Fakultät für Mathematik, Informatik und Statistik
Institut für Statistik
Ludwig-Maximilians-Universität München

BACHELORARBEIT

Ein Vergleich verschiedener multivariater Methoden anhand von Daten zu den Gesundheitsverhältnissen in den OECD-Ländern

Verfasserin: Lena Zelinka

Betreuerin: Dr. Sabine Hoffmann

Ort, Datum: München, den 13.04.2022

Eidesstattliche Erklärung

Hiermit bestätige ich, dass ich die vorliegende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum

Unterschrift

Abstract

Eine hohe Anzahl an Dimensionen ist ein häufiges Problem bei statistischen Analysen. Um dennoch gute Auswertungen zu bekommen und mögliche Assoziationsstrukturen aufzufinden, sind multivariate Analysetechniken sehr hilfreich. Das Ziel der vorliegenden Bachelorarbeit ist ein Vergleich einiger dieser Methoden.

Im Rahmen dieser Arbeit wurden dafür fünf verschiedene Verfahren zur Dimensionsreduktion, auch in Kombination mit Clusterverfahren, betrachtet. Hierfür dienten Daten zu den Gesundheitsverhältnissen in den OECD-Ländern. Diese umfassen neben der medizinischen Gesundheit auch weitere Determinanten, wie die sozialen Umstände, das individuelle Verhalten und die Ausstattung im Gesundheitssektor, welche die Gesundheit der Bevölkerung beeinflussen können.

Die Ergebnisse zeigen, dass es durchaus sinnvoll ist, mehrere multivariate Methoden zu betrachten, um ein Bild über die Struktur der Daten zu bekommen. Die resultierenden Cluster hängen dabei stärker von den Resultaten der Dimensionsreduktionsmethode ab und weniger von dem angewandten Clusterverfahren. Eine objektive Bewertung der Ergebnisse gestaltet sich jedoch meist schwierig, da die „wahre“ Struktur der Daten nicht bekannt ist.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ASW	Average Silhouette Width
CH-Index	Calinski-Harabasz-Index
MAR	Missing At Random
MCAR	Missing Completely At Random
MDS	Multidimensional Scaling
MFA	Multiple Factor Analysis
MICE	Multivariate Imputation by Chained Equations
NMAR	Not Missing At Random
OECD	Organisation for Economic Co-operation and Development
OEEC	Organisation for European Economic Co-operation
PCA	Principal Component Analysis
SNE	Stochastic Neighbor Embedding
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Unter den „Fluch der Dimensionalität“ fallen viele Herausforderungen, die sich im Rahmen von Analysen hochdimensionaler Daten ergeben. Wächst die Anzahl an Variablen, so wächst auch das „Volumen“, das die Beobachtungen einnehmen können, an (vgl. Altman et al., 2018). Um zu vermeiden, dass eine Umgebung des Raums möglicherweise keine Daten enthält, wäre ein sehr großer Stichprobenumfang nötig. Allerdings ist die Erhebung solcher Daten in der Praxis häufig nicht umsetzbar. Eine weitere Problematik hochdimensionaler Daten ist die Komplexität mathematischer und statistischer Eigenschaften. Beispielsweise ist die Entfernung und die Dichte im hochdimensionalen Raum nicht gleichzusetzen mit der im zwei- oder dreidimensionalen Raum (vgl. Schubert et al., 2017, S. 188). Bei der Untersuchung von Mustern und Clusterbildungen in Daten ist die Visualisierung der Ergebnisse oft hilfreich für die Interpretation. Allerdings führt auch dies zu Schwierigkeiten im hochdimensionalen Raum. Eine wichtige Aufgabe in der Datenwissenschaft ist deshalb, einen Weg zu finden, den Fluch der Dimensionalität zu umgehen. Dafür eignen sich Verfahren zur Dimensionsreduktion, mit denen die Daten in einen nieder-dimensionalen Raum konvertiert werden, ohne dass die wesentlichen Informationen dabei verloren gehen. Zudem ermöglichen multivariaten Verfahren die Visualisierung der Datenpunkte, was die Betrachtung der Struktur der Daten erleichtert. Für die Durchführung einiger Methoden sind allerdings subjektive Entscheidungen nötig, was die Interpretation der Daten beeinflusst. Das Ziel der Arbeit ist ein Vergleich verschiedener multivariater Verfahren zur Dimensionsreduktion, sowie in Kombination mit Clusterverfahren.

Die Arbeit gliedert sich wie folgt. Einen kurzen Hintergrund zur Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (engl.: Organisation for Economic Co-operation and Development, OECD) und den Daten bietet das Kapitel 2. Im Rahmen dieser Bachelorarbeit werden fünf multivariate Verfahren verglichen: die Hauptkomponentenanalyse (engl.: Principal Component Analysis, PCA), die Multiple Faktorenanalyse (MFA), die Multidimensionale Skalierung (MDS), Uniform Manifold Approximation and Projection (UMAP) und t-Distributed Stochastic Neighbor Embedding (t-SNE). Dabei ist zu erwähnen, dass diese im Weiteren der

Arbeit häufig nur als *die Methoden* bezeichnet werden. Zudem werden die Methoden auch in Kombination mit zwei verschiedenen Clusterverfahren betrachtet. Im Rahmen der Datenaufbereitung erfolgte eine Schätzung der fehlenden Werte mithilfe eines multiplen Imputationsverfahrens. Alle verwendeten Methoden werden in Kapitel 3 zunächst erläutert und deren Ergebnisse anschließend in Kapitel 4 aufgeführt. Eine Zusammenfassung der wesentlichen Resultate dieser Analysen, sowie Vor- und Nachteile der Methoden und ein Ausblick auf mögliche Weiterführungen, erfolgt in Kapitel 5. Angehängt sind in Anhang A weitere Ergebnisse. Die Umsetzung wurde mithilfe des statistischen Programmes R durchgeführt. Der Anhang B bietet einen Überblick der Struktur des elektronischen Anhangs.

Kapitel 2

Datenbeschreibung

2.1 Organisation für wirtschaftliche Zusammenarbeit und Entwicklung

Das Ziel der OECD ist es, den Wohlstand und die Gleichheit in den Ländern zu fördern. Die internationale Organisation umfasst 38 Mitgliedsstaaten, von Nord- und Südamerika bis nach Europa und in den asiatisch-pazifischen Raum (vgl. *OECD.org* o. D.). Sie bietet umfangreiche Daten und Analysen über verschiedene soziale, wirtschaftliche und ökologische Standards und steht dabei im Austausch mit Parlamenten, Regierungen und Zivilgesellschaften. Die ursprünglich gegründete Organisation für Europäische wirtschaftliche Zusammenarbeit (engl.: Organisation for European Economic Co-operation, OEEC) verwaltete die amerikanische und kanadische Hilfe beim wirtschaftlichen Wiederaufbau Europas nach dem Zweiten Weltkrieg im Rahmen des Marshallplans. Im September 1961 wurde die OEEC in die OECD umgewandelt, um weiterhin Daten für wirtschaftspolitische Fragen zu bieten. Diese umfassen Themen, wie beispielsweise PISA im Bildungsbereich bis hin zu Steuertransparenz und künstlicher Intelligenz (vgl. ebd.). Über die Jahre verfolgte die OECD stets das Ziel: „to become more global, more inclusive and more relevant“ (ebd.).

2.2 Datenaufbereitung

Der gesamte Datensatz setzt sich aus fünf einzelnen Datensätzen von der OECD zusammen (s.h. *OECD.Data.Health* o. D.). Diese enthalten international vergleichbare Informationen über den Gesundheitszustand verschiedener Länder:
Der Datensatz *Inanspruchnahme des Gesundheitswesens* gibt Auskünfte über beispielsweise Impfungen, diagnostische Untersuchungen, durchschnittliche Krankenhausaufenthaltsdauer, Entlassungsquoten und Transplantationen. Des Weiteren sind Angaben zu den *Ressourcen für das Gesundheitswesen*, wie die Gesamtbeschäftigung im Gesundheits- und Sozialwesen,

sowie die Ausstattung an Krankenhäusern und Krankenhausbetten vorhanden. Laufende Konsumausgaben für Gesundheitsgüter und -dienstleistungen sind im Datensatz *Indikatoren für Gesundheitsausgaben* aufgeführt. Im Rahmen dieser Bachelorarbeit wurde dabei als Maßeinheit nur der Anteil am Bruttoinlandsprodukt betrachtet. Auch *nichtmedizinische Determinanten der Gesundheit* wie der Lebensmittelverbrauch, Alkoholkonsum, Tabakkonsum und das Körpergewicht werden zwischen den Ländern verglichen. Verschiedene Inzidenzen, Todesursachen, sowie Mortalität, Lebenserwartung, wahrgenommener Gesundheitszustand und weitere Variablen wurden zu einem Datensatz *Gesundheitszustand* zusammengetragen.
Für eine stabilere Vergleichbarkeit wurden die Werte über drei Jahre gemittelt. Dabei wurde sich für den Zeitraum 2017-2019 entschieden, um mögliche Veränderungen durch die Corona-Pandemie auszuschließen. In dem nun erstellten Datensatz befand sich eine hohe Anzahl an Variablen bzw. Beobachtungen mit fehlenden Angaben. Diese wurden später mithilfe eines multiplen Imputationsverfahrens geschätzt. Mehr Informationen dazu bietet das Kapitel 3.2. Vorher wurden allerdings Spalten bzw. Zeilen mit einem Anteil fehlender Werte ab elf Prozent entfernt. Anschließend wurden die Variablen noch weiter manuell selektiert, um z.B. doppelte Informationen (absolute und relative Anzahl) zu filtern. Daraufhin ergab sich ein großer Datensatz mit 86 Variablen und 34 Ländern. Für einige Variablen wie beispielsweise den Tabakkonsum gab es Angaben für die gesamte Population und getrennt nach Geschlecht. Der Anteil der praktizierenden Ärzt*innen pro Land wurde gesamt, in weiblichem und männlichen Anteil und für verschiedene Altersgruppen angegeben. Die Angabe über den wahrgenommene Gesundheitszustand wurde ebenfalls zusätzlich zwischen den Geschlechtern unterschieden. Zudem wurde die Ausprägung *guter/ sehr guter Gesundheitszustand* neben Geschlecht auch nach Altersgruppen, höchste und niedrigste Einkommensklasse und/oder nach niedrigem, mittleren und hohem Bildungsstand gegliedert. Auf diese Einteilungen wurde verzichtet und ein zusätzlicher kleiner Datensatz mit 52 Variablen und 34 Ländern erstellt, indem nur Variablen für jeweils die gesamte Population ohne Abstufung nach Geschlecht, Alter, Einkommen oder Bildung betrachtet wurden. Somit ergaben sich zwei Datensätze, die zum Vergleich der statistischen Methoden dienten.

Kapitel 3

Methoden

Im folgenden Kapitel werden die verwendeten statistischen Methoden in der Bachelorarbeit näher erläutert. Es wurden herkömmliche Methoden, wie beispielsweise die PCA, betrachtet. Diese wurden neueren und (noch) weniger verbreiteten Methoden, wie zum Beispiel t-SNE und UMAP, gegenübergestellt. Zudem wird ein Verfahren der multiplen Imputation vorgestellt, welches für die Datenaufbereitung genutzt wurde.

3.1 Notation

Für die Beschreibung der Methoden wird eine allgemeine Notation hier festgelegt.
Ein Datensatz mit insgesamt n Beobachtungen und p Variablen entspricht der Matrix

$$\boldsymbol{X} = \begin{pmatrix} x_{11}, & \dots, & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1}, & \dots, & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

Der Wert x_{ij} entspricht dem Wert für die i-te Beobachtung und der j-ten Variable. $\boldsymbol{\Sigma_x}$ ist die (empirische) Kovarianzmatrix der Variablen, σ^2 ist der Wert der Varianz und μ der Mittelwert.

3.2 Multivariate Imputation by Chained Equations

Graham (2012, S. 12ff) unterscheidet zwischen drei Arten von fehlenden Werten: *Missing Completely At Random (MCAR), Missing At Random (MAR)* und *Not Missing At Random (NMAR)*. Ersteres sind Werte, die rein zufällig nicht vorhanden sind und die Wahrscheinlichkeit für das Fehlen ist weder von der Variable, bei der der Wert fehlt, noch von anderen Variablen oder Beobachtungen in den Daten abhängig. Diese können einfach ignoriert werden, da sie keine Verzerrung in der Schätzung verursachen. Die Wahrscheinlichkeit für das Fehlen bei MAR

Werten ist möglicherweise abhängig von anderen beobachtbaren Variablen in den Daten, jedoch unabhängig von der Variable selbst. Betrachtet wird beispielsweise eine Studie, in der das Einkommen und der Bildungsstand erhoben werden. Bei MAR ist die Wahrscheinlichkeit, dass der Wert beim Einkommen fehlt, vom Bildungsniveau der befragten Person abhängig. Die Fehlwahrscheinlichkeit für NMAR Werte hängt hingegen von der fehlenden Variable selbst ab. Somit wäre die Wahrscheinlichkeit, dass der Wert beim Einkommen fehlt, von der Höhe des Einkommens der befragten Person abhängig (vgl. Graham, 2012, S.12ff). Laut Schafer et al. (2002, S. 151) ist „MAR [...] also called ignorable nonresponse, and MNAR is called nonignorable“ (ebd., S. 151), da MNAR zu Verzerrungen in der Analyse führen können. In der Praxis ist es allerdings häufig sehr schwer, die Ursache für das Fehlen genau zu bestimmen. In dieser Arbeit wird für die Imputation angenommen, dass die fehlenden Werte MAR bzw. MCAR entsprechen.

Multivariate Imputation by Chained Equations (MICE) von Stef van Buuren und Karin Groothuis-Oudshoorn ist eine multiple Imputationsmethode zum Umgang mit fehlenden Daten (vgl. van Buuren et al., 2011, S.1f). Bei der multiplen Imputation werden die fehlenden Werte basierend auf den vorhanden Werten mehrmals geschätzt, wodurch sich *m* Datensätze ergeben (vgl. Schafer et al., 2002, S. 165). Die zugrundeliegende Theorie des MICE-Algorithmus wird in van Buuren et al., 2011 genauer beschrieben. Die Imputation durch MICE kann für MAR und MNAR durchgeführt werden, wobei letzteres zusätzliche Modellannahmen erfordert (vgl. ebd., S. 15). Im Allgemeinen wird beim MICE-Verfahren eine Variable mit fehlenden Werten, gemäß ihrer Verteilung durch die anderen Variablen des Datensatzes modelliert, somit kann jede Variable, ob binär, kategorial oder kontinuierlich, imputiert werden (vgl. Azur et al., 2011). Der verkettete Gleichungsprozess kann nach Azur et al. (ebd.) in vier Schritten zusammengefasst werden:

1. Zuerst wird für jeden fehlenden Wert im Datensatz ein „Platzhalter“ eingesetzt, z.B. der Mittelwerte jeder Variable. Dadurch ergibt sich ein vollständiger Datensatz.

2. Anschließend werden die Platzhalter für eine einzelne Variable wieder auf die fehlenden Angaben zurückgesetzt, die beobachteten Werte bleiben dabei unverändert.

3. In diesem Schritt wird ein Regressionsmodell mit den beobachteten Werten der Variable aus Schritt 2 aufgestellt. Diese Variable wird als abhängige Variable betrachtet und alle übrigen Variablen sind unabhängige Variablen. Das Regressionsmodell modelliert die abhängige Variable durch die anderen Variablen in den Daten, unter Berücksichtigung ihrer Verteilung.

4. Daraufhin werden die fehlenden Werte dieser Variable durch Vorhersagen (Imputationen) mithilfe des Regressionsmodells aus Schritt drei geschätzt.

Die Schritte zwei bis vier werden für jede Variable, die fehlenden Angaben enthält, sukzessive durchgeführt. Dabei werden für die unabhängigen Variablen im Regressionsmodell, neben den beobachteten Werten, die bereits imputierten Werte aus Schritt vier verwendet.

Eine Iteration beschreibt das Durchlaufen der Schritte zwei bis vier. Nach einer Iteration wurden alle fehlenden Werte imputiert und der Datensatz ist vollständig. Die Anzahl der Iterationen wird vorher festgelegt und nach jeder Iteration werden die Imputationen aktualisiert (vgl. Azur et al., 2011).

Nach van Buuren et al. (2011, S. 2) ist eine niedrige Anzahl der Iterationen oft ausreichend, wobei die durchschnittliche Anzahl der Iterationen bei zehn liegt, dies ist aber von der jeweiligen Datensituation abhängig (vgl. Azur et al., 2011, S. 42 zitiert nach Raghunathan et al., 2002). Für den Imputationsprozess in dieser Arbeit wurden zehn Iterationen gewählt. Zudem wurde sich für die semi-parametrische Methode *Predictive mean matching* entschieden. Der Vorteil dieser Methode ist zum einen, dass die Imputationenswerte auf die beobachteten Werte beschränkt sind und zum anderen, dass nicht lineare Beziehungen beibehalten werden können (vgl. van Buuren et al., 2011, S. 18). Eine hohe Anzahl m der erstellten Datensätze kann zu einer verbesserten Schätzung führen, mehr Informationen dazu liefert Graham et al., 2007. Für die Hauptkomponentenanalyse kann die gemeinsame Korrelationsmatrix der imputierten Datensätze genutzt werden. Hierfür wurde eine Imputation mit $m = 50$ durchgeführt. Für die weiteren Methoden wurde m dennoch auf fünf gesetzt, da bis auf bei der Hauptkomponentenanalyse die Ergebnisse der imputierten Datensätze nicht kombiniert analysiert werden können. Die Imputation wurde mit den genannten Einstellungen auf dem großen und kleinen Datensatz durchgeführt.

3.3 Ähnlichkeits- und Distanzmaßen

Die Grundlage für einige der multivariaten Methoden sind Ähnlichkeits- bzw. Distanzmaßen. Je größer der Wert des Ähnlichkeitsmaßes bzw. je kleiner der Wert des Distanzmaßes ist, desto ähnlicher sind sich die Objekte. Nach Fahrmeir et al. (2015, S. 440ff) ist ein klassisches Distanzmaß für metrische Merkmale die *Minkowski-q-Metrik* oder L_q*-Distanz*:

$$d_q(x_i, x_l) = \Big(\sum_{j=1}^{p} |x_{ij} - x_{lj}|^q \Big)^{\frac{1}{q}}, \quad q > 1$$

Mit p Variablen und den Merkmalsvektoren $d_q(i, l)$ der Distanz zweier Objekte i und l. Dabei ist x_{lj} der Wert für das Objekt l der Variable j. Häufig angewendet wird die L_1*-Distanz*, die auch *City-Block-Metrik* oder *Manhatten-Metrik* genannt wird

$$d_1(x_i, x_l) = \sum_{j=1}^{p} |x_{ij} - x_{lj}|.$$

Auch die euklidische Distanz ist ein bekanntes Distanzmaß

$$d_2(x_i, x_l) = \Big(\sum_{j=1}^{p} (x_{ij} - x_{lj})^2 \Big)^{\frac{1}{2}} = \|x_i - x_l\|.$$

Der rechnerischen Einfachheit wird oft die quadrierte euklidische Distanz verwendet, welche allerdings nicht zu den *L_q-Distanzen* zählt

$$d_2^2(x_i, x_l) = \|x_i - x_l\|^2$$

(vgl. Fahrmeir et al., 2015, S.440ff).

Eine Distanzmatrix ist die Grundlage für einige der im Folgenden beschrieben Methoden.

3.4 Hauptkomponentenanalyse

Bei multivariaten Analysen ist die Anzahl der Variablen häufig sehr groß. Ziel der PCA, ist es, die Dimension der untereinander korrelierten beobachtbaren Variablen zu reduzieren. Die Methode wurde erstmals von Pearson (1901) eingeführt.
Der Inhalt des folgenden Abschnitts beruht im Wesentlichen auf Fahrmeir et al. (2015, S. 661) und Everitt et al. (2011, S. 61f). Bei der PCA werden die korrelierten Variablen $x^T = (x_1, \ldots, x_p)$ aus den Daten durch neue orthogonale Variablen $y^T = (y_1, \ldots, y_p)$, die sogenannten *Hauptkomponenten*, beschrieben. Um dabei möglichst viel der Gesamtvarianz (also Information) beizubehalten, sind die Hauptkomponenten transformierte lineare Kombinationen aus den beobachtbaren Variablen. Die Hauptkomponenten sind untereinander unkorreliert und werden gemäß ihrem Anteil an erklärender Varianz absteigend geordnet. Somit berücksichtigt die erste Hauptkomponente die meiste Varianz aus den ursprünglichen Daten. Die zweite Hauptkomponente wird dann so gewählt, dass sie die meiste verbleibende Varianz berücksichtigt.

Nach Everitt et al. (ebd., S. 63f) wird bei der ersten Hauptkomponente eine Linearkombination

$$y_1 = a_1^T x, \quad \text{mit} \quad a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

gesucht, welche die größte erklärende Varianz enthält. Also

$$Var(y_1) = Var(a_1^T x) = a_1^T \mathbf{\Sigma_x} a_1 \rightarrow max.$$

Da durch Erhöhen des Koeffizientenvektors $a_1^T = (a_{11}, \ldots, a_{1p})$ die erklärte Varianz von y_1 unbegrenzt ansteigen könnte, muss der Vektor a_1 eingeschränkt werden. Die folgende Nebenbedingung erreicht, dass das Problem wohldefiniert ist (vgl. Jolliffe et al., 2016, S. 2):

$$a_1^T a_1 = 1.$$

Der Gewichtsvektor wird so normiert, dass die Quadratsumme der Koeffizienten den Wert eins annehmen (vgl.Everitt et al., 2011, S.63f).

Um eine Funktion mit einer oder mehreren Nebenbedingungen zu maximieren, wird das Lagrange-Funktional verwendet, welches nach Vidal et al. (2016, S. 27) für die erste Hauptkomponente lautet:

$$\mathcal{L} = a_1^T \mathbf{\Sigma_x} a_1 + \lambda_1(1 - a_1^T a_1).$$

Über die Ableitung ergibt sich

$$\mathbf{\Sigma_x} a_1 = \lambda(a_1) \text{ und } a_1^T a_1 = 1.$$

Das bedeutet, dass a_1 ein Eigenvektor von $\mathbf{\Sigma_x}$ mit dazugehörigem Eigenwert λ ist (vgl. ebd., S. 27). Die Hauptkomponenten können also aus den Eigenvektoren der Kovarianzmatrix $\mathbf{\Sigma_x}$ berechnet werden (vgl. ebd., S.26). Die optimale Lösung für a_1 ist durch den Eigenvektor von $\mathbf{\Sigma_x}$ mit dem dazugehörigen größten Eigenwert $\lambda_1 = \text{Var}(y_1)$ gegeben (vgl. ebd., S.27).

Nach Everitt et al. (2011, S. 64) weist die zweite Hauptkomponente

$$y_2 = a_2^T x, \text{ mit } a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p,$$
$$\text{wobei } a_2^T = (a_{21}, a_{22}, ..., a_{2p})$$

unter folgenden zwei Nebenbedingungen die größte Varianz $Var(y_2)$ auf:

$$(1) \ a_2^T a_2 = 1$$
$$(2) \ a_2^T a_1 = 0 \ \Leftrightarrow \ Cov(y_1, y_2) = 0$$

Die zweite Nebenbedingung gewährleistet, dass die Hauptkomponenten y_1 und y_2 unkorreliert sind. Verallgemeinert entspricht die j-te Hauptkomponente die Linearkombination $y_j = a_j^T x$, die unter folgenden Bedingungen

$$(1) \ a_j^T a_j = 1$$
$$(2) \ a_j^T a_i = 0 \ \ (i < j)$$

die größte Varianz aufweist (vgl. ebd., S. 64). Es können maximal p Hauptkomponenten bestimmt werden.
Unterscheiden sich die Varianzen der Variablen aus den Daten stark, ergibt es durchaus Sinn, die Kovarianzmatrix durch die Korrelationsmatrix zu ersetzten.

Anzahl der Hauptkomponenten

Fahrmeir et al. (2015, S. 668f) fasst vier Kriterien zur Bestimmung der Anzahl der Hauptkomponenten k zusammen, von denen drei für diese Arbeit erwähnt werden. Beim *Eigenwertkriterium* werden nur die Hauptkomponenten berücksichtigt, deren dazugehörigen Eigenwert λ_j größer bzw. gleich den Wert eins hat, also

$$k = max\{j|\lambda_j \geq 1\}.$$

Es ist auch möglich, die Anzahl k anhand eines festgelegten Anteils c an der erklärten Gesamtvarianz q zu bestimmen

$$k = min\{r|\lambda_1 + \cdots + \lambda_r \geq \frac{c}{100}q\}.$$

Die Anzahl für die Hauptkomponenten ergibt sich, sobald der kumulative Anteil der Gesamtvarianz erreicht ist. Auf grafischem Wege lässt sich die optimale Anzahl an Hauptkomponenten ebenfalls ermitteln, mithilfe eines sogenannten *Screeplots*. Ein Beispiel für einen Screeplot ist in Abbildung 3.1 aus Fahrmeir et al. (ebd., S. 669) zu sehen. Der Graph zeigt die nach Größe geordneten Eigenwerte $\lambda_1, \ldots, \lambda_p$. Die Stelle, an der ein Knick entsteht, ist als Kriterium k zu wählen und nur die Hauptkomponenten vor dem Knick werden betrachtet. Bis zu dieser Stelle nehmen die Eigenwerte relativ große Werte an, nach dem Knick flacht die Kurve ab und die Hauptkomponenten tragen weniger zur Varianz bei (vgl. ebd., S.669).

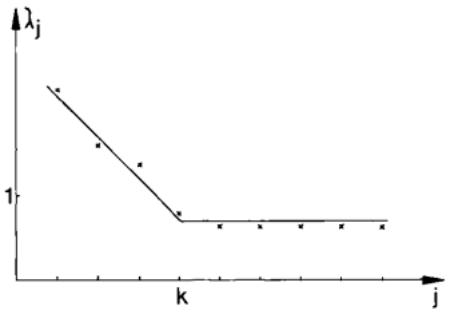

Abbildung 3.1: *Beispielhafter Screeplot zur Verdeutlichung der grafischen Bestimmung der Hauptkomponentenanzahl k (s.h. Fahrmeir et al., 2015, S. 669).*

3.5 Multiple Faktorenanalyse

Die MFA, von Pagès und Escofier (1988–1998), ist sozusagen eine Erweiterung der PCA, die eine Gruppeneinteilung der Variablen berücksichtigt. Im untersuchten Datensatz der OECD sind mehrere Sätze von Variablen (Teildatensätze) für dieselben Beobachtungen (Länder) gegeben. Die Teildatensätze können als Gruppen von Variablen im Gesamtdatensatz betrachtet werden. Die Anzahl und die Art (qualitativ oder quantitativ) der Variablen kann sich zwischen den Gruppen unterscheiden, innerhalb einer Gruppe müssen die Variablen jedoch zur selben Art gehören (vgl. Pagès, 2004, S.1).

Nach Abdi et al. (2013, S.152f) erfolgen drei Schritte bei der MFA:
Im ersten Schritt wird eine PCA auf jede Gruppe, in dem Fall auf jeden der fünf Teildatensätze $g = 1, 2, 3, 4, 5$ durchgeführt. Somit werden die Eigenwerte für jeden Teildatensatz $\lambda_{1,g}, \ldots, \lambda_{p,g}$ berechnet. Damit jede Gruppe denselben Einfluss hat und die MFA nicht von einer der Gruppen mit der stärksten Struktur dominiert wird, wird im zweiten Schritt eine Gewichtung durchgeführt. Jede Variable wird durch die Quadratwurzel des ersten Eigenwertes $\sqrt{\lambda_{1,g}}$ der PCA aus der g-ten Gruppe dividiert. Der erste Eigenwert jedes gewichteten Teildatensatzes hat jetzt ein Wert von eins. Die normalisierten Teildatensätze werden zu einer Matrix $\boldsymbol{X}$ verkettet und unterlaufen nun gemeinsam erneut einer PCA. Im dritten Schritt wird die gemeinsame Struktur analysiert und in einen gemeinsamen Raum projiziert (vgl. ebd., S.152f). Die Faktorwerte der Matrix $\boldsymbol{X}$, welche die Beobachtungen beschreiben, stellen einen Kompromiss bzw. eine gemeinsame Darstellung der Teildatensätze dar (vgl. ebd., S.154). Die MFA-Gewichtung gleicht die Gesamtvarianz der verschiedenen Variablengruppen nicht aus. Nach der Gewichtung ist der erste Eigenwert jeder Gruppe gleich eins, somit wird lediglich sichergestellt, dass die erste Hauptkomponente nicht durch eine Gruppe alleine dargestellt werden kann. Dennoch kann eine Gruppe mit einem großen Einfluss zu zahlreichen Hauptkomponenten beitragen (vgl. Pagès, 2004, S.5).

3.6 Multidimensionale Skalierung

Folgender Abschnitt beruht im Wesentlichen auf Fahrmeir et al. (2015, S.767).
Mit der MDS wird die Beurteilung und Wahrnehmung von Objekten untersucht. Dabei werden Personen meist über die Ähnlichkeit oder Distanz der interessierenden Objekte befragt. Mit der MDS wird versucht, die Objekte aufgrund ihrer Distanzen in einem möglichst niedrig dimensionalen Raum abzubilden. Das Ziel besteht darin, die Merkmale zu finden, welche den Wahrnehmungs- bzw. den Beurteilungsvorgang beeinflussen. Folgende Annahmen sind nach Fahrmeir et al. (ebd., S.767) wichtig, um Informationen über die Merkmale zu erhalten: „Es existiert ein Raum, dessen orthogonale Achsen die gesuchten Merkmale bilden. Ist in dem Merkmalsraum (Wahrnehmungsraum, Urteilsraum) die Distanz zwischen zwei Objekten kleiner als die Distanz zwischen den Objekten eines Vergleichspaares, dann haben die beiden Objekte auch im Urteil der Personen eine größere Ähnlichkeit bzw. eine kleinere Distanz als die Objekte des Vergleichspaares."

Klassische metrische MDS

Die metrische MDS ist auf Torgerson (1952) zurückzuführen. Laut Fahrmeir et al. (2015, S.776ff) erfolgt eine Einteilung in ein *Distanzmodell*, bei dem die Objekte in Distanzen transformiert werden und ein *Raummodell*, das die Distanzen möglichst gut approximieren soll. Nach Fahrmeir et al. (ebd., S. 778f) wird bei der metrischen MDS wie folgt vorgegangen:
Die Distanzen $d_2(x_i, x_l)$ der Punkte x_i und x_l werden meist mit der (quadrierten) euklidischen

Distanz berechnet. Daraufhin wird eine Matrix $\boldsymbol{A}$ mit den Elementen $a_{il} = -\frac{1}{2}d_2(x_i, x_l)$ bestimmt. Anschließend kann die *Skalarproduktmatrix* $\boldsymbol{B} = \boldsymbol{HAH}$ mit der Zentrierungsmatrix $\boldsymbol{H} = I_n - \frac{1}{n}11^T$ gebildet werden.
Durch eine Eigenwertzerlegung der Skalarproduktmatrix in $\boldsymbol{B} = \boldsymbol{P}\Lambda\boldsymbol{P}^T$ werden die Eigenwerte und die dazugehörigen Eigenvektoren bestimmt. Dabei enthält $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ die Eigenwerte und die Spaltenvektoren und $p_1, \ldots, p_n$ der Matrix $\boldsymbol{P}$ sind die orthonormierten Eigenvektoren. Betrachtet werden dabei nur die positiven Eigenwerte ($\lambda_1 \geq \cdots \geq \lambda_t > 0$) mit ihren Eigenvektoren ($y_1 = \sqrt{\lambda_1}p_1, \ldots, y_t = \sqrt{\lambda_t}p_t$). Dabei gilt für jeden positiven Eigenwert λ_s mit $s = 1, \ldots, t$ dass $y_s^T y_s = \lambda_s$. Mit der Matrix $\boldsymbol{Y} = (y_1, \ldots, y_t) \in \mathbb{R}^{n\times t}$ lässt sich die Skalarproduktmatrix $\boldsymbol{B}$ auch als $\boldsymbol{B} = \boldsymbol{YY}^T$ darstellen. Die Konfiguration $\boldsymbol{Y}$ der n Objekte im k-dimensionalen Raum ergibt sich mit den k größten Eigenwerte zu $\boldsymbol{Y} = \boldsymbol{P}_k\Lambda_k^{\frac{1}{2}}$ (vgl. Williams, 2002, S. 12).

Nicht-metrische MDS

Der folgende Abschnitt beruht im Wesentlichen auf Fahrmeir et al. (2015, S.785ff) und Backhaus, Erichson und Weiber (2015, S.361ff). Die nicht-metrische MDS benötigt vollständige Ähnlichkeitsrangordnungen der Objektpaare als Dateninput. Die Ähnlichkeitsrangordnung der Objektpaare und der Objektdistanzen muss folgende Monotoniebedingung erfüllen:

$$u(x_n, x_m) > u(x_i, x_l) \Rightarrow d(x_n, x_m) > d(x_i, x_l).$$

Wenn die Unähnlichkeit $u(x_n, x_m)$ der Objekte x_n und x_m gemäß der Datenerhebung größer sind als die Unähnlichkeit $u(x_i, x_l)$, dann muss auch die Objektdistanz im Merkmalsraum $d(x_n, x_m)$ größer als $d(x_i, x_l)$ sein. Um die Konfiguration in einem Raum mit möglichst geringer Dimensionalität zu finden, die diese Bedingung erfüllt, ist ein iterativer Vorgang nötig. Die nichtmetrische MDS kann entweder mit einer belieben Startkonfiguration beginnen oder es wird das Raummodell der metrischen MDS angewandt. Die Anfangskonfiguration wird schrittweise so verändert, dass sich die Rangfolgen der Distanzen den Unähnlichkeitsmaßen annähern, bis eine möglichst optimale Lösung gefunden wird. Für die Lösung dieser Optimierungsaufgabe wird eine sogenannte monotone Regression der Distanzen auf die Ähnlichkeitsrangordnung aller Objektpaare durchgeführt. Stimmt die Monotoniebedingung bei zwei aufeinanderfolgenden Distanzen nicht, bildet man die *Disparität* $\hat{d}$. Die Disparitäten sind Mittelwerte zwischen den Distanzen der nichtmonotonen Objektpaare. Sie bilden eine schwach monotone Transformation und müssen dabei folgende Monotoniebedingung erfüllen:

$$u(x_n, x_m) > u(x_i, x_l) \Rightarrow \hat{d}(x_n, x_m) \geq \hat{d}(x_i, x_l).$$

Das STRESS-Maß misst die Anpassungsgüte einer Konfiguration an die Monotoniebedingung. Die gefundene Konfiguration wird iterativ so lange weiter verbessert, bis entweder ein minimaler STRESS oder eine vorgegebene Iterationszahl erreicht ist.

3.7 Clusteranalyse

Das Ziel von Clusteranalysen ist, eine Menge von Klassifikationsobjekten, dies können Personen, Aggregate oder Variablen sein, in möglichst homogene Gruppen (Cluster) zusammenzufassen (vgl. Bacher, 2008, S. 15). Innerhalb eines Clusters sind die Objekte also sehr ähnlich zueinander, wohingegen zwischen den Clustern möglichst Heterogenität vorliegen soll. Um die Variation in den Daten zu erklären, sollen die gebildeten Cluster gut an die Daten angepasst sein (vgl. ebd., S. 18). Die Beobachtungen $x_1, \ldots, x_n$ des Datensatzes $\boldsymbol{X}$ werden also in Cluster $C_q = C_1, \ldots, C_k$ geteilt. Dabei ist n_q die Anzahl der Objekte im Cluster C_q mit $q = 1, \ldots, k$.

Es gibt verschiedene Ansätze von Clusterverfahren und im Rahmen dieser Arbeit wird das *hierarchische Klassifikationsverfahren* und das *k-Means Verfahren* vorgestellt.

3.7.1 Hierarchisches Clustering

Ausgangspunkt ist eine symmetrische Distanzmatrix $\boldsymbol{D}$, zum Beispiel mit der euklidischen Distanz berechnet (vgl. Husson, Lê et al., 2017, S.181). Bacher (2008, S. 233f) unterscheidet bei dem hierarchischen Clusterverfahren zwischen dem agglomerativen und dem divisiven Algorithmus. Ersteres Verfahren ordnet zunächst jedes Klassifikationsobjekt einem eigenen Cluster zu, somit ist die Clusterzahl $k = n$. Die Cluster werden nun sukzessive zusammengefasst. Gesucht werden als Erstes zwei Objekte, die die größte Ähnlichkeit bzw. die geringste Unähnlichkeit aufweisen. Diese werden zu einem Cluster zusammengefügt, wodurch sich k um eins verringert. Nun werden erneut die Distanzen des neu gebildeten Clusters zu den übrigen Clustern berechnet und wieder Objekte bzw. Cluster zu einem neuen Cluster agglomeriert. Die Beurteilung der Distanzen zwischen Objekten erfolgt anhand der Distanzmatrix, während die Distanzen zwischen Clustern durch ein *Linkage-Verfahren* berechnet werden. Die Clusterbildung wird so lange fortgeführt, bis alle Objekte schließlich in einem Cluster sind (vgl. ebd., S.233f).

Das divisive Verfahren hingegen startet mit einem Cluster, in dem sich alle Objekte befinden und teilt anschließend die Cluster sukzessive auf, bis sich jedes Objekt in einem eigenen Cluster befindet (vgl. Fahrmeir et al., 2015, S. 453). Mehr zum divisiven Verfahren findet sich beispielsweise im Fahrmeir et al. (ebd.) oder im Bacher (2008).

Linkage-Verfahren

Wie bereits erwähnt, werden durch Linkage-Verfahren Distanzen zwischen Clustern bestimmt. Mögliche Verfahren sind das Single Linkage, Complete Linkage, Average Linkage, Zentroid Linkage oder das Ward Verfahren. In dieser Arbeit wurde häufig das Ward Verfahren gewählt.

Dabei wird die Gesamtvarianz (engl. *total inertia*) folgendermaßen zerlegt:

$$\sum_{j=1}^{p}\sum_{q=1}^{k}\sum_{i=1}^{n_q}(x_{iqj}-\bar{x}_j)^2 = \sum_{j=1}^{p}\sum_{q=1}^{k} n_q(\bar{x}_{qj}-\bar{x}_j)^2 + \sum_{j=1}^{p}\sum_{q=1}^{k}\sum_{i=1}^{n_q}(x_{iqj}-\bar{x}_{qj})^2,$$

$$\text{Gesamtvarianz } \mathbf{T} = \text{Inter-Cluster-Varianz } \mathbf{B} + \text{Intra-Cluster-Varianz } \mathbf{W}$$

(vgl. Husson, Josse et al., 2010, S. 4). Für das Cluster C_q ist x_{iqj} der Wert der Variable j für die Beobachtung i. Der Mittelwert der Variable j entspricht $\bar{x}_j$ und $\bar{x}_{qj}$ ist der Mittelwert der Variable j im Cluster C_q. Die Anzahl der Beobachtungen im Cluster C_q ist durch n_q gegeben. Bei dem Verfahren werden die Cluster fusioniert, die die kleinste Erhöhung der *within-inertia* haben (vgl. ebd., S. 4). Beispielsweise bieten Backhaus, Erichson, Gensler et al. (2021) mehr Informationen zu anderen Linkage-Verfahren.

3.7.2 Optimale Partitionen

Die optimalen Partitionen gehen von einer Startpartition mit einer vorgegebenen Clusteranzahl k aus. Laut Backhaus, Erichson, Gensler et al. (ebd., S. 565) wird die Qualität der Partitionen anhand eines *Gütekriteriums* gemessen. Das Ziel bei diesem Verfahren ist es, eine Gruppierung zu finden, die hinsichtlich des Gütekriteriums optimal ist. Dies wird mithilfe eines Austauschalgorithmus erreicht. Die Objekte werden zwischen den Clustern iterativ getauscht, bis keine Verbesserung mehr im Wert des Gütekriteriums eintritt (vgl. ebd., S.565).
Eine bekannte Methode ist das *k-Means* Verfahren, welches sich nach Backhaus, Erichson, Gensler et al. (ebd., S. 566) in vier Schritte einteilen lässt. Im ersten Schritt werden für die k Gruppen zufällige Clusterzentren μ_q gewählt. Anschließend wird die euklidische Distanz zwischen den Clusterzentren und den Datenpunkten x_i berechnet. Eine Beobachtung wird dann zu dem Cluster geordnet, bei dem das *Varianzkriterium (Z)* am wenigsten vergrößert wird

$$Z = \sum_{q=1}^{k}\sum_{x_i \in C_q} \|x_i - \mu_q^2\|.$$

Im dritten Schritt werden die neuen Clusterschwerpunkte aus den Mittelwerten der Beobachtungen in jedem Cluster C_q berechnet. Die Objekte werden wieder dem Cluster zugeordnet, dessen Clusterzentrum am nächsten liegt. Jedes Mal, nachdem mindestens ein Objekt zwischen den Cluster verschoben wurde, werden die Clusterzentren neu berechnet. Dies wird so lange wiederholt, bis die Varianzen in den Clustern durch weiteres Verschieben nicht mehr verringert werden können und der Prozess beendet ist (vgl. ebd., S. 566ff).

3.7.3 Clustervalidierung

Um die gefundenen Partitionen in den Daten möglichst objektiv vergleichen zu können, werden verschiedene Maße benutzt. Mithilfe des *kophenetische Korrelationskoeffizienten* lassen sich

Partitionslösungen zweier hierarchischen Clusteringverfahren vergleichen.
Zur Beurteilung der Güte der Cluster und der Ermittlung der optimalen Anzahl an Cluster dienen drei bekannte interne Validierungsindizes.

Kophenetischer Korrelationskoeffizient

Der kophenetische Korrelationskoeffizient

$$c = \frac{\sum_{i<l}(d(x_i, x_l) - \bar{d})(t(x_i, x_l) - \bar{t})}{\sqrt{[\sum_{i<l}(d(x_i, x_l) - \bar{d})^2][\sum_{i<l}(t(x_i, x_l)) - \bar{t})^2]}}$$

ist ein Maß zur Bestimmung, wie genau ein Dendrogramm die paarweisen Distanzen zwischen den ursprünglichen Beobachtungen bewahrt (vgl. Saraçli et al., 2013, S. 2). Dabei ist $t(x_i, x_l)$ die Höhe der Knoten im Dendrogramm, bei denen die Beobachtungen x_i und x_l erstmals miteinander verbunden werden und $\bar{t}$ der daraus resultierende Durchschnittswert. Für die Distanzen $d(x_i, x_l)$ zwischen den Beobachtungen wird ebenfalls die Durchschnittsdistanz $\bar{d}$ berechnet. Mit dem kophenetische Korrelationskoeffizient können zwei Dendrogramme verglichen werden. Erstmals eingeführt wurde diese Methode von Sokal et al. (1962). In R wird mit der Funktion *cor.dendlist()* eine kophenetische Korrelationsmatrix zwischen Dendrogrammen bestimmt. Die Werte sind zwischen -1 und 1 beschränkt, wobei Werte nahe an null bedeuten, dass sich die beiden Dendrogramme statistisch nicht sonderlich ähneln (vgl. Kassambara, 2017, S. 82).

Interne Validierungsindizes

Als Maß zur Beurteilung der Clusterlösung, sowie der optimalen Clusteranzahl dienen drei interne Validierungsindizes: der Average Silhouette Width (ASW), der Dunn-Index und der Calinski-Harabasz-Index (CH-Index). Für die Indizes werden keine externe Information über die „wahre“ Clusterbildung benötigt. Die Bewertung erfolgt stattdessen anhand der Homogenität innerhalb der Cluster, deren Trennung und Kompaktheit.

Nach Martinez et al. (2017, S. 211f) wird die durchschnittliche Unähnlichkeit für eine Beobachtung i zu allen anderen Beobachtungen im selben Cluster als a_i bezeichnet. Für alle weiteren Cluster wird die durchschnittliche Distanz der Beobachtungen im Cluster zu der Beobachtung i berechnet. Daraus kann das Minimum b_i aller Durchschnittsdistanzen gewählt werden. Der ASW ist wie folgt definiert (vgl. ebd., S. 212):

$$\text{ASW} = \frac{1}{n}\sum_{i=1}^{n}\frac{b(i) - a(i)}{\max\{a(i); b(i)\}}$$

Der ASW kann Werte zwischen -1 und 1 annehmen. Umso höher der Wert, desto homogener sind die Beobachtungen innerhalb eines Clusters und desto besser ist die Trennung zwischen den Clustern (vgl. Hennig et al., 2015, S. 604).

Mit dem Dunn-Index

$$\text{Dunn} = \min_{q=1,\dots,k} \left(\min_{r=q+1,\dots,k} \left\{ \frac{d(C_q, C_r)}{\max\limits_{q=1,\dots,k} \{\text{diam}(C_q)\}} \right\} \right)$$

wird das Verhältnis zwischen der Trennung von Clustern und der Kompaktheit innerhalb Cluster bestimmt (vgl. Martinez et al., 2017, S. 221). Dabei ist $d(C_q, C_r)$ die Distanz zwischen zwei Clustern und $diam(C_q)$ der Durchmesser eines Clusters bzw. die Variation der Beobachtungen in einem Cluster. Ein großer Dunn-Index spricht für gut getrennte und kompakte Cluster (vgl. Martinez et al. (ebd., S. 221)).

Das Verhältnis der Variation innerhalb des Clusters (B) und zwischen den Clustern (W) wird mit dem CH-Index

$$\text{CH} = \frac{\text{trace}(B)}{\text{trace}(W)} x \frac{n-k}{k-1},$$

bestimmt (vgl. Hennig et al., 2015, S. 599). Die Zerlegung der Gesamtvarianz $T = B + W$ ist in Kapitel 3.7.1 definiert. Auch beim CH-Index wird nach einem Maximum gesucht.

3.8 Uniform Manifold Approximation and Projection

UMAP ist eine neuartige Dimensionsreduktionstechnik von McInnes et al. (2018). Der Algorithmus kann in zwei Schritte eingeteilt werden. Die Erläuterung des Verfahrens basiert auf McInnes et al. (ebd.) und Wang et al. (2021):

Graphkonstruktion für den hochdimensionalen Raum

Im ersten Schritt wird versucht, die Mannigfaltigkeit approximiert darzustellen, indem ein *gewichteter k-Nachbarpraph* erstellt wird (vgl. McInnes et al., 2018, S.4). Für jede Beobachtung x_i wird die Menge der k-nächsten Nachbarn $\{x_{i_1}, \dots, x_{i_k}\}$ um x_i für eine gegebene Distanzmetrik berechnet (vgl. Wang et al., 2021, S. 9). Dafür wird für jede Beobachtung der minimale positive Abstand von x_i zu einem Nachbarn x_{i_j} bestimmt:

$$\rho_i = \min\{d(x_i, x_{i_j}) | 1 \leq j \leq k, \;\; d(x_i, x_{i_j}) > 0\}.$$

Durch ρ_i wird sichergestellt, dass x_i mindestens mit einem weiteren Datenpunkt durch eine Kante mit dem Gewicht eins verbunden ist (vgl. McInnes et al., 2018, S. 15). Es wird außerdem ein Skalierungsparameter σ_i definiert, sodass

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i} \right) = \log_2(k).$$

Der Parameter σ_i normiert die Abstände zwischen den Beobachtungen und ihren Nachbarn, um die relativen hochdimensionalen Ähnlichkeiten zu bewahren (vgl. Wang et al., 2021, S. 9).

Daraufhin kann ein gewichteter und gerichteter Graph $\bar{\boldsymbol{G}} = (\boldsymbol{V}, \boldsymbol{E}, w)$, bestehend aus den Knoten $\boldsymbol{V}$, einer Menge gerichteter Kanten $\boldsymbol{E}$ und einer Gewichtsfunktion w definiert werden (vgl. McInnes et al. (2018, S. 15)). Mit der Menge der gerichteten Kanten $\boldsymbol{E} = \{(x_i, x_{i_j})|1 \leq j \leq k, \ \ 1 \leq i \leq n\}$ und

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right).$$

Die Menge der lokalen Graphen müssen nun zu einer einheitlichen topologischen Darstellung kombiniert werden (vgl. McInnes et al. (ebd., S. 15):
Dafür sei $\boldsymbol{A}$ eine gewichtete Adjazenzmatrix von $\bar{\boldsymbol{G}}$ und $\boldsymbol{B}$ ist eine symmetrische Matrix gegeben durch

$$\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{A}^T - \boldsymbol{A} \circ \boldsymbol{A}^T$$

$\circ$ beschreibt dabei ein paarweises Produkt zweier Matrizen.

Der Wert $\boldsymbol{A}_{ij}$ kann als die Wahrscheinlichkeit, dass die gerichtete Kante von x_i nach x_j existiert, interpretiert werden. Daraus folgt, dass $\boldsymbol{B}_{ij}$ die Wahrscheinlichkeit ist, dass mindestens eine der beiden gerichteten Kanten (von x_i nach x_j oder umgekehrt) existiert (vgl. ebd., S.16). Der Graph $\boldsymbol{G}$ ist somit ein ungerichteter gewichteter Graph mit Adjazenzmatrix $\boldsymbol{B}$, der die Struktur der hochdimensionalen Daten angemessen beschreibt (ebd., S. 16).

Optimierung des niedrig dimensionalen Graphlayouts

In diesem Schritt wird ein *kraft-gerichteter Graph-Layout-Algorithmus* im niedrigdimensionalen Raum angewandt (vgl. ebd., S. 16). Die gewünschten Eigenschaften des k-Nachbargraphen sollen dabei erhalten bleiben (vgl. ebd., S. 14). Der Algorithmus visualisiert die Graphen durch iteratives Annähern von Punkten im niedrigdimensionalen Raum, die im hochdimensionalen Raum nahe beieinander liegen und Auseinanderschieben von Punkten im niedrigdimensionalen Raum, die im hochdimensionalen Raum weiter voneinander entfernt sind (Wang et al., 2021, S. 10). Dabei werden iterativ anziehende Kräfte entlang der Kanten und abstoßende Kräfte zwischen den Knoten angewandt (vgl. McInnes et al., 2018, S. 16).

Die anziehende Kraft zwischen zwei Knoten i und j an den Koordinaten y_i bzw. y_j wird durch

$$\frac{-2ab\|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2} w((x_i, x_j))(y_i - y_j),$$

mit den Hyperparametern a und b bestimmt.

Die abstoßende Kraft ist durch

$$\frac{2b}{(\epsilon + ||y_i - y_j||_2^2)(1 + a||y_i - y_j||_2^{2b})}(1 - w((x_i, x_j)))(y_i - y_j)$$

gegeben, wobei die Konstante ϵ einen sehr kleinen Wert annimmt, um eine Division mit Null zu vermeiden.

Der gewichtete Graph $\boldsymbol{G}$ erfasst die Topologie der Originaldaten im hochdimensionalen Raum. Der gewichtete Graph $\boldsymbol{H}$, aus den Punkten y_i mit $i = 1, \ldots, n$ soll den Graphen $\boldsymbol{G}$ im niedrig dimensionalen Raum so gut wie möglich repräsentieren (vgl. McInnes et al., 2018, S. 17). Mithilfe der anziehenden und abstoßenden Kräfte wird die Kreuzentropie zwischen den beiden topologischen Darstellungen $\boldsymbol{G}$ und $\boldsymbol{H}$ minimiert (vgl. ebd., S. 4, S.17).

Bei der Umsetzung der Methode in R spielen vor allem zwei Hyperparameter eine große Rolle: Die Anzahl der Nachbarn im Radius um x_i wird durch *n_neighbors* und der minimale Abstand zwischen den Punkten im nieder-dimensionalen Raum wird durch *min_dist* festgelegt. Bei kleineren Werten für die Anzahl der Nachbarn werden eher feinere Strukturen in den Daten erfasst, während größere Werte die Struktur als Ganzes abbilden, allerdings mit einem gewissen Verlust an Details, der Parameter steuert also die lokale bzw. globale Struktur (vgl. ebd., S. 23). Mit dem Parameter *min_dist* wird bestimmt, wie eng die Punkte in der niedrig dimensionalen Darstellung geplottet werden. Ein niedriger Wert führt zu einer dichten, gepackten Einbettung und höhere Werte zu einer eher lockeren Verteilung der Punkte, was potenzielle Overplotting-Probleme vermeiden soll (vgl. ebd., S. 23). Für die Gesundheitsdaten der OECD Länder wurde sich hier für einen Wert von *n_neighbors* = 5 und *min_dist* = 0.1 entschieden. Mögliche andere Werte für die Anzahl der Nachbarn werden im Kapitel 4 diskutiert. Für mehr Informationen bietet McInnes (2022, S. 17ff) eine Vorstellung verschiedener Hyperparameter für einen deutlich größeren Datensatz als der hier verwendete.

3.9 t-Distributed Stochastic Neighbor Embedding

Die Methode *t-SNE* von van der Maaten und Hinton (2008) ist eine Variation der *Stochastic Neighbor Embedding (SNE)* Technik. Der Unterschied zwischen t-SNE und SNE liegt in der Kostenfunktion. Bei t-SNE wird eine symmetrisierte Version der SNE-Kostenfunktion mit einfacheren Gradienten verwendet. Zudem wird eine Student-t Verteilung anstatt der ursprünglichen Gauß-Verteilung genutzt, um die Ähnlichkeiten im niedrig dimensionalen Raum zu berechnen (vgl. ebd., S. 2583).
Als Erstes werden nach van der Maaten und Hinton (ebd., S. 2581) die Distanzen im hochdimensionalen Raum in bedingte Wahrscheinlichkeiten $p_{j|i}$ umgewandelt. Diese spiegeln die Ähnlichkeiten zwischen der Beobachtung x_j und x_i wider. Dabei ist $p_{j|i}$ die Wahrscheinlichkeit, dass x_i die Beobachtung x_j als Nachbarn auswählt, unter der Bedingung, dass Nachbarn im Verhältnis zu ihrer Wahrscheinlichkeitsdichte einer zentrierten Gauß-Verteilung um x_i aus-

gewählt werden (vgl. van der Maaten und Hinton, 2008, S. 2581). Die bedingte Wahrscheinlichkeit $p_{j|i}$ berechnet sich wie folgt:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}.$$

Dabei ist σ_i die Varianz der Gauß-Verteilung zentriert um x_i.
Die gemeinsame Wahrscheinlichkeit p_{ij} im hochdimensionalen Raum ist definiert als:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}.$$

Die Wahrscheinlichkeit p_{ij} berücksichtigt dabei Ausreißer im hochdimensionalen Raum (vgl. ebd., S. 2584).

Während im hochdimensionalen Raum Entfernungen mithilfe einer Gaußschen Verteilung in Wahrscheinlichkeiten umgewandelt werden, wird im niedrig dimensionalen Raum eine Student-t-Verteilung mit einem Freiheitsgrad genutzt (vgl. ebd., S. 2585). Somit ergeben sich die gemeinsamen Wahrscheinlichkeiten q_{ji} für y_i und y_j als:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}.$$

Da nur paarweise Ähnlichkeiten modelliert werden, haben p_{ii} und q_{ii} einen Wert von Null. Wären die Wahrscheinlichkeiten gleich ($p_{ij} = q_{ij}$), würden die Abbildungspunkte $\boldsymbol{Y}$ die Ähnlichkeiten zwischen den ursprünglichen Datenpunkten $\boldsymbol{X}$ im hochdimensionalen Raum korrekt modellieren (vgl. ebd., S. 2581).

Ein Maß für die Genauigkeit der Modellierung ist die Kullback-Leibler-Divergenz (vgl. ebd., S. 2581). Dafür wird eine Kostenfunktion C aufgestellt. Bei der Methode t-SNE wird dafür eine symmetrische Version der SNE-Kostenfunktion verwendet, welche eine einfachere Form ihres Gradienten aufweist, der schneller zu berechnen ist. (vgl. ebd., S. 2584). Die Kostenfunktion C minimiert eine einzelne Kullback-Leibler-Divergenz zwischen einer gemeinsamen Wahrscheinlichkeitsverteilung im hochdimensionalen Raums P und einer gemeinsamen Wahrscheinlichkeitsverteilung im niedrig dimensionalen Raums Q:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Sie wird als symmetrisch bezeichnet, da für die Wahrscheinlichkeit $p_{ij} = p_{ji}$ und $q_{ij} = q_{ji}$ gilt (vgl. ebd., S. 2583).

Bei der Methode t-SNE wird sozusagen versucht, eine Verteilung, die die paarweisen Ähnlichkeiten p_{ij} in den ursprünglichen Daten misst, aufzustellen und daraufhin eine Verteilung der paarweisen Ähnlichkeiten q_{ij} im niedrig dimensionalen Raum zu finden, die p_{ij} nach der Kullback-

Leibler-Divergenz möglichst gut modellieren (vgl. van der Maaten, 2014, S. 4).

Bei der Implementierung der Methode in R muss vorher ein Hyperparameter, die *Perplexität* festgelegt werden. Dieser kann laut van der Maaten und Hinton (2008, S. 2582) interpretiert werden, als „a smooth measure of the effective number of neighbors“. Die Werte für den Parameter sollten nach van der Maaten und Hinton (ebd., S. 2582) zwischen fünf und fünfzig liegen. Mit dem R-Packet *Rtsne* wird das Maximum der Werte für die Perplexität mit $3 * \text{Perplexität} < \text{nrow}(X) - 1$ beschränkt, wodurch sich einen Perplexitätswert unter elf für diese Daten ergibt. Zur Gewinnung einer ersten Intuition für den passenden Wert stellt Oskolkov (2019a) eine Faustregel mit $n^{\frac{1}{2}}$ auf. Für $n = 34$ Länder entspricht dies einem gerundeten Wert von 5.8. Für die Durchführung der Methoden wurde hierfür ein Wert von fünf, genau wie bei der Methode UMAP gewählt. Die Wahl der Perplexität wird in Kapitel 4 erläutert.

Kapitel 4

Ergebnisse

Das Ziel dieser Bachelorarbeit ist es, die vorgestellten multivariate Methoden zu vergleichen. Als Erstes werden hierfür die Lösungen im niederdimensionalen Raum von der PCA, der MFA, der MDS, UMAP und t-SNE für den kleinen und den großen Datensatz vorgestellt. Anschließend werden die fünf Methoden in Kombination mit dem hierarchischen Verfahren, sowie dem *k*-Means Verfahren betrachtet. Zur Beurteilung der Übereinstimmung der Dendrogrammen dient der kophenetische Korrelationskoeffizient. Die Ermittlung der optimalen Anzahl an Clustern erfolgte an Hand von drei internen Validierungsindizes: dem *Average Silhouette Width*, dem *Dunn-Index* und dem *Calinski-Harabasz-Index*.

Wie in Kapitel 3.2 erwähnt, werden beim MICE Verfahren mehrere imputierte Datensätze erstellt. Bei der Anwendung multivariater Verfahren gestaltet es sich allerdings schwierig, die imputierte Datensätze gemeinsam zu untersuchen. Aus diesem Grund wurde hauptsächlich nur ein imputierter Datensatz für die Analysen genutzt. Die nachfolgenden Ergebnisse beziehen sich somit nur auf den ersten der fünf erstellten Datensätze. Mögliche Unterschiede zwischen den Imputationen, werden bei der Vorstellung der PCA Ergebnisse kurz angeführt.

4.1 Methoden

PCA

Durch die PCA können beim kleinen Datensatz achtzig Prozent der Varianz mit den ersten elf Hauptkomponenten erklärt werden. Die ersten beiden Hauptkomponenten erhalten dabei zusammen ca. ein Drittel der Varianz. Beim großen Datensatz hingegen erklären bereits die ersten zehn Hauptkomponenten achtzig Prozent der Varianz, dabei ist der erklärte Anteil der ersten beiden Hauptkomponenten bei 43 Prozent. Beim Betrachten der Länder auf den ersten beiden Hauptkomponenten erkennt man einige Unterschiede zwischen den Datensätzen. Die Abbildung 4.1 zeigt die Länder auf den ersten beiden Hauptkomponenten des kleinen Datensatzes, und die Abbildung 4.2 des großen Datensatzes. Die Länder wurden anschließend in sechs Re-

gionen, *Amerika, Ost- und Westeuropa, Ost- und Westasien und Ozeanien* eingeteilt, diese sind in den Abbildungen farblich markiert.

Für den keinen Datensatz hat vor allem die Variable *Neugeborenensterblichkeit* und die Variable *Säuglingssterblichkeit* einen hohen Beitrag zur ersten Dimension. Die weiteren acht Variablen mit dem meisten Beitrag zur ersten Dimension sind: die Dichte der Fachärzt*innen, die Dichte der medizinischen Fachgruppe, der Anteil der Bevölkerung mit einem schlechten/sehr schlechten wahrgenommenen Gesundheitszustand, kurativen (Akut-)Pflegebetten pro 1 000 Einwohner, die Inzidenz an AIDS Erkrankten, die perinatalen Sterblichkeitsrate, die Krankenhausbetten pro 1 000 Einwohner und die Dichte der Chirurg*innen. Die Dichte für eine Variable ist immer pro 1 000 Einwohner, wenn nicht anders angegeben.
Korrelieren Variablen dabei positiv mit der ersten Dimension, befinden sich Länder, die ein vergleichsweise hohen Wert in diesen Variablen haben, auf der rechten Seite der Abbildung 4.1. Korrelieren Variablen hingegen negativ mit der ersten Dimension, sind Länder mit hohen Werten in diesen Variablen eher links und Länder mit niedrigeren Werten eher rechts in der Abbildung 4.1. Die Dichte der Fachärzt*innen und der medizinischen Fachgruppe korreliert positiv am stärksten mit der ersten Dimension. Beispielsweise ist die Dichte der Fachärzt*innen in Japan bei 2.600 und im Vergleich dazu bei Kolumbien bei 0.497. Stark negativ korrelieren vor allem die Neugeborenensterblichkeit und die Säuglingssterblichkeit mit der ersten Dimension. Beispielsweise ist die Neugeborenensterblichkeit pro 1000 Lebendgeburten in Kolumbien bei 6.900 und in Japan bei 0.900. Auf der ersten Dimension scheinen sich die Länder *Kolumbien* und *Japan* am stärksten zu unterscheiden.
Die zehn Variablen mit den größten Beiträgen zur zweiten Dimension, in absteigender Reihenfolge, sind: die Dichte der Gesamtbeschäftigung im sozialen Bereich, der Anteil der täglichen Raucher in der Bevölkerung, der mittlere und gute/ sehr gute wahrgenommene Gesundheitszustand, die Lebenserwartung ab der Geburt, die Fettzufuhr, die Muttersterblichkeit, die Dichte der Psychiater*innen, ein schlechter/sehr schlechter wahrgenommener Gesundheitszustand und die Lebenserwartung der männlichen Achtzigjährigen. Die zweite Dimension korreliert am stärksten positiv mit der Dichte der Gesamtbeschäftigung im Gesundheits- und Sozialwesen und negativ mit dem Anteil der täglichen Raucher. In dieser Dimension unterscheiden sich die Länder Kolumbien und Island am stärksten. Beispielsweise liegt die Lebenserwartung ab der Geburt der kolumbianischen Bevölkerung bei 76.4 Jahren und in Island hingegen bei 82.933 Jahren.
Kolumbien, Chile und die Türkei wirken beim kleinen Datensatz wie Ausreißer, da diese relativ weit von den restlichen Ländern entfernt sind. Zum Vergleich wurden diese Länder entfernt und die PCA auf dem kleinen Datensatz erneut durchgeführt. Die neue Anordnung der Länder zeigt die Abbildung A.1 im Anhang A. Die Länder Japan und Korea sind im Vergleich zu vorher weiter von den anderen Ländern separiert. Für einen einheitlichen Vergleich zwischen allen Methoden werden alle 34 Ländern, also auch Kolumbien, die Türkei und Chile betrachtet.

Zur Erinnerung wurden beim großen Datensatz die Variablen *wahrgenommener Gesundheits-*

zustand und *Lebenserwartung* für verschiedene Varianten, wie zum Beispiel nach Geschlecht oder Altersgruppen aufgenommen. Die erste Hauptkomponente wird vor allem durch den wahrgenommenen Gesundheitszustand beschrieben. Dieser ist in drei Abstufungen gegliedert: guter/sehr guter Gesundheitszustand, mittlerer Gesundheitszustand und schlechter/sehr schlechter Gesundheitszustand. In der Abbildung 4.2 befinden sich Länder, deren Bevölkerung einen hohen Anteil an einem guten Gesundheitszustand hat, weiter links. Hingegen befinden sich Länder, deren Bevölkerung eher einen schlechten oder mittleren Gesundheitszustand haben, weiter rechts in der Abbildung 4.2. Da der Gesundheitszustand für verschiedene Gruppen angegeben ist, fließt die Variable öfter in die PCA ein und ist im Vergleich zu anderen Variablen womöglich über repräsentativ. Dies könnte ein Grund sein, warum die erste Dimension hauptsächlich mit dem Gesundheitszustand korreliert. Allerdings korrelieren die Variablen des Gesundheitszustandes auch im kleinen Datensatz mit der ersten Dimension, sie erklärt also bei beiden Datensätzen einen großen Anteil. Die zweite Dimension wird vor allem durch die Variablen, die die Lebenserwartung erfassen, beschrieben. Länder weiter unten in der Abbildung 4.2 haben tendenziell eine höhere Lebenserwartung.

Allgemein ist zu erkennen, dass sowohl im kleinen, als auch im großen Datensatz die Länder innerhalb einer Region recht nahe beieinander angeordnet werden. Dies ist vor allem bei den Ländern aus Ost-, Westeuropa und Ozeanien der Fall.

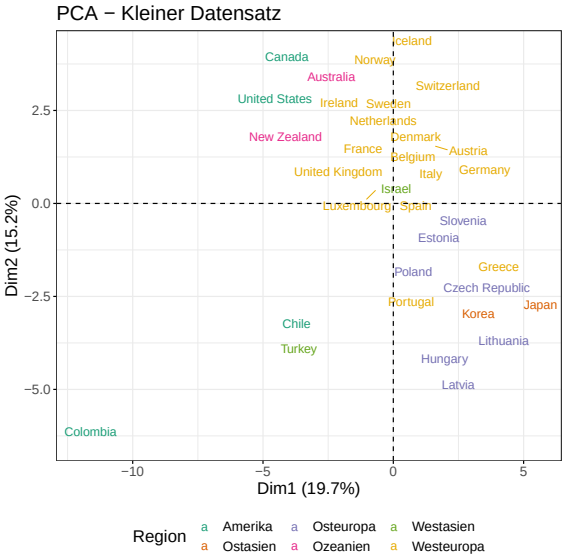

Abbildung 4.1: *Visualisierung der ersten beiden Dimensionen der PCA des kleinen Datensatzes.*

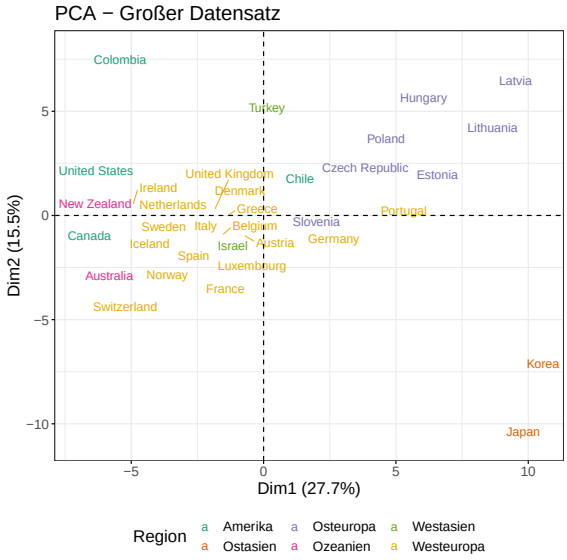

Abbildung 4.2: *Visualisierung der ersten beiden Dimensionen der PCA des großen Datensatzes.*

Die Abbildungen 4.1 und 4.2 zeigen die Ergebnisse für den ersten imputierten Datensatz. Zum Vergleich der Ergebnisse der PCA auf den restlichen vier Imputationen dient die Abbildung A.1 im Anhang A. Hier lassen sich kleine Abweichungen erkennen, im Allgemeinen sind aber keine großen Unterschiede festzustellen. Beim Imputationsverfahren wurde für den kleinen Datensatz ebenfalls eine gemeinsame Kovarianzmatrix von insgesamt m=50 imputierten Datensätzen erstellt, vergleiche Kapitel 3.2. Aus der daraus gebildeten Korrelationsmatrix ließen sich die

Eigenwerte und Eigenvektoren ermitteln. Die Tabelle A.1 zeigt, dass sich die erklärte Varianz nicht groß mit den Ergebnissen von oben unterscheiden. Die Durchführung der Methoden auf nur einen der imputierten Datensätze liefert solide Ergebnisse.

MFA

Sowohl beim kleinen als auch beim großen Datensatz werden mindestens achtzig Prozent der Varianz ab der elften Dimension erklärt. Der Anteil der erklärten Varianz durch die ersten beiden Hauptkomponenten ist dabei beim kleinen Datensatz geringfügig höher. Wie in Kapitel 3.5 erklärt, wurden bei der MFA die Variablen der fünf zusammengefügten Datensätze als fünf Gruppen betrachtet. Die fünf Gruppen umfassen Variablen zu der Inanspruchnahme des Gesundheitswesens, nichtmedizinische Determinanten der Gesundheit, Ressourcen für das Gesundheitswesen, Indikatoren für Gesundheitsausgaben und zum Gesundheitszustand, wie in Kapitel 2 genannt. Die Abbildung A.3 im Anhang A zeigt die Beiträge der Variablengruppen auf den ersten beiden Dimensionen. In beiden Datensätzen spielen Variablen zum Gesundheitszustand auf der ersten Dimension und nichtmedizinische Determinanten auf der zweiten Dimension eine große Rolle. Im großen Datensatz ist die Variablengruppe Inanspruchnahme des Gesundheitswesens ebenfalls auf der ersten Dimension repräsentativ. Im kleinen Datensatz hingegen die Gruppe Ressourcen für das Gesundheitswesen. Die Abbildung 4.3 und die Abbildung 4.4 zeigen die Länder auf den ersten beiden Dimensionen des kleinen Datensatzes bzw. des großen Datensatzes eingefärbt nach den Regionen. Für den kleinen Datensatz stimmt das Ergebnis sehr mit dem der PCA überein. Auch das Ergebnis beim großen Datensatz ähnelt ebenfalls sehr dem der PCA. Zum Beispiel sind in beiden Methoden die Länder Kolumbien, Chile, Japan und Korea weiter entfernt zu den restlichen Ländern. Allerdings werden, im Gegensatz zur PCA, zum einen die westasiatischen Länder Israel und Türkei in Abbildung 4.4 deutlich näher zueinander eingeordnet und zum anderen Griechenland mit einem größeren Abstand zu den restlichen Ländern zugeordnet.

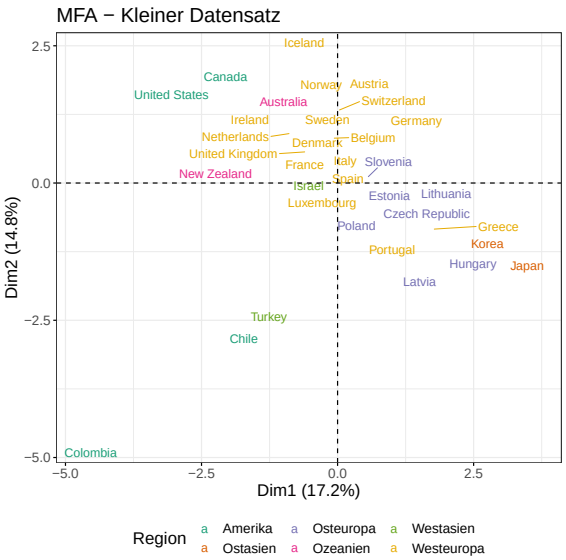

Abbildung 4.3: *Visualisierung der ersten beiden Dimensionen der MFA des kleinen Datensatzes.*

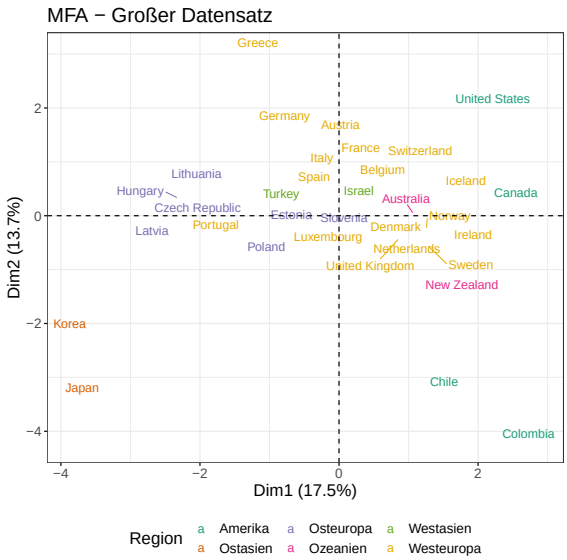

Abbildung 4.4: *Visualisierung der ersten beiden Dimensionen der MFA des großen Datensatzes.*

MDS

Die Distanz zwischen den Objekten der metrischen MDS wurde mit der Manhattan-Metrik berechnet. Die Abbildung 4.5 zeigt die ermittelte Konfiguration der Länder für den kleinen Datensatz und die Abbildung 4.6 für den großen Datensatz. Ähneln sich Länder bezüglich ihrer Variablen, sind diese nah beieinander. Allgemein kommt die MDS zu einer sehr ähnlichen Konfiguration wie die PCA. Die Ermittlung der Konfiguration auf Basis der Korrelationsmatrix, sowie auf Basis der Distanzmatrix mit der Manhattan-Metrik kommen zu sehr ähnlichen Resultaten.

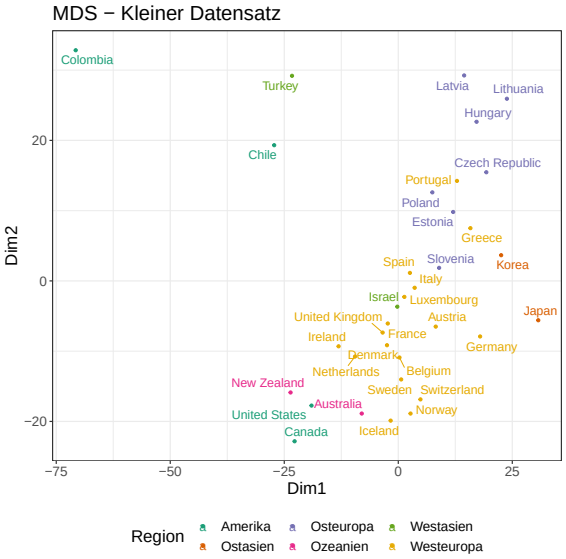

Abbildung 4.5: *Visualisierung der ersten beiden Dimensionen der MDS des kleinen Datensatzes.*

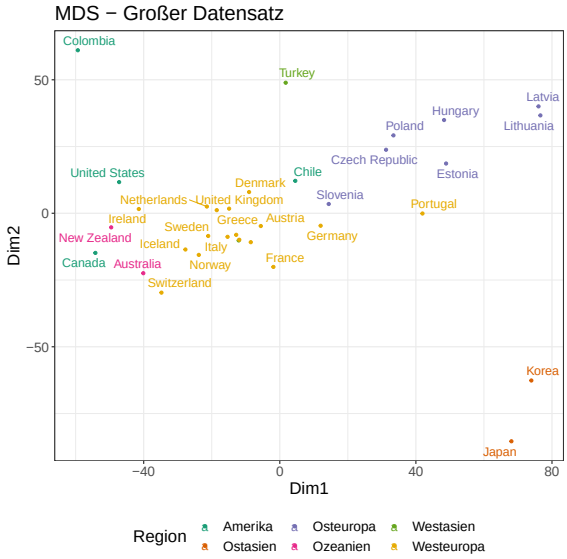

Abbildung 4.6: *Visualisierung der ersten beiden Dimensionen der MDS des großen Datensatzes.*

UMAP

Mit der Methode UMAP kann der hochdimensionale Datensatz auf zwei Dimensionen beschränkt werden. Dies ist in Abbildung 4.7 für den kleinen Datensatz und in Abbildung 4.8 für den großen Datensatz, eingefärbt nach Region, zu sehen. Zu erkennen ist hier, dass sich die Anordnung der Länder deutlich von den vorherigen Methoden unterscheiden. Die Abstände zwischen den Ländern sind hier gleichmäßiger und es scheint keine so eindeutigen Ausreißer wie davor zu geben. Kolumbien, Chile und die Türkei werden bei dieser Methode näher zu den anderen Ländern eingeordnet. In der Abbildung 4.7 sind die Länder *Australien* und *Neuseeland* verglichen mit der Abbildung 4.8 weiter entfernt. Im Allgemeinen wurden für beide Datensätze die Regionen recht nah beieinander gruppiert. Allerdings gilt dies nicht für die Länder aus Westasien und, in Abbildung 4.7, aus Ozeanien.
Das Ergebnis im nieder-dimensionalen Raum hängt dabei sehr stark von der gewählten Anzahl der Nachbarn ab, wie in Kapitel 3.8 erwähnt. Die Abbildung A.6 und die Abbildung A.7 im Anhang zeigen vier Lösungen im zweidimensionalen Raum der Methode UMAP für die Parameter $n\_neighbors = 3, 4, 10, 15$. Hier ist gut zu erkennen, dass für einen sehr kleinen Wert von $n\_neighbors = 3$, die Länder in kleine, weit auseinander stehenden Gruppen geteilt werden. Der Grund hierfür ist, dass sich die Methode UMAP für kleine *n_neighbors* Werte eher auf die lokalen Strukturen konzentriert. Im Gegensatz dazu werden die kleinen Gruppen für immer größer werdende Werte aufgelöst und die Abstände der Länder gleichmäßiger. Da hier die „wahre“ Struktur unbekannt ist, gestaltet sich die Wahl des passenden Parameters als schwierig. Allerdings ist der Beobachtungsumfang im vorliegenden Datensatz mit $n = 34$ Ländern eher gering, warum ein großer Wert für die Anzahl der Nachbarn nicht sinnvoll erscheint. Um dennoch möglichst gut die globale Struktur zu berücksichtigen, wurde ein Parameterwert von fünf gewählt. Für größere Daten kann es durchaus sein, dass ein höherer Wert für den Hyperparameter *n_neighbors* angemessen ist.

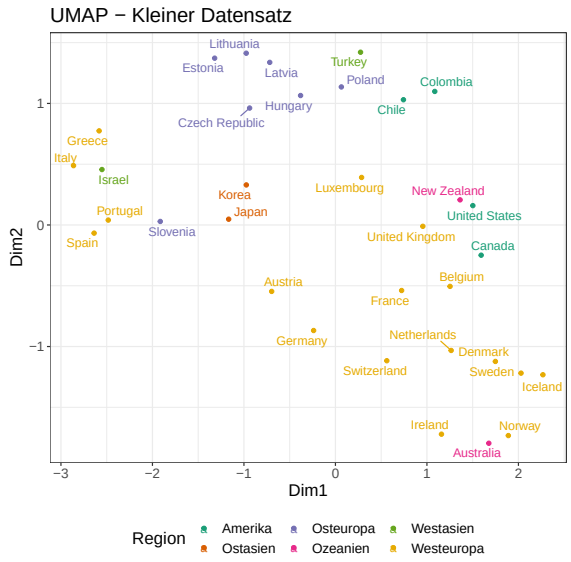

Abbildung 4.7: *Visualisierung der zweidimensionalen Konfiguration von UMAP für den kleinen Datensatz.*

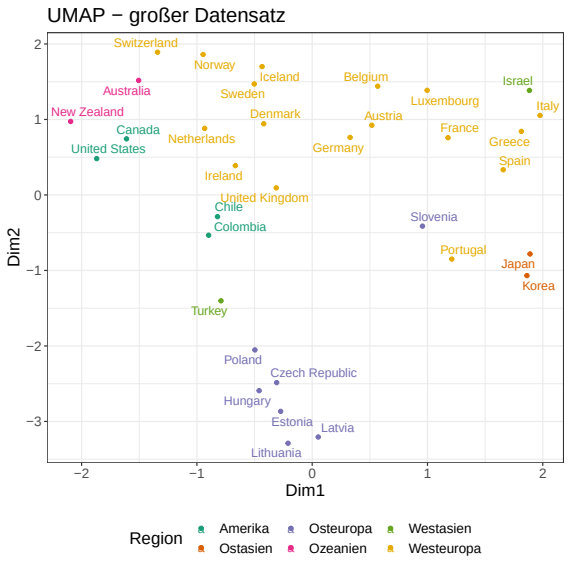

Abbildung 4.8: *Visualisierung der zweidimensionalen Konfiguration von UMAP für den großen Datensatz.*

t-SNE

Die zweidimensionale Darstellung des Datensatzes durch die Methode t-SNE ist in Abbildung 4.9 für den kleinen Datensatz und in Abbildung 4.10 für den großen Datensatz wieder eingefärbt nach den Regionen zu sehen. Für den kleinen Datensatz ähnelt die Struktur der von UMAP. Beim großen Datensatz werden die Länder Japan und Korea erstmals nicht nah zueinander gruppiert und beide Länder sind weit getrennt zu den anderen Ländern dargestellt. Die restlichen Länder sind fast schlauchartig, parallel zur zweiten Dimension gezeigt. Es scheint als wären die restlichen Länder auf der ersten Dimension recht nahe beieinander, allerdings ist die Spanne der Achsen in Abbildung 4.10 sehr viel größer als in Abbildung 4.9. In der oberen Hälfte sind dabei eher osteuropäische Länder, in der unteren Hälfte hingegen eher westeuropäische Länder. Auch in dieser Methode wird die Darstellung stark vom Wert der *Perplexität* beeinflusst. Dieser sollte mindestens den Wert fünf haben und kann maximal den Wert elf annehmen, wie in Kapitel 3.9 erläutert. Die Abbildungen A.4 und A.5 zeigen die Länder für Werte von sieben bis zehn für die Perplexität. Unter Berücksichtigung des Beobachtungsumfangs der OECD Daten erscheint hier ein Wert für die Perplexität von fünf sinnvoll, obwohl meist ein hoher Wert bevorzugt wird. Die Abbildungen 4.9 und 4.10 zeigen die Anordnung der Länder für diese Perplexität. Beim großen Datensatz ergibt sich mit einer Perplexität von sieben eine sehr unterschiedliche Darstellung im Vergleich zu anderen Werten der Perplexität. In A.4 werden Japan und Korea sehr weit weg zu den anderen Ländern dargestellt, sodass die anderen Länder zu einem kleinen Haufen von Datenpunkten werden.

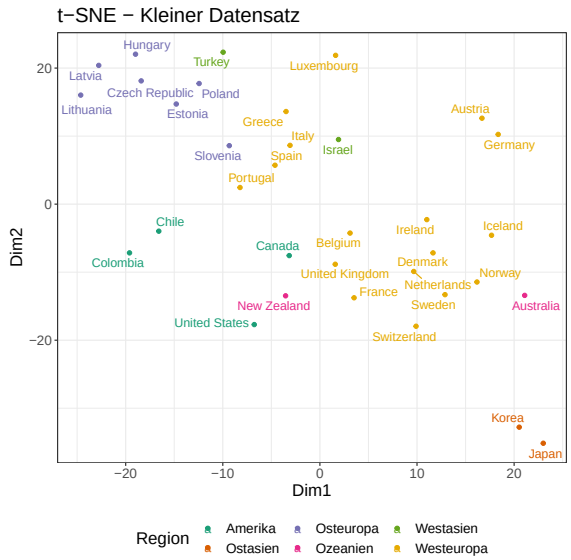

Abbildung 4.9: *Visualisierung der zweidimensionalen Konfiguration von t-SNE für den kleinen Datensatz.*

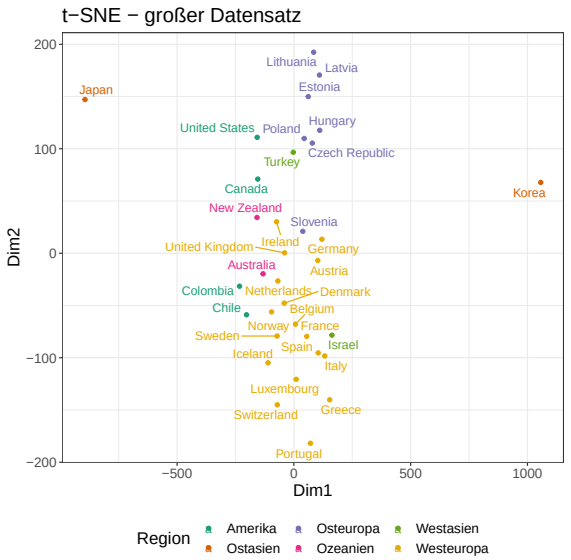

Abbildung 4.10: *Visualisierung der zweidimensionalen Konfiguration von t-SNE für den großen Datensatz.*

4.2 Clusteranalyse

Im folgenden Abschnitt werden die Ergebnisse der Clusteranalyse auf den angeführten Methoden vorgestellt. Bei den Methoden PCA, MFA, und MDS wurde für die Clusteranalyse die Anzahl an Hauptkomponenten genommen, die gemeinsam mindestens achtzig Prozent der Varianz erklären. Dies ergab für den kleinen Datensatz elf Hauptkomponenten bei der PCA und der MFA und neun Hauptkomponenten bei der MDS. Für den großen Datensatz wurden bei der PCA zehn, bei der MFA elf und bei der MDS sieben Dimensionen betrachtet. Für die Methode UMAP ist es möglich mehr als zwei bzw. drei Dimensionen zu analysieren, allerdings gibt es hier kein Kriterium, welche Anzahl an Dimensionen am geeignetsten ist. Deshalb wurde hier das arithmetische Mittel aus den hergenommenen Dimensionen der PCA, der MFA und der MDS gewählt. Dies ergab beim kleinen Datensatz zehn Dimensionen und beim großen Datensatz neun Dimensionen. Die Clusteranalyse auf den Ergebnissen der Methode t-SNE erfolgte dabei auf der zweidimensionalen Lösung für beide Datensätze. Betrachtet wurde das in Kapitel 3.7 vorgestellte k-Means Verfahren und das agglomerative hierarchische Verfahren.

Zur Visualisierung der erzeugten Cluster wurden bei der PCA, der MFA und der MDS die ersten zwei Dimensionen genutzt. Bei UMAP wurden die erzeugten Cluster hingegen auf der zweidimensionalen Lösung gezeigt. Der Grund dafür ist, dass bei dieser Methode nicht eindeutig angenommen werden kann, dass sich die meiste Information auf den ersten beiden Dimensionen widerspiegelt, wie zum Beispiel bei der PCA, sondern über alle Dimensionen verteilt. Bei der Methode t-SNE beruht die Clusteranalyse sowie die Visualisierung der Cluster auf zwei Dimensionen.

Hierarchisches Verfahren

Das (agglomerative) hierarchische Clusterverfahren wurde auf den Ergebnissen der PCA, MFA, MDS, UMAP und t-SNE mit dem Ward-Verfahren angewandt. Die daraus resultierten Dendrogramme sind im Anhang unter dem Kapitel A.5 zu finden. Im Folgenden wird die Übereinstimmung des gesamten Clusterbildungsprozesses für die verschiedenen Methoden vorgestellt. Dafür wurde der paarweise kophenetische Korrelationskoeffizient zwischen den Dendrogrammen der Methoden berechnet, diese sind in der Tabelle 4.1 aufzufinden. Ein Wert nahe an eins steht dabei für eine sehr ähnliche Struktur zwischen zwei Dendrogrammen.

Durch die Tabelle 4.1 wird ersichtlich, dass sich das Dendrogramm der PCA und der MFA am meisten ähneln. Die hierarchische Clusterbildung auf die PCA bzw. MFA Ergebnisse ergeben also sehr ähnlich fusionierte Cluster. Die Dendrogramme der MDS und der MFA korrelieren ebenfalls sehr stark miteinander. Die geringste Übereinstimmung liegt zwischen dem Dendrogramm der PCA und UMAP. Allgemein ist die Korrelation zwischen dem Dendrogramm von UMAP bzw. t-SNE zu einem Dendrogramm einer anderen Methode am geringsten.

	PCA	MFA	MDS	UMAP	t-SNE
PCA	1.000	**0.951**	0.888	**0.397**	0.537
MFA	**0.951**	1.000	0.949	0.432	0.544
MDS	0.888	0.949	1.000	0.416	0.517
UMAP	**0.397**	0.432	0.416	1.000	0.522
t-SNE	0.537	0.544	0.517	0.522	1.000

Tabelle 4.1: *Darstellung des kophenetischen Korrelationskoeffizienten zwischen den Dendrogrammen der hierarchischen Clusterverfahren, angewandt auf die Ergebnisse der Methoden PCA, MFA, MDS, UMAP und t-SNE. Die Durchführung erfolgte hier anhand des kleinen Datensatzes.*

Die Tabelle 4.2 zeigt die Werte für den kophenetischen Korrelationskoeffizienten zwischen den Dendrogrammen des großen Datensatzes. Die größte Ähnlichkeit ist hier deutlich zwischen dem Dendrogramm der PCA und dem der MDS, wohingegen die Korrelation zwischen dem Dendrogramm der PCA und dem der Methode UMAP am geringsten ist. Im Vergleich zu den Ergebnissen des kleinen Datensatzes in Tabelle 4.1 ist zu erkennen, dass nun auf dem großen Datensatz das Dendrogramm der MFA stärker mit dem der Methode UMAP bzw. t-SNE korreliert und deutlich weniger stark mit der Methode PCA bzw. MDS korreliert.

	PCA	MFA	MDS	UMAP	t-SNE
PCA	1.000	0.419	**0.979**	**0.370**	0.427
MFA	0.419	1.000	0.421	0.536	0.614
MDS	**0.979**	0.421	1.000	0.440	0.456
UMAP	**0.370**	0.536	0.440	1.000	0.583
t-SNE	0.427	0.614	0.456	0.583	1.000

Tabelle 4.2: *Darstellung des kophenetischen Korrelationskoeffizienten zwischen den Dendrogrammen der hierarchischen Clusterverfahren, angewandt auf die Ergebnisse der Methoden PCA, MFA, MDS, UMAP und t-SNE. Die Durchführung erfolgte hier anhand des großen Datensatzes.*

k-Means

Die erzeugten Cluster durch das *k*-Means Verfahren wurden anhand drei interner Validierungsindizes bewertet. Im Folgenden wird für jede Methode der ASW, der Dunn-Index und der CH-Index für eine Anzahl an Clustern von $k = 2$ bis $k = 10$ betrachtet. Bei allen drei internen Validierungsindizes ist das Finden eines Maximums erwünscht, wie in Kapitel 3.7.3 erwähnt.
Die Werte können für den ASW zwischen -1 und 1 liegen, wobei höhere Werte für eine homogenere Gruppierung innerhalb der Cluster und eine bessere Trennung zwischen den Gruppen sprechen. Die Abbildung 4.11 zeigt für verschiedene Clusteranzahlen den ASW für die Methoden. Zu erkennen ist hier, dass die optimale Anzahl an Clustern für UMAP und t-SNE bei sieben Clustern liegt, da hierfür der ASW je Methode am größten ist. Für die Clusterergebnisse auf der MFA und der MDS scheinen zwei Cluster und auf der PCA zehn Cluster die optimale Anzahl zu sein. Über alle Clusteranzahlen hinweg haben die Methoden t-SNE und UMAP die höchsten Koeffizienten für den ASW. Der Verlauf des ASWs über die Clusteranzahlen ist bei der PCA, der MFA und der MDS sehr ähnlich, allerdings unterscheiden sich diese vom Verlauf des ASWs bei UMAP und t-SNE.

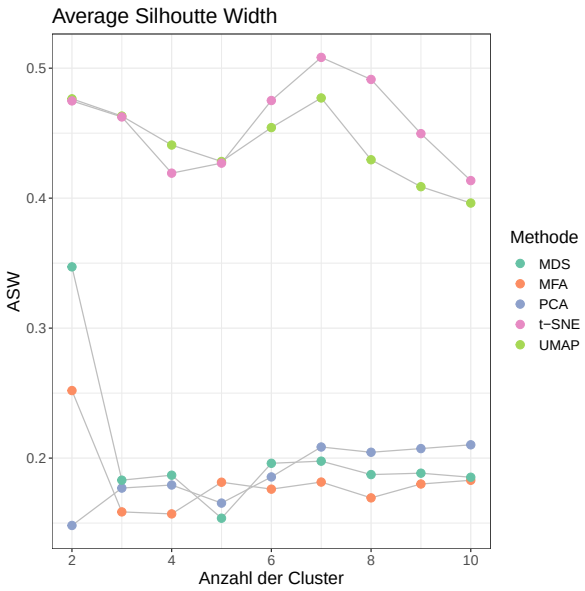

Abbildung 4.11: *Visualisierung des Average Silhoutte Width für jede Methode in Abhängigkeit der Clusteranzahl. Die Durchführung erfolgte hier auf dem kleinen Datensatz.*

Die Abbildung 4.12 zeigt entsprechend den Dunn-Index für jede Methode. Dieser bestimmt das Verhältnis zwischen der Trennung von Clustern und der Kompaktheit innerhalb Gruppen. Der Dunn-Index hat für die PCA, MFA und UMAP ein Maximum bei zehn Cluster. Allerdings sieht es so aus, als würde der Dunn-Index für die Methode UMAP ab sieben Cluster für jedes weitere ansteigen. Für die MDS ergibt sich der größte Dunn-Index bei zwei und für t-SNE bei sieben Clustern. Im Vergleich zum ASW sind die Werte für die Methoden t-SNE und UMAP hier recht unterschiedlich.

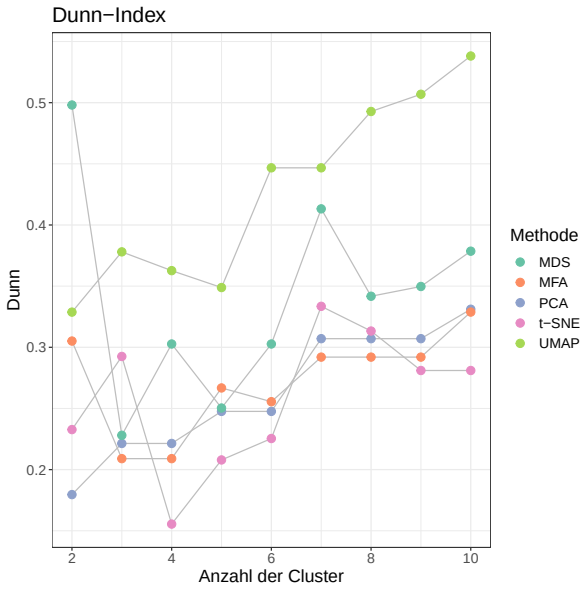

Abbildung 4.12: *Visualisierung des Dunn-Index für jede Methode in Abhängigkeit der Clusteranzahl. Die Durchführung erfolgte hier auf dem kleinen Datensatz.*

Nach dem CH-Index scheint das Verhältnis der Variation innerhalb des Clusters und zwischen

den Clustern für die Ergebnisse der Methode UMAP bei sechs Gruppen am besten zu sein. Für die restlichen Methoden ist der CH-Index bei zehn Clustern am größten. Auch hier unterscheidet sich der Verlauf des CH-Index bei UMAP und t-SNE von den restlichen Methoden, welche alle einen recht gleichmäßigen Indexwert unter zehn über alle Clusteranzahlen hinweg annehmen.

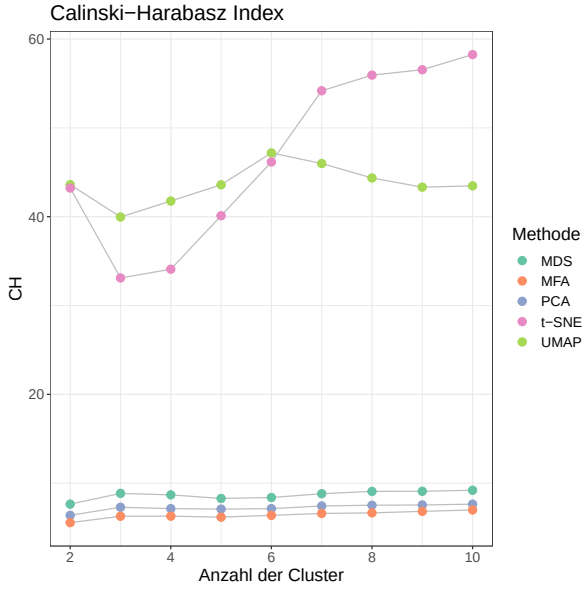

Abbildung 4.13: *Visualisierung des Calinski-Harabasz-Index für jede Methode in Abhängigkeit der Clusteranzahl. Die Durchführung erfolgte hier auf dem kleinen Datensatz.*

Die drei internen Validierungsindizes kommen bei der PCA mit zehn Cluster als optimale Anzahl zum selben Ergebnis. Der Dunn-Index und der ASW stimmen für die Methoden MDS bzw. t-SNE ebenfalls mit zwei bzw. sieben Cluster für die optimale Anzahl überein. Bei den Ergebnissen der MFA sprechen der CH-Index und der Dunn-Index beide für zehn Cluster. Vergleicht man die Indizes bei UMAP, ergibt sich kein Trend für die optimale Anzahl.

Die Abbildung der Validierungsindizes für die Methoden anhand des großen Datensatzes finden sich im Anhang unter dem Kapitel A.6. Die Interpretation erfolgt analog, wie bei den drei Indizes zum kleinen Datensatz.

Vergleich der Clusterverfahren

Auch für das hierarchische Verfahren wurden der ASW, der Dunn-Index und der CH-Index für verschiedene Clusteranzahlen in den Methoden betrachtet. Im Folgenden werden nun die Resultate der drei Indizes des hierarchischen Verfahrens mit denen des k-Means Verfahrens verglichen. Betrachtet man jeden internen Validierungsindex für sich, fällt auf, dass sich die Resultate der zwei Clusterverfahren in Bezug auf die optimale Clusteranzahl im Allgemeinen häufig sehr ähneln. Die Tabellen 4.3 und 4.4 fassen die Anzahl der optimalen Cluster laut den Indizes für jede Methode zusammen. Dabei befinden sich die Ergebnisse des kleinen Datensatzes in der Tabelle 4.3 und des großen Datensatzes in der Tabelle 4.4. Abbildung A.11 zeigt

den Verlauf der Validierungsindizes, links für das hierarchische Verfahren und rechts für das *k*-Means Verfahren, für den kleinen Datensatz und Abbildung A.12 für den großen Datensatz.

In der Tabelle 4.3 ist zu erkennen, dass der CH-Index für das hierarchischen Verfahren und für das *k*-Means Verfahren zu demselben Wert für die optimale Clusteranzahl kommt, betrachtet man jede Methode. Diese Tabelle und die Abbildung A.11 zeigen, dass auch der ASW der zwei Clusterverfahren für eine Methode recht ähnlich ist. In der Spalte für den ASW werden für die zwei Clusterverfahren, nur bei der Methode PCA und UMAP unterschiedliche Werte für die optimale Clusteranzahl vorgeschlagen. Am stärksten variieren die Verläufe des Dunn-Indexes zwischen den zwei Clusterverfahren, wie in Abbildung A.11 zu sehen ist. Die Verläufe aller drei Indizes in Abbildung A.11 sind für die MDS sehr ähnlich.
Im Vergleich dazu zeigt die Abbildung A.12 die Verläufe der drei Indizes für den großen Datensatz. Betrachtet man den Verlauf eines Index, für eine der fünf Methoden, unterscheidet sich dieser nicht stark zwischen dem hierarchischen Verfahren und dem *k*-Means Verfahrens. Beim großen Datensatz unterscheiden sich auch die Verläufe beim Dunn-Index kaum zwischen den zwei Clusterverfahren. Jedoch zeigt die Tabelle 4.4, dass das Maximum eines Index, für eine Methode, nicht immer gleich zwischen den zwei Clusterverfahren ist. Dennoch kommen alle Indizes beim hierarchischen und beim *k*-Means Clustering, für die Ergebnisse der PCA bzw. von UMAP, zu den gleichen Werten für die optimale Anzahl an Cluster.
Grundsätzlich lässt sich für beide Datensätze sagen: Auch wenn sich die optimale Clusteranzahl pro Index innerhalb einer Methode zwischen dem hierarchischen und dem *k*-Means Verfahren unterscheidet, ist der Verlauf dennoch sehr ähnlich. Betrachtet man eine der fünf Methoden, stimmen die Validierungsindizes der zwei Clusterverfahren also meistens überein. Dabei ist anzumerken, dass es durchaus sinnvoll ist, nicht nur eine einzelne Zahl des Indexes, sondern den gesamten Verlauf über mehrere Clusteranzahlen zu betrachten. Häufig unterschied sich das Maximum des Indexes nur gering von den Werten für weniger bzw. mehr Cluster.

	ASW		Dunn-Index		CH-Index	
	HC	*k*-Means	HC	*k*-Means	HC	*k*-Means
PCA	2	10	2	10	10	10
MFA	2	2	2	10	10	10
MDS	2	2	2	2	10	10
UMAP	2	7	10	10	6	6
t-SNE	7	7	9	7	10	10

Tabelle 4.3: *Auflistung der optimalen Clusteranzahlen nach dem ASW, dem Dunn-Index und dem CH-Index für jede Methode, anhand des kleinen Datensatzes.*

	ASW		Dunn-Index		CH-Index	
	HC	*k*-Means	HC	*k*-Means	HC	*k*-Means
PCA	4	4	10	10	2	2
MFA	2	6	2	8	3	10
MDS	4	4	4	10	3	5
UMAP	3	3	10	10	10	10
t-SNE	2	2	3	3	10	10

Tabelle 4.4: *Auflistung der optimalen Clusteranzahlen nach dem ASW, dem Dunn-Index und dem CH-Index für jede Methode, anhand des großen Datensatzes.*

Bei den Methoden, für die die Validierungsindizes beider Clusterverfahren zu derselben optimalen Anzahl an Clustern kamen, ergibt sich folgende Frage: Unterscheidet sich in diesen Fällen, die Zuteilung der Länder in Cluster zwischen dem hierarchischen Verfahren und dem *k*-Means Verfahren?
In Tabelle 4.3 ist zu sehen, dass jeder Index, für die Methode MDS, auf dieselben Werte beim hierarchischen Clustering und beim *k*-Means Verfahren kommt. Tabelle 4.4 zeigt, dass dies beim großen Datensatz für die Methoden PCA, UMAP und t-SNE der Fall ist. Zur Beantwortung der Frage werden in den Abbildungen in Kapitel A.8 die Zuordnungen der Länder zwischen den zwei Clusterverfahren verglichen. Im Allgemeinen ist zuerkennen, dass die Länderzuteilung in Cluster, der zwei Clusterverfahren sehr ähnlich ist. Beispielsweise zeigt Abbildung A.14 die Ländergruppen bei der PCA beim großen Datensatz für zwei, vier und zehn Cluster, also die optimalen Anzahlen bestimmt durch den CH-Index, den ASW und den Dunn-Index aus der Tabelle 4.4. Die linke Seite der Abbildung zeigt die Clusterzuordnung durch das hierarchische Verfahren und rechts durch das *k*-Means Verfahren. Für zwei Cluster ist die Zuordnung der Länder fast identisch, nur die Länder Türkei und Chile werden unterschiedlich zugeordnet. Die Ergebnisse für vier Cluster unterschieden sich ebenfalls nur gering, beim hierarchischen Verfahren bildet Kolumbien ein eigenes Cluster und beim *k*-Means Verfahren wurde das Land mit der Türkei und Chile zusammen gruppiert. Für zehn Cluster wird nur Australien unterschiedlich zugeordnet.
Beim großen Datensatz bei der Methode UMAP ist sowohl für drei Cluster, als auch für zehn Cluster kein Unterschied zwischen den Clusterverfahren. Bei der Methode t-SNE weicht die Zuordnung zwischen den zwei Clusterverfahren nur bei zehn Clustern in den Ländern Türkei und Norwegen ab, für zwei bzw. drei Cluster ist die Zuordnung identisch. Beim kleinen Datensatz für die Methode MDS ist die Zuordnung bei zwei Clustern gleich. Für zehn Cluster hingegen unterscheidet sich die Zuordnung in ein paar Ländern.
Grundsätzlich gleichen sich die Länderzuordnung zwischen den zwei Clusterverfahren, auch wenn kleine Unterschiede existieren.

Die nächste interessante Frage ist, wie sich die Zuteilungen der Länder in die Cluster zwischen

den fünf Methoden unterscheiden. Um dies zu beantworten, wird im nächsten Abschnitt eine mögliche Lösung des *k*-Means Verfahrens zwischen den Methoden grafisch verglichen.

Visualisierung der Clusterzuteilung

Um herauszufinden, ob die Ergebnisse des *k*-Means Verfahrens zu unterschiedlichen Lösungen je Methode führen, wird nun die Einteilung der Länder in die Cluster verglichen. Dies erfolgt beispielhaft anhand des großen Datensatzes.
Eine Möglichkeit wäre, für jede Methode individuell die optimale Anzahl an Cluster anhand der Indizes zu wählen. Somit wäre für den Vergleich die bestmögliche Homogenität, Trennung und Kompaktheit der Cluster in jeder Methode gewährleistet. Allerdings ist es so schwierig, die Unterschiede der Länderzuordnung zwischen den Methoden zu beurteilen.
Eine weitere Möglichkeit wäre bei allen fünf Methoden dieselbe Anzahl an Clustern zu wählen. Betrachtet man die Indizes, ergibt sich allerdings kein allgemeiner Trend in der optimalen Clusteranzahl, der für alle Methoden übereinstimmt. Nach dem ASW liegt die optimale Clusteranzahl beim großen Datensatz zwischen zwei und sechs Clustern. Eine große Anzahl von acht bzw. zehn Clustern werden vom Dunn-Index und vom CH-Index für einige der Methoden als optimal vorgeschlagen. Dennoch wird für einen gerechten Vergleich eine einheitliche Anzahl an Clustern festgelegt. Für die Übersichtlichkeit eignet sich eine niedrigere Anzahl an Cluster besser, weshalb im Folgenden die Einteilung der Länder in vier Cluster anhand des *k*-Means Verfahrens zwischen den fünf Methoden beurteilt wird. Dabei ist anzumerken, dass dies nicht für alle Methoden die optimale Anzahl an Clustern laut den Validierungsindizes ist.

Die Abbildung A.17 im Anhang A zeigt die Einteilung der Länder in die Cluster. Die Clusterzuordnung bei den Ergebnissen für die PCA und die MDS sind identisch und eine recht gute Trennung der Cluster ist zu erkennen. Die Länder Japan und Korea bilden dabei ein Cluster und Kolumbien, Türkei und Chile ein weiteres. Das vierte Cluster umfasst fast alle osteuropäischen Länder. Im dritten und größten Cluster befinden sich die restlichen Länder.
Die Darstellung der Ergebnisse für die MFA zeigt, dass hierbei die Länder Chile und Kolumbien in ein Cluster ohne die Türkei geordnet werden. Die Länder Japan und Korea bilden dabei auch hier ein einzelnes Cluster. Das dritte Cluster besteht hauptsächlich aus westeuropäischen Ländern, sowie Australien, Neuseeland, die USA und Kanada. Das westeuropäische Land Griechenland wird dabei zum vierten Cluster zugeordnet. Das vierte Cluster umfasst zudem alle osteuropäischen Länder, die Türkei und Israel.
Die Einteilung der Länder in die Cluster bei UMAP und t-SNE unterscheiden sich sehr von den anderen Methoden. Bei UMAP sind die Clustergrößen gleichmäßiger verteilt, als bei den Methoden davor. Die Länder Japan, Korea, Portugal, Slowenien und Deutschland bilden ein Cluster. Die restlichen Cluster weisen ebenfalls Länder unterschiedlichster Regionen auf. Das erste Cluster umfasst allerdings hauptsächlich osteuropäische Länder. Dabei ist anzumerken, dass bei der Methoden UMAP sowohl der Dunn-Index als auch der CH-Index ein Maximum bei zehn Clustern aufwies.

Die Clustereinteilung bei t-SNE zeigt, dass Japan und Korea jeweils ein Cluster bilden. Das dritte Cluster umfasst Länder aus Ozeanien, Amerika und Westeuropa. Die Türkei und Israel, sowie alle osteuropäische und weitere westeuropäische Länder bilden das vierte Cluster.
Bei allen Methoden befinden sich die Länder aus Osteuropa, bis auf Slowenien, innerhalb eines Clusters. Die kleineren Regionen mit nur jeweils zwei Ländern, wie Ozeanien, Ostasien und Westasien, werden teilweise getrennten Clustern zugeordnet. Die Clusterzuordnung zeigt, dass eine Differenzierung zwischen den Regionen Süd- und Nordamerika sinnvoll wäre, da Kolumbien und Chile häufig nicht zu einem Cluster mit den USA und Kanada gehören.

Kapitel 5

Fazit

Diese Arbeit vergleicht fünf verschiedene Methoden zur Dimensionsreduktion und deren Ergebnisse in Kombination mit zwei Clusterverfahren. Betrachtet wurde dabei die Hauptkomponentenanalyse, die Multiple Faktorenanalyse, die Multidimensionale Skalierung, sowie zwei eher neuere Verfahren: Uniform Manifold Approximation and Projection und t-Distributed Stochastic Neighbor Embedding.
Die Ergebnisse von UMAP und t-SNE hängen sehr stark von der getroffenen Hyperparameterwahl ab. Aus diesem Grund ist es wichtig, für jeden Datensatz individuell verschiedene Hyperparameter zu betrachten und daraufhin eine sinnvolle Wahl zu treffen. Dies bedeutet allerdings, dass die Entscheidung auf einer subjektiven Beurteilung beruht.
Die Darstellung der Länder auf den ersten beiden Dimensionen zeigt, dass sich die Anordnung der Länder durch die herkömmlichen Methoden PCA, MFA und MDS nicht groß unterscheiden, was für ein stabiles Ergebnis spricht. Ein Unterschied zwischen allen Methoden ist, dass die PCA, die MFA sowie die MDS deterministische Algorithmen sind. Mehrere Durchgänge an den gleichen Daten führen immer zu denselben Ergebnissen. Bei UMAP und t-SNE können, ohne einen gesetzten *Seed*, abweichende Lösungen entstehen.

Mit der PCA und der MFA ist es sehr einfach den Beitrag der Variablen zu den Dimensionen zu bestimmen, was sich für die Interpretation der Komponenten eignet. Mit der MFA ist es zudem möglich, Variablen zu Gruppen zusammenzufassen und deren Einfluss auf die Dimensionen zu gewichten. Beispielsweise wurden beim großen Datensatz die ersten beiden Dimensionen der PCA durch die Variationen der Variablen *Gesundheitszustand* und *Lebenserwartung* dominiert. Mit der Gewichtung der Variablen bei der MFA wird dies verhindert. Ein Vorteil von UMAP und t-SNE zu den anderen Methoden ist, dass diese auch für nicht-lineare Mannigfaltigkeit geeignet sind. Ein Nachteil hingegen kristallisiert sich vor allem bei der Anwendung von t-SNE in Kombination mit Clusterverfahren heraus. Oskolkov (2020) betont, dass die Abstände zwischen Clustern im hochdimensionalen Raum sowohl bei t-SNE als auch bei UMAP im nieder dimensionalen Raum nicht korrekt beibehalten werden können. Oskolkov (2019b) kommt zu

dem Entschluss, dass, auch wenn sich die Qualität der zweidimensionalen Darstellung zwischen t-SNE und UMAP nicht groß unterscheidet, die Anwendung der Methoden doch sehr unterschiedlich ist. Laut Oskolkov (2019b) dient t-SNE eher zu Visualisierungszwecken, wohingegen die Ergebnisse von UMAP durchaus für die Clusteranalyse geeignet sind, da UMAP mehr von der globalen Struktur beibehält. Auch Schubert et al. (2017, S. 192) rät ab, die Ergebnisse der Methode t-SNE für die Untersuchungen von Clustern anzuwenden, da das Verfahren Entfernung oder Dichten möglicherweise nicht ausreichend erhält. Mehr Informationen und einen Vergleich zwischen Unterschieden beim *k*-Means Verfahren angewandt auf den Ergebnissen der PCA, t-SNE und UMAP auf einem größeren Datensatz bietet der Artikel von Oskolkov (2020).
Bei der Anwendung von t-SNE können also leicht falsche Schlüsse gezogen werden, entweder wegen einer nicht passenden Hyperparameterwahl oder durch eine Missinterpretation der Methode in Kombination von Clusterverfahren. Gute Beispiele für verschiedene Szenarien bei t-SNE und deren Ergebnisse gibt der Online-Artikel von Wattenberg et al. (2016).

Um mögliche Clusterstrukturen in den Ländern zu untersuchen, wurden die Ergebnisse im niedrig-dimensionalen Raum genutzt und zwei weitverbreitete Clusteralgorithmen durchgeführt. Der Vorteil beim *k*-Means Verfahren ist, dass die Länder während des Fusionierungsprozesses zwischen den Gruppen getauscht werden können. Die Clusterbildung durch das *k*-Means Verfahren gestaltet sich also recht flexibel. Beim hierarchischen Clusterverfahren hingegen ist das nicht möglich, allerdings wird keine vorher festgelegte Anzahl an Clustern benötigt.
Die Beurteilung der Ergebnisse der Clusterverfahren erfolgte anhand ausgewählter interner Validierungsindizes. Der Vorteil der Indizes besteht darin, dass sie keine weiteren externen Informationen für die Beurteilung der Clustereinteilung benötigen, sondern die Homogenität, Kompaktheit und Trennung der Cluster betrachten. Beim kleinen Datensatz kamen der Average Silhouette Width und der Calinski-Harabasz-Index, vor allem für die PCA, die MFA und die MDS zu sehr ähnlichen Verläufen. Beim großen Datensatz hingegen verlief der Dunn-Index und der Calinski-Harabasz-Index zwischen den drei Methoden ähnlich. Die Verläufe der internen Validierungsindizes für das hierarchischen Verfahren unterscheiden sich nicht groß von dem für das *k*-Means Verfahren in beiden Datensätzen. Das heißt, die zwei Clusterverfahren teilen die Länder ähnlichen Clustern zu, angewandt auf dieselbe Methode. Die Unterschiede bestehen hauptsächlich in der vorher durchgeführten Methode zur Dimensionsreduktion.

Die Ergebnisse der Arbeit zeigen, dass es durchaus sinnvoll ist, für die Analyse dimensionsreduzierter Daten zuerst mehrere Methoden anzuwenden und die Ergebnisse visuell zu betrachten, um einen Überblick über die Struktur der Daten zu bekommen. Da die „wahre“ Zuteilung der Cluster oft nicht bekannt ist, kann es auch hier hilfreich sein, Ergebnisse aus verschiedene Verfahren zu vergleichen.
Außerdem ist es möglich, die PCA mit t-SNE oder UMAP zu kombinieren. Dabei wird zuerst die Dimension linear mit der PCA auf die wichtigsten Hauptkomponenten reduziert, wodurch das *Rauschen* in den Daten verringert wird. Anschließen kann t-SNE oder UMAP genutzt wer-

den, um beispielsweise eine zweidimensionale Konfiguration zu betrachten.

Zu erwähnen ist, dass der vorliegende Datensatz sehr unterschiedliche Variablen zu verschiedenen Themen umfasst. Diese sind teils gemessene/erhobene Werte, aber auch erfragte Werte, wie zum Beispiel der wahrgenommene Gesundheitszustand. Die Variablen decken also sehr verschiedene Bereiche ab, die einen Einfluss auf die Gesundheit haben können. Das könnte ein Grund für die Schwierigkeit sein, die gesamte Varianz durch nur wenige Dimensionen auszudrücken. Bei den Methoden PCA, MFA und MDS hat sich gezeigt, dass zwischen sieben und elf Komponenten benötigt werden, um mindestens achtzig Prozent der Varianz zu erklären. Möglicherweise könnte dagegen die extreme Reduzierung auf nur zwei Dimensionen bei der Methode t-SNE zu einem größeren Informationsverlust als bei den anderen Methoden führen. Eine weitere Herausforderung ist, dass die Anwendung der Methoden auf teilweise imputierten Daten basiert. Der Datensatz wies zu Beginn der Datenaufbereitung fehlende Werte auf, wodurch zum einen Variablen und Länder gar nicht erst betrachtet wurden und zum anderen die fehlenden Angaben geschätzt werden mussten. Auch wenn das MICE-Verfahren eine gute Möglichkeit bietet Werte zu imputieren, ist es dennoch wichtig zu beachtet, dass es teilweise keine echten Angaben, sondern Schätzungen sind.

Es gibt sehr viele verschiedene Möglichkeiten, die für den Vergleich multivariater Methoden in Betracht gezogen werden können. In dieser Arbeit wurden vor allem die Konfiguration im zweidimensionalen Raum, die erklärte Varianz und die Interpretation der Dimensionen anhand der OECD Daten betrachtet. Möglich wäre beispielsweise auch, die Methoden anhand simulierter Daten zu vergleichen. Hierbei könnten auch gefundene Cluster mit einer wahren Clusterstruktur bewertet werden. Ebenso gibt es viele Verfahren, eine Clusteranalyse durchzuführen und die Ergebnisse zu beurteilen. Für die Bewertung der Clusterlösung könnten zusätzlich externe Validierungsindizes genutzt werden, um jeweils zwei Clusterlösungen miteinander zu vergleichen. Neben der Betrachtung einzelner Validierungsindizes, kann die Beurteilung der Clusterlösungen auch anhand eines zusammengesetzten Index gemessen werden. In dem Artikel von Akhanli et al., 2020 wird die Aggregation von verschiedenen Indizes vorgeschlagen, um verschiedene Merkmale einer Clusterbildung zu berücksichtigen. Für die Interpretierbarkeit des zusammengesetzten Index, muss vorher eine Kalibrierung anhand zufälligen Clusterings durchgeführt werden. Akhanli et al. (ebd.) betonen, dass die Auswahl der Indizes wichtig ist und Hintergrundinformationen genutzt werden können, um den zusammengesetzten Index zu definieren. Die Abstände zwischen den Clustern beim hierarchischen Verfahren wurden mithilfe der Methode nach Ward bestimmt. Weitere Möglichkeiten wären beispielsweise das Zentroid-Verfahren, das Weighted-Average-Linkage oder das Complete-Linkage. Ebenfalls interessant wäre alternativ, neben der Betrachtung von entfernungbasierten Clusterverfahren, die Verwendung von dichtebasierten Cluster-Algorithmen. Das Verfahren *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) wurde von Ester et al. (1996) eingeführt. Der Vorteil liegt hierbei, dass Cluster beliebiger Form entdeckt werden, während Rauschpunkte separiert bleiben. Auch die Erweiterung *Density-Based Clustering Based on Hierarchical Density Estimates* (HDBS-

CAN) von Campello et al. (2013) besitzt den Vorteil dichtebasierter Cluster-Algorithmen.
Wie am Anfang der Arbeit angesprochen, ist der Fluch der Dimensionen häufig ein Problem bei statistischen Analysen. Auch die Ermittlung von Ausreißern gestaltet sich im hochdimensionalen Raum oft schwierig. Eine weitere Möglichkeit für den Vergleich der Methoden auf den Gesundheitsdaten wäre eine weiterführende Ausreißeranalyse. Beispielsweise vergleicht Schubert et al. (2017) verschiedene Methoden, unter anderem auch die PCA, die MDS und t-SNE bezüglich der Erkennung von Ausreißern.

Literaturverzeichnis

Abdi, Hervé, Lynne J. Williams und Domininique Valentin (2013). „Multiple factor analysis: principal component analysis for multitable and multiblock data sets“. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.2, S. 149–179. ISSN: 19395108. DOI: 10.1002/wics.1246.

Akhanli, Serhat Emre und Christian Hennig (2020). „Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes“. In: *Statistics and Computing* 30.5, S. 1523–1544. DOI: 10.1007/s11222-020-09958-2.

Altman, Naomi und Martin Krzywinski (2018). „The curse(s) of dimensionality“. In: *Nature Methods* 15.6, S. 399–400. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0019-x.

Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis und Philip J. Leaf (2011). „Multiple imputation by chained equations: what is it and how does it work?“ In: *International journal of methods in psychiatric research* 20.1, S. 40–49. DOI: 10.1002/mpr.329.

Bacher, Johann (2008). *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. 3rd edition. München: Oldenbourg Verlag. ISBN: 978-3-486-58457-8.

Backhaus, Klaus, Bernd Erichson, Sonja Gensler, Rolf Weiber und Thomas Weiber (2021). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN: 978-3-658-32424-7. DOI: 10.1007/978-3-658-32425-4.

Backhaus, Klaus, Bernd Erichson und Rolf Weiber (2015). *Fortgeschrittene Multivariate Analysemethoden*. 3rd edition. Berlin, Heidelberg: Springer Gabler. ISBN: 978-3-662-46086-3. DOI: 10.1007/978-3-662-46087-0.

Campello, Ricardo J. G. B., Davoud Moulavi und Joerg Sander (2013). „Density-Based Clustering Based on Hierarchical Density Estimates“. In: *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, S. 160–172. DOI: https://doi.org/10.1007/978-3-642-37456-2_14.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander und Xiaowei Xu (1996). „A density-based algorithm for discovering clusters in large spatial databases with noise.“ In: *kdd*. Bd. 96. 34. AAAI Press, S. 226–231. URL: https://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf.

Everitt, Brian und Torsten Hothorn (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.

Fahrmeir, Ludwig, Alfred Hamerle und Gerhard Tutz, Hrsg. (2015). *Multivariate statistische Verfahren*. Berlin, Bosten: de Gruyter. ISBN: 9783110816020. DOI: 10.1515/9783110816020.

Graham, John W. (2012). *Missing Data: Analysis and Design*. New York, NY: Springer New York. ISBN: 978-1-4614-4017-8. DOI: 10.1007/978-1-4614-4018-5.

Graham, John W., Allison E. Olchowski und Tamika D. Gilreath (2007). „How many imputations are really needed? Some practical clarifications of multiple imputation theory“. In: *Prevention science : the official journal of the Society for Prevention Research* 8.3, S. 206–213. ISSN: 1389-4986. DOI: 10.1007/s11121-007-0070-9.

Hennig, Christian, Marina Meila, Fionn Murtagh und Roberto Rocci (2015). *Handbook of cluster analysis*. Chapman und Hall/CRC. ISBN: 9780429185472. DOI: https://doi-org.emedien.ub.uni-muenchen.de/10.1201/b19706.

Husson, François, Julie Josse und Jérôme Pagès (2010). „Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?“ In: *Applied mathematics Department*.

Husson, François, Sébastien Lê und Jérôme Pagès (2017). *Exploratory multivariate analysis by example using R*. Second edition. Computer science and data analysis series. Boca Raton, London und New York: CRC Press. ISBN: 9781138196346.

Jolliffe, Ian T. und Jorge Cadima (2016). „Principal component analysis: a review and recent developments“. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, S. 20150202. DOI: 10.1098/rsta.2015.0202.

Kassambara, Alboukadel (2017). „Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning“. In: 1.

Martinez, Wendy L., Angel R. Martinez und Jeffrey L. Solka (2017). *Exploratory Data Analysis with MATLAB: 3rd Edition*. 3rd Edition. New York: Chapman and Hall/CRC. ISBN: 9781315366968. DOI: https://doi.org/10.1201/9781315366968.

McInnes, Leland (2022). „umap Documentation“. In: URL: https://umap-learn.readthedocs.io/_/downloads/en/latest/pdf/.

McInnes, Leland, John Healy und James Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. URL: http://arxiv.org/pdf/1802.03426v3.

OECD.Data.Health (o. D.). https://data.oecd.org/health.htm. zuletzt eingesehen am 20.02.2022. URL: https://data.oecd.org/health.htm.

OECD.org (o. D.). https://www.oecd.org. zuletzt eingesehen am 15.02.2022. URL: https://www.oecd.org.

Oskolkov, Nikolay (19. Juli 2019a). *How to tune hyperparameters of tSNE*. https://towardsdatascienc com/how-to-tune-hyperparameters-of-tsne-7c0596a18868. zuletzt eingesehen am 22.03.2022.

Oskolkov, Nikolay (31. Dez. 2019b). *Why UMAP is Superior over tSNE*. https://towardsdatascience. com/why-umap-is-superior-over-tsne-faa039c28e99. zuletzt eingesehen am 02.04.2022.

Oskolkov, Nikolay (4. März 2020). *tSNE vs. UMAP: Global Structure*. https://towardsdatascience. com/tsne-vs-umap-global-structure-4d8045acba17. zuletzt eingesehen am 02.04.2022.

Pagès, Jérôme (2004). „Multiple Factor Analysis: Main Features and Application to Sensory Data“. In: *Revista Colombiana de Estadistica* 27.1, S. 1–26.

Pagès, Jérôme und Brigitte Escofier (1988–1998). *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation*. Dunod.

Pearson, Karl (1901). „LIII. On lines and planes of closest fit to systems of points in space“. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, S. 559–572. DOI: 10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

Raghunathan, Trivellore E., Peter W. Solenberger und John van Hoewyk (2002). „IVEware: Imputation and Variance Estimation Software User Guide“. In: *Ann Arbor, MI: Institute for Social Research, University of Michigan*.

Saraçli, Sinan, Nurhan Doğan und İsmet Doğan (2013). „Comparison of hierarchical cluster analysis methods by cophenetic correlation“. In: *Journal of Inequalities and Applications* 2013.1. DOI: 10.1186/1029-242X-2013-203.

Schafer, Joseph L. und John W. Graham (2002). „Missing data: Our view of the state of the art“. In: *Psychological Methods* 7.2, S. 147–177. ISSN: 1082-989X. DOI: 10.1037//1082-989X.7.2.147.

Schubert, Erich und Michael Gertz (2017). „Intrinsic t-Stochastic Neighbor Embedding for Visualization and Outlier Detection“. In: *International Conference on Similarity Search and Applications*. Springer International Publishing, S. 188–203.

Sokal, Robert R. und F. James Rohlf (1962). „The Comparison of Dendrograms by Objective Methods“. In: *Taxon* 11.2, S. 33–40. ISSN: 00400262. URL: http://www.jstor.org/stable/1217208.

Torgerson, Warren S. (1952). „Multidimensional scaling: I. Theory and method“. In: *Psychometrika* 17.4, S. 401–419. URL: https://doi.org/10.1007/BF02288916.

van Buuren, Stef und Karin Groothuis-Oudshoorn (2011). „mice : Multivariate Imputation by Chained Equations in R“. In: *Journal of Statistical Software* 45.3. DOI: 10.18637/jss.v045.i03.

van der Maaten, Laurens (2014). „Accelerating t-SNE using Tree-Based Algorithms“. In: *Journal of Machine Learning Research* 15, S. 3221–3245.

van der Maaten, Laurens und Geoffrey Hinton (2008). „Visualizing Data using t-SNE“. In: *Journal of Machine Learning Research* 9.86, S. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Vidal, René, Yi Ma und S. S. Sastry (2016). *Generalized Principal Component Analysis*. Bd. 40. New York, NY: Springer New York. ISBN: 978-0-387-87810-2. DOI: 10.1007/978-0-387-87811-9.

Wang, Yingfan, Haiyang Huang, Cynthia Rudin und Yaron Shaposhnik (2021). „Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization“. In: *Journal of Machine Learning Research* 22.201, S. 1–73. URL: http://jmlr.org/papers/v22/20-1061.html.

Wattenberg, Martin, Fernanda Viégas und Ian Johnson (2016). „How to Use t-SNE Effectively“. In: *Distill*. zuletzt eingesehen am 02.04.2022. DOI: 10.23915/distill.00002.

Williams, Christopher K.I. (2002). „On a Connection between Kernel PCA and Metric Multidimensional Scaling“. In: *Machine Learning* 46.1, S. 11–19. ISSN: 1573-0565. DOI: 10.1023/A:1012485807823.

Anhang A

Ergebnisse

A.1 PCA

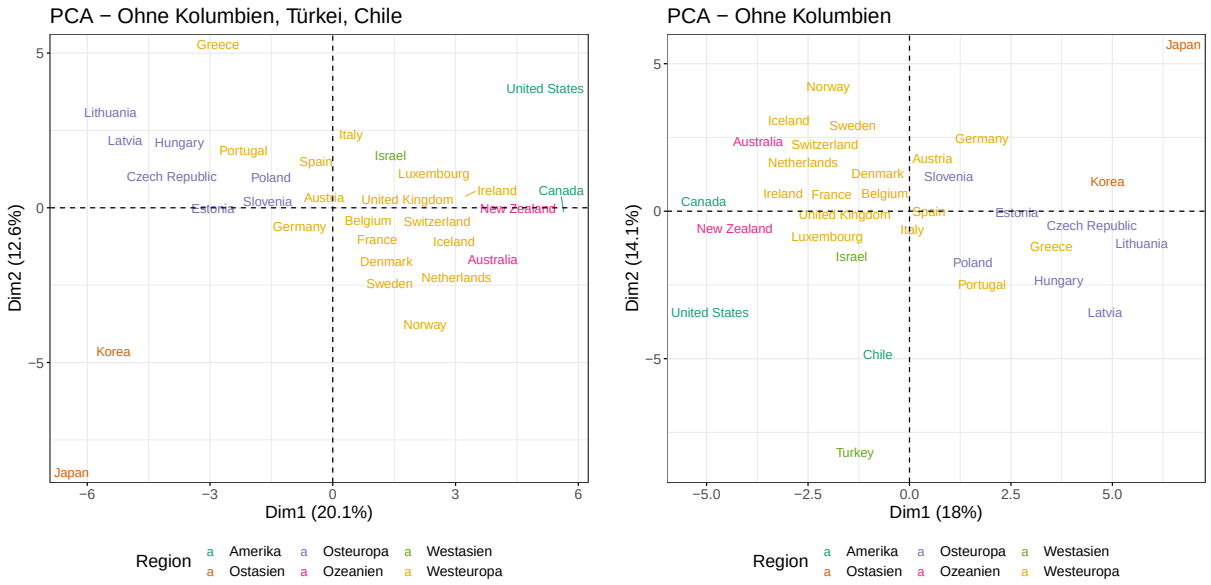

Abbildung A.1: *Visualisierung der ersten beiden Hauptkomponenten der PCA des kleinen Datensatzes. Links sind die Ergebnisse ohne die Länder Kolumbien, Türkei und Chile zu sehen und rechts nur ohne Kolumbien.*

	Hauptkomponenten										
	1	2	3	4	5	6	7	8	9	10	11
Std.abweichung	3.167	2.670	2.248	2.022	1.699	1.594	1.548	1.349	1.314	1.216	1.189
erk. Varianz	0.197	0.140	0.099	0.080	0.057	0.050	0.047	0.036	0.034	0.029	0.028
kum. erk. Varianz	0.197	0.336	0.435	0.516	0.572	0.622	0.669	0.705	0.739	0.768	0.795

Tabelle A.1: Darstellung der Eigenwerte und der erklärten Varianz für die ersten elf Hauptkomponenten der PCA auf der gemeinsamen Korrelationsmatrix. Die Korrelationsmatrix ergab sich aus fünfzig imputierten Datensätzen.

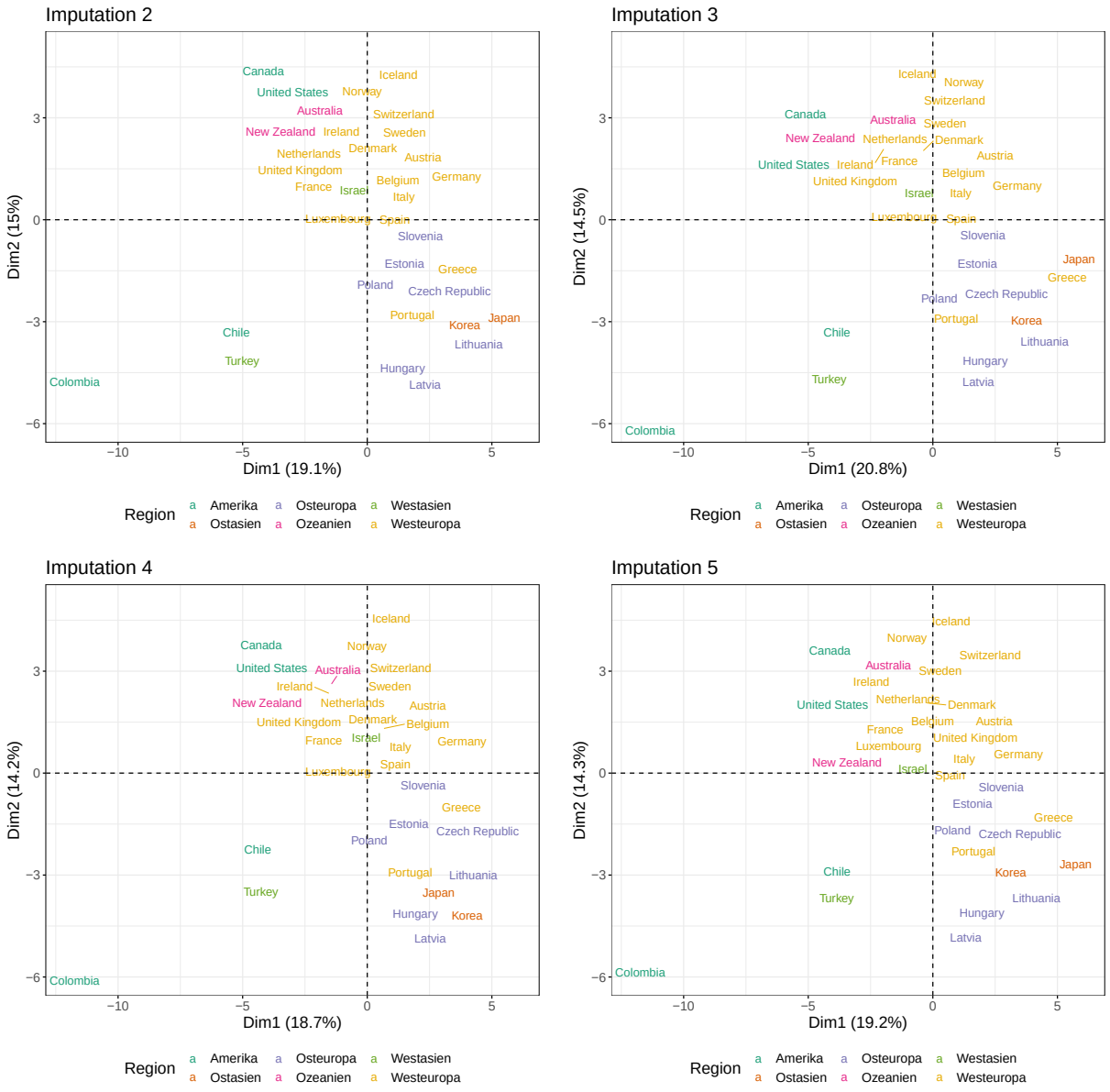

Abbildung A.2: *Visualisierung der zweidimensionalen Lösung der PCA für die vier weiteren imputierten Datensätze. Im Allgemeinen ähneln sich die Ergebnisse und nur geringe Unterschiede sind zu erkennen.*

A.2 MFA

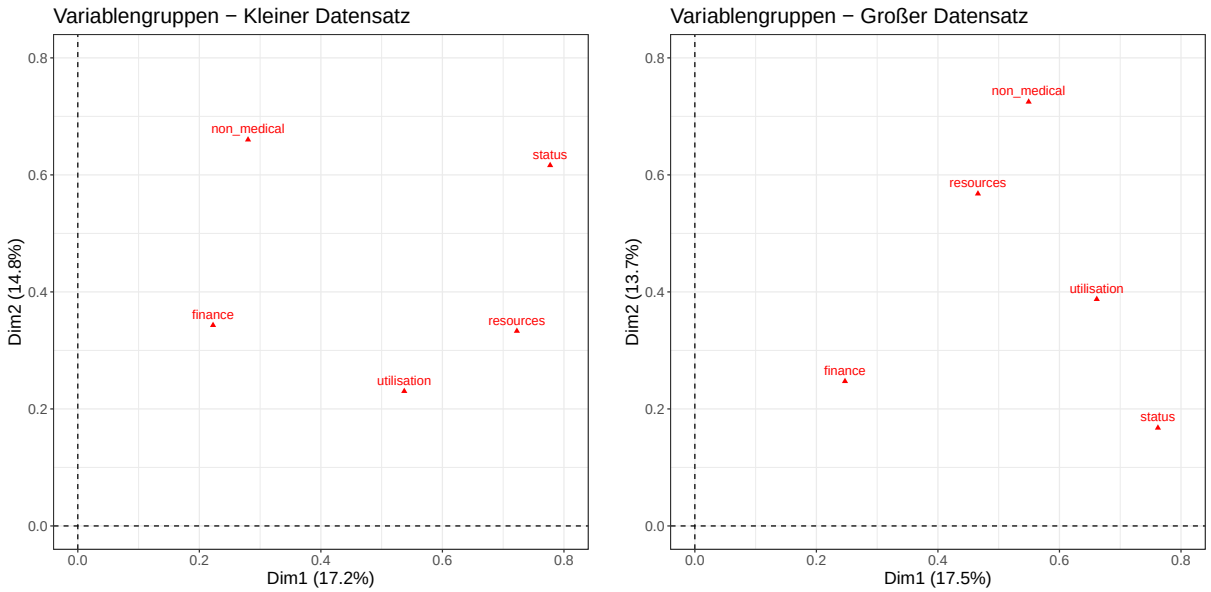

Abbildung A.3: *Visualisierung der Beiträge der Variablengruppen zu den ersten beiden Hauptkomponenten der MFA, links auf dem kleinen Datensatz und rechts auf dem großen Datensatz.*

A.3 t-SNE Hyperparameter

t-SNE Parameterwahl perplexity - Großer Datensatz

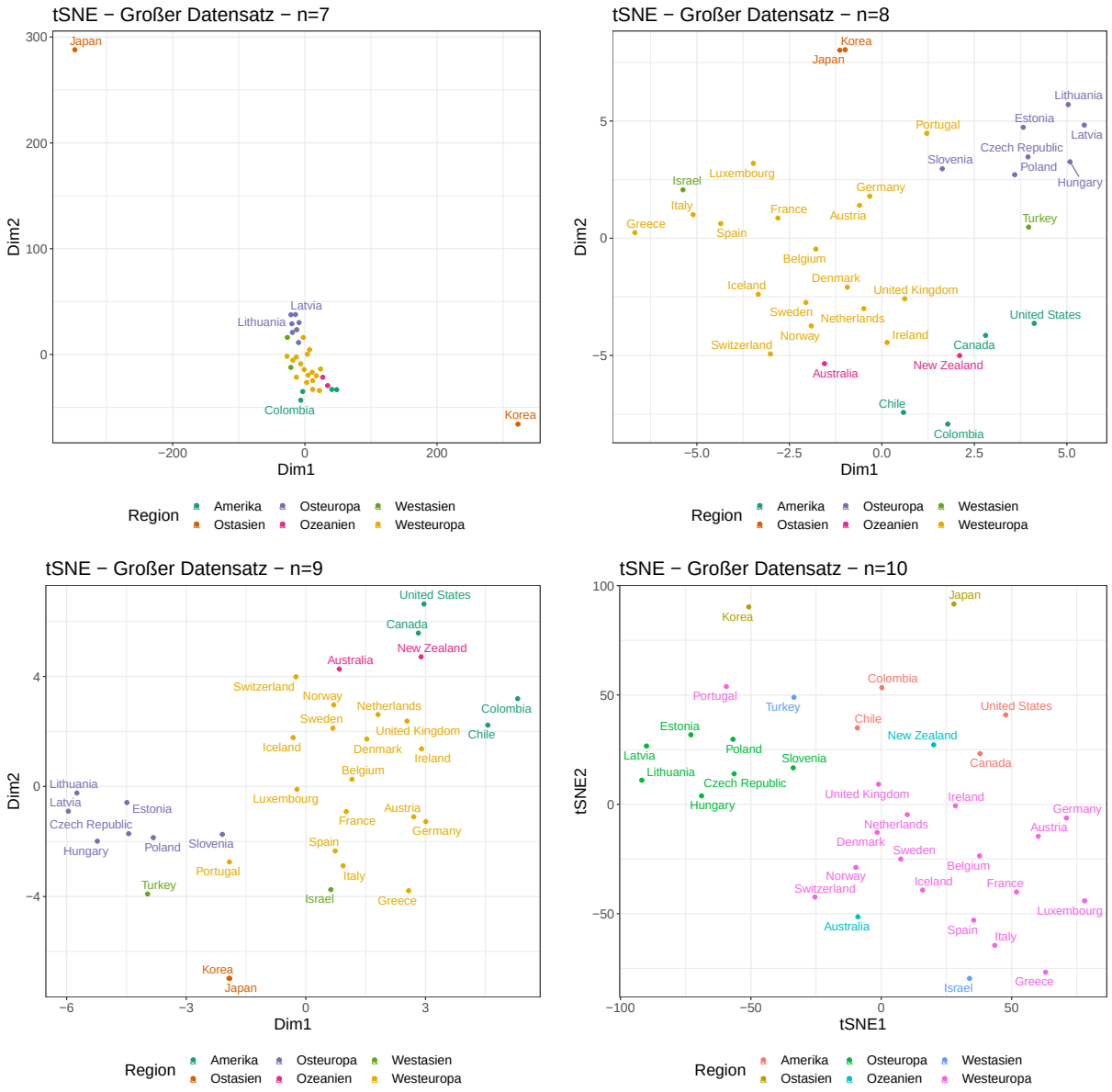

Abbildung A.4: *Visualisierung der zweidimensionalen Konfiguration von t-SNE für den großen Datensatz, für verschiedene Perplexitätswerte mit n = 7,8,9,10.*

t-SNE Parameterwahl perplexity - Kleiner Datensatz

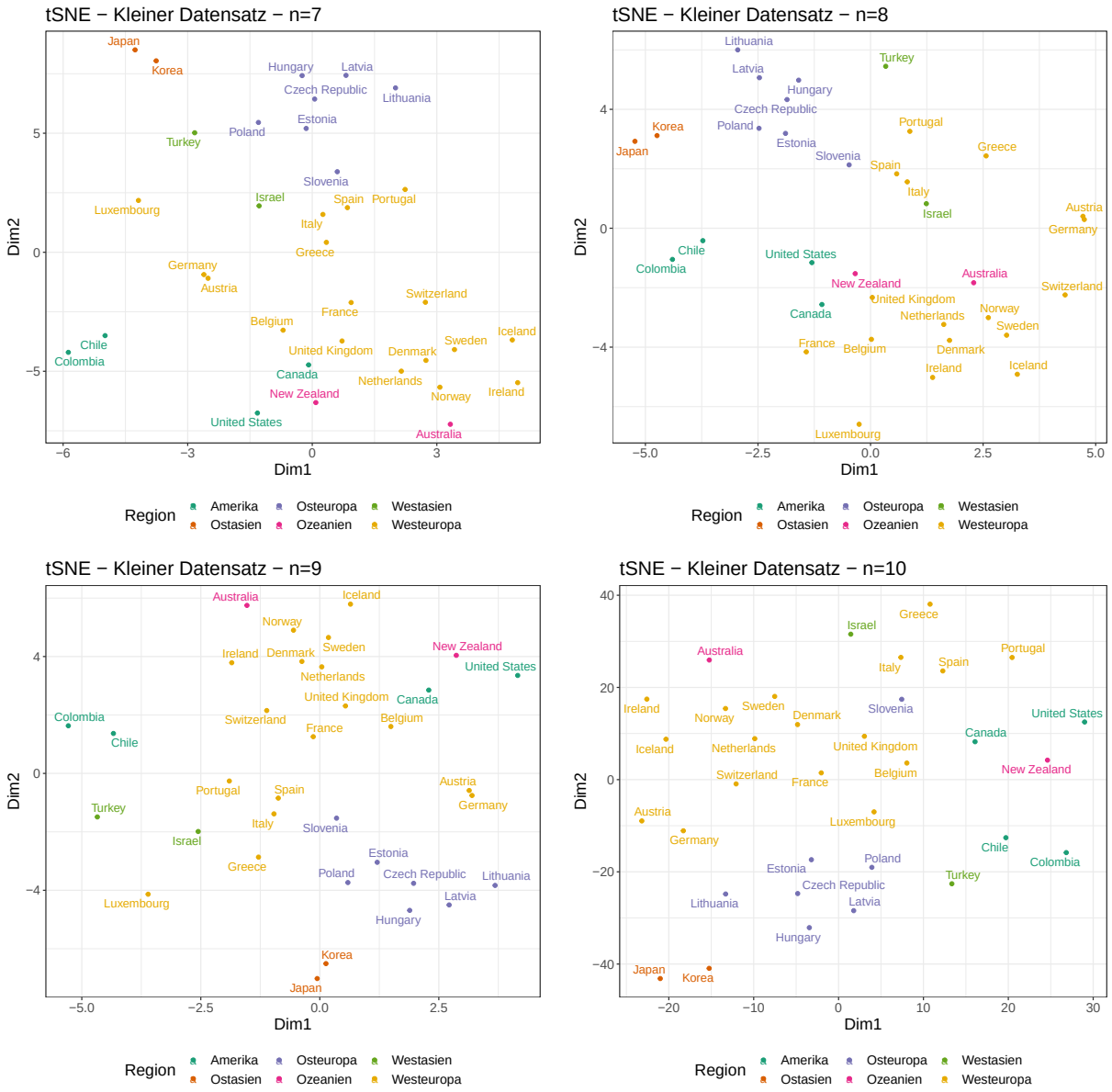

Abbildung A.5: *Visualisierung der zweidimensionalen Konfiguration von t-SNE für den kleinen Datensatz, für verschiedene Perplexitätswerte mit n = 7,8,9,10.*

A.4 UMAP Hyperparameter

UMAP Parameterwahl n_neighbors - Kleiner Datensatz

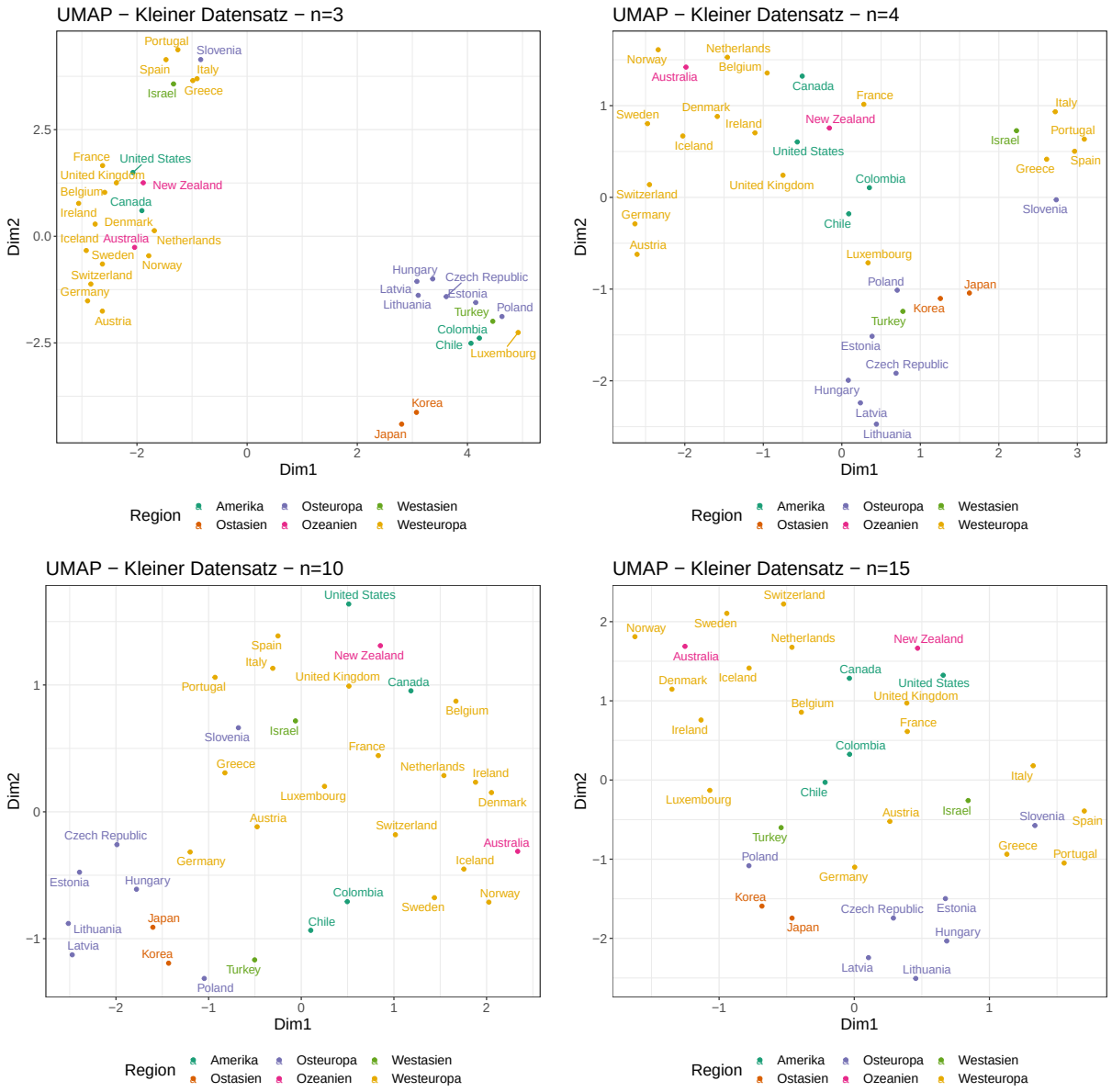

Abbildung A.6: *Visualisierung der zweidimensionalen Konfiguration von UMAP für den kleinen Datensatz, für verschiedene n_neighbors = 3,4,10,15.*

UMAP Parameterwahl n_neighbors - Großer Datensatz

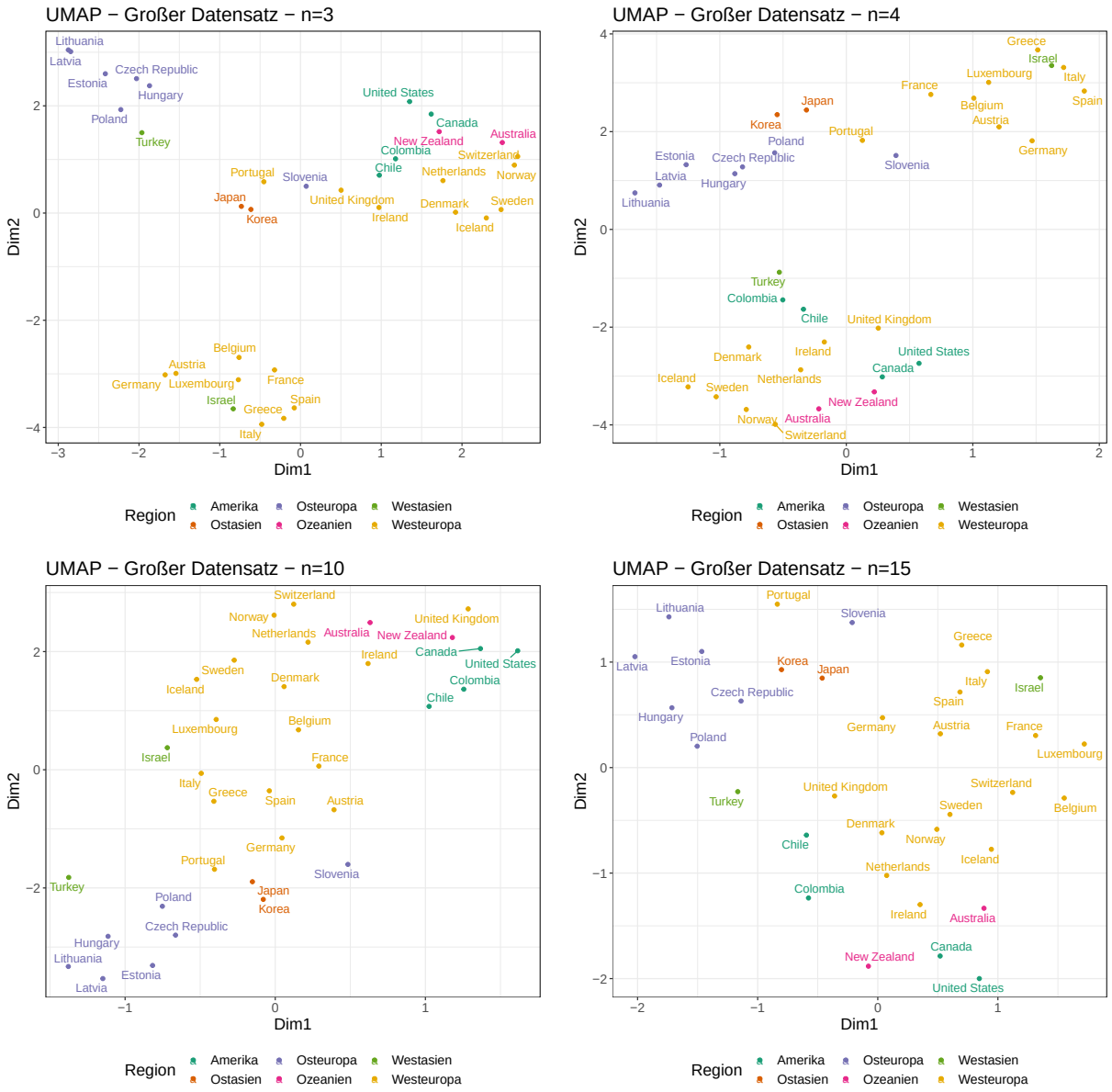

Abbildung A.7: *Visualisierung der zweidimensionalen Konfiguration von UMAP für den großen Datensatz, für verschiedene n_neighbors = 3,4,10,15.*

A.5 Dendrogramme

Hierarchisches Clustering mit euklidischer Distanz und Ward-Verfahren auf den Ergebnissen der PCA, der MFA und der MDS, für den kleinen und den großen Datensatz:

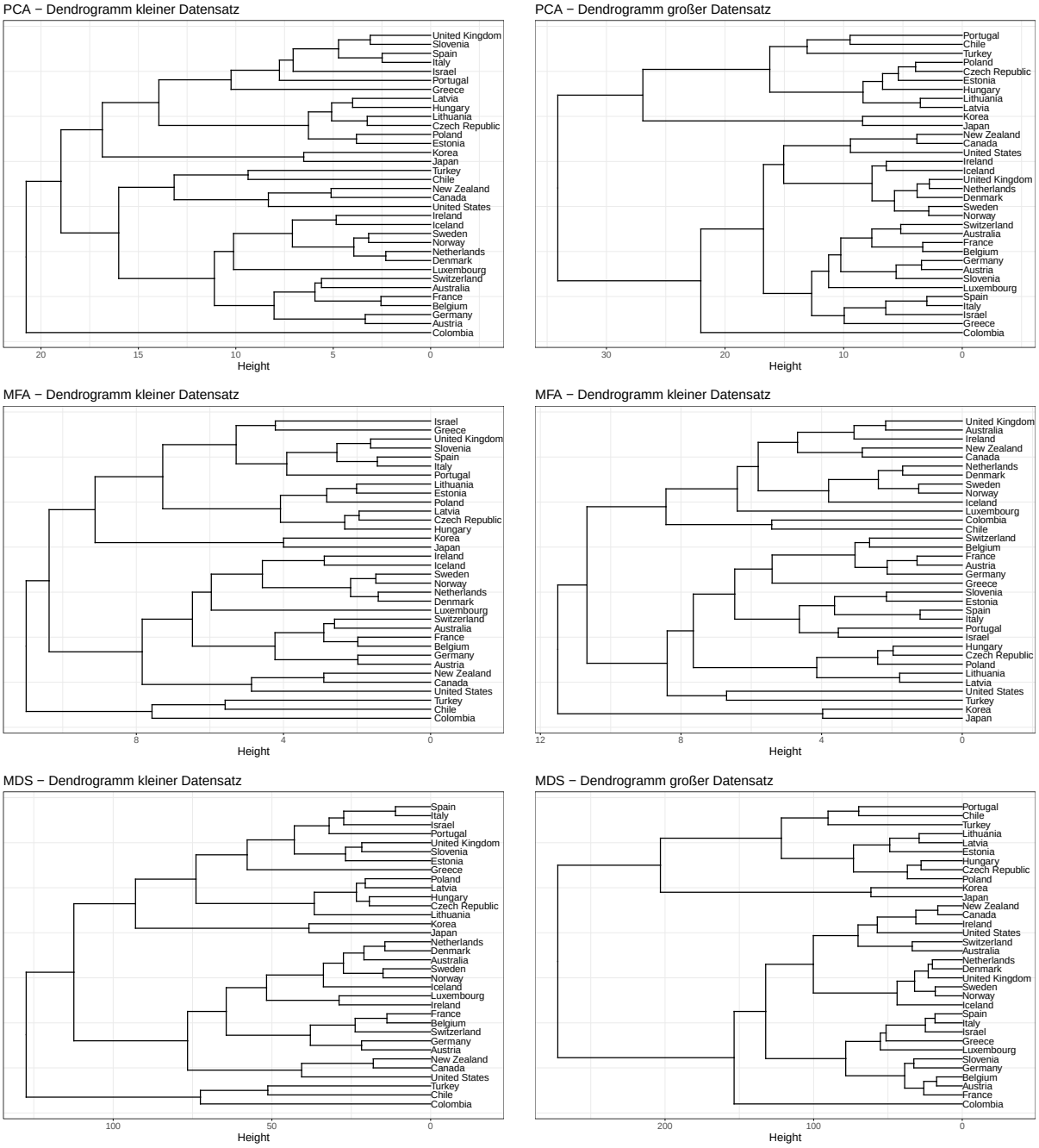

Abbildung A.8: *Visualisierung der Dendrogramme für die PCA, die MFA und die MDS, links für den kleinen Datensatz und rechts für den großen Datensatz.*

Hierarchisches Clustering mit euklidischer Distanz und Ward-Verfahren auf den Ergebnissen von UMAP und t-SNE, für den kleinen und den großen Datensatz:

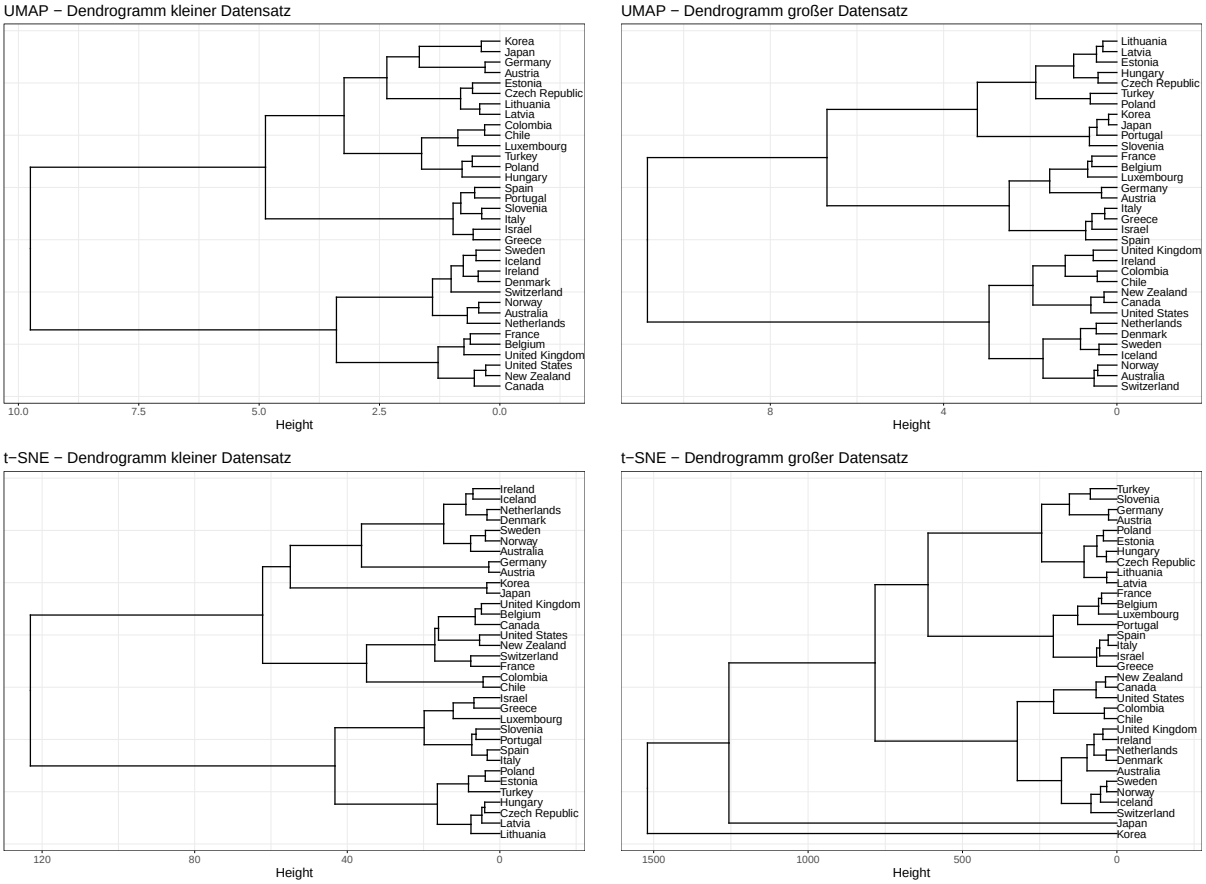

Abbildung A.9: *Visualisierung der Dendrogramme für UMAP und t-SNE, links für den kleinen Datensatz und rechts für den großen Datensatz.*

A.6 Interne Validierungsindizes

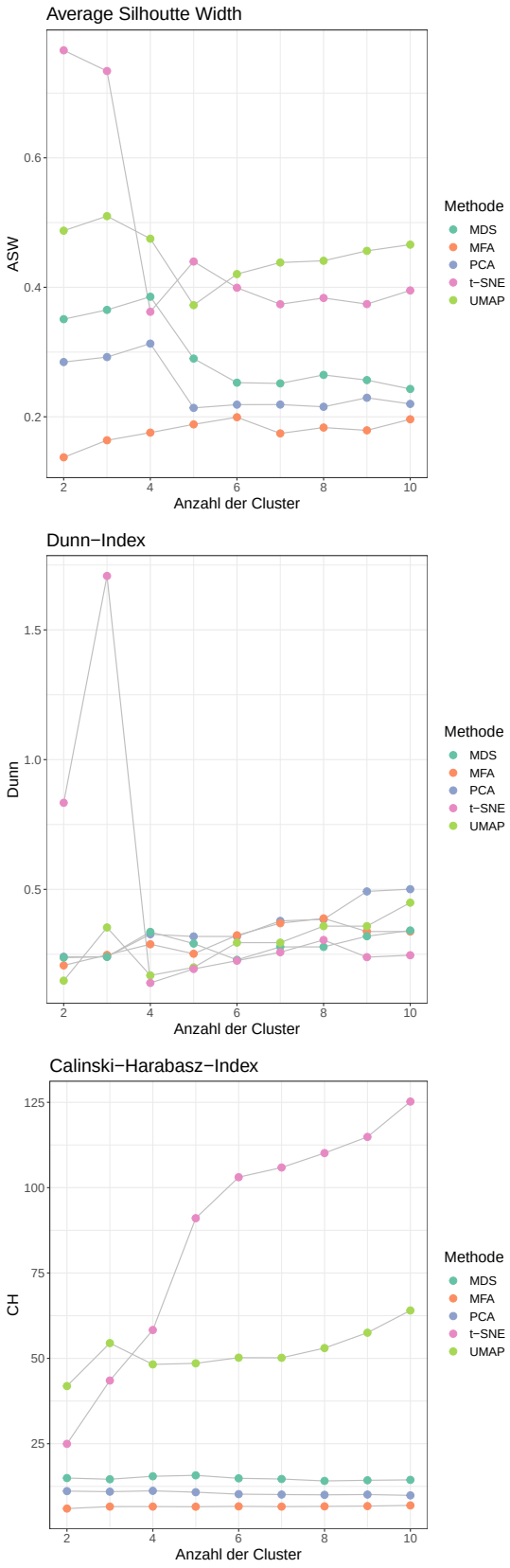

Abbildung A.10: *Visualisierung des ASW, Dunn-Index und des CH-Index der k-Means Ergebnisse für jede Methode in Abhängigkeit der Clusteranzahl. Die Durchführung erfolgte hier auf dem großen Datensatz.*

A.7 Optimale Anzahl an Cluster

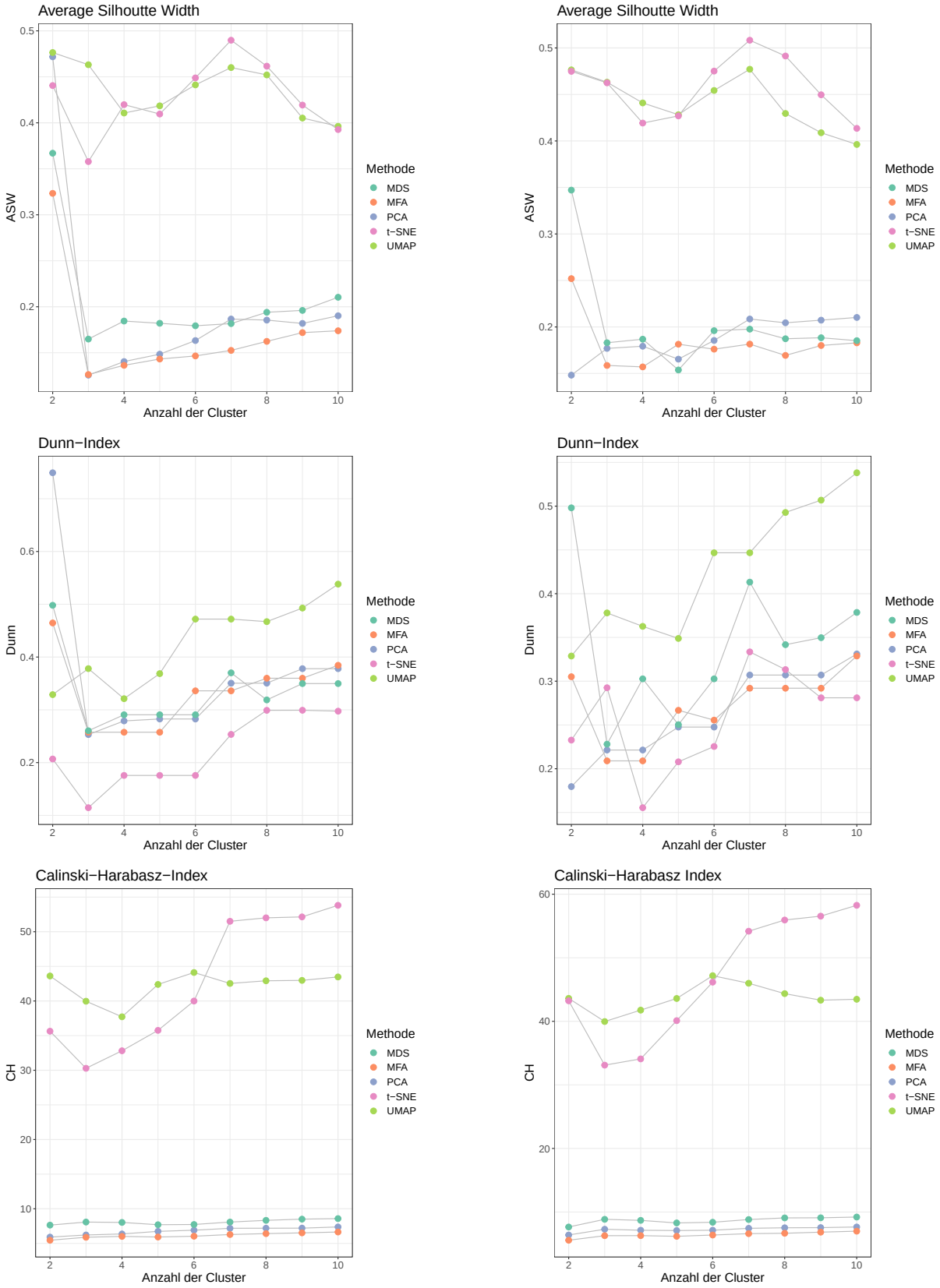

Abbildung A.11: *Vergleich der optimalen Clusteranzahl zwischen dem hierarchischen Verfahren links und dem k-Means Verfahren rechts, anhand dem ASW, dem Dunn-Index und dem CH-Index. Die zwei Clusterverfahren wurden auf die Ergebnisse der Methoden PCA, MFA, MDS, UMAP und t-SNE angewandt, anhand des kleinen Datensatzes.*

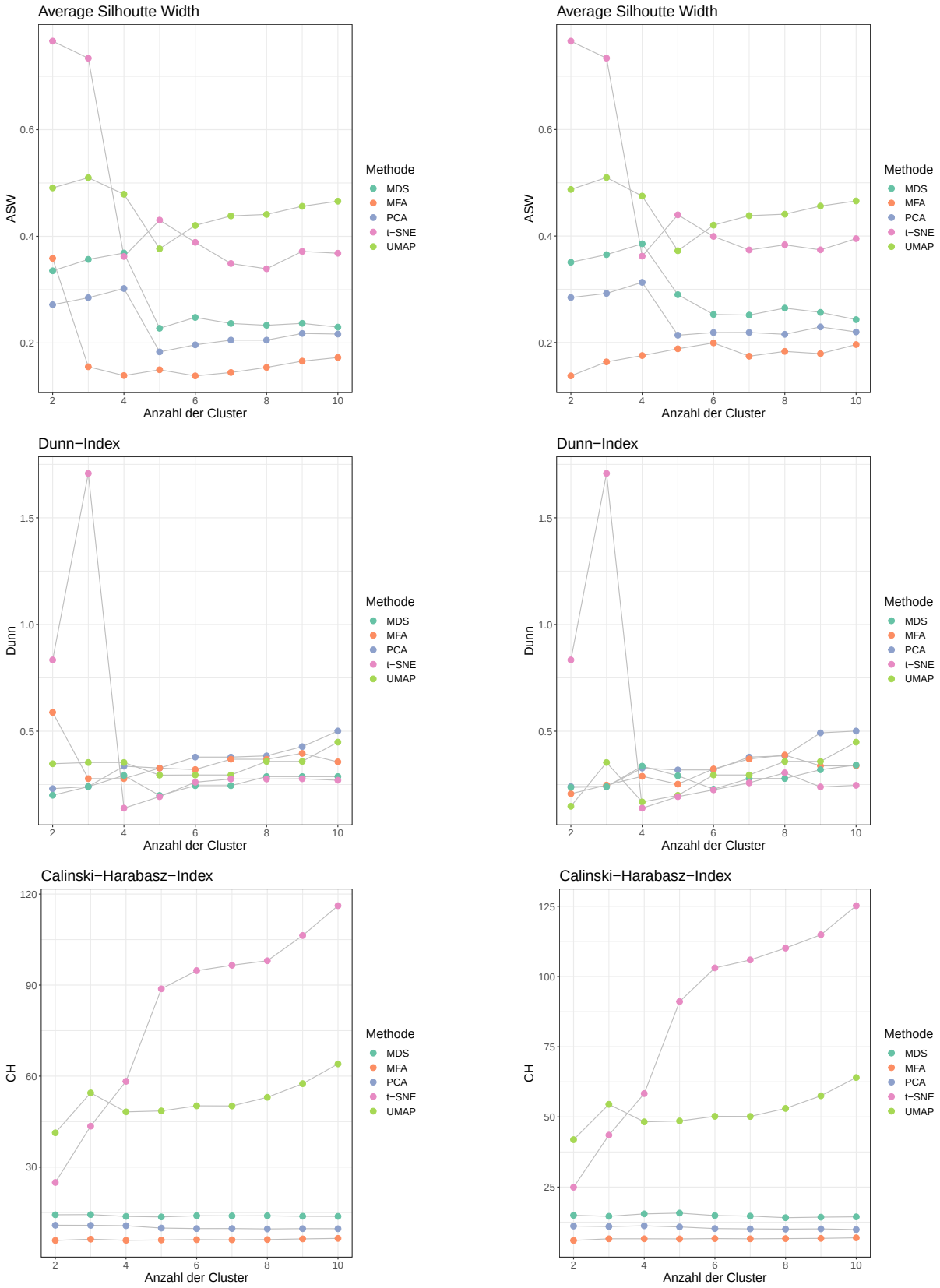

Abbildung A.12: *Vergleich der optimalen Clusteranzahl zwischen dem hierarchischen Verfahren links und dem k-Means Verfahren rechts, anhand dem ASW, dem Dunn-Index und dem CH-Index. Die zwei Clusterverfahren wurden auf die Ergebnisse der Methoden PCA, MFA, MDS, UMAP und t-SNE angewandt, anhand des großen Datensatzes.*

A.8 Länderzuordnung zwischen Clusterverfahren

Unterscheiden sich das hierarchische Verfahren und das *k*-Means Verfahren in der Zuordnung der Länder?

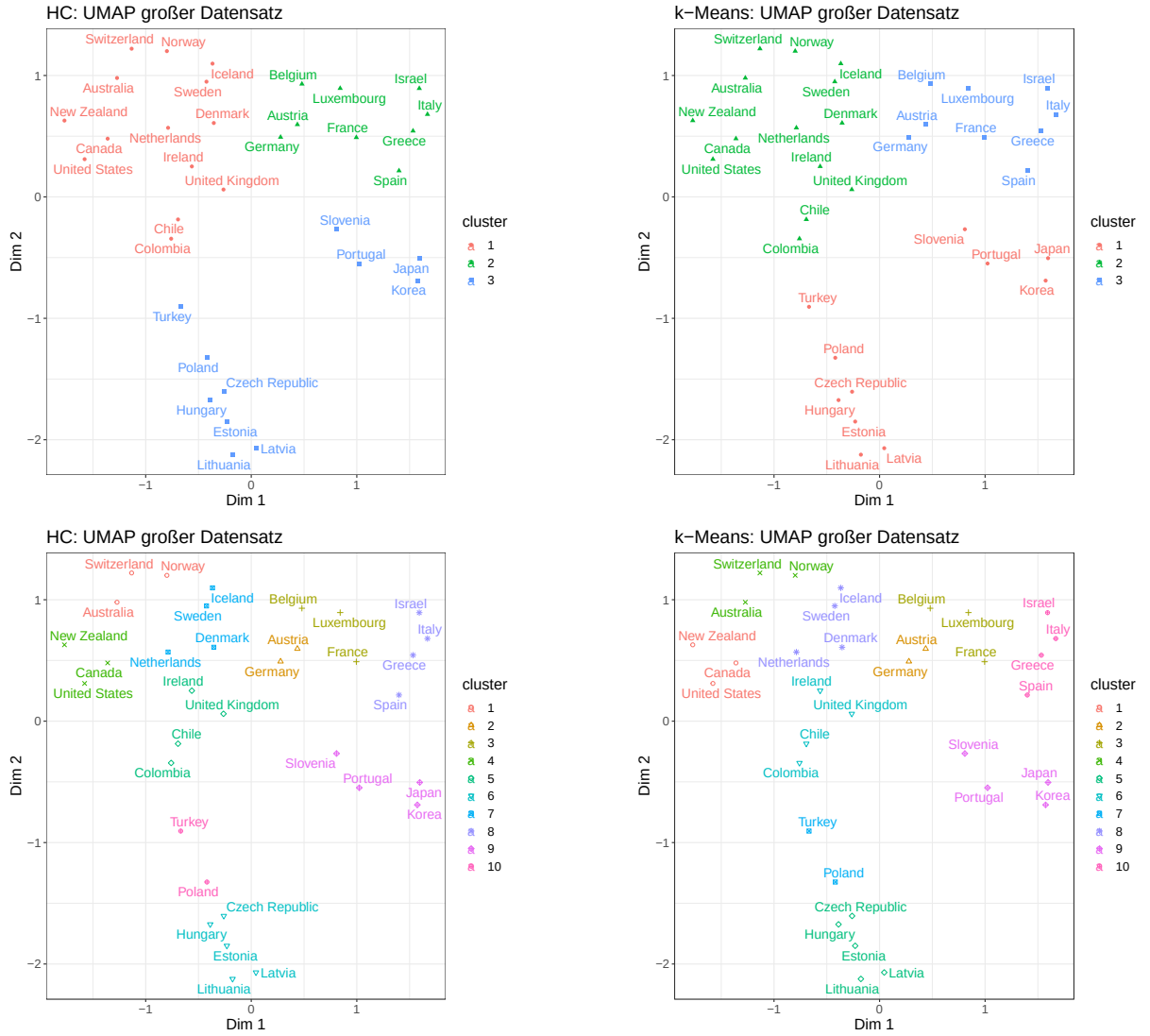

Abbildung A.13: *Vergleich der Zuteilung der Länder in die Cluster zwischen dem hierarchischen Verfahren und dem k-Means Verfahren. Betrachtung anhand der ausgewählten optimalen Clusteranzahl nach den drei Validierungsindizes für die Ergebnisse der Methode UMAP, für den großen Datensatz.*

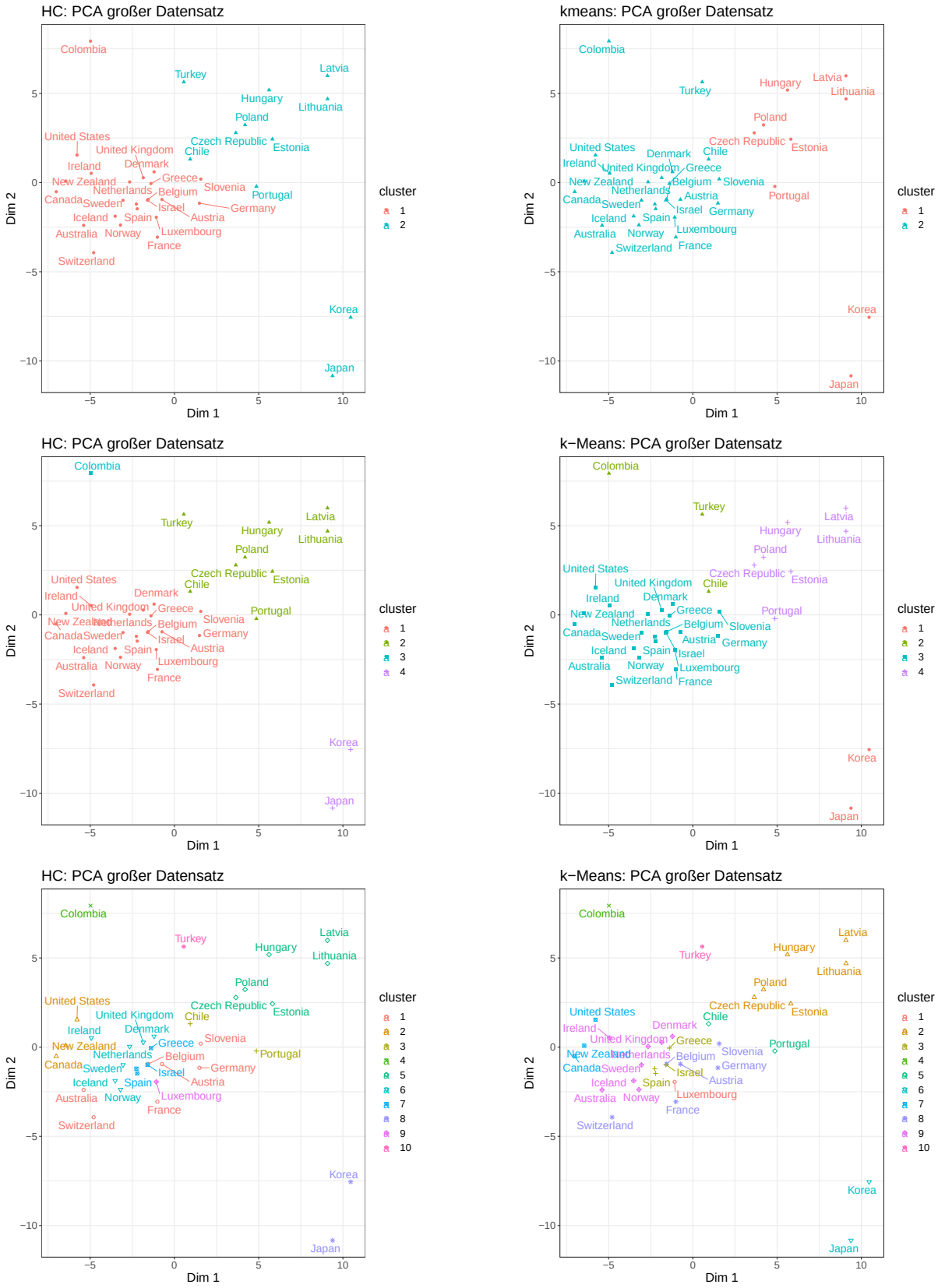

Abbildung A.14: *Vergleich der Zuteilung der Länder in die Cluster zwischen dem hierarchischen Verfahren und dem k-Means Verfahren. Betrachtung anhand der ausgewählten optimalen Clusteranzahl nach den drei Validierungsindizes für die Ergebnisse der PCA, für den großen Datensatz.*

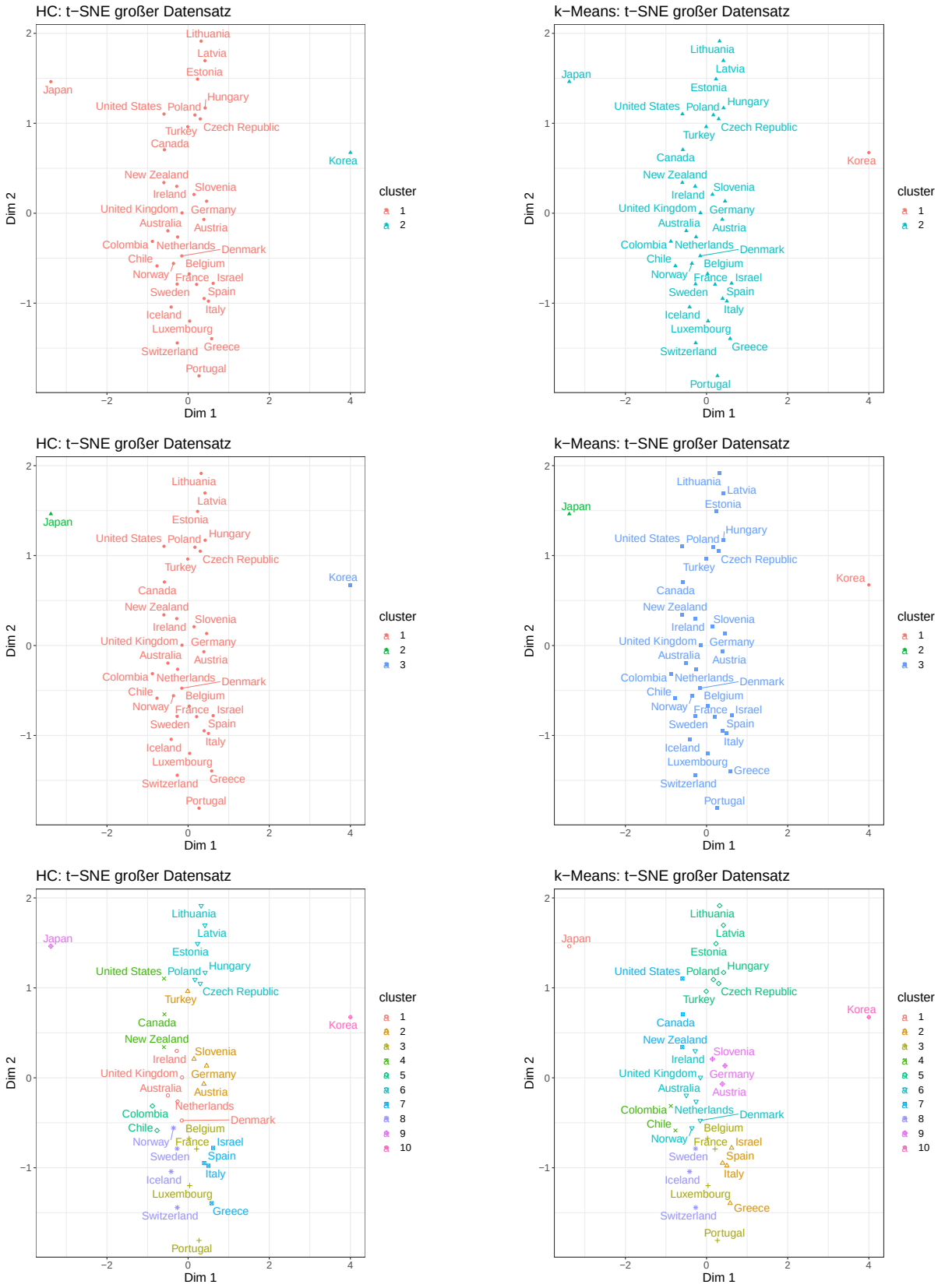

Abbildung A.15: *Vergleich der Zuteilung der Länder in die Cluster zwischen dem hierarchischen Verfahren und dem k-Means Verfahren. Betrachtung anhand der ausgewählten optimalen Clusteranzahl nach den drei Validierungsindizes für die Ergebnisse der Methode t-SNE, für den großen Datensatz.*

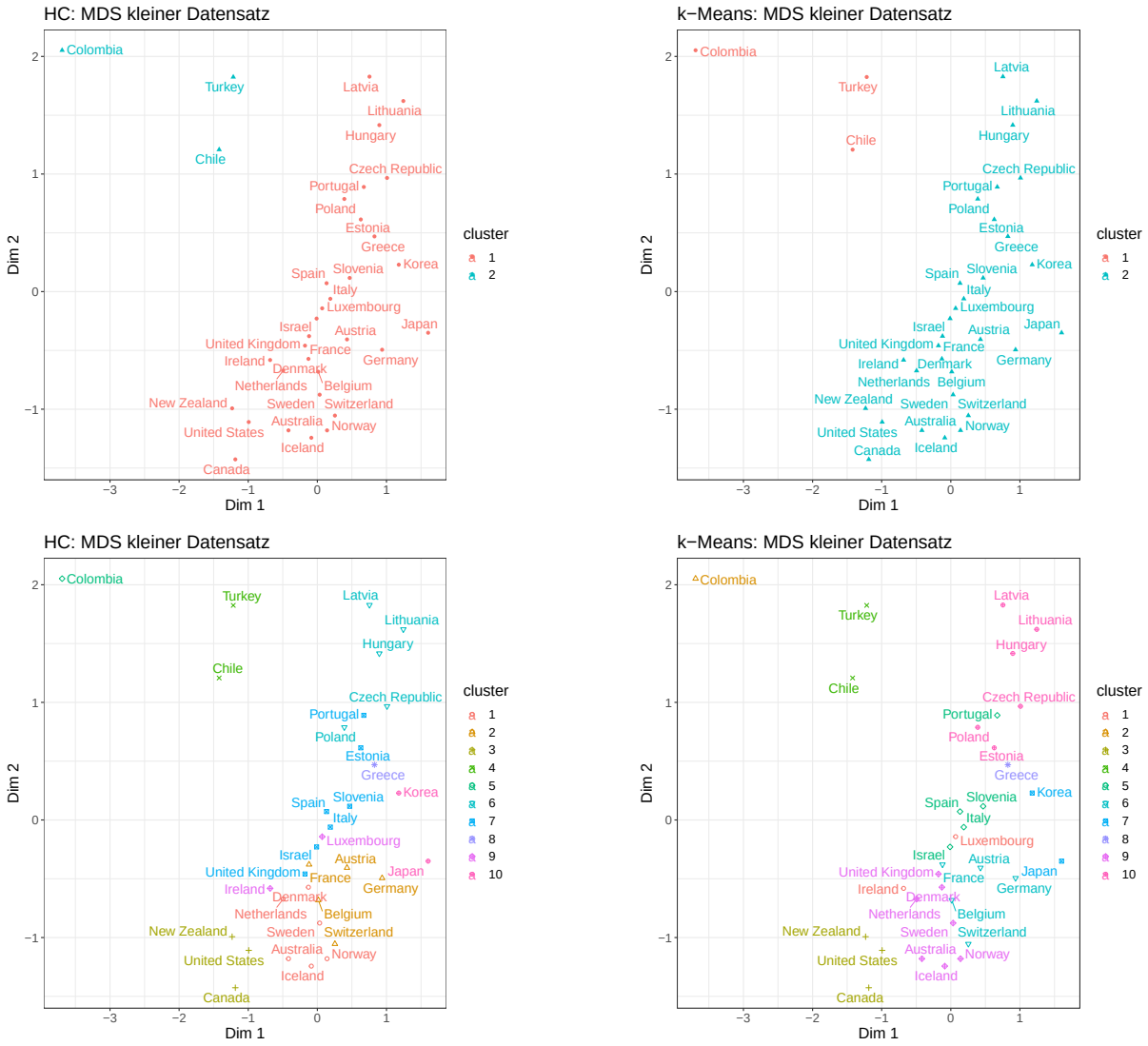

Abbildung A.16: *Vergleich der Zuteilung der Länder in die Cluster zwischen dem hierarchischen Verfahren und dem k-Means Verfahren. Betrachtung anhand der ausgewählten optimalen Clusteranzahl nach den drei Validierungsindizes für die Ergebnisse der MDS, für den kleinen Datensatz.*

A.9 Länderzuordnung zwischen Methoden

Unterscheidet sich die Zuordnung der Länder zwischen den Methoden für eine feste Anzahl an Cluster?

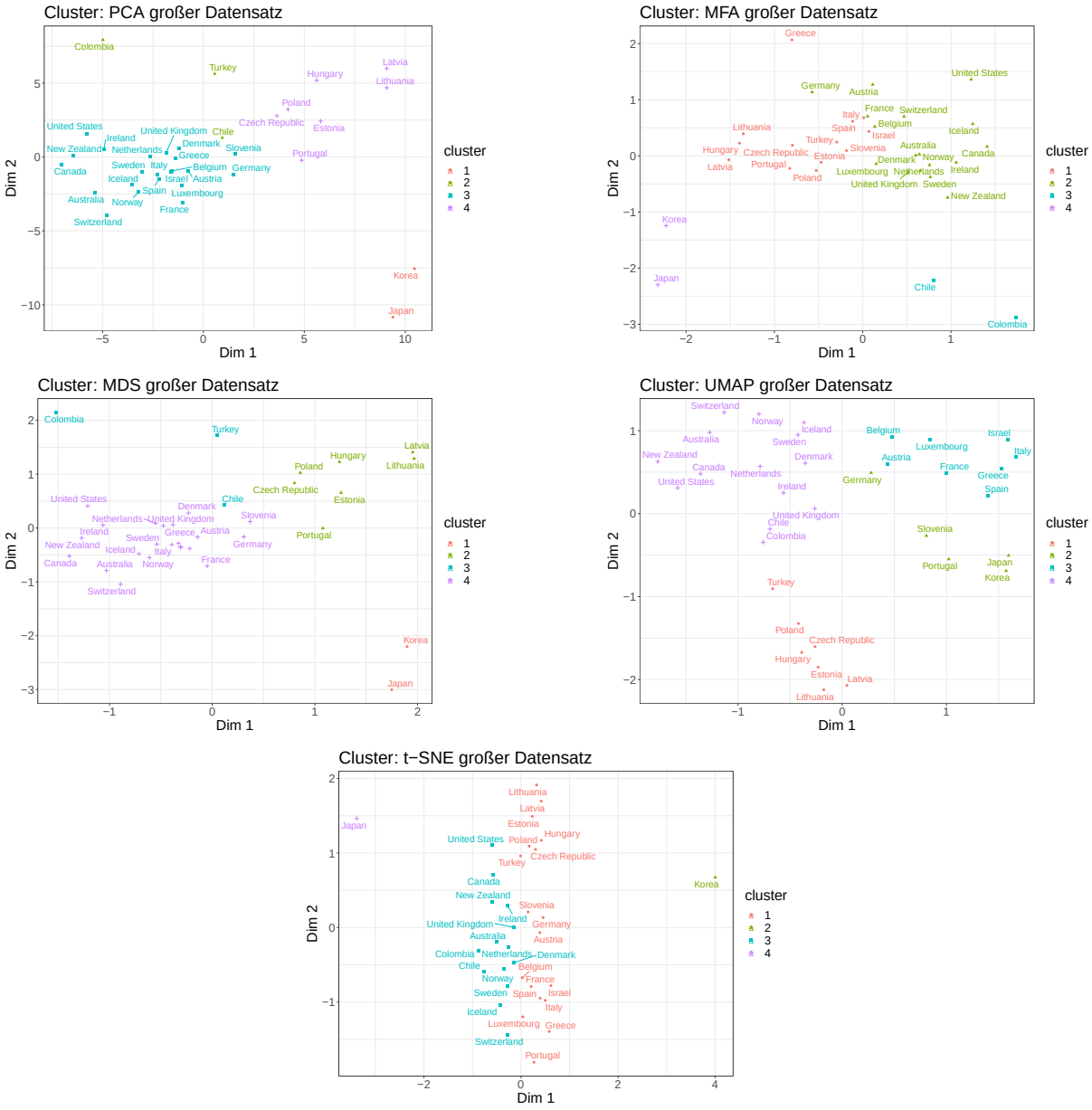

Abbildung A.17: *Vergleich der Zuteilung der Länder in vier Cluster. Betrachtung der Methoden in Kombination mit dem k-Means Verfahren, anhand des großen Datensatzes.*

Anhang B

Elektronischer Anhang

Als elektronischer Anhang zu dieser Arbeit werden die verwendeten Datensätze sowie die Datenaufbereitung und die multivariate Analyse über den Datenverwaltungsdienst *LRZ Sync+Share* des Leibniz-Rechenzentrums zur Verfügung gestellt.
Die Markdown-Files in **R** sind wie folgt strukturiert:

Datenaufbereitung.Rmd
: Einlesen der Teildatensätze und Erstellung des gesamten Datensatzes. Code zu der Analyse der fehlenden Werte, der Variablenselektion, der Variablenumbenennung, der Erstellung der Variablenlabels, der Imputation mit MICE und dem Speichern der Datensätze.

Methodenvergleich1.Rmd
: Durchführung der Methoden PCA und MFA.

Methodenvergleich2.Rmd
: Durchführung der Methoden UMAP, t-SNE und MDS.

Clustering auf Methoden.Rmd
: Durchführung der Clusterverfahren (*k*-Means und hierarchisches Clustering) mit den Ergebnissen der Methoden aus *Methodenvergleich1.Rmd* und *Methodenvergleich2.Rmd*. Berechnung der internen Validierungsindizes und des kophenetischen Korrelationskoeffizienten zwischen den Dendrogrammen, sowie die Visualisierung der Clusterergebnisse.

Anhang.Rmd
: Durchführung der PCA mit der gemeinsamen Korrelationsmatrix (über 50 imputierten Datensätze). Ergebnisse der PCA der vier anderen imputierten Datensätze des kleinen Datensatzes.

Pakete.Rmd
: Auflistung aller verwendeten Pakete.