h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Hochschule Darmstadt

– Fachbereich Informatik –

Quantifizierung des Rezyklatgehaltes in Kunststoffteilen für den Automobilbau mit Hilfe von Machine Learning-Methoden

Abschlussarbeit zur Erlangung des akademischen Grades

Bachelor of Science (B.Sc.)

vorgelegt von

Dennis Imhof

Matrikelnummer: 715862

Referent	:	Prof. Dr. Markus Döhring
Korreferentin	:	Prof. Dr. Inge Schestag

ERKLÄRUNG

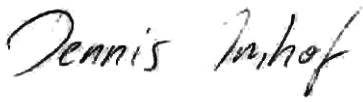
Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Alle referenzierten Internetquellen wurden zuletzt am 13. August 2020 um 19:00 Uhr eingesehen und auf Erreichbarkeit überprüft.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Darmstadt, 13. August 2020



Dennis Imhof

ZUSAMMENFASSUNG

Nachhaltigkeit ist eine Maxime, der gesellschaftlich, politisch und wirtschaftlich eine stetig wachsende Bedeutung zukommt. Auch an kunststofferzeugende und -verarbeitende Industrien werden damit zusätzliche, ökologische Anforderungen gestellt. Dies manifestiert sich unter anderem in einer durch die Europäische Union verabschiedeten Richtlinie, die für Einwegplastikerzeugnisse einen Mindestgehalt von 25% ab 2025 und 30% ab 2030 an recyceltem Kunststoff, auch Rezyklat, vorsieht.[1]

Es ergibt sich unmittelbar ein Bedarf für eine Methode mit der sich der Rezyklatgehalt eines industriellen Kunststofferzeugnisses möglichst effizient analysieren lässt. Dies dient der rechtlichen Absicherung und Compliance mit künftigen ökologischen Normen und Richtlinien einerseits sowie einer Verbesserung der Qualitätskontrolle und der Validierung von Stoffströmen andererseits. Eine solche Methode zur Ermittlung des Rezyklatgehaltes wird in diesem Projekt in Zusammenarbeit mit einem großen, internationalen Automobilproduzenten im Rahmen einer Machbarkeitsstudie entwickelt. Es sollen interdisziplinär Kunststoff-Analytik, Chemometrie und Machine Learning für eine industrielle Anwendung kombiniert werden. Im ersten Schritt werden mithilfe chromatographischer, spektroskopischer, mikroskopischer und nasschemischer Laborverfahren Messdaten vorliegender Kunststoffmuster mit unterschiedlichem Rezyklatgehalt erzeugt. Diese teils hochdimensionalen Daten werden anschließend computergestützt quantitativ ausgewertet. Es wird dabei zunächst überprüft, ob auf Basis der Datengrundlage von insgesamt 16 individuellen Kunststoffproben Modelle gebildet werden können, die sich zu einer Vorhersage des Rezyklatgehaltes eignen. Gleichzeitig wird versucht die hierfür minimal notwendige Datengrundlage und damit minimale Anzahl chemisch-physikalischer Messverfahren zu identifizieren.

Der Fokus dieser Arbeit richtet sich auf den computergestützten Teil der Untersuchung. Dieser setzt sich aus der Verarbeitung der Messdaten im Rahmen von Extract, Transform, Load (ETL)-Prozessen, der qualitativen Untersuchung und Herausarbeitung aussagekräftiger Charakteristiken der Daten im Rahmen *explorativer Datenanalyse* und *Feature Engineerings* sowie der quantitativen, statistischen Analyse des Rezyklatgehalts der Proben zusammen. Dieses Vorgehen wird zunächst separat mit den Messdaten der einzelnen chemisch-physikalischen Messmethoden durchgeführt und darauf aufbauend in einer kombinierten Analyse abgeschlossen. Dabei wird eine Teilmenge der vorliegenden Daten zur Modellbildung als Trainingsdatensatz herangezogen und die entwickelten Modelle mithilfe der verbleibenden Testdaten

1 https://data.consilium.europa.eu/doc/document/PE-11-2019-REV-1/de/pdf (Artikel 6)

validiert.

Ziel der Arbeit ist die Ermittelung einer minimalen Datengrundlage und damit die minimale Anzahl chemisch-physikalischer Messmethoden, die zur Vorhersage des Rezyklatgehaltes der vorliegenden Kunststoffmuster notwendig ist. Es werden deshalb verschiedene *Feature Selection*-Methoden angewendet, um statistisch signifikante Features zu identifizieren und redundante Features zu eliminieren. Zum Vergleich und der Auswahl der besten Modelle auf Basis der Trainingsdaten wird das Bestimmtheitsmaß R^2 herangezogen. Die ermittelten Modelle sind in der Lage den Rezyklatgehalt der Testdaten mit einer maximalen Abweichung von $\pm 3\%$ Genauigkeit vorherzusagen. Weiterhin kann gezeigt werden, dass zu einer Bestimmung des Rezyklatgehaltes bereits ein chemisch-physikalisches Messverfahren ausreichend ist, wobei mit den Daten der Infrarotspektroskopie sowie der Hochleistungsflüssigkeitschromatographie die besten Ergebnisse erzielt werden können.

Schlüsselbegriffe: Chemometrie, Recycling, Preprocessing, Quantifizierung

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

ATR Attenuated Total Reflection

ATR-IR Abgeschwächte Totalreflexion-Infrarotspektroskopie

CSV Comma Separated Values

E/P Ethylen-Propylen-Copolymer

ELSD Lichtstreuungsdetektor

ETL Extract, Transform, Load

GPC Gel-Permations-Chromatographie

HDPE High-Density Polyethylen

HPLC Hochleistungsflüssigkeitschromatographie

ICA Independent Component Analysis

IR Infrarot

LASSO Least Absolute Shrinkage and Selection Operator

MSE Mean Squared Error

NMF Non-negative Matrix Factorization

OLS Ordinary Least Squares

PCA Principal Component Analysis

PE Polyethylen

PP Polypropylen

PUR Polyurethan

PVC Polyvinylchlorid

RMSE Root Mean Squared Error

SEC Größenausschluß-Chromatographie

SQR Residuenquadratsumme

SQT Totale Quadratsumme

1 EINLEITUNG

Das Recycling von Kunststofferzeugnissen stellt einen wichtigen Beitrag im Sinne einer nachhaltigen Wirtschaft dar. Die Wiederverwertung von Kunststoffen kann stofflich, chemisch und thermisch geschehen. Erstrebenswert ist hierbei vorallem die stoffliche Wiederverwertung, bei der das Material ohne eine Änderung seiner chemischen Eigenschaften in den Materialkreislauf zurückgeführt und durch mechanische Aufbereitung in eine verarbeitbare Form gebracht wird. Diese Form der Aufbereitung wird auch physikalisches Recycling genannt und beinhaltet das Schreddern, Mahlen und Filtern des gebrauchten Materials (Gruden [14], S.307). Das so als Granulat erhaltene Rezyklat kann dem Verarbeitungsprozess bei ausreichender Reinheit wieder zugeführt werden.

Kunststoffrecycling nimmt auch für die Automobilindustrie eine bedeutende Rolle ein. Das moderne Auto besteht zu einem stetig wachsenden Anteil aus Kunststoffen, die verschiedenste Anforderungsprofile abdecken. Kunststoffe weisen, gemessen an metallischen Äquivalenten, ein deutlich geringeres Gewicht auf und lassen sich durch Extrusions- und Spritzgussverfahren zu äußerst komplexen, ein- oder mehrteiligen Bauteilen formen. In Abbildung 1.1 ist die durchschnittliche Entwicklung des Massenanteils von Kunststoffen und Metallbauteilen in Personenkraftwägen über die Jahre 1995, 2000 und 2017 dargestellt.

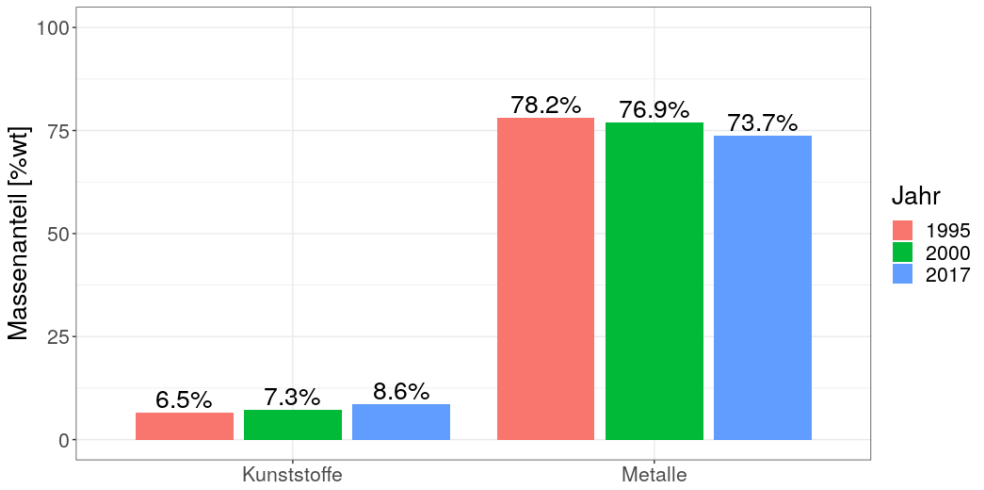


Abbildung 1.1: Entwicklung des durchschnittlichen Massenanteils von Kunststoffen und Metallen in Personenkraftwägen. (Datengrundlage Davis, Diegel und Boundy [9], S.4-26)

Kunststoffen wird zunehmend Funktionalität, die zuvor durch Metallerzeugnisse abgedeckt wurde, zuteil. Mit dieser Entwicklung gehen komplexe ökonomische und ökologische Fragestellungen einher. Die EG Altfahrzeugrichtlinie [1] und die deutsche Altfahrzeugverordnung [2] sehen seit 2015 eine Wiederverwendung oder Verwertung von 95 Gewichtsprozent des Leergewichts von Altfahrzeugen sowie eine Wiederverwendung oder *stoffliche* Verwertung von mindestens 85 Gewichtsprozent des Leergewichts von Altfahrzeugen vor. Mit einem steigendem Kunststoffanteil steht dieser auch bei Recycling betreffenden Fragestellungen immer mehr im Fokus.

Bereits zwei Drittel des in einem PKW verarbeiteten Kunststoffes werden durch drei Kunststoffsorten abgedeckt. Diese sind Polypropylen (PP) (32%), Polyurethan (PUR) (17%) und Polyvinylchlorid (PVC) (16%) (Biron [7], S.886). Polypropylen, das den mit Abstand höchsten Gewichtsanteil der Kunststoffe im Automobilbau einnimmt, wird unter anderem für größere Bauteile wie Armaturenbretter, Scheinwerfergehäuse und Stoßstangenverkleidungen verwendet. Der hohe Gewichtsanteil und die Größe einiger durch PP angefertigter Bauteile machen das Material für Recyclinganwendungen attraktiv. Die Wiederverwertung des Materials ist jedoch mit einigen Hürden verbunden. PP wird häufig im Verbund mit anderen Materialien verwendet, was die Rückgewinnung des Reinstoffs deutlich erschwert. Weiterhin zeigen Untersuchungen von neuwertigem und rezykliertem Polypropylen eine Abnahme der mechanischen Belastungsfähigkeit des Materials mit steigendem Rezyklatgehalt (Bahlouli u. a. [5]).

1.1 MOTIVATION

Recycling von Kunststoffen im Allgemeinen und speziell das Recycling von Polypropylen in der Automobilindustrie hat maßgebliche Auswirkungen auf Designentscheidungen innerhalb des Herstellungsprozesses von Fahrzeugen. Es resultieren hieraus konkrete Fragestellungen nach den Eigenschaften und der Qualität rezyklathaltiger Materialien, was wiederum Methoden zur Kategorisierung und Quantifizierung dieser bedingt. Es sind einerseits Methoden erforderlich, mit denen sich physikalische und chemische Eigenschaften von mit Rezyklat versetztem Material bekannter Konzentration bestimmen und somit gegenüber dem Rohmaterial vergleichen lassen. Hierzu finden sich Ergebnisse hinsichtlich der mechanischen Belastungseigenschaften von rezyklathaltigem PP in den Arbeiten von Bahlouli u. a. [5] und Ragosta u. a. [28].

Andererseits sind bei einer zunehmenden (Wieder-)Verwendung rezyklathaltiger Kunststoffe im Fertigungsprozess Methoden notwendig, mit denen möglichst effizient, schnell und zuverlässig der Anteil des enthaltenen Re-

1 https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:02000L0053-20200306&from=DE, Artikel 7.
2 http://www.gesetze-im-internet.de/altautov/AltfahrzeugV.pdf, § 5.

zyklats vorliegender Bauteile quantitativ bestimmt werden kann. Dies kann sich darin begründen, dass der Rezyklatanteil und dessen Zusammensetzung unbekannt ist oder die vorhandenen Angaben hierzu überprüft werden sollen. Eine Ermittlung des Rezyklatgehaltes ermöglicht Aussagen über die stofflichen Eigenschaften des vorliegenden Materials und stellt ein wichtiges Instrument zur Einhaltung qualitativer Standards sowie ökologischer Richtlinien und politischer Verordnungen dar.

Ein effizientes Verfahren zur Vorhersage des Rezyklatgehaltes in Kunststoffen sollte den Rezyklatgehalt unbekannter Proben auf Basis einer minimalen Anzahl chemisch-physischer Messmethoden bestimmen können, um industriellen Zeit- und Kostenansprüchen zu genügen. Konkret ergibt sich das Untersuchungsziel dieser Arbeit wie folgt:

> Ziel dieser Arbeit ist die Untersuchung der Quantifizierbarkeit des Rezyklatgehalts vorliegender Polypropylenmuster und dabei im Speziellen die Identifizierung einer minimalen hierzu notwendigen Datengrundlage.

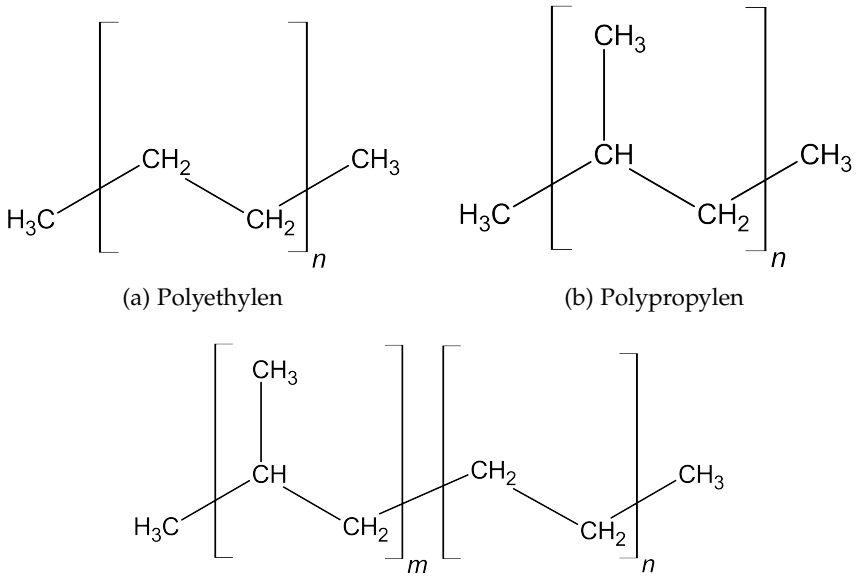
1.2 DURCHFÜHRUNG

Insgesamt liegen 16 nach DIN ISO 3167 Typ 1A zu Zugstäben verarbeitete Muster an Polypropylen zur Untersuchung vor. Vier der vorliegenden Zugstäbe sind in Abbildung 1.2 dargestellt.

Abbildung 1.2: Vier Polypropylen-Zugstäbe mit Rezyklatgehalt 0%

Die Polypropylenmuster weisen zu Gruppen von je vier Proben Rezyklatgehalte von 0%, 30%, 50% und 100% auf. Die exakte Zusammensetzung des Rezyklates ist nicht bekannt. Erwartet wird eine Zusammensetzung aus aufbe-

reitetem Polypropylen sowie eventueller Verunreinigungen durch Polyethylen (PE) und Ethylen-Propylen-Copolymer. Die Strukturformeln von PE, PP und Ethylen-Propylen-Copolymer sind in Abbildung 1.3 dargestellt.



(a) Polyethylen

(b) Polypropylen

(c) Ethylen-Propylen-Copolymer

Abbildung 1.3: Strukturformeln von Polyethylen, Polypropylen und Ethylen-Propylen-Copolymer

Es werden drei unterschiedliche, chemisch-physikalische Messmethoden verwendet, um eine möglichst umfassende Datengrundlage für die anschließende, computergestützte Analyse der Proben zu generieren. Die verwendeten Messmethoden sind:

- Gel-Permations-Chromatographie (GPC)
- Hochleistungsflüssigkeitschromatographie (HPLC)
- Abgeschwächte Totalreflexion-Infrarotspektroskopie (ATR-IR)

Die generierten Messdaten liegen in unterschiedlichen Dateiformaten vor und werden im Rahmen des *Data Loadings* in eine verarbeitbare Datenstruktur überführt. Hierzu wird das R-Softwarepaket *hyperSpec* (Beleites und Sergo [6]) verwendet. Abhängig von der zugrundeliegenden Messmethode werden verschiedene *Preprocessing*-Techniken angewendet, um Störfaktoren zu vermindern und eine bessere Vergleichbarkeit der Spektraldaten und Chromatogramme zu erreichen. Das anschließende *Feature Engineering* dient der Reduktion der hochdimensionalen Daten und der Extraktion von *Features*, die einen Zusammenhang mit dem Rezyklatgehalt der Proben aufweisen. Hierzu werden konventionelle Methoden wie die Berechnung von Signalmaxima und Signalmittelwerten sowie die nicht-negative Matrixfaktorisierung der Daten herangezogen. Die erzeugten Features finden als unabhängige Variablen in Regressionsmodellen zur quantitativen Bestimmung des

Rezyklatgehaltes der Proben Anwendung. Mithilfe von teilautomatisierten *Feature Selection*-Methoden werden die informativsten unabhängigen Variablen identifiziert und damit eine minimale Anzahl notwendiger Variablen zur Vorhersage des Rezyklatgehaltes bestimmt.

Die Messdaten der individuellen Methoden werden zunächst isoliert analysiert. Mithilfe von *Feature Selection* werden die am besten zur Vorhersage des Rezyklatgehaltes geeigneten Features jeder Einzelanalyse identifiziert. Im Anschluß werden diese in einer kombinierten Analyse gemeinsam ausgewertet. Hierbei wird untersucht, welche der verwendeten chemisch-physikalischen Messmethoden sich am besten für eine Vorhersage des Rezyklatgehaltes eignen. Es kommen dabei erneut die Methoden der *Feature Selection* zum Einsatz, um eine minimale Anzahl notwendiger Variablen und damit gleichzeitig eine minimale Anzahl notwendiger chemisch-physikalischer Messmethoden zu bestimmen.

In Abbildung 1.4 ist der vollständige Prozess der Einzelanalysen in Form eines Flowcharts dargestellt. Dabei sind die von der jeweiligen Messmethode abhängigen Verarbeitungpfade in unterschiedlichen Farben dargestellt und beginnen jeweils bei den vorliegenden Rohdaten der jeweiligen Messmethode, die oben links in der Abbildung dargestellt sind.

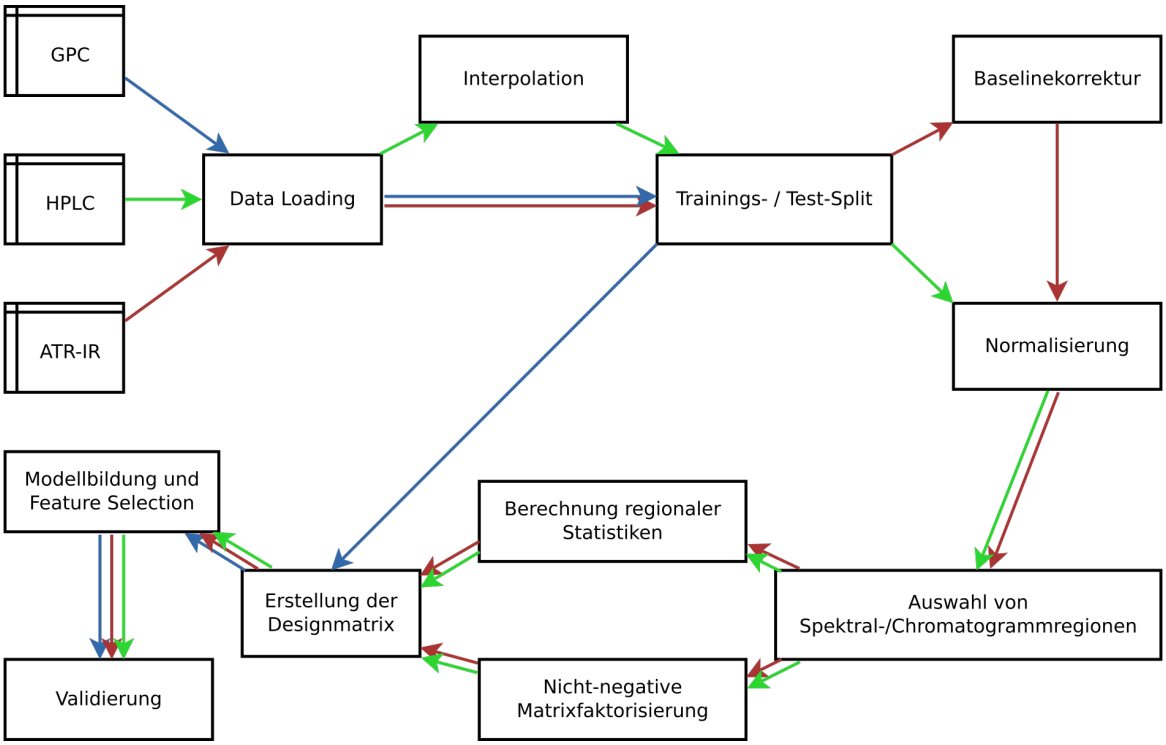


Abbildung 1.4: Darstellung des Analyseprozesses als Flowchart. Data Loading und Preprocessing in der oberen Hälfte. Feature Engineering und quantitative Analyse in der unteren Hälfte. Analysepfade von GPC, HPLC und ATR-IR dargestellt in unterschiedlichen Farben mit Ausgangspunkt oben links.

Weiterhin sind die Verarbeitungsschritte in der oberen Hälfte der Darstellung dem *Preprocessing* der Daten und die Verarbeitungsschritte der unteren Hälfte der Darstellung dem *Feature Engineering* und der quantitativen Analyse zuzuordnen.

1.3 AUFBAU DER ARBEIT

Auf das in den Untersuchungsgegenstand einführende Kapitel 1 folgend, werden in Kapitel 2 die theoretischen Grundlagen der verwendeten Messverfahren und angewendeten Algorithmen dargestellt. Hierbei wird zunächst auf die Funktionsweise der verwendeten chemisch-physikalischen Messmethoden eingegangen. Anschließend werden die mathematischen Grundlagen verschiedener, im Analyseprozess zur Anwendung kommender Vorverarbeitungsmethoden und Algorithmen vorgestellt.

In den sich anschließenden beiden Kapiteln zur Datenvorverarbeitung (Kapitel 3) und quantitativen Analyse (Kapitel 4) werden die Arbeitsschritte des vollständigen, in Abbildung 1.4 dargestellten Analysesprozesses vorgestellt. Dabei wird in Kapitel 3 zunächst die Vorverarbeitung der Rohdaten besprochen, was durch die Vorstellung der vorliegenden Datenformate und Einlesetechniken in Abschnitt 3.1 eingeleitet wird. In den sich anschließenden Abschnitten werden die im Rahmen der Datenvorverarbeitung angewendeten Transformationstechniken vorgestellt, die sich aus der *Interpolation von Spektraldaten und Chromatogrammen* (Abschnitt 3.2), *Korrektur des Untergrundes* (Abschnitt 3.3), *Spektralbereicheingrenzung* (Abschnitt 3.4), *Spektraltransformation* (Abschnitt 3.5) und *Feature Engineering* (Abschnitt 3.6) zusammensetzen.

In Kapitel 4 wird die Anwendung der *multiplen linearen Regression* zur quantitativen Analyse sowie die Methoden der *stufenweisen Regression* (Unterabschnitt 4.2.1) und *LASSO*-Regression (Unterabschnitt 4.2.2) zur Auswahl optimaler Featurekombinationen vorgestellt.

Im Experimentalteil (Kapitel 5) werden die Ergebnisse der Einzelanalysen, die auf den Daten der chemisch-physikalischen Messmethoden der ATR-IR (Abschnitt 5.1), GPC (Abschnitt 5.2) und HPLC (Abschnitt 5.3) basieren sowie das Ergebnis der kombinierten Analyse (Abschnitt 5.4) vorgestellt. In jeder der individuellen Analysen kann dabei ein linearer Zusammenhang zwischen dem Rezyklatgehalt der Proben und der jeweiligen Datengrundlage festgestellt werden. Eine quantitative Vorhersage des Rezyklatgehaltes des Testdatensatzes kann mithilfe jener auf Basis der Trainingsdaten erstellten Modelle mit einer maximalen Abweichung von $\pm 8,5\%$ durchgeführt werden. In der kombinierten Analyse wird ein ausschließlich auf den Daten der ATR-IR basierendes, lineares Modell mit einem maximalen Vorhersagefehler von $\pm 3\%$ als bestes Modell identifiziert. Ein weiteres Modell, das ausschließlich auf Daten der HPLC basiert, weist einen maximalen Vorhersagefehler

von $\pm 5\%$ auf. Das Ziel dieser Arbeit, eine minimale Datengrundlage zur Vorhersage des Rezyklatgehaltes der vorliegenden Polypropylenmuster zu bestimmen, kann erfolgreich abgeschlossen werden, wobei sich durch das ATR-IR sowie das HPLC Verfahren erzeugte Daten hierzu eignen.

In Kapitel 6 erfolgt eine ausführliche Diskussion der Ergebnisse und potenzieller, an den Ergebnissen dieser Arbeit anknüpfender Untersuchungsgegenstände.

2 GRUNDLAGEN

Zur Erzeugung der im Rahmen dieser Arbeit verarbeiteten Daten wurden mehrere nass-chemische und spektroskopische Methoden verwendet. Die angewandten Messmethoden nutzen dabei unterschiedliche Materialeigenschaften wie das Molekulargewicht, die Polarität oder die Kristallinität des Materials aus. Die theoretischen Grundlagen der chemisch-physikalischen Messmethoden werden in diesem Kapitel nur stark verkürzt dargestellt. Für eine ausführliche Abhandlung der Funktionsweise der chromatographischen Messmethoden sei auf Harris [15] (S.534ff) verwiesen.

Weiterhin finden in dieser Arbeit verschiedene statistische Methoden und Algorithmen Anwendung. Im Rahmen des Preprocessings wird das *Asymmetric Least Squares Smoothing*-Verfahren zur Untergrundkorrektur der Spektraldaten und das Non-negative Matrix Factorization (NMF)-Verfahren zur Dimensionsreduktion herangezogen. Zur quantitativen Analyse der transformierten Daten dienen multiple lineare Regressionsmodelle. In den folgenden Abschnitten werden zunächst die verwendeten Messmethoden und anschließend die mathematischen Grundlagen der herangezogenen Algorithmen erläutert.

2.1 GEL-PERMEATIONS-CHROMATOGRAPHIE

GPC oder auch Größenausschluß-Chromatographie (SEC) ist eine chromatographische Methode, die die Größe, genauer das hydroskopische Volumen, der Stoffkomponenten der zu analysierenden Probe ausnutzt, um diese zu trennen. Das gelöste Probenmaterial fließt als mobile Phase mit konstanter Flussgeschwindigkeit durch eine mit feinporigem Gel, der stationären Phase, ausgestattete Säule. Kleinere Moleküle durchdringen die Poren des Gels und halten sich dadurch länger in der stationären Phase auf. Größere Moleküle können nicht in die Poren eindringen und eluieren dadurch schneller. In festen Volumen- respektive Zeitintervallen wird das aus der Säule austretende Eluat mithilfe eines Detektors (Infrarotdetektor oder Lichtstreuungsdetektor) spektroskopisch ausgewertet.

Das aus einer Messung resultierende *Chromatogramm* beinhaltet die Auftragung des kumulativen Detektorsignals der einzelnen Messungen gegen den Elutionszeitpunkt, der aus Elutionsvolumen und der konstanten Flussgeschwindigkeit berechnet wird. Eine erfolgreiche Trennung der in der Probe enthaltenen Komponenten lässt sich an zeitlich sauber getrennten Peaks im Chromatogramm erkennen. Das Zeitintervall zwischen Injektion der Probe und dem Maxium eines Komponentenpeaks wird Retentionszeit der Kom-

ponente genannt. Ein beispielhaftes Chromatogramm basierend auf synthetischen Daten ist in Abbildung 2.1 dargestellt.

Die einheitenlose, kumulierte Intensität des Detektors (a.u. für *arbitrary unit*) wird gegen die Elutionszeit aufgetragen, dh. den Austrittszeitpunkt des jeweiligen Teils der gelösten Probe (Eluat) aus der Chromatographiesäule.

Es sind drei Chromatogramme abgebildet, die jeweils drei Signalpeaks, die zu Alkanen, einfache Kohlenstoffverbindngen, unterschiedlicher Molekülgröße korrespondieren, enthalten. Es ist eine steigende Retentionszeit bei abnehmender Molekülgröße der Alkane zu erkennen, wobei Octan das größte Molekül darstellt und Pentan das kleinste Molekül darstellt.

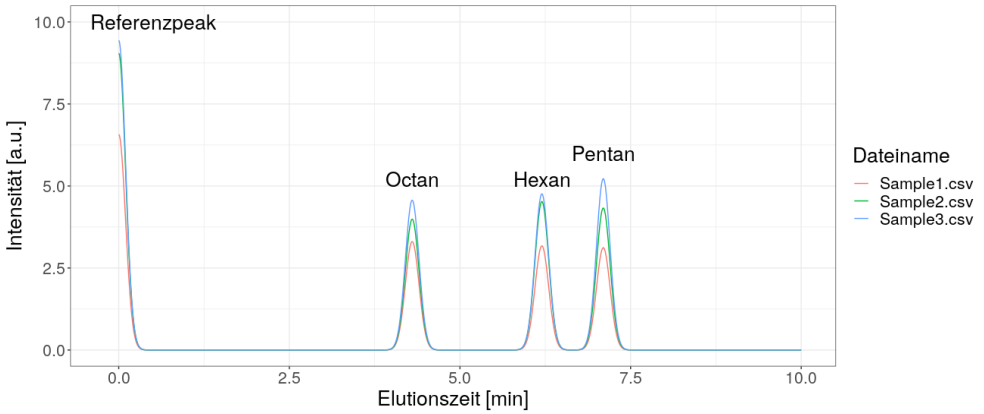


Abbildung 2.1: Chromatogramm basierend auf synthetischen Daten

2.2 HOCHLEISTUNGSFLÜSSIGKEITSCHROMATOGRAPHIE

HPLC ist wie auch die GPC ein chromatographisches Verfahren. Im Gegensatz zur GPC nutzt die HPLC jedoch chemische Eigenschaften der gelösten Stoffkomponenten aus, um eine Trennung dieser zu erreichen. Als stationäre Phase der Trennsäule wird dabei ein polares Packmaterial und als mobile Phase ein unpolares Lösungsmittel eingesetzt. Polar bedeutet hierbei, dass die Elektronen des Moleküls ungleichmässig verteilt und damit ortsabhängig eine stärker positive oder negative Ladungsverteilung auftritt.

Die Trennung der gelösten Komponenten resultiert aus der unterschiedlich starken Adhäsion der jeweiligen Komponenten an der stationären Phase. Polare Moleküle verbleiben aufgrund der höheren Adhäsion länger in der mit polarem Packmaterial gefüllten Trennsäule. Unpolare Molekühle eluieren hingegen schneller. Das Eluat wird wie bei der GPC in festen Elutionsintervallen respektive Zeitabständen mit einem Detektor ausgewertet, woraus das Chromatogramm der Messung resultiert.

2.3 ABGESCHWÄCHTE TOTALREFLEXION-INFRAROTSPEKTROSKOPIE

ATR-IR ist ein spezieller Typ der Infrarotspektroskopie. Die zu untersuchende Probe wird auf einem reflektiven Kristall angebracht und dieser mit Licht im Infrarotbereich durchstrahlt. Teile des Lichts werden vom Probenmaterial absorbiert. Das nicht-absorbierte Licht wird an der Grenzfläche zwischen Kristall und Probenmaterial reflektiert und mit einem Detektor gemessen. Durch Absorption spezifischer Wellenlängen werden Schwingungen im Molekül angeregt. Wellenlängenbereiche, in denen eine Anregung stattfindet, kommen mit verringerter Intensität am Detektor an.

Ein Spektrogramm besteht aus der Auftragung der gemessenen Transmission (oder der hieraus berechneten Absorption) gegen die zugehörige Wellenzahl. Intensität und Wellenzahl der auftretenden Signalpeaks sind dabei charakteristisch für die Molekülstruktur des untersuchten Stoffes. Ein beispielhaftes Spektrogramm einer Infrarotmessung von Polypropylen ist in Abbildung 2.2 dargestellt.

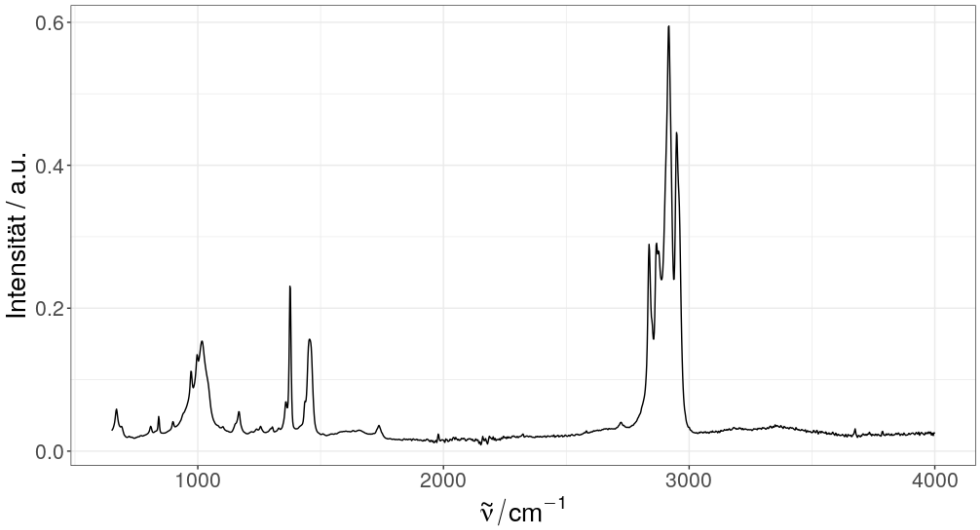


Abbildung 2.2: ATR-IR Spektrogramm von Polypropylen

Die multiplen Signalpeaks im Bereich von 2800 - 3000 cm^{-1} lassen sich beispielsweise auf Streckschwingungen der CH_2- und CH_3-Gruppen zurückführen (Asensio u. a. [3]).

2.4 MULTIPLE LINEARE REGRESSION

Lineare Regression stellt ein relativ einfaches und gut interpretierbares Verfahren zur Modellierung des Zusammenhangs zwischen einer *abhängigen Variablen* und einer oder mehrerer *unabhängiger Variablen* dar. In Gleichung 2.1 ist die mathematische Beschreibung eines linearen Modells mit einer *abhängigen Variablen* y, *Regressionskoeffizieten* β_k mit $k \in \{0 \dots p\}$, *unabhängigen Va-*

riablen x_j mit $j \in \{1 \dots p\}$ und Fehlerterm ϵ dargestellt. β_0 entspricht hierbei dem Achsenabschnitt.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{2.1}$$

Zur Vereinheitlichung der Indizes der *unabhängigen Variablen* x und Regressionskoeffizienten β kann eine Dummyvariable $x_0 = 1$ eingeführt werden.

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{2.2}$$

Dies ist vorallem bei der Darstellung des Modells in Matrixnotation äußerst nützlich. In Gleichungen 2.1 und 2.2 wurde der lineare Zusammenhang für ein Stichprobenelement dargestellt. Üblicherweise enthält die Stichprobenmenge mehr als ein Element. Liegt eine Stichprobe der Größe n vor, ergibt sich die mathematische Beschreibung des linearen Modells wie folgt.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}_{n\times(p+1)} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1} \tag{2.3}$$

Hierbei stellt die mit $\times$ gekennzeichnete Multiplikation eine Matrixmultiplikation dar. Mithilfe von Matrixnotation vereinfacht sich dies zur folgenden Gleichung.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.4}$$

Fettgedruckte Kleinbuchstaben stellen hierbei Spaltenvektoren und fettgedruckte Großbuchstaben stellen Matrizen dar.

2.4.1 *Methode der kleinsten Quadrate*

Ziel der linearen Regression ist es, optimale Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$ aus den vorliegenden Daten $\boldsymbol{X}$ und $\boldsymbol{y}$ zu bestimmen. $\hat{\boldsymbol{\beta}}$ sind dabei Punktschätzer der wahren, aber unbekannten Werte $\boldsymbol{\beta}$. Die in den Gleichungen (2.1) - (2.4) dargestellten Formulierungen stellen somit nur hypothetische Modelle dar. Aus der statistischen Schätzung der Regressionskoeffizienten folgt das in Gleichung 2.5 dargestellte, praktisch anwendbare Regressionsmodell.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{2.5}$$

Die Werte $\hat{\boldsymbol{y}}$ sind dabei die *gefitteten Werte*. Die Differenzen zwischen den wahren Werten $\boldsymbol{y}$ und den *gefitteten Werten* $\hat{\boldsymbol{y}}$ werden als Residuale $\hat{\boldsymbol{\epsilon}}$ bezeichnet.

$$\begin{aligned} \hat{\boldsymbol{\epsilon}} &= \boldsymbol{y} - \hat{\boldsymbol{y}} && (2.6) \\ &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} && (2.7) \end{aligned}$$

Mithilfe der Methode der kleinsten Quadrate wird die Summe des quadratischen Abstands zwischen den wahren Werten $\boldsymbol{y}$ und den *gefitteten Werten* $\hat{\boldsymbol{y}}$ und damit die Residuenquadratsumme (SQR) minimiert, wobei die Residuenquadratsumme folgendermaßen definiert ist.

$$\begin{aligned} \text{SQR} &= \sum_{i=1}^{n} \hat{\epsilon}_i^2 && (2.8) \\ &= \hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}} && (2.9) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) && (2.10) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} - (\boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}})^T + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} && (2.11) \end{aligned}$$

Die beiden mittleren Ausdrücke der Gleichung 2.11 sind Skalare und somit identisch, weshalb diese zusammengefasst werden können. Es ergibt sich folgender Ausdruck für die Residuenquadratsumme.

$$\text{SQR} = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \quad (2.12)$$

Leitet man die Residuenquadratsumme nach $\hat{\boldsymbol{\beta}}$ ab und setzt die Ableitung gleich Null, können damit die Regressionsparameter $\hat{\boldsymbol{\beta}}$ bestimmt werden, die die Residuenquadratsumme minimieren.

$$\frac{\partial \text{SQR}}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0 \quad (2.13)$$

Durch Umformung erhält man die folgende Gleichung, die auch Normalengleichung genannt wird.

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \quad (2.14)$$

Löst man die Gleichung nach $\hat{\boldsymbol{\beta}}$ auf, ergibt sich für $\hat{\boldsymbol{\beta}}$ der folgende Ausdruck.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad (2.15)$$

Unter Einhaltung meherer Annahmen, der *Gauß-Markow-Annahmen*, kann gezeigt werden, dass die durch die Methode der kleinsten Quadrate bestimmten Schätzwerte der Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$ die *besten linearen erwartungstreuen Schätzer* darstellen (Neter u. a. [26], S.18). Das bedeutet, dass deren Erwartungswerte den wahren Werten $\boldsymbol{\beta}$ entsprechen und dabei verglichen mit allen anderen linearen erwartungstreuen Schätzern die kleinste Varianz aufweisen.

2.4.2 *Güte des Fits, Bestimmtheitsmaß und adjustiertes Bestimmtheitsmaß*

Das Bestimmtheitsmaß R^2 sowie das adjustierte Bestimmtheitsmaß sind Gütemaße des Fits eines Regressionsmodells. Das Bestimmtheitsmaß stellt den Anteil der Gesamtvarianz der Daten dar, der durch das statistische Modell erklärt wird und wird mathematisch folgendermaßen ausgedrückt:

$$R^2 = 1 - \frac{\text{SQR}}{\text{SQT}} \tag{2.16}$$

SQR ist wie in Gleichung 2.8 definiert und die Totale Quadratsumme (SQT) ergibt sich wie folgt:

$$\text{SQT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.17}$$

Die SQT ist also die Summe der quadrierten Abstände der *abhängigen Variable* zu ihrem arithmetischen Mittelwert und entspricht dem Residualsummenquadrat eines linearen Modells, das außer eines konstanten Terms keine Parameter besitzt. Je besser das betrachtete Modell die Daten *fittet* oder auch *erklärt*, desto kleiner wird das Verhältnis von SQR zu SQT, womit sich das Bestimmtheitsmaß eins annähert.

Das adjustierte Bestimmtheitsmaß berücksichtigt zusätzlich die Anzahl der im Modell verwendeten unabhängigen Variablen, wobei die Zunahme unabhängiger Variablen bestraft wird. Mathematisch ergibt sich das adjustierte Bestimmtheitsmaß folgendermaßen:

$$\text{adj. } R^2 = 1 - \frac{\text{SQR}/(n-p-1)}{\text{SQT}/(n-1)} \tag{2.18}$$

Dabei entspricht n der Anzahl der Datenpunkte und p der Anzahl unabhängiger Variablen ohne konstanten Term. Werden zusätzliche unabhängige Variablen in das Modell aufgenommen, die nicht zu einer besseren Erklärung der Varianz und damit einer Verkleinerung der SQR beitragen, bewirkt die höhere Anzahl p der unabhängigen Variablen eine Verringerung des adjustierten R^2.

2.4.3 *Multikollinearität*

Besonderes Augenmerk sollte auf den Ausdruck $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ in Gleichung 2.15 gelegt werden. Die Inverse der Matrix $\boldsymbol{X}^T\boldsymbol{X}$ existiert nur dann, wenn diese vollen Rang aufweist. Liegt zwischen den Prediktoren der Designmatrix exakte Multikollinearität vor, bedeutet dies, dass die Prediktoren nicht linear unabhängig sind. Daraus folgt, dass es sich bei der Matrix $\boldsymbol{X}^T\boldsymbol{X}$ um eine singuläre Matrix handelt (Strang [29], S.206). Eine solche Matrix ist nicht invertierbar und es ist somit nicht mehr möglich eine exakte Lösung der Gleichung 2.15 zu bestimmen.

Mithilfe der *Pseudoinversen* (Strang [29], S.396f) sowie numerischer Optimierungsverfahren können die Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$ zwar immer noch bestimmt werden, jedoch sind diese nicht mehr eindeutig.

Liegt keine exakte, jedoch starke Multikollinearität vor, reagieren die Regressionskoeffizienten sehr empfindlich auf Änderungen der Datengrundlage. Die Interpretierbarkeit der Regressionskoeffizienten ist dadurch nicht mehr gegeben. Dies wird auch durch die hohen Standardfehler der Regressionskoeffizienten ersichtlich.

2.5 ASYMMETRIC LEAST SQUARES SMOOTHING

Spektraldaten sollten eine Grundlinie von Null aufweisen, von der sich nur die Signalpeaks abheben. Häufig setzen sich die Spektraldaten jedoch aus den Signalpeaks sowie weiterer Bestandteile zusammen, die durch systematische und statistische Störeinflüsse verursacht werden. Bei Raman- und Infrarotmessungen kann häufig ein durch Fluoreszenz verursachter Untergrund beobachtet werden.

Eine weitere Schwierigkeit ergibt sich darin, dass der durch die systematischen und statistischen Störeinflüsse verursachte Untergrund von Spektrum zu Spektrum variieren kann. Dies wirkt sich äußerst negativ auf die Vergleichbarkeit der Spektren und folglich auch auf die Güte einer auf jenen Spektraldaten basierenden quantitativen Analyse aus. Es ist deshalb notwendig, den Untergrund mit adequaten Verfahren zu entfernen und damit die *Baseline* des Spektrums zu korrigieren.

In dieser Arbeit wird das von Eilers und Boelens entwickelte *Asymmetric Least Squares Smoothing*-Verfahren (Eilers und Boelens [10]) zur Entfernung des Untergrundes im Rahmen von Infrarotmessungen angewendet. Es handelt sich dabei um ein Regressionsverfahren mit zusätzlichem Bestrafungsterm. Mathematisch kann die Bestimmung einer optimalen, von den Parametern p und λ sowie dem Spektrum $\mathbf{y}$ abhängigen Baseline $\mathbf{b}_{opt}$ mithilfe der in Gleichung 2.19 dargestellten Minimierungsvorschrift beschrieben werden.

$$\mathbf{b}_{opt} = \arg\min_{\mathbf{b}} S(\mathbf{b}|\mathbf{y}, p, \lambda) = \sum_i w(y_i, b_i|p)(y_i - b_i)^2 + \lambda \sum_i (\Delta^2 b_i)^2 \quad (2.19)$$

Eine optimale Baseline $\mathbf{b}_{opt}$ weist bei gegebenen Parametern p und λ den geringsten Fehler S auf. Der Fehler des Regressionsfits setzt sich hierbei aus zwei Teilen zusammen. Der erste Teil stellt die Summe des durch $w(y_i, b_i|p)$ gewichteten quadratischen Abstands zwischen den Datenpunkten y_i des Spektrums und den korrespondierenden Datenpunkten der aktuell betrachteten Baseline b_i dar. Der zweite Teil beschreibt den durch Parameter λ gewichteten Glättungsgrad der Baseline.

Der Parameter p steuert den asymmetrischen Einfluss der positiven und negativen Residuale auf den Fehler des Fits und wird mit einem Wert aus dem reellwertigen Intervall $(0,1)$ initialisiert. Die unterschiedliche Gewichtung positiver und negativer Residuale wird durch die in Gleichung 2.20 dargestellte Funktion w erreicht. Hierbei ist I die Indikatorfunktion, die abhängig davon, ob der aktuelle Residualwert, also die Differenz zwischen dem Spektralwert y_i an Index i und dem Wert der aktuellen Baseline b_i an Index i, kleiner oder größer gleich Null ist.

$$w(y_i, b_i|p) = I[(y_i - b_i) < 0](1-p) + I[(y_i - b_i) \geq 0]p \tag{2.20}$$

Für p wird typischerweise ein Wert zwischen 0.001 und 0.05 gewählt (Fulton und Lunev [13]). Dies hat zur Folge, dass positive Residuale, die Peaks des Spektrums, einen deutlich geringeren Einfluss auf den Fehler des Fits haben als negative Residuale. Die Baseline nimmt dadurch im Vergleich zu einem ungewichteten, symmetrischen Regressionsverfahren niedrigere Werte an.

Mit Parameter λ lässt sich der Grad der Glättung steuern. $\Delta^2 b_i$ ist der zweifach auf b_i angewendete Differenzoperator mit dem die Krümmung der Baseline an Position i berechnet wird. Krümmungswerte, die von Null abweichen, können durch λ beliebig stark bestraft werden. Je höher der Wert für λ gewählt wird, desto mehr nähert sich die Baseline einer Geraden an.

In Abbildung 2.3 sind mehrere Baselinefits, die mithilfe unterschiedlicher Parametrisierungen des *Asymmetric Least Squares Smoothing*-Verfahrens berechnet wurden, zu sehen. Diesen liegt ein Infrarotspektrum mit Fluoroeszenzstöreinfluss zugrunde, das dem Datensatz `chondro` des R-Softwarepaketes *hyperSpec* (Beleites und Sergo [6]) entnommen wurde.

Die resultierenden Baselines und die durch Subtraktion der Baseline korrigierten Spektren weisen erhebliche Unterschiede auf. Es ist deshalb wichtig, die durch p und λ parametrisierte Baselinekorrektur zu validieren. Dies ist bei einer überschaubaren Anzahl zu verarbeitender Spektren qualitativ durch visuelle Überprüfung der Spektren möglich. Zur Validierung der Baselinekorrektur bei größeren Datenmengen finden sich Untersuchungen zu einer teilautomatisierten, statistischen Überprüfung des Fits mithilfe von Kreuzvalidierung und Hyperparametertuning (Liland, Almøy und Mevik [20]).

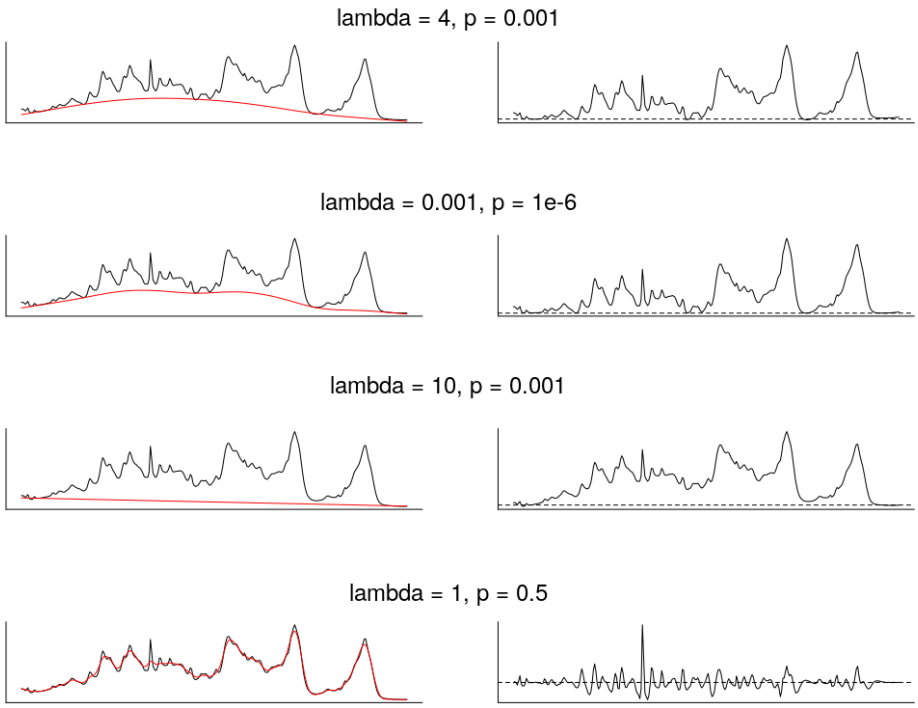


Abbildung 2.3: Asymmetric Least Squares Smoothing von Infrarotspektren mit unterschiedlichen Parametrisierungen. Originalspektrum und Baseline (links). Korrigiertes Spektrum (rechts).

2.6 NICHT-NEGATIVE MATRIXFAKTORISIERUNG

Die NMF ist ein Verfahren, das der Datenreduktion sowie gleichzeitig der Extraktion latenter Variablen dient. Mit latenten Variablen sind in diesem Kontext Komponentenspektren, die zu chemischen Bestandteilen der untersuchenten Probe korrellieren, gemeint. In dieser Arbeit werden durch das NMF-Verfahren Spektraldaten oder Chromatogramme beinhaltende Matrizen faktorisiert und dadurch in Gewichtungsfaktoren und Komponentenspektren aufgespalten. Die Gewichtungsfaktoren dienen als quantitative Features innerhalb von Regressionsmodellen.

Die NMF findet auch in anderen Bereichen wie dem Text-Mining (Pauca u. a. [27]), Empfehlungssystemen (Hernando, Bobadilla und Ortega [17]) oder der Bildanalyse (Hastie, Tibshirani und Friedman [16], S.555) Anwendung. Dabei können zu extrahierende, latente Variablen im Text wiederkehrende Themen oder Gesichtscharakteristiken bei der Verarbeitung von Portraitfotos entsprechen. Ein visualisiertes Beispiel zur Faktorisierung von Portraitfotos mithilfe einer NMF und weiterer Verfahren findet sich unter anderem in Hastie, Tibshirani und Friedman [16] (S.555).

Mathematisch betrachtet dient eine nicht-negative Matrixfaktorisierung der

Lösung des in Gleichung 2.21 aufgeführten Optimierungsproblems. Eine nicht-negative Matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ wird in zwei nicht-negative Matrizen $\mathbf{W} \in \mathbb{R}_+^{m \times r}$, die Gewichtungsfaktormatrix oder auch *Score*-Matrix, und $\mathbf{H} \in \mathbb{R}_+^{r \times n}$, die *Loading*-Matrix faktorisiert. Die Faktormatrizen $\mathbf{W}$ und $\mathbf{H}$ sind jeweils von Rang $r < \min(m, n)$, wobei dieser einen Parameter des NMF-Verfahrens darstellt und manuell festgelegt werden muss. Eine konzeptuelle Darstellung der nicht-negativen Matrixfaktorisierung ist in Abbildung 2.4 zu sehen.

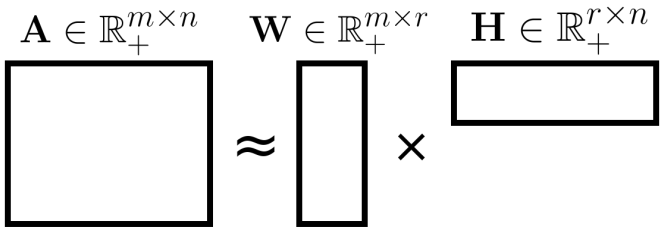


Abbildung 2.4: Konzeptuelle Darstellung einer nicht-negativen Matrixfaktorisierung

Eine optimale Wahl der beiden Matrizen $\mathbf{W}$ und $\mathbf{H}$ hat eine Minimierung der Abweichung zwischen Ursprungsmatrix $\mathbf{A}$ und deren Rekonstruktion durch $\mathbf{WH}$ zur Folge. Zur Bestimmung der Abweichung können unterschiedliche Abweichungs- respektive Distanzmaße angewendet werden. Die in dieser Arbeit verwendete Implementierung nutzt, wie in Gleichung 2.21 dargestellt, die quadratische Frobeniusnorm als Abweichungsmaß. Deren Anwendung hat den Vorteil einer relativ einfachen Berechnung. Es werden dabei lediglich die quadrierten Elemente der Abweichungsmatrix $(\mathbf{A} - \mathbf{WH})$ aufsummiert.

$$\min_{\mathbf{W}>0,\mathbf{H}>0} \|\mathbf{A} - \mathbf{WH}\|_F^2 = \min_{\mathbf{W}>0,\mathbf{H}>0} \sum_{i=1}^{m} \sum_{j=1}^{n} [(\mathbf{A} - \mathbf{WH})_{ij}]^2 \tag{2.21}$$

Die Randbedingung der nicht-Negativität der NMF bietet einen entscheidenden Vorteil gegenüber anderen Faktorisierungsmethoden, die dieser nicht unterliegen. Dies sind unter anderem die Independent Component Analysis (ICA) (Mishra u. a. [24], Wang, Ding und Hou [30]) und Principal Component Analysis (PCA) (Wehrens [31], S.43), die ebenfalls im chemometrischen Arbeitsfeld angewendet werden. Die nicht-Negativität der aus der Faktorisierung resultierenden Komponentenspektren, die sich in der *Loading*-Matrix $\mathbf{H}$ befinden, macht deren chemische Interpretation deutlich intuitiver als dies bei den Komponentenspektren anderer Faktorisierungsmethoden der Fall ist. Dies liegt daran, dass Absorptionsspektrogramme und Chromatogramme, physikalisch bedingt, ausschließlich nicht-negative Werte aufweisen.

3 VORVERARBEITUNG DER DATEN

In diesem Abschnitt werden die individuellen Arbeitsschritte zur Vorverarbeitung der vorliegenden Messdaten, auch *Preprocessing* genannt, vorgestellt. Diese umfassen das Überführen der Rohdaten in eine geeignete Datenstruktur sowie die Transformation der Spektraldaten und Chromatogramme zur Eliminierung von Störfaktoren und der Herstellung einer besseren Vergleichbarkeit dieser. In der oberen Hälfte des in Abbildung 1.4 dargestellten Flowcharts des vollständigen Analyseprozesses sind die individuellen Preprocessingschritte sowie deren Abfolge grafisch dargestellt.

3.1 LADEN DER DATEN

Die resultierenden Messdaten der verschiedenen Messmethoden liegen in Form unterschiedlicher Dateiformate und -strukturen vor. Dies sind strukturierte Comma Separated Values (CSV)- und Exceldateien sowie durch Leerzeichen getrennte, Gleitkommazahlen beinhaltende Textdateien. Um die Datenanalyse in einer einheitlichen und reproduzierbaren Art und Weise durchführen zu können, ist es deshalb notwendig Routinen für die verschiedenen Formate zu entwickeln, die die Messdaten in eine geeignete Datenstruktur überführen. Als Datenstruktur werden in dieser Arbeit *hyperSpec*-Objekte verwendet. Diese sind Teil des R Softwarepakets *hyperSpec* (Beleites und Sergo [6]) und bieten zusätzlich zur Verwaltung der numerischen Spektraldaten und Chromatogramme Möglichkeiten zugehörige Metainformationen zu verwalten. Die vorliegenden Messdaten werden als numerische Matrix `spc` des *hyperSpec* Objekts gespeichert. Diese enthält Spektren oder Chromatogramme als Zeilen. Zusätzlich werden die zur jeweiligen Spalte der Matrix korrespondierenden Elutionszeitpunkte, im Falle der Chromatogramme, und die Wellenzahlen, im Falle der Spektren, in einem zusätzlichen Vektor, dem *k*-Vektor, gespeichert.

Zum Einlesen von CSV und Exceldateien (`.XLS`) werden die Softwarebibliotheken `readr` und `readxl`, die beide Teil der Softwaresammlung `tidyverse` (Wickham u. a. [33]) sind, verwendet. Die darin enthaltenen Funktionen bieten Konfigurationsmöglichkeiten für Trennungs- und Dezimalzeichen sowie regionsabhängige Einstellungen (bspw. im Bezug auf Datum und Zeit). Attenuated Total Reflection (ATR)-IR-, HPLC und die GPC-Rohdaten liegen im CSV-Format vor. Zu den GPC-Messungen liegen zusätzlich bereits durch die Software *Omnic* berechnete Massenschwerpunkte als Excel-Arbeitsblatt vor. Die CSV-Dateien unterscheiden sich hauptsächlich durch die An- oder Abwesenheit von Spaltenbezeichnern und darin, ob eine Datei Einzelmessungen oder mehrere Messungen jeweils als Spalte enthalten.

Die eingelesenen Daten werden in einen Trainings- und Testdatensatz unterteilt. Als Testdatensatz fungieren die zu Probennummer drei korrespondierenden Messdaten mit Rezyklatgehalten von 0, 30%, 50% und 100%. Hierzu wird jeweils ein *hyperSpec*-Objekt für die Trainings- und Testdaten angelegt und alle weiteren Analyseschritte ausschließlich mit dem Trainingsdatensatz ausgeführt. Der Testdatensatz wird nur zur abschließenden Validierung der statistischen Modelle herangezogen.

3.2 INTERPOLATION VON SPEKTRALDATEN UND CHROMATOGRAMMEN

Eine Problematik, die bei der Verarbeitung von Spektren unterschiedlicher Herkunft sowie Chromatogrammen im Allgemeinen auftritt, ist die Uneinheitlichkeit der x-Koordinaten der Spektren und Chromatogramme.

Die Datenpunkte eines Spektrums sind Tupel (x, y), die sich aus einer Wellenzahl x und zugehörigem Absorptions- oder Transmissionswert y zusammensetzen. Datenpunkte von Chromatogrammen sind ebenfalls Tupel (x, y), wobei ein x-Wert hierbei einem Elutionszeitpunkt respektive Elutionsvolumen entspricht.

Wie in Abschnitt 3.1 beschrieben, werden Spektraldaten und Chromatogramme im Zuge des Data Loadings in eine gemeinsame Datenmatrix eines *hyperSpec*-Objekts überführt, wobei die Zeilen der Matrix individuelle Spektren oder Chromatogramme repräsentieren. Jede Spalte der Matrix enthält hierbei die zu einem spezifischen x-Wert, also einer Wellenzahl respektive Elutionszeitpunkt, korrespondierenden y-Werte der individuellen Spektren und Chromatogramme.

Um unterschiedliche Spektren oder Chromatogramme in eine gemeinsame Datenmatrix zu überführen, bieten sich bei uneinheitlichen x-Koordinaten zwei Möglichkeiten an. Es kann eine dünnbesetzte Matrix gebildet werden, die eine Spalte für jede individuell auftretende x-Koordinate enthält. Dies resultiert jedoch in einer sehr großen Matrix, die viele fehlende Werte (`NA`-Werte in R) enthält. Es existieren Datenformate, die eine speichereffiziente Verwaltung von dünnbesetzten Matrizen ermöglichen. Diese sind jedoch nicht immer mit anderen Softwarepaketen kompatibel.

Eine weitere Möglichkeit die Daten in einer gemeinsamen Matrix zu verwalten, stellt die Interpolation der Spektren oder Chromatogramme über gemeinsame x-Koordinaten dar. Hierfür kann die Funktion `spc.loess` des *hyperSpec* Softwarepakets verwendet werden, die zur Interpolation ein lokales Regressionsverfahren (Cleveland [8]) nutzt. Hierbei wird die Interpolationsfunktion aus mehreren, lokalen Regressionen konstruiert. Die resultierende, zusammengesetzte Funktion wird wiederum genutzt, um die zur spezifizierten, gemeinsamen x-Achse korrespondierenden Datenpunkte zu berechnen.

In Abbildung 3.1 ist die Interpolation mithilfe des LOESS-Verfahrens dargestellt. Im oberen Teil der Abbildung sind zwei synthetische Messreihen mit unterschiedlichen x-Koordinaten in rot respektive blau abgebildet. Es wird eine Interpolationsfunktion auf Basis der roten Datenpunkte erzeugt. Der Interpolationsfunktion werden anschließend die x-Koordinaten der blauen Datenpunkte übergeben und die zugehörigen Funktionswerte berechnet. Im unteren Teil der Abbildung ist die unveränderte blaue Messreihe sowie die interpolierte, rote Messreihe dargestellt. Beide Messreihen weisen nun identische x-Koordinaten auf und können in einer gemeinsamen, dichten Matrix verwaltet werden.

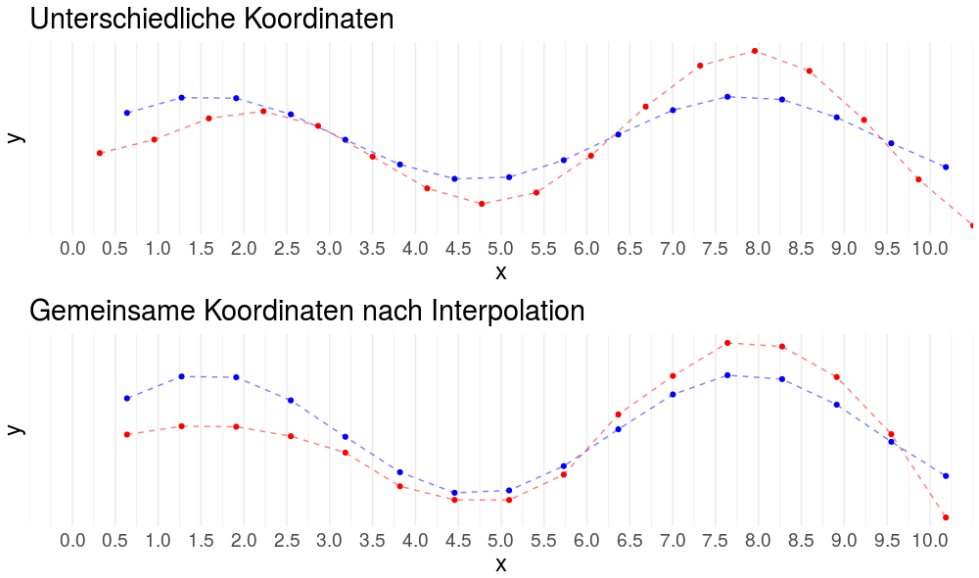


Abbildung 3.1: Interpolation mit LOESS

3.3 KORREKTUR DES UNTERGRUNDES

Zur Durchführung der Baselinekorrektur von ATR-IR-Spektren wird das in Abschnitt 2.5 vorgestellte *Asymmetric Least Squares Smoothing*-Verfahren angewendet. Hierzu wird die Funktion `baseline.als` des *baseline* Softwarepakets (Liland, Almøy und Mevik [20]) verwendet. Geeignete Parameter λ und p werden durch qualitative Validierung der korrigierten Spektren des Trainingsdatensatzes bestimmt.

3.4 EINGRENZUNG RELEVANTER SPEKTRALBEREICHE

Häufig sollen Spektren auf relevante Spektralbereiche reduziert werden. Beispielsweise können durch ein Beschneiden der Spektren Spektralbereiche, die ein starkes Rauschen oder Artefakte aufweisen, entfernt werden. Ein weiterer Anwendungsfall ergibt sich, wenn Statistiken über spezifische Spektralbanden berechnet werden sollen. Dies sind beispielsweise die maximale

Peakhöhe oder das numerische Integral und somit die Fläche unter dem Peak von Interesse. Die Extraktion regionaler Features findet im Rahmen des *Feature Engineerings* Anwendung und wird in Abschnitt 3.6 beschrieben.

Soll eine spezifische Region ausgewählt werden, geschieht dies unter Angabe einer unteren Grenze k_{min} und einer oberen Grenze k_{max} des Spektralbereichs. Da der Wellenzahlvektor $\mathbf{k}$ nur eine endliche Menge diskreter Werte enthält, müssen mithilfe der Berechnungsvorschrift (3.1) der zu k_{min} korrespondierende Index x_{min} des Wellenzahlvektors sowie mithilfe der Berechnungsvorschrift (3.2) der zu k_{max} korrespondierende Index x_{max} des Wellenzahlvektors ermittelt werden.

$$x_{min} = \arg\min_i |k_i - k_{min}| \tag{3.1}$$

$$x_{max} = \arg\min_i |k_i - k_{max}| \tag{3.2}$$

Die Elemente an Position x_{min} und x_{max} des Wellenzahlvektors weisen die kleinste absolute Abweichung zu den Intervallgrenzen k_{min} respektive k_{max} auf. Sind die zu k_{min} und k_{max} korrespondierenden Indizes x_{min} und x_{max} ermittelt, werden diese zur Indexierung der relevanten Spalten der Spektralmatrix verwendet.

Programmatisch lässt sich der Spektralbereich der in *hyperSpec*-Objekten verwalteten Spektraldaten durch den *Bracket-Operator* `[]` eingrenzen. Dieser stellt eine für *hyperSpec*-Objekte überladene Implementierung dar und kann zusätzlich zur Auswahl des Spektralbereichs auch zur Auswahl von Spektren und Metadaten angewendet werden.

3.5 TRANSFORMATION DER SPEKTREN

Für eine Transformation der Spektren gibt es mehrere Gründe. Dies kann die intuitiv besser handhabbare Skalierung der Intensitäten auf einen Wertebereich um eins sein, indem jedes Spektrum durch seinen mittleren Spektralwert geteilt wird. Auch als Preprocessingschritt für anschließende Faktorisierungsmethoden ist eine Normalisierung dringend erforderlich. Eine NMF setzt beispielsweise voraus, dass die zu verarbeitende Matrix (hier die Spektralmatrix) ausschließlich nicht-negative Werte enthält. Es ist deshalb notwendig vor der Anwendung der NMF auf eine Spektralmatrix, das Minimum jedes Spektrums von allen dem Spektrum zugehörigen Spektralwerten abzuziehen. Solche vektorisierten Berechnungsvorschriften können mithilfe der Basis-R Funktionen der *apply-Familie sowie mit der Funktion `sweep` des *hyperSpec*-Pakets durchgeführt werden.

3.6 FEATURE ENGINEERING

Liegen die Spektren oder Chromatogramme nach Anwendung der zuvor beschriebenen Preprocessing-Schritte in einem bereinigten und vergleichba-

ren Format vor, können verschiedene Datenreduktionstechniken angewandt werden, um deren Dimensionalität zu verringern. Spektren und Chromatogramme können sich aus mehreren tausend Datenpunkten zusammensetzen, wobei diese zu spezifischen Wellenzahlen oder Elutionszeitpunkten korrespondieren. Jede individuelle Wellenzahl respektive Elutionszeitpunkt stellt dabei eine Dimension des Spektrums respektive Chromatogramms dar. Durch die Berechnung verschiedener zusammenfassender Statistiken wird versucht einen möglichst hohen Teil der in den Spektren vorliegenden Information, die in Zusammenhang mit der Zielvariablen, also dem Rezyklatgehalt der Probe, steht, zu extrahieren.

Eine äußerst hilfreiche Technik zur qualitativen Untersuchung des Zusammenhanges von Spektral- und Chromatogrammbereichen mit dem Rezyklatgehalt der Probe ist die Visualisierung der Daten sowie Teilen dieser. Der bekannte Rezyklatgehalt der Trainingsdaten wird dabei als Farbvariable verwendet. Es lassen sich dadurch relativ schnell Bereiche identifizieren, die einen mit dem Rezyklatgehalt korrelierenden Farbgradienten aufweisen. In Abbildung 3.2 ist beispielhaft der Wellenzahlbereich von 1140 - 1200 cm^{-1} der ATR-IR-Messdaten dargestellt. Die Farbe der Spektren hängt dabei vom Rezyklatgehalt der korrespondierenden Probe ab. Die Höhe des Signalmaximums bei 1170 cm^{-1} weist eine negative Korrelation mit dem Rezyklatgehalt auf.

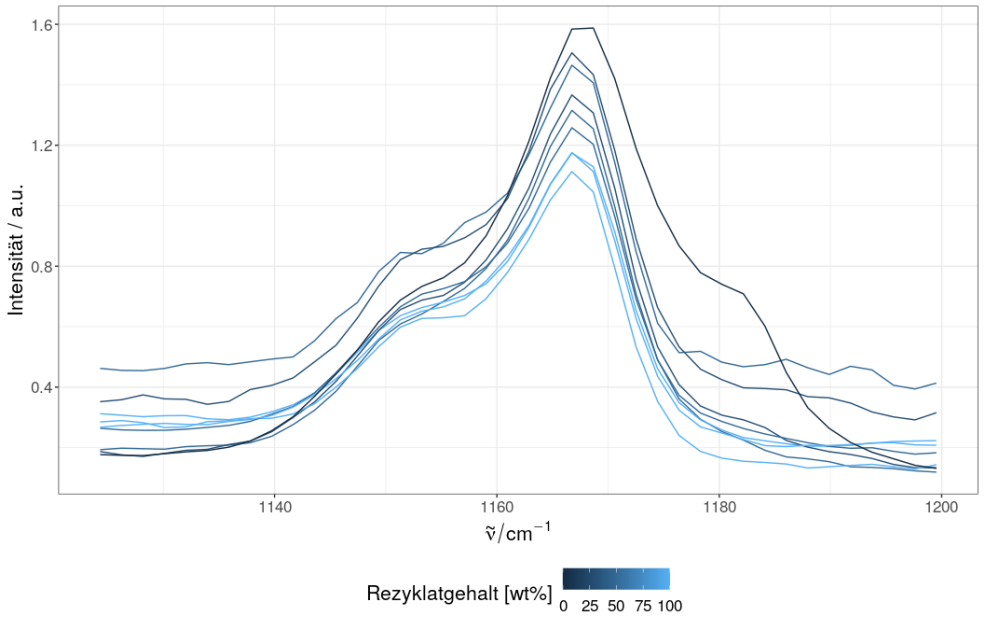


Abbildung 3.2: Signalmaxima eines Spektralbereichs der ATR-IR-Daten mit vom Rezyklatgehalt abhängigem Farbgradienten

Sind interessante Bereiche wie der beispielhaft in Abbildung 3.2 dargestellte Bereich identifizert, können diese zur Berechnung von zusammenfassenden

Statistiken wie deren Maximum, Mittelwert oder numerischen Integrals herangezogen werden. Diese werden in einem *data.frame*, der *Designmatrix* $\mathbf{X}$, gespeichert und stellen die Features (auch Prediktoren oder unabhängige Variablen) im Rahmen der quantitativen Analyse dar. Die Features der *Designmatrix* werden in dieser Arbeit vor allem durch die Berechnung regionaler Statistiken sowie durch die nicht-negative Matrixfaktorisierung relevanter Spektralbereiche und Verarbeitung der dabei resultierenden Gewichtungsfaktorenmatrix erzeugt. Auf diese Methoden wird in den folgenden beiden Unterkapiteln näher eingegangen.

3.6.1 *Extraktion regionaler Statistiken*

Zur Berechnung regionaler Statistiken werden unter anderem die in Abschnitt 3.4 und Abschnitt 3.5 vorgestellten Techniken verwendet. Die Spektren oder Chromatogramme werden zunächst mithilfe von Visualisierungen qualitativ auf informative Regionen untersucht und anschließend auf jene Regionen eingegrenzt. Diese werden wiederum zur Berechnung zusammenfassender Statistiken wie dem Maximum, Mittelwert oder numerischen Integral herangezogen. Die berechneten Statistiken werden als Features und somit als Spalten der Designmatrix $\mathbf{X}$ gespeichert.

Häufig werden die erzeugten Features weiteren Transformationen unterzogen. Dies kann beispielsweise die Anwendung des Logarithmus oder der Quadratwurzelfunktion sein, sollte ein exponentieller oder polynomialer Zusammenhang des Features mit der Targetvariablen, also dem Rezyklatgehalt, vermutet werden. Die transformierten Features können der Designmatrix dabei zusätzlich hinzugefügt werden oder das ursprüngliche Feature ersetzen.

Eine weitere Technik ist das Kombinieren mehrerer Features. Es wird beispielsweise die relative Höhe eines Signalmaximums bezogen auf einen Referenzwert berechnet und der Designmatrix als Feature hinzugefügt.

Die Features werden in einem iterativen Prozess erstellt. Dieser setzt sich aus der Visualisierung der verschiedenen Spektralregionen, Berechnung der Statistiken und Visualisierung des Rezyklatgehaltes gegen die berechneten Statstiken als Scatter- und Lineplot zusammen. Abschließend werden die iterativ erarbeiteten Schritte zur Erzeugung der Features in eine Funktion überführt, sodass diese einheitlich auf Trainings- und Testdaten angewendet werden können.

3.6.2 *Nicht-negative Matrixfaktorisierung*

Die Ausgangsmatrix $\mathbf{A}$ des in Abschnitt 2.6 beschriebenen NMF-Verfahrens besteht im Rahmen dieser Arbeit aus zeilenweisen Spektraldaten oder Chromatogrammen. Jedes Spektrum respektive Chromatogramm setzt sich dabei aus üblicherweise mehreren Tausend Datenpunkten zusammen. Da es sich

bei den Daten um Absorptionswerte handelt, sind diese physikalisch bedingt nicht-negativ. Negative Werte können allerdings trotzdem, durch statistische und systematische Störeinflüsse verursacht, auftreten. Es ist deshalb häufig notwendig die in Abschnitt 3.5 beschriebenen Methoden zur Positivierung von Spektraldaten auf die Datenmatrix **A** anzuwenden bevor eine NMF durchgeführt werden kann.

In der quantitativen Analyse des Rezyklatgehaltes finden die Gewichtungsfaktoren der aus der Faktorisierung resultierenden *Score*-Matrix **W** Anwendung. Die Gewichtungsfaktoren aussagekräftiger Komponentenspektren werden in die Designmatrix **X** aufgenommen und können als Features in Regressionsmodellen verwendet werden. Eine beispielhafte Darstellung der resultierenden Komponentenspektren, die aus der Anwendung einer NMF auf ATR-IR-Spektren von rezyklathaltigem Polypropylen in einem Spektralbereich von 1100 - 1320 cm^{-1} resultieren, ist in Abbildung 3.3 zu sehen.

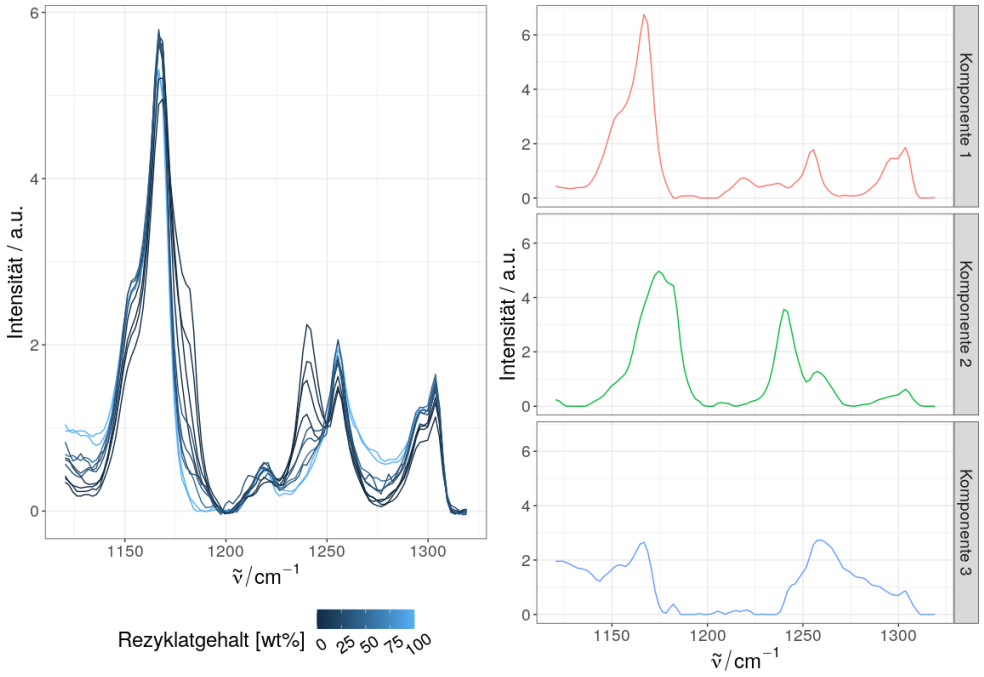


Abbildung 3.3: ATR-IR-Spektren von rezyklathaltigem Polypropylen im Spektralbereich 1100 - 1320 cm^{-1} (links). Resultierende Komponentenspektren einer nicht-negativen Matrixfaktorisierung des Spektralbereichs (rechts).

Die Festlegung des Rangs r der resultierenden Faktormatrizen und damit die Festlegung der Anzahl der resultierenden Komponentenspektren zu $r = 3$ begründet sich in der Anzahl der erwarteten chemischen Bestandteile der Probe. Die Originalspektren weisen mehrere Signalmaxima auf, wobei erwartet wird, dass diese auf PP, PE und Ethylen-Propylen-Copolymer (E/P) zurückzuführen sind. In der abgebildeten Faktorisierung ist unter anderem zu erkennen, dass das negativ mit dem Rezyklatgehalt korrelierte Signal bei ca. 1140 cm^{-1} nahezu ausschließlich durch die zweite Komponente erfasst

wird. Im Rahmen der quantitativen Analyse ist zu untersuchen, ob sich die zur zweiten Faktorisierungskomponente korrespondierenden Gewichtungsfaktoren als unabhängige Variablen in einem Regressionsmodell eignen.

4

QUANTITATIVE ANALYSE

Zur quantitativen Bestimmung des Rezyklatgehaltes der vorliegenden Proben werden Regressionsmodelle herangezogen. Als Datengrundlage dient dabei die Designmatrix $\mathbf{X}$, die in einem iterativen Prozess aus Preprocessing (Kapitel 3), Feature Engineering (Abschnitt 3.6) sowie quantitativer Analyse und Modellvalidierung konstruiert wird.

Die Designmatrix $\mathbf{X} \in \mathbb{R}^{m \times p}$ enthält m Zeilen, die zu den ursprünglichen Spektren der Spektralmatrix $\mathbf{A}$ korrespondieren, sowie p Spalten, die den Features, die im Rahmen des Preprocessings und Feature Engineerings erzeugt wurden, entsprechen.

Die in der Designmatrix enthaltenen Features werden zunächst auf Basis qualitativer Kritieren erstellt. Diese Kriterien fußen einerseits auf chemisch-physikalischem Domänenwissen, wie der Wahl von Spektralbanden und Signalmaxima ausgehend von in der Literatur festgehaltenen Referenzwerten, sowie auf rein datenfokussierten, statistischen Zusammenhängen, die beispielsweise im Rahmen von Visualisierungen festgestellt werden. Im Fall des Polypropylens findet sich beispielsweise in Andreassen [2] eine große Auswahl an relevanten Infrarot- und Ramanspektralbereichen sowie die zu jedem Bereich korrespondierenden funktionalen Gruppen des Moleküls und deren Vibrationsmodi. Die Grundlagen der Vibrationsspektroskopie und eine quantenmechanische Betrachtung dieser finden sich unter anderem in Atkins und Paula [4] (S. 430ff), Harris [15] (S.93ff) und vielen weiteren Lehrbüchern der physikalischen und analytischen Chemie.

Die Erstellung von Features kann auch auf rein datenfokussierten, statistischen Erwägungen basieren. Statistisch relevante Spektralbereiche können mithilfe von Visualisierungen der Spektren des Trainingsdatensatzes identifiziert werden. Es wird dabei, wie in Abbildung 3.2 dargestellt, der bekannte Rezyklatgehalt der Proben als Farbästhetik festgelegt. Zusammenfassende Statistiken jeglicher Spektralbereiche, die auf einen funktionalen Zusammenhang der Absorptionshöhe oder Signalfläche mit dem Rezyklatgehalt schließen lassen, werden anschließend als Feature in die Designmatrix $\mathbf{X}$ aufgenommen.

Das beschriebene Vorgehen des Feature Engineerings führt aufgrund der hohen Zahl erzeugter Features typischerweise dazu, dass die resultierende Designmatrix ein hohes Maß an redundanter Information enthält. Diese redundante Information äußert sich in stark korrellierten Features. Im ungünstigsten Fall liegt exakte Multikollinearität vor, was bedeutet, dass eine oder

mehrere Features der Designmatrix als Linearkombination einer oder mehrerer anderer Features dargestellt werden können. Multikollinearität führt, wie in Unterabschnitt 2.4.3 beschrieben, dazu, dass der Einfluss der einzelnen Prediktoren auf die Targetvariable nicht mehr sinnvoll interpretiert werden kann. Mithilfe der *Feature Selection*-Methoden der *stufenweiser Regression* und der *LASSO-Regression* werden die am besten zur Vorhersage des Rezyklatgehaltes geeigneten Features identifiziert und die Anzahl der unabhängigen Variablen reduziert.

Abgesehen von der Vermeidung von Multikollinearität hat die Reduktion der im Vorhersagemodell verwendeten Variablen eine weitere für diese Arbeit äußerst wichtige Funktion. Erstrebt wird eine Vorhersage des Rezyklatgehaltes unbekannter Proben bei einer möglichst niedrigen Anzahl hierfür benötigter chemisch-physikalischer Messmethoden. Feature Selection dient im Rahmen dieser Arbeit also auch zur Reduktion der zur Datengenerierung erforderlichen Messmethoden und damit zur Reduktion von Zeit und benötigter Resourcen bei der zukünftigen industriellen Anwendung des Vorhersagemodells.

4.1 MULTIPLE LINEARE REGRESSION

Aufgrund ihrer guten Interpretierbarkeit werden multiple lineare Regressionsmodelle zur Vorhersage des Rezyklatgehaltes herangezogen. Die mathematischen Grundlagen dieser Modelle sind in Abschnitt 2.4 dargestellt. Die Spezifikation und der Fit eines linearen Modells stellen sich in R relativ einfach dar. Die Funktion `lm()`, die bereits im Grundumfang Rs enthalten ist, kann hierzu verwendet werden, wobei dieser dazu eine Modellspezifikation und die Designmatrix als Parameter übergeben werden. In diesem Abschnitt wird die Auswertung zweier Modelle auf Basis von ATR-IR Messungen beispielhaft besprochen. Beide Modelle weisen jeweils den Rezyklatgehalt als *abhängige Variable* und eine respektive zwei Gewichtungsfaktoren einer NMF über dem Wellenzahlbereich um 950 cm^{-1} als unabhängige Variablen auf. Weiterhin wird automatisch ein konstanter Term oder auch Achsenabschnitt gefittet.

Mit der `summary()`-Funktion, die ebenfalls Teil der Grundfunktionalität Rs ist, können statistischen Modellen mehrere Kennzahlen zur Auswertung und Diagnose entnommen werden. In Tabelle 4.1 ist die formatierte Ausgabe der `summary()`-Funktion für beide Modelle dargestellt.

Im mittleren Abschnitt der Tabelle sind die mithilfe der Methode der kleinsten Quadrate bestimmten Regressionskoeffizienten $\hat{\beta}$ und deren Standardabweichung angegeben. Die Standardabweichung der Regressionskoeffizienten sind hierbei in Klammern gesetzt. Die Anzahl der Sterne, die einigen der Regressionskoeffizienten angestellt sind, zeigen die Signifikanz des jeweiligen Regressionskoeffizienten an und beschreiben, vereinfacht gesagt,

wie wahrscheinlich es ist, dass der tatsächliche Regressionkoeffizient β_i den Wert null besitzt. Dies würde bedeuten, dass die unabhängige Variable x_i keinen Einfluss auf die abhängige Variable hat und demnach aus dem Modell entfernt werden kann. Eine vertiefende Beschreibung findet sich unter anderem in Faraway [11] (S.46).

Tabelle 4.1: Ausgabe der `summary()`-Funktion für zwei lineare Modelle

	Abhängige Variable:	
	Rezyklatgehalt [%wt]	
	(1)	(2)
NMF 950 K1 cm^{-1}		32.238
		(25.081)
NMF 950 K2 cm^{-1}	-59.136^{***}	-31.180
	(3.148)	(21.957)
Konstante	101.112^{***}	54.300
	(2.968)	(36.531)
R^2	0.981	0.985
Adj. R^2	0.978	0.980
Hinweis:	*p<0.1; **p<0.05; ***p<0.01	

Vergleicht man die Regressionskoeffizienten der NMF-Komponenten des ersten Modells mit denen des zweiten Modells, fällt ein sehr starker Anstieg des Standardfehlers auf. Weiterhin liegen die p-Werte der beiden Regressionskoeffizienten des zweiten Modells nicht einmal mehr unter einem Signifikanzniveau von 10%. Dies suggeriert, dass eine hohe Wahrscheinlichkeit besteht, dass die beiden unabhängigen Variablen keinen Einfluss auf die abhängige Variable haben und aus dem Modell entfernt werden können. Tatsächlich ist die hohe Ungewissheit aber auf Kollinearität der beiden Variablen zurückzuführen. In Abbildung 4.1 sind die univariaten Auftragungen des Rezyklatgehaltes gegen die Gewichtungsfaktoren der beiden NMF-Komponenten dargestellt.

Die beiden Prediktoren stellen nahezu Spiegelungen ein und derselben Komponente dar. Das Modell „weiß“ demnach nicht von welcher der beiden Variablen der erklärende Einfluss stammt. Bei der Hinzunahme der ersten NMF-Komponente im zweiten Modell ist ein Anstieg des Bestimmtheitsmaßes sowie des adjustierten Bestimmtheitsmaßes festzustellen. Dieser ist aller-

dings relativ gering und es sollte aus Gründen der Interpretiertbarkeit das erste Modell verwendet werden.

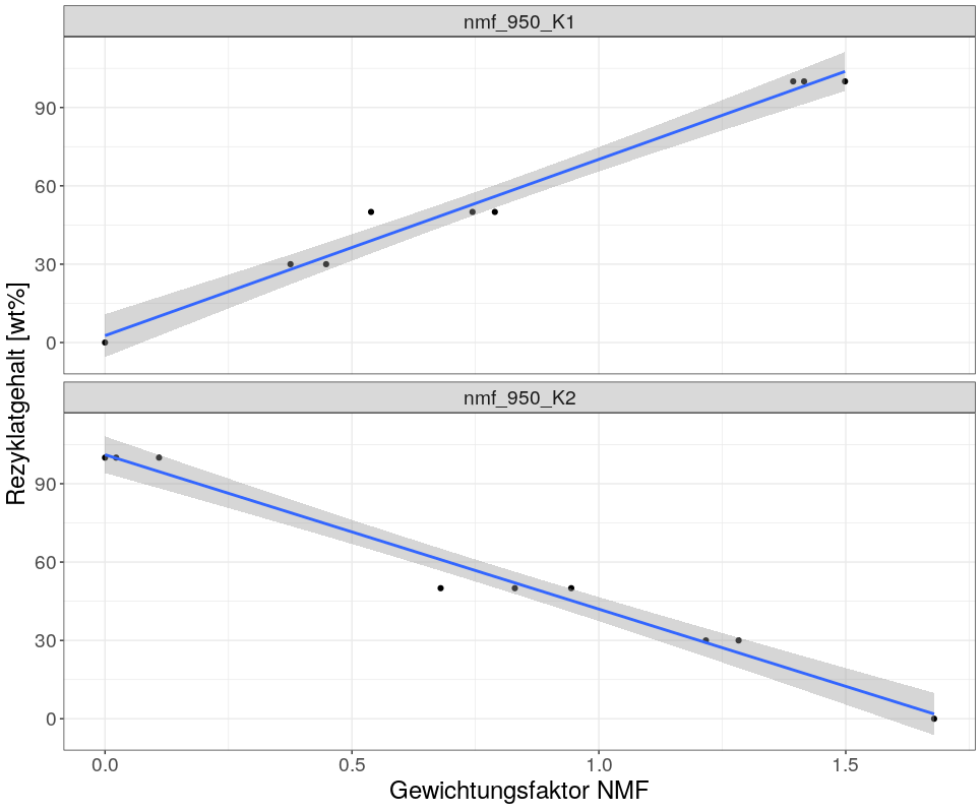


Abbildung 4.1: Auftragung des Rezyklatgehaltes gegen zwei NMF Komponenten des Bereichs um 950 cm^{-1}

Im Rahmen des *Feature Engineerings* wird, wie in Abschnitt 3.6 beschrieben, eine hohe Zahl Prediktoren erzeugt. Das Auftreten von den hier dargestellten (Multi-)Kollinearitätseffekten ist damit sehr wahrscheinlich. Zur Minimierung solcher Effekte werden die in Abschnitt 4.2 beschriebenen Methoden zur *Feature Selection* verwendet.

4.2 FEATURE SELECTION

Das Ziel dieser Arbeit ist, wie in Abschnitt 1.1 dargestellt, die Vorhersage des Rezyklatgehaltes in Kunststofferzeugnissen auf Basis einer Datengrundlage, die mit möglichst geringem Aufwand erzeugt werden kann. „Geringer Aufwand“ ist dabei gleichbedeutend mit der Reduktion der Anzahl der zur Datenerzeugung notwendigen chemisch-physikalischen Messmethoden. Es werden deshalb zunächst in den Analysen, die auf der Datengrundlage einzelner Messmethoden basieren, Methoden der *Feature Selection* angewandt, um die aussagekräftigsten Features zu extrahieren. Die abschließende kombinierte Analyse basiert auf diesen als „beste“ Features identifizierten Features aller Messmethoden. Auch hier werden *Feature Selection*-Methoden angewandt, um die Anzahl der notwendigen Features und damit auch direkt die

Anzahl notwendiger Messmethoden zu reduzieren.

In Abschnitt 4.1 wurde bereits die manuelle Auswertung linearer Modelle auf Basis der durch die *summary*-Funktion bereitgestellten Kennzahlen besprochen. Unterschreitet der zu einem Regressionskoeffizienten korrespondierende p-Wert das gewünschte Signifikanzniveau, deutet dies daraufhin, dass die zugehörige, unabhängige Variable signifikanten Einfluss auf die abhängige Variable ausübt und im Modell verbleiben sollte. Eine manuelle Auswertung ist bei Modellen mit einer geringen Anzahl unabhängiger Variablen noch durchführbar, wird bei einer wachsenden Zahl potenzieller, unabhängiger Variablen jedoch schnell sehr unübersichtlich. Weiterhin sind die p-Werte der Regressionskoeffizienten bei vorliegender (Multi-)Kollinearität nicht mehr aussagekräftig, wie im Beispiel in Tabelle 4.1 dargestellt.

Aufgrund der beschriebenen Schwierigkeiten, die bei einer manuellen *Feature Selection* auftreten, werden (teil-)automatisierte Methoden verwendet. Die Ergebnisse dieser Methoden werden gegenübergestellt und die dabei identifizierten relevanten Features werden zur Nutzung in der abschließenden, kombinierten Analyse gespeichert. Die verwendeten Methoden zur *Feature Selection* werden in den folgenden Unterabschnitten vorgestellt.

4.2.1 *Stufenweise Regression*

Die stufenweise Regression kann auf unterschiedliche Arten durchgeführt werden. Bei der Methode der *Forward Selection* wird zunächst ein Regressionsmodell ohne Variablen an die Daten gefittet. Es wird daraufhin für jede zur Auswahl stehende Variable ein Regressionsmodell mit nur dieser Variable und konstantem Term gefittet. Erreicht eines dieser Modelle eine signifikante, statistische Verbesserung des Fits, gemessen am Root Mean Squared Error (RMSE), wird die in diesem Modell verwendete Variable beibehalten. Der RMSE ergibt sich mathematisch nach folgender Gleichung.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

Hierbei stellen die Werte y_i die beobachteten Werte der Zielvariablen und $\hat{y}_i$ die durch das Modell berechneten Werte dar.

Im nächsten Schritt des *Forward Selection*-Verfahrens werden Modelle mit zwei Variablen gefittet, wobei eine der beiden Variablen die im vorherigen Schritt bestimmte Variable des besten Modells darstellt. Dieser Prozess kann bis zu einem vollständigen Miteinbezug aller Variablen fortgeführt werden.

Das *Backward Selection*-Verfahren bezieht zunächst alle Variablen mitein und entfernt diese iterativ bis eine signifikante Verschlechterung des Fits auftritt. Ein weiteres Verfahren, das *Sequential Replacement*-Verfahren, ist ein *erschöp-*

fendes Verfahren. Es wird zunächst wie beim *Forward Selection*-Verfahren mit einem Modell ohne Variablen begonnen und iterativ Variablen hinzugefügt. Jedoch wird für jede Variablenanzahl die beste Kombination ermittelt, indem alle Kombinationen gefittet werden. Diese Verfahren ist sehr rechenintensiv und sollte nur genutzt werden, wenn die maximale Variablenanzahl im Modell sowie die maximale Anzahl zur Auswahl stehender Variablen überschaubar ist.

In R sind stufenweise Regressionsverfahren unter anderem im Softwarepaket *leaps* (Alan Miller [1]) implementiert. In Verbindung mit dem Softwarepaket *caret* (Kuhn [19]) kann ein stufenweises Regressionsverfahren mithilfe von Bootstrapping oder Kreuzvalidierung durchgeführt werden. Das heißt, dass das Verfahren mehrfach auf Basis unterschiedlicher Teilmengen der Daten durchgeführt und die Ergebnisse der Durchgänge aggregiert werden. Im Falle der Kreuzvalidierung werden die Teilmengen der Daten auch *Folds* genannt. Vertiefende Informationen zur Kreuzvalidierung und Bootstrapping finden sich unter anderem in Hastie, Tibshirani und Friedman [16] (S.241ff).

4.2.2 *LASSO-Regression*

Least Absolute Shrinkage and Selection Operator (LASSO) ist ein Verfahren zur Regularisierung und Feature Selection im Rahmen von Regressionsmodellen (Hastie, Tibshirani und Friedman [16], S.68). Wird LASSO im Kontext der Ordinary Least Squares (OLS)-Regression verwendet, wird nicht mehr nur versucht die Residuenquadratsumme wie in Gleichung 2.8 dargestellt zu minimieren, sondern dieser ein zusätzlicher Bestrafungsterm hinzugefügt. Dieser Bestrafungsterm stellt die Summe der Absolutbeträge der Regressionskoeffizienten dar. Die zu minimierende Funktion ergibt sich damit wie in Gleichung 4.2 dargestellt.

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.2}$$

Der Parameter λ steuert den Einfluss des Bestrafungsterms und muss durch den Anwender des LASSO-Verfahrens spezifiziert werden. Regressionskoeffizienten von Features, die wenig zur Erklärung der Varianz der abhängigen Variable beitragen, werden bei der Wahl eines moderaten bis hohen Wertes λ bis auf null geschrumpft und damit effektiv aus dem Modell entfernt. Der Effekt den verschiedene Werte λ auf ein Modell ausüben, kann mithilfe von Regularisierungspfaden dargestellt werden. In Abbildung 4.2 sind die mit dem Softwarepaket *glmnet* (Friedman, Hastie und Tibshirani [12]) erzeugten Regularisierungspfade für ein auf ATR-Daten basierendes Modell dargestellt.

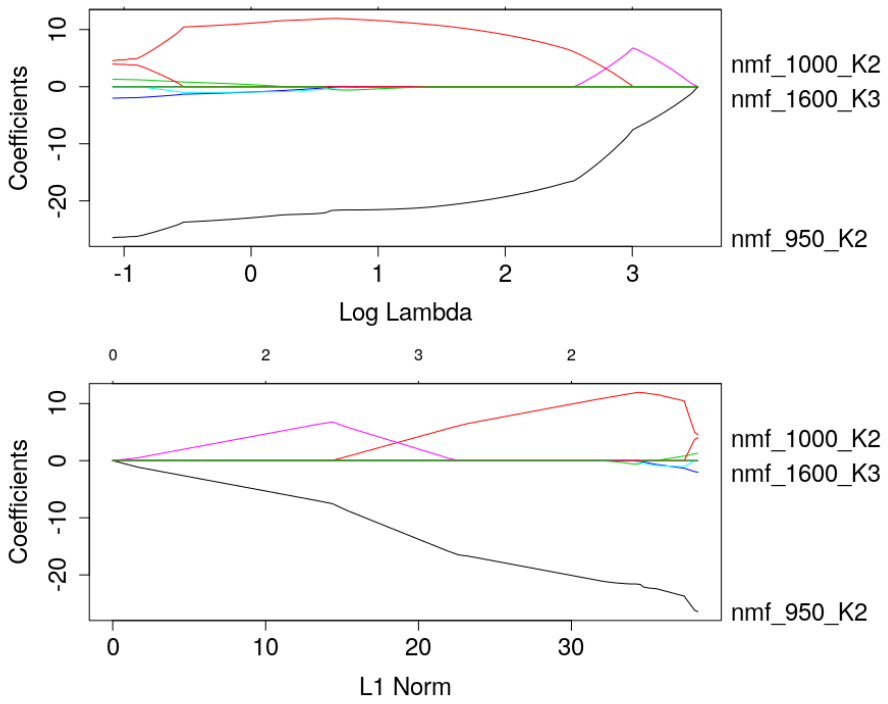


Abbildung 4.2: Zwei Darstellungen der Regularisierungspfade

In der oberen Grafik sind die Regressionskoeffizienten β_j gegen den dekadischen Logarithmus von λ aufgetragen. In der unteren Grafik sind die Regressionskoeffizienten gegen die L_1-Norm der Regressionskoeffizienten $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$, also die Summe ihrer Absolutbeträge, aufgetragen. Man sieht unter anderem, dass der Regressionskoeffizient, der zu Variable `nmf_950_K2` korrespondiert, durchgängig einen Wert ungleich null annimmt, solange $\|\boldsymbol{\beta}\|_1 > 0$. Es lässt sich daraus ableiten, dass diese Variable die Varianz der abhängigen Variable verhältnismäßig gut erklärt und in einem Modell beibehalten werden sollte.

5

EXPERIMENTALTEIL

In diesem Kapitel werden die experimentellen Ergebnisse der Untersuchung des Rezyklatgehaltes der insgesamt 16 vorliegenden Kunststoffproben dargestellt. Zur Auswertung der aus drei verschiedenen Messmethoden stammenden Datengrundlage wurden die in den vorherigen Abschnitten dargestellten Techniken zur Vorverarbeitung und Analyse spektrographischer und chromatographischer Daten angewendet. Weiterhin wurden die Proben vor der Durchführung der qualitativen und quantitativen Analysen in einen Trainings- und Testdatensatz aufgeteilt.

Der Trainingdatensatz besteht aus den Messungen mit Probennummern eins, zwei und vier jeder Rezyklatuntergruppe (0%, 30%, 50% und 100%) und damit aus insgesamt 12 Proben. Der Testdatensatz beinhaltet alle Proben mit Probennummer drei und besteht somit aus vier Messungen. Zur explorativen Datenanalyse sowie des Preprocessings und der Modellbildung wurden ausschließlich die Trainingsdaten herangezogen. Die Testdaten dienten der Validierung der finalen Modelle.

Es werden in den folgenden Abschnitten zunächst die Ergebnisse der Analysen, die auf Basis der Daten individueller Messmethoden durchgeführt wurden, dargestellt und interpretiert. Abschließend werden die Ergebnisse der kombinierten Analyse, in der die im Rahmen des *Feature Engineerings* und anschließender *Feature Selection* erzeugten Features der Einzelmethoden kombiniert ausgewertet wurden, diskuttiert. Die Ergebnisse der abschließenden, kombinierten Auswertung werden herangezogen, um die chemisch-physikalischen Messmethoden zu ermitteln, die sich am besten zur Bestimmung des Rezyklatgehaltes der Proben eignen.

5.1 AUSWERTUNG DER ATR-IR-MESSUNGEN

Die individuellen ATR-Spektren der Messungen liegen als `.csv`-Dateien mit dem Wellenzahlvektor und den Spektraldaten als jeweils separate Spalte vor. Nach Einlesen der Daten mithilfe des Softwarepaketes *readr* in ein *hyperSpec*-Objekt wurden die Vollspektren der Rohdaten zunächst visualisiert und qualitatitiv ausgewertet. Die auf den Rohdaten basierenden Spektren weisen erhebliche Unterschiede in ihrer durchschnittlichen Intensität sowie in der Höhe des Untergrunds auf. Im ersten Preprocessingschritt wurden die Untergründe der Spektren deshalb mit dem *Asymmetric Least Squares Smoothing*-Verfahren, wie in Abschnitt 3.3 beschrieben, korrigiert. Hierbei konnte mit den Parametern $\lambda = 6$ und $p = 0.001$ eine einheitliche Baseline ohne unnatürliche Kurvatur hergestellt werden. Anschließend wurde eine Division

der Spektren durch ihren Mittelwert mithilfe der in Abschnitt 3.5 beschriebenen Transformationsmethoden durchgeführt. Die Intensitäten spezifischer Spektralregionen sind durch diese Normalisierung von Spektrum zu Spektrum vergleichbar. Die Vollspektren der Rohdaten und der Daten nach den beschriebenen Preprocessingschritten sind in Abbildung 5.1 dargestellt.

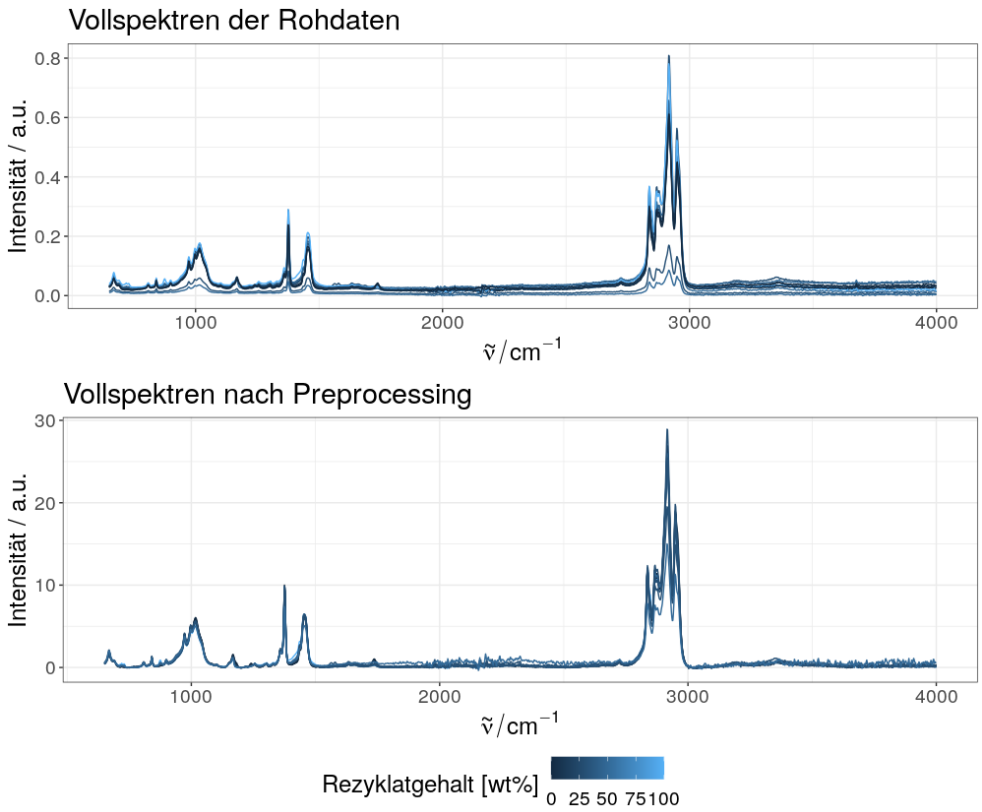


Abbildung 5.1: Vollspektren der Rohdaten (oben). Vollspektren nach Preprocessing (unten).

5.1.1 *Extraktion regionaler Statistiken*

Multiple Spektralbanden der normalisierten Spektren wurden zunächst qualitativ mithilfe von Visualisierungen untersucht. Die stärksten Korrelationseffekte der Signalintensitäten mit dem Rezyklatgehalt konnten bei den Signalmaxima um 722 cm^{-1}, 730 cm^{-1}, 875 cm^{-1} und 1167 cm^{-1} festgestellt werden. Die Spektralregionen sowie eine Auftragung des Rezyklatgehaltes gegen die ermittelten Intensitätswerte der Signalmaxima und -mittelwerte sind in Abbildung 5.2 dargestellt. Die Auftragung des Rezyklatgehaltes gegen die Signalmaxima und -mittelwerte ist dabei einerseits als *Scatterplot* sowie zusätzlich als Regressionsgerade mit 95%igem Konfidenzintervall dargestellt. Zur Berechnung und Darstellung der Regressionsgeraden mit Konfidenzintervall wurde die Funktion `geom_smooth()` des *ggplot2*-Softwarepakets (Wickham [32]) verwendet.

Das Konfidenzintervall der Regression des Rezyklatgehaltes gegen das Signalmaximum in der Region um 1166 cm^{-1} ist deutlich breiter als die Konfidenzintervalle der übrigen beiden Regressionsgeraden. Weiterhin ist der Signalmittelwert der Region um 725 cm^{-1} sowie das Signalmaximum der Region um 875 cm^{-1} positiv und das Signalmaximum der Region um 1166 cm^{-1} negativ mit dem Rezyklatgehalt korreliert. Dies bestätigt die Annahmen der chemischen Zusammensetzung der Proben. Der positiv mit dem Rezyklatgehalt korrelierte Signalmittelwert des Doppelpeaks in der Region 722 - 730 cm^{-1} kann der Gegenwart von kristallinem PE zugeordnet werden (Luda, Brunella und Guaratto [23], S.5). Die lineare Intensitätszunahme des Doppelpeaks mit dem Rezyklatgehalt deutet darauf hin, dass das verwendete Rezyklat einen konstanten PE-Anteil aufweist.

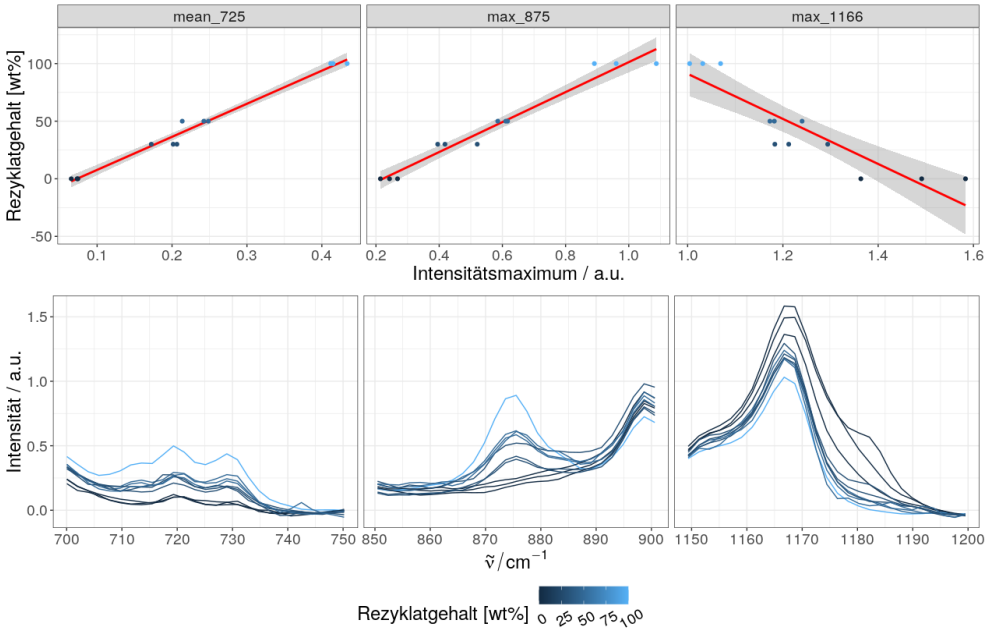


Abbildung 5.2: Auftragung des Rezyklatgehaltes gegen das Intensitätsmaximum mehrerer Regionen (oben). Zugehörige Spektralbereiche (unten).

Das ebenfalls positiv mit dem Rezyklatgehalt korrelierte Signalmaximum um 875 $cm^{-}1$ könnte auf eine Vinyliden-Gruppe des PEs zurückzuführen sein. Diese entsteht bei der Verarbeitung von PE unter hohem Druck (Long [21], S. 268).

Das negativ mit dem Rezyklatgehalt korrelierte Signalmaximum bei 1166 cm^{-1} ist der Schaukelschwingung der Methylgruppe (-CH_3) von kristallinem Polypropylen zuzuordnen (Luda, Brunella und Guaratto [23], S.5). Dass eine Verunreinigung des Reinstoffs zu einer Abnahme dieses Signals führt, ist demnach naheliegend.

Die extrahierten Signalmaxima und Signalmittelwerte der drei Regionen wurden in eine Designmatrix $\boldsymbol{X}$ überführt. Zusätzlich wurden Signalmaxi-

ma weiterer, weniger auffälliger Regionen extrahiert und der Designmatrix hinzugefügt.

5.1.2 *Nicht-negative Matrixfaktorisierung*

Ein Nachteil der Verwendung von Signalmaxima als Feature ist die starke Anfälligkeit Streuungseffekten gegenüber. Das Signalmaximum hängt definitionsbedingt nur von einem Punkt des Spektrums ab. Ein statistischer Streueinfluss auf diesen Punkt kann deshalb starken, negativen Einfluss auf die Güte eines linearen Regressionsmodells ausüben, das das Maximum als unabhängige Variable enthält.

Die Gewichtungsfaktoren der NMF-Komponenten werden hingegen auf Basis der gesamten Datenpunkte eines Spektralbereichs berechnet, was den Einfluss statistischer Streuung potenziell abschwächt. Es wurden insgesamt vier nicht-negative Matrixfaktorisierungen mit je drei Komponenten durchgeführt und die resultierenden Gewichtungsfaktoren der Designmatrix hinzugefügt. Die dafür herangezogenen Regionen sind in Tabelle 5.1 aufgeführt.

Tabelle 5.1: Zur NMF herangezogene Regionen

Featurename	Untere Grenze $[cm^{-1}]$	Obere Grenze $[cm^{-1}]$
`nmf_722`	715	735
`nmf_870`	850	890
`nmf_950`	700	1000
`nmf_1166`	1150	1175

Die Spektralbereiche um 722 cm^{-1} und 875 cm^{-1} wurden individuell mit `nmf_722` respektive `nmf_870` sowie gemeinsam mit `nmf_950` faktorisiert. Aus den vier nicht-negativen Matrixfaktorisierungen mit jeweils drei Komponenten resultierten 12 Gewichtungsfaktoren, die der Designmatrix hinzugefügt wurden.

5.1.3 *Feature Selection*

Die Designmatrix umfasste insgesamt 16 Features, wovon vier Features regionale Statistiken darstellten und 12 der Features aus den nicht-negativen Matrixfaktorisierungen resultierten. Es lagen damit mehr Features $p = 16$ als Spektren $n = 12$ des Trainingsdatensatzes vor. Um die Anzahl der Variablen auf ein Minimum zu reduzieren, wurden die in Abschnitt 4.2 vorgestellten Methoden angewendet.

Sequential Replacement

Das *Sequential Replacement*-Verfahren wurde mithilfe von zehnfach wiederholter *3-Fold* Kreuzvalidierung durchgeführt und die in Abbildung 5.3 dargestellten RMSE-Werte mit den besten Modellen einer spezifischen Variablenanzahl erreicht. Die Trainingsdaten wurden dabei bei jedem der zehn Wiederholungen in drei Teilmengen (*Folds*) aufgeteilt. Das Modell wurde auf Basis von zwei der drei Teilmengen gefittet und der Validierungsfehler aus dem Vorhersagefehler auf Basis der dritten Teilmenge berechnet.

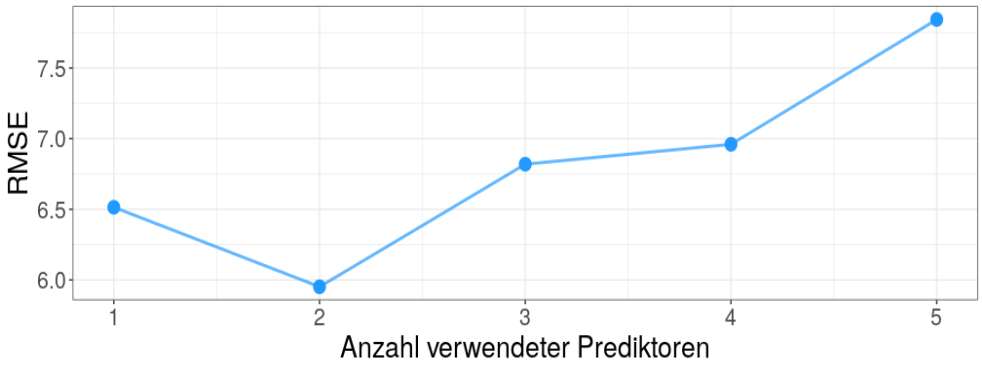


Abbildung 5.3: Validierungsfehler der besten durch Sequential Replacement ermittelten Modelle.

Tabelle 5.2: summary()-Ausgabe der ersten beiden Sequential Replacement Modelle

	Abhängige Variable:	
	Rezyklatgehalt [%wt]	
	(1)	(2)
Max. 722	283.817^{***}	
	(10.076)	
Max. 875		121.916^{***}
		(2.549)
NMF 870 K2		-2.503^{***}
		(0.236)
Konstante	-29.205^{***}	9.374^{**}
	(2.931)	(4.033)
R^2	0.988	0.997
Adj. R^2	0.986	0.997
Hinweis:	*p<0.1; **p<0.05; ***p<0.01	

Die Hinzunahme von mehr als zwei Variablen führt zu Overfitting. Das heißt, dass das Modell zwar die Trainingsdaten sehr genau beschreibt, jedoch nicht mehr gut generalisiert, was einen höheren Fehler bei der Anwendung des Modells auf die unbekannten Validierungsdaten zur Folge hat. Die durch die `summary()`-Funktion erzeugten Statistiken der durch *Sequential Replacement* ermittelten Modelle mit einer und zwei Variablen sind in Tabelle 5.2 aufgeführt.

Beide Modelle weisen hohe (adjustierte) Bestimmtheitsmaße auf und erfassen dementsprechend ein hohes Maß der Varianz der abhängigen Variable. Die Vorhersagen des Rezyklatgehaltes der Trainingsdaten weisen eine Abweichung von maximal $\pm 3\%$ auf.

LASSO-Regression

Die LASSO-Regression wurde mit dem Softwarepaket *glmnet* (Friedman, Hastie und Tibshirani [12]) durchgeführt. Hierzu wurde für den Regularisierungsparameter λ eine Parametersequenz im Bereich $\lambda \in [10^{-3}, 10^{2}]$ angelegt und zu jedem individuellen Parameter λ wurden 3-fold kreuzvalidierte Fits des LASSO-Modells berechnet.

In Abbildung 5.4 ist der Mean Squared Error (MSE) der Validierung gegen den Logarithmus des für das jeweilige Modell eingesetzten Parameters λ dargestellt. Die vertikale, gepunktete Linie links markiert den Wert λ_{min} mit dem kleinsten zugehörigen Validierungsfehler. In dem zugehörigen Modell verbleiben insgesamt acht abhängige Variablen mit Regressionskoeffizienten ungleich null. Dies ist den Werten oberhalb der Abbildung zu entnehmen. Die zweite, vertikale, gepunkte Linie markiert den Wert λ_{1se}, der den höchsten Wert für λ, dessen zugehöriger Fehler noch innerhalb einer Standardabweichung von λ_{min} liegt, darstellt.

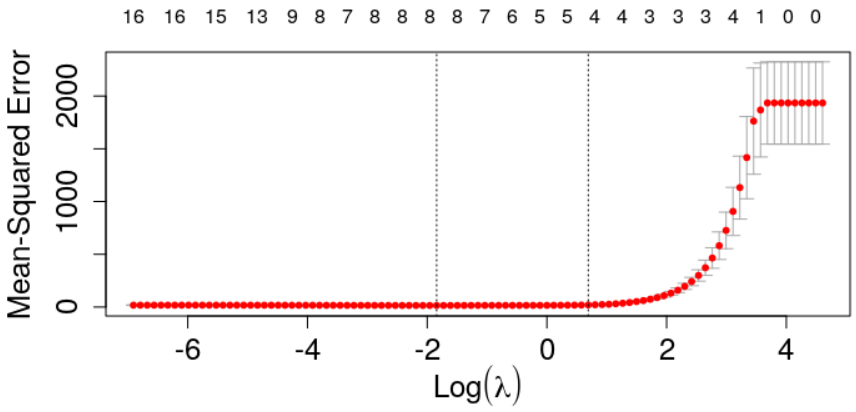


Abbildung 5.4: Auftragung des mittleren, quadratischen Validierungsfehlers gegen den natürlichen Logarithmus des Regularisierungskoeffizienten λ.

Im stärker regularisierten Modell mit Regularisierungsparameter $\lambda_{1se} \approx 1,99$ verbleiben vier Regressionskoeffizienten, die nicht durch die Regularisierung auf null gesetzt werden. Die Regularisierungspfade der Regressionskoeffizienten, abhängig von der Wahl des Regularisierungsparameters λ, sind in Abbildung 5.5 dargestellt.

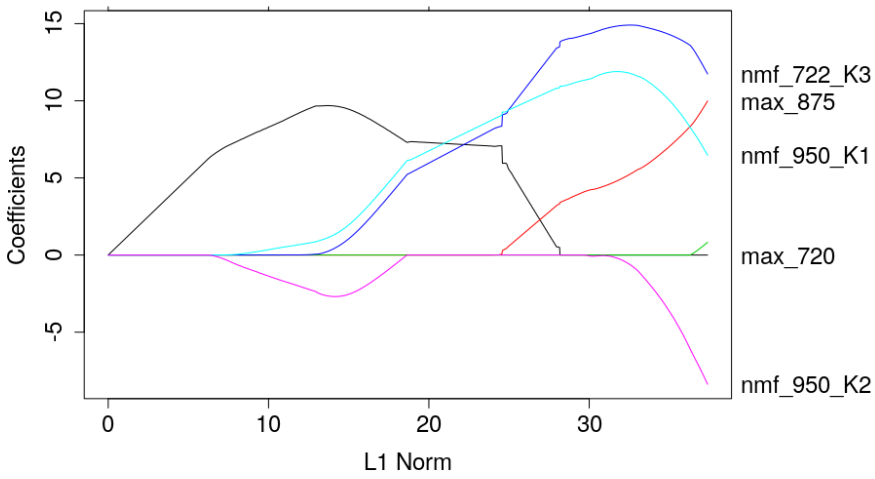


Abbildung 5.5: Regularisierungspfad der LASSO-Regression

Es ist hierbei auffällig, dass sich nur eine der Variablen, die durch *Sequential Replacement* identifiziert wurden, im LASSO-Modell wiederfinden. Bei den Variablen `nmf_722_K3` und `max_722` ist dies auf deren Austauschbarkeit aufgrund sehr hoher Korrelation ($\approx 0,99$) zurückzuführen. Im weniger stark regularisierten Modell mit Regularisierungsparamter $\lambda_{min} \approx 0,18$ verbleiben insgesamt zehn Variablen mit Regressionskoeffzienten ungleich null.

5.1.4 *Auswertung der Modelle mithilfe der Testdaten*

In Tabelle 5.3 sind die Vorhersagen des Rezyklatgehaltes der Testdaten dargestellt. Es wurden hierzu die beiden Modelle, die durch das *Sequential Replacement*-Verfahren identifiziert wurden, sowie die beiden LASSO-Modelle mit Regularisierungsparametern λ_{min} und λ_{1se} herangezogen.

Tabelle 5.3: Vorhersage des Rezyklatgehaltes der Testdaten

Wahr	SeqRep1	SeqRep2	$LASSO_{\lambda_{min}}$	$LASSO_{\lambda_{1se}}$
100.00	91.83	103.53	98.07	94.08
0.00	7.57	-0.03	3.61	11.07
30.00	25.77	27.17	26.78	27.77
50.00	52.10	53.67	51.55	47.07

Die Vorhersagen für die Reinstoffprobe weisen die stärkste Streuung auf. Die hierbei beste Vorhersage wird durch das zweite *Sequential Replacement*-

Modell getroffen. Weiterhin ist festzustellen, dass die Ergebnisse der LASSO-Modelle keine besseren Ergebnisse liefern, als die auf *Sequential Replacement* basierenden Modelle. Jedoch beziehen diese deutlich mehr Variablen mit ein.

Es werden deshalb nur die Features, die durch das *Sequential Replacement*-Verfahren identifiziert wurden, in die Modellbildung der kombinierten Analyse miteinbezogen.

5.2 AUSWERTUNG DER GPC-MESSUNGEN

Die GPC-Messdaten der Kunststoffproben liegen als Tupel vor, die sich aus der molaren Masse und dem zur molaren Masse korrespondierenden mittleren Massenanteil zusammensetzen. Zusätzlich liegen die mithilfe der proprietären Software *Omnic* bestimmten, mittleren molaren Massen zu jeder Probe vor und können unmittelbar als Feature der Designmatrix X verwendet werden. Die Trainingsdaten setzen sich wie auch bei den ATR-Messungen aus den Messungen zu den Probennummern eins, zwei und vier zusammen. Als Testdaten fungieren die Messdaten zu Probe drei sowie zwei weitere Blendmessungen, die sich aus einer Mischung der Reinstoffprobe und der zu 100% aus Rezyklat bestehenden Probe zusammensetzen. Die Rezyklatgehälter der beiden Blendproben ergeben sich zu $72,4\%$ und $84,4\%$.

Die Auftragung des Massenanteils gegen die molare Masse ist in Abbildung 5.6 dargestellt.

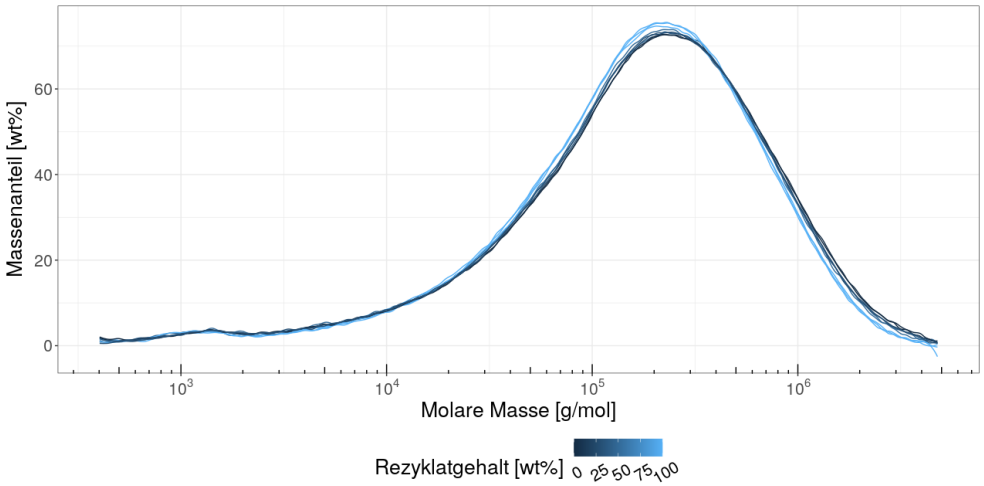


Abbildung 5.6: Auftragung des Massenanteils gegen die molare Masse

Mit steigendem Rezyklatgehalt ist eine Abnahme der mittleren molaren Masse zu verzeichnen. Dies ist bereits an der Linksverschiebung des Maximums der in Abbildung 5.6 dargestellten Kurve erkennbar.

Die Auftragung des Rezyklatgehaltes gegen die mittleren molaren Massen, die mithilfe der proprietären Software *Omnic* bestimmt wurden, bildet die-

sen Zusammenhang deutlicher ab. Diese findet sich als Scatter- und Regressionsplot mit Standardabweichung in Abbildung 5.7. Es ist ein deutlicher, linearer Zusammenhang der mittleren molaren Masse mit dem Rezyklatgehalt der Proben zu erkennen.

Die Abnahme der mittleren molaren Masse mit steigendem Rezyklatgehalt kann durch Alterungseffekte und den Einfluss des Verarbeitungsprozesses auf das Rezyklat erklärt werden. Alterung, mechanische Einwirkung sowie Wärme- und Oxidationseinflüsse bei der Verarbeitung führen zu Kettenspaltungen des wiederaufbereiteten Polymers. Dies hat eine durchschnittlich kürzere Kettenlänge des Polymers und damit unmittelbar eine geringere mittlere molare Masse zur Folge. Der Einfluss von Recycling auf die chemische Beschaffenheit von High-Density Polyethylen (HDPE) findet sich unter anderem in Loultcheva u. a. [22].

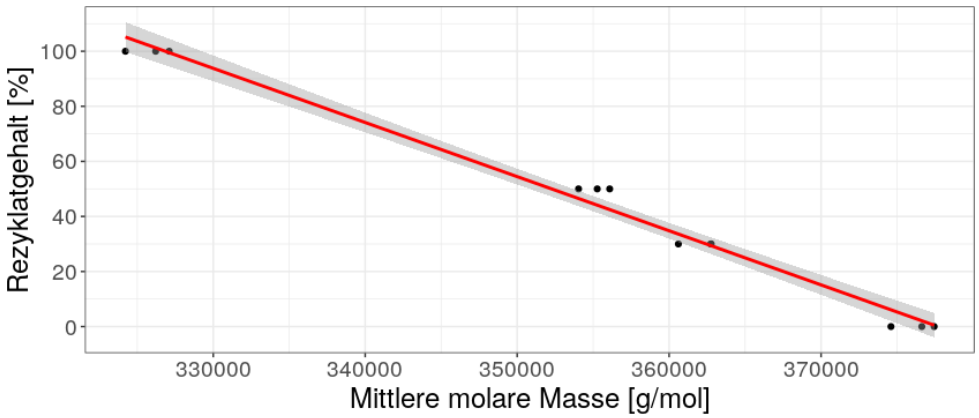


Abbildung 5.7: Auftragung des Massenanteils gegen die molare Masse

Tabelle 5.4: Ausgabe der `summary()`-Funktion für das lineare Modell

	Abhängige Variable:
	Rezyklatgehalt [%wt]
Mittlere molare Masse $\bar{M}_w$	37.705***
	(1.542)
Konstante	45.000***
	(1.476)
R^2	0.984
Adj. R^2	0.982
Hinweis:	*p<0.1; **p<0.05; ***p<0.01

In Tabelle 5.4 ist die Ausgabe der `summary()`-Funktion für das in Abbildung 5.7 dargestellte, lineare Modell als Tabelle aufgeführt. Das Bestimmtheitsmaß R^2 des Modells liegt unter den Bestimmtheitsmaßen der Modelle, die auf Basis der ATR-Daten ermittelt wurden. Die Vorhersage des Rezyklatgehaltes der Trainingsdaten weist ebenfalls eine höhere Abweichung mit maximal $\pm 6.5\%$ auf.

5.2.1 *Auswertung der Testdaten*

Die Auswertung der Testdaten mit dem in Tabelle 5.4 dargestellten Modell liefert die in Tabelle 5.5 tabellarisch aufgelisteten Ergebnisse. Ein Vorhersa-

Tabelle 5.5: Vergleich des echten Rezyklatgehaltes mit den Vorhersagen des Modells auf Basis der GPC-Testdaten

Wahr	Vorhersage
0.00	8.43
30.00	31.45
50.00	43.31
100.00	105.86
72.40	78.84
84.40	80.16

gefehler von bis zu 8,5% entspricht dem Ergebnis, das mit dem einfachen Modell `SeqRep1` auf Basis der ATR-IR-Daten, wie in Tabelle 5.3 dargestellt, erreicht wurde. Die mit der proprietären Software *Omnic* berechneten mittleren molaren Massen werden als potenzielles Feature in die Designmatrix der kombinierten Analyse mitaufgenommen.

5.3 AUSWERTUNG DER HPLC-MESSUNGEN

Wie bereits in Abschnitt 2.2 beschrieben, ist die HPLC ein nasschemisches Verfahren, bei dem die unterschiedliche chemische Beschaffenheit der Probenkomponenten ausgenutzt wird, um die Probe aufzutrennen. Zusätzlich zu den 16 Proben unterschiedlichen Rezyklatgehaltes wurde ein Standard aus bekannten Anteilen reinen Polypropylens und Polyethylens gemessen. Die Chromatogramme der Proben und des Standards sind in Abbildung 5.8 dargestellt. Die Proben sind hierbei in schwarz und die Standardmessung in rot dargestellt. Zur Detektion wurde ein Lichtstreuungsdetektor (ELSD) verwendet.

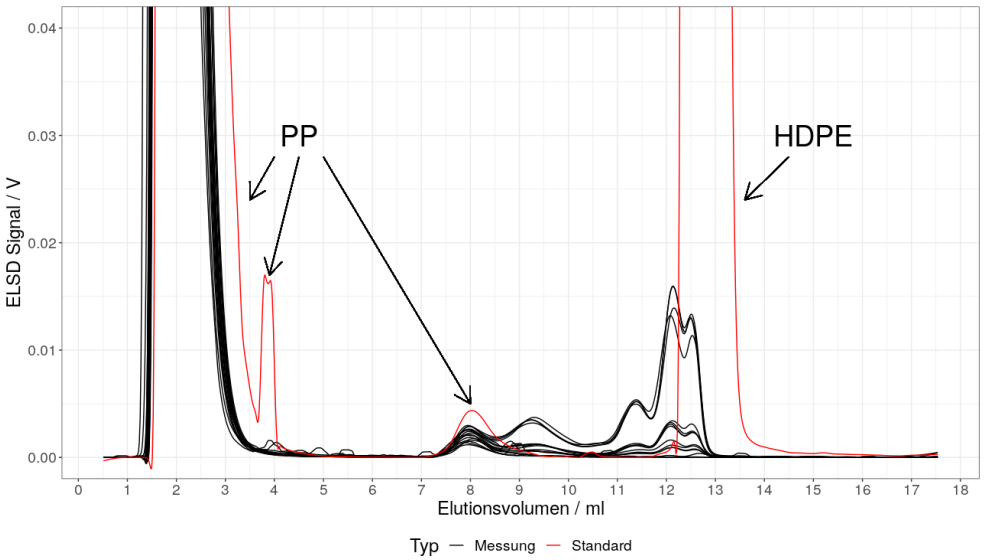


Abbildung 5.8: Chromatogramme der Probenmessungen (schwarz) und Standardmessung (rot).

Die Signale der Standardmessung bei 1,5 bis 4 ml und um 8 ml Elutionsvolumen sind hierbei Polypropylen zuzuordnen. Das Signal bei 12 bis 14 ml ist wiederum dem HDPE-Anteil des Standards zuzuordnen. Polypropylen weist im verwendeten Lösungsmittel 1-Decanol eine wesentlich kürzere Retentionszeit auf als Polyethylen (Monrabal [25], S.108).

Hohe Korrelationseffekte zwischen dem detektierten Signal und dem Rezyklatgehalt der Proben lassen sich bei Elutionsvolumina von 7 bis 13 ml feststellen. Eine Auftragung des Elutionsbereichs mit durch den Rezyklatgehalt bestimmter Färbung der Chromatogramme ist in Abbildung 5.9 abgebildet.

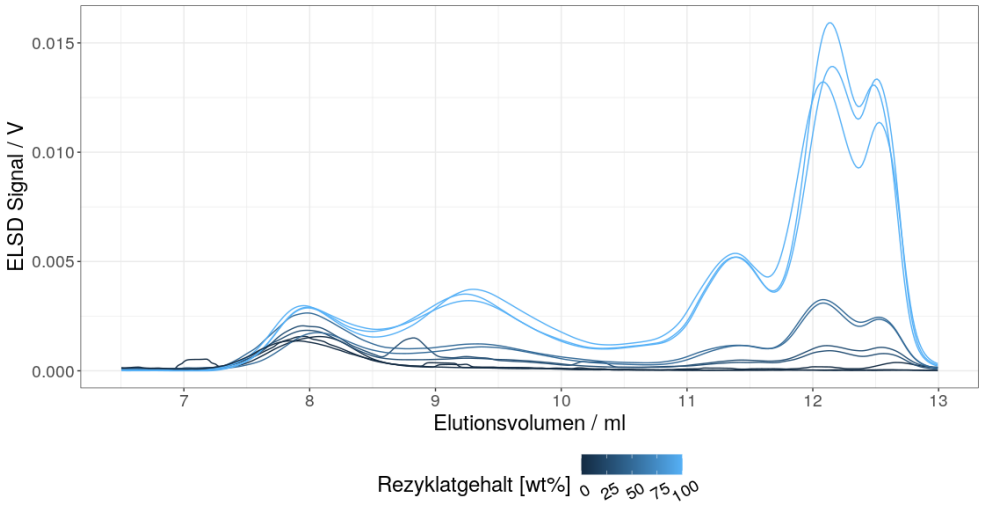


Abbildung 5.9: Elutionsbereich von 7 ml bis 13 ml mit vom Rezyklatgehalt abhängiger Färbung der Chromatogramme

Der Anstieg der Signalhöhen der beiden Doppelpeaks im Bereich 12 bis 13 ml weist einen nicht-linearen Zusammenhang zwischen Rezyklatgehalt und Signalhöhe auf. Der quadratische Zusammenhang des detektierten Signals ist dabei auf den Lichtstreuungsdetektor zurückzuführen (Héron, Dreux und Tchapla [18]).

Die Signale im Bereich 11 bis 13 ml sind dem HDPE-Anteil des Rezyklats zuzuordnen. Die Signale im Bereich von 7 bis 11 ml werden vermutlich durch Ethylen-Propylen Copolymere unterschiedlicher Mischungsverhältnisse verursacht. Dabei führt ein höherer Anteil Polyethylens zu einer höheren Retentionszeit, wie unter anderem in Monrabal [25] (S.136) dargestellt.

5.3.1 *Extraktion regionaler Statistiken*

Im Rahmen des *Feature Engineerings* wurden mehrere Signalmaxima und Signalmittelwerte extrahiert. In der Auftragung des Rezyklatgehaltes gegen die individuellen Features ist, wie angenommen, ein nicht-linearer Zusammenhang erkennbar. In den Auftragungen des Rezyklatgehaltes gegen die Prediktoren `max_114` und `mean_95`, die das Maximum des Elutionsbereichs um 11,4 ml respektive den Mittelwert des Elutionsbereichs um 9,5 ml beschreiben, ist ein quadratischer Zusammenhang zu vermuten.

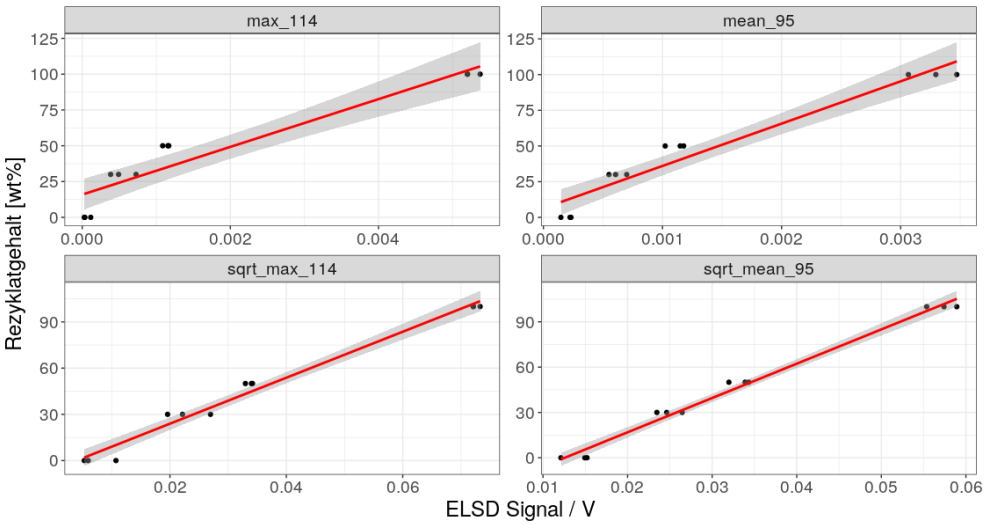


Abbildung 5.10: Auftragung des Rezyklatgehaltes gegen Signalmaximum und Signalmittelwert (oben). Auftragung des Rezyklatgehaltes gegen die Prediktoren nach einer Transformation durch Ziehen der Quadratwurzel (unten).

Die Auftragungen des Rezyklatgehaltes gegen die durch Ziehen der Quadratwurzel transformierten Prediktoren weisen einen linearen Zusammenhang auf. Weiterhin ist der als graue Schattierung dargetellte Standardfehler der Regressionsgeraden im Falle der transformierten Prediktoren deutlich kleiner. In die Designmatrix, die zur *Feature Selection* und Modellbildung

herangezogen wird, wurden ausschließlich transformierte Prediktoren aufgenommen.

5.3.2 *Nicht-negative Matrixfaktorisierung*

Es wurden mehrere Elutionsbereiche durch nicht-negative Matrixfaktorisierungen entmischt. Die resultierenden Gewichtungsfaktoren wiesen wie die regionalen Statistiken einen quadratischen Zusammenhang mit dem Rezyklatgehalt auf. Eine Linearisierung konnte durch eine Transformation durch Ziehen der Quadratwurzel erreicht werden. In Abbildung 5.11 ist das Ergebnis der nicht-negativen Matrixfaktorisierung des Elutionsbereichs von 7 bis 14 ml dargestellt. In der oberen Abbildung sind die beiden Komponentenchromatogramme, auch Loadings, zu sehen. Die erste Komponente bildet dabei die zu Ethylen-Propylen-Copolymer und Polyethylen zugehörigen Signale ab.

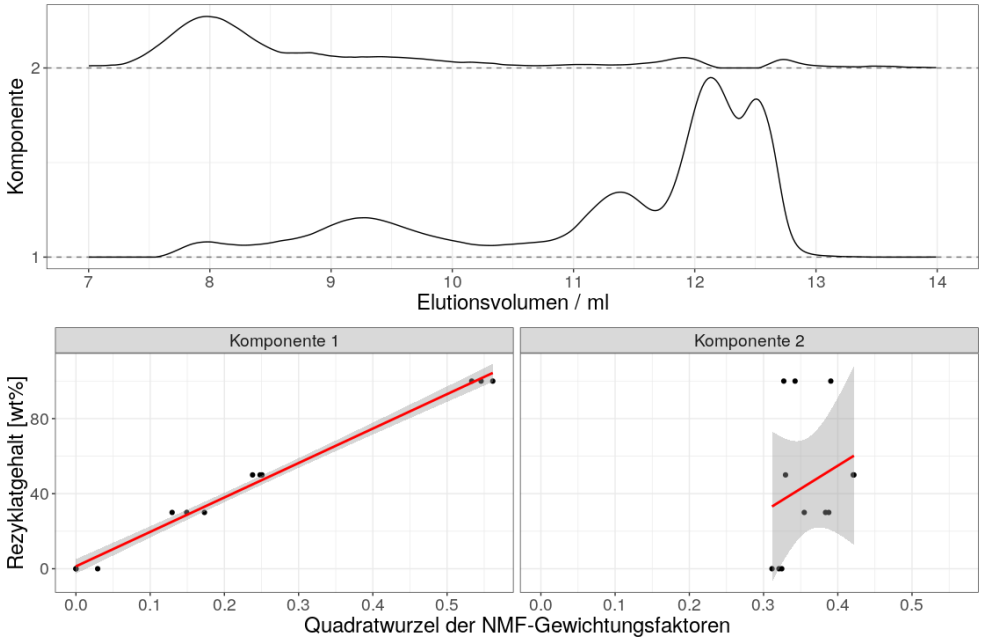


Abbildung 5.11: Nicht-negative Matrixfaktorisierung des Elutionsbereichs von 7 bis 14 ml mit zwei Komponenten. Oben: Komponentenchromatogramme. Unten: Auftragung des Rezyklatgehaltes gegen die transformierten Gewichtungsfaktoren der beiden Komponenten.

Die unteren beiden Abbildungen stellen die Auftragung des Rezyklatgehaltes der Proben gegen die transformierten Gewichtungsfaktoren der beiden NMF-Komponenten dar. Die Regressionsgerade des Rezyklatgehaltes gegen die Gewichtungsfaktoren der ersten Komponente weist dabei einen sehr guten Fit auf. Die Gewichtungsfaktoren wurden wie auch die regionalen Statistiken zunächst durch Ziehen der Quadratwurzel transformiert und der Designmatrix hinzugefügt. Insgesamt wurden vier, teilweise überlappende Re-

gionen mithilfe von nicht-negativen Matrixfaktorisierungen entmischt und die transformierten Gewichtungsfaktoren der Designmatrix hinzugefügt.

5.3.3 *Feature Selection*

Die Designmatrix bestand zunächst aus insgesamt 14 Features, wobei davon sechs Features transformierte Signalmaxima und Signalmittelwerte darstellten. Die weiteren acht Features stellten transformierte Gewichtungsfaktoren nicht-negativer Matrixfaktorisierungen dar. Zur Bestimmung der zur Vorhersage des Rezyklatgehaltes am besten geeigneten Prediktoren wurde das *Sequential Replacement*-Verfahren und LASSO-Regression angewendet.

Sequential Replacement

Mithilfe des *Sequential Replacement*-Verfahrens wurden ein bis zwei Prediktoren als die Anzahl notwendiger Prediktoren ermittelt, die zu einem minimalen Validierungsfehler des Regressionsmodells führen. Durch zehnfach wiederholte Kreuzvalidierung mit Aufteilung der Trainingsdaten in zwei Trainingsdatensätze und einen Validierungsdatensatz wurden die in Abbildung 5.12 dargestellten, gemittelten Validierungsfehler ermittelt.

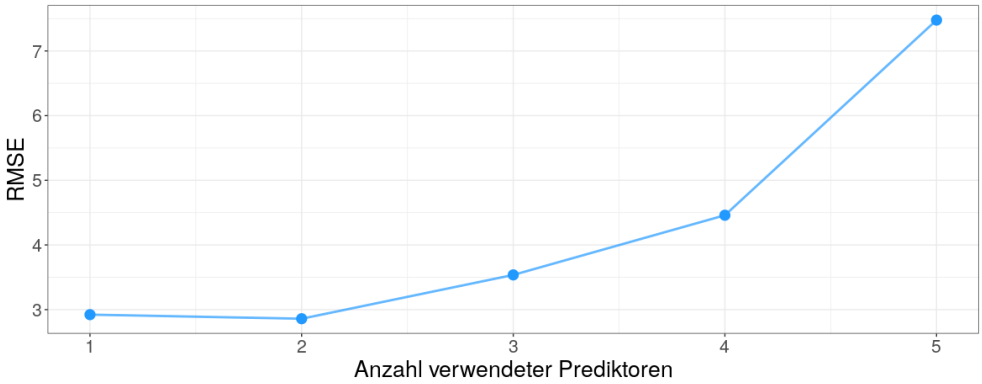


Abbildung 5.12: Validierungsfehler der besten durch Sequential Replacement ermittelten Modelle.

Die unabhängigen Variablen, die die Varianz des Rezyklatgehaltes am besten beschreiben, sind beide der NMF-Entmischung entstammende, durch Ziehen der Quadratwurzel transformierte Gewichtungsfaktoren. In Tabelle 5.6 ist die Ausgabe der `summary()` Funktion für die Modelle mit einer und zwei unabhängigen Variablen dargestellt.

Das auf HPLC-Daten basierende Modell (1) weist das höchste Bestimmtheitsmaß aller in dieser Arbeit untersuchten Einvariablenmodelle auf.

Tabelle 5.6: Ausgabe der `summary()`-Funktion der Sequential Replacement Modelle

	Abhängige Variable:	
	Rezykaltgehalt [%wt]	
	(1)	(2)
NMF 7,5 - 10 K2	232.801***	141.419***
	(5.489)	(30.506)
NMF comp. 7 - 14 K1		72.838**
		(24.097)
Konstante	−18.466***	−10.905***
	(1.724)	(2.810)
R^2	0.994	0.997
Adj. R^2	0.994	0.997
Hinweis:	*p<0.1; **p<0.05; ***p<0.01	

LASSO-Regression

Die Variablenselektion mithilfe des LASSO-Verfahrens führte zu sehr eindeutigen Ergebnissen. Wie den in Abbildung 5.13 abgebildeten Regressionspfaden zu entnehmen, führt das Regularisierungsverfahren zu einer Eliminierung aller Variablen bis auf jene, die auch durch das *Sequential Replacement*-Verfahren als am besten geeignet identifiziert werden konnten.

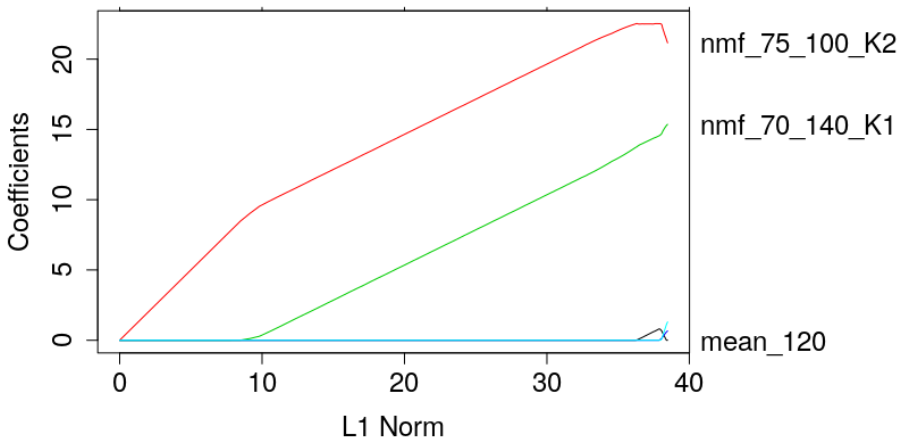


Abbildung 5.13: Auftragung der Regressionskoeffizienten gegen die Summe der Absolutbeträge der Regressionskoeffizienten (L1 Norm).

Durch Kreuzvalidierung wurden die beiden Regularisierungskoeffizienten λ_{min}, der zum niedrigsten Validierungsfehler korrespondiert, und λ_{1se}, dem höchsten Wert λ, dessen zugehöriger Validierungsfehler maximal eine Standardabweichung über dem Minimum liegt, ermittelt. Diese weisen deutlich höhere Werte auf als die ermittelten Regularisierungskoeffizienten im Fall der Analyse der ATR-Messdaten. Der höhere Regularisierungsgrad führt zu einer eindeutigen Identifizierung der beiden NMF-Komponenten als stärkste Features.

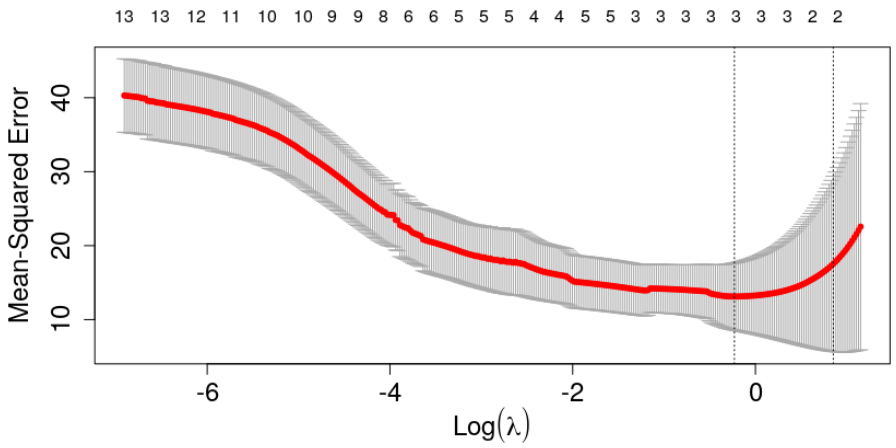


Abbildung 5.14: Auftragung des mittleren, quadratischen Validierungsfehlers gegen den natürlichen Logarithmus des Regularisierungskoeffizienten λ.

5.3.4 *Auswertung der Testdaten*

Die Vorhersagen der Rezyklatgehälter der Testdaten mithilfe der Ein- und Zweivariablenmodelle sind in Tabelle 5.7 dargestellt.

Tabelle 5.7: Vorhersage des Rezyklatgehaltes der Testdaten

Wahr	SeqRep1	SeqRep2
50	51.57	51.09
100	100.59	102.37
0	7.63	4.95
30	32.59	29.64

Die auf den HPLC-Daten basierenden Modelle weisen die niedrigsten mittleren Abweichungen von den tatsächlichen Werten auf. Die höchste Abweichung ist bei den rezyklatfreien Proben festzustellen. Dies kann darauf zurückzuführen sein, dass die angewendeten nicht-negativen Matrixfaktorisierungen keine vollständige Trennung der dem Polypropylen und Ethylen-

Propylen-Copolymers zuzuordnenden Signale um 8 ml erreichen. Betrachtet man die in Abbildung 5.11 dargestellten Komponentenspektren (oben), lassen sich in Komponente eins und zwei Signale um 8 ml feststellen. Dies wiederum könnte zu einer fehlerhaften Registrierung von Ethylen-Propylen-Copolymer und damit einer zu hohen Vorhersage des Rezyklatgehaltes der Reinstoffprobe führen.

5.4 KOMBINIERTE ANALYSE

In der abschließenden, kombinierten Analyse wurden insgesamt sechs Prediktoren berücksichtigt. Dies sind die Prediktoren, die im Rahmen der individuellen Analysen durch *Sequential Replacement* und LASSO-Regression bestimmt wurden. Diese setzen sich aus Signalmaxima und Gewichtungsfaktoren regionsspezifischer nicht-negativer Matrixfaktorisierungen zusammen und sind in Tabelle 5.8 aufgelistet.

Tabelle 5.8: Im Rahmen der kombinierten Analyse verwendete Prediktoren

Bezeichner	Beschreibung
ATR Max. 875	Signalmaximum bei 875 cm^{-1}
ATR Max. 722	Signalmaximum bei 722 cm^{-1}
ATR NMF 870 K2	Gewichtungsfaktoren der 2. NMF-Komponente der Region um 870 cm^{-1}
GPC $\bar{M}_w$	Mittlere molare Masse
HPLC NMF 70 - 140 K1	Gewichtungsfaktoren der 2. NMF-Komponente der Region von 7 bis 14 ml
HPLC NMF 70 - 100 K2	Gewichtungsfaktoren der 2. NMF-Komponente der Region von 7 bis 10 ml

5.4.1 *Feature Selection mit Sequential Replacement*

Mit dem zehnfach kreuzvalidiert durchgeführten *Sequential Replacement* Verfahren wurden die in Abbildung 5.15 dargestellten, mittleren Validierungsfehler berechnet. Der mittlere Validierungsfehler weist von Modell zu Modell nur eine sehr geringe Variation auf. Dies ist in der Ausgabe der `summary()` Funktion der Modelle mit einer bis vier unabhängigen Variablen, wie in Tabelle 5.9 dargestellt, besonders gut an der Variation des Bestimmtheitsmaßes und des adjustierten Bestimmtheitsmaßes zu sehen.

Eine weitere interessante Eigenschaft der Modelle mit zwei und drei unabhängigen Variablen ist die homogene Herkunft der unabhängigen Variablen. Modell (2) beinhaltet ausschließlich unabhängige Variablen, die aus

den HPLC-Messdaten stammen und Modell (3) beinhaltet ausschließlich Variablen, die den ATR-Messdaten entstammen.

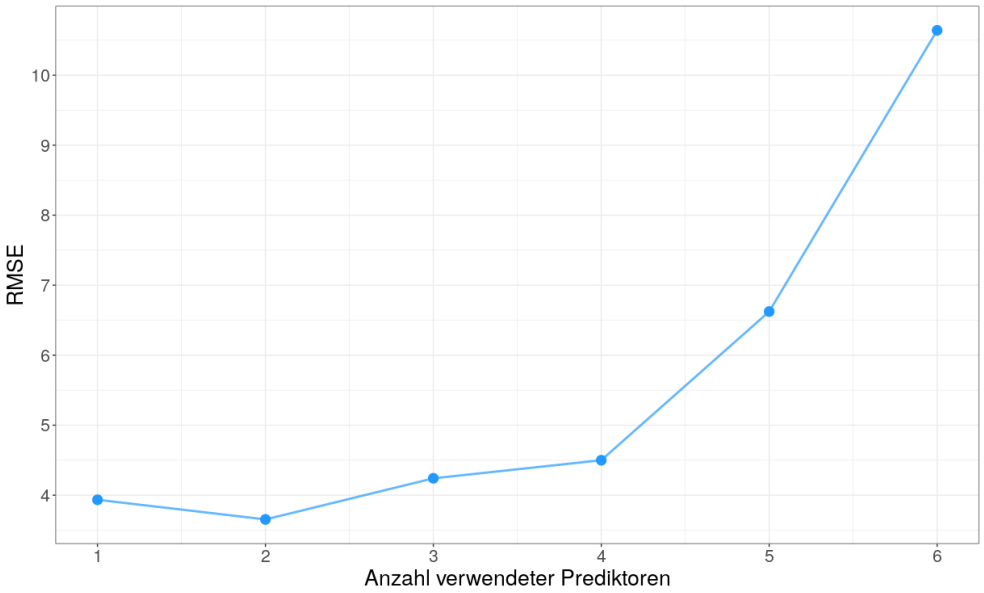


Abbildung 5.15: Validierungsfehler der besten durch Sequential Replacement ermittelten Modelle.

Modell (4) setzt sich aus unabhängigen Variablen zusammen, die alle der verwendeten Messverfahren miteinbeziehen. Die p-Werte der HPLC- und GPC-Variablen unterschreiten jedoch nicht einmal das 10% Signifikanzniveau. Die GPC-Variable weist zusätzlich einen Standardfehler auf, der dem Zweifachen des zugehörigen Regressionskoeffizienten entspricht. Das heißt, dass der tatsächliche Beitrag der GPC-Variablen zur Erklärung der Varianz des Rezyklatgehaltes äußerst ungewiss ist und null im Bereich der erwartbaren Werte für jenen Regressionskoeffizienten liegt.

Tabelle 5.9: Ausgabe der `summary()`-Funktion der Sequential Replacement Modelle

	Abhängige Variable:			
	Rezyklatgehalt [%wt]			
	(1)	(2)	(3)	(4)
HPLC NMF 7,5 - 10 K2	232.801***	141.419***		74.412
	(5.489)	(30.506)		(41.242)
HPLC NMF 7 - 14 K1		72.838**		
		(24.097)		
ATR Max. 875			114.723***	96.632**
			(21.696)	(37.069)
ATR Max. 722			16.889	
			(50.549)	
ATR NMF 870 K2			−2.358***	−2.069**
			(0.500)	(0.865)
GPC $\bar{M}_w$				0.0002
				(0.0004)
Konstante	−18.466***	−10.905***	7.094	−82.636
	(1.724)	(2.810)	(8.037)	(156.533)
R^2	0.994	0.997	0.997	0.998
Adj. R^2	0.994	0.997	0.996	0.997
Hinweis:				*p<0.1; **p<0.05; ***p<0.01

5.4.2 *LASSO-Regression*

In Abbildung 5.16 sind die Regularisierungspfade der mit den sechs Prediktoren durchgeführten LASSO-Regression abgebildet. Mit einer steigenden Regularisierung und damit einer kleineren L1-Norm tendiert das LASSO-Modell dazu, ausschließlich die den HPLC-Messdaten entstammenden, unabhängigen Variablen im Modell zu behalten. Dies deckt sich mit den Ergebnissen des *Sequential Replacement*-Verfahrens, das für das Zweivariablenmodell, welches ausschließlich den HPLC-Messdaten entstammende Variablen beinhaltet, den niedrigsten Validierungsfehler ermittelt.

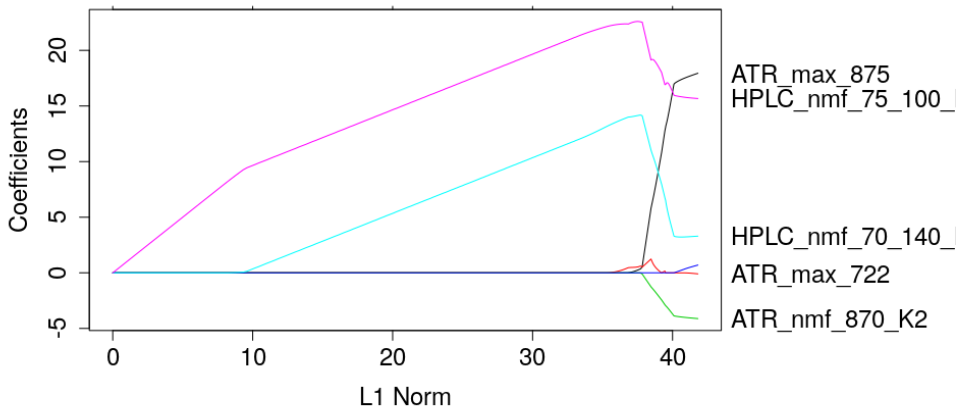


Abbildung 5.16: Regularisierungspfad der LASSO-Regression

5.4.3 *Vorhersage des Rezyklatgehaltes der Testdaten*

Zur Vorhersage des Rezyklatgehaltes der Testdaten wurden die vier, in Tabelle 5.9 dargestellten Modelle verwendet. Die Vorhersageergebnisse sind in Tabelle 5.10 aufgeführt. Die höchste Abweichung tritt bei der Vorhersage des Rezyklatgehaltes des Reinstoffs durch `SeqRep1`, dem Einvariablenmodell, auf. In den Modellen mit mehr als einer unabhängigen Variable liegt das Abweichungsmaximum bei ca. 5%.

Tabelle 5.10: Vorhersagen der Modelle der kombinierten Analyse

Wahr	SeqRep1	SeqRep2	SeqRep3	SeqRep4
0.00	7.63	4.95	0.39	1.37
30.00	32.59	29.64	27.07	28.33
50.00	51.57	51.09	53.59	54.54
100.00	100.59	102.37	102.88	102.53

Zur Wahl eines *besten* Modells sind mehrere Faktoren miteinzubeziehen. Aufgrund der hohen Abweichung bei der Vorhersage des Rezyklatgehaltes des Reinstoffs durch Modell `SeqRep1` und aufgrund des Miteinbezugs von insgesamt drei Messverfahren im Falle des Modells `SeqRep4` sind diese beiden Modelle niedriger zu priorisieren als die Modelle `SeqRep2` und `SeqRep3`.

Letztere beiden Modelle liefern Ergebnisse vergleichbarer Güte. Jedoch basiert Modell `SeqRep2` auf den Messdaten des HPLC-Verfahrens und `SepRep3` auf spektroskopischen ATR-IR Messungen. Aus ökonomischen Gesichtspunkten fällt die Wahl damit klar auf Modell `SeqRep3`, da ATR-IR Messungen nur einen Bruchteil des Aufwands, der für die Durchführung von HPLC-Messungen notwenig ist, bedürfen. Dabei kann das Kunststofferzeugnis, wie

in Abschnitt 2.3 beschrieben, ohne besondere Vorbereitung auf dem ATR-Element des Infrarotmessgeräts fixiert und vermessen werden.

Im Fall der HPLC-Messung muss ein Teil der Proben entnommen und mithilfe eines passenden Lösungsmittels in Lösung gebracht werden. Erst dann können die gelösten Proben eluiert und gemessen werden. Dieses Verfahren ist zusätzlich zu dem höheren zeitlichen Aufwand ein destruktives Verfahren, das mit jedem Messvorgang einen Teil der Probe verbraucht.

6 DISKUSSION DER ERGEBNISSE

Mit den in dieser Arbeit vorgestellten Methoden konnte gezeigt werden, dass der Rezyklatgehalt vorliegender Polypropylenmuster durch lineare Regressionmodelle mit minimalem Aufwand mit einer maximalen Abweichung von $\pm 3\%$ Genauigkeit vorhergesagt werden kann.

Die Analyse der GPC-Messdaten ergab einen negativen linearen Zusammenhang des Rezyklatgehaltes mit der mittleren molaren Masse der Probe, was auf eine durch die Aufbereitung des Materials verursachte Verkürzung der durchschnittlichen Polymerkettenlänge zurückzuführen ist (Loultcheva u. a. [22]).

Im Rahmen der Analyse der ATR-IR-Messdaten konnte eine positive Korrelation des Rezyklatgehaltes mit Polyethylen zuzuordnenden Signalen festgestellt werden. Auch bei der Analyse der HPLC-Messdaten wurde eine hohe Korrelation zwischen dem Rezyklatgehalt der Proben und HDPE sowie Ethylen-Propylen-Copolymer zuzuordnenden Signalen festgestellt.

Die Bestimmung des Rezyklatgehaltes der vorliegenden Polypropylenstäbe konnte mithilfe der ermittelten Modelle mit Erfolg durchgeführt werden. Jede der individuellen Analysen lieferte dabei bereits Ergebnisse mit einer maximalen Abweichung von $\pm 8,5\%$, wobei die besten Ergebnisse mit den Messdaten der ATR-IR- und HPLC-Messungen erzielt werden konnten. Dies folgte unmittelbar aus der Analyse der kombinierten Messdaten und hierbei im Speziellen aus der *Feature Selection*. Aufgrund der einfacheren Handhabbarkeit des ATR-IR-Messverfahrens bietet sich ein auf jenen Messdaten basierendes Modell zur Verwendung im industriellen Kontext an.

Mithilfe der Programmiersprache R konnten die notwendigen Verarbeitungsschritte des *Data Loadings*, *Preprocessings* und *Feature Engineerings* stark vereinfacht werden. Zusätzliche Daten können somit durch wenige Funktionsaufrufe in ein Format überführt werden, das eine Vorhersage des Rezyklatgehaltes erlaubt. Für eine skalierbare Anwendung in einer industriellen Produktionsumgebung sollte jedoch eine Portierung der Methoden zu einer robusteren Software in Erwägung gezogen werden. Auch die Verwaltung der anfallenden Daten sollte hierbei versioniert und automatisiert geschehen.

Trotz der positiven Ergebnisse dieser Arbeit muss berücksichtigt werden, dass die überschaubare Stichprobengröße von 16 Proben eine endgültige Aussage über die Generalisierbarkeit und Güte des Vorhersageverfahrens

nur bedingt zulässt. Hierzu ist eine größere Datengrundlage notwendig, die im weiteren Verlauf des Projekts geschaffen wird.

6.1 AUSBLICK

Es sind vielfältige, an den Ergebnissen dieser Arbeit anknüpfende Schritte denkbar. Hervorzuheben ist hierbei die weitere Validierung des Verfahrens mit zusätzlichen Daten, die bestenfalls eine höhere Diversität des vorliegenden Rezyklatgehaltes aufweisen. Weiterhin ist zu untersuchen, wie gut das erarbeitete Verfahren auf Kunststoffproben unterschiedlicher Zusammensetzung generalisiert werden kann. Mit einer anderen Zusammensetzung kann dabei sowohl das Grundmaterial, das in dieser Arbeit Polypropylen darstellte, als auch die Zusammensetzung des Rezyklats gemeint sein.

Ein erster Prototyp für die industrielle Nutzung des in dieser Arbeit vorgestellten Analyseverfahrens konnte bereits realisiert werden. Es handelt sich dabei um eine Webapplikation zur Vorhersage des Vinylacetatgehaltes vorliegender Polymermischungen. Die Weboberfläche der Appliaktion ist in Abbildung 6.1 dargestellt.

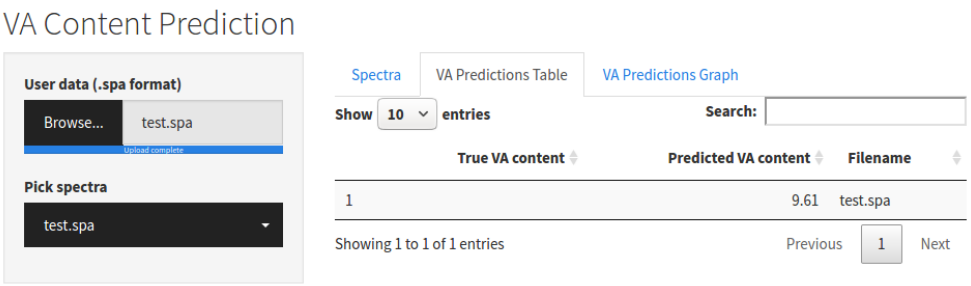


Abbildung 6.1: Oberfläche eines ersten industriell verwendeten Prototyps

Der Applikation können im binären Dateiformat *.spa* vorliegende Infrarot (IR)-Spektraldaten übergeben werden, die daraufhin, wie in Kapitel 3 beschrieben, vorverarbeitet werden. Mithilfe der transformierten Daten und eines trainierten Regressionsmodells wird der Vinylacetatgehalt der korrespondierenden Proben vorhergesagt. Es erfolgt daraufhin eine tabellarische sowie grafische Ausgabe dieser. Die Anwendung wird bereits industriell zur Qualitätskontrolle von Stichproben eingesetzt.

Wie bereits in Kapitel 6 beschrieben, ist die Verwendung eines auf ATR-IR-Messdaten basierenden Modells im industriellen Kontext aufgrund der guten Handhabbarkeit und des niedrigen Aufwands des Verfahrens erstrebenswert. Ein äußerst interessanter Forschungsgegenstand wäre die Untersuchung einer Eignung durch IR-Handgeräte aufgenommener Daten zur Vorhersage des Rezyklatgehaltes. Dies würde den Prozess der Datengenerierung deutlich vereinfachen.

Abgesehen von Verbesserungen des statistischen Verfahrens selbst, eröffnen sich eine Vielzahl logistischer Fragestellungen im Hinblick auf die Inbetriebnahme jenes Verfahrens im industriellen Kontext. Es wäre zu untersuchen wie ein solches Vorhersagemodell möglichst automatisiert in den Betriebsablauf eingebunden und hierbei beispielsweise zur Anomaliedetektion im Rahmen der Qualitätssicherung verwendet werden kann.

LITERATUR

[1] Thomas Lumley based on Fortran code by Alan Miller. *leaps: Regression Subset Selection*. R package version 3.1. 2020. URL: https://CRAN.R-project.org/package=leaps.

[2] Erik Andreassen. "Infrared and Raman spectroscopy of polypropylene". In: *Polypropylene : an A-Z reference*. Jan. 1999, S. 320–328. ISBN: 978-94-011-4421-6. DOI: 10.1007/978-94-011-4421-6_46.

[3] Ruth Asensio, Margarita Moya, José Roja und Marisa Gómez. "Analytical characterization of polymers used in conservation and restoration by ATR-FTIR spectroscopy". In: *Analytical and bioanalytical chemistry* 395 (Okt. 2009), S. 2081–96. DOI: 10.1007/s00216-009-3201-2.

[4] Peter Atkins und Julio Paula. *Atkins' physical chemistry*. Oxford University press, 2008. ISBN: 9780195685220. URL: http://www.worldcat.org/isbn/9780195685220.

[5] N. Bahlouli, D. Pessey, S. Ahzi und Y. RÉmond. "Mechanical behavior of composite based polypropylene: Recycling and strain rate effects, 8th International Conference on Mechanical and Physical Behaviour of Materials under Dynamic Loading." In: *Journal of Physics D: Applied Physics* 134 (2006), pp. 1319–1323. URL: https://hal.archives-ouvertes.fr/hal-00097033.

[6] Claudia Beleites und Valter Sergo. *hyperSpec: a package to handle hyperspectral data sets in R*. R package version 0.99-20200213.1. 2020. URL: https://github.com/cbeleites/hyperSpec.

[7] M. Biron. "Thermoplastics and Thermoplastic Composites: Second Edition". In: *Thermoplastics and Thermoplastic Composites: Second Edition* (Nov. 2012), S. 1–1044.

[8] William S. Cleveland. "Robust Locally Weighted Regression and Smoothing Scatterplots". In: *Journal of the American Statistical Association* 74.368 (1979), S. 829–836. DOI: 10.1080/01621459.1979.10481038.

[9] Stacy C Davis, Susan W Diegel, Robert G Boundy u. a. *Transportation energy data book*. Techn. Ber. Oak Ridge National Laboratory, 2020.

[10] Paul Eilers und Hans Boelens. "Baseline Correction with Asymmetric Least Squares Smoothing". In: *Unpubl. Manuscr* (Nov. 2005).

[11] Julian James Faraway. *Linear models with R*. Bd. 63. Texts in statistical science. 2005, S. x + 229. ISBN: 1-58488-425-8.

[12] Jerome Friedman, Trevor Hastie und Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), S. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[13] George Fulton und Artem Lunev. "Probing the correlation between phase evolution and growth kinetics in the oxide layers of tungsten using Raman spectroscopy and EBSD". In: *Corrosion Science* 162 (Jan. 2020), S. 108221. DOI: 10.1016/j.corsci.2019.108221.

[14] D. Gruden. *Umweltschutz in der Automobilindustrie: Motor, Kraftstoffe, Recycling*. ATZ/MTZ-Fachbuch. ISBN: 9783834895264.

[15] D.C. Harris. *Quantitative Chemical Analysis*. Macmillan Learning, 2015. ISBN: 9781464135385. URL: https://books.google.de/books?id=b3UhrgEACAAJ.

[16] Trevor Hastie, Robert Tibshirani und Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[17] Antonio Hernando, Jesús Bobadilla und Fernando Ortega. "A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model". In: *Knowl. Based Syst.* 97 (2016), S. 188–202.

[18] Sylvie Héron, Michel Dreux und Alain Tchapla. "Post-column addition as a method of controlling triacylglycerol response coefficient of an evaporative light scattering detector in liquid chromatography –evaporative light-scattering detection". In: *Journal of Chromatography A* 1035.2 (2004), S. 221 –225. ISSN: 0021-9673. DOI: https://doi.org/10.1016/j.chroma.2004.02.052. URL: http://www.sciencedirect.com/science/article/pii/S0021967304002870.

[19] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-86. 2020. URL: https://CRAN.R-project.org/package=caret.

[20] Kristian Liland, Trygve Almøy und Bjørn-Helge Mevik. "Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra". In: *Applied spectroscopy* 64 (Sep. 2010), S. 1007–16. DOI: https://doi.org/10.1366/000370210792434350.

[21] D. A. Long. "Infrared and Raman characteristic group frequencies. Tables and charts George Socrates John Wiley and Sons, Ltd, Chichester, Third Edition, 2001. Price £135". In: *Journal of Raman Spectroscopy* 35.10 (2004), S. 905–905. DOI: 10.1002/jrs.1238. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrs.1238. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrs.1238.

[22] M.Kostadinova Loultcheva, M. Proietto, N. Jilov und F.P. [La Mantia]. "Recycling of high density polyethylene containers". In: *Polymer Degradation and Stability* 57.1 (1997), S. 77 –81. ISSN: 0141-3910. DOI: https://doi.org/10.1016/S0141-3910(96)00230-3. URL: http://www.sciencedirect.com/science/article/pii/S0141391096002303.

[23] Maria Luda, Valentina Brunella und D. Guaratto. "Characterisation of Used PP-Based Car Bumpers and Their Recycling Properties". In: *ISRN Materials Science* 2013 (März 2013). DOI: 10.1155/2013/531093.

[24] Puneet Mishra, Christophe B.Y. Cordella, Douglas Rutledge, Pilar Barreiro, Jean-Michel Roger und Belén Diezma. "Application of independent components analysis with the JADE algorithm and NIR hyperspectral imaging for revealing food adulteration". In: *Journal of Food Engineering* 168 (Juli 2016), S. 7–15. DOI: 10.1016/j.jfoodeng.2015.07.008.

[25] Benjamin Monrabal. "Harald Pasch and Muhammad Imran Malik: Advanced separation techniques for polyolefins". In: *Analytical and Bioanalytical Chemistry* 407.12 (2015), S. 3269–3270.

[26] J. Neter, M. H. Kutner, C. J. Nachtsheim und W. Wasserman. *Applied Linear Statistical Models*. Chicago: Irwin, 1996.

[27] V.Paul Pauca, Farial Shahnaz, Michael Berry und Robert Plemmons. "Text Mining Using Non-Negative Matrix Factorizations." In: *SIAM Proceedings Series* (Apr. 2004). DOI: 10.1137/1.9781611972740.45.

[28] Giuseppe Ragosta, Pellegrino Musto, Pietro Russo, Giovanni Camino und Luciano Di Maio. "Recycling of Polypropylene Based Car Bumpers: Mechanical and Morphological Analysis". In: *Progress in Rubber, Plastics and Recycling Technology* 19 (Feb. 2003), S. 1–15. DOI: 10.1177/147776060301900101.

[29] Gilbert Strang. *Introduction to Linear Algebra*. Fourth. Wellesley, MA: Wellesley-Cambridge Press, 2009. ISBN: 9780980232714.

[30] Guoqing Wang, Qingzhu Ding und Zhenyu Hou. "Independent component analysis and its applications in signal processing for analytical chemistry". In: *TrAC Trends in Analytical Chemistry* 27 (Apr. 2008), S. 368–376. DOI: 10.1016/j.trac.2008.01.009.

[31] Ron Wehrens. *Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Heidelberg: Springer, 2011. DOI: 10.1007/978-3-642-17841-2.

[32] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[33] Hadley Wickham u. a. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), S. 1686. DOI: 10.21105/joss.01686.