 Clusteranalyse von Wetterdaten zur Identifikation von Wetterlagetypen

Katja Gutmair, Noah Hurmer, Stella Akouete und Anne Gritto

01. März 2021

to be written.

# Einleitung

Die Fridays for Future Bewegung ist eine globale soziale Bewegung, die vor allem von Schülern und Studenten ausgeht. Diese gehen jeden Freitag auf die Straße und protestieren, um auf den Klimawandel und ihre Folgen aufmerksam zu machen.  
Es ist bekannt, dass der Klimawandel viele Veränderungen mit sich bringt. Magdalena Mittermeier vom Department für Geografie und Maximilian Weigert vom Statistischen Institut der LMU untersuchen, wie sich das Auftreten verschiedener Großwetterlagen (GWL) unter dem Einfluss des Klimawandels verändert. Im Rahmen des Statistischen Praktikums unterstützen wir dieses Projekt, indem wir beobachtete Wetterdaten auf täglicher Basis in Cluster einteilen.

## Großwetterlagen

Es gibt zwei Kategorien, um die Großwetterlagen zu unterteilen, einmal die objektiven und subjektiven Großwetterlagen. Die uns zur Verfügung gestellte Einteilung in Großwetterlagen beruht auf subjektiven Einteilung, nach dem Katalog nach Hess und Brezowsky, der im Jahr 1952 veröffentlicht wurde. Es wurden 29 Großwetterlagen über Europa und dem Atlantik definiert. Zudem gibt es die Kategorie Unbestimmt (U), die angegeben wurde, wenn sich ein Tag keiner GWL zuordnen ließ (siehe Peter C. Werner (2010)). Die jeweiligen Tage lassen sich anhand von mehreren Variablen einteilen. Die Einteilung erfolgt nach den Zirkulationsformen in Bodennäheund in mittleren tropesphärischen Niveau. Die Zirkulationsformen sind zyklonal und antizyklonal Die Lage “West zyklonal” kommt am häufigsten vor.

## Fragestellung

…..

# Methodik

## Daten

Seit 1900 wird viermal täglich an 160 Standorten das Geopotential auf 500 hPa in und der Luftdruck (Mslp) in Pascal auf Meeresspiegelhöhe erhoben. Diese Variablen sind Teil des Reanalyse-Datensatzes ERA-20C, der für die Analysen zur Verfügungung steht. Außerdem ist uns die GWL für jeden Tag im Zeitraum von 1900 bis 2010 bekannt. Im Rahmen des Statistischen Praktikums wird mit dem Reanalyse-Datensatz geclustert.

## Clusterbewertungskriterien

Um den Erfolg einer Clusterlösung bewerten zu können und somit verschiedene Clusteransätze vergleichen zu können, mussten vorerst Bewertungskriterien etabliert werden. Hierbei soll zum einen 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.

### 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.

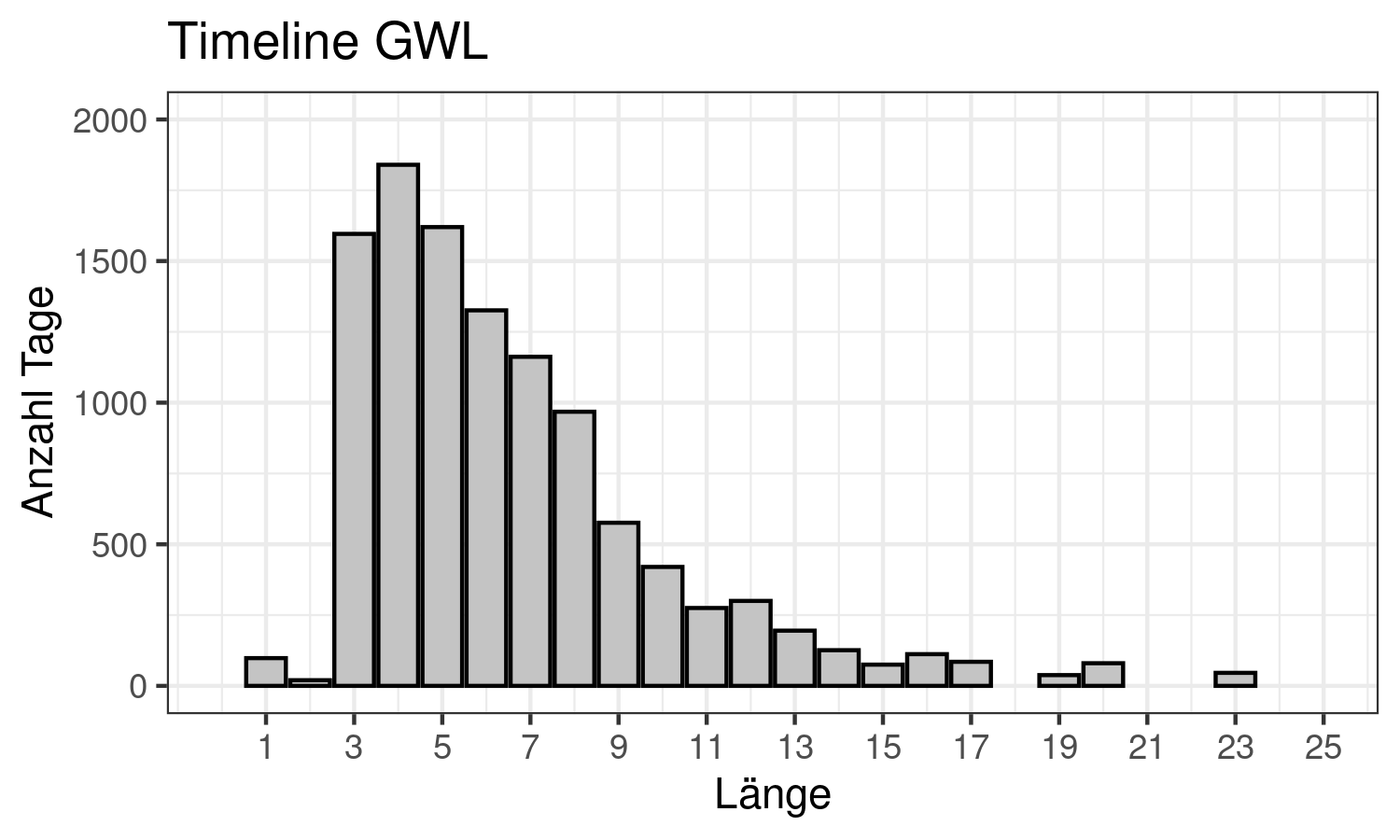
Der Silhouettenkoeffizient ist definiert als die Summe von Silhouetten. Gehört eine Beobachtung zum Cluster , so ist die Silhouette von definiert als

wobei durchschnittliche Distanz zu allen anderen Objekten desselben Clusters und durchschnittliche Distanz zu allen anderen Objekten des nächstgelegenen Clusters ist. Der Silhouettenkoeffizient ist dann definiert durch die durchschnittliche Silhouettenwerte von allen Beobachtungen in einem Datensatz , also

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

### 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 betrachten. 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.

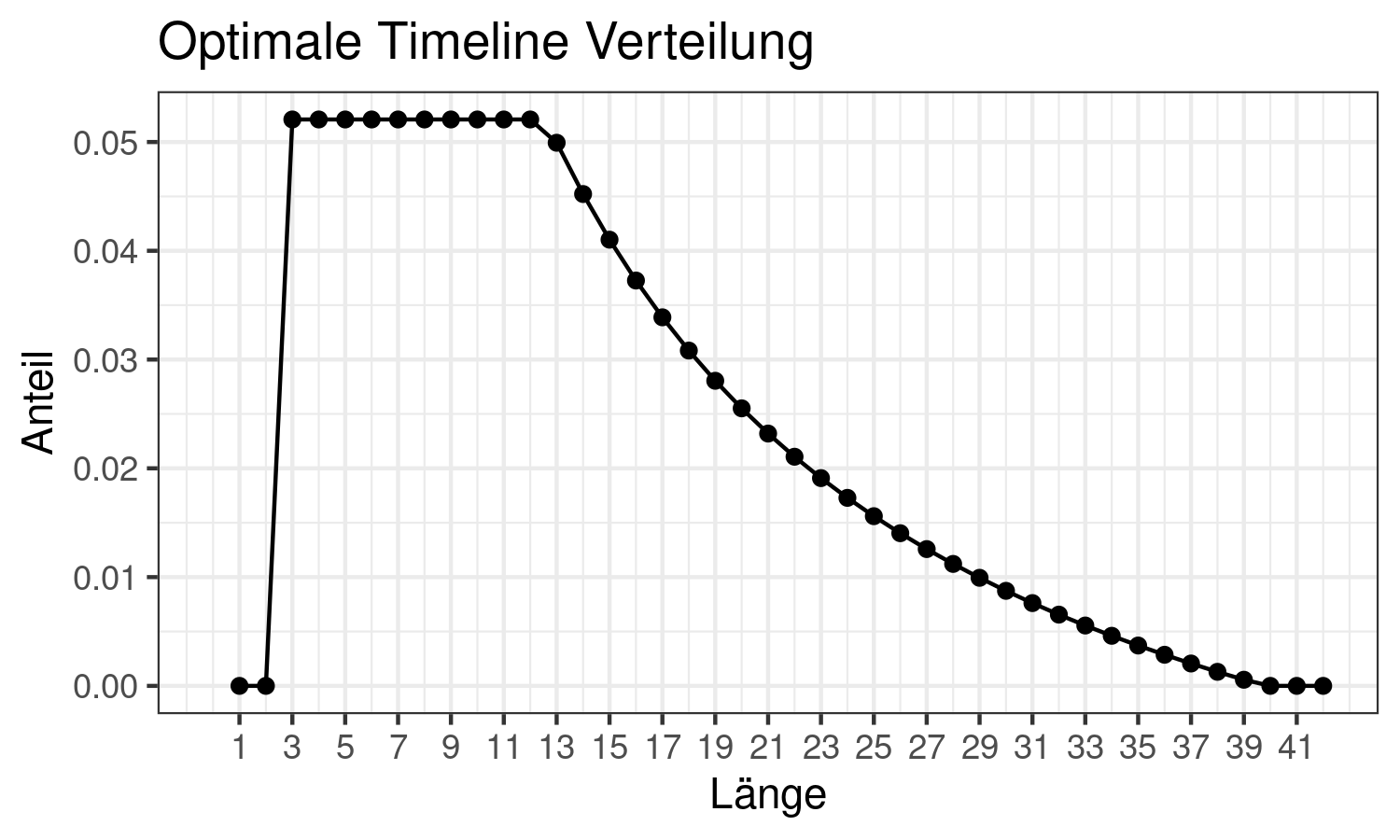


Timeline der GWL 1971-2000

Bei der Darstellung in Abbildung ist die Anzahl der GWL jeweils mit der Länge dessen multipliziert, 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 ). Der *Timeline-Score* (TLS) ergibt sich dann aus der Summe der punktweisen Abweichungen der Timeline der Clusterlösung zur optimalen Timeline Verteilung. Um den Wertebereich auf [0,1] zu beschränken, wird hier noch mit dem maximal möglichen Abstand normiert.



Erwünschte Verteilung der Timeline einer Clusterlösung

### Vergleich zur GWL-Aufteilung

Die Aufteilung der GWL nach Hess und Brezowsky über die Cluster wird im Folgenden nicht benutzt um Modelle zu bewerten bzw. zu vergleichen, 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: ) 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.

## Clustern mit den Originaldaten

Die Grundidee des Clustern ist, Objekte so in Gruppen aufzuteilen, dass Beobachtungen innerhalb der Cluster möglichst homogen und die Gruppen untereinander möglichst heterogen sind. Um ein Clustering durchzuführen wird zum einen ein Distanzmaß benötigt, das die Distanzen zwischen Beobachtungen in einem Datensatz angibt. Zum anderen muss ein Clusteralgorithmus verwendet werden, der resultierende Cluster aus dem Datensatz ausgibt. Sowohl Distanzmaß als auch der Algorithmus können selber implementiert werden. Allerdings gibt es auch schon viele verschiedene Metriken und Algorithmen, die für eine Clusteranalyse geeignet sind.

**Distanzmaß** Ein Distanzmaß gibt die Distanz von allen Beobachtungen eines Datensatzes untereinander an. Es ist somit metrisch und ein Distanzmaß für die Menge an Beobachtungen bildet auf die reellen Zahlen ab . Die Distanz zwischen zwei Beobachtungen und wird also geschrieben als . Für metrische Distanzmaße postuliert man nach Folien Annika Hoyer

**Clusteralgorithmus** Es gibt verschiedene Ansätze zu clustern. In diesem Projekt werden Partitionierende Verfahren und Dichtebasiertes Clustering betrachtet. Partitionierende Verfahren, oder auch Optmial Partitionen, messen die Qualität von Partitionen anhand eines Gütekriteriums. Gesucht sind dabei die Partitionen, die hinsichtlich des Gütekriteriums optimal sind, nämlich . Allerdings können dabei nicht alle zulässigen Partitionen betrachtet werden, da die Dimesionen dabei zu groß werden. Daher arbeitet ein Algorithmus mit numerischen Lösungsansätzen durch Austauschverfahren. Das bedeutet aber auch, dass die ausgegebene Lösung nicht die bestmögliche sein muss, sondern eventuell nur ein lokales Optimum. s. Hoyer Folien. Der K-Means oder K-Medoid Algorithmus sind bekannte Beispiele für die Partitionierenden Verfahren.

Dichtebasiertes Clustern…

### Filter Ansatz

#### Motivation

Beim genaueren Blick in die Beschreibungen einzelner GWL ist zu erkennen, dass diese häufig duch Position oder Formation 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 (“Großwetterlagen übersicht,” n.d.). 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.

#### 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. Außerdem 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 beim Vergleich der Tage nicht mit einbezogen werden, 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 also “gefiltert” werden. Da sonst bestimmte Hyperparameter oder Grenzwerte fest angegeben werden müssten, lässt sich dies durch ein seperates Clusterverfahren über die 160 Standorte pro Tag erreichen. Mit der Idee anhand beider Parameter-Messwerte (Mslp und Geopotential) “Gebiete-Muster” zu erkennen und, da durch experimentieren herausgefunden wurde, dass die beiden Parameter vereinigt keine besonders sauberen Muster zu erkennen ließen, 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.

#### DBSCAN und Fuzzy

Um der Anforderung der nicht uniformen Gebiete gerecht zu werden, erscheint ein dichtebasiertes 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 ein sog. 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. (“DBSCAN Clustering,” n.d.) *DBSCAN* benötigt keine Angabe der Clusteranzahl, sondern nur der Hyperparameter *minPoints* (minimale Anzahl an Punkte 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 präsentiert, selbst wenn der Nachbarschaftsparameter pro Tag anhand dem Wert der größten Wölbung eines kNN-plots (k = minPoints) spezifisch berechnet wird. 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 Algorythmus’ - nicht möglich. Deswegen lässt sich beobachten, dass die Extrema oft nicht in einem der eingeteilten Clustern befinden, da sie sich 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 as nicht von besonderem Interesse vermutet und deshalb wird *DBSCAN* hierfür ausgeschlossen.

Ein weiterer vielversprechender Ansatz ist das *Fuzzy*-Clustering. *Fuzzy* entpricht vom Prinzip k-Means, 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.

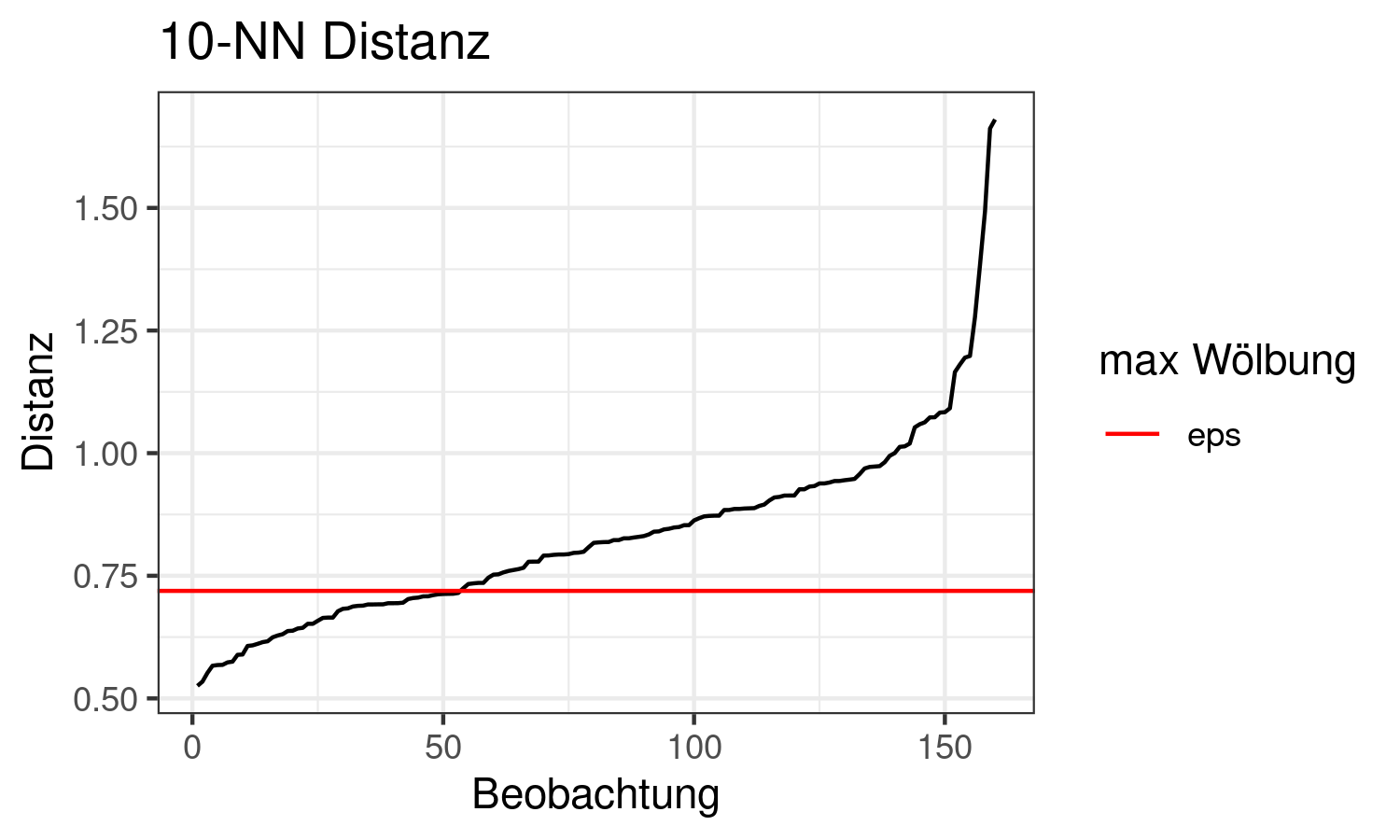
Allerdings ist dieses Verfahren hier auch nicht optimal, da, neben dem, dass *Fuzzy* üblicherweise rechentechnisch sehr teuer implementiert ist, kommt allerdings hinzu, dass dieses Verfahren auf einem Mittelpunkt pro Cluster beruht und die Distanz dazu anhand eines gegebenen Distanzmaß’ bestimmt wird. 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.

#### 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 verhinderen, dass die Cluster gar nicht oder zu gering wachsen. Allerdings führt aber zu großer 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.

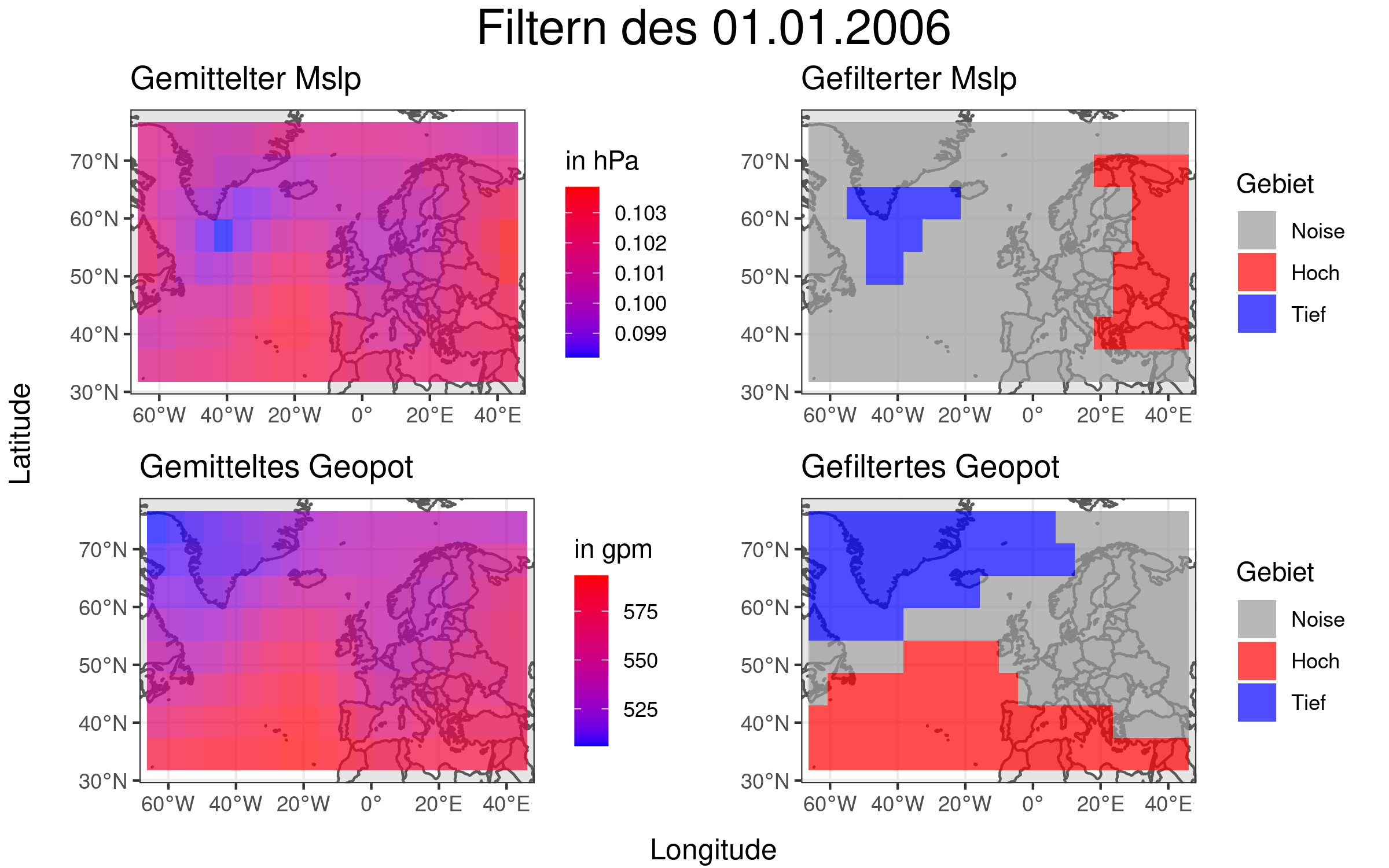
Der Nachbarschaftsparameter wird hier für jede Tag-Parameter Kombination berechnet, indem der Punkt der größten Wölbung eines kNN-Distanzplot berechnet wird. Dieser visualisiert die Distanzen von Beobachtungen zu seinen Nachbarn. wird hier als 10 festgesetzt, mit dem Ziel, am Ende Gebiete der Mindestgröße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



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 lieferte.

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



Beispiel der Gebietszuteilung eines Tages durch SCAPOI

#### 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 mieinander verglichen werden können. Der *Rand-Index* präsentiert eine solche Möglichkeit. 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. Zu dem sind natürlich 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 mit 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 eine mögliche Ausprägung der Gebiete als Teilmengen voneinander ebenfalls nicht eine Distanz von 0 zuzuweisen.

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

#### Ergebnisse

## Clustern mit extrahierten Daten

Ein weiterer Ansatz ist, dass Informationen aus dem Reanalyse Datensatz extrahiert werden und diese dann Variablen eines neuen Datensatzes werden. Diese Idee knüpft an den des vorher beschriebenen Ansatz des Filterns an. Auch hier ist die Lage von Extremwerten von Interesse. Zwar wird hier die Form der jeweiligen Tief- und Hochdruckgebiete vernachlässigt, aber es lassen sich weitere interessante Informationen mit hinzunehmen.

Dabei versteht man unter interessanten Informationen unter anderem die Verteilung der Parameter Luftdruck und Geopotential sowie das Einbeziehen der räumlichen Ebene. Das bedeutet, der neue Datensatz beinhaltet zwei Kategorien, die Verteilung der Parameter, die bestimmte Messwerte pro Tag enthält und die räumliche Ebene, die diese dann örtlich einordnet.

Wir erhoffen uns von dieser Methodik, dass die Dimensionen weiter reduziert werden können und dass wichtige Größen spezifisch gewichtet werden können.

### Extrahieren der Variablen

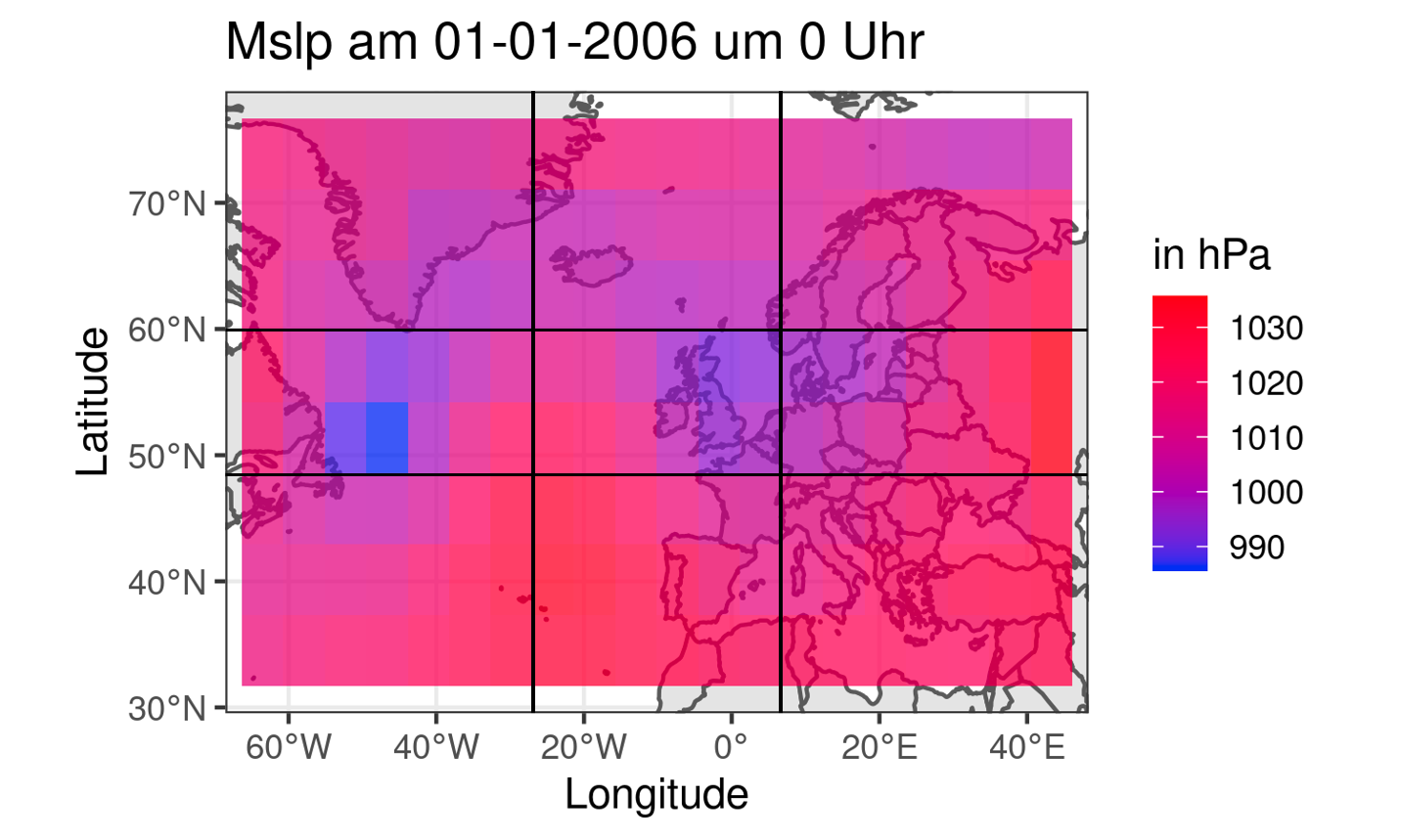
Die Ausgangslage beim Extrahieren der Variablen ist dabei größtenteils der Datensatz mit 320 Dimensionen, also der, bei dem die vier Messzeitpunkte für jeden Tag gemittelt wurden. Davon extrahieren wir verschiedene Größen, die jeweils eine für uns interessante Variable über alle Standorte zusammengefasst 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, dadurch 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. Dadurch lassen sich die Tief- und Hochgebiete am Tag miteinander vergleichen. 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 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 jetzt 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 Bild sehen kann.



Aufteilung 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.

. 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 $Zeile\_{maxMslp, i} \in \{1, 2, 3\}\ und\\ Spalte\_{maxMslp, i} \in \{1, 2, 3\}\ mit\ i = 1, ... , 10958$.

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.

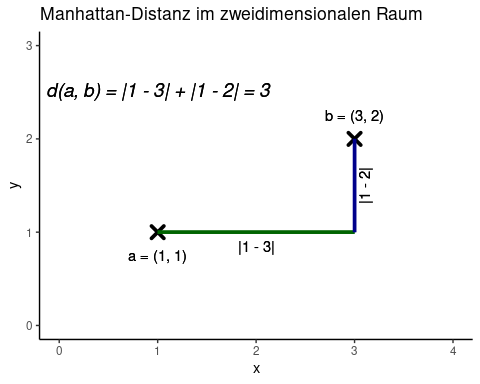
Zuletzt werden bei der räumlichen Ebene für beide Parameter die Mittelwerte in allen 9 Quadranten angegeben werden. Das sind somit 18 weitere Variablen.

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

### Distanzmetrik

Cluster können gefunden werden, indem die Distanzen von allen Beobachtungen in einem Datensatz miteinander verglichen werden. . Dafür muss ein Distanz- oder Ähnlichkeitsmaß festgelegt werden oder neu implementiert werden. Die Wahl für diese Clusteranalyse fiel auf die Manhattan-Distanz, die für zwei Beobachtungen *a* und *b* definiert ist als

Die Manhattan Distanz wird in Abbildung für einen zweidimensionalen Fall, also , beispielhaft veranschaulicht.



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. Außerdem wurde die Mahalanobis-Distanz in Betracht gezogen. Diese berücksichtigt Korrelationen zwischen verschiedenen Variablen. Der extrahierte Datensatz weist zum Teil hohe Korrelationen auf, z.B. liegt der Wert des Korrelationskoeffizienten nach Pearson für den Mittelwert und Median des Luftdrucks bei ungefähr 0.9.

#### Skalierung

Die Variablen haben verschiedene Skalen. Werte des Luftdrucks liegen beispielsweise immer zwischen und die des Geopotentials zwischen . 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. . Dabei wird für jede Beobachtung einer Variable i der Erwartungswert dieser Variablen abgezogen und schließlich durch die Standardabweichung der Variable dividiert. Dies wird für alle im Datensatz enthaltenen, also 48 Variablen durchgeführt. Nach der Skalierung sind die resultierenden Werte vergleichbar, da sie die gleichen Skalen haben und somit sind auch die Distanzen zwischen Beobachtungen vergleichbar.

#### 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 Kategorien summieren sich jeweils auf eins und die Variablen werden innerhalb einer Kategorie gleich gewichtet.

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

### Clusterverfahren

Um eine Clusteranalyse durchzuführen, gibt es viele verschiedene Methodik bzw. Algorithmen. 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 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 Brian 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.

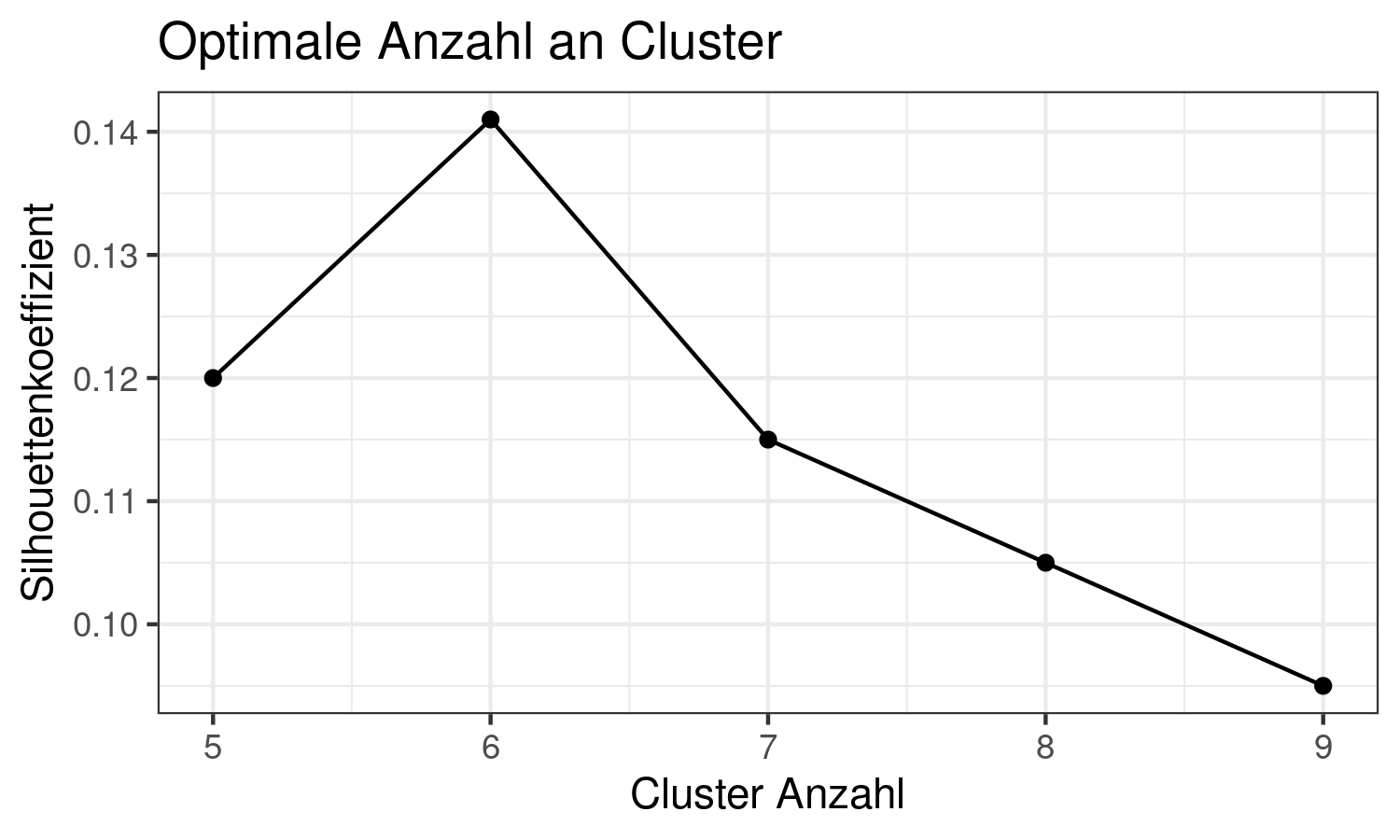
#### Partitioning Around Medoids

Bei Partitioning Around Medoids (PAM) wird angegeben, wie viele Cluster gebildet werden sollen. Der Algorithmus sucht dann anfänglich 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 der Distanzen von allen Datenpunkten zu deren jeweiligen Medoids. Anhand dieses Gütekriteriums werden die nicht-repräsentativen Datenpunkte dem Cluster zugefügt, das die Summe am stärksten minimiert. Dann werden zufällig ein weiterer Datenpunkt gewählt und es werden die Kosten berechnet, die entstehen, wenn man diesen mit den repräsentativen Punkten tauschen würde. Dabei sind die Kosten definiert als die Veränderung des Gütekriteriums , wenn die 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 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.

#### Wahl der Clusteranzahl

Die Wahl der Clusteranzahl muss vor dem Ausführen des Algorithmus statt finden. Aber um die optimale Anzahl an Clustern zu finden, wird der Silhouettenkoeffizient für verschiedene betrachtet. Der Silhouettenkoeffizient ist wie in Abschnitt 2.1 beschrieben, unabhägig von der Clusteranzahl. Daher kann er für die Wahl der Clusteranzahl heran gezogen werden. In Abbildung sieht man die Silhouettenkoeffizienten für . Den größten Wert bekommt man mit einer Clusteranzahl . Somit wird das resultierende Clusterergebnis mit dem Algorithmus PAM, der Distanzmetrik Manhattan und 6 Clustern gebildet.



Optimale Anzahl an Cluster, hier .

## Weitere Versuche

### Clara, K-Means und PCA

Der zur Verfügung gestellte Reanalyse Datensatz beträgt ursprünglich 1280 Dimensionen, da für jeden Tag zu vier Uhrzeiten an 160 Standorten zwei Parameter, Luftdruck und Geopotential, gemessen wurden. Da diese Anzahl an Dimensionen rechentechnisch sehr groß ist, wird, wie in Abschnitt 1.2 beschrieben, der Mittelwert für beide Parameter über die 4 Uhrzeiten betrachtet, sodass die Dimensionen auf 320 reduziert werden. Um den Rechenaufwand noch weiter einzuschränken, werden im Folgenden ausschließlich die Jahre 1971 bis 2000 betrachtet.

Anfangs wurden verschiedene Clusterversuche mit diesem Datensatz durchgeführt. **CLARA**

**Mahalanobis**

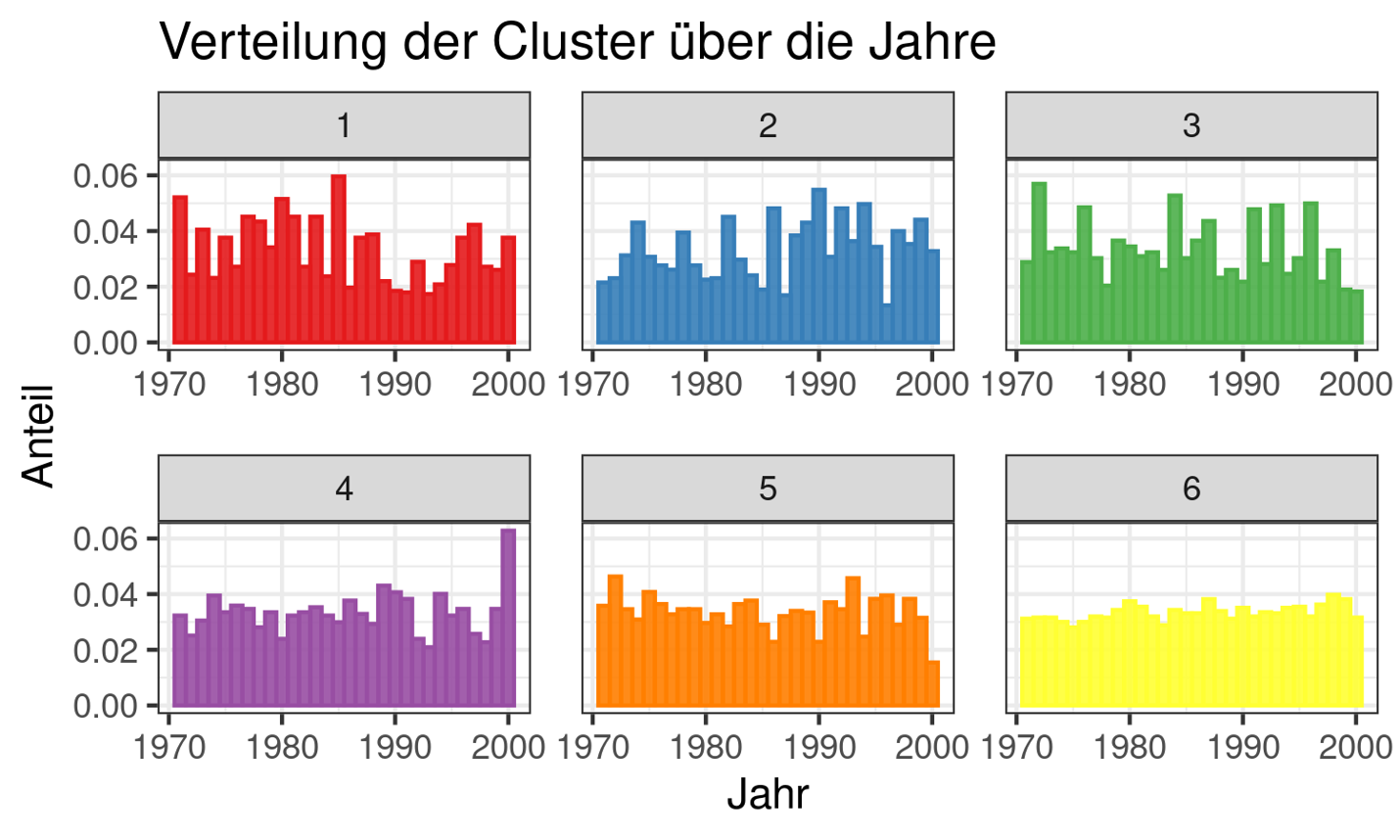
**PCA**

# Ergebnisse

Dieser Abschnitt präsentiert die Ergebnisse der deskriptiven Analyse der finalen Clusterlösung. Hierbei wird in Abschnitt 3.1 betrachtet, wie die Cluster eins bis sechs über die Jahre 1971 bis 2010 verteilt sind. Zudem wird das Verhältnis von Sommer- und Wintertagen in den einzelnen Clustern betrachtet. Abschnitt 3.2 befasst sich mit den Ähnlichkeiten und Unterschieden der extrahierten 48 Variablen in den Clustern. Abschließend wird in Abschnitt 3.3 die Clusterlösung mit der GWL Einteilung nach Hess und Brezowsky verglichen, also in welchem Ausmaß die GWLs über die Cluster verteilt sind.

## Verteilung der Cluster über die Zeit

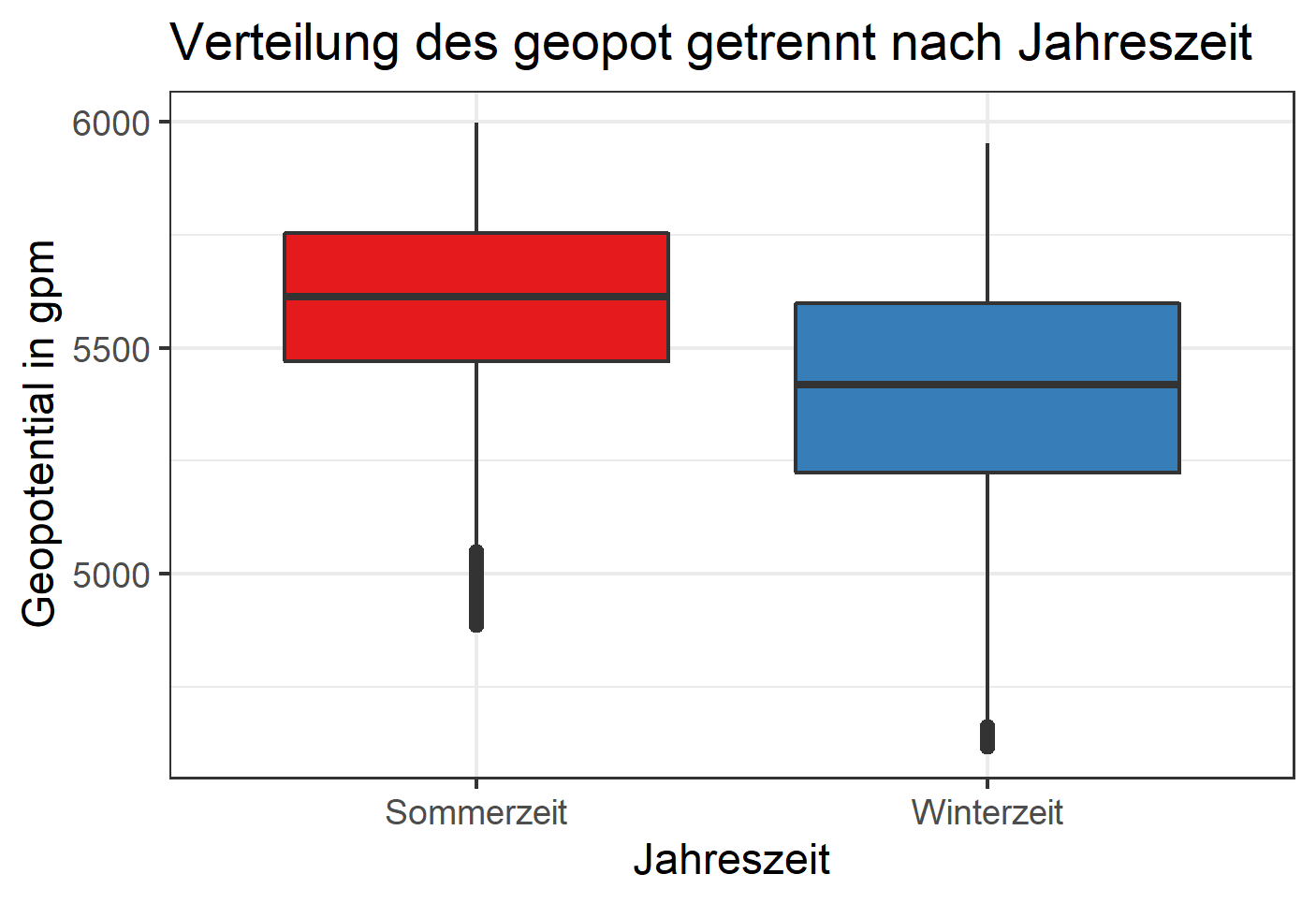
Abb. stellt dar, wie häufig jedes Cluster im Zeitraum 1971 bis 2000 vorkommt. Zu erkennen ist, dass sich ein Cluster nicht auf eine bestimmte Zeitperiode beschränkt, sondern über den gesamten Zeitraum erstreckt. Die Cluster 4,5 und 6 sind hierbei gleichmäßiger über den Zeitraum von 1971 bis 2000 verteilt 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 vier bis sechs 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.



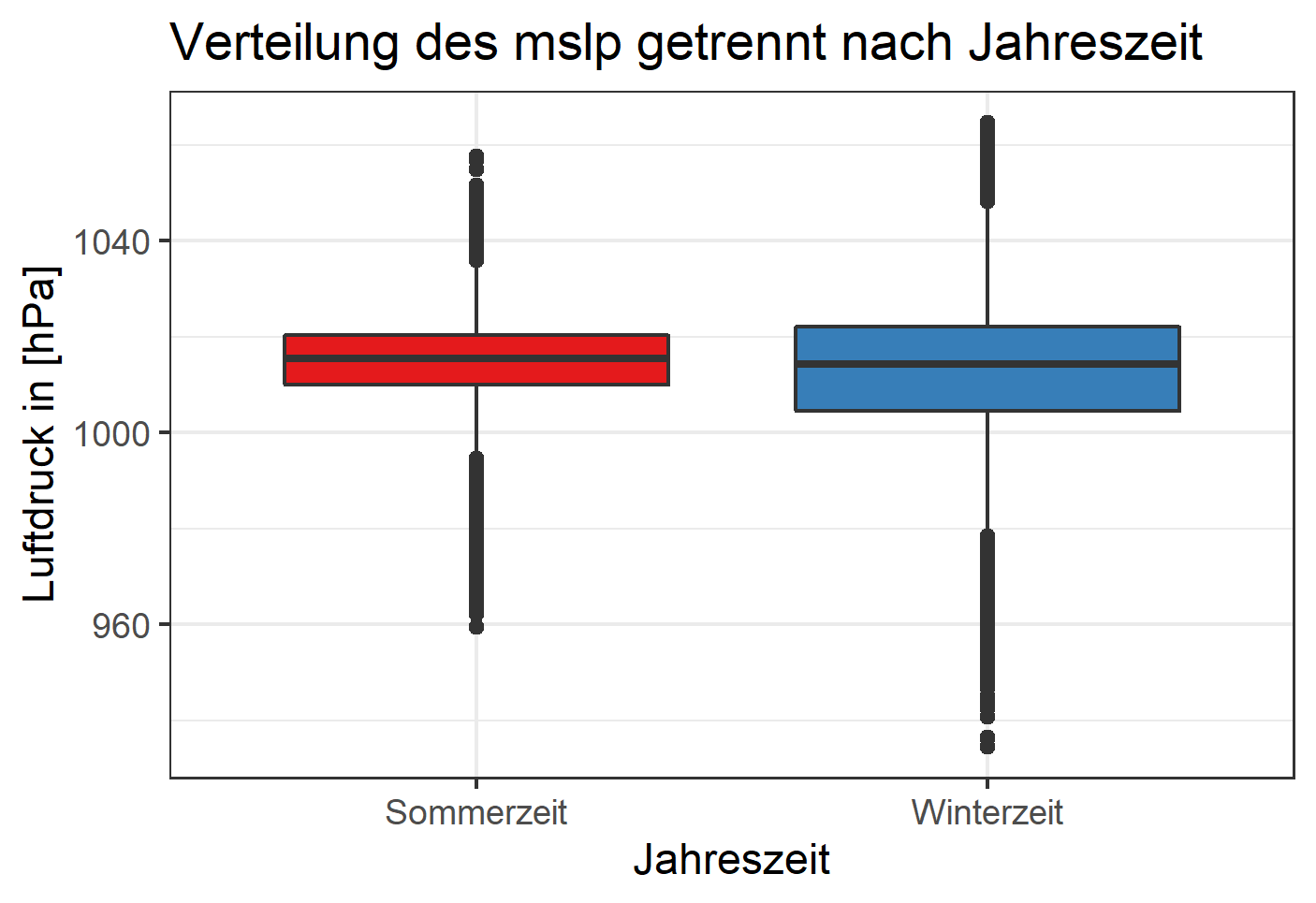
Darstellung der Häufigkeit, wie viele Tage pro Jahr im Cluster 1 bis 6 enthalten sind

## Verhältnis von Sommer- und Wintertagen in den Clustern

Die Werte des Luftdrucks im Winter streuen mehr als im Sommer (Abb. ) und die Werte des Geopotentials im Sommer höher sind als im Winter (Abb. ) 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. Diese Aufteilung des Jahres erfoglt nach Vorlage der Publikation von James P.M. 2006. Alle Tage im Zeitraum 16. Oktober bis 15. April werden als Wintertage definiert, die restlichen Tage folglich als Sommertage.(James P.M., 2006)

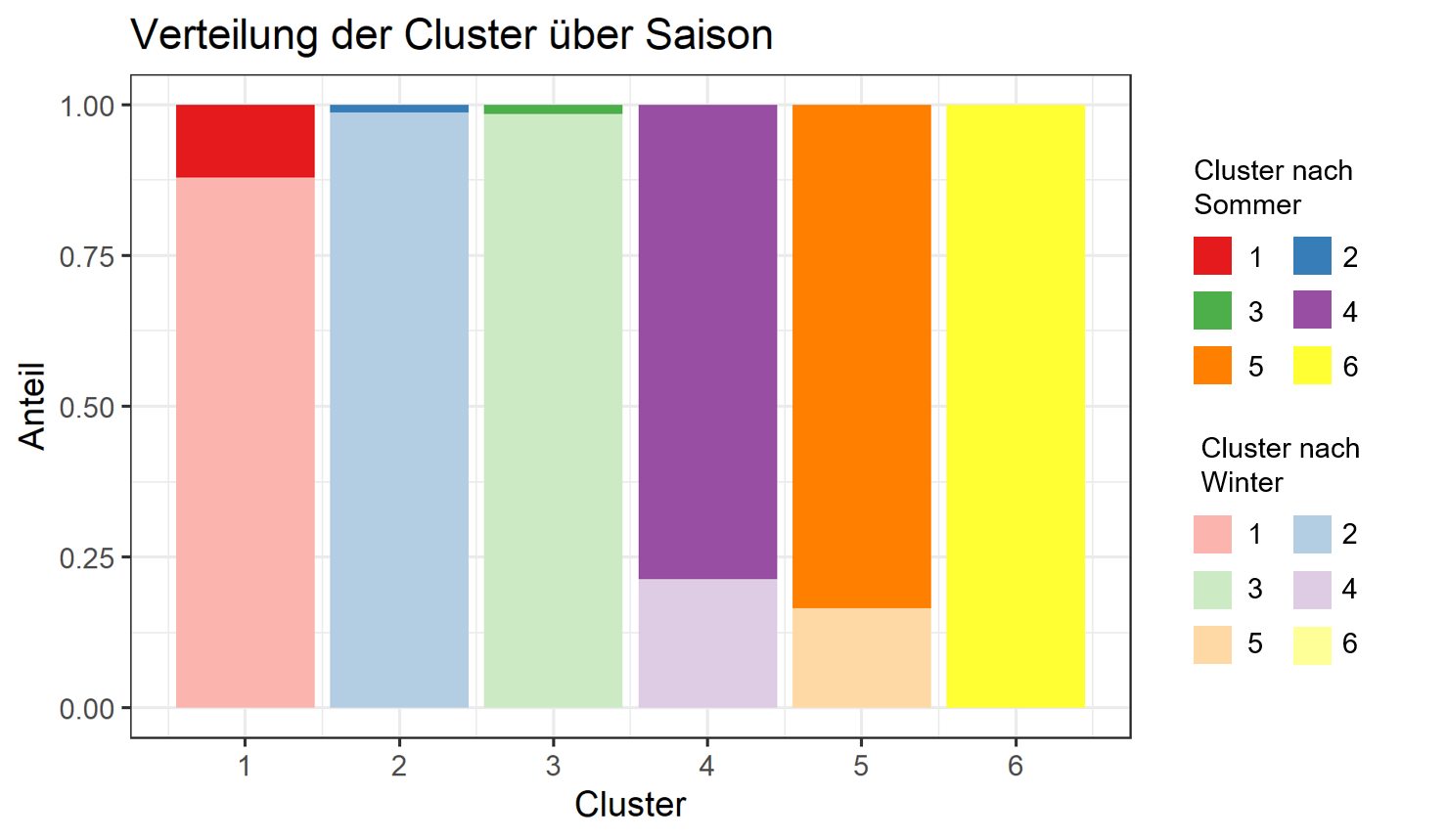


Darstellung der Verteilung des Luftdrucks über die 160 Messpunkte aufgeteilt in Sommer- und Winterzeit im Zeitraum 1971 - 2000



Darstellung der Verteilung des Geopotentials über die 160 Messpunkte aufgeteilt in Sommer- und Winterzeit im Zeitraum 1971 - 2000

Die Graphik in Abb. visualisiert die relativen Häufigkeiten der Winter- und Sommertage in jedem Cluster. Cluster 1 bis 3 enthält 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.

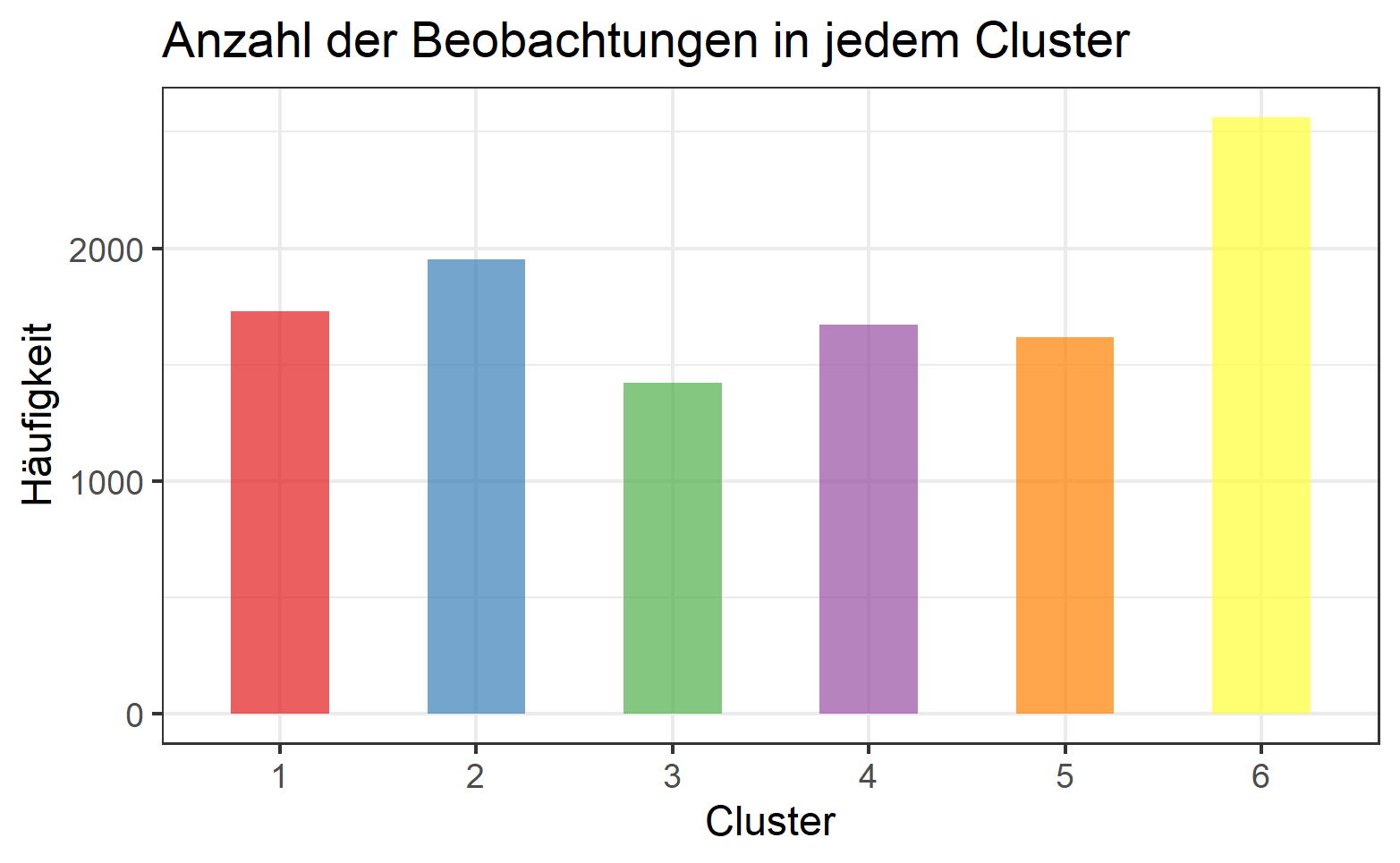


Vorkommen von Winter- und Sommertagen in jedem Cluster

## Unterschiede und Ähnlichkeiten in den Clustern

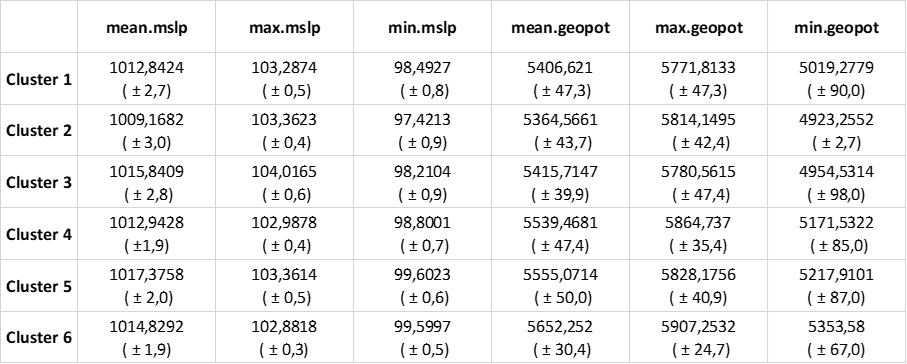
Dieser Abschnitt befasst sich mit der Fragestellung, wie sich die Werte der 49 extrahierten Variablen, mit denen geclustert wurde, zwischen den Clustern unterscheiden. Dazu werden Mittelwerte, Standardabweichung und Verteilungen ausgewählter, repräsentativer Variablen und dessen räumliche Verteilung über die 160 Messorte betrachtet.

Abb. 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.



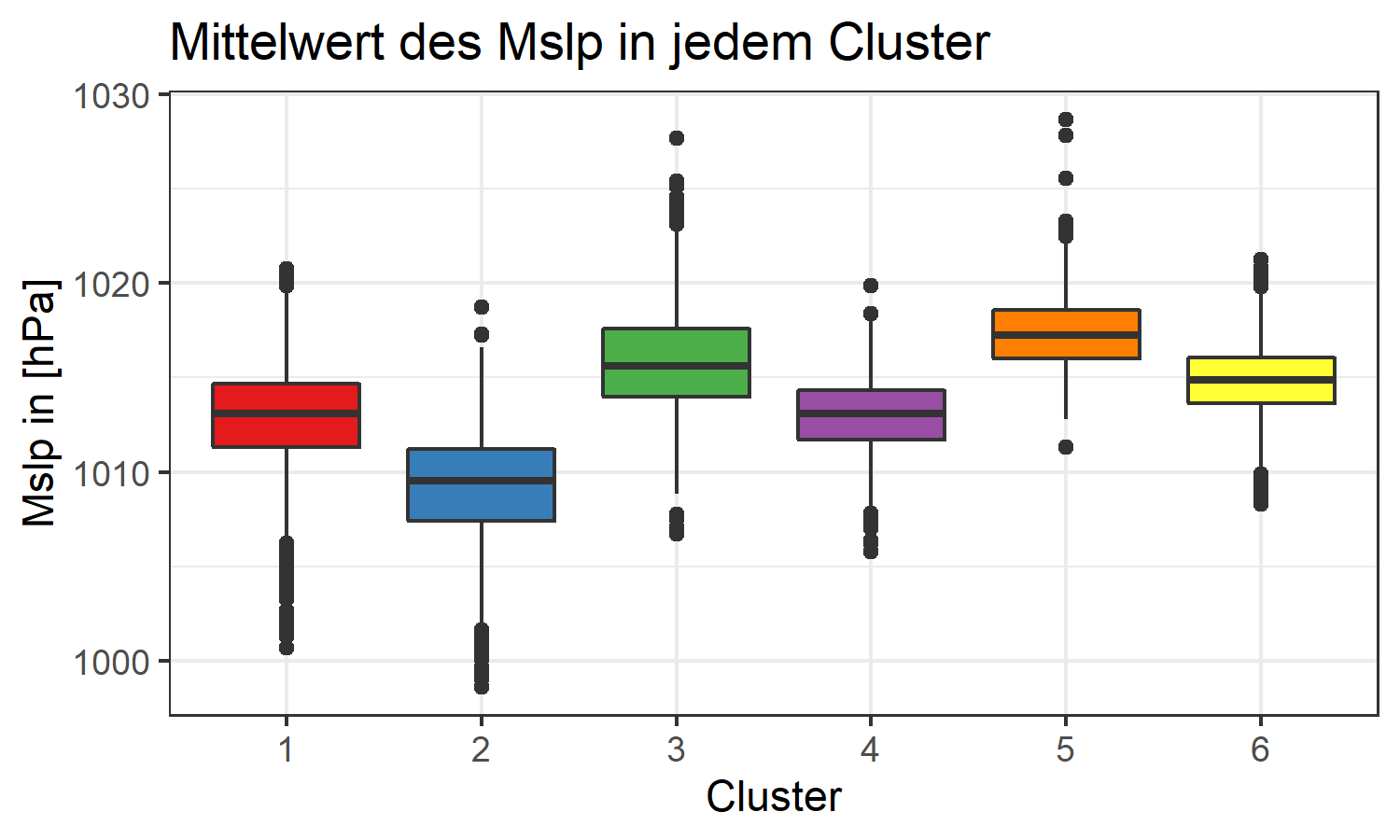
Verteilung der Tage im Zeitraum 1971 - 2000 auf die Cluster 1 bis 6

Tabelle bildet Mittelwert und Standardabweichung der Variablen mean.mslp, max, mslp, mean.geopot, max.geopt und min.geopot in jedem Cluster ab. Hierbei unterschieden sich die Werte einer Variablen in jedem Cluster nur gering voneinander. Zum Beispiel beträgt die maximale Abweichung der Variable mean.mslp in allen Clustern nur 8,2 hPa und die maximale Abweichung der Variable mean.geopot 287,7 gpm.

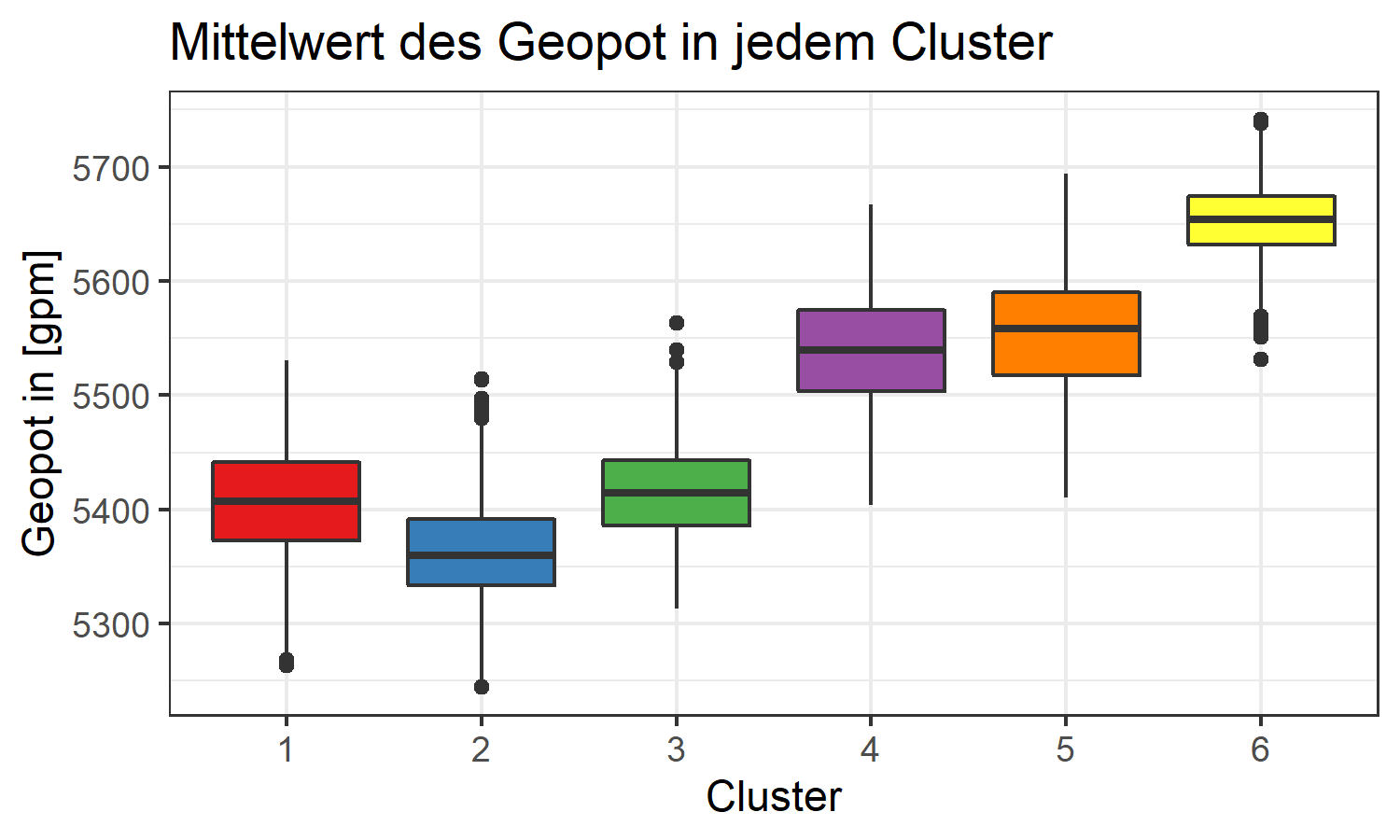


Mittelwert und Standardabweichung ausgewählter Variablen pro Cluster. Alle Werte sind in der Einheit hPa für die Variablen mslp und gpm für die Variablen geopot

Abb. und Abb. bilden die Verteilung der Variablen mean.mslp und mean.geopot je Cluster ab. Verglichen mit der Variable mean.geopot unterschiedet sich der Median und die Verteilung von mean.mslp in jedem Cluster wenig. Hierbei beinhaltet Cluster 2 mit einem Median von 1009,5 hPa und einem Interquartilsabstand(IQR) von 3,8 hPa (1. Quartil = 1007,4 3.Quartil = 1011,2 hPa) tendenziell die kleinsten Werte auf, während Cluster 5 mit einem Median von 1017 hPa und einem Interquartilsabstand von 3 hPa (1.Quartil = 1016 hPa, 3. Quartil = 1019 hPa) tendenziell die höchsten Werte aufweist. Bei der Variable mean.geopot 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 mean.mslp beinhaltet auch das Cluster 2 bei der Variable mean.geopot tendenziell die niedrigsten Werte mit einem Median von 5360 gpm und einem IQR von 58 gpm (1.Quartil = 5334 gpm, 3. Quartil = 5392 gpm). Cluster 6 weist tendenziell die höchsten Werte mit einem Median von 5654 gpm und einem IQR von 42 gpm (1.Quartil = 5632 gpm , 3. Quartil = 5674 gpm). Die Variable mean.geopot weist mehr Ausreißer auf als die Variable mean.mslp.



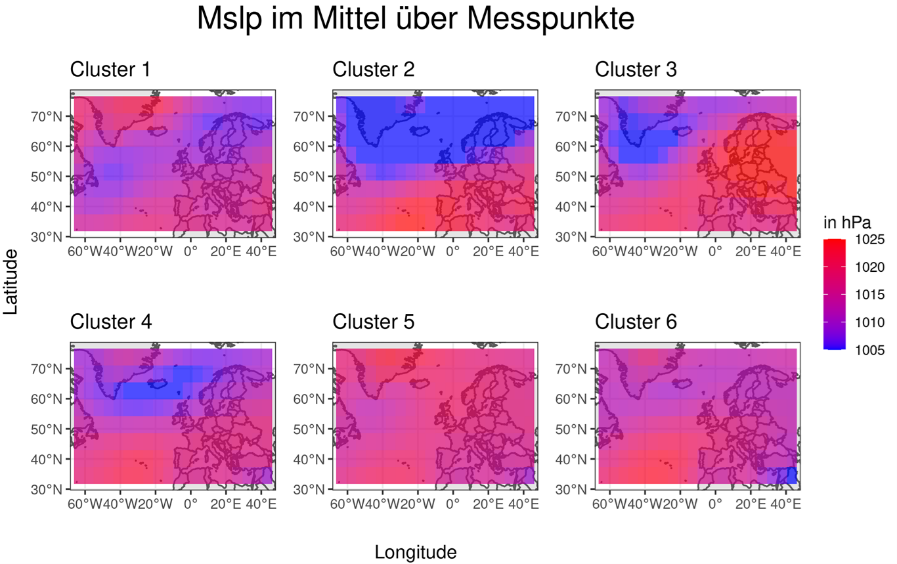
Verteilung der Variable mean.mslp in jedem Cluster



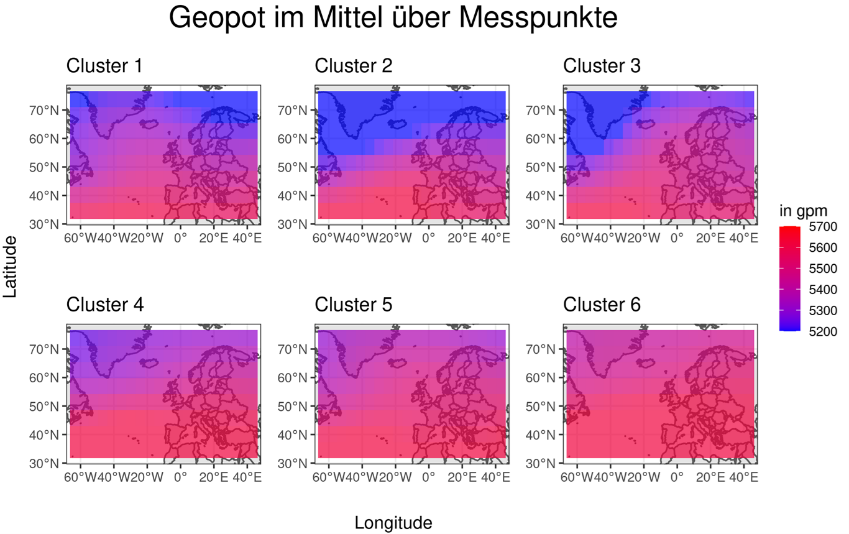
Verteilung der Variable mean.geopot in jedem Cluster

Alle anderen extrahierten Variablen, die in die Clusteranalyse miteingegangen sind und die die Verteilung beschreiben, ähneln den beschriebenen Ergebnissen von mean.mslp und mean. geopot. Die Verteilungsvariablen des Luftdrucks sind immer höher als die Verteilungsvariablen des Geopotentials, so dass mean. mslp und mean. geopot representativ für die übrigen Verteilungsvariablen.

Nun wird betrachtet, wie sich der Mittelwert des Luftdrucks und der Mittelwert des Geopotentials räumlich unterscheidet. Abb bzw. Abb. stellt pro Cluster und für jeden der insgesamt 160 Standorte das arithmetrische Mittel der mslp bzw. geoppot Werte im Zeitraums 1971 – 2000 dar. Blau Flächen visualisieren Gebiete mit niedrigerenn, rote Flächen visualisieren Gebiete mit höheren Werten.



Räumliche Verteilung des gemittelten Luftdrucks über 30 Jahre an 160 Standorten je Cluster



Räumliche Verteilung des gemittelten Geopotentials über 30 Jahre an 160 Standorten je Cluster

Die räumliche Verteilung der gemittelten Werte des Luftdrucks ist pro Cluster unterschiedlich. In Cluster 1 befinden sich nordwestlich sehr hohe Werte während in Cluster 2 ähnlich hohe Werte des Luftdrucks im Süden zu finden sind. Zudem sind dort die hohen Werte über eine größere Fläche verteilt als in Cluster 1. In Cluster 3 ist hingegen das Gebiet mit hohen Luftdruckwerten östlich gelegen, in Cluster 4 und 6 südwestlich. Cluster 5 und 6 sind gekennzeichnet von tendenziell hohen Luftdruckwerten, ein ausgeprägtes Gebiet mit tiefen Luftdruckwerten ist in Cluster 5 nicht enthalten, und in Cluster 6 nur südöstlich. Im Gegensatz dazu befindet sich in Cluster 2 die größte Fläche mit niedrigen Luftdruckwerten, das nördlich gelegen ist. Insgesamt hat jedes Cluster eine eigene, charakteristische Verteilung der Luftdruckwerte.

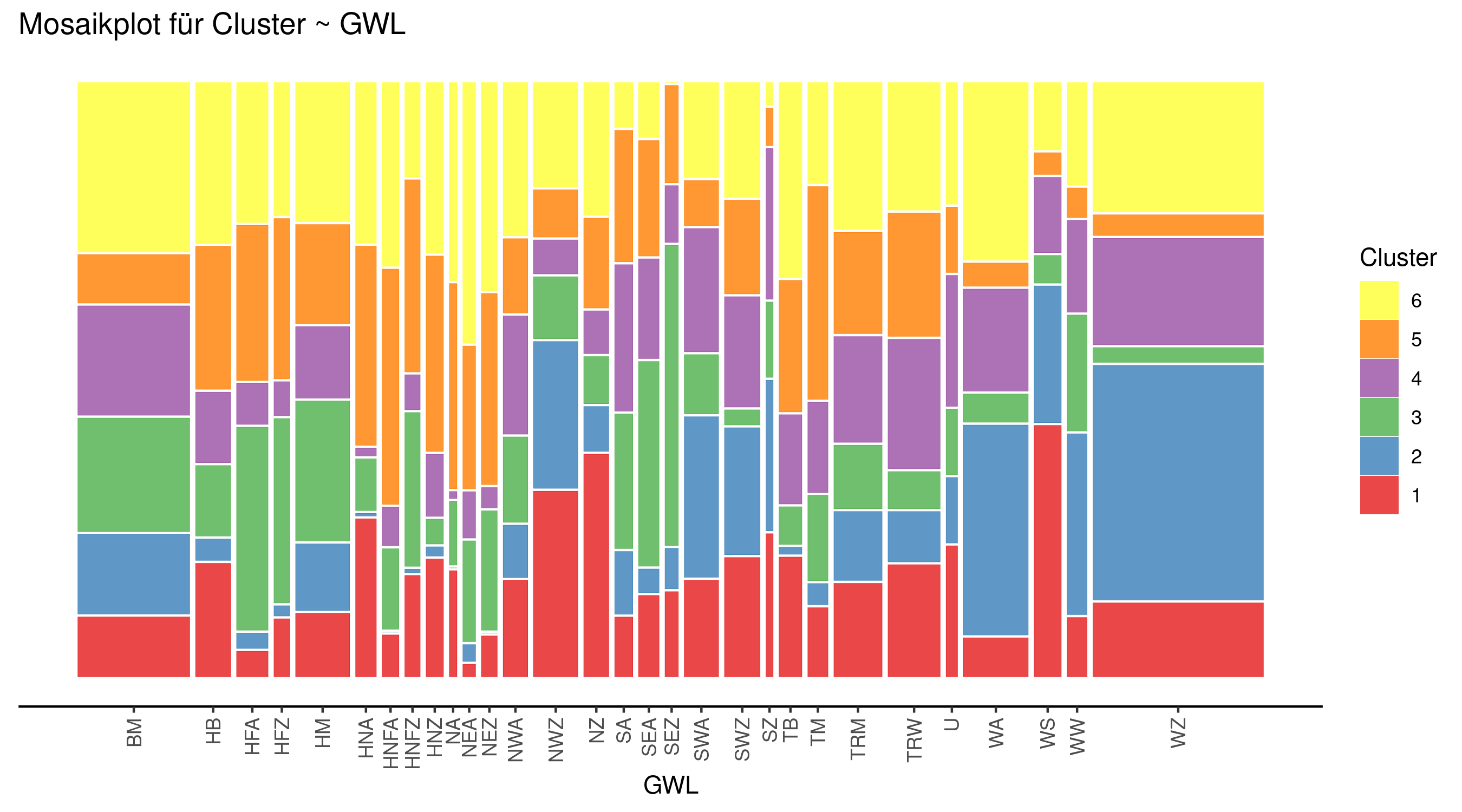
Bei der räumlichen Verteilung der Werte des Geopotentials aufgeteilt nach Cluster ergibt sich hingegen ein anderes Bild im Vergleich zum Luftdruck. Tendenziell befinden sich in jedem Cluster nördlich, bzw. nordwestlich niedrigere Werte , südlich, bzw. südwestlich sind höhere Werte. In den Clustern 1 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 Varaible 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.

## Vergleich der Clusterlösung mit den GWL

Der Mosaikplot in Abb. stellt die Aufteilung der GWLs auf die 6 Cluster dar. Zudem wird gezeigt, wie häufig eine bestimmte GWL in den 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 GWLs verteilen sich hierbei gleichmäßig auf die Cluster. Dennoch gibt es auch GWLs, 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.

Zudem wechseln in manchen Fällen die GWLs und die Zugehörigkeit der Tage, die zu einem Cluster gehören, am selben Tag.



Mosaikplot, der darstellt, mit welchem Anteil die GWLs auf die Cluster 1 bis 6 aufgeteilt sind

## Diskussion

### 3er gwl tage entfernen und anschauen

Ein auffälliger Unterschied in der Art der Unterteilung der Tage zwischen den Großwetterlagen von ist die Anforderung einer GWL mindestens 3 Tage lang zu sein. Man könnte naive 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 die GWL, die länger als 3 Tage andauern, über die Cluster verteilt werden.

### wie sind GWL&Saison kombos in den Clustern verteilt?

### weitere ergebnis ansaetze

### Saison und einzenld betrachtet

#### pam + filter

### Geopot raeumlich importance

### Gewichtungsvektor wahl

### Ausblick

#### variablen

#### zeitliche struktur

# Schluss

Brian Everitt, Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R*. Springer, New York.

“DBSCAN Clustering.” n.d. <https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556>.

“Großwetterlagen übersicht.” n.d. <http://www.sklima.de/wetterlagen_uebersicht.php>.

Hellbrück, R. 2016. *Angewandte Statistik Mit R: Eine Einführung Für ökonomen Und Sozialwissenschaftler*. Springer Gabler, Wiesbaden.

Peter C. Werner, Friedrich-Wilhelm 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.”

X. Jin, J. Han. 2017. “K-Medoids Clustering.”