Methodologisches Konzept des Deutschen Lernatlas-Index (DLA-Index)

Beschreibung

Dieser Methodikbericht wurde erstellt von:

Dr. Björn Christensen und Dr. Kerstin Reimer, Analytix GmbH, Kiel
Fernando Cartwright, Polymetrika, Ottawa, Kanada
Martin Rohde et al., OFFIS Institut für Informationstechnologie, Oldenburg

Kontakt:

Dr. Ulrich Schoof
Project Manager
Programm: Zukunft der Beschäftigung / Good Governance
Bertelsmann Stiftung
Telefon 05241 81-81 384
Mobile 0172 2958064
Fax 05241 81-681 384
E-Mail ulrich.schoof@bertelsmann-stiftung.de
www.bertelsmann-stiftung.de

Inhalt

1. Einführung

Die Entscheidungen im Zusammenhang mit der Konstruktion von Composite-Indizes können in drei breite Kategorien unterteilt werden: Methodologie, Methoden und Instrumentation. Die konkreten Verfahren der Konstruktion eines Composite-Index aus Rohdaten sind die Methoden. Die Wahl der Methoden wird durch die zugrunde liegenden Annahmen und Entscheidungen bestimmt, die die Methodologie ausmachen. Auch wenn mehrere Methoden zur Auswahl stehen können, die den Anforderungen der Methodologie genügen, werden die brauchbarsten Methoden im Allgemeinen durch die Form der Rohdaten bestimmt, die durch die Instrumentation erzeugt werden. Die Instrumentation beinhaltet die Erhebung von Beobachtungen und ihre Aufbereitung in eine numerische Rohform, wie etwa Umfragen und die dazugehörigen Methoden.

Die Erstellung eines Composite Indicators ist in erster Linie eine Messaufgabe. In den Sozialwissenschaften ist die Messung eine praktische Aufgabe zur quantitativen Beschreibung von Phänomenen ohne natürliche Metrik. Die Validität einer Messung oder eines Messinstruments hängt hauptsächlich davon ab, wie passend die resultierenden Zahlen den Rückschlüssen zugeordnet werden, die sie stützen müssen. Obwohl in unserem Fachgebiet die Meinungsverschiedenheiten darüber anhalten, welche Belege erforderlich sind, um die Validität von Messungen zu „beweisen“, müssen zwei allgemeine Bedingungen erfüllt sein.

Erstens sollten zufällige oder systematische Fehler nicht zu willkürlichen Rückschlüssen führen. Anders ausgedrückt: Rückschlüsse, die auf der Interpretation der Messungen basieren, sollten eher mit dem Urteil sachkundiger Experten in Einklang stehen als solche, die ohne Messungen gezogen werden. Oftmals gibt es kein Expertenurteil ohne das Composite, was die methodische Klarheit und Transparenz der Konstruktion des Composite-Index noch wichtiger macht.

Zweitens sollte die aus den Komponentenvariablen und ihrer Zusammensetzung resultierende stipulative Definition des Messziels mit der mit Interpretationen der Messung verbundenen normativen Definition übereinstimmen. Dies sei anhand eines Negativbeispiels verdeutlicht: Im Erziehungsbereich korreliert die Leseleistung stark mit der Mathematikleistung, und aufgrund des dämpfenden Effekts von Messfehlern können die Ergebnisse eines sehr genauen Mathematiktests enger mit der Leseleistung korrelieren als die eines ungenauen Lesetests. Aufgrund systematischer begrifflicher Unterschiede zwischen der Lese- und der Mathematikleistung lassen die Ergebnisse eines Mathematiktests jedoch keine validen Rückschlüsse auf die Leseleistung der Grundgesamtheit zu.

Im Hinblick auf Methodologie, Instrumentation und Methoden wurden jeweils spezifische Entscheidungen getroffen, um diesen Kriterien bei der Konstruktion des DLA-Index gerecht zu werden. Kapitel 2 dieses Arbeitspapiers beschreibt einen methodologischen Rahmen, der die methodologischen Paradigmen der Messung und der Prognose zu einer Synthese verbindet. Diese Paradigmen schaffen zusammen einen Rahmen für den intern konsistenten Einsatz unterschiedlicher Konstruktionsmethoden. Kapitel 3 behandelt die Rolle der Instrumentation bei der Erstellung von Composite-Indizes sowie mehrere bei der Aufbereitung der Daten wichtige Überlegungen, die die Bandbreite der durch einen Composite-Index gestützten Rückschlüsse eingrenzen oder ausdehnen. Kapitel 4 bietet einen detaillierten Überblick über die zur Konstruktion des DLA verwendeten Methoden.

2. Methodologie

Alle Composite-Indizes sind in dem Sinne statistische Aggregate, dass Daten aus verschiedenen Quellen gemäß einer parametrischen Funktion so kombiniert werden, dass das resultierende Composite eine schlüssige Gesamtaussage erlaubt. Statistische Aggregate gehören im Allgemeinen zu einem von zwei Paradigmen: Messung oder Prognose. Obwohl sie wegen der Ähnlichkeit ihrer Berechnungsmethoden oft verwechselt werden, bilden Messung und Prognose aufgrund ihrer unterschiedlichen zugrunde liegenden Annahmen in vielerlei Hinsicht Gegensätze. Die zwei Paradigmen werden im Folgenden getrennt behandelt und dann mit Blick auf ihre Bedeutung für die Erstellung eines Composite-Index untersucht.

2.1 Das Messparadigma

Das charakteristische Kennzeichen des Messparadigmas ist die Annahme, dass es ein Merkmal oder Konstrukt gibt, das zwar nicht selbst direkt beobachtbar ist, aber beobachtbare Phänomene beeinflusst. Wenn es möglich ist, ein Konstrukt zu messen, dann sollten die Variablen, die verschiedene Arten von mit diesem Konstrukt in Beziehung stehenden Beobachtungen abbilden, miteinander korrelieren, da sie dieselbe zugrunde liegende Dimension darstellen. Der Grad, in dem eine bestimmte Variable nicht mit den anderen Variablen übereinstimmt, ist eine Indikation für ihren Messfehler. Zur Optimierung des Schätzaggregats in einem Messparadigma muss jede Komponentenvariable mit dem Inversen ihres Messfehlers gewichtet werden.

In der Praxis ist die Bestimmung des Messfehlers problematisch, weil er zunächst unbekannt ist. Wir sind gezwungen, jede Variable (fälschlicherweise) als messfehlerfrei anzunehmen. Bei jeder Menge aus mehr als zwei Variablen wird das Kriterium zur Bestimmung des Messfehlers durch die verbleibenden Variablen definiert und ist daher für jede Variable in der Menge unterschiedlich. Um ein einheitliches Kriterium festzulegen, wird das Konzept der Übereinstimmung zwischen Variablen und Konstrukten öfter als die Übereinstimmung zwischen jeder Variablen und einem globalen Faktor operationalisiert, der allen Variablen gemein ist. Die Festlegung von Kriterien für die Qualität einer Variablen im Messparadigma ist also ein zweistufiger Prozess: Erstens, Ermittlung des gemeinsamen Faktors der verschiedenen Beobachtungen und, zweitens, Ermittlung der Verwandtschaft oder Ähnlichkeit jeder Variablen mit diesem gemeinsamen Faktor. Mit zunehmender Variablenzahl werden die Schätzungen des gemeinsamen Faktors stabiler und die des Messfehlers jeder Variablen genauer. Im Messparadigma hängt die Genauigkeit eines Aggregats von der Zahl der Komponentenvariablen und dem Grad ihrer Redundanz ab. Wenn sowohl die Zahl als auch die Redundanz steigt, nimmt die Genauigkeit der Messfehlerschätzung zu und in der Folge können den Variablen im Aggregationsprozess optimale Gewichte zugeordnet werden.

Das Kausalmodell der Messvarianz ist im linken Teil von Abbildung 1 illustriert. Die Streuungsursache in einem Messparadigma ist ein latentes Konstrukt, das auch als Dimension oder Faktor bezeichnet wird und die unabhängige Komponente des Modells darstellt. Die Existenz von Messkonstrukten wird angenommen, kann aber nicht bewiesen werden. Die beobachteten Variablen Var1, Var2 und Var3 sind Manifestationen des Messkonstrukts X und die abhängigen Komponenten. Diese Annahme wird gestützt, wenn in einer Variablenmenge jede Variable mit jeder anderen in Beziehung steht. Oft gibt es Teilmengen von Variablen mit Wechselbeziehungen, die nicht zufällig genauso stark sind wie die Wechselbeziehungen zwischen den übrigen Variablen. Dieses Phänomen ist im rechten Teil von Abbildung 1 illustriert. Während die meisten Variablen in unter-

schiedlichem Maße innerhalb eines gemeinsamen Raumes zueinander in Beziehung stehen, haben viele Variablen mehr mit einem sekundären Raum gemein. Diese Variablen wären relativ schlechte Maße für die erste Dimension, stellen jedoch genaue Maße für die zweite Dimension dar.

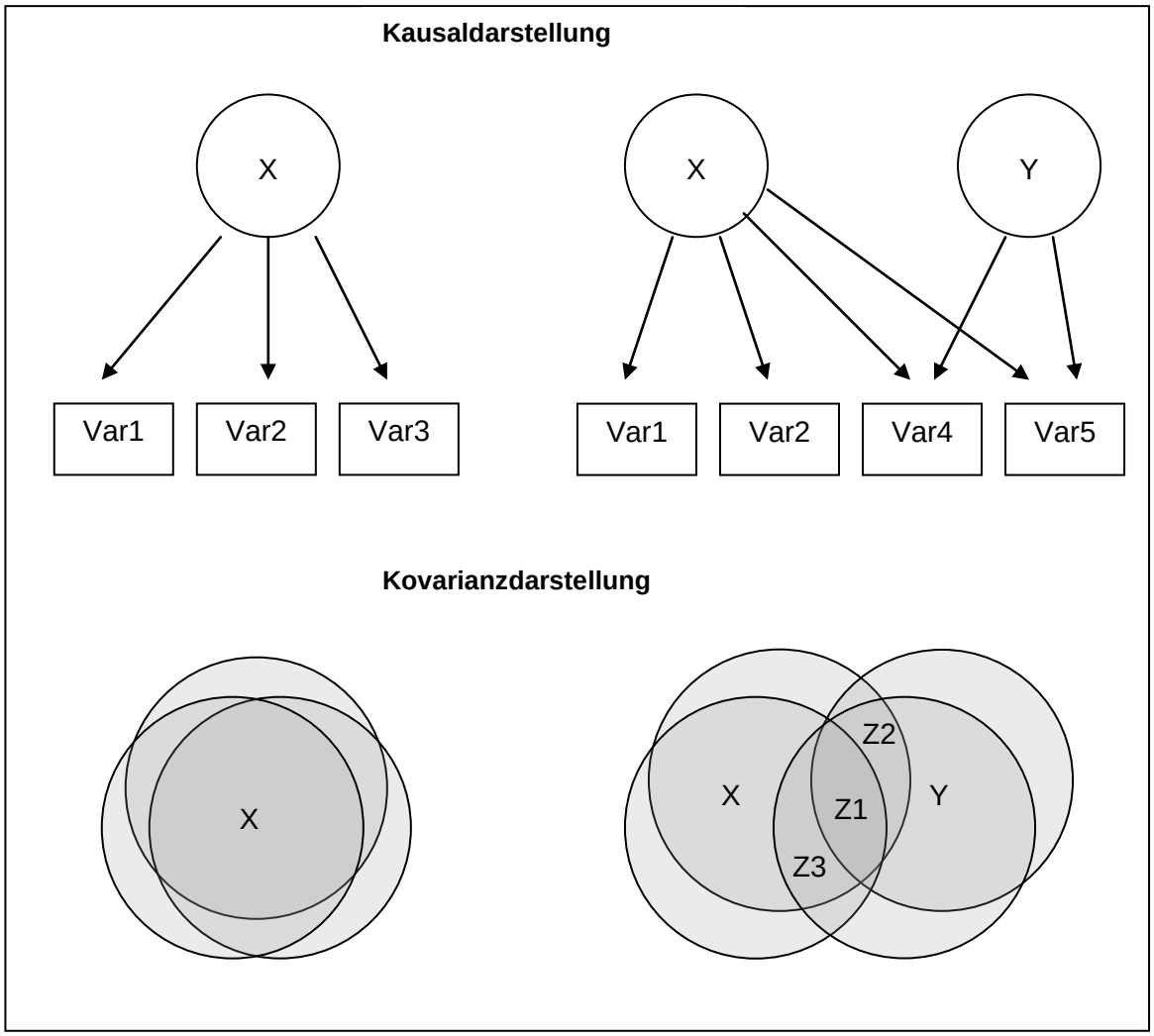

Abbildung 1. Kausal- und Kovarianzdarstellungen von Messmodellen

Um die theoretische Annahme zu testen, dass ein Konstrukt existiert und gemessen werden kann, beginnt die statistische Datenanalyse im Messparadigma gewöhnlich mit einer Untersuchung der Dimensionalität einer Menge von Messvariablen. Stellen diese mehr als eine Dimension dar, muss bestimmt werden, wie viele Dimensionen für das Untersuchungsphänomen relevant sind. Oftmals beschreiben sekundäre Dimensionen Faktoren, die mit dem untersuchten Gebiet in keiner Beziehung stehen, wie z.B. die Messverzerrung. Sie sollten daher isoliert und entfernt werden.

Unter dem Blickwinkel der Datenanalyse ist die wahre Kausalstruktur leider nie bekannt, sodass durch die Kommunalität (gemeinsame Kovarianz) zwischen Variablen ein statistisches Konstrukt definiert werden muss. Die unteren Figuren in Abbildung 1 illustrieren mittels Venn-Diagrammen die Kovarianz zwischen den Variablen. Die Kreise stellen einzelne Variablen dar und der Grad ihrer Überlappung bildet die Stärke ihrer Kovarianz ab. Im Falle des eindimensionalen Modells auf der linken Seite stellt die Kommunalität genau eine einzige Dimension dar, da die gemeinsame

Kovarianz in dem mit „X“ gekennzeichneten Bereich den größten Teil der Varianz aller drei Variablen ausmacht. Hingegen ist in der Figur rechts unten der gemeinsame Faktor aller Variablen der mit „Z1“ gekennzeichnete Bereich, der bei allen Variablen nur einen sehr kleinen Teil der Varianz ausmacht. Daher ist es bei dieser Menge von Variablen nicht möglich, ein einheitliches Konstrukt adäquat zu messen. Die zwei größten Bereiche gemeinsamer Varianz sind X und Y, was darauf hindeutet, dass sie getrennt werden können, um zwei Konstrukte zu messen. Die Situation wird durch die Bereiche Z1, Z2 und Z3 verkompliziert, die ebenfalls mehreren Variablen gemein sind. Jede inkorrekte Zuordnung dieser Bereiche zu X oder zu Y kann zu systematischen Fehlern führen. Aus diesem Grund kann das Messparadigma nicht nur eine statistische Aufgabe sein – es erfordert eine sorgfältige inhaltliche Analyse der Variablen selbst.

Bei der Entwicklung eines Composite Indicators ist das Messparadigma nützlich, wenn a) ein Teilbereich mangelhaft oder normativ definiert ist, aber keine natürliche Metrik besitzt; oder b) mehrere Messungen konzeptionell ähnlich sind, aber keine kanonische Methode existiert, sie als Gesamtheit zu interpretieren. Zu den statistischen Aggregationsmethoden, die sich auf dieses Paradigma stützen, gehören finite Mischmodelle, die klassische und die probabilistische Testtheorie, die Faktorenanalyse und die Generalisierbarkeitstheorie. Jede dieser Methoden geht von der Existenz eines oder mehrerer latenter Konstrukte aus, die die beobachteten Daten erzeugen.

2.2 Das Prognoseparadigma

Während ein Aggregat im Messparadigma nützlich ist, weil es ein unbeobachtbares Konstrukt abbildet, ist es im Prognoseparadigma von Nutzen, weil es eine aktuell nicht verfügbare Kriteriumsvariable vertreten kann. Die Kriteriumsvariable (gegebenenfalls eine Menge von Variablen, aber der Klarheit halber verwenden wir den Singular) ist möglicherweise nicht verfügbar, weil sie eine zukünftige Gegebenheit, wie zum Beispiel das Wirtschaftswachstum, darstellt oder ein nur mit sehr hohen Kosten direkt messbares Konstrukt, wie etwa das öffentliche Vertrauen in die Regierung. In jedem Fall ist das Aggregat ein nützlicher Proxy, um Rückschlüsse auf die nicht verfügbare Variable zu ziehen. Die Kriteriumsvariable ist bei der Konstruktion des Aggregats von entscheidender Bedeutung, wird aber kein Teil von ihm.

In einem Prognoseparadigma werden die zu aggregierenden Variablen normalerweise nach ihrer relativen Bedeutung für das Kriterium oder der relativen Stärke ihrer Beziehung zu diesem gewichtet. Die wichtigste Voraussetzung dafür, die Aggregation der Variablen zu spezifizieren, ist die Festlegung einer klaren Definition der Kriteriumsvariablen. Ist diese mangelhaft definiert, wird die Prognose instabil. Deshalb erfordert die optimale Aggregation im Prognoseparadigma ebenfalls eine zweistufige Vorgehensweise. Im ersten Schritt wird (werden) die Kriteriumsvariable(n) definiert und im zweiten Schritt wird die relative Stärke jeder zu aggregierenden Variablen als Prädiktor des Kriteriums berechnet.

Im Gegensatz zum Messparadigma trifft das Prognoseparadigma keine Annahmen über die Wechselbeziehungen zwischen den verschiedenen Variablen. Die Prognose wird sogar vereinfacht, wenn keinerlei Beziehung zwischen den zu aggregierenden Variablen besteht. Im oberen Teil von Abbildung 2 sind zwei Kausaldarstellungen von Prognosemodellen illustriert. Die Modelle scheinen zwar oberflächlich betrachtet den Modellen in Abbildung 1 zu ähneln, beinhalten jedoch keine Annahmen über die Existenz zugrunde liegender Konstrukte. Außerdem ist die Richtung der

Kausalität umgekehrt, sodass das Aggregat nicht die unabhängige, sondern die abhängige Komponente des Modells ist.

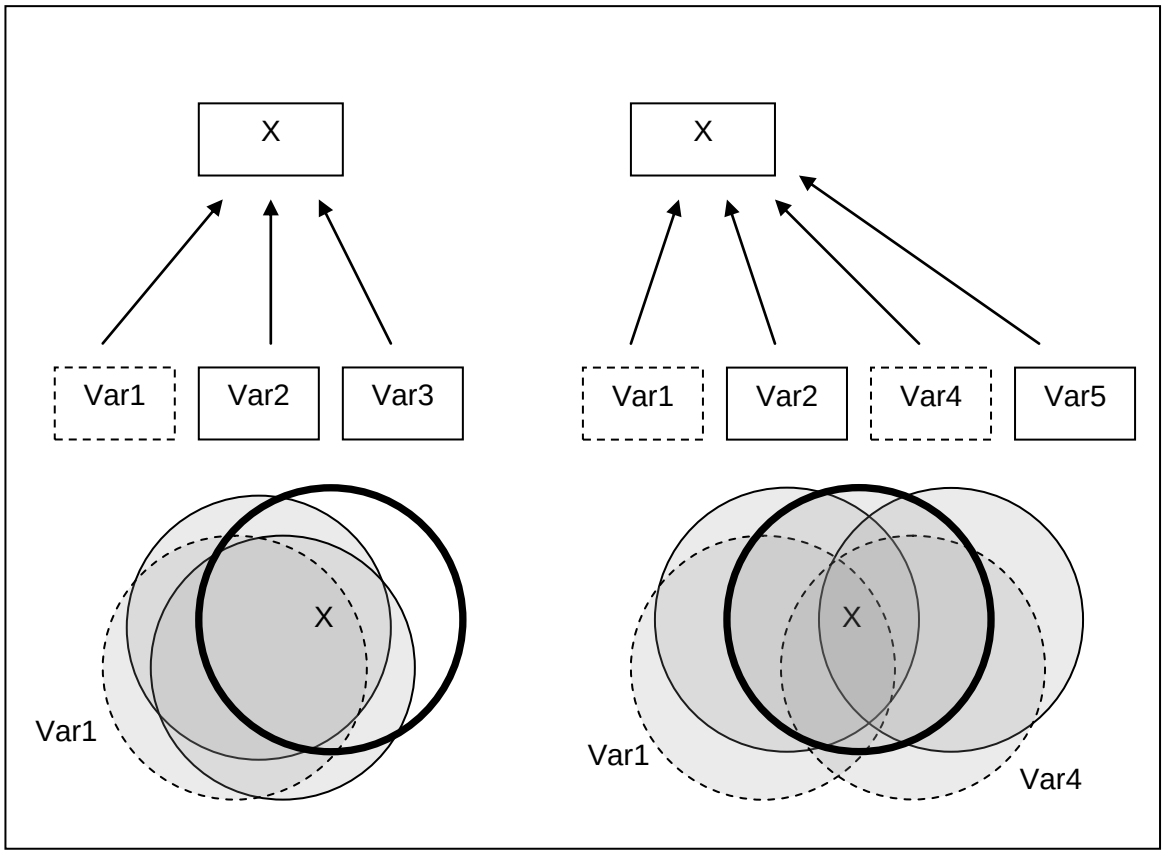

Abbildung 2. Kausal- und Kovarianzdarstellungen von Prognosemodellen

Die Kovarianzdarstellungen im unteren Teil von Abbildung 2 verdeutlichen die Bedeutung der Komplementarität. Die Prädiktorvariablen weisen dieselben Kovarianzmuster auf wie die Messvariablen in Abbildung 1. Die Variablen auf der linken Seite sind weitgehend redundant, die auf der rechten Seite eher komplementär. Die zwei Mengen von Variablen sollen die Varianz der Kriteriumsvariablen X erklären. Die Menge der redundanten Variablen lässt einen erheblichen Teil dieser Varianz unerklärt. Eine Variable trägt keine Zusatzinformation zur Prognose bei, weil ihr Beitrag von den Beiträgen der anderen zwei Variablen vollständig überlagert wird. Pragmatisch gesehen könnte die Menge auf einen einzigen Prädiktor reduziert werden, ohne merklich an Prognosegenauigkeit zu verlieren. Die Menge der komplementären Variablen hingegen erklärt X fast vollständig. Obwohl die Variablen eine gewisse Redundanz aufweisen, wären immer noch zwei Prädiktorvariablen erforderlich, um die Prognosegenauigkeit aufrechtzuerhalten.

Dieses Beispiel zeigt die Wichtigkeit der Verwendung heterogener Prädiktorvariablen bei der Konstruktion von Aggregaten in einem Prognoseparadigma. Auch wenn die Genauigkeit eines Prognoseaggregats tendenziell mit der Zahl der aggregierten Variablen zunimmt, ist es wichtiger, ihre univariaten Beziehungen zum Kriterium und die Wechselbeziehungen zwischen ihnen zu beachten. Alle univariaten Beziehungen sollten stark und alle Wechselbeziehungen schwach ausgeprägt sein. Wenn eine dieser zwei Bedingungen nicht erfüllt ist, lässt sich die Genauigkeit des Aggregats durch zusätzliche Variablen nicht weiter steigern.

Composite Indicators sind im Prognoseparadigma wesentlich verbreiteter als im Messparadigma. Das Übergewicht des Prognoseparadigmas ist wohl die Folge zweier Faktoren: der leichteren Kommunizierbarkeit von Modellen mit möglichst wenigen Variablen und der intuitiven Anziehungskraft der Prognose. Da Prognosemodelle durch redundante Prädiktoren nicht verbessert werden und sie daher weniger Variablen erfordern, können ihre Struktur und Zusammensetzung einem Laienpublikum leichter vermittelt werden. Außerdem verwenden Prognosen eine vertraute logische Struktur nach dem Motto „wenn A und/oder B, dann C“. Daher können Fachexperten den Variablen auch ohne statistische Methoden, sondern einfach durch Kausalüberlegungen und Vergleiche, Aggregationsgewichte zuweisen. So ist es bei einer der verbreitetsten Anwendungen dieses Paradigmas ein Experten- oder Stakeholder-Panel, das die Wichtigkeit jeder Variablen beurteilt und ihr ein entsprechendes Gewicht zuweist.

Die Hauptschwäche des Prognoseparadigmas liegt darin, dass seine intuitive Anziehungskraft von einer Versuchung zu methodologischen Unsauberkeiten begleitet ist. Im verbreiteten Fall panelbasierter Gewichte müssen die Panelmitglieder die Wichtigkeit beurteilen, was die Existenz eines gemeinsamen Kriteriums impliziert. Doch „Wichtigkeit“ kann je nach den Interessen des einzelnen Panelmitglieds unterschiedlich definiert werden. Politiker erachten vielleicht diejenigen Variablen als wichtig, die die öffentliche Meinung beeinflussen, Wissenschaftler gewichten möglicherweise forschungsrelevante Variablen höher und Sozialaktivisten weisen unter Umständen Variablen ein höheres Gewicht zu, die die Lage an den Rand gedrängter Minderheiten beschreiben. Tatsächlich wird mit zunehmender Heterogenität des Panels – die die meisten Composite-Indicator-Methodologien befürworten – die Definition der Kriteriumsvariablen immer mangelhafter. Dies wiederum verringert die Genauigkeit des Aggregats. Die konzeptionelle Unklarheit vieler Composite-Indizes ist nicht dem Prognoseparadigma an sich anzulasten, sondern seiner falschen Anwendung. Damit das Prognoseparadigma genaue Aggregate liefert, die valide Rückschlüsse erlauben, müssen die Aggregationsmethodologien wohldefinierte und transparente Kriterien haben.

Zu den statistischen Methoden, die sich explizit auf das Prognoseparadigma stützen, gehören Modellierungstechniken wie Regression und Klassifikation. Ungenauigkeiten bei der Definition von Kriterien sind in rein statistischen Modellen weniger häufig, da das Kriterium dort im Allgemeinen eine abhängige Variable ist, die explizit definiert und gemessen werden muss. Die statistische Schätzung eines Aggregationsmodells nutzt einen speziellen Datensatz, in dem sowohl die Prädiktor- als auch die Kriteriumsvariablen verfügbar sind. Mithilfe dieses Datensatzes wird das Aggregationsverfahren trainiert, d.h. die Aggregationsgewichte werden so bestimmt, dass das Aggregat die Kriterien bestmöglich approximiert. Die Gewichte werden dann allgemein auch außerhalb des Trainingsdatensatzes unter der Annahme benutzt, dass das Aggregat die Kriterien auf Basis der verfügbaren Variablen bestmöglich approximiert.

2.3 Synthese der Paradigmen

Die Konstruktion eines Composite-Index' ist konzeptionell eher eine Messaufgabe, da das Composite eine normative Idee oder Perspektive (z.B. den Zustand einer Wirtschaft, das Wohlergehen von Gesellschaften usw.) abbilden muss. Doch viele Composite-Indizes beinhalten das Messparadigma innerhalb eines Prognoseparadigmas. Beispielsweise könnten Politiker die Zufriedenheit einer Wahlbevölkerung prognostizieren wollen, aber nur am Beitrag spezieller, politisch steuerbarer Einflussfaktoren auf diese Zufriedenheit interessiert sein, die mangelhaft definiert sind. Solche Faktoren können die Wahrnehmung der wirtschaftlichen Chancengleichheit, der sozialen Gerechtigkeit und der Qualität des Gesundheitswesens sein. Jeder dieser Faktoren ist ein Messkonstrukt, das die Anwendung einer Messmethodologie erfordert, um genaue Daten zu erzeugen, doch um sie zu aggregieren, muss der individuelle Beitrag jedes Konstrukts zum Kriterium Zufriedenheit geschätzt werden, das ebenfalls ein Messkonstrukt sein kann.

Ein gemeinsamer Mangel von Methodologien zur Erstellung von Composite Indicators ist die Nichtberücksichtigung des Messparadigmas. Oft werden beobachtete Variablen mit den Konstrukten gleichgesetzt, die sie abbilden sollen. So werden beispielsweise Absolventenzahlen weiterführender Schulen als Qualität des öffentlichen Schulwesens interpretiert, Krankenhauswartezeiten als Zulänglichkeit der Gesundheitsversorgung und so weiter. Doch die Nichtberücksichtigung des Messparadigmas bei der Composite-Konstruktion gefährdet das Anführen schlüssiger Validitätsargumente bei der Kommunikation des Composites. Oft werden Variablen wegen ihres individuellen Schlagzeilenwerts oder ihrer politischen Attraktivität anstatt aus streng methodologischen Gründen in Composites aufgenommen. Als Folge davon wird ein reines Prognosemodell leicht durch Redundanz verschlechtert. Wenn diese Redundanz nicht methodologisch behandelt wird, müssen die Variablen im Composite Indicator entweder willkürlich gewichtet werden, wodurch Dimensionen, die auf der Zahl der sie repräsentierenden Variablen anstatt auf ihrem prognostischen Wert beruhen, systematisch über- oder unterrepräsentiert werden, oder der kommunikative Wert dieser redundanten Indikatoren geht verloren, weil ihnen allesamt nur ein vernachlässigbares Aggregationsgewicht zugeordnet wird.

Um die Beiträge beider Paradigmen zum Konstruktionsprozess zu maximieren, wendet der DLA eine neunstufige Methodologie an.

1. Definition des Kriteriums,
2. Definition der Bandbreite der Prädiktoren,
3. Identifikation von Variablen für das Kriterium,
4. Identifikation von Variablen für jeden Prädiktor,
5. Anwendung eines Messmodells zur Erstellung einer operationalen Definition des Kriteriums,
6. Anwendung von Messmodellen zur Erstellung einer operationalen Definition der Prädiktoren,
7. Verwendung der Prädiktoren zur Schätzung des Kriteriums in einem Prognosemodell,
8. Kommunikation des Composites als Aggregation der Prädiktoren mit den Parametern des im vorherigen Schritt definierten Modells,
9. Feststellung und Prüfung der Validität möglicher Rückschlüsse mit Korrekturen an den Methoden, falls erforderlich.

Es gibt mehrere qualitative und quantitative Methoden, die dieser Methodologie gerecht werden können. Die Beiträge informierter Stakeholder liefern die Anfangsspezifikationen in den Schritten 1 bis 4. Diese Anfangsspezifikationen lassen sich jedoch mittels qualitativer oder quantitativer Methoden verfeinern oder zur Berücksichtigung weiterer Informationen modifizieren. Die Schritte 5 bis 8 sind im Wesentlichen quantitativer Art. Allerdings sind Iterationen der zwei Paradigmen erforderlich, um Variablen mit unerwünschten statistischen Eigenschaften (z.B. Varianz oder Kovarianz von null, hohe Schiefe) zu identifizieren und zu entfernen.

Der letzte Schritt des Prozesses unterscheidet sich konzeptionell von den vorangegangenen, da es bei ihm um die Interpretation und nicht um die Konstruktion des Composites geht. Selbst wenn die Methodologie intern robust ist, sind viele Rückschlüsse und Interpretationen aufgrund außerhalb der Methodologie liegender Faktoren möglicherweise nicht valide. Bei diesem Schritt ist Expertenrat erforderlich, um die Vertrauenswürdigkeit des Composites hinsichtlich bestimmter Rückschlüsse zu beurteilen.

3. Instrumentation (Datenaufbereitung)

Der Begriff *Instrumentation* beschreibt die Datenaufbereitungsmethoden, die bei der Konstruktion, nicht bei der Analyse von Daten zum Einsatz kommen. Es gibt drei Typen von Quelldaten:

- Individualbefragungsdaten,
- aggregierte Statistiken und
- geografische Daten.

Unabhängig vom ursprünglichen Datentyp müssen alle Daten einer geografischen Einheit zugeordnet werden, um einen Composite-Index erstellen zu können. Diese geografische Einheit wird die Analyseeinheit für die Konstruktion des Index bilden. Um die in den ursprünglichen Quelldaten enthaltene Information optimal zu nutzen, sollte diese Analyseeinheit die feinste geografische Einheit sein, für die Daten verfügbar sind.

Individualbefragungsdaten sind Antworten von Einzelpersonen, Unternehmen oder Institutionen auf spezielle Fragen. Umfragedaten werden normalerweise in einer nicht zufälligen Stichprobe erhoben. Um Statistiken zu erzeugen, die eine geografische Region genau abbilden, müssen Individualdaten daher im Verhältnis zu ihrer Zufallsstichprobenauswahlwahrscheinlichkeit gewichtet werden. Der größte Vorbehalt gegen die Verwendung von Individualbefragungsdaten besteht darin, dass die Befragungsdichte, und damit der Schätzfehler, je nach geografischer Region variiert. Dies bedeutet im Allgemeinen, dass eine Gleichgewichtung aller geografischen Regionen nicht zu gleichen Beiträgen führen wird, da der wahre Beitrag mit steigendem Schätzfehler sinkt. In bestimmten geografischen Regionen kann der Schätzfehler zu groß für eine brauchbare Schätzung sein. In diesem Fall sollten Daten aus zusätzlichen geografischen Regionen in die Schätzungen einfließen, entweder über ein höheres Aggregationsniveau oder durch ein aufwendigeres Imputationsverfahren.

Aggregierte statistische Daten sind im Allgemeinen aus Umfrage- oder Zensusdaten abgeleitet. Im Falle von Composite Indicators werden diese Quelldaten jedoch gewöhnlich als Primärdaten behandelt. In vielen Fällen sind aggregierte Daten für die gewünschte Analyseeinheit verfügbar. Wo dies nicht der Fall ist, können die Daten auf die richtige Einheit transformiert werden (s. Kap. 3.2). Der DLA-Algorithmus (in Anlehnung an den ELLI-Algorithmus[1]) benötigt eine volle Datenmatrix mit standardisierten Werten, eine Matrix also, die für alle (vergleichbaren geografischen) Analyseeinheiten standardisierte Werte bereitstellt. Dazu werden die Daten auf strukturelle Unterschiede untersucht, nach deren Abschluss die Regionen jeweils eindeutig einem Cluster vergleichbarer Analyseeinheiten angehören. Fehlende Werte werden durch statistische Imputationsverfahren ergänzt. Am Ende der Datenaufbereitungsphase steht eine mit standardisierten Werten gefüllte Matrix, die auf ihre Funktionsfähigkeit bzw. Konsistenz geprüft wird – also das Datenset bzw. die Datenstruktur.

[1] Methodologisches Konzept des ELLI-Index, Bertelsmann Stiftung 2010.

Um eine volle Datenmatrix mit standardisierten Werten für den DLA-Algorithmus zu erhalten, werden folgende Schritte zur Aufbereitung der Daten durchlaufen:

1. Bestimmung der (geografischen) Analyseeinheiten,
2. Standardisierung,
3. Statistische Imputation fehlender Werte (EM-Algorithmus),
4. Finale Strukturierung der Datenmatrix.

3.1 Identifizierung der Analyseeinheiten

Die erhobenen Daten liegen zum Teil auf unterschiedlichen geografischen Ebenen vor (zum Beispiel PISA Daten nur auf Bundesländer-Ebene und Vereinsdichte-Daten nur auf Kreisebene). Zur Konstruktion der Datensets müssen daher die vorliegenden Daten auf die jeweilige geografische Ebene, für den DLA-Index vor allem auf die Kreisebene, angepasst werden, so dass die erhobenen Daten vergleichbar sind (siehe 3.2). Die Analyseeinheiten (Unit of Analysis), z.B. Kreise, bestimmen die Zeilenanzahl M des Datensets.

Zusätzlich wird die vorliegende Datenmatrix auf strukturelle Unterschiede analysiert Für einen Großteil der Daten lässt sich beispielsweise ein Ost-West-Gefälle identifizieren, das es zu berücksichtigen gilt. Die Datenmatrix wird daher in zwei Teilmatrizen unterteilt, auf die bestimmte Analyseschritte des DLA-Algorithmus separat angewendet werden (müssen). Schließlich können in dem Datenset noch einige Datenlücken existieren, die dann durch statistische Imputation mit Hilfe des Expectation Maximisation (EM) Algorithmus (Abschnitt 3.4) unter Berücksichtigung der Ost-West-Strukturunterschiede gefüllt werden.

3.2 Umgang mit Daten unterschiedlicher geografischer Ebenen

Die Ursache für Daten unterschiedlicher geografischer Ebenen kann zum einen inhaltlicher Natur oder durch die Art der Erhebung begründet sein, zum anderen jedoch auch z.B. an Restriktionen hinsichtlich der Veröffentlichung, Gebietsreformen oder einfach an fehlender/lückenhafter statistischer Erfassung liegen. Die Daten können daher nicht nur zwischen den Kennzahlen, sondern in vereinzelten Fällen auch innerhalb einer Kennzahl auf unterschiedlicher geografischer Ebene vorliegen. Neben der im Fokus des DLA-Index stehenden Ebene der Kreise/kreisfreien Städte und der bereits genannten Bundeslandebene gibt es zusätzlich die Regierungsbezirksebene, Raumordnungsebene, Ebene der Regionalen Anpassungsschichten und die Arbeitsagenturbezirksebene als der Kreisebene übergeordnete Einheiten sowie die Gemeindeebene und die PLZ-Ebene als untergeordnete Einheiten. Das heißt, die nicht auf der Ebene der Kreise und kreisfreien Städte existierenden Daten werden, soweit das durch die Datenlage möglich ist, entweder durch verfügbare Daten einer übergeordneten geografischen Einheit ersetzt oder durch die Summe der Werte der eines Kreises/einer kreisfreien Stadt zugehörigen untergeordneten Ebenen berechnet. Letzteres ist nur möglich, wenn die Rohdaten als absolute Zahlen vorliegen. Erst im gewünschten Aggregationszustand, der Kreisebene, werden die Daten in Relation zur Bevölkerung oder ähnlichem Maß gesetzt, um die Vergleichbarkeit zu anderen Regionen innerhalb einer Kennzahl herzustellen.

3.3 Standardisierung der Daten

Aufgrund unterschiedlicher Mess-Skalen besteht die Gefahr, dass Beiträge bestimmter Variablen über- oder unterbewertet werden. Die Standardisierung ist die einzige Möglichkeit, Variablen mit unterschiedlichen Maßeinheiten vergleichbar zu machen. Außerdem werden dadurch die Korrelationsrechnung und die Rechenschritte der Faktorenanalyse im Rahmen des DLA-Algorithmus erleichtert und die Interpretation der Ergebnisse vereinfacht. Bevor die fehlenden Werte durch statistische Imputation eingesetzt werden, wie im Abschnitt 3.4 beschrieben, sind die Ausgangsdaten deshalb zu standardisieren.

Sei $X = (x_{ij})_{\substack{i=1,..,M \\ j=1,..,N}}$ die Matrix mit rohen Ausgangsdaten. Dabei ist M die Anzahl der (geografischen) Analyseeinheiten und N die Anzahl der Variablen. Die z-standardisierte Datenmatrix $Z = (z_{ij})_{\substack{i=1,..,M \\ j=1,..,N}}$ hat die Einträge

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \text{mit} \quad \bar{x}_j = \sum_{i=1}^{M} x_{ij} \quad \text{und} \quad s_j = \sqrt{\sum_{i=1}^{M} x_{ij}^2 - \bar{x}_j} \; .$$

3.4 Statistische Imputation

Nach der Standardisierung der Daten werden mit dem Expectation-Maximisation (EM) Algorithmus die fehlenden Werte (max. 25% pro Variable) geschätzt und in die Datenmatrix eingetragen. Die Imputation erfolgt dabei nach Ost und West getrennt, mithilfe einer Dummy-Variable, die die Gruppenzugehörigkeit bestimmt. Das Verfahren basiert auf der Unabhängigkeit der Modell-Parameter und der fehlenden Werte. EM ist das Akronym für Erwartung und Maximierung. Die E- und M-Schritte werden iterativ aufgerufen, bis das Verfahren konvergiert. Als Konvergenzkriterien werden der Wert ε als Genauigkeit der Annäherung (aktuell auf 0.001 gesetzt) sowie die maximale Anzahl der Iterationsschritte (aktuell auf 25 gesetzt) definiert.

Zunächst werden in der Datenmatrix alle fehlenden Werte x_{ij} mit dem jeweiligen Mittelwert der Spalte j (der Variable X_j) ersetzt.

Für jede Variable X_{j_0}, $j_0 = 1,..,N$, werden nun folgende Schritte durchgeführt:

1. Berechne Korrelationen $c_{j_0 k}$, $k = 1,..,N$, mit allen anderen Variablen und bilde eine Matrix

$$C_{j_0} = \begin{pmatrix} c_{j_0 1} & 0 & \cdots & 0 \\ 0 & c_{j_0 2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & c_{j_0 N} \end{pmatrix}$$

2. Für jeden als fehlend markierten Wert $x_{i_0 j_0}$ der laufenden Variable:
   a. Berechne $H = X \cdot C_{j_0}$, $H = (h_{ij})_{\substack{i=1,..,M \\ j=1,..,N}}$ mit $h_{ij} = x_{ij_0} \cdot c_{j_0 k}$ und

$$D_{i_0} = \begin{pmatrix} d_{i_0 1} & 0 & \cdots & 0 \\ 0 & d_{i_0 2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & d_{i_0 M} \end{pmatrix} \quad \text{mit} \quad d_{i_0 m} = \frac{1}{\sqrt{\sum_{j=1}^{N} x_{i_0 j} \cdot h_{mj}}}, \; m = 1,..,M.$$

   $d_{i_0 m}$ bezeichnet den inversen euklidischen Abstand der i_0-ten Zeile der Matrix X zur Zeile m der Matrix H, in der jede Spalte des Zeilenpaares mit deren Korrelation zu der laufenden Variable gewichtet ist.
   b. Ersetze den als fehlend markierten Wert $x_{i_0 j_0}$ durch den gewichteten Mittelwert des Vektors X_{j_0}, $\frac{\sum_{m=1}^{M} d_{i_0,m} \cdot x_{m,i_0}}{\sum_{m=1}^{M} d_{i_0,m}}$
   c. Ist die Differenz zwischen dem neu eingesetzten Wert und dem vorherigen Wert kleiner als ε, markiere den Wert $x_{i_0 j_0}$ als nicht-fehlend. Ansonsten wiederhole den Schritt 1, bis die maximale Anzahl der Iterationsschritte erreicht ist.

3.5 Finale Strukturierung des Datensets/der Matrix

Am Ende muss die nun mit standardisierten Werten gefüllte Matrix – also das Datenset bzw. die Datenstruktur – noch durch multivariate Verfahren auf ihre Funktionsfähigkeit bzw. Konsistenz geprüft werden. Dabei fallen mitunter Variablen aus der Matrix heraus, insbesondere solche die

1. nach der Imputation keine echte Varianz besitzen,
2. deren Korrelation mit einer anderen Variablen größer oder gleich 0,90 ist,
3. keine signifikante Korrelation mit den Outcome-Variablen aufweisen.

Expertenwissen und Fachkenntnis sind bei der Prüfung der Variablen in Bezug auf ihre Konsistenz und Funktionsfähigkeit erforderlich. Ebenso ist dieses Expertenwissen erforderlich bei der Prüfung der Variablen auf ihre Verwendbarkeit als Manifestationen der vorgesehenen Indikatoren, Lerndimensionen oder Konstrukte. Eine Überprüfung durch Sachverständige ist erforderlich, um festzustellen, ob die in dem DLA-Datenset identifizierten Beziehungen mit den vorab theoretisch festgelegten Beziehungen konsistent sind. Inkonsistenzen können auftreten, wenn durch Aggregierung bzw. Disaggregierung der DLA-Daten oder durch Datenimputation die Assoziation zwischen einer Variable und ihrem zugeordneten Konstrukt unterbrochen wird. Falls notwendig, werden Variablen aus der Datenmatrix entfernt oder ihre vorgesehene Zuordnung zu einer Dimension und den Konstrukten geändert.

Die genannten Schritte für die finale Strukturierung der Datenmatrix sind nicht algorithmisch in der ELLI-Plattform implementiert worden, auf welcher auch der DLA-Index zu finden ist, sondern wer-

den parallel zu den anderen automatisierten Schritten der Datenaufbereitung durch ein Expertengremium vorgenommen.

Am Ende steht eine volle Datenmatrix, die nachfolgend wieder mit X bezeichnet wird und die für statistische Analysen bzw. den DLA-Algorithmus genutzt werden kann. Außerdem sind alle Variablen einer Lerndimension und zusätzlich einem Konstrukt oder mehreren Konstrukten zugeordnet.

3.6 Anmerkungen zur Datenqualität

Die Validität der Interpretation der Ergebnisse des DLA-Index basiert darauf, dass die bei der Erstellung dieses Index' verwendeten Variablen nur mit vernachlässigbaren zufälligen oder systematischen Fehlern behaftet sind. Die bei der Aufbereitung der DLA-Daten durchlaufenen Schritte wurden gewählt, um fehlende Werte zu ersetzen und statistische Verzerrungen in den Daten zu reduzieren. Mangelhafte Rohdaten lassen sich jedoch nicht auf methodologischer Ebene ersetzen oder korrigieren. Schließlich sollte bei der Erklärung der Sensitivität eines Composite-Index berücksichtigt werden, wie sich verschiedene Formen der Instrumentation (einschließlich des Datentyps, der Berichtseinheit, des Schätzfehlers) auf die Variabilität der Schlussfolgerungen auf der Ebene der methodologischen Annahmen auswirken.

4. Methoden

Der DLA-Algorithmus, welcher in sehr enger Anlehnung an den ELLI-Algorithmus für Europa entwickelt wurde[2], ist eine Abfolge von statistischen Verfahren, die eine Menge von Quelldaten in einen zusammengesetzten Index (oder mehrere Indizes) transformieren. Dabei durchläuft er sowohl das Bemessungs- als auch das Vorhersageparadigma in mehreren Iterationen. Sowohl für die Variablen in jeder Lerndimension als auch für die Outcome-Variablen werden Faktorenanalysen durchgeführt. Auf den so berechneten Faktoren werden Regressionsverfahren angewandt, um die Gewichte der Faktoren zur Berechnung der Indexwerte für die verschiedenen Lerndimensionen zu erhalten. Mit der Hilfe der Faktorladungen, der Regressionskoeffizienten und Bestimmtheitsmaße, die zu den einzelnen Lerndimensionen berechnet worden sind, werden anschließend die finalen Gewichte der Variablen zur Berechnung der Composite-Index-Werte und der Konstrukt-Werte bestimmt.

4.1 Faktorenanalyse der Outcome-Kennzahlen

Die Faktorenanalyse (FA) wird eingesetzt, um eine große Menge von Variablen gemäß ihrer korrelativen Beziehungen in voneinander unabhängige Gruppen zu ordnen. Dies führt zu einer Vereinfachung des Modells durch die Reduktion der Variablen auf komplexere Hintergrundfaktoren.

In diesem Abschnitt sei $X = (x_{ij})_{\substack{i=1,..,M \\ j=1,..,N}}$ die Matrix mit Spalten aus Outcome-Variablen mit eingesetzten fehlenden Werten und standardisiert wie im Kapitel 3 beschrieben. Dabei ist M die Anzahl der (geografischen) Analyseeinheiten und N sei die Anzahl der Outcome-Variablen.

Die Grundannahme hinter der gesamten Faktorenanalyse ist, dass jeder Wert einer Ausgangsvariable sich als Linearkombination von Q hypothetischen Faktoren beschreiben lässt:

$$x_{ij} = f_{i1}a_{j1} + f_{i2}a_{j2} + ... + f_{iQ}a_{jQ} = \sum_{q=1}^{Q} f_{iq}a_{jq} ,$$

dabei ist f_{iq} der sog. Faktorwert, der angibt, wie stark die in dem Faktor q zusammengefassten Merkmale bei der Einheit i ausgeprägt sind, und a_{jq} die sog. Faktorladung, die angibt, wie gut eine Variable zu einer Variablengruppe passt. Die Faktorladung entspricht der Korrelation zwischen der Variable j und dem Faktor q.

In Matrixschreibweise lässt sich die Gleichung (0.1) auf folgende Weise darstellen:

$$X = F \cdot A^t$$

[2] Methodologisches Konzept des ELLI-Index, Bertelsmann Stiftung 2010.

Die obige Gleichung hat theoretisch unendlich viele Lösungen. Wir suchen nach einer speziellen Lösung mit folgenden Eigenschaften für die Faktoren:

1. Die Faktoren sind wechselseitig voneinander unabhängig.
2. Die Faktoren repräsentieren die zu messenden latenten Variablen.
3. Die Faktoren erklären sukzessiv maximale Varianz.

Im DLA-Berechnungsverfahren wird für die Outcome-Variablen nur ein Faktor extrahiert (Abbildung). Die einzelnen Schritte werden im Folgenden beschrieben. Im **ersten** Schritt wird für alle Outcome-Variablen die sogenannte Korrelationsmatrix erstellt. Der Korrelationsmatrix lässt sich entnehmen, welche Variablen in der weiteren Analyse unberücksichtigt bleiben sollen, da sie mit den übrigen Variablen nur minimal korrelieren und somit keinem gemeinsamen Hintergrundfaktor zugeordnet werden können. Im **zweiten** Schritt findet die Faktorextraktion statt. Im **dritten** Schritt wird ermittelt, welche Werte die untersuchten Variablen hinsichtlich des extrahierten Faktors annehmen.

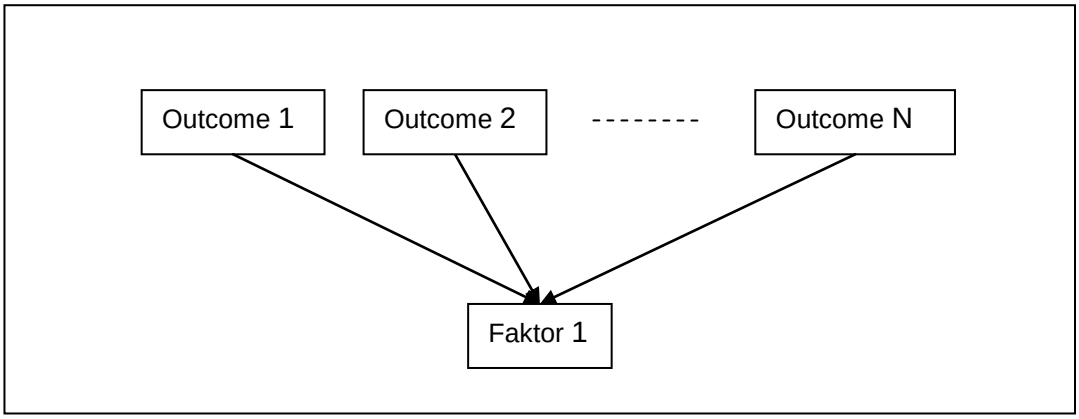

Abbildung 4. Faktorenanalyse der Outcome-Variablen

4.1.1 Erstellung der Korrelationsmatrix

Faktoren sind als „hinter den Variablen stehende Größen“ zu begreifen. Sie repräsentieren damit den Zusammenhang zwischen den verschiedenen Ausgangsvariablen. Dieser Zusammenhang kann durch eine Korrelationsrechnung messbar gemacht werden. Korrelationen zeigen den Grad des Zusammenhangs zwischen Variablen an, wodurch diese im Sinne der Faktorenanalyse als „bündelungsfähig“ oder „nicht bündelungsfähig“ identifiziert werden können.

Im ersten Schritt werden die Korrelationen der Outcome-Variablen berechnet. Die Korrelationen werden in einer Korrelationsmatrix $C = (c_{jk})_{\substack{j=1,..,N \\ k=1,..,N}}$ zusammengefasst.

$$C = X^t X, \quad \text{bzw.} \quad c_{jk} = (x_{ij})^t_{i=1,..,M} \cdot (x_{ik})_{i=1,..,M},$$

4.1.2 Faktorextraktion

Im nächsten Schritt werden die Faktoren, in diesem Fall nur ein Faktor, der sog. Outcome, extrahiert. Hinter der Ziehung der Faktoren stehen folgende Überlegungen. Summiert man die quadrierten Ladungen der Variablen auf einem Faktor q, ergibt sich die Varianz λ_q, die durch diesen Faktor aufgeklärt wird. Je höher die Korrelation der Variablen, desto größer die Varianz λ_q desjenigen Faktors q, der die meiste Varianz erklärt.

Die Varianzen λ_q sind identisch mit den Eigenwerten der Korrelationsmatrix C. Die zu den Eigenwerten zugehörigen Eigenvektoren, die zusätzlich bestimmte Auswahlkriterien erfüllen, bestimmen die gesuchten Faktoren. Ordnet man die einzelnen Eigenwerte der Größe nach, ergeben sich die mit diesen Werten assoziierten Faktoren, die sukzessiv die maximale Varianz erklären.

Sei $\mathrm{EW} = \lambda_1, \lambda_2, ..., \lambda_N$ die Menge der Eigenwerte der Korrelationsmatrix C.

Sei $\lambda = \max(\mathrm{EW})$ der maximale Eigenwert und V der zugehörige Eigenvektor.

Als nächstes wird die Kommunalität der Variablen berechnet. Dazu werden die Werte des auf Länge 1 normierten Eigenvektors quadriert. Die Variablen mit Kommunalität <0.01 werden aus der Variablenliste entfernt und die Matrix X neu bestimmt.

4.1.3 Überprüfung der Korrelationen auf inhaltliche Konsistenz

In diesem zusätzlichen, wichtigen Schritt werden die Korrelationen, sprich Faktorladungen, dahingehend geprüft, ob die von Experten im Vorwege festgelegte erwartete Wirkungsrichtung zwischen einer Variable und dem extrahierten Faktor (Outcome) bestätigt werden kann und zusätzlich von relevanter Bedeutung ist, d.h. wenn - in diesem Fall - eine Korrelation von mindestens 0,5 auf einem Signifikanzniveau von 0,05 vorliegt. Ist eine der beiden Bedingungen nicht erfüllt, wird die Variable aus dem relevanten Set eliminiert.[3]

Die Schritte 1 (Erstellung der Korrelationsmatrix), 2 (Faktorextraktion) und 3 (Überprüfung der Korrelationen) werden solange wiederholt, bis die Liste der Variablen sich nicht mehr ändert.

4.1.4 Bestimmung der Faktorwerte

Im letzten Schritt der Faktorenanalyse stellt sich noch die Frage, welche Werte die untersuchten Analyseeinheiten bezüglich des Faktors annehmen. Die Matrix F der Faktorwerte (factor scores) wird folgendermaßen berechnet:

$$\mathrm{F} = \mathrm{X} \cdot (\mathrm{C}^{-1}\mathrm{V})$$

[3] In Ausnahmefällen kann eine Variable auch aufgrund ihrer zentralen inhaltlichen Bedeutung als relevant eingestuft werden, ohne dass sie das Korrelationskriterium erfüllt.

4.2 Erste Faktorenanalyse und Korrelationsanalyse der Lerndimensionen-Kennzahlen

Bevor die Korrelationsanalyse zur Überprüfung der inhaltlichen Konsistenz zwischen den Kennzahlen und dem Outcome-Faktor durchgeführt wird, erfolgt für jede Lerndimension zunächst eine Faktorenanalyse nach im vorherigen und auch im folgenden Abschnitt detailliert beschriebenem Verfahren. Die Anzahl der extrahierten Faktoren wird hier jedoch nicht vorher festgelegt, sondern modellseitig bestimmt. Um die strukturellen Unterschiede zwischen Ost und West adäquat zu berücksichtigen, wird die Faktorenanalyse für beide Regionen separat implementiert. Diese erste Analyse dient dazu, diejenigen Variablen zu identifizieren, die zu mindestens einem der extrahierten Faktoren einen relevanten Zusammenhang aufweisen, was erfüllt ist, wenn eine der Faktorladungen einen Wert von größer oder gleich 0,4 besitzt.[4] Damit wird sichergestellt, dass eine Variable grundsätzlich die Voraussetzung für den weiteren Analyseprozess erfüllt. Die Faktorenanalyse wird in mehreren Iterationen (hier: 25) durchgeführt, um schließlich eine stabile Lösung zu erhalten. Das Ergebnis, d.h. das nach Durchlauf aller Iterationen resultierende Set relevanter Variablen, kann dabei für Ost und West unterschiedlich ausfallen.

Die daran anschließende Korrelationsanalyse der durch die Faktorenanalyse selektierten Variablen einer Lerndimension mit dem Outcome-Faktor wird ebenfalls für beide Regionen getrennt durchgeführt. Um ein für Ost und West gleichermaßen relevantes und inhaltlich konsistentes Variablenset zu generieren, werden die Korrelationsergebnisse beider Regionen pro Variable (aus der Schnittmenge) einander gegenübergestellt und mittels folgender Regel untersucht: Wenn mindestens einer der beiden Korrelationskoeffizienten die zuvor von Experten definierte Wirkungsrichtung auf einem Signifikanzniveau von kleiner als 0,1 aufweist und der andere Koeffizient im Fall nicht-erwarteter Richtung nur ein Signifikanzniveau von größer als 0,5 besitzt, d.h. stark insignifikant/inhaltlich von keiner Bedeutung ist, wird die Variable für das finale Set ausgewählt. Alle so selektierten und damit modelltechnisch und inhaltlich geeigneten Variablen werden dann noch einmal durch Experten auf mögliche inhaltliche Redundanz untersucht und gegebenenfalls eliminiert. Die darauffolgende weitere Faktorenanalyse (siehe 4.3) verwendet ausschließlich das aus dem oben beschriebenen Analyseprozess resultierende Variablenset pro Lerndimension.

4.3 Faktorenanalyse der Lerndimensionen-Kennzahlen

Auch diese Faktorenanalyse wird für Ost und West separat und wiederum für jede der Lerndimensionen einzeln durchgeführt.

Wir bezeichnen mit $X = (x_{ij})_{\substack{i=1,..,M \\ j=1,..,N}}$ die Matrix mit Spalten entsprechend den Lerndimensionen. M sei die Anzahl der Analyseeinheiten und N die Anzahl der Variablen. Gesucht ist die Lösung der Gleichung

$$X = F \cdot A^t,$$

[4] In Ausnahmefällen kann eine Variable auch aufgrund ihrer zentralen inhaltlichen Bedeutung als relevant eingestuft werden, ohne dass sie das Auswahlkriterium erfüllt.

wobei F die Matrix der Faktorwerte und A die Matrix der Faktorladungen bezeichnen.

Geometrisch gesehen lässt sich das Ziel der Faktorenanalyse folgendermaßen veranschaulichen. Die Zeilen der Matrix X lassen sich als Koordinaten einzelner Punkte in einem Koordinatensystem interpretieren, in dem die Variablen die Koordinatenachsen bilden. Das Ziel der Faktorenanalyse ist, die Anzahl der Achsen (Dimension des Raumes) zu reduzieren und das neue Koordinatensystem so zu drehen, dass

1. die Korrelation zwischen je zwei neuen Achsen Null wird und
2. die Punkte auf der 1. neuen Achse maximalen Anteil an der Gesamtvarianz haben, auf der 2. Achse die maximale Restvarianz aufweisen, die durch die erste Achse nicht aufgeklärt wird, und die dritte Achse den maximalen Anteil von der restlichen Varianz aufklärt, die durch die beiden ersten Achsen nicht erfasst wird usw.

Dieses Vorgehen bezeichnet man als eine sukzessiv varianzmaximierende, orthogonale Rotationstransformation.

In folgenden Unterabschnitten werden die einzelnen Schritte der Faktorenanalyse der Lerndimensionen beschrieben.

Im **ersten** Schritt wird für alle in 4.2 final selektierten Variablen die Korrelationsmatrix erstellt. Im **zweiten** Schritt findet die Faktorextraktion statt. Aufgrund verschiedener statistischer Kennzahlen kann in dieser Stufe entschieden werden, wie viele Faktoren ausgewählt werden. Die im zweiten Schritt extrahierten Faktoren sind in der Regel nur sehr schwer oder auch gar nicht zu interpretieren. Um die Ergebnisinterpretation zu erleichtern, werden die Faktoren im **dritten** Schritt einer speziellen Transformation unterzogen, die als Faktorrotation bezeichnet wird. Im **vierten** und letzten Schritt wird ermittelt, welche Werte die untersuchten Variablen hinsichtlich der extrahierten und rotierten Faktoren annehmen. Dies dient der inhaltlichen Interpretation der Faktoren, insbesondere der Klärung, welche Variablen welchen Faktoren zuzuordnen sind.

4.3.1 Variablenselektion und Erstellung der Korrelationsmatrix

Zunächst werden alle Variablen, deren quadrierte Korrelation mit dem Outcome-Faktor weniger als 0,001 beträgt, aus der Variablenliste entfernt. Die restlichen Variablen bilden die Spalten der Datenmatrix, die wir wiederum mit $X = (x_{ij})_{\substack{i=1,..,M \\ j=1,..,N}}$ bezeichnen wollen.

Anschließend wird die Korrelationsmatrix $C = (c_{jk})_{\substack{j=1,..,N \\ k=1,..,N}}$ nach Abschnitt 4.1.1 berechnet.

4.3.2 Faktorextraktion

Jeder Eigenwert der Korrelationsmatrix gibt für den mit diesem Eigenwert assoziierten Faktor die Varianz an, die sich als Summe der quadrierten Ladungen der Variablen auf diesen Faktor ergibt. Großer Eigenwert bedeutet somit einen großen Anteil an Varianz, die durch den zugehörigen Faktor aufgeklärt wird.

Sei also EW die Menge der Eigenwerte der Korrelationsmatrix C, die beginnend mit dem größten Wert der Größe nach geordnet sind:

$$EW = \lambda_1, \lambda_2, ..., \lambda_N \text{ , mit } \lambda_j > \lambda_k \text{ für } j < k \text{ , } j, k = 1, ..., N$$

Aus dieser Menge werden die ersten Q Eigenwerte nach folgenden Kriterien ausgewählt:

1. Auswahl aller Eigenwerte, die größer oder gleich 1 sind
2. die Liste der ausgewählten Eigenwerte besteht aus mindestens 3 Elementen (vorausgesetzt, dass mindestens drei Variablen vorhanden sind).

Die zu den ausgewählten Eigenwerten zugehörigen Eigenvektoren V_q, $q = 1, .., Q$, bestimmen die gesuchten Faktoren für die weitere Analyse, bzw. geometrisch interpretiert bilden die Achsen des neuen Koordinatensystems. Die Eigenvektoren V_q, $q = 1, .., Q$, werden spaltenweise zu einer Matrix $V = (v_{jq})_{\substack{j=1,..,N \\ q=1,..,Q}}$ zusammengefasst.

4.3.3 Faktorrotation

Die beobachteten Variablen sind Ausdruck komplexer Hintergrundfaktoren. Die Beziehungen der einzelnen Variablen zu diesen Hintergrundfaktoren zeigt sich dann an den jeweiligen Faktorladungen, wobei große Faktorladungen eine hohe Bedeutung und niedrige Faktorladungen eine geringe Bedeutung des Faktors für die jeweilige Variable anzeigen. Korreliert ein Faktor dabei mehr oder weniger stark mit vielen Variablen, wie in Abbildung 5 dargestellt, ist er sehr schwer zu interpretieren.

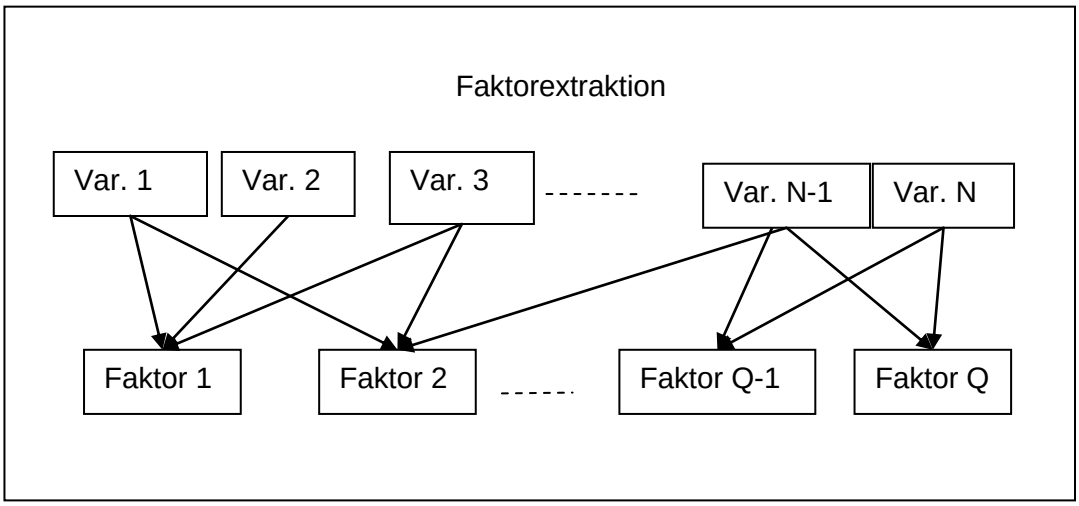

Abbildung 5. Faktorenanalyse vor der Rotation

Die Pfeile zeigen starke Korrelation der Variablen mit den jeweiligen Faktoren an.

Die Faktoren können aber einer als Rotation bezeichneten Transformation unterworfen werden, die die Zuordnung der Variablen zu den Hintergrundfaktoren erheblich erleichtert (s. Abbildung 6).

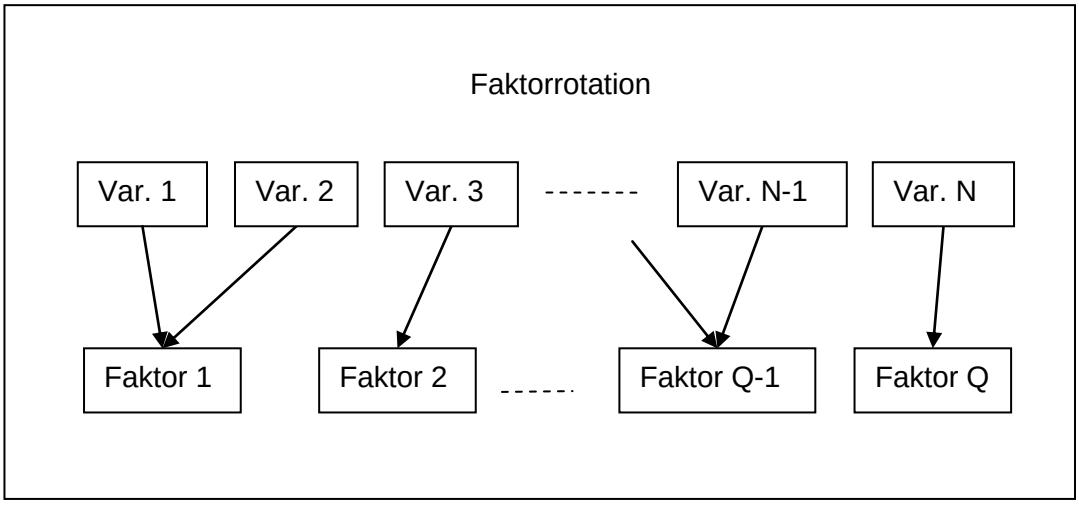

Abbildung 6. Faktorenanalyse nach der Rotation

Die Pfeile zeigen starke Korrelation der Variablen mit den jeweiligen Faktoren an.

Geometrisch veranschaulicht rotiert man die Koordinatenachsen so in ihrem Ursprung, dass sich die Faktorladungen besser auf die Faktoren verteilen. Im DLA-Verfahren wird eine orthogonale (rechtwinklige) Rotation verwendet, die unter der Bezeichnung Varimax-Methode bekannt ist. Bei der orthogonalen Rotation geht man davon aus, dass die Faktoren nicht untereinander korrelieren und ihre Vektoren daher stets senkrecht zueinander stehen. Sämtliche Faktorachsen bleiben daher während der Rotation ebenfalls im rechten Winkel zueinander. Bei der Varimax-Methode werden die Achsen dabei so rotiert, dass sich die Anzahl der Variablen mit hohen multiplen Faktorladungen reduziert.

Für den Rotationsalgorithmus werden zunächst die Spalten der Matrix V auf Länge 1 normiert:

Sei $\Gamma = \begin{pmatrix} \gamma_1 & 0 & \cdots & 0 \\ 0 & \gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_Q \end{pmatrix}$ mit $\gamma_q = \sqrt{\sum_{j=1}^{N} v_{jq}^2}$ (Länge des Spaltenvektors der Matrix V)

Setze $\tilde{V} = V \cdot \Gamma^{-1}$. Somit ist $\tilde{V} = \tilde{v}_{jq \;\; \substack{j=1,..,N \\ q=1,..,Q}}$ mit $\tilde{v}_{jq} = \frac{v_{jq}}{\gamma_q}$.

Als nächstes wird der Varimax-Index einer Matrix $V = v_{jq \;\; \substack{j=1,..,N \\ q=1,..,Q}}$ definiert:

$$\text{Varimax-Index} = \sum_{q=1}^{Q} s_q^2 \quad \text{mit} \quad s_q^2 = \frac{1}{N}\sum_{j=1}^{N} \tilde{v}_{jq}^4 - \frac{1}{N^2}\left(\sum_{j=1}^{N} \tilde{v}_{jq}^2\right)^2$$

Mit dem Varimax-Index wird die gesamte Varianz der quadrierten Ladungen auf den Faktoren berechnet. Da die Varianz der quadrierten Ladungen möglichst groß werden soll, wird nach einer Rotationslösung gesucht, die den Varimax-Index maximiert.

Beim Ortho-Varimax-Rotationsalgorithmus werden nacheinander alle Paare von Spaltenvektoren der Matrix jeweils so lange rotiert, bis der Varimax-Index maximal wird. Für jedes Spaltenpaar V_p und V_q der Matrix V wird dabei der Varimax-Index der Matrix berechnet, die Spalten um einen bestimmten Winkel θ rotiert und anschließend der Varimax-Index der neu entstandenen Matrix bestimmt. Unterscheidet sich der neue Varimax-Index um kleiner als 0,001 vom vorherigen, geht man zum nächsten Spaltenpaar über. Ansonsten wird der Rotationswinkel θ korrigiert und der Rotationszyklus erneut durchgeführt, solange bis das vordefinierte Abbruchkriterium erfüllt ist.

Rechnerisch erfolgt die Rotation zweier Spalten p und q auf folgende Art:

Man berechne $V_p' \quad V_q' = V_p \quad V_q \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$

und ersetze die Spalten V_p und V_q der Matrix V durch Spalten V_p' und V_q'.

Nach der Faktorrotation wird für jede Variable die Kommunalität, d.h. Summe der quadrierten Ladungen (Zeileneinträge der Matrix), berechnet.

4.3.4 Bestimmung der Faktorwerte

Im Anschluss an die Ziehung und die Rotation der Faktoren wird die Matrix F der Faktorwerte folgendermaßen berechnet:

$$F = X \cdot (C^{-1}V),$$

wobei X die Matrix der Variablen, die nach der Rotation nicht entfernt wurden, und V die Matrix der Eigenvektoren nach der Rotation darstellt.

Nach dieser Berechnung ist die Faktorenanalyse abgeschlossen. Die für die folgenden Schritte des DLA-Algorithmus relevanten Ergebnisse sind die Faktoren, welche als Prädiktorvariablen in die sich anschließende Regressionsanalyse (siehe 4.4) eingehen, und die rotierte Faktorladungsmatrix, welche zur Berechnung der Gewichte der einzelnen Variablen am jeweiligen Index verwendet wird.

4.4 Regressionsanalyse

Die Regressionsanalyse wird eingesetzt, um Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen (den Prädiktorvariablen) festzustellen. Ein Spezialfall von Regressionsmodellen sind lineare Modelle. Man spricht von der multiplen linearen Regression, wenn folgendes Modell gewählt wird:

$$Y = b_0 + b_1 X_1 + ... + b_Q X_Q + \varepsilon$$

Dabei ist ε eine Zufallsvariable und stellt eine Störgröße dar. Im DLA-Algorithmus bilden die Q Varimax-Faktoren aus der Faktorenanalyse (der Variablen einer Lerndimension) die Prädiktorvariablen der dimensionsspezifischen Regressionsanalyse. Die Regressionsanalyse wird wiederum für Ost und West getrennt berechnet. In diesem Abschnitt werden die Faktoren der Lerndimensionen mit X_q bezeichnet. Der Faktor aus der Faktorenanalyse der Outcome-Variablen stellt die abhängige Variable dar und wird im Folgenden mit Y bezeichnet. Mit den M Beobachtungen (M Analyseeinheiten) liegt ein Gleichungssystem vor, das sich in Matrix-Schreibweise folgendermaßen darstellen lässt:

$$Y = XB + \varepsilon \text{ mit } Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} \in \mathbb{R}^M,\ X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1Q} \\ 1 & x_{21} & \dots & x_{2Q} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{M1} & \dots & x_{MQ} \end{pmatrix} \in \mathbb{R}^{M\times(Q+1)},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_M \end{pmatrix} \in \mathbb{R}^M$$

Im multiplen linearen Regressionsmodell können die Regressionskoeffizienten mit Hilfe der Methode der gewichteten kleinsten Quadrate geschätzt werden.

Dazu wird die Summe der Quadrate der Residuen minimiert, also die Summe der Differenzen zwischen $X \cdot B$ und den Messwerten Y. Die Lösung dieses Minimierungsproblems ergibt den Vektor der geschätzten Regressionskoeffizienten:

$$B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_Q \end{pmatrix} = (X^t X)^{-1} X^t Y \quad .$$

Dieser Schätzer ist der beste (erwartungstreu mit kleinster Varianz) lineare unverzerrte Schätzer.

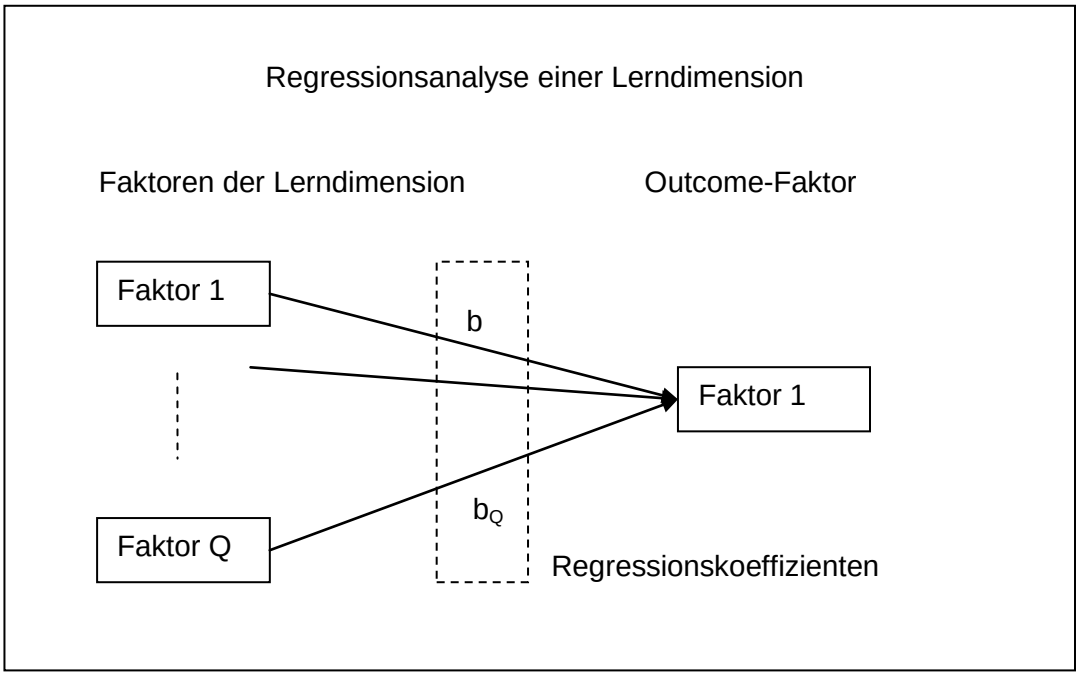

Abbildung 7. Regression

Hat man die Regressionskoeffizienten ermittelt, ist man auch an der Güte dieser Regression interessiert. Häufig wird als Maß für die Güte das **Bestimmtheitsmaß** R^2 verwendet. Das Bestimmtheitsmaß ist ein Maß für den erklärten Anteil der Variabilität (Varianz) der abhängigen Variable Y durch ein statistisches Modell. Indirekt wird damit auch der Zusammenhang zwischen der abhängigen und den unabhängigen Variablen gemessen. Generell gilt, je näher der Wert des Bestimmtheitsmaßes bei 1 liegt, desto größer ist die Güte der Regression. Wenn $\bar{y}$ den Mittelwert

aller Outcome-Faktoren y_i bezeichnet und $\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_Q x_{iQ}$ den geschätzten Regresswert aus dem Regressionsmodell für die Region i, dann ergibt sich

$$R^2 = 1 - \frac{\sum_{i=1}^{M} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{M} (y_i - \bar{y})^2}$$

Bei einer multiplen Regression entspricht das R^2 dem Quadrat des multiplen Korrelationskoeffizienten zwischen der abhängigen und den unabhängigen Variablen, also der Korrelation zwischen Y und $b_1 X_1 + ... + b_Q X_Q$.

4.5 Berechnung der Lerndimension-Indizes

Mit Hilfe der in der Faktoren- und der Regressionsanalyse (pro Lerndimension und Ost bzw. West) gewonnenen rotierten Faktorladungsmatrix und der Regressionskoeffizienten werden die Anteile der einzelnen Kennzahlen am jeweiligen Dimensions-Index berechnet. Dies erfolgt zunächst für Ost und West getrennt.

Es wird angenommen, dass insgesamt $N \in \mathbb{N}$ Kennzahlen und $M \in \mathbb{N}$ Regionen/Analyseeinheiten in die DLA-Gesamtindex-Berechnung eingehen. Dazu wird zunächst eine der vier Lerndimensionen $k \in 1,...,4$ betrachtet: Seien m_k Variablen für die Lerndimension k relevant, nämlich die Variablen $j_{m_1}, ..., j_{m_k} \in 1,...,N$. In der Faktorenanalyse seien Q_k Faktoren bestimmt worden. Seien $b_0^k, ..., b_{Q_k}^k$ die Regressionskoeffizienten der Faktoren und $A^k \in \mathbb{R}^{m_k \times Q_k}$ die rotierte Ladungsmatrix der Faktorenanalyse der Lerndimension k, d.h. a_{xy}^k sei der Eintrag dieser Matrix zu der Variable x und dem Faktor y. Dann wird mit

$$\begin{pmatrix} t_{j_1}^k \\ \vdots \\ t_{j_{m_k}}^k \end{pmatrix} = A^k b^k = \begin{pmatrix} a_{11}^k & \dots & a_{1Q_k}^k \\ \vdots & \ddots & \vdots \\ a_{m_k 1}^k & \cdots & a_{m_k Q_k}^k \end{pmatrix} \begin{pmatrix} b_1^k \\ \vdots \\ b_{Q_k}^k \end{pmatrix}$$

für jede einzelne Lerndimension differenziert nach Ost und West zunächst die Anteile t_j^k der einzelnen Variablen j am Index der Lerndimension k berechnet. Es sei $t_j^k = 0, \text{ falls } j \notin \{j_{m_1}, ... j_{m_k}\}$. Um einen gemeinsamen, inhaltlich plausiblen Anteil für Ost und West zu erhalten, wird ein - mit der Anzahl Analyseeinheiten pro Region - gewichteter Mittelwert gebildet, sofern t_j^k für beide Teile die erwartete Wirkungsrichtung hinsichtlich des Outcome besitzt. Liegt nur bei einem t_j^k die erwartete Richtung vor, wird stellvertretend für beide Teile auch nur dieser Wert als Anteil für die gesamte Kennzahl herangezogen. Für den sehr seltenen Fall, dass die Anteile sowohl in Ost als auch in

West nicht die erwartete Wirkungsrichtung aufweisen, erhält diese Kennzahl ein Gesamtgewicht von null.

Im nächsten Schritt wird der Index pro Lerndimension $\hat{Y}^k$ gebildet, indem der Vektor der Anteile der Kennzahlen einer Lerndimension k mit deren dazugehöriger Datenmatrix $X^k = x_{ij}^k \in \mathbb{R}^{M \times N}$, $k \in 1,...,4$, $i \in 1,...,M$, $j \in 1,...,N$, (mit imputierten und standardisierten Werten) multipliziert wird:

$$\hat{Y}^k = X^k t^k = \begin{pmatrix} x_{11}^k & \cdots & x_{1j_{m_k}}^k \\ \vdots & \ddots & \vdots \\ x_{M1}^k & \cdots & x_{Mj_{m_k}}^k \end{pmatrix} \begin{pmatrix} t_{j_1}^k \\ \vdots \\ t_{j_{m_k}}^k \end{pmatrix} \quad .$$

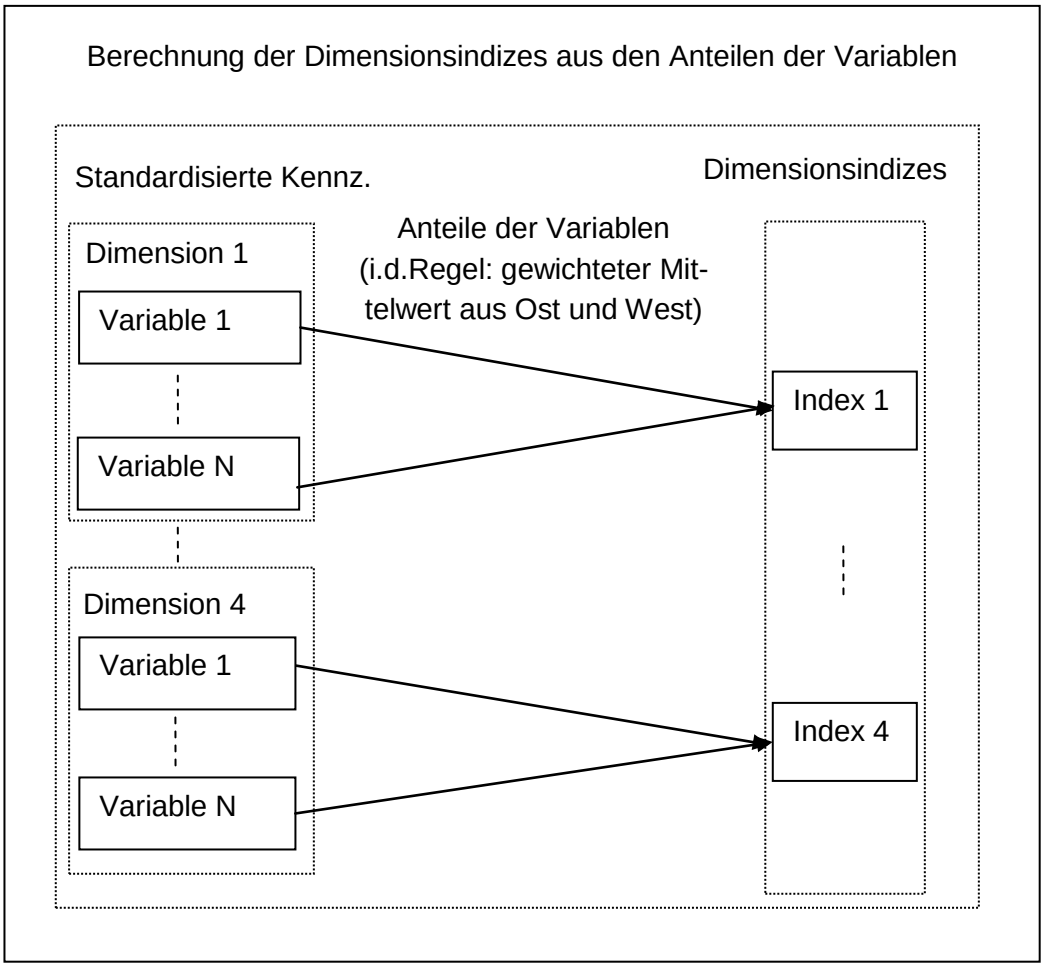

Abbildung 8. Berechnung der Dimensionsindizes

4.6 Berechnung des DLA-Gesamtindexes

Um den aggregierten DLA-Gesamtindexwert zu berechnen, werden alle vier Lerndimensionen betrachtet. Bezeichne $\hat{y}_i^k$ den Index der Analyseeinheit i zu der Lerndimension k, $k \in 1,...,4$, $i \in 1,...,M$.

Dann wird der DLA-Gesamtindexwert e_i einer Region i definiert als Summe der durch die Bestimmtheitsmaße gewichteten Indexwerte der einzelnen Lerndimensionen:

$$e_i = \sum_{k=1}^{4} R_k^2 \hat{y}_i^k \quad \text{für } i \in 1,...,M$$

Die Bestimmtheitsmaße pro Lerndimension sind ebenfalls (mit der Anzahl der Analyseeinheiten pro Region) gewichtete Mittelwerte aus den beiden einzelnen Regressionsanalysen für Ost und West.

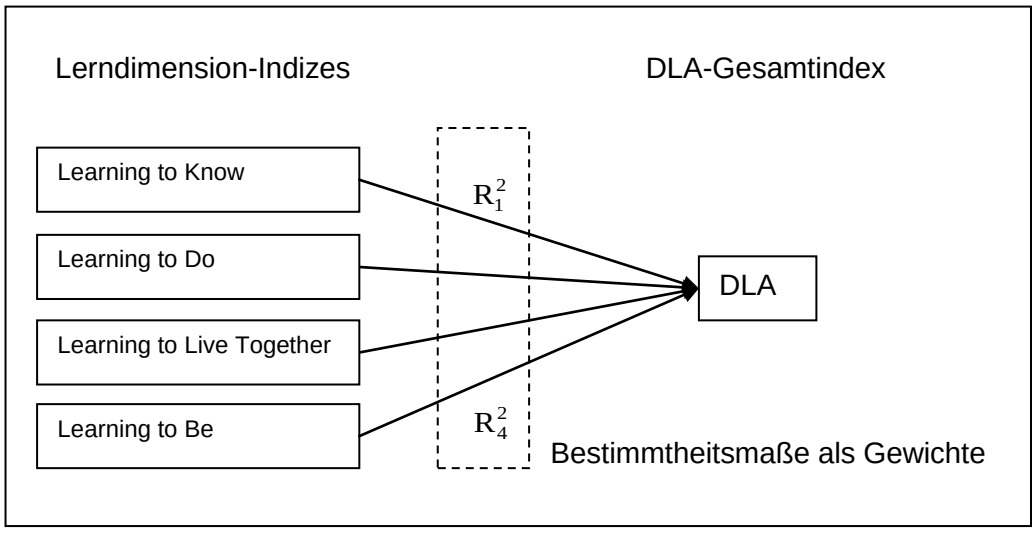

Abbildung 9. Aggregation der DLA-Indexwerte

Die Bestimmtheitsmaße repräsentieren den analytischen Erklärungsgehalt der in den Faktoren subsummierten Kennzahlenauswahl einer Dimension für den Outcome-Faktor. Somit erhält jede Dimension ihr individuelles Gewicht im Gesamtindex, d.h. sie fließt mit unterschiedlich starker Bedeutung in den DLA-Gesamtindex ein, um so die reale Lernsituation angemessen widerspiegeln zu können.

4.7 Finale Skalierung

Damit der DLA-Index ebenso wie die Dimensionsindizes mit den Lernindizes auf europäischer Ebene vergleichbar sind, werden alle fünf Indices abschließend an das europäische Skalenniveau angepasst, indem sie durch Transformation den gleichen Mittelwert und die gleiche Standardabweichung wie der Gesamtindex und die Dimensionsindizes für Deutschland auf EU-Ebene erhalten.