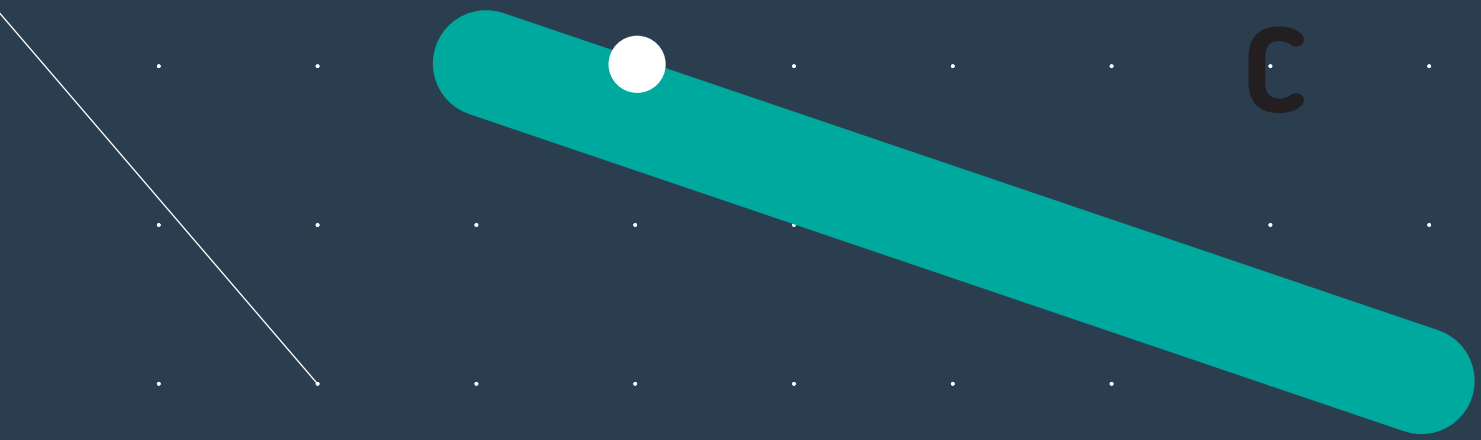
Hochschule Konstanz
Department of Computer Science

Submitted by
Jonas Mayer
Student Number 305630

jonas.mayer@htwg-konstanz.de



Bachelor Thesis

Datenanalyse PEP

Konstanz, TODO

Hochschule Konstanz
Department of Computer Science

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

1

Einleitung

Hier Einleitung schreiben
TESTSTETS

2

Theoretische Grundlagen

! TODO: Kapitel kürzen, nur das was ich nachher verwende, weniger komplex !

Hier werden die theoretischen Grundlagen für die vorliegende Arbeit gelegt. Ausgangspunkt ist die Beobachtung, dass Smart-Home-/IoT-Produkte Auswirkungen auf die Umwelt in ihrer Nutzungsphase, Fertigung, Distribution und Entsorgung haben. Für die standardisierte Berichterstattung solcher Wirkungen 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 einer praxistauglichen Heuristik für Produkte ohne PEP eingesetzt werden.

2.1. PEP-Ecopassport

Die einzige 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 klar von unternehmensbezogenen Treibhausgas-Bilanzierungen: Es erfasst ausschließlich produktspezifische Umweltwirkungen entlang des Lebenszyklus und folgt dabei den methodischen Vorgaben der ISO 14040-Reihe. Für umfassende *GHG-Assessments* 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.

! TODO: Einheitliche Bezeichnungen, nur die für später relevanten (?) !

Tabelle 2.1: Umweltindikatoren

Indikator	Beschreibung
Acidification	Versauerung von Böden und Gewässern durch säurebildende Emissionen
Climate Change (Fossil)	Treibhauspotenzial durch fossile CO_2-Emissionen
Climate Change (Land Use and Land Use Change)	Treibhauspotenzial infolge von Landnutzungsänderungen (LULUC)
Climate Change (Total)	Gesamtes Treibhauspotenzial aus allen Quellen
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. Das Portable Document Format (PDF)

Das *Portable Document Format (PDF)* ist eines der beliebtesten elektronischen Dokumentenformate und ist primär ein *layoutbasiertes Format*. Es wurde entwickelt, um das Erscheinungsbild der Originaldokumente plattform- und anwendungsübergreifend zu bewahren. [LB95] Das Format beschreibt Objekte auf einer niedrigen Strukturebene und legt die *Positionen und Schriftarten der einzelnen Zeichen* fest, aus denen der sichtbare Text zusammengesetzt ist. Zu den beschriebenen Objekten gehören:

- Gruppen von Zeichen (Textobjekte)
- Linien, Kurven und Bilder
- Stilattribute wie Schriftart, Farbe, Strichführung, Füllung und geometrische Formen.

[BK17]

Obwohl PDF die visuelle Darstellung eines Dokuments zuverlässig bewahrt, fehlt den meisten Dateien eine explizite logische Struktur auf höherer Ebene. Die folgenden semantischen Einheiten sind im Format nicht direkt enthalten und werden nur durch die oben genannte niedrige Strukturebene zusammengesetzt:

- logische Komponenten wie Wörter, Textzeilen, Absätze, Tabellen oder Abbildungen [CF04]
- Informationen über die *semantischen Rollen* des Textes (z. B. Haupttext, Fußnote oder Bildunterschrift), [BK17]
- eine eindeutige Lese- und Wortreihenfolge, insbesondere bei mehrspaltigen Layouts oder eingebetteten Elementen. [BK17]

Hinzuzufügen ist, dass PDF-Dokumente mit semantischen Informationen *getaggt* werden können. In der Praxis sind diese zusätzlichen Informationen selten gegeben. [BK17] Die für diese Arbeit relevanten PEP-Ecopassport-PDFs sind alle nicht getaggt.

Das Fehlen dieser semantischen Informationen erschwert die Erkennung, Wiederverwendung oder Bearbeitung des Layouts und Inhalts erheblich. [CZ17] Die automatische Extraktion dieser Metadaten und Textinhalte ist daher eine zentrale, aber fehleranfällige Aufgabe, da es keine allgemein verbindlichen Standards für die Strukturierung solcher Informationen in PDF-Dokumenten gibt. [Lip+13]

2.2.2. Herausforerungen bei der automatisierten Extraktion

Die Rekonstruktion des Textflusses und der semantischen Einheiten aus den Positionen einzelner Zeichen ist komplex:

1. Wortidentifikation Die korrekte Bestimmung von Wortgrenzen ist nicht trivial:

- *Abstände:* Die Abstände zwischen Zeichen können innerhalb einer Zeile variieren, sodass keine feste Regel existiert, um Wortgrenzen ausschließlich anhand der Zeichenpositionen zu bestimmen. [BK17]
- *Silbentrennung:* In mehrspaltigen Layouts getrennte Wörter müssen korrekt wieder zusammengeführt werden. [BK17]
- *Ligaturen:* Zeichenkombinationen wie „fl“ oder „fi“ werden im PDF oft als einzelnes Zeichen gespeichert und müssen beim Extrahieren in mehrere Zeichen aufgelöst werden. [Lip+13]
- *Diakritische Zeichen:* Buchstaben mit Diakritika (z. B. à, ä, ã) können als zwei separate Zeichen gespeichert sein und müssen beim Parsing zu einem Zeichen zusammengeführt werden. [BK17]

2. Lesereihenfolge (Reading Order) Die korrekte Lesereihenfolge ist entscheidend für die Verständlichkeit des Textes und der weiterführenden Interpretation. [BK17] In mehrspaltigen Layouts sind Textzeilen im PDF häufig in einer verschränkten Reihenfolge gespeichert. Ohne Korrekturmechanismen führt dies zu unleserlichem, inhaltlich falsch zusammengesetztem Text.[LB95]

3. Absatzgrenzen (Paragraph Boundaries) Die Erkennung von Absatzanfängen und -enden ist besonders schwierig:

- *Unterbrechungen:* Text, der zu einem Absatz gehört, kann durch Formeln, Tabellen oder Abbildungen unterbrochen und später auf derselben Seite fortgesetzt werden.

- *Seitenumbrüche:* Absätze können am Seiten- oder Spaltenende abgeschnitten und auf der folgenden Seite fortgeführt werden, ohne dass dies im PDF strukturell kenntlich gemacht wird.

[BK17]

4. Technologische und Layout-Herausforderungen

- *Überlagerungen (Overlays):* In grafisch komplexen Dokumenten können Text- und Bildelemente überlappen, etwa wenn Beschriftungen in Abbildungen eingebettet sind. Dies erschwert die korrekte Segmentierung. [CF04]

- *Segmentierungsfehler:* Bei Tabellen, Karten oder Diagrammen kann Text aus unterschiedlichen logischen Einheiten fälschlicherweise in dieselbe Gruppe aggregiert werden. [CF04]

- *Type-3-Fonts:* Manche Zeichen (insbesondere Ligaturen und Sonderzeichen) werden im PDF nicht als Textobjekte, sondern als Vektorgrafiken gespeichert. Solche Elemente sind mit herkömmlicher Textextraktion nicht identifizierbar und erfordern erweiterte, teils OCR- oder ML-basierte Verfahren. [BK17]

2.2.3. Extraktionsansätze

Da diese Probleme weit verbreitet und bekannt sind, gibt es mehrere Extraktionsansätze, um PDF-Dateien in ein strukturierts 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. Sie lassen sich schnell einsetzen, da sie ohne domänenspezifische Daten auskommen. [AA25] Mehrere weit verbereutete Softwarebibliotheken implementieren regelbasierte Parser, darunter *pdfplumber*, *pypdfium2* und *pypdf*. Da die Arbeit von Selg auf *pdfplumber* aufbaut, wird diese Bibliothek näher besprochen. [Sel25]

Sie ist vollständig in Python implementiert und baut auf der weit verbreiteten Bibliothek *pdfminer.six* auf. Das Werkzeug wurde speziell für die Text- und Tabellenextraktion aus PDF-Dokumenten entwickelt und gilt als eine der benutzerfreundlichsten Lösungen im Python-Ökosystem. [AA25]

pdfplumber ist ein Python-Werkzeug, das PDF-Dateien so einliest, dass Text und einfache grafische Elemente jeder Seite als Python-Objekte vorliegen. Jede Seite wird

dabei als Sammlung von Textfragmenten, Linien, Rechtecken und Bildern mit ihren Positionen beschrieben. [YCZ25] Für die Tabellenerkennung nutzt *pdfplumber* vor allem sichtbare horizontale und vertikale Linien als potenzielle Zellgrenzen. Über Optionen lässt sich diese Heuristik an unterschiedliche Layouts anpassen. [YCZ25]

Der regelbasierte Ansatz von *pdfplumber* führt bei klar strukturierten, editierbaren PDF-Dokumenten 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*). [AA25]

Die Grenzen des Werkzeugs zeigen sich jedoch vor allem bei komplexen oder unregelmäßig formatierten PDF-Dateien. Insbesondere wissenschaftliche Dokumente mit mathematischen Ausdrücken, eingebetteten Formeln oder 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. Auch bei Patenten oder Dokumenten mit grafischen Strukturen (z. B. chemischen Formeln oder Bauzeichnungen) stößt der regelbasierte Ansatz an seine Grenzen. [AA25]

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 erkennt *pdfplumber* nicht zuverlässig.

Damit steht *pdfplumber* exemplarisch für klassische Extraktionstools, die ohne maschinelles Lernen oder tiefere Dokumentenverständnis-Modelle arbeiten und deshalb bei komplexen, visuell strukturierten Dokumenten wie den PEP-Ecopassports an ihre methodischen Grenzen stoßen.

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

TODO: GROBID Doku suchen

Für diese Arbeit wird *Docling* näher betrachtet, da es ein vollständig offenes, lokal ausführbares Toolkit darstellt, das eine vollständige End-to-End-Pipeline für Layout-Analyse, Strukturerkennung und Tabellensegmentierung bereitstellt. Es integriert mo-

derne Deep-Learning-Modelle, berücksichtigt gleichzeitig geometrische Informationen und und lässt sich einfach in eine Python Pipeline einbauen. [Aue+24]

Es wurde mit dem Ziel entwickelt, PDF-Dokumente und andere Formate in eine maschinell verarbeitbare, strukturierte Repräsentation zu überführen. Im Gegensatz zu Werkzeugen wie *pdfplumber*, die auf geometrischen Heuristiken basieren, kombiniert *Docling* klassische Parsing-Verfahren mit tiefen neuronalen Modellen für Layout- und Strukturerkennung. [Aue+24]

Technisch basiert *Docling* auf einer linearen Verarbeitungs-Pipeline, die mehrere spezialisierte Komponenten kombiniert. Nach dem initialen Parsen durch ein PDF-Backend (z. B. *qpdf* oder *pypdfium*) werden für jede Seite Bitmap-Abbilder erzeugt, auf denen KI-Modelle für Layout- und Strukturerkennung ausgeführt werden. [Aue+24] Das zugrunde liegende Layout-Analysemodell *DocLayNet* identifiziert auf Basis eines trainierten Objektdetektors verschiedene Seitenelemente und deren Begrenzungsrahmen als Absätze, Überschriften, Listen, Abbildungen oder Tabellen. [Aue+24] Diese visuellen Einheiten werden mit den extrahierten Text-Tokens verknüpft und zu konsistenten Dokumentstrukturen zusammengeführt. Für erkannte Tabellenobjekte kommt anschließend das Vision-Transformer-Modell *TableFormer* zum Einsatz, das die logische Zeilen- und Spaltenstruktur einer Tabelle rekonstruiert und die Zellen semantisch klassifiziert (z. B. Kopf- oder Körperzellen). Für gescannte oder bildbasierte Dokumente steht optional eine OCR-Komponente auf Basis von *EasyOCR* zur Verfügung. [Aue+25]

Nach Abschluss aller Erkennungsschritte werden die Ergebnisse zu einem vollständigen *DoclingDocument* zusammengeführt. *DoclingDocument* ist eine vereinheitlichte interne Repräsentation, die sämtliche Inhalte eines Dokuments (Text, Tabellen, Bilder, Layoutinformationen, Hierarchieebenen und Metadaten) in strukturierter Form abbildet und damit das Herzstück der *Docling*-Extraktion. Diese Dokumente können in verschiedene Formate übersetzt und exportiert werden, darunter JSON, Markdown und HTML. Im Post-Processing ergänzt ein sprachsensitives Modell weitere Merkmale wie die Korrektur der Lesereihenfolge, die automatische Spracherkennung und die Extraktion zentraler Metadaten (Titel, Autoren, Referenzen). [Aue+24]

Durch diese Architektur adressiert *Docling* mehrere der in 2.2.2 beschriebenen Extraktionsprobleme, die klassische Tools nur unzureichend lösen können. Es rekonstruiert eine konsistente Lesereihenfolge auch bei mehrspaltigen Layouts, erkennt logische Dokumentstrukturen und kann Tabellen semantisch interpretieren, anstatt sie rein geometrisch zu segmentieren. [Aue+25] Darüber hinaus bietet es eine robuste Metadaten- und Inhaltsklassifizierung, die zwischen Fließtext, Überschriften, Listen, Bildunterschriften und Formeln unterscheidet. Die erzeugten Ausgaben sind reich strukturiert und dienen als Grundlage für weiterführende Analysen oder Datenpipelines, etwa zur Wissensextraktion, semantischen Suche oder automatisierten Inhaltsklassifikation. [Aue+24]

Tabelle 2.2 fasst die Extraktionsfähigkeiten der beiden Ansätze zusammen.

Tabelle 2.2: Vergleich der Extraktionsfähigkeiten von *pdfplumber* und *Docling*

Aspekt	**pdfplumber**	**Docling**
Zeichen- und Worterkennung	gut – präzise Koordinatenanalyse für editierbare PDFs	gut – kombiniert geometrische und visuelle Merkmale
Lesereihenfolge (Reading Order)	schlecht – keine Korrektur v.a. bei mehrspaltigem Layout	gut – erkennt Spalten und Lesefluss kontextsensitiv
Absatz- und Textstruktur	teilweise – heuristisch aus Zeilenabständen abgeleitet	gut – erkennt Absätze, Überschriften und Listen
Tabellenerkennung	teilweise – zuverlässig bei klaren Linien, sonst fehlerhaft	gut – rekonstruiert Tabellen semantisch mit KI-Modell
Grafiken und eingebettete Objekte	nicht – keine Analyse oder Erkennung	teilweise – erkennt Abbildungen und Beschriftungen
Metadatenextraktion	nicht – keine Unterstützung	gut – extrahiert Titel, Autoren, Referenzen
Nicht-textuelle Inhalte (OCR)	nicht – nur editierbare PDFs	gut – optionale OCR für gescannte Dokumente
Komplexe Layouts (mehrspaltig, technisch)	schlecht – häufige Fehlsegmentierung	gut – robuste Layout-Analyse durch *DocLayNet*
Semantische Rollen (z. B. Caption, Footnote)	nicht – keine Klassifizierung	gut – unterscheidet semantische Dokumentelemente

2.2.4. Zielformat JSON

Die aus PDF-Dokumenten extrahierten Inhalte, beispielsweise in Markdown- oder Textform, bieten trotz ihrer besseren Lesbarkeit keine strukturierte Grundlage für eine automatisierte Datenanalyse. Weder die mit *pdfplumber* gewonnenen Textsegmente noch die von *Docling* erzeugten Markdown-Dateien enthalten eine einheitliche logische Struktur, die eine konsistente Zuordnung von Umweltindikatoren, Materialien oder Metadaten über verschiedene PEPs hinweg erlaubt. Für weiterführende Analysen ist daher ein fest definiertes, maschinenlesbares Zielformat erforderlich, das alle relevanten Inhalte in klar benannten Feldern abbildet.

Das in dieser Arbeit verwendete textbasierte Austauschformat *JavaScript Object Notation (JSON)* hat sich als Standard für den strukturierten Datenaustausch etabliert und ist sowohl für Menschen lesbar als auch für Maschinen leicht zu verarbeiten [Pez+16]. Obwohl es historisch aus der JavaScript-Syntax hervorgegangen ist, wird JSON sprachunabhängig in nahezu allen modernen Programmiersprachen eingesetzt [Pez+16]. JSON kombiniert einfache Datentypen (*string, number, boolean, null*) mit komplexen Strukturen wie *objects* (Schlüssel–Wert-Paare) und *arrays* (geordnete Listen), wodurch hierarchische, verschachtelte Informationen kompakt und eindeutig dargestellt werden können [Pez+16].

In dieser Arbeit dient JSON als einheitliches Zielformat für die harmonisierte Speicherung der extrahierten PEP-Ecopassport-Daten. Das Format ermöglicht eine konsistente, maschinenlesbare Repräsentation komplexer Strukturen wie Umweltindikatoren, Materialkompositionen und Energieverbrauchsmodellen und lässt sich nahtlos in nachgelagerte Analyseumgebungen (z. B. Python, R oder Datenbanken) integrieren [Pez+16]. Damit bildet JSON die Grundlage für eine standardisierte und reproduzierbare Datenanalyse.

2.2.5. Informationsextraktion Markdown/Text -¿ JSON

! TODO: Kapitel benennen !

Die Informationsextraktion *(Information Extraction, IE)* dient dazu, aus unstrukturierten Texten, wie die Markdown Dateien in dieser Arbeit, gezielt Informationen zu gewinnen und diese in eine strukturierte Form, wie JSON, zu bringen [Gri15]. Grundsätzlich lassen sich zwei methodische Ansätze unterscheiden: (1) klassische, regel- oder modellbasierte Pipeline-Systeme und (2) moderne, auf großen Sprachmodellen (LLMs) basierende Verfahren.

Regelbasierte Ansätze Traditionelle IE-Systeme folgen einer mehrstufigen Verarbeitungspipeline. Typischerweise werden dabei in aufeinanderfolgenden Schritten benannte Entitäten erkannt, syntaktische Strukturen analysiert, Koreferenzen aufgelöst und schließlich Relationen zwischen Entitäten extrahiert. Solche Systeme nutzen überwiegend probabilistische Sequenzmodelle wie Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) oder Feature-basierte Klassifikatoren. [Gri15] Der Vorteil liegt in der hohen Präzision und der Nachvollziehbarkeit einzelner Verarbeitungsschritte und ihrer Deterministik. Ihre Schwächen zeigen sich jedoch bei komplexen oder stark heterogenen Textformaten, wie sie in aus PDF-Dokumenten extrahierten Markdown- oder Textsegmenten vorkommen: Fehler in einer frühen Pipeline-Stufe können sich fortpflanzen (Fehlersummierung) und die Erstellung regelbasierter Komponenten ist zeit- und ressourcenintensiv, vor allem bei großen Unterschieden in der Struktur des inputs, wie es in dem Kontext dieser Arbeit gegeben ist. [Gri15]

LLM-basierte Ansätze Eine neuere Möglichkeit stellen große Sprachmodelle (LLMs) dar, um Informationsextraktion als semantisches Verständnisproblem zu formulieren. LLMs können Textpassagen kontextsensitiv interpretieren und strukturierte Ausgaben, etwa in JSON-Form, direkt generieren. [Nad+24] Sie sind in der Lage, Entitäten, Relationen und numerische Werte inhaltlich zuzuordnen, ohne dass ein manuelles Regelwerk oder ein domänenspezifisch annotiertes Trainingskorpus erforderlich ist. [Mor+25] Zudem ermöglichen sie die Extraktion aus komplexen Layouts, indem sie zuvor durch Tools wie *Docling* generierte Markdown- oder Textsegmente semantisch analysieren. Damit entfällt die sequentielle Verarbeitung einzelner Pipeline-Stufen. [Nad+24] Der Hauptnachteil besteht in der verlorenen Deterministik und möglichen Halluzinationen (falsch generierten Werten), die durch präzises Prompt-Design und Validierungsschritte minimiert, aber nicht vollständig ausgeschlossen werden können. [Mor+25]

Vergleich Während klassische regelbasierte Verfahren beim Transfer von Markdown-Dateien in strukturierte JSON-Formate durch ihre klare Logik und Nachvollziehbarkeit überzeugen, stoßen sie bei komplexen Textstrukturen und uneinheitlichen Formulierungen an Grenzen. LLM-basierte Methoden bieten hier eine deutlich höhere Flexibilität, da sie Inhalte kontextsensitiv interpretieren und direkt in das definierte JSON-Schema überführen können. Allerdings erfordern sie eine Validierung der Modellantworten, da geringere Transparenz und vereinzelte Fehlinterpretationen möglich sind. Da diese Arbeit auf eine quantitative Analyse der PEP-Ecopassport PDFs abzielt, wird trotzdem ein LLM für die Extraktion und Überführung in das harmonisierte JSON-Zielformat eingesetzt.

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. Die *Hauptkomponentenanalyse* wird dafür verwendet die wichtigsten Merkmale zu identifizieren. 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 und explorative Statistik

Die deskriptive und explorative Statistik bilden die Grundlage der quantitativen Datenanalyse. Beide dienen der Zusammenfassung, Beschreibung und Visualisierung 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]

Deskriptive Statistik Die deskriptive Statistik umfasst numerische und grafische 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*, *Median* und *Modalwert*. 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. Der Modalwert ist der Wert, der in der Stichprobe am häufigsten vorkommt. [Dim+19] 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]

Verteilungsformen und Schiefe Ein zentrales Merkmal numerischer Daten ist die Form ihrer Verteilung. In symmetrischen Verteilungen fallen Mittelwert, Median und Mo-

dalwert 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. In der explorativen Praxis werden zudem log-transformierte Werte betrachtet, um stark asymmetrische Verteilungen zu symmetrisieren und visuell leichter interpretierbar zu machen.[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. 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 ein Datensatz normalverteilt ist.

2.3.3. Mathematische Transformationen

Um reale Daten einer Normalverteilung anzunähern und somit die Nutzung parametrischer Tests zu ermöglichen, werden nach [MJ10] und [FM09] mathematische Transformationen angewendet. Die logarithmische Transformation (Log-Transformation) wird als eine der gängigsten Methoden, um Rechtsschiefe zu korrigieren, 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

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

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 [25e].

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.

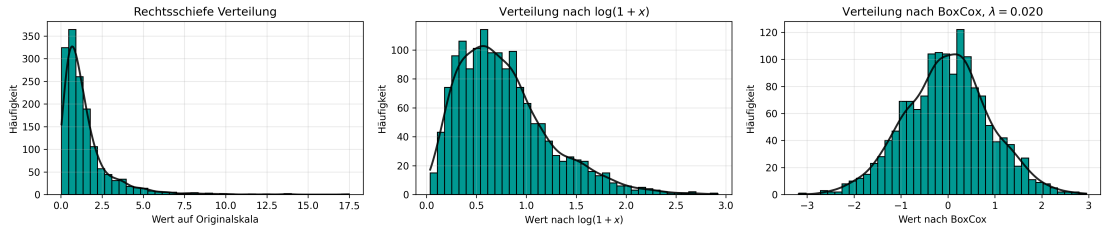


Abbildung 2.1: Beispiel einer rechtsschiefen Verteilung vor und nach Transformationen. Die überlagerte Dichtekurve verdeutlicht die Form der Verteilung.

2.3.4. Hauptkomponentenanalyse (PCA)

Die *Hauptkomponentenanalyse* (Principal Component Analysis, PCA) ist eine weit verbreitete multivariate statistische Methode zur Vereinfachung komplexer Datensätze, Extraktion der wesentlichen Informationen und der Erkennung von Strukturen in den Daten. Ihre mathematischen Grundlagen beruhen auf fundamentalen Konzepten der linearen Algebra. [AW10]

Die PCA verfolgt zwei zentrale Ziele:

Reduktion der Dimensionalität Die ursprünglichen Variablen werden durch eine kleinere Menge neuer, orthogonaler Variablen ersetzt, die *Hauptkomponenten*. Diese werden als lineare Kombinationen der Ausgangsvariablen konstruiert und erklären maximal mögliche Varianz. Die erste Hauptkomponente trägt dabei den größten Anteil der Gesamtvarianz. [MR93]

Vereinfachung und Interpretation Durch Komprimierung ohne wesentlichen Informationsverlust ermöglicht die PCA eine vereinfachte Darstellung des Datensatzes sowie die Analyse von Beziehungen zwischen Beobachtungen und Variablen. Muster von Ähnlichkeiten können grafisch auf wenigen Achsen dargestellt werden. [AW10]

Mathematische Grundlagen Die PCA beruht darauf, aus einer Menge korrelierter Variablen neue, unkorrelierte Variablen (*Hauptkomponenten*) zu erzeugen. Dafür wird die Varianz-Kovarianz-Matrix S zerlegt [MR93]. Die Hauptkomponenten ergeben sich aus den Eigenwerten und Eigenvektoren dieser Matrix. Dazu wird

$$|S - \lambda I| = 0$$

gelöst. Die Eigenwerte λ_i geben an, wie viel Varianz jede Hauptkomponente erklärt. Die zugehörigen Eigenvektoren liefern die Gewichte der linearen Kombinationen, aus denen die Hauptkomponenten entstehen:

$$V = A'X.$$

[MR93] Die Hauptkomponenten sind unkorreliert [AW10].

Singulärwertzerlegung Statt über Eigenwerte kann die PCA auch mit der Singulärwertzerlegung (SVD) berechnet werden, wie es in modernen Softwarebibliotheken (z.B. Python scikit-learn [25d] und Matlab [25b]) üblich ist:

$$X = P\Lambda Q^T.$$

Die quadrierten Singulärwerte entsprechen dabei den Eigenwerten, und die Hauptkomponenten der Beobachtungen ergeben sich zu

$$F = P\Lambda.$$

Die SVD zeigt, dass die PCA eine optimale Niedrigrang-Approximation der Datenmatrix im Sinne der kleinsten Quadrate liefert. [AW10]

PCA in dieser Arbeit Die Hauptkomponentenanalyse bildet die Grundlage der Principal Component Regression (PCR), bei der die ursprünglichen Regressorvariablen durch die Hauptkomponenten ersetzt werden. Dies ist insbesondere dann vorteilhaft, wenn starke Multikollinearität zwischen den Variablen besteht, da die Schätzung der Regressionskoeffizienten durch die Entkorrelierung wesentlich stabiler wird. [Jol82] Da die in den PEP-Daten enthaltenen Materialanteile teils ausgeprägt korrelieren, wird die lineare Regression in dieser Arbeit auf Basis der zuvor berechneten Hauptkomponenten durchgeführt.

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.

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

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

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]

Zentrale Annahmen Für die lineare Regression gelten folgende Grundannahmen:

- **Linearität:** Die Beziehung zwischen abhängiger und unabhängigen Variablen ist näherungsweise linear.
- **Erwartungswert Null:** 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.

[**Su2012**] 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.

Parameterschätzung 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:

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

Die resultierenden Schätzer sind unter den genannten Modellannahmen unverzerrt und besitzen die kleinste Varianz unter allen linearen, unverzerrten Schätzverfahren (Gauss-Markov-Eigenschaft). [**Su2012**]

Modellinterpretation 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]

Anwendungsrahmen 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. Der Teil der Normalisierung und Datenbereinigung wird in 3.1.3 detailreicher dargestellt.

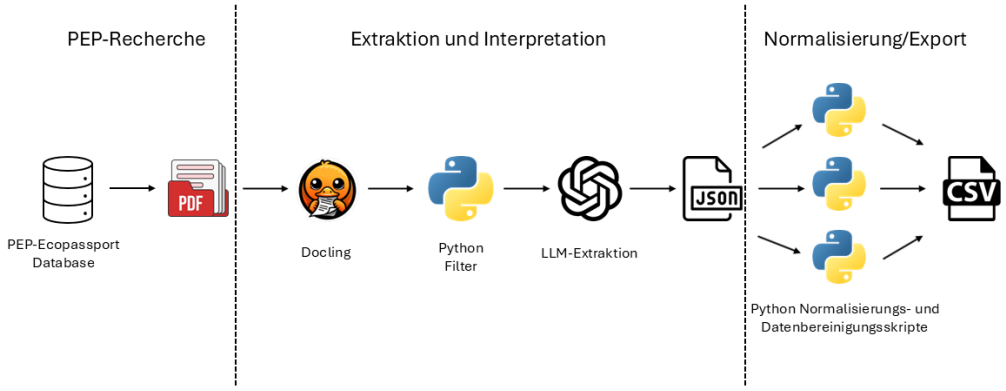


Abbildung 3.1: Schematischer Aufbau der Pipeline: von der PEP-Erfassung bis zur strukturierten Datenbasis.

3.1.1. 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 bildet dabei die Quelle der Untersuchung. Für jedes gefundene Dokument wurden Produktinformationen, Material- und Energiedaten sowie Metadaten in strukturierter Form erfasst.

Zielsetzung Im ersten Schritt der Arbeit sollte eine Datenbasis zu IoT-Produkten in der PEP-Datenbank aufgebaut werden. Dazu wurden die Dokumente automatisiert re-

cherchiert, 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.

Keyword-basierte Suche 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. Insbesondere der Endpoint `/xhr/searchPep` lieferte HTML-Snippets mit Produktnamen und Links. Diese wurden mithilfe von JavaScript iterativ abgefragt, geparst und in einer CSV-Datei gespeichert. Die Implementierung verwendete `fetch`-Requests mit kopierten FormData-Parametern und CSRF-Tokens. Für die Paginierung wurde die Gesamtanzahl der Treffer aus dem Response genutzt, um alle Ergebnisse in Schleifen abzurufen. Um gezielt IoT-nahe Produkte zu erfassen, wurden die Abfragen mit spezifischen Suchbegriffen in den FormData-Parametern erweitert (z. B. *controller*, *sensor*, *gateway*, *wifi*, *knx*, *zigbee*, *cloud*). Dadurch konnten 184 PEP-Einträge, welche potentiell Gebäudeautomatisierungsgeräte beschreiben identifiziert und als CSV exportiert werden. Anschließend erfolgte eine manuelle Prüfung und Klassifikation der Ergebnisse, da die Suchbegriffe im Produkttitel nicht ausnahmslos 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.

Klassifikation der Produkte Die ermittelten Produkte wurden in einer Excel-Datei manuell kategorisiert. Neben dem Produktnamen und der URL enthielt die Datei eine Spalte *IoT-Einschätzung* mit vordefinierten Auswahloptionen. Farbcodierungen erleichterten die visuelle Trennung zwischen den Gruppen. Zur Validierung wurden zusätzlich die jeweiligen PEP-PDFs gelesen und, falls erforderlich, weitere Produktinformationen aus Herstellerportalen herangezogen. So konnten 102 Smart-Home-relevante Geräte eindeutig als IoT oder IoT-nah eingestuft werden.

Kategorisierung Die Zuordnung erfolgte nach funktionalen Kriterien:

- **Gebäudeautomatisierung:** Geräte mit Konnektivität (z. B. ZigBee, WiFi, KNX) oder Cloud-Anbindung, wie Gateways, smarte Sensoren oder Steuerungen.
- **eher ja:** Komponenten mit indirekter IoT-Relevanz, etwa Erweiterungsmodule. (Teilweise aussortiert)
- **eher nein:** Elektronik mit digitaler Funktion, jedoch ohne Vernetzung. (Aussortiert)

- **keine IoT-Relevanz:** Produkte ohne Kommunikationsfähigkeit (z. B. Kabel, Trafos, LED-Panels). (Aussortiert)

Manuelle Ergänzungen 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. Dadurch erweiterte sich der Datensatz um weitere 145 PEPs. Insgesamt umfasst die erstellte Datenbasis 247 PEP-Dokumente, die anschließend in der Parsing-Pipeline verarbeitet und vereinheitlicht wurden.

3.1.2. 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, Materialanteile und Metadaten zu erzeugen, welche entsrechend der Zielsetzung der Arbeit analysiert wereden können. Die finale Lösung kombiniert eine robuste Layoutanalyse mit Docling und eine LLM-basierte, schemagesteuerte Inhaltsinterpretation.

Entwicklung und Vorläufer Zu Beginn wurde eine auf `pdfplumber` basierende Pipeline eingesetzt, die auf der von Selg [Sel25] entwickelten Pipeline aufbaut. Sie erkennt mithilfe 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 entwickelt 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 ausreichend anzuwenden. 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 Downstream-Verarbeitung deutlich verbessert

werden.

Methodisches Konzept 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. Hierbei werden nicht inhaltsrelevante Textsegmente wie Kopf- und Fußzeilen, Unternehmensinformationen oder generische Beschreibungen der PEP-Standards über eine Blacklist entfernt. Diese Vorverarbeitung reduziert die Eingabelänge und verbessert die inhaltliche Fokussierung der anschließenden LLM-basierten Extraktion.
- **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.

Wahl des Modells und der LLM Schnittstelle Für die semantische Extraktion wurde *GPT-5* verwendet, angesprochen über die *Responses-API*. Diese neue Schnittstelle unterstützt native strukturierte Ausgaben und optional eine Schema-Validierung. Der Aufruf erfolgt im `response_format=json_object`-Modus, wodurch fehlerhafte JSON-Formate ausgeschlossen sind. Gegenüber dem früher verwendeten *gpt-3.5-turbo* zeigte sich GPT-5 deutlich stabiler in der Erkennung von numerischen Werten, Einheiten und Modulzuordnungen (A1–A3, A4, A5, B*, C*, D). Zudem reduziert sich der Post-Processing-Aufwand erheblich, da keine nachträgliche JSON-Reparatur erforderlich ist. 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.

Extraktionslogik

- **Indikatoren:** Matching über Name und Einheit auf Basis der EF 3.1-Labels (z. B. $\mathrm{kg\,CO_2\,eq}$, $\mathrm{kg\,Sb\,eq}$, MJ). Header-Kontext wird mitgeparst, B-Phasen werden nicht aggregiert.
- **Plausibilität:** Flatline-Filter (identische Modulwerte), Prüfungen von *Total* vs. Modulsumme, Toleranz für negative D-Werte.

- **Einheiten:** Zahlen ohne Einheitenzeichen. Einheiten werden ausschließlich in das Feld `unit` extrahiert, ohne heuristische Ergänzung oder Raten.

Robustheit und Grenzen Die neue Pipeline konnte die Anzahl fehlerhafter oder unvollständiger Einträge deutlich reduzieren. Fallback-Mechanismen greifen bei fehlerhaften Tabellen automatisch auf den Fließtext zurück, wodurch auch reine Text-PEPs ausgewertet werden können. Für PDFs mit reinen Rastertabellen bleibt jedoch eine Einschränkung bestehen, da ohne OCR keine Inhaltsextraktion möglich ist. Der Einsatz eines LLMs führt aufgrund der stochastischer Modellkomponenten zudem zu einer eingeschränkten Reproduzierbarkeit und Transparenz: Obwohl das Risiko minimiert wurde, erzeugen identische Eingaben aufgrund von Halluzinationen nicht immer identische Ausgaben. Diese Einschränkungen sind angesichts der PDF-Heterogenität nicht zu umgehen und methodisch vertretbar.

3.1.3. Normalisierung der Daten

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

Vokabularanalyse Zur systematischen Erfassung der vorhandenen Begriffe wurde ein Hilfsskript (`pep_vocab_scan.py`) 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. Die Ausgabe diente der Identifikation inkonsistenter Schreibweisen, Abkürzungen und Synonyme, die anschließend über Zuordnungstabellen (*Mapping Dateien*) vereinheitlicht wurden.

Normalisierung mittels Mapping-Tabellen Für jede Datendomäne (z. B. Einheit, Material, Land, Energiequelle, Phase) wurde eine Zuordnungstabelle erstellt, die reguläre Ausdrücke den vereinheitlichten Standardbegriffen zuordnet. Diese Mappings wurden schrittweise verfeinert, bis alle identifizierten Abweichungen abgedeckt waren. Beispielsweise wurden die Einheitenvarianten *„kg CO_2-eq“*, *„kg CO2e“* und *„kg CO2 equiv.“* auf den Standardbegriff *„kg CO2 eq“* gemappt. Ebenso wurden Synonyme wie *„Aluminum“* und *„Aluminium“* oder *„Paper“* und *„Carton“* zu gemeinsamen Bezeichnungen zusammengeführt. Auch Länderangaben wurden vereinheitlicht, indem Bezeichnungen

und ISO-Kürzel (z. B. „France“, „FR“) zu konsistenten Formen zusammengefasst wurden.

Phasen- und Energiemodell-Normalisierung Für Lebenszyklusphasen wurden die in den PEPs auftretenden Varianten (z. B. „A1–A3“, „Production“, „Manufacturing“) auf ein einheitliches Schema („manufacturing“, „distribution“, „use“, „end_of_life“) abgebildet. Analog wurden Angaben zu Strommixen standardisiert, wobei landesspezifische Modelle (z. B. „FR“, „France Mix“, „French grid“) zu klar benannten Einträgen wie *„France grid mix“* zusammengeführt wurden. Diese Normalisierung ist entscheidend, um energiebezogene und phasenabhängige Indikatoren konsistent vergleichen zu können.

Entscheidungen und Vereinfachungen Bei der Vereinheitlichung wurden einige pragmatische Entscheidungen getroffen: *Steel* und *Iron* wurden beispielsweise zur Kategorie *Steel (metals)* zusammengefasst, da beide ähnliche Materialeigenschaften und Umweltwirkungen aufweisen. *Paper* und *Carton* wurden zu *Paper/Cardboard* kombiniert. Für Kunststoffe wurde eine vereinfachte Zusammenführung vorgenommen. Ein Python-Skript *apply_mappings.py* wendet die auf dem Vokabular basierenden Mapping Dateien auf alle extrahierten JSON Dateien an.

Konsolidierung von Materialeinträgen Durch das Mapping entstehen teilweise doppelte Materialeinträge, welche innerhalb der Feldstruktur `material_composition` mithilfe von *merge_material.py* zusammengeführt werden, sofern sie identische Bezeichnungen aufwiesen. Dabei wurden nur exakt gleiche Strings (nach Vereinheitlichung von Groß-/Kleinschreibung und Leerzeichen) berücksichtigt Prozent- und Gewichtsangaben wurden bei Dubletten addiert und auf drei Nachkommastellen gerundet. Diese Maßnahme reduziert Redundanz und erleichtert die spätere Aggregation der Materialanteile.

Ergebnis 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.

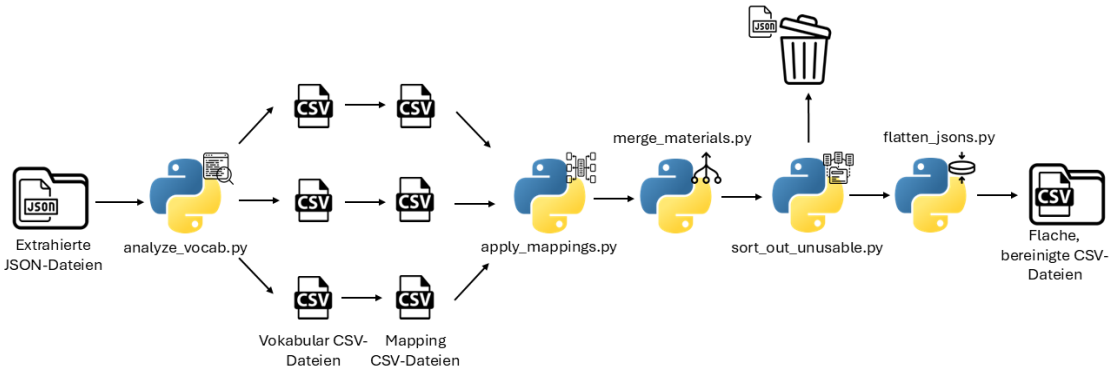


Abbildung 3.2: Übersicht der Skriptabfolge zur Normalisierung und Bereinigung der extrahierten JSON-Daten.

3.1.4. Datenbereinigung und Validierung

Ausschluss unvollständiger Datensätze Nach der automatischen Extraktion und Normalisierung werden 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.

Qualitätssicherung und Validierung Die Datenbereinigung wurde durch automatisierte Prüfmechanismen begleitet. Dazu zählten Validierungen auf fehlende oder ungültige Einheiten, numerische Typfehler (z. B. String statt numerischer Wert) sowie Plausibilitätsprüfungen, etwa auf Null- oder Extremwerte bei zentralen Variablen. Fehlerhafte Regex-Definitionen, die während der Normalisierung auftraten, wurden iterativ korrigiert und mit Testdateien überprüft. Ein zusätzlicher Kontrolllauf identifizierte Datensätze mit nicht plausiblen Summen oder Flatline-Indikatoren (identische Werte in allen Phasen), die von der weiteren Analyse ausgeschlossen wurden.

Entscheidungen und Grenzen 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.

Datenformatierung für die statistische Auswertung 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.

Ergebnis 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

Explorative Modellentwicklung

Dieses Kapitel beschreibt die Wahl des Regressionsmodells. Im Mittelpunkt steht die explorative Entwicklung und experimentelle Erprobung verschiedener Modellvarianten, aus denen das in Abschnitt 5.3 eingesetzte Hauptmodell für die CO_2-Äquivalente sowie die übrigen Indikatoren abgeleitet wurde.

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 erwiesen haben und welche Ansätze verworfen wurden.

4.1. Experimentelle Fragestellungen

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

1. **Beitrag der Materialinformationen:** Verbessert sich die Vorhersagegüte gegenüber einem Modell, das nur Gewicht und Stromverbrauch nutzt, wenn zusätzlich Materialinformationen einbezogen werden?
2. **Rohmaterialien vs. Material-PCA:** Ist es günstiger, die zahlreichen Materialspalten direkt zu verwenden, oder führt eine PCA des Materialblocks zu stabileren Modellen?

3. **Wahl des Regressionsverfahrens und Regularisierung:** Unterscheiden sich OLS, `Ridge` und `Lasso` hinsichtlich Stabilität und erzielbarer Gütemaße bei den vorhandenen PEP-Daten, insbesondere bei vielen korrelierten Regressoren?

4. **Wahl der Zieltransformation:** Führt eine Log-Transformation (`log1p`) 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. 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 CO_2-Äquivalente vollständig vorliegen.

Verglichen wurden drei Feature-Sets, jeweils in einem linearen Regressionsmodell auf der Transformationsskala $\log(1 + \mathrm{CO2}_{\mathrm{total}})$:

- **Basis:** nur Gewicht und Stromverbrauch.
- **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 %).

Für alle drei Varianten wurden dieselben Train/Test-Splits verwendet, so dass die Gütemaße direkt vergleichbar sind. Tabelle 4.1 zeigt die erzielten Ergebnisse.

Das Basismodell aus Gewicht und Stromverbrauch erklärt bereits einen großen Anteil der Varianz der CO_2-Äquivalente. 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 $\mathrm{RMSE}_{\mathrm{Test}}$ leicht. Dies ist ein Hinweis auf Überanpassung durch die vielen, teilweise korrelierten Materialvariablen.

Tabelle 4.1: Vergleich verschiedener Feature-Sets für den Indikator *Climate change (total)* auf der Skala $\log(1 + \text{CO2}_{\text{total}})$.

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

Deutlich besser schneidet das Modell mit einem Minimalvorkommen an Materialien ab. R^2_{Test} ist hier durchscnittlich 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 SGD-Regressionsansatz gewählt.

4.3. Vergleich der Regressionsverfahren

Die im vorherigen Abschnitt beschriebenen Experimente zum Vergleich der Feature-Sets wurden für den Indikator *Climate change (total)* zunächst mit einer klassischen linearen Regression auf Basis der der `LinearRegression` aus `scikit-learn` nachgebildet und in eine Pipeline mit `StandardScaler`, optionaler PCA und Train/Test-Aufteilung eingebettet.

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 äußere Wiederholungen des Experiments ähnliche R^2- und RMSE-Werte. Beide Verfahren schnitten konstant besser ab als `LinearRegression`. Die Wahl des Regressors war allerdins, wie die Zahlen zur Modellgüte in 4.2 zeigen, damit deutlich weniger bedeutend als der Einfluss des Feature-Sets.

Tabelle 4.2: Vergleich der Regressionsverfahren auf dem festgelegten Feature-Set.

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

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

5

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.

5.1. 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.

5.1.1. Vollständigkeit der Werte

Zur Bewertung der Datenvollständigkeit wurde der Anteil fehlender Werte pro Variable berechnet und in einem Balkendiagramm dargestellt (Abb. 5.1). 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.

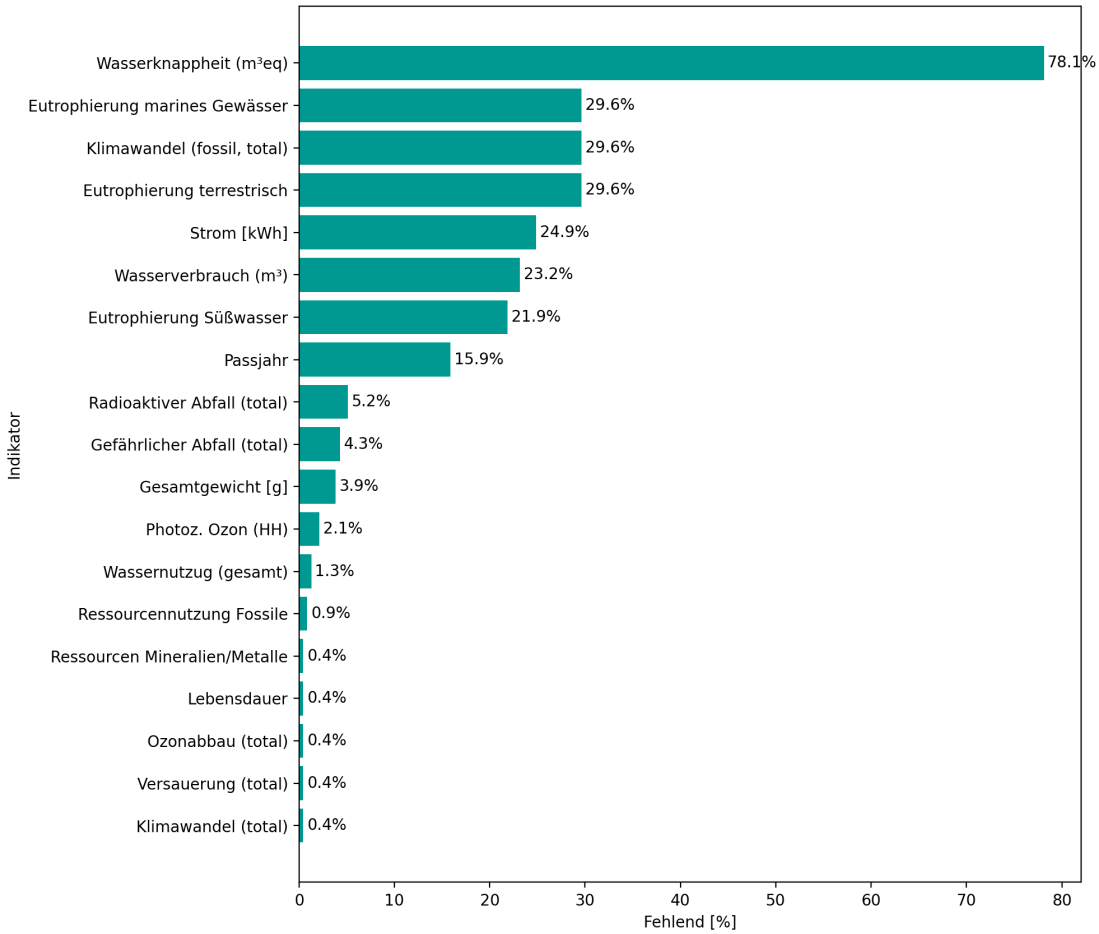


Abbildung 5.1: Anteil fehlender Werte pro Umweltindikator (%, $N = 233$ Produkte).

Zusätzlich wurden die Erscheinungsjahre der PEP-Ecopassports untersucht. Die Veröffentlichungen reichen von 2020 bis 2025 und verteilen sich wie in Abb. 5.2 dargestellt. Der deutliche Anstieg ab 2022 zeigt die zunehmende Etablierung des Formats und eine stärkere Datenverfügbarkeit in den letzten Jahren.

5.1.2. 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.

Die in Tab. 5.1 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

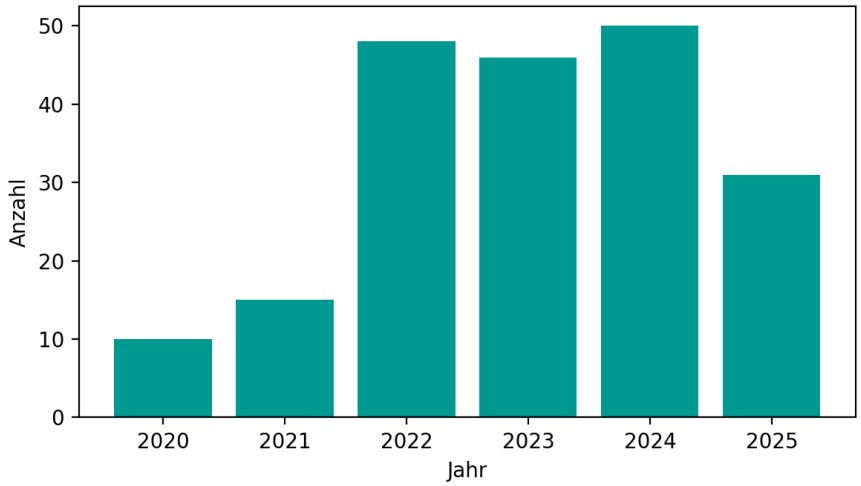


Abbildung 5.2: Erscheinungsjahre der analysierten PEP-Dokumente ($N = 233$ Produkte).

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 5.1: Robuste deskriptive Kennzahlen der Basisvariablen.

Connected dimmer mit Bluetooth interface (PEP-Link), während das größte Produkt, ein *Flüssigkeitskühler mit drehzahlgeregeltem Schraubenverdichter und Greenspeed*^{TM}*-Technologie* (PEP-Link), mehr als 13 t erreicht. Der Mittelwert liegt mit 278 kg weit über dem Median (2.18 kg), was die ausgeprägte Rechtsschiefe bestätigt.

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

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

Im Datensatz zeigt sich, dass der Großteil der verwendeten Energiemodelle auf allgemeine europäische Strommixe (EU27) und Frankreich entfällt. 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 in der Regel ein repräsentativer 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.

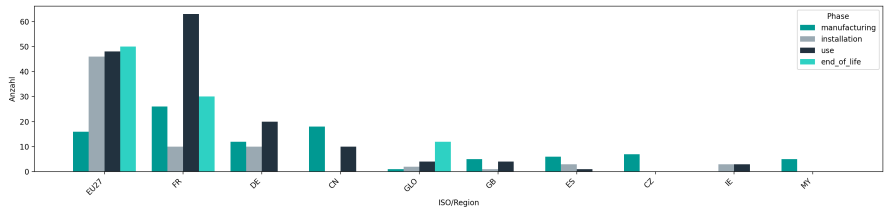


Abbildung 5.3: Verteilung der verwendeten Energiemodelle (ISO-Regionen) über die Lebenszyklusphasen

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 5.2 wird aufgeführt, aus welchen Materialien das durchschnittliche PEP-Produkt aus dem Datensatz besteht (Mittelwert). *N* gibt 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 5.2 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

Tabelle 5.2: Durchschnittliche Materialanteile nach Hauptkategorien (Mittelwert in %).

(a) Metalle

Material	Mittelwert
Stahl	$2.65 \times$
Aluminium	6.13
Kupfer	5.41
Messing	$8.60 \times$
Zamak	$4.50 \times$
Nickel	$1.00 \times$
Zinn	$6.00 \times$
Zink	$5.00 \times$
Bronze	$1.00 \times$
Neodym	$1.00 \times$
Hartlot	0.00

(b) Kunststoffe

Material	Mittelwert
Polycarbonat (PC)	8.23
ABS	2.84
Polyamid (PA)	2.37
PVC	2.08
PS	1.13
PP	$8.00 \times$
Gummi	$6.90 \times$
PMMA	$6.60 \times$
Epoxidharz	$6.00 \times$
Polyesterharz	$5.70 \times$
PE	$4.40 \times$
PU	$4.00 \times$
PBT	$2.30 \times$
PET	$1.30 \times$
POM	$9.00 \times$
TBBPA	$7.00 \times$
HIPS	$6.00 \times$
Silikon	$4.00 \times$
EPDM	$2.00 \times$
PPS	$2.00 \times$
Sonstige	$3.00 \times$

(c) Andere

Material	Mittelwert	n
Papier	1.57×10^{1}	197
Elektronik	3.69	102
Holz	3.12	81
Glas	2.90	67
PCBA	1.79	24
PCB	1.27	49
Kabel	3.50×10^{-1}	38
Kältemittel	3.50×10^{-1}	53
Ferrit	2.80×10^{-1}	38
Elektromotoren	2.70×10^{-1}	9
Lack / Farbe	1.50×10^{-1}	39
Tinte	8.00×10^{-2}	15
Silizium	8.00×10^{-2}	9
Batterie	8.00×10^{-2}	10
Thionylchlorid	8.00×10^{-2}	5
Öl	7.00×10^{-2}	8
Mineralwolle	6.00×10^{-2}	13
Bitumen	4.00×10^{-2}	7
Titandioxid	4.00×10^{-2}	14
Quarz	2.00×10^{-2}	7
Flussmittel	2.00×10^{-2}	6
Filz	1.00×10^{-2}	11
Aluminiumoxid	1.00×10^{-2}	5
Haftkleber	1.00×10^{-2}	4
Sonstige	4.80×10^{-1}	–

vernachlässigbar ist. Insgesamt spiegelt sich in der Verteilung die Diversität der erfassten Produktgruppen wider.

5.1.3. 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.

! TODO: Übersetzen, EINHEITLICH! Layout anpassen !

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_2H_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 5.3: Gesamtindikatoren (Total) mit Median/IQR und Mittelwert (gerundet auf zwei Nachkommastellen).

Wie Tabelle 5.3 zeigt, weisen alle Umweltindikatoren deutlich **rechtsschiefe Verteilungen** auf: Der Mittelwert liegt bei allen Größen um ein Vielfaches über dem Median. Besonders ausgeprägt ist dies bei *Climate change (total)*, *Resource use (fossils)*, *Water use* und *Hazardous waste disposed*, bei denen einzelne Extremwerte die Verteilungen dominieren. Dagegen zeigen *Ozone depletion* und *Resource use (minerals/metals)* geringere Abstände zwischen Mittelwert und Median, bleiben aber ebenfalls schief. Insgesamt bestätigt sich eine stark heterogene Datenbasis mit wenigen Produkten, die sehr hohe Umweltwirkungen aufweisen.

5.2. PCA der Materialien

Zur explorativen Analyse und zur Reduktion der Dimensionalität der Materialdaten wurde eine Hauptkomponentenanalyse (PCA) durchgeführt. Benutzt wurde die Python-Bibliothekt *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 5.4 dargestellt sind, zeigen, dass die Daten teilweise multimodal verteilt sind und viele Beobachtungen

im Bereich sehr kleiner Anteile liegen. Eine einfache z-Standardisierung würde diese Verteilungen nicht normalisieren und wäre durch Ausreißer stark beeinflusst.

TODO: Titel hinzufügen

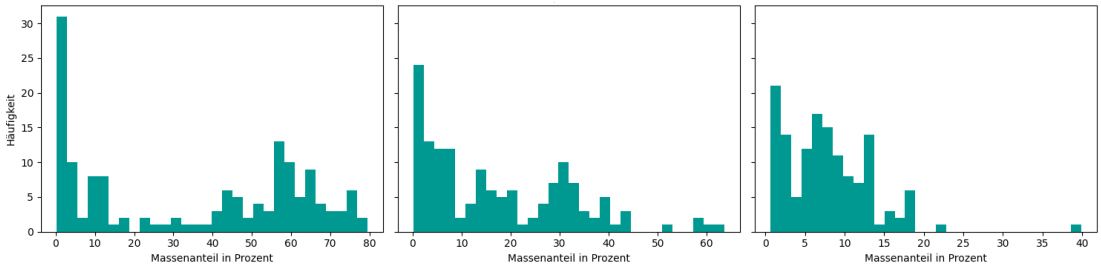


Abbildung 5.4: Beispiele für Histogramme ausgewählter Materialanteile (nur PEPs mit positivem Anteil).

Für die PCA auf dem Materialblock wird daher der *scikit-learn RobustScaler* verwendet, der jede Materialsäule um den Median zentriert und durch die Interquartilsbreite skaliert. Dadurch werden die typischen Wertebereiche der Materialien stärker gewichtet und einzelne extreme PEPs sind weniger dominant. Die PCA selbst wird anschließend auf der robust skalierten Materialmatrix durchgeführt.

5.2.1. Ergebnis der PCA

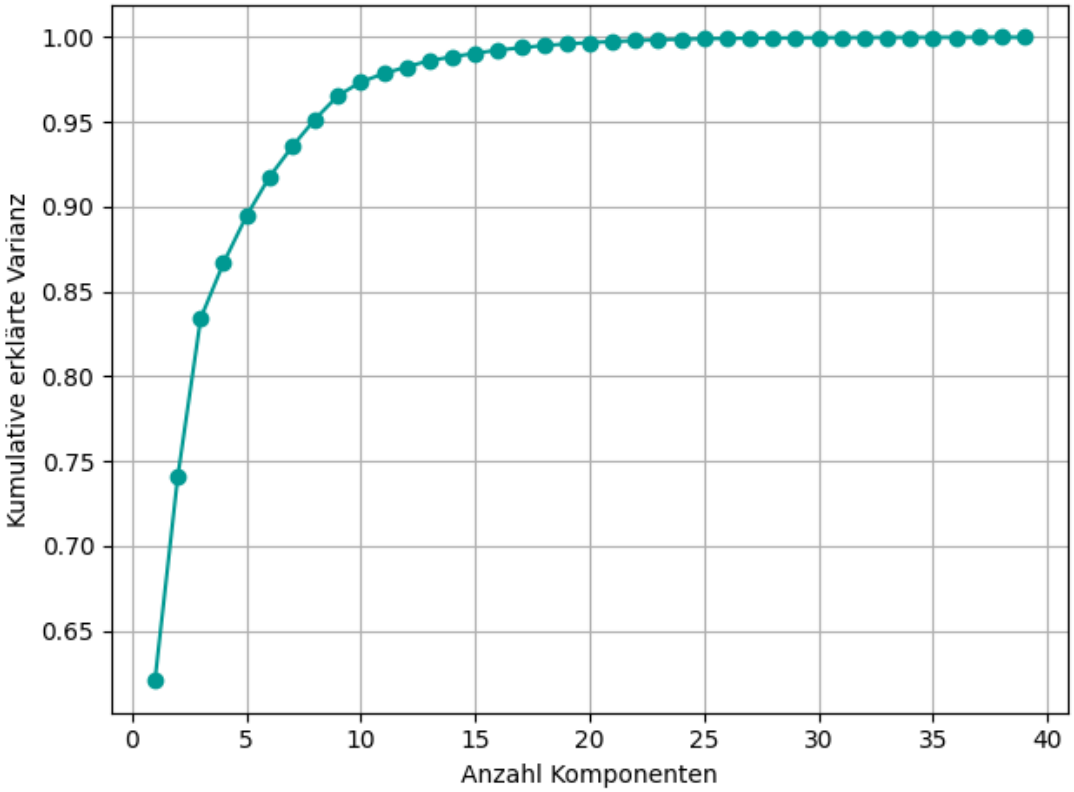


Abbildung 5.5: Kumulative erklärte Varianz

In Abbildung 5.5 ist die durch die Hauptkomponenten erklärte kumulative Varianz dargestellt. Die ersten 10 Hauptkomponenten erklären bereits ungefähr 55 % der Varianz, 20 Komponenten etwa 80 % und 30 Komponenten ungefähr 90 %. Ab etwa 30 Komponenten nehmen zusätzliche Komponenten nur noch geringe Varianzanteile auf (z. B. ca. 97 % bei 40 Komponenten). Für die weitere Modellierung wurde daher ein Varianzschwellenwert von 90 % gewählt, was in diesem Datensatz etwa $k = 30$ Hauptkomponenten entspricht. Dieser Wert ist ein Kompromiss zwischen Dimensionsreduktion und Informationsverlust.

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

5.2.2. 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:

- **PC_1 (*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.

- **PC_2 (*PE (Polyethylen)* versus übrige Kunststoffe):** PC_2 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.

- **PC_3 (unspezifische Kunststoffmischungen):** In PC_3 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.

- **PC_4 (Leiterplatten und Elektronik versus PVC):** PC_4 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.

- **PC_5 (Elektronikschwerpunkt):** In PC_5 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 bzw. Karton 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 PC_4 und PC_5 Leiterplatten und Elektronik gegenüber PVC lastigen Gehäusen abgrenzen. Diese fünf Hauptkomponenten erklären zusammen knapp 90 % der Varianz im Materialblock und bilden damit die prägendsten Materialmuster der PEPs ab. In den folgenden Regressionsmodellen werden sie als kompakte Materialindikatoren genutzt, um den Einfluss des Materialmixes auf die Umweltindikatoren zu erfassen, ohne alle einzelnen Materialspalten separat berücksichtigen zu müssen.

5.3. Lineare Regression der CO_2-Äquivalente

Ziel der folgenden Analyse ist es 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 CO_2-Indikator und ein lineares Regressionsmodell. Weitere Umweltindikatoren und alternative Modellklassen werden in späteren Abschnitten betrachtet.

5.3.1. Datenbasis und Transformation

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

Die Verteilungen der CO_2-Äquivalente, des Produktgewichts und des Stromverbrauchs sind stark rechtsschief und decken mehrere Größenordnungen ab. Um den Einfluss extremer Werte zu verringern und die Größenordnungen besser vergleichbar zu machen, werden diese Variablen mit der Funktion `log1p` transformiert. Es werden

die folgenden Größen definiert:

$$\text{log\_cc} = \log(1+\text{CO2}_{\text{total}}), \quad \text{log\_w} = \log(1+\text{weight}), \quad \text{log\_e} = \log(1+\text{electricity\_consumption})$$

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.

5.3.2. Modellformulierung

TODO: Klar machen, dass Materialien nicht geloggt werden

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

$$\text{log\_cc} = \beta_0 + \beta_1 \cdot \text{log\_w} + \beta_2 \cdot \text{log\_e} + \sum_{j=1}^{k} \gamma_j \cdot \text{PC\_mat}_j + \varepsilon,$$

wobei PC_mat_j die Material-Hauptkomponenten aus der PCA bezeichnen und k die Anzahl der verwendeten Komponenten ist. Diese Hauptkomponenten fassen jeweils ein charakteristisches Muster aus Materialanteilen zusammen (vgl. 5.2.2) und fungieren als verdichtete Materialindikatoren im Regressionsmodell. Der Fehlerterm ε umfasst alle nicht modellierten Einflüsse sowie Mess- und Rundungsfehler.

5.3.3. Schätzverfahren, Validierung und Ergebnisse

Zur Bewertung der Modellgüte wird ein Train/Test-Split mit einem Testanteil von 10 % verwendet. Das Modell wird ausschließlich auf die 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.

Ergebnisse der CO_2-Regression Die Modelle werden über das Bestimmtheitsmaß R^2 und den Root-Mean-Square-Error (RMSE) bewertet. Tabelle 5.4 fasst die Testgüte des CO_2 Regressionsmodells zusammen.

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.

Tabelle 5.4: Gütekennzahlen des linearen Regressionsmodells (`Climate Change (total)` als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.896
$\text{RMSE}_{\text{Test}}$	$25116.20\,\text{kg CO}_2$

Der absolute RMSE-Wert von $25116.20\,\text{kg CO}_2$ 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:** $266000.0\,\text{kg CO}_2$
- **Vorhersage:** $37447.76\,\text{kg CO}_2$
- **Absoluter Fehler:** $228552.24\,\text{kg CO}_2$
- **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 $2220.80\,\text{kg}$ 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 CO_2-Äquivalente im Mittel korrekt einordnet. Eine präzise Abschätzung

über alle Produkte hinweg ist mit dem verwendeten, bewusst kompakten Feature-Set nur eingeschränkt möglich.

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

Visualisierung der Vorhersagequalität Zur Veranschaulichung der Modellgüte zeigt Abbildung 5.6 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.

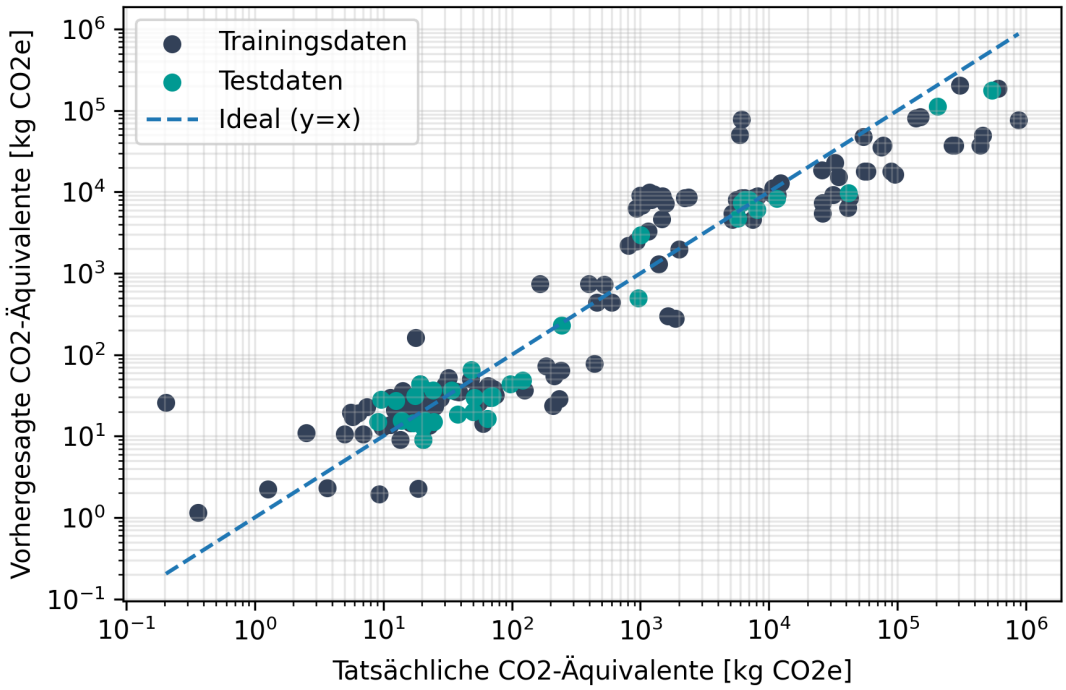


Abbildung 5.6: Vorhergesagte und tatsächliche Werte von `Climate Change total` für einen exemplarischen Train und Test Split des CO_2-Modells.

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 $10^5\,\mathrm{kg}$) bzw. überschätzt (z. B. Cluster bei $10^3, \mathrm{kg}$) werden.

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

Abbildung 5.7 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.

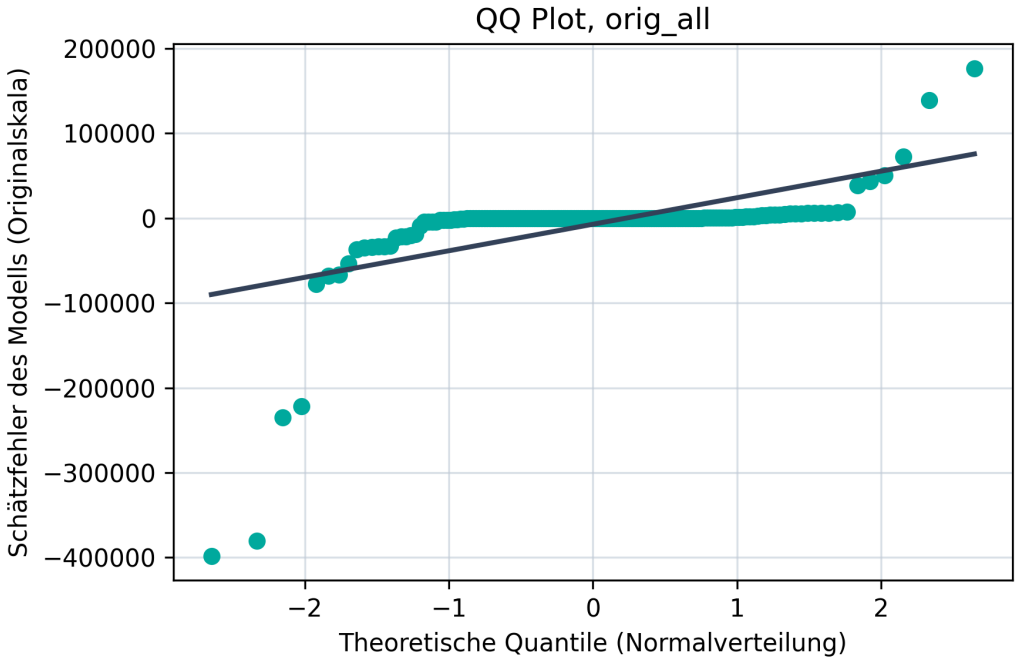


Abbildung 5.7: QQ Plot der Residuen des CO_2-Modells auf Originalskala im Testset.

Abbildung 5.8 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.

Die Normalitätsannahme ist auf Originalskala für die CO_2-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

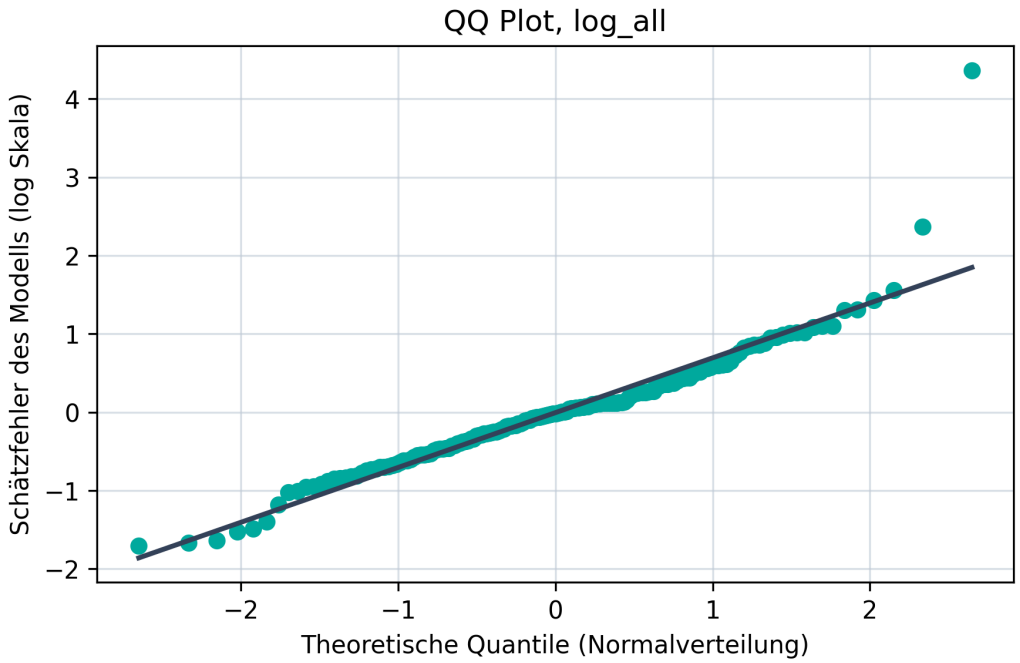


Abbildung 5.8: QQ Plot der Residuen des CO_2-Modells im Testset auf der Transformationsskala `log_cc`.

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 Fehlwerteanteile, und die Erklärbarkeit der jeweiligen Indikatoren. Für die übrigen Indikatoren werden die getesteten Transformationen (keine Transformation, `log1p`, Box-Cox) und die resultierende Auswahl jeweils knapp zusammengefasst, und die Fehlerdiagnostik wird auf die wichtigsten Befunde reduziert. Für jeden Indikator werden diese Transformationsoptionen systematisch verglichen. Die Auswahl erfolgt datengetrieben anhand der Testgüte, wobei Fehlermaße auf Originalskala nach Rücktransformation berichtet werden, damit sie in der Einheit interpretierbar bleiben.

5.4. Lineare Regression der anderen Indikatoren

Die für CO_2 aufgebaute Regressions-Pipeline wird nun auf weitere Umweltindikatoren angewendet, um zu prüfen, wie gut sich diese mit denselben Produktmerkmalen erklären lassen.

5.4.1. Regression des Indikators Acidification

Neben den CO_2-Äquivalenten wurde das lineare Regressionsmodell auch auf den Indikator *Acidification* angewendet. Es wurden keine Transformation, `log1p` und eine Box-Cox Transformation verglichen. Die beste Testgüte wird mit `log1p` 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 $\text{log\_acid} = \log(1 + \text{acidification}_{\text{total}})$. Es konnten 177 PEPs benutzt werden. Tabelle 5.5 fasst die Testleistung nach Rücktransformation zusammen.

Tabelle 5.5: Gütekennzahlen des linearen Regressionsmodells (`Acidification` als Zielvariable).

Größe	Wert (Test)
R^2_{Test}	0.845
$\text{RMSE}_{\text{Test}}$	$490.87\,\text{kg SO}_2$
Median absoluter Fehler	$0.2864\,\text{kg SO}_2$
$\text{MdARE}_{\text{Test}}$ (Median rel. Fehler)	0.9579
$\text{MARE}_{\text{Test}}$ (Mittelwert rel. Fehler)	3.8258

Das Modell erklärt rund 85% der Varianz mit einem RMSE von $490{,}87\,\text{kg SO}_2$, was auf eine insgesamt gute Vorhersagegüte für den Indikator `Acidification` hinweist. Die robusten Fehlermaße ergänzen dieses Bild. Der Median des absoluten Fehlers liegt weit unter dem RMSE bei etwa $0{,}29\,\text{kg SO}_2$ 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 5.9 ein Streudiagramm der vorhergesagten gegenüber den tatsächlichen Werten von `Acidification`.

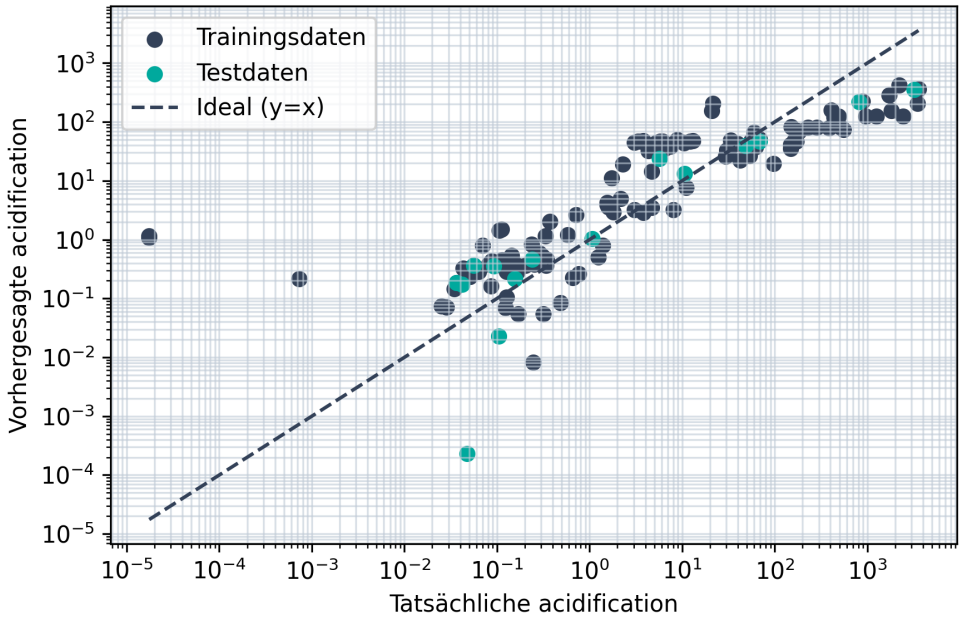


Abbildung 5.9: Vorhergesagte gegenüber tatsächlichen Acidification-Werten. Beide Achsen sind logarithmisch skaliert.

Die meisten Punkte liegen in der Nähe der Diagonalen, insbesondere im Bereich mittlerer Acidification-Werte, was auf eine gute Abbildung des allgemeinen Trends hinweist. Hohe Werte werden tendenziell leicht unterschätzt und im Bereich von 10^1 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 CO_2-Fall übernommene Modell auch für den Indikator *Acidification* robuste und plausible Vorhersagen liefert.

Abbildung 5.10 zeigt einen QQ Plot der Schätzfehler des Modells im Vergleich zu einer Normalverteilung. Wie beim CO_2 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, 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.

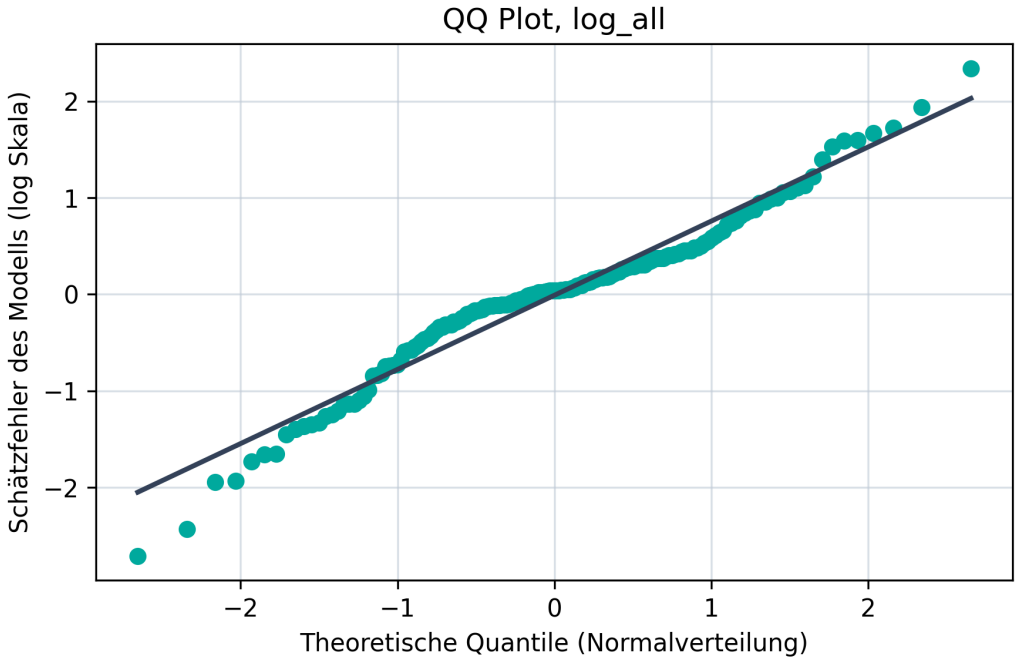


Abbildung 5.10: QQ Plot der Schätzfehler des Acidification Modells (Test- und Trainingsset).

5.4.2. Regression des Indikators Hazardous Waste Disposed

Für den Indikator *Hazardous waste disposed* konnten $n = 168$ PEPs verwendet werden. Es wurden keine Transformation, `log1p` und eine Box Cox Transformation verglichen. Die beste Testgüte wird mit `log1p` erreicht, daher wird im Folgenden dieses Modell betrachtet. Die Zielvariable ist damit log_hwd $= \log(1+\text{hazardous\_waste\_disposed}_{\text{total}})$. Tabelle 5.6 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 $177\,\text{kg}$ und somit erneut weit unter dem RMSE. Der Median der relativen Fehler ($\text{MdARE} \approx 66\%$) zeigt eine vergleichsweise gut 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.

Tabelle 5.6: Gütekennzahlen des linearen Regressionsmodells (`Hazardous waste disposed` als Zielvariable).

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

Zur Veranschaulichung zeigt Abbildung 5.11 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 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 5.12 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.

5.4.3. Regression des Indikators Water Use

Für den Indikator *Water use* wird ebenfalls das lineare Regressionsmodell mit Gewicht, Stromverbrauch und Material-PCs verwendet. Die Zielvariable ist der log-transformierte Gesamtwert $\log(1 + \text{water\_use}_{\text{total}})$. Tabelle 5.7 fasst die Testleistung über $R = 100$ äußere Train/Test-Splits zusammen.

Im Mittel erklärt das Modell damit rund 60 % der Varianz der log-transformierten Water-use-Werte, allerdings mit einer sehr hohen Streuung zwischen den einzelnen Train/Test-Aufteilungen. Der beste Lauf erreicht ein Test-R^2 von 0,872 bei einem RMSE von 1,431 (log1p-Einheiten), liegt damit aber klar unter der Güte, die für *Climate change*

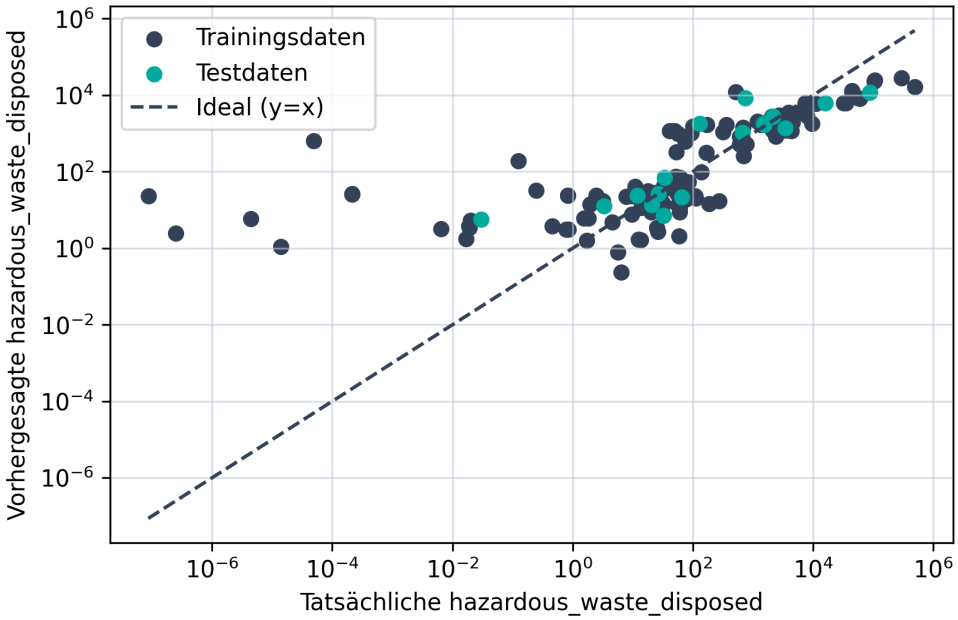


Abbildung 5.11: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Hazardous waste disposed*. Beide Achsen sind logarithmisch skaliert.

Tabelle 5.7: Gütekennzahlen der Regression für den Indikator *Water use* (Zielvariable auf der Skala $\log(1+$ water_use$)$) über $R = 100$ äußere Train/Test-Splits.

Größe	Mittelwert (Test) $\pm$ Std.	Bester Lauf (Test)
R^2_{Test}	0.603 ± 0.588	0.872
$\text{RMSE}_{\text{Test}}$	2.196 ± 0.753	1.431

und *Acidification* erreicht wurde.

Abbildung 5.12: QQ Plot der Schätzfehler des *Hazardous waste disposed* Modells auf der Transformationsskala.

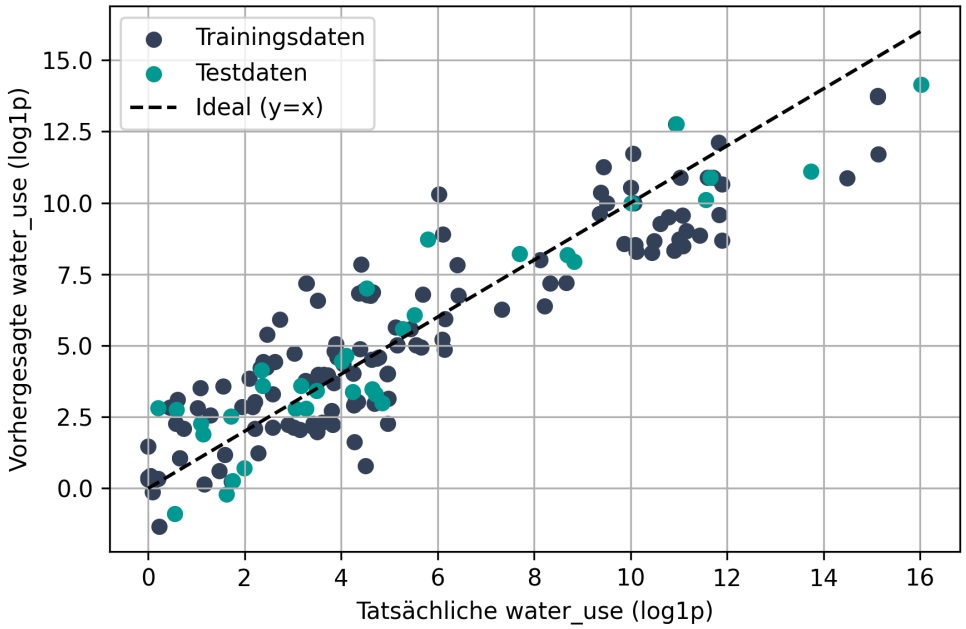


Abbildung 5.13: Vorhersagte gegenüber tatsächlichen Werten des Indikators *Water use* (log1p-Skala) für den besten Lauf der Regression.

Das Streudiagramm in Abbildung 5.13 zeigt, dass die meisten Schätzungen in der

Nähe der Diagonalen $y = x$ liegen und der Wasserverbrauch damit insgesamt moderat getroffen wird. Für kleine bis mittlere Werte (unter etwa 6 log1p-Einheiten) ist die Streuung relativ groß und es gibt sowohl Über- als auch Unterschätzungen. Im Bereich höherer Werte liegen die Werte dagegen enger an der Diagonalen, so dass besonders wasserintensive Produkte vergleichsweise zuverlässig vorhergesagt werden. Insgesamt ist die Modellgüte für *Water use* damit als moderat und deutlich weniger stabil und präzise einzuschätzen als für die besser erklärbaren Indikatoren.

5.4.4. Regression des Indikators Photochemical Ozone Formation (HH)

Für den Indikator *Photochemical ozone formation, human health* (`photochemical_ozone_formation_hh`) wird dasselbe Modell wie in Abschnitt 5.3 auf insgesamt $n = 171$ PEPs angewendet. Tabelle 5.8 zeigt die Testgüte über $R = 100$ äußere Train/Test-Splits.

Tabelle 5.8: Gütekennzahlen der Regression für den Indikator *Photochemical ozone formation, human health* (Zielvariable auf der Skala $\log(1 + \text{photochemical\_ozone\_formation\_hh})$) über $R = 100$ äußere Train/Test-Splits.

Größe	Mittelwert (Test) $\pm$ Std.	Bester Lauf (Test)
R^2_{Test}	0.626 ± 0.182	0.880
$\text{RMSE}_{\text{Test}}$	1.102 ± 0.211	0.717

Im Mittel werden damit rund 63 % der Varianz der log-transformierten Indikatorwerte erklärt. Die Streuung der Testgüte ist deutlich größer als bei *Climate change* und *Acidification*, aber ähnlich wie bei *Hazardous waste disposed*. Der beste Lauf erreicht ein Test-R^2 von 0,880 bei einem RMSE von 0,717 (log1p-Einheiten).

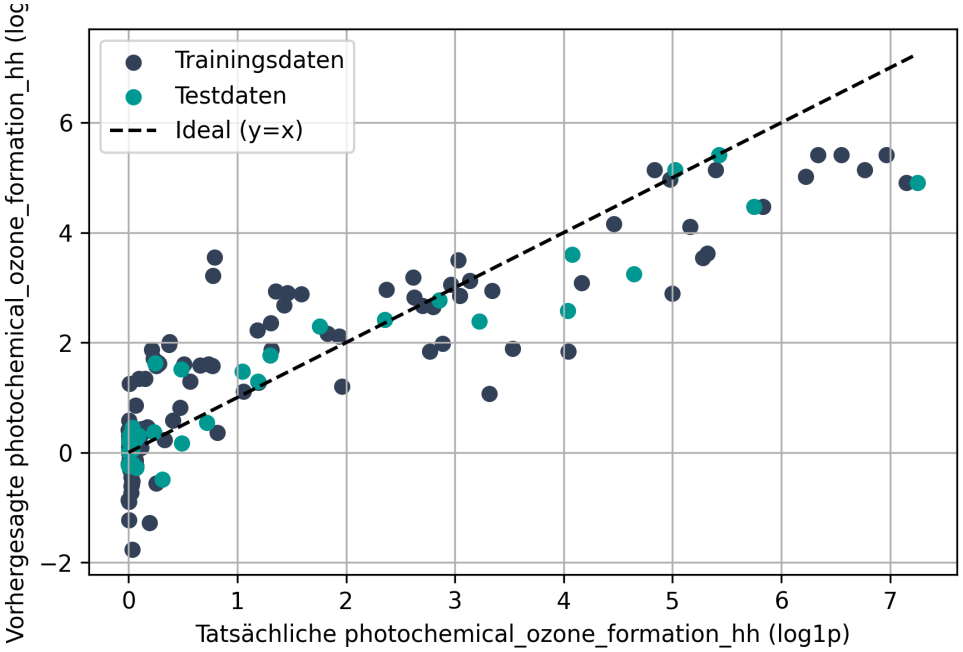


Abbildung 5.14: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Photochemical ozone formation, human health* (log1p-Skala) für den besten Lauf der Regression ($n = 171$ PEPs).

Das Streudiagramm in Abbildung 5.14 zeigt, dass für sehr kleine Werte (nahe 0 auf der $\log(1 + x)$-Skala) eine relativ hohe Streuung vorliegt und der Indikator ungefähr gleich häufig zu hoch wie zu niedrig geschätzt wird. Im mittleren Bereich (etwa $0{,}5$ bis 2 log1p-Einheiten) liegen die Punkte tendenziell oberhalb der Diagonalen und das Modell überschätzt die Werte leicht. Für größere Werte ab ungefähr 3 log1p-Einheiten liegen die Punkte überwiegend unterhalb der Diagonalen, so dass hohe Werte eher unterschätzt werden.

5.4.5. Regression des Indikators Resource Use, Fossils

Tabelle 5.9: Gütekennzahlen der Regression für den Indikator *Resource Use, Fossils* (Zielvariable auf der Skala $\log(1 + \text{photochemical\_ozone\_formation\_hh})$) über $R = 100$ äußere Train/Test-Splits.

Größe	Mittelwert (Test) $\pm$ Std.	Bester Lauf (Test)
R^2_{Test}	0.790 ± 0.145	0.940
$\text{RMSE}_{\text{Test}}$	1.669 ± 0.463	0.990

Für den Indikator *Resource use, fossils* erreicht das Modell über alle im Mittel ein

Test-R^2 von $0{,}790 \pm 0{,}145$ bei einem $\mathrm{RMSE}_{\text{Test}}$ von $1{,}669 \pm 0{,}463$ (log1p-Skala). Im besten Lauf werden sogar ein Test-R^2 von 0,940 und ein RMSE von 0,990 erreicht. Der fossile Ressourcenverbrauch wird ähnlich gut geschätzt wie *Acidification*.

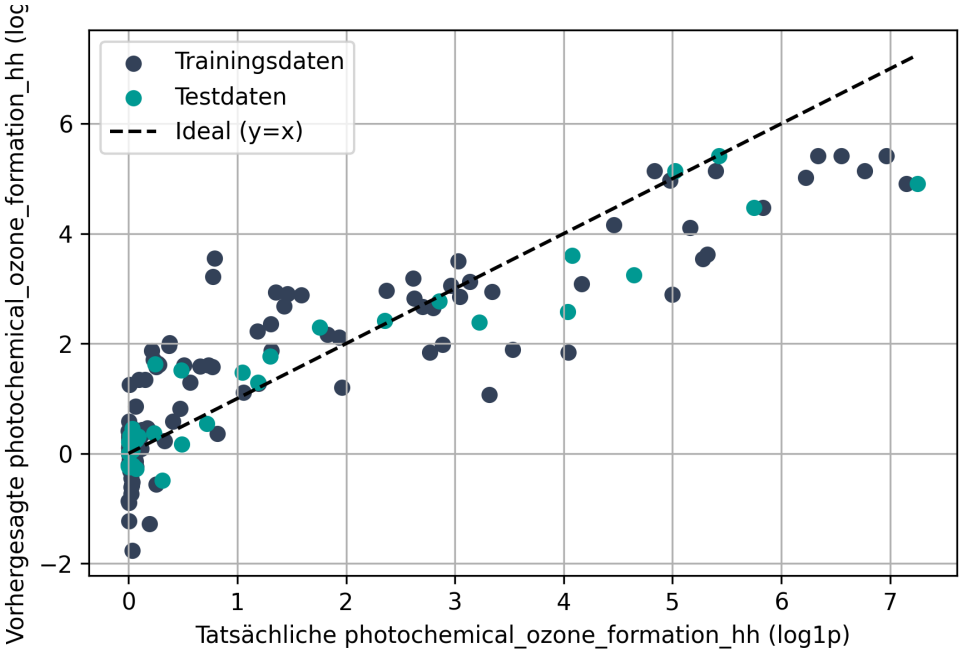


Abbildung 5.15: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Resource Use, fossils* (log1p-Skala) für den besten Lauf der Regression ($n = 171$ PEPs).

Das Streudiagramm in Abbildung 5.15 zeigt, dass die meisten Schätzungen dicht an der Diagonalen $y = x$ liegen. Kleine und mittlere Werte weisen lediglich eine moderate Streuung mit geringfügigen Über- und Unterschätzungen auf, während hohe Werte besonders gut getroffen werden. Eine deutliche systematische Abweichung ist hier nicht erkennbar.

5.4.6. Regression des Indikators Eutrophication (terrestrial)

Für den diesen Indikator wurde das Regressionsmodell wie zuvor auf $n = 107$ PEPs angewendet. Über $R = 100$ äußere Train/Test-Splits ergibt sich ein mittleres Test-R^2 von $0{,}773 \pm 0{,}154$ bei einem $\mathrm{RMSE}_{\text{Test}}$ von $1{,}155 \pm 0{,}312$ (log1p-Skala). Im besten Lauf werden ein Test-R^2 von $0{,}932$ und ein RMSE von $0{,}788$ erreicht und liegen damit auf einem ähnlich hohen Niveau wie bei *Resource use, fossils*.

Tabelle 5.10: Gütekennzahlen der Regression für den Indikator *Eutrophication, terrestrial* (Zielvariable auf der Skala $\log(1 + \text{photochemical\_ozone\_formation\_hh})$) über $R = 100$ äußere Train/Test-Splits.

Größe	Mittelwert (Test) ± Std.	Bester Lauf (Test)
R^2_{Test}	0.773 ± 0.154	0.932
$\text{RMSE}_{\text{Test}}$	1.155 ± 0.312	0.788

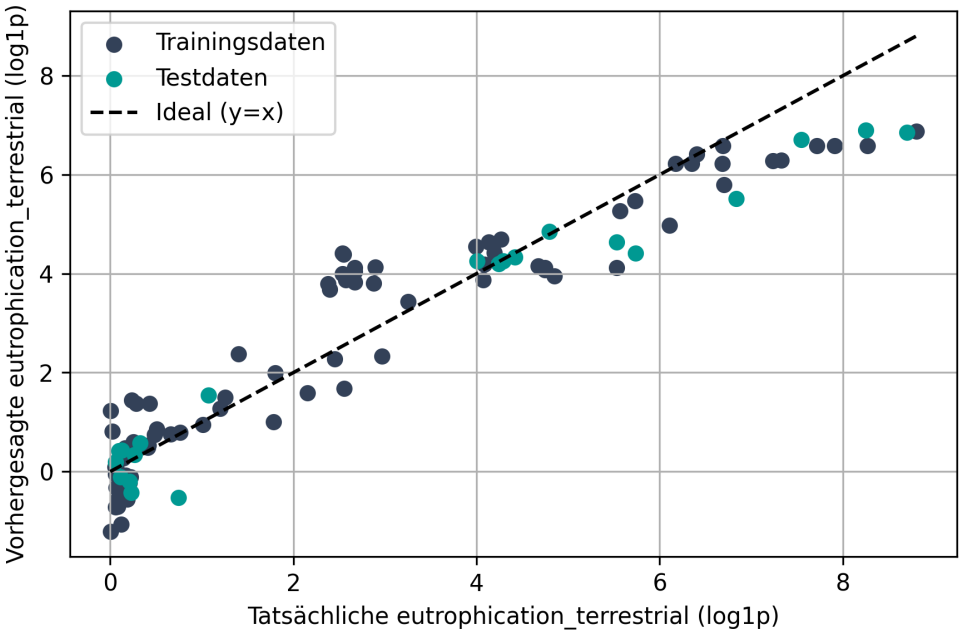


Abbildung 5.16: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Eutrophication (terrestrial)* (log1p-Skala) für den besten Lauf der Regression ($n = 171$ PEPs).

Das Streudiagramm in Abbildung 5.16 zeigt, dass der Großteil der Punkte nahe der Diagonalen $y = x$ liegt. Sehr kleine Werte (nahe $\log(1 + x) = 0$) streuen relativ stark, die absoluten Fehler bleiben dort jedoch gering. Im mittleren Bereich um 2–3 log-Einheiten treten etwas stärkere Über- und Unterschätzungen auf, während hohe Eutrophierungswerte ab etwa 4 log-Einheiten leicht unterschätzt werden.

5.4.7. 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 5.11 fasst diese Indikatoren zusammen.

!TODO: Die Tabelle passt Layout technisch noch nicht!

Tabelle 5.11: Indikatoren mit geringer Modellgüte ($R^2_{\text{Test, mean}} < 0{,}5$) in der log1p-Skala der Zielvariablen über $R = 100$ äußere Train/Test-Splits.

Indikator	R^2_{Test} (Mean ± Std.)	$\text{RMSE}_{\text{Test}}$ (Mean ± Std.)	Anzahl analysierter PEPs
Eutrophication (freshwater)	0.323 ± 0.276	0.618 ± 0.137	133
Eutrophication (marine)	0.239 ± 1.395	1.242 ± 0.552	107
Ozone depletion	-0.100 ± 1.070	0.008 ± 0.004	170
Radioactive waste disposed	-0.018 ± 2.604	1.263 ± 0.445	158
Resource use (minerals & metals)	0.401 ± 0.311	0.271 ± 0.057	175

Als besonders kritisches Beispiel zeigt Abbildung 5.17 den Indikator *Ozone depletion*. Obwohl die absoluten Fehler aufgrund der sehr kleinen Werte gering bleiben, liegen die Werte weit von der Diagonalen entfernt und die Streuung ist hoch. Entsprechend ist das durchschnittliche Test-R^2 sogar leicht negativ.

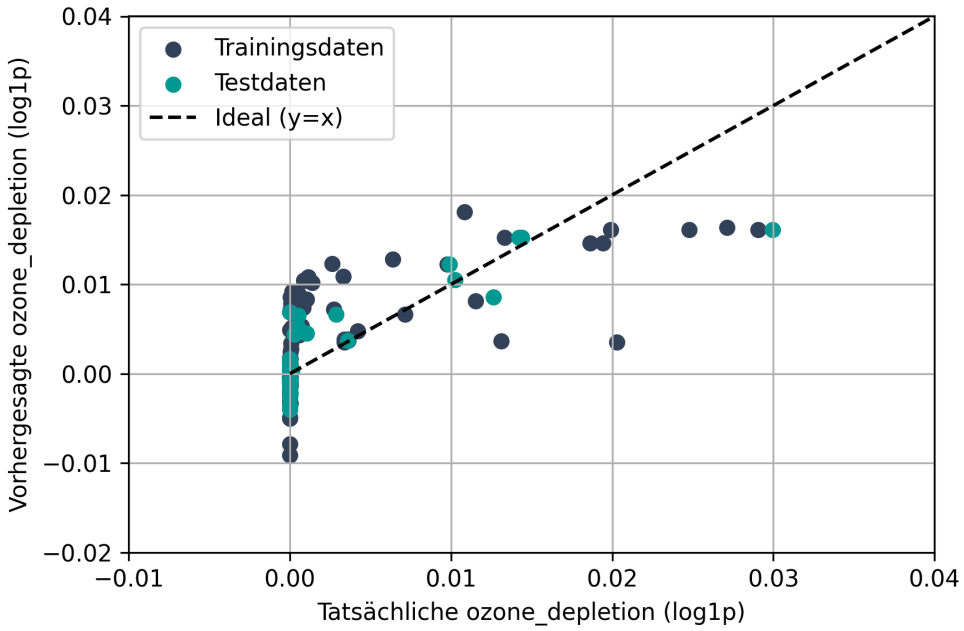


Abbildung 5.17: Vorhergesagte gegenüber tatsächlichen Werten des Indikators *Ozone depletion* (log1p-Skala) für den besten Lauf der Regression.

Ähnliche Muster zeigen sich bei weiteren Indikatoren. Weitere Streudiagramme für alle hier aufgeführten Indikatoren sind im Anhang **??** dargestellt.

! TODO: Evtl. Anhang für weitere Scatter (und Rohdaten?) !

6

Diskussion

Aus den in Kapitel 5.3 und 5.4 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 diskutiert die Ergebnisse und bereitet damit die spätere Ableitung einer einfachen Heuristik zur Abschätzung der Indikatoren vor.

6.1. Einordnung der CO2-Regressionsergebnisse

Die Diskussion bezieht sich auf das in Abschnitt 5.3 eingeführte Regressionsmodell, in dem die Zielvariable als log_cc $= \log(1 +$ cc_total$)$ definiert ist. Als Prädiktoren gehen das log-transformierte Produktgewicht (log_w), der Stromverbrauch (log_e) sowie die aus dem Materialblock abgeleiteten PCA-Hauptkomponenten ein. Die Ergebnisse basieren auf rund 173 PEPs und einer SGD-basierten linearen Regression mit mehrfach wiederholten äußeren Train/Test-Splits.

Interpretationsebene Die Kombination aus Gewicht, Stromverbrauch und Material-Hauptkomponenten erklärt mit 88% einen großen Anteil der Varianz der ausgewiesenen PEP-CO_2-Werte. Praktisch bedeutet dies, dass das Modell auf Basis weniger, relativ einfach zugänglicher Merkmale typischerweise die Größenordnung der CO_2-Äquivalente trifft. 12% der Varianz liegt damit noch außerhalb des Modells und kommt vermutlich aus verschiedenen Rechenmethoden innerhalb der PEPs und feineren Einflussfaktoren wie die Lebensdauer und das genaue benutzte Energiemodell. Der RMSE

liegt grob im Bereich eines Faktors zwei bis drei auf der Originalskala und erlaubt daher eher eine heuristische Abschätzung als eine exakte Bilanzierung einzelner Geräte.

- **Bereiche mit guter bzw. schwächerer Modellpassung**
  - Mittlere CO_2-Bereiche:
    - Im Bereich typischer CO_2-Werte liegen die Punkte im Streudiagramm dicht um die Diagonale, sowohl für Trainings- als auch für Testdaten. Es ist keine ausgeprägte systematische Über- oder Unterschätzung erkennbar.
    - Die Streuung der Testpunkte ist dabei sehr ähnlich zur Streuung der Trainingspunkte, was darauf hindeutet, dass das Modell den Zusammenhang in diesem Wertebereich gut erfasst und vernünftig generalisiert.
  - Extreme bzw. sehr große Produkte:
    - Einige wenige PEPs mit sehr hohem Gewicht (z. B. ≥ 720 kg) treten als deutliche Ausreißer auf und liegen sichtbar abseits der Diagonalen. Diese Beobachtungen tragen überproportional zur Gesamtsumme der Fehler bei und verzerren die Fit-Kennzahlen etwas.
    - Naheliegende Ursachen sind abweichende Systemgrenzen oder besondere Anwendungsfälle (z. B. Spezialanlagen), zusätzliche Prozessschritte wie aufwendige Installation oder Transporte sowie mögliche Inkonsistenzen in den zugrunde liegenden PEP-Daten.
    - Insgesamt spricht dies dafür, dass sich das Modell primär für typische Produkte der Stichprobe eignet und sehr große, außergewöhnliche Geräte nur eingeschränkt abbilden kann, weil sich deren Berechnung der Emission noch einmal deutlich vom Rest der PEPs unterscheidet.
  - Sehr kleine CO_2-Werte:
    - Im Bereich sehr kleiner Emissionswerte ist die relative Streuung auf der log1p-Skala zwar sichtbarer, die absoluten Abweichungen bleiben jedoch gering.
    - Für die Gesamtbewertung der Modellgüte sind diese Unterschiede daher weniger kritisch als Fehlerschätzungen im mittleren und hohen CO_2-Bereich, in dem der Großteil der betrachteten PEPs liegt und der Beitrag zum Gesamtemissionsniveau am größten ist.

6.2. Grenzen und Unsicherheiten des Modells

Die Aussagekraft der CO_2-Regression wird wesentlich durch die Qualität und Homogenität der zugrunde liegenden PEP-Daten bestimmt. Die PEPs selbst beruhen Hintergrunddatensätzen, Annahmen, unterschiedlichen Berechnungsmodellen und Systemgrenzen, womit sie keine ultimative Wahrheitsquelle darstellen können. 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 Muster erkennen kann und damit nicht nur für einen einzelnen Gerätetyp gültig ist. Gleichzeitig bedeutet die Produktvielfalt, dass für spezielle Nischenprodukte 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.

Auch methodisch ergeben sich Einschränkungen. Die verwendete `log1p`-Transformation stabilisiert die Regression, reduziert den Einfluss extremer Werte und führt zu einem weitgehend linearisierbaren Zusammenhang. Dadurch wird allerdings die direkte Interpretation der Fehler in kg CO_2-Äquivalenten weniger intuitiv, da Fehler auf der Log-Skala grob multiplikativen Abweichungen auf der Originalskala entsprechen. Zudem kann das lineare Modell auf der Log-Skala theoretisch negative Werte vorhersagen, die nach der Rücktransformation zu sehr kleinen, aber dennoch positiven Emissionen führen. In der praktischen Anwendung könnte daher eine Untergrenze der Vorhersagen sinnvoll sein.

Ein weiterer wichtiger Punkt ist die Indikatorabhängigkeit der Ergebnisse. Die höchste und stabilste Modellgüte wird für den Indikator *Climate change (total)* erreicht. Für einige andere Indikatoren wie *Acidification*, *Hazardous waste disposed*, *Water use*, *Photochemical ozone formation*, *Resource use (fossils)* oder *Eutrophication (terrestrial)* liefert die gleiche Pipeline zwar noch brauchbare Ergebnisse, die Test-R^2-Werte und die Streuung über die Splits sind jedoch schwächer als im CO_2-Fall. Für mehrere weitere Indikatoren mit geringer Modellgüte (vgl. Tab. 5.11) gelingt es kaum, wenn überhaupt einen moderaten Anteil der Varianz zu erklären. Hier scheinen die betrachteten Eingangsgrößen (Gewicht, Strom, Materialmix) nur einen begrenzten Einfluss zu haben oder die Datenlage ist zu verrauscht.

Insgesamt kann die CO_2-Regression daher als obere Schranke der erreichbaren

Genauigkeit unter den gegebenen Daten- und Featureeinschränkungen gelesen werden. Die im Mittel beobachtete Fehlerspanne bildet die Grundlage für die in Kapitel 7 abgeleitete Heuristik zur Abschätzung von CO_2-Äquivalenten ohne PEP. Auch in dieser Abschätzung kann man keine genauen Ergebnisse erwarten. Das Modell kann nur die Größenordnung zuverlässig einschätzen. Außerdem sollte diese Heuristik ausdrücklich als auf den Indikator *Climate change (total)* beschränkt verstanden werden. Eine direkte Übertragung auf andere Umweltindikatoren ist aufgrund der beschriebenen Unterschiede in der Modellgüte nur eingeschränkt möglich.

Literatur

[Jol82] Ian T. Jolliffe. “A Note on the Use of Principal Components in Regression”. In: (1982). DOI: 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. 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. 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. 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. 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. URL: https://www.sciencedirect.com/science/article/pii/S1078817410000027.

[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. 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. URL: https://dl.acm.org/doi/abs/10.1145/2467696.2467753.

[Gri15] Ralph Grishman. “Information Extraction”. In: (2015). DOI: 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. 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. URL: https://ieeexplore.ieee.org/abstract/document/7991564.

[CZ17] Andreiwid Sheffer Corrêa und Pär-Ola Zander. “Unleashing Tabular Content to Open Data: A Survey on PDF Table Extraction Methods and Tools”. In: (2017). DOI: 10.1145/3085228.30852. 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. URL: https://journals.lww.com/ijam/fulltext/2018/04010/Descriptive_statistics.7.aspx.

[Sel18] Howard J. Seltman. “Experimental Design and Analysis”. In: (2018). DOI: 10560/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. URL: https://ieeexplore.ieee.org/document/9002177.

[ARC21] Anthony C. Atkinson, Marco Riani und Aldo Corbellini. “The Box–Cox Transformation: Review and Extensions”. In: (2021). DOI: 10.1214/20-STS778. URL: https://projecteuclid.org/journals/statistical-science/volume-36/issue-2/The-BoxCox-Transformation-Review-and-Extensions/10.1214/20-STS778.full.

[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. 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. 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. 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. URL: https://arxiv.org/abs/2501.17887.

[25a] *Layout-Parser Dokumentation*. Dokumentation der Layout-Parser Bibliothek. 2025. URL: https://layout-parser.github.io/ (besucht am 23. 11. 2025).

[25b] *Matlab PCA Dokumentation*. Dokumentation der Matlab pca Bibliothek. 2025. URL: https://de.mathworks.com/help/stats/pca.html (besucht am 22. 11. 2025).

[Mor+25] José Teófilo Moreira-Filho u. a. “Automating Data Extraction From Scientific Literature and General PDF Files Using Large Language Models and KNIME: An Application in Toxicology”. In: (2025). DOI: 10.1002/wcms.70047. 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-kJ3qnsF1xLwSIzTYOv-ZEkCqc/mbesqrsCBZbWbKJq6-kJ3lQmBuGvAHsLUfQU9idjOpk (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] *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. URL: https://dl.acm.org/doi/full/10.1145/3696474.3696731.