

H T
W I
G N

Hochschule Konstanz
Department of Computer Science

Submitted by
Jonas Mayer
Student Number 305630

jonas.mayer@htwg-konstanz.de

B

C

Bachelor Thesis

Automatisierte Extraktion und
Modellierung von Umweltindikatoren
aus PEP-Ecopassport-Dokumenten

S

Konstanz, 02.01.2026

Bachelor Thesis

Automatisierte Extraktion und Modellierung von Umweltindikatoren aus PEP-Ecopassport-Dokumenten

by

Jonas Mayer

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science

in Applied Computer Science

at the Hochschule Konstanz University of Applied Sciences,

Student Number: 305630

Date of Submission: 02.01.2026

Supervisor: **Prof. Dr. Doris Bohnet**

Second Examiner:

- An electronic version of this thesis is available at <https://github.com/jonez187/bachelorarbeit-htwg-latex>.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, Jonas Mayer, geboren am 06.11.2002 in Spaichingen,

1. dass ich meine Bachelorarbeit mit dem Titel

„Automatisierte Extraktion und Modellierung von
Umweltindikatoren aus PEP-Ecopassport-Dokumenten“

an der HTWG Konstanz unter Anleitung von Prof. Dr. Doris Bohnet selbstständig
und ohne fremde Hilfe angefertigt habe und keine anderen als die angeführten
Hilfen benutzt habe,

2. dass ich die Übernahme wörtlicher Zitate, von Tabellen, Zeichnungen, Bildern und
Programmen aus der Literatur oder anderen Quellen (Internet) sowie die Verwen-
dung der Gedanken anderer Autoren an den entsprechenden Stellen innerhalb
der Arbeit gekennzeichnet habe,

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Konstanz, 02.01.2026

Abstract

Diese Arbeit entwickelt eine durchgängige Pipeline zur automatisierten Aufbereitung von PEP-Ecopassport-Deklarationen und nutzt die daraus gewonnene Datenbasis zur heuristischen Abschätzung zentraler Umweltindikatoren für Produkte der Gebäudeautomation. PEPs liefern standardisierte Umweltkennzahlen, liegen in der Praxis jedoch als heterogene PDF-Dokumente vor und sind daher nur eingeschränkt automatisiert auswertbar.

Auf Basis des aufbereiteten Datensatzes wird ein Regressionsansatz entwickelt, der ausschließlich Merkmale nutzt, die auch ohne PEP typischerweise verfügbar oder abschätzbar sind. Dazu zählen insbesondere Gesamtgewicht, über die Lebensdauer aggregierter Stromverbrauch und verdichtete Materialinformationen, die durch eine PCA aus dem verwendeten Material abgeleitet werden. Die Modellgüte wird mit strikt getrennten Trainings- und Testdaten und zusätzlichen robusten Fehlermaßen bewertet.

Die stabilsten Ergebnisse werden für *Climate change (total)* erzielt. Für mehrere weitere Indikatoren ergeben sich ebenfalls gute Schätzungen, während *Eutrophication (freshwater)*, *Eutrophication (marine)* und *Radioactive waste disposed* nur eingeschränkt erklärbar sind. Methodische Unterschiede zwischen den PEPs, wie verschiedene Berechnungsmethoden, begrenzen die Vergleichbarkeit und wirken als Rauschen. Insgesamt ermöglicht der Ansatz eine schnelle, datengetriebene Einordnung für neue Produkte ohne PEP und kann frühe Entscheidungen in Entwicklung und Beschaffung unterstützen.

Inhaltsverzeichnis

Abstract	ii
1 Einleitung	1
1.1 Motivation und Relevanz	1
1.2 Zielsetzung der Arbeit	3
1.3 Wissenschaftliche Fragestellung	3
1.4 Aufbau der Arbeit	4
2 Theoretische Grundlagen	5
2.1 PEP-Ecopassport	5
2.1.1 PEP-Standard	6
2.1.2 Aufbau typischer PEP-Dokumente	6
2.2 Datenextraktion aus PDF-Dokumenten	8
2.2.1 PDF-Struktur und Herausforderungen der Textextraktion	9
2.2.2 Extraktionsansätze	10
2.2.3 Zielformat JSON	11
2.2.4 Informationsextraktion von Markdown nach JSON	12
2.3 Statistische Grundlagen	13
2.3.1 Deskriptive Statistik	13
2.3.2 Explorative Datenanalyse und Visualisierungen	14
2.3.3 Mathematische Transformationen	14
2.3.4 Hauptkomponentenanalyse (PCA)	15
2.3.5 Lineare Regression	17
3 Pipeline und Datenbasis (Methodik)	20
3.1 Überblick über die Pipeline	20
3.2 Datenerhebung und PEP-Recherche	21
3.3 PDF-Parsing und Extraktion	22
3.4 Normalisierung der Daten	24
3.5 Datenbereinigung und Validierung	25
4 Analyse der erarbeiteten Daten	27
4.1 Deskriptive Analyse der PEP-Daten	27
4.1.1 Vollständigkeit der Werte	28

4.1.2	Überblick der <i>Input</i> -Variablen	31
4.1.3	Überblick der Umweltindikatoren	35
4.2	Explorative Modellentwicklung	37
4.2.1	Experimentelle Fragestellungen	37
4.2.2	Vergleich der Feature-Sets	38
4.2.3	Vergleich der Regressionsverfahren	40
4.3	PCA der Materialien	40
4.3.1	Ergebnis der PCA	42
4.3.2	Interpretation der Material-Hauptkomponenten	43
4.4	Lineare Regression des Indikators <i>Climate change (total)</i>	44
4.4.1	Datenbasis und Transformation	44
4.4.2	Modellformulierung	45
4.4.3	Schätzverfahren, Validierung und Ergebnisse	45
4.4.4	Visualisierung der Vorhersagequalität	47
4.5	Lineare Regression der anderen Indikatoren	50
4.5.1	Regression des Indikators Acidification	51
4.5.2	Indikatoren mit geringer Modellgüte	53
5	Diskussion und Fazit	56
5.1	Vorgehen und Methodischer Beitrag	56
5.2	Einordnung der Ergebnisse im Kontext der Forschungsfrage	57
5.3	Grenzen und Limitationen	58
5.4	Ausblick und zukünftiger Forschungsbedarf	60
A	Anhang	62
A.1	Visualisierung weiterer Regressionsmodelle (Hohe Modellgüte)	62
A.1.1	Regression des Indikators Hazardous waste disposed	62
A.1.2	Regression des Indikators Water use	64
A.1.3	Regression des Indikators Photochemical ozone formation (HH)	65
A.1.4	Regression des Indikators Resource use, fossils	67
A.1.5	Regression des Indikators Eutrophication (terrestrial)	68
A.1.6	Regression des Indikators Ozone depletion	70
A.1.7	Regression des Indikators Resource use, minerals and metals	71
A.2	Visualisierung weiterer Regressionsmodelle (Geringe Modellgüte)	71
A.2.1	Regression des Indikators Eutrophication (freshwater)	73
A.2.2	Regression des Indikators Eutrophication Marine	74
Literatur		77

1

Einleitung

1.1. Motivation und Relevanz

Gebäudeautomatisierung und IoT-Komponenten gewinnen stark an Bedeutung. Sensoren, Gateways und Steuerungen werden in Gebäuden zunehmend eingesetzt, um Energieflüsse zu optimieren, Komfort zu erhöhen und Prozesse zu automatisieren. Mit der wachsenden Verbreitung solcher Geräte steigt auch ihr Anteil an Materialverbrauch, Energieeinsatz und Abfallaufkommen entlang des Lebenszyklus.

Der Gebäudesektor zählt neben Verkehr und der industriellen Produktion zu den wesentlichen Verursachern von CO₂-Emissionen. In einer Studie von Bitkom e.V., dem Branchenverband der deutschen Informations- und Telekommunikationsbranche, wird ein Drittel der Emissionen in Deutschland dem Gebäudesektor zugerechnet [21]. Digitale Gebäudetechnologien werden in dieser Studie als relevanter Hebel betrachtet, durch bedarfsgerechte, intelligente Steuerung und Monitoring den Energieverbrauch und die Emissionen von Gebäuden zu senken.

Neben erhöhtem Komfort können Gebäudeautomatisierung und IoT-Komponenten somit tatsächlich ökonomischen und ökologischen Nutzen liefern.

Die Größenordnung dieses Potenzials lässt sich anhand einer Szenariorechnung aus der Bitkom-Studie verdeutlichen. Abbildung 1.1 zeigt das modellierten Potential zur Einsparung von thermischer Energie digitaler Gebäudetechnologien im Gebäudesektor in Deutschland, wobei die Gebäudeautomation gemäß DIN EN 15232 in Effizienzklassen eingeteilt wird. Klasse D dient dabei als Referenz und entspricht Gebäuden ohne automatische Regelung.

Gleichzeitig sind die eingesetzten Geräte und die dadurch ermöglichten Einsparun-

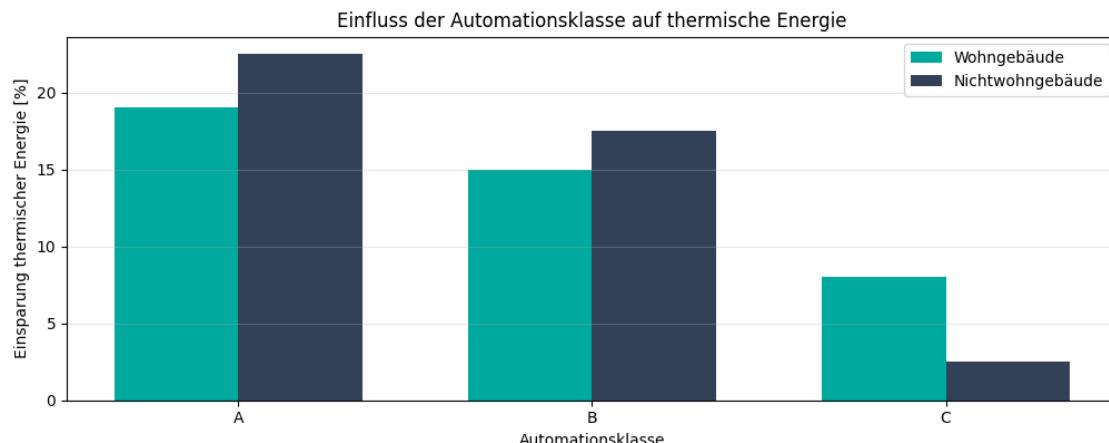


Abbildung 1.1: Einsparungen thermischer Energie nach Automationsklassen, relativ zur Klasse D (Zahlen aus der Studie [21]). [Eigene Darstellung]

gen ökologisch nicht „gratis“. Herstellung, Materialeinsatz, Transport sowie die Nutzung verursachen eigene Treibhausgasemissionen und weitere Umweltwirkungen. Damit entsteht ein Zielkonflikt: Einerseits besteht ein Einsparpotenzial auf Gebäudeebene. Andererseits müssen die zusätzlichen Umweltwirkungen der Hardware transparent und vergleichbar quantifiziert werden, um ökologische Hotspots zu identifizieren und fundierte Entscheidungen in Entwicklung, Beschaffung und Portfoliomanagement zu ermöglichen. Besonders in frühen Phasen liegen jedoch oft nur wenige robuste Produktmerkmale vor, während vollständige Ökobilanzen typischerweise aufwendig sind und detaillierte Annahmen erfordern.

PEP-Ecopassport-Dokumente liefern standardisierte Umweltindikatoren und stellen damit eine wertvolle Datenquelle dar. In der Praxis sind sie jedoch heterogen aufgebaut und als PDF-Dateien veröffentlicht. Dadurch werden automatisierte Auswertung, Skalierung auf größere Datenmengen und konsistente Vergleichbarkeit erschwert. Es entsteht eine Lücke zwischen dem Bedarf an schnellen, nachvollziehbaren Umweltabschätzungen und der tatsächlich verfügbaren Datengrundlage.

Eine kürzlich abgeschlossene Bachelorarbeit von Selg [Sel25] entwickelte eine Extraktionspipeline, mit der relevante Daten aus PEP-Ecopassport-PDF-Dateien automatisiert ausgelesen und strukturiert gespeichert werden können, und führte erste statistische Analysen durch. Die vorliegende Arbeit knüpft an diese Vorarbeit an, erweitert die Pipeline für eine größere und robustere Datenbasis und nutzt die extrahierten Variablen für eine vertiefte quantitative Analyse sowie die Entwicklung eines Modells.

1.2. Zielsetzung der Arbeit

Ziel dieser Arbeit ist der Aufbau einer durchgängigen Pipeline, die PEP-Ecopassport-Dokumente automatisiert verarbeitet und in eine strukturierte, analysierbare Datenbasis überführt. Die Pipeline wird so ausgelegt, dass sie heterogene PDF-Layouts robust verarbeitet und eine skalierbare Datengrundlage für die nachfolgende Analyse und Modellentwicklung bereitstellt.

Im Mittelpunkt steht die Ableitung eines robusten und interpretierbaren Modells zur heuristischen Abschätzung von Umweltauswirkungen, insbesondere der CO₂-Emissionen. Dazu werden in den PEP-Daten wiederkehrende Muster zwischen wenigen, allgemein verfügbaren Produktmerkmalen und den ausgewiesenen Umweltindikatoren identifiziert und modelliert. Der Ansatz ist so gewählt, dass das Modell auch auf neue Produkte ohne verfügbare PEPs übertragbar ist und eine erste Einordnung der zu erwartenden Umweltauswirkungen ermöglicht.

Die Genauigkeit des Modells soll analysiert und bewertet werden. Ergänzend sollen die Grenzen dieser Herangehensweise diskutiert, insbesondere im Hinblick auf Datenheterogenität, Ausreißer und die eingeschränkte Übertragbarkeit auf einzelne Produktkategorien und Nutzungsszenarien werden.

1.3. Wissenschaftliche Fragestellung

Aus der Zielsetzung ergibt sich die folgende zentrale Fragestellung:

Wie können Umweltindikatoren aus PEP-Ecopassport-Deklarationen für Produkte der Gebäudeautomatisierung robust extrahiert und analysiert werden, und inwieweit lassen sich daraus interpretierbare Modelle ableiten, die Umweltauswirkungen für Produkte ohne PEP heuristisch abschätzen?

Die Beantwortung dieser Fragestellung soll wissenschaftliche Erkenntnisse über die zentralen Umweltindikatoren innerhalb einer heterogenen Produktgruppe liefern und zeigen, welche Produktmerkmale die größten Erklärungsbeiträge leisten. Gleichzeitig wird untersucht, unter welchen Voraussetzungen vereinfachte, interpretierbare Regressionsmodelle eine praktikable Erstabschätzung ermöglichen, und wo die Grenzen liegen. Damit werden sowohl methodische Grundlagen für eine skalierbare Auswertung von PEP-Daten als auch praktische Anhaltspunkte für datenbasierte Nachhaltigkeitsbewertungen in frühen Entscheidungsphasen geschaffen.

1.4. Aufbau der Arbeit

Zur Beantwortung der Forschungsfrage wird ein methodischer Ablauf umgesetzt, der von der theoretischen Einordnung über den Aufbau einer belastbaren Datenbasis bis zur Modellierung und Diskussion der Ergebnisse reicht. Die Arbeit ist in fünf Kapitel gegliedert.

Kapitel 2 beschreibt die theoretischen Grundlagen. Zunächst werden der PEP-Ecopassport Standard und der typische Aufbau von PEP-Dokumenten erläutert. Anschließend werden Herausforderungen der strukturierten Informationsextraktion aus PDF-Dokumenten und geeignete Extraktionsansätze diskutiert. Das Kapitel schließt mit statistischen Grundlagen, die für die spätere explorative Analyse, Transformationen und Regressionsmodelle benötigt werden.

Kapitel 3 stellt die entwickelte Pipeline und die erzeugte Datenbasis dar. Es beschreibt die Recherche und Auswahl geeigneter PEP-Dokumente, das PDF-Parsing und die Extraktion strukturierter Inhalte, die Normalisierung zentraler Begriffe und Einheiten sowie die Datenbereinigung und Validierung. Ziel ist eine konsistente, analysierbare Datenbasis als Grundlage der Modellierung.

Kapitel 4 analysiert den erarbeiteten Datensatz und leitet daraus Modellentscheidungen ab. Zunächst werden Vollständigkeit, Verteilungen der Input-Variablen und zentrale Umweltindikatoren deskriptiv ausgewertet. Darauf aufbauend folgt eine explorative Modellentwicklung, in der Feature-Sets, PCA-Varianten und lineare Regressionsverfahren verglichen werden. Zusätzlich wird die Material-PCA beschrieben und interpretiert. Abschließend werden Regressionsmodelle für *Climate change (total)* und für weitere Indikatoren geschätzt und hinsichtlich Vorhersagegüte und Fehlerstruktur bewertet.

Kapitel 5 fasst die zentralen Ergebnisse zusammen und ordnet sie im Kontext der Forschungsfrage ein. Darüber hinaus werden Limitationen der Datenbasis und des Modellansatzes, methodische und technische Grenzen sowie die Übertragbarkeit der heuristischen Schätzung diskutiert. Das Kapitel schließt mit einem Ausblick auf mögliche Erweiterungen der Pipeline und weiteren Forschungsbedarf.

2

Theoretische Grundlagen

In diesem Kapitel werden die theoretischen Grundlagen für die vorliegende Arbeit gelegt. Für die standardisierte Berichterstattung von Umweltwirkungen existieren Formate wie die PEP-Ecopassports, die Indikatoren entlang des Lebenszyklus ausweisen. Damit die Angaben für quantitative Analysen verwendet werden können, müssen Begriffe, Einheiten und Moduldefinitionen vereinheitlicht und strukturiert werden, da diese in den PEP-Ecopassport-PDFs inkonsistent vorliegen. Ebenso ist ein grundlegendes Verständnis statistischer Verfahren erforderlich, um Muster und Zusammenhänge zuverlässig zu erkennen. Dieses Kapitel führt daher zunächst in Struktur und Inhalte von PEP-Deklarationen ein, beschreibt die Grundlagen der Datenextraktion aus PDF-Dateien und skizziert anschließend die methodischen Bausteine (u. a. PCA und lineare Regression), die in den folgenden Kapiteln zur Reduktion von Variablen, zur Erklärung von Indikatorvarianz und zur Ableitung eines Modells zur heuristischen Abschätzung der Umweltauswirkungen von Produkten ohne PEP eingesetzt werden.

2.1. PEP-Ecopassport

Die Datenquelle dieser Arbeit bilden die *PEP-Ecopassports*, welche ausschließlich im PDF-Format vorliegen. In diesem Kapitel werden die Standards, Inhalte und Strukturen dieser Dokumente beschrieben, um die spätere Datenerhebung und -verarbeitung nachvollziehbar zu machen.

2.1.1. PEP-Standard

Der *PEP-Ecopassport* ist ein international anerkanntes Programm für die Erstellung standardisierter Umweltproduktdeklarationen für elektrische, elektronische sowie Heizungs-, Lüftungs-, Klima- und Kälteprodukte (HVAC). Träger des Programms ist die *P.E.P. Association*, eine gemeinnützige Organisation, deren Ziel es ist, ein gemeinsames und verlässliches Referenzsystem für Umweltinformationen dieser Produktkategorien bereitzustellen. Das Programm versteht sich als Branchenspezialisierung innerhalb des Rahmens der *Environmental Product Declarations (EPD)* gemäß ISO 14025 und der Lebenszyklusnormen nach ISO 14040 und basiert somit auf international festgelegten Normen [Ass24].

Die PEP-Deklarationen basieren auf quantitativen Ergebnissen einer Lebenszyklusanalyse (*Life Cycle Assessment, LCA*) und dienen der vergleichenden Bewertung von Produkten mit identischer Funktion. Die Datenerhebung und Berechnung erfolgt nach vordefinierten Parametern, die in sogenannten *Product Category Rules (PCR)* und bei Bedarf in *Product Specific Rules (PSR)* festgelegt sind. Jede PEP-Deklaration unterliegt einer unabhängigen Überprüfung der angewandten Methodik und der zugrunde liegenden LCA-Daten [Has+13].

Die Teilnahme am PEP-Programm ist freiwillig, gewinnt jedoch in der Praxis an Bedeutung, da Umweltproduktdeklarationen zunehmend als Nachweis oder Auswahlkriterium in Ausschreibungen und Produktbewertungen herangezogen werden. Eine gesetzliche Verpflichtung zur Erstellung besteht bislang nur in Einzelfällen, beispielsweise in Frankreich, wenn ein Hersteller aktiv mit Umweltvorteilen wirbt [Ass24].

Das PEP-Programm unterscheidet sich nach Angaben der *P.E.P. Association* klar von unternehmensbezogenen Treibhausgas-Bilanzierungen. Es erfasst ausschließlich produktspezifische Umweltwirkungen entlang des Lebenszyklus und folgt dabei den methodischen Vorgaben der ISO 14040-Reihe. Für eine umfassende Treibhausgasbilanz auf Organisationsebene sind PEP-Daten daher nicht geeignet [Ass24].

2.1.2. Aufbau typischer PEP-Dokumente

Ein vollständiges PEP umfasst typischerweise etwa zehn Seiten und gliedert sich in mehrere inhaltlich definierte Abschnitte.

Titel- und Metadatenblatt Das Deckblatt enthält grundlegende Angaben zum Produkt (Name, Version, Sprache, Hersteller), zum Veröffentlichungs- und Revisionsdatum. Darüber hinaus sind Kontaktinformationen, Firmenadresse und Registrierungsnummer enthalten.

Allgemeine Produktinformationen Dieser Abschnitt beschreibt die funktionale Einheit (*functional unit*), in welcher auch der Stromverbrauch dargestellt ist. Weiterhin werden Referenzlebensdauer, hier meist 10 bis 20 Jahre, die Produktfunktion, Anwendungsbe-reiche und gegebenenfalls weitere Varianten aufgeführt.

Materialzusammensetzung Die Zusammensetzung des Produkts wird teils in Ta-bellenform, teils grafisch als Kreisdiagramm nach Hauptgruppen ausgewiesen, z. B. Kunststoffe, Metalle und weitere Materialien (Papier/Karton, Elektronik, Sonstiges). In diesem Abschnitt ist meist auch das Gesamtgewicht des Produktes zu finden.

Szenarien und Lebenszyklusphasen PEP-Dokumente sind entlang der Phasen des Produktlebenszyklus strukturiert, die den Vorgaben der EN 15804 entsprechen:

- **Herstellung (A1–A3)**: Produktion und Vormaterialien
- **Distribution (A4)**: Transport vom Werk zum Markt, häufig standardisierte Annahmen (z. B. 1 000 km Schiff, 3 300 km Lkw)
- **Installation (A5)**: Montage, meist nur Verpackungsabfälle berücksichtigt
- **Nutzungsphase (B)**: Betrieb des Geräts mit angegebenem Energieverbrauch, z. B. 126 kWh über 20 Jahre.
- **End-of-Life (C1–C4)**: Entsorgungsszenario gemäß PCR-Vorgaben (Recycling-, Deponie-, Transportanteile).
- **Optionale Phase (D)**: Rückgewinnung und Wiederverwendung außerhalb der Systemgrenzen.

In der weiteren Datenaufbereitung werden diese Phasen zu den Kategorien *manufacturing*, *distribution*, *installation*, *use* und *end_of_life* zusammengefasst.

Energiemodelle Zusätzlich werden die verwendeten Energiemodelle angegeben (z. B. *France Grid Mix*), welche die Herkunft und Zusammensetzung des im Lebenszyklus des Produkts genutzten Stroms beschreiben. Die Genauigkeit dieser Angaben variiert deutlich zwischen den Dokumenten. In einigen Fällen ist jeder einzelnen Produktlebenszyklusphase ein spezifisches Land inklusive des Jahres zugeordnet, während andere PEPs für alle Phasen einen einheitlichen europäischen Strommix angeben.

Umweltindikatoren Die Umweltwirkungen werden für jede Lebenszyklusphase und als Gesamtwert angegeben. Die für diese Arbeit relevanten Indikatoren sind in der Tabelle 2.1 aufgeführt.

Tabelle 2.1: Umweltindikatoren

Indikator	Beschreibung
Acidification	Versauerung von Böden und Gewässern durch säurebildende Emissionen
Climate Change (Total)	Gesamtes Treibhauspotenzial aus allen Quellen, CO ₂ -Äquivalente
Eutrophication (Freshwater)	Nährstoffanreicherung in Binnengewässern
Eutrophication (Marine)	Nährstoffanreicherung in marinen Ökosystemen
Eutrophication (Terrestrial)	Nährstoffanreicherung in Böden
Hazardous Waste Disposed	Entsorgung gefährlicher Abfälle
Ozone Depletion	Abbau der stratosphärischen Ozonschicht durch FCKW-Emissionen
Photochemical Ozone Formation (Human Health)	Bildung von bodennahem Ozon (Sommersmog)
Radioactive Waste Disposed	Entsorgung radioaktiver Abfälle
Resource Use (Fossils)	Nutzung fossiler Energieressourcen
Resource Use (Minerals and Metals)	Verbrauch abiotischer Ressourcen (Metalle und Mineralien)
Water Use	Entnahme und Verbrauch von Frischwasser

Verifikations- und Anhangsangaben Im abschließenden Teil werden die angewendeten Regelwerke und Datenquellen genannt, z. B. *PCR-ed3-EN-2015_04_02* und *PSR-0005-ed2-EN-2016_03_29*, die eingesetzte Software (z. B. SimaPro 9.3 mit Ecoinvent 3.8), die Verifizierungsstelle und deren Akkreditierungsnummer.

Obwohl der inhaltliche Mindestumfang und die zu berichtenden Umweltindikatoren durch die zugrundeliegenden ISO- und PCR-Vorgaben festgelegt sind, besteht keine feste formale Struktur. Das Layout, die grafische Aufbereitung und die Anordnung der Tabellen können je nach Hersteller, Software und Version variieren. So enthalten einige PEPs tabellarische Aufstellungen sämtlicher Indikatoren, während andere ergänzend oder teilweise ausschließlich Diagramme und grafische Vergleichsdarstellungen beinhalten.

2.2. Datenextraktion aus PDF-Dokumenten

Da die PEP-Ecopassport-Umweltdaten ausschließlich in PDF-Dateien veröffentlicht werden, besteht der erste Schritt darin, die Informationen zu extrahieren, um sie für die quantitative Analyse in ein einheitlich strukturiertes und maschinenlesbares Format zu

bringen. Dieser Prozess bildet die Grundlage für die weitere Verarbeitung, Strukturierung und Analyse der Umweltindikatoren.

2.2.1. PDF-Struktur und Herausforderungen der Textextraktion

In [LB95] wird das PDF-Format als ein weit verbreitetes, primär layoutbasiertes Dokumentenformat beschrieben. Das Format enthält die Platzierung und Darstellung von Textobjekten sowie grafischen Elementen wie Linien, Kurven und Bildern inklusive Stilattributen etwa Schriftart, Farbe oder Strichführung. Dadurch bleibt das visuelle Erscheinungsbild eines Dokuments zuverlässig über alle Geräte erhalten.

Nach [BK17] und [CF04] fehlt dem PDF-Format allerdings eine explizite logische Struktur auf höherer Ebene. Semantische Einheiten wie Wörter, Textzeilen, Absätze oder Tabellen sind nicht direkt enthalten. Auch die Rolle eines Textblocks, etwa als Haupttext, Überschrift oder Fußnote, ist nicht eindeutig hinterlegt. Ebenso ist die Lese- und Wortreihenfolge, insbesondere bei mehrspaltigen Layouts oder eingebetteten Elementen, nicht definiert.

Wie [CZ17] zeigt, erschwert das fehlende semantische Strukturgerüst in vielen PDFs die automatische Erkennung und Wiederverwendung von Layout und Inhalt deutlich, da die Rekonstruktion des Textflusses und der semantischen Einheiten ausschließlich auf den Positionen einzelner Zeichen basiert.

Die folgenden Herausforderungen der PDF-Extraktion sind bekannt und in der Literatur gut dokumentiert (vgl. [BK17; Lip+13; CF04; LB95]). Explizit ergeben sich dabei vier typische Problemklassen:

- **Wort- und Zeichenrekonstruktion:** Wortgrenzen sind aufgrund variierender Zeichenabstände uneindeutig. Zusätzlich müssen Silbentrennungen, Ligaturen und diakritische Zeichen korrekt zusammengeführt werden.
- **Lesereihenfolge:** In mehrspaltigen oder komplexen Layouts ist die interne Reihenfolge der Textobjekte häufig nicht identisch mit dem menschlichen Lesefluss, was ohne Korrektur zu falsch zusammengesetztem Text führt.
- **Absatz- und Blockstruktur:** Absatzgrenzen sind nicht explizit kodiert. Textblöcke können durch Tabellen, Formeln oder Abbildungen unterbrochen sein und über Seiten oder Spalten hinweg fortgesetzt werden.
- **Layout- und Rendering-Artefakte:** Überlagerungen, Segmentierungsfehler bei Tabellen und Diagrammen sowie als Grafiken gerenderte Zeichen (z. B. Type-3-Fonts) erschweren die zuverlässige Extraktion.

2.2.2. Extraktionsansätze

Da diese Probleme weit verbreitet und bekannt sind, gibt es mehrere Extraktionsansätze, um PDF-Dateien in ein strukturiertes Format zu bringen, mit dem Ziel, sie anschließend weiter zu analysieren.

Klassische Verfahren (regelbasierte Parser) Regelbasierte PDF-Parsing-Methoden arbeiten mit fest definierten Regeln und benötigen kein Training. Mehrere weit verbreitete Softwarebibliotheken implementieren regelbasierte Parser, darunter *pdfplumber*, *pypdfium2* und *pypdf*. Da die Arbeit von Selg auf *pdfplumber* aufbaut, wird diese Bibliothek näher besprochen [Sel25].

Die Eigenschaften und Grenzen des Tools werden in [YCZ25] und [AA25] beschrieben:

Jede Seite wird als Sammlung von Textfragmenten, Linien, Rechtecken und Bildern mit ihren Positionen in Python-Objekte gespeichert. Für die Tabellenerkennung werden horizontale und vertikale Linien als potenzielle Zellgrenzen interpretiert.

Bei klar strukturierten, editierbaren PDF-Dokumenten führt diese Methodik zu guten Ergebnissen. In Studien zur Leistungsbewertung von Extraktionstools erzielte es in Domänen wie juristischen oder technischen Dokumenten hohe F1-Scores (beispielsweise 0,98 im Bereich *Law*).

Die Grenzen des Werkzeugs zeigen sich bei komplexen oder unregelmäßig formatierten PDF-Dateien. Insbesondere wissenschaftliche Dokumente mit mathematischen Ausdrücken und verschachtelten Tabellen führen zu deutlichen Leistungseinbußen. In der Kategorie *Scientific* sank der F1-Score auf 0,76, was vor allem auf unvollständige Tabellenerkennung und fehlerhafte Segmentierung zurückzuführen ist.

Aus Sicht der zuvor beschriebenen Herausforderungen adressiert *pdfplumber* also Probleme auf der Ebene der Zeichen- und Worterkennung weitgehend. Die Wiederherstellung der Lesereihenfolge erfolgt allerdings rein geometrisch, ohne semantisches Verständnis, wodurch Textpassagen aus mehrspaltigen Layouts häufig in falscher Reihenfolge extrahiert werden. Absatzgrenzen, semantische Rollen (z. B. Überschriften, Fließtext, Bildunterschriften) und komplexe Tabellenstrukturen werden nicht zuverlässig erkannt.

Damit steht *pdfplumber* exemplarisch für klassische Extraktionstools, die ohne maschinelles Lernen und Modellen zum Dokumentenverständnis arbeiten und deshalb bei komplex strukturierten Dokumenten wie den PEP-Ecopassports an ihre methodischen Grenzen stoßen.

Erweiterte Verfahren (z.B. Docling) Neben klassischen regelbasierten Parsern etablierten sich moderne, KI-gestützte Dokumentenanalyse-Frameworks. Dazu gehören komplexe Layout-Modelle wie *LayoutParser* [25b], *GROBID* [25a] und *Docling*. Sie kombinieren visuelle Merkmale, Textinformationen und semantische Modelle, um Dokumente mit anspruchsvoller Struktur automatisch zu analysieren und in maschinenlesbare Formate zu überführen.

Für diese Arbeit wird *Docling* näher betrachtet, da es ein lokal ausführbares Open-Source-Toolkit ist, das eine vollständige End-to-End-Pipeline für Layout-Analyse, Strukturerkennung und Tabellensegmentierung bereitstellt.

Die Funktion und Hintergründe von *Docling* werden in der Dokumentation (vgl. [Aue+24; Aue+25]) ausführlich beschrieben und im Folgenden zusammengefasst:

Im Gegensatz zu regelbasierten Werkzeugen wie *pdfplumber*, kombiniert *Docling* klassische Parsing-Verfahren mit tiefen neuronalen Modellen für Layout- und Strukturerkennung.

Docling verarbeitet Dokumente in einer linearen Pipeline. Ein auf *DocLayoutNet* trainiertes Layoutmodell ist Teil dieser Pipeline und detektiert Seitenelemente wie Absätze, Überschriften, Listen, Abbildungen und Tabellen, die anschließend mit den extrahierten Text-Tokens zu einer konsistenten Dokumentstruktur zusammengeführt werden. Tabellen werden mit *TableFormer* in ihrer Zeilen- und Spaltenlogik rekonstruiert und die Zellen semantisch klassifiziert.

Die Ergebnisse werden in einem *DoclingDocument* gebündelt, das Inhalte, Layout und Metadaten einheitlich repräsentiert und in Formate wie JSON, Markdown oder HTML exportiert. Ein Post-Processing verbessert zudem die Lesereihenfolge, erkennt die Sprache und extrahiert zentrale Metadaten.

Durch diese Architektur können zentrale Herausforderungen adressiert werden (vgl. 2.2.1). *Docling* rekonstruiert den Lesefluss auch bei mehrspaltigen Layouts, erkennt logische Dokumentelemente und interpretiert Tabellen strukturell statt rein geometrisch. Zusätzlich klassifiziert es Inhalte nach Rollen wie Fließtext, Überschrift oder Bildunterschrift, wodurch die Ausgaben als strukturierte Basis für weiterführende Analysen nutzbar sind.

2.2.3. Zielformat JSON

Die aus den PEP-PDFs extrahierten Inhalte liegen zunächst als Text oder Markdown vor. Diese Formate sind zwar lesbar, liefern aber kein stabiles Schema, um Umweltindikatoren, Materialien und Metadaten über viele Dokumente hinweg eindeutig und konsistent zuzuordnen. Für die quantitative, automatisierte Auswertung und Analyse ist daher ein fest definiertes, maschinenlesbares Zielformat erforderlich.

Als Zielformat wird *JavaScript Object Notation (JSON)* verwendet, das sich als Standard für strukturierten Datenaustausch etabliert hat. JSON bildet Daten über einfache Typen und hierarchische Strukturen wie Objekte und Arrays ab und eignet sich damit für verschachtelte Informationen [Pez+16]. In dieser Arbeit dient JSON als einheitliche Repräsentation der extrahierten PEP-Daten und ermöglicht eine reproduzierbare Weiterverarbeitung in Analyseumgebungen wie Python oder R.

2.2.4. Informationsextraktion von Markdown nach JSON

In [Gri15] wird Informationsextraktion (IE) als Aufgabe beschrieben, aus unstrukturiertem Text gezielt relevante Informationen zu gewinnen und sie in eine strukturierte Repräsentation wie ein JSON-Schema zu überführen. Für diese Aufgabe lassen sich zwei Ansätze unterscheiden. Erstens klassische, regelbasierte Pipeline-Systeme. Zweitens Verfahren auf Basis großer Sprachmodelle (LLMs).

[Gri15] beschreibt die Grundsätze und Herausforderungen traditioneller IE-Systeme. Traditionelle IE-Systeme sind meist mehrstufig aufgebaut, zum Beispiel mit Entitätserkennung, syntaktischer Analyse, Koreferenzauflösung und Relationsextraktion. Ein Vorteil dieser Pipeline-Ansätze ist die hohe Nachvollziehbarkeit und die weitgehend deterministische Verarbeitung. Bei heterogenen Textformaten können allerdings Fehler aus frühen Stufen leicht in spätere Schritte übertragen werden und die Entwicklung robuster Regeln oder Feature-Sets ist mit erheblichem Aufwand verbunden.

In [Nad+24] werden große Sprachmodelle (LLMs) als neuere Möglichkeit beschrieben, Informationsextraktion semantisch zu lösen. LLMs können Textpassagen kontextsensitiv interpretieren und strukturierte Ausgaben direkt erzeugen. Wie [Mor+25] zeigt, lassen sich dabei Entitäten, Relationen und numerische Werte häufig ohne manuelles Regelwerk oder domänenspezifisch annotiertes Trainingskorpus zuordnen. Zudem können LLMs zuvor durch Tools wie *Docling* erzeugte Markdown- oder Textsegmente semantisch auswerten, sodass eine strikt sequentielle Abarbeitung einzelner Pipeline-Stufen teilweise entfällt [Nad+24]. Ein zentraler Nachteil ist die fehlende Deterministik und das Risiko von Halluzinationen, also falsch generierten Werten, das sich durch präzises Prompt-Design und Validierungsschritte zwar reduzieren, aber nicht vollständig ausschließen lässt [Mor+25].

Zusammenfassend sind regelbasierte Verfahren gut nachvollziehbar, stoßen bei den heterogenen und layoutbedingt uneinheitlichen PEP-Texten jedoch schnell an Grenzen. LLM-basierte Methoden sind flexibler und können Inhalte direkt in das vorgegebene JSON-Schema überführen, erfordern dafür aber Validierungsschritte. Da diese Arbeit auf eine quantitative Analyse der PEP-Ecopassport PDFs abzielt, wird trotzdem ein LLM für die Extraktion und Überführung in das harmonisierte JSON-Zielformat eingesetzt,

ergänzt durch Plausibilitätsprüfungen und Validierung.

2.3. Statistische Grundlagen

Die in dieser Arbeit verwendeten statistischen Verfahren bilden die methodische Grundlage zur Analyse und Modellierung der aus PEPs extrahierten Daten. Dazu werden zunächst *deskriptive und explorative* Verfahren eingesetzt, um Strukturen, Streuungen und Ausreißer in den Daten sichtbar zu machen. Darauf aufbauend wird die *lineare Regression* als einfaches, interpretierbares Modell genutzt, um heuristische Beziehungen zwischen Einflussgrößen und den resultierenden Umweltindikatoren zu identifizieren. Diese Kombination ermöglicht eine robuste, nachvollziehbare und datengetriebene Einschätzung ökologischer Wirkzusammenhänge im Datensatz.

2.3.1. Deskriptive Statistik

Die deskriptive Statistik bildet die Grundlage der quantitativen Datenanalyse. Sie dient der Zusammenfassung und Beschreibung von Datensätzen, um zentrale Merkmale einer Verteilung zu charakterisieren und potenzielle Muster oder Auffälligkeiten zu erkennen [FM09]. Der Schwerpunkt liegt nicht auf Hypothesentests, sondern auf dem Verständnis der vorhandenen Daten [Dim+19].

Die deskriptive Statistik umfasst numerische Verfahren zur Beschreibung der *Lage- und Streuungskennzahlen* von Daten [FM09]. Ziel ist die Abbildung großer Datenmengen auf wenige aussagekräftige Kennzahlen. Zu den typischen Lagemaßen gehören *Mittelwert* und *Median*. Der Mittelwert beschreibt die durchschnittliche Ausprägung, während der Median die geordnete Verteilung in zwei gleich große Hälften teilt. Der Median gilt als *robustes Lagemaß*, da er, im Gegensatz zum Mittelwert, wenig durch Ausreißer beeinflusst wird. Für die Streuung werden Standardabweichung, Spannweite und insbesondere der *Interquartilsabstand (IQR)* verwendet. Der IQR beschreibt die mittleren 50 % der Daten und ist ein robustes Maß, das gegenüber Extremwerten stabil bleibt. Für ordinale Merkmale ist der Median das geeignete Lagemaß. Der IQR, ergänzt um Minimum und Maximum, quantifiziert die Streuung [FM09].

Ein zentrales Merkmal numerischer Daten ist die Form ihrer Verteilung. In symmetrischen Verteilungen fallen Mittelwert und Median zusammen. Bei *rechtsschiefen* Verteilungen liegen einzelne hohe Werte weit über dem zentralen Bereich, sodass der Mittelwert größer als der Median ist. Bei *linksschiefen* Verteilungen gilt das umgekehrte Muster [KSY18]. Schiefe beeinflusst die Interpretation von Lage- und Streumaßen und

motiviert den Einsatz robuster Kennwerte wie Median und IQR [MJ10].

2.3.2. Explorative Datenanalyse und Visualisierungen

Die explorative Datenanalyse ergänzt die deskriptive Statistik durch strukturentdeckende Verfahren. Sie dient der visuellen Erkundung und Bewertung von Mustern, Ausreißern oder Zusammenhängen zwischen Variablen, ohne dass zuvor Hypothesen formuliert werden müssen. Zentrale Visualisierungen sind Histogramme, Boxplots und QQ-Plots [KSY18].

Histogramme stellen Häufigkeitsverteilungen kontinuierlicher Merkmale über Klassen dar. Sie erlauben Rückschlüsse auf Symmetrie, Schiefe und Mehrgipfligkeit und dienen zur Prüfung von Verteilungsannahmen [MJ10].

Boxplots visualisieren Median (Q_2), Quartile (Q_1, Q_3) und potenzielle Ausreißer. Quartile teilen die Verteilung in vier gleich große Teile. Die sogenannten *Whisker* markieren üblicherweise den Bereich bis zum 1,5-fachen Interquartilsabstand. Werte außerhalb gelten als potenzielle Ausreißer [MJ10]. Diese Darstellungsform ermöglicht die Beurteilung von Streuung, Schiefe und Extremwerten und eignet sich für den Vergleich mehrerer Merkmale [KSY18].

QQ-Plots (Quantile-Quantile-Plots) untersuchen grafisch, wie gut ein Datensatz einer bestimmten theoretischen Verteilung folgt [Sel18]. In dieser Arbeit werden sie eingesetzt, um zu prüfen, ob die Residuen eines Modells normalverteilt sind.

2.3.3. Mathematische Transformationen

Transformationen dienen nach [MJ10] und [FM09] dazu, Schiefe zu reduzieren, Varianz zu stabilisieren und parametrische Tests zu ermöglichen, damit die Modellierung auf der Transformationsskala sinnvoller wird. Als eine gängige Transformation, um Rechtsschiefe zu korrigieren, wird die logarithmische Transformation (Log-Transformation) erwähnt. Da der Logarithmus extrem hohe Werte stärker staucht als kleine Werte, kann die Transformation nach [MJ10] den Einfluss von Ausreißern auf das Gesamtergebnis reduzieren. Für sehr kleine Werte von x nahe 0 gilt die Näherung $\log(1 + x) \approx x$. Kleine Werte werden demnach kaum verändert.

Neben der festen Log-Transformation kann nach [ARC21] auch die Box-Cox-Transformation verwendet werden, die eine Familie von Potenztransformationen mit dem Parameter λ

darstellt. Für positive Daten ist sie definiert als

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0. \end{cases}$$

Damit ist die Log-Transformation als Spezialfall für $\lambda = 0$ enthalten. Ansonsten gilt $\lambda = 1$ entspricht keiner Transformation und $\lambda = 1/2$ entspricht der Quadratwurzeltransformation. Im Unterschied zu $\log(1 + x)$ ist die Form der Transformation hier nicht fest vorgegeben, sondern wird über λ angepasst. Ziel ist es, eine Skala zu finden, auf der sich ein lineares Modell einfacher und mit besser erfüllten Verteilungsannahmen der Fehler beschreiben lässt [ARC21]. In Softwarebibliotheken wird der Parameter λ datengetrieben bestimmt. In Python übernimmt dies beispielsweise die Bibliothek `scipy`, indem λ intern per Maximum-Likelihood geschätzt wird [25f].

Die Wirkung der Transformationen wird in Abbildung 2.1 exemplarisch gezeigt. Links ist die rechtsschiefe Verteilung auf Originalskala dargestellt, in der Mitte dieselben Werte nach der Transformation $\log(1 + x)$ und rechts nach einer Box-Cox-Transformation.

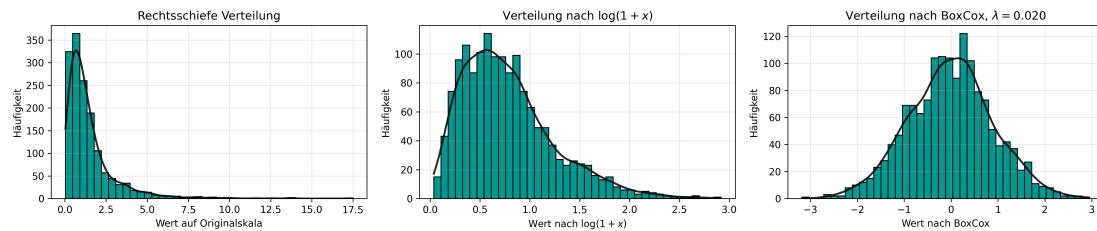


Abbildung 2.1: Beispiel einer rechtsschiefen Verteilung vor und nach Transformationen. Die überlagerte Dichtekurve verdeutlicht die Form der Verteilung. [Eigene Darstellung]

In dieser Arbeit wird vor allem die Transformation $\log(1 + x)$ verwendet (`log1p`), um Indikatorwerte und Residuen auf einer besser interpretierbaren Skala zu analysieren.

2.3.4. Hauptkomponentenanalyse (PCA)

Die Hauptkomponentenanalyse (Principal Component Analysis, PCA) reduziert die Komplexität multivariater Daten, indem neue Variablen, die *Hauptkomponenten*, als Linearkombinationen der ursprünglichen Variablen konstruiert werden. In [MR93] wird dieses Prinzip so beschrieben, dass die erste Komponente die größtmögliche Varianz der Daten erfasst, während weitere Komponenten orthogonal dazu definiert werden und jeweils möglichst viel der verbleibenden Varianz erklären.

Eine anschauliche Interpretation liefert die geometrische Sichtweise. In [AW10] wird erläutert, dass die Werte der Hauptkomponenten für jede Beobachtung als Projektionen

der Datenpunkte auf neue Achsen verstanden werden können. Diese Hauptkomponenten lassen sich als gedrehte Achsen im Merkmalsraum interpretieren. Genau das illustriert Abbildung 2.2. Die Standardachsen (x_1, x_2) beschreiben die Daten zunächst in der ursprünglichen Basis. Die PCA wählt stattdessen eine neue orthogonale Basis (v_1, v_2), wobei v_1 entlang der Richtung maximaler Streuung der Punktwolke liegt. Die zweite Achse v_2 steht senkrecht dazu und beschreibt die Reststreuung.

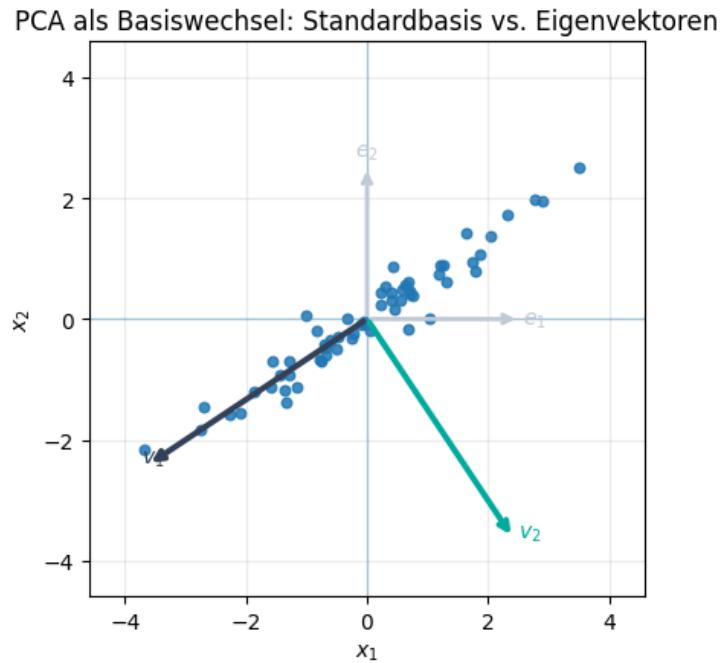


Abbildung 2.2: PCA als Basiswechsel: Standardachsen (x_1, x_2) und Hauptachsen (v_1, v_2) der Punktwolke. [Eigene Darstellung]

Für viele Anwendungen genügt es, nur die ersten wenigen Komponenten weiterzuverwenden, weil sie den dominanten Teil der Varianz tragen. Dadurch vereinfachen sich Visualisierung und Modellierung, ohne dass die Datenstruktur vollständig verloren geht. In [AW10] werden diese Ziele als Informationsverdichtung und Strukturanalyse der Variablen beschrieben.

In dieser Arbeit wird PCA als Baustein für die Principal Component Regression genutzt. Dabei werden korrelierte Regressoren, hier insbesondere Materialanteile, durch unkorrelierte Hauptkomponenten ersetzt. In [Jol82] wird dieses Vorgehen damit begründet, dass die Regression durch orthogonale Prädiktoren stabiler werden kann. Dadurch lässt sich Multikollinearität reduzieren, während das Modell weiterhin auf denselben Eingangsinformationen basiert.

2.3.5. Lineare Regression

Die lineare Regression dient in dieser Arbeit als methodische Grundlage zur Modellierung der Umweltwirkungen von Produkten auf Basis quantitativer Einflussgrößen. Ziel ist es, Zusammenhänge zwischen erklärenden Variablen wie *Produktgewicht*, *Materialzusammensetzung*, *Stromverbrauch* und *verwendetem Energiemix* und den resultierenden *Umweltindikatoren* zu erkennen und zur Abschätzung unbekannter Werte nutzbar zu machen.

Regressionsmodell und Grundannahmen

Das Regressionsmodell beschreibt den linearen Zusammenhang zwischen einer abhängigen Variable y (z. B. einem Umweltindikator) und mehreren unabhängigen Variablen x_1, x_2, \dots, x_k (z. B. Gewicht, Stromverbrauch, Materialanteile):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon.$$

Dabei ist β_0 der Achsenabschnitt, β_i die Regressionskoeffizienten der jeweiligen Einflussgrößen und ε ein zufälliger Fehlerterm, der unerklärte Varianzanteile abbildet. Die Koeffizienten β_i quantifizieren die Richtung und Stärke des Einflusses einzelner Variablen auf den Zielindikator [MPV22].

Für die lineare Regression gelten folgende Grundannahmen:

- **Linearität:** Die Beziehung zwischen der abhängigen und den unabhängigen Variablen ist näherungsweise linear.
- **Erwartungswert der Fehler:** Die Fehlerterme haben einen Erwartungswert von null $E(\varepsilon) = 0$ und sind normalverteilt.
- **Homoskedastizität:** Die Varianz der Fehler ist konstant und unabhängig von den erklärenden Variablen.
- **Unabhängigkeit:** Die Fehlerterme sind voneinander unkorreliert. [SYT12]

Diese Annahmen sichern die Unverzerrtheit und Effizienz der Parameterschätzungen. Für explorative Anwendungen, wie sie in dieser Arbeit verfolgt werden, steht jedoch die Strukturentdeckung im Vordergrund. Moderate Abweichungen von den Idealannahmen sind daher akzeptabel, sofern sie dokumentiert werden.

Schätzung der Regressionskoeffizienten

Die Schätzung der Regressionskoeffizienten erfolgt nach der Methode der kleinsten Quadrate (*Ordinary Least Squares*, OLS). Dabei werden die Parameter so bestimmt, dass

die Summe der quadrierten Abweichungen zwischen beobachteten und modellierten Werten minimal wird:

$$S(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

[SYT12]

Dabei ist n die Anzahl der Beobachtungen, y_i der beobachtete Zielwert und \hat{y}_i die Modellvorhersage für Beobachtung i . Die Koeffizienten β_i geben an, wie stark sich der Zielindikator y im Mittel verändert, wenn sich die Einflussgröße x_i um eine Einheit ändert, während alle anderen Variablen konstant bleiben.

In praktischen Anwendungen können die OLS-Schätzer trotz unverzerrter Erwartung eine hohe Varianz aufweisen, insbesondere bei Multikollinearität. In diesen Fällen können *Regularisierungsverfahren* die Vorhersagegüte verbessern, indem sie große Koeffizienten gezielt bestrafen und dadurch stabilere Modelle erzeugen [SYT12].

Die *Ridge-Regression* (L2-Regularisierung) erweitert das OLS-Kriterium um einen quadratischen Strafterm:

$$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^k \beta_j^2.$$

Der Regularisierungsparameter $\lambda \geq 0$ steuert die Stärke der Schrumpfung. Mit wachsendem λ werden die Koeffizienten stärker in Richtung null gezogen, bleiben jedoch typischerweise ungleich null. Dies reduziert die Varianz der Schätzung, akzeptiert dafür eine geringe Verzerrung und kann den Testfehler senken [MPV22].

Die *Lasso-Regression* (L1-Regularisierung) verwendet einen absoluten Strafterm:

$$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^k |\beta_j|.$$

Durch die L1-Struktur können Koeffizienten exakt null werden. Lasso ermöglicht damit zusätzlich eine implizite Variablenelektion und führt häufig zu sparsameren, leichter interpretierbaren Modellen [SYT12].

Der Parameter λ wird in der Regel in beiden Regularisierungsmethoden datengetrieben bestimmt, etwa mittels Cross Validation auf Trainingsdaten. Dadurch wird ein Kompromiss zwischen Unteranpassung bei zu großer Regularisierung und Überanpassung bei $\lambda = 0$ gewählt [SYT12].

Residuen und Gütemaße

Ein *Residuum* e_i ist die Abweichung zwischen einem beobachteten Wert y_i und dem vom Modell vorhergesagten (angepassten) Wert \hat{y}_i und wird definiert als

$$e_i = y_i - \hat{y}_i.$$

Residuen können als beobachtbare Werte der Modellfehler verstanden.

Das *Bestimmtheitsmaß* R^2 beschreibt den Anteil der Varianz des Zielindikators, der durch die erklärenden Variablen erklärt wird, und dient als zentrales Maß der Modellgüte. Der *Root Mean Square Error* (RMSE) misst die durchschnittliche Abweichung zwischen beobachteten und vorhergesagten Werten. Er hat die gleiche Einheit wie die Zielvariable und ist dadurch leicht interpretierbar und ebenfalls ein Indikator für die Modellgüte [MPV22].

In dieser Arbeit wird die multiple lineare Regression verwendet, um Heuristiken zur Abschätzung der Umweltwirkungen von Elektro- und Elektronikprodukten zu entwickeln. Das Modell dient der quantitativen Erfassung von Zusammenhängen zwischen Produktmerkmalen und Umweltindikatoren und daraus schließend der möglichst präzisen Prognose der Umweltindikatoren anhand der Input-Variablen. Damit bildet die lineare Regression eine nachvollziehbare, statistisch fundierte Basis für die Entwicklung eines vereinfachten Bewertungsmodells innerhalb der PEP-Datenanalyse.

3

Pipeline und Datenbasis (Methodik)

Dieses Kapitel beschreibt den Aufbau der Datenpipeline, die Extraktion der relevanten Variablen aus PEP-Ecopassport-Dokumenten und die Struktur und Aufbereitung der Datenbasis.

3.1. Überblick über die Pipeline

Ziel der entwickelten Pipeline ist die automatisierte Extraktion strukturierter Daten aus PEP-Ecopassport-Dokumenten im PDF-Format. Die PEPs bilden die zentrale Quelle für produktbezogene Umweltinformationen, enthalten jedoch uneinheitlich formatierte Tabellen und Textblöcke, die eine direkte Auswertung erschweren.

Die Pipeline wandelt die heterogenen PDF-Dokumente in ein einheitliches, maschinenlesbares Datenformat um. Als Input dienen die PEP-PDFs, der Output ist eine strukturierte CSV-Datei, die sämtliche relevanten Variablen zu Produkt, Materialien, Energieverbrauch und Umweltindikatoren enthält. Der Prozess umfasst mehrere aufeinanderfolgende Schritte:

- **Recherche und Erfassung:** Recherche, Speicherung und Bewertung der verfügbaren PEP-Dokumente mit Gebäudeautomatisierungsbezug aus der öffentlichen PEP-Datenbank [Ass25].
- **Extraktion:** Umwandlung der PDF-Dateien in Rohtext und Tabelleninhalte mittels Dokumentenparser, Layout- und Tabellenstrukturen werden erkannt.

- **Interpretation:** Zuordnung der erkannten Inhalte zu einem definierten Schema mithilfe regelbasierter und LLM-gestützter Methoden.
- **Normalisierung:** Harmonisierung von Einheiten, Materialnamen und Energiemodellen zur Sicherstellung der Vergleichbarkeit.
- **Export:** Zusammenführung aller Informationen in eine flache, analysierbare CSV-Datei als Grundlage der nachfolgenden statistischen Auswertung.

Abbildung 3.1 zeigt den groben schematischen Aufbau des Gesamtprozesses von der Rohdatenerfassung bis zur strukturierten Datenbasis.

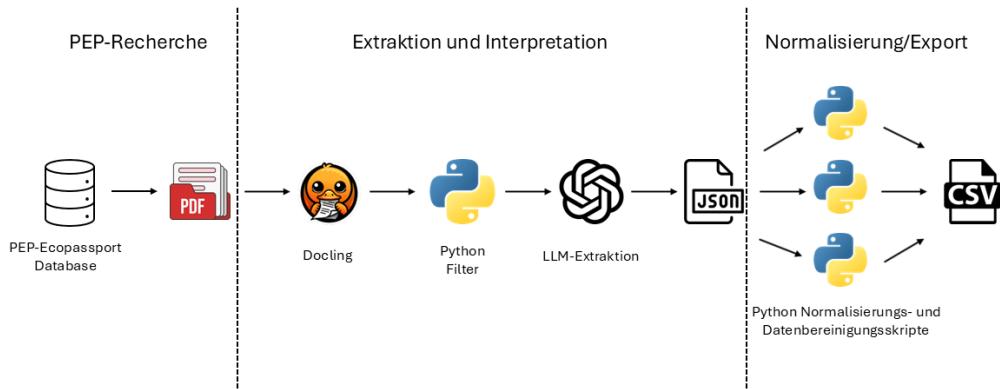


Abbildung 3.1: Schematischer Aufbau der Pipeline: von der PEP-Erfassung bis zur strukturierten Datenbasis. [Eigene Darstellung]

3.2. Datenerhebung und PEP-Recherche

Ziel der Datenerhebung ist die Identifikation von PEP-Ecopassport-Dokumenten, die sich auf Geräte der Gebäudeautomatisierung oder IoT-Komponenten beziehen. Die PEP-Datenbank bildet dabei die Quelle der Untersuchung. Für jedes gefundene Dokument wurden Produktinformationen, Material- und Energiedaten sowie Metadaten in strukturierter Form erfasst [Ass25].

Im ersten Schritt wurden die Dokumente automatisiert recherchiert, die zugehörigen PDFs heruntergeladen, analysiert und nach Relevanz für den Smart-Home- bzw. IoT-

Bereich klassifiziert. Das Ergebnis wurde in CSV-Dateien gespeichert und diente als Grundlage für die weitergehende Analyse.

Zur Ermittlung der verfügbaren PEP-Dokumente wurde zunächst die Suchfunktion der PEP-Datenbank analysiert. Über die Browser-Entwicklertools konnten die zugrunde liegenden Netzwerkanfragen identifiziert werden. Dadurch konnte mithilfe von *JavaScript* HTML-Snippets der Produkte mit den gesuchten Daten abgefragt, geparsst und in CSV-Dateien gespeichert werden. Um dabei gezielt IoT-relevante Produkte zu erfassen, wurden die Abfragen mit spezifischen Suchbegriffen in den Parametern erweitert (z. B. *controller, sensor, gateway, wifi, knx, zigbee, cloud*). Anschließend erfolgte eine manuelle Prüfung und Klassifikation der Ergebnisse, da die Suchbegriffe im Produkttitel nicht direkt auf Komponenten der Gebäudeautomation schließen lassen. Zudem wendet die Suchfunktion der PEP-Plattform nicht alle Filter korrekt an und die Produkte auf verschiedenen Ergebnisseiten überschneiden sich teilweise.

Die ermittelten Produkte wurden in einer Excel-Datei manuell kategorisiert. Die Zuordnung erfolgte nach folgenden Kriterien:

- **Gebäudeautomatisierung:** Geräte mit klarer Konnektivität (z. B. ZigBee, WiFi, KNX) oder Cloud-Anbindung, wie Gateways, smarte Sensoren oder Steuerungen.
- **Keine IoT-Relevanz (Aussortiert):** Produkte ohne Kommunikationsfähigkeit (z. B. Kabel, Trafos, LED-Panels).

Zusätzlich zur halbautomatisierten Suche wurden IoT-relevante Unternehmen gezielt identifiziert (z. B. ABB, Siemens, Schneider Electric, Legrand, Somfy, Daikin, Bosch, Honeywell). Deren PEP-Dokumente wurden manuell durchsucht und ergänzt. Dieses Vorgehen ist aufwändiger als die automatisierte Suche über Netzwerkabfragen, konnte die Datenbasis aber deutlich vergrößern.

3.3. PDF-Parsing und Extraktion

Die Extraktion strukturierter Daten aus PEP-PDFs stellt einen technisch anspruchsvollen Teil der Arbeit dar. Ziel ist es, aus den heterogenen Dokumenten eine konsistente, maschinenlesbare Repräsentation der Umweltindikatoren, Gerätedaten und Metadaten zu erzeugen, welche entsprechend der Zielsetzung der Arbeit analysiert werden können. Die finale Lösung kombiniert eine robuste Layoutanalyse mit Docing und eine LLM-basierte, schemagesteuerte Inhaltsinterpretation.

Zu Beginn wurde eine auf `pdfplumber` basierende Pipeline eingesetzt, die auf der von Selg [Sel25] entwickelten Pipeline aufbaut. Sie erkennt mithilfe von Regex- und

Textheuristiken Tabellen und Materialisten. Obwohl dieser Ansatz für einzelne PDFs funktionierte, erwies sich die Übertragbarkeit als unzureichend. Ursache waren typische Strukturprobleme von PDF-Dateien: eine verzerrte Zeilen- und Wortreihenfolge im Textlayer, stark variierende Layouts, Tabellen als Rasterbilder sowie uneinheitliche Bezeichnungs- und Einheitenformate. Bereits kleine Abweichungen in Tabellenköpfen führten zu fehlerhaften Zuordnungen von Indikatoren oder Spalten.

Die Vielzahl individueller Ausnahmen entwickelte sich zu einem unübersichtlichen Netz von abzufangenden Ausnahmefällen, das neue Konflikte zwischen bestehenden und neu hinzugefügten Layouts verursachte. Auch die manuelle Ergänzung einzelner Werte ist bei der erforderlichen Datenmenge in dieser Arbeit nicht mehr praktikabel. Eine vollständige Generalisierung des pdfplumber-Parsers war im Rahmen der Arbeit nicht realistisch umsetzbar.

Diese Limitierungen führten zur Entwicklung einer neuen, modularen Pipeline, die auf dem Open-Source-Framework *Docling* von IBM basiert. Docling erlaubt die strukturierte Segmentierung von PDF-Inhalten in Absätze, Tabellen, Listen und Bilder und exportiert diese in Markdown oder JSON. Dadurch konnte die textuelle Logik vom Layout entkoppelt und die Zuverlässigkeit der Verarbeitung deutlich verbessert werden.

Die Pipeline trennt klar zwischen Layoutanalyse und Inhaltsinterpretation:

- **Docling-Konvertierung:** PDF-Dateien werden in eine Markdown-Struktur überführt. OCR und Bildbeschreibung sind deaktiviert, um Laufzeit und Speicherverbrauch zu reduzieren. Tabellen- und Abschnittsgrenzen bleiben erhalten.
- **Regelbasierter Filter:** Um Kontextverluste des nachgelagerten Sprachmodells zu vermeiden, wurde ein regelbasierter Python-Filter auf die aus Docling generierten Markdown-Dateien angewendet. Irrelevante Segmente (z. B. Kopf- und Fußzeilen, Unternehmensinformationen, Titelblätter) werden über eine Blacklist entfernt.
- **LLM-basierte Extraktion:** Der konvertierte Text wird in Abschnitten an ein Sprachmodell übergeben, das definierte Variablen extrahiert und im JSON-Format zurückgibt. Die Promptstruktur erzwingt strikte Datentypen und klare Feldbezeichnungen.

Für die semantische Extraktion wurde *GPT-5* verwendet, angesprochen über die *Responses-API*. Diese neue Schnittstelle unterstützt strukturierte Ausgaben und optional eine Schema-Validierung. Durch einen Aufruf im `response_format=json_object`-Modus, werden nur gültige JSON-Formate geliefert, was den Post-Processing-Aufwand erheblich senkt. GPT-5 konnte Numerische Werte mit zugehörigen Einheiten stabil erkennen und Module (A1–A3, A4, A5, B*, C*, D) zuverlässig zuordnen. Die Temperatur des LLMs wurde auf 0 gesetzt, um Zufälligkeit und Kreativität möglichst zu vermeiden.

Die Kombination aus Docling und GPT-5 führte somit zu einem skalierbaren Verfahren, das auch bei komplexen Layouts konsistente Ergebnisse liefert.

Die neue Pipeline konnte die Anzahl fehlerhafter oder unvollständiger Einträge deutlich reduzieren. Für PDFs mit reinen Rastertabellen bleibt jedoch eine Einschränkung bestehen, da ohne OCR keine Inhaltsextraktion möglich ist. Der Einsatz eines LLMs führt aufgrund der stochastischen Modellkomponenten zudem zu einer eingeschränkten Reproduzierbarkeit und Transparenz. Obwohl das Risiko minimiert wurde, könnte es vorkommen, dass identische Eingaben aufgrund von Halluzinationen nicht immer identische Ausgaben liefern. Diese Einschränkungen sind angesichts der PDF-Heterogenität nicht zu umgehen und methodisch vertretbar.

3.4. Normalisierung der Daten

Nach der Extraktion lag ein heterogener Datensatz mit uneinheitlichen Bezeichnungsformen für Länder, Materialien, Lebenszyklusphasen, Energiequellen und Einheiten vor. Um eine konsistente Auswertung zu ermöglichen, wurden sämtliche Schreibweisen vereinheitlicht. Ziel war es, strukturell vergleichbare Werte zu schaffen und gruppierte Analysen über mehrere PEPs hinweg zu ermöglichen.

Zur systematischen Erfassung der vorhandenen Begriffe wurde ein Hilfsskript entwickelt, das die in den JSON-Dateien vorkommenden Rohwerte systematisch zusammenfasst. Für jedes relevante Feld werden die Häufigkeiten einzelner Strings erfasst und als Übersichtstabellen ausgegeben.

Für jede Datendomäne wurde darauf basierend eine Zuordnungstabelle (*Mapping-Datei*) erstellt, die reguläre Ausdrücke den vereinheitlichten Standardbegriffen zuordnet. Vereinheitlicht wurden:

- **Einheiten** (z. B. *kg CO₂-eq*, *kg CO₂e* und *kg CO₂ equiv* zu *kg CO₂ eq*),
- **Materialien** (z. B. *Paper*, *Cardboard* und *Carton* zu *Paper*),
- **Lebenszyklusphasen** (z. B. *A1 bis A3* und *Manufacturing* zu *manufacturing*),
- **Strommixe** (z. B. *France grid mix*, *France Mix* und *French grid* zu *FR*).

Diese Mappings wurden schrittweise verfeinert, bis alle Rohwerte abgedeckt waren.

Bei der Vereinheitlichung wurden einige pragmatische Entscheidungen getroffen: *Steel* und *Iron* wurden beispielsweise zur Kategorie *Steel* zusammengefasst, *Carton* und *Cardboard* werden zu *Paper* zusammengefasst, da beide ähnliche Materialeigenschaften aufweisen. Für Kunststoffe wurde eine vereinfachte Zusammenführung vorgenommen.

Ein Python-Skript wendet diese Mapping-Dateien auf alle extrahierten JSON-Dateien an.

Doppelte Materialeinträge, die durch das Mapping innerhalb eines Produkts entstanden sind, werden, sofern sie identische Bezeichnungen aufweisen, mithilfe eines weiteren Python-Skripts zusammengeführt. Dabei werden Prozentangaben und Gewichtsangaben bei Dubletten addiert.

Durch die Vokabularanalyse und anschließende Normalisierung entsteht so ein standardisierter Datensatz mit konsistenten Schreibweisen und eindeutiger Begriffssystematik. Diese Vereinheitlichung bildet die methodische Grundlage für die nachfolgende quantitative Auswertung.

3.5. Datenbereinigung und Validierung

Nach der automatischen Extraktion und Normalisierung werden zuerst fehlerhafte oder unvollständige Datensätze mithilfe eines Python-Skripts ausgeschlossen. Ein Datensatz galt als unbrauchbar, wenn sämtliche Umweltindikatoren fehlten oder ausschließlich Nullwerte enthielten, oder wenn die zentralen Felder `total_weight`, `electricity_consumption`, `material_composition` und `energy_model` gleichzeitig leer waren. Diese Kriterien führten zur Aussortierung von 8 der insgesamt 252 Datensätze. Diese PEPs enthielten zwar Metadaten, jedoch keine quantitativen Werte und wurden daher nicht in die Analyse einbezogen.

Tabelle 3.1: Anzahl der PEP-Dokumente von der Recherche bis zum Analysedatensatz.

Schritt	Anzahl
Automatisierte Keyword-Suche (identifiziert)	+ 184
Aussortiert ohne Gebäudeautomatisierungsrelevanz	- 77
Manuell ergänzte PEPs	+ 145
Gesamtzahl recherchierter PEPs	252
Aussortiert als unbrauchbar oder fehlerhaft	- 18
Finaler Analysedatensatz	234

Die Datenbereinigung wurde durch automatisierte Prüfmechanismen begleitet. Dazu zählten Validierungen hinsichtlich fehlender oder ungültiger Einheiten, numerische Typfehler (z. B. String statt numerischer Wert) und Plausibilitätsprüfungen, etwa auf Null- oder Extremwerte bei zentralen Variablen. Ein zusätzlicher Kontrolllauf identifizierte Datensätze mit nicht plausiblen Summen (z. B. Summe der Materialanteile) oder Flatline-Indikatoren (identische Werte in allen Phasen), die von der weiteren Analyse ausgeschlossen wurden.

Mehrere alternative Ansätze wurden im Verlauf der Datenbereinigung geprüft und bewusst verworfen. Die Aktivierung von OCR für alle PDFs hätte den Aufwand und die Laufzeit erheblich erhöht, ohne die Datenqualität signifikant zu verbessern. Ebenso wurde auf ein vollautomatisches Mapping über Sprachmodelle verzichtet, da dieses zu unkontrollierten Korrekturen führte. Eine Hierarchisierung der Materialien (z. B. *Iron* als Unterkategorie von *Metals*) oder eine gesonderte Behandlung von Verpackungsmaterialien wurde aus Gründen der Vergleichbarkeit nicht umgesetzt.

Die bereinigten JSON-Dateien werden anschließend in ein flaches, tabellenbasiertes Format überführt. Während die ursprünglichen JSON-Strukturen sowohl menschen- als auch maschinenlesbar angelegt waren, wurde das Format nun in eine einheitliche, analysierbare Datenstruktur überführt, die sich für statistische Auswertungen und Visualisierungen eignet.

Nach Abschluss der Bereinigung standen 234 strukturierte Datensätze zur Verfügung. Diese bilden die Grundlage für die nachfolgende statistische Analyse. Die zentrale Datenbasis umfasst vereinheitlichte Material-, Energie- und Länderbezeichnungen sowie geprüfte Indikatorwerte, wodurch eine vergleichbare quantitative Auswertung der Umweltwirkungen ermöglicht wird.

4

Analyse der erarbeiteten Daten

Auf Grundlage der in Kapitel 3 beschriebenen Datenbasis wird im Folgenden die Analyse durchgeführt. Zunächst erfolgt eine deskriptive Analyse der erhobenen Daten, um einen Überblick über deren Struktur und Verteilungen zu gewinnen. Darauf aufbauend werden lineare Regressionsanalysen durchgeführt, um Zusammenhänge zwischen den Input-Variablen und den Umweltindikatoren zu ermitteln. Diese Analysen bilden die Grundlage für die in dieser Arbeit entwickelte Heuristik. Sie soll eine Abschätzung von Umweltindikatoren für Produkte ohne PEP-Ecopassport ermöglichen und trägt damit zur zentralen Zielsetzung der Arbeit bei.

4.1. Deskriptive Analyse der PEP-Daten

Wie bereits in Kapitel 2.3 erläutert, ist eine deskriptive Betrachtung der Daten ein notwendiger erster Schritt, um die Qualität und Aussagekraft des Datensatzes zu beurteilen. Ziel dieses Abschnitts ist es, einen Überblick über die Vollständigkeit der vorliegenden Daten zu geben. Dazu werden zunächst die Anteile der fehlenden Werte analysiert, gefolgt von einer Beschreibung der Input-Variablen Gesamtgewicht, Stromverbrauch, Materialzusammensetzung und Energiemodelle. Abschließend werden die Verteilungen der Umweltindikatoren untersucht, um erste strukturelle Muster und Auffälligkeiten innerhalb des Datensatzes zu identifizieren.

4.1.1. Vollständigkeit der Werte

Zur Bewertung der Datenvollständigkeit wurde der Anteil fehlender Werte pro Variable berechnet und in einem Balkendiagramm dargestellt (Abb. 4.1). Die Missingness umfasst sowohl Werte, die die Datenpipeline nicht extrahieren konnte, als auch Werte, die in den PEPs nicht berichtet werden.

Die Analyse zeigt deutliche Unterschiede zwischen den Indikatoren: Für *Wasserknappheit* fehlen rund 78 % der Werte, während mehrere weitere Indikatoren wie *Eutrophierung mariner Gewässer*, *Klimawandel (fossil, total)* und *Eutrophierung terrestrisch* Fehlstände von etwa 30 % aufweisen.

Für den in der Regression verwendeten Stromverbrauch (*electricity_consumption*) fehlen knapp 25 % der Werte. In diesen PEPs wird zwar ein Energienutzungsmodell beschrieben und eine Formel zur Berechnung des Verbrauchs angegeben, der tatsächliche Gesamtstromverbrauch über die Lebensdauer wird jedoch nicht als konkreter Zahlenwert ausgewiesen, sondern basiert auf externen Katalogdaten (z. B. Verlustleistung P_{use}). Diese Informationen stehen in der vorliegenden Datenpipeline nicht zur Verfügung und können daher nicht automatisch in *electricity_consumption* überführt werden. In der weiteren Analyse stehen somit nur die PEPs mit explizit angegebenem Stromverbrauch zur Verfügung, was den Stichprobenumfang für die Regression reduziert.

Ein Sonderfall betrifft die Indikatoren, die sich auf den Wasserverbrauch beziehen. In den PEP-Daten werden hierfür nicht durchgängig dieselben Indikatorbezeichnungen berichtet. Ein Teil der Dokumente beinhaltet den Indikator *Water use*, während andere stattdessen *Water scarcity* enthalten. Dadurch entsteht in der Vollständigkeitsanalyse zunächst der Eindruck fehlender Werte, obwohl in vielen Fällen lediglich ein alternativer, methodisch anders definierter Indikator vorliegt.

Diese Indikatoren sind nicht ohne Weiteres miteinander vergleichbar, wenn sie auf unterschiedlichen Versionen des *Product Environmental Footprint* (PEF) basieren. Konkret wurden PEPs identifiziert, die auf PEF 3.0 und weitere, die auf PEF 3.1 beruhen. PEF 3.1 verwendet für die Wassermethode aktualisierte Charakterisierungsfaktoren, wodurch sich die resultierenden Indikatorwerte systematisch unterscheiden. Aus diesem Grund werden *Water use* und *Water scarcity* im Folgenden nicht zusammengeführt, und Analysen von *Water scarcity* aufgrund zu weniger Daten nicht durchgeführt.

Zusätzlich wurden die Erscheinungsjahre der PEP-Ecopassports untersucht. Die Veröffentlichungen reichen von 2020 bis 2025 und verteilen sich wie in Abb. 4.2 dargestellt. Der deutliche Anstieg ab 2022 zeigt die zunehmende Etablierung des Formats und eine stärkere Datenverfügbarkeit in den letzten Jahren.

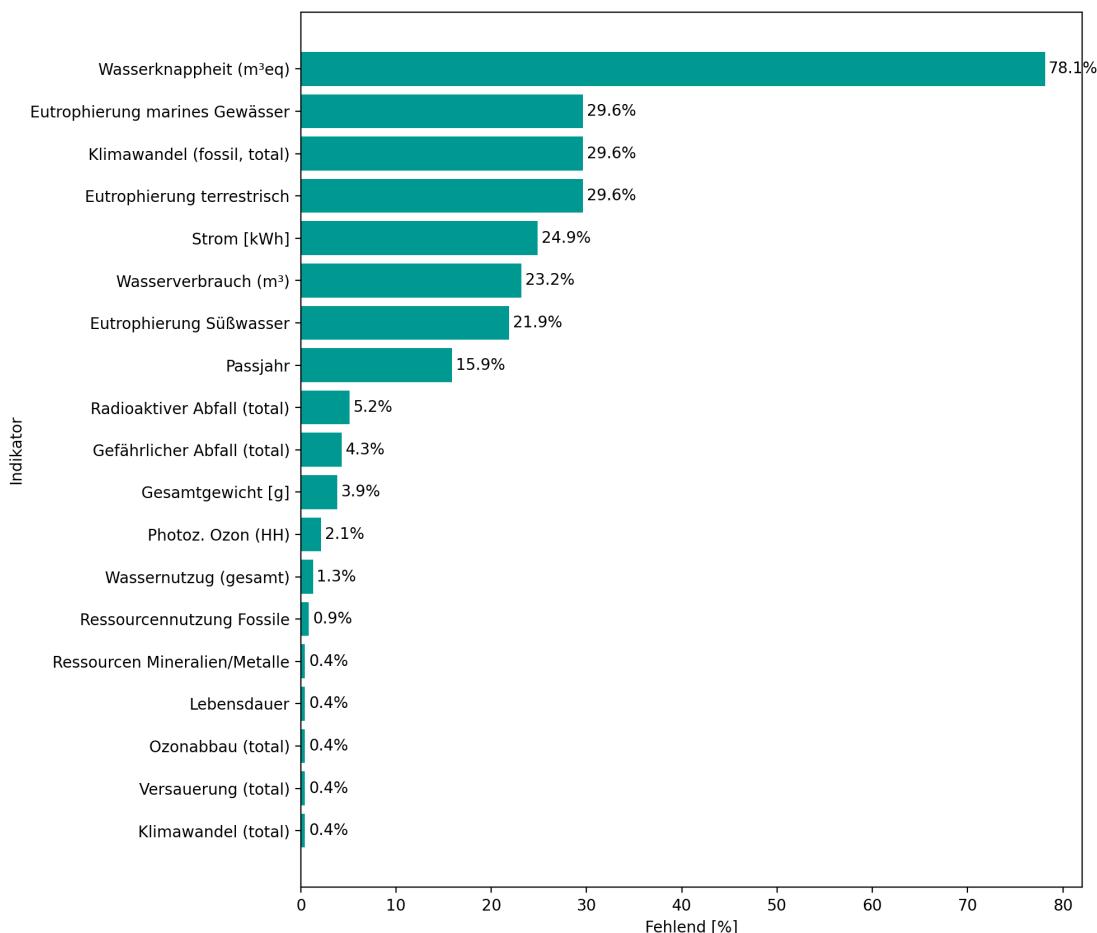


Abbildung 4.1: Anteil fehlender Werte pro Umweltindikator (%). N = 234 Produkte). [Eigene Darstellung]

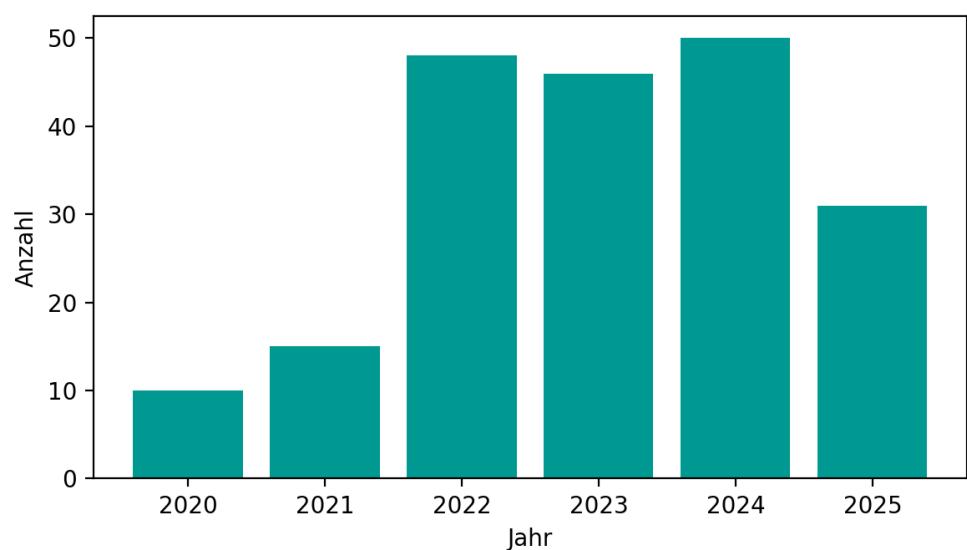


Abbildung 4.2: Erscheinungsjahre der analysierten PEP-Dokumente ($N = 233$ Produkte). [Eigene Darstellung]

4.1.2. Überblick der *Input*-Variablen

Für die Input-Variablen werden robuste Kennzahlen (**Median**, **IQR**) gezeigt und durch den **Mittelwert** ergänzt, um die Wirkung der Schiefe (v. a. Rechtsschiefe) zu verdeutlichen. Ausreißer werden nicht entfernt, ihre Einflüsse spiegeln sich im Mittelwert wider.

Variable	Einheit	Min	Median	Max	IQR	Mittelwert
Gesamtgewicht	kg	0.0395	2.178	13022.6	125.210	278.023
Stromverbrauch	kWh	0.026	326.511	8203569.5	86147.1	228061.654

Tabelle 4.1: Robuste deskriptive Kennzahlen der Basisvariablen.

Die Kennzahlen in Tabelle 4.1 beziehen sich jeweils auf die verfügbaren, nicht fehlenden Werte der entsprechenden Variable. Sie zeigen deutlich **rechtsschiefe Verteilungen** mit großen Interquartilsabständen (IQR). Beim *Gesamtgewicht* reicht die Spannweite von 0.04 kg bis über 13 000 kg, was die starke Heterogenität der betrachteten Produkte verdeutlicht. Das kleinste Produkt ist ein leichtes elektronisches Gerät, ein *Connected dimmer mit Bluetooth interface* (PEP-Link), während das größte Produkt, ein *Flüssigkeitskühler mit drehzahlgeregelter Schraubenverdichter und Greenspeed™-Technologie* (PEP-Link), mehr als 13 t erreicht. Der Mittelwert liegt mit 278 kg weit über dem Median (2.18 kg), was die ausgeprägte Rechtsschiefe bestätigt.

Auch der *Stromverbrauch* weist eine extreme Streuung auf (ca. 86147 kWh), mit Werten zwischen 0.026 kWh und über 8.2e6 kWh. Damit ist das kleinste Produkt nahezu stromlos im Betrieb, während das größte Produkt eine mehrjährige oder großtechnische Nutzung abbildet. Der Mittelwert (228000 kWh) übersteigt den Median (327 kWh) um mehrere Größenordnungen, was die starke Rechtsverschiebung der Verteilung verdeutlicht.

Abbildung 4.3 und Abbildung 4.4 ergänzen Tab. 4.1 um eine visuelle Einordnung. Beide Verteilungen sind auf Originalskala stark rechtsschief. Wenige sehr große Werte dominieren die Skala, wodurch der Großteil der Beobachtungen im linken Randbereich zusammengefasst wird. Um die typische Größenordnung dennoch sichtbar zu machen, ist die Originalskala in den Histogrammen auf einen Maximalwert begrenzt. Die Anzahl der nicht dargestellten Werte oberhalb dieser Grenze ist jeweils in der Grafik angegeben.

Die log1p Transformation reduziert diese Skalenprobleme, indem große Werte komprimiert und Unterschiede im unteren Bereich besser aufgelöst werden. Dadurch wird die Verteilung zwar besser interpretierbar, sie wird jedoch nicht symmetrisch oder normalverteilt. Für die nachfolgenden Regressionsmodelle ist dies auch nicht erforderlich. Entscheidend ist hier vor allem, dass die extremen Ausreißer im Datensatz weniger do-

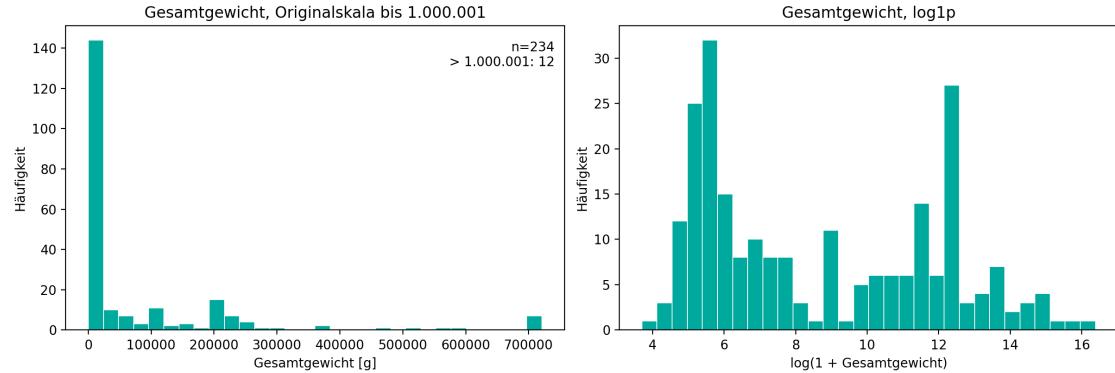


Abbildung 4.3: Histogramme des Gesamtgewichts auf Originalskala (links, bis zu 1.000.000 g) und nach log1p-Transformation (rechts). [Eigene Darstellung]

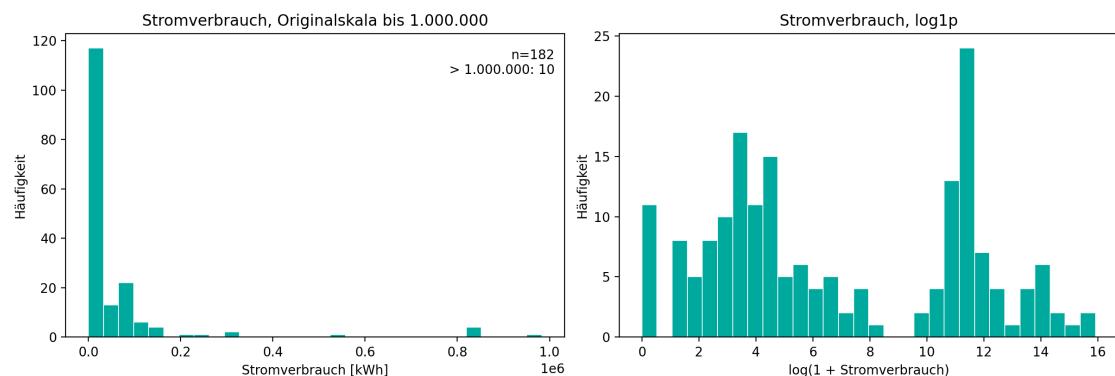


Abbildung 4.4: Histogramme des Stromverbrauchs auf Originalskala (links, bis zu 1.000.000 kWh) und nach log1p-Transformation (rechts). [Eigene Darstellung]

minieren und Beziehungen zwischen Variablen auf einer geeigneteren Skala modelliert werden können.

Die Zusammenhänge zwischen Stromverbrauch und Umweltauswirkungen können von der Art der Stromerzeugung abhängen. Da sich die Strommixe regional unterscheiden, variieren auch die resultierenden Emissionen je nach Herkunftsland des Energiebezugs.

Wie 4.5 zeigt, fällt der Großteil der verwendeten Energiemodelle auf allgemeine europäische Strommixe (EU27) und Frankreich. Besonders in den Phasen Nutzung und End-of-Life ist der Anteil europäischer Modelle deutlich höher. Dies liegt vermutlich daran, dass die Produkte häufig europaweit vertrieben und verwendet werden. Daher ist es schwierig, den tatsächlichen Energiebezug eines spezifischen Landes realistisch abzubilden, weshalb öfter ein europäischer Durchschnitt angenommen wird.

Der hohe Anteil von Frankreich ist auf eine große Anzahl an PEP-Dokumenten aus Frankreich zurückzuführen, die zu national vermarkteteten Produkten gehören. Von dort stammt die Association P.E.P und das Format ist dort am meisten etabliert.

Auch in der Herstellungsphase dominiert ein europäischer Energiemix, ergänzt durch

einzelne Modelle aus Deutschland und China, was auf internationale Produktionsketten hinweist. Insgesamt verdeutlicht die Verteilung, dass die meisten PEP-Deklarationen von europäischen Strommixen ausgehen, wodurch die berechneten Umweltauswirkungen tendenziell niedrigere fossile Anteile aufweisen, als es bei stärker kohleabhängigen Regionen (z. B. China) der Fall wäre.

Aufgrund der europäischen Prägung des Datensatzes ist die Aussagekraft der anschließenden Regression für Produkte außerhalb Europas eingeschränkt. Entsprechende Auswertungen werden mit erhöhter Unsicherheit und geringerer Datenqualität verbunden sein.

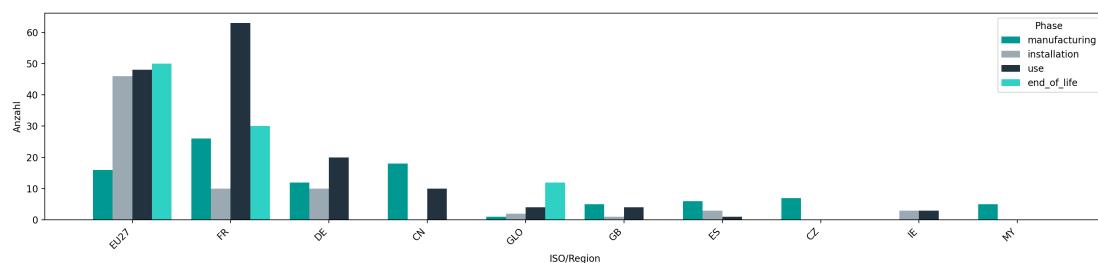


Abbildung 4.5: Verteilung der verwendeten Energiemodelle (ISO-Regionen) über die Lebenszyklusphasen. [Eigene Darstellung]

Eine weitere Variable, die die Umweltindikatoren stark beeinflussen, und damit in der zu entwickelnden Heuristik eine Rolle spielt, ist die Zusammensetzung des Produkts aus den verschiedenen Materialien. In der Tabelle 4.2 wird aufgeführt, aus welchen Materialien das durchschnittliche PEP-Produkt aus dem Datensatz besteht. Für die Mittelwerte wurde pro Produkt ein fehlendes Material als Anteil 0 % behandelt, sodass die Kennzahlen die durchschnittliche Zusammensetzung über den gesamten Datensatz abbilden. N gibt zusätzlich die Anzahl der Produkte an, in welchen das aufgeführte Material auftaucht. Wie in den meisten PEP-Dokumenten werden die modularen Materialien in die Gruppen *Metalle*, *Plastik* und *andere* gegliedert.

Die in Tabelle 4.2 dargestellten Materialanteile zeigen eine sehr heterogene Zusammensetzung der untersuchten Produkte. Mit durchschnittlich rund 26 % ist *Stahl* das mengenmäßig dominierende Einzelmaterial, gefolgt von *Papier* (15.7 %), welches vor allem für Verpackungen verwendet wird, und *Polycarbonat (PC)* (8.2 %). Während Metalle in nahezu allen PEPs vertreten sind, treten bestimmte Kunststoffe und Spezialmaterialien (z. B. PMMA, PBT, PPS) nur in wenigen Fällen auf. Die Kategorie *Andere* enthält zahlreiche kleinvolumige Komponenten, deren summiertes Anteil jedoch nicht vernachlässigbar ist. Insgesamt spiegelt sich in der Verteilung die Diversität der erfassten Produktgruppen wider.

Tabelle 4.2: Durchschnittliche Materialanteile über alle Produkte, gruppiert nach Hauptkategorien (Mittelwert in %).

(a) Metalle			(b) Kunststoffe			(c) Andere		
Material	Mittelwert	n	Material	Mittelwert	n	Material	Mittelwert	n
Stahl	26.46	199	Polycarbonat (PC)	8.23	113	Papier	15.73	197
Aluminium	6.13	140	ABS	2.84	105	Elektronik	3.69	102
Kupfer	5.41	161	Polyamid (PA)	2.37	118	Holz	3.12	81
Messing	0.86	82	PVC	2.08	86	Glas	2.90	67
Zamak	0.45	15	PS	1.13	62	PCBA	1.79	24
Nickel	0.10	12	PP	0.80	75	PCB	1.27	49
Zinn	0.06	20	Gummi	0.69	70	Kabel	0.35	38
Zink	0.05	9	PMMA	0.66	19	Kältemittel	0.35	53
Bronze	0.01	5	Epoxydharz	0.60	42	Ferrit	0.28	38
Neodym	0.01	5	Polyesterharz	0.57	23	Elektromotoren	0.27	9
Hartlot	0.00	7	PE	0.44	68	Lack / Farbe	0.15	39
			PU	0.40	40	Tinte	0.08	15
			PBT	0.23	17	Silizium	0.08	9
			PET	0.13	25	Batterie	0.08	10
			POM	0.09	16	Thionylchlorid	0.08	5
			TBBPA	0.07	9	Öl	0.07	8
			HIPS	0.06	4	Mineralwolle	0.06	13
			Silikon	0.04	6	Bitumen	0.04	7
			EPDM	0.02	5	Titandioxid	0.04	14
			PPS	0.02	5	Quarz	0.02	7
			Sonstige	0.03	—	Flussmittel	0.02	6
						Filz	0.01	11
						Aluminiumoxid	0.01	5
						Haftkleber	0.01	4
						Sonstige	0.48	—

4.1.3. Überblick der Umweltindikatoren

Die Umweltindikatoren bilden die Output-Variablen, auf deren Basis später die Heuristik entwickelt wird. Eine deskriptive Betrachtung verdeutlicht bereits die Verteilungsstruktur der Daten.

Wie Tabelle 4.3 zeigt, weisen alle Umweltindikatoren deutlich rechtsschiefe Verteilungen auf. Der Mittelwert liegt bei allen Größen um ein Vielfaches über dem Median. Besonders ausgeprägt ist die Schiefe bei *Climate change (total)*, *Resource use (fossils)*, *Water use* und *Hazardous waste disposed*.

Abbildung 4.6 ergänzt diese Zusammenfassung, indem sie die Streuung innerhalb der Indikatoren sowie die Ausreißerstruktur in Boxplots auf logarithmischer Skala zeigt. Bei den meisten Indikatoren erstreckt sich die Verteilung über mehrere Größenordnungen. Alle Indikatoren weisen zahlreiche Ausreißer nach oben auf. Die logarithmische Skala komprimiert große Werte, weshalb extrem hohe Beobachtungen im Plot optisch näher zusammenrücken, obwohl sie sich in den Originaleinheiten stark unterscheiden.

Insgesamt bestätigt sich eine heterogene, rechtsschiefe Datenbasis mit einigen Produkten, die sehr hohe Umweltwirkungen aufweisen.

Indikator (total)	Min	Median	Max	IQR	Mean	Einheit
Acidification	1.70×10^{-5}	4.30×10^{-1}	3.65×10^3	1.03×10^1	1.11×10^2	kg SO ₂ eq
Climate change (total)	3.10×10^{-3}	8.68×10^1	1.04×10^6	1.98×10^3	2.27×10^4	kg CO ₂ eq
Eutrophication (freshwater)	1.00×10^{-6}	2.66×10^{-2}	2.36×10^2	3.14×10^{-1}	2.62	kg P eq
Hazardous waste disposed	1.00×10^{-4}	3.93×10^1	4.89×10^5	5.97×10^2	6.44×10^3	kg
Ozone depletion	0	7.00×10^{-6}	1.92×10^{-1}	2.86×10^{-4}	3.23×10^{-3}	kg CFC-11 eq
Photochemical ozone formation (HH)	2.00×10^{-6}	1.81×10^{-1}	1.41×10^3	3.13	4.06×10^1	kg C ₂ H ₄ eq
Resource use (fossils)	3.26×10^{-2}	1.62×10^3	1.06×10^8	9.51×10^4	1.58×10^6	MJ
Resource use (minerals/metals)	1.00×10^{-6}	3.92×10^{-3}	5.87	4.95×10^{-2}	2.28×10^{-1}	kg Sb eq
Radioactive waste disposed	0	6.56×10^{-2}	3.26×10^3	5.38×10^{-1}	2.26×10^1	kg
Water use	9.30×10^{-5}	4.24×10^1	5.77×10^6	3.84×10^2	9.88×10^4	m ³

Tabelle 4.3: Gesamtindikatoren (Total) mit Median, IQR und Mittelwert (gerundet auf zwei Nachkommastellen).

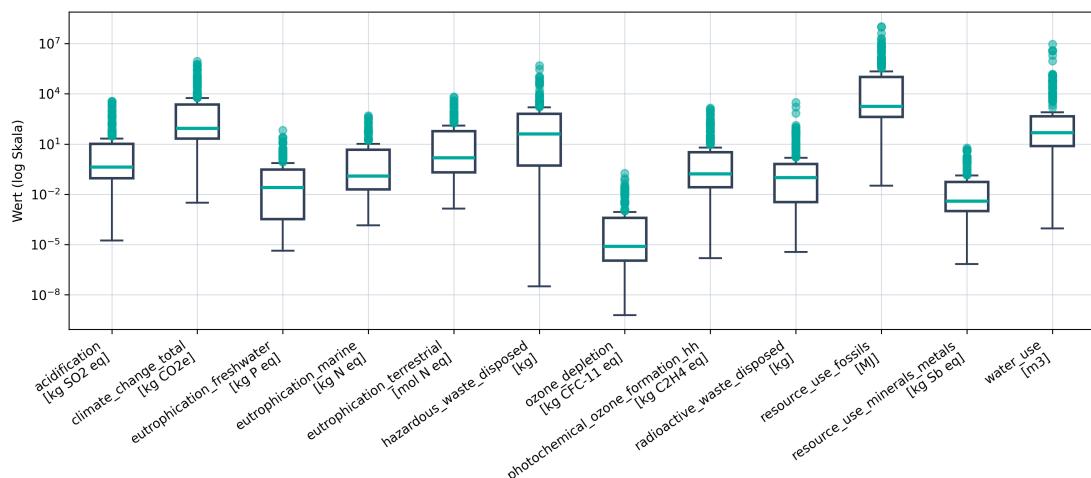


Abbildung 4.6: Boxplots der Total-Werte je Indikator auf logarithmischer Skala. [Eigene Darstellung]

4.2. Explorative Modellentwicklung

Dieses Kapitel beschreibt die Wahl des Regressionsmodells. Im Mittelpunkt steht die explorative Entwicklung und Auswertung verschiedener Modellvarianten, aus denen das in Abschnitt 4.4 eingesetzte Hauptmodell für die CO₂-Äquivalente und die übrigen Indikatoren abgeleitet wird.

Die Modellentwicklung erfolgte iterativ und datengetrieben: Unterschiedliche Feature-Sets, verschiedene PCA-Varianten sowie alternative lineare Schätzer wurden ausprobiert und anhand einheitlicher Gütemaße bewertet. Ziel ist es nachvollziehbar zu machen, welche Kombinationen sich in der Praxis als robust erweisen und welche Ansätze verworfen wurden.

Vor der Modellschätzung wurden die kontinuierlichen Eingangsvariablen standardisiert. Hierfür wurde der StandardScaler aus *scikit-learn* verwendet, der jedes Feature gemäß $x' = (x - \mu_{\text{train}})/\sigma_{\text{train}}$ transformiert. Dabei werden Mittelwert μ_{train} und Standardabweichung σ_{train} ausschließlich auf dem Trainingsdatensatz geschätzt und anschließend unverändert auf Validierungs- und Testdaten angewendet (`with_mean=True`, `with_std=True`). Im *Basisblock* betrifft dies `log_w` und `log_e`. Bei der globalen PCA wird die Standardisierung entsprechend auf alle Regressoren gemeinsam angewendet.

4.2.1. Experimentelle Fragestellungen

Ausgehend von der Zielsetzung, mit möglichst wenigen und robust erfassbaren Merkmalen brauchbare Vorhersagen zu erhalten, leiten sich für die explorative Modellentwicklung insbesondere folgende Fragestellungen ab:

1. **Beitrag der Materialinformationen:** Verbessert sich die Vorhersagegüte gegenüber einem Modell, das nur Gewicht und Stromverbrauch nutzt, wenn zusätzlich Materialinformationen einbezogen werden?
2. **Rohmaterialien vs. Material-PCA:** Ist es günstiger, die zahlreichen Materialspalten direkt zu verwenden, oder führt eine PCA des Materialblocks zu stabileren Modellen?
3. **Beitrag der Energiemodelle:** Verbessert sich die Vorhersagegüte, wenn zusätzlich Energiemodelle einbezogen werden?
4. **Wahl des Regressionsverfahrens und Regularisierung:** Unterscheiden sich OLS, Ridge und Lasso hinsichtlich Stabilität und erzielbarer Gütemaße bei den vorhandenen PEP-Daten?

5. **Wahl der Zieltransformation:** Führt eine Log-Transformation (\log_{10}) oder eine Box-Cox-Transformation zu besserer Vorhersagegüte und plausibleren Residuen?

Die nachfolgenden Abschnitte stellen die dafür durchgeführten Experimente vor und leiten aus den beobachteten Unterschieden einfache Heuristiken für die weitere Modellierung ab.

4.2.2. Vergleich der Feature-Sets

Als erster Schritt der experimentellen Modellentwicklung wurde untersucht, welchen Beitrag unterschiedliche Eingangsmerkmale zur Vorhersage von *Climate change (total)* leisten. Grundlage sind hierbei $n = 173$ PEPs, für die Gesamtgewicht, Stromverbrauch und die CO₂-Äquivalente vollständig vorliegen.

Verglichen wurden mehrere Feature-Sets, jeweils in einem linearen Regressionsmodell auf der Transformationsskala $\log(1 + \text{CO}_2_{\text{total}})$:

- **Basis:** nur Gewicht und Stromverbrauch.
- **Basis + Energiemodelle:** Gewicht und Stromverbrauch + Energiemodelle.
- **Basis + Rohmaterialien:** zusätzlich alle ausgewählten Materialien als separate Regressoren.
- **Basis + Rohmaterialien mit Minimalvorkommen:** Materialien, die in mehr als 10 PEPs vorkommen, als separate Regressoren.
- **Basis + PCA-Materialien:** statt der Rohmaterialsäulen werden k Hauptkomponenten aus einer PCA auf dem Materialblock verwendet (Varianzschwelle 90 %).
- **Globale PCA (Basis + Rohmaterialien mit Minimalvorkommen):** Die PCA wird auf alle Features global angewendet.

Für alle Varianten wurden dieselben Train/Test-Aufteilungen verwendet, so dass die Gütemaße direkt vergleichbar sind. Tabelle 4.4 zeigt die erzielten Ergebnisse.

Das Basismodell aus Gewicht und Stromverbrauch erklärt bereits einen großen Anteil der Varianz der CO₂-Äquivalente.

Obwohl die Einbeziehung des verwendeten Energiomodells intuitiv eine höhere Modellgüte erwarten lässt, liegt die Güte weit unter der des Basismodells. Ein plausibler Grund ist die starke Dominanz weniger Strommixe, insbesondere des europäischen und französischen Mixes (siehe Abb. 4.5). Für die anderen Energiemodelle sind in den betrachteten PEPs zu wenig Datenpunkte um deren Effekt zuverlässig zu schätzen.

Tabelle 4.4: Vergleich verschiedener Feature-Sets für den Indikator *Climate change (total)* auf der Skala $\log(1 + \text{CO}_2_{\text{total}})$. RMSE nach Rücktransformation

Modellvariante	R^2_{Train}	R^2_{Test}	RMSE _{Test}
Basis (Gewicht, Strom)	0.817	0.770	50104.33 kg CO ₂
Basis + Energiemodelle	0.810	0.501	159778.91 kg CO ₂
Basis + Rohmaterialien	0.904	0.567	101227.08 kg CO ₂
Basis + Rohmaterialien ($n \geq 10$)	0.842	0.832	41910.11 kg CO ₂
Basis + PCA-Materialien ($n \geq 10$)	0.887	0.882	33106.20 kg CO ₂
PCA auf alle Variablen	0.822	0.738	53991.70 kg CO ₂

Ein weiterer Grund ist die unzureichende Abdeckung der Jahreszahlen der benutzten Strommixe in den PEPs. Sie sind nur in wenigen Datensätzen vorhanden und können daher nicht robust als erklärende Variable modelliert werden. Gleichzeitig kann sich die Stromerzeugungsstruktur zwischen Jahren deutlich verändern [24]. Wenn unterschiedliche Jahre im Datensatz vermischt werden, entsteht zusätzliche Varianz in den Zielwerten, die durch das benutzte Feature-Set nicht erklärt werden kann und sich als Rauschen im Modell zeigt.

Die Erweiterung um Rohmaterialanteile erhöht zwar das Trainings- R^2 deutlich, bringt aber auf dem Testdatensatz keinen Mehrwert und verschlechtert R^2_{Test} und den RMSE_{Test} leicht. Dies ist ein Hinweis auf Überanpassung durch die vielen, teilweise korrelierten Materialvariablen.

Deutlich besser schneidet das Modell mit einem Minimalvorkommen an Materialien ab. R^2_{Test} ist hier durchschnittlich deutlich höher als im Basismodell.

Auffällig ist, dass die Variante *PCA auf alle Variablen* deutlich schlechter abschneidet als die material-spezifische PCA. Eine PCA über alle Variablen scheint die starken Basiseffekte von Gewicht und Strom mit den vielen, teils verrauschten Materialvariablen zu vermischen und damit genau diese klaren Zusammenhänge abzuschwächen.

Die PCA-Variante mit Material-Hauptkomponenten erreicht das höchste Test- R^2 und einen deutlich geringeren Test-RMSE als die anderen Modelle. Die Materialinformationen tragen also erkennbar zur Vorhersage bei, müssen dafür aber in verdichteter Form (PCA) in das Modell eingehen. Auf Basis dieser Experimente wurde das „Basis + PCA-Materialien“-Feature-Set als Ausgangspunkt für die weitere Modellentwicklung und den späteren Regressionsansatz gewählt.

4.2.3. Vergleich der Regressionsverfahren

Die im vorherigen Abschnitt beschriebenen Experimente zum Vergleich der Feature-Sets wurden für den Indikator *Climate change (total)* zunächst mit einer klassischen linearen Regression auf Basis der der `LinearRegression` aus `scikit-learn` nachgebildet.

Im nächsten Schritt wird das lineare Modell anschließend mit Ridge und Lasso implementiert, die über einen Regularisierungsparameter λ die Modellkomplexität steuern und dadurch insbesondere bei vielen, korrelierten Regressoren stabilere Schätzungen liefern können. In `scikit-learn` entspricht dies dem Parameter `alpha` ($\alpha \equiv \lambda$).

Bei festem Feature-Set (Gewicht, Stromverbrauch, PCA-Materialkomponenten) lieferten Ridge und Lasso über mehrere Wiederholungen des Experiments ähnliche R^2 - und RMSE-Werte. Beide Verfahren schnitten konstant besser ab als `LinearRegression`. Die Wahl des Regressors war allerdings, wie die Zahlen zur Modellgüte in 4.5 zeigen, damit deutlich bedeutend als der Einfluss des Feature-Sets.

Tabelle 4.5: Vergleich der Regressionsverfahren auf dem festgelegten Feature-Set.

Regressionsverfahren	R^2_{Train}	R^2_{Test}	$\text{RMSE}_{\text{Test}}$
LinearRegression	0.887	0.882	33106.20 kg CO ₂
Ridge	0.891	0.896	25117.31 kg CO ₂
Lasso	0.891	0.896	25116.20 kg CO ₂

Für die weiteren Analysen wurde das Ridge Modell verwendet, da es sich über mehrere Wiederholungen mit verschiedenen zufälligen Test-/Trainingssplits als minimal robuster und konstanter erwies.

4.3. PCA der Materialien

Zur explorativen Analyse und zur Reduktion der Dimensionalität der Materialdaten wurde eine Hauptkomponentenanalyse (PCA) durchgeführt. Benutzt wurde die Python-Bibliothek `scikit-learn` [25d]. Alle Materialien, die in einem Produkt nicht vorkommen, wurden als 0 interpretiert.

Vor der PCA wurden die Verteilungen der Materialanteile grafisch untersucht. Die Histogramme, von welchen die drei häufigsten Materialien in Abbildung 4.7 dargestellt sind, zeigen, dass die Daten teilweise multimodal verteilt sind und viele Beobachtungen

im Bereich sehr kleiner Anteile liegen. Bei einer einfachen z-Standardisierung wäre die Verteilung durch Ausreißer weiterhin stark beeinflusst.

Für die PCA der Materialien wird daher der *scikit-learn RobustScaler* [25e] verwendet. Dieser zentriert jede Materialspalte um den Median und skaliert durch den IQR, wobei die Quantile standardmäßig als $Q_{0.25}$ und $Q_{0.75}$ definiert sind (`quantile_range=(25, 75)`). Im Gegensatz zur z-Standardisierung wird damit die Skalierung deutlich weniger durch Ausreißer beeinflusst. Die PCA wird anschließend auf den robust skalierten Materialien durchgeführt. In den nachfolgenden Regressionsmodellen werden die Skalierungsparameter jeweils nur auf den Trainingsdaten geschätzt und anschließend auf Validierungs- bzw. Testdaten angewendet, um Data Leakage zu vermeiden.

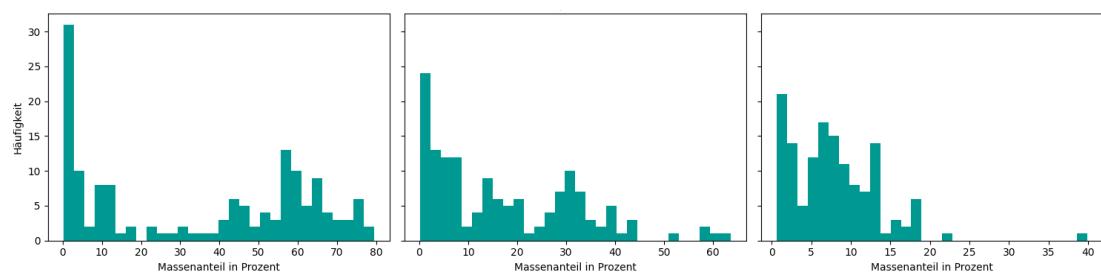


Abbildung 4.7: Histogramme der Materialien mit den größten Anteilen über alle Produkte. [Eigene Darstellung]

4.3.1. Ergebnis der PCA

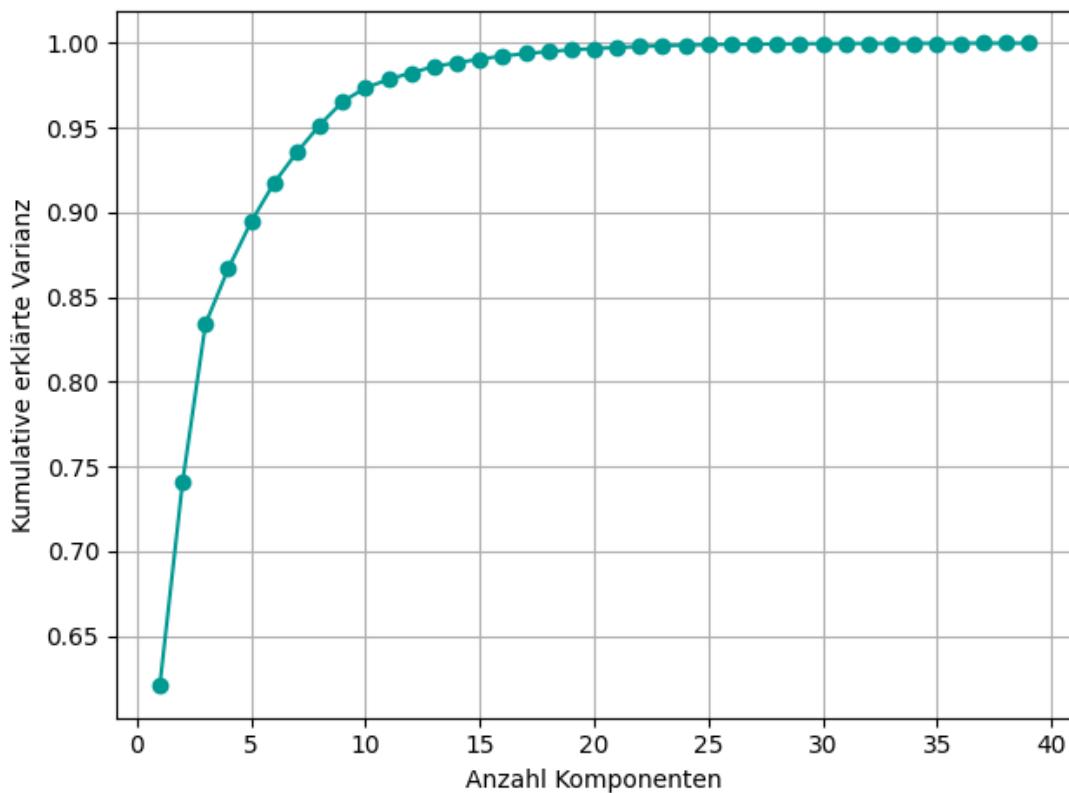


Abbildung 4.8: Kumulative erklärte Varianz. [Eigene Darstellung]

In Abbildung 4.8 ist die durch die Hauptkomponenten erklärte kumulative Varianz dargestellt. Die ersten 5 Hauptkomponenten erklären bereits ungefähr 90 % der Varianz. Ab etwa 20 Komponenten nehmen zusätzliche Komponenten nur noch geringe Varianzanteile auf. Für die weitere Modellierung wurde daher ein Varianzschwellenwert von 95 % gewählt, was in diesem Datensatz etwa $k = 8$ Hauptkomponenten entspricht. Dieser Wert ist ein Kompromiss zwischen Dimensionsreduktion und Informationsverlust.

Anzumerken ist, dass die hier dargestellte PCA auf dem vollständigen Datensatz basiert und nur der explorativen Darstellung der Materialien dient. Für die nachfolgende Modellschätzung wird die PCA jeweils ausschließlich auf dem Trainingsdatensatz gefittet und anschließend auf die Testdaten angewendet, um zu vermeiden, dass das Modell durch die PCA die Testdaten lernt.

4.3.2. Interpretation der Material-Hauptkomponenten

Jede Hauptkomponente ist eine lineare Kombination der einzelnen Materialanteile. Die zugehörigen *Loadings* geben an, wie stark ein bestimmtes Material zur jeweiligen Komponente beiträgt. Große Beträge der Loadings (unabhängig vom Vorzeichen) weisen auf Materialien hin, die diese Komponente besonders prägen. Das Vorzeichen bestimmt lediglich die Richtung der Achse (Zunahme vs. Abnahme eines Materials), ist für die inhaltliche Einordnung des Musters aber weniger wichtig als die absolute Größe.

Auf Basis der Top-Loadings pro Komponente lassen sich die ersten Hauptkomponenten wie folgt interpretieren:

- **PC₁ (*PS (Polystyrol)* dominiert):** Die erste Hauptkomponente wird fast vollständig durch ps geprägt, mit deutlich kleineren Beiträgen von pe und anderen Kunststoffen. Sie unterscheidet damit Produkte mit hohen Polystyrolanteilen (zum Beispiel bestimmte Gehäuse oder Schäume) von Produkten, bei denen PS kaum eine Rolle spielt.
- **PC₂ (*PE (Polyethylen)* versus übrige Kunststoffe):** PC₂ hat ein sehr hohes positives Loading auf pe, während other_plastics und ps eher negativ geladen sind. Diese Komponente beschreibt also eine Achse zwischen Produkten mit ausgeprägtem PE-Anteil und solchen, bei denen eher andere Kunststoffe oder unspezifische Kunststoffmischungen dominieren.
- **PC₃ (unspezifische Kunststoffmischungen):** In PC₃ dominiert other_plastics, ergänzt durch positive Beiträge von pe, electronics und other sowie leicht negative Beiträge von pvc. Diese Komponente steht für Produkte mit einem breiten Kunststoffmix und einem gewissen Elektronikanteil, die sich von eher PVC-basierten Gehäusen abgrenzen.
- **PC₄ (Leiterplatten und Elektronik versus PVC):** PC₄ wird stark durch pcba (Leiterplattenbestückung) geprägt, mit zusätzlichen positiven Beiträgen von Glas und Elektronik und einem deutlich negativen Loading auf pvc. Sie trennt damit Produkte mit hohem Leiterplatten und Elektronikanteil von solchen, bei denen PVC-Gehäusematerial im Vordergrund steht.
- **PC₅ (Elektronikschwerpunkt):** In PC₅ weist electronics das höchste positive Loading auf, während pvc, pcba, pe und other_plastics überwiegend negativ geladen sind. Diese Komponente beschreibt Produkte, bei denen Elektronikbauteile in der Masse dominieren und klassische Gehäuse und Strukturkunststoffe relativ weniger Gewicht haben.

Bemerkenswert ist, dass klassische Strukturmetalle wie Stahl oder Kupfer und Materialien wie Papier in den ersten Hauptkomponenten nicht mit den höchsten Ladungen auftreten. Das liegt daran, dass ihre Anteile über viele PEPs hinweg vergleichsweise stabil sind und dadurch weniger zur Gesamtvarianz beitragen als die stark schwankenden Kunststoff und Elektronikanteile. Ihr Einfluss verteilt sich daher auf spätere Hauptkomponenten mit geringerem Varianzanteil.

Insgesamt zeigt die Material-PCA, dass sich die sehr unterschiedlichen Materiallisten auf wenige dominante Muster verdichten lassen. Die ersten drei Komponenten beschreiben vor allem verschiedene Kunststoffmischungen (PS, PE und andere Kunststoffe), während PC₄ und PC₅ Leiterplatten und Elektronik gegenüber PVC lastigen Gehäusen abgrenzen. Diese fünf Hauptkomponenten erklären zusammen knapp 90 % der Varianz im Materialblock und bilden damit die prägendsten Materialmuster der PEPs ab. In den folgenden Regressionsmodellen werden sie genutzt, um den Einfluss des Materialmixes auf die Umweltindikatoren zu erfassen, ohne alle einzelnen Materialien separat berücksichtigen zu müssen.

4.4. Lineare Regression des Indikators *Climate change (total)*

Ziel der folgenden Analyse ist es entsprechend der Zielsetzung dieser Arbeit zu untersuchen, inwieweit sich die in den PEPs ausgewiesenen Treibhausgasemissionen (*Climate Change, total*) durch wenige, aus den Dokumenten verfügbare Produktmerkmale erklären lassen, die grundsätzlich auch für Produkte ohne PEP messbar sind. Im Fokus steht in diesem Kapitel ausschließlich der CO₂-Indikator und ein lineares Regressionsmodell. Weitere Umweltindikatoren werden in späteren Abschnitten betrachtet.

4.4.1. Datenbasis und Transformation

Für die Regression werden nur Datensätze berücksichtigt, bei welchen cc_total, total_weight und electricity_consumption vorhanden und positiv sind. Nach dieser Filterung verbleiben insgesamt $n = 173$ PEPs. Die Materialinformationen liegen als Massenanteile vor.

Die Verteilungen der CO₂-Äquivalente, des Produktgewichts und des Stromverbrauchs sind stark rechtsschief und decken mehrere Größenordnungen ab. Um den Einfluss extremer Werte zu verringern und die Größenordnungen besser vergleichbar zu machen, werden diese Variablen mit der Funktion log1p transformiert. Es werden die

folgenden Größen definiert:

$$\log_{\text{cc}} = \log(1+\text{CO2}_{\text{total}}), \quad \log_{\text{w}} = \log(1+\text{weight}), \quad \log_{\text{e}} = \log(1+\text{electricity_consumption})$$

Die lineare Regression wird auf der Transformationsskala von \log_{cc} durchgeführt. Bei Bedarf lassen sich die Vorhersagen über die inverse Funktion `expm1` wieder auf die Originalskala der Emissionen zurückführen.

4.4.2. Modellformulierung

Das endgültig betrachtete Modell nutzt drei Arten von erklärenden Variablen: das log-transformierte Gesamtgewicht, den log-transformierten, über die Lebensdauer aggregierten Stromverbrauch und verdichtete Materialinformationen aus einer PCA der Materialien. In der log-transformierten Skala hat das Modell die Form

$$\log_{\text{cc}} = \beta_0 + \beta_1 \cdot \log_{\text{w}} + \beta_2 \cdot \log_{\text{e}} + \sum_{j=1}^k \gamma_j \cdot \text{PC_mat}_j + \varepsilon,$$

wobei PC_mat_j die Material-Hauptkomponenten aus der PCA bezeichnen und k die Anzahl der verwendeten Komponenten ist. Die Materialanteile selbst werden nicht log transformiert. Sie liegen als Massenanteile vor, werden skaliert und anschließend per PCA zu Hauptkomponenten zusammengefasst. Diese Hauptkomponenten fassen jeweils ein charakteristisches Muster aus Materialanteilen zusammen (vgl. 4.3.2) und fungieren als verdichtete Materialindikatoren im Regressionsmodell. Der Fehlerterm ε umfasst alle nicht modellierten Einflüsse sowie Mess- und Rundungsfehler.

4.4.3. Schätzverfahren, Validierung und Ergebnisse

Zur Bewertung der Modellgüte wird ein Train/Test-Split mit einem Testanteil von 10 % verwendet. Das Modell wird ausschließlich durch die 90% Trainingsdaten angepasst und anschließend auf dem unabhängigen Testset ausgewertet. Als Regressor wird Ridge (L2-Regularisierung) verwendet, wobei der Regularisierungsparameter λ ausschließlich auf den Trainingsdaten durch Cross-Validation bestimmt wird. Auf diese Weise erhält man Gütemaße, die angeben, wie gut das Modell auf die nie zuvor gesehenen Testdaten generalisieren kann.

Die Modelle werden über das Bestimmtheitsmaß R^2 und den Root-Mean-Square-Error (RMSE) bewertet. Der RMSE wird nach Rücktransformation der Vorhersagen auf der Originalskala berichtet. Tabelle 4.6 fasst die Testgüte des CO₂ Regressionsmodells

zusammen.

Tabelle 4.6: Gütekennzahlen des linearen Regressionsmodells (Climate Change (total) als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.896
RMSE _{Test}	25116.20 kg CO ₂

Das Modell erklärt damit etwa 90 % der Varianz von log_cc auf dem Testset und lässt nur etwa 10 % unerklärt. Dieser Wert spricht für eine hohe Modellgüte.

Der absolute RMSE-Wert von 25116.20 kg CO₂ erscheint hoch. Dies ist vor allem eine Folge der größten Produkte im Datensatz. Abweichungen der Schätzung bei großen Produkten dominieren den RMSE, da dieser einzelne große Fehler quadratisch stärker gewichtet.

Zur Illustration, wie stark einzelne Produkte die Fehlermaße beeinflussen können, wird der größte Fehler im Testset separat betrachtet. Das Produkt mit der größten Abweichung ist:

- **Produkt:** Daikin Applied Europe SpA Wärmepumpe [25c]
- **Tatsächlicher Wert:** 266000.0 kg CO₂
- **Vorhersage:** 37447.76 kg CO₂
- **Absoluter Fehler:** 228552.24 kg CO₂
- **Relativer Fehler:** 85.9%

Dieses Beispiel verdeutlicht, dass einzelne sehr große Produkte einen großen absoluten Fehler verursachen können und damit einen überproportionalen Einfluss auf RMSE-basierte Kennzahlen haben, obwohl sich der relative Fehler in Grenzen hält. Daher werden im Folgenden neben R^2 und RMSE auch robuste und relative Fehlermaße berichtet, um die Modellgüte über verschiedene Größenordnungen hinweg nachvollziehbar zu interpretieren.

Folgende Maße werden definiert:

- **Median abs. Fehler (MdAE):** MdAE = median($|y_i - \hat{y}_i|$).
- **MdARE (Median relative Errors):** MdARE = median($\frac{|y_i - \hat{y}_i|}{y_i}$) (für $y_i > 0$).
- **MARE (Mean relative Errors):** MARE = $\frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}$ (für $y_i > 0$).

Der MdAE wird auf Originalskala in der jeweiligen Einheit angegeben. MdARE und MARE sind einheitenlos und können als Anteil interpretiert werden (z. B. $0.66 \approx 66\%$).

Der MdAE liegt bei 2220.80 kg CO_2 und beschreibt die typische absolute Abweichung eines repräsentativen Produkts. Der MdARE liegt bei 51.49%. Der MARE beträgt 120.27% und fällt deutlich höher aus, was auf einzelne sehr große relative Abweichungen hinweist. Diese Kombination aus Median und Mittelwert zeigt, dass die Fehlerverteilung durch Ausreißer geprägt ist und eine rein RMSE-basierte Interpretation die typische Modellleistung verzerren kann.

Eine mittlere relative Abweichung von 120% bedeutet, dass das Modell die Größenordnung der verursachten CO_2 -Äquivalente im Mittel korrekt einordnet. Eine präzise Abschätzung über alle Produkte hinweg ist mit dem verwendeten, bewusst kompakten Feature-Set nur eingeschränkt möglich.

Eine weitere Auswertung der größten absoluten Abweichungen zeigt, dass diese vor allem bei besonders schweren Produkten auftreten. Die fünf stärksten Ausreißer stammen aus PEPs mit einem Gesamtgewicht von mindestens 720 kg.

4.4.4. Visualisierung der Vorhersagequalität

Zur Veranschaulichung der Modellgüte zeigt Abbildung 4.9 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von Climate Change total. Trainings- und Testdaten werden getrennt dargestellt, und eine Diagonale $y = x$ markiert die ideale Übereinstimmung zwischen Vorhersage und Realität.

Die meisten Punkte liegen nahe der Diagonalen, und die Streuung ist für Trainings- und Testdaten ähnlich. Dies passt zu den ausgewiesenen Gütemaßen und spricht dafür, dass das Modell die Daten abbildet, ohne stark zu überanpassen. Gleichzeitig ist in einigen Wertebereichen sichtbar, dass der Zielwert tendenziell unter- (z. B. Cluster bei 10^4 bis 10^5 kg) bzw. überschätzt (z. B. Cluster bei 10^3 , kg) wird.

Für die OLS Theorie wird angenommen, dass die Fehler auf der Modellskala näherungsweise normalverteilt sind. Für die hier verfolgte Vorhersage und Fehleranalyse ist diese Annahme jedoch nicht zwingend erforderlich, dient aber als diagnostischer Hinweis.

Abbildung 4.10 zeigt einen QQ Plot der Residuen auf der Originalskala, nach Rücktransformation verglichen mit den theoretischen Quantilen der Normalverteilung. Hier sind deutliche Abweichungen von der Referenzgeraden sichtbar, insbesondere in den Rändern, was auf Schiefe und schwere Verteilungsschwächen hindeutet. Dies ist plausibel, da auf Originalskala einzelne sehr große Produkte die Fehler dominieren und Fehler häufig multiplikativ wirken, was auf Originalskala zu stark asymmetrischen Residuen führt.

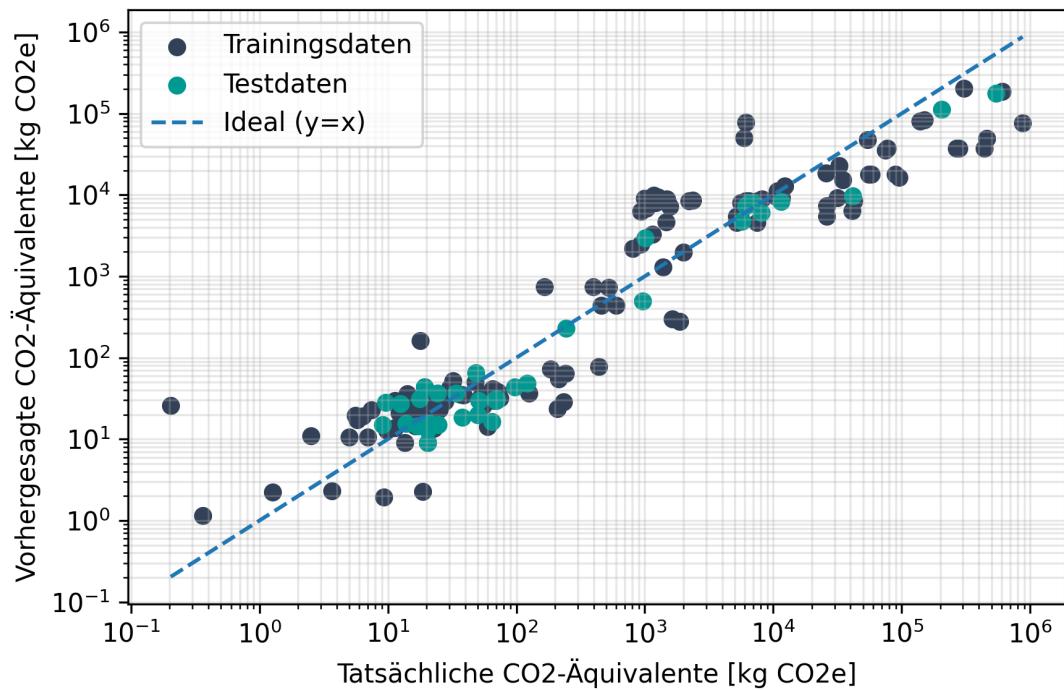


Abbildung 4.9: Vorhergesagte und tatsächliche Werte von *Climate Change total* für einen exemplarischen Train und Test Split des CO₂-Modells. [Eigene Darstellung]

Abbildung 4.11 zeigt den QQ Plot der Residuen auf der Transformationsskala. Hier liegen die Punkte deutlich näher an der Referenzgeraden als auf Originalskala, was eine näherungsweise Normalität im mittleren Bereich unterstützt. In den äußersten Quantilen bleiben Abweichungen sichtbar, was auf eine erhöhte Wahrscheinlichkeit großer Fehler hinweist.

Die Normalitätsannahme ist auf Originalskala für die CO₂-Daten klar verletzt, was aufgrund der großen Wertebandbreite und der dominierenden Ausreißer bei sehr großen Produkten erwartbar ist. Auf der Transformationsskala liegen die Fehler deutlich näher an einer Normalverteilung, während in den Rändern weiterhin Abweichungen verbleiben.

Für die Zielsetzung dieser Arbeit, nämlich robuste Vorhersagen für neue Produkte, ist diese Diagnose dennoch konsistent mit einem brauchbaren Modell. Die Modellgüte wird ausschließlich auf strikt getrennten Testdaten berichtet, und zusätzlich werden robuste und relative Fehlermaße verwendet, um die Leistung über verschiedene Größenordnungen hinweg fair zu bewerten. Die verbleibenden Abweichungen von der Normalität werden deshalb nicht als Ausschlusskriterium interpretiert, jedoch als Hinweis, dass klassische Annahmen der OLS-Theorie nur eingeschränkt auf diesen Datensatz übertragbar sind und Fehlermaße auf der Originalskala stark durch wenige große Produkte geprägt werden.

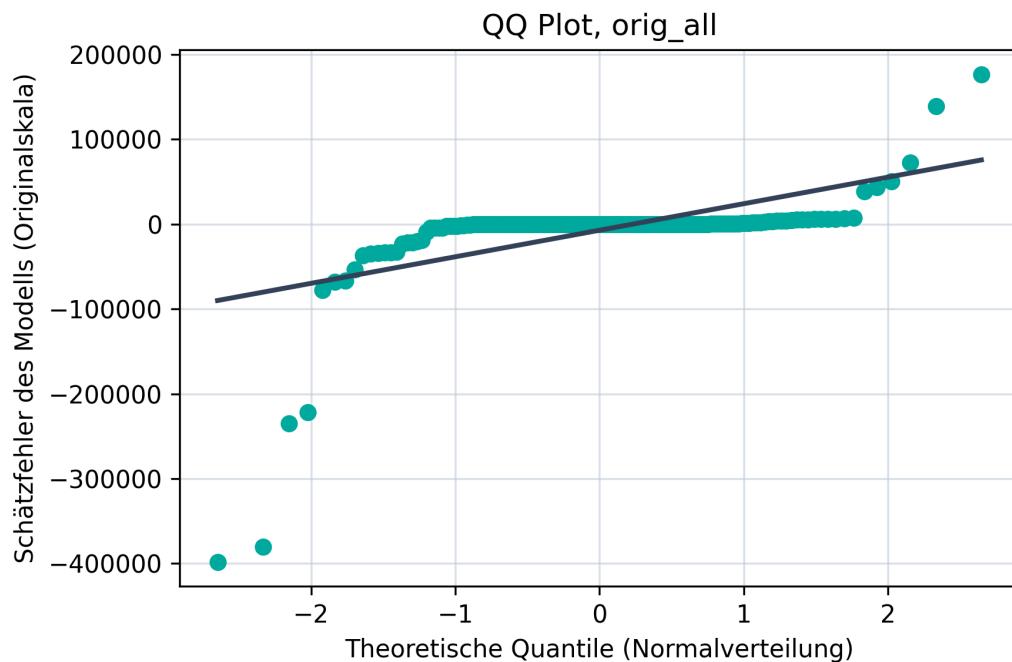


Abbildung 4.10: QQ Plot der Residuen des CO₂-Modells auf Originalskala im Testset. [Eigene Darstellung]

Die in diesem Abschnitt beschriebene Pipeline, bestehend aus transformierter Zielgröße, technischen Basismerkmalen (Gewicht, Stromverbrauch) und verdichteten Materialinformationen (PCA) wird im nächsten Schritt auf weitere Umweltindikatoren übertragen. Dabei ändert sich die verfügbare Datenbasis, bedingt durch unterschiedliche Fehlerteanteile, und die Erklärbarkeit der jeweiligen Indikatoren. Für die übrigen Indikatoren werden die getesteten Transformationen (keine Transformation, log1p, Box-Cox) und die resultierende Auswahl jeweils knapp zusammengefasst, und die Fehlerdiagnostik wird auf die wichtigsten Befunde reduziert. Für jeden Indikator werden diese Transformationsoptionen systematisch verglichen. Die Auswahl erfolgt datengetrieben anhand der Testgüte, wobei Fehlermaße auf Originalskala nach Rücktransformation berichtet werden, damit sie in der Einheit interpretierbar bleiben.

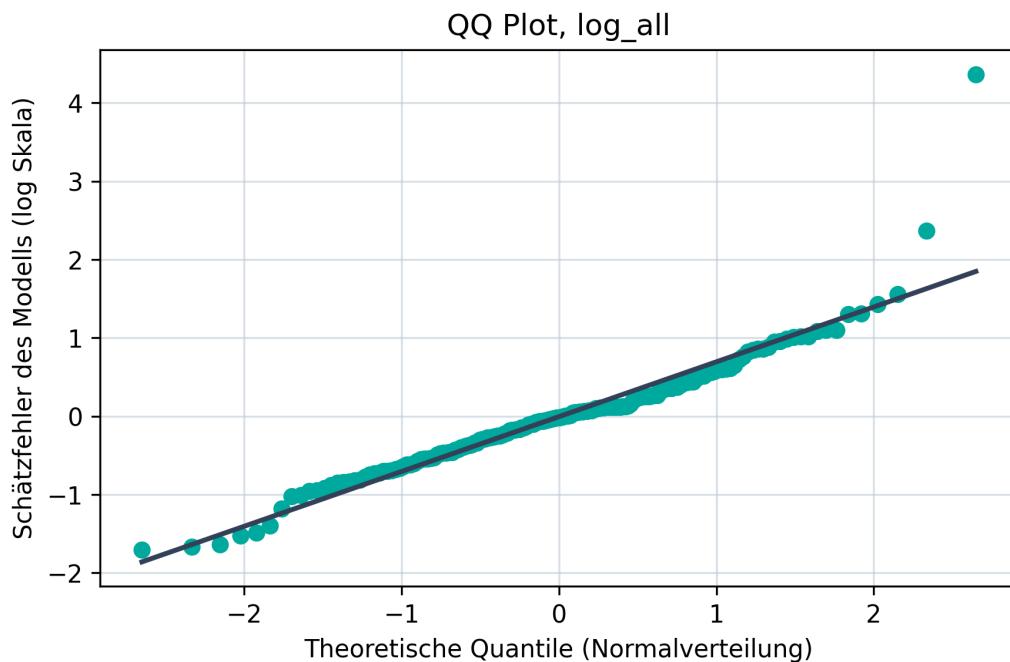


Abbildung 4.11: QQ Plot der Residuen des CO₂-Modells im Testset auf der Transformationsskala log_cc. [Eigene Darstellung]

4.5. Lineare Regression der anderen Indikatoren

Die für *Climate change (total)* aufgebaute Regressionspipeline wird im Folgenden auf weitere Umweltindikatoren angewendet, um zu prüfen, in welchem Umfang sich diese mit denselben Produktmerkmalen erklären lassen. Die zentralen Gütemaße aller Indikatoren sind in Tabelle 4.7 zusammengefasst. Ergänzende Streudiagramme und QQ-Plots für die einzelnen Indikatoren sind in Anhang A.1 dargestellt.

Insgesamt zeigen die meisten Indikatoren eine moderate bis hohe Testgüte ($R^2_{\text{Test}} \approx 0.73$ bis 0.87). Die Kombination aus RMSE und Median absoluter Fehler verdeutlicht eine typische Schiefe der Fehlerverteilungen. Der RMSE wird von wenigen sehr großen Abweichungen dominiert, während der Median den typischen Fehler eines repräsentativen Produkts beschreibt.

Die relativen Fehlermaße MdARE und MARE sind einheitenlos und als Anteil interpretierbar, zum Beispiel entspricht MdARE = 0.66 einer typischen relativen Abweichung von etwa 66 %. Dass MARE in allen aufgezeigten Fällen deutlich größer als MdARE ist, weist auf eine stark rechtsschiefe Verteilung der relativen Fehler hin. Dies tritt insbesondere bei sehr kleinen Zielwerten auf, da dort bereits kleine absolute Abweichungen zu sehr großen relativen Fehlern führen.

Tabelle 4.7: Übersicht der Testgüte der Regressionsmodelle für weitere Indikatoren. RMSE und Median absoluter Fehler sind auf der Originalskala angegeben.

Indikator	<i>n</i>	Transform.	R^2_{Test}	$\text{RMSE}_{\text{Test}}$	Median abs. Fehler	MdARE	MARE
Acidification	177	log1p	0.845	490.87 kg SO ₂	0.2864 kg SO ₂	0.9579	3.8258
Hazardous waste disposed	168	log1p	0.813	18489.74 kg	176.7890 kg	0.6639	13.1431
Water use	169	Box-Cox	0.726	28682.36 m ³	1353.56 m ³	0.9297	2.3278
Photochemical ozone formation (HH)	171	Box-Cox	0.802	109.1708 kg C ₂ H ₄	0.0255 kg C ₂ H ₄	0.6665	1.2811
Resource use, fossils	171	log1p	0.871	1119662.28 MJ	14460.95 MJ	0.6370	11.1394
Eutrophication (terrestrial)	107	Box-Cox	0.793	99.33 mol N	0.0854 mol N	0.5611	1.1653
Ozone depletion	170	Box-Cox	0.858	0.0029 kg CFC – 11	0.00001 kg CFC – 11	0.8635	1.0411
Resource use, minerals and metals	175	Box-Cox	0.866	0.7218 kg Sb	0.0010 kg Sb	0.7798	1.0003

4.5.1. Regression des Indikators Acidification

Als repräsentatives Beispiel für Indikatoren mit hoher Modellgüte wird im Folgenden *Acidification* detaillierter dargestellt. Es wurden keine Transformation, log1p und eine Box-Cox Transformation verglichen. Die beste Testgüte wird mit log1p erreicht. Die Zielvariable ist damit $\log(1 + \text{acidification}_{\text{total}})$. Es konnten $n = 177$ PEPs verwendet werden. Tabelle 4.8 fasst die Testleistung nach Rücktransformation zusammen.

Tabelle 4.8: Gütekennzahlen des linearen Regressionsmodells (Acidification als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.845
$\text{RMSE}_{\text{Test}}$	490.87 kg SO ₂
Median absoluter Fehler	0.2864 kg SO ₂
MdARE _{Test} (Median rel. Fehler)	0.9579
MARE _{Test} (Mittelwert rel. Fehler)	3.8258

Das Modell erklärt rund 85% der Varianz mit einem RMSE von 490,87 kg SO₂, was auf eine insgesamt gute Vorhersagegüte für den Indikator Acidification hinweist. Die robusten Fehlermaße ergänzen dieses Bild. Der Median des absoluten Fehlers liegt weit unter dem RMSE bei etwa 0,29 kg SO₂ und beschreibt damit die typische Abweichung eines repräsentativen Produkts. MdARE $\approx 0,96$ bedeutet, dass die typische

Abweichung in der Größenordnung des Zielwerts liegt. Der deutlich größere Mittelwert der relativen Fehler ($MARE \approx 3,83$) weist zugleich auf eine stark schiefe Fehlerverteilung mit einzelnen sehr großen relativen Abweichungen hin. Dies ist insbesondere bei kleinen Zielwerten plausibel, da dort bereits kleine absolute Fehler zu sehr großen relativen Fehlern führen. Insgesamt ist das Modell damit für Acidification gut geeignet, einzelne Produkte können jedoch deutlich schlechter getroffen werden.

Zur Veranschaulichung zeigt Abbildung 4.12 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von Acidification.

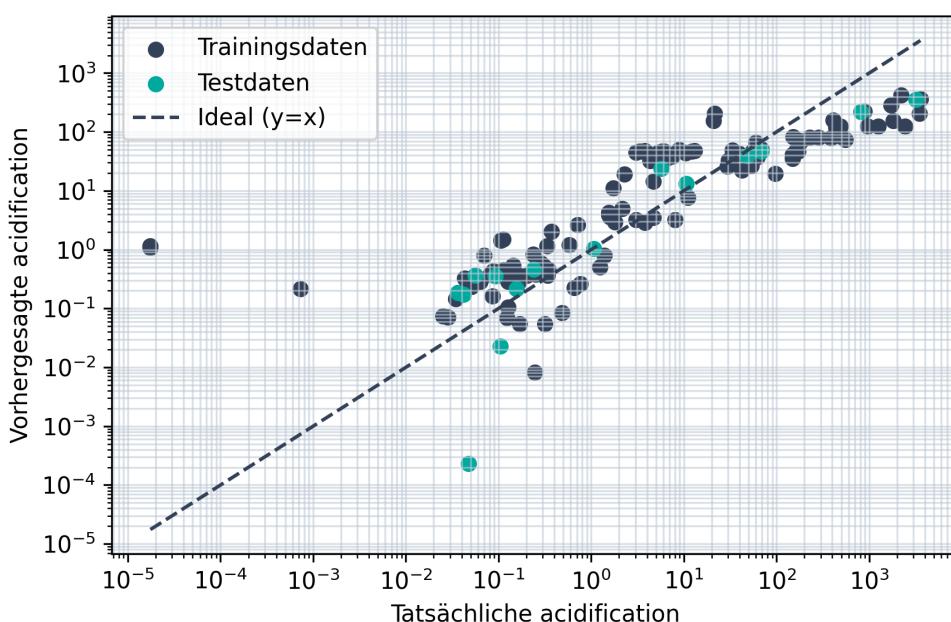


Abbildung 4.12: Vorhergesagte gegenüber tatsächlichen Acidification-Werten. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

Die meisten Punkte liegen in der Nähe der Diagonalen, insbesondere im Bereich mittlerer Acidification-Werte, was auf eine gute Abbildung des allgemeinen Trends hinweist. Hohe Werte werden tendenziell leicht unterschätzt und im Bereich von 10^1 ist ein Cluster sichtbar, der tendenziell überschätzt wird.. Im Bereich der sehr kleinen Werte schätzt das Modell relativ betrachtet sehr ungenau, die absoluten Fehler bleiben dort jedoch gering. Das Fehlerverhalten von Trainings- und Testdaten ist vergleichbar, so dass keine starke Überanpassung erkennbar ist. Insgesamt bestätigt die Analyse, dass das aus dem CO₂-Fall übernommene Modell auch für den Indikator Acidification robuste und plausible Vorhersagen liefert.

Abbildung 4.13 zeigt einen QQ Plot der Residuen des Modells im Vergleich zu einer Normalverteilung. Wie beim CO₂ Indikator zeigen sich auf der Originalskala typischerweise starke Abweichungen, weshalb die Fehlerdiagnostik hier auf der Transformationsskala

berichtet wird. Daher wird hier nur die Transformationsskala dargestellt. Im mittleren Bereich liegen die Punkte nah an der Referenzgeraden, was darauf hindeutet, dass der Großteil der Fehler näherungsweise normalverteilt ist. In den Randbereichen sind jedoch Abweichungen erkennbar. Insgesamt ist die Fehlerverteilung im Zentrum gut durch eine Normalverteilung approximierbar, während die Extrembereiche durch schwerere Verteilungsschwänze geprägt sind.

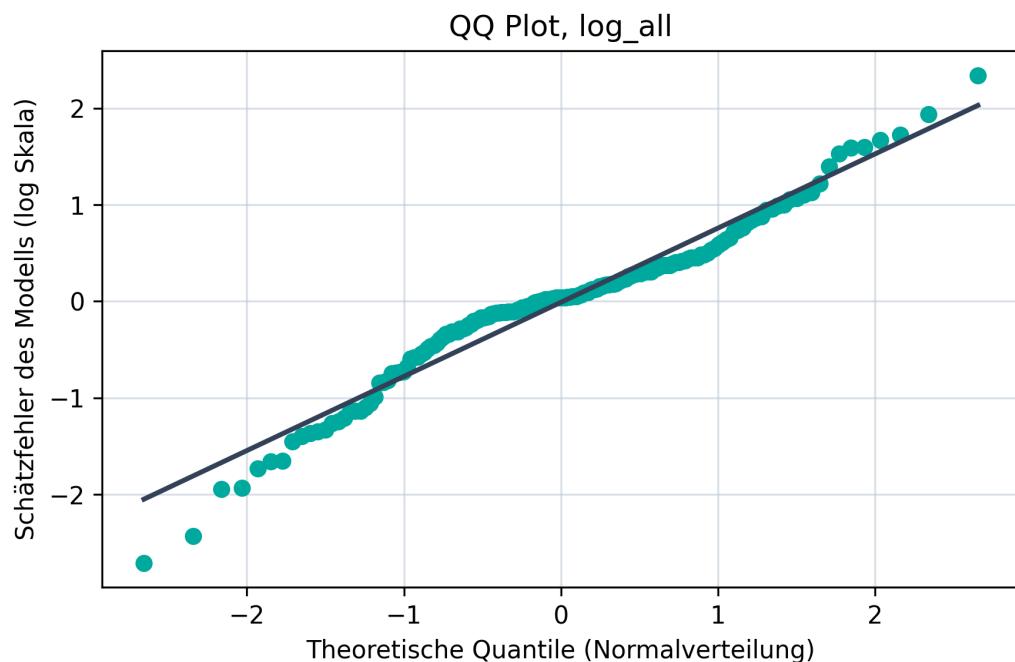


Abbildung 4.13: QQ Plot der Residuen des Acidification Modells (Test- und Trainingsset). [Eigene Darstellung]

Eine ausführliche Visualisierung der übrigen Indikatoren, einschließlich QQ-Plots, ist in Anhang A.1 dokumentiert.

4.5.2. Indikatoren mit geringer Modellgüte

Neben den in Tabelle 4.7 aufgeführten Indikatoren mit moderater bis hoher Modellgüte wurden alle weiteren Umweltindikatoren mit derselben Regressionspipeline geschätzt. Für einige Zielgrößen bleibt das erreichte Test- R^2 jedoch unter 0,5, sodass hier nicht von einem zuverlässigen Vorhersagemodell gesprochen werden kann. Tabelle 4.9 fasst diese Indikatoren zusammen.

Für diese schwächer erklärbaren Zielgrößen wurden ebenfalls verschiedene Transformationen der Zielvariable verglichen. Box-Cox verbessert die Testgüte teilweise leicht. Insgesamt bleiben die Zugewinne jedoch begrenzt und führen nicht zu stabilen Vorher-

sagen.

Tabelle 4.9: Indikatoren und Gütemaße mit geringer Modellgüte.

Indikator	R^2_{Test}	$\text{RMSE}_{\text{Test}}$	Anzahl analysierter PEPs
Eutrophication (freshwater)	0.434	1.3165	133
Eutrophication (marine)	0.322	25.0197	107
Radioactive waste disposed	0.492	176.8942	158

Abbildung 4.14 zeigt den Indikator *Radioactive waste disposed* als ein Beispiel. Der Indikator ist nur eingeschränkt erklärbar ($R^2_{\text{Test}} \approx 0,49$), und die relativen Fehlermaße ($\text{MdARE} \approx 1,85$, $\text{MARE} \approx 19,01$) weisen auf eine stark instabile Vorhersage hin. Gleichzeitig bleibt der Median des absoluten Fehlers mit 0,5193 vergleichsweise klein, was zur sehr kleinen Größenordnung des Indikators passt und zeigt, dass die extremen relativen Fehler vor allem bei kleinen Zielwerten entstehen. Insgesamt wird die Streuung im Streudiagramm nicht ausreichend abgebildet, und eine Box-Cox-Transformation kann die Fehlerstruktur nicht stabilisieren.

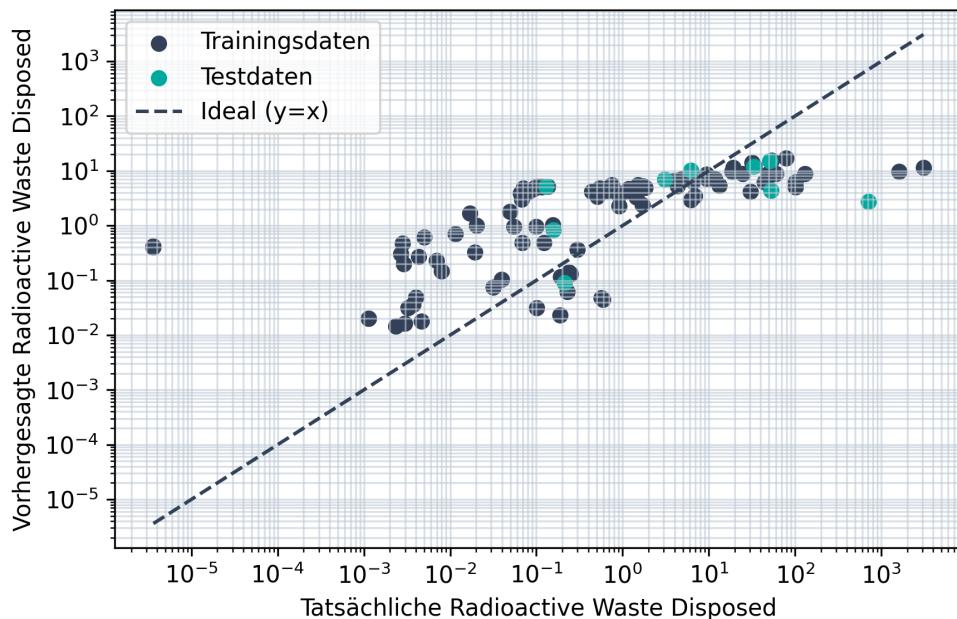


Abbildung 4.14: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Radioactive waste disposed*. [Eigene Darstellung]

Der QQ Plot in Abbildung 4.15 zeigt die Residuen auf der Transformationsskala. Im Zentrum liegen die Quantile nur näherungsweise auf der Referenzgeraden, während die äußereren Bereiche deutlich abweichen. Dies weist auf schwere Verteilungsschwänze und systematische Modellfehler hin.

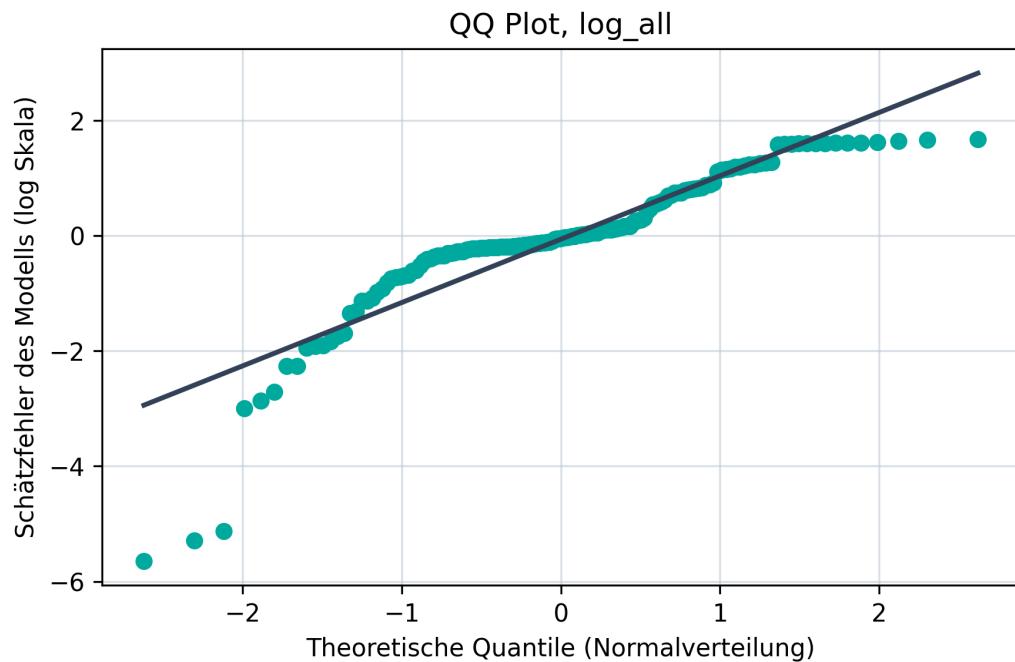


Abbildung 4.15: QQ Plot der Residuen des Modells für *Radioactive waste disposed* auf der Transformationsskala. [Eigene Darstellung]

Ähnliche Muster zeigen sich auch bei den beiden Eutrophication-Indikatoren in Tabelle 4.9. Sie sind im Anhang A.2 ebenfalls visualisiert.

5

Diskussion und Fazit

Aus den in Kapitel 4.4 und 4.5 vorgestellten Regressionsmodellen ergibt sich ein differenziertes Bild. Einige Indikatoren lassen sich sehr gut, andere nur eingeschränkt durch Gewicht, Stromverbrauch und Materialinformationen erklären. Dieses Kapitel ordnet die Ergebnisse ein und diskutiert Unsicherheiten und Grenzen.

5.1. Vorgehen und Methodischer Beitrag

Ziel der Arbeit war es, eine belastbare und skalierbare Datengrundlage aus PEP-Ecopassport-Dokumenten aufzubauen und darauf aufbauend ein kompaktes, gut übertragbares Modell zur heuristischen Abschätzung zentraler Umweltindikatoren zu entwickeln. Im Vordergrund stand dabei nicht die exakte Reproduktion vollständiger Ökobilanzen, sondern eine schnelle Größenordnungsabschätzung, die auch dann möglich ist, wenn für ein neues Produkt kein PEP vorliegt und nur wenige robuste Produktmerkmale bekannt sind.

Ausgangspunkt der Analyse war eine systematische Recherche nach geeigneten PEPs. Insgesamt wurden 252 PEP-Dokumente identifiziert. Nach Ausschluss fehlerhafter oder unbrauchbarer Dokumente verblieb ein Analysedatensatz von 234 PEPs. Für die Modellierung wurden je nach Zielindikator nur PEPs berücksichtigt, in denen die Zielgröße und die benötigten Eingangsvariablen numerisch verfügbar waren. Für *Climate change (total)* ergibt sich dadurch beispielsweise eine Modellstichprobe von $n = 173$.

Nach der Konvertierung der PDFs in strukturierte Markdown-Segmente wurden die Inhalte schema-gesteuert in JSON extrahiert. Die eigentliche Extraktion der Variablen

erfolgte schema-gesteuert mit einem Sprachmodell (LLM), das die gewünschten Felder in einem JSON-Format zurückgab.

Um uneinheitliche Schreibweisen in Ländern, Materialien, Phasen, Energiequellen und Einheiten zu vereinheitlichen, wurden die Daten auf Basis einer Vokabularanalyse regelbasiert normalisiert.

Auf der so erzeugten Datenbasis wurde ein bewusst kompakter Regressionsansatz gewählt, der zum Modellierungsziel passt. Als Eingangsgrößen dienen Variablen, die auch ohne PEP typischerweise messbar oder abschätzbar sind, insbesondere Gesamtgewicht und über die Lebensdauer aggregierter Stromverbrauch, sowie PCA-verdichtete Materialinformationen.

Die Modellgüte wird konsequent über getrennte Trainings- und Testdaten beurteilt. Dieses Vorgehen ist für die Zielsetzung relevant, weil es die Aussagekraft zur Übertragbarkeit auf neue, bislang ungesehene Produkte stärkt und Überanpassung durch eine zu starke Optimierung auf die vorhandenen PEP-Daten reduziert.

Der methodische Beitrag der Arbeit liegt damit in einem reproduzierbaren Prozess vom heterogenen PDF-Dokument bis zu einem konsistenten Datensatz und in einem Modellansatz, der mit wenigen, robust erfassbaren Eingangsgrößen eine plausible Einordnung von Umweltindikatoren ermöglicht. Die nachfolgenden Abschnitte ordnen die erzielten Ergebnisse vor diesem Hintergrund in Bezug auf die Forschungsfrage ein und diskutieren die Grenzen der Übertragbarkeit.

5.2. Einordnung der Ergebnisse im Kontext der Forschungsfrage

Die Forschungsfrage zielt darauf ab, Umweltwirkungen von Produkten der Gebäudeautomation systematisch auf Basis von PEP-Deklarationen zu analysieren und zugleich zu prüfen, inwieweit sich daraus ein Modell ableiten lässt, das auch für Produkte ohne PEP belastbare Schätzungen ermöglicht. Die Ergebnisse dieser Arbeit liefern hierzu ein differenziertes Bild.

Erstens zeigt die Arbeit, dass eine systematische, vergleichbare Auswertung der in PEPs ausgewiesenen Umweltindikatoren grundsätzlich möglich ist. Sie setzt jedoch eine Standardisierung der heterogenen Dokumente und Begriffe voraus.

Zweitens lassen die Modellresultate erkennen, dass einige Indikatoren mit wenigen, allgemein verfügbaren Produktmerkmalen gut erklärbar sind. Am deutlichsten gilt dies für *Climate change (total)*. Gewicht, Stromverbrauch über die Lebensdauer und verdichtete Materialinformationen sind zentrale Treiber der Treibhauswirkung, sodass die Vorhersagen auf dem Testdatensatz vergleichsweise stabil ausfallen. Gleichzeitig

zeigen die Fehlermaße, dass auch dieses Modell keine präzise Reproduktion einzelner PEP-Werte liefert, sondern vor allem eine Einordnung der Größenordnung ermöglicht.

Für mehrere weitere Indikatoren lässt sich der Ansatz zwar übertragen, die Modelle erreichen jedoch nicht die Stabilität der CO₂-Regression. Dies betrifft insbesondere *Acidification*, *Hazardous waste disposed*, *Water use*, *Photochemical ozone formation (HH)*, *Resource use (fossils)*, *Eutrophication (terrestrial)*, *Ozone depletion* und *Resource use (minerals and metals)*. Für diese Zielgrößen werden in der Regel noch brauchbare Ergebnisse erzielt, die Streuung der Vorhersagefehler ist jedoch größer und die erklärte Varianz fällt geringer aus als bei *Climate change (total)*.

Für *Eutrophication (freshwater)*, *Eutrophication (marine)* und *Radioactive waste disposed* gelingt es dagegen nur, einen begrenzten Anteil der Varianz zu erklären. Dies deutet darauf hin, dass die verwendeten Eingangsgrößen (Gewicht, Stromverbrauch und Materialmuster) wesentliche Einflussfaktoren dieser Indikatoren nicht abbilden, oder dass die PEP-Daten für diese Zielgrößen stärker durch unterschiedliche Systemgrenzen, Berechnungsannahmen und verwendete Hintergrunddatensätze geprägt sind.

Die stark rechtsschiefen Verteilungen der Inputgrößen und Indikatoren unterstreichen, dass Transformationen für eine stabilere Modellierung hilfreich sind..

Vor diesem Hintergrund ist der Begriff *belastbare Schätzung* in dieser Arbeit als *heuristische Abschätzung* zu verstehen. Das Modell eignet sich insbesondere für eine schnelle, datengetriebene Orientierung, etwa zur Vorpriorisierung von Produkten oder Varianten, zur groben Einordnung erwartbarer Größenordnungen und zur Plausibilitätsprüfung von PEP-Angaben. Als Ersatz für eine vollständige Ökobilanz oder für feingranulare Vergleiche ähnlicher Produkte ist der Ansatz dagegen nicht geeignet.

Insgesamt beantwortet die Arbeit die Forschungsfrage somit in zwei Teilen. Die PEP-Daten können durch die entwickelte Pipeline systematisch ausgewertet und in eine konsistente Datenbasis überführt werden. Auf dieser Grundlage lassen sich für ausgewählte Indikatoren, insbesondere *Climate change (total)*, aus wenigen, ohne PEP erfassbaren Produktmerkmalen Modelle ableiten, die eine robuste Einordnung der Umweltauswirkungen in Größenordnungen ermöglichen. Für andere Indikatoren zeigt sich hingegen, dass zusätzliche erklärende Variablen oder alternative Modellierungsstrategien erforderlich wären, um vergleichbare Schätzqualität zu erreichen.

5.3. Grenzen und Limitationen

Trotz der erzielten Ergebnisse ist die Aussagekraft der Modelle durch mehrere Faktoren begrenzt. Ein zentraler Punkt ist die Datengrundlage selbst. Die 234 recherchierten

PEPs stellen keine zufällige Stichprobe aller Produkte der Gebäudeautomation dar, sondern spiegeln nur Produktgruppen und Hersteller wider, für die PEP-Deklarationen verfügbar sind. Entsprechend sind die abgeleiteten Zusammenhänge primär innerhalb des beobachteten Datenraums belastbar. Für Produktarten, die in der Stichprobe nur selten oder gar nicht vorkommen, ist mit deutlich größeren Prognosefehlern zu rechnen.

Ein Teil der verbleibenden Modellunsicherheit ist nicht durch das Regressionsverfahren, sondern durch die Struktur der PEP-Daten selbst bedingt. Besonders relevant ist die Abbildung des Strommixes. In vielen PEPs ist der verwendete Strommix entweder nur grob (z. B. als Länder- oder Regionenmix) angegeben oder es fehlen Referenzjahre des Mixes. Gleichzeitig kann sich die Emissionsintensität der Stromerzeugung über die Zeit verändern. Dadurch entsteht eine zusätzliche, nicht beobachtbare Varianz in der Zielgröße, die mit den verfügbaren Inputvariablen (Gewicht, Stromverbrauch, Materialmix) nicht erklärt werden kann.

Ähnliche Effekte treten auf, wenn PEPs unterschiedliche Annahmen zu Nutzungsszenarien, Systemgrenzen oder Methodenversionen verwenden. Beispielsweise treten in den Daten verschiedene PEF-Versionen auf (PEF 3.0 vs. PEF 3.1). Bei einem Wechsel der PEF-Version ändern sich die Berechnungsmethoden, was zu systematischen Änderungen der Indikatorwerte führen kann. Solche methodisch bedingten Unterschiede sind mit den hier genutzten Eingangsgrößen nicht verknüpft. Das Regressionsmodell kann sie deshalb nicht als physikalische Beziehungen zwischen Gewicht, Stromverbrauch, Materialmix und Indikatorwerten lernen, sondern nur als zusätzliches Rauschen wiederfinden. Dies begrenzt die maximal erreichbare Vorhersagegüte, selbst wenn das Modell formal eine hohe erklärte Varianz auf Teilen der Stichprobe erzielt.

Auch die Fehlerstruktur der Modelle setzt Grenzen. Die Verteilungen der Zielgrößen sind stark rechtsschief und enthalten Ausreißer. Transformationen wie \log_{10} p oder Box-Cox stabilisieren die Modellierung, führen jedoch nicht zu vollständig normalverteilten Residuen. Damit sind klassische Schlussfolgerungen der OLS-Theorie, etwa Standardfehler oder p-Werte, nur eingeschränkt belastbar. Für die Zielsetzung dieser Arbeit ist aber die testbasierte Generalisierung zentral. Ergänzend zu R^2 und RMSE werden robuste Fehlermaße wie der Median absoluter Fehler sowie Median und Mittelwert der relativen absoluten Fehler genutzt, um die typische Modellleistung trotz Ausreißern nachvollziehbar zu charakterisieren.

Die Materialmodellierung ist ebenfalls mit einer Abwägung verbunden. Die Nutzung einer PCA auf dem Materialblock erhöht die Stabilität, reduziert Multikollinearität und verhindert Überanpassung durch viele korrelierte Materialspalten. Gleichzeitig sinkt die Interpretierbarkeit einzelner Materialeffekte, da die Hauptkomponenten eher Muster im Materialmix abbilden als direkte kausale Beiträge einzelner Materialien.

Schließlich bestehen technische Limitationen in der Extraktionspipeline. Rastertabel-

len können ohne OCR nicht zuverlässig ausgelesen werden, was zu fehlenden Werten oder zum Ausschluss einzelner PEPs führt. Darüber hinaus kann es trotz robuster Layoutanalyse in Einzelfällen zu Extraktionsfehlern kommen, die sich bis in die Auswertung fortpflanzen. Die schema-gesteuerte Extraktion mit einem Sprachmodell garantiert keine vollständige Deterministik und Transparenz einzelner Extraktionsentscheidungen.

Insgesamt eignen sich die Modelle insbesondere zur Einordnung typischer Produkte innerhalb des betrachteten Datenraums.

5.4. Ausblick und zukünftiger Forschungsbedarf

Ein naheliegender nächster Schritt ist eine Kategorisierung nach Produktgruppen und eine getrennte Modellierung je Kategorie. Viele der beobachteten Streuungen dürften dadurch reduziert werden, da sich Zusammenhänge zwischen Gewicht, Stromverbrauch, Material und Indikatoren in homogeneren Teilmengen stabiler verhalten. Eine solche Segmentierung könnte über PEP-Metadaten oder Produktbeschreibungen erfolgen und würde gleichzeitig die Grundidee eines kompakten Feature-Sets beibehalten.

Methodisch bietet sich außerdem an, robustere Modellklassen zu testen, die mit Ausreißern und Heterogenität besser umgehen können, ohne die Interpretierbarkeit vollständig zu verlieren. Beispiele wären robuste lineare Modelle (z. B. Huber-Regressoren oder Quantilsregression) oder auch baumbasierte Verfahren als Vergleichsbasis. Dabei sollte die Evaluation weiterhin strikt über getrennte Trainings- und Testdaten erfolgen, um die Generalisierbarkeit auf neue Produkte abzusichern.

Auf Pipeline-Ebene kann der gezielte Einsatz von OCR für Rastertabellen eine sinnvolle Erweiterung sein. Damit ließen sich zusätzliche PEPs und insbesondere schwer zugängliche Tabellenwerte erschließen, was die Vollständigkeit der Datenbasis verbessern würde. Gleichzeitig erhöht OCR den Rechenaufwand und kann neue Fehlerquellen einführen. Ein sinnvoller Kompromiss wäre eine fallweise Aktivierung nur dann, wenn DocLink keine verwertbaren Tabellenextrakte liefert.

Schließlich ist für die Zukunft eine automatisierte Standardisierung und Validierung der Datenbasis relevant. Dazu gehören Plausibilitätschecks, Einheitenprüfungen und konsistente Aggregationsregeln.

Die systematische Dokumentation von PCR-Versionen, Systemgrenzen und Datenbankbezügen, könnte die Modelle zusätzlich verbessern. Damit ließe sich besser trennen, ob Streuung primär aus echten Produktunterschieden oder aus methodischen Unterschieden zwischen PEPs resultiert. Eine Aufnahme in das Feature-Set widerspricht allerdings einem kompakten, messbaren Modell. Zudem tauchen diese Informationen in

den PEPs nur unregelmäßig auf.

Insgesamt zeigt der Ansatz ein praktikables Potenzial für eine schnelle, datengetriebene Einordnung von Umweltwirkungen. Zukünftige Forschungen können darauf aufbauen, indem sie die Pipeline um zusätzliche Datenquellen erweitern, die Modelle stärker segmentieren und die Robustheit gegenüber Heterogenität systematisch erhöhen, ohne die Übertragbarkeit auf Produkte ohne PEP zu verlieren.

A

Anhang

A.1. Visualisierung weiterer Regressionsmodelle (Hohe Modellgüte)

Die folgenden Abschnitte ergänzen die im Haupttext berichteten Ergebnisse um zusätzliche Indikatoren. Dargestellt sind jeweils Streudiagramme und QQ Plots zur Einordnung der Fehlerstruktur.

A.1.1. Regression des Indikators Hazardous waste disposed

Das Modell erreicht eine gute Testgüte, allerdings ist die Fehlerverteilung stark schief. Der RMSE wird durch wenige große Abweichungen dominiert, während der Medianfehler deutlich kleiner ausfällt. Abbildung A.1 zeigt die insgesamt gute Trendabbildung bei mittleren und großen Werten, während kleine Zielwerte zu großen relativen Fehlern führen können.

Der QQ Plot in Abbildung A.2 zeigt eine näherungsweise Linearität im Zentrum, mit Abweichungen in den Randbereichen, was auf schwere Verteilungsschwäne hinweist.

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

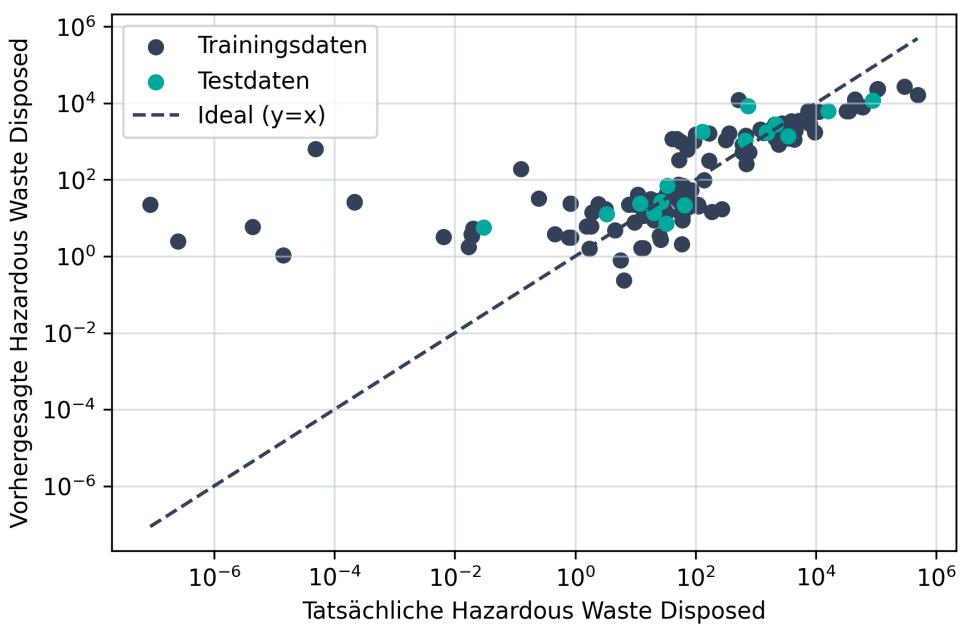


Abbildung A.1: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Hazardous waste disposed*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

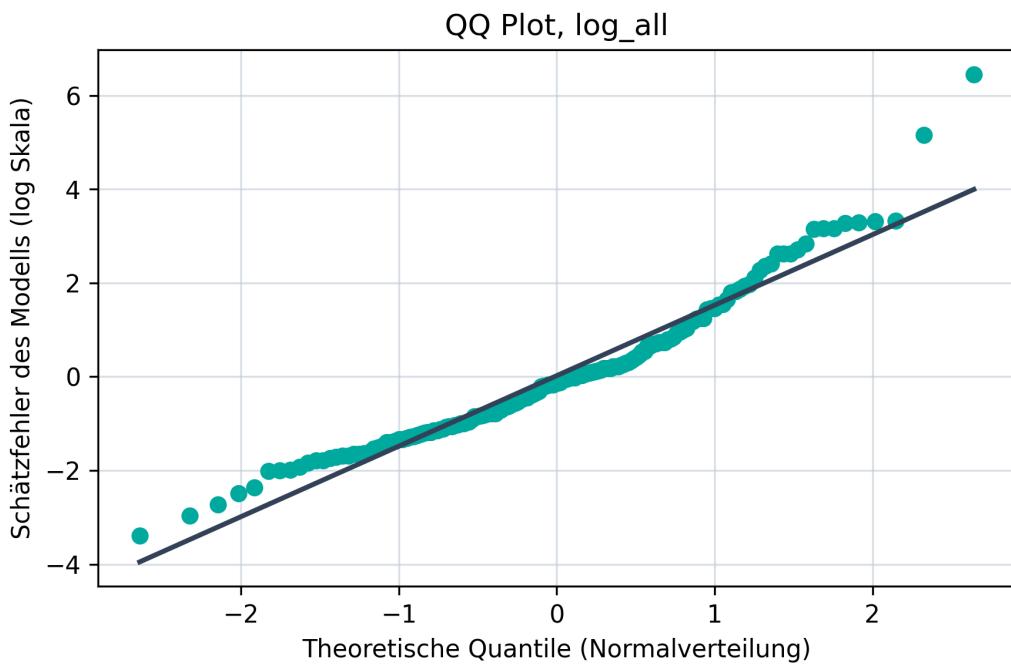


Abbildung A.2: QQ Plot der Residuen des *Hazardous waste disposed* Modells auf der Transformationsskala. [Eigene Darstellung]

A.1.2. Regression des Indikators Water use

Die Testgüte ist moderat, mit deutlich größerer Streuung als bei den stärksten Indikatoren. Abbildung A.3 zeigt eine tendenzielle Unterschätzung im oberen Wertebereich. Die Fehlerverteilung ist durch Ausreißer geprägt, was sich auch im QQ Plot widerspiegelt.

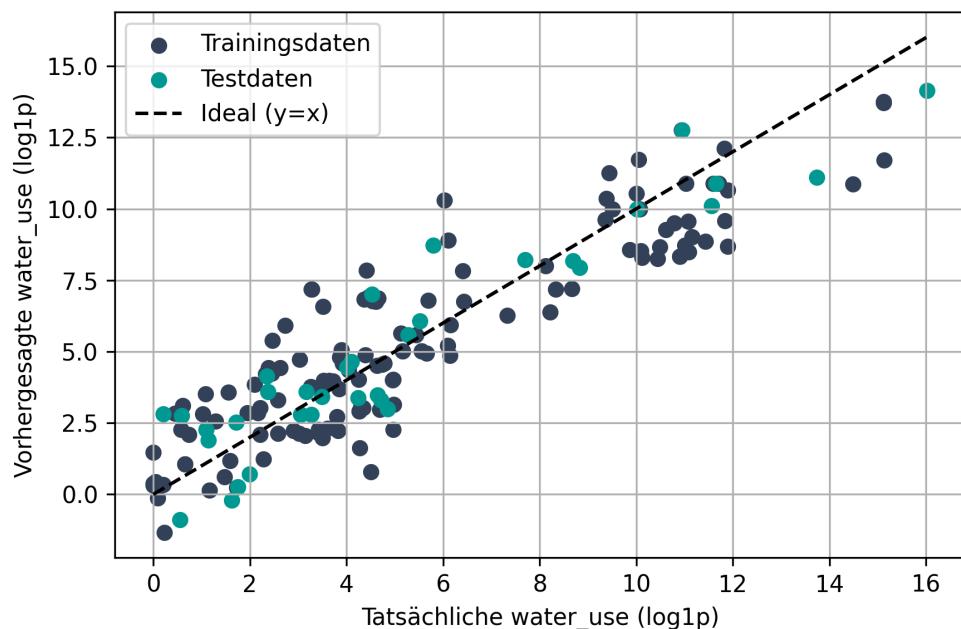


Abbildung A.3: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Water use*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

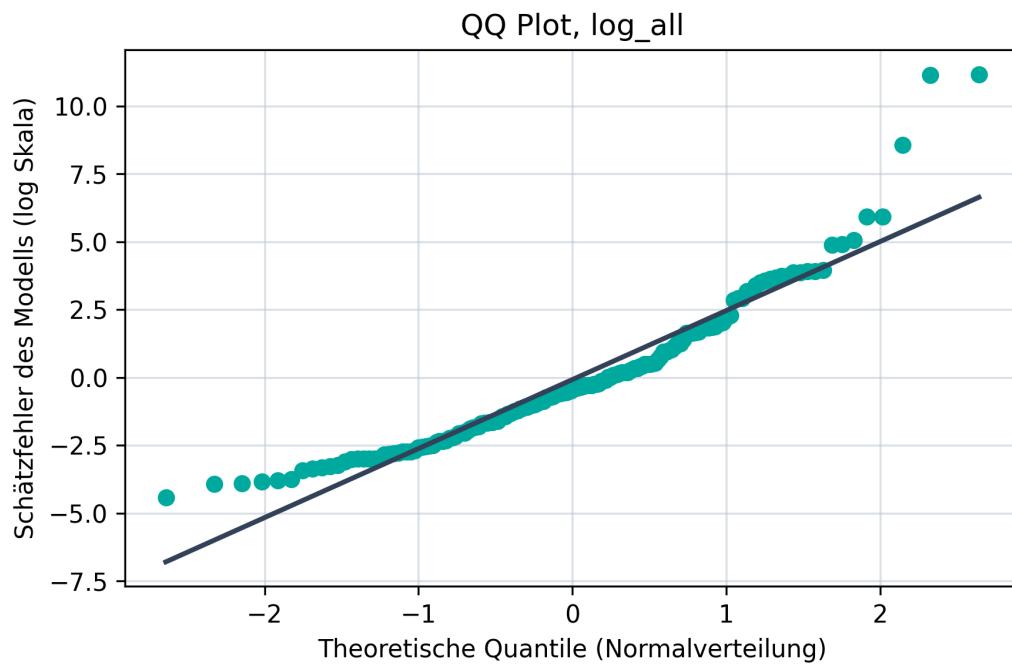


Abbildung A.4: QQ Plot der Residuen des *Water use* Modells auf der Transformationsskala. [Eigene Darstellung]

A.1.3. Regression des Indikators Photochemical ozone formation (HH)

Die Vorhersagen erfassen den Haupttrend gut, mit größerer Streuung bei kleinen Werten und einer leichten Unterschätzung großer Zielwerte, vgl. Abbildung A.5. Die Residuen zeigen in den Randbereichen deutliche Abweichungen von Normalität, vgl. Abbildung A.6.

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

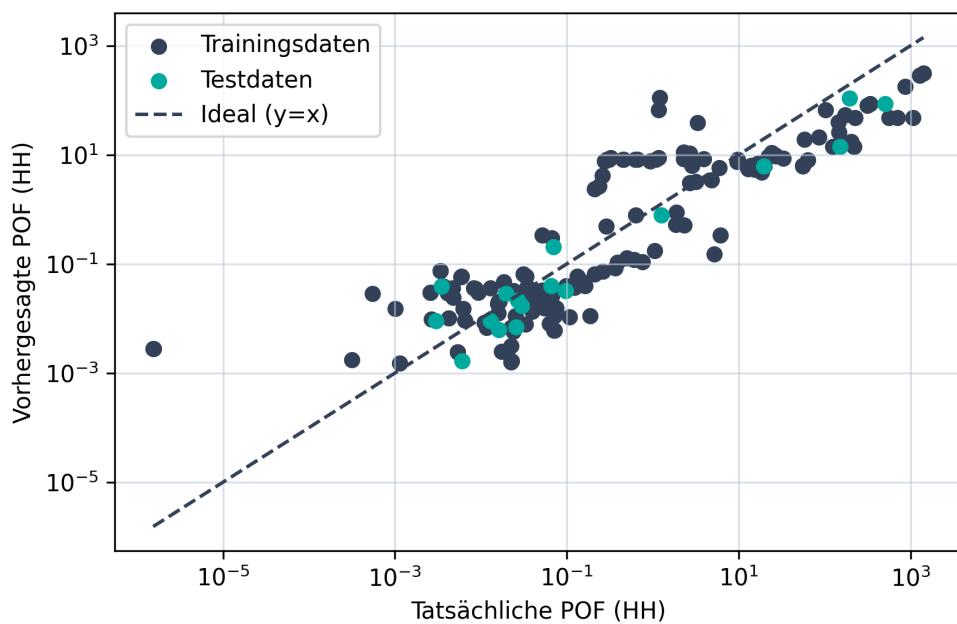


Abbildung A.5: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Photochemical ozone formation, human health*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

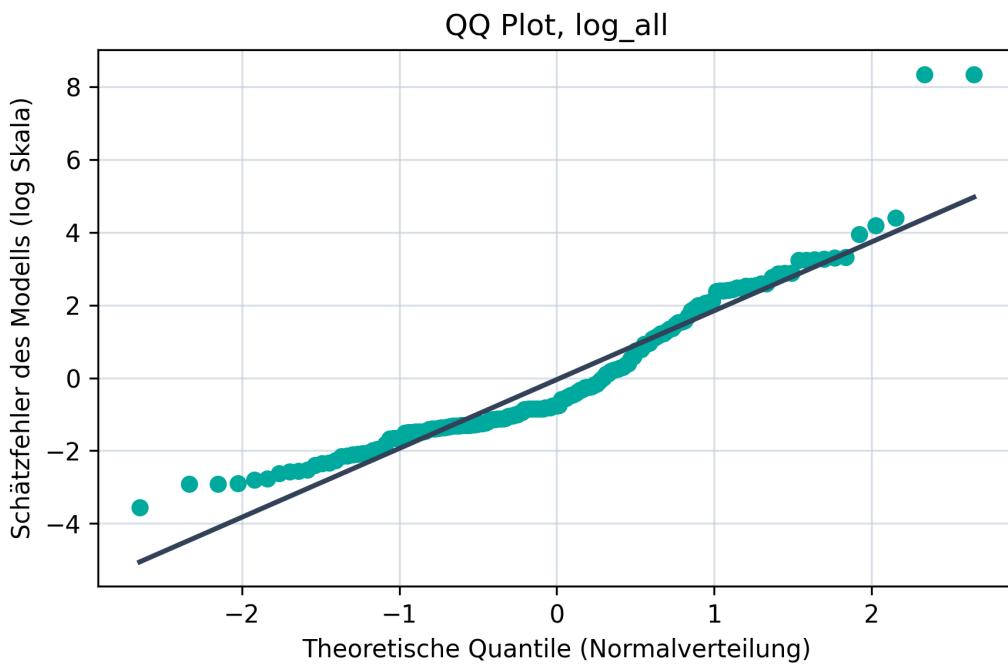


Abbildung A.6: QQ Plot der Residuen des *Photochemical ozone formation, human health* Modells auf der Transformationsskala. [Eigene Darstellung]

A.1.4. Regression des Indikators Resource use, fossils

Abbildung A.7 zeigt eine enge Punktwolke entlang der Diagonalen mit einzelnen starken Überschätzungen. Der QQ Plot weist auf schwere Verteilungsschwäne hin, vgl. Abbildung A.8.

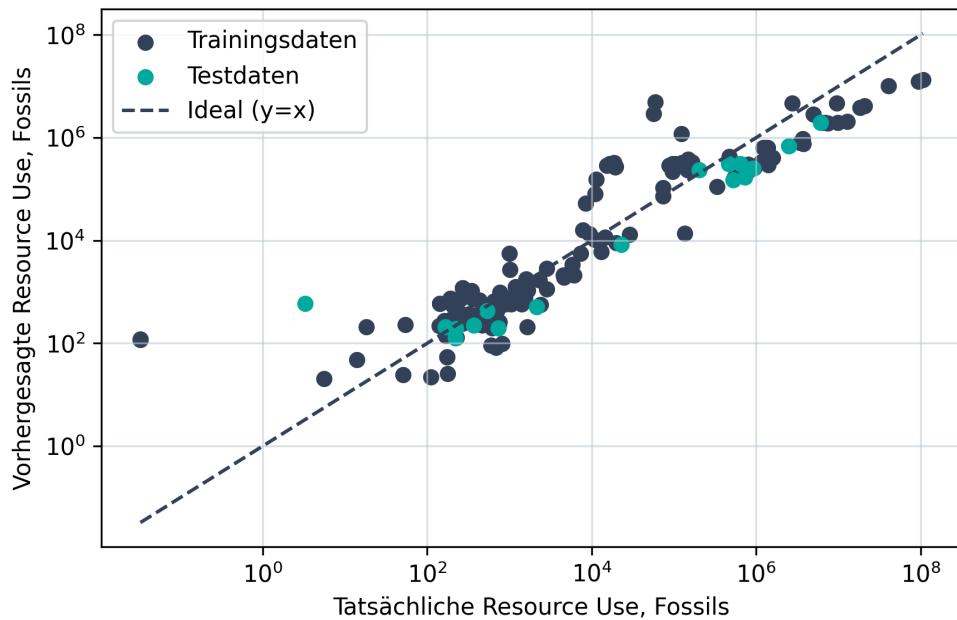


Abbildung A.7: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Resource use, fossils*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

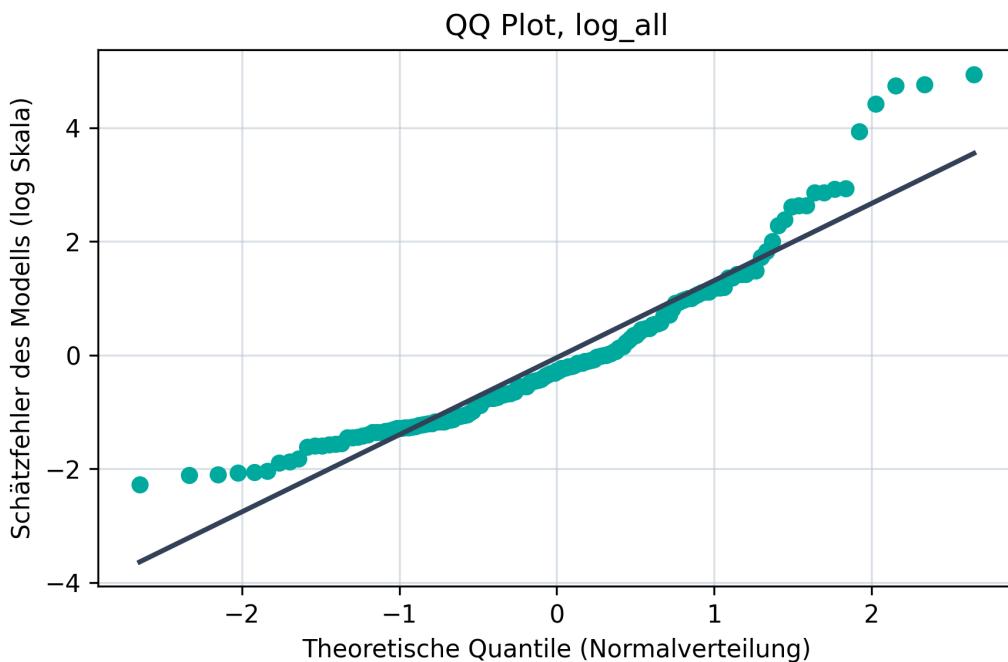


Abbildung A.8: QQ Plot der Residuen des *Resource use, fossils* Modells auf der Transformationsskala.
[Eigene Darstellung]

A.1.5. Regression des Indikators Eutrophication (terrestrial)

Das Modell erreicht eine solide Testgüte, mit erhöhter Streuung bei kleinen Werten. Im oberen Wertebereich ist eine leichte Unterschätzung sichtbar, vgl. Abbildung A.9. Der QQ Plot zeigt Abweichungen in den Randbereichen, vgl. Abbildung A.10.

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

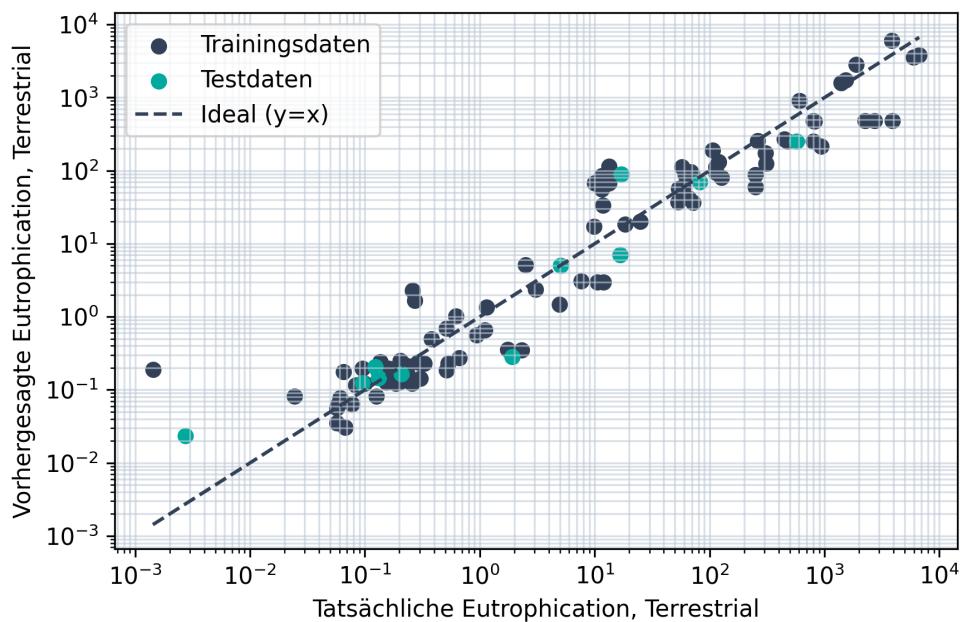


Abbildung A.9: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Eutrophication, terrestrial*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

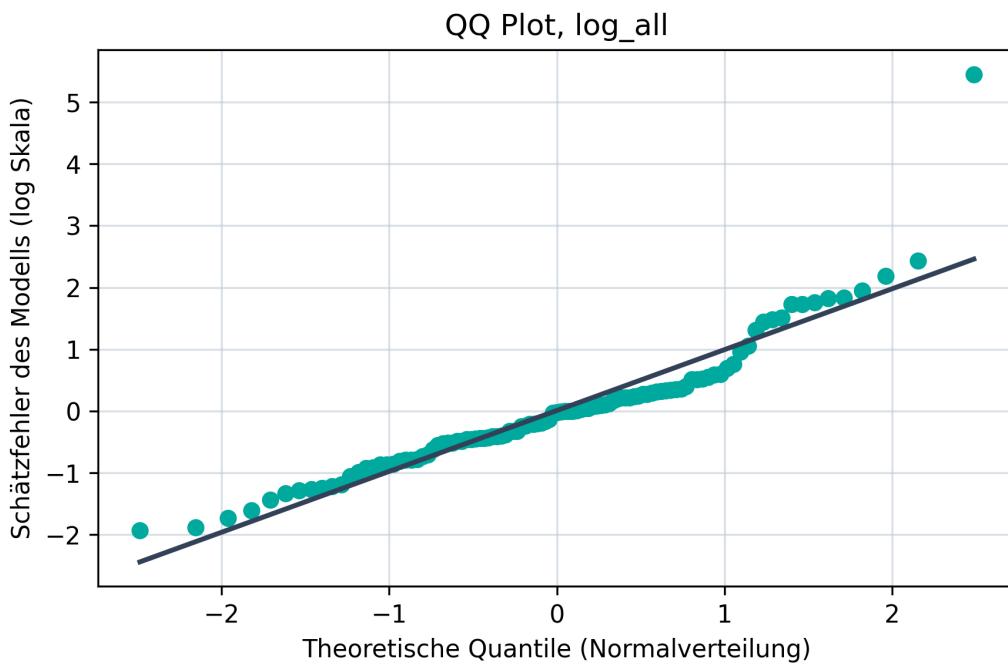


Abbildung A.10: QQ Plot der Residuen des *Eutrophication, terrestrial* Modells auf der Transformationsskala. [Eigene Darstellung]

A.1.6. Regression des Indikators Ozone depletion

Der Indikator liegt auf einer sehr kleinen Skala, daher erscheinen absolute Fehler in gerundeter Darstellung teilweise als Null. Relativ betrachtet zeigen sich dennoch Ausreißer, insbesondere bei sehr kleinen Zielwerten. Abbildung A.11 und Abbildung A.12 verdeutlichen die Streuung und eine rechtsschiefe Fehlerstruktur in den oberen Quantilen.

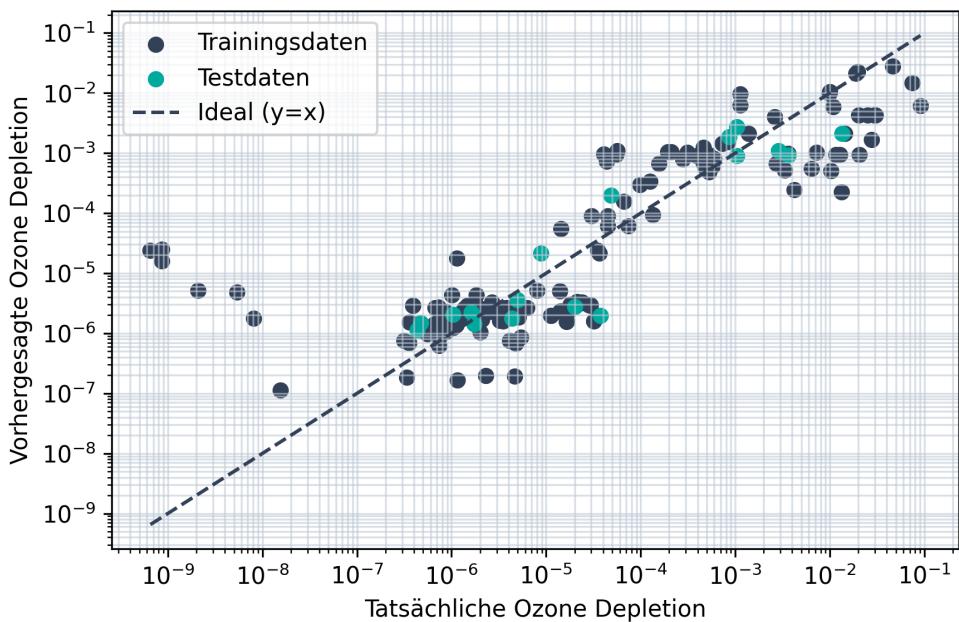


Abbildung A.11: Vorhergesagte gegenüber tatsächlichen Werten von Ozone depletion. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

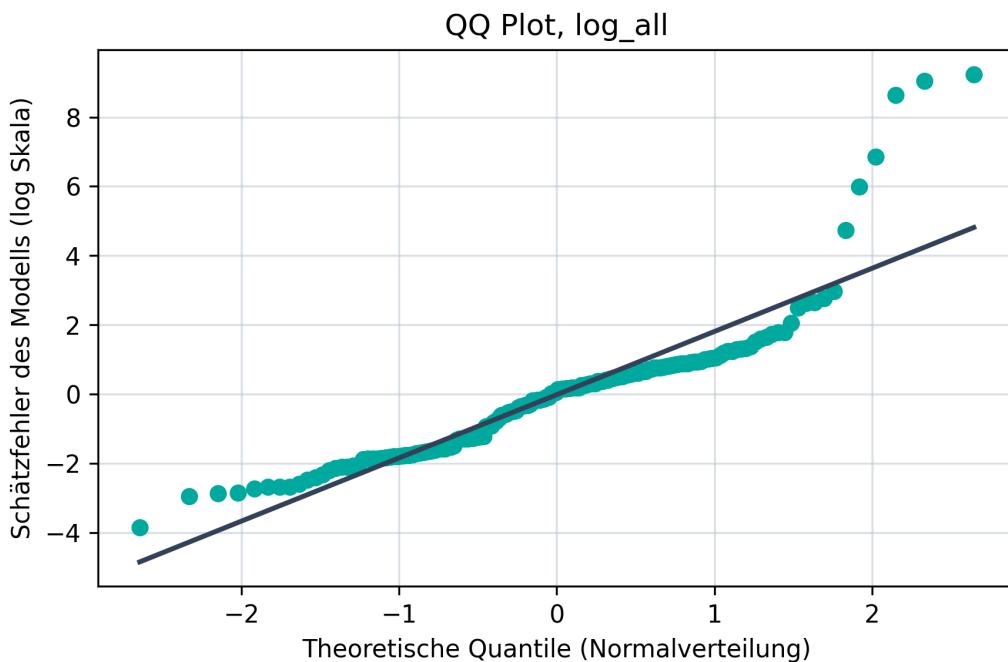


Abbildung A.12: QQ Plot der Residuen des *Ozone depletion* Modells auf der Transformationsskala. [Eigene Darstellung]

A.1.7. Regression des Indikators Resource use, minerals and metals

Die Testgüte ist hoch und der Haupttrend wird gut abgebildet. Einzelne Ausreißer bleiben bestehen, vgl. Abbildung A.13, und die Residuen zeigen schwere Schwänze, vgl. Abbildung A.14.

Die zugehörigen Gütekennzahlen und die gewählte Zieltransformation sind in Tabelle 4.7 zusammengefasst.

A.2. Visualisierung weiterer Regressionsmodelle (Geringe Modellgüte)

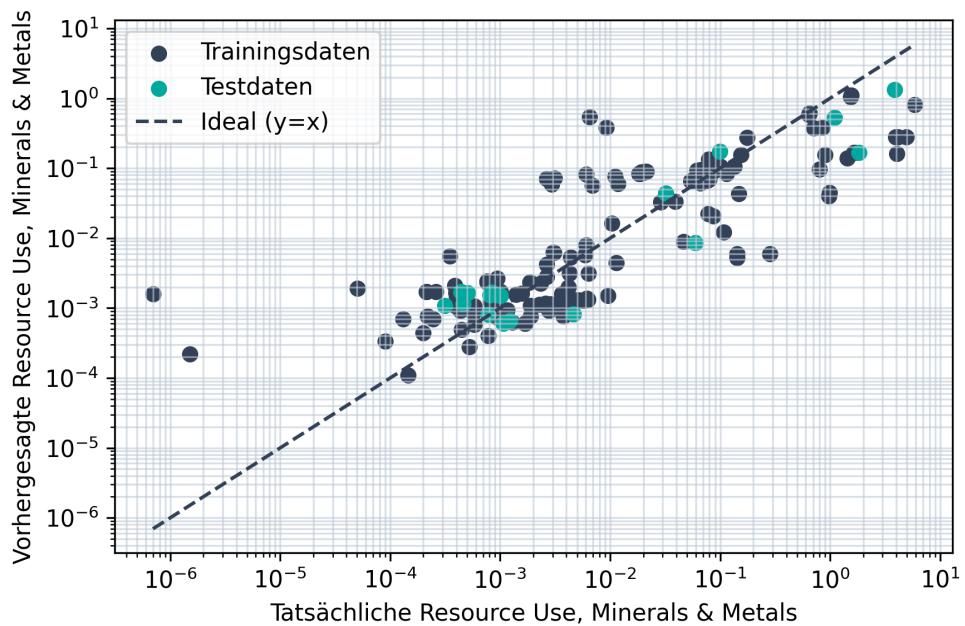


Abbildung A.13: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Resource use, minerals and metals*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

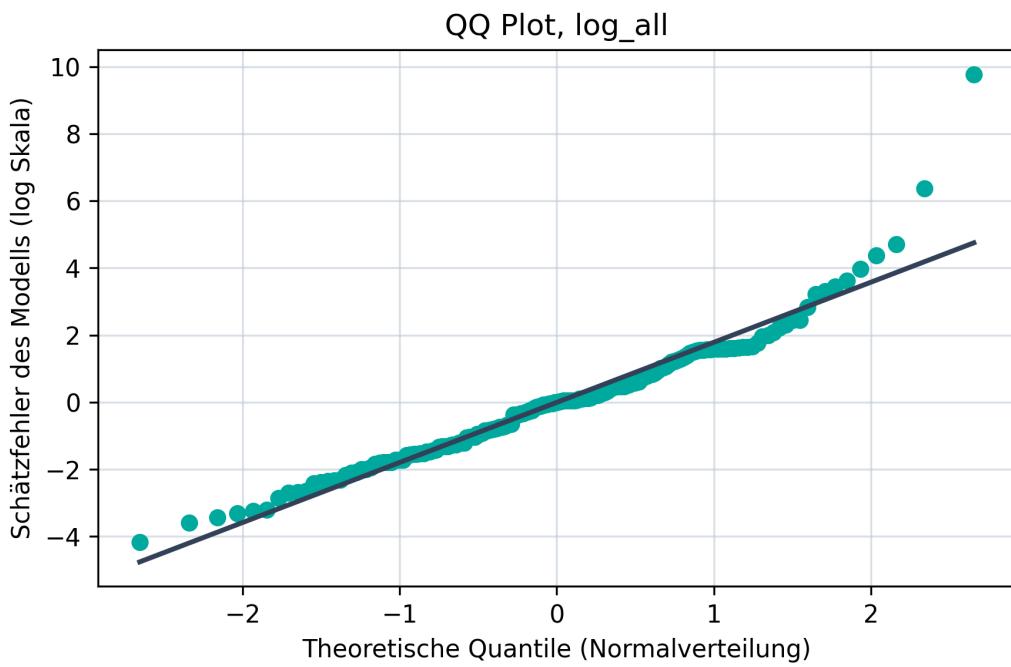


Abbildung A.14: QQ Plot der Residuen des Modells für *Resource use, minerals and metals* auf der Transformationsskala. [Eigene Darstellung]

A.2.1. Regression des Indikators Eutrophication (freshwater)

Für den Indikator *Eutrophication (freshwater)* wird die zuvor aufgebaute Regressionspipeline analog angewendet. Trotz des einheitlichen Modellansatzes fällt die Vorhersagegüte deutlich geringer aus als bei den besser erklärbaren Zielgrößen. Auf dem Testdatensatz wird nur ein begrenzter Anteil der Varianz erklärt ($R^2_{\text{Test}} = 0.434$ bei $n = 133$).

Die robusten Fehlermaße verdeutlichen die Instabilität der Vorhersagen. Der Median der relativen Fehler liegt bei $\text{MdARE} \approx 5.04$, der Mittelwert bei $\text{MARE} \approx 208.43$. Damit treten sehr große relative Abweichungen auf.

Insgesamt ist *Eutrophication (freshwater)* mit dem gewählten Feature-Set aus Gewicht, Stromverbrauch und Material-PCA nur eingeschränkt erklärbar.

Zur Veranschaulichung zeigt Abbildung A.15 das Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten für *Eutrophication (freshwater)* auf logarithmischen Achsen. Es ist erkennbar, dass das Modell große Teile der Zielwerte nur in einem relativ engen Band vorhersagt. Sehr kleine tatsächliche Werte werden deutlich überschätzt, während sehr große Werte tendenziell unterschätzt werden. Damit bildet das Modell eher einen mittleren Größenbereich ab, als die volle Spannweite der Daten, was zur geringen erklärten Varianz und zu den hohen relativen Fehlern passt.

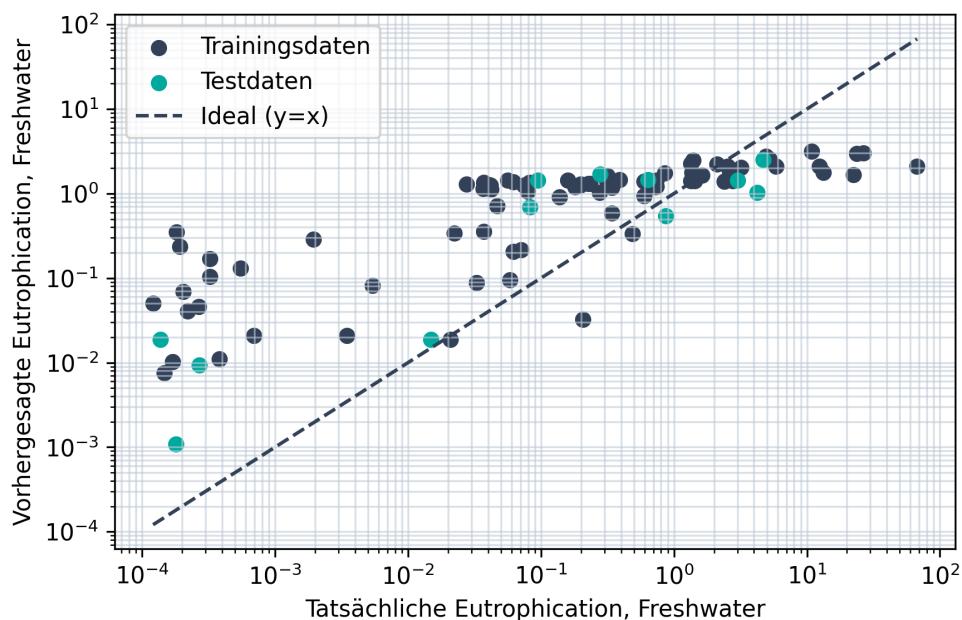


Abbildung A.15: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Eutrophication (freshwater)*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

Abbildung A.16 zeigt den QQ Plot der Residuen auf der Transformationsskala. Ein Teil der Punkte folgt im Zentrum näherungsweise der Referenzgeraden. Die meisten

weichen jedoch, insbesondere in den äußeren Quantilen, deutlich ab. Dies spricht gegen eine Normalverteilung der Fehler und deutet auf schwere Verteilungsschwänze und Ausreißer hin.

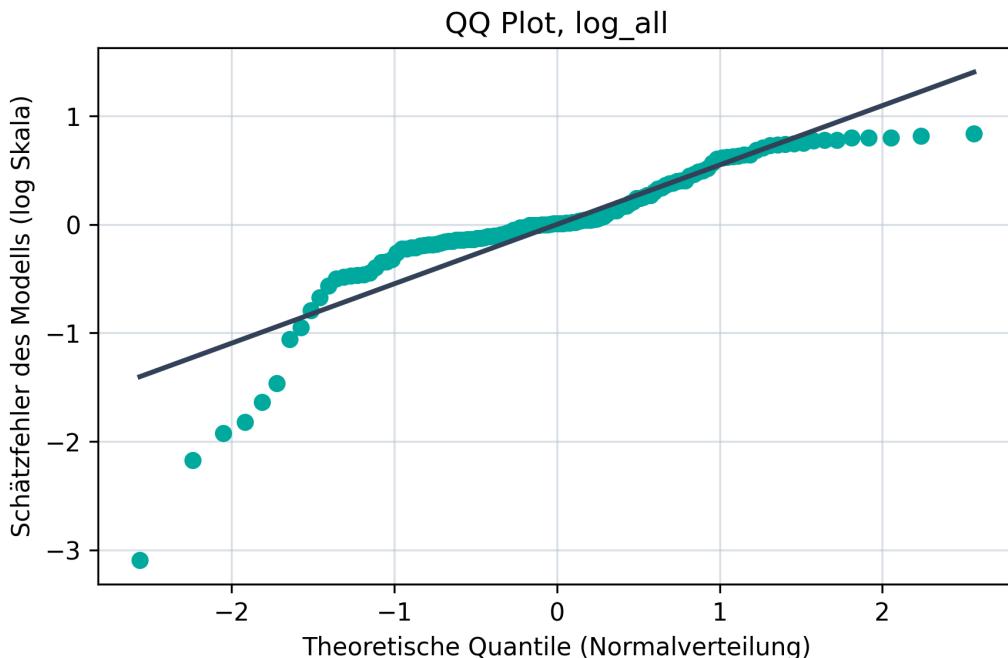


Abbildung A.16: QQ Plot der Residuen des Modells für *Eutrophication (freshwater)* auf der Transformationsskala. [Eigene Darstellung]

A.2.2. Regression des Indikators Eutrophication Marine

Für den Indikator *Eutrophication (marine)* wurde dieselbe Regressionspipeline wie zuvor angewendet. Auch hier bleibt die Testgüte niedrig, und die relativen Fehler sind extrem. Tabelle A.1 fasst die Gütemaße zusammen.

Die geringe erklärte Varianz zeigt, dass Gewicht, Stromverbrauch und die verdichtenen Materialinformationen die Streuung dieses Indikators nur unzureichend abbilden. Die sehr hohen relativen Fehler deuten zusätzlich auf starke Instabilität hin, insbesondere bei kleinen Zielwerten. Dass der RMSE ohne die Top 20 Beobachtungen deutlich kleiner ausfällt, spricht dafür, dass wenige extreme Fälle die Fehlerstatistik dominieren.

Abbildung A.17 zeigt, dass das Modell den groben Trend nur schwach trifft und viele Punkte weit von der Ideallinie entfernt liegen, selbst auf logarithmischer Skala. Der QQ Plot in Abbildung A.18 bestätigt eine deutliche Abweichung von der Normalverteilung auf der Transformationsskala, was auf schwere Schwänze und systematische Modellfehler hinweist.

Tabelle A.1: Gütekennzahlen des linearen Regressionsmodells (Eutrophication (marine) als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.322
RMSE _{Test}	25.0197
Rel. RMSE (RMSE/ \bar{y})	1.7841
MdARE _{Test} (Median rel. Fehler)	7.4875
MARE _{Test} (Mittelwert rel. Fehler)	186.5045
RMSE ohne Top 20	1.2260

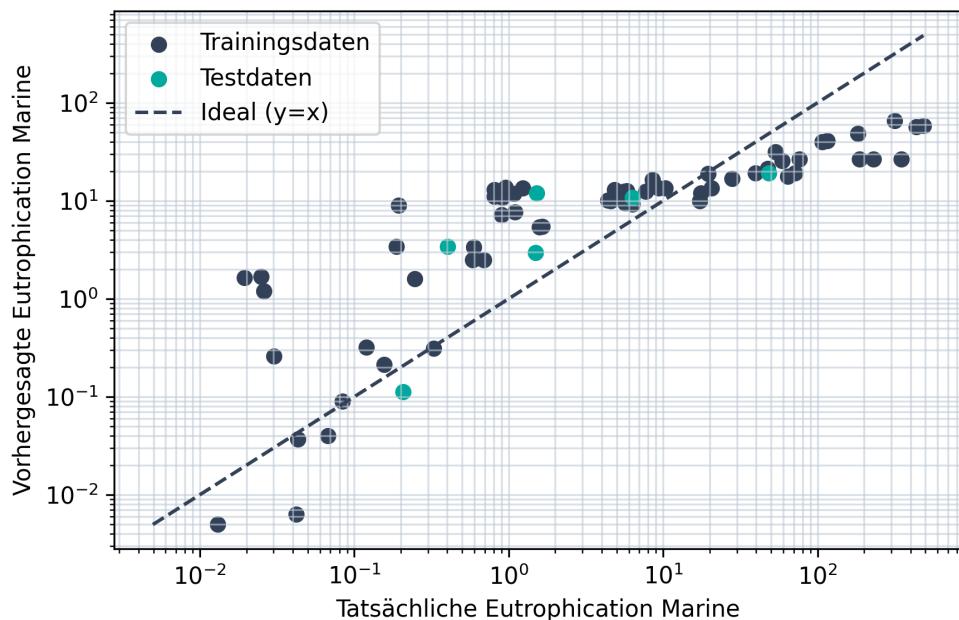


Abbildung A.17: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Eutrophication (marine)*. Beide Achsen sind logarithmisch skaliert. [Eigene Darstellung]

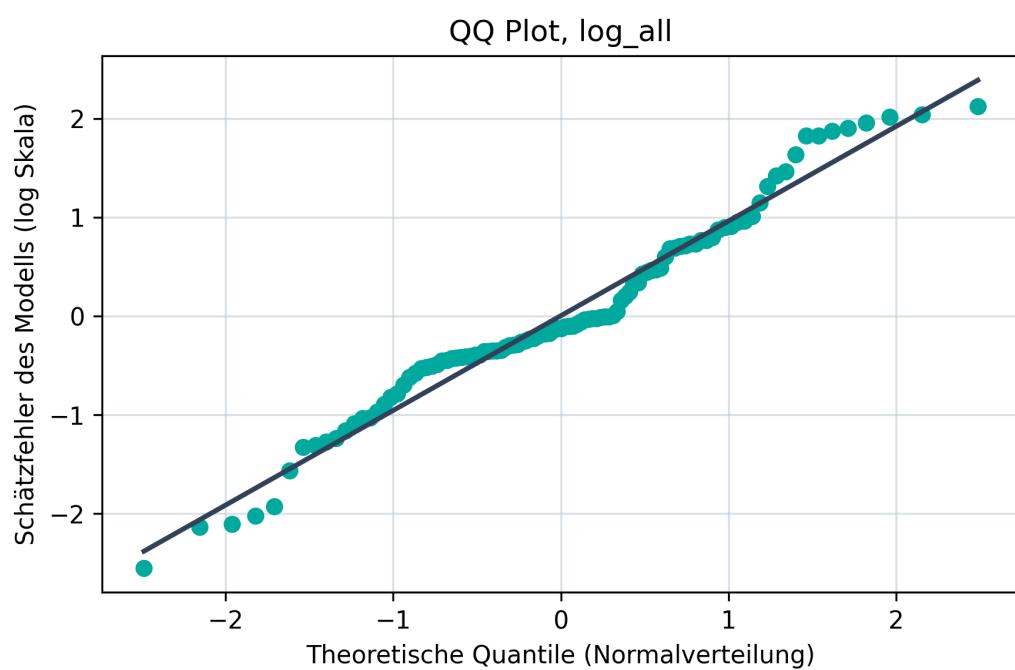


Abbildung A.18: QQ Plot der Residuen des Modells für *Eutrophication (marine)* auf der Transformationsskala.
[Eigene Darstellung]

Literatur

- [Jol82] Ian T. Jolliffe. "A Note on the Use of Principal Components in Regression". In: (1982). DOI: [10.2307/2348005](https://doi.org/10.2307/2348005). URL: <https://academic.oup.com/jrsssc/article/31/3/300/6985100>.
- [MR93] Andrzej Maćkiewicz und Waldemar Ratajczak. "Principal components analysis (PCA)". In: (1993). DOI: [10.1016/0098-3004\(93\)90090-R](https://doi.org/10.1016/0098-3004(93)90090-R). URL: <https://www.sciencedirect.com/science/article/pii/009830049390090R>.
- [LB95] William S. Lovegrove und David F. Brailsford. "Document analysis of PDF files: methods, results and implications". In: (1995). URL: <https://nottingham-repository.worktribe.com/output/1024553>.
- [CF04] Hui Chao und Jian Fan. "Layout and Content Extraction for PDF Documents". In: (2004). DOI: [10.1007/978-3-540-28640-0_20](https://doi.org/10.1007/978-3-540-28640-0_20). URL: https://link.springer.com/chapter/10.1007/978-3-540-28640-0_20.
- [FM09] Murray J. Fisher und Andrea P. Marshall. "Understanding descriptive statistics". In: (2009). DOI: [10.1016/j.aucc.2008.11.003](https://doi.org/10.1016/j.aucc.2008.11.003). URL: <https://www.sciencedirect.com/science/article/abs/pii/S1036731408001732>.
- [AW10] Herve Abdi und Lynne J. Williams. "Principal component analysis". In: (2010). DOI: [10.1002/wics.101](https://doi.org/10.1002/wics.101). URL: <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.101>.
- [MJ10] Gill Marshall und Leon Jonker. "An introduction to descriptive statistics: A review and practical guide". In: (2010). DOI: [10.1016/j.radi.2010.01.001](https://doi.org/10.1016/j.radi.2010.01.001). URL: <https://www.sciencedirect.com/science/article/pii/S1078817410000027>.
- [SYT12] Xiaogang Su, Xin Yan und Chih-Ling Tsai. "Linear regression". In: (2012). DOI: [10.1002/wics.1198](https://doi.org/10.1002/wics.1198). URL: <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.1198>.
- [Has+13] Mehrdad Hassanzadeh u. a. "Environmental declaration in compliance with ISO 14025 thanks to a collaborative program of electrical and electronic industry: The PEP ecopassport program". In: (2013). DOI: [10.1049/cp.2013.0577](https://doi.org/10.1049/cp.2013.0577). URL: <https://ieeexplore.ieee.org/abstract/document/6683180>.

- [Lip+13] Mario Lipinski u. a. "Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents". In: (2013). DOI: [10.1145/2467696.2467753](https://doi.org/10.1145/2467696.2467753). URL: <https://dl.acm.org/doi/abs/10.1145/2467696.2467753>.
- [Gri15] Ralph Grishman. "Information Extraction". In: (2015). DOI: [10.1109/MIS.2015.68](https://doi.org/10.1109/MIS.2015.68). URL: <https://ieeexplore.ieee.org/abstract/document/7243219>.
- [Pez+16] Felipe Pezoa u. a. "Foundations of JSON Schema". In: (2016). DOI: [10.1145/2872427.2883029](https://doi.org/10.1145/2872427.2883029). URL: <https://dl.acm.org/doi/abs/10.1145/2872427.2883029>.
- [BK17] Hannah Bast und Claudius Korzen. "A Benchmark and Evaluation for Text Extraction from PDF". In: (2017). DOI: [10.1109/JCDL.2017.7991564](https://doi.org/10.1109/JCDL.2017.7991564). URL: <https://ieeexplore.ieee.org/abstract/document/7991564>.
- [CZ17] Andreiwid Sheffer Corrêa und Pär-Ola Zander. "Unleashing Tabular Content to Open Data: A Survey on PDF Table Extraction Methods and Tools". In: (2017). DOI: [10.1145/3085228.3085278](https://doi.org/10.1145/3085228.3085278). URL: <https://dl.acm.org/doi/abs/10.1145/3085228.3085278>.
- [KSY18] Parampreet Kaur, Jill Stoltzfus und Vikas Yellapu. "Descriptive statistics". In: (2018). DOI: [10.4103/IJAM.IJAM_7_18](https://doi.org/10.4103/IJAM.IJAM_7_18). URL: https://journals.lww.com/ijam/fulltext/2018/04010/Descriptive_statistics.7.aspx.
- [Sel18] Howard J. Seltman. "Experimental Design and Analysis". In: (2018). DOI: [10.5070/islandora:1012018](https://doi.org/10.5070/islandora:1012018). URL: <https://repository.iit.edu/islandora/object/islandora%3A1012018>.
- [Dim+19] Gabrijela Dimić u. a. "Descriptive Statistical Analysis in the Process of Educational Data Mining". In: (2019). DOI: [10.1109/TELSIKS46999.2019.9002177](https://doi.org/10.1109/TELSIKS46999.2019.9002177). URL: <https://ieeexplore.ieee.org/document/9002177>.
- [ARC21] Anthony C. Atkinson, Marco Riani und Aldo Corbellini. "The Box–Cox Transformation: Review and Extensions". In: (2021). DOI: [10.1214/20-STS778](https://doi.org/10.1214/20-STS778). URL: <https://projecteuclid.org/journals/statistical-science/volume-36/issue-2/The-BoxCox-Transformation-Review-and-Extensions/10.1214/20-STS778.full>.
- [21] *Studie zum Klimaschutz und Energieeffizienz durch digitale Gebäudetechnologien*. Durchgeführt von Bitkom e.V. 2021. URL: https://telematik-markt.de/sites/default/files/news/attachments/211110-bitkom-klimaschutz-und-energieeffizienz-durch-digitale-gebäudetechnologien_0.pdf (besucht am 30.12.2025).

- [MPV22] Douglas C. Montgomery, Elizabeth A. Peck und G. Geoffrey Vining. "Introduction to Linear Regression Analysis". In: (2022). URL: <http://wiley.com/erie/Introduction+to+Linear+Regression+Analysis%2C+6e+Solutions+Manual-p-9781119578765>.
- [Ass24] Association P.E.P. *PEP Ecopassport*. Offizielle Website der Initiative für Umweltdeklarationen elektronischer Produkte. 2024. URL: <https://www.pep-ecopassport.org/> (besucht am 18.10.2025).
- [Aue+24] Christoph Auer u. a. "Docling Technical Report". In: (2024). DOI: [10.48550/arXiv.2408.09869](https://doi.org/10.48550/arXiv.2408.09869). URL: <https://arxiv.org/abs/2408.09869>.
- [24] *France energy mix*. Berichtet von IEA (International Energy Agency). 2024. URL: <https://www.iea.org/countries/france/energy-mix> (besucht am 02.01.2025).
- [Nad+24] Rohaan Nadeem u. a. "Extraction of User-Defined Information from PDF". In: (2024). DOI: [10.1109/DASA63652.2024.10836169](https://doi.org/10.1109/DASA63652.2024.10836169). URL: <https://ieeexplore.ieee.org/document/10836169>.
- [AA25] Narayan S. Adhikari und Shradha Agarwal. "A Comparative Study of PDF Parsing Tools Across Diverse Document Categories". In: (2025). DOI: [10.48550/arXiv.2410.09871](https://doi.org/10.48550/arXiv.2410.09871). URL: <https://arxiv.org/abs/2410.09871>.
- [Ass25] Association P.E.P. *PEP Ecopassport Database*. Offizielle PEP Datenbank. 2025. URL: <https://register.pep-ecopassport.org/pep/consult> (besucht am 11.11.2025).
- [Aue+25] Christoph Auer u. a. "Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion". In: (2025). DOI: [10.48550/arXiv.2501.17887](https://doi.org/10.48550/arXiv.2501.17887). URL: <https://arxiv.org/abs/2501.17887>.
- [25a] *Grobid Github Dokumentation*. Grobid Bibliothek auf Github. 2025. URL: <https://github.com/kermitt2/grobid> (besucht am 23.12.2025).
- [25b] *Layout-Parser Dokumentation*. Dokumentation der Layout-Parser Bibliothek. 2025. URL: <https://layout-parser.github.io/> (besucht am 23.11.2025).
- [Mor+25] José Teófilo Moreira-Filho u. a. "Automating Data Extraction From Scientific Literature and General PDF Files Using Large Language Models and KNIME: An Application in Toxicology". In: (2025). DOI: [10.1002/wcms.70047](https://doi.org/10.1002/wcms.70047). URL: <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcms.70047>.

- [25c] *PEP Passport der Wärmepumpe von Daikin Applied Europe SpA.* 2025. URL: <https://register.pep-ecopassport.org/pep/consult/mbesqrsCBZbWbKJq6-kJ3qnsF1xLwSIzTY0v-ZEkCqc/mbesqrsCBZbWbKJq6-kJ3lQmBuGvAHsLUfQU9idj0pk> (besucht am 24. 12. 2025).
- [25d] *Scikit-learn PCA Dokumentation.* Dokumentation der Python Bibliothek scikit-learn zur PCA. 2025. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html> (besucht am 22. 11. 2025).
- [25e] *Scikit-learn RobustScaler Dokumentation.* Dokumentation der Python Bibliothek scikit-learn zu RobustScaler. 2025. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html> (besucht am 30. 12. 2025).
- [25f] *Scipy boxcox Dokumentation.* Dokumentation der Python Bibliothek scipy zur boxcox Funktion. 2025. URL: <https://docs.scipy.org/doc/scipy-1.16.2/reference/generated/scipy.stats.boxcox.html> (besucht am 23. 12. 2025).
- [Sel25] Lenny Selg. "Analyse und Vergleich von Umweltparametern vernetzter Geräte auf Basis von PEP Deklarationen". In: (2025).
- [YCZ25] Wen Yang, Feifei Cao und Xueli Zhao. "Extraction of PDF Table Data Based on the Pdfplumber Method". In: (2025). DOI: [10.1145/3696474.3696731](https://doi.org/10.1145/3696474.3696731). URL: <https://dl.acm.org/doi/full/10.1145/3696474.3696731>.