Same Same but Different? - Machine Learning Algorithmen zur Indikatoren basierten Regionenbildung

Thomas ZIPPERLE[1], Clara ORTHOFER[1]

Technische Universität München, Arcissstraße 21, 80333 München, +49 (89) 289 - 23948, thomas.zipperle@tum.de, www.ewk.ei.tum.de

Kurzfassung:

Um globale Herausforderungen, wie den Klimawandel, mathematisch zu analysieren und auf Lösungswege zu untersuchen, braucht es globale und komplexe Systemmodelle. Um die Lösbarkeit solch großer Modelle zu gewehrleisten, ist es notwendig die modellierten Länder zu Regionen zusammenzufassen. Traditionell werden solche Regionen für Energiesystem-modelle auf Basis geografischer Zusammengehörigkeit sowie der historischen (energie-) wirtschaftlichen sowie politischen Entwicklung gebildet. In der vorliegenden Analyse wird geprüft, ob diese historisch definierten Regionen auch in Anbetracht des drastischen energiewirtschaftlichen Wandels vor dem die Weltgemeinschaft steht, (sollten die Bestrebung, die globale Erwärmung auf deutlich unter 2 °C im Vergleich zu vorindustriellen Levels zu begrenzen, in Taten umgesetzt werden) noch Gültigkeit haben. Dabei wird eine Regionenbildung mittels Machine Learning Algorithmen auf Basis der Solar- und Wind-energiepotentiale vorgeschlagen und mit den Regionen traditioneller Energiesystemmodelle verglichen. Im Zuge der Analyse werden sechs ausgewählte und teils eigens für die Analyse generierte Features zu zwei verschiedenen Sets kombiniert. Beide Sets werden sowohl mittels *KMeans* als auch *GaussianMixture* geclustert. Die Gegenüberstellung der Ergebnisse deuten darauf hin, dass die Methoden des *Machine Learning* als unterstützendes Tool, jedoch nicht als Ersatz für die traditionelle Bildung von Metaregionen eingesetzt werden können.

Keywords: Energiesystemmodellierung, Machine Learning, Clusteranalyse, Metaregionen, Klimawandel, MESSAGE

1 Einleitung

Im Rahmen des bei der UN-Klimakonferenz 2015 in Paris geschlossenen Klimaabkommens, hat sich die ein Großteil der Weltgemeinschaft dazu verpflichtet, gemeinsam aktiv zu werden um die anthropogene globale Erderwärmung auf deutlich unter 2°C gegenüber vor-industrieller Zeit zu begrenzen [27]. Abbildung 1 zeigt, dass die Umsetzung dieses Ziels nach drastischen Maßnahmen verlangt, und dass eine Trendumkehr in der Entwicklung der glo-balen Treibhausgasemissionen unabdingbar ist, um eine Klimakatastrophe zu vermeiden. Dem Energiesektor, mit einem Anteil von über 70% an den globalen Emissionen, kommt dabei eine besonders wichtige Rolle zu [28].

[1] Jungautoren

Bei der letzten Klimakonferenz in Katowice wurde ein Regelwerk zur weltweite Umsetzung des Pariser Klimaabkommens getroffen. Dabei wurde vereinbart, wie künftig Emissionen gemessen, dokumentiert und berichtet werden. Jedoch konnte auch bei dieser Konferenz noch keine Einigung darüber erzielt werden, welcher Staat wie viele Emissionen künftig ausstoßen darf beziehungsweise einsparen muss. Klar ist aber, dass diese Größe an die nationalen Gegebenheiten, also wirtschaftliche Stärke und Emissionsintensität, aber auch Potentiale zur erneuerbaren Energieversorgung beinhalten werden.

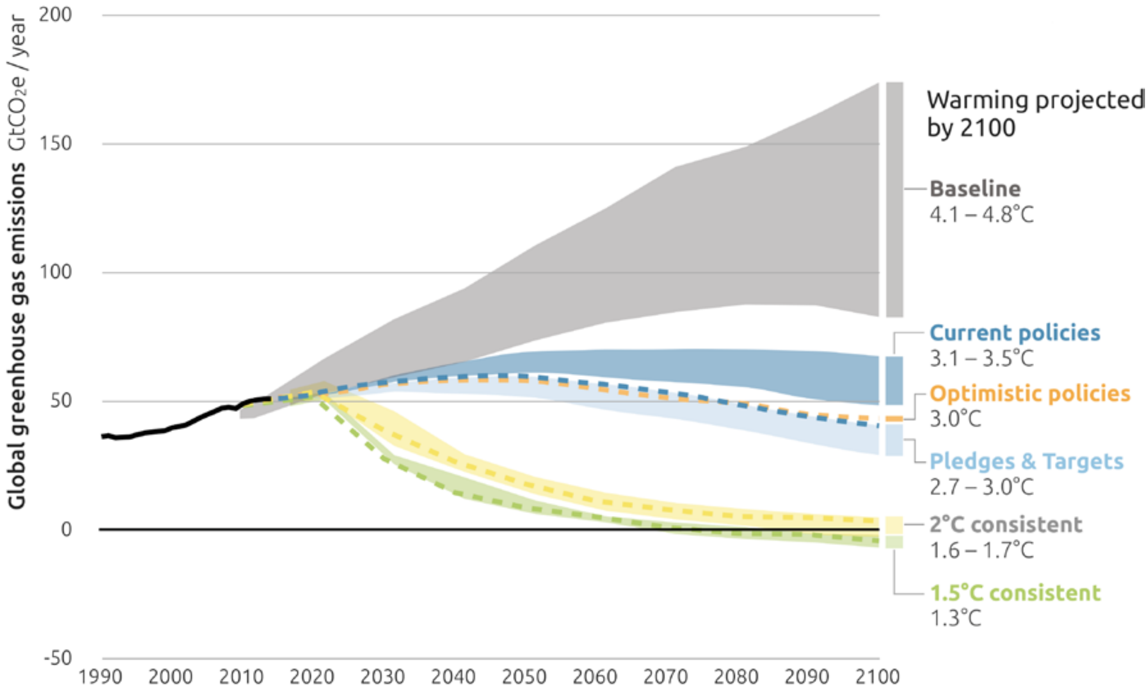

Abbildung 1: Szenarios zur Entwicklung der globalen Treibhausgasemissionen bis 2100 – Die berechneten Emissionen und erwartete korrelierende Erwärmung basieren auf den nichtbindenden Zusagen und aktuellen Richtlinien der berücksichtigten Staaten. Climate Action Tracker [25].

Globale Energiesystemmodelle welche im Gegensatz zu regional begrenzten Modellen in der Lage sind globale Rückkopplungseffekte abzubilden, helfen dabei, Szenarios zur Erreichung des 2°C-Ziels zu entwickeln. Um aussagekräftige Ergebnisse zu erzielen, müssen sie aber nicht nur alle Länder der Welt, sondern auch lange Zeithorizonte umfassen. Um die Berechenbarkeit solch großer Modelle zu gewährleisten, ist es notwendig Nationalstaaten zu Metaregionen zusammenzufassen.

Traditionell werden in der Energiesystemmodellierung Regionen von Experten auf Basis ihrer langjährigen Erfahrung in der Modellierung gebildet. Die Staaten werden dabei entsprechend ihrer geographischen (z.B. „*Afrika*“ / „*Sub-Sahara Afrika*“ [1-6]) beziehungsweise politisch-historischen Zusammengehörigkeit (z.B. „*Former Soviet Union*“ [1-4, 6]) zusammengefasst (Tabelle 2). Aus diesen Metaregionen werden wiederum auf Basis der historischen Energieverbrauchsstruktur, Staatsform und Staatshistorie, sowie Mitgliedschaften in Handelsabkommen, ökonomische Situation, demografische Entwicklung, politische Ambitionen und energie- sowie klimapolitische Ziele eines Staates, Regionen entwickelt (vgl. [3,4,6]). Abbildung 2 zeigt beispielhaft die häufig genutzte Aufteilung der Welt in elf Regionen, wie sie für viele globale Szenarioanalysen mit dem Energiesystemmodells MESSAGE genutzt wird [6].

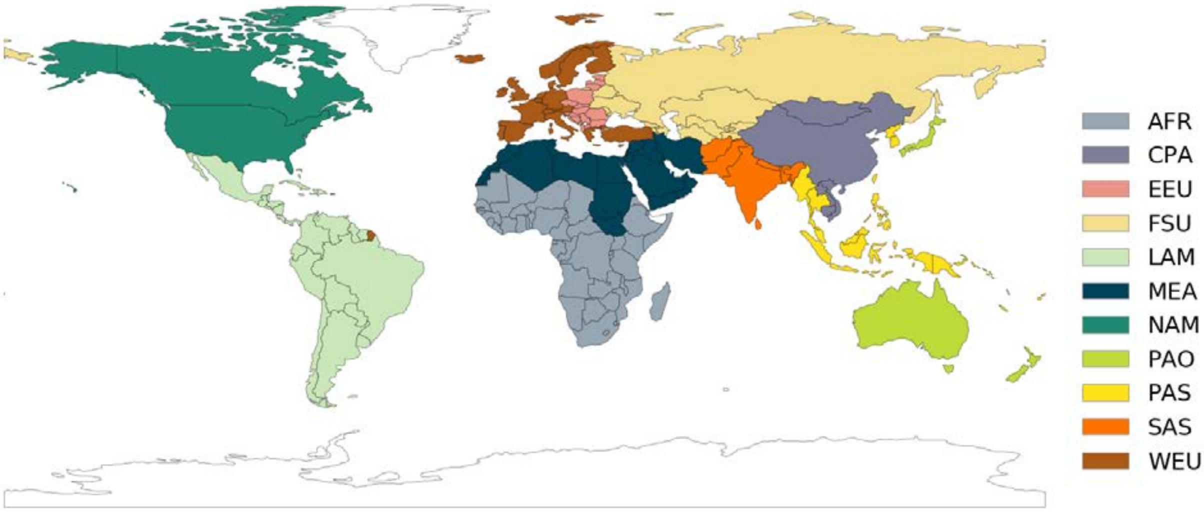

Abbildung 2: Metaregionen des Energiesystemmodells MESSAGE [6], eigene Darstellung.

Solch klassische Regionenstrukturen haben sich in der Enrgiesystemmodllierung über Dekaden bewährt und wurden daher von Modellgeneration zu Modellgerneration „vererbt". In anbetracht der drastischen Veränderungen, vor denen die meisten Energiesysteme der Welt heute stehen, sind diese Regionen jedoch zu reevaluieren und auf ihre Gültigkeit zu überprüfen. Denn, um das 2°C Ziel zu erreichen, muss die heute noch von fossilen Energieträgern dominierte globale Energiebereitstellung auf so gut wie aussschließlich CO_2-arme erneuerbare Energien umgestellt werden. Dieser Wandel wird neben den Enerigesysteme der einzelnen Nationalstaaten auch den internationalen Energiemarkt stark beeinflussen. Denn, im Gegensatz zu leicht handel- und transportierbaren fossilen Energieträgern, sind die erneuerbaren Energiepotenitale ortsgebunden und damit nur in transfomierter Form handelbar (z.B. Strom aus Wind- und Solarenergie, Wasserstoff aus Grünstrom, ect.).

Im Gegensatz zu bisher, als die historisch gewachsenen Handelsbeziehungen und –infrastrukturen sowie die geographisch sehr konzentriert auftretenden fossilen Energiereserven die dominierenden Faktoren in der Entwicklung der Energiesysteme waren, könnten in Zukunft die erneuerbaren Enerigenpotentiale die limitierenden und bestimmenden Faktoren werden. Um also für die Energiesystemmodellierung auch zukünftig noch relevante Ländergruppen zu identifizieren, wird in dieser Arbeit ein neuer Ansatz zur Bildung von Metaregionen untersucht. Dabei wird auf Basis einer breiten mit Länderdaten befüllten Datenbank eine Clusteranalyse der erneuerbaren Energieressourcen, energiewirtschaftlichen Veränderungspotentiale sowie der wichtigsten makroökonomischen Parametern durchgeführt.

2 Methodik

Die vorliegende Analyse basiert auf einer Datenbank, welche mit online verfügbar Datensätzen zu Energiebereitstellung, -verbrauch und -handel, zu erneuerbaren Energiepotentialen, zu konventionellen und unkonventionellen fossilen Ressourcen und Reserven sowie zu den wichtigsten makroökonomischen Parametern befüllt wurde. Einen detaillierten Überblick über die recherchierten Daten gibt Tabelle 3.

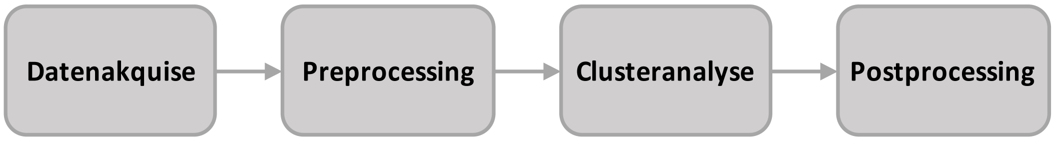

Abbildung 3: Schematisches Ablaufdiagramm der Clusteranalyse zur Bildung von Metaregionen mittels Machine Learning Algorithmen.

2.1 Preprocessing: Explorative Datenanalyse & Feature Engineering

Hauptbestandteil des Preprocessings ist das sogenannte Feature Engineering. Ein Feature, auch Merkmal genannt, beschreibt dabei eine messbare Größe (z.B. Endenergieverbrauch) oder eine zuordenbare Eigenschaft (z.B. Kontinent). Der Prozess der Wandlung der Rohdaten in einen Feature-Vektor wird als Feature Engineering bezeichnet [26].

Im Rahmen dieser Arbeit wurden die Rohdaten, welche in unterschiedlichen Formaten vorlagen, auf ein einheitliches Datenformat gebracht und in einer HDF-Datei (*Hierarchical Data Format)* abgelegt. Informationen welche in Form von Textwerten (z.B. Kontinente) vorlagen wurden anhand des *OneHotEncoder* von [23] für die Clusteranalyse aufbereitet. Bei dieser Zeichenkodierung wurde für jedes Element des Merkmals eine Spalte angelegt und mit 0 (nicht vorhanden) oder 1 (vorhanden) kodiert.

Im Anschluss wurden die gesammelten Daten im Rahmen einer explorativen Datenanalyse visuell aufbereitet und so auf Konsistenz geprüft. Identifizierte Datenlücken wurden soweit möglich mittels Extrapolation oder Expertenabschätzungen geschlossen. Fehlerhafte Daten wurden, sofern korrekte Werte verfügbar waren, korrigiert oder andernfalls aus dem Datensatz entfernt. Ein weiteres Ziel der visuellen explorativen Datenanalyse ist es ein Verständnis über mögliche Zusammenhänge zu entwickeln und so relevante Einflussparameter aus den Eingangsdatensätzen herauszuarbeiten und so die energiewirtschaftlich relevanten Parameter zu identifizieren und neue, kombinierte Features aus den bestehenden Datensätzen abzuleiten. Zusätzlich wurde eine Hauptkomponentenanalyse durchgeführt um die mehrdimensionalen Daten zu strukturieren und zu veranschaulichen. Da aber keine offensichtlichen visuellen Strukturen in den Daten identifiziert werden konnten, wurden manuell einzelne Merkmale zu komplexen Features kombiniert um die dem Clusteralgorithmus zugeführte Feature-Zahl auf ein relevantes Minimum zu reduzieren ohne dabei übermäßig relevante Informationen zu verlieren. Diese Reduktion ist essentiell um die Interpretierbarkeit der Ergebnisse des Prozesses der Clusterbildung zu gewährleisten.

Ziel der durchgeführten Analyse war, die Clusterung von Ländern aufgrund ihrer strukturell ähnlichen erneuerbaren Energiepotentiale und, so die Hypothese, einer potentiell ähnlichen zukünftigen Energiesystementwicklung. Dieser Hypothese folgend, wurde die Clusteranalyse mit Fokus auf die Wind- und Solarenergiepotentiale ausgeführt, da diese beiden Potentiale einerseits global die größten technischen potentiale aufweisen und andererseits aktuell die am schnellsten wachsende Energiesparte ausmachen. Da die Beschreibung der Wind und Solarenergiepotentiale sehr komplex ist und aus einer länderspezifischen Kombination mehrere duzend Einzelparametern besteht, wurden im Rahmen des Feature Engineering fünf Meta-Features herausgearbeitet.[2]

Zur Kategorisierung der Länder nach der zur Verfügung stehenden erneuerbaren Energiemengen wurden die vorhanden Potentiale zu einem **Solar**- und einem **Windenergiefaktor** zusammengefasst. Diese Faktoren beschreiben jeweils die Relation zwischen absolut verfügbaren Energiepotentialen eines erneuerbaren Energieträgers zum Endenergieverbrauch. Dazu wurden die einzelnen Potentialkategorien summiert und das Verhältnis zum Endenergieverbrauch berechnet.

Als Indikatoren zur Unterscheidung der Länder entsprechend der „Qualität" ihrer Solar- und Windenergiepotentiale wurden die mittleren **gewichteten Vollaststunden** und daraus wiederum die resultierenden **gemittelten Kosten** für die einfache sowie die fünffache Deckung des Endenergiebedarfs berechnet. Dazu wurden für jede Potentialkategorie entsprechend der Region, des Technologietyps, der Vollaststunden sowie Distanz zum Verbraucher spezifische Stromgestehungskosten berechnet. Anschließend wurden die Potentiale den spezifischen Kosten entsprechend sortiert und alle bis zur bis zur Deckung des einfachen bzw. fünffachen Endenergieverbrauchs benötigten Potentialkategorien identifiziert. Aus den so als „benutzt" identifizierten Kategorien wurde anschließend die gewichtete mittlere Vollaststundenzahl der Stromerzeugung sowie die gemittelten Stromgestehungskosten je Land berechnet. Aus dem Verhältnis zwischen genutzten und ungenutzten Potentialen wurde außerdem ein **Landnutzungskoeffizient** berechnet, welcher als Indikator dafür dient, wie viel zusätzliche Fläche noch genutzt werden könnte – also wie viel „überschüssiges" Potential pro Land zur Verfügung steht.

Abschließend wurden aus allen zu Verfügung stehenden Eingangsdatensätzen jene fünf Features für die Clusteranalyse ausgewählt, welche als die für die Fragestellung relevantesten identifiziert wurden. Diese wurden mit dem Ziel einer Verbesserung der Analyse aufbereitet und mittels Normierung auf die Clusterung vorbereitet.

2.2 Machine Learning Algorithmen - Clusteranalyse

Machine Learning (dt. Maschinelles Lernen) ist ein Teilgebiet der *Artificial Intelligence* (dt. künstliche Intelligenz) welches sich in der zweiten Hälfte des 20. Jahrhunderts entwickelt hat. Hierbei handelt es sich um selbstlernende Algorithmen welche aus bekannten Daten neue Erkenntnisse extrahieren und darauf basierend Vorhersagen treffen zu können.

Nach [22] werden die Algorithmen des *Machine Learning* drei verschiedenen Gattungen eingeteilt:

[2] Das Solarenergiepotential wird je Land nach PV und CSP Potential unterschieden und jeweils in neun Kapazitätskategorien welche wiederum in drei Verbraucherdistanzkategorein unterteilt sind angegeben (54 Indikatoren je Land) [15]. Die onshore Windenergiepotentiale wurden gleichen Parametrisierung wie die Solarpotentiale angegeben (27 Indikatoren je Land) [11]. Die offshore Windenergiepotentiale wurden zusätzlich nach Wassertiefe gestaffelt (81 Indikatoren je Land) [11].

- *Supervised learning* (dt. überwachtes Lernen): Algorithmen versuchen anhand gekennzeichneter Trainingsdaten (bekannte Ausgabewerte bzw. Zielvariable) Muster in den Eingangsdaten zu erlernen, um so für unbekannte oder zukünftige Daten Vorhersagen treffen zu können.
- *Unsupervised learning* (dt. unüberwachtes Lernen): Verfahren dieser Gattung generieren Informationen aus der Struktur der unbekannten Daten ohne bekannten Zielvariablen oder Belohnungsfunktion.
- *Reinforcement learning* (dt. verstärkendes Lernen): Hierbei ist es das Ziel ein Modell zu entwickeln, welches seine Leistung mittels Interaktionen mit der Systemumgebung anhand einer Belohnungsfunktion verbessert.

In dieser Arbeit wurden ausschließlich Methoden des *unspervised learnings* verwendet, da es bei Bildung von Metaregionen keine vordefinierte und als „richtig“ oder „falsch“ identifizierbare Zuordnung gibt.

Bei der Clusteranalyse handelt es sich um ein exploratives Datenanalyseverfahren welches zur Gattung des *unsupervised learning* zählt. Es kann eingesetzt werden um tief liegende Strukturen bzw. Gruppierungen aus einem unbekannten Datensatz zu extrahieren. Die identifizierten Cluster beschreiben Datenpunkte welche anhand von bestimmten gemeinsamen Eigenschaften gruppiert werden können und sich hinreichend zu den anderen Clustern unterscheiden. [22]

Für diese Arbeit wurde die Software Bibliothek „Scikit-learn“ verwendet [23]. Die Bibliothek stellt unter anderem verschiedene vordefinierte Clusteralgorithmen zur Verfügung, welche für die Datenanalyse verwendet werden können. Im Rahmen dieser Untersuchung wurden die Algorithmen *KMeans* und *GaussianMixture* verwendet. Der Clusteralgorithmus *KMeans* zählt zu den am meist verwendeten Ansatz der Clusteranalyse. Der Grund dafür liegt in der mathematischen Formulierung von *KMeans,* welche sehr einfach ist, weshalb auch die resultierenden Ergebnisse einfacher interpretierbar sind als die Ergebnisse komplexerer Verfahren. Der *KMeans* Algorithmus berechnet innerhalb eines multidimensionalen Datensatzes für eine vorgegebene Anzahl von Clustern die optimale Clusterzuordnung entsprechend der folgenden zwei Gesetze:

- Das arithmetische Mittel aller zum Cluster gehörenden Punkte bildet das Clusterzentrum.
- Jeder Punkt ist näher an seinem eigenen Clusterzentrum als an anderen Clusterzentren. [24]

Entsprechend dieses Ansatzes benutzt der *KMeans* Algorithmus ein deterministisches Verfahren, um jeden Datenpunkt zu exakt einem Cluster hinzuzufügen (*hard assignment*). Dies ist jedoch problematisch, wenn es Überlappungen einzelner Features gibt.

Der zweite verwendete Ansatz ist das *GaussianMixture* Verfahren. *Dieser* Algorithmus erweitert die Grundidee von *KMeans* um einen probabilistischen Ansatz (*soft assignment*) und umgeht so die Probleme welche überlappende Daten und die Benutzung von *hard assignment* Algorithmen hervorrufen können. Der *GaussianMixture* Ansatz benutzt dazu ein probabilistisches Modell, das davon ausgeht, dass alle Datenpunkte aus einer endlichen

Anzahl von Gauß-Verteilungen mit unbekannten Parametern erzeugt werden. Die optimalen Cluster werden hierbei wie folgt gebildet:

- Die Cluster werden gebildet in dem für jeden Datenpunkt die Wahrscheinlichkeit der Zugehörigkeit zu jedem Cluster berechnet wird und dem wahrscheinlichsten Cluster zugeordnet wird. [23]

Ein Nachteil dieser beiden Clusteralgorithmen ist, dass die Anzahl der Cluster im Voraus fest vorgegeben werden muss und keiner der beiden Algorithmen Ausreißer ignorieren kann [24]. Der dichtebasierte Clusteralgorithmus *DBSCAN* [23] kann mit diesen zwei Nachteilen umgehen. Da der DBSCAN Algorithmus jedoch Eingangsdaten gleichmäßiger Dichte benötigt, konnten im Rahmen dieser Arbeit keine zufriedenstellenden Ergebnisse mittels *DBSCAN* erzeugt werden.

Die Bewertung der Ergebnisse von unüberwacht lernenden Verfahren, wie der Clusteranalyse, unterscheidet sich von der trivialen Bewertung überwachten Lernens Verfahren (Klassifikationsalgorithmus). Da die Trainingsdaten bei der Clusterbildung keine bekannten Ausgabewerte haben, ist es nicht möglich einen eingängigen „Fehlerkoeffizienten“, also einen Quotienten aus Anzahl der Vorhersagefehler zu berechneten Datenpunkten, auszugeben. Die Metriken zur Bewertung der Clusteranalyse untersuchen vielmehr die Separationen und Zugehörigkeit der Datenpunkte der identifizierten Cluster. [23]

In dieser Arbeit wird daher der für unüberwachtes Verfahren gängige Silhouetten-Koeffizient (egl. Silhouette-Score) als Kennzahl für die Qualität der Clusteranalyse gewählt. Dieser Koeffizient ist unabhängig von der gewählten Clusteranzahl und deckt einen Wertebereich zwischen minus Eins und plus Eins ab, wobei plus Eins für gut strukturierte Cluster steht und negative Werte darauf hinweisen, dass Datenpunkte dem falschen bzw. einem unähnlicheren Cluster zugewiesen wurden. Werte nahe 0 deuten darauf hin, dass Cluster nicht scharf getrennt sind und sich möglicherweise überlappen. [23]

3 Ergebnisse

In diesem Kapitel werden die Ergebnisse der explorativen Datenanalyse, der Feature-Auswahl sowie die Ergebnisse der Clusteranalyse zur Bildung neuer Metaregionen vorgestellt und diskutiert.

3.1 Preprocessing:

3.1.1 Explorative Datenanalyse

Abbildung 4 zeigt vier Diagramme in welchen jeweils zwei gewählte Features einander gegenübergestellt sind. Diese Darstellung ermöglicht die Identifikation von Zusammenhängen zwischen zwei Features. Die farbliche Kodierung liefert hierbei die geografische Information der Zugehörigkeit eines Landes zu einem Kontinent. Die Größe der Kreise stellt ein Indikator für den Primärenergiebedarf eines Landes dar. Wobei ein größerer Kreis einen höheren Bedarf darstellt.

Das Diagramm oben links zeigt den prozentuellen Anteil von Solar- und Windenergie am gesamten Primärenergiebedarf eines Lands gegenüber den berechneten mittleren gewichteten Stromgestehungskosten zur einmaligen Deckung des Primärenergiebedarfs (vgl. Kapitel

2.1). Die Grafik zeigt, dass einerseits Wind- und Solarenergie heute in der Deckung des Primärenergiebedarfs eine untergeordnete Rolle spielen und dass andererseits die Nutzung der Wind- und Solarenergiepotentiale vor allem in Europa stak ausgeprägt ist, obwohl die gewichteten Kosten in Europa im Vergleich zu anderen Regionen relativ hoch sind.

Das Bild oben rechts zeigt die gewichteten Stromgestehungskosten gegenüber den mittleren Volllaststunden der Solar- und Windstromerzeugung. Hier ist zu erkennen, dass sich die regionalen Potentiale nicht nur in der Verfügbarkeit, sondern auch in der Volllaststundenzahl sowie der gemittelten Stromerzeugungskosten deutlich unterscheiden. Wenig überraschend lässt sich aus den Daten auch ein eindeutig negativer Zusammenhang zwischen den gemittelten Vollaststunden und den gewichteten Stromgestehungskosten identifizieren (ausgenommen der vier Ausreißer).

Das Bruttoinlandsprodukt pro Kopf gegenüber dem Primärenergiebedarf pro Kopf ist in der Abbildung unten links abgebildet. Die Darstellung zeigt einen klar positiven Zusammenhang zwischen einem höheren Bruttoinlandsprodukt und einem höheren Energiebedarf. Dieser Effekt ist unabhängig von der geografischen Zuordnung.

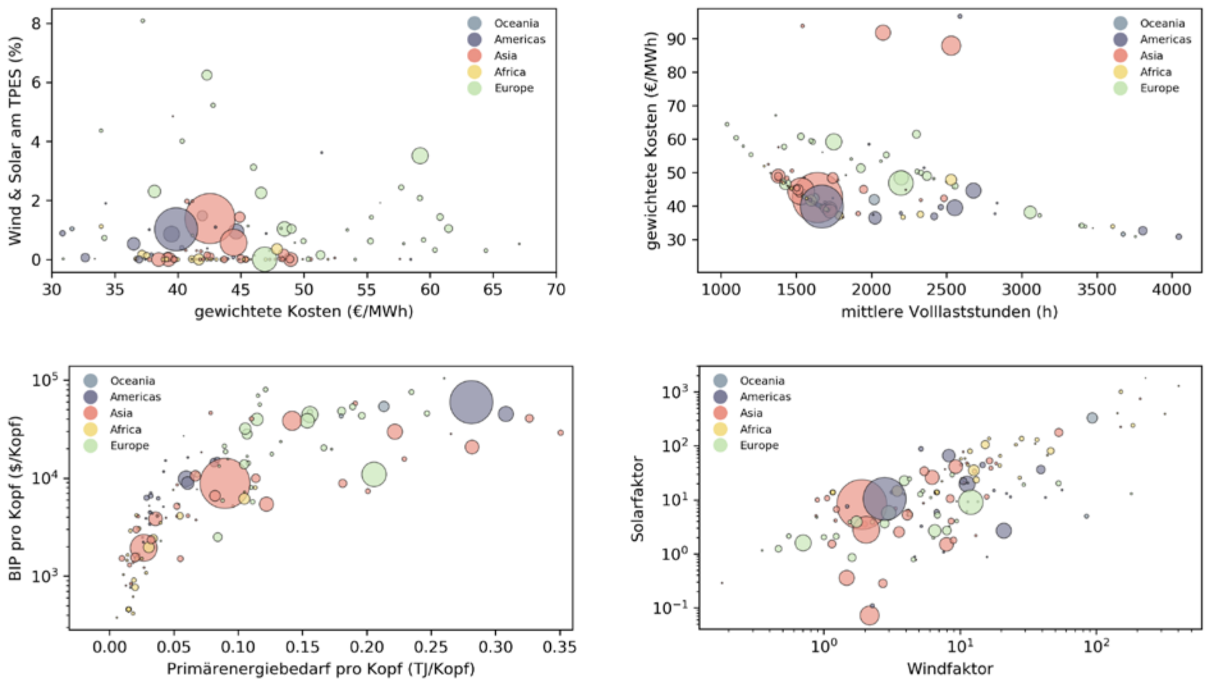

Abbildung 4: Darstellung von relevanten Features der explorative Datenanalyse. Je verfügbarem Land ist ein Datenpunkt abgebildet. Die Größe der Datenpunkte korreliert mit den Primärenergieverbräuchen der dargestellten Länder. Die Achsen der Grafik von Solar- und Windenergiefaktor sind logarithmisch skaliert. (Datenquellen vgl. Tabelle 3 – Definition der Solar-/Windfaktoren sowie der Berechnung der gemittelten gewichteten Stromgestehungskosten und Volllaststunden vgl. Kapitel 2.1)

Dass vierte Diagramm der Grafik (unten links) zeigt die logarithmische Gegenüberstellung des Solar- und Windenergiefaktors.[3] Die Darstellung zeigt, dass die Solar und Wind-

---

[3] Der Solarenergiefaktor beschreibt den Quotienten aus Solarenergiepotential zu Endenergiebedarf (2015) pro Land. Diese Definition gilt analog für den Windenergiefaktor. (Vgl. Kapitel 2.1)

potentiale global verteilt sind (wobei Länder mit geringem Energiebedarf höhere Potentiale ausweisen) und in beinahe allen Ländern ausreichend Wind- beziehungsweise Solarenergiepotentiale vorhanden sind um sogar die Deckung eines vielfachen des heutigen Energiebedarfs zu bewerkstelligen.

3.1.2 Feature Auswahl

Um das bestmögliche Ergebnis aus der Clusteranalyse zu ziehen, wurde die Feature-Anzahl sowie die Kombination der gewählten Features variiert. Außerdem wurden zwei verschiedene Algorithmen (*KMeans* und *GaussianMixture*) zur Clusterung der aufbereiteten Daten eingesetzt, um so die Eignung der beiden Algorithmen für den vorliegenden Anwendungsfall zu überprüfen. Abschließend wurden die Clusterergebnisse unter anderem in Form von Landkarten und Boxplots aufbereitet um die Interpretation der Clusteranalyse zu erleichtern.

Für die Clusteranalyse wurden folgende sechs beschreibenden Features ausgewählt:

- Bruttoinlandsprodukt pro Kopf (*GDP_per_cap*),
- Primärenergieverbrauch pro Kopf (*TPES_per_cap*),
- gewichtete Volllaststunden resultierend aus der Deckung der Endenergie mittels Solar und Windenergie (*wav_flh*),
- gewichtetet Kosten - resultierend aus der Deckung der Endenergie mittels Solar und Windenergie (*wav_lcoe*),
- Landnutzungskoeffizient – Flächenbedarf zur Bereitstellung des fünffachen Endenergieverbrauchs mittels Solar- und Windenergie (*used_area_TCF5*)
- Kontinent (*region*)

Aus dieser Auswahl wurden zwei Sets generiert, welche gegeneinander getestet wurden. Die ausgewählten Sets unterscheiden sich lediglich um das Feature Kontinent welches im *Set 1* Berücksichtigung fand, im *Set 2* jedoch nicht enthalten war. Dadurch sollte der geografische Einfluss auf die Clusterung reduziert werden, da die geographische Zugehörigkeit der Länder ohnehin in indirekter Form in den anderen Features enthalten ist.

Die Datenbasis der gewählten Features für das Bruttoinlandsprodukt, Bevölkerungszahlen sowie den Energieverbräuchen ist das Jahr 2015. Für das *Set 1* und *Set 2* konnte ein vollständiger Feature-Vektor für 121 Länder erzeugt werden. Länder für welche lediglich unvollständige Datensätze vorliegen wurden in der Clusterung nicht berücksichtigt und sind in den Ergebnisdarstellungen mit „*no-value*“ gekennzeichnet.

Die Abhängigkeiten und Verteilung der einzelnen Features ist in Abbildung 5 dargestellt. Die farbliche Hervorhebung entspricht hierbei den gewählten Regionen des Energiesystemmodells MESSAGE. Eine visuelle Clusterung der Daten bezüglich des Potentials der erneuerbaren Ressourcen basierend auf der Unterscheidung der Verteilungsfunktionen oder Abhängigkeiten der anderen Features ist für die MESSAGE Regionen in dieser Abbildung nicht möglich, da keine einheitlichen Gruppierungen erkennbar sind. Folglich wird im nachfolgenden Kapitel eine Clusteranalyse durchgeführt um eventuelle mathematische Zusammenhänge in den Daten zu identifizieren. Abbildung 5 hilft jedoch dabei, eindeutige Ausreißer im Feature-Vektor zu erkennen und auf eventuelle Fehler in den Eingangsdaten hinzuweisen. In Abbildung 5 ist zum Beispiel ein klarer Ausreißer bei den gemittelten Stromgestehungskosten (*wav_loce*) erkennbar. Dabei handelt es sich jedoch nicht um einen Fehler in den Daten, sondern lediglich der Darstellung Südkoreas teurer und im Verhältnis zum Energieverbrauch geringer Wind und Solarstrom Erzeugungspotentiale.

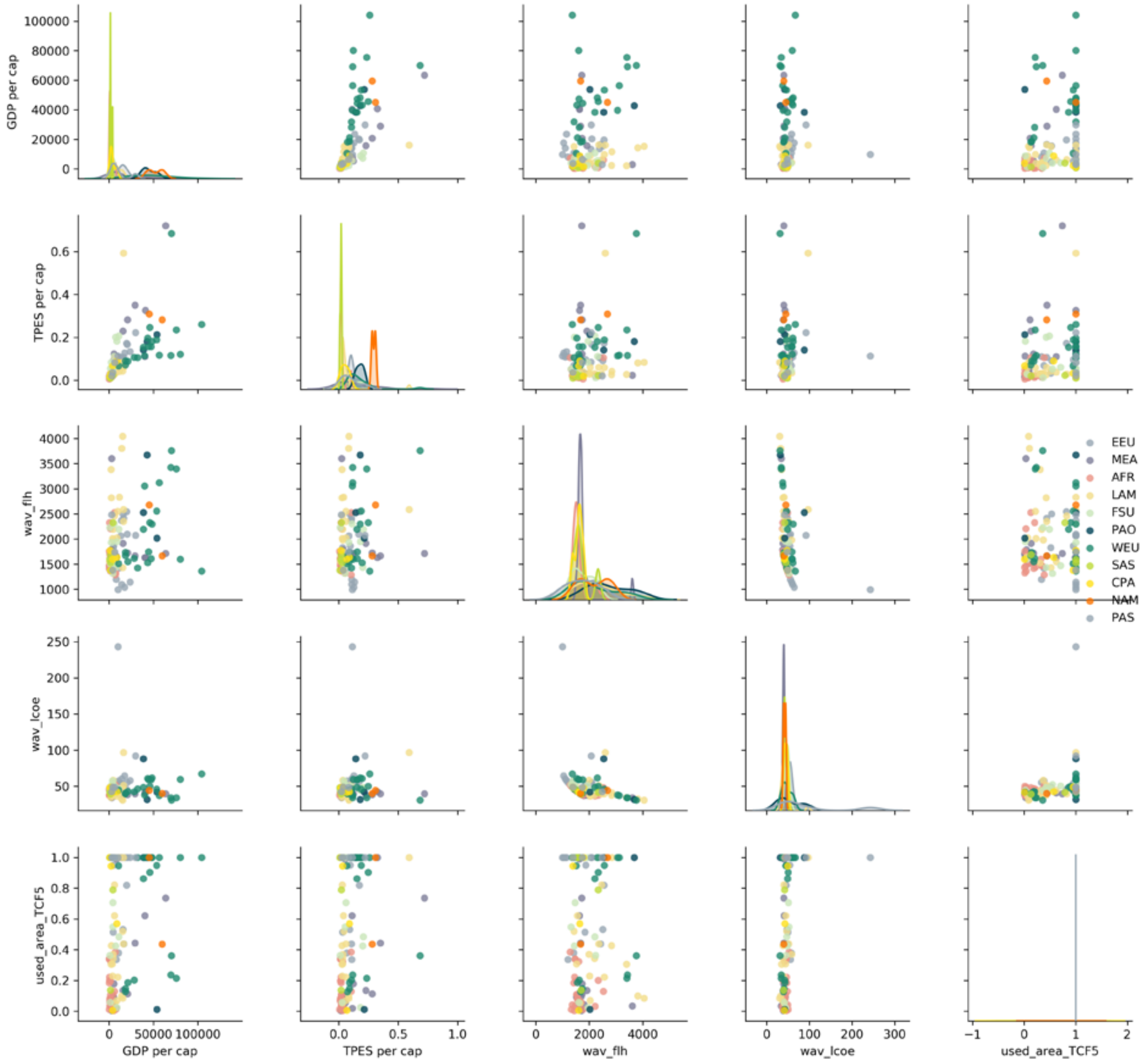

Abbildung 5: Gegenüberstellung aller für die Clusterung ausgewählten und verarbeiteten Features. Die Farbliche Hervorhebung entspricht den MESSAGE-Regionen aus Abbildung 2. Die Diagramme auf der Hauptdiagonalen zeigen die geschätzte Verteilungsfunkten der Features. Die anderen Diagramme zeigen die Abhängigkeit der Features zueinander.

3.2 Clusteranalyse

Ein Vergleich der Ergebnisse der Clusteranalyse in Form des Silhouetten-Koeffizienten der gewählten Algorithmen und Sets ist in Tabelle 1 dargestellt. Der Koeffizient liegt im Wertebereich zwischen 0,294 und 0,457 was auf einen schwach strukturierten Datensatz hinweist.

Grundsätzlich ist für beide Sets zu erkennen, dass der Algorithmus *KMeans* besser abgegrenzte Cluster findet als der *GaussianMixture* Ansatz. Dies kann darauf zurückgeführt werden, dass lediglich bei sehr großen Datenmengen der zentrale Grenzwertsatz greift und die Daten daher nur in sehr großer Menge vorliegend auch wirklich annähernd normalverteilt sind. Die Ergebnisse weisen darauf hin, dass die verwendete Datenbasis (121 Länder – fünf

Features) für eine angemessene Clusterung mit dem GaussianMixture Verfahren zu klein zu sein scheint.

Tabelle 1: Vergleich des Silhouetten-Koeffizient der ausgewählten Clusteralgorithmen und Feature-Kombinationen für die gewählte Clusteranzahl von 10.

Algorithmus	***Set 1***	***Set 2***
KMeans	0,457	0,382
GaussianMixture	0,402	0,294

Weiteres zeigt sich in beiden Verfahren, dass das Weglassen der geografischen Information bezüglich der Zugehörigkeit zu einem Kontinent (Unterschied zwischen *Set 1* und *Set 2*) eine Verschlechterung des Silhouetten-Koeffizienten herbeiführt. Trotzdem werden in der weiteren Interpretation der Ergebnisse die Clusterergebnisse nur für das *Set 2* durchgeführt da in dieser Arbeit die Potentiale der erneuerbaren Energien unabhängig ihrer geografischen Lage untersucht werden sollen.

3.2.1 KMeans

In Abbildung 6 ist das Ergebnis der Clusterung des *Set 2* mittels *KMeans* und einem Silhouetten-Koeffizienten von 0,382 geografisch dargestellt. Die Clusteranalyse zeigt eine Gruppierung der Länder Nordamerikas mit Ländern in Mittel- und Nordeuropa zu Cluster 1.

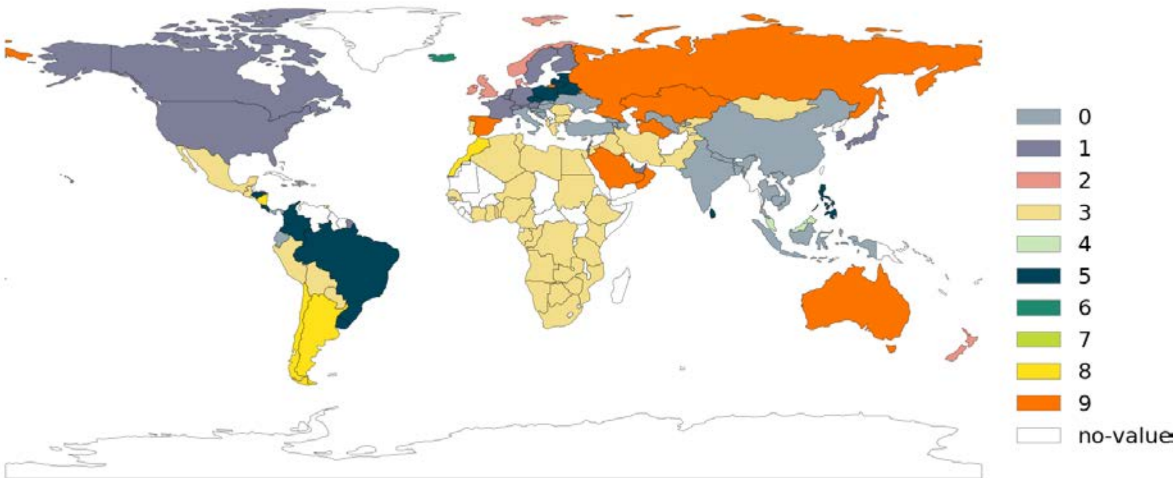

Abbildung 6: Geografische Darstellung der Ergebnisse der Clusteranalyse der Set 2-Features mittels KMeans Algorithmus.

Der Vergleich mit Abbildung 7 zeigt, dass diese Länder jeweils im oberen Drittel des Bruttoinlandprodukts und Primärenergieverbrauchs liegen (beide pro Kopf). Außerdem benötigen 75 % der Länder dieses Clusters, zur Deckung des fünffachen Endenergiebedarfs mittels Solar- und Windverstromung, mehr als 90 % der zur erneuerbaren Energieerzeugung zu Verfügung stehenden Landfläche.

Ein weiterer großer Zusammenschluss von Ländern ist Cluster Nummer 3 welcher im Wesentlichen aus den Ländern Afrikas besteht. Dieser Cluster unterscheidet sich zu Cluster 1 im Bruttoinlandprodukt pro Kopf und Primärenergieverbrauch pro Kopf. Im Gegensatz zu Cluster 1 liegen die Länder dieses Clusters im unteren Drittel beider Indikatoren. Aufgrund des aktuell geringeren Energiebedarfs sowie der präferierten natürlichen Gegebenheiten benötigen 75 % der Länder des Clusters 1 lediglich ca. 20 % der Landfläche zur Deckung des fünffachen Endenergiebedarfs. Die Cluster 4, 6 und 7 gruppieren jeweils lediglich wenige Länder zu einem Cluster. Wobei hier Cluster in mindestens einem der gewählten Features einen Ausreißer beinhalten. Zum Beispiel wird in dieser Clusterung Island keiner Region zugeteilt sondern stellt als Ausreißer (ein vielfach höherer Primärenergiebedarf pro Kopf) einen eigenen Cluster dar (Cluster 6).

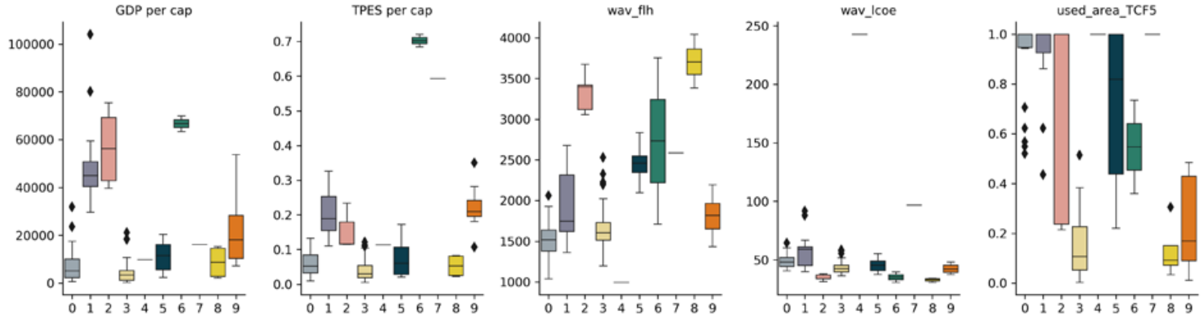

Abbildung 7: Darstellung der Ergebnisse der Clusteranalyse der Set 2-Features mittels KMeans Algorithmus. Die Definition der hier dargestellten Features findet sich in Kapitel 3.2.

Eine Gegenüberstellung der beschreibenden Features zur visuellen Einschätzung des Clusterergebnisses findet sich in im Anhang (Abbildung 10).

3.2.2 GaussianMixture

Das Ergebnis der *GaussianMixture* Clusteranalyse unterscheidet im Vergleich zu *KMeans* die Länderzuordnung innerhalb von Nordamerika und Afrika (Abbildung 8). Hingegen werden Länder aus Mittel- und Südamerika, Afrika, Russland und ein Großteil Asiens zu einem sehr großen Cluster 6 zusammengefasst. Ein möglicher Grund dafür ist eine nur schwach vorhandene Struktur innerhalb der Eingangsdaten, worauf auch der geringere Silhouetten-Koeffizienten von lediglich 0,294 hinweist.

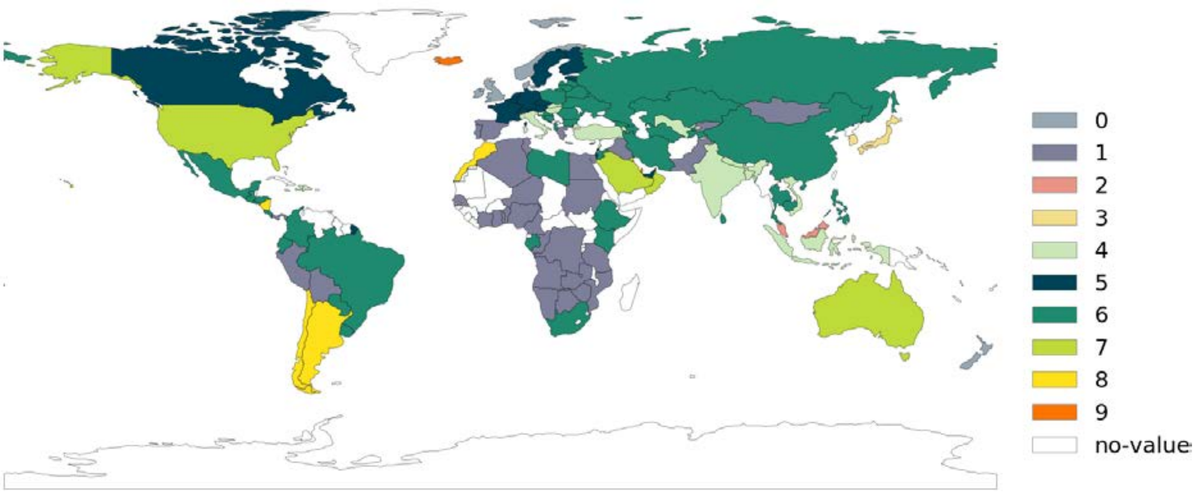

Abbildung 8: Geografische Darstellung der Ergebnisse der Clusteranalyse der Set 2-Features mittels GaussianMixture Algorithmus.

Für den erwähnten, in Größe dominanten und dadurch unspezifischen Cluster 6, reicht die Streuung der für die Bereitstellung des fünffachen Endenergiebedarfs benötigten Fläche von 0,3 % bis 100 % (Abbildung 9). Diese erhöhte Streuung zeigt sich, wenn auch weniger stark, bei allen Features, vor allem im Vergleich zu den Ergebnissen der *KMeans* Clusterung. Weiteres werden auch hier Cluster mit Ausreißer (2 und 6) identifiziert.

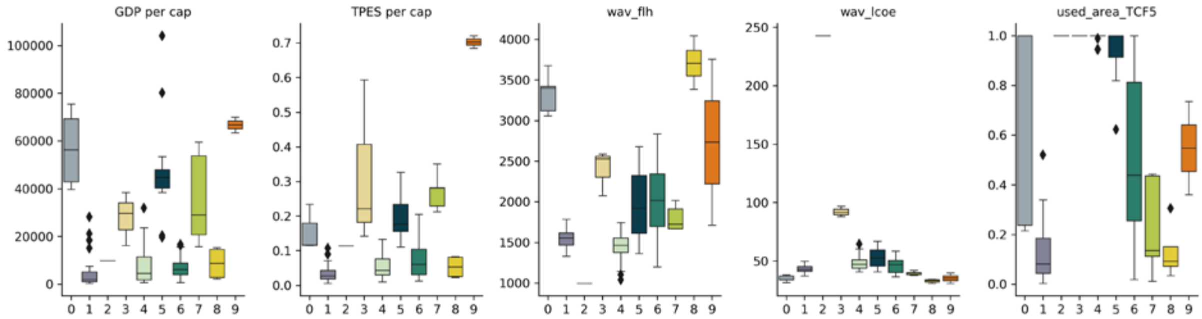

Abbildung 9: Darstellung der Ergebnisse der Clusteranalyse der Set 2-Features mittels Gaussian-Mixture Algorithmus. Die Definition der hier dargestellten Features findet sich in Kapitel 3.2.

Auch für diese Analyse findet sich im Anhang eine Gegenüberstellung der beschreibenden Features zur visuellen Einschätzung des Clusterergebnisses (Abbildung 11).

4 Fazit, Diskussion und Ausblick

Um die im Pariser Klimaabkommen vereinbarten Ziele noch zu erreichen, müssen effektive transnationale Strategien zur Herbeiführung einer radikalen Emissionswende entwickelt und rasch umgesetzt werden. Nationale Energiesystemmodelle sind nicht in der Lage die für diese Fragestellung so relevanten internationalen Rückkopplungseffekte und Marktdynamiken abzubilden. Globale Energiesystemmodellen, welche diese Dynamiken abbilden können, müssen dafür, um handhabbar und in annehmbarer Zeit mathematisch lösbar zu bleiben, auf geografische Detailschärfe verzichten. Anstatt einzelner Länder bilden Modelle dieser Art, Regionen ab. Die dadurch bedingte Unschärfe kann durch eine geschickte und auf die konkrete Fragestellung angepasste Zusammenfassung von Ländern zu Metaregionen ausgeglichen werden. Traditionell wurden die Länder dazu anhand der Expertise der Modellierer basierend auf ihrer geographischen sowie politisch- historischen Nähe zueinander zu Regionen aggregiert. In dieser Arbeit wurde dieser Ansatz reevaluiert. Dazu wurden mittels mathemaischer Clusterbildung unter Anwendung der *KMeans* und *Gaussian-Mixture* Algorithmen und einem Fokus auf die Wind- und Solarenergiepotenziale der Länder, Regionen gebildet und die Ergebnisse untersucht. Alle Ergebnisse weisen jedoch Silhouetten-Koeffizienten kleiner 0,5 was auf eine schwache strukturelle Abgrenzung der Cluster und damit auf die schwachen Strukturen der Eingangsdaten hinweist.

Auch die traditionelle Bildung von Metaregionen basiert im Grunde auf Daten. Jedoch können hier qualitative Informationen in den Prozess der Erzeugung von Metaregionen mit einfließen oder es können bestimmte Daten für bestimmte Länder bewusst ignoriert werden, wenn diese aus einem erklärbaren Grund von dem erwartet Wert abweichen. Dies ist bei einem ausschließlich datenbasierten Verfahren wie der Clusteranalyse nicht möglich - *Same Same but Different*. Die Clusteranalyse wird daher den herkömmlichen Ansatz nicht ersetzten können. Jedoch können Methoden des *Machine Learning* als unterstützendes Tool für die traditionelle Bildung von Metaregionen eingesetzt werden. Denn es erlaubt eine schnelle prototypische Umsetzung der Clusterbildung von Länder basierend auf einer flexiblen Auswahl von Eingangs-Features. Das Ergebnis der Clusteranalyse kann ein erster Startpunkt sein oder helfen um mögliche Ausreißer zu identifizieren.

5 Authors Contribution

TZ war hauptverantwortlich für die Umsetzung der Clusteranalyse in Python. CO war federführend in der Findung und Definition der Themenstellung sowie der Interpretation der Ergebnisse.

6 Anhang

6.1 Tabellen

Tabelle 2: Übersicht der Regionen der meistgenutzten globalen Energiesystemmodelle

Modell	Version	Anzahl an Regionen	Quelle	Regionen
TIMES -TIAM	UCL	16	[1]	*Africa*, Australia, Canada, Central and South America, China, Eastern Europe, *Former Soviet Union*, India, Japan, Mexico, Middle-East, Other Developing Asia, So-Korea, United Kingdom, USA, Western Europe
	NZ	16	[2]	*Africa*, Australia, New-Zealand, Canada, Central and South America, China, Eastern Europe, *Former Soviet Union*, India, Japan, Mexico, Middle-East, Other Developing Asia, So-Korea, USA, Western Europe
	ETSAP	15	[3]	*Africa*, Australia, Canada, Central and South America, China, Eastern Europe, *Former Soviet Union*, India, Japan, Mexico, Middle-East, Other Developing Asia, So-Korea, USA, Western Europe
OSeMOSYS	GENeSYS-MOD	10	[4]	*Africa*, China, Europe, *Former Soviet Union*, India, Middle East, North America, Oceania, Rest of Asia and South America
POLES	JRC	66	[5]	Argentina, Austria, Australia, Belgium, Brazil, Bulgaria, Canada, Croatia, Chile, Cyprus, China, Czech, Republic, Egypt, Denmark, Iceland, Estonia, India, Finland, Indonesia, France, Iran, Germany, Japan, Greece, Malaysia, Hungary, Mexico, Ireland, New, Zealand, Italy, Norway, Latvia, Russia, Lithuania, Saudi, Arabia, Luxembourg, South, Africa, Malta, South, Korea, Netherlands, Switzerland, Poland, Thailand, Portugal, Turkey, Romania, Ukraine, Slovakia, United, States, Slovenia, Vietnam, Spain, Sweden, United, Kingdom Rest Central America, Rest South America, Rest Balkans, *Rest CIS*, Mediter. Middle East, Rest of Persian Gulf, Morocco and Tunisia, Algeria and Libya, *Rest Sub-Saharan Africa*, Rest South Asia, Rest South East Asia, Rest Pacific
MESSAGE		11	[6]	*Sub-Saharan Africa*, Centrally planned Asia and China, Central and Eastern Europe, *Former Soviet Union*, Latin America and the Caribbean, Middle East and North Africa, North America, Pacific OECD, Other Pacific Asia, South Asia, Western Europe

Tabelle 3: Quellenzuordnung der Eingangsdaten für die Clusteranalyse

Parameter		**Quellen**
Energieressourcen	Solar - PV	[15]
	Solar - CSP	[15]
	Wind – onshore	[11]
	Wind – offshore	[11,12]
	Biomasse	[20]
	Wasserkraft – Laufwasser	[21]
	Wasserkraft – Pumpspeicher	[21]
	Kohle	[8]
	Erdöl	[8]
	Erdgas	[8]
	Unkonventionelles Gas	[17, 18]
	Unkonventionelles Erdöl	[18]
Lastgänge	Solar	[19]
	Wind	[19]
	Temperatur	[19]
Energiebilanzen	Primärenergieverbrauch (fossil)	[8]
	Import/Export Saldo	[6,14]
	Transformationseffizienz (fossil)	[8]
	Endenergieverbrauch (fossil)	[8]
	Primärenergieverbrauch (erneuerbar)	[7,8]
	Transformationseffizienz (erneuerbar)	[7]
	Endenergieverbrauch (erneuerbar)	[7,8]
Installierte Kapazitäten	Kraftwerke (konventionell)	[9]
	Kraftwerke (erneuerbar)	[7,8,9,14]
Emissionen	Treibhausgas-Emissionsbilanzen	[6,8,13]
	UNFCCC-NDCs	[10]
Makroökonomische Parameter	Bruttoinlandsprodukt	[6]
	Bevölkerungsentwicklung	[6]
	HDI / GINI Index	[6]
	Bevölkerungsdichte	[6]
	Gesellschaftliche Strukturentwicklung	[6]

6.2 Abbildungen

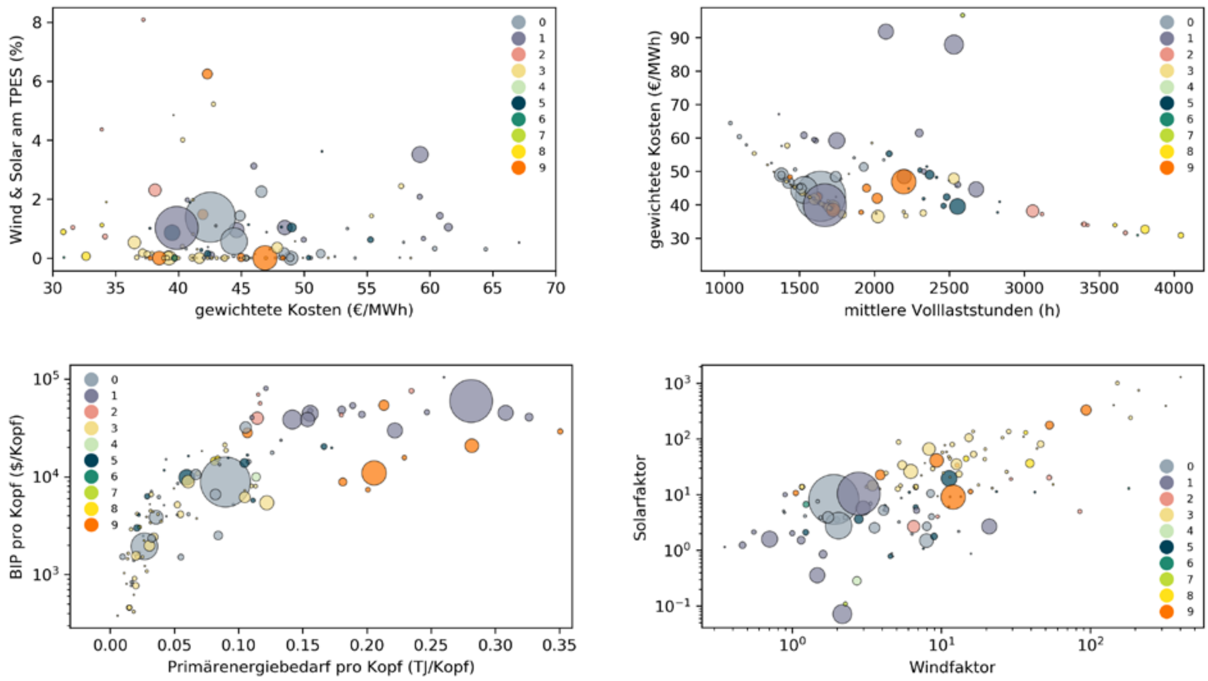

Abbildung 10: Gegenüberstellung von ausgewählten Features. Farbliche Hervorhebung der 10 Cluster nach KMeans für das Set 2.

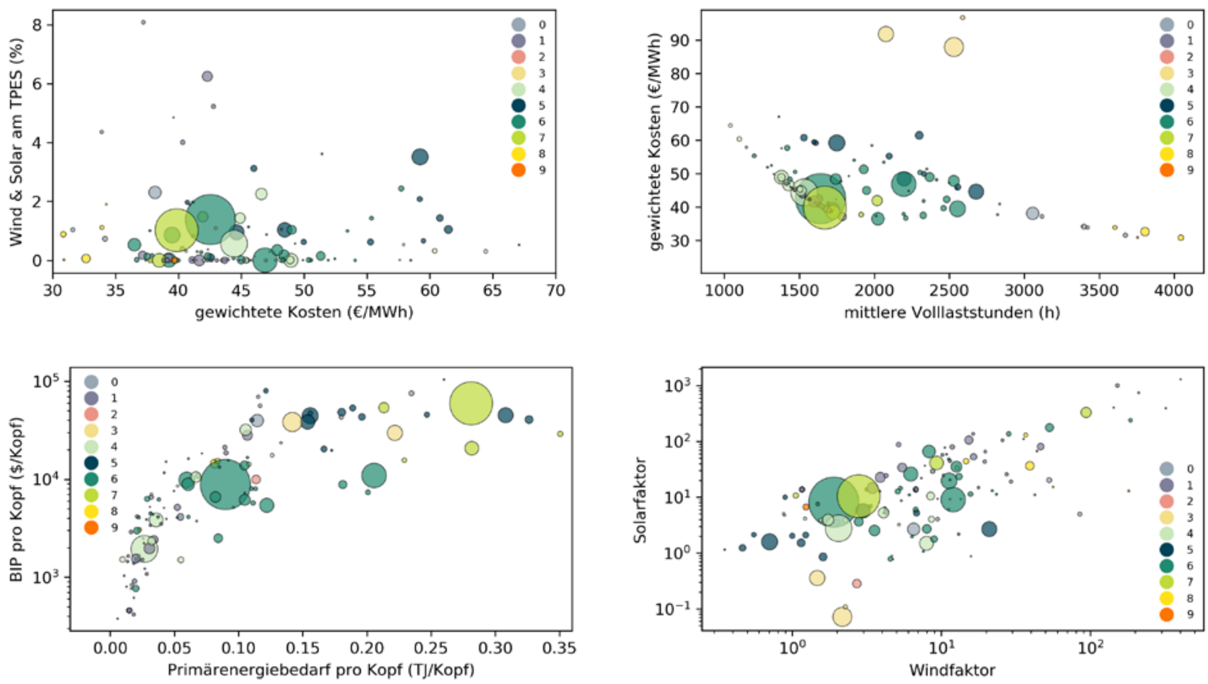

Abbildung 11: Gegenüberstellung von ausgewählten Features. Farbliche Hervorhebung der 10 Cluster nach GaussianMixture für das Set 2.

7 Literatur

[1] Anandarajah, G., Pye, S., Usher, W., Kesicki, F. & McGlade, C. E. TIAM-UCL Global Model Documentation. http://www.ucl.ac.uk/energy-models/models/tiam-ucl/tiam-ucl-manual (University College London, 2011)

[2] Loulou, Richard & Labriet, Maryse ETSAP-TIAM: the TIMES integrated assessment model Part I: Model structure. https://link.springer.com/content/pdf/10.1007%2Fs10287-007-0046-z.pdf (Springer-Verlag, 2007)

[3] Vaillancourt, Kathleen, Labriet, Maryse, Loulou, Richard, Waaub, Jean-Philippe The Role of Nuclear Energy in Long-Term Climate Scenarios: An Analysis with the World-TIMES Model. https://iea-etsap.org/TIAM_f/4_Nucleaire_EnergyPolicy_ORMMES06.pdf (Les Cahiers du GERAD, 2007)

[4] Löffler, Konstantin, Hainsch, Karlo, Burandt, Thorsten, Oei, Pao-Yu, Kemfert, Claudia & von Hirschhausen, Christian Designing a Model for the Global Energy System—GENeSYS-MOD: An Application of the Open-Source Energy Modeling System (OSeMOSYS) (Energies (10), 2017)

[5] Keramidas, K., Kitous, A., Després, J., Schmitz, A., POLES-JRC model documentation. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC107387/kjna28728enn.pdf (JRC Technical Reports, 2017)

[6] The World Bank (WB), World Indicators. https://data.worldbank.org/indicator (2018)

[7] International Renewable Energy Agency (IRENA), Renewable Energy Dashboard. http://resourceirena.irena.org/gateway/dashboard/ (2018)

[8] British Petroleum (BP), Statistical Review of World Energy 2017. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html (2017)

[9] Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, Enipedia, World Resources Institute. Global Power Plant Database. Published on Resource Watch and Google Earth Engine. http://resourcewatch.org/ & https://earthengine.google.com/ (2018)

[10] World Resources Institute (WRI) and CAIT Climate Data Explorer, CAIT Paris Contributions Map. http://cait.wri.org/indcs/ (2018)

[11] National Renewable Energy Laboratory, Global CFDDA-based Onshore and Offshore Wind Potential Supply Curves by Country, Class, and Depth (quantities in GW and PWh). https://openei.org/datasets/dataset/global-cfdda-based-onshore-and-offshore-wind-potential-supply-curves-by-country-class-and-depth-q (2018)

[12] National Renewable Energy Laboratory, Offshore Wind Resource. https://openei.org/datasets/dataset/offshore-wind-resource (2018)

[13] World Resources Institute (WRI), CAIT Climate Data Explorer, Climate Analysis Indicators Tool: WRI's Climate Data Explorer. http://datasets.wri.org/dataset/cait-unfccc-annex-i-ghg-emissions-data (2013)

[14] Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), BGR Energiestudie 2017. https://www.bgr.bund.de/DE/Themen/Energie/Produkte/energiestudie2017_Zusammenfassung.html?nn=1542226 (2017)

[15] Pietzcker, R. C., Stetter, D., Manger, S., Luderer, G., Using the sun to decarbonize the power sector: The economic potential of photovoltaics and concentrating solar power. (Applied Energy 135, 704–720. 2014)

[16] World Energy Council, World Energy Resources 2016 https://www.worldenergy.org/publications/2016/world-energy-resources-2016/ (2016)

[17] U.S. Energy Information Agency (eia), World Shale Resource Assessments. https://www.eia.gov/analysis/studies/worldshalegas/ (2015)

[18] Hongjun WANG, Feng MA, Xiaoguang TONG, Zuodong LIU, Xinshun ZHANG, Zhenzhen WU, Denghua LI, Bo WANG, Yinfu XIE, Liuyan YANG, Assessment of global unconventional oil and gas resources. (Petroleum Exploration and Development, 43, 6, 925-940, 2016)

[19] National Aeronautics and Space Administration (NASA), Global Modeling and Assimilation Office: Modern-Era Retrospective analysis for Research and Applications (MERRA). https://gmao.gsfc.nasa.gov/reanalysis/MERRA/ (2018)

[20] Moreira, José Roberto, Global Biomass Energy Potential. (Mitigation and Adaptation Strategies for Global Change 11, 2, 13–342, 2006)

[21] World Energy Council (WEC), World Energy Resources Hydropower 2016, https://www.worldenergy.org/wp-content/uploads/2017/03/WEResources_Hydropower_2016.pdf (2016)

[22] S. Raschka, Machine Learning mit Python: Das Praxis-Handbuch für Data Science, Predictive Analytics und Deep Learning, 1. Auflage, (2017)

[23] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, (2011)

[24] Jake VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media, https://jakevdp.github.io/PythonDataScienceHandbook, (2016)

[25] Climate Action Tracker, https://climateactiontracker.org/global/temperatures (2019)

[26] Google Developers, Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/ (2019)

[27] United Nations Framework Convention on Climate Change, The Paris Agreement 2015, https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement, (2019)

[28] N. Stern, Stern review: the economics of climate change, United Kingdom (2006)