5 Datensätze und Software

Dieses Kapitel enthält die Beschreibung der Datensätze und Software, die in der vorliegenden Arbeit zum Einsatz kamen. Die Abschnitte 5.1 und 5.2 befassen sich näher mit simulierten und realen Datensätzen und deren Eigenschaften. Bei allen Datensätzen handelt es sich um digitalisierte Spektren, die für (quantitative) multivariate Kalibrationen und deren Validation genutzt werden. Jedem der Spektren werden hierbei ein oder mehrere Eigenschaftswerte (Properties, im allgemeinen die Konzentration eines Stoffes) zugeordnet.

Abschnitt 5.3 bescheibt die für die Berechnungen eingesetzte Software (kommerzielle und Eigenentwicklung) und die verwendete Hardware.

5.1 Simulierte Datensätze

Zur Austestung und Optimierung des Genetischen Algorithmus im Rahmen der PCR wurde auf Datensätze mit simulierten Spektrenreihen zurückgegriffen. Die Spektren weisen zwei Banden auf. Es wurden Störungen wie Rauschen und einige Basislinieneffekte simuliert.

5.1.1 Spektrenberechnung

Die Erzeugung der Spektren erfolgte mit einem BASIC-Programm [94], das für die automatische Berechnung von Spektren erweitert wurde und Datensätze im JCAMP-Format liefert.

Die Spektren umfassen jeweils 257 Datenpunkte in einem willkürlich gewählten Wellenzahlenbereich von $1600\,\mathrm{cm}^{-1}$ bis $1344\,\mathrm{cm}^{-1}$ mit einer digitalen Auflösung von $1\,\mathrm{cm}^{-1}$. Es wird ein 2–Komponentensystem (A / B) angenommen, mit Banden bei $1500\,\mathrm{cm}^{-1}$ (A) und $1460\,\mathrm{cm}^{-1}$ (B). In der IR-Spektroskopie ist es üblich, zur Berechnung der Banden ein Lorentzprofil heranzuziehen. Die allgemeine Gleichung lautet:

$$y = \frac{1}{1 + a \cdot x^2} \quad . \tag{5.1}$$

Auf die in der IR–Spektrochemometrie üblichen Bandenparameter bezogen, ergibt

sich:

$$A(W_j) = A(W_m) \cdot \frac{H^2}{H^2 + 4 \cdot (W_j - W_m)^2} \quad . \tag{5.2}$$

Dabei sind:

$$\begin{aligned} a &= \text{Konstante,} \\ A(W_j) &= \text{Absorption an der Wellenzahl } j, \\ A(W_m) &= \text{Absorption im Bandenmaximum,} \\ H &= \text{Halbwertsbreite.} \end{aligned}$$

Die gewählten Parameter der einzelnen Banden sowie die minimalen bzw. maximalen Absorptionswerte sind in Tabelle 5.1 aufgeführt. Für jede Reihe wurden 20 Kalibrations- und zweimal 10 Validationsspektren erzeugt. Um reale Situationen

Tabelle 5.1: Banden-Parameter

Bande		A	B
Wellenzahl	/ cm^{-1}	1500	1460
Halbwertsbreite	/ cm^{-1}	70	40
min. Absorption	/ AE	0.440	0.303
max. Absorption	/ AE	1.020	0.680

AE = Absoptionseinheiten

zu simulieren, wurden die Spektren mit unterschiedlichen Störungen wie Rauschen, Basislinieneffekten und interferierenden Banden überlagert. Ein Beispiel für einen solchen Kalibrationsdatensatz ist in Abbildung 5.1 wiedergegeben.

5.1.2 Basislinieneffekte

Um Variationen der Basislinie zu simulieren, wurden Polynome nullten bis dritten Grades generiert und zu den Spektren addiert. Innerhalb eines Spektrensatzes wurden die Koeffizienten der einzelnen Funktionen zufälligen Schwankungen innerhalb einer bestimmten Intervallbreite unterworfen. Diese betrug für Polynome 2. Grades $\pm 10\%$ des Ausgangswertes, für Polynome 3. Grades $\pm 1\%$ des Ausgangswertes.

Die folgenden 15 Basislinieneffekte wurden generiert:

- Offsets (2 Spektrensätze),
- lineare Basislinien: drei verschiedene Steigungen, jeweils mit positivem und negativen, Vorzeichen (6 Spektrensätze),

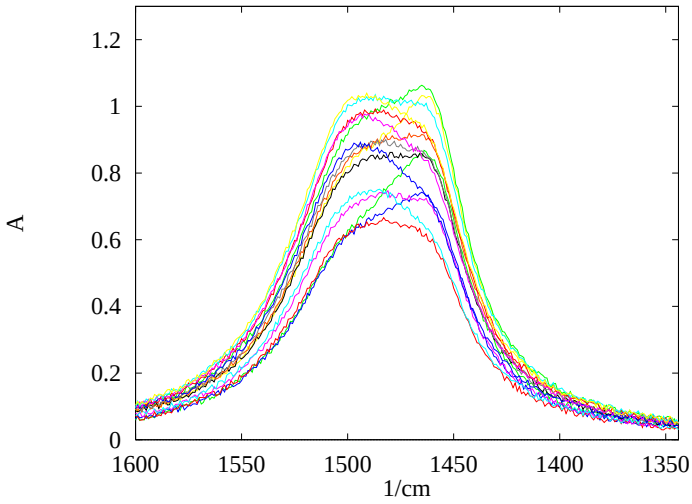


Abbildung 5.1: Spektrensatz mit zusätzlicher Störbande bei 1480 cm^{-1}.

- Polynome 2. Grades, je zwei mit gleichen Koeffizienten, aber entgegengesetzten Vorzeichen (4 Spektrensätze),
- Polynome 3. Grades, mit gleichen Koeffizienten, aber entgegengesetzten Vorzeichen (2 Spektrensätze),
- eine zusätzlich unterlegte Störbande (1 Spektrensatz).

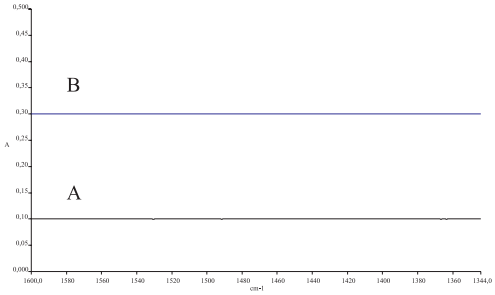
In Abbildung 5.2 werden diese Basislinieneffekte demonstriert. Zu den Spektren wurde darüberhinaus gleichverteiltes Rauschen mit einer Amplitude von ± 0.01 AE addiert.

5.2 Reale Datensätze

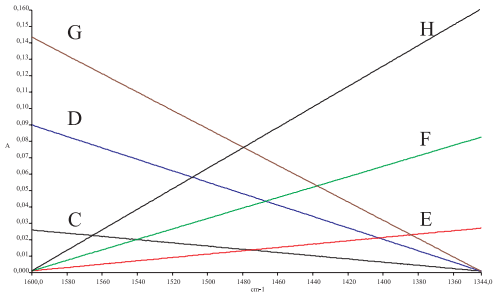
Die simulierten Datensätze repräsentieren vereinfachte und künstliche Situationen. Um einen Eindruck davon zu bekommen, wie der Genetische Algorithmus sich bezüglich der Faktor–Selektion bei realen Daten verhält, wurden Spektrensätze aus der NIR–spektrometrischen Praxis herangezogen.

5.2.1 Weizen–Spektren

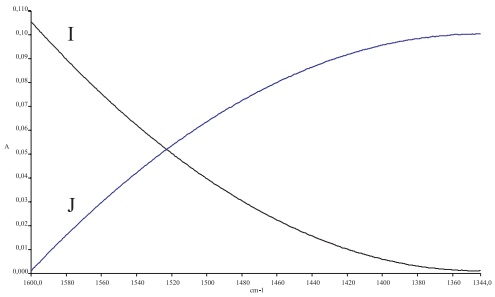
Bei den Untersuchungen wurde auf einen von *J. H. Kalivas* veröffentlichten Weizen-Datensatz [95, 96] zurückgegriffen (*s.* Abb. 5.3). Motivation für die Verwendung dieser über das Internet frei verfügbaren Daten war, daß die Kalibration der Feuch-



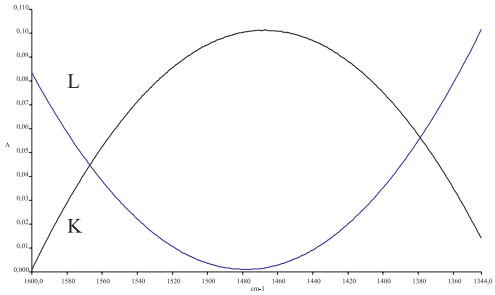
Offsets



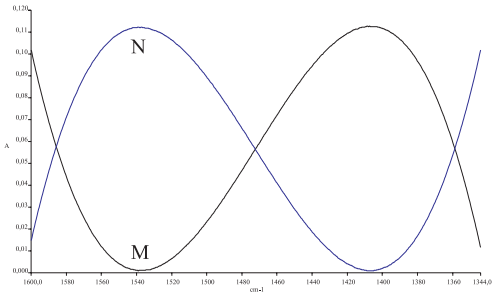
Polynom 1. Grades



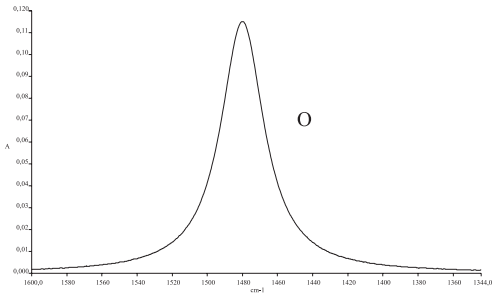
Polynom 2. Grades



Polynom 2. Grades



Polynom 3. Grades



Bande

Abbildung 5.2: Basislinieneffekte.

tigkeit und des Proteingehaltes von Getreiden eine klassische Anwendung der NIR–Spektrochemometrie darstellt und daher hohe praktische Relevanz aufweist.

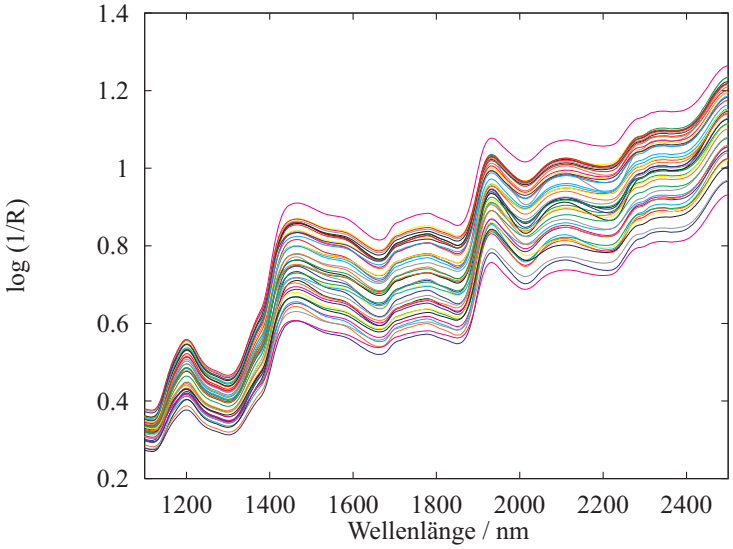


Abbildung 5.3: Spektrensatz von Weizenkörnern.

Der Datensatz beinhaltet 100 Spektren von Weizen–Proben mit spezifiziertem Feuchtigkeits- und Proteingehalt, die im Spektralbereich von 1100 bis 2500 nm in 2 nm Intervallen in diffuser Reflexion vermessen wurden. Von diesen Spektren wurden 87 (analog zu [95]) in drei Sets aufgeteilt: Ein Kalibrationssatz (WHT_K, 50 Spektren) und zwei Validationssätze (WHT_{V1} und WHT_{V2}, zu je 20 Spektren). Die Aufteilung der Standards kann der Tabelle A.1 (im Anhang) entnommen werden. Der erste Validationsdatensatz wird als 'interner', der zweite als 'externer' Validationsdatensatz bezeichnet. Durch letzteren erfolgt die abschließende Validation der Kalibrationsmodelle.

5.2.2 Olfen–Spektren

Dieser Datensatz (*s.* Abb. 5.4) besteht aus insgesamt 107 NIR-Spektren von Olfen–Tabletten[1] [97]. Die Kalibration soll dazu dienen, den Gehalt des Wirkstoffs Diclofenac ([*o*-(2,6–Dichloranilin)phenyl]essigsäure – Natrium, *s.* Abb. 5.5) in den Tabletten zu bestimmen. Der Gehalt beläuft sich auf 40 bis 60 mg pro Tablette. Auch dieser Datensatz wurde in drei Sets aufgesplittet, nämlich einen Kalibrationssatz (43 Spektren) und zwei Validationssätze (je 32 Spektren).

Die Spektren wurden in einem Bereich von 6.000 cm^{-1} bis 11.520 cm^{-1} in Intervallen von 12 cm^{-1} mit dem Bühler Tabletten-Autosampler aufgenommen. In diesem

[1] Olfen © ist eine Warenbezeichnung der Mepha Ltd., Schweiz.

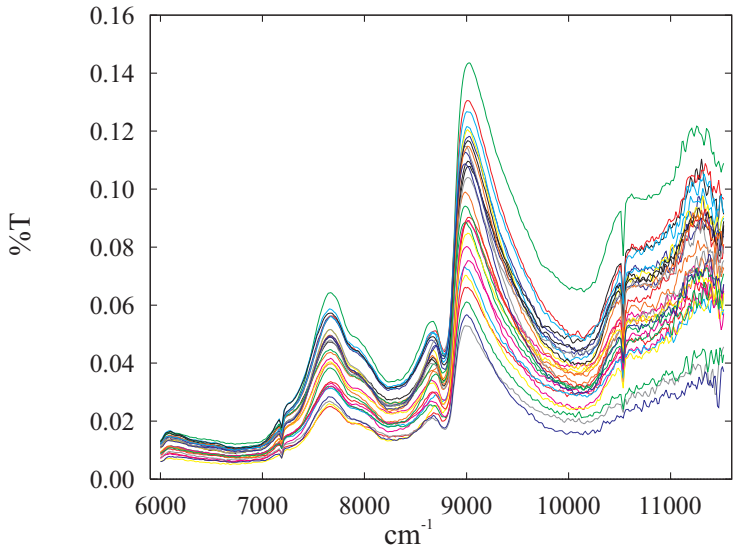


Abbildung 5.4: Olfen–Spektrensatz (30 Scans pro Spektrum).

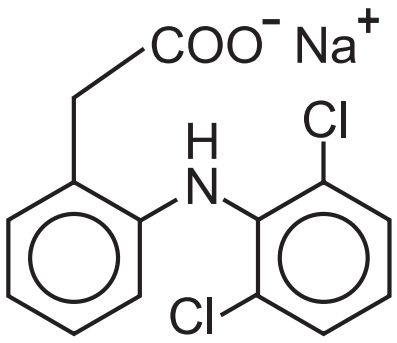


Abbildung 5.5: Struktur des Wirkstoffs in den Olfen–Tabletten: Diclofenac.

Gerät werden die Tabletten in Transmission gemessen. Hierbei ist der Lichtdurchsatz jedoch sehr gering, so daß relativ lange gescanned werden muß, um dennoch ein gutes Signal–Rausch–Verhältnis zu erlangen (mittels Signal–Averaging, das heißt Akkumulieren und anschließendes Mitteln der Spektren). Die einzelnen Datensätze unterscheiden sich allerdings deutlich in der Anzahl akkumulierter Spektren: Im Kalibrationsdatensatz betrug die Scanzahl pro Spektrum 30, im ersten Validationsdatensatz 10 und im zweiten Validationsdatensatz nur 1 Scan[2]. Die Wahl dieser Scanzahlen war an den praktischen Anforderungen einer modernen Routine-Analytik orientiert, in der jede einzelne Tablette so schnell und präzise wie möglich vermessen werden soll. Da bei der Kalibration selbst der Zeitfaktor für die Aufnahme der Spektren keine so große Rolle spielt, scheint es sinnvoll, für die Aufstellung des Kalibrationsmodells mit einem besseren Signal–Rauschverhältnis in den Spektren zu arbeiten, als bei der späteren betrieblichen Durchführung von Analysen. Es wird daher überprüft, welche Auswirkungen das Herabsetzen der Scanzahl auf 10 bzw. 1 auf den Analysenfehler hat.

5.2.3 Wäßrige Systeme

Bei den wäßrigen Systemen handelt es sich um drei verschiedene Datensätze von wäßrigen Lösungen anorganischer Salze (*s.* Tab. 5.2). Die Lösungen von Natrium- und Kaliumchlorid decken einen Konzentrationsbereich zwischen $0.0\,\mathrm{mol\,l^{-1}}$ und $0.124\,\mathrm{mol\,l^{-1}}$ ab. Die Lösungen des Zweikomponentensystems aus Kalium- und Aluminiumchlorid liegen im Konzentrationsbereich zwischen $0.0\,\mathrm{mol\,l^{-1}}$ und $0.7\,\mathrm{mol\,l^{-1}}$ je Komponente. Die genannten Salze haben kein eigenes Spektrum, sondern beeinflussen z.B. durch die Ausbildung von Hydrathüllen die Struktur des Wassers, so daß sich ihre Konzentrationen nur indirekt durch die Änderung der Bandenstruktur der NIR–Spektren des Wassers ermitteln läßt [61, 93, 98, 99]. Abbildung 5.6 zeigt dies am Beispiel des Zweikomponentensystems aus Aluminum- und Kaliumchlorid.

Die Spektren wurden in einem Bereich von $4\,000\ \mathrm{cm^{-1}}$ bis $10\,000\ \mathrm{cm^{-1}}$ mit einer digitalen Auflösung von $1\ \mathrm{cm^{-1}}$ an einem 1700X FT-NIR Spektrometer von Perkin-Elmer gemessen. Die Aufteilung der einzelnen Datensätze ist in Tabelle 5.2 aufgelistet. Die Datensätze wurden im Arbeitskreis bereits in [61, 93] ausführlich analysiert und beschrieben. Hier sollen sie vorrangig zur Überprüfung der Leistungsfähigkeit des Genetischen Algorithmus herangezogen werden.

5.3 Eingesetzte Software

Im Rahmen der Arbeit wurde eine spezielle Software (VG) entwickelt und in Zusammenarbeit mit *U. Depczynski*[3] in C++ realisiert: Das Programm VG ('virtuell

[2]Meßzeit für 30 Scans ca. 1 Minute, für einen Scan ca. 2 Sekunden.
[3]Institut für Angewandte Mathematik, Universität Hohenheim

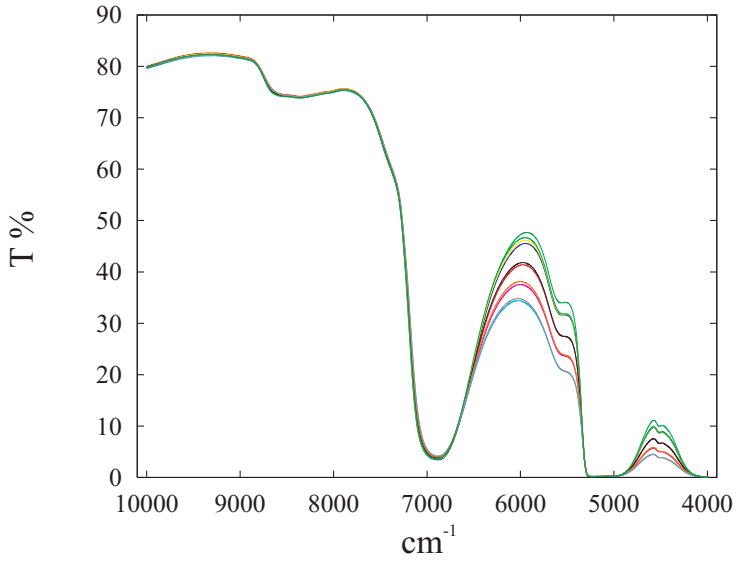


Abbildung 5.6: Spektren des Kalium- / Aluminiumchlorid–Datensatzes.

Tabelle 5.2: Aufteilung der Spektren wäßriger Lösungen anorganischer Salze.

Lösung	Anzahl der Spektren		
	Kalibration	int. Validation	ext. Validation
NaCl	120	54	17
KCl	132	45	16
KCl / $AlCl_3$	52	10	11

genetics') ist für die Optimierung von Faktor–Selektionen in PCR-Kalibrationen auf Basis eines Genetischen Algorithmus ausgelegt und wurde (wenn nicht explizit anders ausgewiesen) für alle in dieser Arbeit dargestellten Berechnungen herangezogen. Von dem Programm werden Spektren, Properties, etc. eingelesen und die spektralen Daten gegebenenfalls einer Eigenwertzerlegung und Regression unterzogen. Die Ergebnisse der Kalibrationsoptimierung durch Faktor–Selektion werden dann in Dateien (z.B.: Report-Files, CSV-Files) und am Bildschirm ausgegeben. Die Ergebnisausgabe in Dateiform und die umfangreichen Steuerungsmöglichkeiten über Kommandozeilenoptionen ermöglichen eine vollständige Automatisierung der Berechnungen. Eine genaue Beschreibung der Funktionalität und des strukturellen Aufbaus des in die VG-Software integrierten Parallelen Genetischen Algorithmus findet sich in Kapitel 7. Darüber hinaus wurden folgende Software-Pakete eingesetzt:

- NIRCAL–Software Versionen 1.0 bis 3.0

  Die Weiterentwicklung von Teilen der NIRCAL–Software war Gegenstand des Kooperationsprojekts zwischen der Gerhard-Mercator-Universität – GH Duisburg und der Firma BÜHLER AG.

- MATLAB Version 5.2

  Im Rahmen dieser Arbeit wurden vor allem die Funktionen zur Eigen- und Singulärwertzerlegung genutzt, um die Software-Entwicklung zu validieren.

- MATHEMATICA Version 3.0

  Auch hier stand der Einsatz der Funktionen zur Eigen- und Singulärwertzerlegung in erster Linie unter dem Aspekt der Validierung und der Visualisierung von Ergebnissen.

- QUANT$^+$–Software Version 3.0

  Die QUANT$^+$–Software wurde zur Berechnung von PLS-Kalibrationsmodellen herangezogen.

Die Software-Entwicklung fand unter Anwendung von Visual C++ der Firma Microsoft in den Versionen 4.0 und 5.0 statt.

Hardware

Der überwiegende Teil der Berechnungen wurde mit einem 200 MHz Pentium–Pro Prozessor[4] durchgeführt. Die anfallenden Rechenzeiten waren dabei abhängig von der Anzahl der Datenpunkte im Kalibrationsdatensatz, der Problem–Komplexität (Zahl der Freiheitsgrade) und den Parameter–Einstellungen des Genetischen Algorithmus.

[4]Pentium–Pro © ist ein Warenzeichen der Firma INTEL Inc.