

B

C

Bachelor Thesis

Datenanalyse PEP

S

Bachelor Thesis

Datenanalyse PEP

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: TODO

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

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

Abstract

Hier Abstract schreiben

Inhaltsverzeichnis

Abstract	i
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)	19
3.1 Überblick der Pipeline	19
3.2 Datenerhebung und PEP-Recherche	20
3.3 PDF-Parsing und Extraktion	21
3.4 Normalisierung der Daten	23
3.5 Datenbereinigung und Validierung	24
4 Analyse der erarbeiteten Daten	26
4.1 Deskriptive Annäherung an die PEP-Daten	26
4.1.1 Vollständigkeit der Werte	27

4.1.2	Überblick der <i>Input</i> -Variablen	29
4.1.3	Überblick der Umweltindikatoren	33
4.2	Explorative Modellentwicklung	34
4.2.1	Experimentelle Fragestellungen	34
4.2.2	Vergleich der Feature-Sets	35
4.2.3	Vergleich der Regressionsverfahren	37
4.3	PCA der Materialien	38
4.3.1	Ergebnis der PCA	39
4.3.2	Interpretation der Material-Hauptkomponenten	40
4.4	Lineare Regression des Indikators <i>Climate change (total)</i>	41
4.4.1	Datenbasis und Transformation	41
4.4.2	Modellformulierung	42
4.4.3	Schätzverfahren, Validierung und Ergebnisse	42
4.4.4	Visualisierung der Vorhersagequalität	44
4.5	Lineare Regression der anderen Indikatoren	47
4.5.1	Regression des Indikators Acidification	47
4.5.2	Regression des Indikators Hazardous Waste Disposed	50
4.5.3	Regression des Indikators Water Use	53
4.5.4	Regression des Indikators Photochemical Ozone Formation (HH)	55
4.5.5	Regression des Indikators Resource Use, Fossils	58
4.5.6	Regression des Indikators Eutrophication (terrestrial)	60
4.5.7	Regression des Indikators Ozone Depletion	63
4.5.8	Regression des Indikators Resource Use, Minerals and Metals	65
4.5.9	Indikatoren mit geringer Modellgüte	68
5	Diskussion	70
5.1	Einordnung der Ergebnisse	70
5.2	Fehlerstruktur und Modellannahmen	71
5.3	Grenzen des Modells	71
6	Fazit	72
6.1	Zusammenfassung des methodischen Vorgehens	72
6.2	Einordnung der Ergebnisse	73
6.3	Ausblick und zukünftiger Forschungsbedarf	74
	Literatur	75

1

Einleitung

1.1. Motivation und Relevanz

Gebäudeautomation 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 Anteil von etwa einem Drittel dem Gebäudesektor zugerechnet [21]. Digitale Gebäudetechnologien werden in dieser Studie als relevanter Hebel betrachtet, durch bedarfsgerechte, intelligente Steuerung und fortgeschrittenem Monitoring den Energieverbrauch und die Emissionen von Gebäuden zu senken.

Neben eines erhöhten Komforts können Gebäudeautomation 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 die modellierten CO₂-Minderungspotenziale 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 Einspa-

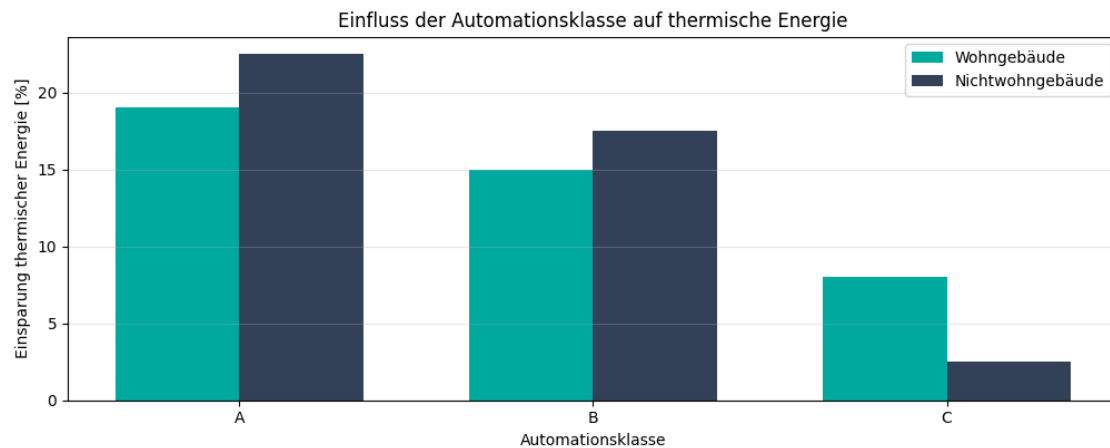


Abbildung 1.1: Einsparungen thermischer Energie nach Automationsklassen, relativ zur Klasse D (Zahlen aus der Studie [21]).

rungen 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 sind 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 und einer Modellentwicklung.

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 sowie die eingeschränkte Übertragbarkeit auf einzelne Produktkategorien und Nutzungsszenarien werden.

1.3. Wissenschaftliche Fragestellung

Aus der Zielsetzung ergibt sich die folgenden zentrale Fragestellung:

Wie lassen sich Umweltwirkungen von Produkten der Gebäudeautomatisierung auf Basis von PEP-Ecopassport- Deklarationen systematisch extrahieren und statistisch analysieren, und inwieweit lassen sich daraus robuste, interpretierbare Modelle ableiten, die Umweltindikatoren auch für Produkte ohne PEP heuristisch abschätzen können?

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 sechs Kapitel gegliedert.

Kapitel 2 beschreibt die theoretischen Grundlagen. Zunächst werden der PEP-Eco-passport Standard und der typische Aufbau von PEP-Dokumenten erläutert. Anschließend werden Herausforderungen der strukturierten Informationsextraktion aus PDF-Dokumenten sowie 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 und Verteilungen der Input-Variablen sowie 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)* sowie für weitere Indikatoren geschätzt und hinsichtlich Vorhersagegüte und Fehlerstruktur bewertet.

Kapitel 5 diskutiert die Ergebnisse im Hinblick auf praktische Aussagekraft, Fehlermuster und Modellannahmen. Zudem werden Grenzen der Übertragbarkeit, der Einfluss von Datenheterogenität sowie mögliche Bias-Quellen reflektiert.

Kapitel 6 fasst das methodische Vorgehen und die wichtigsten Ergebnisse zusammen und gibt einen Ausblick auf Weiterentwicklungen der Pipeline, zusätzliche Merkmale und möglichen weiteren Forschungsbedarf.

2

Theoretische Grundlagen

Hier werden die theoretischen Grundlagen für die vorliegende Arbeit gelegt. Für die standardisierte Berichterstattung von Umweltwirkungen existieren deklarative 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 rohen 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, Inhalt und Struktur 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, Anwendungsbereiche und gegebenenfalls weitere Varianten aufgeführt.

Materialzusammensetzung Die Zusammensetzung des Produkts wird teils in Tabellenform, 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 des Systemgrenzenmodells.

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 sowie 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 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) sowie 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 ist nicht trivial und 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], sowie *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. Das Layoutmodell *DocLayoutNet* 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 [Pez+16]. JSON bildet Daten über einfache Typen sowie 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 [Pez+16].

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 erheblichen 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 ein-

gesetzt.

2.3. Statistische Grundlagen

Die in dieser Arbeit verwendeten statistischen Verfahren bilden die methodische Grundlage zur Analyse und Modellierung der aus PEP-Ecopassports 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 der *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 Schätzfehler 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. Eine gängige Transformation um Rechtsschiefe zu korrigieren, ist 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.

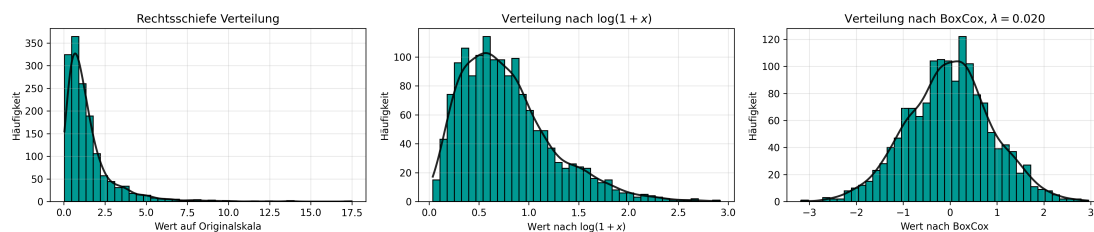


Abbildung 2.1: Beispiel einer rechtsschiefen Verteilung vor und nach Transformationen. Die überlagerte Dichtekurve verdeutlicht die Form der Verteilung.

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 Projektio-

nen 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.

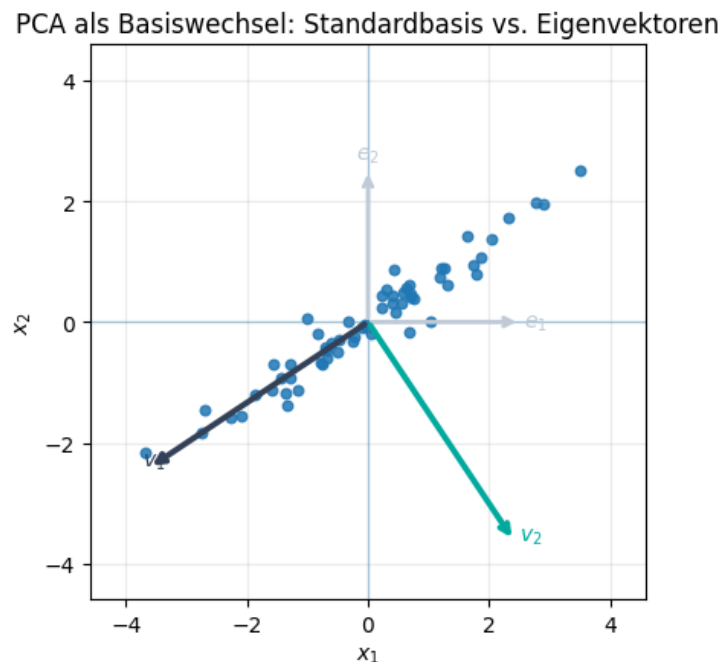


Abbildung 2.2: PCA als Basiswechsel: Standardachsen (x_1, x_2) und Hauptachsen (v_1, v_2) der Punktwolke.

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 (PCR) 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 quantifizieren und zur Abschätzung unbekannter Werte nutzbar zu machen.

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.

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.

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 [MPV22]. 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 sowie die Struktur und Aufbereitung der resultierenden Datenbasis.

3.1. Überblick der 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, als Output entsteht 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 definierten Variablen 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.

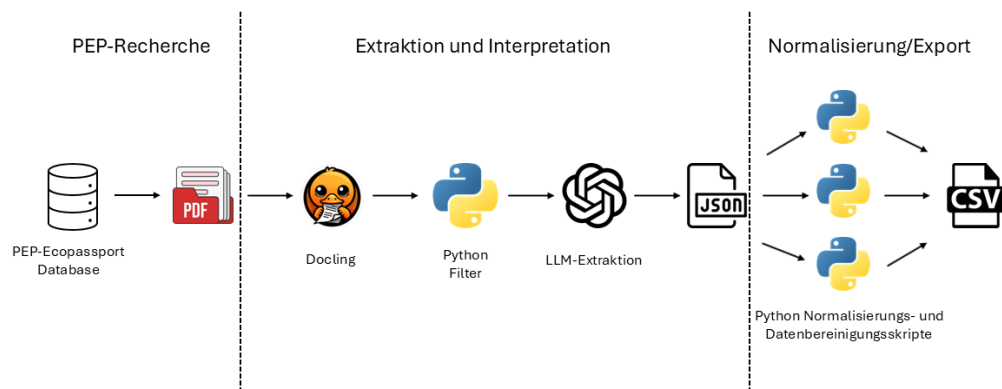


Abbildung 3.1: Schematischer Aufbau der Pipeline: von der PEP-Erfassung bis zur strukturierten Datenbasis.

3.2. Datenerhebung und PEP-Recherche

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

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-Formaten 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, geparkt 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 Produktitel nicht direkt auf Komponenten der Gebäudeautomation schließen können. Zudem wendet die Suchfunktion der PEP-Plattform nicht alle Filter korrekt an und es treten Überschneidungen zwischen den Seiten auf.

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 enorm vergrößern.

3.3. PDF-Parsing und Extraktion

Die Extraktion strukturierter Daten aus PEP-PDFs stellte den technisch anspruchsvollsten 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 Docling 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 Materiallisten. 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) 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 und Einheiten stabil erkennen, sowie 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 Bezeichnungen 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 inventorisiert. Für jedes relevante Feld (z. B. Indikatoreinheiten, Materialbezeichnungen oder Energiequellen) 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. Folgende Datendomänen wurden vereinheitlicht: 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-A3 und *Manufacturing* zu *manufacturing*) und Strommixe (z. B. *France grid mix*, *France Mix*, *French grid* zu *FR*). Diese Mappings wurden schrittweise verfeinert, bis alle identifizierten Abweichungen 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 *sort_out_unusable.py* 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 Aussonderung von 8 der insgesamt 242 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 auf fehlende oder ungültige Einheiten, numerische Typfeh-

ler (z. B. String statt numerischer Wert) sowie Plausibilitätsprüfungen, etwa auf Null- oder Extremwerte bei zentralen Variablen. Ein zusätzlicher Kontrolllauf identifizierte Datensätze mit nicht plausiblen Summen 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 wurden anschließend mit `flatten_jsons.py` 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 Länderschreibweisen sowie geprüfte Indikatorwerte, wodurch eine zuverlässige 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 Annäherung an die 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 sind die Grundlage der Heuristik, die eine Abschätzung von Umweltindikatoren für Produkte ohne PEP-Ecopassport ermöglichen soll und adressiert damit die zentrale Zielsetzung der Arbeit.

4.1. Deskriptive Annäherung an die 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 sowie über zentrale Eingangs- und Ausgangsvariablen 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 Nullwerte aus der Datenpipeline als auch Indikatoren, 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 marines 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.

Der Indikator *Wasserknappheit* wird aufgrund der hohen Ausfallrate von der Auswertung ausgeschlossen, da keine belastbaren statistischen Aussagen getroffen werden können.

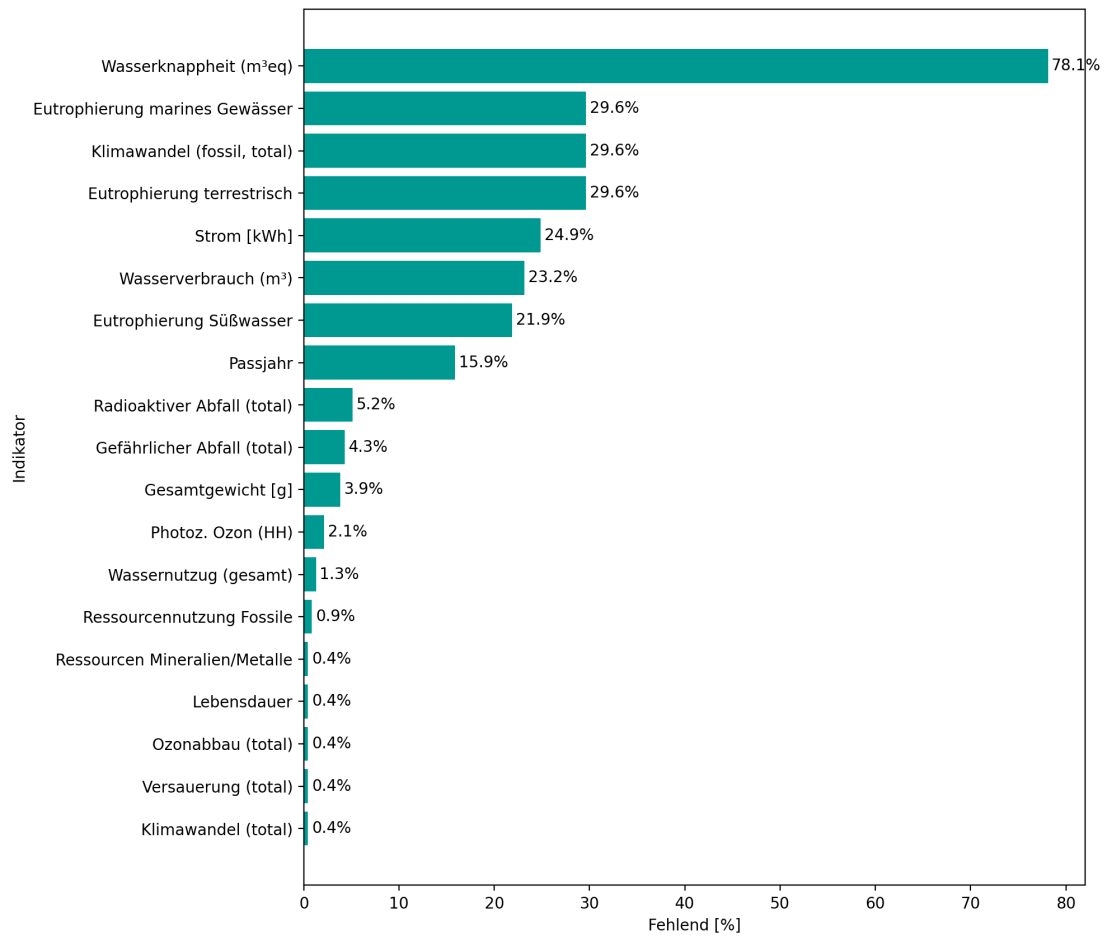
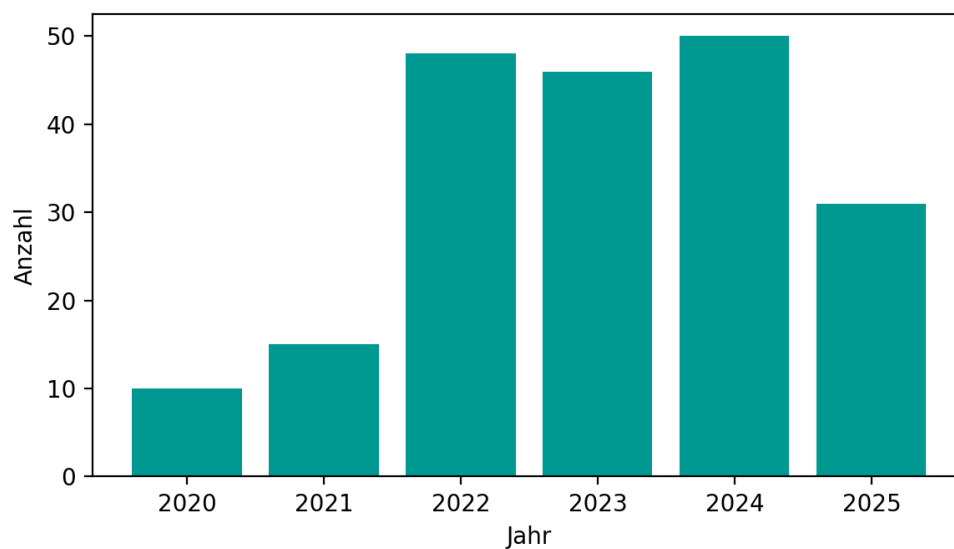


Abbildung 4.1: Anteil fehlender Werte pro Umweltindikator (%; $N = 233$ Produkte).

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 4.2: Erscheinungsjahre der analysierten PEP-Dokumente ($N = 233$ Produkte).

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 in Tab. 4.1 dargestellten Basisvariablen 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 drehzahlgeregeltem 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 nahe-

zu 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.

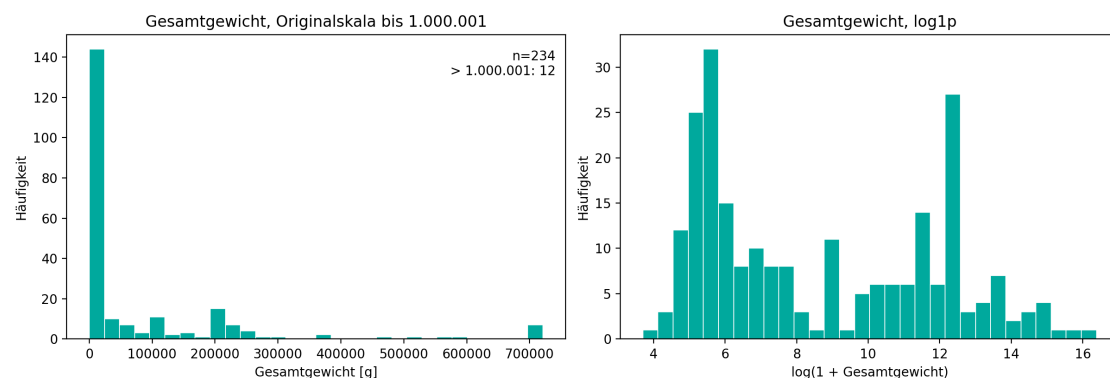


Abbildung 4.3: Histogramme des Gesamtgewichts auf Originalskala (links, bis zu 1.000.000 kWh) und nach log1p-Transformation (rechts).

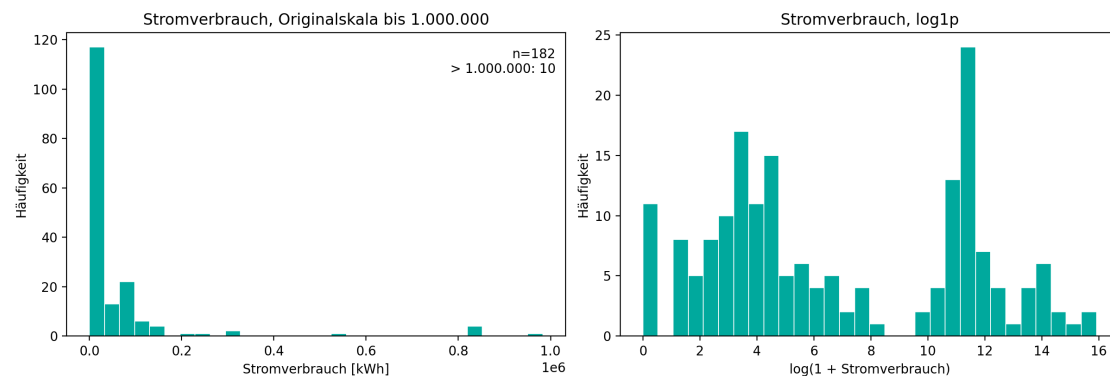


Abbildung 4.4: Histogramme des Stromverbrauchs auf Originalskala (links, bis zu 1.000.000 g) und nach log1p-Transformation (rechts).

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 dominieren 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 vermarkteten 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.

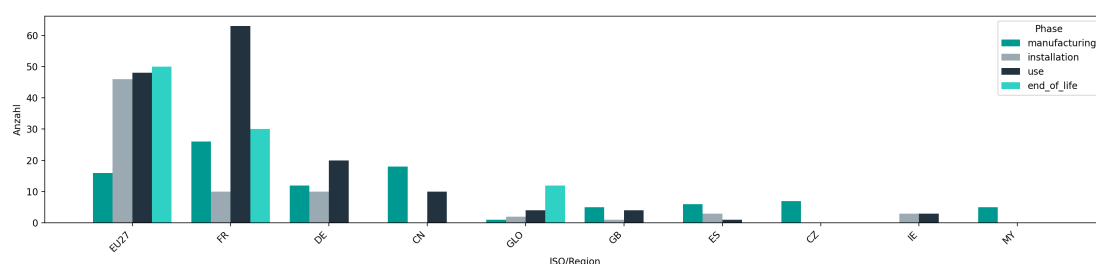


Abbildung 4.5: Verteilung der verwendeten Energiemodelle (ISO-Regionen) über die Lebenszyklusphasen

Eine weitere Variable, die die Umweltindikatoren stark beeinflussen, und damit in der zu entwickelnden Heuristik eine Rolle spielen muss, 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.

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	Epoxidharz	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	–

Dazu wurde der Mittelwert jedes Materials über alle Produkte hinweg berechnet. *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 summierter Anteil jedoch nicht vernachlässigbar ist. Insgesamt spiegelt sich in der Verteilung die Diversität der erfassten Produktgruppen wider.

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).

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 Origineleinheiten stark unterscheiden.

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

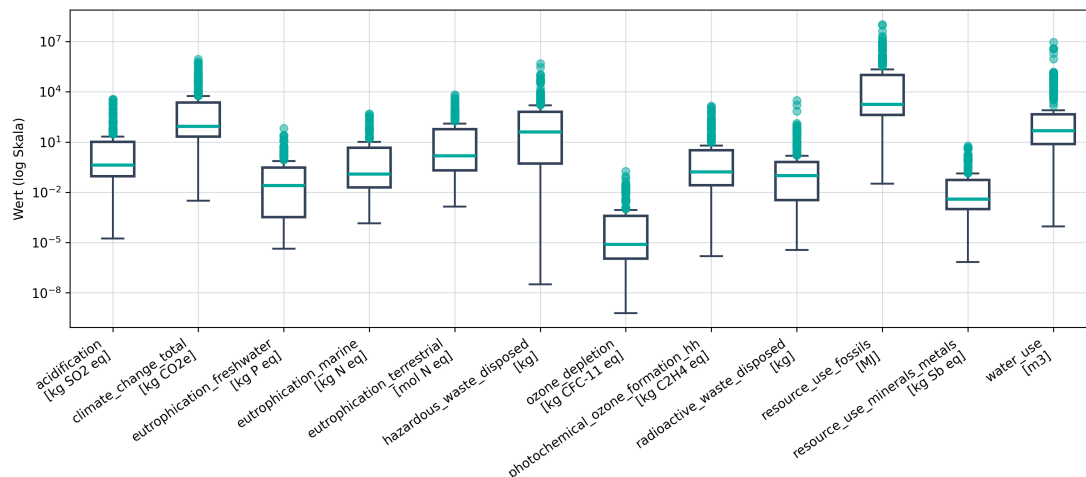


Abbildung 4.6: Boxplots der Total-Werte je Indikator auf logarithmischer Skala.

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 sowie 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 Modellent-

wicklung 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_{1p}) 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 drei 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 Rohmaterialspalten 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 drei Varianten wurden dieselben Train/Test-Splits verwendet, so dass die Gütemaße direkt vergleichbar sind. Tabelle 4.4 zeigt die erzielten Ergebnisse.

Tabelle 4.4: Vergleich verschiedener Feature-Sets für den Indikator *Climate change (total)* auf der Skala $\log(1 + \text{CO2}_{\text{total}})$.

Modellvariante	R^2_{Train}	R^2_{Test}	RMSE _{Test}
Basis (Gewicht, Strom)	0.817	0.770	1.904
Basis + Energiemodelle	0.668	0.501	2.302
Basis + Rohmaterialien	0.904	0.567	2.140
Basis + Rohmaterialien ($n \geq 10$)	0.842	0.832	1.152
Basis + PCA-Materialien ($n \geq 10$)	0.887	0.882	1.095
PCA auf alle Variablen	0.822	0.738	1.579

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

Obwohl die Einbeziehung des verwendeten Energiemodells 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. 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. 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 `LinearRegression` aus `scikit-learn` nachgebildet.

Im nächsten Schritt wird das lineare Modell anschließend mit Ridge und Lasso implementiert, die über einen Regularisierungsparameter α die Modellkomplexität steuern und dadurch insbesondere bei vielen, korrelierten Regressoren stabilere Schätzungen liefern können.

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.

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

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

Regressionsverfahren	R^2_{Train}	R^2_{Test}	RMSE _{Test}
LinearRegression	0.887	0.882	77565.84
Ridge	0.906	0.896	59635.57
Lasso	0.906	0.896	59636.03

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 3 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 die Interquartilsbreite (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 der robust skalierten Materialmatrix 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.

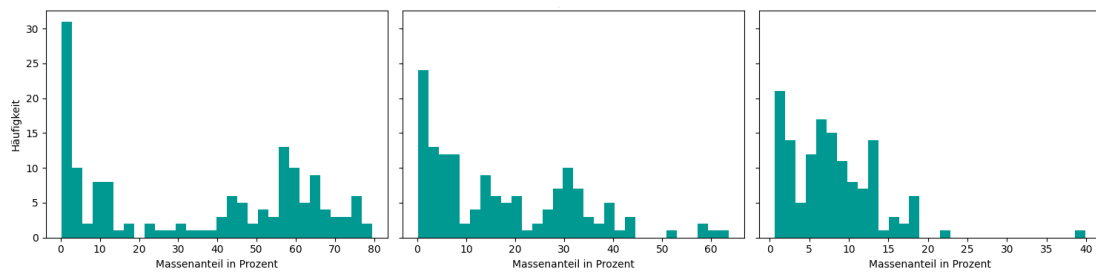


Abbildung 4.7: Histogramme der Materialien mit den größten Anteilen über alle Produkte.

4.3.1. Ergebnis der PCA

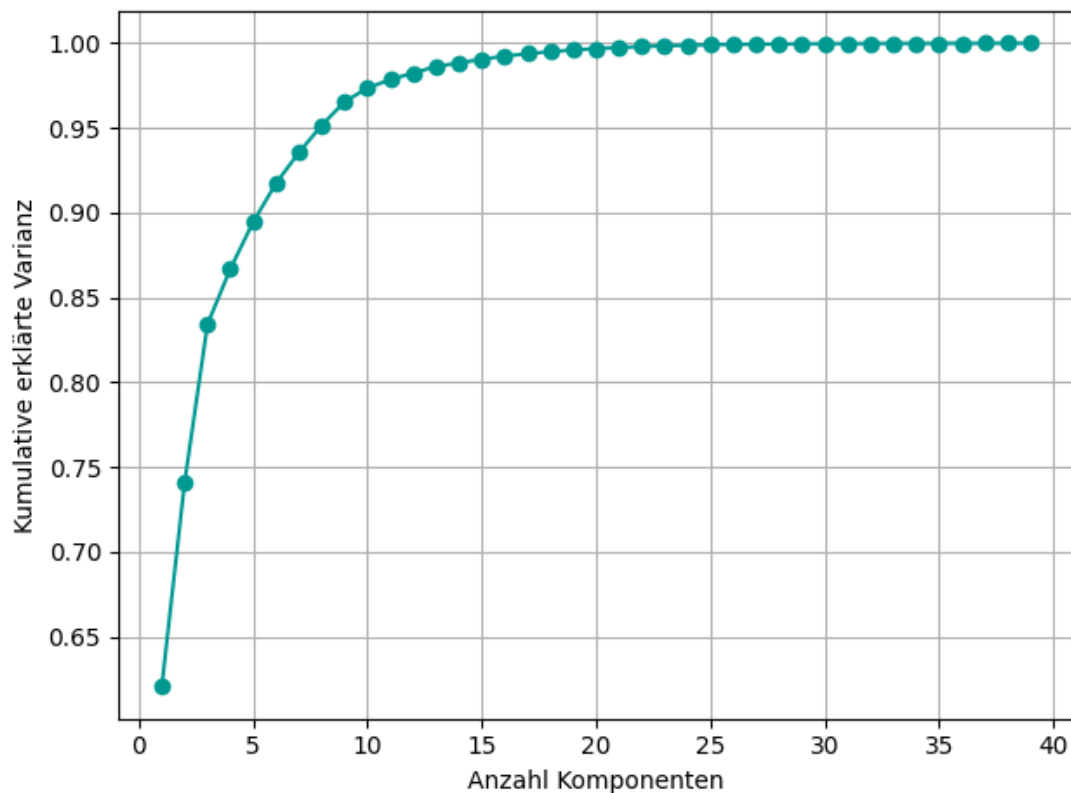


Abbildung 4.8: Kumulative erklärte Varianz

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. Die-

ser 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 sowie 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 95 % 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 ver-

bleiben 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 \log_{1p} transformiert. Es werden die folgenden Größen definiert:

$$\log_cc = \log(1 + \text{CO2}_{\text{total}}), \quad \log_w = \log(1 + \text{weight}), \quad \log_e = \log(1 + \text{electricity_consumption})$$

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

$$\log_cc = \beta_0 + \beta_1 \cdot \log_w + \beta_2 \cdot \log_e + \sum_{j=1}^k \gamma_j \cdot \text{PC_mat}_j + \varepsilon,$$

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 weitere Aufteilungen der 90% Trainingsdaten angepasst und anschließend auf dem unabhängigen Testset ausgewertet. 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. 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.

Der Median der absoluten Fehler liegt bei 2220.80 kg und beschreibt die typische absolute Abweichung eines repräsentativen Produkts. Der Median der relativen Fehler

liegt bei 51.49%. Der Mittelwert des relativen Fehlers 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 CO₂-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.

Eine genannte Annahme der linearen Regression ist, dass die Schätzfehler des Modells näherungsweise normalverteilt sind.

Abbildung 4.10 zeigt einen QQ Plot der Schätzfehler 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 Verteilungsschwänze 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 4.11 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 äußeren Quantilen bleiben Abweichungen sichtbar, was auf eine erhöhte Wahrscheinlichkeit großer Fehler hinweist.

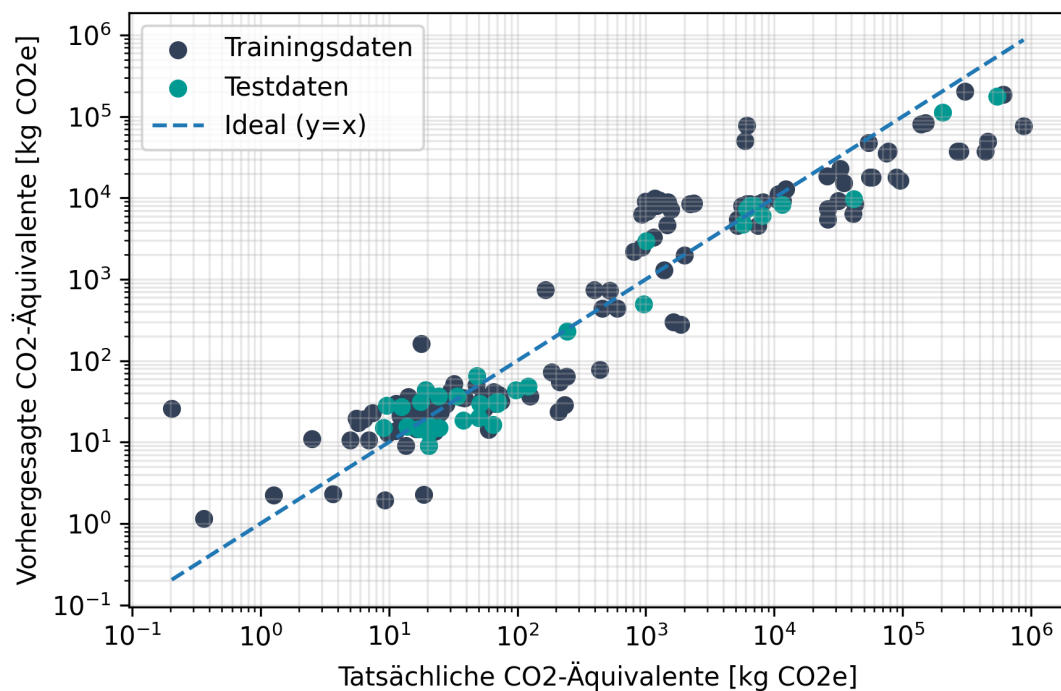


Abbildung 4.9: Vorhergesagte und tatsächliche Werte von *Climate Change total* für einen exemplarischen Train und Test Split des CO₂-Modells.

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.

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 Fehler-

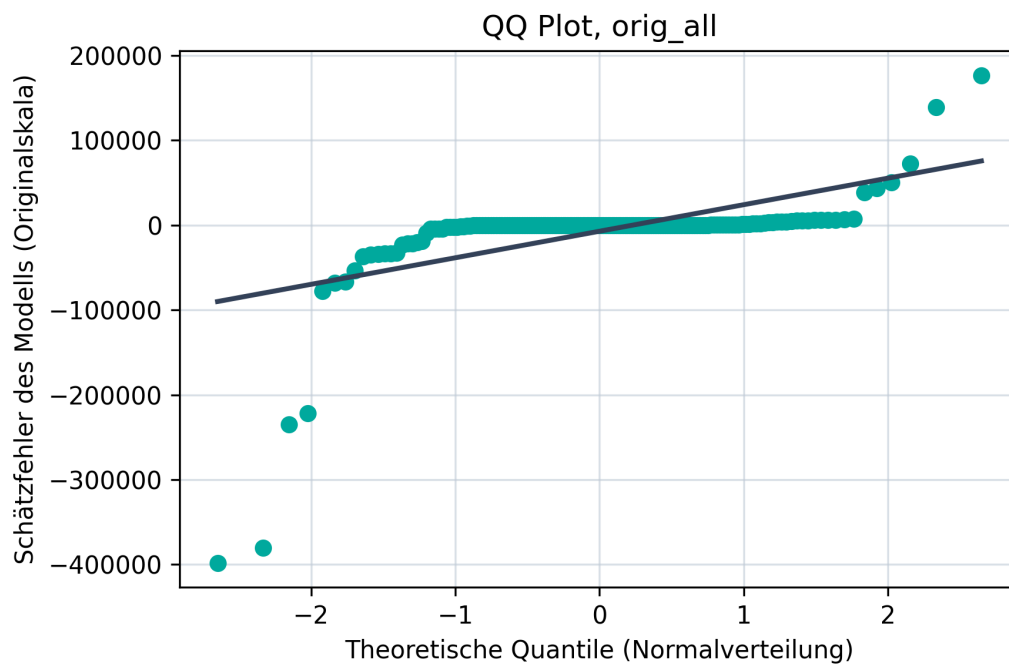


Abbildung 4.10: QQ Plot der Residuen des CO₂-Modells auf Originalskala im Testset.

teanteile, und die Erklärbarkeit der jeweiligen Indikatoren. Für die übrigen Indikatoren werden die getesteten Transformationen (keine Transformation, \log_{1p} , 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.

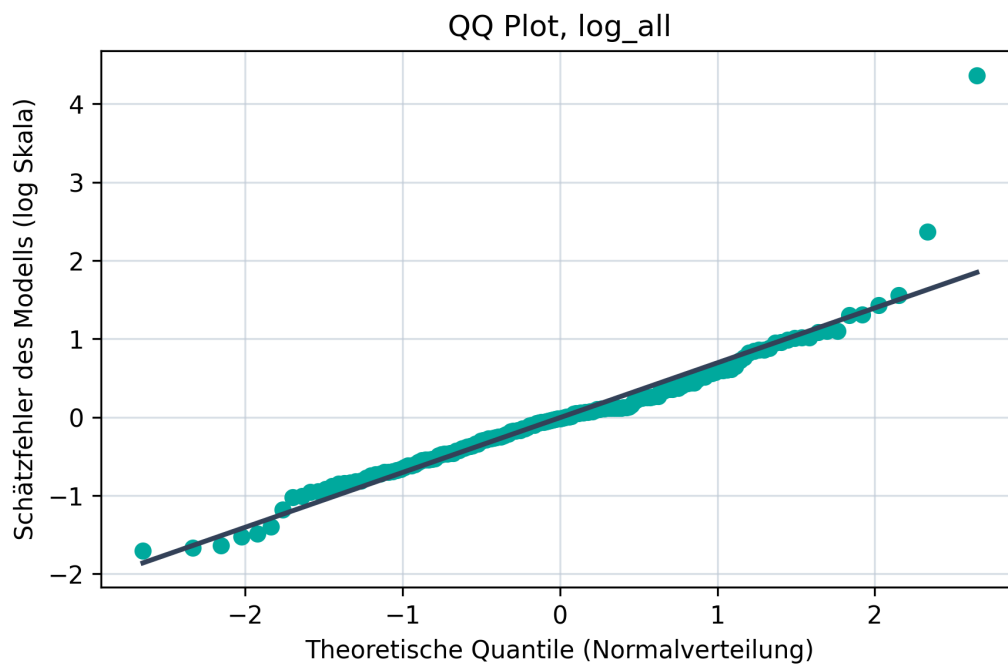


Abbildung 4.11: QQ Plot der Residuen des CO₂-Modells im Testset auf der Transformationsskala \log_{cc} .

4.5. Lineare Regression der anderen Indikatoren

Die für CO₂ aufgebaute Regressions-Pipeline wird nun auf weitere Umweltindikatoren angewendet, um zu prüfen, wie gut sich diese mit denselben Produktmerkmalen erklären lassen.

4.5.1. Regression des Indikators Acidification

Neben den CO₂-Äquivalenten wurde das lineare Regressionsmodell auch auf den Indikator *Acidification* angewendet. Es wurden keine Transformation, \log_{1p} und eine Box-Cox Transformation verglichen. Die beste Testgüte wird mit \log_{1p} erreicht, daher wird im Folgenden dieses Modell berichtet. Daher wird im Folgenden dieses Modell berichtet. Die Zielvariable ist dabei der log-transformierte Gesamtwert des Indikators $\log_{acid} = \log(1 + \text{acidification}_{total})$. Es konnten 177 PEPs benutzt werden. Tabelle 4.7 fasst die Testleistung nach Rücktransformation zusammen.

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

Tabelle 4.7: 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
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0.9579
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	3.8258

weit unter dem RMSE bei etwa 0,29 kg SO₂ und beschreibt damit die typische Abweichung eines repräsentativen Produkts. Der Median der relativen Fehler ($\text{MdARE} \approx 0,96$) zeigt, dass die Vorhersage für ein typisches Produkt häufig in der Größenordnung des wahren Werts liegt. Der deutlich größere Mittelwert der relativen Fehler ($\text{MARE} \approx 3,83$) 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*.

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 10¹ gibt es ein Cluster, der 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 Schätzfehler des Modells im Vergleich zu einer Normalverteilung. Wie beim CO₂ Indikator zeigen sich auf der Originalskala der Fehler typischerweise sehr starke Abweichungen, die eine schwache Interpretationsebene bereitstellen. Daher wird hier nur die Transformationsskala dargestellt. Im mittleren Bereich liegen die Punkte nah an der Referenzgeraden, was darauf hindeutet,

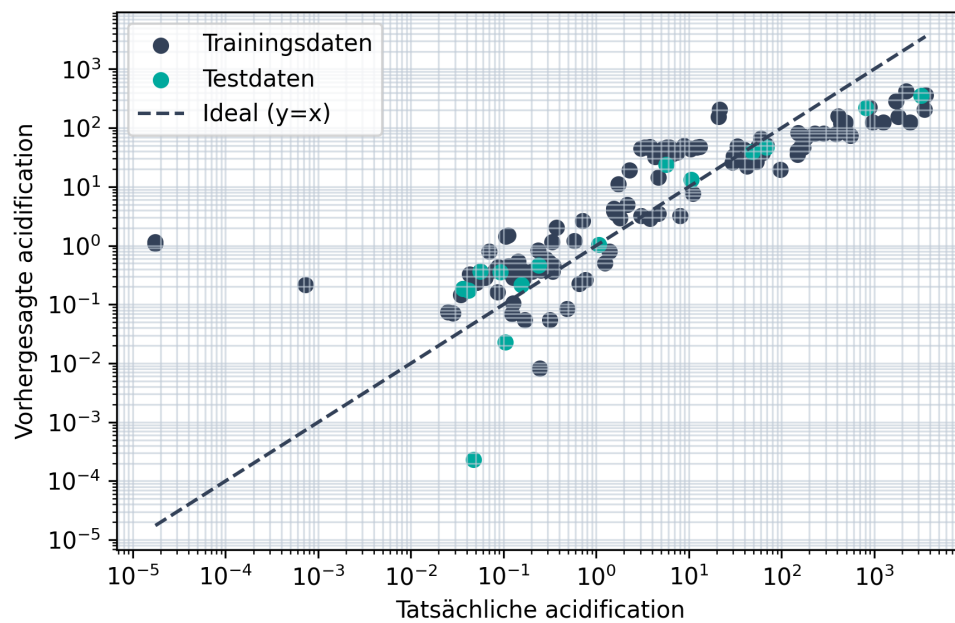


Abbildung 4.12: Vorhergesagte gegenüber tatsächlichen Acidification-Werten. Beide Achsen sind logarithmisch skaliert.

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.

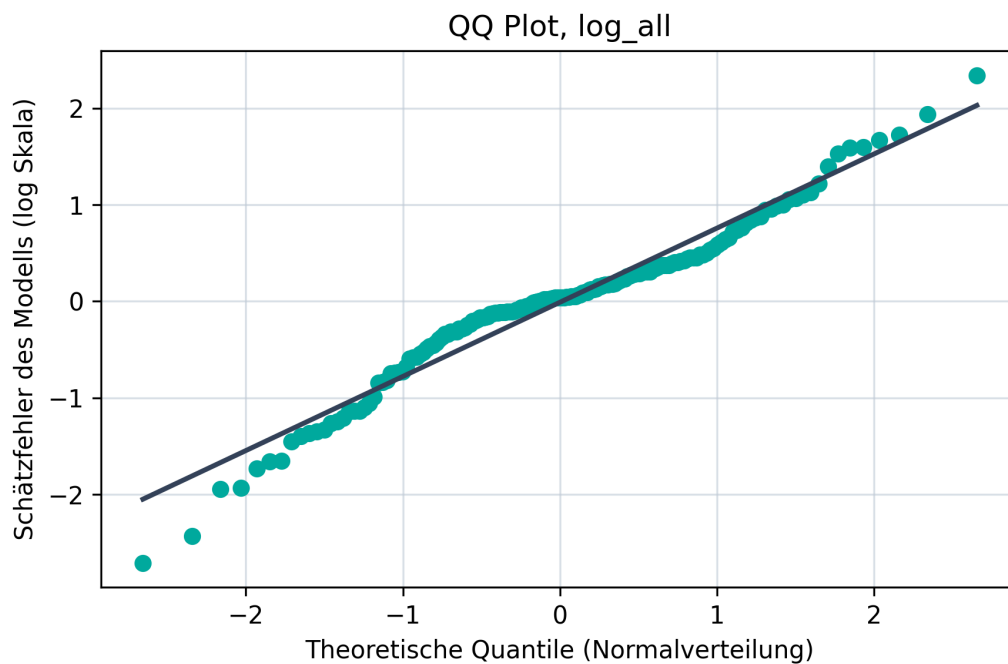


Abbildung 4.13: QQ Plot der Schätzfehler des Acidification Modells (Test- und Trainingsset).

4.5.2. Regression des Indikators Hazardous Waste Disposed

Für den Indikator *Hazardous waste disposed* konnten $n = 168$ PEPs verwendet werden. Es wurden keine Transformation, \log_{1p} und eine Box Cox Transformation verglichen. Die beste Testgüte wird mit \log_{1p} erreicht, daher wird im Folgenden dieses Modell betrachtet. Die Zielvariable ist damit $\log_hwd = \log(1 + \text{hazardous_waste_disposed}_{\text{total}})$. Tabelle 4.8 fasst die Testleistung nach Rücktransformation zusammen.

Das Modell erklärt damit rund 81% der Varianz auf dem Testset. Der RMSE auf Originalskala ist vergleichsweise hoch, was auf einzelne sehr große Abweichungen hinweist. Der Median des absoluten Fehlers liegt bei etwa 177 kg und somit erneut weit unter dem RMSE. Der Median der relativen Fehler ($\text{MdARE} \approx 66\%$) zeigt eine vergleichsweise gute Genauigkeit. Der deutlich größere Mittelwert der relativen Fehler ($\text{MARE} \approx 1314\%$) weist allerdings auf eine stark schiefe Fehlerverteilung mit wenigen extremen relativen Abweichungen hin, was insbesondere bei sehr kleinen Zielwerten plausibel ist.

Zur Veranschaulichung zeigt Abbildung 4.14 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von *Hazardous waste disposed*. Beide Achsen sind logarithmisch skaliert.

Die meisten Punkte liegen im mittleren bis hohen Wertebereich nahe der Diagonalen, so dass der generelle Trend erfasst wird. Bei sehr kleinen tatsächlichen Werten ist

Tabelle 4.8: Gütekennzahlen des linearen Regressionsmodells (*Hazardous waste disposed* als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.813
$\text{RMSE}_{\text{Test}}$	18489.74 kg
Median absoluter Fehler	176.7890 kg
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0.6639
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	13.1431

die Streuung deutlich größer und es treten ausgeprägte Über und Unterschätzungen auf. Dies passt zu den hohen relativen Fehlern, da bereits kleine absolute Abweichungen dort sehr große relative Fehler erzeugen.

Abbildung 4.15 zeigt den QQ Plot der Residuen auf der Transformationsskala. Im Zentrum folgen die Punkte der Referenzgeraden weitgehend, in den Randbereichen sind Abweichungen sichtbar. Damit ist die Normalitätsannahme näherungsweise im mittleren Bereich erfüllt, während einzelne große Fehler auf schwerere Verteilungsschwänze hinweisen.

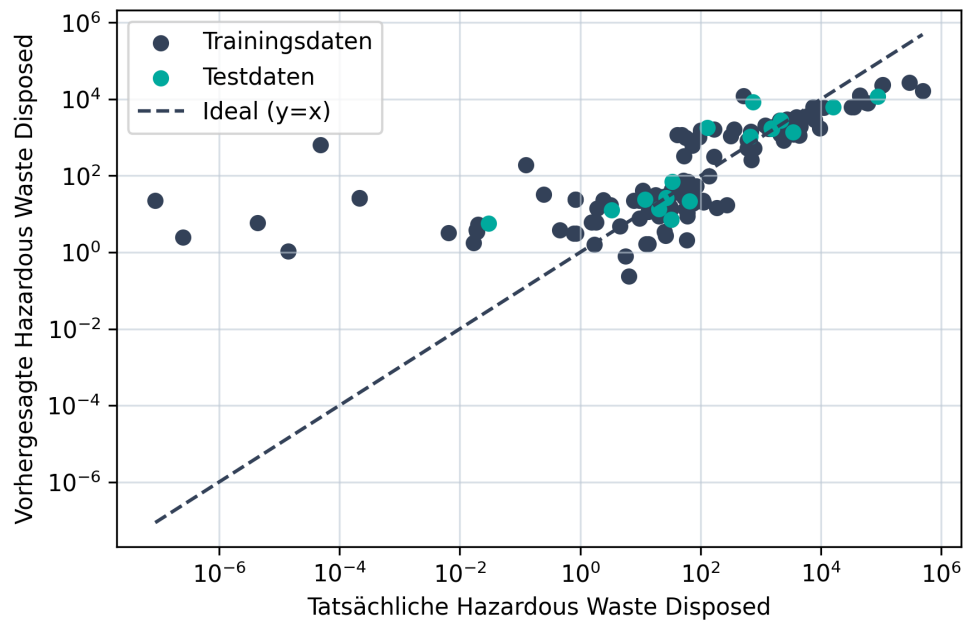


Abbildung 4.14: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Hazardous waste disposed*. Beide Achsen sind logarithmisch skaliert.

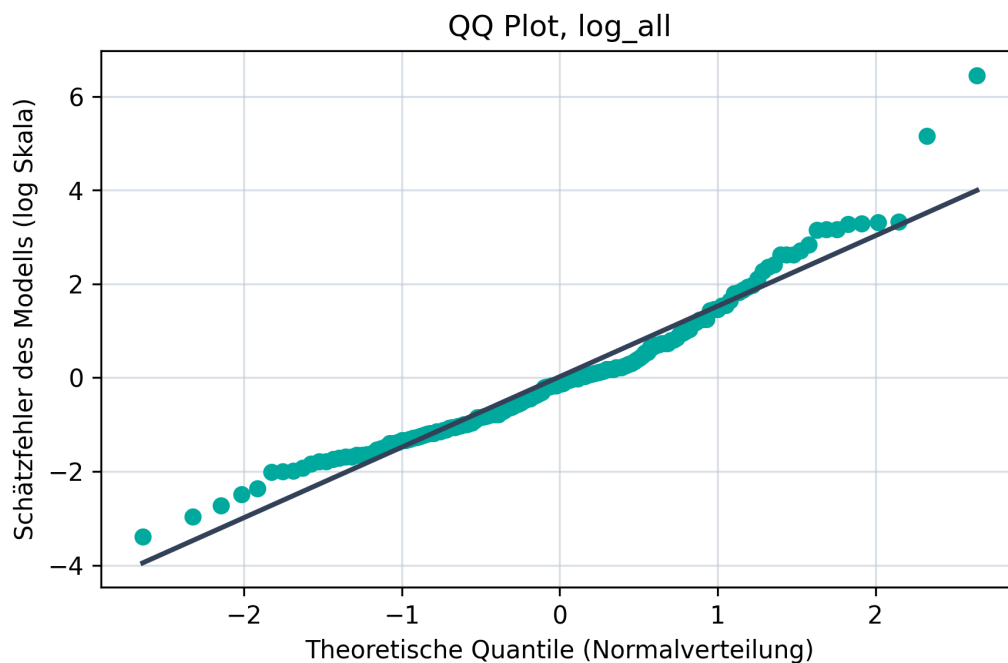


Abbildung 4.15: QQ Plot der Schätzfehler des *Hazardous waste disposed* Modells auf der Transformationskala.

4.5.3. Regression des Indikators Water Use

Für den Indikator *Water use* wird dieselbe Regressionspipeline wie zuvor verwendet. Es wurden keine Transformation, \log_{1p} und eine Box Cox Transformation verglichen. Die beste Testgüte wird mit der Box Cox Transformation erreicht, daher wird im Folgenden dieses Modell berichtet. Tabelle 4.9 fasst die Testleistung nach Rücktransformation zusammen.

Tabelle 4.9: Gütekennzahlen des linearen Regressionsmodells (*Water use* als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0,726
$\text{RMSE}_{\text{Test}}$	28682,36
Median absoluter Fehler	1353,5625
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0,9297
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	2,3278

Das Modell erklärt damit rund 73% der Varianz auf dem Testset. Der RMSE auf Originalskala beträgt 28682,36 und wird durch größere Abweichungen einzelner Produkte geprägt. Der Median des absoluten Fehlers liegt bei 1353,56 und beschreibt die typische Abweichung eines repräsentativen Produkts. Der Median der relativen Fehler ($\text{MdARE} \approx 93\%$) zeigt, dass die Vorhersage für ein typisches Produkt häufig in der Größenordnung des wahren Werts liegt. Der Mittelwert der relativen Fehler ($\text{MARE} \approx 233\%$) ist deutlich höher und weist auf eine schiefe Fehlerverteilung mit einzelnen deutlich schlechter vorhergesagten Beobachtungen hin.

Zur Veranschaulichung zeigt Abbildung 4.16 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von *Water use*. Beide Achsen sind logarithmisch skaliert.

Im mittleren Wertebereich liegen viele Punkte in der Nähe der Diagonalen. Der generelle Zusammenhang wird somit erfasst. Bei sehr großen tatsächlichen Werten liegen mehrere Punkte unterhalb der Diagonalen, was auf eine Unterschätzung besonders wasserintensiver Produkte hindeutet. Bei sehr kleinen tatsächlichen Werten ist die Streuung auf der Log Darstellung relativ groß, wodurch sich sehr große relative Fehler ergeben können. Allgemein wird eine größere Streuung beobachtet als bei den vorherigen Indikatoren, welche sich in den Gütekennzahlen widerspiegelt.

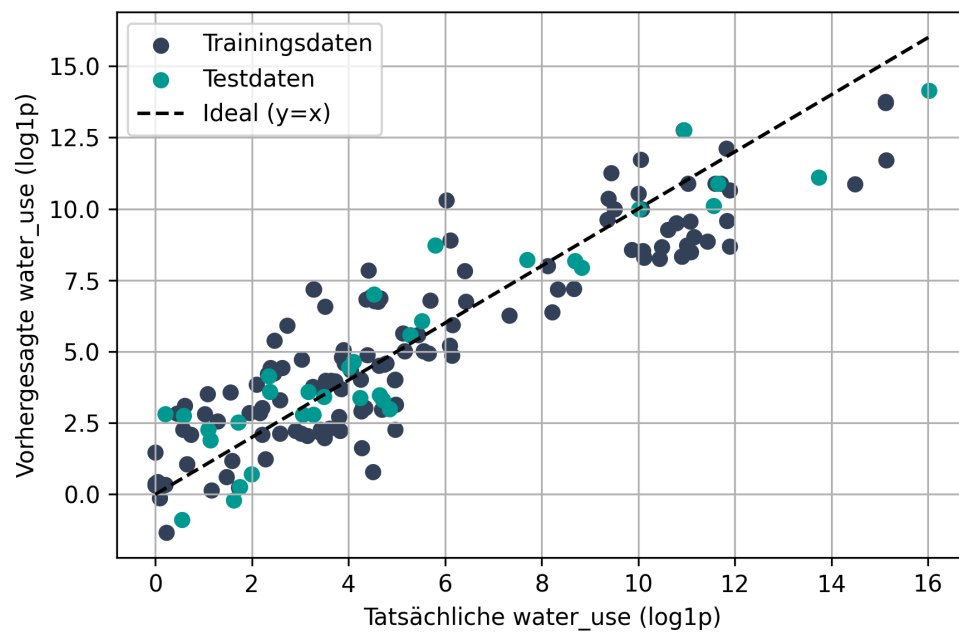


Abbildung 4.16: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Water use*. Beide Achsen sind logarithmisch skaliert.

Abbildung 4.17 zeigt den QQ Plot der Residuen auf der Transformationsskala. Im Zentrum liegen die Punkte näher an der Referenzgeraden, in den Randbereichen treten deutliche Abweichungen auf. Besonders in dem größten dargestellten Quantil kann kein Zusammenhang mit einer Normalverteilung beobachtet werden. Dies weist auf schwerere Verteilungsschwänze und einzelne extreme Residuen hin.

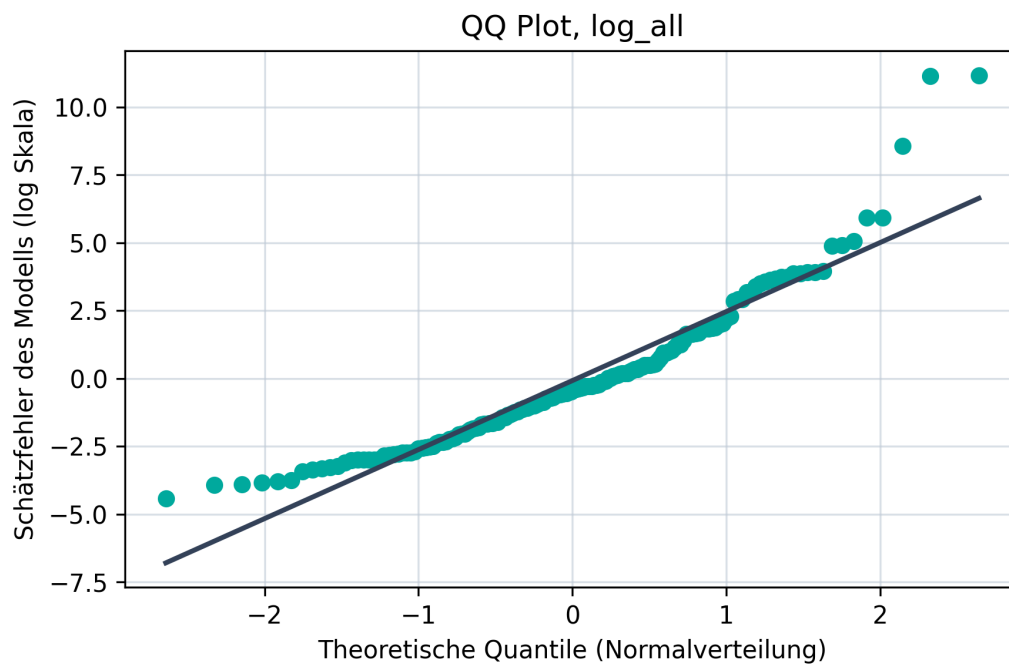


Abbildung 4.17: QQ Plot der Residuen des *Water use* Modells auf der Transformationsskala.

4.5.4. Regression des Indikators Photochemical Ozone Formation (HH)

Für den Indikator *Photochemical ozone formation, human health* (`photochemical_ozone_formation_hh`) wird dieselbe Regressionspipeline wie zuvor verwendet. Es wurden keine Transformation, \log_{1p} und eine Box-Cox Transformation verglichen. Die beste Testgüte wird mit der Box-Cox Transformation erreicht, daher wird im Folgenden dieses Modell berichtet.

Es konnten $n = 171$ PEPs verwendet werden. Tabelle 4.10 fasst die Testleistung nach Rücktransformation zusammen.

Das Modell erklärt damit rund 80% der Varianz auf dem Testset. Der Median des absoluten Fehlers liegt bei 0,0255 kg C_2H_4 und beschreibt die typische Abweichung eines repräsentativen Produkts. Der Median der relativen Fehler ($MdARE \approx 67\%$) deutet auf eine moderate typische relative Genauigkeit hin. Der höhere Mittelwert der relativen Fehler ($MARE \approx 128\%$) weist auf eine schiefe Fehlerverteilung mit einigen deutlich schlechter getroffenen Beobachtungen hin.

Zur Veranschaulichung zeigt Abbildung 4.18 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von PDF_{HH} . Beide Achsen sind logarithmisch skaliert.

Im mittleren Wertebereich liegen viele Punkte nahe der Diagonalen, während bei sehr kleinen Werten eine deutlich größere Streuung sichtbar ist. Für größere tatsächliche Werte liegen mehrere Punkte unterhalb der Diagonalen, was auf eine tendenzielle Un-

Tabelle 4.10: Gütekennzahlen des linearen Regressionsmodells (Photochemical ozone formation, human health als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0,802
$\text{RMSE}_{\text{Test}}$	109,1708 kg C ₂ H ₄
Median absoluter Fehler	0,0255 kg C ₂ H ₄
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0,6665
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	1,2811

terschätzung hoher Werte hindeutet. Einige Punkte liegen deutlich oberhalb der Diagonalen, was auf wenige starke Überschätzungen einzelner Beobachtungen hinweist.

Abbildung 4.19 zeigt den QQ Plot der Residuen auf der Transformationsskala. Einige Punkte liegen nahe an der Referenzgeraden, allerdings weichen die meisten Punkte, vor allem in den Randbereichen, deutlich ab. Dies spricht für schwerere Verteilungsschwänze und einzelne extreme Residuen.

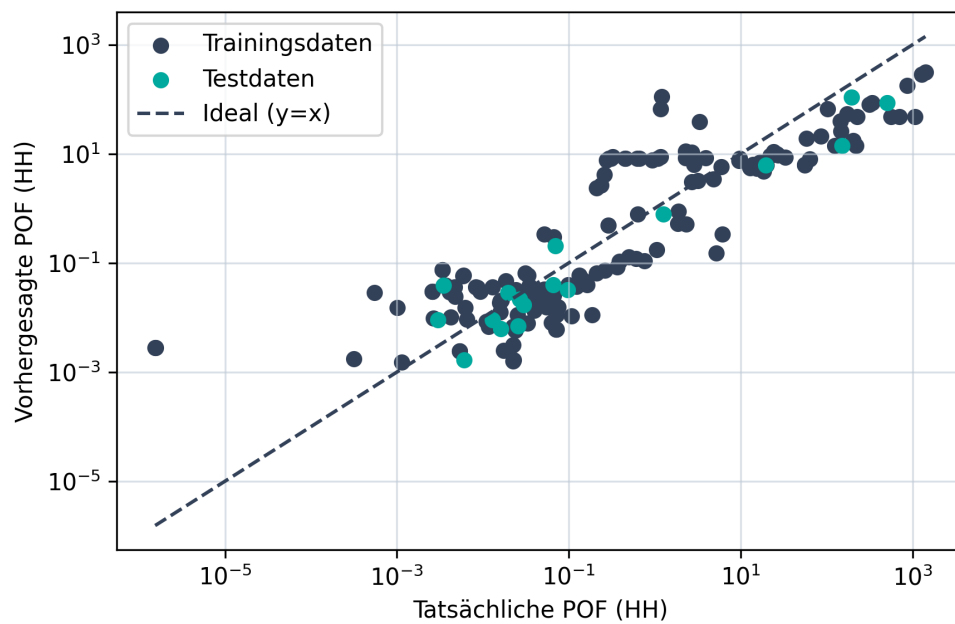


Abbildung 4.18: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Photochemical ozone formation, human health*. Beide Achsen sind logarithmisch skaliert.

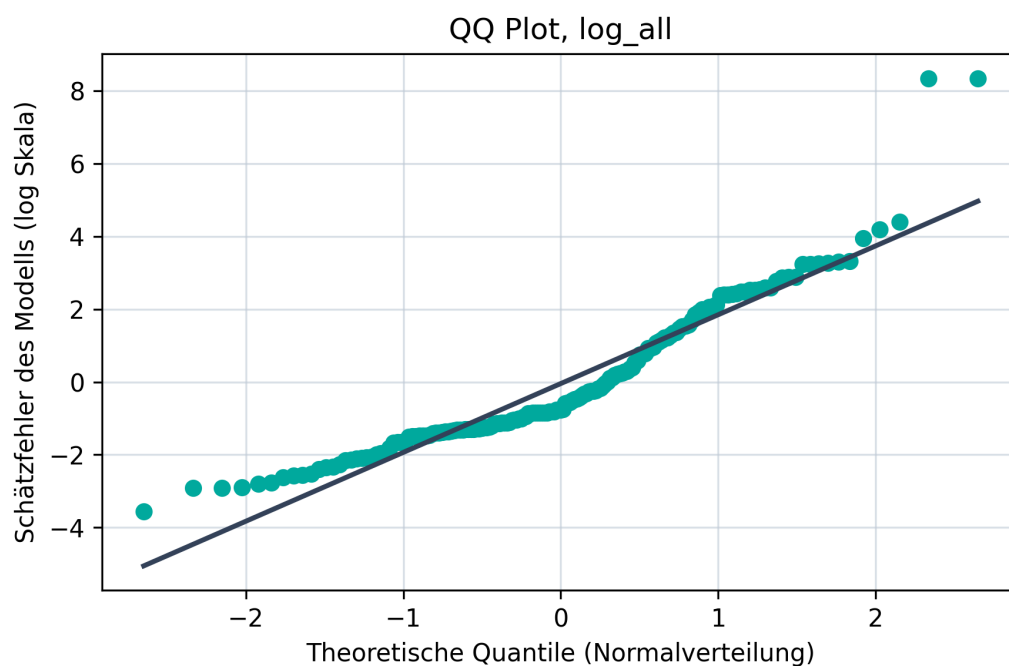


Abbildung 4.19: QQ Plot der Residuen des *Photochemical ozone formation, human health* Modells auf der Transformationsskala.

4.5.5. Regression des Indikators Resource Use, Fossils

Für den Indikator *Resource use, fossils* wird dieselbe Regressionspipeline wie zuvor verwendet. Es wurden keine Transformation, \log_{1p} und eine Box-Cox Transformation verglichen. Die beste Testgüte wird mit \log_{1p} erreicht, daher wird dieses Modell benutzt. Tabelle 4.11 fasst die Testleistung nach Rücktransformation zusammen.

Tabelle 4.11: Gütekennzahlen des linearen Regressionsmodells (*Resource use, fossils* als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0,871
$\text{RMSE}_{\text{Test}}$	1119662,28 MJ
Median absoluter Fehler	14460,95 MJ
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0,6370
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	11,1394

Das Modell erklärt damit rund 87% der Varianz auf dem Testset und erreicht eine insgesamt hohe Vorhersagegüte. Der RMSE auf Originalskala wird durch einzelne sehr große Abweichungen dominiert. Der Median des absoluten Fehlers liegt bei 14460,95 MJ und beschreibt die typische Abweichung eines repräsentativen Produkts. Der Median der relativen Fehler ($\text{MdARE} \approx 64\%$) deutet auf eine moderate typische relative Genauigkeit hin. Der deutlich größere Mittelwert der relativen Fehler ($\text{MARE} \approx 1114\%$) weist auf eine stark schiefe Fehlerverteilung mit wenigen extremen relativen Abweichungen hin, was insbesondere bei sehr kleinen Zielwerten plausibel ist.

Zur Veranschaulichung zeigt Abbildung 4.20 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von *Resource use, fossils*. Beide Achsen sind logarithmisch skaliert.

Die meisten Punkte liegen über einen breiten Wertebereich nahe der Diagonalen, so dass der generelle Trend gut erfasst wird. Im mittleren Bereich ist eine moderate Streuung sichtbar, während sehr große Werte überwiegend in der Nähe der Diagonalen liegen. Einzelne Punkte liegen deutlich oberhalb der Diagonalen, was auf wenige starke Überschätzungen hinweist und zur Schiefe der Fehlerverteilung beiträgt.

Abbildung 4.21 zeigt den QQ Plot der Residuen auf der Transformationsskala. Im Zentrum folgen die Punkte der Referenzgeraden weitgehend, in den Randbereichen

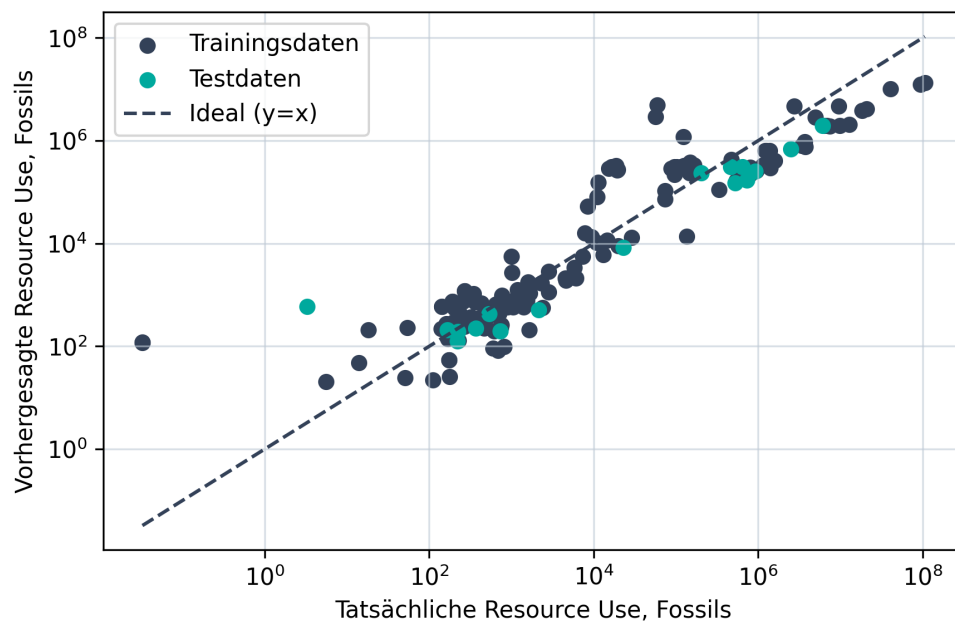


Abbildung 4.20: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Resource use, fossils*. Beide Achsen sind logarithmisch skaliert.

treten Abweichungen auf. Dies weist auf schwerere Verteilungsschwänze und einzelne extreme Residuen hin.

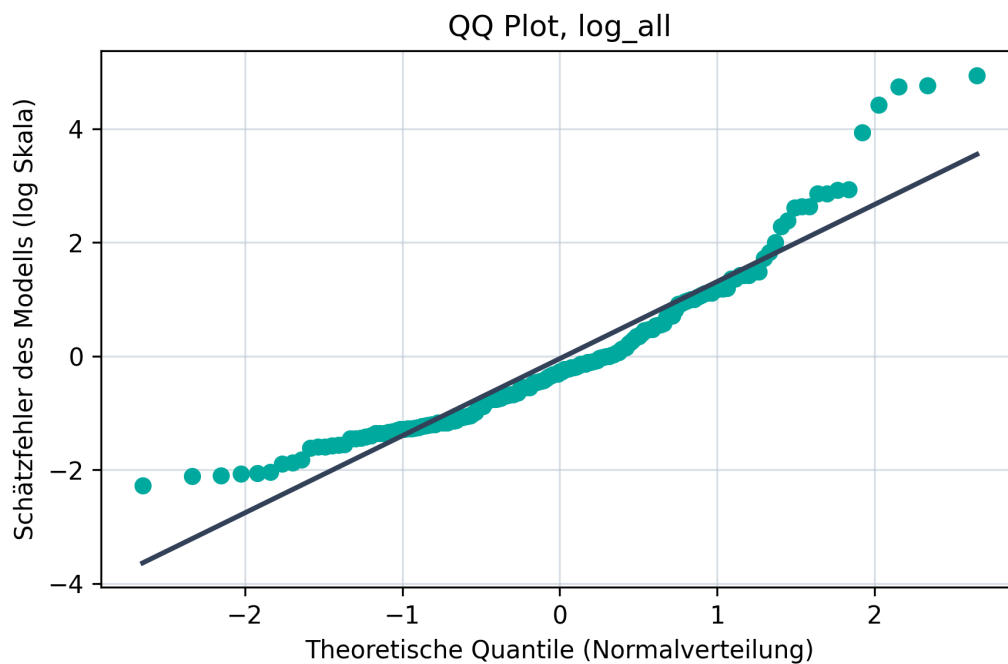


Abbildung 4.21: QQ Plot der Residuen des *Resource use, fossils* Modells auf der Transformationsskala.

4.5.6. Regression des Indikators Eutrophication (terrestrial)

Für den Indikator *Eutrophication, terrestrial* konnten $n = 107$ PEPs verwendet werden. Es wurden keine Transformation, \log_{1p} und eine Box Cox Transformation verglichen. Die beste Testgüte wird mit der Box Cox Transformation erreicht, daher wird im Folgenden dieses Modell berichtet.

Tabelle 4.12 fasst die Testleistung nach Rücktransformation zusammen.

Das Modell erklärt damit rund 79% der Varianz auf dem Testset. Der RMSE beträgt 99,33 mol N und wird durch größere Abweichungen einzelner Produkte geprägt. Der Median des absoluten Fehlers liegt bei 0,0854 mol N und beschreibt die typische Abweichung eines repräsentativen Produkts. Der Median der relativen Fehler ($MdARE \approx 56\%$) deutet auf eine moderate typische relative Genauigkeit hin. Der höhere Mittelwert der relativen Fehler ($MARE \approx 117\%$) weist auf eine schiefe Fehlerverteilung mit einigen deutlich schlechter getroffenen Beobachtungen hin.

Zur Veranschaulichung zeigt Abbildung 4.22 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten. Beide Achsen sind logarithmisch skaliert.

Die meisten Punkte liegen nahe der Diagonalen, insbesondere im mittleren Wertebereich. Bei sehr kleinen tatsächlichen Werten ist die Streuung größer, was zu höheren relativen Fehlern führen kann. Einige größere Werte liegen unterhalb der Diagonalen, was auf eine leichte Unterschätzung im oberen Bereich hindeutet.

Tabelle 4.12: Gütekennzahlen des linearen Regressionsmodells (Eutrophication, `terrestrial` als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0,793
$\text{RMSE}_{\text{Test}}$	99,33 mol N
Median absoluter Fehler	0,0854 mol N
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0,5611
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	1,1653

Abbildung 4.23 zeigt den QQ Plot der Residuen auf der Transformationsskala. Im Zentrum folgen die Punkte der Referenzgeraden weitgehend, in den Randbereichen sind Abweichungen sichtbar. Dies weist auf schwerere Verteilungsschwänze und einzelne extreme Residuen hin.

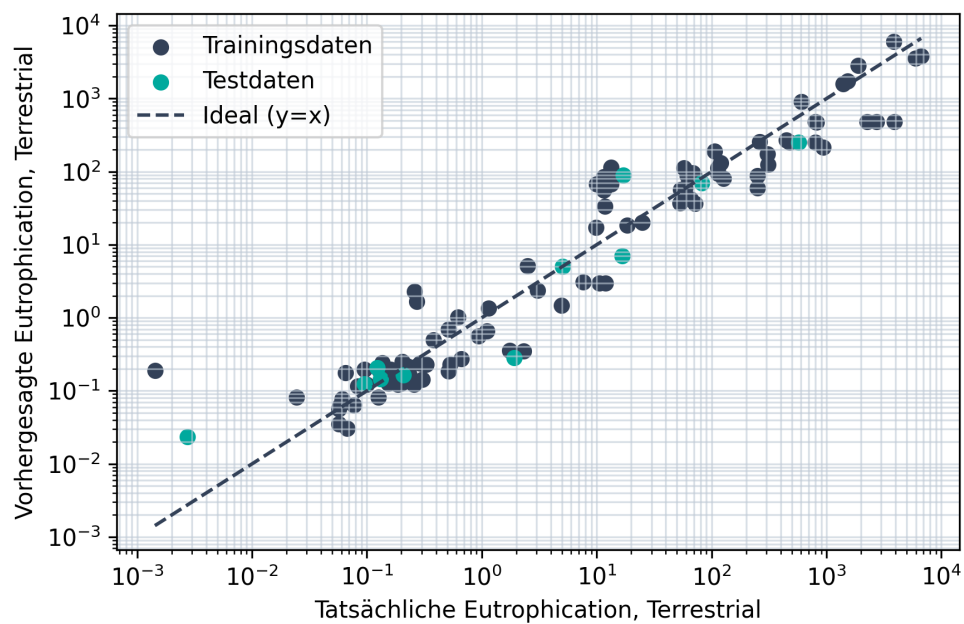


Abbildung 4.22: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Eutrophication, terrestrial*. Beide Achsen sind logarithmisch skaliert.

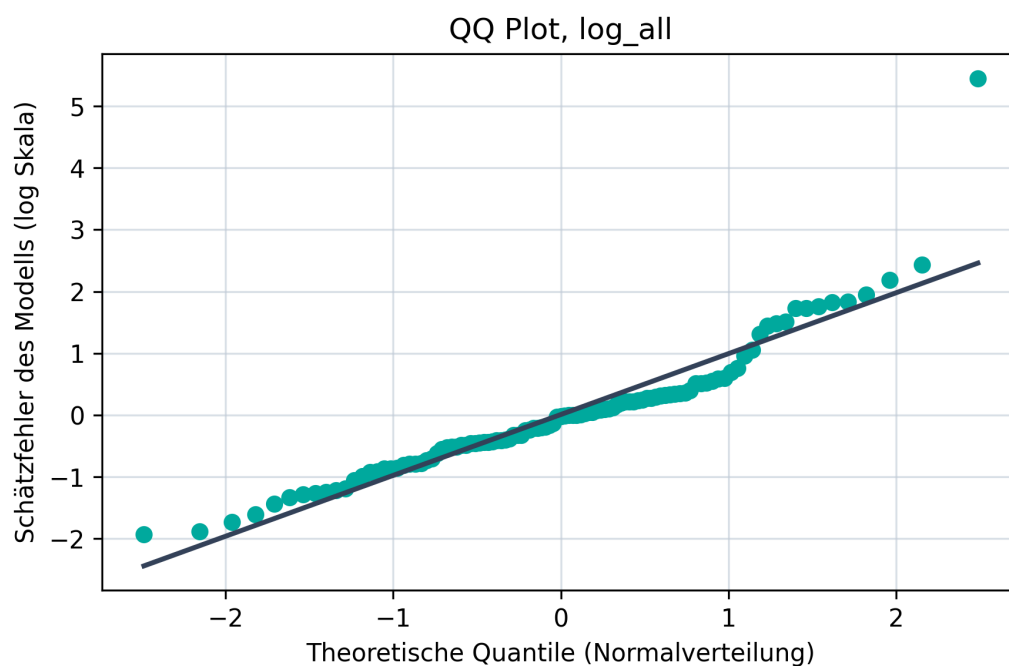


Abbildung 4.23: QQ Plot der Residuen des *Eutrophication, terrestrial* Modells auf der Transformationskala.

4.5.7. Regression des Indikators Ozone Depletion

Das lineare Regressionsmodell wird analog zu den vorherigen Indikatoren auf `ozone_depletion` angewendet. Es wurden keine Transformation, `log1p` und eine Box-Cox Transformation verglichen. Die beste Testgüte ergibt sich mit der Box-Cox Transformation. Daher wird im Folgenden dieses Modell berichtet. Es konnten $n = 170$ PEPs benutzt werden. Tabelle 4.13 fasst die Testleistung nach Rücktransformation zusammen.

Tabelle 4.13: Gütekennzahlen des linearen Regressionsmodells (`ozone_depletion` als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.858
$\text{RMSE}_{\text{Test}}$	0.0029
Median absoluter Fehler	0.0000
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0.8635
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	1.0411

Das Modell erklärt auf dem Testset rund 86% der Varianz. Der RMSE auf Originalskala ist klein, was zur sehr kleinen Größenordnung des Indikators passt. Der Median des absoluten Fehlers wird in der Ausgabe auf vier Dezimalstellen gerundet, daher erscheint er als 0.0000. Die relativen Fehlermaße zeigen dennoch eine deutlich heterogene Fehlerstruktur. Ein typisches Produkt weist einen relativen Fehler von etwa $\text{MdARE} \approx 0.86$ auf. Der größere Mittelwert $\text{MARE} \approx 1.04$ deutet auf einzelne Beobachtungen mit sehr großen relativen Abweichungen hin. Dies ist bei `ozone_depletion` plausibel, da bei sehr kleinen Zielwerten bereits kleine absolute Fehler zu großen relativen Fehlern führen.

Zur Veranschaulichung zeigt Abbildung 4.24 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von `ozone_depletion` auf logarithmischen Achsen.

Das Streudiagramm zeigt eine grobe Ausrichtung entlang der Diagonalen. Bei großen Zielwerten liegen viele Punkte unterhalb der Referenzlinie, was auf eine Tendenz zur Unterschätzung im oberen Wertebereich hinweist. Gleichzeitig existieren einzelne Fälle mit starker Überschätzung bei sehr kleinen Zielwerten, die auf der Log Skala als große vertikale Abweichungen erscheinen und die relativen Fehler stark erhöhen.

Abbildung 4.25 zeigt den QQ Plot der Residuen auf der Transformationsskala. Im

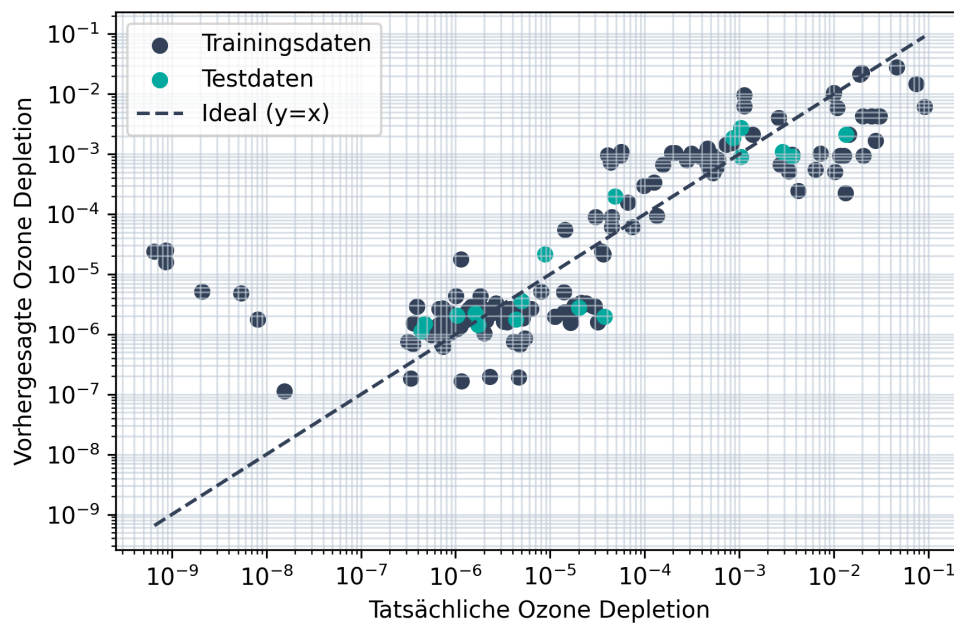


Abbildung 4.24: Vorhergesagte gegenüber tatsächlichen Werten von Ozon depletion. Beide Achsen sind logarithmisch skaliert.

Zentrum liegen die Punkte näher an der Referenzgeraden, während die oberen Quantile deutlich nach oben abweichen. Dies weist auf eine rechtsschiefe Fehlerverteilung mit wenigen sehr großen positiven Fehlern hin, konsistent mit den extremen Überschätzungen bei kleinen Zielwerten.

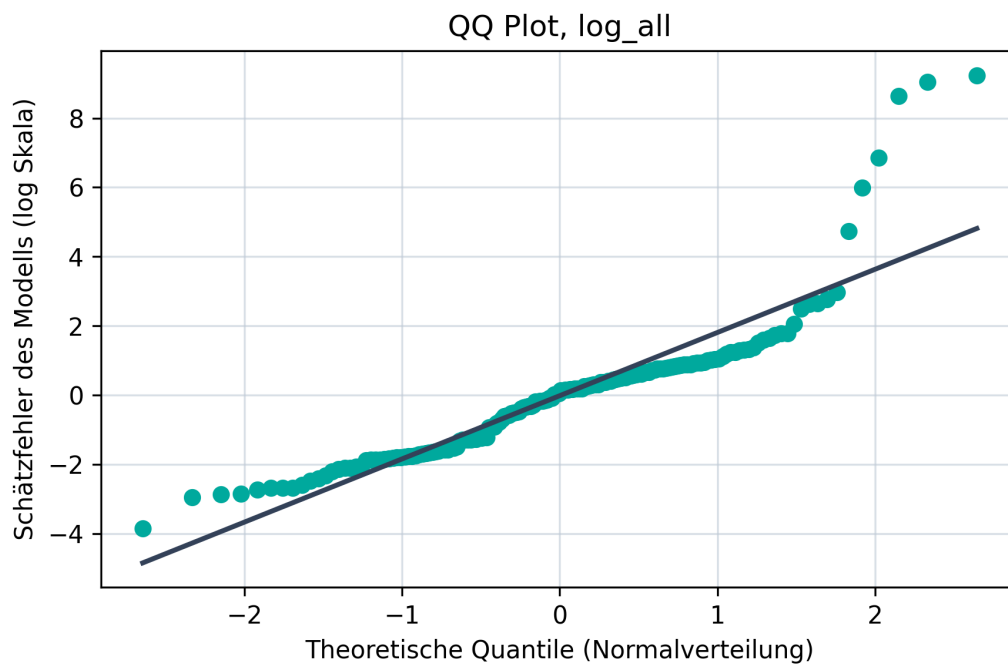


Abbildung 4.25: QQ Plot der Schätzfehler des Ozone depletion Modells auf der Transformationsskala.

4.5.8. Regression des Indikators Resource Use, Minerals and Metals

Für den Indikator *Resource use, minerals and metals* wurde das lineare Regressionsmodell mit Gewicht, Stromverbrauch und Material-PCs angewendet. Es wurden keine Transformation, \log_{1p} und Box-Cox verglichen. Die beste Testgüte wurde mit Box-Cox erreicht, daher wird dieses Modell im Folgenden berichtet.

Tabelle 4.14 fasst die Testleistung nach Rücktransformation auf die Originalskala zusammen.

Das Modell erreicht eine hohe Testgüte ($R^2 \approx 0,87$) und bildet den Indikator insgesamt gut ab. Der Median der relativen Fehler zeigt, dass ein typisches Produkt oft in der richtigen Größenordnung getroffen wird ($\text{MdARE} \approx 0,78$). Der größere Mittelwert der relativen Fehler ($\text{MARE} \approx 1,00$) weist auf einzelne Fälle mit deutlich größerer relativer Abweichung hin, bleibt aber insgesamt in einem moderaten Rahmen.

Abbildung 4.26 zeigt die vorhergesagten gegenüber den tatsächlichen Werten auf logarithmischen Achsen.

Der Hauptteil der Punkte liegt nahe der Diagonalen, während einzelne Beobachtungen deutlich abweichen. Das passt zur Kombination aus gutem R^2 und einer Fehlerverteilung mit Ausreißern.

Abbildung 4.27 zeigt einen QQ Plot der Residuen auf der Transformationsskala. Im Zentrum folgen die Residuen näherungsweise der Normalverteilung, während die

Tabelle 4.14: Gütekennzahlen des linearen Regressionsmodells (Resource use, minerals and metals als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.866
$\text{RMSE}_{\text{Test}}$ (Originalskala)	0.7218
Median absoluter Fehler	0.0010
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0.7798
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	1.0003

äußeren Quantile schwere Schwänze zeigen.

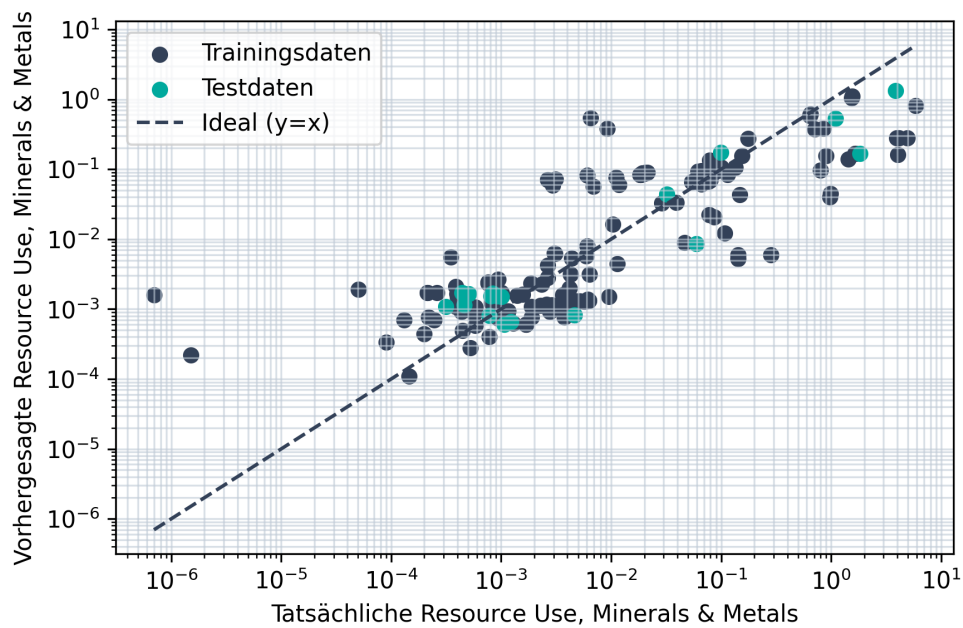


Abbildung 4.26: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Resource use, minerals and metals*. Beide Achsen sind logarithmisch skaliert.

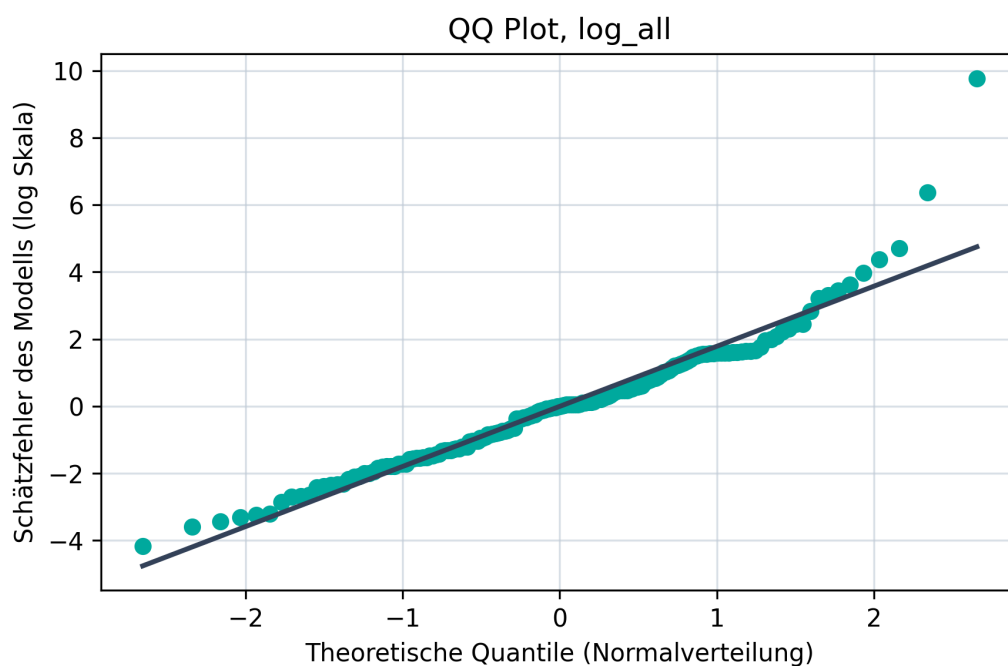


Abbildung 4.27: QQ Plot der Residuen des Modells für *Resource use, minerals and metals* auf der Transformationsskala.

4.5.9. Indikatoren mit geringer Modellgüte

Neben den oben beschriebenen 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, so dass hier nicht von einem zuverlässigen Vorhersagemodell gesprochen werden kann. Tabelle 4.15 fasst diese Indikatoren zusammen.

Für diese schwächer erklärbaren Zielgrößen wurden ebenfalls verschiedene Transformationen der Zielvariable verglichen. Box-Cox kann die Testgüte für alle restlichen Indikatoren leicht verbessern, die Zugewinne bleiben jedoch insgesamt begrenzt und reichen nicht aus, um die Indikatoren in den Bereich stabiler Vorhersagegüte zu überführen.

!TODO: Die Tabelle passt Layout technisch noch nicht!

Tabelle 4.15: Indikatoren und Gütemaße mit geringer Modellgüte ($R^2_{\text{Test, mean}} < 0,5$).

Indikator	R^2_{Test}	RMSE _{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

Als besonders kritisches Beispiel zeigt Abbildung 4.28 den Indikator *Radioactive Waste Disposed*. Obwohl die absoluten Fehler aufgrund der sehr kleinen Werte gering bleiben, liegen viele Punkte weit von der Diagonalen entfernt und die Streuung ist hoch. Entsprechend ist das durchschnittliche Test- R^2 klein. Eine Box-Cox Transformation kann die Verteilung der Fehler nicht glätten.

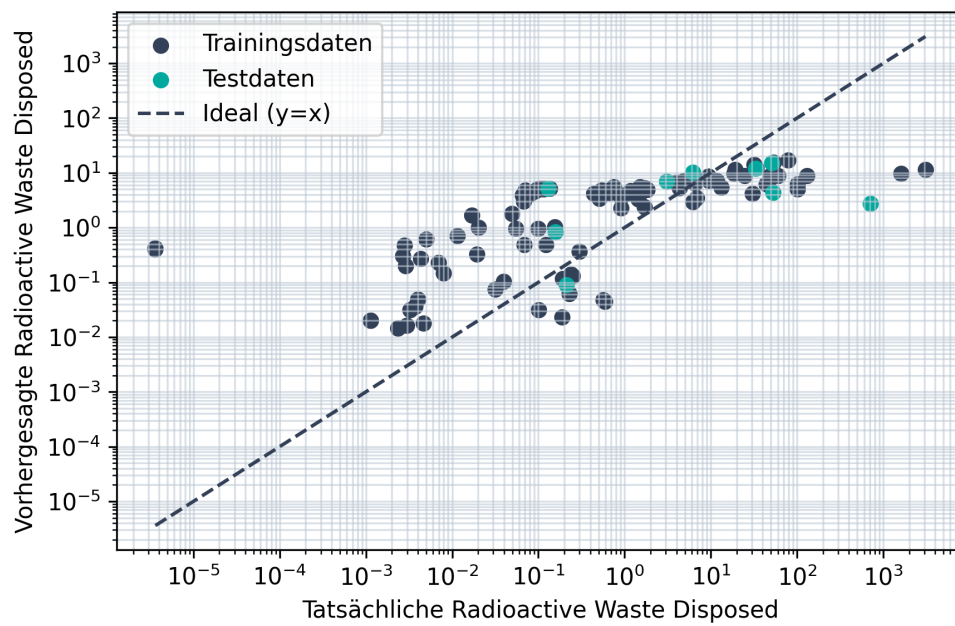


Abbildung 4.28: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Radioactive Waste Disposed*.

Ähnliche Muster zeigen sich bei den weiteren in Tabelle 4.15 aufgezählten Indikatoren.

5

Diskussion

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 der benutzten Materialien erklären. Dieses Kapitel ordnet die Ergebnisse ein und diskutiert Unsicherheiten und Grenzen.

5.1. Einordnung der Ergebnisse

Die höchste und stabilste Modellgüte wird für den Indikator *Climate change (total)* erreicht. Insgesamt kann die CO₂ Regression als obere Schranke der erreichbaren Genauigkeit unter den gegebenen Featureeinschränkungen interpretiert werden.

Das gewählte Feature-Set kann etwas 90% der Varianz der CO₂-Werte erklären, was eine hohe Modellgüte beschreibt. Da die geschätzten Werte im Mittel trotzdem um 120% abweichen, ist auch dieses Modell nicht präzise. Für die anderen Indikatoren wurde die Pipeline im Wesentlichen nur über die Transformation der Zielvariablen angepasst. Eine ähnlich ausführliche explorative Entwicklung wie für CO₂ fand dort nicht statt.

Trotzdem lässt sich das Modell auf mehrere weitere Indikatoren wie *Acidification*, *Resource use (fossils)* oder *Water use* übertragen und liefert brauchbare Ergebnisse, die Streuung ist jedoch größer. Für Indikatoren mit geringer Modellgüte (vgl. Tab. 4.15) gelingt es dagegen nur, einen begrenzten Anteil der Varianz zu erklären. Dies deutet darauf hin, dass die verwendeten Eingangsgrößen entweder nur einen schwachen Einfluss haben oder dass die Datenlage für diese Zielgrößen stark verrauscht ist.

5.2. Fehlerstruktur und Modellannahmen

Ein zentrales Diagnoseergebnis ist, dass die Schätzfehler der Modelle nicht normalverteilt sind. Dies zeigt sich in den QQ Plots durch systematische Abweichungen von der Referenzgeraden, insbesondere in den äußeren Quantilen. Die Residuen weisen damit schwere Verteilungsschwänze und Ausreißer auf.

Für die Interpretation bedeutet dies Folgendes. Klassische Schlussfolgerungen der OLS Theorie, etwa Standardfehler und p Werte sind nur eingeschränkt zuverlässig. Zudem ist die Fehlerverteilung trotzdem plausibel, da die Indikatoren über mehrere Größenordnungen streuen und einzelne Produkte sehr große Umweltwirkungen aufweisen. Eine Transformation wie \log_{10} oder Box Cox die Fehlerstruktur typischerweise, sie führt aber nicht zu perfekter Normalität, weil Datenheterogenität und Ausreißer weiterhin bestehen.

Für die Zielsetzung dieser Arbeit, nämlich robuste Vorhersagen für neue Produkte mit einem simplen Modell, ist daher eine testbasierte Bewertung zentral. Ergänzend zu RMSE werden robuste Kennzahlen wie Median absoluter Fehler und Median relativer Fehler berichtet. Dadurch wird die Modellleistung weniger durch Extremwerte verzerrt und über verschiedene Größenordnungen hinweg besser interpretierbar.

5.3. Grenzen des Modells

Die Aussagekraft der CO₂-Regression wird wesentlich durch die Qualität und Homogenität der zugrunde liegenden PEP-Daten bestimmt. Die PEPs selbst beruhen auf Hintergrunddatensätzen, Annahmen, unterschiedlichen Berechnungsmodellen und Systemgrenzen. Damit können sie keine ultimative Wahrheitsquelle darstellen. Das Modell lernt damit nicht nur physikalische Zusammenhänge zwischen Gewicht, Stromverbrauch, Materialmix und Emissionen, sondern auch diese Heterogenität mit. Ein Teil der beobachteten Streuung ist daher als Unsicherheit zu verstehen und dürfte selbst mit komplexeren Modellen kaum vollständig eliminierbar sein.

Hinzu kommt, dass die Stichprobe zwar 173 PEPs umfasst, diese aber unterschiedliche Produktkategorien abdecken. Positiv ist, dass das Modell der Zielsetzung entsprechend geräteübergreifende die Muster und Trends erkennen kann und damit nicht nur für einen einzelnen Gerätetyp gültig ist. Gleichzeitig bedeutet die Produktvielfalt, dass Produkte ohne vergleichbare Vertreter in der Stichprobe mit deutlich größeren Prognosefehlern gerechnet werden muss. Das Modell ist somit vor allem als Näherung für typische Produkte innerhalb des betrachteten Datenraums zu verstehen.

6

Fazit

6.1. Zusammenfassung des methodischen Vorgehens

Diese Arbeit untersucht, inwieweit Umweltindikatoren aus PEP-Ecopassport- Dokumenten mit wenigen, allgemein verfügbaren Produktmerkmalen geschätzt werden können, um auch für Produkte ohne PEP eine belastbare Größenordnungsabschätzung zu ermöglichen. Im Fokus steht damit nicht die exakte Rekonstruktion einer vollständigen Ökobilanz, sondern ein kompakter, praxistauglicher Ansatz zur Vorhersage zentraler Umweltwirkungen.

Als Grundlage wurde eine automatisierte Pipeline aufgebaut, die heterogene PEP-PDFs in ein einheitliches, maschinenlesbares Datenformat überführt. Die Pipeline umfasst die Recherche und Erfassung relevanter PEPs, das PDF-Parsing und die strukturierte Extraktion von Produkt- und Umweltinformationen sowie die Normalisierung und Validierung der resultierenden Daten. Dadurch entsteht eine Datenbasis, die eine systematische statistische Auswertung der Indikatoren und die anschließende Modellierung ermöglicht.

Für die statistische Auswertung werden ausschließlich Datensätze berücksichtigt, bei denen die jeweiligen Zielindikatoren sowie zentrale Produktmerkmale (z. B. Gesamtgewicht und Stromverbrauch) plausibel und numerisch verfügbar sind. Bereits die deskriptive Analyse zeigt, dass sämtliche Umweltindikatoren stark rechtsschief verteilt sind und mehrere Größenordnungen abdecken. Einzelne Produkte weisen sehr hohe Werte auf und prägen Mittelwerte und RMSE-basierte Kennzahlen überproportional.

Um diese Skaleneffekte zu reduzieren und die Modellierung zu stabilisieren, wurden für alle Indikatoren systematisch verschiedene Transformationen der Zielvariablen ver-

glichen. Je nach Indikator wurden keine Transformation, \log_{1p} sowie Box-Cox-Transformationen getestet. In den meisten Fällen führte eine logarithmische Transformation zu der besten oder zumindest stabilsten Testgüte, während Box-Cox die Ergebnisse teilweise noch leicht verbessern konnte. Die endgültig berichteten Fehlermaße werden nach Möglichkeit nach Rücktransformation auf der Originalskala angegeben, damit sie in der jeweiligen Einheit interpretierbar bleiben.

Als Modellansatz wurde eine lineare Regression mit bewusst kompaktem Feature-Set gewählt, um die Zielsetzung der Arbeit zu verfolgen. Als erklärende Variablen dienen das log-transformierte Produktgewicht, der log-transformierte, über die Lebensdauer aggregierte Stromverbrauch sowie verdichtete Materialinformationen, die über eine PCA aus dem Materialblock abgeleitet werden. Dadurch bleibt das Modell auch bei vielen Materialspalten handhabbar und nutzt wenige, interpretierbare Hauptkomponenten statt viele einzelne Materialinformationen.

Die Modellgüte wird mit einer streng getrennten Train/Test-Auswertung beurteilt. Dazu werden die Daten einmalig in Trainings- und Testdaten aufgeteilt und das Modell ausschließlich auf den Trainingsdaten angepasst. Hyperparameter, etwa bei Ridge oder Lasso, werden über wiederholte innere Train/Valid-Splits im Trainingsdatensatz bestimmt. Die abschließende Bewertung erfolgt genau einmal auf dem unabhängigen Testset. Als Kennzahlen werden neben R^2 und RMSE auch robuste Fehlermaße berichtet, insbesondere der Median des absoluten Fehlers sowie Median und Mittelwert der relativen absoluten Fehler, um die typische Modellleistung trotz Ausreißern nachvollziehbar zu charakterisieren.

6.2. Einordnung der Ergebnisse

Die Ergebnisse der Regressionmodelle zeigen, dass sie als *heuristische Abschätzung* zu verstehen sind und nicht als Ersatz für eine vollständige Berechnung. Auch bei hoher erklärter Varianz können einzelne Produkte deutlich verfehlt werden, insbesondere in Bereichen mit sehr kleinen oder sehr großen Zielwerten und bei Nischenprodukten. Die Ergebnisse liefern damit keine belastbare Grundlage, um PEP Werte präzise zu replizieren oder einzelne Produkte auf Dezimalstellen genau zu vergleichen.

Der Mehrwert liegt vielmehr in der schnellen, datengetriebenen Einordnung von Größenordnungen und in der Priorisierung. Das Modell kann beispielsweise genutzt werden, um frühe Design und Beschaffungsentscheidungen grob zu unterstützen, Produkte oder Varianten hinsichtlich erwarteter Umweltwirkungen vorzusortieren, sowie Auffälligkeiten in PEP Angaben zu identifizieren, wenn ein ausgewiesener Wert stark

von der erwarteten Größenordnung abweicht. Damit passt der Ansatz besonders zu Anwendungen, in denen nur wenige Merkmale verfügbar sind, aber trotzdem eine erste Orientierung benötigt wird.

6.3. Ausblick und zukünftiger Forschungsbedarf

Ein zentraler Entwurfsgedanke war die Beschränkung auf Merkmale, die auch ohne PEP typischerweise messbar oder zumindest plausibel abschätzbar sind. Es wäre allerdings prinzipiell möglich, zusätzliche erklärende Variablen in das Modell aufzunehmen, etwa die verwendete Berechnungsmethode und die verwendete Datenbank. Solche Informationen könnten die Genauigkeit erhöhen, würden jedoch die Einfachheit und Übertragbarkeit des Ansatzes verringern, da sie für neue Produkte häufig nicht verlässlich oder nicht in standardisierter Form vorliegen. Damit stünde eine solche Erweiterung im Spannungsfeld zur Zielsetzung, mit einem bewusst kompakten Feature Set robuste, breit nutzbare Abschätzungen zu ermöglichen.

Eine naheliegende Erweiterung, die diese Zielsetzung weitgehend erhält, ist die Modellierung getrennt nach Produktkategorien. Eine Kategorisierung lässt sich meist aus Metadaten oder Produktbeschreibungen ableiten und führt gleichzeitig zu homogeneren Teilmengen, in denen sich die Beziehungen zwischen Gewicht, Stromverbrauch, Materialmix und Indikatoren stabiler verhalten dürften. Dadurch könnte die Vorhersagegüte insbesondere für Indikatoren mit mittlerer Stabilität deutlich steigen, ohne dass das Modellkonzept wesentlich komplizierter wird und damit die Zielsetzung verloren geht.

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] Andreiuid 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: [10.560/islandora:1012018](https://doi.org/10.560/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-ST5778](https://doi.org/10.1214/20-ST5778). URL: <https://projecteuclid.org/journals/statistical-science/volume-36/issue-2/The-BoxCox-Transformation-Review-and-Extensions/10.1214/20-ST5778.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-gebaudetechnologien_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/en-ie/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://arxiv.org/abs/2408.09869). URL: <https://arxiv.org/abs/2408.09869>.
- [Nad+24] Rohaan Nadeem u. a. "Extraction of User-Defined Information from PDF". In: (2024). DOI: [10.1109/DASA63652.2024.10836169](https://ieeexplore.ieee.org/document/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://arxiv.org/abs/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://arxiv.org/abs/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 KNI-ME: An Application in Toxicology". In: (2025). DOI: [10.1002/wcms.70047](https://wires.onlinelibrary.wiley.com/doi/full/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>.