



# Clusteranalyse von Wetterdaten zur Identifikation von Wetterlagetypen

Bericht zum Statistischen Praktikum im WS 2020/21

Betreuer: Prof. Dr. Helmut Küchenhoff

Projektpartner: M.Sc. Maximilian Weigert und M.Sc. Magdalena Mittermeier

Katja Gutmair, Noah Hurmer, Stella Akouete und Anne Gritto

22. März 2021

## Zusammenfassung

Um Tage anhand ihrer Wetterdaten in Gruppen oder sogenannte Wetterlagen einzuteilen, wird hier explorativ nach unsupervised Verfahren gesucht, welche dies erreichen. Dafür liegen pro Tag Messdaten zum Luftdruck auf Meeresspiegelhöhe sowie Geopotential auf 500 hPa an mehreren Standorten vor. Mit diesen Informationen werden verschiedene Clusteranalysen durchgeführt und dessen Ergebnisse verglichen sowie Bezug auf eine vorhandene Einteilung in Großwetterlagen nach Hess und Brezowsky genommen. Um Clusterergebnisse vergleichen zu können, werden die Kriterien Silhouettenkoeffizient und Timeline-Score vorgestellt und diskutiert, die den Erfolg und die fachliche Sinnhaftigkeit repräsentieren. Zwei Verfahren werden hier genauer beschrieben. Dabei befasst sich ein folgend Filter-Ansatz genanntes Verfahren mit der räumlichen Struktur bestimmter Gebiete, indem ein Algorithmus implementiert wird (*SCAPOI*), bei dem durch ein zweistufiges Clustern pro Tag Gebiete von Interesse gefunden werden, um Tage folglich anhand dessen zu vergleichen. Der zweite Ansatz extrahiert aus dem Datensatz repräsentative Variablen, um die räumliche Struktur der Daten zu entfernen, somit eine Clusteranalyse zu erleichtern sowie bestimmten Ausprägungen mehr Einfluss zuordnen zu können. Folgend wird darauf der Clusteralgorithmus *PAM* mit der Manhattan-Metrik angewandt. Für letzteren Ansatz werden Ausprägungen der Cluster deskriptiv analysiert und die Aufteilung der Großwetterlagen in den Clustern genauer betrachtet. Dabei lassen sich durchaus Unterschiede zwischen den Clustern feststellen eine saubere Aufteilung der Großwetterlagen auf Cluster allerdings nicht.

# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
1.1 Großwetterlagen . . . . .	1
<b>2 Methodik</b>	<b>3</b>
2.1 Daten . . . . .	3
2.2 Clusterbewertungskriterien . . . . .	5
2.2.1 Silhouettenkoeffizient . . . . .	5
2.2.2 Timeline . . . . .	6
2.2.3 Vergleich zur GWL-Aufteilung . . . . .	8
2.3 Filter Ansatz . . . . .	8
2.3.1 Motivation . . . . .	8
2.3.2 Prinzip Filtern . . . . .	8
2.3.3 DBSCAN und Fuzzy . . . . .	9
2.3.4 SCAPOI . . . . .	10
2.3.5 Distanzmetrik . . . . .	12
2.3.6 Kriterien Filtern . . . . .	13
2.4 Clustern mit extrahierten Daten . . . . .	13
2.4.1 Extrahieren der Variablen . . . . .	14
2.4.2 Clusterverfahren . . . . .	16
2.4.2.1 Partitioning Around Medoids . . . . .	17
2.4.2.2 Skalierung . . . . .	17
2.4.2.3 Gewichtung . . . . .	18
2.4.2.4 Distanzmetrik . . . . .	19
2.4.2.5 Wahl der Clusteranzahl . . . . .	19
2.5 Weitere Versuche . . . . .	20
2.5.1 CLARA, K-Means und PAM mit Originaldaten . . . . .	20
2.5.2 K-Means und PAM mit Extrahierten Daten . . . . .	22
<b>3 Ergebnisse PAM mit extrahierten Variablen</b>	<b>23</b>
3.1 Verteilung der Cluster über die Zeit . . . . .	24
3.2 Verhältnis von Sommer- und Wintertagen in den Clustern . . . . .	25
3.3 Unterschiede und Ähnlichkeiten in den Clustern . . . . .	27
3.4 Vergleich der Clusterlösung mit den GWL . . . . .	31
<b>4 Diskussion und Ausblick</b>	<b>32</b>
4.1 Bewertungskriterien . . . . .	32
4.2 Saison . . . . .	32
4.2.1 Clusterergebnis nach Saison . . . . .	34
4.2.2 Clustern getrennt nach Saison . . . . .	35

4.3	CWL-Mindestlänge 3 Tage . . . . .	36
4.4	GWL-Mindestlänge 4 Tage . . . . .	37
4.5	Erweiterung Variablen . . . . .	37
4.6	PAM und Filter . . . . .	38
4.7	Räumliche Struktur . . . . .	39
4.8	Gewichtungsvektor . . . . .	40
4.8.1	Cluster-Boosting . . . . .	40
4.9	Zeitliche Struktur . . . . .	41
<b>5</b>	<b>Schluss</b>	<b>42</b>
<b>6</b>	<b>Referenzen</b>	<b>43</b>

## Tabellenverzeichnis

1	Großwetterlagen nach Hess und Brezowsky . . . . .	2
2	Extrahierte Variablen . . . . .	16
3	Gewichte für Variablen . . . . .	18
4	Mittelwert und Standardabweichung ausgewählter Variablen pro Cluster. Alle Werte, die den Luftdruck betreffen, sind in der Einheit hPa, alle Werte, die das Geopotential betreffen, sind in der Einheit gpm . . . . .	28
5	R Libraries . . . . .	43

## Abbildungsverzeichnis

1	Standorte der 160 Messpunkte der Variablen Luftdruck und Geopotential auf der Weltkarte . . . . .	3
2	Messwerte des Mslp an den 160 Messpunkten am 01.01.1980 um 0 Uhr . . . . .	4
3	Messwerte des Geopotentials an den 160 Messpunkten am 01.01.1980 um 0 Uhr . . . . .	4
4	Über den Tag gemittelte Messwerte an den 160 Messpunkten am 01.01.1980. Hierbei wurde bei den Variablen Luftdruck und Geopotential das arithmetische Mittel eines Standorts über die vier Tageszeitpunkte (0 Uhr, 6 Uhr, 12 Uhr, 18 Uhr) gebildet. . . . .	5
5	Timeline der GWL 1971-2000. Die Anzahl der GWL wurde mit deren Länge multipliziert, sodass jeder Tag eine Beobachtung darstellt . . . . .	6
6	Erwünschte Verteilung der Timeline einer Clusterlösung . . . . .	7
7	10-NN Distanz mit Punkt der maximalen Wölbung . . . . .	11
8	Beispiel der Gebietszuteilung eines Tages durch SCAPOI . . . . .	12
9	Aufteilung der 160 Standorte in 9 Quadranten . . . . .	15
10	Beispiel für die Manhattan-Distanz im zweidimensionalen Raum . . . . .	19
11	Optimale Anzahl an Cluster. Die Clusteranzahl mit dem maximalen Silhouettenkoeffizient ist die optimale Clusteranzahl, hier $k = 6$ . . . . .	20
12	Visualisierung der Clusterlösung mittels Principal Component Analysis. . . . .	22
13	Silhouettenplot für Clusterergebnis mit extrahierten Variablen mit $k = 6$ . . . . .	23
14	Timeline für Clusterergebnis mit extrahierten Variablen mit $k = 6$ . Die rote Linie zeigt die optimale Timeline Verteilung. . . . .	24
15	Aufteilung der Tage auf die Jahre getrennt nach Cluster. Pro Cluster wird dargestellt, welcher relative Anteil aller Tage, die diesem Cluster zugeordnet sind, sich in einem Jahr befinden. . . . .	25
16	Darstellung der Verteilung der Variable Mittelwert des Luftdrucks getrennt nach Sommer- und Winterzeit im Zeitraum 1971-2000 . . . . .	26
17	Darstellung der Verteilung der Variable Mittelwert des Luftdrucks getrennt nach Sommer- und Winterzeit im Zeitraum 1971-2000 . . . . .	26

18	Gestapeltes Balkendiagramm, der den relativen Anteil an Winter- und Sommertagen je Cluster abbildet . . . . .	27
19	Verteilung der Tage im Zeitraum 1971 - 2000 auf die Cluster 1 bis 6 . . . . .	28
20	Verteilung der Variable Mittelwert des Luftdrucks in jedem Cluster . . . . .	29
21	Verteilung der Variable Mittelwert des Geopotentials in jedem Cluster . . . . .	29
22	Räumliche Verteilung des Luftdrucks je Cluster. Für jeden der 160 Standorte wurde das arithmetische Mittel der Variable Geopotential über alle Tage im Zeitraum von 1971-2000 berechnet. . . . .	30
23	Räumliche Verteilung des Geopotentials je Cluster. Für jeden der 160 Standorte wurde das arithmetische Mittel der Variable Geopotential über alle Tage im Zeitraum von 1971-2000 berechnet. . . . .	31
24	Mosaikplot, der darstellt, mit welchem Anteil die GWL auf die Cluster 1 bis 6 aufgeteilt sind . . . . .	32
25	Werte des Luftdrucks der Jahre 1971 - 2000 aufgeteilt nach 4 Jahreszeiten . . . . .	33
26	Werte des Geopotentials der Jahre 1971 - 2000 aufgeteilt nach 4 Jahreszeiten . . . . .	33
27	Mosaikplot nur für Sommertage, der darstellt, mit welchem Anteil die GWL auf die Cluster 1 bis 6 aufgeteilt sind . . . . .	34
28	Mosaikplot nur für Wintertage, der darstellt, mit welchem Anteil die GWL auf die Cluster 1 bis 6 aufgeteilt sind . . . . .	35
29	Mosaikplot, der darstellt, mit welchem Anteil die GWL auf die Cluster 1 bis 6 aufgeteilt sind. Hierbei sind nur Tage abgebildet, wo mindestens drei aufeinanderfolgende Tage dem selben Cluster zugeordnet sind . . . . .	36
30	Timeline des Clusterergebnis von der Kombination der extrahierten Daten mit dem Filtern der Msdp Daten. Die rote Line bezeichnet die erwünschte Timeline Verteilung.	38
31	Mosaikplot der Cluster - GWL Einteilung des Clusterergebnis von der Kombination der extrahierten Daten mit dem Filtern der Msdp Daten. . . . .	39
32	Verteilung der vertikalen Position der Geopotential Extrema von 1971-2000 . . . . .	40

## Abkürzungen

Ausdruck	Kurzform
Cluster-Wetterlage (aufeinanderfolgende Tage mit demselben Clusterindex)	<i>CWL</i>
Clusteranzahl	<i>k</i>
Clustering Large Applications	<i>CLARA</i>
Density-Based Spatial Clustering of Applications with Noise	<i>DBSCAN</i>
Geopotential auf 500 hPa in $m^2/s^2$	<i>Geopotential</i>
geopotentieller Meter	<i>gpm</i>
Großwetterlage	<i>GWL</i>
GWL Aufteilungswert	<i>HB<sub>diff</sub></i>
Interquartilsabstand	<i>IQR</i>
Kovarianzmatrix	$\Sigma$
Luftdruck in Pascal auf Meeresspiegelhöhe	<i>Mslp</i>
Partitioning Around Medoids	<i>PAM</i>
Principle Component (Analysis)	<i>PC / PCA</i>
Spatial Clustering around Points of Interest	<i>SCAPOI</i>
Timeline-Score	<i>TLS</i>

# 1 Einleitung

Auf den Klimawandel und seine Auswirkungen wird seit den letzten Jahren immer mehr Fokus gelegt. Durch diesen können sich die Häufigkeit und Länge der auftretenden Großwetterlagen (*GWL*) ändern. Lang anhaltende *GWL* können problematisch sein, da sie zu Wetteranomalien führen und Variablen wie zum Beispiel Temperatur, Regen und Wind beeinflussen können (James (2007)). Magdalena Mittermeier vom Department für Geographie und Maximilian Weigert vom Statistischen Institut der LMU untersuchen, wie sich das Auftreten verschiedener Großwetterlagen unter dem Einfluss des Klimawandels verändern kann. Dafür soll ein Modell zur automatisierten Klassifikation der *GWL* nach Hess und Brezowsky anhand von bestimmten Klimaparametern entwickelt werden. Zudem sollen Großwetterlagen charakterisiert und eine Analyse von Trends in deren Entwicklung über die Zeit durchgeführt werden.

In diesem Bericht wird untersucht, inwieweit die 29 *GWL* nach Hess und Brezowsky zu allgemeineren Typen von Wetterlagen zusammengefasst werden können. Dabei werden Tage anhand von beobachteten Wetterdaten in Cluster aufgeteilt. Die Klimaparameter für das Clustering sind der Luftdruck auf Meeresspiegelhöhe sowie das Geopotential auf 500hPa. Es soll untersucht werden, ob sich die einzelnen Tage sinnvoll in Cluster einteilen lassen, ohne die Information der herrschenden *GWL* zu benutzen.

## 1.1 Großwetterlagen

Nach dem Katalog von Hess und Brezowsky, welcher 1952 veröffentlicht wurde, sind 29 Großwetterlagen definiert (P. C. Werner (2010)). Diese werden nach Wetterstation Peißenberg (2021) als “ein bestimmter atmosphärischer Zustand, der in seiner charakteristischen Strömungsanordnung mehrere Tage im Wesentlichen gleich bleibt (Definition nach BAUR)” beschrieben. Im Allgemeinen gilt, dass eine *GWL* mindestens drei Tage oder länger andauert. Falls es vorkommt, dass der Übergang von einer in die andere Großwetterlage nicht sehr eindeutig ist, werden diese Tage zum Teil der nachfolgenden oder vorherigen *GWL* zugeteilt, wenn diese schon längere Zeit andauerte. Sonst werden diese als “unbestimmt” (U) definiert (P. C. Werner (2010)).

Hess und Brezowsky definieren *GWL* auf räumlicher Ebene über Europa und Teilen des Atlantik. Grundlagen dieser Klassifikation sind Zirkulationsformen, die sich aus der gemischten, meridionalen und zonalen zusammensetzen (Ernst (1995)). Diese werden durch die “Lage der steuernden Zentren (Höhenhoch- und Höhentiefdruckgebiete und Tröge) und durch die Erstreckung der Frontalzonen bestimmt” (P. C. Werner (2010)). Außerdem sind Strömungsanordnungen in der Höhe, sowie der Luftdruck auf Meeresspiegelhöhe wichtig für die Zuordnung zu einer Zirkulationsform. Zudem wurde für die Klassifikation von Großwetterlagen die Zugrichtung wandernder Druckgebilde betrachtet (P. C. Werner (2010)). In Tabelle 1 sind alle Großwetterlagen nach Hess und Brezowsky zusammengefasst (Wetterstation Peißenberg (2021)).

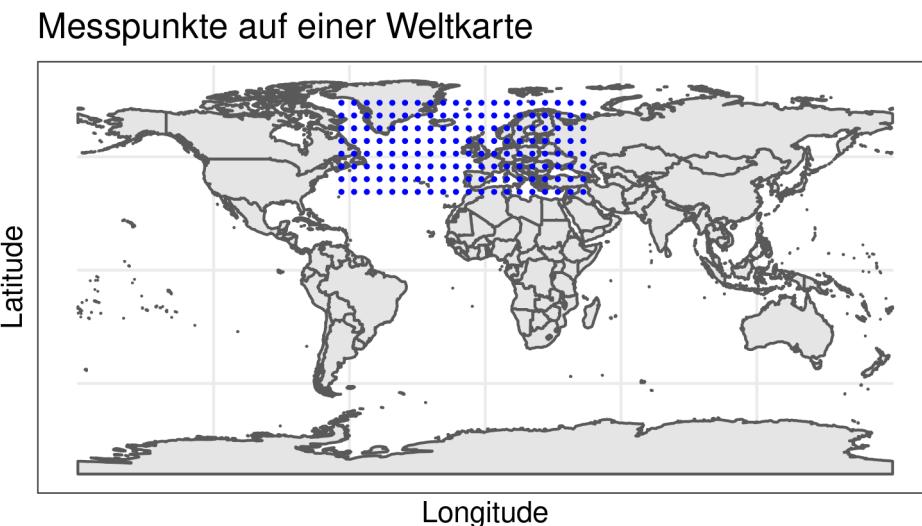
**Tabelle 1:** Großwetterlagen nach Hess und Brezowsky

Großwetterlage	Abkürzung
Westlage, antizyklonal	WA
Westlage, zyklonal	WZ
Südliche Westlage	WS
Winkelförmige Westlage	WW
Südwestlage, antizyklonal	SWA
Südwestlage, zyklonal	SWZ
Nordwestlage, antizyklonal	NWA
Nordwestlage, zyklonal	NWZ
Hoch Mitteleuropa	HM
Hochdruckbrücke (Rücken) Mitteleuropa	BM
Tief Mitteleuropa	TM
Nordlage, antizyklonal	NA
Nordlage, zyklonal	NZ
Hoch Nordmeer-Island, antizyklonal	HNA
Hoch Nordmeer-Island, zyklonal	HNZ
Hoch Britische Inseln	HB
Trog Mitteleuropa	TRM
Nordostlage, antizyklonal	NEA
Nordostlage, zyklonal	NEZ
Hoch Fennoskandien, antizyklonal	HFA
Hoch Fennoskandien, zyklonal	HFZ
Hoch Nordmeer-Fennoskandien, antizyklonal	HNFA
Hoch Nordmeer-Fennoskandien, zyklonal	HNFZ
Südostlage, antizyklonal	SEA
Südostlage, zyklonal	SEZ
Südlage, antizyklonal	SA
Südlage, zyklonal	SZ
Tief Britische Inseln	TB
Trog Westeuropa	TRW
Übergang/unbestimmt	U

## 2 Methodik

### 2.1 Daten

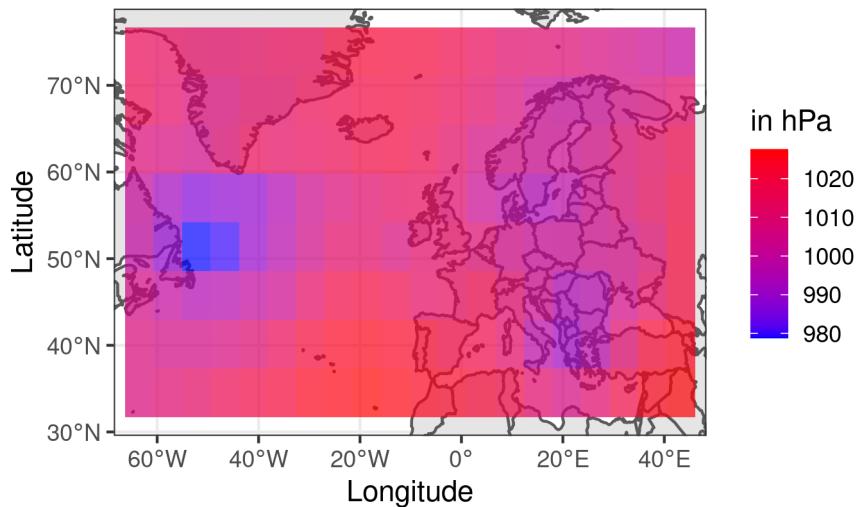
Seit 1900 wird viermal täglich an 160 Standorten der Luftdruck in Pascal auf Meeresspiegelhöhe (Mslp) und das Geopotential auf  $500\text{hPa}$  in  $\frac{\text{m}^2}{\text{s}^2}$  erhoben. Im Folgenden wird für Visualisierungen der Luftdruck in hPa ( $1\text{hPa} = 100\text{Pa}$ ) und das Geopotential in gpm ( $1\text{gpm} = 9.80665\frac{\text{m}^2}{\text{s}^2}$ ) (Karren (Zugriffen: 2021-03-19)) dargestellt. Die Standorte befinden sich in einem  $8 \times 20$  Grid über Europa und Teilen des Atlantik, wie in Abb. 1 dargestellt ist.



**Abbildung 1:** Standorte der 160 Messpunkte der Variablen Luftdruck und Geopotential auf der Weltkarte

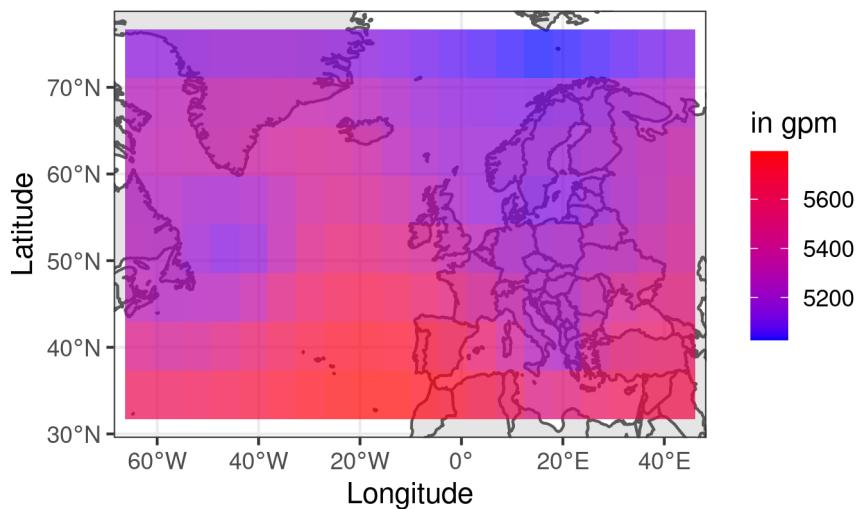
Die Variablen sind Teil des Reanalyse-Datensatzes ERA-20C, der für die Analysen zur Verfügung steht. Zudem liegen die *GWL* für jeden Tag im Zeitraum von 1900 bis 2010 vor. Für die folgenden Clusteranalysen wird nur mit dem Reanalyse-Datensatz geclustert. Die Abbildungen 2 und 3 visualisieren die Werte des Luftdrucks, bzw. Geopotentials am 01.01.1980 um 0 Uhr über alle 160 Standorte. Dieser Tag wird im weiteren Verlauf als Beispieltag benutzt.

Mslp am 01-01-1980 um 0 Uhr



*Abbildung 2:* Messwerte des Mslp an den 160 Messpunkten am 01.01.1980 um 0 Uhr

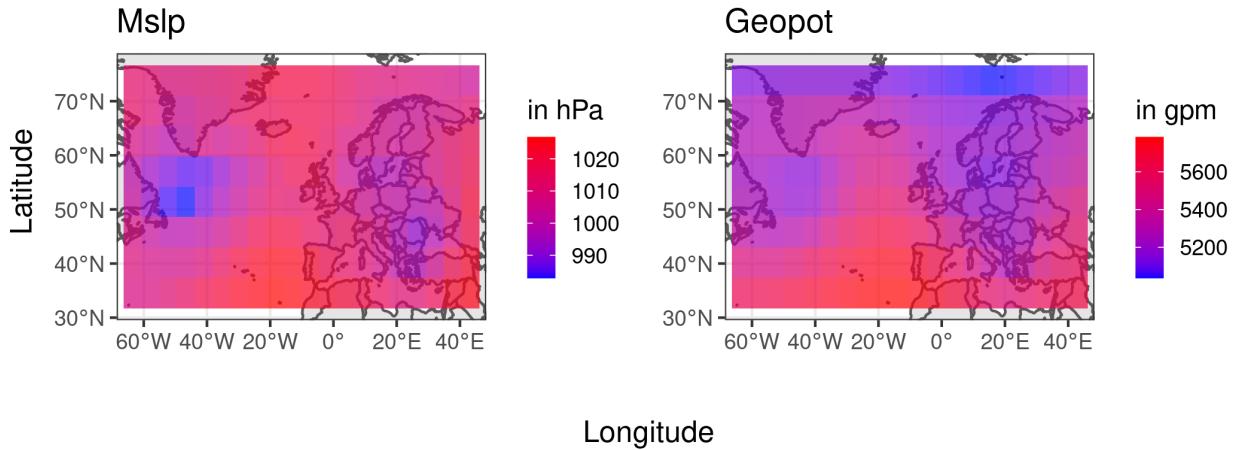
Geopot am 01-01-1980 um 0 Uhr



*Abbildung 3:* Messwerte des Geopotentials an den 160 Messpunkten am 01.01.1980 um 0 Uhr

Vor allem aus rechentechnischen Begrenzungen wurde für dieses Projekt zum einen entschieden, die Werte an den vier Uhrzeiten für jeden Tag zu mitteln. Abbildung 4 visualisiert dies für Werte des Luftdrucks und Geopotentials am 01. Januar 1980 über die 160 Standorte an einem Tag. Dadurch hat der resultierende Datensatz 320 Dimensionen, die aus 2 Parametern je für 160 Messpunkte entstehen.

Mittelwerte am 01.01.1980



**Abbildung 4:** Über den Tag gemittelte Messwerte an den 160 Messpunkten am 01.01.1980. Hierbei wurde bei den Variablen Luftdruck und Geopotential das arithmetische Mittel eines Standorts über die vier Tageszeitpunkte (0 Uhr, 6 Uhr, 12 Uhr, 18 Uhr) gebildet.

Zum anderen werden, auch auf Grund von rechentechnischen Kapazitäten, die zu betrachtenden Jahre eingeschränkt. Gemäß den Empfehlungen der WMO (Weltdorganisation für Meteorologie) benutzt man häufig zur Erfassung des Klimas einen Zeitraum von 30 Jahren (“WMO Guidelines on the Calculation of Climate Normals” (2017)). Hier wird spezifisch die Klimaperiode der Jahre 1971-2000 für Analysen betrachtet.

## 2.2 Clusterbewertungskriterien

Um den Erfolg einer Clusterlösung bewerten und somit verschiedene Clusteransätze vergleichen zu können, mussten vorerst Bewertungskriterien etabliert werden. Hierbei soll beantwortet werden, ob grundsätzlich das Clustering erfolgreich ein Muster erkennt aber zugleich beachtet werden, ob dieses Muster gemäß der Daten auch sinnvoll ist. Repräsentierend für das erste Kriterium wurde der Silhouettenkoeffizient betrachtet, für letzteres die Verteilung der Anzahl von aufeinanderfolgenden Tagen jeweils im selben Cluster, das im Folgenden Timeline genannt wird.

### 2.2.1 Silhouettenkoeffizient

Der Silhouettenkoeffizient ist eine Maßzahl für die Qualität eines Clusterings. Außerdem ist dieser unabhängig von der Anzahl der Cluster, weshalb der Silhouettenkoeffizient auch zum Festlegen der Clusteranzahl bei folgenden Analysen verwendet wird.

Der Silhouettenkoeffizient ist definiert als die Summe von Silhouetten. Gehört eine Beobachtung  $o$  zum Cluster  $A$ , so ist die Silhouette von  $o$  definiert als

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

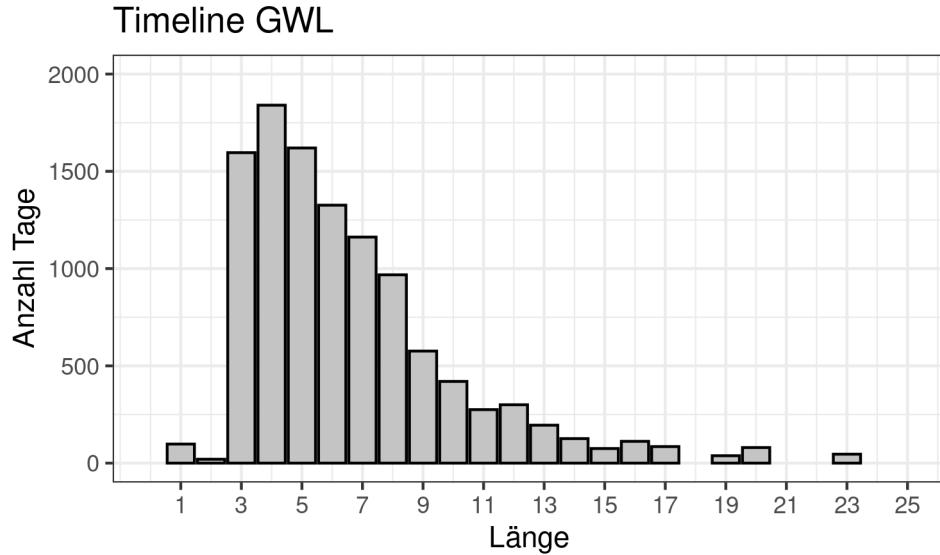
wobei  $dist(A, o)$  die durchschnittliche Distanz zu allen anderen Objekten desselben Clusters und  $dist(B, o)$  die durchschnittliche Distanz zu allen anderen Objekten des nächstgelegenen Clusters ist. Der Silhouettenkoeffizient  $s$  ist dann definiert durch die durchschnittlichen Silhouettenwerte von allen  $n$  Beobachtungen  $o$  in einem Datensatz  $D$ , also

$$s = \frac{1}{n} \sum_{o \in D} S(o), \text{ mit } -1 \leq S(o) \leq 1$$

Sowohl die Silhouetten der Beobachtungen als auch der Silhouettenkoeffizient selber können zwischen -1 und 1 liegen. Ist die Silhouette für ein Objekt  $o$  nahe der eins, so bedeutet das inhaltlich, dass die Distanz zu dem nächstgelegendem Cluster, dem  $o$  nicht zugehört, deutlich größer ist, als die Distanz zu seinem eigenen Cluster. Ist  $S(o)$  negativ, so wäre die Beobachtung eher dem anderen Cluster zuzuordnen (Hellbrück (2016)).

### 2.2.2 Timeline

Da die zeitliche Struktur der Daten, nämlich das Aufeinanderfolgen der Tage in spezifischer Reihenfolge, bei der weiteren Analyse nicht mitbeachtet wird, lässt sich hiermit die Sinnhaftigkeit einer Clustereinteilung gut bewerten. Eine Aufteilung, bei der keine zeitliche Struktur erkennbar ist, beispielsweise ein konstanter Wechsel der Clusterzugehörigkeit in aufeinanderfolgenden Tagen, ist hier nicht als sinnvoll zu bewerten. Gleichermaßen unerwünscht ist jedoch eine Aufteilung, die sehr lange Intervalle von Tagen gleicher Clusterzugehörigkeit aufweist. Die am längsten anhaltende *GWL* nach Hess und Brezowsky in der hier zu untersuchenden Zeitperiode beträgt 23 Tage. Allgemein zeigen Großwetterlagen jedoch kürzere Längen auf.



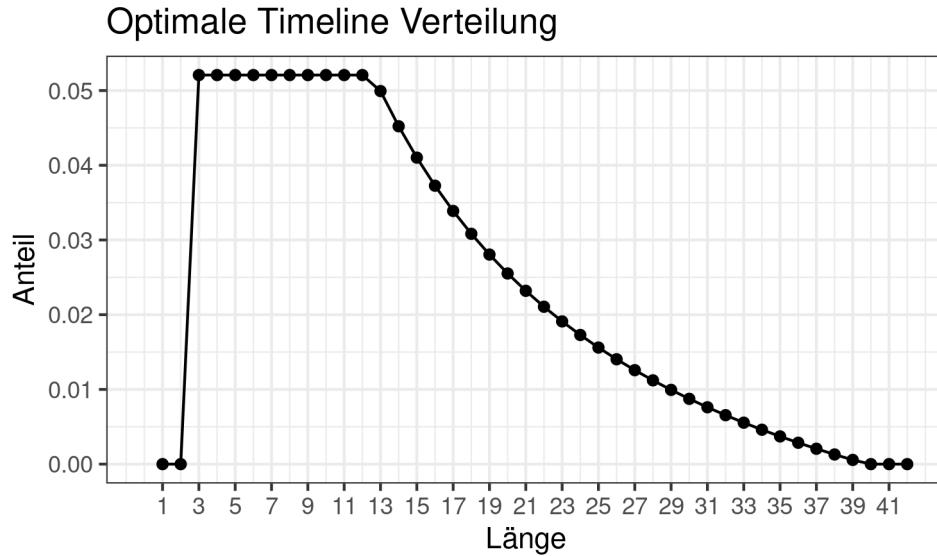
**Abbildung 5:** Timeline der *GWL* 1971-2000. Die Anzahl der *GWL* wurde mit deren Länge multipliziert, sodass jeder Tag eine Beobachtung darstellt

Bei der Darstellung in Abbildung 5 ist die Anzahl der *GWL* jeweils mit der Länge dessen multi-

pliziert, sodass jeder Tag eine Beobachtung in der Darstellung ergibt. Die *GWL*, die hier nur eine Länge von 1 oder 2 Tagen aufweisen, sind jeweils als *U* (Undefiniert/Übergang) definiert. Es ist zu erkennen, dass die meisten Tage sich in Großwetterlagen der Längen zwischen 3 und 8 Tagen befinden, längere Großwetterlagen immer seltener werden und ab einer Länge von 15 Tagen nur noch einige wenige zu beobachten sind. Anhand dieser Erkenntnisse werden nun Anforderungen an die Timeline der Clusterergebnisse gestellt.

Folglich sollen Cluster-Wetterlagen (*CWL*) auch eine Mindestlänge von 3 Tagen besitzen, also Längen von 1 oder 2 Tagen sind bestenfalls nicht zu beobachten. Gleichermaßen sollen *CWL* nicht zu lang werden. Da in der Zeitperiode 1971-2000 die maximal beobachtete Länge einer *GWL* 23 Tage beträgt, die Clusterzahl aber geringer als die Anzahl der *GWL* (29) sein soll, wird hier festgelegt, dass optimalerweise *CWL* nicht länger als 40 Tage zu beobachten sind. Darüber hinaus sollen *CWL* eher mit den Längen 3 bis 12 Tagen auftreten. Diese Optimierungsannahmen wurden hier allein in Bezug auf den Vergleich der *GWL* Timeline getroffen.

Um die Timeline einer Clusterlösung quantifizierbar und vergleichbar zu gestalten, wird der Timeline eine Verteilung unterlegt (Abbildung 6). Der Timeline-Score (*TLS*) ergibt sich dann aus der Summe der punktweisen Abweichungen der Timeline der Clusterlösung zur optimalen Timeline Verteilung.



**Abbildung 6:** Erwünschte Verteilung der Timeline einer Clusterlösung

$$TLS(X) = 1 - \sum_{i=1}^l \left| \frac{x_i}{N} - w_i \right|$$

wobei :  $X = (x_1, \dots, x_l)$  = Anzahl Tage der jeweiligen *CWL*-Länge

$N$  = Anzahl Tage gesamt

$W = (w_1, \dots, w_l)$  = Ausprägungen der optimalen Timeline Verteilung

### 2.2.3 Vergleich zur GWL-Aufteilung

Die Aufteilung der *GWL* nach Hess und Brezowsky über die Cluster wird im Folgenden nicht benutzt um eine Modellentscheidung zu treffen, da in der Analyse allgemein auf die Information der herrschenden *GWL* verzichtet werden soll. Allerdings wird hier eine Möglichkeit präsentiert, diese Aufteilung zu messen, um dies später in der Ergebnisanalyse zu nutzen. Dafür wird für jede *GWL* der größte Anteil innerhalb eines Clusters aufsummiert und durch 29 geteilt, um die Maßzahl (im Folgenden:  $HB_{diff}$ ) nach oben auf 1 zu beschränken. Die *GWL*-Kategorie  $U$  wird dabei ausgeschlossen. Somit liegt diese Maßzahl bei 1, falls alle *GWL* sich jeweils nur in einem Cluster befinden.

$$HB_{diff} = \frac{1}{29} * \sum_{i=1}^{29} \gamma$$

wobei :  $\gamma$  = größter Anteil der *GWL* innerhalb eines Clusters

$$= \frac{1}{n_i} \max_{k \in K} \left( \sum_{t=1}^{n_i} \mathbf{I}_k(Tag_t) \right)$$

mit :  $K$  = Menge der Cluster;  $n_i$  = Anzahl Tage in der *GWL*  $i$

## 2.3 Filter Ansatz

### 2.3.1 Motivation

Beim genaueren Blick in die Beschreibungen einzelner *GWL* ist zu erkennen, dass diese häufig durch Position oder Form bestimmter Gebiete mit erkennbar höheren oder tieferen Messwerten definiert sind.

Beispielsweise wird die *GWL* Trog Westeuropa (TRW) definiert durch ein sich vertikal erstreckendes Tiefdruckgebiet von Skandinavien bis zur Iberischen Halbinsel, flankiert von Hochdruckgebieten über dem Atlantik und Westrussland. Hingegen die *GWL* Hoch Britische Inseln (HB) ist, wie der Name bereits vermuten lässt, beschrieben durch ein Hochdruckgebiet über dem Vereinigten Königreich und Irland, umgeben von mehreren Tiefdruckgebieten Wetterstation Peißenberg (2021). Ähnlich ist dies bei allen weiteren Großwetterlagen zu beobachten.

Demnach lässt sich vermuten, dass die *GWL* sich anhand der Position und Form, der an dem Tag respektiven Hoch- und Tief(druck)gebiete, sinnvoll gruppieren ließen.

### 2.3.2 Prinzip Filtern

Diesen Grundgedanken verfolgend, sind die Tage optimalerweise in interessierende Gebiete zu unterteilen und anhand der Positionen und Formen dieser Gebiete miteinander zu vergleichen. „Interessierende Gebiete“ wurden hierbei vorerst angenommen als die Gebiete um die täglich gemessenen Extrema. Folglich also einem Gebiet höherer Messwerte um den am Tag maximal gemessenen Wert und einem Gebiet tieferer Messwerte um den minimalen Messwert. Dabei ist zu beachten, dass diese Gebiete nicht zu groß werden aber auch nicht nur aus einzelnen wenigen Punkten bestehen. Außer-

dem sollte ihre Form nicht durch den Clusteralgorithmus bestimmt bzw. beeinflusst werden, da die Gebiete sonst alle ähnliche Formen aufweisen würden, und nicht treu repräsentiert werden würden. Beides führte später zu Problemen bei dem Vergleich zwischen den Tagen. Alle Standorte die nicht in den interessanten Gebieten liegen sollten bei dem Vergleich der Tage nicht mit einbezogen werden und demnach bei der Gebietseinteilung als Rauschen bezeichnet werden.

Um diese Gebietseinteilung eines Tages durchzuführen, soll also ein metrischer Messwert eines bestimmten Standortes mit einer Gebietszugehörigkeit ersetzt werden. Die Messinformationen des Tagen sollen somit “gefiltert” werden. Da sonst bestimmte Hyperparameter oder Grenzwerte fest angegeben werden müssten, lässt sich dies durch ein separates Clusterverfahren über die 160 Standorte pro Tag erreichen. Mit der Idee, anhand beider Parameter-Messwerte (Mslp und Geopotential) “Gebiete-Muster” zu erkennen wird jeder Tag zwei mal auf diese Weise geclustered; jeweils pro Parameter ein Mal. Die Feature-Variablen dieses Clusterverfahrens sind demnach Longitude, Latitude und der Parametermesswert.

### 2.3.3 DBSCAN und Fuzzy

Um der Anforderung der nicht uniformen Gebiete gerecht zu werden, erscheint ein dichtebasierter Clusterverfahren von Vorteil. Ein implementiertes Verfahren, das auch die Möglichkeit des Rauschens beinhaltet, ist *DBSCAN* (Density-Based Spatial Clustering of Applications with Noise). Ein solches Verfahren nutzt einen sogenannte Nachbarschaftsparameter, um Beobachtungen zueinander als Nachbarn einzuteilen oder eben nicht. Demnach liegt hier nicht ein Mittelpunkt eines Clusters vor, zu dem der Abstand einer jeweiligen Beobachtung entscheidend dafür ist, ob sie sich jeweils in diesem Cluster befindet, sondern Beobachtungen innerhalb eines Clusters weisen jeweils geringe Distanzen zu ihren Nachbarpunkten im selben Cluster auf. Dadurch werden Cluster erstellt, die sehr unförmig sein können. (Yildirim (2020))

*DBSCAN* benötigt keine Angabe der Clusteranzahl, sondern nur der Hyperparameter *minPoints* (minimale Anzahl an Punkten pro Cluster) und *eps* (Nachbarschaftsparameter). Hier lässt sich erhoffen, dass eventuell auch mehr als nur die zwei Gebiete um die Extrema erkannt werden. Allerdings musste schnell erkannt werden, dass sich der Algorithmus mit diesen Daten sehr sensitiv gegenüber den Hyperparametern präsentierte, selbst wenn der Nachbarschaftsparameter pro Tag anhand dem Wert der größten Wölbung eines *kNN*-plots ( $k = \text{minPoints}$ ) spezifisch berechnet wird (siehe Absatz 2.3.4). Dies führt dazu, dass sehr viele Tage nur zu Noise, einem einzigen Cluster oder zu riesigen Clustern gefiltert werden; was widerrum nicht erwünscht ist, da die Vergleichbarkeit der Tage im Nachhinein damit nahezu unmöglich wird.

Ein Bestimmen der Startpunkte und somit ein Festsetzen der Cluster-Orte ist - zumindest in vorhandenen Implementationen dieses Algorithmus' - nicht möglich. Deswegen lässt sich beobachten, dass sich die Extrema oft nicht in einem der eingeteilten Clustern befinden, da sie oft stark von durchschnittlichen Messwerten abweichen. *DBSCAN* führt somit im besten Fall zu Gebieten, dessen Messpunkte sich eher im Mittelfeld der Skala des Tages befinden und innerhalb der Gebiete kaum Veränderungen aufweisen. Solche Gebiete sind aber hier als nicht von besonderem Interesse vermutet

und deshalb wird *DBSCAN* hierfür ausgeschlossen.

Ein weiterer vielversprechender Ansatz ist das *Fuzzy-Clustering* (Cebeci (2017)). *Fuzzy* entspricht vom Prinzip k-Means, es werden also pro Cluster Mittelpunkte festgelegt. Allerdings wird für jede Beobachtung eine Zugehörigkeitswahrscheinlichkeit für jedes Cluster berechnet, anstatt einer Cluster-Id. Hier können Startpunkte angegeben werden und anhand der Clusterzugehörigkeitswahrscheinlichkeit jeder Beobachtung, bestimmte Beobachtungen mithilfe eines Schwellenwertes im Nachhinein zu Rauschen verwandelt werden.

Allerdings ist dieses Verfahren hier auch nicht optimal. Neben dem, dass *Fuzzy* üblicherweise rechen-technisch sehr teuer implementiert ist, kommt hinzu, dass dieses Verfahren auf einen Mittelpunkt pro Cluster beruht und die Distanz dazu anhand eines gegebenen Distanzmaß' bestimmt wird. Nachdem die Koordinaten als Variablen aufgenommen werden, führt dies zu Clustern gleicher Form und verletzt somit die Anforderung, die Form eines Gebietes möglichst getreu darzustellen.

### 2.3.4 SCAPOI

Im Folgenden wird der benutzte Algorithmus beschrieben, der eine abgeänderte Version des *DBSCAN* darstellt. Dieser beinhaltet fixe Startpunkte und ein iterierend strenger werdendes Nachbarschaftskriterium. Er wird im folgenden immer als *SCAPOI* (Spatial Clustering around Points of Interest) benannt.

Die Startpunkte werden hier jeweils als die Extrema der Messpunkte gewählt. Der Nachbarschaftsparameter *eps* muss groß genug gewählt (bzw. berechnet) werden, um zu berücksichtigen, dass die Extrema im Vergleich zu anderen Messwerten stark abweichen und somit zu verhindern, dass die Cluster gar nicht oder zu gering wachsen. Allerdings führt aber ein zu groß gewählter *eps* dann zu einem zu starken Wachsen der Cluster, da der Abstand eines Messpunktes zu seinem Nachbar selten größer ist, als der Abstand der Extrema zu seinen Nachbarn. Deshalb wird dieser Nachbarschaftsparameter *eps* pro Iteration verkleinert.

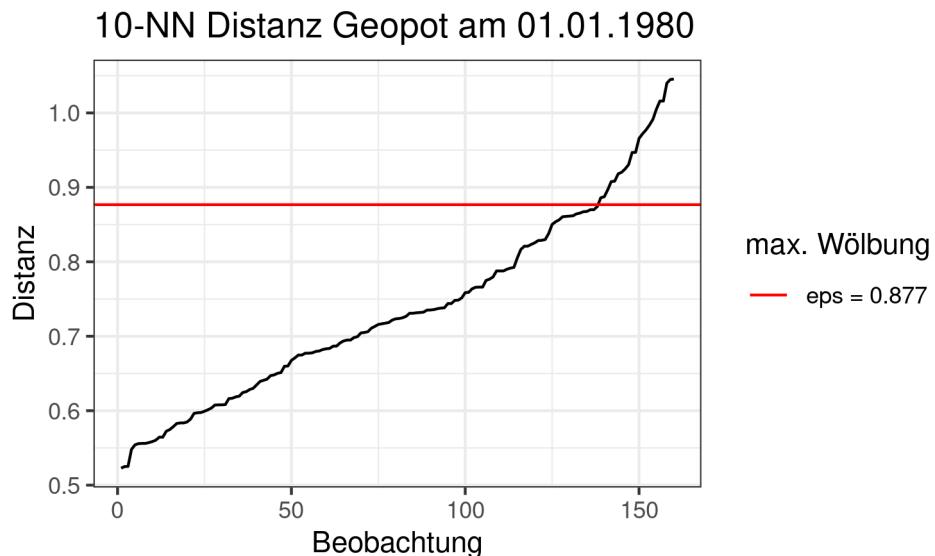
Der Nachbarschaftsparameter wird hier für jede Tag-Parameter Kombination berechnet, indem der Punkt der größten Wölbung eines kNN-Distanzplot berechnet wird. Dieser visualisiert die Distanzen von Beobachtungen zu seinen *k* Nachbarn. *k* wird hier als 10 festgesetzt, mit dem Ziel, am Ende Gebiete der Mindestgröße 10 zu erhalten.

**Data:** Messwerte eines Parameters der 160 Standorte am Tag

**Result:** Gebietszugehörigkeitsvektor pro Tag

```
1 ;
2 eps0 = Berechnete Distanz des Punktes der größten Krümmung in Bezug auf kNN;
3 Startpunkte = Orte des gemessenen Minimums und Maximums;
4 for jeden Startpunkt do
5   eps = eps0;
6   beginne ein Cluster um den Startpunkt;
7   while Neue Punkte gefunden werden, die hinzugefügt werden do
8     Prüfe ob es Punkte gibt, die < eps von einem im Cluster existierenden Punkt entfernt
      sind;
9     if ein Punkt bereits einem anderen Cluster angehört then
10       Füge es dem Cluster hinzu, dessen Startpunkt es am nächsten liegt;
11     else
12       füge es dem Cluster hinzu;
13     end
14     eps = reduzierter eps;
15   end
16 end
17 nicht zugeteilte Punkte bleiben Noise;
```

**Algorithm 1:** SCAPOI - Algorithmus



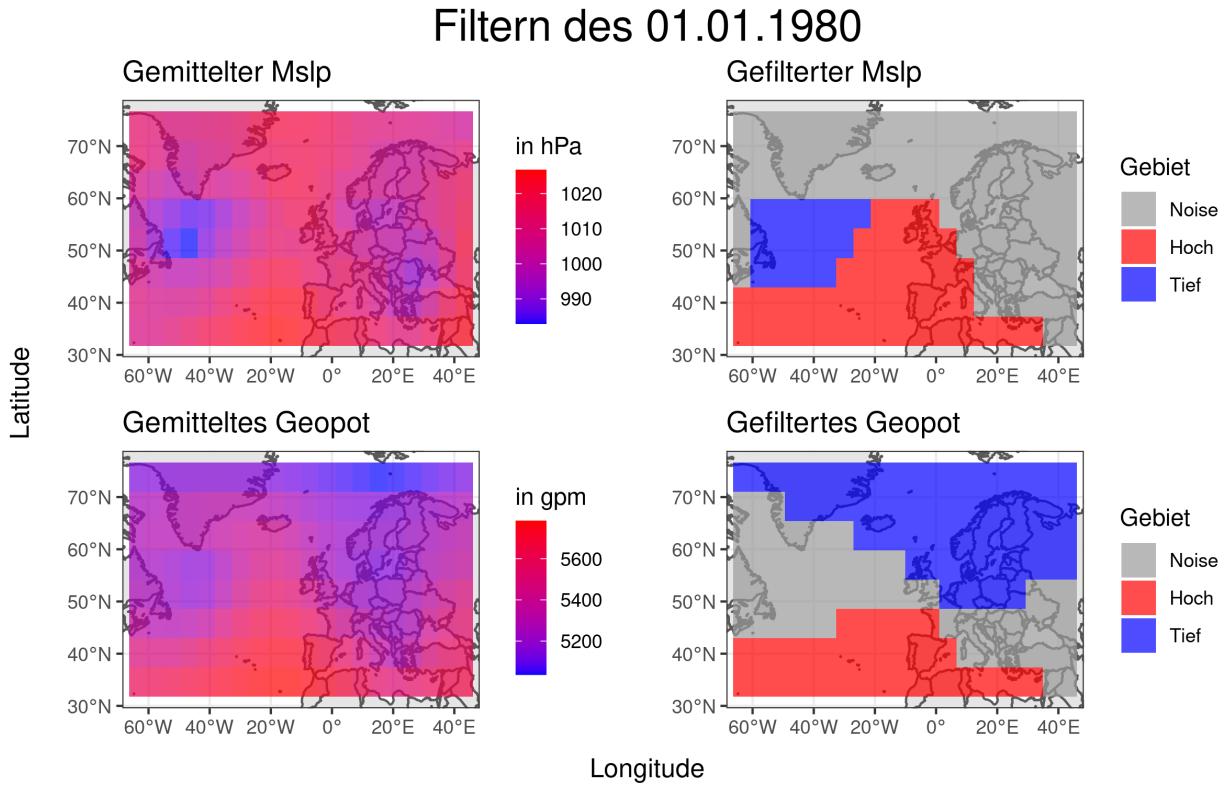
**Abbildung 7:** 10-NN Distanz mit Punkt der maximalen Wölbung

Eine Verkleinerung des Nachbarschaftsparameters pro Iteration findet hier in linearer Form statt, da dies, rein visuell bewertet, die repräsentativsten Gebiete als Ergebnis liefert.

$$\text{eps}_{t+1} = \text{eps}_t - \frac{(t-1)\text{eps}_t}{6}$$

wobei :  $t$  = Iterationsindex

Im Vergleich zu *DBSCAN*, mit dem Gruppierungen von Beobachtungen gesucht werden, die einer bestimmten Mindestdichte, sowie einer Mindestgröße gerecht werden, wird mit *SCAPOI* versucht um bestimmte, definierte Beobachtungen Gruppen zu bilden, die eine gewisse Dichte aufweisen. Im Fall hier (2 Cluster um Minimum und Maximum respektive), erhält man nun einen Gebietszugehörigkeitsvektor mit den Klassen: Noise, High und Low.



*Abbildung 8: Beispiel der Gebietszuteilung eines Tages durch SCAPOI*

### 2.3.5 Distanzmetrik

Um nun wieder auf Tagesebene clustern zu können, sprich mit den Tagen als Beobachtungseinheit Cluster zu bilden, benötigt man eine Metrik, mit der diese Gebietszugehörigkeitsvektoren zweier Tage miteinander verglichen werden können. Der *Rand-Index* präsentiert eine solche Möglichkeit (Tang (2017)). Dieser vergleicht jeweils, ob pro Clusterlösung Paare zweier Beobachtungen jeweils im selben Cluster liegen. Obwohl dies eher gedacht ist, um Lösungen verschiedener Clusterverfahren mit denselben Beobachtungen zu vergleichen, ist der *Rand-Index* hier möglich, da die Messpunkte konstant sind. Allerdings ist dabei irrelevant, in welchem Cluster sich das Paar jeweils befindet, was in diesem Fall nicht erwünscht ist: Zwei Tage mit identischen Gebietsformen und -orten

aber gespiegelter Zugehörigkeit sollen nicht eine Distanz von 0 zueinander aufweisen. Zudem sind Messpunkte, die als Noise definiert wurden, nicht von Interesse und sollten demnach auch nicht mit einbezogen werden.

Deshalb wurde folgend eine Distanzmetrik definiert, die über zwei Tage alle Messpunkte, die jeweils als Noise definiert wurden, nicht betrachtet und bei den Verbleibenden vergleicht, welche Gebietszugehörigkeit die Messpunkte jeweils aufweisen. Danach wird noch durch die größte Anzahl an Messpunkten nicht in Noise dieser zwei Tagen geteilt, um den Wertebereich auf [0,1] zu beschränken sowie einer möglichen Ausprägung der Gebiete als Teilmengen voneinander ebenfalls nicht eine Distanz von 0 zuzuweisen.

$$d(A, B) = 1 - \left( \frac{\sum_{i=1}^{160} \mathbf{I}(a_i = b_i \neq 0)}{\max_{x \in A, B} (\sum_{i=1}^{160} \mathbf{I}(x_i \neq 0))} \right)$$

wobei :  $A = (a_1, \dots, a_{160})$ ;  $B = (b_1, \dots, b_{160})$

und :  $\forall_{x \in A, B} : x \in \{0(\text{Noise}), 1(\text{Hoch}), 2(\text{Tief})\}$

### 2.3.6 Kriterien Filtern

Dieser Ansatz liefert für beide Parameter kombiniert einen Silhouettenkoeffizient von 0.0832 und einen *TLS* von -0.1527. Dabei wurden die Distanzmatrizen des Filterns der beiden Parameter addiert und damit eine Clusteranalyse wie in Abschnitt 2.4.2.1 (*PAM*) beschrieben durchgeführt.  $HB_{diff}$  ergibt sich hier zu 0.4100.

Nach Parametern getrennte Analysen weisen je einen höheren Silhouettenkoeffizient auf, aber einen geringeren *TLS*. (Mslp:  $s = 0.1369$ ,  $TLS = -0.2027$ ,  $HB_{diff} = 0.3980$  ; Geopot:  $s = 0.1541$ ,  $TLS = -0.3898$ ,  $HB_{diff} = 0.3162$ )

Clusteranzahl ist dabei 5 für die kombinierte Analyse, sowie Mslp alleine, Geopotential alleine ergibt 6 Cluster.

## 2.4 Clustern mit extrahierten Daten

Ein weiterer Ansatz ist, dass Informationen aus dem Reanalyse Datensatz extrahiert werden und diese dann Variablen eines neuen Datensatzes werden. Die Grundidee beruht auf der naiven Annahme, dass Großwetterlagen zum einen über bestimmte Messwerte am Tag und zum anderen über die Lage dieser Werte charakterisiert werden. Die *GWL Tief Mitteleuropa* (TM) zeichnet sich beispielsweise durch ein Tief über Mitteleuropa aus (Tiefgraber (2013)). Daher sind zwei Kategorien von Interesse, die Verteilung der Messwerte der Parameter, Mslp und Geopotential, und deren räumliche Lage.

Es wird von dieser Methodik erhofft, dass die Dimensionen weiter reduziert werden können und dass wichtige Größen spezifisch gewichtet werden können.

### 2.4.1 Extrahieren der Variablen

Die Ausgangslage beim Extrahieren der Variablen ist dabei größtenteils der Datensatz mit 320 Dimensionen, also der, bei dem die vier Messzeitpunkte für jeden Tag gemittelt wurden. Davon werden verschiedene Größen extrahiert, die jeweils eine interessante Variable über alle Standorte zusammengefasst verkörpert, wie zum Beispiel der Mittelwert des Luftdrucks über alle Standorte pro Tag. Dieser ist damit unabhängig von den Standorten und gehört zu der Kategorie "Verteilung der Parameter" am Tag. Weitere Variablen dieser Kategorie sind das Minimum und Maximum, der Median, die 0.25- und 0.75-Quantile, die Intensität und die Veränderung über den Tag jeweils für beide Parameter Luftdruck und Geopotential.

Für das Minimum, Maximum, Median, Mittelwert und die beiden Quartile wird je ein Tag mit den 160 Standorten betrachtet, wovon diese Variablen für den Luftdruck sowie für das Geopotential extrahiert werden.

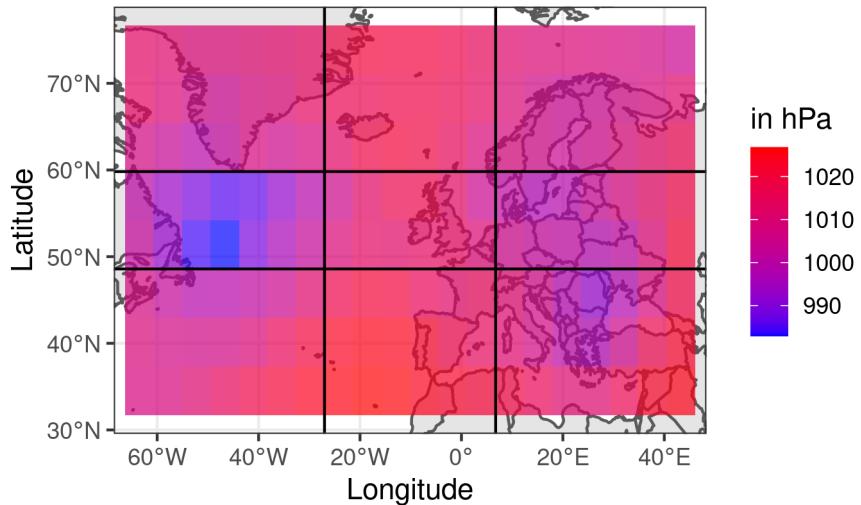
Die Intensität wird in "Intensität Hoch" und "Intensität Tief" aufgeteilt und ist die Anzahl der Messwerte am Tag, die unter bzw. über dem 0.25- bzw. 0.75-Quantil, über alle Tage zusammen betrachtet, liegen. Sind beispielsweise an einem Tag 10 Messwerte des Geopotentials unterhalb des 0.25-Quantils über alle Tage betrachtet, so ist die Variable "Intensität Tief Geopotential" für diesen Tag 10. Die Intention dahinter ist, zum einen, die Größe von Hoch- und Tiefgebiete am Tag zu bestimmen. Hochgebiete sind hier einfachhalber durch hohen Luftdruck und hohes Geopotential definiert, wobei diese Parameter getrennt voneinander betrachtet werden und analog für ein Tiefgebiet. Das bedeutet, insgesamt gibt es 4 Variablen, die die Intensität beschreiben - Intensität Hoch und Intensität Tief je für Mslp und Geopotential.

Zum anderen kann die Größe und Intensität der Gebiete über alle Tage verglichen werden, da sie in Bezug auf die Quartile über alle Tage gebildet werden. So kann es zum Beispiel sein, dass an einem Tag die Intensität des Luftdrucks für ein Hochgebiet 0 ist, da an diesem Tag generell niedrige Mslp Werte beobachtet wurden.

In Abschnitt 2.1 wurde bereits beschrieben, dass der Mittelwert über vier Messzeitpunkte pro Tag gebildet wurde. Da dies mit einem Informationsverlust einhergeht, wird die Variable "Veränderung über den Tag" eingeführt. Sie ist definiert als die summierten, absoluten Differenzen des maximalen und minimalen Messwertes für jeden Standort am Tag. Diese Variable wird folglich mit Hilfe des Originaldatensatzes, ohne Informationsverlust, für beide Parameter Mslp und Luftdruck extrahiert.

Da bereits Variablen extrahiert wurden, die die Verteilung der Parameter an verschiedenen Tagen beschreiben, ist des Weiteren noch die räumliche Ebene von Interesse. Dafür wird das 8x20 Grid in 9 Quadranten unterteilt, also in Nord - Süd, Ost - West und jeweils die Mitte bzw. das Zentrum, wie man in Abbildung 9 sehen kann.

## Gemittelter Mslp am 01.01.1980



**Abbildung 9:** Aufteilung der 160 Standorte in 9 Quadranten

Es wird für jeden Tag angegeben, in welchem der 9 Quadranten sich die Extremwerte, also Minimum sowie Maximum für je Luftdruck und Geopotential, befinden. Ursprünglich waren diese Variablen kategorial, da die Quadranten von eins bis neun durch nummeriert wurden, z.B.

$Quadrant_{maxMslp,i} \in \{1, 2, \dots, 9\}$  mit  $i = 1, \dots, 10958$  (Anzahl der Tage). Allerdings ist der Datensatz dadurch sowohl mit numerischen, als auch mit kategorialen Variablen und die Möglichkeiten zu clustern sind damit eingeschränkt. Deshalb wurden zwei “Dummy-Variablen” eingeführt, sodass die Lage auch numerisch angegeben werden kann. Diese Variablen sind Spalte und Zeile für die vier Extremwerte am Tag, beispielsweise  $Zeile_{maxMslp,i} \in \{1, 2, 3\}$  und  $Spalte_{maxMslp,i} \in \{1, 2, 3\}$  mit  $i = 1, \dots, 10958$ .

Zudem werden die Distanzen zwischen Extrempunkten mit einbezogen. Zum einen die Distanzen zwischen dem Maximum und Minimum für je Geopotential und Luftdruck. Zum anderen die Distanzen vom Minimum bzw. Maximum des Geopotential zu den jeweiligen Extremwerten des Luftdrucks für jeden Tag. Alle Distanzen werden mit der euklidischen Distanz gebildet, wobei die Longituden und Latituden der Extremwerte am Tag zur Berechnung betrachtet werden.

Zuletzt werden bei der räumlichen Ebene für beide Parameter die Mittelwerte in allen 9 Quadranten angegeben. Das sind somit 18 weitere Variablen.

Insgesamt umfasst der neu extrahierte Datensatz 48 Variablen. Die Variablen wurden in Absprache mit dem Projektpartner definiert und in Tabelle 2 sind die extrahierten Variablen für je das Geopotential sowie den Luftdruck zusammengefasst.

**Tabelle 2:** Extrahierte Variablen

Variable	Definition
<b>Verteilungsvariablen</b>	
Minimum	Minimaler Wert pro Tag
Maximum	Maximaler Wert pro Tag
Mittelwert	Mittelwert pro Tag
Median	Median pro Tag
Quartile	Quartile pro Tag
Intensität Hoch	Anzahl der Messpunkte am Tag, die über alle Daten über dem 0.75-Quantil liegen
Intensität Tief	Anzahl der Messpunkte am Tag, die über alle Daten unter dem 0.25-Quantil liegen
Veränderung über den Tag	Summierte Differenzen von vier Messzeitpunkten am Tag für alle Standorte
<b>Räumliche Ebene</b>	
Spalte Minimum	Spalte $x_{min}$ , in dem sich Minimum befindet; $x_{min} = 1, 2, 3$
Zeile Minimum	Zeile $y_{min}$ , in dem sich Minimum befindet; $y_{min} = 1, 2, 3$
Spalte Maximum	Spalte $x_{max}$ , in dem sich Maximum befindet; $x_{max} = 1, 2, 3$
Zeile Maximum	Spalte $y_{max}$ , in dem sich Maximum befindet; $y_{max} = 1, 2, 3$
Distanz zwischen Extrema	Euklidische Distanz zwischen Minimum und Maximum
Distanz der beiden Minima	Euklidische Distanz vom Minimum Geopotential zu Minimum Mslp
Distanz der beiden Maxima	Euklidische Distanz vom Maximum Geopotential zu Maximum Mslp
Mittelwerte in den Quadranten	Mittelwerte in jeweils 9 Quadranten

#### 2.4.2 Clusterverfahren

Hier ist die Wahl auf den Clusteralgorithmus *PAM* (Partitioning Around Medoids) gefallen, der 1990 von Kaufman und Rousseeuw eingeführt wurde. Dabei werden die Beobachtungen in  $k$  disjunkte Partitionen aufgeteilt. Dieser wird auch “k-Medoid” Algorithmus genannt, weil die Beobachtung, die am zentralsten innerhalb eines Clusters liegt, das Zentrum dieses Clusters ist und somit repräsentativ für andere Beobachtungen desselben Clusters ist. Da hier eine reale Beobachtung repräsentativ für ein Cluster ist, ist nach Q. Zhang (2005) dieses Verfahren robuster als beispielsweise ein k-Means Algorithmus, bei dem der Mittelwert von allen Punkten eines Cluster repräsentativ für dieses ist (B. Everitt (2011)). Ein weiterer Grund für PAM ist, dass der implementierte Algorithmus `pam()` in R aus dem Package “cluster” sehr vielfältig ist. Er akzeptiert als Input sowohl direkt eine Distanzmatrix als auch einen Dataframe mit Daten.

**2.4.2.1 Partitioning Around Medoids** Bei *PAM* wird angegeben, wie viele Cluster  $k$  gebildet werden sollen. Der Algorithmus sucht dann anfänglich  $k$  repräsentative Beobachtungen, die das Zentrum der entstandenen Cluster darstellen, um danach iterativ bessere Repräsentanten zu finden. Alle möglichen Kombinationen von repräsentativen und nicht-repräsentativen Beobachtungen werden analysiert und die Qualität der jeweiligen Clusterings wird anhand eines Gütekriteriums evaluiert. Dieses Gütekriterium ist hier die Summe  $S$  der Distanzen von allen Datenpunkten zu deren jeweiligen Medoids. Anhand dieses Gütekriteriums werden die nicht-repräsentativen Datenpunkte dem Cluster zugefügt, das die Summe  $S$  am stärksten minimiert. Dann wird zufällig ein weiterer Datenpunkt gewählt und es werden die Kosten  $C$  berechnet, die entstehen, wenn man diesen mit den repräsentativen Punkten tauschen würde. Dabei sind die Kosten definiert als die Veränderung des Gütekriteriums  $S$ , wenn ein repräsentativer Punkt mit einem anderen Datenpunkt getauscht wird. Das bedeutet, die Kosten können sowohl negativ als auch positiv sein. Sind sie negativ, so wird die Summe  $S$  der Distanzen von allen Datenpunkten zu deren Medoids kleiner und die Punkte werden vertauscht (X. Jin (2017)).

Der Algorithmus wird hier nach X. Jin (2017) als Pseudocode dargestellt.

**Data:** Datensatz  $D$  mit  $n$  Beobachtungen  $o$

**Result:** Set mit  $k$  Clustern

- 1 Suche zufällig  $k$  Punkte aus  $D$  als erste repräsentative Objekte;
- 2 **while** Stop-Kriterium nicht erreicht wurde **do**
- 3     **for** jeden weiteren Datenpunkt  $o$  in  $D$  **do**
- 4         Finde den am nächsten liegenden repräsentativen Datenpunkt und teile  $o$  diesem Cluster zu
- 5     **end**
- 6     Wähle zufällig einen nicht-repräsentativen Punkt  $o_{rand}$ ;
- 7     Berechne die Kosten  $C$  für einen möglichen Tausch von  $o_{rand}$  und einem repräsentativen Datenpunkt  $o_i$ ;
- 8     **if**  $C < 0$  **then**
- 9         Tausche  $o_i$  mit  $o_{rand}$  und forme ein neues Set mit  $k$  repräsentativen Punkten
- 10     **end**
- 11 **end**
- 12 Gebe das Cluster Ergebnis aus;

**Algorithm 2:** PAM Pseudocode

**2.4.2.2 Skalierung** Die Variablen haben verschiedene Skalen. Werte des Luftdrucks liegen beispielsweise immer zwischen ungefähr 935 und 1065 hPa und die des Geopotentials ca. zwischen 4530 und 5870 gpm. Außerdem gibt es auch Variablen, wie zum Beispiel die Zeilen oder Spalten die nur Werte von eins, zwei oder drei aufweisen. Der Abstand von verschiedenen Variablen kann somit sehr groß sein und da zum Berechnen einer Distanzmatrix diese Werte betrachtet werden, muss

der Datensatz vor dem Clustering skaliert werden. Die Skalierung erfolgt hier durch eine Standardisierung, d.h.  $x_{p,neu} = \frac{x_p - \mu_p}{\sigma_p}$  mit  $p = 1, \dots, 48$ . Dabei wird für jede Beobachtung einer Variable  $p$  der Erwartungswert dieser Variablen abgezogen und schließlich durch die Standardabweichung der Variable dividiert. Dies wird für alle im Datensatz enthaltenen durchgeführt.

**2.4.2.3 Gewichtung** Des Weiteren soll der extrahierte Datensatz gewichtet werden. Das bedeutet, es wird fachlich entschieden, wie wichtig bestimmte Variablen sind. Da die Gewichtung Einfluss auf das Bilden einer Distanzmatrix hat, sollen Variablen, die von besonderem Interesse sind, höher gewichtet werden, sodass dies bei der Distanzmetrik berücksichtigt werden kann.

Dafür werden die Variablen zuerst in verschiedene Kategorien aufgeteilt. Die Kategorien sollen alle insgesamt das gleiche Gewicht haben. Hier wurden 8 Kategorien gewählt: Die Verteilungsvariablen Minimum, Maximum und Mittelwert ergeben je eine Kategorie für das Geopotential und den Luftdruck. Die restlichen Verteilungsvariablen, Median, Quartile, Intensität und Veränderung über den Tag sind 2 weitere Kategorien. Die fünfte und sechste Kategorie setzen sich aus den räumlichen Variablen Zeile, Spalte, Distanzen zwischen Extrema zusammen und die Mittelwerte in den Quadranten bilden die letzten zwei. Die Gewichte der Kategorien summieren sich jeweils auf eins und die Variablen werden innerhalb einer Kategorie gleich gewichtet.

In Tabelle 3 sieht man die Kategorien mit ihren Variablen und Gewichten.

**Tabelle 3:** Gewichte für Variablen

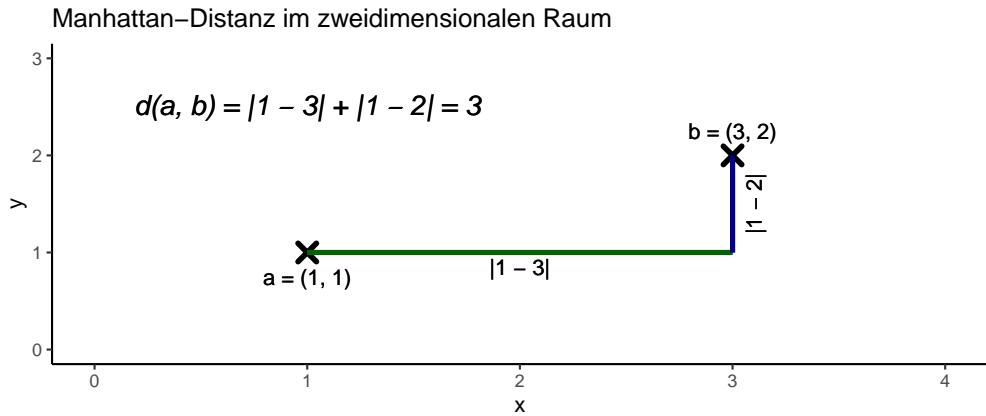
Variable	Gewichte
<b>1 Verteilungsvariablen Mslp I</b> Minimum, Maximum, Mittelwert	$\frac{1}{3}$
<b>2 Verteilungsvariablen Geopotential I</b> Minimum, Maximum, Mittelwert	$\frac{1}{3}$
<b>3 Verteilungsvariablen Mslp II</b> Median, Quartile, Intensität Hoch, Intensität Tief, Veränderung über den Tag	$\frac{1}{6}$
<b>4 Verteilungsvariablen Geopotential II</b> Median, Quartile, Intensität Hoch, Intensität Tief, Veränderung über den Tag	$\frac{1}{6}$
<b>5 Räumliche Ebene für Mslp</b> Spalte und Zeile für Minimum/Maximum, Distanz zwischen Extrema, Distanz der beiden Minima/Maxima	$\frac{1}{6}$
<b>6 Räumliche Ebene für Geopotential</b> Spalte und Zeile für Minimum/Maximum, Distanz zwischen Extrema, Distanz der beiden Minima/Maxima	$\frac{1}{6}$
<b>7 Mittelwerte in den Quadranten für Mslp</b> Mittelwerte in den 9 Quadranten	$\frac{1}{9}$
<b>8 Mittelwerte in den Quadranten für Geopotential</b> Mittelwerte in den 9 Quadranten	$\frac{1}{9}$

**2.4.2.4 Distanzmetrik** Die Wahl für diese Clusteranalyse fiel auf die Manhattan-Distanz, die für zwei Beobachtungen  $a$  und  $b$  definiert ist als

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

wobei  $a = (a_1, \dots, a_p)$  und  $b = (b_1, \dots, b_p)$  mit  $p = 48$  (Anzahl der Variablen)

Die Manhattan Distanz wird in Abbildung 10 für einen zweidimensionalen Fall, also  $p = 2$ , beispielhaft veranschaulicht.

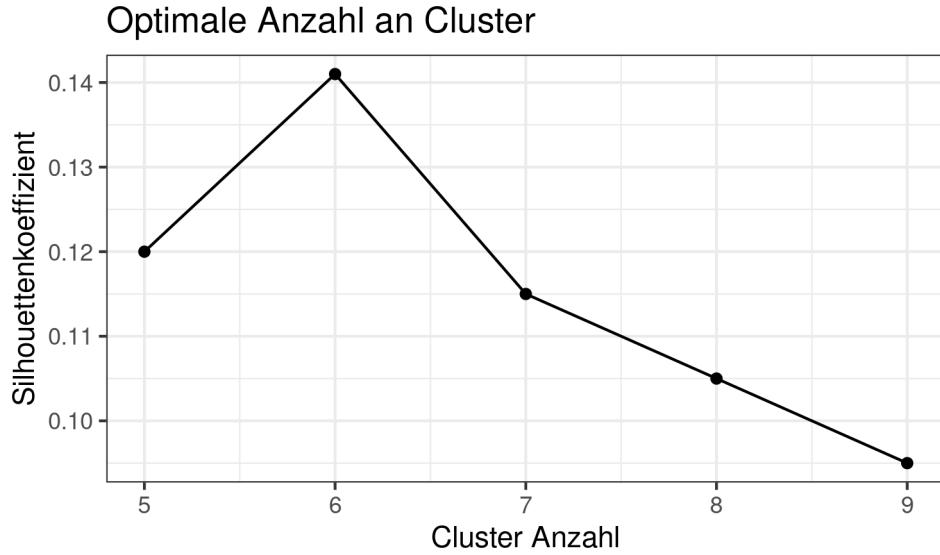


**Abbildung 10:** Beispiel für die Manhattan-Distanz im zweidimensionalen Raum

Weitere Distanzmetriken, die ausprobiert wurden, sind unter anderem die Euklidische Distanz und die Gower Distanz, die aber beide zu schlechteren Ergebnissen, gemessen anhand der definierten Bewertungskriterien, geführt haben. In Abschnitt 2.5 werden diese weiteren Versuche grob dargestellt.

**2.4.2.5 Wahl der Clusteranzahl** Die Wahl der Clusteranzahl  $k$  muss vor dem Ausführen des Algorithmus statt finden. Aber um die optimale Anzahl an Clustern zu finden, wird der Silhouettenkoeffizient für verschiedene  $k$  betrachtet. Der Silhouettenkoeffizient ist wie in Abschnitt 2.2.1 beschrieben, unabhängig von der Clusteranzahl. Daher kann er für die Wahl der Clusteranzahl herangezogen werden. In Abbildung 11 sieht man die Silhouettenkoeffizienten für  $k = \{5, 6, 7, 8, 9\}$ . Den größten Wert bekommt man mit einer Clusteranzahl  $k = 6$ . Somit wird das resultierende Clusterergebnis mit dem Algorithmus *PAM*, der Distanzmetrik Manhattan und 6 Clustern gebildet.

Dieses Clusterergebnis liefert einen Silhouettenkoeffizienten  $s = 0.1411$  und einen Timeline-Score von 0.3357. Die Güte der Verteilung der *GWL* zu den Clustern  $HB_{diff}$  beträgt 0.3309.



**Abbildung 11:** Optimale Anzahl an Cluster. Die Clusteranzahl mit dem maximalen Silhouettenkoeffizient ist die optimale Clusteranzahl, hier  $k = 6$ .

## 2.5 Weitere Versuche

Das endgültig gewählte Methode wurde mit der Distanzmetrik Manhattan und dem Clusteralgorithmus *PAM* anhand des extrahierten Datensatzes erzielt. Es wurden anfangs aber verschiedene Methodiken ausprobier und anhand der definierten Bewertungskriterien beurteilt. Im Folgenden werden weitere Versuche kurz dargestellt und zusammengefasst.

### 2.5.1 CLARA, K-Means und PAM mit Originaldaten

Bevor die Variablen extrahiert wurden, gab es auch Clusteranalysen mit dem Reanalyse-Datensatz, bei dem für beide Parameter die vier Beobachtungen am Tag jeweils gemittelt wurden, also der Datensatz mit 320 Dimensionen für die Jahre 2006 bis 2010.

#### CLARA

Zum einen wurde ein Clustering mit dem Algorithmus Clustering Large Applications (*CLARA*) durchgeführt. Dieser basiert auf dem k-Medoid-Ansatz, ist aber speziell für Datensätze, die sehr viele Beobachtungen enthalten, z.B. mehrere tausend Objekte. Dieser Algorithmus wird von L. Kaufman (2005) eingeführt und erklärt.

Dieser Algorithmus wurde mit der Manhattan-Distanz, als auch mit der Euklidischen Distanz ausprobiert. Für eine Clusteranzahl von  $k = 5$  für *CLARA* mit Manhattan als Distanzmaß, beträgt der Silhouettenkoeffizient  $s = 0.1302$ , der Timeline-Score  $TLS = 0.1670$  und der Wert für die Aufteilung der *GWL* in den Clustern  $HB_{diff} = 0.4919$ .

Führt man die Clusteranalyse mit *CLARA*, der Euklidischen Distanz und  $k = 6$  Clustern durch, so erhält man für den Silhouettenkoeffizient einen Wert von  $s = 0.1239$ , der Timeline-Score beträgt

$TLS = 0.1764$  und  $HB_{diff} = 0.4882$ .

## K-Means

K-Means ist ein bekannter Algorithmus für Clusteranalyse, bei dem jedes Cluster dabei von einem Centroid repräsentiert wird, das im Gegensatz zu *PAM* nicht eine echte Beobachtung ist, sondern der Mittelwert von allen Objekten im gleichen Cluster. Der k-Means Algorithmus wird von Wu (2012) genauer beschrieben.

Der Silhouettenkoeffizient beträgt dabei  $s = 0.1286$ , der Timeline-Wert ist  $TLS = 0.2433$  und die Maßzahl für die Güte der Aufteilung der *GWL* in den Clustern  $HB_{diff} = 0.5167$ .

## PAM mit Mahalanobis-Distanz

Da beim Clustering der Originaldaten nicht beachtet wird, dass es Korrelationen zwischen den Variablen gibt, wurde die Mahalanobis-Distanz in Betracht gezogen. Diese beachtet bei der Berechnung der Distanzen die Korrelation zwischen den verschiedenen Variablen. Korrelationen gibt es in dem Datensatz vor allem zwischen den Parametern Luftdruck und Geopotential sowie dem Standort. Es lässt sich intuitiv vermuten, dass wenn z.B. an einem bestimmten Standpunkt hoher Luftdruck vorliegt, dass der Luftdruck des nebenliegenden Standorts wahrscheinlich auch eher hoch ist.

Für  $p$ -dimensionale Daten  $x = (x_1, x_2, \dots, x_p)^T$  mit einem Vektor mit Mittelwerten  $\mu = (\mu_1, \mu_2, \dots, \mu_p)^T$  und Kovarianzmatrix  $\Sigma$  ist die Mahalanobis-Distanz generell definiert als  $D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$ .

Ist  $\Sigma$  die Einheitsmatrix, so entspricht die Mahalanobis-Distanz der Euklidischen Distanz (X. Li (2019)).

Mit der resultierenden Distanzmatrix wurde dann eine Clusteranalyse mit *PAM* durchgeführt für  $k = 6$ . Der Silhouettenkoeffizient liegt dabei bei  $-0.0014$  und ist somit negativ. Auch der Timeline-Wert liegt unter 0 mit  $TLS = -0.6657$ . Der Wert für die Güte der Aufteilung der *GWL* in den Clustern beträgt  $HB_{diff} = 0.4127$ .

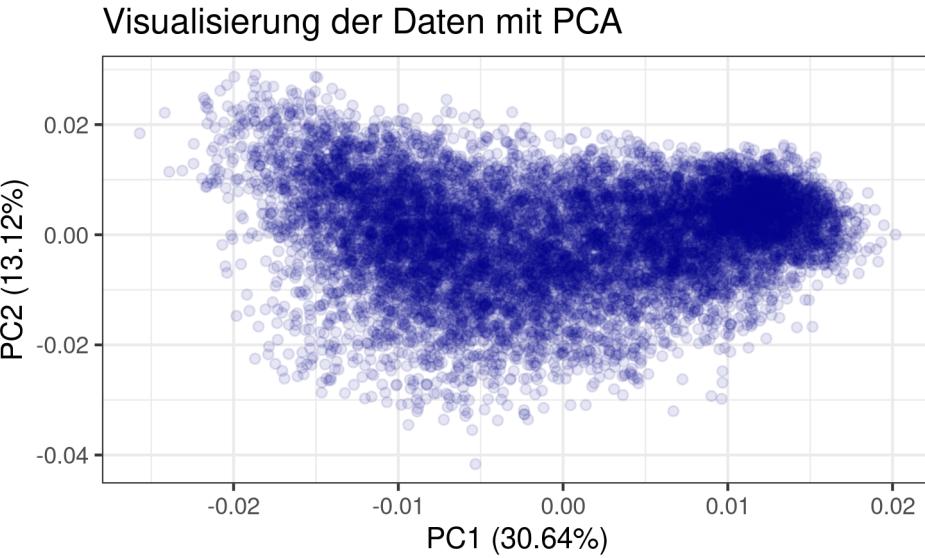
## PCA

Mit Hilfe einer Principle Component Analysis (*PCA*) lassen sich die Dimensionen eines Datensatzes effektiv reduzieren. Hierbei wird versucht die Variablen durch Linearkombinationen zu ersetzen, die jeweils bestmöglichst die Varianz der Beobachtungen beschreiben. Dadurch wird erreicht, dass ein Großteil der Varianz durch eine bestimmte Anzahl der ersten Principle Components (*PC*) erklärt werden kann, folglich eine Dimensionsreduktion mit minimalen Informationsverlust stattfindet (Jaadi (2020), Tuladhar (2020)).

In Abbildung 12 sind die ersten beiden *PC* dieses Datensatzes abgebildet und es ist zu beobachten, dass diese ca. 44% der Varianz erklären können. Eine auffällige Aufteilung der Beobachtungen in Gruppen ist hier allerdings nicht zu beobachten.

Ein Ansatz ist hier, mit der Anzahl an *PC* weiter zu clustern, die mindestens 85% der Varianz der Daten erklären (Harrison (2014)). Hier ergibt sich dies zu 12 *PC*. Eine Analyse mit *k-means* ( $k =$

6) ergab einen Silhouettenkoeffizient von 0.1566 und ein  $TLS$  von 0.1373. Außerdem lässt sich ein  $HB_{diff}$  von 0.4042 beobachten.



*Abbildung 12:* Visualisierung der Clusterlösung mittels Principal Component Analysis.

## 2.5.2 K-Means und PAM mit Extrahierten Daten

### K-Means

Obwohl die Entscheidung auf *PAM* als Clusteralgorithmus gefallen ist, wurde auch ein k-Means Clustering für den Zeitraum 1971 – 2000 durchgeführt. Diese benutzt die Euklidische Distanz und für die Berechnung der Distanzmatrix wurden diesselben Variablen und Gewichte herangezogen, wie in Abschnitt 2.4.2 erklärt.

Der Silhouettenkoeffizient betrug dabei 0.1191, der Timeline-Score  $TLS = 0.4455$  und  $HB_{diff} = 0.3379$ . Der Silhouettenkoeffizient ist hier niedriger als bei dem vorgestellten Clustering mit *PAM*, somit spricht das für die Wahl von *PAM*.

### PAM

Auch bei *PAM* wurden verschiedene Clusterings anhand des extrahierten Datensatzes für die Jahre 1971 – 2000 durchgeführt. Hier wurden unterschiedliche Distanzmetriken angewandt. Neben der Manhattan-Distanz wurde zum einen auch die Euklidische Distanz verwendet. Der Silhouettenkoeffizient beträgt dabei für eine Clusteranzahl  $k = 5$  0.1072, der Timeline-Score 0.3168 und  $HB_{diff} = 0.3683$ . Somit sind die Kriterien  $s$  und  $TLS$  kleiner als die beim Clustering mit der Manhattan-Distanz und *PAM* ( $s = 0.1411$ ,  $TLS = 0.3357$ ).

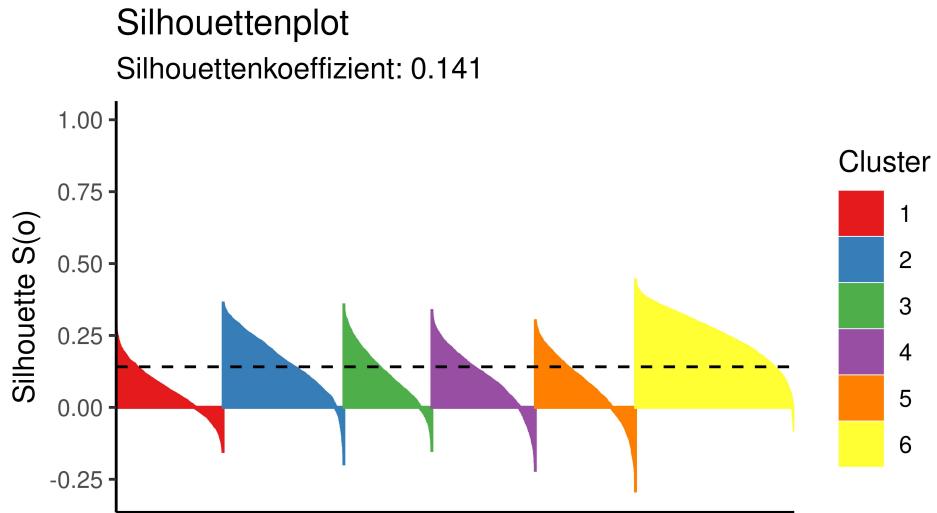
Zum anderen wurde die Gowerdistanz in Erwägung gezogen. Diese ergibt, ebenfalls mit  $k = 5$ , für den Silhouettenkoeffizienten  $s = 0.1309$ , den Timeline-Score 0.4067 und für  $HB_{diff} = 0.3616$ . Der Silhouettenkoeffizient ist demnach hier ebenfalls niedriger verglichen mit dem Clusterergebnis

mit der Manhattan-Distanz. Der Timeline-Score ist dagegen höher. Allerdings existieren hierbei fast 1000 Tage, die einzeln vorkommen. Dies ist eher nicht erwünscht, wie in Abschnitt 2.2.2 beschrieben.

### 3 Ergebnisse PAM mit extrahierten Variablen

In diesem Abschnitt werden die Ergebnisse des Clusterings mit den extrahierten Variablen dargelegt. Die Optimale Clusteranzahl lag hier bei  $k = 6$ .

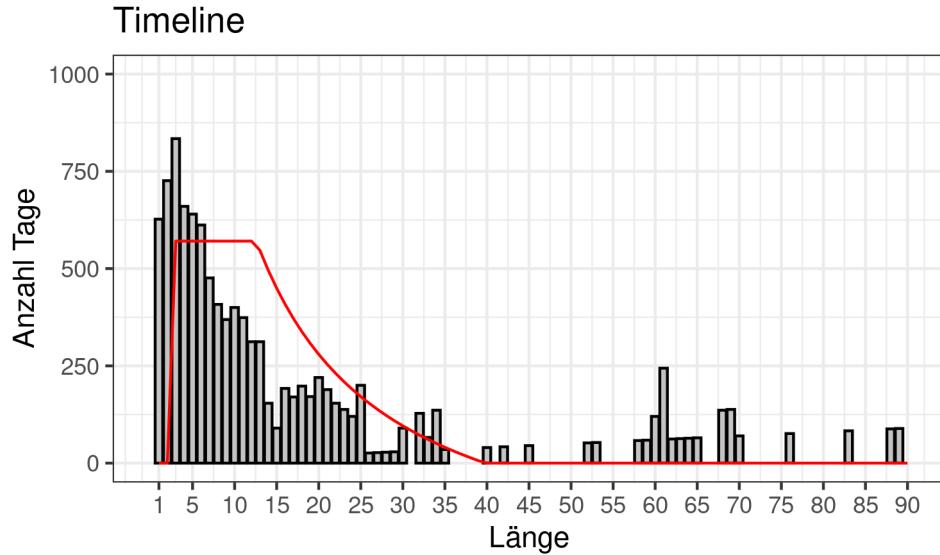
In Abbildung 13 ist der Silhouettenplot der resultierenden Cluster zu sehen. Auf der x-Achse sind hier alle Beobachtungen, aufgeteilt nach Cluster und innerhalb der Cluster nach abnehmender Silhouette geordnet, abgebildet. Der Silhouettenkoeffizient beträgt  $s = 0.141$ . Es lässt sich erkennen, dass Cluster 6 die meisten Beobachtungen aufweist. Außerdem beinhaltet es kaum Beobachtungen, die eine negative Silhouette aufweisen. Der höchste Silhouettenwert lässt sich zudem in Cluster 6 finden. Cluster 3 enthält die wenigsten Beobachtungen. Es fällt zudem auf, dass Cluster 5 die kleinste Silhouette beinhaltet, mit  $S(o) < -0.25$ .



**Abbildung 13:** Silhouettenplot für Clusterergebnis mit extrahierten Variablen mit  $k = 6$

Abbildung 14 zeigt die Timeline für dieses Clustergebnis. Es lässt sich erkennen, dass die meisten Tage in aufeinanderfolgenden Clustern der Länge 1 bis 13 sind. In Bezug auf die Timeline der *GWL* ist dies auch erwünscht, mit Ausnahme der Längen 1 und 2. Einzelne Tage kommen in den Jahren 1971 - 2000 ungefähr 600 Mal vor. Am häufigsten lassen sich aufeinanderfolgende Cluster der Länge 3 beobachten, mit ca. 800 Tagen. Hier sind jenseits der Länge 40 nur wenige Tage zu finden. Die *CWL*, die am längsten andauert, beträgt 89 Tage. Auffällig ist außerdem, die Länge 61 vier Mal auftritt. Insgesamt gehören also ungefähr 250 Tage zu einer *CWL*, die 61 Tage anhält.

Der Timeline-Score beträgt  $TLS = 0.3357$ .

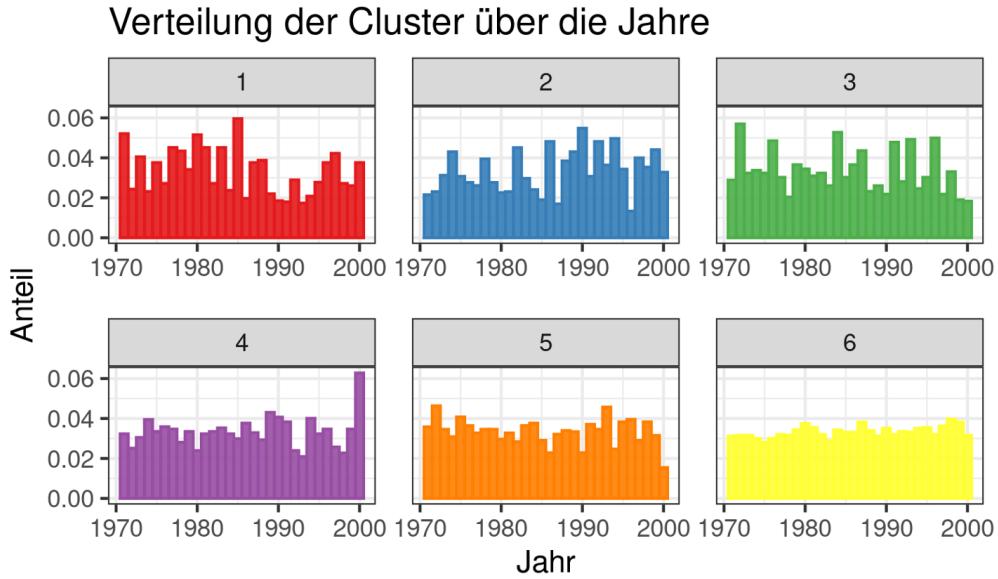


**Abbildung 14:** Timeline für Clusterergebnis mit extrahierten Variablen mit  $k = 6$ . Die rote Linie zeigt die optimale Timeline Verteilung.

Die folgenden Abschnitte präsentieren die Ergebnisse der deskriptiven Analyse der finalen Clusterlösung. Hierbei wird in Abschnitt 3.1 betrachtet, wie die Cluster 1 bis 6 über die Jahre 1971 bis 2010 verteilt sind. Zudem wird das Verhältnis von Sommer- und Wintertagen in den einzelnen Clustern betrachtet. Abschnitt 3.3 befasst sich mit den Ähnlichkeiten und Unterschieden der extrahierten 48 Variablen in den Clustern. Abschließend wird in Abschnitt 3.4 die Clusterlösung mit der *GWL* Einteilung nach Hess und Brezowsky verglichen, also in welchem Ausmaß die *GWL* über die Cluster verteilt sind.

### 3.1 Verteilung der Cluster über die Zeit

Abb. 15 stellt dar, wie häufig jedes Cluster im Zeitraum 1971 bis 2000 vorkommt. Zu erkennen ist, dass sich ein Cluster nicht auf eine bestimmte Zeitperiode beschränkt, sondern sich über den gesamten Zeitraum erstreckt. Zudem ist kein Trend in der Aufteilung der Tage innerhalb eines Clusters auf die Jahre 1971 bis 2000 erkennbar. Die Cluster 4,5 und 6 sind hierbei gleichmäßiger über den Zeitraum von 1971 bis 2000 verteilt als die Cluster 1, 2 und 3. Beispielsweise beinhaltet das Jahr 1985 6% aller Tage, die Cluster 1 zugeordnet sind, das Jahr 1990 beinhaltet hingegen nur 2% aller Tage, die Cluster 1 zugeordnet sind. Damit verglichen sind Cluster 4 bis 6 sehr gleichmäßig über alle Jahre verteilt. In Cluster 4 sticht das Jahr 2000 hervor. In diesem sind 6,3% aller Tage, die Cluster 4 zugeordnet sind, vertreten, während in allen anderen Jahren durchschnittlich je 3% aller Tage, die Cluster 4 zugeordnet sind, vertreten sind. Somit beinhaltet das Jahr 2000 ca. doppelt so viele Cluster 4-Tage verglichen mit den anderen Jahren. Von allen Clustern ist Cluster 6 am gleichmäßigsten über alle Jahre verteilt.



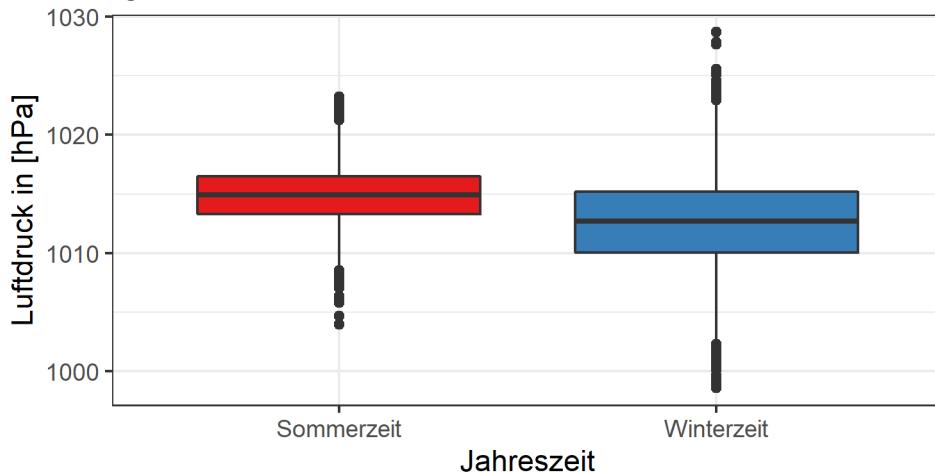
**Abbildung 15:** Aufteilung der Tage auf die Jahre getrennt nach Cluster. Pro Cluster wird dargestellt, welcher relative Anteil aller Tage, die diesem Cluster zugeordnet sind, sich in einem Jahr befinden.

### 3.2 Verhältnis von Sommer- und Wintertagen in den Clustern

Die Werte des Mittelwerts des Luftdrucks (Abb. 16) und des Mittelwert des Geopotentials (Abb. 17) sind in der Sommerzeit tendenziell höher als in der Winterzeit. Hierbei sind die saisonalen Unterschiede bei dem Mittelwert des Geopotentials deutlich stärker ausgeprägt als bei dem Mittelwert des Luftdrucks. Diese Aufteilung des Jahres in Sommer- und Winterzeit erfolgt nach Vorlage der Publikation von James P.M. 2006 (James (2007)). Um diese saisonalen Unterschiede zu berücksichtigen und die Verteilung der Winter- und Sommertagen in den Clustern zu betrachten, wird das Kalenderjahr in eine Winter- und in eine Sommerzeit aufgeteilt. Alle Tage im Zeitraum 16. Oktober bis 15. April werden als Wintertage definiert, die restlichen Tage folglich als Sommertage (James (2007)).

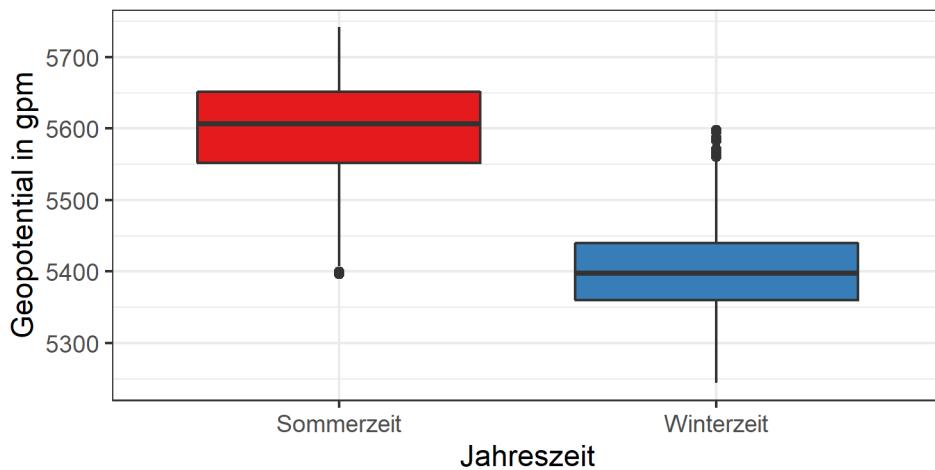
Die Graphik in Abb. 18 visualisiert die relativen Häufigkeiten der Winter- und Sommertage in jedem Cluster. Cluster 1 bis 3 enthalten überwiegend Wintertage; in Cluster 1 sind 88% aller Tage Wintertage, in Cluster 2 99% und in Cluster 3 98% aller Tage Wintertage. In Cluster 4 bis 6 sind hingegen überwiegend Sommertage vertreten; in Cluster 4 sind 79% aller Tage Sommertage und in Cluster 5 83% aller Tage Sommertage. Ein Sonderfall ist Cluster 6, da dieses ausschließlich aus Sommertagen besteht.

### Verteilung des Mittelwerts des Luftdrucks getrennt nach Jahreszeit

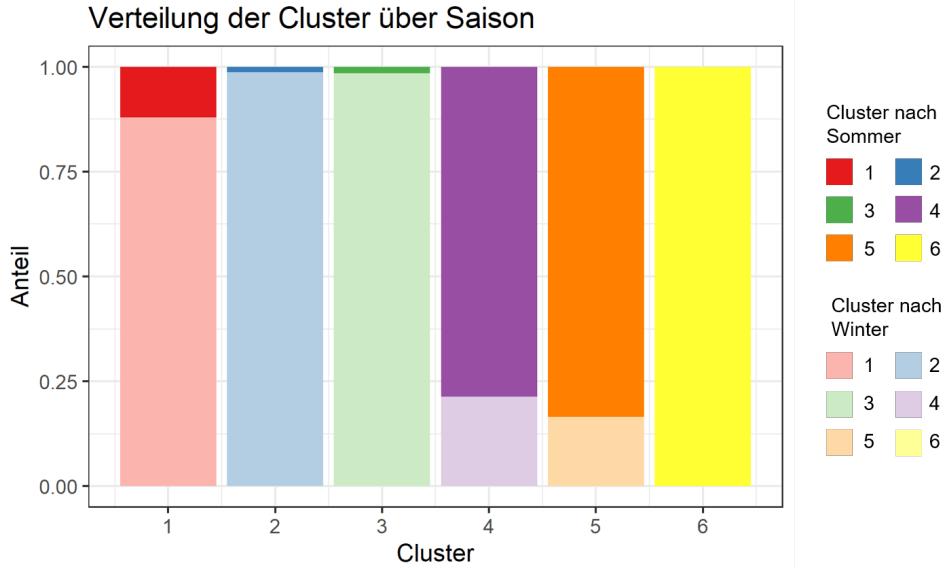


**Abbildung 16:** Darstellung der Verteilung der Variable Mittelwert des Luftdrucks getrennt nach Sommer- und Winterzeit im Zeitraum 1971-2000

### Verteilung des Mittelwerts des Geopotentials getrennt nach Jahreszeit



**Abbildung 17:** Darstellung der Verteilung der Variable Mittelwert des Luftdrucks getrennt nach Sommer- und Winterzeit im Zeitraum 1971-2000



**Abbildung 18:** Gestapeltes Balkendiagramm, der den relativen Anteil an Winter- und Sommertagen je Cluster abbildet

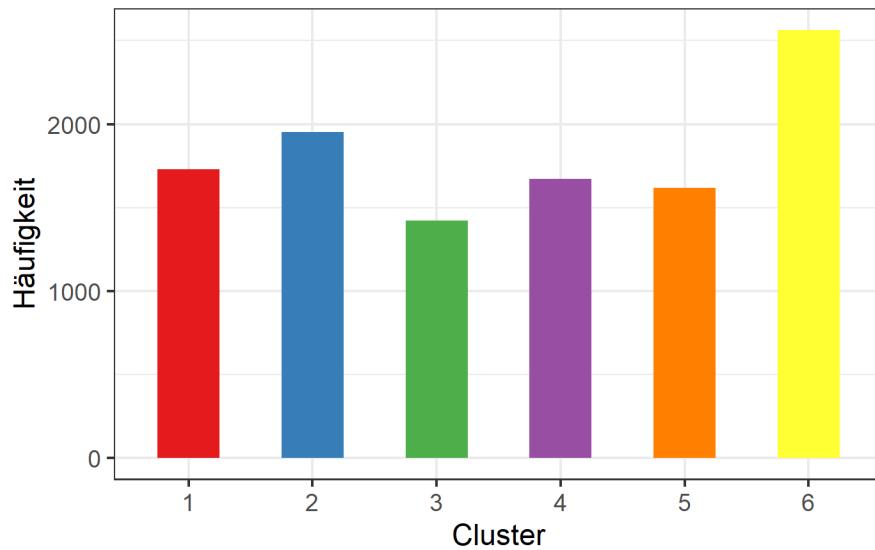
### 3.3 Unterschiede und Ähnlichkeiten in den Clustern

Dieser Abschnitt befasst sich mit der Fragestellung, wie sich die Werte der 48 extrahierten Variablen, mit denen geclustert wurde, zwischen den Clustern unterscheiden. Dazu werden Mittelwerte, Standardabweichung und Verteilungen ausgewählter, repräsentativer Variablen und dessen räumliche Verteilung über die 160 Messorte betrachtet.

Abb. 19 zeigt die Aufteilung der Tage im Zeitraum 1971 bis 2000 auf die Cluster 1 bis 6. Hierbei bildet Cluster 3 mit einer Anzahl an 1422 Tagen das kleinste Cluster. Cluster 6 beinhaltet die meisten Tage (2565 Tage). Während in Cluster 2 1952 Tage vertreten sind, beinhalten Cluster 1,4 und 5 eine ähnliche Anzahl an Tagen.

Tabelle 4 bildet Mittelwert und Standardabweichung der Variablen Mittelwert, Maximum und Minimum des Luftdrucks und Mittelwert, Maximum und Minimum des Geopotentials in jedem Cluster ab. Hierbei unterschieden sich die Werte einer Variablen über alle Cluster nur gering voneinander. Zum Beispiel beträgt die maximale Abweichung der Variable Mittelwert des Luftdrucks in allen Clustern nur  $8.2 hPa$  und die maximale Abweichung der Variable Mittelwert des Geopotentials  $287.7 gpm$ .

### Anzahl der Beobachtungen in jedem Cluster



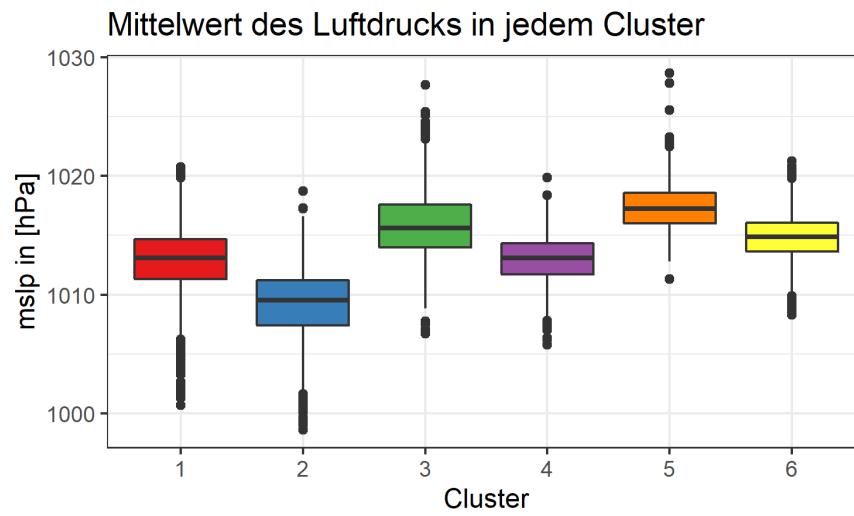
**Abbildung 19:** Verteilung der Tage im Zeitraum 1971 - 2000 auf die Cluster 1 bis 6

**Tabelle 4:** Mittelwert und Standardabweichung ausgewählter Variablen pro Cluster. Alle Werte, die den Luftdruck betreffen, sind in der Einheit hPa, alle Werte, die das Geopotential betreffen, sind in der Einheit gpm

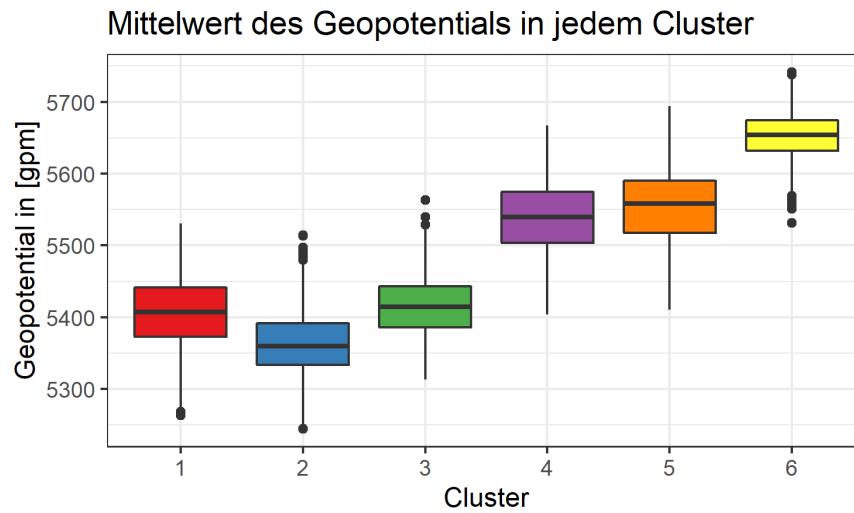
cluster	mean.mslp	max.mslp	min.mslp	mean.geopot	max.geopot	min.geopot
1	1012.8	1032.9	984.9	5406.6	5771.8	5019.3
	± 2.7	± 5.1	± 8.1	± 47.1	± 47.3	± 90.0
2	1009.2	1033.6	974.2	5364.6	5814.1	4923.3
	± 3.0	± 4.5	± 9.0	± 43.8	± 42.4	± 87.7
3	1015.8	1040.2	982.1	5415.7	5780.6	4954.5
	± 2.8	± 5.9	± 9.3	± 39.9	± 47.4	± 98.0
4	1012.9	1029.9	988.0	5539.5	5864.7	5171.5
	± 1.9	± 3.9	± 7.4	± 47.4	± 35.4	± 85.0
5	1017.4	1033.6	996.0	5555.1	5828.2	5217.9
	± 1.9	± 4.7	± 6.3	± 49.6	± 40.9	± 86.9
6	1014.8	1028.8	996.0	5652.3	5907.3	5353.6
	± 1.9	± 3.2	± 4.6	± 30.4	± 24.7	± 66.9

Abb. 20 und Abb. 21 bilden die Verteilung der Variablen Mittelwert des Luftdrucks und Mittelwert des Geopotentials je Cluster ab. Verglichen mit der Variable Mittelwert des Geopotentials unterschiedet sich der Median und die Verteilung der Variable Mittelwert des Luftdrucks in jedem Cluster wenig. Hierbei beinhaltet Cluster 2 mit einem Median von 1009.5hPa und einem Interquartilsab-

stand ( $IQR$ ) von  $3.8hPa$  (1. Quartil =  $1007.4hPa$ , 3. Quartil =  $1011.2hPa$ ) tendenziell die kleinsten Werte auf, während Cluster 5 mit einem Median von  $1017hPa$  und einem Interquartilsabstand von  $3hPa$  (1. Quartil =  $1016hPa$ , 3. Quartil =  $1019hPa$ ) tendenziell die höchsten Werte aufweist. Bei der Variable Mittelwert des Geopotentials weisen die Werte der Cluster 4 bis 6 deutlich höhere Werte auf als die Werte in den Clustern 1 bis 3. Wie bei der Variable Mittelwert des Luftdrucks beinhaltet auch das Cluster 2 bei der Variable Mittelwert des Luftdrucks tendenziell die niedrigsten Werte mit einem Median von  $5360gpm$  und einem  $IQR$  von  $58gpm$  (1. Quartil =  $5334gpm$ , 3. Quartil =  $5392gpm$ ). Cluster 6 weist tendenziell die höchsten Werte mit einem Median von  $5654 gpm$  und einem  $IQR$  von  $42gpm$  (1. Quartil =  $5632gpm$ , 3. Quartil =  $5674gpm$ ). Die Variable Mittelwert des Geopotentials weist weniger Ausreißer auf als die Variable Mittelwert des Luftdrucks.



**Abbildung 20:** Verteilung der Variable Mittelwert des Luftdrucks in jedem Cluster

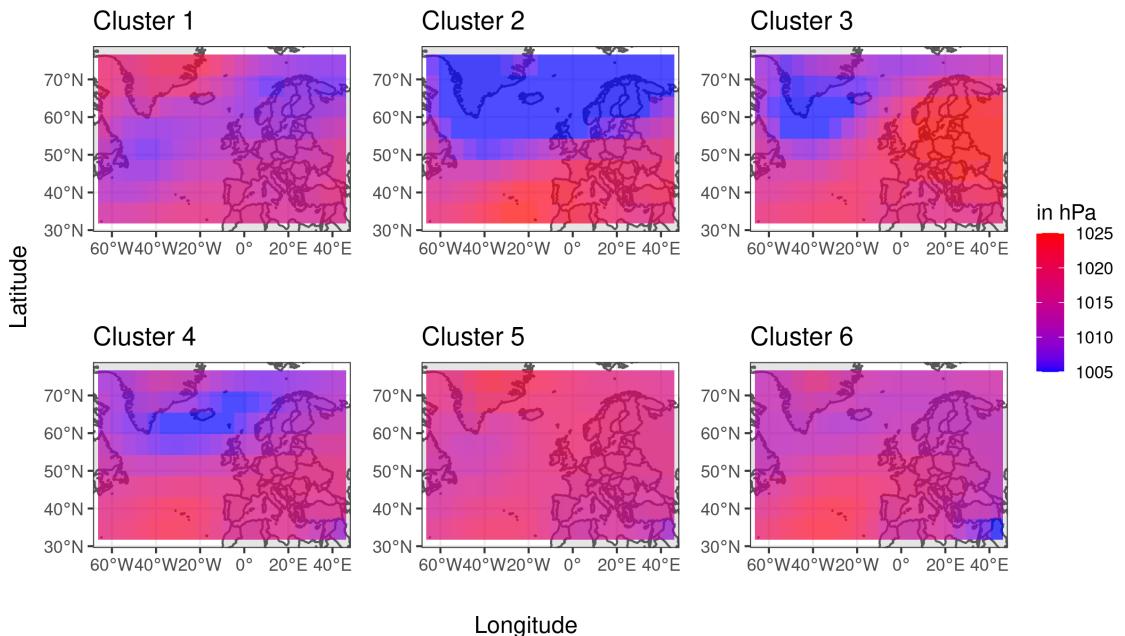


**Abbildung 21:** Verteilung der Variable Mittelwert des Geopotentials in jedem Cluster

Alle anderen extrahierten Variablen, die in die Clusteranalyse miteingegangen sind und die Verteilung beschreiben, ähneln den beschriebenen Ergebnissen von Mittelwert des Luftdrucks und Mittelwert des Geopotentials hinsichtlich der Verteilung der Werte in und zwischen den Cluster.

Nun wird betrachtet, wie sich der Mittelwert des Luftdrucks und der Mittelwert des Geopotentials räumlich unterscheidet. Abb. 22 bzw. Abb. 23 stellt pro Cluster und für jeden der insgesamt 160 Standorte das arithmetische Mittel des Luftdrucks bzw. des Geopotentials im Zeitraums 1971 – 2000 dar. Dies kann man sich vorstellen, wie ein repräsentierendes Beispielbild pro Cluster. Blaue Flächen visualisieren Gebiete mit niedrigeren, rote Flächen visualisieren Gebiete mit höheren Werten.

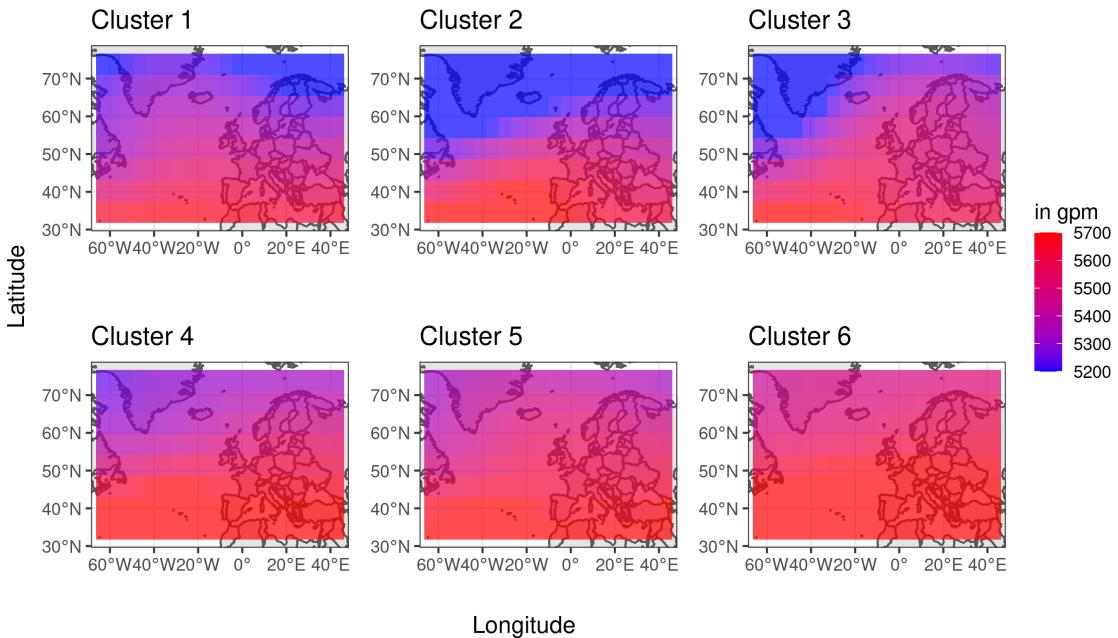
### Mslp im Mittel über Messpunkte



**Abbildung 22:** Räumliche Verteilung des Luftdrucks je Cluster. Für jeden der 160 Standorte wurde das arithmetische Mittel der Variable Geopotential über alle Tage im Zeitraum von 1971-2000 berechnet.

Die räumliche Verteilung der gemittelten Werte des Luftdrucks ist pro Cluster unterschiedlich. In Cluster 1 befinden sich nordwestlich höhere Werte während in Cluster 2 ähnlich hohe Werte des Luftdrucks im Süden zu finden sind. Zudem sind dort Gebiete mit höheren Werten größer in Cluster 1. In Cluster 3 ist hingegen das Gebiet mit höheren Luftdruckwerten östlich gelegen, in Cluster 4 und 6 südwestlich. Cluster 5 und 6 sind gekennzeichnet von tendenziell höheren Luftdruckwerten, ein ausgeprägtes Gebiet mit tieferen Luftdruckwerten ist in beiden Clustern nicht enthalten. Im Gegensatz dazu befindet sich in Cluster 2 die größte Fläche mit niedrigeren Luftdruckwerten, das nördlich gelegen ist. Insgesamt hat jedes Cluster eine eigene, charakteristische Verteilung der Luftdruckwerte.

## Geopot im Mittel über Messpunkte



**Abbildung 23:** Räumliche Verteilung des Geopotentials je Cluster. Für jeden der 160 Standorte wurde das arithmetische Mittel der Variable Geopotential über alle Tage im Zeitraum von 1971-2000 berechnet.

Bei der räumlichen Verteilung der Werte des Geopotentials aufgeteilt nach Cluster ergibt sich hingegen ein anderes Bild im Vergleich zum Luftdruck. Tendenziell befinden sich in jedem Cluster nördlich, bzw. nordwestlich niedrigere Werte, südlich, bzw. südwestlich sind höhere Werte. In den Clustern 1 bis 3 ist der Unterschied zwischen den niedrigsten und höchsten Werten des Geopotentials stärker ausgeprägt als in den Clustern 4 bis 6.

Das Cluster 6 ist sowohl bei der Variable Luftdruck als auch bei der Variable Geopotential von höheren Werten geprägt im Vergleich zu den restlichen Clustern. Eine mögliche Erklärung für dieses Phänomen ist, dass Cluster 6 als einziges Cluster ausschließlich Tage der Sommerzeit enthält.

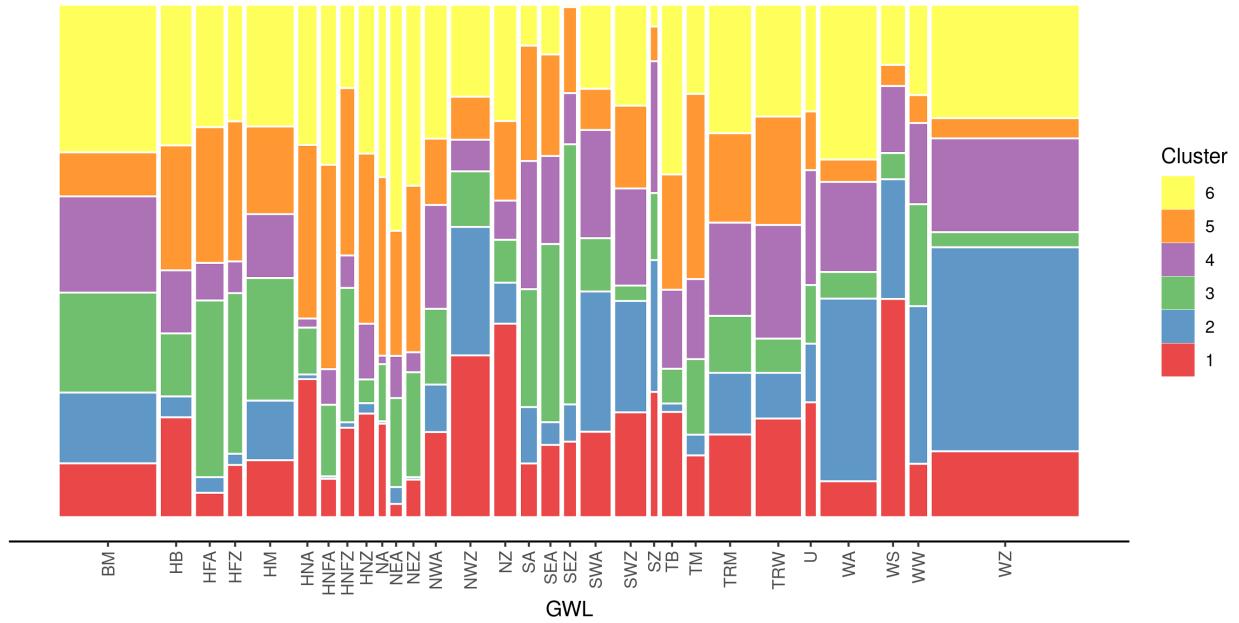
### 3.4 Vergleich der Clusterlösung mit den GWL

Der Mosaikplot in Abb. 24 stellt die Aufteilung der *GWL* auf die 6 Cluster dar. Zudem wird gezeigt, wie häufig eine bestimmte *GWL* in dem betrachteten Zeitraum von 1971 bis 2000 vorkommt. Im Allgemeinen ist die Anzahl der auftretenden Wetterlagen heterogen. Hierbei ist die Großwetterlage WZ (Westlage zyklonal) am häufigsten vertreten, dicht gefolgt von BM (Hochdruckbrücke Mitteleuropa). Selten vertreten sind die *GWL* NA (Nordlage antizyklonal), SZ (Südlage zyklonal) und undefinierte Tage, also Tage, denen keine *GWL* zugeordnet werden kann. Außer HNFA, NA und NEZ (nicht in Cluster 2), SEZ (nicht in Cluster 6) ist jede *GWL* in allen Clustern vertreten. Der Großteil der *GWL* verteilen sich hierbei gleichmäßig auf die Cluster. Dennoch gibt es auch *GWL*, die hauptsächlich in einem Cluster vorherrschen. WZ und HNFA sind zu je 40% in Cluster 2, bzw. in Cluster 5 vertreten. 52% der *GWL* SEZ befindet sich in Cluster 3 und 43% der *GWL* WS in

Cluster 1.

$HB_{diff}$  beträgt hier 0.3309.

Mosaikplot für Cluster ~ GWL



**Abbildung 24:** Mosaikplot, der darstellt, mit welchem Anteil die GWL auf die Cluster 1 bis 6 aufgeteilt sind

## 4 Diskussion und Ausblick

### 4.1 Bewertungskriterien

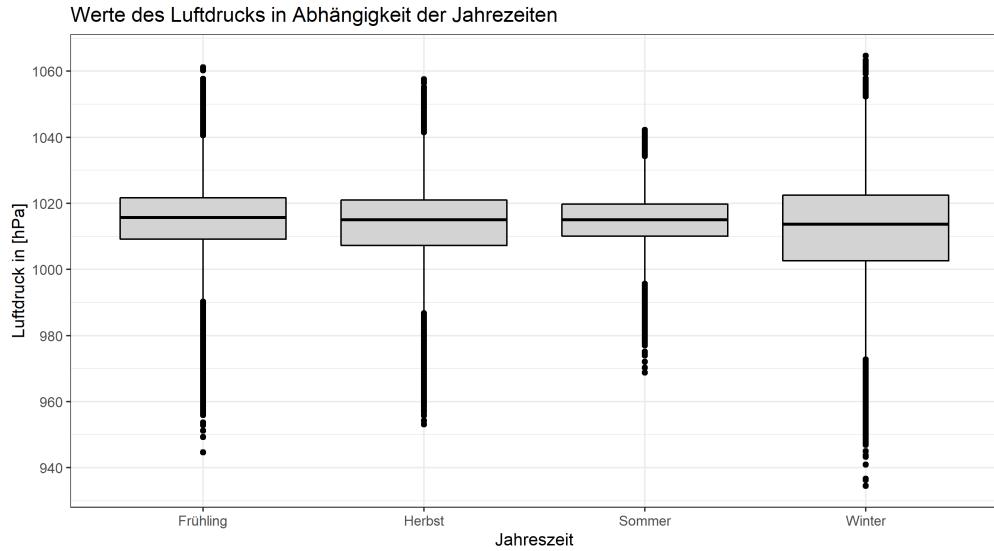
Der in Abschnitt 2.2 beschriebenen Timeline-Score (*TLS*) ist im Gegensatz zu dem Silhouettenkoeffizient nicht unabhängig von der Anzahl der Cluster und somit problematisch um Clusterlösungen mit verschiedener Clusteranzahl zu vergleichen. Die hier vorgestellten Clusterverfahren schwanken in der gewählten Clusteranzahl *nur* zwischen 5 und 6.

Gleiches Problem existiert für  $HB_{diff}$ . Hier wird zudem die Häufigkeit der *GWL* nicht mit in Betracht gezogen. Das bedeutet, dass eine *GWL* wie NEA (Nordostlage, antizyklonal), die sehr selten vorkommt, dasselbe Gewicht auf die Maßzahl wie WZ (Westlage, zyklonal), die am häufigsten vorkommende *GWL*, hat.

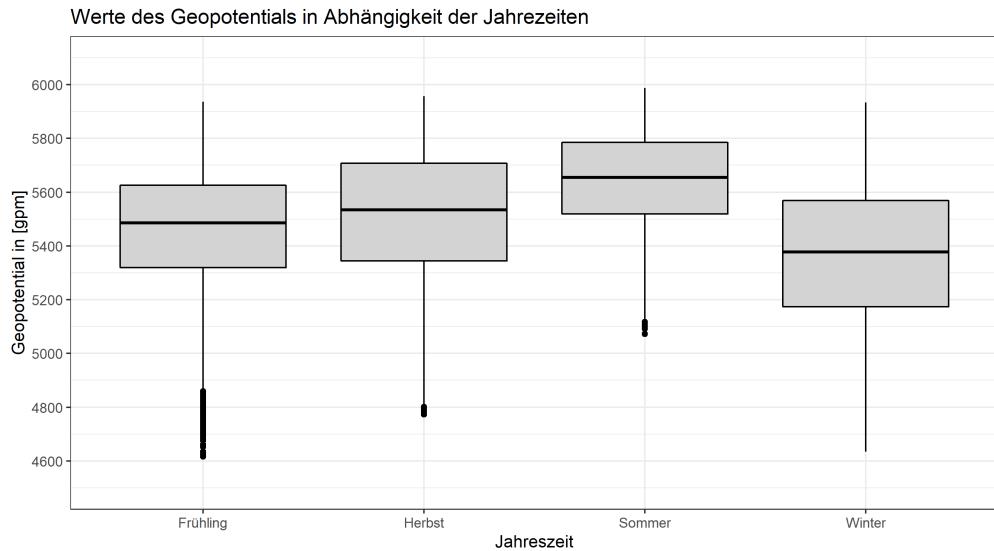
### 4.2 Saison

Es wurde auch in Betracht gezogen, eine allgemeine Saisonbereinigung durchzuführen. Dabei wurde ursprünglich angedacht, die Saisonbereinigung anhand der vier Jahreszeiten, Winter, Frühling, Sommer und Herbst durchzuführen. In Abbildungen 25 und 26 sind die Werte des Luftdrucks bzw. Geopotentials aufgeteilt nach den Jahreszeiten dargestellt. Vor allem der Luftdruck zeigt keine deutlichen Unterschiede in Bezug auf die Saison. Der Median liegt für alle bei etwas unter 1020hPa. Nur

die Streuung fällt im Sommer etwas geringer aus als in den anderen Jahreszeiten. Wie im Artikel von James (2007) vorgeschlagen, erscheint demnach eine Aufteilung in 2 Jahreszeiten sinnvoller. Da in den Clustern 1, 2 und 3 vorwiegend Wintertage und in den Clustern 4, 5 und 6 hauptsächlich Sommertage sind, siehe Abschnitt 3.2, erscheint eine Analyse mit zwei getrennten Jahreszeiten hier plausibel.



**Abbildung 25:** Werte des Luftdrucks der Jahre 1971 - 2000 aufgeteilt nach 4 Jahreszeiten

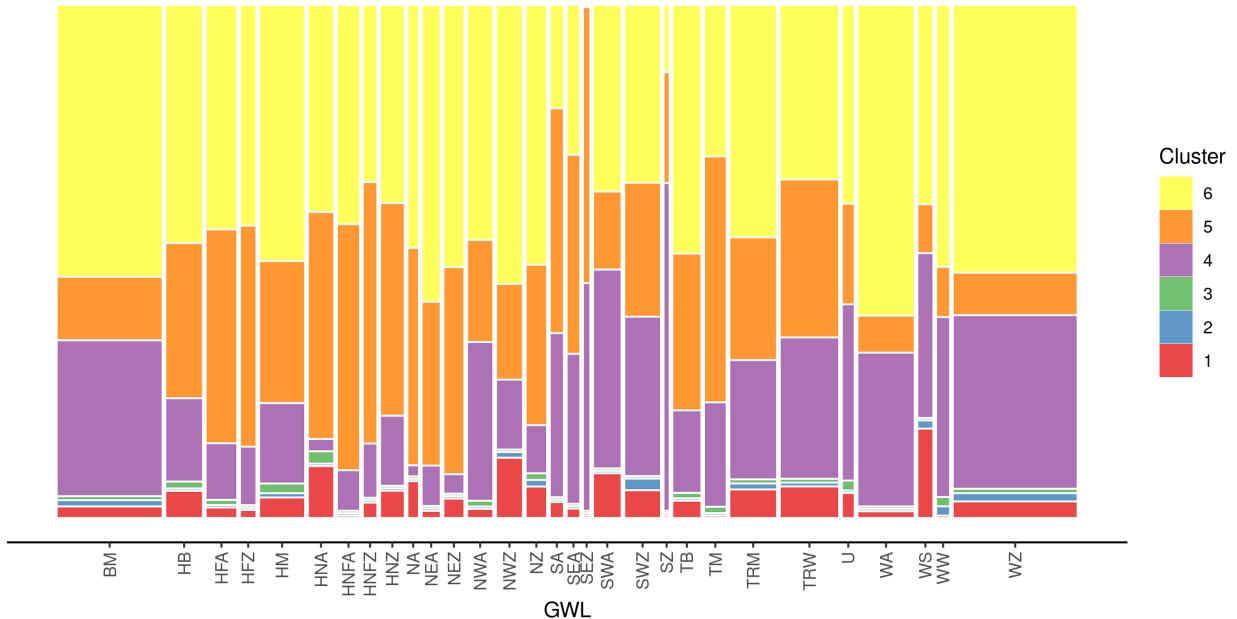


**Abbildung 26:** Werte des Geopotentials der Jahre 1971 - 2000 aufgeteilt nach 4 Jahreszeiten

#### 4.2.1 Clusterergebnis nach Saison

Abbildung 27 stellt einen Mosaikplot nur für die Sommertage dar, in dem die Verteilung der *GWL* innerhalb der Cluster zu sehen ist. Schon erwähnte Ergebnisse spiegeln sich hier wieder. Die Sommertage liegen hauptsächlich in den Clustern 4, 5 und 6. Es lässt sich noch immer sagen, dass fast alle Cluster in den *GWL* vertreten sind, aber viele treten häufiger in bestimmten Clustern auf. So liegen zum Beispiel über 50% der Beobachtungen von der *GWL* BM in Cluster 6. Der Wert für die Güte der Aufteilung der *GWL*  $HB_{diff}$  beträgt hier 0.4849. Er ist also größer, als  $HB_{diff}$  für beide Saisonen zusammen betrachtet, der bei 0.3309 liegt.

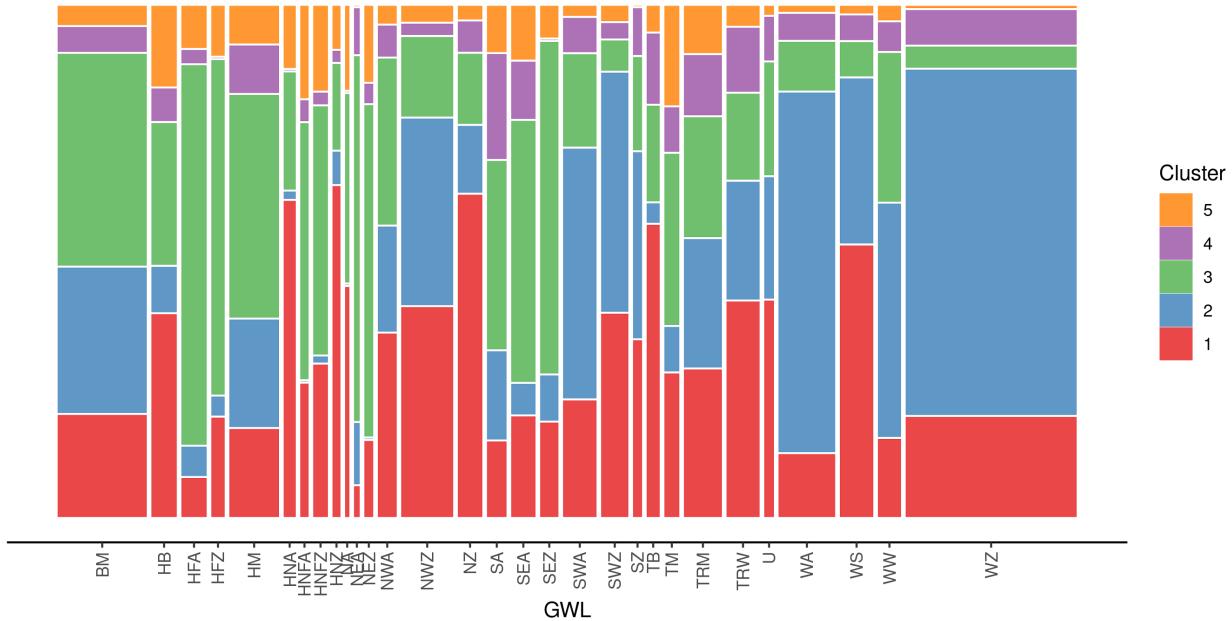
Mosaikplot für Cluster ~ *GWL* im Sommer



**Abbildung 27:** Mosaikplot nur für Sommertage, der darstellt, mit welchem Anteil die *GWL* auf die Cluster 1 bis 6 aufgeteilt sind

In Abbildung 28 sieht man wieder einen Mosaikplot, der nun nur Wintertage beinhaltet. Cluster 6 ist hier nicht enthalten, da alle Beobachtungen in diesem Cluster Sommertage sind. Hier sieht man erneut, dass die meisten Wintertage in den Clustern 1, 2 und 3 liegen. Auch hier lässt sich erkennen, dass manche *GWL* mit mehr als 50% in bestimmten Clustern vertreten sind, beispielsweise liegen ca 75% der Beobachtungen von der Großwetterlage WZ in Cluster 2.  $HB_{diff}$  liegt hier bei 0.5246.

Mosaikplot für Cluster ~ GWL im Winter



**Abbildung 28:** Mosaikplot nur für Wintertage, der darstellt, mit welchem Anteil die GWL auf die Cluster 1 bis 6 aufgeteilt sind

#### 4.2.2 Clustern getrennt nach Saison

Des Weiteren wurde auch überlegt, den extrahierten Datensatz in Sommer- und Winterdaten aufzuteilen und diese getrennt zu clustern (James (2007)). Dafür wurde wiederum das Jahr am 15/16 April, sowie dem 15/16 Oktober getrennt und folgend eine Clusteranalyse mit dem *PAM*-Algorithmus und der Manhattan-Distanzmetrik durchgeführt.

Im Gegensatz zu dem ungetrennten Clustern (Absatz 3), bei dem  $k = 6$  Cluster entstehen, resultiert das nach Saison getrennte Clustern für beide Jahreszeiten in nur  $k = 5$  Cluster. Dabei ergibt sich für den Sommer ein Silhouettenkoeffizient von  $s = 0.1126$ , einen Timeline-Score von  $TLS = 0.4767$  sowie ein  $HB_{diff}$  von 0.3448. Für den Winter  $s = 0.0909$ ,  $TLS = 0.2303$  und  $HB_{diff} = 0.2970$ .

Auffällig ist dabei, dass der Sommer in allen Kriterien bessere Werte aufweist als der Winter. Dies könnte damit zusammenliegen, dass gewisse Variablen des Luftdrucks im Winter mehr zu streuen scheinen (vgl. Abb. 16 und Abb. 25).

Allerdings scheint dieses saisongetrennte Clusterverfahren im Vergleich zur ungetrennten Methode ( $s = 0.1411$ ,  $TLS = 0.3357$ ,  $HB_{diff} = 0.3309$ ) keine Vorteile aufzuweisen, da nur der Sommer bei bestimmten Bewertungskriterien ( $TLS$  und  $HB_{diff}$ ) höhere Werte aufweist, diese aber auch noch durch die andere Anzahl an Cluster beeinflusst werden (vgl. Absatz 4.1). Der Silhouettenkoeffizient hingegen, der nicht von der Clusterzahl  $k$  beeinflusst wird, weist bei der normalen, wie in Absatz 2.4 beschriebenen Clustermethode den höchsten Wert auf.

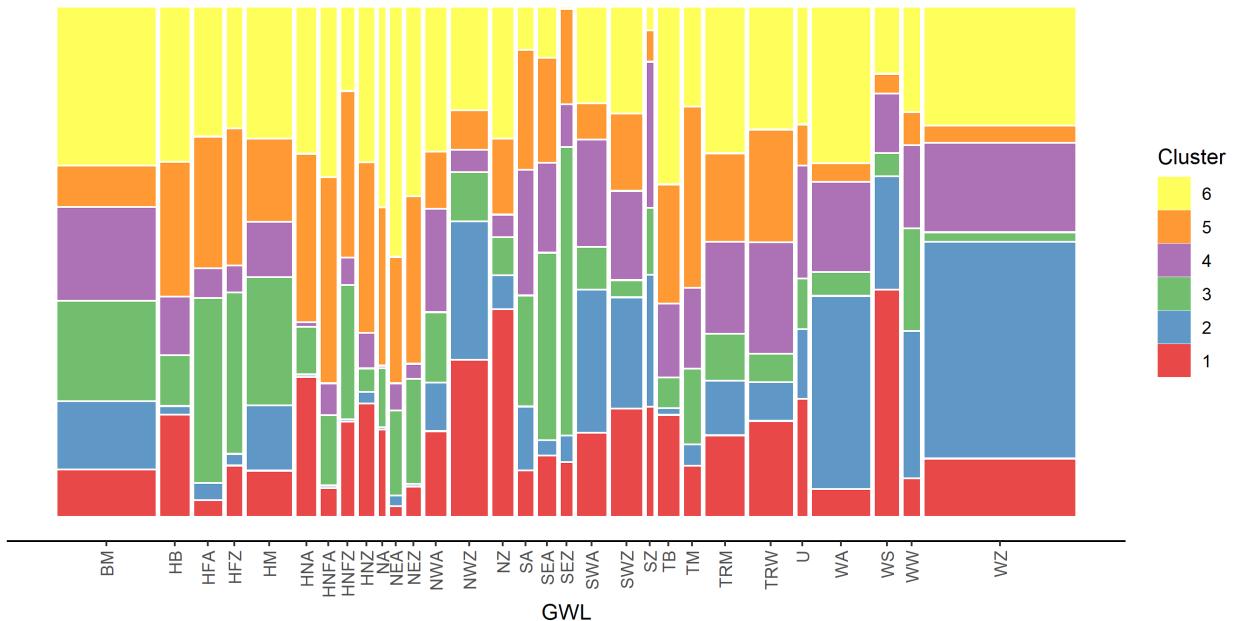
Deshalb wurde für dieses Verfahren keine Saison-getrennte Analyse bevorzugt.

### 4.3 CWL-Mindestlänge 3 Tage

Es lässt sich im Clusterergebnis beobachten, dass häufig der Wechsel zwischen *CWL* nicht sauber bzw. schlagartig stattfindet sondern, dass ein Wechsel häufig mit einem Hin- und Herschwanken zwischen den beiden *CWL* begleitet wird. Dieses Phänomen wirkt sich folglich negativ auf die Aufteilung der *GWL* in die Cluster aus, somit könnten, ähnlich wie dies bei der *GWL* Einteilung erfolgt, *CWL* mit einer Länge unter drei Tagen als *Übergang* bezeichnet werden und nicht mit in die Analyse eingehen.

Im Gegensatz zu dem Mosaikplot aus Abb. 24 sind in Abb. 29 nur Tage vertreten, bei denen drei aufeinanderfolgende Tage demselben Cluster zugeordnet sind. Während im Fall der ungefilterten Tage (weniger als drei aufeinanderfolgende Tage, die dem selben Cluster angehören und nicht entfernt wurden) die vier *GWL* HNFA, NA, NEZ und SEZ jeweils nur in fünf Clustern waren, waren bei den gefilterten Tagen (weniger als drei aufeinanderfolgende Tage, die demselben Cluster angehören und entfernt wurden) fünf *GWL*, HNA, HNFA, HNFZ, NEZ und SEZ in nur fünf Cluster und die *GWL* NA in nur vier Cluster vertreten. Wie schon erwähnt, sind bei den ungefilterten Tagen WZ und HNFA zu je 40% in Cluster 2, bzw Cluster 5, SEZ zu 52% in Cluster 3 und WS zu 43% in Cluster 1. Bei den gefilterten Tagen ist nun WZ statt 40% zu 43% in Cluster 2 und HNFA statt 40% zu 41% in Cluster 5 vertreten. SEZ befindet sich nun zu 58% in Cluster 3 statt zu 52%. Und Cluster WS ist zu 45% statt zu 43% in Cluster 1. Bei den gefilterten Tagen ist NA zu 40% in Cluster 5, bei den gefilterten Tagen befanden sich hingegen nur 34% in diesem Cluster.

Mosaikplot für Cluster ~ *GWL*



**Abbildung 29:** Mosaikplot, der darstellt, mit welchem Anteil die *GWL* auf die Cluster 1 bis 6 aufgeteilt sind. Hierbei sind nur Tage abgebildet, wo mindestens drei aufeinanderfolgende Tage dem selben Cluster zugeordnet sind

Letztendlich scheint das Herausfiltern von Tagen, bei denen weniger als drei aufeinanderfolgende Tage demselben Cluster zugeordnet sind, zu einer leichten heterogenen Aufteilung der *GWL* auf die Cluster zu führen. Eine Erklärung für diese nur leicht heterogene Aufteilung ist, dass der relative Anteil der Tage, bei denen drei aufeinanderfolgende Tage demselben Cluster zugeordnet sind, nur ca. 12% aller Tage im Zeitraum von 1971 bis 2000 ausmacht.

$HB_{diff}$  beträgt hier 0.3477 im Vergleich zu 0.3309 ohne Elimination der *CWL* der Längen 1 und 2.

#### 4.4 GWL-Mindestlänge 4 Tage

Ein auffälliger Unterschied in der Art der Unterteilung der Tage zwischen den Großwetterlagen von Hess und Brezowsky und der angewendeten Clusteranalyse ist die Anforderung einer *GWL* mindestens 3 Tage lang zu sein. Man könnte naiv vermuten, dass Tage einer bestimmten *GWL* zugeordnet werden, obwohl sie diese nicht deutlich verkörpern, um eine Länge von 3 Tagen zu erreichen, oder gar eine länger andauernde *GWL* nicht zu unterbrechen. Die geringe Anzahl der als *U* definierten Tage unterstützt diesen Gedankengang etwas.

Folglich wird hier untersucht, wie sich die Tage die sich in *GWL* befinden, die länger als 3 Tage andauern, über die Cluster verteilen lassen. Dafür werden vor dem Clustern alle Tage entfernt, an denen die herrschende *GWL* weniger als 4 Tage andauert. Danach eine Clusteranalyse mit *PAM* und der Manhattan-Metrik durchgeführt. Dies führt mit  $k = 5$  zu einem Silhouettenkoeffizienten von  $s = 0.2557$  (im Vergleich dazu  $s = 0.1411$  ohne das Entfernen bestimmter Tage). Der Timeline-Score beträgt  $TLS = -0.1083$ , da deutlich mehr *CWL* der Längen 1 und 2 gebildet werden. Die zu beobachtende  $HB_{diff}$  ist 0.3499 im Vergleich zu 0.3309, *GWL* werden hier also nicht sauberer in die Cluster unterteilt.

#### 4.5 Erweiterung Variablen

**Temperatur** - Da der Reanalyse Datensatz bereits Informationen zur Temperatur an den Messpunkten eines Tages beinhaltet, könnte dieser Parameter recht einfach in das Modell als Variable aufgenommen werden oder auch nur als Ausprägung, mithilfe der die bereits vorliegenden Parameter bereinigt werden können.

**Strömungen** - In *GWL*-Beschreibungen tauchen häufig Ausprägungen von Strömungsrichtungen auf. Ein naiver Ansatz, dies zum Teil zu replizieren, ist, aus den gegebenen Daten eines Tages zum Mslp und Geopotential eine Art “Bewegung” zu bestimmen, indem zum Beispiel anhand der Veränderung des Standortes von dem maximal gemessenen Luftdruck über die 4 Messzeiten am Tag ein Vektor berechnet wird, mit dem Aussagen zur Bewegungsrichtung und -stärke des Hochdruckgebietes getroffen werden können.

## 4.6 PAM und Filter

Die Methode, wie in Absatz 2.3 beschrieben, weist in dessen Status Quo womöglich einen zu hohen Informationsverlust auf, um alleinig angewendet zu werden. Intuitiv lässt sich vermuten, dass ein Zusammenführen mit dem im Absatz 2.4 beschriebenen Verfahren zu einer insgesamten Verbesserung führen könnte. Folgendes kann durch Addieren der Distanzmetriken vor dem Clustern auf Tagesebene erreicht werden. Die darauf folgende Clusteranalyse wurde hier auch mit *PAM* durchgeführt.

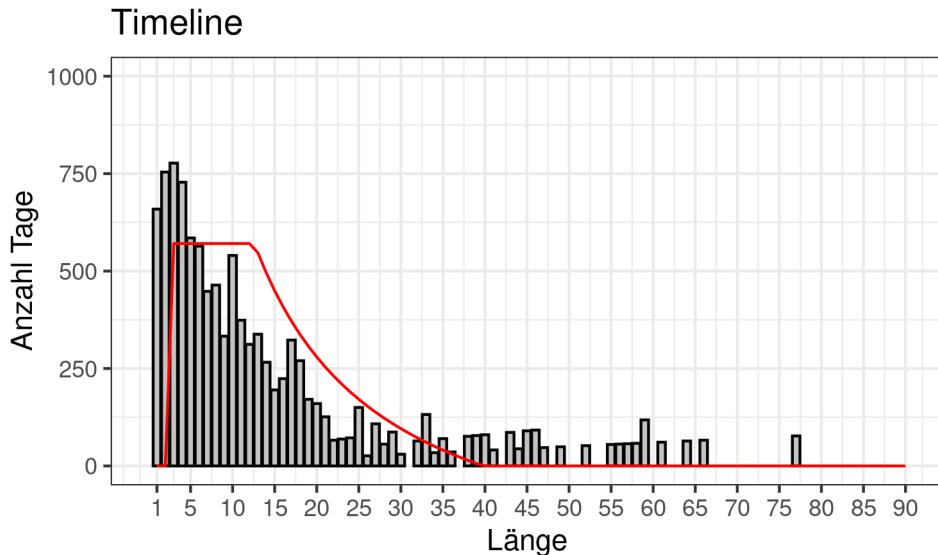
Ein einfaches Addieren mit Filtern von beiden Parametern, Mslp und Geopotential, führt zu einer Verschlechterung des Silhouettenkoeffizient von  $s = 0.141$  zu  $s = 0.099$  aber zu einer Verbesserung des *TLS* von  $TLS = 0.3357$  zu  $TLS = 0.5663$ . Dabei beträgt die Summe über alle Elemente der Distanzmatrix des Filterns ca.  $\frac{1}{7}$  der Distanzmatrix der extrahierten Methode. Also grob kann gesagt werden, dass dem Filtern hier ein Gesamtgewicht von etwas größer als  $\frac{1}{7}$  zugeteilt wird.

Es ist ein  $HB_{diff}$  von 0.3878 zu beobachten, welcher etwas höher als der  $HB_{diff}$  der Clusterlösung der extrahierten Daten mit  $HB_{diff} = 0.3309$  ist.

Allerdings fällt hierbei die Clusteranzahl von  $k = 6$  auf  $k = 5$ , was den Vergleich des *TLS* sowie des  $HB_{diff}$  wie in Absatz 4.1 beschrieben, problematisiert.

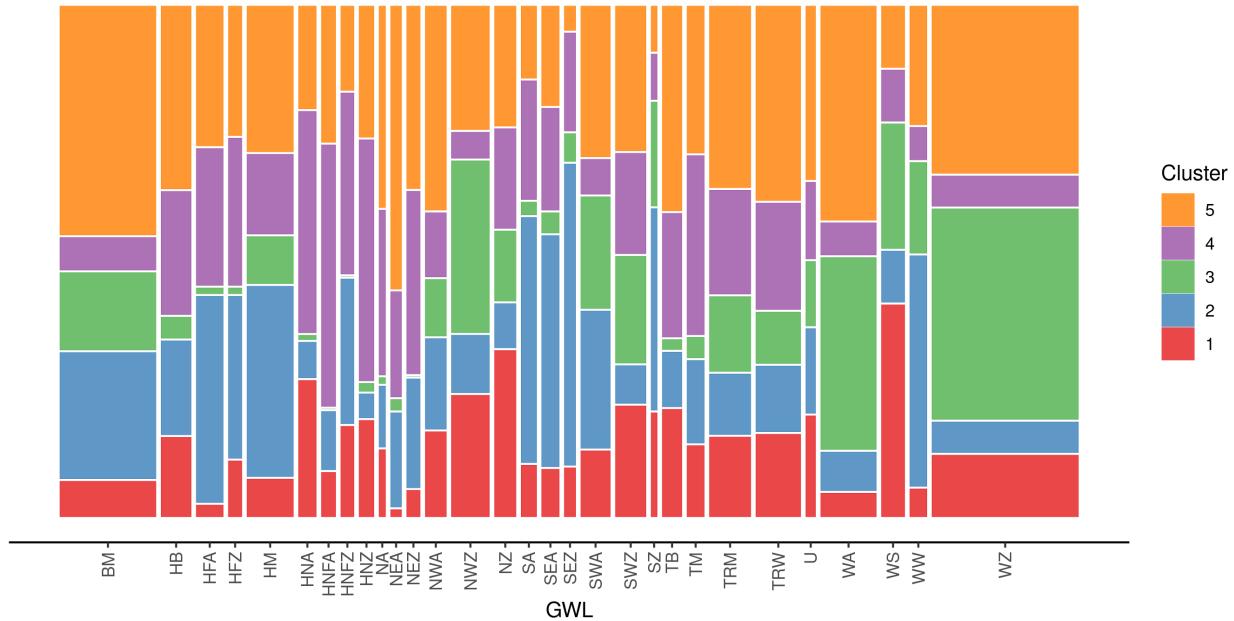
Die auf den Clusterbewertungskriterien basierte beste Kombination der beiden Ansätze verspricht die Addition der Distanzmatrix der extrahierten Daten mit der Distanzmatrix des Filterns nur mit Mslp. Dabei wird die Filter Distanzmatrix mit 3 multipliziert, wodurch ein Summenverhältnis von ca. 3.8 und somit einem ungefähren Gesamtgewicht des Filterns von  $\frac{1}{4}$  erreicht wird.

Hier ergibt sich mit  $k = 5$  Clustern ein Silhouettenkoeffizient von  $s = 0.137$  sowie einem Timeline-Score von  $TLS = 0.4118$  (siehe Abb. 30). Zudem lässt sich ein  $HB_{diff}$  von 0.4122 beobachten (siehe Abb. 31).



**Abbildung 30:** Timeline des Clusterergebnis von der Kombination der extrahierten Daten mit dem Filtern der Mslp Daten. Die rote Linie bezeichnet die erwünschte Timeline Verteilung.

Mosaikplot für Cluster ~ GWL



**Abbildung 31:** Mosaikplot der Cluster - GWL Einteilung des Clusterergebnis von der Kombination der extrahierten Daten mit dem Filtern der Mspl Daten.

Grundsätzlich lässt sich aber auch anmerken, dass der Filter-Ansatz noch sehr unausgereift ist und durch Ausweitung zur Fähigkeit mehrere Gebiete zu erkennen oder ein grundsätzliches Optimieren des Algorithmus', z.B. der Abfallrate des Nachbarschaftsparameters  $\text{eps}$  oder der Wahl des Clusterverfahren auf Tagesebene, verbessert werden könnte.

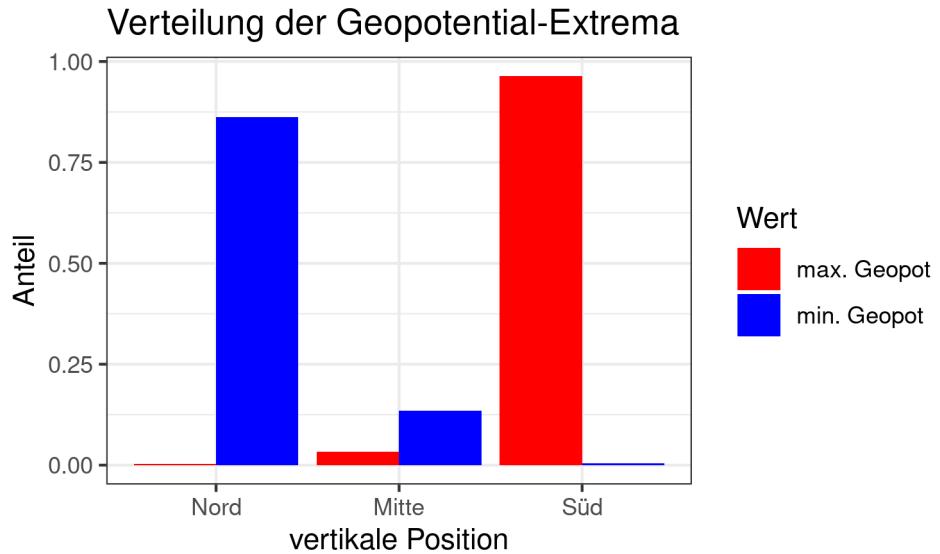
Allerdings bleibt zu untersuchen, ob ein Image-Clustering Convolutional Neural Network (*CNN*) wie in Varghese (2019) beschrieben, hierzu nicht allgemein geeigneter ist.

## 4.7 Räumliche Struktur

Anhand der geringen Clusterbewertungskriterien des Filter-Ansatz' mit Geopotential (siehe Absatz 2.3.6) sowie den über die Cluster gemittelten Tagesbilder des Clusters der extrahierten Daten mit PAM, bei dem sich die Bilder des Geopotentials räumlich sehr wenig unterscheiden (Abb. 23), lässt sich die Vermutung aufstellen, dass die Ausprägungen des Geopotentials, zumindest so wie sie vorliegen, räumlich kaum Unterschiede aufweisen und somit sich zwischen Tagen keine verschiedenen Strukturen finden lassen (siehe Abb. 32).

Auf der Basis ist es eventuell sinnvoller, nur die räumliche Komponente des Mspl beim Clusterverfahren zu betrachten oder zumindest diese höher zu gewichten. Etwas unterstützt wird dieser Gedanke davon, dass in den Beschreibungen der *GWL* nach Hess und Brezowsky häufig Bezug auf Ort und Form eines Druckgebietes am Boden genommen wird (Wetterstation Peißenberg (2021)).

Möglicherweise ließen sich auch deutlicher räumliche Strukturen in den Ausprägungen des Geopotentials finden, wenn diese im vorhinein räumlich bereinigt werden. Zum Beispiel unter der Annahme, dass im Norden immer geringere Messwerte erscheinen als im Süden.



*Abbildung 32:* Verteilung der vertikalen Position der Geopotential Extrema von 1971-2000

## 4.8 Gewichtungsvektor

Grundsätzlich lässt sich feststellen, dass die Wahl des Gewichtungsvektors in der Clustermethode der extrahierten Daten (siehe Abschnitt 2.4) einflussreich gegenüber dem Clusterergebnis ist, bzw. das gezeigte Verfahren sensitiv gegenüber den Variablen-Gewichtungen ist. Dieser Vektor wurde, wie im Abschnitt 2.4.2.3 beschrieben, in Absprache mit den Projektpartnern rein auf fachlich sinnvoller Ebene festgelegt.

Es lässt sich allerdings anzweifeln, ob so eine starre und ausgeglichene Gewichtungswahl zu dem besten Clusterergebnis führt. Wie schon im Abschnitt 4.7 angeschnitten, könnte die räumliche Komponente des Geopotentials heruntergewichtet werden. Eventuell sind bestimmte Gebiete auch relevanter als andere (z.B. der Raum über dem Festland Europa relevanter, als der über dem Atlantik), was unter anderem durch ungleiche Gewichte der Quadrantenmittelwerte dargestellt werden könnte.

### 4.8.1 Cluster-Boosting

Ein Ansatz die Gewichte der Variablen flexibler zu wählen ist, sie durch ein maschinelles Verfahren automatisch zu bestimmen. Zum Beispiel könnte eine Art komponentenweises Boosting durchgeführt werden (ähnlich wie komponentenweises Boosting in Regression vgl. L. Fahrmeir (2013)), das iterativ das Gewicht einer Variable erhöht um einen bestimmten Wert zu optimieren.

$$W^{(t)} = \operatorname{argmax}_{W_j, j=1, \dots, k} \left( f \left( \operatorname{cluster}(W_j^{(t-1)} X) \right) \right)$$

$$\text{wobei } W_j^{(t-1)} = W^{(t-1)} + c_j$$

$X$  = skalierter Datensatz ,  $W = (w_1, \dots, w_k)^T$  = Gewichtungsvektor

$c$  = Schrittweite ,  $f$  = Bewertungsfunktion ,  $t$  = Iterationsindex

Um die benötigte Rechenleistung zu reduzieren sowie auch zu verhindern, dass das Verfahren sich festläuft, indem z.B. immer dasselbe Gewicht erhöht wird, könnte eine Sampling-Strategie durchgeführt werden, sodass pro Iteration immer nur ein Teildatensatz geclustered wird. Ansatzweise werden hier 3 5-Jahres Perioden pro Iteration zufällig gewählt, die den Datensatz einer Iteration darstellen:  $D = d_1, d_2, d_3$ , wobei  $d_i$  mit  $i \in \{1, 2, 3\}$  ein 5 Jahres-Datensatz ist. Die zu optimierende Funktion könnte wie folgt zum Beispiel die Summe aus dem Durchschnitt des *Silhouettenwertes* der 3 Datensätze, dem Durchschnitt des *TLS* der 3 Datensätze und einem *Stabilitätswert* (*stab*) sein.

$$f(D) = \operatorname{avgSil}(D) + \operatorname{avgTLS}(D) + \frac{1}{3} \operatorname{stab}(D)$$

$$\operatorname{avgSil}(D) = \frac{1}{3} \sum_{i=1}^3 s(d_i)$$

$$\operatorname{avgTLS}(D) = \frac{1}{3} \sum_{i=1}^3 \operatorname{TLS}(d_i)$$

$$\operatorname{stab}(D) = 1 - (\max_{i,j=1,2,3} (|s(d_i) - s(d_j)|) + \max_{i,j=1,2,3} (|\operatorname{TLS}(d_i) - \operatorname{TLS}(d_j)|))$$

mit :  $s$  = Silhouettenkoeffizient

## 4.9 Zeitliche Struktur

Allgemein muss erkannt werden, dass jeglicher hier beschriebener Versuch Wetterlagen anhand einer Clusteranalyse der Tage einzuteilen, vernachlässigt, dass die Tage eine zeitliche Reihenfolge mit sich bringen. Anhand der Vorgabe, dass eine *GWL* nach dem Katalog von Hess und Brezowky mindestens 3 Tage lang sein muss, lässt sich vermuten, dass einzelne Tage keine Wetterlage definieren können, sondern nur einen Teil davon bilden.

Eine Möglichkeit dem entgegenzuwirken, liefert eventuell das Einbringen einer “3-Tage Regel” in das Clusterverfahren, die Tage nur einem bestimmten Cluster zuteilt, wenn in eine Richtung die folgenden zwei Tage ebenfalls diesem Cluster zugeteilt werden würden. Womöglich vorstellbar ist dies mit einem Verfahren wie (z.B.) *PAM*, dass seinen Medoid versetzt, wenn dadurch ein bestimmtes Gütekriterium verbessert wird, zumindest annähernd möglich. Das Gütekriterium könnte dann zum Beispiel zum Teil aus dem hier vorgestellten *TLS* bestehen.

Vorstellbar ist auch, dass eine Wetterlage nicht durch ein festes Muster, dem alle Tage innerhalb der Wetterlage ähneln, definiert werden kann, sondern dass eine Wetterlage ein fluides Konzept ist, dessen Anfangstage, Endtage und Tage dazwischen, sich anders ausprägen. Somit wäre hier eine Methode benötigt, in der solches zeitlich abhängiges Ausprägungsmuster über mehrere Tage definiert werden kann.

Grundsätzlich lassen sich die Daten aber auch als Format eines Videos betrachten, da eine Abfolge von Bildern mit zeitlich vorgegebener Reihenfolge vorliegt. Vielleicht ist es im Allgemeinen sinnvoller, Abschnitte in diesem Video zu suchen, die anderen ähneln. Dafür müsste jedoch auf ein Modell ausgewichen werden, das ein solches Datenformat benutzen kann.

## 5 Schluss

Insgesamt lässt sich festhalten, dass hier vorgestellte Clusterverfahren in Bezug auf den Silhouettenwert Strukturen bis zu ca.  $s = 0.15$  aufweisen. Der zeitliche Verlauf der Tageseinteilung der Clustergebnisse erreicht dabei teilweise annähernd sinnvolle Strukturen mit *CWL* zusammenhängender Tage einer als sinnvoll angenommenen Länge. Und das, obwohl die zeitliche Komponente der Daten in der Clusterbildung nicht berücksichtigt wird. Letzteres wurde als einer der eventuell größeren Mängel der vorgestellten Methoden erkannt.

Die *GWL* nach Hess und Brezowsky finden sich jeweils meistens in allen Ergebnisclustern wieder. Zum Teil heben sich dennoch *GWL* hervor, die mit einem größeren Anteil einem bestimmten Cluster zugeteilt werden können. Eine besonders saubere Teilung der 29 *GWL* in allgemeinere Wettertypen lässt sich allerdings nicht feststellen.

Im Kapitel 4 werden zwar einige vielversprechende Ansätze beschrieben, die Methoden zu verbessern bzw. auszubauen. Diese benötigen aber eventuell eine tiefgründige Analyse zur fachlichen Sinnhaftigkeit vor einer Implementation.

## Anmerkungen

Reproduzierbarer Code und relevante Dateien sind in einem Github Repository der Organisation *weather-frog* zu finden. Für Zugang, Kontakt zu den Autoren aufnehmen.

Wir, Katja Gutmair, Noah Hurmer, Stella Akouete und Anne Gritto, möchten uns herzlich bei M.Sc. Maximilian Weigert und M.Sc. Magdalena Mittermeier für die angenehme Zusammenarbeit und bei Prof. Dr. Helmut Küchenhoff für die Beratung bedanken.

**Tabelle 5:** R Libraries

Verfahren	Funktion
<b>cluster</b>	
CLARA	clara()
PAM	pam()
<b>stats</b>	
k-Means	kmeans()
<b>dbSCAN</b>	
DBSCAN	dbscan()
kNN-Distanz	kNNDist()
<b>e1071</b>	
FUZZY	cmeans()

## 6 Referenzen

- B. Everitt, T. Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with r*. Springer, New York.
- Cebeci, Zeynel. 2017. “Partitioning Cluster Analysis Using Fuzzy c-Means.” <https://cran.r-project.org/web/packages/ppclust/vignettes/fcm.html>.
- Ernst, Dittmann. 1995. “Objektive Wetterlagenklassifikation.” *Berichte Des Deutschen Wetterdienstes* 197.
- Harrison, Myles. 2014. “PCA and k-Means Clustering of Delta Aircraft.” <https://www.r-bloggers.com/2014/06/pca-and-k-means-clustering-of-delta-aircraft/>.
- Hellbrück, R. 2016. *Angewandte Statistik Mit r: Eine Einführung für ökonomen Und Sozialwissenschaftler*. Springer Gabler, Wiesbaden.
- Jaadi, Zakaria. 2020. “A Step-by-Step Explanation of Principle Component Analysis.” <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>.
- James, P. 2007. “An Objective Classification Method for Hess and Brezowsky Grosswetterlagen over Europe.” *Theoretical and Applied Climatology* 88: 17–42. <https://doi.org/10.1007/s00704-006-0239-3>.
- Karran, Denny. Zugegriffen: 2021-03-19. “Meteorologische Grundlagen Das Geopotential,” Zugriffen: 2021-03-19. <https://www.synoptische-meteorologie.de/meteorologische-grundlagen/geopotential/>.
- L. Fahrmeir, S. Lang, T. Kneib. 2013. *Regression: Models, Methods and Applications*. Springer, Berlin, Heidelberg.
- L. Kaufman, P.Rousseeuw. 2005. “Finding Groups in Data: An Introduction to Cluster Analysis,” 126.
- P. C. Werner, F. W. Gerstengarbe. 2010. “PIK Report No. 119.” <https://www.pik-potsdam.de/en/output/publications/pikreports/.files/pr119.pdf>.

- Q. Zhang, I. Couloigner. 2005. “A New and Efficient k-Medoid Algorithm for Spatial Clustering.”
- Tang, Dave. 2017. “The Rand Index.” <https://davetang.org/muse/2017/09/21/the-rand-index/>.
- Tiefgraber, M. 2013. “Großwetterlagen.” <https://14-tage-wettervorhersage.de/news/thema/131104/>.
- Tuladhar, Sunny K. 2020. “K-Means and PCA for Image Clustering: A Visual Analysis.” <https://towardsdatascience.com/k-means-and-pca-for-image-clustering-a-visual-analysis-8e10d4abba40>.
- Varghese, Danny. 2019. “Image Clustering Using Transfer Learning.” <https://towardsdatascience.com/image-clustering-using-transfer-learning-df5862779571>.
- Wetterstation Peißenberg, private. 2021. “Großwetterlagen übersicht.” [http://www.sklima.de/wetterlagen\\_uebersicht.php](http://www.sklima.de/wetterlagen_uebersicht.php).
- “WMO Guidelines on the Calculation of Climate Normals.” 2017 WMO-No.1203.
- Wu, Junjie. 2012. *Advances in k-Means Clustering: A Data Mining Thinking*. Springer, Berlin, Heidelberg.
- X. Jin, J. Han. 2017. “K-Medoids Clustering.”
- X. Li, L. F. Li, S. R. Deng. 2019. “Outlier Detection Based on Robust Mahalanobis Distance and Its Application,” 3. <https://doi.org/10.4236/ojs.2019.91002>.
- Yildirim, Soner. 2020. “DBSCAN Clustering — Explained.” <https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556>.