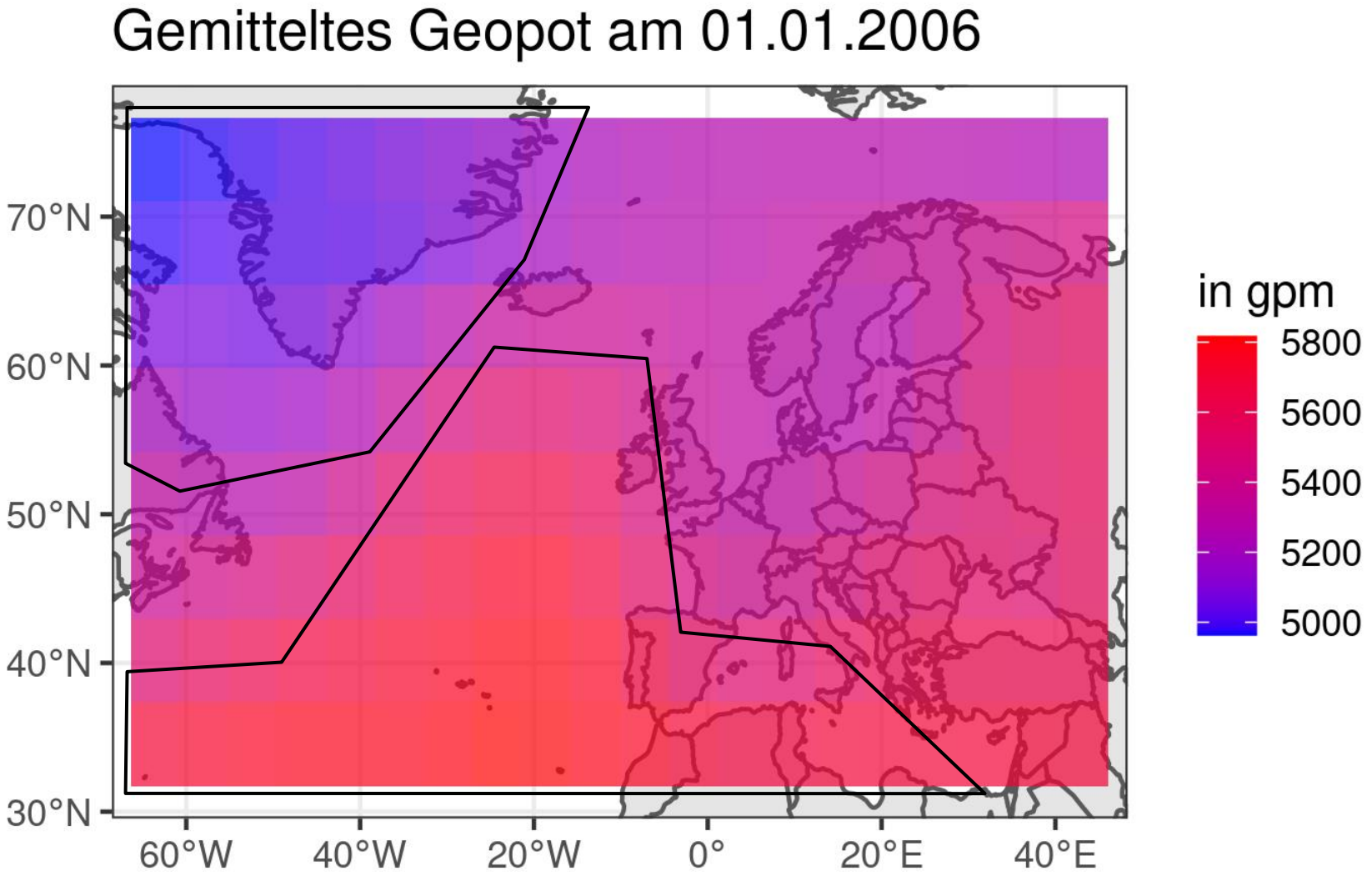
1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Algorithmus
    iii. Ergebnisse
3. Deskriptive Analyse
    i. Verteilung über die Zeit
    ii. Unterschiede und Ähnlichkeiten in den Clustern
    iii. Vergleich zur gegebenen GWL-Einteilung
4. Ausblick

Anderer Ansatz

- Muster-Erkennung in den Bildern der Tage
  - Vorfiltern der Daten pro Tag
    - ⇨ Clustern mit dem Standort als Beobachtungseinheit
  - Verwandlung Messdaten/Standort zu “Gebietszugehörigkeit”/Standort



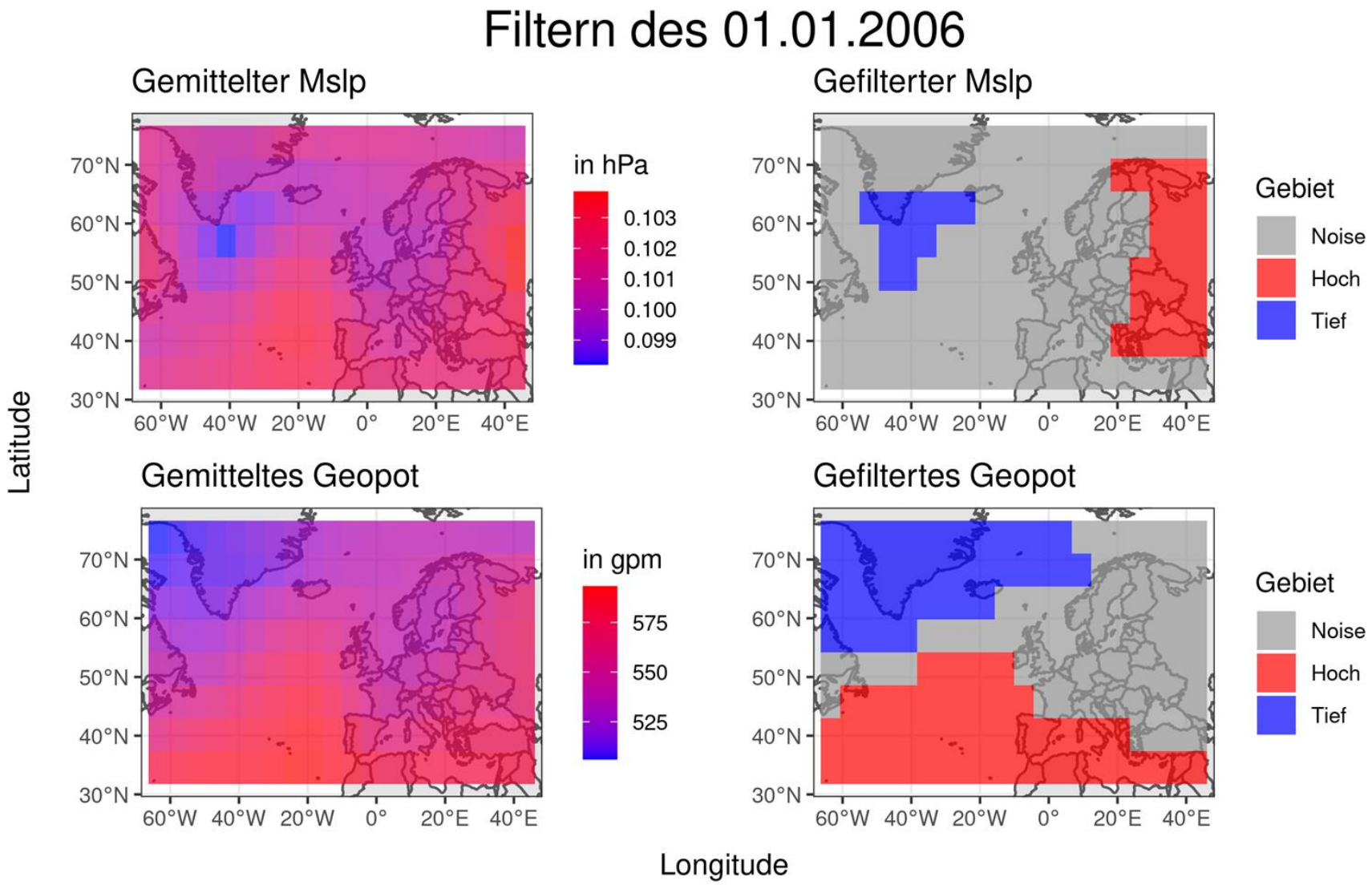
- Position und Form der “Hoch-” und “Tiefgebiete”



- Position und Form der “Hoch-” und “Tiefgebiete”

Anderer Ansatz

- Muster-Erkennung in den Bildern der Tage
  - Vorfiltern der Daten pro Tag
    - ⇨ Clustern mit dem Standort als Beobachtungseinheit
  - Verwandlung Messdaten/Standort zu “Gebietszugehörigkeit”/Standort
- Clusterverfahren
  - Dichtebasiertes Clustern mit Noise
  - Startpunkte der Cluster fix
  - Cluster iterierend wachsen lassen mit zunehmend strengem Nachbarschaftsparameter


Filtern des 01.01.2006
Gemittelter Mslp
Gefilterter Mslp
Gemitteltes Geopot
Gefiltertes Geopot
Latitude
Longitude
70°N
60°N
50°N
40°N
30°N
60°W
40°W
20°W
0°
20°E
40°E
in hPa
0.103
0.102
0.101
0.100
0.099
in gpm
575
550
525
Gebiet
Noise
Hoch
Tief

Anderer Ansatz

- Distanzberechnung zwischen Tagen

$$d(a,b) = 1 - \left(\frac{\sum I(a_i = b_i)}{\sum I(a_i)}\right)$$

wobei: $\sum I(a_i) \coloneqq Summe\ der\ Beobachtungen\ nicht\ in\ Noise$

- Weiteres Cluster auf Tagesebene mit erhaltener Distanzmatrix

Anderer Ansatz

- Probleme
  - Instabil durch Hyperparameter eps und dessen Verkleinerung
  - Sehr teuer
  - Starkes Reduzieren der gegebenen Information

Limitationen & Ausblick

- Wahl des Gewichtsvektors und der Variablen
  - Ausschlaggebend auf die Clusterbewertungskriterien
  - Fachlich sinnvoll
  - Evtl durch mit mehr Vorinformation über die Daten entscheiden
- Saison
  - Saisonbereinigung
  - Datensatz aufteilen und getrennt analysieren

Limitationen & Ausblick

- Einbeziehen der zeitlichen Struktur
  - Einführen einer 3-Tage-Regel beim Clusterverfahren
  - Datenformat als video betrachten statt Ansammlung von Bildern
- Einbeziehen weiterer Variablen
  - Anderer vorhandenen Messdaten (z.B. Temperatur)
  - Berechnung der Stömungsrichtung anhand des Bewegens bestimmter Gebiete über den Tag

Gliederung

1. Einführung
    i. Vorstellen des Projekts
    ii. Datensätze
2. Methodik
    i. Preprocessing
    ii. Wahl des Algorithmus
    iii. Ergebnisse
3. Deskriptive Analyse
    i. Verteilung über die Zeit
    ii. Unterschiede und Ähnlichkeiten in den Clustern
    iii. Vergleich zur gegebenen GWL-Einteilung
4. Ausblick
5. Fazit

Fazit

Lassen sich Tage anhand von ihren Wettermesswerten sinnvoll clustern?

- Es ist wenig Struktur erkennbar
- Instabil

Wie unterscheiden sich die entstandenen Cluster voneinander?

- Starke Unterteilung der Sommer- und Winterzeit und in den von ihnen abhängigen Variablen
- räumliche Unterscheidung auf Mslp Ebene erkennbar, beim Geopotential eher nicht

Referenzen

- Fattouh, L. & Alharbi, M. Using Modified Partitioning Around Medoids Clustering Technique in Mobile Network Planning. *International Journal of Computer Science Issues* **9** (2013).
- Hoyer, A. (ed Ludwig-Maximilians-Universität München) (Sommersemester 2020).
- James, P. M. An objective classification method for Hess and Brezowsky Grosswetterlagen over Europe. *Theoretical and Applied Climatology* **88**, 17-42, doi:10.1007/s00704-006-0239-3 (2007).
- Neuen, A. *Großwetterlagen: Die antizyklonale Westlage (WA)*, <https://wetterkanal.kachelmannwetter.com/grosswetterlagen-die-antizyklonale-westlage-wa/> (11.11.2015).
- Schwarzer. *SKlima.de, private Wetterstation Peißenberg*, <http://sklima.de/impressum.php> (2021).