VERGLEICH UND EVALUIERUNG VERSCHIEDENER CLUSTERING ALGORITHMEN UND METHODEN ZUR ANWENDUNG AUF WETTERDATEN ZUM DEFINIEREN VON WETTEREREIGNISPROFILEN UND DEREN CHARAKTERISTIKEN

Julian Lukas Ruben Erath
Duale Hochschule
Baden-Württemberg, Stuttgart
Wirtschaftsinformatik
Paulinenstraße 50
70178 Stuttgart
E-Mail: julian.erath@ibm.com

ABSTRACT

In der vorgelegten Arbeit werden Clusteranalysen an Wetterdaten aus Ontario, Kanada durchgeführt, um Profile für verschiedene Wetterereignisse zu erstellen. Dabei werden verschiedene Clustering-Algorithmen, darunter KMeans, hierarchisches agglomeratives Clustering, Gauß'sche Mischmodelle und DBSCAN, mit Hilfe der Forschungsmethode Design Science Research (DSR) untersucht und angewandt. Das Ziel ist die Identifizierung der am besten geeigneten Methode für die Definition von Wetterereignisprofilen anhand deren Parameter sowie die Optimierung der Analysen und Methoden, um genaueste Ergebnisse zu erzielen. Die Qualität der Ergebnisse wird gemessen anhand der Genauigkeit, Leistungsfähigkeit und den vorgestellten Evaluationsmetriken. Die Wetterereignisprofile (WEP) könnten in der Wettervorhersage zur Bestimmung von verschiedenen Arten von Wetterereignissen und zur Anzeige relevanter Informationen in Dashboards verwendet werden. Ebenso können Modelle für WEP zur Anomaliedetektion entwickelt werden. Die Daten stammen aus sieben Jahren historischer Wetterdaten von IBM Deutschland GmbH und haben das Potenzial für zukünftige Forschung, insbesondere im Bereich der Identifizierung hochkorrelierter Wettermerkmale und der Auswahl relevanter Merkmale.

In the work presented clustering analyses are performed on weather data from Ontario, Canada to generate profiles for various weather events. Different clustering algorithms, including KMeans, hierarchical agglomerative clustering, Gaussian mixed models, and DBSCAN, are investigated and applied using the Design Science Research (DSR) research method. The goal is to identify the most appropriate method for defining weather event profiles based on their parameters and to optimize the analyses and methods to obtain the most accurate results. The quality of the results is measured by the accuracy, performance and evaluation metrics presented. Weather event profiles (WEP) could be used in weather forecasting to determine different types of weather events and to display relevant information in dashboards. Likewise, models for WEP can be developed for anomaly detection. The data comes from seven years of historical weather data from IBM Deutschland GmbH and has the potential for future research, especially in the area of identifying highly correlated weather features and selecting relevant features.

SCHLÜSSELWÖRTER

Clusteranalyse, Clustermodelle, Wetterdaten, Meteorologie, Wetterdaten Gruppierung, Wetterereignisprofile

1. EINFÜHRUNG

1.1. Motivation

Das Wetter nimmt in vielen Aspekten des menschlichen Lebens eine zentrale Rolle ein. Seit dem Beginn zivilisierten Lebens streben Menschen danach, die Abläufe und Muster von Wetterveränderungen zu verstehen und vorherzusagen. Fortgeschrittene und intelligente Analysetechniken helfen den Menschen dabei, die Auswirkungen des Wetters typisieren und vorhersagen zu können. Im Zeitalter der Informationstechnologien entsteht nun ein Trend, Machine Learning (ML) Techniken und automatisierte Big Data Analysen auf Wetterdaten anzuwenden, um das Wetter noch effektiver kategorisieren, analysieren und vorhersagen zu können (Grabbe et al. 2014). In der Entdeckung von Mustern und Abläufen im Wetter wurden bereits große Fortschritte erzielt, wie in Fathi u. a. 2021 systematisch erörtert wurde. Es gibt jedoch eine Vielzahl von Wetterbedingungen, welche die genaue Analyse, Kategorisierung von Wetterereignissen und deren Vorhersage erschweren. Daher ist es von Relevanz, den Einsatz von ML-Techniken im Bereich der Meteorologie genauer zu untersuchen.

1.2. Problemstellung

Die bisher zumeist eingesetzten Methoden zur Wetteranalyse beruhen auf physikalischen Gleichungen, wie den numerischen Wettervorhersagemodellen. Der Einsatz dieser Modelle ist lediglich für eine genaue Vorhersage des Wetters von bis zu zwei Wochen bewährt. Über diesen Zeitraum hinaus werden physikalische Modelle aber sehr ungenau. Zudem ist deren Berechnung sehr komplex und erfordert viel Rechenleistung (Holmstrom et al. 2016). Darüber hinaus können sie aufgrund ihrer begrenzten räumlichen und zeitlichen Auflösung bestimmte Wetterereignisse wie lokale Gewitter, Stürme oder andere Wetterkatastrophen nicht genau kategorisieren und vorhersagen (de Lima et al. 2013; Ferstl et al. 2017). Das enorme Volumen an meteorologischen Daten macht auch eine manuelle, vollumfängliche Analyse und Kategorisierung durch Meteorologen unmöglich und rentiert sich aus betriebswirtschaftlicher Sicht nicht (de Lima et al. 2013). Daher setzen Wissenschaftler und Data Scientisten vermehrt ML und Big Data Analysetechniken zur Analyse von Wetterdaten ein. Diese haben den Vorteil, dass sie teil- oder vollautomatische Analysen von Wetterdaten ermöglichen und das Klassifizieren und Auswerten von Wetterdaten, übernehmen können (Wei Fang et al. 2014). Für viele Anwendungsfälle werden klassifizierte Wetterdaten benötigt. Daher wird untersucht, wie die Kategorisierung von Wetterdaten mittels unüberwachten ML-Ansätzen teil- oder auch vollautomatisiert werden kann. Hier muss zunächst genauer untersucht werden, aus welchen Parametern sich solche WEP zusammensetzen und wie diese genau definiert werden können. Der Einsatz von ML-Ansätzen hat auch den Vorteil, dass durch diese Zusammenhänge in den Wetterdaten gefunden werden können, die einem Menschen nicht auffallen würden. Dies kann sowohl für die Kategorisierung der Wetterdaten als auch für eine weitere Auswertung genutzt werden (Pooja et al. 2020).
Das Thema dieser Thesis ist vor allem im Hinblick auf die momentane Entwicklung des Klimawandels interessant. Mithilfe von Big Data und ML-Methoden analysierte Klimadaten können insofern effektiver ausgewertet werden, da sie im Sinne der Forschung bezüglich des Klimawandels dokumentiert und ihr zeitlicher Trend analysiert werden kann. Interessant ist insbesondere die Betrachtung zur Entwicklung der definierten WEP und deren Merkmale, gerade hinsichtlich extremer Wetterereignisse.
Bisher existiert in der Literatur keine klare Definition von WEP durch Clustergruppen. Mit dieser Problemstellung befasst sich die vorgelegte Arbeit. Im Praxisteil dieser Arbeit sollen Wetterdaten geclustert und ähnliche Wetterereignisse in Gruppen organisiert werden.

Bei der Literaturrecherche dieser Arbeit wurde deutlich, dass die Forschung in relevanten Features bei der Clusteranalyse auf Wetterdaten bereits sehr vorangeschritten ist. Die Ergebnisse bisheriger Forschung können als Basis der weitergehenden Forschung in dieser Arbeit genutzt werden. Bisherige Forschung beschäftigt sich viel mit der Auswertung der Ergebnisse von Clusteranalysen, im Spezifischen auch Clusteranalysen bezogen auf Wetterdaten, bspw. in Form von Performanz, Evaluationsmetriken und auch der Qualität der Ergebnisse. Diese Forschung wird ebenfalls als Basis für den Praxisteil dieser Arbeit genutzt. Des Weiteren ging aus der Literaturanalyse hervor, dass einige Forschende bereits erfolgreich Clusteranalysen eingesetzt haben, um Wettercluster zu generieren und auszuwerten. Die identifizierten Wettercluster werden zumeist für eine direkte weitere Analyse, bspw. zur Evaluation effektiver Maßnahmen bei bestimmten Wetterereignissen genutzt. Einige Forschende beschreiben zum Teil auch, wie sich die identifizierten Wetterereignisse zusammensetzen und beschreiben relevante Parameter, jedoch werden keine klaren WEP in größerer Menge und über eine größere geographische Fläche definiert. Auch wurde in der Forschung bereits eine Vielzahl von Clusteringalgorithmen und Methoden angewandt. Jedoch wurden bisher nicht mehrere spezifische Algorithmen ausgewählt und unter spezifischen Gesichtspunkten verglichen, um die effizientesten Algorithmen und Methoden zum Clustering von WEP zu identifizieren.

1.3. Zielsetzung

Die Zielsetzung der vorgelegten Bachelorarbeit ist die Beantwortung der folgenden Forschungsfrage: *„Welche Algorithmen und Methoden der Clusteranalyse eignen sich am besten zum Definieren von WEP, gemessen anhand der identifizierten Kriterien und welche Möglichkeiten gibt es, diese weiter zu optimieren?“*
Zur Einlösung der Zielsetzung und somit der Lösung der erkannten Problemstellungen bei der Analyse von Wetterdaten, soll der ML-Ansatz des Clusterings verwendet werden, um Wetterdaten zu gruppieren und so WEP zu erstellen. Dieser Ansatz soll dazu beitragen, die räumliche und zeitliche Variabilität des Wetters unter Einfluss einer Vielzahl von Wetterparametern zu berücksichtigen und genauere Vorhersagen zu treffen. Durch die Erstellung von WEP sollen so auch die Charakteristiken verschiedener Wetterereignisse analysiert werden können. Dazu sollen verschiedene Clustering-Algorithmen angewandt und verglichen werden. Die Arbeit wird sich auf die Identifizierung von charakteristischen Mustern und Merkmalen von WEP konzentrieren, sowie auf den Vergleich relevanter Clustering Algorithmen. Diese sollen verglichen werden, um anhand adäquater Kriterien die leistungsstärksten, effizientesten und genauesten Algorithmen zum Definieren von WEP zu evaluieren. Als Ergebnis dieser Arbeit sollen klare WEP für konkrete geographische Regionen definiert werden. Auch sollen mehrere spezifische Algorithmen ausgewählt und unter spezifischen Gesichtspunkten verglichen werden, um die effizientesten Algorithmen und Methoden zum Clustering von WEP zu identifizieren.
Die Ergebnisse dieser Arbeit tragen dazu bei, die bestehenden und zukünftigen Big Data und ML-Techniken im Bereich der Meteorologie zu verbessern und damit das tägliche Leben der Menschen effektiv zu erleichtern.
Der Beitrag der vorgelegten Arbeit zur Wissenschaft ist dabei, wie in Gregor und Hevner 2013 beschrieben, ein

Verbesserungsbeitrag, indem die genaue Anwendung der Clusteringmethoden auf Wetterdaten analysiert und evaluiert wird und die ML-Methodik in Form von positiver Steigerung der Effizienz, Produktivität und Qualität optimiert wird (Gregor und Hevner 2013).

1.4. Vorgehen und kritische Auswahl der Forschungsmethodik

Im Theorieteil der Arbeit wird eine Literaturrecherche mit Konzeptmatrix nach Webster und Watson 2002 durchgeführt. Im Praxisteil der vorgelegten Arbeit wird Design Science Research (DSR) nach Gregor und Hevner 2013 als Forschungsmethodik zur Entwicklung eines Prototyps angewandt (siehe auch: Hevner und Chatterjee 2010; Hevner und March 2004).
Die Methode des iterativen Prototypings nach Wilde und Hess 2007 soll im Rahmen des DSR erweiternd eingesetzt werden. Sie beinhaltet die schrittweise Entwicklung und Verfeinerung eines Prototyps oder Artefakts, um die Forschungsfragen zu beantworten und das Designproblem zu lösen. Es wird in dem iterativen Prozess ein Prototyp entwickelt, getestet und evaluiert, bis das Artefakt die Anforderungen erfüllt. Dieser Ansatz ermöglicht es, Designentscheidungen zu verfeinern, Fehler zu korrigieren und sicherzustellen, dass das endgültige Artefakt den praktischen Anforderungen entspricht.
Zur Evaluierung wird dabei die Kreuzvalidierung nach Shao 1993 verwendet. Diese ist eine Technik in der Statistik und ML, um die Leistung eines Modells zu bewerten und sicherzustellen, dass es generalisierbare Ergebnisse liefert. Dabei wird der Datensatz in verschiedene Teile aufgeteilt, wobei Teile für das Training und andere für die Validierung des Modells verwendet werden. Dies hilft, Overfitting zu verhindern und die Zuverlässigkeit der Modellvorhersagen zu erhöhen.
In dem Forschungsrahmen für DSR werden diese Methoden effektiv zusammen verwendet, indem das iterative Prototyping eingesetzt wird, um das technisches Artefakt zu entwickeln und kontinuierlich zu verbessern. Während dieses Prozesses werden die Daten aus den Prototyp-Tests gesammelt und zur Evaluierung und Optimierung genutzt. Die Kreuzvalidierung wird dann verwendet, um die Leistung des entwickelten Artefakts zu bewerten, indem die Genauigkeit und Vorhersagefähigkeit des entwickelten Artefakts überprüft wird. Dies trägt dazu bei, sicherzustellen, dass das Artefakt nicht nur während der Entwicklung, sondern auch in der Validierungsphase angemessen funktioniert und generalisierbare Ergebnisse liefert. Die Kombination dieser Methoden ermöglicht es, DSR in der Entwicklung und Evaluierung von technischen Artefakten fundiert und methodisch durchzuführen, wodurch die Qualität und Nützlichkeit dieser Artefakte verbessert wird.
Schlussendlich wird so ein Prototyp zum Clustering der Wetterdaten zum Definieren von WEP implementiert werden und so zur Verprobung und Evaluierung der Methodik, Parameter und Algorithmen genutzt werden. Die Forschungsrelevanz des Projektes ist dadurch gegeben, dass die vorgelegte Arbeit mit ihrer Implementierung des Artefakts und dessen Ergebnissen zur Wissensgrundlage für weitergehende Forschung und Nutzung in der Praxis beiträgt.

2. DISKUSSION DES AKTUELLEN STANDS DER FORSCHUNG

2.1. Meteorologie

In der Literatur werden verschiedene Datenquellen (wie ERA5) diskutiert. Eine Beschreibung der verschiedenen Datenquellen, relevante Institutionen und wichtige Instrumente sowie deren Vor- und Nachteile werden im Weiteren diskutiert.
Eine der wichtigsten Institutionen bei der Forschungsarbeit mit Wetterdaten ist die European Centre for Medium-Range Weather Forecasts (ECMWF). Die ECMWF ist eine unabhängige regierungsübergreifende Organisation, welche international unterstützt wird. Ihr Ziel ist es, Wettervorhersagen und Klimaprognosen von hoher Qualität zu erstellen. Hierfür definieren sie internationale Standards für Big Data Themen im Bereich der Meteorologie. Das ECMWF arbeitet auch mit nicht europäischen Nutzern und Staaten zusammen (Pelosi et al. 2020). Von besonderer Relevanz ist ERA5 („the fifth generation of atmospheric reanalysis of the global climate to be produced“) (Hersbach et al. 2020), welches die aktuelle Version des globalen Klima- und Wetter-Reanalyse-Datensatzes des ECMWF ist. ERA5 ist ein umfassendes und standardisiertes globales Klima- und Wetterdatenarchiv mit hoher zeitlicher und räumlicher Auflösung (ECMWF 2023a). ERA5 besteht aus einer Vielzahl von stündlich gemessenen atmosphärischen, land- und ozeanbezogenen Klimaparametern. Die Daten decken die Erden in einem Gitter mit einem Abstand von 30km ab und teilen die Atmosphäre in 137 Ebenen, beginnend mit der Erdoberfläche, bis zu einer Höhe von 85km ein. Qualitätsgesicherte monatliche Aktualisierungen von ERA5 (vom Jahr 1959 bis heute) werden innerhalb von 3 Monaten nach der Echtzeit vom ECMWF veröffentlicht. Vorläufige tägliche Aktualisierungen des Datensatzes stehen den Nutzern innerhalb von 5 Tagen nach der Echtzeit zur Verfügung (ECMWF 2023b). Eine genaue Auflistung der enthaltenen Parameter, detaillierte Erklärungen zu diesen und weitere Standardisierungen sind in der ECMWF Dokumentation zu ERA5 zu entnehmen (ECMWF 2023c). Aus Tabelle 2 sind die relevanten meteorologischen Parameter (zusammengetragen aus der Literatur), welche die Wettereignisse definieren, zu entnehmen.

2.2 Einsatz von Clustering in der Meteorologie für die Definition von WEP

In diesem Kapitel wird der Stand der Forschung zum Einsatz von Clustering in der Meteorologie für die Definition von WEP diskutiert. Es wird darauf eingegangen, welche Fortschritte Forschende in diesem Bereich schon gemacht haben, welche Algorithmen, Methoden, Themen und Aspekte dazu relevant sind, derzeitige Problemstel-

lungen im Bereich identifiziert und welche Aspekte dieses Themas derzeit und zukünftig noch zu bearbeiten sind.
Die Notwendigkeit ML-Techniken in der Meteorologie zu verwenden, wird auch durch Liljequist und Cehak 1984 deutlich. Sie erklären, dass in der Meteorologie die Schwierigkeit besteht, dass meteorlogische Phänomene - kontrastär zu bspw. der Experimentalphysik - nicht simuliert und befreit von störenden Einflüssen untersucht werden können. So müsse „der Meteorologe seine Beobachtungen und Messungen oft durchführen und sein Beobachtungsmaterial nachher statistisch bearbeiten“, um genug Material für seine Forschung zu erlangen. Zudem ist viel Material aufgrund von störenden Einflüssen unbrauchbar (Liljequist und Cehak 1984). Zusätzlich müssen auch genug Messungen vorhanden sein, damit die Resultate nicht von Zufälligkeiten verfälscht sind. Um diese Probleme zu lösen, kann Maschinelles Lernen, bspw. in Form einer Clusteranalyse verwendet werden, um anfallende meteorlogische Daten dauerhaft und automatisiert zu speichern und zu analysieren. Diese Daten dann durch vordefinierte WEP zu labeln, bietet für Meteorologen und Data Scientisten einen enormen Mehrwert, was die Bedeutung dieser Forschung weiter verdeutlicht.

Grabbe et al. 2014 haben den Einsatz von Clustering Techniken auf Wetterdaten bereits untersucht, um den Einfluss des Wetters auf geplante Ankunftszeiten von Flugzeugen an Flughäfen und deren Verspätung zu analysieren und um zu evaluieren, welche Maßnahmen bei bestimmten Wetterereignissen effektiv sind, um das Verkehrsflussmanagement zu optimieren. Sie haben herausgefunden, dass es möglich ist, historische Wetterdaten und Flugverkehrsarchivdaten zu clustern, um Auskunft darüber geben zu können, welche Art von Verkehrsmanagement als Reaktion auf verschiedene Wetterereignisse effektiv ist. Genutzt wurde hierfür die Clustering Technik des Expectation Maximation (EM) Algorithmus als Erweiterung des KMeans Algorithmus, um für jede Stunde ein bestimmtes Wetterereignis identifizieren zu können. Hierzu wurden Wetterdaten des Localized Aviation MOS Program (Ghirardelli 2005) genutzt. Insbesondere die Wetterfaktoren Windrichtung, Windgeschwindigkeit, Niederschlag, Niederschlagswahrscheinlichkeit, Gewitterwahrscheinlichkeit, Wahrscheinlichkeit von gefrorenem Niederschlag, Wahrscheinlichkeit für Schnee, Wolkenbildung und Sichtbarkeit wurden als die relevantesten Wetterparameter identifiziert, welche den größten Einfluss auf die Verspätung von Flugzeugen und am Flughafen und den Einsatz von Verkehrsmanagementinitiativen haben. Grabbe et al. 2014 konnten am Newark Liberty International Airport fünf Wettercluster für 2012 identifizieren. Auffällig war, dass ein einzelner Cluster (s. Tabelle 1, Cluster 2) ca. 57% aller Datenpunkte enthielt, welcher grundsätzlich unauffälliges Wetter und normale Abläufe am Flughafen enthält. Zwei Wettercluster wurden identifiziert, welche Unwetter enthielten, welche mit Verspätungen am Flughafen und den Einsatz von Verkehrsmanagementinitiativen korrelieren. Einer dieser Cluster ist charakterisiert durch erhöhte Niederschlagswahrscheinlichkeit und einer niedrigen Wolkendecke (s. Tabelle 1, Cluster 0), der andere Cluster ist charakterisiert durch schlechtes Wetter, starke Winde und viel Niederschlag. Diese Cluster machen jeweils ca. 7,5% und 3,4% aller Datenpunkte aus. Die Cluster-Ergebnisse der Arbeit sind in Tabelle 1 dargestellt. Die Autoren haben jedoch nur teilweise spezifische WEP definiert. Auch wurden nicht mehrere spezifische Clustering Algorithmen unter Variation ihrer Parameter verglichen und auf ihre Performanz und die Qualität ihrer Ergebnisse überprüft.

Tabelle 1 Clusterergebnisse von Grabbe et al. 2014: Beschreibung der Wettercluster am Chicago O'Hare International Airport 2012

Cluster Index	Number of Members	Description
0	660	Reduced ceilings and elevated probability of precipitation
1	1833	Fair daytime weather hours
2	5045	Early morning and nighttime operations
3	923	Moderate daytime weather hours
4	299	Bad weather – reduced ceilings and visibility, increased winds, increased probability of precipitation

Pooja et al. 2020 stellen eine Selektionsmethode von Features für das Clustering von Wetterdaten vor, bei der mittels Tanimoto Correlation Coefficient Ähnlichkeiten zwischen Features gefunden werden. Damit können relevante Features mit höherer Selektionsgenauigkeit für die Clusteranalyse gewählt werden. Die Korrelation zwischen Features wird berechnet, um größere Ähnlichkeiten zwischen Wetterfeatures zu finden und die Genauigkeit der Clusteranalyse zu erhöhen. Features mit kleiner Ähnlichkeit werden entfernt. Zur genauen Zusammensetzung Berechnung des Tanimoto Correlation Coefficient wird auf Pooja et al. 2020 verwiesen. Mittels des Tanimoto Correlation Coefficient ermittelte Features, welche über eine hohe Ähnlichkeit verfügen, werden dann für die Clusteranalyse ausgewählt, um eine möglichst genaue Wetteranalyse zu gewährleisten. Nach der Selektion relevanter Features wird eine Clusteranalyse ausgeführt, um Wetterdaten in Clustern zu gruppieren. Pooja et al. 2020 verwenden hierfür den EM-Algorithmus, wobei bei der Clusteranalyse die erwartete Likelihood-Wahrscheinlichkeit zwischen den Clusterzentren und den Datenpunkten berechnet wird, um mittels der Maximum a posteriori function die erwartete Wahrscheinlichkeit zum möglichst exakten Zuweisen der Datenpunkte in die richtigen Cluster zu erreichen. Die Clusterergebnisse werden anschließend durch linear program boosting classification kategorisiert.

Tabelle 2 Relevante meteorologische Parameter

Faktor	Symbol	Einheit	Messung	Berechnung	Beschreibung
Luftfeuchtigkeit / Wasserdampfkonzentration	f	g/m^3	Hygrometer	$f = \frac{m_W}{V}$	Anteil des Wasserdampfs am Gasgemisch der Luft
Temperatur	°	°C oder °K	Thermometer	°K = 273,15 + °C	Wärmegrad der Luft
Luftdruck	Pa oder bar	Pa oder mbar	Barometer / Manometer / Anemometer	1 Pa = 1 Nm^{-2} 1 mbar = 100Nm^{-2} = 1 hPa	Die auf eine Fläche wirkende Kraft
Luftdichte	ρ	kg/m^3	Barometer / Manometer	$\rho = kg/m^3 \quad \rho = \frac{p*M}{R*T}$	Masse Luft, die in einem bestimmten Volumen enthalten ist
Luftströme (Vertikal) Stärke	v ↕	km/h	Anemometer	1 = km/h	Gibt an, wie schnell sich die Luft an einem spezifischen Punkt in vertikaler Richtung vorbeibewegt
Luftströme (Horizontal) Stärke	v ↔	km/h	Anemometer	1 = km/h	Gibt an, wie schnell sich die Luft an einem spezifischen Punkt in horizontaler Richtung vorbeibewegt
Luftströme (Vertikal) Richtung			Anemometer		Gibt die genaue Richtung an, in die sich die Luft auf vertikaler Achse an einem spezifischen Punkt vorbei bewegt
Luftströme (Horizontal) Richtung	Winkel in ° oder Himmelsrichtung	° oder Himmelsrichtung	Anemometer	0 – 360 ° / sechzehn 22,5°-Schritte (Nord (N), Nordnordost (NNE), Nordost (NE), Ostnordost (ENE), usw.)	Gibt die genaue Richtung an, in die sich die Luft auf horizontaler Achse an einem spezifischen Punkt vorbei bewegt
Böen	v ↔	km/h	Anemometer	1 = km/h	Starke Luftbewegung von kurzer Dauer
Niederschlag	mm	mm	Ombrometer / Hyetometer	$1 = \frac{l}{m^2} = 1\frac{dm^3}{m^2} = \frac{(0{,}1\,m)^3}{m^2} = 0{,}001\,\frac{m^3}{m^2} = 0{,}001\,m = 1\,mm$	Gibt die Menge an Wasser an, das sich aus Wolken, Nebel oder Dunst oder Luftfeuchtigkeit in einer definierten Zeitspanne auf einer $1m^2$ großen Fläche sammelt
Sonneneinstrahlung / Sonnenintensität	E	W/m^2	Pyranometer	W/m^2	Bestrahlungsstärke der Sonne auf die Erdoberfläche
Schneedichte	-	kg/m^3	Bestimmte Menge Schnee wird aus der Schneedecke ausgestochen und gewogen	kg/m^3	Masse Schnee pro Quadratmeter in der Schneeschicht
Schneefläche / Schnealbedo	-	0 - 1	Satellitenbild Auswertung	0 - 1	Misst die Reflektion des schneebedeckten Teils einer Fläche der Erde als Fraktion der Solarstrahlung (Kurzwellen), die vom Schnee aus dem solaren Spektrum reflektiert werden
Taupunkt	°	°C oder °K	Taupunktspiegel-hygrometer	°K = 273,15 + °C	Kondensationspunkt von Wasser in der Luft
Bedeckungsgrad / Bewölkungsgrad		ISO/CIE-Standard ISO 15469:2004(E) / CIE S 011/E:2003	ISO/CIE-Standard ISO 15469:2004(E) / CIE S 011/E:2003		Ansammlung von kondensiertem Wasser in der Erdatmosphäre

Zusammengetragen aus
Liljequist und Cehak 1984;
Grabbe et al 2014;
Fathi et al. 202;
Hasan et al. 2016;
ECMWF 2023

Zur Evaluation der Ergebnisse wird auf vier Parameter zurückgegriffen. Die Genauigkeit der Feature Selektion, die Genauigkeit der Cluster, die False Positive Rate und die Durchlaufzeit. Obwohl Pooja et al. 2020 eine Zwei-Schritt-Methodik einführen, vergleichen und evaluieren sie nicht verschiedene Clustering Algorithmen unter Variation ihrer Parameter. Allerdings wurden effektive Messkriterien für die Performanz und die Qualität der Ergebnisse eingeführt und effektive Methodiken zur Selektion relevanter Features. Auch die identifizierten Wettercluster mit relevanten Kriterien und Charakteristiken werden nicht als WEP vorgestellt.

Das und Sun 2014 nutzen Assoziationsregeln, um den Zusammenhang zwischen bestimmten Wetterereignissen und Unfallstatistiken zu analysieren und evaluieren die Effektivität verschiedener Maßnahmen. Mit diesen konnten sie einfache Regelsätze einführen, welche Assoziationen zwischen Wetterparametern, Wetterereignissen und geeigneten Maßnahmen zum Vorbeugen von Unfällen erzeugen. Außerdem konnten versteckte Muster in den Daten analysiert werden, um ein besseres Verständnis über die Wetterdaten zu erlangen. Obwohl Das und Sun 2014 keine Clusteralgorithmen einsetzen, identifizieren sie Wetterereignisse über Assoziationsregeln und stellen u.a. fest, dass bei Regen eine erhöhte Unfallgefahr herrscht. Es ist denkbar, diese Assoziationsregeln auch zum konkreten Identifizieren von WEP oder die Methodik erweiternd mit der Clusteranalyse einzusetzen.
Xu et al. 2015 stellen eine kaskadierte Clusteringmethode bestehend aus KMeans Algorithmen vor, mit welchem sechs Wettercluster generiert werden. Hierfür wird ein kaskadiertes KMeans Clustermodell auf zehn ausgewählte Wetterparameter angewandt. Die maximale Euklidische Distanz zwischen den Datenpunkten eines Clusters und dem Clusterzentrum wird als Kante eines Clusters zum Berechnen der optimalen Zuweisung der Datenpunkte in die Cluster genutzt. Xu et al. 2015 analysieren viele Parameter ihrer vorgestellten kaskadierten KMeans Clustermethode. Allerdings werden keine genauen WEP definiert und auch keine anderen Clusteralgorithmen miteinander verglichen.

Ferstl et al. 2017 nutzen agglomeratives hierarchisches Clustering (HAC) für die Analyse der zeitlichen Veränderung in Ensembles von Wettervorhersagen. Sie vergleichen keine verschiedenen Clusteringmethoden und definieren auch keine WEP, aber nutzen HAC erfolgreich, um Wetterensembles anhand verschiedener Parameter zeithierarchisch für Wettervorhersagen zu clustern.
De Lima et al. 2013 stellen zwei Clusteringansätze vor, die durch Klassifikation ergänzt werden. Es wird HAC zum Clustern der Wetterdaten verwendet und anschließend eine Klassifikation auf die erstellten Testdaten angewandt. Zur Auswertung der Ergebnisse wird eine neu vorgestellte Similarity Metric verwendet. Der zweite Ansatz kombiniert eine Klassifikation mit einer dichtebasierten Clusteranalyse, bei welcher drei Cluster identifiziert werden. Ausgewertet werden die Ergebnisse mit einer Kreuzvalidierung und Klassifikations-Performanz-Indexes. Erfolgreich konnten extreme Wetterereignisse identifiziert werden. Die genauen Charakteristiken der Extremwetterereignisse wurden nicht vorgestellt. Die Details zu den Methodiken sind sehr ausführlich beschrieben.

Nach eingehender Literaturanalyse geht hervor, dass sich zukünftige Arbeiten im Gebiet des Clusterings von Wetterdaten sowohl mit der Evaluierung und dem Vergleich von verschiedenen Clusteralgorithmen und der Variation ihrer Parameter beschäftigen muss als auch mit der genauen Beschreibung der identifizierten WEP.

2.3 Machine Learning

2.3.1 Normalisierung

Die Daten-Normalisierung (oft auch als Daten-Standardisierung bezeichnet) ist eine Methode zur Aufbereitung von Analysedaten. Sie bezeichnet den Prozess, bei dem Parameterwerte verschiedener Intervalle mit verschiedenen physikalischen Einheiten auf einen einheitlichen Intervall (bspw. -1 bis 1 oder 0 bis 1) skaliert werden (Zhou 2021). Dieser Schritt ist insbesondere bei der Anwendung von Algorithmen von Relevanz, die auf der Skaleninvarianz basieren, so wie die meisten Clusteralgorithmen, welche die Distanz der Datenpunkte zueinander berechnen (Jo 2019). Zur Berechnung der Distanz wird zumeist die Euklidische Distanz, in manchen Fällen aber auch die Hausdorff Distanz, Manhatton Distanz oder Cosine Distanz verwendet. Für weitere Details sei auf Liu und Deng 2020 verwiesen.
Jo, J.M. 2019 diskutiert verschiedene Normalisierungsmethoden wie simple feature scaling, min-max scaling, maximum-absolute scaling, Z-score scaling und robust scaling. Diese basieren auf verschiedenen mathematischen Vorgehensweisen, um die Datenparameter auf eine gemeinsame Skala zu bringen, indem die Standardnormalverteilung aller Parameter so umgerechnet werden, dass sie einen Mittelwert von 0 und eine Standardabweichung von 1 haben. Für die detaillierten mathematischen Hintergründe sei auf Zhou 2021 sowie Jo, J.M. 2019 verwiesen.

2.3.2 Auswahl der Anzahl an Cluster

Ein Großteil, der in der Literatur in der Arbeit mit Wetterdaten verwendeten Clustermethoden nutzen Algorithmen, welche nicht automatisch die optimale Anzahl der zu generierenden Cluster berechnen. Sie benötigen eine Methode zum Finden der optimalen Clusteranzahl *k*. Im Sinne des Praxisteils der Arbeit ist damit die Anzahl der zu identifizierenden WEP zu verstehen. Es ist davon auszugehen, dass für verschiedene Regionen der Erde unterschiedliche Anzahlen von WEP identifiziert werden. In diesem Kapitel werden Methoden diskutiert, mit welchen die optimale Anzahl an Clustern in einem Datensatz - unter Einsatz einer bestimmten Clustering Methode - identifiziert werden können.
In Liu und Deng 2020 wird zu diesem Zweck die ‚Ellenbogen-Methode' (engl. ‚Elbow Method') vorgestellt. Diese stellt ein Verfahren zur Bestimmung der optimalen

Anzahl an Clustern bei einer Clusteranalyse dar. Eine ausgewählte Clustering Methode wird iterativ durchgeführt, wobei die Anzahl der zu generierenden Cluster *k* mit der ersten Iteration 1 beträgt und für jede Iteration *i* um 1 erhöht wird, bis jeder Datenpunkt *x* ein eigenes Cluster darstellt, also $k = xmax$. Bei jeder Iteration wird die Summe der quadrierten Distanzen aller Datenpunkte jedes Clusters zum Clusterzentrum berechnet. Der Durchschnitt aller quadratischen Distanzen der Clusterzentren zu allen Datenpunkten der Cluster, kann als Wert einer Kostenfunktion verstanden werden, welcher zu minimieren ist. Üblicherweise wird für diese Berechnung die Euklidische Distanz verwendet (Syakur et al. 2018). Für diese Berechnung kann die Kostenfunktion $L = \sum_{i=1}^{k} \sum_{x \in C_i} |x - C_i|^2$ verwendet werden, wobei *L* die Kostenfunktion ist, mit *x* ein Element des Clusters C_i und *k* der Anzahl an Cluster $|C_i|$ (Liu und Deng 2020). Mit steigender Anzahl der Cluster *k* wird der Wert *L* der Kostenfunktion kleiner. Vor der Stelle der optimalen Clusteranzahl *k* ist die Verminderung des Wertes *L* der Kostenfunktion groß und nach der Stelle der optimalen Clusteranzahl *k* ist dieser stark verringert. In einem sogenannten Elbow-Plot wird der Wert der Kostenfunktion in Abhängigkeit zu *k* dargestellt. Der Elbow-Plot beschreibt typischerweise einen Bogen. An der Stelle der optimalen Clusteranzahl *k* ist eine starke Beugung des Graphen festzustellen, woher die visuelle Elbow-Methode ihren Namen nimmt. An dieser Stelle, an der der Winkel des Plots minimal ist, erreicht *k* also seinen optimalen Wert.
In Yuan und Yang 2019 wird außerdem die Silhouette-Methode vorgestellt, welche als Evaluationsmetrik für Clusteranalysen verwendet wird. Die Details zu Evaluationsmetriken für Clusteranalysen und zu Silhouette-Methode werden in Kapitel 2.3.6. vorgestellt. Allerdings sei vorweggegriffen, dass Evaluationsmetriken auch zur Bestimmung der optimalen Clusteranzahl *k* in einem Clusteringmodell verwendet werden können. Ähnlich wie bei der Elbow Methode wird eine ausgewählte Clustering Methode iterativ durchgeführt, indem für *i=1* auch *k=1* und für jede Iteration *i* um 1 erhöht wird, bis $k = x_{max}$. Bei jeder Iteration wird der durchschnittliche Silhouette Koeffizient über alle Datenpunkte berechnet und jenes Clustermodell mit Clusteranzahl *k* gewählt, bei dessen Iteration der durchschnittliche Silhouette Koeffizient maximal ist.

2.3.3. Möglichkeiten zum Messen und Evaluieren der Performanz der ML-Algorithmen

Die diskutierten ML-Algorithmen werden im Praxisteil der Arbeit auf die Wetterdatensätze angewandt werden, um Datenpunkte in WEP zu clustern. Für jeden Algorithmus wird dessen Effizienz evaluiert. Es sind drei Messwerte von Relevanz: Die Performanz des Algorithmus, der Wert des Evaluationsindexes und die Qualität der Ergebnisse in Form von der Genauigkeit zu identifizierten WEP (Pooja et al. 2020; De Lima et al. 2013). Ein Evaluationsindex gibt an, wie gut ein Algorithmus gleiche Datenpunkte in gleiche Cluster und ungleiche Datenpunkte in ungleiche Cluster sortiert. Gesucht wird also eine hohe Intra-Cluster-Ähnlichkeit und eine geringe Inter-Cluster-Ähnlichkeit (Zhou 2021).
Zu diesem Zweck soll der bereits in Kapitel 2.3.2. vorgestellte Silhouette Koeffizient verwendet werden. Neben diesem werden in Zhou 2021 auch noch weitere Validitätsindizes vorgestellt (Jaccard Coefficient, Fowlkes und Mallows Index, Rand Index, Davies−Bouldin Index, Dunn Index). Aufgrund der Prominenz des Silhouette Koeffizienten in der Forschungsliteratur wird in dieser Arbeit auf diesen zurückgegriffen.
Die Silhouette-Methode wird in Yuan und Yang 2019 vorgestellt. Bei dieser wird berechnet, wie gut ein Datenpunkt in das ihm zugewiesene Cluster passt. Hierzu werden die Faktoren Kohäsion und der Silhouette Koeffizient verwendet. Kohäsion beschreibt die Ähnlichkeit, zwischen einem Datenpunkt und dessen Cluster. Wird der Datenpunkt mit einem fremden Cluster verglichen, wird dieser Wert Separation genannt. Dieser Vergleich wird über den Silhouette Koeffizient berechnet, der Werte zwischen -1 und 1 annehmen kann. Nimmt der Silhouette Koeffizient den Wert 1 an, so ist ein Datenpunkt dem zugewiesenen Cluster sehr ähnlich, bei -1 sind sich Datenpunkt und Cluster sehr unähnlich. Der Silhouette Wert kann auch genutzt werden, um zu evaluieren, ob das genutzte Clustermodell akzeptabel und passend ist. Der Silhouette Koeffizient wird wie folgt berechnet:

$$s(x) = \frac{b(x) - d(x)}{max\{d(x), b(x)} = \begin{cases} 1 - \frac{d(x)}{b(x)}, & d(x) < b(x) \\ 0, & d(x) = b(x) \\ \frac{b(x)}{d(x)} - 1, & d(x) > b(x) \end{cases}$$

Formel 1 Algorithmus zum Berechnen des Silhouette Koeffizienten

Kalkuliert wird die durchschnittliche Distanz *d(x)* eines Datenpunktes *x* zu den anderen Datenpunkten desselben Clusters. Je kleiner *d(x)*, desto mehr passt *x* zu dem jeweiligen Cluster, es besteht also eine hohe Kohäsion. *d(x)* entspricht der Intra-Cluster-Unähnlichkeit des Datenpunktes *x*, welche minimiert werden soll. Der Durchschnitt von *d(x)* über alle Datenpunkte eines Clusters *C* wird die Cluster-Unähnlichkeit von *C* genannt. In einem zweiten Schritt wird die durchschnittliche Distanz aller Datenpunkte *b(x)* eines fremden Clusters zu *x* berechnet. Diese wird als Inter-Cluster-Unähnlichkeit bezeichnet, welche maximiert werden soll. Je größer *b(x),* desto weniger passt *x* zu dem jeweiligen Cluster. Es besteht also eine große Separation.

Ein weiterer wichtiger Faktor zum Messen der Performanz eines Algorithmus ist dessen Komplexität *O*. Diese beschreibt die Menge an Ressourcen, die bei Anwendung eines Algorithmus benötigt werden. Es wird zwischen Zeitkomplexität und Speicherkomplexität unterschieden. Die Zeitkomplexität gibt an, wie viel Zeit ein Algorithmus bei der Ausführung benötigt, während die Speicherkomplexität angibt, wie viel Speicher- und Rechenressourcen für die Ausführung benötigt werden (Firdaus und Uddin 2015).

3 EXPERIMENTELLES DESIGN

Im folgenden Kapitel der Arbeit wird auf das Design und die Implementierung des erstellten Artefakts eingegangen. Zuerst werden die akquirierten und verwendeten Daten genauer erklärt. Anschließend werden die im Theorieteil der Arbeit identifizierten Methoden und Algorithmen in einem Python Programm mittels Jupyter Notebooks unter Verwendung der Python Libraries Pandas und Sklearn implementiert. Wie im vorangegangenen Kapitel diskutiert, wird das zu entwickelnde Artefakt iterativ verbessert und evaluiert.

3.1 Population, Datenquellen und Datensätze

Die für die Erstellung des Designartefakts / Prototypen der Arbeit verwendeten Wetterdaten werden von der IBM Deutschland GmbH und ihren Partnern im Bereich der Meteorologie zur Verfügung gestellt. Die zur Verfügung gestellten Datensätze enthalten aggregierte und abgeleitete Wettermerkmale aus sieben Jahren (Juli 2015 bis Dezember 2022) historischer ERA5-Wetterdaten, aufgeteilt auf die Region Ontario im Süden von Zentralkanada. Für die Analyse wurde die kanadische Provinz Ontario in 55 Regionen aufgeteilt, welche in Abb. 1 dargestellt sind und für jede Region wurde mittels der historischen aggregierten Wetterdaten vom Jahr 2015 bis 2022 die Clusteranalyse unter Anwendung des implementierten Modells durchgeführt, um für jede Region bedeutende WEP zu definieren. Die Aggregationen der Wetterparameter sind räumlich nach den Regionen und zeitlich nach den Messstunden erzeugt. Die Wetterdaten werden in stündlicher Granularität gemessen.

Abbildung 1 Karte der Regionen der genutzten aggregierten historischen Wetterdaten in Ontario, Kanada

Die für die Nutzung in der Analyse des Praxisteils dieser Arbeit ausgewählten ERA5 Parameter, aufgeteilt nach Regionen, sind die folgenden Wetterparameter:
• maximale Böenstärke (max_windgust), welcher die höchste Windböenintensität der betrachteten Stunde des Datenpunktes in m/s angibt
• durchschnittliche Temperatur (avg_temp) in Kelvin
• maximale Schneedichte (max_snow_density_6) in kg/m^3
• maximaler kumulierter Niederschlag (max_cumulative_precip) in mm
• maximale kumulierte Eisschicht (max_cumulative_ice) in mm
• maximaler kumulierter Schnee (max_cumulative_snow) in mm
• maximale durchschnittliche Windgeschwindigkeit (avg_windspd) in m/s
• durchschnittliche Änderung des Luftdrucks (avg_pressure_change) in Pa
• durchschnittliche Windrichtung (avg_winddir) in Kreisgrad (für die Analyse wurde der Sinus (avg_winddir_sin) und Kosinus (avg_winddir_cos) des Gradwertes genutzt)

Der ursprüngliche Datensatz enthielt weitere Parameter, die für die eigentliche Analyse redundant waren, bspw. Bodenfeuchte, der Taupunkt oder die erweiterte Schneedichte. Die Parameter Taupunkt, Feuchtkugeltemperatur und Temperaturveränderung sind redundant, da sie stark mit dem Parameter Lufttemperatur korrelieren (s. Abb. 2). Bei Verwendung mehrerer stark korrelierender Temperaturparameter wird das Ergebnis der Clusteranalyse verfälscht. Zudem sind im ursprünglichen Datensatz für jeden Parameter der minimale, der maximale und der Durchschnittswert für die betrachtete Stunde enthalten. Es wird jeweils nur einer dieser Werte benutzt, der bei der PCA als der relevanteste identifiziert wurde. Die Relevanz der Parameter wird durch die Höhe der entsprechenden Parameterwerte in den Eigenvektoren der Hauptkomponenten wiedergegeben. Mittels der Literaturanalyse und einer PCA wurden die relevantesten Parameter bestimmt und in das Datenobjekt zur Anwendung im Clustermodell aufgenommen. Zudem wurden weitere benötigte Parameter integriert. Die durchschnittliche Temperaturänderung wurde aus bestehenden Daten berechnet und der kategorische Wert der Windrichtung mittels Berechnung von Sinus und Cosinus der Einheit Grad in zwei numerische Werte für die Windrichtung umgerechnet und in zwei neue Spalten des Datenobjektes aufgenommen. Bei der Durchführung der Analyse werden nur die in diesem Kapitel erläuterten Wetterparameter genutzt, während bei der Visualisierung der Ergebnisse alle in Abb. 2 dargestellten Parameter genutzt werden, um weitere Zusammenhänge zwischen den Wetterparametern auswerten zu können, ohne die eigentliche Analyse durch redundante Daten zu verfälschen. Durch Darstellung der Datenparameter über den gesamten Zeitverlauf in Zeitreihendiagrammen und durch Erkundung des Datenobjektes können weitere wichtige Erkenntnisse über die Daten und die Analyse gewonnen werden. Durch eine solche Auswertung wurde der Input des Modells iterativ optimiert. Mittels der ‚StandardScaler' Klasse der scikit-learn Bibliothek werden diese normalisiert (Scikit Learn 2023b), um trotz unterschiedlicher physikalischer Einheiten und variierender Skalen der Einheitswerte eine effektive Analyse zu ermöglichen. So wurden die Daten durch Normalisierung auf eine vergleichbare Skala gebracht.

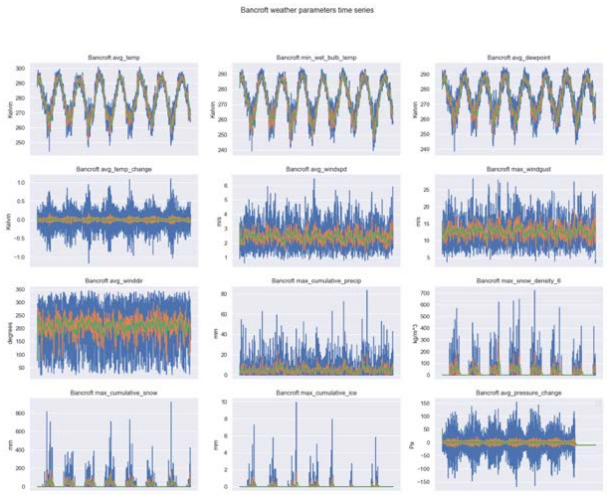


Abbildung 2 Wetterparameter der Region Bancroft von den Jahren 2015 bis 2022 dargestellt in Zeitreihendiagrammen

3.2 Aufbau des Modells und Durchführung der Analyse

Nach Durchführung der Datenakquise und der Datenvorverarbeitung, inklusive Feature Engineering zum Entwickeln neuer Parameter und Optimieren der bestehenden Parameter, beginnt der Aufbau des Clustering Modells. An dieser Stelle variiert die Implementierung der verschiedenen Clusteringmethoden. Zur Vorbereitung des KMeans Clustering Modells, des HAC-Modells und der Gauß'schen Mischmodelle muss zuerst die optimale Anzahl der zu generierenden Cluster für den geladenen Wetterdatensatz gefunden werden. Hierfür wird die in Kapitel 2.3.3 eingeführte Ellenbogen-Methode, unterstützt durch den Silhouette Score, eingesetzt. Die Ergebnisse dieser Methoden sind für den KMeans Algorithmus, angewandt auf die Region Bancroft, beispielhaft in Abb. 3 dargestellt.

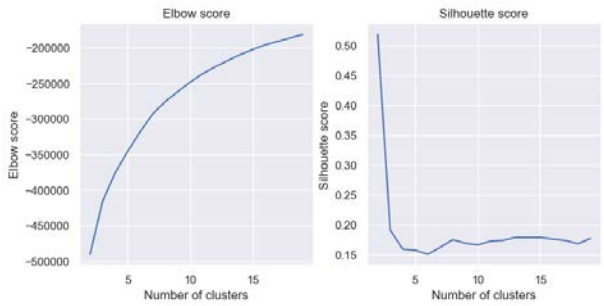


Abbildung 3 Beispielhafte Ergebnisse der Ellenbogen und Silhouette-Methode zum Identifizieren der optimalen Anzahl Cluster eines Modells (KMeans, reguläre Methode)

In Abb. 3 wird ersichtlich, dass die mathematisch optimale Anzahl an Clustern zwei beträgt, an dieser Stelle beträgt der Silhouette Score 0,52. In der Literaturanalyse wurde erörtert, dass für eine bestimmte Region kleiner als 30km^2 eine Wetterclusteranzahl von ca. 5 bis 12 zu erwarten ist. Auch nach iterativer manueller und visueller Evaluation der Ergebnisse, weist die Genauigkeit der Ergebnisse für eine Clusteranzahl zwischen 5 und 12 eine sehr hohe Qualität auf. Daher wurde der weitere Aufbau der Analyse in drei Methoden unterteilt: Reguläres Clustering, kaskadiertes Clustering und eine Clusteranalyse nach durchgeführter PCA zur Dimensionsreduktion. Die Details zu den jeweiligen Methoden werden in den folgenden Kapiteln diskutiert.

3.2.1 Reguläres Clustering Model

Sollte für die Ergebnisse der Ellenbogen- und Silhouette-Methode eine optimale Clusteranazahl von zwei herauskommen, so wird bei der Anwendung des regulären Modells für die verwendete Anzahl der Cluster das zweithöchste lokale Maximum des Silhouette Score genutzt. Ist für das Ergebnis dieser Methoden die optimale Clusteranzahl höher als zwei, wird das globale Maximum genutzt. Mit diesem Wert für die Anzahl der Cluster wird das jeweilige Modell anschließend auf den Wetterdaten der einzelnen Regionen trainiert. Die Ausführungszeit, die beanspruchten Rechenressourcen und der Silhouette Score wird gemessen. Für jeden Parameter wird der Durchschnittswert, der maximale Wert, der minimale Wert sowie der Median für jedes Cluster berechnet und in Radardiagrammen dargestellt. Die so dargestellten Cluster verfügen über normalisierte Werte, da für die Visualisierung in Radardiagrammen eine vergleichbare Skala nötig ist. Aus den Visualisierungen geht zudem hervor, aus wie vielen Datenpunkte jeder Cluster besteht. Die in Radardiagrammen visualisierten Daten sind beispielhaft für die Region Bancroft unter Verwendung des KMeans Algorithmus in Abb. 4 dargestellt.
Die Radardiagramme können visuell evaluiert werden, um grobe Informationen über die Wettercluster zu erlangen wie bspw. die generelle Form, die Anzahl der enthaltenen Datenpunkte und ausschlaggebende Parameter einzelner Cluster (bspw. kumulatives Eis, kumulativer Schnee und Schneedichte für Cluster 2 und 4). Für eine detailliertere Evaluierung der Wettercluster werden Distogramme zu jedem Parameter aller Cluster erzeugt. In diesen können die Details der einzelnen Parameter betrachtet und ausgewertet werden. Insbesondere fallen Korrelationen zwischen bestimmten Clustern und Parametern auf, bspw. den verschiedenen Temperaturparametern. Auch können durch Visualisierung der Ergebnisse in Form von Distogrammen signifikante, typische und charakteristische Parameter für einzelne Cluster identifiziert werden. Beispielhaft sei hier der Cluster 2 in Abb. 5 genannt, welcher durch hohes kumulatives Eis, eine hohe Schneedichte und eine nordöstliche bis östliche Windrichtung gekennzeichnet ist.
Zuletzt werden für die reguläre Clustering Methode auch die Clusterstatistiken zur Detailanalyse (s. Tabelle 3) sowie Durchlaufzeit, CPU-Ausführzeit und der genutzte physikalische Speicher (RAM) des Modells ausgegeben, welche in Tabelle 6 zusammengeführt sind.

Tabelle 3 Cluster-Statistiken der Wetterdaten (für die Parameter Windböenstärke und Lufttemperatur. KMeans, reguläre Methode)

	max_windgust	max_windgust	max_windgust	max_windgust	max_windgust	max_windgust	max_windgust	max_windgust	avg_temp	avg_temp	avg_temp	avg_temp	avg_temp	avg_temp	avg_temp	avg_temp
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
Cluster																
0	15299	16,61398508	2,62952922	11,786448	14,752408	16,170511	17,965321	27,842741	15299	278,7891077	9,204299382	246,8967621	272,3597642	278,256279	285,9888897	300,8444696
1	17652	10,88417489	2,352073859	3,918856	9,241031	11,0439855	12,560998	21,397372	17652	267,4314554	6,668863133	243,8493934	262,9244489	268,585182	272,8438728	278,9632008
2	498	16,39884962	3,698682178	9,565561	14,585312	16,704868	17,962448	27,842741	498	270,7082114	3,459636788	260,4995251	268,7308082	271,6053702	273,2308144	276,3089349
3	3636	15,94988902	3,494824118	7,964065	13,0162925	15,266787	18,124078	28,236084	3636	270,3980325	5,153416272	248,7658252	267,7856857	271,9464159	273,7894735	284,0694475
4	23298	10,35639391	2,158539944	4,089015	8,8051305	10,3743355	11,908469	17,777832	23298	289,0160281	5,644858223	275,7352288	284,9921038	289,9529886	293,2788173	300,034144
5	4962	14,68120223	3,168057642	7,166526	12,417266	14,00773	16,903349	27,293554	4962	288,51096	5,644653274	273,6833886	284,3417776	290,0551678	292,8742362	299,3861519

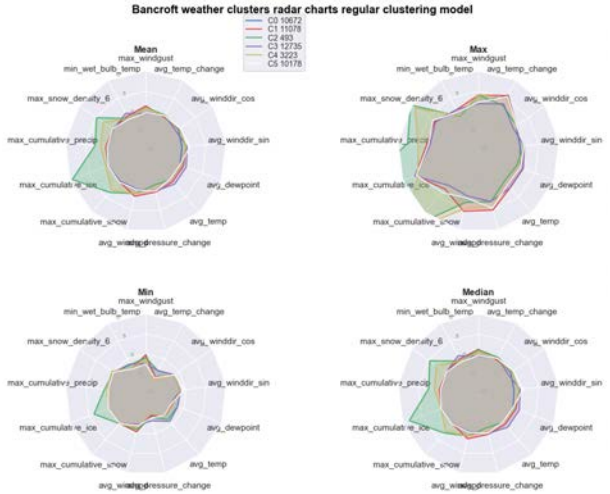


Abbildung 4 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, reguläre Methode) in Radardiagrammen

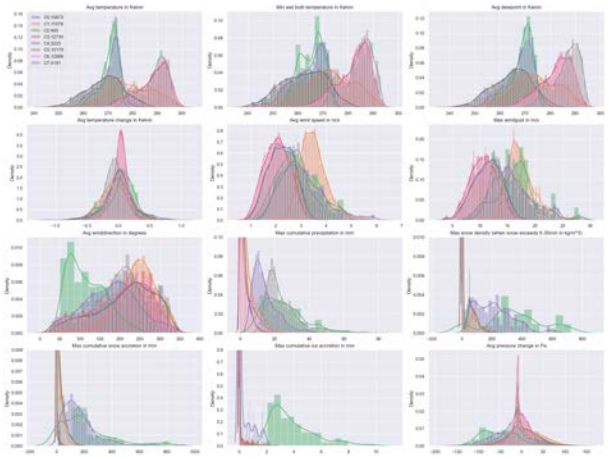


Abbildung 5 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, reguläre Methode) in Distogrammen

3.2.2 Kaskadiertes Clustering Model

Ergibt sich für die Ellenbogen- und Silhouette-Methode eine optimale Clusteranzahl von zwei, so wird zusätzlich zu dem regulären Clustermodell das kaskadierte Modell eingesetzt. Bei diesem werden die Wetterdaten einer Region zuerst in zwei Cluster gruppiert. Durch iterative Implementierung und Evaluation konnte herausgefunden werden, dass ein Cluster zumeist ein Wettercluster ist, der Extremwetterereignisse enthält, wenn dieser 5% bis 20% der Datenpunkte der Gesamtmasse enthält. Die Ergebnisse dieses ersten Schrittes des kaskadierten Clusterings sind in Abb. 6 und 7 zu sehen.

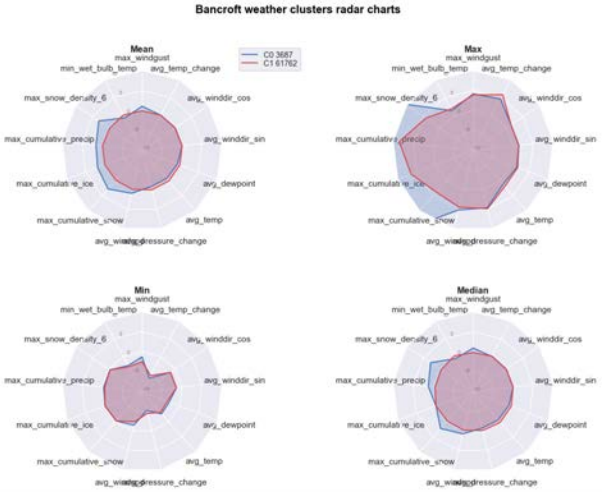


Abbildung 6 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, erster Schritt der kaskadierten Methode) in Radardiagrammen

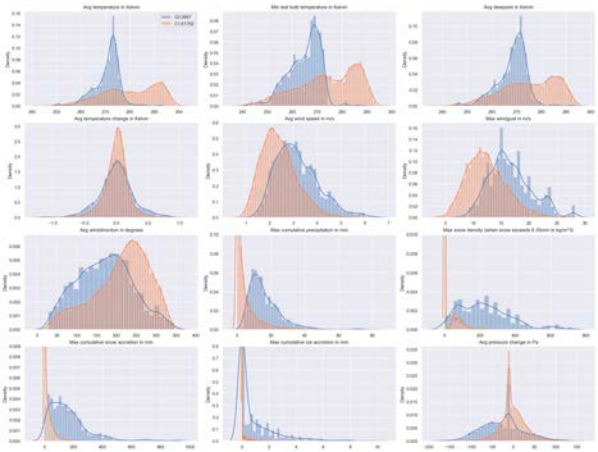


Abbildung 7 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, 1. Schritt der kaskadierten Methode) in Distogrammen

Tritt wie in den Abb. 6 und 7 der Fall ein, dass ein Wettercluster 5% bis 20% der Datenpunkte enthält, wird dieser als ein Wettercluster betrachtet, der Extremwetterereignisse enthält. Auf diesen wird in einem zweiten Schritt die Ellenbogen- und Silhouette-Methode an-gewandt und mit der hierbei identifizierten optimalen Anzahl an Clustern, das Extremwettercluster in einem zweiten Schritt regulär geclustert. In Abb. 8 sind für die Region Bancroft unter Verwendung des KMeans Algorithmus beispielhaft die Ellenbogen- und Silhouette-Methode für den zweiten Schritt des kaskadierten Modells zu sehen.

Die endgültigen Ergebnisse der kaskadierten Methode werden ebenfalls in Radar- und Disto-grammen dargestellt (s. Abb. 9 und 10). Es fällt auf, dass die Wetterereignisse viel genauer und klarer definiert sind und der endgültige Silhouette Score viel höher ist, die Cluster sind mathematisch also optimaler getrennt. Allerdings ist zu bedenken, dass in der kaskadierten Methode nur die Extremwetterereignisse und deren Charakteristiken eindeutig definiert und analysiert werden können.

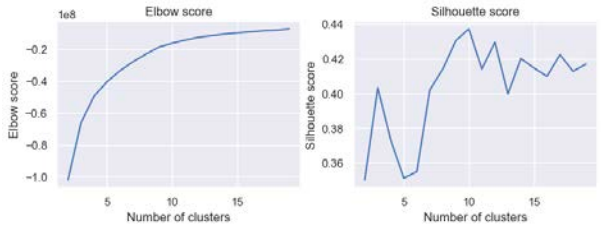


Abbildung 8 Beispielhafte Ergebnisse der Ellenbogen und Silhouette-Methode (KMeans, 2. Schritt kaskadiertes Modell)

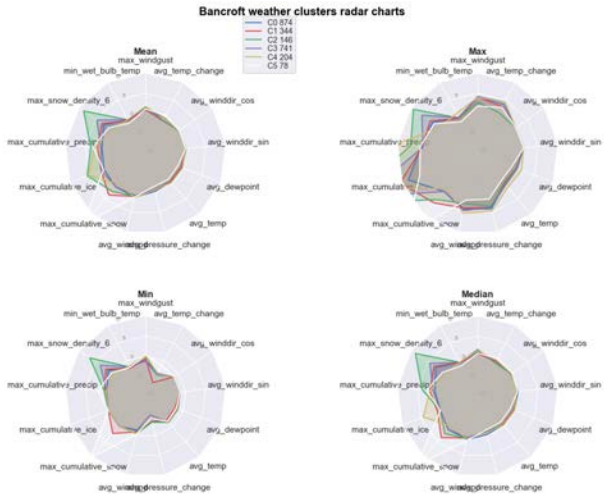


Abbildung 9 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, 2. Schritt der kaskadierten Methode) in Radardiagrammen

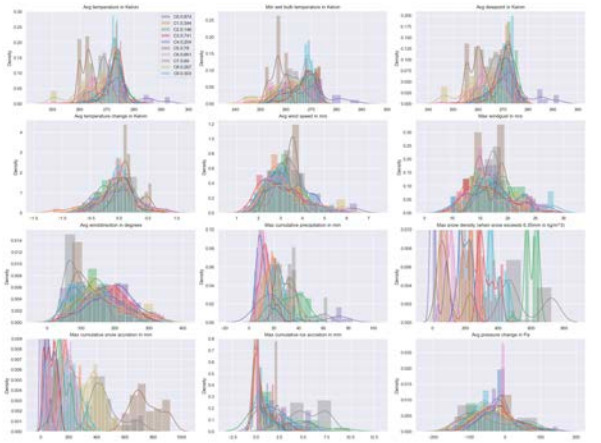


Abbildung 10 Beispielhafte Visualisierung der Cluster der (2. Schritt der kaskadierten Methode) in Distogrammen

3.2.3 PCA Clustering Model

Die in Kapitel 2.3.1 eingeführte PCA nach Zhou 2021 oder Abdi und Williams 2010 wird genutzt, um die hohe Dimensionalität der Wetterdatensätze zu reduzieren, indem die große Anzahl der Wetterparameter in ein kleineres Set von Hauptkomponenten transformiert wird. Der größte Teil des Informationsgehalts der Parameter bleibt erhalten, der allerdings nur durch Aufschlüsselung der Hauptkomponenten für Menschen interpretierbar wird (für Details sei auf Zhou 2021 oder Abdi und Williams 2010 verwiesen). Für die PCA wird eine Anzahl von drei Hauptkomponenten gewählt, damit das Clusterergebnis neben den bisher aufgeführten Visualisierungen auch in einem dreidimensionalen Raum dargestellt und visuell ausgewertet werden kann. Zudem kann die Zusammensetzung der Hauptkomponenten untersucht werden, um die Relevanz der ursprünglichen Wetterparameter für die einzelnen Hauptkomponenten selbst zu untersuchen und so schlussendlich auch die Relevanz einzelner Wetterparameter im Allgemeinen für die Analyse evaluieren zu können. Dieses Wissen wurde iterativ zur Verbesserung der Modelle genutzt. Die Ergebnisse der PCA-Methode werden zusätzlich zu den Visualisierungen in Radardiagrammen (Abb. 11) und Distogrammen (Abb. 12) zur Bewertung der Clusterergebnisse auch in einem dreidimensionalen Graphen visualisiert. Zusätzlich werden die Varianzpfeile der Eigenvektoren der Parameter im 3D-Plot angegeben (Abb. 13 und 14). Diese Auswertung unterstützt iterativ dabei, die Daten und Ergebnisse sowie die Zusammenhänge der Parameter und deren Relevanz besser zu verstehen und die Clusteranalyse generell zu optimieren. Die PCA-Methode ermöglicht das Betrachten neuer Aspekte bei der Clusteranalyse, wie bspw. die Verteilung im dreidimensionalen Raum und insbesondere das Verständnis der Wetterextreme (die orangenen Datenpunkte / Cluster 3 in Abb. 14). Allerdings müssen die Hauptkomponenten zur Ergebnisbetrachtung und zur Definition der WEP in die ursprünglichen Wetterparameter zurückgeschlüsselt werden.

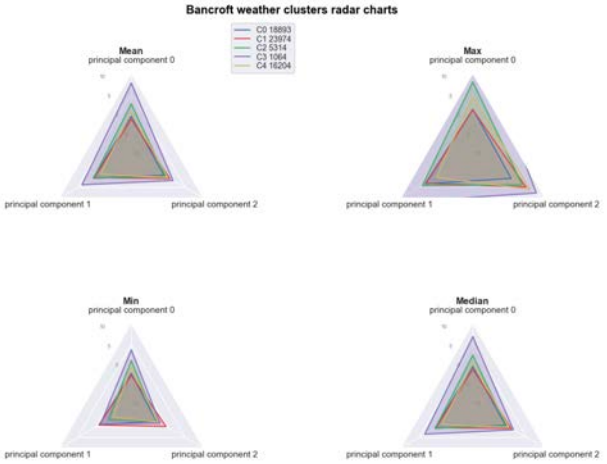


Abbildung 11 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, reguläre PCA-Methode) in Radardiagrammen

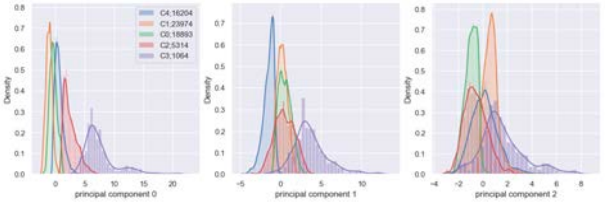


Abbildung 12 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, reguläre PCA-Methode) in Distogrammen

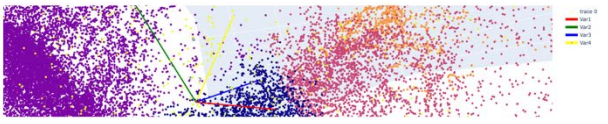


Abbildung 13 Varianzpfeile der regulären PCA-Methode im dreidimensionalen Raum

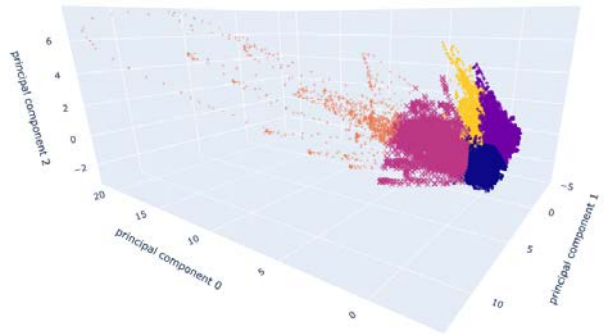


Abbildung 14 Beispielhafte Visualisierung der Cluster der Wetterdaten (KMeans, reguläre PCA-Methode) im dreidimensionalen Raum

3.2.4 Definition der Wetterereignisprofile

Zur Definition der WEP werden alle Datenpunkte unter Betrachtung des genauen Datums und der genauen Uhrzeit, zu der sie eingetreten sind, in Verbindung mit dem zugewiesenen Clusterlabel betrachtet. Zeitlich benachbarte Datenpunkte, die dem gleichen Wettercluster angehören, werden dem gleichen Wetterereignis zugeordnet. Die Nachbarschaft hat eine Toleranz von drei Nachbarn / Stunden. Unter Betrachtung der zeitlichen Dauer und der Ergebnis- und Parametervisualisierung der Wetterereignisse können diese aus meteorlogischer Sicht qualitativ beschrieben, definiert und benannt werden. Zur qualitativen Messung der Genauigkeit werden die generierten WEP mit konkreten historischen Wetterereignissen verglichen, die von der kanadischen Regierung für die jeweiligen Regionen erfasst und veröffentlicht wurden (Government of Canada 2022). Insbesondere ist relevant, ob die genannten Wetterparameter und ihre Intensität, die geographische Einordnung, sowie die zeitliche Abgrenzung übereinstimmen. Die Genauigkeit der WEP werden darin gemessen, wie viel Prozent der verglichenen Wetterereignisse übereinstimmen. Die Qualität der WEP gibt an, wie hoch der Detailgrad der WEP ist und wie gut die Modelle die Wetterparameter ein- und voneinander abgrenzen können, um das WEP zu beschreiben.
Dieser Schritt wird für alle durchgeführten Methoden, unter Implementierung aller genannten Clustering Algorithmen durchgeführt. Das iterativ erarbeitete und evaluierte Artefakt, das in diesem Kapitel beschrieben ist, wird auf alle in Abb. 1 dargestellten Regionen unter sukzessiver Implementierung der vorgestellten Clustering Algorithmen angewandt. Diese ergeben die identifizierten WEP als Kernergebnisse dieser Arbeit (Tabelle 4), welcher nach Aggregation zeitlicher benachbarter Datenpunkte die Wetterereignisse in Tabelle 5 (beispielhaft für die GMM auf der Region Bancroft) ergeben. Beispielhaft wird an dieser Stelle das WEP des Clusters 2 des kaskadierten KMeans Clustermodells erläutert, um eine beispielhafte Ausführung der durch die vorgestellten Modelle generierten Ergebnisse als Kernergebnisse dieser Arbeit zu geben. Dieses stellt Schneestürme mit viel Niederschlag dar, welches mit 270 Datenpunkten 0,41% der Gesamtdatenmenge ausmacht. Dieses WEP ist geprägt durch sehr niedrige Temperaturen unter dem Gefrierpunkt, starke Südost-winde bis zu 6 m/s, starke Windböen mit bis zu 23 m/s, hohen Niederschlag mit bis zu 40 mm, niedrige Schneedichte mit 0 bis 200 kg/m^3 und einer sehr hohen kumulative Schneedecke mit bis zu 500 mm. Eine genauere Analyse und die Ergebnisauswertung erfolgen im folgenden Kapitel.

4 ERGEBNISAUSWERTUNG UND DISKUSSION

4.1 Untersuchung und Evaluation der Definition von WEP

Über die Zusammensetzung und Ausprägungen der Parameter eines Clusters lassen sich Charakteristiken für einzelne WEP beschreiben. Mittels der Clusteranalyse konnte somit über die WEP und deren Charakteristiken ein konkretes Wetterereignis sowie dessen Zeitraum und Dauer bestimmt werden. In diesem Kapitel wird beispielhaft der Fokus auf die Region Bancroft gelegt. Diese befindet sich geographisch mittig in Ontario, mit einigem Abstand zu den sehr kalten Gebieten im Norden und mit genügend Abstand zu den Gebieten im Süden, wo der sogenannte ‚Lake effect' das Klima maßgeblich beeinflusst (Hjelmfelt 1990). Für die Region Bancroft wurden unter Anwendung der beschriebenen Modelle und Methoden die in Tabelle 4 identifizierten WEP für die Region Bancroft hier beispielhaft die GMM tabellarisch dargestellten WEP identifiziert. Dabei wurden die erzeugten Radardiagramme, Distogramme und Clusterstatistiken der jeweiligen Algorithmen und Methoden manuell qualitativ ausgewertet und mithilfe der Literatur und Ergebnissen der Literaturanalyse aus Kapitel 2 qualitativ benannt und charakterisiert. Es fällt auf, dass sich gleiche WEP, identifiziert durch unterschiedliche Algorithmen, in der Anzahl der enthaltenen Datenpunkte, der Dauer und der Ausprägung der Parameter ähneln. Zudem wurden durch die regulären Clustering Methoden allgemeinere Wetterereignisse identifiziert, wie generell Schneefall, warme Tage oder windige Wetterereignisse. Mittels der kaskadierten Methode konnten detaillierte Wetterextreme gefunden werden. Die PCA-Methode wurde zur Verdeutlichung der Verteilung der Datenpunkte in einem dreidimensionalen Raum genutzt, eine Aufschlüsselung der Hauptkomponenten zur Detailanalyse ist aus Zeitgründen nicht erfolgt. Das größere Cluster, nach dem ersten Clustering Schritt der kaskadierten Methoden, wird als gewöhnliches Wetterevent mit allgemein moderat ausgeprägten Parametern betrachtet. Es fällt auf, dass die verschiedenen Algorithmen grundsätzlich ähnliche bis gleiche WEP erkennen. Das in Kapitel 4.3.4 erläuterte WEP ‚Schneesturm mit viel Niederschlag' wird mit ähnlichen Charakteristiken durch das kaskadierte KMeans Modell und durch das kaskadierte GMM-Modell erkannt, ebenso wie die WEP ‚Moderater Schneefall', Starker Schneefall mit viel liegendem Schnee', ‚Frontdurchlauf / Langanhaltender gefrierender Regen' und ‚Schneesturm mit sehr niedrigen Temperaturen'. Die WEP ‚Sturm mit gefrierendem Regen' und ‚Milder Schneefall' wurden von allen kaskadierten Modellen erkannt. Durch die regulären Clustermodelle wurden ebenfalls ähnliche WEP identifiziert.

Tabelle 4 Tabellarische Auflistung der identifizierten WEP für die Region Bancroft für die GMM

GMM Reguläres Clustering	Ereignis Name	Ereignis Beschreibung	**GMM Kaskadiertes Clustering**	Ereignis Name	Ereignis Beschreibung
Cluster 0 20633 Datenpunkte 31,53% der Gesamt-menge	Windiger und warmer Sturm	□ Hohe Temperaturen bis zu 298°K □ Moderater bis starker Wind (2 bis 4 m/s) □ Moderate bis starke Windböen (12 bis 22 m/s) □ Moderater Niederschlag (um die 20 mm) □ Kein Schnee oder Eis	**Cluster 0** 2920 Datenpunkte 4,46% der Gesamt-menge	Moderater Schnee-fall	□ Niedrige Temperaturen nur wenig höher als 0°C □ Hohe Schneedichte (um die 200 kg/m^3) □ Moderate kumulative Schneedecke (200 bis 400 mm) □ Wenig Niederschlag (10 bis 20 mm) □ Mittelstarke Südwestwinde (um 2,5 m/s)
Cluster 1 10133 Datenpunkte 15,48% der Gesamt-menge	Gewöhnliches Ereignis	□ Moderate Ausprägung der meisten Parameter □ Kein Schnee oder Eis	**Cluster 1** 1040 Datenpunkte 1,59% der Gesamt-menge	Starker Schneefall mit viel liegendem Schnee	□ Niedrige Temperaturen nur wenig höher als 0°C □ Moderater bis hoher Niederschlag (mit 20 bis 40 mm) □ Moderate Winde um die die 2,5 m/s) □ Hohe Schneedichte (um die 350 bis 500 kg/m^3) □ Hohe kumulative Schneedecke (200 bis 400 mm) □ Sehr hohe kumulative Eisdecke (bis zu 8 mm)
Cluster 2 10551 Datenpunkte 16,12% der Gesamt-menge	Gewöhnliches Ereignis	□ Moderate Ausprägung der meisten Parameter □ Kein Schnee oder Eis	**Cluster 2** 430 Datenpunkte 0,66% der Gesamt-menge	Sturm mit gefrierendem Regen / Schwerer Schnee- und Eissturm	□ Niedrige Temperaturen nur wenig um die 0°C □ Starke Ostwinde (mit bis zu 4,5 m/s) □ Starke Windböen (mit bis zu 20 m/s) □ Hoher Niederschlag (mit bis zu 60 mm) □ Hohe Schneedichte (bis 700 kg/m^3) □ Hohe kumulative Schneedecke (bis 700 mm) □ Hohe kumulative Eisdecke (bis zu 8 mm) □ Sehr hohe negative Änderungen des Luftdrucks (bis zu 100 Pa)
Cluster 3 9578 Datenpunkte 14,63% der Gesamt-menge	Gewöhnliches Ereignis	□ Moderate Ausprägung der meisten Parameter □ Kein Schnee oder Eis	**Cluster 3** 2630 Datenpunkte 4,02% der Gesamt-menge	Moderater Regen mit liegendem Schnee	□ Niedrige Temperaturen nur wenig höher als 0°C □ Keine Schneedichte □ Niedrige kumulative Schneedecke (bis 50 mm) □ Wenig Niederschlag (mit bis zu 15 mm) □ Schwache Westwinde (um die 3 m/s) □ Niedrige kumulative Schneedecke (bis zu 100 mm) □ Moderate kumulative Eisdecke (bis zu 2 mm)
Cluster 4 3900 Datenpunkte 5,96% der Gesamt-menge	Moderater Schneesturm	□ Niedrige Temperaturen wenig unter 0°C □ Moderate Winde bis zu 4 m/s □ Hohe Windböenstärke bis zu 23,5 m/s □ Niedriger bis moderater Niederschlag (10 bis 22 mm) □ Hohe Schneedichte um die 300 bis 450 kg/m^3 □ Moderate kumulative Schneedecke (200 bis 400 mm	**Cluster 4** 261 Datenpunkte 4,02% der Gesamt-menge	Schneesturm mit viel Niederschlag	□ Sehr niedrige Temperaturen unter 0°C □ Starke Südostwinde (bis zu 4 m/s) □ Starke Windböen (mit bis zu 25 m/s) □ Hoher Niederschlag (mit bis zu 40 mm) □ Niedrige Schneedichte (0 bis 200 kg/m^3) □ Sehr hohe kumulative Schneedecke (bis zu 900 mm) □ Moderate kumulative Eisdecke (bis zu 2 mm)
Cluster 5 7716 Datenpunkte 11,79% der Gesamt-menge	Kälte mit niedrigem Niederschlag und Wind unter dem Gefrierpunkt	□ Sehr niedrige Temperaturen unter 0°C □ Schwache Westwinde (mit 2 bis 4 m/s) □ Schwache Windböen (um die 15 m/s) □ Niedriger Niederschlag (5 bis 10 mm) □ Niedrige Schneedichte (bis zu 200 kg/m^3) □ Niedrige kumulative Schneedecke (bis zu 100 mm)	**Cluster 5** 4497 Datenpunkte 6,87% der Gesamt-menge	Milder Schneefall	□ Niedrige Temperaturen um die 0°C □ Mittlelhohe bis niedrige Schneedichte (um die 100 kg/m^3) □ Mittelstarke Westwinde (2 bis 4 m/s) □ Mittelstarke Windböen (mit bis zu 22 m/s) □ Wenig Niederschlag (5 bis 10 mm) □ Moderate kumulative Schneedecke (bis zu 180 mm)
Cluster 6 1528 Datenpunkte 2,33% der Gesamt-menge	Sturm mit gefrierendem Regen / Schwerer Schnee- und Eissturm	□ Sehr niedrige Temperaturen unter 0°C □ Starke Südostwinde (mit bis zu 4,5 m/s) □ Starke Windböen (mit bis zu 25 m/s) □ Hoher Niederschlag (mit bis zu 45 mm) □ Hohe Schneedichte (500 bis 700 kg/m^3) □ Hohe kumulative Schneedecke (400 bis 700 mm) □ Hohe kumulative Eisdecke (bis zu 8 mm) □ Sehr hohe Änderungen des Luftdrucks (bis zu 100 Pa)	**Cluster 6** 934 Datenpunkte 1,43% der Gesamt-menge	Frontdurchlauf / Langanhaltender gefrierender Regen	□ Mittlere Temperaturen (von 270°K bis 280°K) □ Starke Ostwinde (mit bis zu 6 m/s) □ Starke Windböen (mit bis zu 28 m/s □ Hoher Niederschlag (bis zu 35 mm) □ Moderate kumulative Schneedecke (bis zu 400 mm) □ Hohe kumulative Eisdecke (bis zu 4 mm) □ Lange Dauer (bis zu mehreren Tagen)
Cluster 7 1306 Datenpunkte 2,00% der Gesamt-menge	Gefrierender Regen ohne Wind und viel Eis und Schnee	□ Niedrige Temperaturen um die 0°C □ Starke Südostwinde (mit bis zu 4,5 m/s) □ Starke Windböen (mit bis zu 25 m/s) □ Moderater Niederschlag (mit bis zu 20 mm) □ Moderate Schneedichte (bis zu 300 kg/m^3) □ Niedrige kumulative Schneedecke (bis zu 100 mm) □ Moderate kumulative Eisdecke (bis zu 2,5 mm) □ Sehr hohe Änderungen des Luftdrucks (bis zu 100 Pa)			

Die WEP ‚Warmer Sturm', Milder Schneefall', ‚Schneesturm' und ‚Gewöhnliches Ereignis' wurden bis auf DBSCAN von allen regulären Clustermodellen erkannt. Die WEP ‚Gewöhnliches windiges Ereignis', ‚Gewöhnliches Schneeereignis mit liegendem Schnee' und ‚Gewöhnlicher warmer Tag' wurden von den regulären KMeans und GMM-Modellen erkannt. Die Details zu den exakten durch die jeweiligen Modelle identifizierten WEP inklusive konkreter relevanter Parameter als prägnante Eigenschaften und prozentualer Anteil der Datenpunkte ist in Tabelle 4 (beispielhaft für die GMM) gegeben. Auch fallen Korrelationen einiger Parameter der WEP auf, so zum Beispiel, dass Westwinde mildes Wetter mit sich bringen. Durch die PCA-Methode wird zudem deutlich, dass die meisten Datenpunkte sehr dicht beieinander liegen und die Ausprägungen derer Parameter nicht sehr stark variieren. Wetterextreme hingegen sind auffällig weniger dicht verteilt und geprägt durch extremere Ausprägungen der Parameter. Auch ist zu erkennen, dass Wetterextreme vermehrt starke Ausprägungen der Parameter Schneedichte, Schneedecke, Eisdecke, Niederschlag und Windstärke aufweisen. Unter Einbezug zeitlich benachbarter Datenpunkte zur Definition zeitlich abgegrenzter Wetterereignisse, wird deutlich, dass die meisten Wetterextreme auf den Frühling und Winter fallen. Dies kann darin begründet sein, dass Kanada im Sommer grundsätzlich milderes Wetter aufweist oder dass die Wetterextreme im Sommer nicht erkannt wurden. Da im regulären Clustermodell jedoch Sommerereignisse identifiziert wurden, ist von Ersterem auszugehen. Meteorologen können die vorgestellten Modelle und Methoden für detailliertere Auswertungen des Wetters nutzen, eine detaillierte meteorologische Analyse liegt nicht im Rahmen dieser Arbeit. Die Genauigkeit der einzelnen Algorithmen und Methoden, welche in den vorangegangenen Kapiteln vorgestellt wurden, fällt unterschiedlich aus. In bestimmten Situationen hat jedes Modell Vor- und Nachteile. Diese werden in den folgenden Kapiteln diskutiert. Nach erfolgreicher Definition der WEP durch die jeweiligen Algorithmen und Methoden wie in Tabelle 5 dargestellt, wurden zeitlich benachbarte Datenpunkte, die dem gleichen WEP angehören, dem gleichen Wetterereignis unter Zuweisung einer Wetterereignisnummer gruppiert. So konnten konkrete zeitlich begrenzte Wetterereignisse definiert werden, wie sie beispielhaft für das kaskadierte KMeans Modell für Bancroft in Tabelle 5 zu sehen sind.

Tabelle 5 Aggregation zeitlich benachbarter Datenpunkte zur Definition von Wetterereignissen (Auszug der Ergebnisse der KMeans Modelle für Bancroft für das Jahr 2022)

index	year	storm_id2	storm_id	label2	storm_start_dh	storm_end_dh	min_temp	max_windgust	max_windspd	max_snow_density	max_cumulative_precip	max_cumulative_snow	max_cumulative_ice	wep	duration	duration_hour
305	2022	c123	5_c003	5	16.01.22 15:00	17.01.22 13:00	252,159256	15,30119	4,900168543	57,1838256	37,828	920,464		0 Snowstorm with very low temperature	0 days 22:00:00	22
308	2022	c124	7_c072	7	18.01.22 20:00	19.01.22 06:00	253,790497	17,28235	5,841605511	61,1281422	7,008	163,533		0 Cold with low precipitation and wind below freezing point	0 days 10:00:00	10
309	2022	c125	0_c094	0	01.02.22 10:00	01.02.22 23:00	260,352356	16,29994	5,481623166	184,146529	15,3	133,29		0 Mild snowfall	0 days 13:00:00	13
310	2022	c125	7_c073	7	02.02.22 09:00	02.02.22 20:00	260,417999	14,0039	4,275904797	156,314139	17,867	187,312		0 Cold with low precipitation and wind below freezing point	0 days 11:00:00	11

4.2 Auswertung und Vergleich der verschiedenen Clustering Algorithmen mit Evaluierung der verschiedenen Methodiken

Im folgenden Kapitel werden die implementierten Clustering Algorithmen - untergliedert in die drei vorgestellten Methoden - ausgewertet und verglichen. Relevant sind der bereits diskutierte Silhouette Score als Evaluationsmatrix (für die Bewertung eines Algorithmus im Gesamten wird der Durchschnitt der einzelnen Methoden genommen), die Performanz in Form von Durchlaufzeit, CPU-Ausführungszeit und genutztem physikalischem Speicher (RAM) sowie die Qualität und Genauigkeit der Ergebnisse. Für letztere werden die definierten WEP unter Betrachtung ihrer zeitlichen Einordnung und Parameterausprägung mit offiziellen Wettermeldungen verglichen. Die kanadische Regierung stellt auf ihrer Internetseite eine Auflistung regionaler ‚Wetter-Highlights' für jedes Jahr zur Verfügung, welche für diesen Zweck genutzt werden (Government of Canada 2022). ‚The Weather Network' führt ein ähnliches Register (The Weather Network 2022). Mit den genannten Quellen wird verglichen, wie viele von den bedeutendsten zehn der historisch dokumentierten Wetterereignisse von den Modellen richtig vorhergesagt wurden. Die Auswertung anhand der genannten Kriterien für die Algorithmen und Methoden ist tabellarisch in Tabelle 6 dargestellt. Da die Genauigkeit und Qualität der Ergebnisse über eine manuelle Auswertung erfolgt, wird diese aufgrund der zeitlichen Limitation nur für die Region Bancroft vorgenommen. Für den DBSCAN Algorithmus ist aufgrund seiner Funktionsweise kein kaskadierter Ansatz möglich (s. Kapitel 2.3.4). Die PCA-Methode wurde nur zur Verdeutlichung der Verteilung der Datenpunkte in einem dreidimensionalen Raum genutzt, eine Aufschlüsselung der Hauptkomponenten zur Detailanalyse für eine qualitative Auswertung ist aus Zeitgründen nicht erfolgt. Eine Auswertung der Ergebnisse in Tabelle 6 macht deutlich, dass der Silhouette Score für die DBSCAN Modelle am höchsten ist. Jedoch identifiziert DBSCAN als dichtebasiertes Modell nur zwei verschiedene Cluster. Diese können in gewöhnliche und Extremwetterereignisse eingeteilt werden, was insbesondere durch die Visualisierung der Ergebnisse des DBSCAN PCA Modells verdeutlicht wird. Bezüglich der Modelle mit höherer Genauigkeit hat KMeans einen höheren Silhouette Score. Jedoch ist der Silhouette Score über alle Modelle hinweg eher gering.

Tabelle 4 Auswertung und Vergleich der vorgestellten Clusteralgorithmen und Methoden anhand des Silhouette Scores, der Performanz, der Genauigkeitund der Qualität der generierten WEP

	Silhouette Score	*Performanz in Durchlaufzeit, CPU-Ausführungszeit und genutzte Rechenressourcen*			*Anzahl Cluster*	*Genauigkeit*	*Qualität / Detail*
		Execution time	*CPU-execution time*	*Genutzte Rechenressourcen (physikalischer Speicher)*			
KMeans Regular method	*0,195426366*	*1.923,3 sec*	*2.996,6 sec*	*743.153,66 MB*	*5*	*7/10*	*Moderat*
KMeans Cascaded method	*0,676872128*	*2.207,19 sec*	*3.447,26 sec*	*875.134,98 MB*	*10*	*6/10*	*Hoch*
KMeans PCA method	*0,339207292*	*1.971,96 sec*	*3.031,43 sec*	*442.576,896 MB*	*-*	*-*	*-*
KMeans total	*0,403835262*	*6.102,45 sec*	*9.475,28 sec*	*2.060.865,54 MB*	*5 to 10*	*6.5/10*	*Hoch*
HAC Regular method	*0,158407099*	*70.046,84 sec*	*28.367,62 sec*	*2.486.848*	*4*	*5/10*	*Moderat*
HAC Cascaded method	*0,579056567*	*91.060,9 sec*	*39.714,67 sec*	*2.932.902,40*	*3*	*2/10*	*Niedrig*
HAC PCA method	*0,336365936*	*84.056,21 sec*	*40.565,7 sec*	*2.784.217,60*	*-*	*-*	*-*
HAC total	*0,357943201*	*245.163,95 sec*	*108.647,99 sec*	*5.203.968,749 MB*	*3 to 4*	*3.5/10*	*Niedrig bis moderat*
GMM Regular method	*0,057888212*	*3.184,08 sec*	*10.144,34 sec*	*890.085,38*	*8*	*7/10*	*Moderat*
GMM Cascaded method	*0,312381864*	*3.728,01 sec*	*12.232,32 sec*	*921.899,01*	*7*	*7/10*	*Sehr hoch*
GMM PCA method	*0,16188301*	*1.606,51 sec*	*4.070,52 sec*	*498.655,23*	*-*	*-*	*-*
GMM total	*0,177384362*	*8.518,60 sec*	*26.447,18 sec*	*2.310.639,62 MB*	*7 to 8*	*7/10*	*Sehr hoch*
Regular method	*0,49252297*	*4.185,9 sec*	*7.106,79 sec*	*965.820,42*	*2*	*7/10*	*Low*
PCA method	*0,68362696*	*2.816,01 sec*	*4.596,45 sec*	*725.385,22*	*-*	*-*	*-*
DBSCAN	*0,58807496*	*7.120,20 sec*	*11.816,37 sec*	*1.691.205,63 MB*	*2*	*7/10*	*Low*

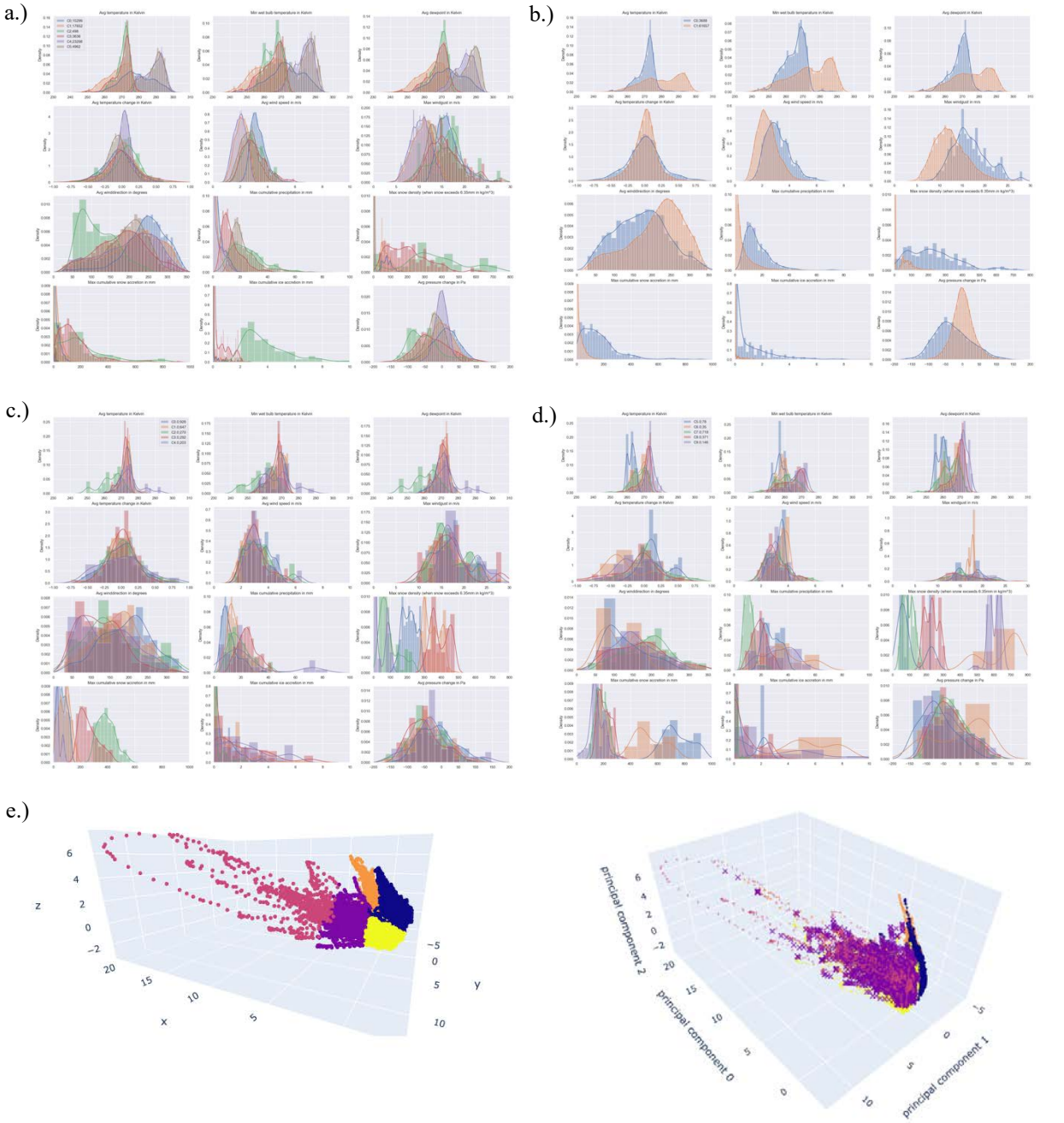


Abbildung 15 Ausprägungen der Wetterparameter der WEP für KMeans für die Region Bancroft
a.) regulären Methode b.) erster Schritt der kaskadierten Methode c.) kaskadierten Methode (Cluster 0 bis 4)
d.) kaskadierten Methode (Cluster 5 bis 9) e.) PCA-Methode

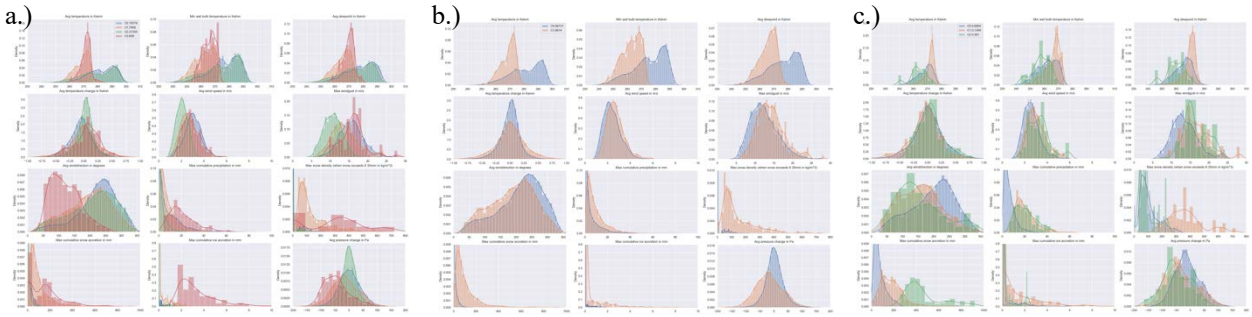
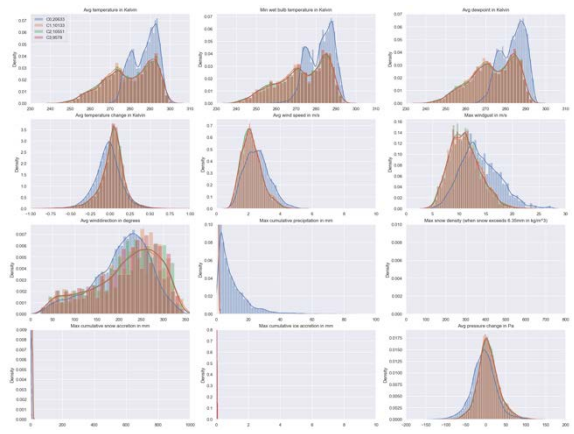
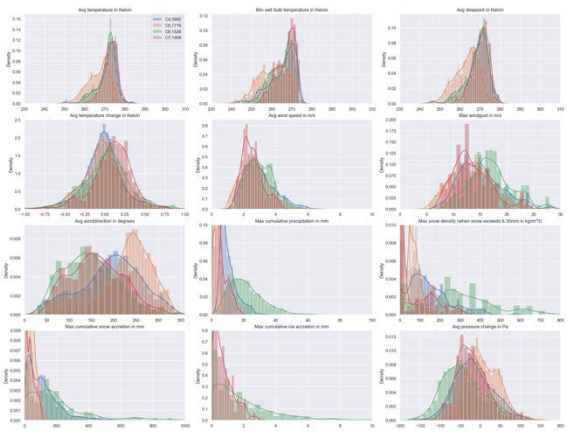


Abbildung 16 Ausprägungen der Wetterparameter der WEP für HAC für die Region Bancroft
a.) regulären Methode b.) erster Schritt der kaskadierten Methode c.) kaskadierten Methode

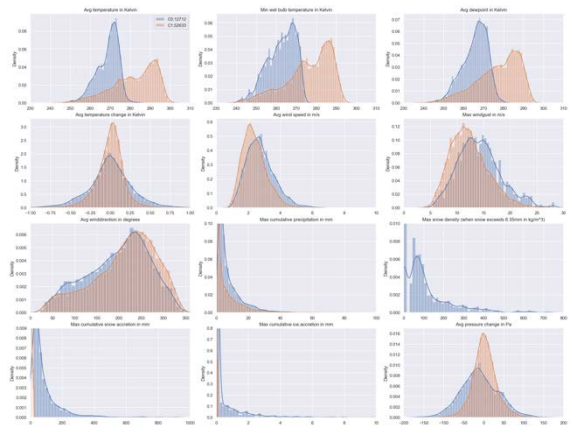
a.)



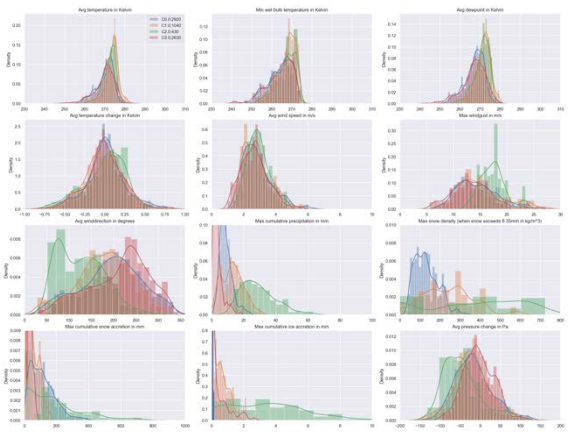
b.)



c.)



d.)



e.)

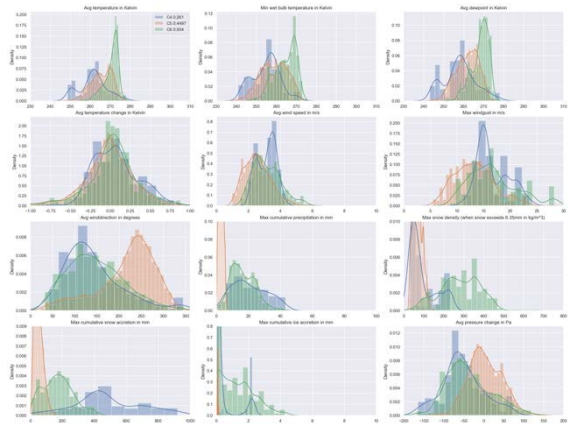
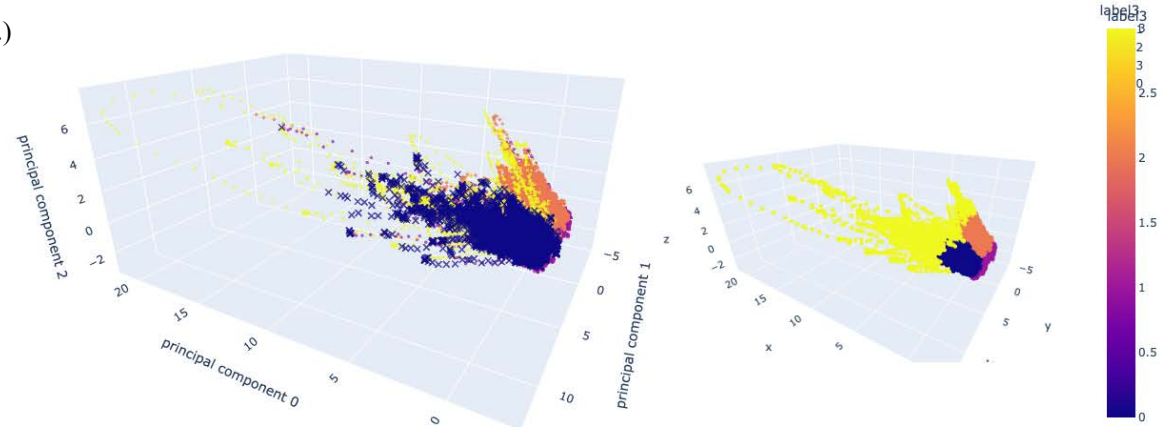
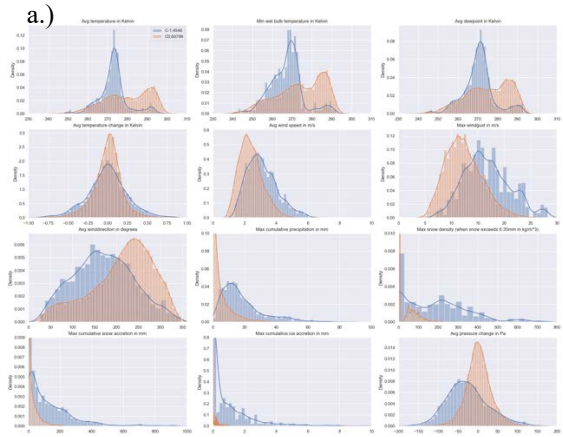


Abbildung 17 Ausprägungen der Wetterparameter der WEP für GMM für die Region Bancroft
a.) regulären Methode (Cluster 0 bis 3)
b.) regulären Methode (Cluster 4 bis 7)
c.) erster Schritt der kaskadierten Methode
d.) kaskadierte Methode (Cluster 0 bis 3)
e.) kaskadierte Methode (Cluster 4 bis 6)
f.) PCA-Methode

f.)



a.)



b.)

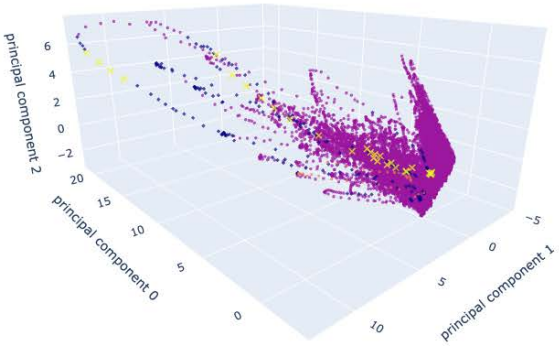


Abbildung 18 Ausprägungen der Wetterparameter der WEP für DBSCAN für die Region Bancroft
a.) reguläre Methode b.) PCA-Methode

Generell ist der Silhouette Score für die kaskadierten Methoden sehr viel höher als für die regulären und für die PCA-Methoden. Die niedrigsten Durchlaufzeiten und damit die höchste Performanz wurden durch die KMeans und DBSCAN Modelle erreicht. Diese besit-zen, wie im Theorieteil bereits deutlich wurde, auch die niedrigste Komplexität. Zudem wurden durch diese Modelle auch am wenigsten Rechenressourcen gebraucht. Unter der Genauigkeit der Clusterergebnisse wird verstanden, wie richtig die Clusterergebnisse verglichen mit reellen historisch dokumentierten Wetterereignissen sind. Mit der Qualität der Ergebnisse ist gemeint, wie detailliert ein WEP durch ein Modell beschrieben werden kann. Für beide Kriterien erzielen KMeans und GMM die besten Ergebnisse. Zudem ist zu erkennen, dass GMM trotz sehr niedrigem Silhouette Score qualitativ sehr hochwertige Clusterergebnisse erzielt. HAC schneidet im Vergleich am schlechtesten ab. Nicht nur sind die Clusterergebnisse qualitativ nicht hochwertig, auch die verbrauchten Rechenressourcen und die Durchlaufzeiten sind unverhältnismäßig hoch.

4.3 Bewertung der Ergebnisse

Die vorgestellten Algorithmen und Methoden wurden erfolgreich eingesetzt, um WEP zu finden und zu definieren. Die WEP konnten erfolgreich charakterisiert und anhand relevanter Parameter und deren Ausprägungen, der Dauer der Wetterereignisse eines WEP und der Anzahl der zugehörigen Datenpunkte charakterisiert werden. Anschließend wurden über Aggregation zeitlicher Nachbarn mit gleichem Label erfolgreich konkrete Wetterereignisse identifiziert. Insbesondere WEP, welche im Frühling und Winter stattfinden und über starke Ausprägungen der Wind- und Niederschlagsparameter charakterisiert sind, konnten erfolgreich identifiziert werden. Einige Wetterereignisse wie Fluten oder Dürren konnten nicht erfasst werden, obwohl diese laut Literatur stattgefunden haben. Auch wurden wenige bis keine Extremwetterereignisse im Sommer identifiziert. Dies kann daran liegen, dass im kanadischen Sommer generell gemäßigteres Wetter stattfindet. Da aber in der Literatur von Stürmen, Fluten, Dürren und Hitzewellen im Sommer berichtet wurde, ist auch möglich, dass diese Wetterereignisse vor allem durch Parameter identifiziert werden, welche nicht in den genutzten Datensätzen vorhanden sind. In der Literatur wird diesbezüglich insbesondere die Luft- und Bodenfeuchte diskutiert. In den Ergebnissen zu den WEP und Wetterereignissen ist zu erkennen, dass einige Wetterereignisse häufig miteinander korrelieren und gemeinsam nacheinander auftreten. Es ist zu analysieren, inwiefern diese Ereignisse zusammenhängen. Denkbar ist, dass einige WEP verschiedene Stufen des gleichen Ereignisses darstellen, bspw. drei Phasen eines Sturmes. Insbesondere ist die Dauer der Ereignisse relevant, welche durch die implementierten Modelle gut analysiert werden kann. Zum Identifizieren von extremen WEP wurde erkannt, dass der erste Schritt der kaskadierten Clustermethode den Datensatz in zwei Cluster teilen sollte, wobei hingegen der größere Cluster ca. 90% der Datenpunkte besitzt. Insgesamt wurden KMeans und GMM als die Algorithmen mit den qualitativ hochwertigsten Ergebnissen zum Definieren von WEP erkannt, wobei KMeans eine geringfügig bessere Performanz und einen höheren Silhouette Score aufweist. HAC ist für den diskutierten Anwendungsfall aufgrund der qualitativ minderwertigen Ergebnisse und der sehr schlechten Performanz irrelevant. DBSCAN ist grundsätzlich zum Identifizieren spezifischer WEP ungeeignet, ist aber hervorragend dazu in der Lage, Wetterextreme von gewöhnlichen Wetterereignissen zu separieren.

4.4 Beantwortung der Forschungsfragen

Im Weiteren wird die in Kapitel 1.2 gestellte Forschungsfrage beantwortet und die Antworten diskutiert.
Welche ML-Algorithmen eignen sich für das Clustering von Wetterdaten zum Definie-ren von WEP? Und welche Algorithmen und Methoden der Clusteranalyse eignen sich am besten zum Definieren von WEP, gemessen anhand der identifizierten Kriterien und welche Möglichkeiten gibt es, diese weiter zu optimieren?

Für das Clustering von Wetterdaten eignen sich insbesondere der KMeans Algorithmus und die Gauß'schen Mischmodelle. Zum Identifizieren grober Wetterereignisse ist die vorgestellte reguläre Methode geeignet, zum Identifizieren von spezifischen Wetterereignisse oder Wet-terextremen hingegen die kaskadierte Methode. Die Möglichkeiten zur weiteren Optimierung der vorgestellten und diskutierten Modelle werden in Kapitel 6 und 7 diskutiert. Als Evaluationskriterien werden die folgenden genutzt. Die Performanz, gemessen über Durchlaufzeiten, CPU-Ausführungszeiten und genutzter RAM, ist geeignet zum Bewerten der eingesetzten Algorithmen. Auch die qualitative Evaluierung über Qualität und Genauigkeit der Ergebnisse ist von hoher Relevanz und geeignet zum Bewerten der Modelle. Hierzu sollten eine erweiterte Referenztabelle oder manuell gelabelte Datensätze erstellt werden, um eine qualitativ hochwertige Auswertung zu ermöglichen. Der Silhouette Score als Evaluationsmatrix zum Bewerten der Aufteilung der Datenpunkte in die Cluster ist nicht optimal geeignet. Es sollten weitere Evaluationsmetriken (wie der Jaccard Coefficient, Fowlkes Index, Mallows Index, Rand Index, Davies−Bouldin Index oder Dunn Index) verprobt und evaluiert werden, um zu analysieren, welche Evaluationsmatrix für diesen Anwendungsfall am besten geeignet ist. Diese Anforderung ist insbesondere deswegen notwendig, da die qualitativen Clusterergebnisse für die GMM-Modelle eine sehr hohe Qualität aufweisen, der Silhouette Score für diese Modelle aber sehr niedrig ist.

4.5 Rückbezug zur Theorie

Der in Grabbe et al. 2014 und Pooja et al. 2020 genutzte EM-Algorithmus konnte im Rahmen dieser Praxisarbeit in Form der Gauß'schen Mischmodelle erfolgreich eingesetzt werden, um Wetterdaten zu clustern und WEP zu definieren. Der Algorithmus konnte erfolgreich auf die diskutierten Methoden (kaskadiert und PCA) angewandt werden. Korrelationsanalysen der Wettercluster wie sie von Grabbe et al. 2014 angemerkt und durchgeführt wurden, werden im Zukunftsausblick dieser Arbeit diskutiert und sind grundsätzlich möglich. Wie auch bereits Grabbe et al. 2014 demonstrieren, werden durch Clustering von Wetterdaten zumeist ein bis wenige Cluster generiert, die einen Großteil der Datenpunkte enthalten und welche unauffälliges und gewöhnliches Wetter enthalten und mehrere Cluster, die deutlich weniger Datenpunkte enthalten, dafür aber auffälliges Wetter oder Extreme aufweisen.
Pooja et al. 2020 nutzen zur Evaluierung der Clusterergebnisse eine ähnliche Methode wie diese Arbeit (Genauigkeit der Feature Selektion, die Genauigkeit der Cluster, die False Positive Rate und die Durchlaufzeit), es sind aber in zukünftiger Forschung weitere und optimalere Evaluationsmethoden zu untersuchen, wie im vorherigen Kapitel diskutiert wurde. Die Erweiterung der Clustermethodik durch bspw. Klassifikation wie in Pooja et al. 2020 oder De Lima et al. 2013 ist in weitergehender Forschung und zur Anwendung in der Praxis sinnvoll. Auch sollten kaskadierte Zwei-Schritt Methodiken untersucht werden, die mehrere Modelle (wie DBSCAN und GMM oder Clustering und Klassifikation) kombinieren. In dieser Arbeit wurde diesbezüglich die bereits von Xu et al. 2015 vorgestellte kaskadierte Methode untersucht, welche insbesondere erfolgreich zum Definieren von Wetterextremen angewandt wurde. Die in Ferstl et al. 2017 und De Lima et al. 2013 genutzten HAC-Modelle erwiesen sich für den konkreten Anwendungsfall der vorgelegten Arbeit als nicht effizient.
De Lima et al. 2013 kombinieren in ihrer Arbeit zudem Dichte-basierte Modelle (wie DBSCAN) mit Klassifikation. In weiterer Forschung sollte der kombinierte Ansatz über DBSCAN auch weiter untersucht werden. Die in Kapitel 2.1.4 diskutierten Wetterkategorien wie sie von Liljequist und Cehak 1984 sowie von Jahn 2015 beschrieben werden, stellten sich im Praxisteil der Arbeit als zu oberflächlich heraus. Es wurden insbesondere detailliertere Wetterereignisse erkannt, die zudem speziell auf die analysierte Region zugeschnitten sind. Mittels der vorgestellten Modelle zum Definieren von WEP können Listen von Wetterkategorien vereinfacht und detailliert für weitere spezifische Regionen der Erde generiert und untersucht werden. Diese können dann den im Theorieteil dieser Arbeit diskutierten Wetterkategorien untergeordnet werden.

4.6 Herausforderungen und Limitationen

Die erzielten Ergebnisse dieser Arbeit werden durch einige Herausforderungen und Limitationen begrenzt. Diese werden in diesem Kapitel diskutiert. Im Rahmen der vorgelegten Arbeit war es aufgrund zeitlicher Einschränkungen und dem begrenzten Umfang nicht möglich, alle Daten der in Kapitel 4.1 vorgestellten Regionen Kanadas zu analysieren. Diese könnten jedoch weitere wertvolle Erkenntnisse zum Thema liefern und sollten in zukünftiger Forschung ausgewertet werden. Zudem wurden in dieser Arbeit nur Wetterdaten analysiert, welche typisch für das Klima in Ontario und durch Kälte, Schnee, Eis und starke Winde gekennzeichnet sind. Eine erweiterte Anwendung und Auswertung der Modelle auf weitere Klimaregionen der Erde ist erforderlich, um die Modelle weiter zu evaluieren und zu optimieren. In weitergehenden Analysen sollte zudem ein vollständiges Spektrum an Wetterparametern genutzt werden, wie es in Kapitel 2.1.3 vorgestellt wurde. Insbesondere die im Praxisteil dieser Arbeit fehlenden Parameter Luftfeuchte, Bodenfeuchte und Sonneneinstrahlung könnten große Auswirkungen auf die Analyse haben und besonders relevant zum Identifizieren von Wetterereignissen wie Fluten, Dürren oder Hitzewellen sein. Auch die manuelle, qualitative Auswertung der Analyseergebnisse im Praxisteil dieser Arbeit erwies sich aufgrund fehlenden meteorologischen Fachwissens als Herausforderung. Die generierten Ergebnisse in Form von WEP sollten im Weiteren von meteorologischen Experten ausgewertet und charakterisiert werden. Zuletzt war im Umfang der Arbeit auch eine vollumfängliche Auswertung der PCA-Methode angedacht. Aufgrund der zeitlichen Beschränkung der Arbeit konnte diese allerdings nicht fertig umgesetzt werden.

4.7 Kritische Analyse der erzielten Ergebnisse

Vorteile der vorgestellten Modelle sind, dass diese insbesondere Wetterextreme im Detail charakterisieren und identifizieren können. Die vorgestellten Clusteranalysen können auf Grundlage der vorherrschenden Wetterparameter zu anderen Ergebnissen für verschiedene Klimaregionen der Erde führen. Obwohl in dieser Arbeit ein Schwerpunkt auf die Ergebnisse der kaskadierten Methode gelegt wurde, ist auch eine detailliertere Betrachtung der Clusterergebnisse der regulären Clusteringmethode interessant. So ist denkbar, diese nach Jahreszeiten zu sortieren, um noch aufschlussreichere Aussagen über diese Wetterereignisse treffen zu können. Bspw. wäre es denkbar, dass die größeren Cluster der regulären Methoden in gewöhnliche milde Ereignisse verschiedener Jahreszeiten gliederbar sind. Eine Analyse, aufgegliedert nach Jahreszeiten, ist in weitergehender Forschung notwendig. In der Literatur wird eine Vielzahl weiterer Cluster-Algorithmen genannt, welche aufgrund der geringeren Relevanz nicht weiter betrachtet wurden. Eine Anwendung und Evaluierung dieser Algorithmen auf den spezifischen Anwendungsfall ist in weiterer Forschung notwendig. Hier sind insbesondere Affinity Propagation, Mean-Shift, Ward Hierarchical Clustering, OPTICS und BIRCH (Scikit Learn 2023a) interessant. Bei der Betrachtung der Ergebnisse fällt zudem auf, dass die GMM, trotz ihres geringen Silhouette Scores, qualitativ sehr hochwertige und genaue Ergebnisse liefern. Es liegt daher nahe, dass der Silhouette Score nicht die optimale Evaluationsmetrik zur Bewertung der Clusterergebnisse von Wetterdaten ist. In weitergehender Forschung ist die Anwendung weiterer Evaluationsmetriken zu analysieren, um jene zu identifizieren, die für den Anwendungsfall des Clusterns von Wetterdaten am geeignetsten sind. Interessant sind hierbei insbesondere der Jaccard Coefficient, der Fowlkes Index, der Mallows Index, der Rand Index, der Davies–Bouldin Index, der Dunn Index (Zhou 2021) und die von De Lima et al. 2013 vorgestellte Similarity Metric. Auch sollte in zukünftiger Forschung die Skalierbarkeit der Modelle in Bezug auf die geographische Größe der zu analysierenden Regionen untersucht werden, um festzustellen, welche räumliche Granularität am geeignetsten zur Definition von WEP ist. Dies ist umso mehr der Fall, da ähnliche WEP für verschiedene Regionen Ontarios identifiziert wurden. Zu-dem besteht die Notwendigkeit der Auswertung der Analyseergebnisse durch meteorlogische Fachexperten, um komplexe Zusammenhänge der Wetterparameter und auch geographische Ursachen wie den nordamerikanischen Lake Effect adäquat auswerten zu können. Auch sollte der Ansatz des Clusterings mit weiteren Methoden zum Definieren von WEP detaillierter verglichen werden. Insbesondere die Klassifikation und Assoziationsregeln sind hierfür relevant. Nach der erfolgreichen Anwendung der kaskadierten Methode in der vorgelegten Arbeit, ist zudem zu untersuchen, welche weiteren Zwei-Schritt-Methoden zur Optimierung der Analyse genutzt werden können. Als Beispiel sei hier die bereits diskutierte Verbindung von Clustering und Klassifikation oder eine Verbindung verschiedener Clusteringalgorithmen genannt. Die qualitativ hochwertigen Ergebnisse der regulären DBSCAN Methode könnten bspw. als erster Schritt des kaskadierten Modells genutzt werden, um die Stärken des Dichtebasierten Modells zur Identifikation von Ausreißern auszunutzen. In einem zweiten Schritt könnte dann KMeans oder GMM genutzt werden, welche besonders eine Stärke darin zeigen, qualitative und genaue Ergebnisse zur Identifizierung von Wetterextremen zu erzielen. Zuletzt muss auch der Einsatz der PCA-Methode vollumfänglich, auch unter Einsatz weiterer Algorithmen und Methoden, verprobt und evaluiert werden.

5 ZUSAMMENFASSUNG UND AUSBLICK

In der vorgelegten Arbeit wurden auf Grundlage der durchgeführten Literaturrecherche zur Clusteranalyse von Wetterdaten zum Definieren von WEP relevante Clusteringalgorithmen und Methoden. unter Betrachtung und Optimierung der in der Literatur identifizierten relevanten Faktoren, auf verschiedene Regionen Kanadas angewandt. Anschließend wurde mittels maschinellen Lernens von Clusteranalysen Profile für verschiedene Wetterereignisse zu identifiziert. Zur Einlösung der Zielsetzung wurde die Forschungsmethodik Design Science Research ergänzt durch iteratives Prototyping und die Kreuzvalidierung verwendet. Durch iterative Implementierung von regulären, kaskadierten und Hauptkomponentenanalyse-Modellen wurden erfolgreich WEP identifiziert. Dies geschah unter Anwendung der Clustering Algorithmen KMeans, hierarchisches agglomeratives Clustering, Gauß'sche Mischmodelle und DBSCAN. Geographisch eingeschränkt wurde die Analyse auf die Regionen Ontarios unter Betrachtung relevanter identifizierter Parameter und der zeitlichen Granularität der Wetterereignisse. Die verschiedenen eingesetzten Algorithmen und Methodiken wurden unter den Aspekten der Performanz, anhand einer Evaluationsmatrix sowie der Qualität und Genauigkeit der Ergebnisse verglichen. Gemessen wurden diese Kriterien in Form der Durchlaufzeiten, der CPU-Ausführungszeiten und der genutzten Rechenressourcen, des Silhouette Scores und einer manuellen, qualitativen Auswertung der Qualität und Genauigkeit der Modelle. So konnten Stärken und Schwächen der jeweiligen Modelle eruiert werden und das kaskadierte KMeans und das kaskadierte GMM-Modell als die geeignetsten für die Definition von extremen WEP identifiziert werden. Mit den regulären KMeans und den regulären GMM-Modellen konnten insbesondere gemäßigte und gewöhnliche Wetterereignisse identifiziert werden.

In weitergehender Arbeit mit Wetterdaten können die identifizierten WEP und die implementierten Modelle für eine Reihe von Anwendungsfällen in der Forschung und Praxis genutzt werden. Die WEP mit ihren Charakteristiken können bspw. in Anwendung auf Wettervorhersagen untersucht werden, um detaillierte Informationen (wie ausschlaggebende und gefährliche Ausprägungen bestimmter Parameter oder Handlungsempfehlungen bei

Wetterextremen) zu vorhergesagtem Wetter zu bekommen und diese Informationen in einem Dashboard als Warnung anzuzeigen. Weiterhin können bestehende WEP über eine Korrelationsanalyse mit weiteren Faktoren untersucht werden, um bspw. effiziente Maßnahmen gegen die Auswirkungen von Katastrophen und Schäden bei bestimmen Wetterereignissen zu untersuchen. Hierzu können Modelle für diverse WEP gebaut werden. Denkbar sind Modelle zur Vorhersage und zur Vorbeugung von Stromausfällen bei bestimmten Wetterereignissen (Eskandarpour und Khodaei 2016). Weiterhin kann in zukünftiger Forschung die Erkennung und Vorhersage der definierten WEP mittels Wetterdaten untersucht werden und die Effektivität von ML-Modellen zur Wettervorhersage mit physikalischen Methoden verglichen werden. Es besteht also eine Vielzahl von Anwendungsfällen, in denen die implementierten Artefakte dieser Arbeit, sowie die identifizierten WEP in zukünftiger Forschung genutzt werden können.

LITERATUR

Abdi, H. / Williams, L.J. (2010): Principal component analysis, in: Wiley interdisciplinary re-views: computational statistics, 2(4), pp.433-459.

Ackerman, S. / Knox, J. (2011): Meteorology. Jones & Bartlett Publishers.

Akande, A. / Costa, A.C. / Mateu, J. / Henriques, R. (2017): Geospatial analysis of extreme weather events in Nigeria (1985–2015) using self-organizing maps, in: Advances in Meteorology, 2017.

Ban, Z. / Liu, J. / Cao, L. (2018): Superpixel segmentation using Gaussian mixture model, in: IEEE Transactions on Image Processing, 27(8), pp.4105-4117.

Bellman, R. (1957): Dynamic programming. Princeton University, NJ, Princeton University Press, New Jersey.

Chandola, V. / Banerjee, A. / Kumar, V. (2009): Anomaly detection: A survey, in: ACM com-puting surveys (CSUR). Jul 30;41(3):1-58.

Chu, X. / Ilyas, I.F. / Krishnan, S. / Wang, J. (2016): Data cleaning: Overview and emerging challenges, in: Proceedings of the 2016 international conference on management of data (pp. 2201-2206).

Cui, M. (2020): Introduction to the k-means clustering algorithm based on the elbow method. Accounting, Auditing and Finance, 1(1), pp.5-8.

Das, S. / Sun, X. (2014): Investigating the pattern of traffic crashes under rainy weather by association rules in data mining, in: Transportation Research Board 93rd Annual Meeting (No. 14-1540). Transportation Research Board Washington DC.

de Lima, Glauston, R.T. / Stephan, S. (2013): A new classification approach for detecting severe weather patterns, in: Computers & geosciences 57 (2013): 158-165.

ECMWF (2023a): ERA5: data documentation. URL: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation, Abruf: 01.03.2023, 13:36 Uhr

ECMWF (2023b): ERA5: reanalysis datasets for forecasts. URL: https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, Abruf: 01.03.2023, 13:44 Uhr

ECMWF (2023c): ERA5: data documentation parameterlistings. URL: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5: datadocumentation-Parameterlistings, Abruf: 01.03.2023, 13:59 Uhr

Epstein, E.S. (1969): A scoring system for probability forecasts of ranked categories, in: Journal of Applied Meteorology (1962-1982), 8(6), pp.985-987.

Eskandarpour, R. / Khodaei, A. (2016): Machine learning based power grid outage predic-tion in response to extreme events, in: IEEE Transactions on Power Systems, 32(4), pp.3315-3316.

Fathi, M. / Haghi Kashani, M. / Jameii, S. M. / Mahdipour, E. (2022): Big Data Analytics in Weather Forecasting: A Systematic Review, in: Archives of Computational Methods in Engineering 29.2 (2022, Springer): 1247–1275

Ferstl, F. / Kanzler, M. / Rautenhaus, M. / Westermann, R. (2017): Time-hierarchical clus-tering and visualization of weather forecast ensembles, in: IEEE transactions on vis-ualization and computer graphics, 23(1), pp.831-840.

Firdaus, S. / Uddin, M.A. (2015): A survey on clustering algorithms and complexity analysis, in: International Journal of Computer Science Issues (IJCSI), 12(2), p.62.

Géron, A. (2019): Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

Ghirardelli, J.E. (2005): An Overview of the Redeveloped Localized Aviation Mos Program (Lamp) For Short-Range Forecasting.

Giordani, P. / Ferraro, M.B. / Martella, F. (2020): Introduction to Clustering. Springer Singa-pore.

Grabbe, S.R. / Sridhar, B. / Mukherjee, A. (2014): Clustering days with similar airport weather conditions, in: 14th AIAA Aviation Technology, Integration, and Operations Conference (p. 2712).

Gregor, S. / Hevner, A.R. (2013): Positioning and Presenting Design Science Research for Maximum Impact, in: MIS Quarterly, Jg. 37, Nr. 2, S. 337-355

Government of Canada (2022): Canada's top 10 weather stories of 2022. URL: https://www.canada.ca/en/environment-climate-change/services/top-ten-weather-stories/2022.html, Abruf: 20.04.2023, 14:05 Uhr

Hasan, N. / Uddin, M.T. / Chowdhury, N.K. (2016): Automated weather event analysis with machine learning, in: International Conference on Innovations in Science 2016, Engi-neering and Technology (ICISET) (pp. 1-5). IEEE.

Hegland, M. (2003): Algorithms for Association Rules, in: Mendelson, S., Smola, A.J. (eds) Advanced Lectures on Machine Learning. Lecture Notes in Computer Science(), vol 2600. Springer, Berlin, Heidelberg.

Hersbach, H. / Bell, B. / Berrisford, P. / Hirahara, S. / Horányi, A. / Muñoz-Sabater, J. / Nicolas, J. / Peubey, C. / Radu, R. / Schepers, D. / Simmons, A. (2020): The ERA5 global reanalysis, in: Quarterly Journal of the Royal Meteorological Socie-ty, 146(730), pp.1999-2049.

Hevner, A. / Chatterjee, S. (2010): Design Research in Information Systems, Theory and Practice. Hrsg. von R. Sharda/S. Voß. Bd. 22. Integrated Series in Information Sys-tems. New York, NY, USA: Springer New York, NY.

Hevner, A. / March, S.T. / Park, J. / Ram, S. (2004): Design Science in Information Systems Research, in: MIS Quaterly 28.1, S. 75–105.

Hjelmfelt, M.R. (1990): Numerical study of the influence of environmental conditions on lake-effect snowstorms over Lake Michigan, in: Monthly Weather Review, 118(1), pp.138-150.

Holmstrom, M. / Liu, D. / Vo, C. (2016): Machine learning applied to weather forecasting. Meteorology. Appl. Dec; 10: 1-5.

Horenko, I. / Dolaptchiev, S.I. / Eliseev, A.V. / Mokhov, I.I. / Klein, R. (2008): Metastable decomposition of high-dimensional meteorological data with gaps, in: Journal of the atmospheric sciences, 65(11), pp.3479-3496.

Hupfer, P. / Kuttler, W. (2005): Witterung und Klima. Eine Einführung in die Meteorologie und Klimatologie, 11. Auflage

Jahn, M. (2015): Economics of extreme weather events: Terminology and regional impact models. Weather and Climate Extremes, 10, pp.29-39.

Jo, J.M. (2019): Effectiveness of normalization pre-processing of big data to the machine learning performance, in: The Journal of the Korea institute of electronic communica-tion sciences, 14(3), pp.547-552.

Kassambara, A. (2017): Practical guide to cluster analysis in R: Unsupervised machine learning, 1. Auflage, Sthda.

Kotsiantis, S. / Kanellopoulos, D. (2006): Association rules mining: A recent overview, in: GESTS International Transactions on Computer Science and Engineering. 2006 Jan;32(1): 71-82.

Liljequist, G.H. / Cehak, K. (1984): Allgemeine Meteorologie. 3. Auflage, Springer-Verlag.

Liu, F. / Deng, Y. (2020): Determine the number of unknown targets in open world based on elbow method, in: IEEE Transactions on Fuzzy Systems, 29(5), pp.986-995.

Liu, F. / Ting, K.M. / Zhou, Z.H. (2012): Isolation-based anomaly detection, in: ACM Trans Knowl. Discov. Data 6(1): Article 3

Mitchell, T. (1997): Machine learning. McGraw Hill, New York

Moon, T.K. (1996): The expectation-maximization algorithm, in: IEEE Signal processing mag-azine, 13(6), pp.47-60.

Pelosi, A. / Terribile, F. / D'Urso, G. / Chirico, G.B. (2020): Comparison of ERA5-Land and UERRA MESCAN-SURFEX reanalysis data with spatially interpolated weather ob-servations for the regional assessment of reference evapotranspiration. Water, 12(6), p.1669.

Pooja, S.B. / Balan, R.S. / Anisha, M. / Muthukumaran, M.S. / Jothikumar, R. (2020): Techniques Tanimoto correlated feature selection system and hybridization of clus-tering and boosting ensemble classification of remote sensed big data for weather forecasting. Computer Communications, 151, pp.266-274.

Poteraş, C.M. / Mihăescu, M.C. / Mocanu, M. (2014): An optimized version of the K-Means clustering algorithm, in 2014 Federated Conference on Computer Science and Infor-mation Systems (pp. 695-699). IEEE.

Ray, P. (ed) (2015): Mesoscale meteorology and forecasting. Springer.

Runkler, T.A. (1999): Probabilistische und Fuzzy Methoden für die Clusteranalyse, in: Seis-ing, R. (eds) Fuzzy Theorie und Stochastik. Computational Intelligence. Vie-weg+Teubner Verlag, Wiesbaden.

Scikit Learn (2023a): Clustering. URL: https://scikit-learn.org/stable/modules/clustering.html, Abruf: 07.03.2023, 14:33 Uhr

Scikit Learn (2023b): Preprocessing. URL: https://scikit-learn.org/stable/modules/preprocessing.html, Abruf: 19.04.2022, 16:37 Uhr

Sagiroglu, S. / Sinanc, D. (2013): Big data: A review, in: International conference on collabo-ration technologies and systems (CTS) 2013 May 20 (pp. 42-47). IEEE.

Savaresi, S.M. / Boley, D.L. / Bittanti, S. / Gazzaniga, G. (2002): Cluster selection in divi-sive clustering algorithms, in: Proceedings of the 2002 SIAM International Confer-ence on Data Mining (pp. 299-314). Society for Industrial and Applied Mathematics.

Spektrum Akademischer Verlag, Heidelberg, (2000): Lexikon Der Geowissenschaften: Atmosphäre. URL: https://www.spektrum.de/lexikon/geowissenschaften/atmosphaere/1060, Abruf: 23.02.2023, 13:46 Uhr

Syakur, M. A. / Khotimah, B. K. / Rochman, E. M. S. / Satoto, B. D. (2018): Integration k-means clustering method and elbow method for identification of the best customer profile cluster, in: IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.

The Weather Network (2022): The Weather Network. URL: https://www.theweathernetwork.com/en/news/weather/, Abruf: 24.04.2023, 15:06 Uhr

Thudumu, S. / Branch, P. / Jin, J. / Singh, J. (2020): A comprehensive survey of anomaly detection techniques for high dimensional big data, in: Journal of Big Data. Dec;7: 1-30.

Fang, W. / Sheng, V.S. / Wen, X. / Pan, W. (2014): Meteorological data analysis using mapreduce, in: The Scientific World Journal, 2014.

Webster, J. / Watson, R.T. (2002): Analyzing the past to prepare for the future: Writing a literature review, in: MIS quarterly. Jun 1: xiii-xiii.

Xu, Q. / He, D. / Zhang, N. / Kang, C. / Xia, Q. / Bai, J. / Huang, J. (2015): A short-term wind power forecasting approach with adjustment of numerical weather prediction in-put by data mining, in: IEEE Transactions on sustainable energy, 6(4), pp.1283-1291.

Yuan, C. / Yang, H. (2019): Research on K-value selection method of K-means clustering algorithm. J, 2(2), pp.226-235.

Zhou, Z.H. (2021): Machine learning. Springer Nature.

KONTAKT

JULIAN L.R. ERATH, geboren am 21.01.2001 in Sindelfingen, Deutschland, absolvierte 2023 seinen Bachelor of Science in Wirtschaftsinformatik Data Science an der Dualen Hochschule Baden-Württemberg in Stuttgart. Seit 2020 arbeitet er bei der IBM Deutschland GmbH, mit welcher er sein duales Studium abschloss. Seit Abschluss seines Bachelorstudiums arbeitet er als Data Consultant bei der IBM Deutschland GmbH und absolviert berufsbegleitend seinen Master of Science in Data Science an der Hochschule der Medien in Stuttgart. Julian.Erath@ibm.com