„Der Einfluss statistischer Ausreißer auf die Schätzung der natürlichen Variabilität in Daten zu Biota“

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Wirtschaftsmathematiker

René Lehmann

aus Schönebeck/Elbe

Berichter: apl. Professor Hans Toni Ratte

Universitätsprofessor Erhard Cramer

Tag der mündlichen Prüfung: 15.03.2012

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

Lebenslauf:

Persönliche Daten:

Name: Lehmann

Vorname: René

Geburtstag: 31.01.1982

Geburtsort: Schönebeck/Elbe

Staatsangehörigkeit: deutsch

Qualifikationen:

2000: Abitur

2001 bis 2008: Wirtschaftmathematik, Diplom

2010: Beginn der Promotion

Danksagung

Die vorliegende Arbeit wurde von mir als externer Doktorand am Institut für Umweltforschung der naturwissenschaftlichen Fakultät der RWTH Aachen angefertigt.

Herrn Professor Dr. Hans Toni Ratte

danke ich für seine spontane Bereitschaft, die Betreuung dieser Arbeit zu übernehmen und seine stete Unterstützung sowohl in fachlicher, als auch in organisatorischer Hinsicht.

Mein besonderer Dank gilt

Professor Dr. Erhard Cramer

vom Institut für Statistik und Wirtschaftsmathematik der RWTH Aachen für seine fachliche Betreuung der Arbeit aus mathematischer Sicht, den regen Austausch und die vielen wertvollen Hinweise und Anregungen.

Professor Dr. Francisco Javier Prieto danke ich für den fachlichen Austausch zu seinen Veröffentlichungen und die Unterstützung bei der programmiertechnischen Umsetzung seiner Algorithmen.

Ich danke meinen Eltern für ihren steten Rückhalt während der Erstellung dieser Arbeit und die immer wieder neuen Motivationsschübe. Ich weiß, dass man sich etwas Schöneres vorstellen kann, als 400 Seiten auf Rechtschreibung und Ausdruck zu prüfen…

Meiner Freundin Samantha danke ich dafür, dass sie sich geduldig meine fortwährenden Ausführungen und Erklärungen zu den Inhalten und Problemstellungen meiner Arbeit anhörte und mich in den entscheidenden Momenten zu stützen wusste, wenn das Licht am Ende des Tunnels in weite Ferne rückte.

Meinen Kollegen und Freunden Patrick Ackermann und Philipp Gärtner danke ich für die vielen schlechten Wortspiele, welche mich oft bis in die Abendstunden hinein erheiterten: die Zeit des Forschens und Schreibens dieser Arbeit.

Die Freude an der Arbeit ist die halbe Strecke auf dem Weg zum Ziel.

Für meine Eltern

Inhaltsverzeichnis

1 Motivation **1**

2 Polychlorierte Dibenzodioxine und Dibenzofurane, polychlorierte Biphenyle und dioxinähnliche polychlorierte Biphenyle **8**

2.1 Dioxine 9

2.2 PCB 11

2.3 Dioxinähnliche PCB 13

3 Statistische Ausreißeranalytik **15**

3.1 Allgemeines zum statistischen Vorgehen bei der Identifizierung von Ausreißern 15

3.1.1 Adjustiertes α und multiples Testen: die Simes-Hochberg-Prozedur 16

3.1.2 Kerndichteschätzungen 19

3.1.3 Robuste Schätzung linearer Modelle mit der LMS Methode 22

3.1.4 Datentransformation und Kompositionsdaten 24

3.1.5 Nulldaten und Messungen unterhalb der Bestimmungsgrenze 30

3.1.6 Fehlende Werte in Kompositionsquerschnittsdaten 32
3.1.7 Natürliche Variabilität 42
3.1.8 Der Exakte Test von Fisher 52
3.1.9 Kontaminationswahrscheinlichkeit und Kontaminationsanteil 53
3.1.10 Graphische und deskriptive Methoden zur Bestimmung von Ausreißern 56
3.2 Bestimmung von Ausreißern in univariaten Querschnittsdaten . . 65
3.2.1 Der robuste Ausreißertest nach Hampel 66
3.3 Bestimmung von Ausreißern in multivariaten Querschnittsdaten . 68
3.4 Die robuste Mahalanobis-Distanz 69
3.4.1 Das PP-Verfahren nach Pena und Prieto 74
3.5 Allgemeines zu Zeitreihen 80
3.5.1 Das Zeitreihenmodell 81
3.5.2 Schätzung eines Zeitreihenmodells 99
3.5.3 Robuste LMS Schätzung eines Zeitreihenmodells nach Rousseeuw und Leroy 106
3.5.4 Trends in Zeitreihen 107
3.5.5 Test auf Trends in Zeitreihen 114
3.5.6 Entfernung von Trends aus Zeitreihen 128
3.5.7 Algorithmus zur Ausreißeranalytik bei nichtstationären Zeitreihen 138
3.5.8 Fehlende Werte in Zeitreihen Kompositionsdaten 141
3.6 Ausreißer in univariaten Zeitreihen 143

3.6.1 Schätzung des Effektes eines Ausreißers in univariaten Zeitreihen 144
3.6.2 Bereinigung univariater Zeitreihen von level shifts 150
3.6.3 Bestimmung von Ausreißern des Typs „A“, „T“ und „I“ in univariaten Zeitreihen 157
3.7 Ausreißer in multivariaten Zeitreihen 163
3.7.1 Projektion einer multivariaten Zeitreihe 163
3.7.2 Schätzung des Effektes eines Ausreißers in multivariaten Zeitreihen 173
3.7.3 Bereinigung multivariater Zeitreihen von level shifts . . . 176
3.7.4 Bestimmung von Ausreißern des Typs „A“, „T“ und „I“ in multivariaten Zeitreihen 181

4 Auswertung und Diskussion spezieller Datensätze 185
4.1 Querschnittsdaten 186
4.1.1 Univariate Querschnittsdaten 186
4.1.2 Multivariate Querschnittsdaten 193
4.2 Multivariate Zeitreihen 211
4.2.1 PCB Leitkongenere 211

5 Allgemeine Auswertung und Diskussion 240
5.1 Univariate Querschnittsdaten 241
5.2 Multivariate Querschnittsdaten 242
5.3 Stichprobenumfang und Dimension der Daten 248
5.4 Probleme mit der Mahalanobis-Distanz 250
5.5 Kongenerenprofile 251

5.6 Multivariate Zeitreihen 252

6 Zusammenfassung **261**

Abbildungsverzeichnis **265**

Tabellenverzeichnis **273**

Literaturverzeichnis **282**

A Messprogramme **307**

B Auswertungstabellen zu Querschnittsdaten **309**

C Auswertungstabellen zu Zeitreihendaten **336**

D Kongenerenprofile dl PCB Aal **348**

E Kongenerenprofile dl PCB Brasse **349**

F Kongenerenprofile dl PCB Grünkohl **350**

G Kongenerenprofile dl PCB Weidelgras **351**

H Kongenereprofile PCB Leitkongenere Ei des Wanderfalken **352**

I Kongenereprofile PCB Leitkongenere Aal **354**

J Kongenereprofile PCB Leitkongenere Brasse **356**

K Kongenereprofile PCB Leitkongenere Brassen-Filets **359**

L Kongenereprofile PCB Leitkongenere Grünkohl **363**

M Kongenereprofile PCB Leitkongenere Weidelgras **365**

N Kongenerenprofile PCDD Aal **367**

O Kongenerenprofile PCDD Brasse **368**

P Kongenerenprofile PCDD Miesmuschel **371**

Q Kongenerenprofile PCDD Fichte-Nadel **372**

R Kongenerenprofile PCDD Grünkohl **375**

S Kongenerenprofile PCDD Weidelgras **378**

T Kongenerenprofile PCDF Brasse **379**

U Kongenerenprofile PCDF Fichte-Nadel **380**

V Kongenerenprofile PCDF Grünkohl **383**

W Kongenerenprofile PCDF Weidelgras **386**

Kapitel 1

Motivation

Biometrische und chemometrische Daten werden erhoben, um neue Erkenntnisse über Sachverhalte und Kausalitäten bezüglich der Wirkungen chemischer Stoffe auf biologische Organismen, Akkumulationseigenschaften oder auch Toxizitätsindizes zu erlangen. Dabei werden Messwerte unter möglichst homogenen Bedingungen generiert, um die Variabilität der Ergebnisse nicht unnötig zu steigern. Man verwendet standardisierte Prozeduren und geeichte Messinstrumente, um die, durch das Design des Experimentes oder die Versuchsumgebung induzierte, Streuung der Messwerte einzuschränken. Dennoch werden selbst unter kontrollierten Bedingungen häufig extreme oder ungewöhnliche Messwerte bzw. Messwerttupel beobachtet. Diese können gegebenenfalls als Messfehler, Einheitenfehler oder statistische Ausreißer charakterisiert und von den weiteren Versuchsauswertungen ausgeschlossen werden.

Wild lebende Pflanzen und Tiere, sogenannte Biota, unterliegen hingegen keiner Kontrolle durch vorgegebene Testdesigns, wie sie in Laboren eingesetzt werden. Sie befinden sich in einem Zustand ständiger Wechselwirkungen mit ihrer

Umwelt, beispielsweise durch unkontrollierte Chemikalieneinträge in den Boden durch Regen und Wind. Dort werden PCB und dioxinähnliche PCB resorbiert und lagern sich in Pflanzen ein. Tiere sind den Chemikalien insbesondere bei der Nahrungsaufnahme ausgesetzt, seien es Pflanzen- oder Fleischfresser.

Die Untersuchung der Konzentration verschiedener PCB und dioxinähnlicher PCB in wild lebenden Pflanzen und Tieren ist somit einerseits abhängig von den Umwelteinflüssen, andererseits aber auch vom Individuum selbst. Jeder biologische Organismus lagert Substanzen individuell ein und baut sie genauso individuell ab. Auch wenn diese Vorgänge einer gewissen Gesetzmäßigkeit folgen, so belegen Laborstudien, dass der Einfluss des jeweiligen Individuums auf die Bioakkumulation und den Abbau einer Chemikalie erheblich ist [vgl. EHM$^+$06, AGRH07].

Es ist folglich zu erwarten, dass die zufällige Streuung gemessener Werte, z.B. der Konzentration von PCB28 und PCB52 (in $\mu g/kg$) im tierischen Fett bei Biota erheblich höher ist, als bei Experimenten unter kontrollierten Laborbedingungen. Dementsprechend komplizierter ist es, die natürliche Variabilität aus den Daten zu schätzen und extreme Werte ausfindig zu machen.

Das Problem statistischer Ausreißer in Biota wurde unter anderem diskutiert Tracey und Hanson. Sie analysierten Biota-Sedimentdaten zu PCB auf statistische Ausreißer [TH96]. Kozlov et al. [KEKZ09] untersuchten den Einfluss industrieller Umweltverschmutzung auf terrestrische Biota, z.B. in Form von Sprosslängen und Nadeln von Pinien (Pinus sylvestris). Die Kontamination der Daten durch statistische Ausreißer wurde untersucht und ca. 0.2% der Daten als Ausreißer verworfen. Die EPA (Environmental Protection Agency, USA) stellte der wissenschaftlichen Welt einen Biota-Sediment Datensatz, bestehend aus mehr als 20000 Werten zu

Bioakkumulation in Sedimenten, zur Verfügung. Die Daten umfassen unter anderem PCB und Dioxine. Sie wurden vor der Veröffentlichung von statistischen Ausreißern bereinigt, soweit diese detektierbar waren. Die EPA weist jedoch explizit darauf hin, dass der Datensatz dennoch Ausreißer enthalten kann, die noch nicht entfernt wurden [BSA09].

Was aber ist zu tun, wenn ein Wert oder Tupel von Messwerten unplausibel erscheint, sei es aufgrund der extremen Ausprägung, oder aber durch die ungewöhnlich erscheinende Kombination der Daten? War die Messung fehlerhaft, oder das untersuchte Individuum einer ungewöhnlich hohen Konzentration ausgesetzt? Ist der Messwert bei den weiteren Analysen und Auswertungen zu berücksichtigen und damit als Indikator der natürlichen Variabilität zu sehen, oder muss er als Ausreißer verworfen und von den weiteren Auswertungen ausgeschlossen werden? Was sind Ausreißer? Wie unterscheidet man zwischen „natürlicher Streuung bzw. Variabilität“ und „Ausreißern“. Welche Arten von Ausreißern sind sinnvoll zu unterscheiden und welche Auswirkungen auf die Schätzung der natürlichen Variabilität hätte es, wenn man strikt alle Ausreißer in Biota Daten von den Analysen ausschlösse? Welche Kontrollmechanismen gibt es, um sicherzustellen, dass korrekte Daten nicht fälschlicher Weise als Ausreißer deklariert werden? Wie groß ist die natürliche Variabilität der Konzentration von PCB und dioxinähnlichen PCB in Biota einzuschätzen und welche unterschiedlichen Schätzungen ergeben sich unter Berücksichtigung der Präsenz statistischer Ausreißer?

Um diese Fragen zu beantworten, ist der Begriff des Ausreißers in zwei verschiedenen Formen zu erfassen. Es wird zwischen „statistischen Ausreißern“ und „sachlogischen Ausreißern“ unterschieden. Die Argumentation zum Einfluss von Ausreißern auf die natürliche Variabilität der Belastung durch Dioxine, PCB und

dioxinähnliche PCB in Biota beschränkt sich rein formal auf statistische Ausreißer. Die sachlogische Analyse auf Ausreißer und eine Argumentation zum Ausschluss von Messwerten aus Datensätzen auf dieser Basis ist nicht Gegenstand der Arbeit, da solche Entscheidungen stets kontextspezifisch und unter gegenseitiger Kontrolle verschiedener Instanzen zu treffen sind.

Es wird zunächst der aktuelle Stand der Forschung auf dem Gebiet der statistischen Ausreißeranalytik kurz dargestellt. Dabei wird zwischen univariaten und multivariaten Verfahren für Querschnittsdaten und für Zeitreihen unterschieden. Darüber hinaus wird ein kurzer Überblick zum Begriff der natürlichen Variabilität unter Berücksichtigung des sachspezifischen Kontextes gegeben.

Im weiteren Verlauf steht die anwendungsorientierte Darstellung der theoretischen Verfahren zur Ausreißererkennung anhand von Algorithmen im Vordergrund. Die Ansätze von Pena und Prieto ([vgl. PP01b]) und Mahalanobis [vgl. FH08] für multivariate Querschnittsdaten und Galeano et al. ([vgl. GPT06]) für multivariate Zeitreihen sind dabei ein zentrales Thema. Die dort vorgestellten Algorithmen werden möglichst dynamisch programmiert, um dem Anwender ein praktisches Werkzeug an die Hand zu geben. Die Quellcodes sind in **R** verfasst. Für univariate Querschnittsdaten findet der Hampel-Test Verwendung [vgl. Sac09].

Nachdem die statistischen Methoden beleuchtet wurden, ist es das erklärte Ziel, einen repräsentativen Teil der Datensätze der Dioxin-Datenbank des Bundes und der Länder zu Biota aufzubereiten und der Analyse mit Hilfe der vorgestellten statistischen Verfahren zu unterziehen. Die natürliche Variabilität wird sowohl ohne Ausschluss, als auch unter Ausschluss der gefundenen statistischen Ausreißer geschätzt und die Ergebnisse verglichen.

Für die Bewertung der Belastungssituation verschiedener Umweltkompartimen-

te durch Dioxine, PCB und dioxinähnliche PCB werden sogenannte Kongenerenprofile erstellt (wie z.B. im Dioxinskandal in Deutschland 2010/2011). Kongenerenprofile sind Balkendiagramme der mittleren Konzentrationen der verschiedenen Stoffe in einem Umweltkompartiment. Um adäquate politische Maßnahmen ableiten zu können, ist es notwendig diese nicht durch Ausreißer zu verzerren. Ein Kompartiment, welches von besonderem politischen Interesse während der letzten Jahre und Dioxinskandale war, sind Biota. Wildlebende Tiere sind selten industriell gefertigtem Tierfutter ausgesetzt und daher wenig gefährdet, durch diese eine Kontamination mit Dioxinen, PCB und dioxinähnlichen PCB zu erfahren. Kongenerenmuster aus Biotadaten können folglich als Referenzmuster für die Hintergrundbelastung von Pflanzen und Tieren durch Dioxine, PCB und dioxinähnliche PCB herangezogen werden. Daher werden in dieser Arbeit aus den bereinigten Daten die entsprechenden Kongenerenprofile erstellt.

Weiterhin angestrebt ist eine Schätzung der Wahrscheinlichkeit, dass ein Datensatz zur Belastung von Biota durch Dioxine, PCB und dioxinähnliche PCB durch statistische Ausreißer kontaminiert ist. In einem letzten Schritt wird der Kontaminationsanteil der Datensätze durch statistische Ausreißer geschätzt. Im Anschluss werden die Ergebnisse zusammengefasst und diskutiert, um einen Überblick über die allgemeinen und die speziellen Erkenntnisse zu geben.

Sämtliche Datensätze entstammen der Dioxindatenbank von Bund und Ländern `www.pop-dioxindb.de`. Die verwendeten Messprogramme sind in Anhang A gegeben. Die statistische Ausreißeranalyse gestaltet sich mitunter schwierig, da häufig hochdimensionale Querschnittsdaten (Dimension $6 \leq d \leq 12$) mit geringen Stichprobenumfängen (häufig $n \approx 2d$) untersucht werden. Aus diesem Grund wird bereits an dieser Stelle darauf hingewiesen, dass eine statistische Ausreißer-

analytik für multivariate Daten oft möglich ist, die Ergebnisse jedoch stets kritisch hinterfragt werden müssen, da Ausreißer in kleinen Datensätzen hoher Dimension jederzeit auch ein Indikator für die natürliche Streuung sein können. Es ist daher stets zu prüfen, ob neben den statistischen auch sachlogische Gründe vorliegen, einen Datenpunkt als Ausreißer zu kennzeichnen. Diesem Anspruch wird in Abschnitt 4 und 5 Rechnung getragen.

Die Arbeit ist wie folgt aufgebaut: In Abschnitt 2 werden Dioxine, PCB und dioxinähnliche PCB vorgestellt. Der Abschnitt 3.1 befasst sich mit den allgemeinen mathematischen Grundlagen, wie der Analyse und Transformation von Daten des vorliegenden Typs (sog. Kompositionsdaten), dem Umgang mit Messwerten unterhalb der Bestimmungsgrenze und der Schätzung fehlender Werte. Es werden die Begriffe „natürliche Variabilität“, „Kontaminationsanteil“ und „Kontaminationswahrscheinlichkeit“ eingeführt. Zusätzlich erfolgt die Vorstellung hilfreicher graphischer Analysemethoden. Die statistische Ausreißeranalytik für univariate Querschnittsdaten wird in Abschnitt 3.2 anhand des robusten Ausreißertests nach Hampel dargestellt. Die dort verwendetet Teststatistik stellt die Grundlage für das in Abschnitt 3.3 vorgestellte Projektionsverfahren von Pena und Prieto zum Auffinden statistischer Ausreißer in multivariaten Querschnittsdaten dar. Zusätzlich wird die Verwendung der robusten Mahalanobis-Distanz zum Auffinden statistischer Ausreißer illustriert. Für den Anwender liegen zwei Implementierungen der Projektionsmethode von Pena und Prieto vor. Die erste Implementierung enthält die Vorgehensweise, wie sie von Pena und Prieto beschrieben wurde [PP01b], die zweite Implementierung ist eine Übersetzung des Originalcodes von Pena und Prieto aus Matlab 4 in **R**. Diese verwendet aktuellere kritische Werte als in der Veröffentlichung von Pena und Prieto 2001 angegeben wurden und dient daher zur

finalen Auswertung der Querschnittsdaten in dieser Arbeit. In Abschnitt 3.5 werden die Grundlagen der Modellierung und Schätzung statistischer Zeitreihen vorgestellt. Sowohl die Integration statistischer Ausreißer in Zeitreihenmodelle, als auch die robuste Schätzung fehlender Werte und die Elimination von Trends werden erläutert und, sofern noch nicht vorhanden, implementiert. Die Schätzung des Effektes und die entsprechende Elimination von Ausreißern in Zeitreihen sind Gegenstand von Abschnitt 3.6 und 3.7. Anschließend erfolgt eine spezielle und eine allgemeine Auswertung und Diskussion der Ergebnisse in den Abschnitten 4 und 5. Dort werden neben dem Kontaminationsanteil und der Kontaminationswahrscheinlichkeit auch die Erstellung von Kongenerenprofilen und Probleme bei den Auswertungen diskutiert. Eine Zusammenfassung der Ergebnisse und Erkenntnisse ist in Abschnitt 6 gegeben.

Kapitel 2

Polychlorierte Dibenzodioxine und Dibenzofurane, polychlorierte Biphenyle und dioxinähnliche polychlorierte Biphenyle

Polychlorierte Dibenzodioxine und Dibenzofurane (auch PCDD und PCDF), polychlorierte Biphenyle (auch PCB) und dioxinähnliche polychlorierte Biphenyle (auch dl PCB, engl. dl=dioxin like) sind chemische Verbindungen. Polychlorierte Dibenzodioxine und Dibenzofurane werden in den deutschen Medien oft als „Dioxine“ bezeichnet, obwohl die Stoffgruppe der Dioxine mehr als nur PCDD und PCDF umfasst. In den folgenden Abschnitten werden Definitionen zu den genannten chemischen Stoffen anhand von Strukturformeln gegeben, sowie Eigenschaften dieser Stoffe (z.B. Persistenz oder Wirkungen auf den menschlichen Organismus) genannt.

2.1 Dioxine

Im „Handbuch Boden“ werden PCDD und PCDF wie folgt definiert: „Polychlorierte Dibenzodioxine und Dibenzofurane sind tricyclische, aromatische Ether, die jeweils eine unterschiedliche Anzahl von Chloratomen (1-8) in ihren Molekülen aufweisen.

Für die Gruppe der PCDD ergeben sich 75 verschiedene Chlorkongenere, für die Gruppe der PCDF insgesamt 135. Es kommen demnach 210 verschiedene polychlorierte Dibenzodioxine und -furane vor, die sich in ihrer Toxizität erheblich unterscheiden.

Toxikologisch relevant sind vor allem tetra-, penta- und hexachlorierte Dibenzodioxine und Dibenzofurane, die an den Positionen 2,3,7,8 substituiert sind. Diese Verbindungen werden auch als "dirty dozen"bezeichnet und machen über 80 % aller Polychlorierten Dibenzodioxine und -furane aus, die in der Umwelt gefunden werden.

Die giftigste Verbindung unter den PCDD/F ist das 2,3,7,8 Tetrachlordibenzodioxin (2,3,7,8 TCDD), das sogenannte Seveso-Dioxin. Die Eigenschaften dieser Verbindung sind bisher am umfangreichsten untersucht worden. Sie dient als Bezugsgröße für sämtliche Toxizitätsbewertungssysteme.“[vgl. Wol95, S.4].

Dioxine entstehen in der industriellen Fertigung, speziell bei chemischen Verbrennungsprozessen, sofern die notwendigen Ausgangsstoffe (Kohlenwasserstoff und Chlor) bei entsprechender Temperatur vorhanden sind [vgl. UBA11]. Als Pri-

märquellen für die Entstehung von PCDD und PCDF werden im „Handbuch Boden“ industrielle Prozesse, wie z.B. der Einsatz von Chlor in der Papierproduktion, das Aktivkohlerecycling oder die Generierung von Metallchloriden genannt. In Verbrennungsprozessen können aus PCP oder PCB Dioxine entstehen, ebenso durch Pyrolyse [vgl. Wol95, S.6]. Die chemische Struktur von PCDD und PCDF ist in Abbildung 2.1 gegeben.

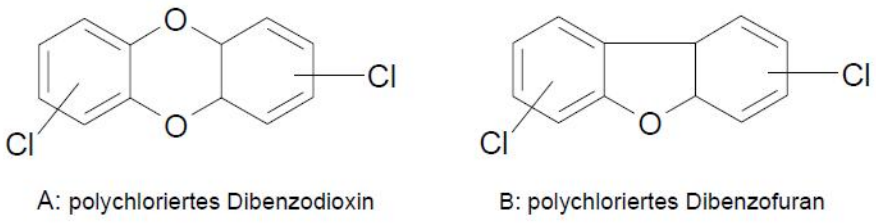

Abbildung 2.1: Allgemeine Strukturformel für PCDD/F [vgl. Umw08, S.1]

Die Leitkongenere für die Gruppe der PCDD werden bezeichnet mit 2,3,7,8-TCDD, 1,2,3,7,8-PeCDD, 1,2,3,4,7,8-HxCDD, 1,2,3,6,7,8-HxCDD, 1,2,3,7,8,9-HxCDD, 1,2,3,4,6,7,8-HpCDD und OCDD. Die Leitkongenere für die Gruppe der PCDF werden bezeichnet mit 2,3,7,8-TCDF, 1,2,3,7,8-PeCDF, 2,3,4,7,8PeCDF, 1,2,3,4,7,8-HxCDF, 1,2,3,6,7,8-HxCDF, 1,2,3,7,8,9-HxCDF, 2,3,4,6,7,8-HxCDF, 1,2,3,4,6,7,8-HpCDF, 1,2,3,4,7,8,9-HpCDF und OCDF. Die Abkürzungen bedeuten dabei T=Tetra, Pe=Penta, Hx=Hexa, Hp=Hepta, O=Okta. Diese Leitkongenere werden stets herangezogen, um die Belastung eines zu untersuchenden Stoffes oder Umweltkompartimentes durch Dioxine zu bewerten. In diesem Zusammenhang berechnet man aus mehreren Proben eine „mittlere“ Belastung durch die jeweiligen Leitkongenere und erstellt ein sogenanntes Kongenerenprofil, welches statistisch betrachtet einem Balkendiagramm der „mittleren“ Konzentrationen entspricht.

2.2 PCB

Hartmann et al. beschreiben das Vorkommen und die Wirkung polychlorierter Biphenyle wie folgt.

„Die industrielle Produktion der PCB begann bereits 1929 mit einem Verfahren durch direkte Chlorierung von Biphenyl. Theoretisch sind durch die 10 möglichen Chlorierungsstellen bis zu 209 Kongenere mit 10 unterschiedlichen Chlorierungsgraden denkbar. Die chemische und physikalische Stabilität der PCB, ihr elektrischer Widerstand, niedriger Dampfdruck, Widerstandsfähigkeit gegen Abbau und Zersetzung bei hohen Temperaturen trugen zu ihrer starken kommerziellen Nutzung bei.

PCB wurden auf zwei unterschiedliche Arten eingesetzt:

- Verwendung in geschlossenen Systemen vor allem in elektrischen und hydraulischen Geräten. Die wichtigsten Freisetzungsquellen bei dieser Verwendungsart waren Leckagen, Brände, Unfälle, rechtswidrige Ablagerungen und unangemessene Entsorgung.

- Verwendung in offenen Systemen als Pestizidfüllstoff, Flammhemmstoff, Weichmacher in Dichtungsmassen oder in Farben. Die wichtigsten Freisetzungsquellen bei dieser Verwendungsart waren Deponierung, Migration und Emission in der Luft durch Verdunstung.

Andere, wichtige Quellen für die ubiquitäre Verbreitung der PCB in der Umwelt sind Abfallverbrennung, Ausbringen von belasteten Klärschlämmen und Altölverbrennung.

Die PCB weisen wie die Dioxine und Furane toxische und persistente Eigenschaften auf und gehören zu den von der UNEP international anerkannten 12 persistenten organischen Schadstoffen (Persistent Organic Pollutants – POP), die durch Lipophilität, geringe Flüchtigkeit und Abbaubeständigkeit charakterisiert sind. Aufgrund dieser Eigenschaften weisen die PCB eine lange Persistenz in der Umwelt auf und zeigen unter typischen Umweltbedingungen eine hohe Fähigkeit zur Biomagnifikation und Biokonzentration.

1985 erfolgte aufgrund der Reproduktionstoxizität und der bioakkumulativen Auswirkungen das Verbot der Vermarktung und Verwendung von PCB.

Dennoch wurden in Deutschland in Klärschlämmen PCB mit Durchschnittsgehalten von ca. 0,16 mg/kg gefunden (UMK 2000).“[vgl. HSW03, S.18f]

Nach Hartmann et al. ist eine Aufnahme von PCB durch Pflanzen kaum möglich, jedoch besteht die Möglichkeit einer Anlagerung an deren Oberfläche. Wildlebende Pflanzen stellen für viele wildlebende Tiere eine Nahrungsquelle dar. Folglich gelangt PCB bei der Nahrungsaufnahme in deren Organismus. Da PCB eine geringe Abbaurate im Gewebe besitzen, wirken sie stark akkumulierend. Weiter werden nach Hartmann et al. über Pflanzen oder die Atemluft aufgenommene PCB größtenteils im Gewebe eingelagert. In der Folge reichern sich PCB über die Zeit im tierischen Gewebe an [vgl. HSW03, S.20f].

„PCB werden als vermutlich karzinogen für den Menschen eingestuft und erzeugen ein breites Spektrum nachteiliger Auswirkungen bei Tieren, darunter Reproduktionstoxizität, Immuntoxizität und Karzinogenität.“[vgl. HSW03, S.22]

Ballschmitter und Zell entwarfen eine Nomenklatur der verschiedenen PCB Typen [vgl. BZ80, S.21]. Die 6 PCB Leitkongenere nach Ballschmitter und Zell sind PCB28, PCB52, PCB101, PCB138, PCB153 und PCB180. Ihre Summe findet Verwendung bei der Bewertung der Belastung einer Probe durch PCB [vgl. SSM03, VN05]. Gelegentlich wird diese Summe zusätzlich mit dem Faktor 5 multipliziert, um die Gesamtbelastung durch PCB bewerten zu können [103, SSM03, JM06, vgl.].

Die allgemeine chemische Struktur von PCB ist in Abbildung 2.2 gegeben.

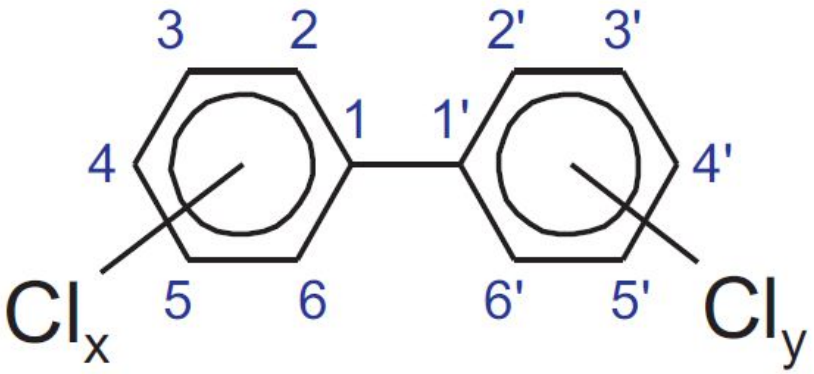

Abbildung 2.2: Allgemeine Strukturformel für PCB [vgl. PBL$^+$09, S.4]

2.3 Dioxinähnliche PCB

Groh und Borho beschreiben dl PCB wie folgt: „Polychlorierte Biphenyle (PCB) sind synthetische Produkte ohne natürliches Vorkommen. Je nach Chlorierungsgrad und Stellung der Chloratome am Biphenylgerüst sind 209 mögliche Vertreter zu unterscheiden, darunter die zwölf dl PCB Kongenere. Die industriell genutzten PCB-Gemische können sich in ihrer Zusammensetzung erheblich unterscheiden.

Sie haben sich infolge ihrer Verwendung in offenen und geschlossenen Systemen in der Umwelt ubiquitär verbreitet und aufgrund ihrer Persistenz angereichert. Man findet diese Stoffgruppe daher in verschiedenen Umweltkompartimenten wie Wasser, Boden, Luft, Sedimenten, Pflanzen und Tieren."[vgl. GB10, S.17]

Es handelt sich bei der Gruppe der dl PCB folglich um 12 spezielle PCB Kongenere: PCB77, PCB81, PCB105, PCB114, PCB118, PCB123, PCB126, PCB156, PCB1157, PCB167, PCB169 und PCB189. Die Strukturformeln dieser dl PCB sind in Abbildung 2.3 gegeben.

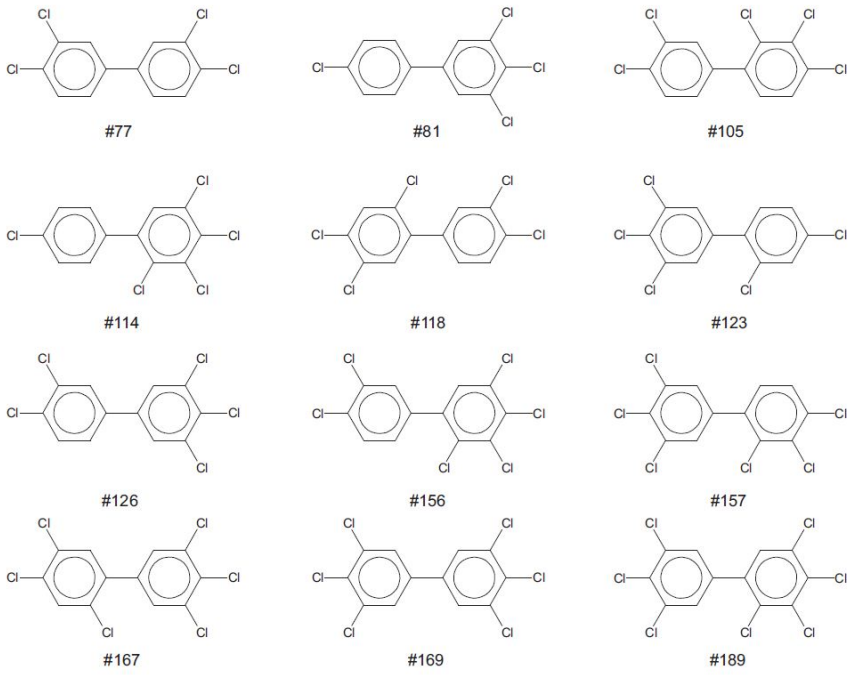

Abbildung 2.3: Strukturformel der 12 dl PCB [vgl. PBL$^+$09, S.5]

Kapitel 3

Statistische Ausreißeranalytik

3.1 Allgemeines zum statistischen Vorgehen bei der Identifizierung von Ausreißern

Die korrekte Auswertung von Umweltdaten bedarf verschiedener statistischer Methoden und Ansätze. Je nach Datentyp sind z.B. Transformationen notwendig, bevor eine Auswertung begonnen werden kann. Was geschieht aber, wenn bestimmte Datenpunkte nicht transformiert werden können, z.B. wenn $\ln 0$ gebildet werden soll? Gibt es Möglichkeiten die Daten entsprechend zu modifizieren? Wie können lineare Modelle robust geschätzt werden in Gegenwart statistischer Ausreißer? Dieser Abschnitt dient der Bereitstellung statistischer Grundlagen, um die im weiteren Verlauf der Arbeit vorgestellten Methoden der statistischen Ausreißeranalytik anwenden und verstehen zu können.

3.1.1 Adjustiertes α und multiples Testen: die Simes-Hochberg-Prozedur

Methoden zur Erkennung statistischer Ausreißer arbeiten häufig sequentiell, d.h. ein Datensatz wird analysiert und die gefundenen Ausreißer entfernt. Der bereinigte Datensatz wird erneut auf Ausreißer geprüft. Das Verfahren wird solange wiederholt, bis eine Abbruchbedingung (z.B. „die maximale Anzahl der Iterationen wurde erreicht“, „es wird kein weiterer Ausreißer erkannt“, „mehr als 50% der Daten wurden als Ausreißer erkannt“ o.ä.) erfüllt ist.

Wird ein Datensatz mehrfach mit statistischen Verfahren untersucht, so ist die Problematik der sogenannten α-Fehlerkumulierung oder α-Inflation zu beachten. Sollen z.B. k verschiedene Nullhypothesen zum gleichen Signifikanzniveau α getestet werden (sog. multiples Testen), ist die Wahrscheinlichkeit einen Fehler erster Art zu begehen für jeden einzelnen Test durch α beschränkt (d.h. lokal wird das Signifikanzniveau α eingehalten). Die Wahrscheinlichkeit in allen durchgeführten Tests keinen Fehler erster Art begangen zu haben ist jedoch nicht mehr länger durch α beschränkt.

Um dies einzusehen, überlege man sich folgendes. Seien $Y_1, \ldots, Y_k$ stochastisch unabhängige und identisch binomial-$(1, \alpha)$-verteilte Zufallsvariablen. Sie sind folgendermaßen zu interpretieren:

$$\{Y_i = 1\} \sim \text{ in Testproblem } i \text{ wurde ein Fehler erster Art begangen}$$

$$\{Y_i = 0\} \sim \text{ in Testproblem } i \text{ wurde kein Fehler erster Art begangen } .$$

Folglich gilt $P(Y_i = 1) = \alpha$ und $P(Y_i = 0) = 1 - \alpha$ für $i = 1, \ldots, k$. Das

Ereignis „in keinem der k Testprobleme wird ein Fehler erster Art begangen“ ist äquivalent zu

$$A := \{Y_1 = 0, Y_2 = 0, \ldots, Y_k = 0\} \ .$$

Unter Nutzung der stochastischen Unabhängigkeit der $Y_1, \ldots, Y_k$ tritt das Ereignis A mit Wahrscheinlichkeit $P(A) = (1 - \alpha)^k$ ein. Folglich ist die Wahrscheinlichkeit, in mindestens einem der k Testprobleme einen Fehler erster Art zu begehen, gegeben durch $1 - P(A) = 1 - (1 - \alpha)^k$.

Multiples Testen beinhaltet demnach das Problem, die Wahrscheinlichkeit des Auftretens eines Fehlers erster Art schlechter kontrollieren zu können. Wenn das Signifikanzniveau α für jeden einzelnen Test eingehalten wird (d.h. lokal), so folgt daraus nicht, dass es für alle Testprobleme insgesamt eingehalten wird (global). Man unterscheidet folglich die lokalen Signifikanzniveaus und das globale Signifikanzniveau. Um das globale Signifikanzniveau einzuhalten, müssen die lokalen Signifikanzniveaus entsprechend gewählt werden. Man arbeitet mit sogenannten adjustierten Signifikanzniveaus oder p-Werten. Wird ein Test mehrmalig durchgeführt, so wählt man bei jeder Anwendung ein neues, kleineres lokales Signifikanzniveau, oder äquivalent dazu wird der zugehörige p-Wert nach oben korrigiert.

In diesem Kontext wird die Simes-Hochberg-Prozedur verwendet, welche nach Sachs [vgl. Sac09, S.499f.] im Vergleich zur oft verwendeten Holm-Prozedur eine höhere Power aufweist. Vom Bonferroni-Verfahren wird hier kein Gebrauch gemacht, da nach Sachs [vgl. Sac09, S.499f.] die Power einer multiplen Testprozedur durch ein schrittweises Verfahren, wie die Simes-Hochberg-Prozedur, deutlich verbessert werden kann.

Die Simes-Hochberg-Prozedur ist eine sogenannte step up procedure, bei der

die p-Werte adjustiert werden. Äquivalent dazu kann das lokale Signifikanzniveau α schrittweise nach unten angepasst werden. Werden m Nullhypothesen getestet, so bestimmt man zunächst aus der aufsteigenden Ordnungsstatistik der m p-Werte $(p_{(1)}, \ldots, p_{(m)})$ einen kritischen Index.

$$j^*_{SH} := \max \left\{ j \mid p_{(j)} \leq \frac{\alpha}{m - j + 1} \right\}$$

Im Anschluss daran errechnen sich die adjustierten p-Werte aus folgender Formel:

$$p^{SH}_{(j)} := \min_{k=j^*_{SH},\ldots,m} \left\{ \min \left\{ (m - k + 1) \cdot p_{(k)}, 1 \right\} \right\} \quad .$$

Wie man sieht, wird jeder p-Wert mit dem Korrekturfaktor $(m - k + 1)$ multipliziert. Die so berechneten, adjustierten p-Werte werden dann mit dem Signifikanzniveau α verglichen. Äquivalent dazu ist es möglich, α mittels Division durch den jeweiligen Korrekturfaktor anzupassen. Man erhält somit neue, adjustierte Signifikanzniveaus.

Das Verfahren wird in Tabelle 3.1 exemplarisch dargestellt. Das Signifikanzniveau sei $\alpha = 0.05$. Es wird zunächst der größte p-Wert mit 1 multipliziert, folglich gilt $p^{SH}_{(5)} = p_{(5)} \cdot 1 = 0.065 > 0.05$. Im Anschluss daran wird der nächst kleinere p-Wert mit 2 multipliziert, also $p^{SH}_{(4)} = p_{(4)} \cdot 2 = 0.088 > 0.05$. Dieser Wert ist ebenfalls größer als das Signifikanzniveau α und daher wird nun gesetzt: $p^{SH}_{(4)} := p^{SH}_{(5)}$. Weiter wird berechnet $p^{SH}_{(3)} = p_{(3)} \cdot 3 = 0.099 > 0.05$ und folglich $p^{SH}_{(3)} := p^{SH}_{(5)}$. Im vierten Schritt ist $p^{SH}_{(2)} = p_{(2)} \cdot 4 = 0.048 < 0.05$ und damit kleiner als das Signifikanzniveau α. Daher bleibt $p^{SH}_{(2)} = 0.048$ unverändert bestehen. Im letzten Iterationsschritt ergibt sich $p^{SH}_{(1)} = p_{(1)} \cdot 5 = 0.005 < 0.05$ und auch

hier ist demzufolge die Setzung $p_{(1)}^{SH} = 0.005$ korrekt.

Nummer	p_j	Rang	$p_{(j)}^{SH}$
1	0.044	4	0.065
2	0.065	5	0.065
3	0.012	2	0.048
4	0.033	3	0.065
5	0.001	1	0.005

Tabelle 3.1: Beispiel zur Simes-Hochberg Prozedur ([vgl. Sac09, S.500]

Das dazu äquivalente Vorgehen bei der schrittweisen Adjustierung des Signifikanzniveaus α beschreibt sich wie folgt. Anstatt im ersten Schritt den größten p-Wert mit 1 zu multiplizieren, wird das zugehörige α durch 1 dividiert, also $\alpha_{(5)} = 0.05/1$. Im zweiten Schritt teilt man dann α durch 2, anstatt den p-Wert mit 2 zu multiplizieren. Analog geht man in Schritt 3, 4 und 5 vor. Somit erhält man $\alpha_{(3)} = 0.05/3$, $\alpha_{(4)} = 0.05/4$ und $\alpha_{(5)} = 0.05/5$. Nun können die ursprünglichen p-Werte mit den zugehörigen $\alpha_{(k)}$ verglichen werden. Diese nennt man „adjustierte α“.

3.1.2 Kerndichteschätzungen

Robuste Methoden der statistischen Ausreißeranalytik univariater Daten verwenden als Lage- und Streuungsmaß häufig den Median und den Median der absoluten Abweichungen vom Median, kurz MAD (Median Absolute Deviation). Diese können zu beobachteten Daten $x_1, \ldots, x_n \in \mathbb{R}$ auf mehrere Arten geschätzt werden, z.B. kann der Median $\tilde{x}$ als 50%-Quantil der empirischen Häufigkeitsverteilung oder als zentraler Wert der Ordnungsstatistik geschätzt werden. Die MAD ist

dann definiert als Median der absoluten Abweichungen vom Median.

$$MAD := median\{|x_i - \tilde{x}|, \ldots, |x_n - \tilde{x}|\}$$

Eine weitere Möglichkeit besteht darin, die Daten $x_1, \ldots, x_n$ zunächst als Realisationen von stochastisch unabhängigen und identisch verteilten reellen Zufallsvariablen $X_1, \ldots, X_n$ aufzufassen. Die Dichtefunktion der unbekannten Verteilung P^{X_1} kann mit Hilfe von Kerndichteschätzungen geschätzt werden. Daraus lassen sich im nächsten Schritt Schätzungen für den Median und die MAD ableiten. Die Vorgehensweise wird im Folgenden skizziert.

Seien univariate Daten $x_1, \ldots, x_n$ gegeben. Diese werden aufgefasst als Realisationen von stochastisch unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$. Die Wahrscheinlichkeitsverteilung P^{X_1} sei unbekannt. Es wird angenommen, dass P^{X_1} durch eine stetige Dichtefunktion beschrieben werden kann. Im Falle normal-(0,1)-verteilter Daten hat diese Dichte die Gestalt

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},\ x \in \mathbb{R}.$$

Hier ist f unbekannt und soll geschätzt werden. Diverse Autoren diskutierten verschiedene Möglichkeiten eine Dichteschätzung durchzuführen. Hier soll der klassische Kerndichteschätzer vorgestellt werden. Er wurde erstmalig von Rosenblatt eingeführt, um Kerndichteschätzungen durchzuführen und theoretische Eigenschaften herzuleiten bzw. zu diskutieren. Breiman et al. verwenden zur Schätzung der Dichtefunktion die „variable kernel method“ [BMP77]. Diese berück-

sichtigt für jeden Datenpunkt x_i dessen Abstand zum nächstgelegen Datenpunkt in der Stichprobe. Whittle formuliert eine allgemeine Klasse von Gewichtungsfunktionen, um Kerndichteschätzungen durchzuführen [Whi58]. B.W. Silverman entwickelte einen statistischen Test, um die zugrunde liegende unbekannte Verteilung auf Multimodalität zu untersuchen [Sil81]. Hall gibt analytische und numerische Approximationen der Verteilung dieser Teststatistik [HW96]. Die Methode der Kerndichteschätzung wurde stetig weiterentwickelt im Hinblick auf ihre Konvergenzrate, Unverzerrtheit und die optimale Wahl des Glättungsparameters [Hal90a, HW96, Hal87, HW95, HW88, Hal90b, HT99, Hal90c, Kur06, MVM10, Tha09].

Der Kerndichteschätzer $\hat{f}$ einer Dichtefunktion f habe allgemein die folgende Form:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \ .$$

Dabei ist K eine reelle Funktion (sog. Glättungskern), für die gilt:

$$K : M \to [0, \infty) \ \text{ mit } \ \emptyset \neq M \subset \mathbb{R}, \ \int_M K(x)dx = 1 \ .$$

Der Glättungsparameter h wird auch Fensterbreite genannt. In der praktischen Anwendung ersetzt man K z.B. durch die Dichtefunktion der Normal-(0,1)-Verteilung (sog. Gauß-Kern). Weiterhin oft verwendet werden der Cauchy-Kern, der Picard-

Kern und der Epanechnikov-Kern.

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ (Gauß-Kern)}$$

$$K(x) = \frac{1}{\pi(1+x^2)} \text{ (Cauchy-Kern)}$$

$$K(x) = \frac{1}{2} e^{-|x|} \text{ (Picard-Kern)}$$

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{wenn } x \in [-1,1] \\ 0 & \text{sonst} \end{cases} \text{ (Epanechnikov-Kern)}$$

Verschiedene Glättungskerne finden sich in der einschlägigen Literatur, z.B. bei Pan et al. [PFF00] und bei Hall [Hal92]. Auf diese Weise erhält man eine Schätzung der Verteilung, die den Daten zugrunde liegt. In **R** steht zu diesem Zweck die Funktion `density` im Paket `stats` zur Verfügung [Tea10]. Sie lässt unter anderem die Wahl zwischen mehreren bereits implementierten Glättungskernen und der Fensterbreite h.

3.1.3 Robuste Schätzung linearer Modelle mit der LMS Methode

Statistische Ausreißer können die Kleinste-Quadrate Schätzung (Ordinary Least Squares, kurz OLS) eines linearen Modells (z.B. Zeitreihenmodelle, Trendgeraden usw.) verzerren. Vielfältige Beispiele geben Rousseeuw und Leroy [vgl. RL87]. Sie schlugen die Verwendung der sogenannten LMS Methode vor (Least Median of Squares), um die Modellparameter robust zu schätzen. Wie die OLS Methode angewendet werden kann um Zeitreihenmodelle zu schätzen, wird in Abschnitt 3.5.2.1 erläutert. Hier soll zunächst kurz der klassische OLS Ansatz

vorgestellt werden. Zunächst sei ein lineares Modell gegeben [vgl. RL87, S.1].

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i, \;\; i = 1, \ldots, n \tag{3.1}$$

Die Störgröße ε_i wird häufig als normal-$(0, \sigma)$-verteilt angenommen, wobei $\sigma \in \mathbb{R}$ ein unbekannter Parameter ist. Im Modell ist y_i die abhängige Variable und $x_{i1}, \ldots, x_{ip}$ sind die erklärenden Variablen [vgl. RL87, S.1]. Die unbekannten Parameter $\beta_1, \ldots, \beta_p \in \mathbb{R}$ sollen anhand gegebener Daten geschätzt werden. Ihre Schätzungen seien bezeichnet mit $\hat{\beta}_1, \ldots, \hat{\beta}_p$. Angenommen es läge eine Schätzung der Modellparameter vor. Setzt man diese in die Modellgleichung ein, so erhält man

$$\hat{y}_i = x_{i1}\hat{\beta}_1 + \ldots + x_{ip}\hat{\beta}_p, \;\; i = 1, \ldots, n \;\; .$$

Unter Verwendung der OLS Methode werden die Parameter $\beta_1, \ldots, \beta_p$ derart geschätzt, dass

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.2}$$

minimal wird [vgl. RL87, S.1 ff.]. Es ist bekannt, dass dieser Ansatz nicht robust gegenüber Ausreißern ist [vgl. RL87, HR09, HRRS05]. Um dieses Problem zu lösen, schlagen Rousseeuw und Leroy die LMS Methode vor. Diese liefert eine robuste Schätzung des Modells 3.1. Sie verwendet statt der Summe in Gleichung (3.2) den Median [vgl. RL87, S.14].

$$\min_{\beta} median\{y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n\} \tag{3.3}$$

Rousseeuw und Leroy stellen fest, dass es keine explizite Formel gibt, um β direkt aus den Daten zu schätzen. Vielmehr ist der Einsatz entsprechender Software notwendig, um das Optimierungsproblem (3.3) zu lösen [vgl. RL87, S.29]. In **R** steht zu diesem Zweck im Paket `MASS` die Funktion `lqs` zur Verfügung [Rip11].

3.1.4 Datentransformation und Kompositionsdaten

Viele Umweltdaten werden in Prozent, ppm (parts per million) o.ä. angegeben. Diese Daten nennt man auch „Kompositionsdaten“ (engl. „compositional data“) [vgl. Ait86]. Sie werden als Teile eines Ganzen verstanden, da sie stets bestimmte Werte nicht unter- bzw. überschreiten können (z.B. 0% und 100%) und ihre Summe stets einen bestimmten Maximalwert annimmt (z.B. 100%).

Genauer gesagt, enthalten Kompositionsdaten nur relative Informationen, d.h. Informationen über die Verhältnisse der einzelnen Anteile in der Probe zueinander. Würden beispielsweise in der chemisch-analytischen Untersuchung von Weidelgras (Lolium perenne) die Konzentrationen aller möglichen Stoffe und Chemikalien in der Probe (z.B. Chlorophyll, PCB 28, PCB 52 etc.) in Prozent erfasst, so ergäbe sich deren Summe zu 100% - man hätte die Probe gewissermaßen in ihre sämtlichen Einzelteile zerlegt und die jeweiligen Anteile der vorhandenen Stoffe in Prozent oder ggf. in $\mu g/kg$ o.ä. angegeben.

In der Praxis ist dieses Vorgehen selbstverständlich nicht umsetzbar. Es wird nicht gelingen eine Probe vollständig zu zerlegen, jedoch stellt dies kein Problem für die mathematische Modellierung dar. Alle nicht in der Analyse der Probe erfassten Stoffe werden dem „Rest“ zugeschrieben. Analysiert man eine Weidelgrasprobe beispielsweise auf PCB und nicht auf Dioxine, so erhält man aus

der chemischen Analytik entsprechende Konzentrationen für die PCB-Kongenere, vgl. Abschnitt 2. Dies bedeutet aber nicht, dass in der Probe keine Dioxine vorhanden sind. Vielmehr kennt man nun die Anteile der einzelnen PCB-Kongenere an der Probe, sowie ihre Verhältnisse untereinander und auch ihre Verhältnisse im Vergleich zum Rest der in der Probe zwar vorhandenen, aber nicht einzeln nachgewiesenen Stoffe.

Wird die Konzentration einer einzigen Chemikalie in der Probe untersucht, so werden genau genommen immer zwei Werte erfasst, denn neben der gemessenen Konzentration erhält man auch die Information über den Anteil aller anderen Chemikalien am Ganzen, sprich den Rest.

Für die Praxis folgt daher: wenn bekannt ist, dass eine Probe aus k Stoffen besteht, so müssen nur die Konzentrationen $c_1, \ldots, c_{k-1}$ der ersten $k-1$ Stoffe bestimmt werden. Die Konzentration des k-ten Stoffes c_k ergibt sich dann aus der Differenz zum Ganzen: $c_k = 100\% - c_1 - \ldots - c_k$. Aus mathematischer Sicht bedeutet dies, dass eine Reduktion der Dimension der Daten im Sinne einer Verminderung der Anzahl zu messender Größen (bzw. Stoffe wie z.B. PCB 28, PCB 52 etc.) möglich ist, wenn die Messung eines Merkmals (z.B. PCB 28) ausgelassen und der entsprechende Wert als Differenz der Gesamtsumme aller anderen gemessenen Größen zu 100% angegeben wird.

Zur Illustration ein kleines Beispiel. Angenommen eine Probe bestünde aus drei verschiedenen Stoffen: zu 20% aus Stoff A, zu 50% aus Stoff B und zu 30% aus Stoff C. Alle zusammen ergeben 100%, also die gesamte Probe. Es ist prinzipiell nicht notwendig die Probe auf alle drei Stoffe zu untersuchen. Es ist ausreichend nur die Konzentrationen der Stoffe A und B zu bestimmen, denn die Konzentration von Stoff C ergibt sich aus der Differenz 100%-20%-50%=30%. Die Messung der

dritten Konzentration ist folglich überflüssig.

Weitere Beispiele geben Buccianti und Pawlowsky-Glahn [vgl. BPG05]. Filzmoser et al. beschreiben die Problematik der Notwendigkeit einer Transformation für Umweltdaten [vgl. FHR09]. Die Wahl der passenden Transformation und deren Einfluss auf statistische Kenngrößen wird dort diskutiert.

Dieser Abschnitt befasst sich mit der Frage, warum Datentransformationen sinnvoll und notwendig für Kompositionsdaten sind. Welcher Voraussetzungen bedarf es, um eine logratio Transformation anwenden zu können? Warum werden transformierte Daten anstelle der Originaldaten einer statistischen Analyse unterzogen? Welchen Einfluss hat die Transformation der Daten auf die Verlässlichkeit statistischer Kenngrößen? Sind statistische Analysen auch sinnvoll, wenn die Daten nicht transformiert werden?

Die euklidische Geometrie ist auf Kompositionsdaten nicht anwendbar, d.h. der Abstand zwischen zwei in einer Probe gemessenen Konzentrationen von Stoffen ergibt sich nicht über deren Differenz. Um dies einzusehen, beantworte man sich selbst folgende Frage. Welche der beiden Aussagen ist hilfreicher, wenn man versucht die Zusammensetzung einer Probe zu beschreiben, die von einem Stoff A $1mg/kg$ und von einem Stoff B $0.1mg/kg$ enthält?

Aussage 1: *In der vorliegenden Probe sind* $0.9mg/kg$ *mehr von Stoff A, als von Stoff B enthalten.*

Aussage 2: *In der vorliegenden Probe ist Stoff A zehnmal stärker vertreten als Stoff B.*

Wenn die konkreten Daten ($1mg/kg$ und $0.1mg/kg$) nicht bekannt wären, so wäre es schwierig mit Aussage 1 überhaupt etwas anzufangen. Ist der Unterschied von $0.9mg/kg$ bedeutsam oder nicht? Betragen die Konzentrationen von Stoff A und Stoff B $400mg/kg$ und $399.1mg/kg$, dann ist es schwer vorstellbar, dass ein Unterschied von $0.9mg/kg$ stark ins Gewicht fällt. Aussage 2 hingegen wäre auch dann informativ, wenn die konkreten Messdaten ($1mg/kg$ und $0.1mg/kg$) nicht vorhanden wären. Es ist also festzustellen, dass nicht der Betrag der Differenz der Konzentrationen die wahre Natur der Daten wiedergibt, sondern ihr Verhältnis. Der Betrag der Differenz beschreibt aber gerade die euklidische Metrik. Diese ist für die Anwendung auf Kompositionsdaten folglich nutzlos [vgl. AE05, FHR09, Ait03b].

Aitchison entwickelte einen Abstandsbegriff für Kompositionsdaten, die sog. Aitchison-Geometrie [vgl. Ait86]. Die Bestimmung vieler statistischer Kenngrößen wie Mittelwerte, aber auch Darstellungsformen wie Boxplots und Histogramme, ist nur sinnvoll, wenn den verwendeten Daten die euklidische Geometrie zugrunde liegt. Ein Ausweg für die Analyse von Kompositionsdaten wäre, alle Kenngrößen und Verfahren in der Aitchison Geometrie darzustellen. Diese Vorgehensweise ist jedoch hochgradig ineffizient. Leichter ist es hingegen die Daten zu transformieren, so dass den dadurch gewonnenen Daten die euklidische Geometrie zugrunde liegt. Diese können mit den bekannten statistischen Verfahren analysiert und die Ergebnisse rücktransformiert werden. Man benötigt somit nur eine Vorschrift zur Transformation und eine zur Rücktransformation.

Filzmoser et al. betrachten insbesondere die Auswirkungen der alr, clr und der ilr Transformation [vgl. FHR09, S.6101 ff.]. Es sei $x = (x_1, \ldots, x_D)$ eine D-dimensionale Beobachtung (z.B. werden die Konzentrationen $D-1$ verschiedener

Chemikalien in Weidelgras gemessen und der x_D stellt den „Rest“ dar). Die alr, clr und ilr Transformation sind im Folgenden gegeben durch $alr(x)$, $clr(x)$ und $ilr(x)$.

$$alr(x) = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j} \right) \text{ für ein } j \in \{1, \ldots, D\}$$

$$clr(x) = \left(\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)$$

$$ilr(x) = (z_1, \ldots, z_{D-1}), \text{ mit } z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \quad i = 1, \ldots, D-1$$

Die ilr Transformation liefert nach Filzmoser et al. stets konsistente Darstellungen von Kompositionsdaten bzgl. der euklidischen Geometrie. Dort können statistische Standardmethoden zur Parameterschätzung, Boxplots, Histogramme und Verfahren zur Schätzung der Verteilungsdichte eingesetzt werden [vgl. FHR09, S.6102 ff.]. Die alr und die clr Transformation sind unter gewissen Voraussetzung auf Kompositionsdaten mit $D = 2$ ebenfalls anwendbar, allerdings nicht in einer so universellen Art und Weise wie die ilr Transformation. Im Falle $D \geq 3$ hat die alr Transformation Nachteile gegenüber der ilr Transformation. So ist z.B. die Wahl des Nenners x_j in der Transformationsformel subjektiv. Die Wahl von x_j beeinflusst die Gestalt der resultierenden alr transformierten Daten und kann in der späteren Auswertung zu unterschiedlichen Ergebnissen führen [vgl. PGETD08, FHR09]. Ein großer Vorteil der ilr Transformation gegenüber der alr und clr Transformation liegt in ihrer Isometrie bezüglich Aitchison Geometrie und euklidischer Geometrie. Diese Isometrieeigenschaft besagt folgendes: seien x, y Kompositionsdaten. Der Abstand zweier Punkte in der Aitchison-Geometrie sei

mit $d_A(\cdot,\cdot)$ und der euklidische Abstand mit $d_E(\cdot,\cdot)$ bezeichnet. Dann gilt:

$$d_A(x,y) = d_E(ilr(x), ilr(y)) \ .$$

Hier wird die Verwendung der ilr Transformation empfohlen. Die oft verwendete rein logarithmische Transformation von Kompositionsdaten liefert fehlerhafte Resultate und wird von Filzmoser et al. nicht empfohlen [vgl. FHR09, S.6103].

Nachdem Kompositionsdaten einer ilr-Transformation unterzogen und statistisch untersucht wurden, ist es ggf. notwendig die so erhaltenen Ergebnisse einer inversen ilr-Transformation zu unterziehen, um sie im Raum der Ursprungsdaten darzustellen. Dies ist z.B. bei zeitreihenanalytischen Methoden notwendig, wenn die ilr-transfomierten Zeitreihen von Ausreißern bereinigt wurden. Wie sehen die korrespondierenden, von Ausreißern bereinigten, Zeitreihen in Kompositionsdarstellung aus? Eine Antwort liefert die inverse ilr-Transformation. Sie wird in zwei Schritten ausgeführt. Sei y ein $(D-1)$-dimensionaler ilr-transformierter Datenpunkt. Der zugehörige Datenpunkt x in Kompositionsdarstellung ergibt sich dann wie folgt.

$$\begin{aligned}
y &= (y_1, \ldots, y_{D-1})', \ \ x = (x_1, \ldots, x_D)' \\
z_i &:= \sum_{j=i}^{D} \frac{y_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}} y_{i-1}, \ \ y_0 := y_D := 0, \ \ i = 1, \ldots, D \\
x_i &:= \frac{e^{z_i}}{e^{z_1} + \ldots + e^{z_D}}, \ \ i = 1, \ldots, D
\end{aligned}$$

3.1.5 Nulldaten und Messungen unterhalb der Bestimmungsgrenze

Häufig kommt es vor, dass die Präsenz von Chemikalien in Biota nicht nachgewiesen werden kann, obwohl sachlogisch davon auszugehen ist, dass zumindest ein geringer Anteil im Prüforganismus vorhanden ist. Die statistische Literatur beschäftigte sich intensiv mit dem Problem der 0 in Kompositionsdaten, dazu siehe Aitchison [Ait86], Fry et al. [FFM00] und Bacon-Shone [BS92]. Misst man beispielsweise die Konzentration von PCB 28, PCB 52 etc. in Weidelgras, so kann es vorkommen, dass die Konzentration von PCB 28 so gering ist, dass sie unterhalb der Bestimmungsgrenze liegt und folglich nicht angegeben werden kann. Die gemessene Konzentration beträgt dann 0. Dies entspricht offenbar keinesfalls der Realität, wenn davon auszugehen ist, dass in der Tat eine Kontamination von Weidelgras durch PCB 28 vorliegt.

Wie aber gibt man diesen sehr geringen, nicht messbaren Wert an? Ist es notwendig ihn überhaupt anzugeben? Welche Konsequenzen hätte es, wenn weiterhin als Messwert die 0 beibehalten würde? Ist es sinnvoller statt 0 die Bestimmungsgrenze zu wählen?

Fakt ist, dass eine Identifikation der unbekannten Konzentration mit dem Wert 0 eine Unterschätzung der tatsächlichen Konzentration darstellt. Die Streuung der Daten wird verzerrt (vergrößert) und eine logratio Transformation der Daten ist nicht möglich, da gilt $\lim_{x \searrow 0} \ln x = -\infty$. Ersetzt man hingegen die als 0 angenommene Konzentration durch die Bestimmungsgrenze, so überschätzt man die tatsächlich vorhandene Konzentration und verzerrt (verringert) die tatsächliche Streuung der Daten. Beide Alternativen sind daher abzulehnen.

In der statistischen Literatur werden verschiedene Gründe für Nulldaten unterschieden. Zum einen entstehen solche Werte durch Rundung, zum anderen kann es sich um echte Nullwerte handeln. In obigem Beispiel sind sowohl ein Rundungsfehler als auch eine echte Null plausibel. Lediglich der sachlogische Hintergrund suggeriert, dass es sich um einen Rundungsfehler handelt, da Weidelgras in der Umwelt vielen Belastungen ausgesetzt ist, insbesondere durch PCB. Warum sollte sich PCB 28 nicht in der vorliegenden Probe angereichert haben, wenn andere Messungen bestätigen, dass oft eine Belastung durch PCB 28 vorliegt?

Aitchison argumentierte, dass, abhängig von dem Grund der gemessenen Nulldaten, eine Modifikation der Messwerte durchgeführt werden kann. Fry et al. zeigten, dass, unabhängig vom Grund für das Auftreten der Nulldaten, eine Modifikation zur Umwandlung dieser Messwerte angewendet werden kann [vgl. FFM00].

Die in der statistischen Literatur vorgestellten Methoden umfassen unter anderem statistische Ränge, die sog. Verschmelzung (engl. amalgamation) und Ersetzungsverfahren.

Die Verwendung von Rangdaten zeigt einen deutlichen Nachteil, da Informationen verloren gehen. Dieser Ansatz scheint daher wenig praktikabel.

Bei der Verschmelzung werden Merkmale zusammengefasst, wörtlich „verschmolzen“. Weist ein Merkmal (z.B. die Konzentration von PCB 28 in Weidelgrasproben) eine 0 auf, so kann man dieses Merkmal in einer Gruppe mit anderen Merkmalen zusammenfassen und ein gemeinsames, „verschmolzenes“ Merkmal erstellen. Diese Vorgehensweise birgt einige Schwachstellen. Informationen über die ursprünglichen Merkmale und deren Verhältnisse untereinander gehen bei der Verschmelzung verloren und Ergebnisse aus angepassten Modellen sind folglich schwerer zu interpretieren, da auch Wechselwirkungen zwischen den verschmol-

zenen Merkmalen nicht mehr erkennbar sind.

Das Ersetzungsverfahren hat zum Ziel, sämtliche Nullwerte durch sorgfältig ausgewählte kleine Werte zu ersetzen und parallel dazu die Nichtnulldaten anzupassen, um weiterhin die Verhältnisse der Daten untereinander zu erhalten. Diese Methode ist bekannt unter dem Namen „modified Aitchison“, siehe [vgl. AE05] und [vgl. Ait03a]. Angenommen es liegt ein Kompositionsdatenpunkt $x = (x_1, \ldots, x_D)$ vor. Dieser enthalte L Nullen und folglich $D - L$ Werte größer als 0. Die Eigenschaften von Kompositionsdaten sind in Abschnitt 3.1.4 erläutert. Der maximale Rundungsfehler sei δ. Weiterhin seien $\tau_A = \delta(L+1)(D-L)/D^2$ und $\tau_S = \delta L(L+1)/D^2$. Die Nulldaten werden jeweils durch τ_A ersetzt und parallel dazu alle Nichtnulldaten x_i um $x_i \cdot \tau_S$ verringert. Der maximale Rundungsfehler δ entspricht für gewöhnlich der Bestimmungsgrenze, da zum einen alle Werte kleiner oder gleich dieser Bestimmungsgrenze als 0 gemessen werden und somit höchstens ein Fehler in Höhe der Bestimmungsgrenze auftreten kann. Andererseits ist der Fehler bei der Messung für gewöhnlich kleiner als die Bestimmungsgrenze. Die Wahl von δ ist somit kontextspezifisch. Allgemein sei der Messfehler bei der Datenerfassung mit Δ und die Bestimmungsgrenze mit Λ bezeichnet. Dann ergibt sich $\delta = \max\{\Delta, \Lambda\}$. Liegen Kompositionsdaten $x = (x_1, x_2)$ vor, so werden Nullwerte durch die halbe Bestimmungsgrenze ersetzt.

3.1.6 Fehlende Werte in Kompositionsquerschnittsdaten

Bei der chemischen Analyse wird eine Probe nie vollständig in all ihre Bestandteile aufgeschlüsselt. Zum einen ist dies praktisch nicht möglich, zum anderen sprechen Kosten und zeitlicher Aufwand gegen eine zu detaillierte Untersuchung.

So kann es vorkommen, dass z.B. nur der Gehalt von PCB28 und PCB52 in einer Weidelgrasprobe untersucht wird, nicht aber die Gehalte der restlichen PCB-Typen. Auch können Daten bei der Übertragung in Datenbanken verloren gehen.

Fehlende Werte in multivariaten Daten stellen ein großes Problem bei der statistischen Analyse dar. Liegen für einen multivariaten Datenpunkt keine vollständigen Informationen vor, so kann dieser nicht ohne Weiteres in der Analyse der Gesamtdaten berücksichtigt werden. Ist beispielsweise ein Datensatz von n Weidelgrasproben zu PCB Gehalten der Gestalt

$$x_i = \begin{pmatrix} PCB28_i \\ PCB52_i \\ PCB101_i \\ PCB138_i \\ PCB153_i \\ PCB180_i \\ Rest_i \end{pmatrix}$$

gegeben, und liegen zu einem Datenpunkt nur Werte für $PCB28$ und $PCB180$ vor, müssen die fehlenden Messwerte zu $PCB52$, $PCB101$, $PCB138$, $PCB153$ und $Rest$ geschätzt werden, um den Datenpunkt nicht vollständig von der multivariaten Analyse ausschließen zu müssen. Selbstverständlich ist der Ausschluss eines Datenpunktes, der fehlende Werte aufweist, stets eine Option, aber in gleichem Maße ist es erstrebenswert fehlende Werte zu schätzen um den Datenpunkt nicht zu verwerfen, da jedes Datum stets eine wertvolle Information darstellt. Durch fehlende Werte in einem Datum wird die multivariate Interpretation des

Punktes zwar erschwert bzw. verhindert, jedoch enthalten die nicht fehlenden Messwerte immer noch Informationen über das entsprechende univariate Merkmal.

In dieser Arbeit werden ausschließlich Kompositionsdaten betrachtet. Aus diesem Grund werden speziell Verfahren zur Schätzung fehlender Werte in Kompositionsquerschnittsdaten und Kompositionszeitreihendaten betrachtet. Um den Einfluss statistischer Ausreißer auf die Schätzung zu mildern, werden insbesondere robuste Verfahren vorgestellt.

Um fehlende Werte in Kompositionsquerschnittsdaten zu ersetzen, stehen zwei Methoden zur Verfügung. Ein Ansatz beschreibt die Schätzung durch die Betrachtung der k nächst gelegenen Nachbarn (k nearest neighbours [vgl. HTF10a]). Eine weitere Möglichkeit besteht darin die Daten modellbasiert iterativ zu schätzen. Die Methode der k nächsten Nachbarn ist ein praktikables Verfahren, welches als Initialisierungsschritt für den modellbasierten Ansatz genutzt werden kann [vgl. HTF10a]. Mit Hilfe des modellbasierten Ansatzes (siehe Abschnitt 3.1.6.2) können fehlende Werte iterativ sehr schnell geschätzt werden. Die Methode ist robust gegenüber statistischen Ausreißern und berücksichtigt im Gegensatz zur Methode der k nächsten Nachbarn die multivariaten Beziehungen zwischen den Daten.

3.1.6.1 Die Methode der k nächsten Nachbarn

Die Anwendung statistischer Standardmethoden auf Kompositionsdaten führt häufig zu Trugschlüssen und unverlässlichen Ergebnissen. Dies betrifft sowohl Standardmethoden, wie die Bildung eines arithmetischen Mittelwertes oder die Berechnung von Korrelationen [vgl. FHR09, FH09], als auch Methoden zur Schätzung fehlender Werte [vgl. MFBVPG03, BTDB06]. Die Methode der k nächsten

Nachbarn verwendet die statistischen Informationen den nächsten Nachbarn eines Datenpunktes, der fehlende Werte aufweist. Die nächsten Nachbarn eines Punktes hängen selbstverständlich vom verwendeten Abstandsbegriff ab. Kompositionsdaten unterliegen der sogenannten Aitchison-Metrik, daher wird dieser Abstandsbegriff zum Auffinden benachbarter Datenpunkte verwendet (s. Abschnitt 3.1.4).

3.1.6.1.1 Algorithmus zur Schätzung fehlender Werte mit der Methode der k nächsten Nachbarn Der folgende Algorithmus wurde inspiriert durch die Vorgehensweise von Hron et al. [vgl. HTF10a, S.5 ff]. Seien Kompositionsdaten $x_1, \ldots, x_n$ gegeben mit $x_i = (x_{i1}, \ldots, x_{iD})$, $i = 1, \ldots, n$.

1. Wähle ein $k \in \{1, \ldots, n\}$.

2. Für jeden Datenpunkt x_i sei M_i die Menge der Indizes der fehlenden Werte von x_i und O_i die Menge der Indizes der vorhandenen Werte in x_i, z.B. für $x_1 = (0.1, fehlt, 0.3, 0.15)$ wäre $M_1 = \{2\}$ und $O_1 = \{1, 3, 4\}$.

3. Setze $i := 1$.

4. (a) Der aktuell betrachtete Datenpunkt ist x_i. Wenn $M_i = \emptyset$, d.h. in x_i fehlen keine Werte, und wenn $i < n$ ist, dann setze $i := i + 1$ und gehe zu Schritt 4(a). Wenn in x_i keine Werte fehlen und $i = n$ gilt, so wurden alle Datenpunkte auf fehlende Werte untersucht und der Algorithmus endet. Wenn $M_i \neq \emptyset$, d.h. in x_i fehlen Werte, so betrachte nacheinander alle Stellen $j \in M_i$, sprich alle Stellen, an denen Werte in x_i fehlen. Für jedes $j \in M_i$ führe die Schritte 4(b)-4(e) aus.

   (b) Wähle alle Datenpunkte $x \in \{x_1, \ldots, x_n\}$, denen sowohl kein Wert

an der Stelle j, als auch kein Wert an den Stellen O_i fehlt. Bezeichne diese Datenpunkte mit $y_1, \ldots, y_r$.

(c) Bestimme den Aitchison Abstand jedes dieser Datenpunkte zu x_i und wähle die k Datenpunkte mit den geringsten Abständen aus. Bezeichne diese k Datenpunkte mit $z_1, \ldots, z_k$.

(d) Beachte: Der aktuell betrachtete Datenpunkt ist immer noch x_i. Dieser hat an der Stelle j einen fehlenden Wert. Die k nächsten Nachbarn von x_i sind $z_1, \ldots, z_k$. An allen Stellen $o \in O_i$ haben $x_i, z_1, \ldots, z_k$ keinen fehlenden Wert, d.h. $x_{io}, z_{1o}, \ldots, z_{ko}$ sind keine fehlenden Werte für jedes $o \in O_i$. Für $l = 1, \ldots, k$ bestimme die Werte

$$f_l = \frac{median\{x_{io} \mid o \in O_i\}}{median\{z_{lo} \mid o \in O_i\}}$$

(e) Der fehlende Wert x_{ij} des Vektors x_i wird nun ersetzt.

$$x_{ij} = median\{f_1 z_{1j}, f_2 z_{2j}, \ldots, f_k z_{kj}\}$$

5. Wenn $i = n$ gilt, so endet der Algorithmus, ansonsten setze $i := i + 1$ und gehe zu Schritt 4(a).

In **R** steht das Paket robCompositions zur Verfügung. Dieses beinhaltet die bereits implementierte Prozedur `impKNNa()`, mit deren Hilfe das hier beschriebene Verfahren umgesetzt werden kann [vgl. THF10].

3.1.6.2 Der modellbasierte Ansatz zur Schätzung fehlender Werte

3.1.6.2.1 Algorithmus zur Schätzung fehlender Werte mit dem modellbasierten Ansatz Der folgende Algorithmus findet sich in leicht verkürzter Form bei Hron et al. [vgl. HTF10a, S.7 ff]. Sei $x_1, \ldots, x_n$ ein Datensatz von Kompositionsdaten mit $x_i = (x_{i1}, \ldots, x_{iD})'$, $i = 1, \ldots, n$. Jeder Datenpunkt x_i besitzt folglich D Einträge, auch Merkmale oder Variablen genannt. Die Daten $x_1, \ldots, x_n$ werden zeilenweise in eine Datenmatrix X^* vom Format $n \times D$ eingetragen, d.h. die i-te Zeile von X^* entspricht dem Kompositionsdatenpunkt x_i und Spalte j entspricht dem Merkmal j ($i = 1, \ldots, n$ und $j = 1, \ldots, D$). Sei M_j die Anzahl der fehlenden Werte in Spalte j. Ordne nun die Spalten von X^* so um, dass die Spalte mit den meisten fehlenden Werten zur ersten Spalte wird, die Spalte mit den zweitmeisten fehlenden Werten die zweite Spalte wird usw. . Durch die Umordnung der Spalten der Datenmatrix wurden lediglich die Merkmale umgeordnet, was für Kompositionsdaten nicht weiter problematisch ist. Welchen Unterschied macht es beispielsweise, wenn sich die Reihenfolge der gemessenen Konzentrationen von $(PCB28,\ PCB52,\ PCB101,\ Rest)$ zu $(PCB52,\ PCB101,\ PCB28,\ Rest)$ ändert? Für die Analyse von Kompositionsdaten ist die Reihenfolge der Einträge der Merkmale unbedeutend, da sämtliche Informationen in den Verhältnissen (Quotienten) der Messwerte liegen und nicht in ihren Absolutwerten oder deren Reihenfolge. Das hier beschriebene Verfahren endet, wenn eine vorher festgelegte Grenze von δ nicht mehr überschritten wird, z.B. $\delta = 0.001$

1. Setze $X := X^*$.

2. Verwende die Methode der k nächsten Nachbarn, um eine Initialisierungsschätzung der fehlenden Werte in den Datenpunkten $x_1, \ldots, x_n$ zu erhalten.

3. Für $j = 1, \ldots, D$ sei M_j die Anzahl fehlender Werte in Spalte j von Matrix X. Es muss gelten $M_1 \geq M_2 \geq \ldots \geq M_D$.

4. Setze $l := 1$.

5. Bestimme zu $x_1, \ldots, x_n$ die korrespondierenden ilr transformierten Daten $y_1, \ldots, y_n$. Fasse diese zeilenweise in einer Datenmatrix Y vom Format $n \times (D-1)$ zusammen.

6. Definiere m_l als die Menge der Indizes der fehlenden Werte in Spalte l der Matrix X^* und o_l sei die Menge der Indizes vorhandener Werte in Spalte l, sprich $o_l = \{1, \ldots, n\} \backslash m_l$. Weiterhin seien $y_l^{o_l} := \{y_{il} \mid i \in o_l\}$ und $y_l^{m_l} := \{y_{il} \mid i \in m_l\}$, sowie $Y_{-l}^{o_l}$ die Matrix der y_i mit $i \in o_l$, wobei die l-te Spalte dieser Matrix entfernt wird, analog dazu $Y_{-l}^{m_l}$.

   Ein Beispiel dazu

   Sei ein Kompositionsdatensatz x_1, x_2, x_3, x_4, x_5 mit $D = 3$ gegeben.

$$x_1 = \begin{pmatrix} 0.1 \\ 0.4 \\ 0.5 \end{pmatrix}, x_2 = \begin{pmatrix} - \\ 0.2 \\ - \end{pmatrix}, x_3 = \begin{pmatrix} - \\ - \\ 0.4 \end{pmatrix}, x_4 = \begin{pmatrix} 0.1 \\ 0.3 \\ 0.6 \end{pmatrix}, x_5 = \begin{pmatrix} 0.3 \\ 0.4 \\ 0.3 \end{pmatrix}$$

Damit hat X^* die folgende Gestalt:

$$X^* = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ - & 0.2 & - \\ - & - & 0.4 \\ 0.1 & 0.6 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix} .$$

Unter Verwendung der Methode der k nächsten Nachbarn werden für $k = 3$ die fehlenden Werte geschätzt.

$$X = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 2/30 & 0.2 & 0.25 \\ 0.08 & 0.32 & 0.4 \\ 0.1 & 0.6 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

Die ilr Transformation liefert nun eine Matrix $Y = (ilr(X_{i.}))_{i=1,\ldots,5}$ der Gestalt

$$Y = \begin{pmatrix} 1.2230026 & 0.1577863 \\ 0.9881112 & 0.1577863 \\ 1.2230026 & 0.1577863 \\ 1.1799893 & 0.4901291 \\ 0.1174457 & -0.2034219 \end{pmatrix} ,$$

wobei $X_{i.}$ die i-te Zeile von X darstellt. Für $l = 1$ ergeben sich dann $m_1 =$

$\{2,3\}$ und $o_1 = \{1,4,5\}$. Damit ist

$$y_1^{o_1} = \{y_{11}, y_{41}, y_{51}\} = \{1.2230026, 1.1799893, 0.1174457\}$$

und

$$y_1^{m_1} = \{y_{11}, y_{21}\} = \{0.9881112, 1.2230026\} \ ,$$

sowie

$$Y_{-1}^{o_1} = \begin{pmatrix} y_{12} \\ y_{42} \\ y_{52} \end{pmatrix} = \begin{pmatrix} 0.1577863 \\ 0.4901291 \\ -0.2034219 \end{pmatrix}, \ Y_{-1}^{m_1} = \begin{pmatrix} y_{22} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 0.1577863 \\ 0.1577863 \end{pmatrix} \ .$$

7. Schätze die Parameter des Modells

$$y_l^{o_l} = Y_{-l}^{m_l}\beta + \varepsilon$$

mittels linearer Regression. Die Schätzung des Parametervektors β sei mit $\hat{\beta}$ bezeichnet. Für den Fall, dass durch statistische Ausreißer kontaminierte Kompositionsdaten vorliegen, empfiehlt sich eine robuste Schätzung des Regressionsmodells.

8. Setze die Schätzung $\hat{\beta}$ in die folgende Gleichung ein

$$\hat{y}_l^{m_l} = Y_l^{m_l}\hat{\beta}$$

und ersetze in den ilr transformierten Daten die Werte $y_l^{m_l}$ durch $\hat{y}_l^{m_l}$.

9. Wende die inverse ilr Transformation auf die $y_1, \ldots, y_n$ an. Wichtig zu bemerken an dieser Stelle ist die Tatsache, dass neben den fehlenden Werten durch die Rücktransformation auch die nicht fehlenden Werte aktualisiert wurden. Diese haben nach Anwendung der inversen ilr Transformation definitiv andere numerische Werte, als noch zu Beginn des Verfahrens. Dies ist allerdings nicht weiter problematisch, da die Verhältnisse zwischen den einzelnen Variablen $1, \ldots, D$ durch das Verfahren unverändert bleiben. Wie in Abschnitt 3.1.4 dargestellt, liegt der wahre Informationsgehalt von Kompositionsdaten in den Quotienten der Variablen, nicht aber in den gemessenen Absolutwerten.

10. Führe die Schritte 5 bis 9 für $l = 2, \ldots, D-1$ aus.

11. Wiederhole die Schritte 4 bis 9 solange, bis der euklidische Abstand der empirischen Kovarianzmatrizen der ilr transformierten Messwerte der aktuellen Iteration und der vorhergehenden Iteration kleiner als δ ist.

In **R** steht das Paket robCompositions zur Verfügung. Dieses beinhaltet die Funktion `impCoda()`, mit deren Hilfe das hier beschriebene Verfahren umgesetzt werden kann [vgl. THF10]. Wie man aus der Formel zur Berechnung der ilr transformierten Daten (siehe auch Abschnitt 3.1.4)

$$ilr(x) = (z_1, \ldots, z_{D-1}), \quad \text{mit} \quad z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \quad i = 1, \ldots, D-1$$

erkennt, beeinflusst ein fehlender Wert an der Stelle j von Datenpunkt i (d.h. x_{ij} ist ein fehlender Wert) lediglich die Berechnung von $y_{i1}, \ldots, y_{ij}$. Das obige iterative Verfahren ersetzt nacheinander für jede Variable die fehlenden Werte,

d.h. es werden zunächst die fehlenden Werte in $x_{11}, \ldots, x_{n1}$ und anschließend die fehlenden Werte in $x_{12}, \ldots, x_{n2}$ usw. ersetzt.

3.1.7 Natürliche Variabilität

In diesem Abschnitt soll definiert werden, wie der Begriff „natürliche Variabilität“ innerhalb dieser Arbeit zu verstehen ist. Zunächst wird ein allgemeiner Begriff der natürlichen Variabilität gegeben und anschließend der spezielle Fall von Kompositionsdaten dargestellt.

Angenommen es wäre irgendwie möglich alle Einflussfaktoren oder Inputs eines Systems, sei es biologisch, chemisch oder technisch, perfekt zu kontrollieren und zu steuern bis ins kleinste Element, so gäbe es nur noch das Prinzip von Wirkung und Wechselwirkung. Gegeben ein bekannter Zustand Z_{t_0} eines Systems zum Zeitpunkt t_0, könnte man unter Kenntnis aller noch so kleinen physikalischen Kräfte und Wechselwirkungen bzw. funktionalen Zusammenhänge, chemischen Reaktionen und Vorhersehbarkeit jeglichen individuellen Verhaltens, den Zustand des Systems zum Zeitpunkt $t_1 > t_0$ exakt vorhersagen.

In der Praxis ist dieser Determinismus nur teilweise beobachtbar. Es gibt unbekannte Wechselwirkungen und chaotisches Verhalten in subatomaren Dimensionen. Individuelle Entscheidungen sind nur teilweise vorhersehbar, aber niemals vollständig bekannt. Darüber hinaus würde die Berücksichtigung aller bekannten Wechselwirkungen, und seien sie scheinbar noch so gering oder unbedeutend, die Analyse des Systems aufgrund seiner Komplexität praktisch unmöglich machen.

Mit anderen Worten: es existiert kaum ein vollständig bekanntes komplexes

System und die Berücksichtigung zu vieler Wechselwirkungen im System macht die Analyse praktisch unmöglich. Dies zeigt sich selbst unter kontrollierten Laborbedingungen. Führt man beispielsweise einen Bioconcentration Fish Test nach OECD Guideline 305 durch, so ist es zwar möglich die dort zu untersuchenden Fische möglichst homogen auszuwählen, aber die Konzentration der Chemikalie, die im Versuchsaufbau in Fisch 1 und Fisch 2 an Tag X gemessen wird, ist für beide Individuen selbstverständlich unterschiedlich. Kein Fisch gleicht exakt dem anderen und kein Messsystem ist so genau, dass es tatsächlich fehlerfrei misst. Wären alle Individuen gleich und würden die Chemikalie im Wasser gleichermaßen aufnehmen, einlagern, abbauen etc., so erhielte man für n Fische, die an Tag X auf die Konzentration der im Fischfett eingelagerten Chemikalie analysiert würden, genau n-mal den gleichen Messwert μ. Dies ist praktisch unmöglich.

Was bleibt, ist immer ein fehlerbehafteter Messwert, entstanden durch Einflüsse aus dem System, dem die zu untersuchenden Objekte entnommen wurden, und durch die Messung selbst. Modelliert man die Messfehler als Zufallsvariablen $M_1, \ldots, M_n$ und die Fehler die im System generiert werden durch Zufallsvariablen $S_1, \ldots, S_n$, so fordert man für Messungen $i = 1, \ldots, n$ stets die stochastische Unabhängigkeit von $M_1, \ldots, M_n, S_1, \ldots, S_n$ und die identische Verteilung von $M_1, \ldots, M_n$ bzw. $S_1, \ldots, S_n$ mit

$$E(S_i) = E(M_i) = 0, \;\; Var(M_i) = \sigma_M^2, \;\; Var(S_i) = \sigma_S^2 \;\; .$$

Es ist praktisch nicht möglich zwischen Messfehler M_i und Systemfehler S_i zu unterscheiden, da am Ende lediglich der reale Messwert x_i vorliegt. Allerdings ist es möglich davon auszugehen, dass der Messfehler M_i aufgrund kontrollierter

und definierter Rahmenbedingungen (z.B. geeichte Messinstrumente nach ISO) gering ist und der Hauptteil an Variabilität durch den Fehler S_i generiert wird. Als Messergebnisse bzw. Daten erhält man folglich Werte $x_1, \ldots, x_n$ die sich folgendermaßen zusammensetzen:

$$x_i = \mu + S_i + M_i, \;\; i = 1, \ldots, n \;\; .$$

Die Zufallsvariablen M_i und S_i werden zusammengefasst zu einer „natürlichen Störgröße“

$$\varepsilon_i = M_i + S_i \;\; .$$

3.1.7.1 Natürliche Variabilität in univariaten Querschnittsdaten

In univariaten Daten $x_1, \ldots, x_n$ sei die Varianz der Störgrößen ε_i bezeichnet als *natürliche Variabilität*. Wegen der stochastischen Unabhängigkeit von M_i und S_i ergibt sich die Varianz von ε_i zu

$$Var(\varepsilon_i) = \sigma_S^2 + \sigma_M^2 = \sigma_\varepsilon^2 \;\; .$$

Die Störgrößen ε_i werden häufig als normalverteilt angenommen. Warum diese Annahme sinnvoll ist, wird im Folgenden motiviert. Geht man davon aus, dass im System selbst viele kleine unbeobachtbare Fehler entstehen, die man wiederum als stochastisch unabhängige Zufallsvariablen auffassen kann, so ergeben diese in ihrer Summe den Systemfehler S_i. Addiert man zu diesem weiterhin den Messfehler M_i, der ebenfalls als Summe vieler kleiner unbeobachtbarer stochastischer Störungen angesehen werden kann, so ist es möglich die Störgröße ε als Summe

vieler stochastisch unabhängiger Zufallsvariablen zu interpretieren.

Der Zentrale Grenzwertsatz der Stochastik besagt nun: sei $Z_1, Z_2, \ldots$ eine Folge stochastisch unabhängiger und identisch verteilter Zufallsvariablen mit existierendem Erwartungswert $\mu = E(Z_1) < \infty$ und existierender Varianz $\sigma^2 = Var(Z_1) < \infty$. Weiter sei $S_n := \sum_{j=1}^{n} Z_j$ für $n = 1, 2, \ldots$, dann gilt für $Y_n := (S_n - n\mu)/(\sigma\sqrt{n})$

$$\lim_{n\to\infty} P(Y_n \leq y) = \Phi(y) \ , \ y \in \mathbb{R},$$

wobei Φ die Verteilungsfunktion der Normal-(0,1)-Verteilung darstellt [vgl. Sac09, Dav01, S.226,237,S.43ff].

Der Erwartungswert der Störgrößen sei $E(\varepsilon_i) = E(M_i + S_i) = E(M_i) + E(S_i) = 0$. Diese Forderung impliziert, dass die Störgrößen nicht als systematische Fehler auftreten. Wäre z.B. $E(\varepsilon_i) = \mu > 0$, so wäre der erwartete Messfehler positiv und die Daten folglich systematisch als zu groß gemessenen.

Die natürliche Variabilität wird als Varianz aus der Summe systeminterner Fehler und der Messfehler interpretiert. Die Varianz einer reellen Zufallsvariable X ist in euklidischen Räumen (z.B. $\mathbb{R}^n$ mit der euklidischen Metrik, welche den Abstand zweier Punkte $x, y \in \mathbb{R}^n$ durch $d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$ angibt) allgemein definiert durch

$$Var(X) = E(|X - E(X)|^2) \ .$$

Anhand von Daten $x_1, \ldots, x_n$, die als Realisationen von stochastisch unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ verstanden werden,

ergeben sich diverse Schätzer, wie z.B.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2, \;\; \hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

Eine Schätzung der Varianz erhält man, indem $X_1, \ldots, X_n$ durch Daten $x_1, \ldots, x_n$ ersetzt werden.

3.1.7.2 Natürliche Variabilität in multivariaten Querschnittsdaten

Seien multivariate Daten $x_1, \ldots, x_n \in \mathbb{R}^p$ gegeben. Diese werden aufgefasst als Realisationen von stochastisch unabhängigen und identisch verteilten multivariaten Zufallsvariablen $X_1, \ldots, X_n$ mit $X_i = (X_{i1}, \ldots, X_{ip})'$ für $i = 1, \ldots, n$. Unter der Annahme der Existenz des Erwartungswertvektors $\mu = E(X_1)$ und der Kovarianzmatrix S der Verteilung P^{X_1}, wird die natürliche Variabilität definiert als die Summe der Elemente der Hauptdiagonale der Kovarianzmatrix S (auch Spur von S genannt):

$$natvar := tr(S) = \sum_{j=1}^{p} Var(X_{1j}) \;\; .$$

Anders gesagt: es wird die natürliche Variabilität einer jeden Komponente des Zufallsvektors $X_1 = (X_{11}, \ldots, X_{1p})'$ betrachtet, unabhängig von den stochastischen Wechselwirkungen der Komponenten untereinander. Diese p natürlichen Variabilitäten $Var(X_{11}), \ldots, Var(X_{1p})$ werden danach durch Summation zusammengefasst zu einer gesamten natürlichen Variabilität.

3.1.7.3 Natürliche Variabilität in univariaten stationären Zeitreihen

Sei eine univariate stationäre Zeitreihe $x_1, \ldots, x_n$ gegeben. Diese folge einem AR$[p]$-Modell (Details siehe Abschnitt 3.5.1.1). Ein AR$[p]$-Modell hat die folgende Gestalt:

$$x_t = c_\pi - \pi_1 x_{t-1} - \ldots - \pi_p x_{t-p} + \varepsilon_t$$

d.h. der Zustand des Prozesses zum Zeitpunkt t hängt im Wesentlichen von den Zuständen des Prozesses zu den Zeitpunkten $t-1, t-2, \ldots, t-p$ und der Störgröße ε_t ab. An die Störgrößen ε_t werden gewisse Anforderungen gestellt. Um die Definition der natürlichen Variabilität zu motivieren, genüge an dieser Stelle die Idee, dass es sich um stochastisch unabhängige und identisch verteilte Zufallsvariablen mit $E(\varepsilon_t) = 0$ und $Var(\varepsilon_t) < \infty$ handelt. Die exakten Hintergründe werden in Abschnitt 3.5.1.1 erläutert. Die Störgrößen werden auch Innovationen des Prozesses genannt. Der Zustand des Prozesses zum Zeitpunkt t ergibt sich somit als Summe aus einer Linearkombination der Zustände des Prozesses zu den p vorangegangenen Zeitpunkten $x_{t-1}, \ldots, x_{t-p}$ und einer zufälligen Innovation ε_t. Der Prozess ist demnach „aus sich selbst entstehend“, „auf sich selbst zurückgreifend“ bzw. „Auto Regressive“. Ohne die Störgrößen ε_t wäre der Prozess deterministisch. Die natürliche Variabilität findet sich somit in den Innovationen der Zeitreihe. Sie wird daher definiert als die Varianz der Innovationen der Zeitreihe.

$$natvar = Var(\varepsilon_1)$$

3.1.7.4 Natürliche Variabilität in multivariaten stationären Zeitreihen

Sei eine multivariate Zeitreihe $X_1, \ldots, X_n$ gegeben, $X_i = (X_{i1}, \ldots, X_{ik})'$. Diese folge einem Vektor-AR[p]-Modell (kurz VAR[p], Details siehe Abschnitt 3.5.1.1). Die Interpretation des VAR[p]-Modells ist analog zum univariaten Fall.

$$\begin{aligned} X_t &= c_\Pi - \Pi_1 X_{t-1} - \ldots - \Pi_p X_{t-p} + E_t \\ E_t &= (E_{t1}, \ldots, E_{tk})',\ t = 1, \ldots, n \\ \Pi_j &\in \mathbb{R}^{k \times k},\ \ j = 1, \ldots, p \end{aligned}$$

$\Pi_1, \ldots, \Pi_p$ sind Koeffizientenmatrizen und werden analog zum univariaten Fall interpretiert, vgl. Abschnitt 3.1.7.3. Zu detaillierten Forderungen an die Störgrößen bzw. Innovationen des Modells wird auf Abschnitt 3.5.1.1 verwiesen. Für den Moment genüge die Idee, es handele sich um stochastisch unabhängige und identisch multivariat verteilte Zufallsvariablen mit Erwartungswertvektor $\mu = E(E_1) = 0_k$ und Kovarianzmatrix $S = (Cov(E_{1i}, E_{1j}))_{i,j=1\ldots,k}$. Wie im univariaten AR[$p$]-Modell wäre die Zeitreihe ohne die Störgrößen deterministisch. Daher wird die natürliche Variabilität im multivariaten Zeitreihenmodell definiert als die Summe der natürlichen Variabilitäten der Komponenten der multivariaten Innovationen:

$$natvar = tr(S) = \sum_{j=1}^{k} Var(E_{1j}) \ .$$

3.1.7.5 Natürliche Variabilität in Kompositionsquerschnittsdaten

Die Standardschätzung der Varianz durch

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

ist in Kompositionsdaten nicht länger valide, da diese nicht der euklidischen Metrik, sondern der Aitchison-Metrik unterliegen [vgl. Ait86, FHR09]. Eine sinnvolle Schätzung der natürlichen Variabilität ergibt sich somit erst, nachdem die Daten einer ilr Transformation unterzogen wurden. Die Streuung eines Kompositionszufallsvektors $X = (X_1, \ldots, X_D)'$ wurde von Aitchison durch die Variationsmatrix **T** beschrieben [vgl. Ait86, PGETD08, HK10].

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1D} \\ t_{21} & t_{22} & \ldots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \ldots & t_{DD} \end{pmatrix}, \quad t_{ij} = var\left(\ln\frac{X_i}{X_j}\right), \quad i,j = 1,\ldots,D$$

Ein Maß für die globale Streuung des Kompositionszufallsvektors X ist die totale Variation. Diese lässt sich einerseits anhand der Variationsmatrix **T** definieren mittels

$$totvar(X) := \frac{1}{2D}tr(\mathbf{T}) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}t_{ij} \ , \tag{3.4}$$

andererseits ist es möglich die totale Variation unter Verwendung der ilr Transformation anzugeben [vgl. HK10, Proposition 1]. Um die totale Variation anhand der ilr Transformation zu beschreiben, sei zunächst $X^* = ilr(X)$ der zu X gehörende ilr transformierte Zufallsvektor. Da $X = (X_1, \ldots, X_D)'$ gilt, besitzt X^* eine

Darstellung $X^* = (X_1^*, \ldots, X_{D-1}^*)'$. Im Fall $D = 2$ ist X^* univariat verteilt und die totale Variation sei definiert durch $totvar(X) := Var(X^*)$. Für $D > 2$ folgt der Zufallsvektor X^* einer multivariaten Verteilung. Diese besitze die Kovarianzmatrix $\Sigma^* = (\sigma_{ij}^*)_{i,j=1,\ldots,D-1} = (Cov(X_i^*, X_j^*))_{i,j=1,\ldots,D-1}$. Daraus ergibt sich die totale Variation für $D > 2$ als Spur von Σ^* [vgl. HK10, Proposition 1].

$$totvar(X) := tr(\Sigma^*) = \sum_{i=1}^{D-1} \sigma_{ii}^* \tag{3.5}$$

Die Äquivalenz der Darstellungen 3.4 und 3.5 zeigt Hron [vgl. HK10, Definition 2 und Proposition 1]. Die natürliche Variabilität in Kompositionsquerschnittsdaten sei fortan definiert als die totale Variation.

Eine Schätzung der totalen Variation zu gegebenen Kompositionsquerschnittsdaten $x_1, \ldots, x_n$ mit $x_i = (x_{i1}, \ldots, x_{iD})'$, $i = 1, \ldots, n$ ergibt sich somit wie folgt. Die gegebenen Daten werden identifiziert mit Kompositionszufallsvektoren $X_1, \ldots, X_n$. Die ilr transformierten Daten $x_1^*, \ldots, x_n^*$ werden folglich identifiziert mit Zufallsvariablen $X_1^*, \ldots, X_n^*$. Diese seien multivariat verteilt mit Kovarianzmatrix $\Sigma^* = (\sigma_{ij}^*)_{i,j=1,\ldots,D-1}$. Es muss folglich eine Schätzung von Σ^* bestimmt werden. Diese sei bezeichnet mit $\hat{\Sigma}^*$.

$$\hat{\Sigma}^* = ((\hat{\sigma}_{ij}^*)_{i,j=1,\ldots,D-1}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^* - \bar{x}^*)(x_i^* - \bar{x}^*)', \quad \bar{x}^* = \frac{1}{n} \sum_{i=1}^{n} x_i^*$$

Im letzten Schritt bilde man die Summe der Hauptdiagonalenelemente

$$\widehat{totvar} = \sum_{i=1}^{D-1} \hat{\sigma}_{ii}^*,$$

um eine Schätzung der totalen Variation zu erhalten.

3.1.7.6 Natürliche Variabilität in Zeitreihen von Kompositionsdaten

Gegeben seien Zeitreihenkompositionsdaten $x_1, \ldots, x_T$ mit $x_t = (x_{t1}, \ldots, x_{t,k+1})'$. Diese werden aufgefasst als Realisationen von Zufallsvariablen $X_1, \ldots, X_T$ mit $X_t = (X_{t1}, \ldots, X_{t,k+1})'$ für $t = 1, \ldots, T$. Sei $X_t^* = ilr(X_t),\ t = 1, \ldots, T$ mit $X_t^* = (X_{t1}^*, \ldots, X_{tk}^*)$. Die Zeitreihe $(X_t^*)_{t=1,\ldots,T}$ folge für $k = 1$ einem AR[p]-Zeitreihenmodell, bzw. im Fall $k > 1$ einem VAR[p]-Modell.

$$
\begin{aligned}
X_t^* &= c_\Pi - \Pi_1 X_{t-1}^* - \ldots - \Pi_p X_{t-p}^* + E_t \\
E_t &= (E_{t1}, \ldots, E_{t,k})',\ t = 1, \ldots, T \\
\Pi_j &\in \mathbb{R}^{k \times k},\ \ j = 1, \ldots, p \\
c_\Pi &\in \mathbb{R}^k \ .
\end{aligned}
$$

Das Modell besagt, dass sich der Zustand der Zeitreihe $(X_t^*)_{t=1,\ldots,T}$ zum Zeitpunkt t als Summe aus einer Linearkombination der Zustände der p vorangegangenen Zeitpunkte $X_{t-1}^*, \ldots, X_{t-p}^*$ und einer Innovation E_t ergibt. Die theoretischen Hintergründe zum VAR[p]-Zeitreihenmodell und dessen Interpretation werden in Abschnitt 3.5 erläutert.

Ohne die Innovationen E_t wäre die Zeitreihe $(X_t^*)_{t=1,\ldots,T}$ deterministisch. Daher wird die natürliche Variabilität in Zeitreihenkompositionsdaten analog zur natürlichen Variabilität in uni- und multivariaten Zeitreihen definiert, vgl. die Abschnitte 3.1.7.3 und 3.1.7.4.

$$
natvar = \sum_{j=1}^{k} Var(E_{1j})
$$

3.1.8 Der Exakte Test von Fisher

In dieser Arbeit werden zwei Methoden der statistischen Ausreißeranalytik für multivariate Querschnittsdaten verwendet, diese seien hier kurz als Verfahren A und B bezeichnet. Die Ergebnisse sollen sowohl qualitativ als auch quantitativ verglichen werden. Interessant ist die Frage, ob eines der beiden Verfahren möglicherweise konservativer arbeitet in dem Sinne, dass weniger Daten als statistische Ausreißer erkannt werden. Der exakte Test von Fisher wid verwendet, um die Kontaminationsanteile und Kontaminationswahrscheinlichkeiten (vgl. Abschnitt 3.1.9) der Verfahren A und B zu vergleichen. Ob Datenpunkte als Ausreißer identifiziert werden oder nicht, wird hier mit Hilfe binomialverteilter Zufallsvariablen modelliert. Seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ stochastisch unabhängige binomialverteilte Zufallsvariablen mit $X_i, Y_j \in \{0, 1\}$. Die Interpretation gestaltet sich wie folgt:

- $\{X_i = 1\} \sim$ Datenpunkt i wurde mit Verfahren A als statistischer Ausreißer erkannt
- $\{Y_i = 1\} \sim$ Datenpunkt i wurde mit Verfahren B als statistischer Ausreißer erkannt.

Die Zufallsvariablen $X_1, \ldots, X_n$ seien darüber hinaus identisch verteilt mit Parameter p_x, analog dazu $Y_1, \ldots, Y_m$ mit p_y. Es gilt die Frage zu beantworten, ob beide Verfahren statistisch gesehen den gleichen Anteil an Ausreißern in den zugehörigen Grundgesamtheiten detektieren. Dies lässt sich als statistisches Testproblem formulieren:

$$H_0 : \; p_x = p_y \;\text{ gegen }\; H_1 : \; p_x \neq p_y \;\; .$$

Bildet man die Zufallsvariablen $X := \sum_{i=1}^{n} X_i$ und $Y := \sum_{j=1}^{m} Y_j$, so ist X binomial-(n, p_x)-verteilt und Y ist binomial-(m, p_y)-verteilt [vgl. Kre05]. Unter H_0 gilt mit $t := x + y$:

$$P(X = x|X + Y = t) = \frac{\binom{n}{x}\binom{m}{t-x}}{\binom{n+m}{t}} \ ,$$

d.h. es liegt eine hypergeometrische Verteilung mit Parametern $n + m, n, t$ vor [vgl. Rüg02, HEK05, S.88 ff,S.416]. Ein α-Signifikanztest ist gegeben durch

$$\varphi(x, y) = \begin{cases} 1 & \text{falls } p_F \leq \alpha \\ 0 & \text{falls } p_F > \alpha \ , \end{cases}$$

mit $p_F = 2\min\{1 - P(X \leq x - 1|X + Y = t),\ P(X \leq x|X + Y = t)\}$. Der Exakte Test von Fisher kann in **R** mit Hilfe der Funktion `fisher.test` aufgerufen werden.

3.1.9 Kontaminationswahrscheinlichkeit und Kontaminationsanteil

Um bewerten zu können, wie groß der Grad der Kontamination eines Querschnittsdatensatzes durch statistische Ausreißer ist, benötigt man den Begriff des Kontaminationsanteiles. Darüber hinaus wird in diesem Abschnitt die Kontaminationswahrscheinlichkeit definiert. Sie dient der Bewertung, mit welcher Wahrscheinlichkeit ein Querschnittsdatensatz durch mindestens einen Ausreißer kontaminiert ist. Beide Begriffe finden in der Diskussion von Abschnitt 5 Verwendung.

Die Schätzung des Kontaminationsanteiles erfolgt unter Verwendung eines Bi-

nomialmodells. Seien $y_1, \ldots, y_n$ Querschnittsdaten. Diese werden aufgefasst als Realisationen von stochastisch unabhängigen Zufallsvariablen $Y_1, \ldots, Y_n$. Definiere Zufallsvariablen $X_1, \ldots, X_n$ mit

$$X_i = \begin{cases} 1 & Y_i \text{ ist ein Ausreißer} \\ 0 & Y_i \text{ ist kein Ausreißer} \end{cases}, \ i = 1, \ldots, n.$$

Die entsprechenden Realisationen seien bezeichnet mit $x_1, \ldots, x_n$. Wegen der stochastischen Unabhängigkeit von $Y_1, \ldots, Y_n$ sind auch $X_1, \ldots, X_n$ stochastisch unabhängig. Die Zufallsvariablen $X_1, \ldots, X_n$ seien identisch binomial-$(1, p_{prop})$-verteilt, wobei $p_{prop} \in (0, 1)$ ein unbekannter Parameter ist ($p_{prop} \sim proportion$). Nach Krengel ist $X := \sum_{i=1}^{n} X_i$ binomial-(n, p_{prop})-verteilt [Kre05]. Die entsprechende Realisation von X ergibt sich kanonisch aus $x = \sum_{i=1}^{n} x_i$. Der Kontaminationsanteil sei definiert als der Parameter p_{prop}. Eine Schätzung von p_{prop} ist gegeben durch $\hat{p}_{prop} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ [vgl. Kre05]. Nach Krengel ist

$$\frac{1}{(n + c^2)} \left((X \pm 0.5) + 0.5c^2 \pm c\sqrt{(X \pm 0.5) - n^{-1}(X \pm 0.5)^2 + 0.25c^2} \right) \tag{3.6}$$

ein approximatives Konfidenzintervall für p_{prop}, wobei $c = z_{1-\alpha/2}$ das $(1 - \alpha/2)$-Quantil der Normal-(0,1)-Verteilung ist [vgl. Kre05, S.85]. Ein exaktes Konfidenzintervall ist gegeben durch [vgl. Kre05, S.69 ff.]

$$\left(\frac{q_1}{\frac{n-X+1}{X} + q_1}; \ \frac{q_2}{\frac{n-X}{X+1} + q_2} \right) \quad . \tag{3.7}$$

Dabei gilt $q_1 = F_{\alpha/2, m_1, n_1}$ und $q_2 = F_{1-\alpha/2, m_2, n_2}$, d.h. q_1 und q_2 sind das $\alpha/2$-Quantil bzw. das $(1 - \alpha/2)$-Quantil der F-Verteilung mit den Freiheitsgraden

$m_1 = 2X$, $n_1 = 2(n - X + 1)$, $m_2 = 2(X + 1)$ und $n_2 = 2(n - X)$.

Soll eine Schätzung der Wahrscheinlichkeit des Ereignisses „ein zufällig ausgewählter Datensatz ist durch mindestens einen statistischen Ausreißer kontaminiert“ bestimmt werden, so wird obiges Modell modifiziert. Seien $y_1, \ldots, y_n$ zufällig ausgewählte Datensätze. Diese werden aufgefasst als Realisationen von stochastisch unabhängigen Zufallsvariablen. Praktisch stelle man sich vor, dass aus allen möglichen Datensätzen genau n verschiedene ausgesucht werden. Für jeden Datensatz gilt, dass er entweder durch statistische Ausreißer kontaminiert wurde oder nicht. Es werden Zufallsvariablen $X_1, \ldots, X_n$ definiert, um dies mathematisch wiederzuspiegeln.

$$X_i = \begin{cases} 1 & Y_i \text{ ist durch statistische Ausreißer kontaminiert.} \\ 0 & Y_i \text{ ist durch keinen Ausreißer kontaminiert.} \end{cases}, \ i = 1, \ldots, n.$$

Analog zu obiger Argumentation ist die Zufallsvariable $X := \sum_{i=1}^{n} X_i$ binomial-(n, p_p)-verteilt. Dabei ist $p_p \in (0, 1)$ der unbekannte Parameter der Binomialverteilung ($p_p \sim probability$). Die Kontaminationswahrscheinlichkeit wird definiert als Parameter p_p. Eine Schätzung von p_p bestimmt sich analog zu obiger Vorgehensweise durch $\hat{p}_p = \bar{X}$. Die Konfidenzintervalle entsprechen denen aus Formel (3.6) und (3.7).

3.1.10 Graphische und deskriptive Methoden zur Bestimmung von Ausreißern

Die graphische Analyse liefert einen ersten Eindruck von der Struktur der Daten und gibt Hinweise auf mögliche Wechselwirkungen zwischen Variablen. Darüber hinaus sind potentielle statistische Ausreißer sichtbar. Verteilungseigenschaften und eventuelle funktionale Zusammenhänge können abgeleitet werden. Es ist daher nicht überraschend, dass sich bereits viele Autoren mit graphischen Analysemethoden und Darstellungsformen befassten. Dazu zählen unter anderem Hartung und Elpelt, die einen Überblick zu verschiedenen Visualisierungsmethoden für multivariate Querschnittsdaten geben [HE89]. Heiberger und Holland [HH04] entwickelten einen praktischen Leitfaden zur graphischen Datenanalyse in S-PLUS, **R** und SAS. Maindonald und Braun geben Beispiele zu Visualisierungstechniken in **R** [MB03].

3.1.10.1 Eindimensionale Datensätze

Um die rein numerische Ausreißeranalytik zu unterstützen und tiefere Einblicke in die Struktur der Daten zu bekommen, ist es stets ratsam verschiedene Visualisierungen zu generieren. Dazu zählen insbesondere der einfache Scatterplot, der Boxplot und der Zeitreihenplot. Diese Darstellungsformen geben erste visuelle Hinweise auf eventuell vorhandene statistische Ausreißer.

Zur Veranschaulichung der Vorgehensweise bei der graphischen Analyse wird ein Datensatz zu Aalen (Anguilla anguilla) aus dem Jahr 2002 verwendet. Dieser enthält 105 Messpunkte der Gestalt $x = PCB\ 118$. Wie in Abschnitt 3.1.5 be-

schrieben, werden Nulldaten ersetzt und entsprechend Abschnitt 3.1.4 der „Rest“ gebildet und die ilr Transformation auf die Daten angewendet. Diese Vorgehensweise wird detailliert anhand von Beispielen in Abschnitt 4 illustriert. Zunächst empfiehlt es sich stets, die Daten in einem Scatterplot zu visualisieren. Die Betrachtung der Punktewolke kann erste Hinweise auf Ausreißer enthalten. Die ilr transformierten Daten sind in Abbildung 3.1 dargestellt.

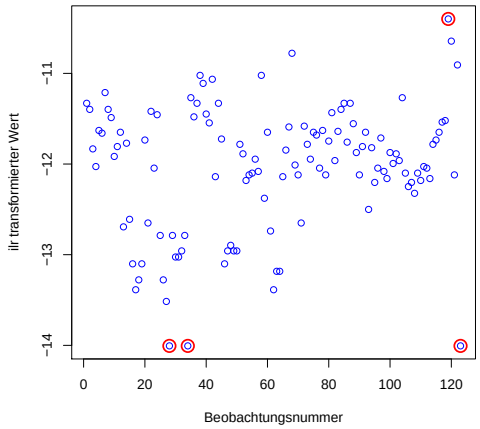

Abbildung 3.1: Scatterplot eindimensionaler ilr transformierter Daten

Es lässt sich bereits anhand dieser simplen Darstellung erkennen, dass die rot markierten Datenpunkte potentielle Kandidaten für Ausreißer sind.

Im zweiten Schritt werden die Daten im Boxplot dargestellt. Die Länge der Whisker beträgt standardmäßig das 1.5-fache des Interquartilabstandes und wird auf den letzten Wert, der noch innerhalb dieser Whisker liegt, verkürzt. Die Box zwischen den Whiskern gibt die zentralen 50% der Daten wieder.

Auch der Boxplot in Abbildung 3.2 lässt vermuten, dass es sich bei den rot markierten Datenpunkten um Ausreißer handelt.

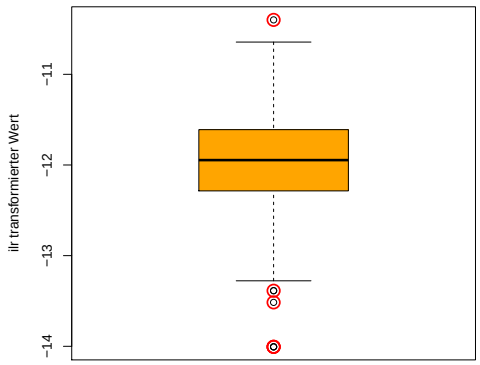

Abbildung 3.2: Boxplot eindimensionaler ilr transformierter Daten

Erst bei Verlängerung der Whisker auf das 3-fache des Interquartilabstandes werden die Werte nicht mehr als Ausreißer markiert, wie aus Abbildung 3.3 ersichtlich wird.

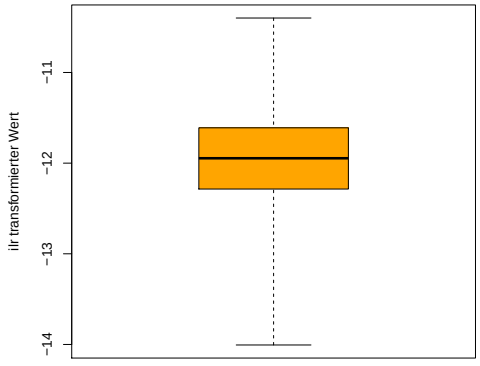

Abbildung 3.3: Boxplot eindimensionaler ilr transformierter Daten mit 3-fachem Interquartilabstand

Diese graphische Vorgehensweise lässt sich auch mathematisch beschreiben. Seien $x_{0.25}$ und $x_{0.75}$ das 25% bzw. 75% Quantil des Datensatzes $D \subset \mathbb{R}$. Sei $O \subset D$ die Menge der Ausreißer in D, dann gilt

$$O := \{x \in D | x \notin [x_{25} - k \cdot (x_{75} - x_{25});\ x_{75} + k \cdot (x_{75} - x_{25})]\}$$

mit $k = 1.5$, oder $k = 3$. Nach Sachs [vgl. Sac09, S.322] werden auffällige Werte, die mit $k = 1.5$ erkannt wurden, als Ausreißer bezeichnet. Für $k = 3$ erkannte Ausreißer gelten als extreme Ausreißer.

3.1.10.2 Mehrdimensionale Datensätze

Seien D_1^*, D_2^* multivariate Datensätze mit Dimension 2 und 3, dann lassen sich diese visualisieren mit Hilfe von Scatterplots, welche die Daten als Punktwolken im zwei- oder dreidimensionalen Raum darstellen. Zur Illustration werden zwei Datensätze zu Aalen aus dem Jahr 2002 verwendet. Datensatz D_1^* enthält Daten der Gestalt $x = (PCB\ 101,\ PCB\ 138)'$ und Datensatz D_2^* enthält Daten der Gestalt $x = (PCB\ 153,\ PCB\ 180,\ PCB\ 28)'$. Die Daten werden entsprechend der Abschnitte 3.1.5 und 3.1.4 aufbereitet und die ilr Transformation auf sie angewendet. Dies liefert die Datensätze D_1 und D_2. Je Datensatz liegen $n = 105$ Messwerttupel vor. Die Darstellung des Datensatzes D_1 im Scatterplot zeigt, dass der rot gekennzeichnete Punkt $(-2.29; -15.36)$ ein potentieller statistischer Ausreißer ist. Es liegt nun nahe, zunächst die einzelnen Merkmale auf Ausreißer zu untersuchen. Dies liefert einen weiteren potentiellen Ausreißer mit den Koordinaten $(-1.70; -14.48)$, welcher im Scatterplot ebenfalls rot gekennzeichnet wird. Ist ein Punkt in mindestens einer Dimension ein Ausreißer, so ist er möglicherwei-

se insgesamt als Ausreißer zu bewerten. Eine Kombination aus Scatterplot und Boxplot vereinfacht die graphische Analyse, siehe Abbildung 3.4.

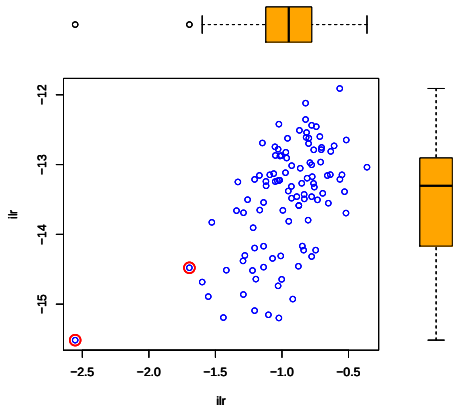

Abbildung 3.4: Boxplots und Scatterplot zweidimensionaler Messdaten

Beide potentiellen Ausreißer sind nur in einer Dimension auffällig. Wie man sieht, deklariert der obere Boxplot den Punkt $(-1.70; -14.48)$ in Abbildung 3.4 als Ausreißer. Im Scatterplot jedoch ist dieser Datenpunkt nicht ausreißerverdächtig. Dies zeigt, dass Datenpunkte, die in einer Dimension als Ausreißer betrachtet werden, in höheren Dimensionen möglicherweise keine Ausreißer mehr sind.

Datensatz D_2 wird in Abbildung 3.5 als Scatterplot dargestellt. Die rot markierten Datenpunkte sind potentielle Ausreißer. Sie besitzen die folgenden Koordinaten: $(0.83; 2.66; -15.12)$ und $(0.42; 0.73; -13.87)$.

An dieser Stelle ist es wieder interessant zu untersuchen, in welchen Merkmalen diese Punkte auffällig werden. Hilfreich ist die Darstellung eines jeden Merkmals im Boxplot.

Der Punkt $(0.83; 2.66; -15.12)$ ist offensichtlich im ersten Merkmal auffäl-

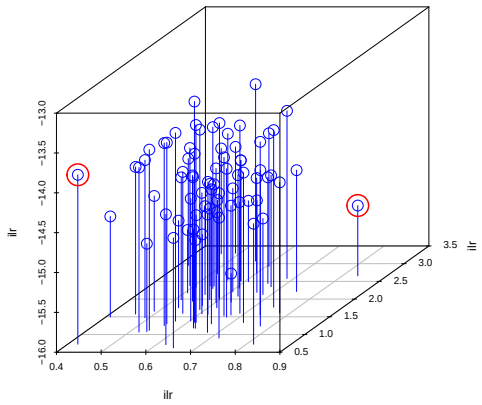

Abbildung 3.5: Scatterplot dreidimensionaler Messdaten

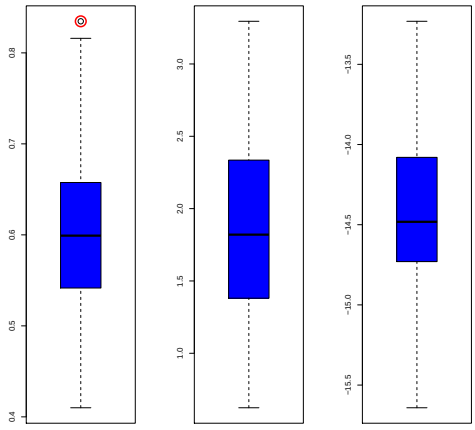

Abbildung 3.6: Boxplot für jedes der drei Merkmale

lig und damit ein potentieller Ausreißer, wie Abbildung 3.6 zeigt. Der Punkt $(0.42; 0.73; -13.87)$ ist in keinem Merkmal im Boxplot auffällig. Dies zeigt, dass potentielle Ausreißer nicht zwingend in mindestens einer Dimension auffällig sein müssen.

Für höhere Dimensionen ist die einfache gemeinsame Darstellung aller Merkmale nicht mehr möglich. In diesen Fällen empfiehlt es sich auf eine gleichzeitige Visualisierung aller Merkmale zu verzichten und sich stattdessen auf die Boxplots der einzelnen Dimensionen zu beschränken.

Warum verwendet man nicht ausschließlich Boxplots zur Identifikation auffälliger Datenpunkte? In der Praxis kommt es nicht selten vor, dass eine Beobachtung in keinem Merkmal einen potentiellen Ausreißer aufweist bzw. nicht alle potentiellen Ausreißer in den Boxplots angezeigt werden, obgleich im Scatterplot Datenpunkte auffällig sind, wie die Abbildungen 3.5 und 3.6 zeigen. Ein weiteres Beispiel dieser Problematik gibt Datensatz D_3^*. Dieser enthält $n = 76$ Datenpunkte der Gestalt $x = (PCB\ 28,\ PCB\ 52)'$ zu Aalen aus dem Jahr 2002. Er wurde analog zu den Datensätzen D_1 und D_2 entsprechend der Abschnitte 3.1.5 und 3.1.4 aufbereitet. Der Datensatz der ilr transformierten Daten sei bezeichnet mit D_3. Die Boxplots der einzelnen Merkmale in Abbildung 3.7 zeigen lediglich einen auffälligen Wert bzw. potentiellen statistischen Ausreißer. Der Scatterplot in Abbildung 3.8 hingegen lässt fünf auffällige Datenpunkte erkennen, welche rot markiert sind. Die Analyse jedes Merkmals für sich gibt folglich noch keinen kompletten Aufschluss darüber, ob ein Datensatz potentielle statistische Ausreißer enthält. Es wäre darüber hinaus denkbar, zusätzlich den untersten Datenpunkt in Abbildung 3.8 nahe der x-Achse als potentiellen Ausreißer zu markieren. Dies

zeigt, dass die graphische Vorgehensweise bei der statistischen Ausreißeranalytik einerseits hilfreich, andererseits jedoch sehr subjektiv ist.

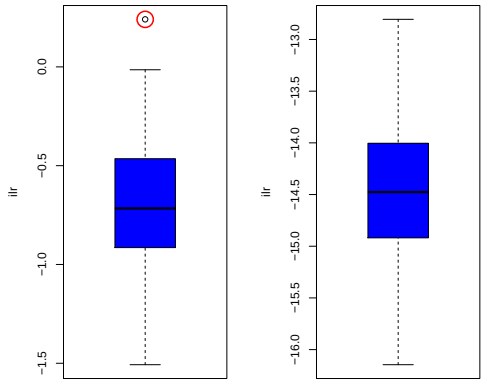

Abbildung 3.7: Boxplots zu Aalen aus dem Jahr 2002

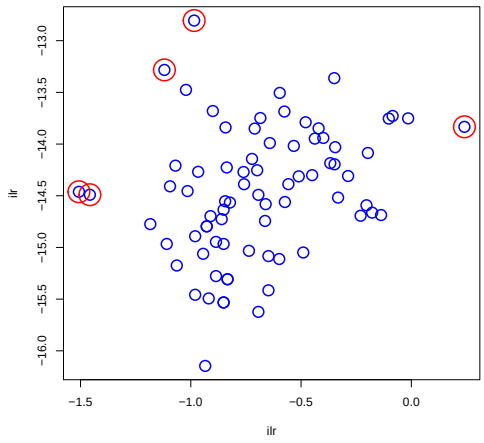

Abbildung 3.8: Scatterplot zu Aalen aus dem Jahr 2002

In höheren Dimensionen ist die Darstellung als einfacher Scatterplot nicht mehr möglich. Die Identifikation von Ausreißerkandidaten wird dadurch erheblich erschwert. Es ist allerdings möglich jede Merkmalsdimension in der Ebene durch

unterschiedliche Symbole wiederzugeben. Vorher sollten jedoch die Merkmale so skaliert werden, dass sie sich alle in einem ähnlichen Bereich aufhalten. Zur Illustration wird ein Datensatz, bestehend aus Daten $n = 76$ der Gestalt $x = (PCB\ 138,\ PCB\ 153,\ PCB\ 180)'$ zu Aalen aus dem Jahr 2003. Die Daten wurden entsprechend der Abschnitte 3.1.5 und 3.1.4 aufbereitet.

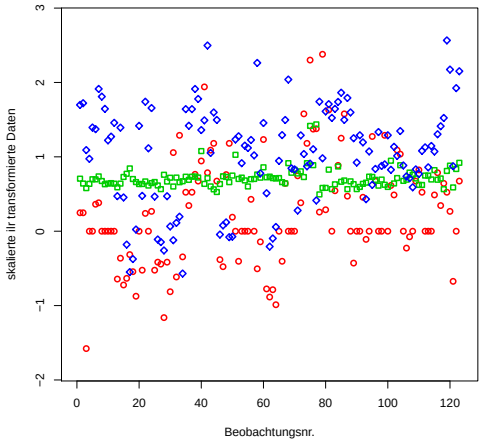

Abbildung 3.9: Scatterplot der skalierten Merkmale des Datensatzes D_2

Als Referenz zur Skalierung diente das Merkmal 2. Die Werte für Merkmal 1 und 2 wurden angepasst. Der dargestellte Datensatz hat die folgende Struktur.

$$\begin{pmatrix} 10 \cdot Merkmal\ 1 \\ Merkmal\ 2 \\ 15 + Merkmal\ 3 \end{pmatrix}$$

Einen Überblick über die wichtigsten Darstellungsformen multivariater Daten geben Hartung und Elpelt [vgl. HE89, S.610 ff.]. Die dort vorgestellten Diagramme sind mit einiger Übung gut überschaubar, solange die Datensätze keine zu großen Dimensionen annehmen, oder zu viele Beobachtungen vorliegen. Die hier

vorgestellten graphischen Methoden sind rein deskriptiver Natur und die Identifikation ausreißerverdächtiger Datenpunkte somit nicht modellbasiert.

Ein weiteres Werkzeug zur Unterstützung der graphischen Analyse sind Scatterplotmatrizen. Scatterplotmatrizen sind sinnvoll für Dimension $p \geq 3$. Dabei werden alle Kombinationsmöglichkeiten von zwei Merkmalen in einem separaten Scatterplot dargestellt. Einen Eindruck liefert Abbildung 4.9 auf Seite 203.

3.2 Bestimmung von Ausreißern in univariaten Querschnittsdaten

Diverse Autoren Befassten sich mit dem Problem statistischer Ausreißer und einer adäquaten Definition des Begriffes der sog. „outlyingness“, darunter z.B. Stahel [Sta81] und Donoho [Don82]. Darüber hinaus wurden die Begriffe der „α-outlier“ und der „α-outlier region“ von Davies und Gather [vgl. DG89] und Gather et al. [vgl. GKP04] geprägt. Stahel und Donoho definierten die outlyingness als Maßzahl, die angibt wie weit ein Messwert vom „Zentrum“ oder „Schwerpunkt“ der Datengesamtheit, unter Berücksichtigung der Streuung, entfernt liegt. Für univariate Daten $x_1, \ldots, x_n$ wurde die outlyingness eines Wertes $x_i \in \{x_1, \ldots, x_n\}$ definiert durch

$$out(x_i) := \frac{x_i - median\{x_1, \ldots, x_n\}}{mad\{x_1, \ldots, x_n\}} \quad ,$$

dazu siehe auch Stahel [vgl. Sta81, S.111] und Donoho [vgl. Don82, S.1804 f.]. Der Begriff der α-outlier definiert nach Gather et al. Daten als Ausreißer, wenn sie nicht zum Modell der „guten Daten“ passen. Demnach wird ein Datum als Ausreißer betrachtet, wenn es unter dem angenommen statistischen Datenmodell

(z.B. Normal-(μ, σ)-Verteilung) in einem gewissen Sinne „unwahrscheinlich“ ist [vgl. GKP04, BG00, S.5,S.120].

3.2.1 Der robuste Ausreißertest nach Hampel

Wenngleich in dieser Arbeit die Majorität der Daten multivariat ist, so werden doch vereinzelt univariate Datensätze auf Ausreißer untersucht. Daher soll hier kurz der bekannteste robuste Ausreißertest für univariate Daten, der sogenannte Hampel-Test (auch häufig als „Hampel identifier“ bezeichnet), vorgestellt werden. Weitere Ausreißertests für univariate Daten sind z.B. der Q-Test nach Dixon und die Ausreißertests nach Grubbs oder Nalimov. Einen Überblick über häufig verwendete Ausreißertests geben z.B. Barnett und Lewis [BL94] und Hawkins [Haw80].

Seien univariate Daten $x_1, \ldots, x_k$ gegeben. Diese werden identifiziert mit stochastisch unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_k$. Der robuste Ausreißertest nach Hampel setzt voraus, dass P^{X_1} eine unimodale symmetrische Verteilung ist. Das Verfahren verwendet robuste Schätzungen für den Erwartungswert und die Standardabweichung. Dies sind der Median $\tilde{x}$ und die MAD („Median Absolute Deviation“).

Der Median und die MAD können auf mehrere Arten geschätzt werden. Es ist z.B. möglich eine Schätzung des Medians anhand der Ordnungsstatistik der Daten $x_{(1)}, \ldots, x_{(k)}$ zu generieren. Bei ungerader Beobachtungszahl k ergibt sich die Schätzung des Medians durch $\tilde{x} := x_{((k+1)/2)}$. Für gerades k ist der Median durch die Ordnungsstatistik nicht eindeutig bestimmt. Alle $x \in [x_{(k/2)}; x_{((k+1)/2)})$ sind potentielle Kandidaten. In diesem Fall wird der Mittelwert der Intervallgrenzen

gebildet: $\tilde{x} := \frac{1}{2}\left(x_{(k/2)} + \tilde{x}_{(k+1)/2}\right)$. Die MAD berechnet sich dann als Median der absoluten Abweichungen der Messwerte vom Median.

$$MAD := Median\{|x_1 - \tilde{x}|, \ldots, |x_k - \tilde{x}|\}$$

Im Fall einer normalverteilten Grundgesamtheit erhält man durch Multiplikation der MAD mit 1.4826 eine robuste, konsistente Schätzung der Standardabweichung σ [vgl. Sac09, S.68]. Wird diese Schätzmethode verwendet, so darf der robuste Ausreißertest nach Hampel nicht eingesetzt werden, wenn mehr als 50% der Daten den gleichen Wert annehmen, da in diesem Fall die aus der Stichprobe geschätzte MAD Null wäre.

Eine weitere Möglichkeit besteht darin, den Median als 50%-Quantil der Verteilung P^{X_1} zu wählen. Zwar ist diese Verteilung unbekannt, jedoch kann die Dichte f der Verteilung P^{X_1} mit Hilfe einer Kerndichteschätzung geschätzt werden (vgl. Abschnitt 3.1.2). Aus der geschätzten Dichte $\hat{f}$ lässt sich mittels numerischer Integration der Median schätzen:

$$\tilde{x} := \inf\left\{ t \;\middle|\; \int_{-\infty}^{t} \hat{f}(x)dx \geq 0.5 \right\} .$$

Die MAD wird analog als Median aus den Daten für $y_1, \ldots, y_n$ geschätzt, wobei $y_i = |x_i - \tilde{x}|$ gilt.

Der Hampel-Test erkennt einen Datenpunkt x_a als Ausreißer, falls

$$\frac{|\tilde{x} - x_a|}{MAD} > C_{k,\alpha}^{Ham}$$

gilt, wobei $C_{k,\alpha}^{Ham}$ ein kritischer Wert nach Hampel ist, siehe Dietrich und Schulze

[DS09, S.208]. In Tabelle 3.2 sind die kritischen Werte für verschiedene Datenanzahlen k und Signifikanzniveaus α gegeben.

α \ k	10	12	14	15	16	18	20	30	40	50	100	200
0.01	6.7	6.2	5.9	5.9	5.6	5.4	5.3	4.8	4.6	4.5	4.3	4.3
0.05	4.6	4.4	4.3	4.3	4.2	4.1	4.1	3.9	3.8	3.8	3.8	3.8

Tabelle 3.2: Kritische Werte des zweiseitigen Hampel Tests [vgl. DS09, S.709]

Nach Lin und Johnson [vgl. LJ03, S.47] ist der Hampel-Test für univariate Daten stets zu bevorzugen, da er im Allgemeinen eine geringere Anfälligkeit für Masking Effekte aufweist als parametrische Verfahren, wie z.B. der Q-Test nach Dixon und die Tests nach Grubbs oder Nalimov.

In der statistischen Ausreißeranalytik werden zwei Effekte beschrieben: das Masking und das Swamping. Rambold beschreibt den Masking Effekt derart, dass eine Gruppe von Ausreißern eventuell die Gegenwart anderer Ausreißer verschleiern (maskieren) kann [vgl. Ram99, S.8 f]. Der Begriff Swamping beschreibt das Phänomen, dass Nichtausreißer, die von echten Ausreißern umgeben sind, fälschlicher Weise ebenfalls als Ausreißer erkannt werden.

3.3 Bestimmung von Ausreißern in multivariaten Querschnittsdaten

Die statistische Ausreißeranalytik ist für multivariate Daten komplizierter als für univariate, was in der Tatsache begründet liegt, dass Ausreißer nicht zwingend in einer Dimension als extreme Werte auftreten müssen, wie aus den Abbildungen

3.7 und 3.8 ersichtlich wird. Graphische Methoden sind für multivariate Daten prinzipiell anwendbar, jedoch ist eine Visualisierung lediglich bis zur Dimension drei möglich. Möglichkeiten zur Darstellung höherdimensionaler Daten bieten z.B. Scatterplotmatrizen und Hartung und Elpelt [vgl. HE89, S.610 ff.]. Diese Verfahren sind allerdings rein deskriptiver Natur. Wünschenswert ist ein statistisches Entscheidungskriterium, z.B. in Form eines statistischen Tests, um objektiv entscheiden zu können, ob ein Datenpunkt als Ausreißer zu betrachten ist oder nicht.

Im Wesentlichen stehen zu diesem Zweck zwei Vorgehensweisen zur Verfügung. Zum einen ist es möglich die Daten rein multivariat zu analysieren, zum anderen wurden Projektionsverfahren entwickelt, durch die aus einem multivariaten Datensatz mehrere univariate Datensätze erstellt und auf Ausreißer untersucht werden können. Der Einsatz von Projektionsverfahren zur Analyse multivariater Daten geht auf Roy [vgl. Roy53], Kruskal [vgl. Kru69] und Switzer [vgl. Swi70] zurück. Gnanadesikan und Kettenring nutzen Projektionsmethoden zum Auffinden statistischer Ausreißer [GK72]. Projektionsverfahren werden auch als PP-Verfahren (Projection Pursuit) bezeichnet. Sie wurden vielfach weiterentwickelt zum Zweck der robusten Schätzung von Lageparametern und Kovarianzmatrizen multivariater Verteilungen und zur Identifizierung multivariater statistischer Ausreißer [PP01b, PFF00, MZ01].

3.4 Die robuste Mahalanobis-Distanz

Mahalanobis entwickelte ein Maß um die Distanz zwischen zwei mehrdimensionalen Punkten zu bestimmen [Mah36]. Seien $x, y \in \mathbb{R}^p$ Realisationen von sto-

chastisch unabhängigen und identisch multivariat verteilten Zufallsvariablen X und Y. Die Kovarianzmatrix der Verteilung von X sei S. Der Abstand von x und y ist dann nach Mahalanobis gegeben durch

$$d(x,y) = \sqrt{(x-y)'S^{-1}(x-y)} \ .$$

Sind Daten $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ gegeben, so werden diese aufgefasst als Realisationen von stochastisch unabhängigen und identisch multivariat verteilten Zufallsvariablen $X_1, \ldots, X_n$. Die sogenannte Mahalanobis-Distanz eines Datenpunktes $x_i \in \{x_1, \ldots, x_n\}$ zum „Schwerpunkt“ oder „Zentrum“ der Daten ist definiert durch

$$\sqrt{(x_i - \hat{\mu})'\hat{S}^{-1}(x_i - \hat{\mu})},$$

wobei $\hat{\mu}$ und $\hat{S}$ die multivariaten Schätzungen des Lageparameters und der Kovarianzstruktur anhand der Daten $\{x_1, \ldots, x_n\}$ sind. Im Fall multivariat normalverteilter Daten ist die quadrierte Mahalanobis-Distanz approximativ χ^2_p-verteilt [vgl. FH08, S.236]. Stochastisch betrachtet, nimmt die quadrierte Mahalanobis-Distanz unter der Nullhypothese „die Daten sind nicht durch statistische Ausreißer kontaminiert“ mit Wahrscheinlichkeit $1-\alpha$ Werte im Intervall $[0, \chi^2_{1-\alpha,p}]$ an, wobei $\chi^2_{1-\alpha,p}$ das $(1-\alpha)$-Quantil der χ^2_p-Verteilung ist. Ist die quadrierte Mahalanobis-Distanz eines Datenpunktes x_i größer als $\chi^2_{1-\alpha,p}$, so wird dieser als potentieller Ausreißer betrachtet [vgl. RZ90, S.633]. In der Praxis wird häufig $\alpha = 0.025$ oder $\alpha = 0.01$ gewählt. Die Mahalanobis-Distanz ist offensichtlich nicht robust gegen Ausreißer, wenn die Standardformeln zur Schätzung von Lageparameter (geschätzt durch das arithmetische Mittel) und Kovarianzmatrix (geschätzt durch die Stichprobenkovarianz) eingesetzt werden. Einen Ausweg aus dem Problem

der geringen Robustheit von Mittelwert und Stichprobenkovarianz gegen Ausreißer bieten die MCD und S Schätzer [vgl. MMY06]. Werden diese Schätzer eingesetzt, so ist es weiterhin möglich einen Datenpunkt als Ausreißer zu klassifizieren, wenn dessen robuste Mahalanobis-Distanz das $(1-\alpha)$-Quantil der χ^2_p-Verteilung überschreitet. Allerdings ist die Approximation der Verteilung der robusten Mahalanobis-Distanz durch die χ^2_p-Verteilung nicht mehr optimal. In diesem Fall sind andere kritische Werte sinnvoller [FGR05, HR05].

Filzmoser und Hron beschreiben die Verwendung der Mahalanobis-Distanz zur Ausreißererkennung in Kompositionsdaten und ihre Konsistenz unter der ilr Transformation [FH08]. Eine automatisierte Berechnung der robusten Mahalanobis-Distanz ist in **R** durch die Funktion `Moutlier` im Paket `chemometrics` gegeben [FV10]. Um die Funktion `Moutlier` nutzen zu können, müssen mindestens $2p$ Datenpunkte vorliegen. Dies fordert die Definition des MCD Schätzers [vgl. BF04, S.2f]. Weiterhin ist es notwendig, dass die zur Schätzung der Kovarianzmatrix verwendeten Datenpunkte nicht in einer Hyperebene des $\mathbb{R}^p$ liegen. Um dies einzusehen, wird der MCD Schätzer kurz beschrieben [vgl. BF04, S.2f].

Seien $n \geq 2p$ und $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$, sowie $h \in \mathbb{N}$ mit $n/2 < h < n$. Dabei werden $x_1, \ldots, x_n$ aufgefasst als Realisationen von stochastisch unabhängigen und identisch verteilten multivariaten Zufallsvariablen $X_1, \ldots, X_n$. Die aus den Daten geschätzte Kovarianzmatrix S hat die Gestalt $\hat{S} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$ mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Die Aufgabe bei der Bestimmung des MCD Schätzer besteht darin, eine Teilmenge $\mathcal{X}' = \{x_{i_1}, \ldots, x_{i_h}\} \subset \mathcal{X}$ derart zu finden, dass die Determinante der daraus geschätzten Kovarianzmatrix $\hat{S}' = \frac{1}{h}\sum_{j=1}^{h}(x_{i_j} - \bar{x}')(x_{i_j} - \bar{x}')'$ mit $\bar{x}' = \frac{1}{h}\sum_{j=1}^{h} x_{i_j}$ minimal wird. Angenommen für die entsprechenden Zufallsvariablen $X_{i_1}, \ldots, X_{i_h}$ gelte $X_{i_j} \in \mathcal{H} = \{x \mid a^{*'}x = b^*,\ a^* \in \mathbb{R}^p,\ b^* \in \mathbb{R}\}$,

d.h. die X_{i_j} besitzen nur Realisationen in einer Hyperebene des $\mathbb{R}^p$. Dann gilt für die Determinante der Kovarianzmatrix S der Verteilung von X_{i_j} die Gleichung $det(S) = 0$. Um dies einzusehen betrachte man für eine Zufallsvariable $Y = (Y_1, \ldots, Y_p)' \in \mathcal{H}$ explizit die Kovarianzmatrix:

$$S = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} ,$$

mit $\sigma_{ij} = Cov(Y_i, Y_j)$ und $\sigma_{ii} = Cov(Y_i, Y_i) = Var(Y_i) = \sigma_i^2$. Da $Y \in \mathcal{H}$ gilt, ergibt sich folgender Zusammenhang:

$$\begin{aligned} & a_1^* Y_1 + \ldots + a_p^* Y_p = b^* \\ \Leftrightarrow\ & Y_p = a_1 Y_1 + \ldots + a_{p-1} Y_{p-1} + b \\ \text{mit } & a_i = -\frac{a_i^*}{a_p},\ b = \frac{b^*}{a_p},\ a_p \neq 0,\ i = 1, \ldots, p-1 \ . \end{aligned}$$

Die Kovarianzen besitzen die folgenden Darstellungen:

$$\begin{aligned} Cov(Y_i, Y_p) &= E(Y_i Y_j) - E(Y_i)E(Y_j) \\ &= E(Y_i(a_1 Y_1 + \ldots + a_{p-1} Y_{p-1} + b)) \\ &\quad - E(Y_i)E(a_1 Y_1 + \ldots + a_{p-1} Y_{p-1} + b) \\ &= \sum_{j=1}^{p-1} a_j E(Y_i Y_j) + bE(Y_i) - \sum_{j=1}^{p-1} a_j E(Y_i)E(Y_j) - bE(Y_i) \\ &= \sum_{j \neq i}^{p-1} (a_j E(Y_i Y_j) - a_j E(Y_i)E(Y_j)) + a_i E(Y_i^2) - a_i E(Y_i)^2 \end{aligned}$$

$$= \sum_{j \neq i}^{p-1} a_j \sigma_{ij} + a_i \sigma_i^2 \ .$$

Damit ergibt sich die Kovarianzmatrix S zu

$$S = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & a_1\sigma_1^2 + \sum_{j\neq 1}^{p-1} a_j\sigma_{ij} \\ \sigma_{21} & \sigma_2^2 & \cdots & a_2\sigma_2^2 + \sum_{j\neq 2}^{p-1} a_j\sigma_{ij} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & a_p\sigma_p^2 + \sum_{j\neq p}^{p-1} a_j\sigma_{ij} \end{pmatrix} \ ,$$

d.h. die p-te Spalte von S ist eine Linearkombination der Spalten $1, \ldots, p-1$:

$$\text{Spalte } p = a_1 \cdot \text{Spalte } 1 + \ldots + a_{p-1} \cdot \text{Spalte } (p-1) \ .$$

Für die Determinante der Kovarianzmatrix gilt folglich $det(S) = 0$, d.h. S^{-1} existiert nicht. Somit kann die Mahalanobis-Distanz nicht bestimmt werden [vgl. Bos08, S.145].

Die hier verwendeten Methoden zur Identifikation mehrdimensionaler Ausreißer beruhen sowohl auf dem Projektionsverfahren von Pena und Prieto [PP01b], als auch auf dem Verfahren einer individuellen robusten Abstandsmessung eines jeden Datenpunktes von der Datenmasse (Mahalanobis-Distanz mit robusten Schätzern für den Lageparameter und die Kovarianzmatrix, [Cam80, VF09]). Projektionsverfahren haben gegenüber rein multivariaten Verfahren den Vorteil, dass keine mehrdimensionalen Modelle geschätzt oder multivariate Verteilungsannahmen getroffen werden müssen. Die projizierten Daten sind eindimensional, was

die Analyse, sowohl rechnerisch als auch graphisch, vereinfacht. Die näheren mathematischen Hintergründe wurden von Pena und Prieto erläutert [vgl. PP01b]. Sie sind hier, des Verständnisses wegen, zum Teil wiedergegeben. Darüber hinaus werden Algorithmen zur effizienten Umsetzung und Implementierung der Verfahren angegeben.

3.4.1 Das PP-Verfahren nach Pena und Prieto

Pena und Prieto schlagen vor, multivariate Ausreißer mit Projektionsverfahren zu identifizieren [PP01b]. Die resultierende Teststatistik ist asymptotisch χ^2-verteilt. Der Arbeitsaufwand besteht im wesentlichen darin, die Lösung eines multivariaten Maximierungs- bzw. Minimierungsproblems zu finden. Pena und Prieto schlagen vor, die Projektionsrichtungen in Abhängigkeit vom Kurtosiskoeffizienten der multivariaten Stichprobe zu wählen [vgl. PP01b, S.288f.]. Sie argumentieren zunächst für univariate Datensätze, dass eine geringe Kontamination der Daten durch Ausreißer den Kurtosiskoeffizienten vergrößern würde. Dies sei unabhängig davon, ob die Kontamination symmetrisch oder asymmetrisch wäre, d.h. ob eine Kontamination vorliegt, die sich eher „im oberen bzw. unteren Bereich“ der Daten manifestiert, oder ob eine Kontamination sowohl durch hohe, als auch durch niedrige Messwerte gleichermaßen vorliegt [vgl. PP01b, S.288]. Nach diesem Ansatz ist sinnvoll, die Projektionsrichtung multivariater Daten so zu wählen, dass die resultierenden univariaten Daten einen möglichst großen Kurtosiskoeffizienten aufweisen, wenn man statistische Ausreißer detektieren möchte. Andererseits ist es nach Pena und Prieto möglich, dass in univariaten Daten eine große asymmetrische Kontamination durch Ausreißer (z.B. ein Cluster von

statistischen Ausreißern) zu einem sehr kleinen Kurtosiskoeffizienten führen kann [vgl. PP01b, S.288]. Dies legt nahe, die Projektionsrichtung multivariater Daten so zu wählen, dass der Kurtosiskoeffizient der projizierten Daten möglichst klein ist. Weiterhin ist es möglich Ausreißer zu detektieren, wenn eine Projektionsrichtung gewählt wird, die senkrecht zur, den Kurtosiskoeffizienten maximierenden oder minimierenden, Richtung ist [vgl. PP01b, S.288f.]. Die projizierten univariaten Daten werden mit einem herkömmlichen robusten univariaten Verfahren zum Auffinden statistischer Ausreißer weiter untersucht und iterativ von gefundenen Ausreißern bereinigt.

3.4.1.1 Algorithmus zum Verfahren nach Pena und Prieto

Der hier vorgestellte Algorithmus findet sich bei Pena und Prieto [vgl. PP01b, S.289ff.]. Er wurde für diese Arbeit auf zwei Arten implementiert. Zum einen liegt er als getreue Umsetzung des hier gegebenen Algorithmus vor, zum anderen wurde eine von Francisco Javier Prieto zur Verfügung gestellte Implementierung des Algorithmus für MATLAB in **R** übertragen [Pri00]. Diese zweite Version verwendet modifizierte kritische Werte für r_i in Gleichung (3.9). Aus diesem Grund findet der Algorithmus in der Implementierung nach Prieto mehr statistische Ausreißer. Es konnte darüber hinaus festgestellt werden, dass die Menge der Ausreißer, welche anhand der eigens implementierten Version in den Referenzdatensätzen von Pena und Prieto [vgl. PP01b, S.295] gefunden wurden, stets eine Teilmenge der Ausreißer war, die nach der Implementierung nach Prieto gefunden wurden. Seien $x_1, \ldots, x_n$ gegebene p-dimensionale multivariate Daten.

1. Schätze den Erwartungswertvektor und die Kovarianzmatrix unter Verwen-

dung der Daten $x_1, \ldots, x_n$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad S := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

2. Bestimme die Matrix $S^{-1/2}$, so dass $S^{-1/2}S^{-1/2} = S^{-1}$, wobei S^{-1} die Inverse Matrix zu S ist.

3. Für $i = 1, \ldots, n$ bestimme $y_i = S^{-1/2}(x_i - \bar{x})$.

4. (a) Setze $y_i^{(1)} = y_i$ und $j := 1$.

   (b) Bestimme d_j so, dass

   $$d_j = \arg\max_d \frac{1}{n}\sum_{i=1}^{n}\left(d' y_i^{(j)}\right)^4, \quad \|d_j\|_2 = 1 \tag{3.8}$$

   gilt. Dies geschieht durch eine Eigenwert- und Eigenvektorbestimmung, wie im Abschnitt 3.4.1.2 beschrieben.

   (c) Berechne $z_i^{(j)} := d_j' y_i^{(j)}$ für $i = 1, \ldots, n$. Dies ist die Projektion von $y_i^{(j)}$ entlang d_j. Falls j<p gilt, berechne $Q_j := I - d_j d_j'$, definiere $y_i^{(j+1)} := Q_j y_i^{(j)}$ und setze j:=j+1. Anschließend gehe zu Schritt 4b). Falls $j = p$ gilt, gehe zu Schritt 5.

5. Führe die Schritte 4a)-4c) aus, mit dem Unterschied, dass in Gleichung (3.8) des Schrittes 4b) minimiert statt maximiert wird. Ist dies getan, liegen insgesamt $2p$ Projektionsrichtungen d_j und die zugehörigen projizierten Daten $z_i^{(j)}$ vor.

6. Es wird untersucht, ob $z_i^{(j)}$ in einer der Projektionsrichtungen auffällig ist,

ein statistischer Ausreißer sein zu können.

$$r_i = \max_{1 \leq j \leq 2p} \frac{\left| z_i^{(j)} - median\left(z^{(j)}\right) \right|}{MAD\left(z^{(j)}\right)} \tag{3.9}$$

Wenn $r_i > \beta_p$ gilt, so ist z_i möglicherweise ein statistischer Ausreißer. Der Wert von β_p ist abhängig von der Dimension p. Man wähle $\beta_5 = 4.1$, $\beta_{10} = 6.9$ oder $\beta_{20} = 10.8$. Ist die Dimension des Datensatzes $p \notin \{5; 10; 20\}$, so kann β_p durch lineare Interpolation der Punkte

$$(\log 5;\ \log \beta_5),\ \ (\log 10;\ \log \beta_{10})\ \text{ und }\ (\log 20;\ \log \beta_{20})$$

bestimmt werden. Dieses Vorgehen folgt dem Vorschlag von Pena und Prieto [vgl. PP01b, S.290].

7. Wenn in Schritt 6 ein möglicher statistischer Ausreißer gefunden wurde, erstelle einen neuen Teildatensatz bestehend aus allen Daten, für die $r_i \leq \beta_p$ gilt und gehe zu Schritt 1. Führe die gesamte Prozedur solange durch, bis in den Daten keine potentiellen statistischen Ausreißer mehr gefunden werden, oder aber durch den schrittweisen Ausschluss von Daten in Schritt 7 der Satz verbleibender Daten weniger als $(n + p + 1)/2$ betragen würde.

8. Berechne die Mahalanobis-Distanz aller als potentielle Ausreißer gekennzeichneten Daten, indem zur Schätzung des Erwartungswertvektors und der Kovarianzmatrix nur die Daten verwendet werden, die nicht als potentielle Ausreißer gekennzeichnet wurden. Sei U die Menge aller Datenpunkte, die

nicht als potentielle Ausreißer gekennzeichnet wurden.

$$\tilde{m} = \frac{1}{|U|} \sum_{i \in U} x_i, \;\; \tilde{S} = \frac{1}{|U| - 1} \sum_{i \in U} (x_i - \tilde{x})(x_i - \tilde{x})'$$

Für alle potentiellen Ausreißer ergibt sich dann die quadrierte Mahalanobis-Distanz zu

$$v_i = (x_i - \tilde{m})' \tilde{S}^{-1} (x_i - \tilde{m}) \quad (i \notin U) \;\; .$$

Alle x_i, für die $v_i < \chi^2_{p,0.99}$ $(i \notin U)$ gilt, werden nicht als Ausreißer betrachtet. Sie werden der Menge U hinzugefügt und Schritt 8 solange wiederholt, bis keine $x_i \notin U$ mehr auftreten, für die gilt $v_i < \chi^2_{p,0.99}$, oder aber $U = \{x_1, \ldots, x_n\}$ gilt.

Anhand der von Ausreißern bereinigten Daten kann nun eine robuste Schätzung des Erwartungswertvektors und der Kovarianzmatrix erfolgen.

$$\tilde{x} = \frac{1}{|U|} \sum_{i \in U} x_i, \;\; \tilde{S}_c = \frac{1}{(|U| - 1) k_p} \sum_{i \in U} (x_i - \tilde{x})(x_i - \tilde{x})'$$

Der Parameter k_p ist dimensionsabhängig und dient zur Generierung einer Kovarianzmatrix mit unverzerrter Spur (Spur=Produkt der Elemente der Hauptdiagonalen). Man wähle $k_5 = 0.98$, $k_{10} = 0.95$ und $k_{20} = 0.92$. Falls die Dimension p der Daten nicht 5, 10 oder 20 ist, so lässt sich ein passendes k_p durch lineare Interpolation der Messpunkte $(\log 5;\ \log k_5)$, $(\log 10;\ \log k_{10})$ und $(\log 20;\ \log k_{20})$ bestimmen [vgl. PP01b, S.292].

3.4.1.2 Algorithmus zur Bestimmung der Projektionsrichtungen nach Pena und Prieto

Pena und Prieto geben ein einfaches iteratives Verfahren zur Bestimmung der Projektionsrichtungen an [vgl. PP01b, S.290]. Der Ausdruck $\frac{1}{n}\sum_{i=1}^{n}\left(d'y_i^{(j)}\right)^4$ muss bezüglich $d \in \mathbb{R}^p$ maximiert (minimiert) werden. Pena und Prieto zeigen, dass diese Maximierung (Minimierung) äquivalent zur Bestimmung des Eigenvektors der Matrix $M(d) = \sum_{i=1}^{n}\left(d'y_i^{(j)}\right)^2 y_i^{(j)}y_i^{(j)'}$, der zum maximalen (minimalen) Eigenwert korrespondiert, ist[vgl. PP01b, S.290].

1. Setze für einen beliebigen Vektor $d \in \mathbb{R}^p$

$$M(d) := \sum_{i=1}^{n}\left(d'y_i^{(j)}\right)^2 y_i^{(j)}y_i^{(j)'} \quad .$$

2. Wähle einen Startvektor $\bar{d}_0$, so dass $\|\bar{d}_0\|_2 = 1$.

3. Setze $l := 0$.

4. Wähle ein beliebiges aber festes $\varepsilon > 0$, z.B. $\varepsilon = 0.001$.

5. Berechne $\bar{d}_{l+1}$ als Eigenvektor der Matrix $\bar{M}(d_l)$, der zum maximalen (minimalen) Eigenwert von $M(\bar{d}_l)$ korrespondiert. Falls nötig normiere d_{l+1}, so dass $\|\bar{d}_{l+1}\|_2 = 1$. Falls $\|\bar{d}_{l+1} - \bar{d}_l\| > \varepsilon$ gilt, setze $l := l+1$ und wiederhole Schritt 5.

6. Wenn $\|\bar{d}_{l+1} - \bar{d}_l\| < \varepsilon$ gilt, setze $d_j := \bar{d}_{l+1}$.

3.5 Allgemeines zu Zeitreihen

Bei der Bestimmung von Ausreißern in Zeitreihen werden andere, als die bisher beschriebenen Methoden verwendet. Dies liegt in der zeitlichen Ordnung der Messwerte und ihrer Korrelationsstruktur begründet. Zeitreihen weisen die Charakteristik auf, dass die Messwerte eines (unbekannten) zugrunde liegenden Prozesses beobachtet werden. Werte zum Zeitpunkt t_0 können durchaus von der Entwicklung des Prozesses abhängen, d.h. vom Zustand des Prozesses zu früheren Zeitpunkten $t < t_0$.

Allgemein habe ein Datensatz die folgende Darstellung.

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{k1} \end{pmatrix}, \ldots, \begin{pmatrix} y_{1n} \\ \vdots \\ y_{kn} \end{pmatrix} \quad \text{bzw. analog} \quad \begin{pmatrix} y_{1t} \\ \vdots \\ y_{kt} \end{pmatrix}_{t=1,\ldots,n}, \quad k \in \mathbb{N}$$

Die Daten werden als Realisationen von Zufallsvariablen $Y_t := (Y_{1t}, \ldots, Y_{kt})'$, $t = 1, \ldots, n$ aufgefasst. Für eine univariate Zeitreihe gilt offenbar $k = 1$.

Zunächst wird die Untersuchung univariater Zeitreihen auf Ausreißer vorgestellt. Das Verfahren zur Identifikation von Ausreißern in multivariaten Zeitreihen beruht im Wesentlichen auf einer Projektionsmethode, durch die mehrere univariate Zeitreihen erzeugt und auf Ausreißer untersucht werden.

Ein multivariater Ausreißer der durch äußere Einflüsse generiert wurde (z.B. Änderung der Rahmenbedingungen), kann sich in den unterschiedlichen Komponenten der multivariaten Zeitreihe unterschiedlich auswirken. Eine Überschwem-

mung kann in einer Region zu einer vorübergehenden erhöhten chemischen Belastung der dortigen Flora und Fauna führen. Dieses Verhalten wird als transient bezeichnet und der entsprechende Ausreißer als sogenannter transienter Ausreißer deklariert. In einem anderen Gebiet hingegen kann ein nachhaltiger Effekt z.B. in Form einer dauerhaften Erhöhung der durchschnittlichen chemischen Belastung in Biota auftreten. Derartige Ausreißer werden als level shifts bezeichnet, da sie die Änderung (shift) des mittleren Niveaus (level) modellieren. Ein Ausreißer kann darüber hinaus auch in Form einer Linearkombination mehrerer Ausreißertypen auftreten.

Alle hier vorgestellten Verfahren und Algorithmen zur Identifikation und Bereinigung einer Zeitreihe von Ausreißern, sind im Detail bei Galeano et al. [vgl. GPT06] nachzulesen. Prioritär ist hier die Fokussierung auf die Anwendung der dort entwickelten Verfahren und die dynamische Implementierung der Algorithmen. Ein numerischer Algorithmus zur Lösung des multivariaten Optimierungsproblemes aus Abschnitt 3.7.1.3 mittels iterativer Bestimmung von Eigenwerten wird entwickelt.

3.5.1 Das Zeitreihenmodell

3.5.1.1 Das VARMA Modell

Im Fokus dieses Abschnittes sind allgemeine Grundlagen und theoretische Ansätze zum Verständnis des VARMA (Vector Auto Regressive Moving Average) Zeitreihenmodells. Praktische Ansätze zur Inversion von Filtern und unterschiedliche Darstellungsformen für Zeitreihen die einem VARMA Modell folgen, sowie die Integration statistischer Ausreißer in das Modell werden vorgestellt [vgl.

SS01, GPT06, Dav01].

Sei $X = (X_1, \ldots, X_n)$ mit $X_t = (X_{1t}, \ldots, X_{kt})'$ für $t = 1, \ldots, n$ eine k-dimensionale Zeitreihe. Diese folge dem Vektor-ARMA[p,q] (kurz VARMA) Modell

$$\Phi(B)X_t = C + \Theta E_t \ .$$

B ist der Backshiftoperator, d.h. $B^s X_t = X_{t-s}$, $\Phi(B) = I - \Phi_1 B - \cdots - \Phi_p B^p$ und $\Theta(B) = I - \Theta_1 B - \cdots - \Theta_q B^q$. Die Φ_i, Θ_j sind Matrizen ($i = 1, \ldots, p$ und $j = 1, \ldots, q$), sowie $(E_t)_{t=1,\ldots,n}$ unabhängig identisch multivariat normal-$(0_k, \Sigma)$-verteilte Zufallsvariablen, die sogenannten Innovationen der Zeitreihe. Die Kovarianzmatrix Σ sei positiv definit und C ein konstanter Vektor. Im univariaten Fall sind 0_k und Σ folglich Skalare $0_k = 0$ und $\Sigma = \sigma_E^2$ und man spricht von einem ARMA, statt von einem VARMA Modell.

Wie ist dieses Modell zu interpretieren? Man betrachte zunächst die linke Seite: $\Phi(B)X_t$. Diese ist offenbar eine Linearkombination der Zufallsvariablen (Beobachtungen) $X_t, \ldots, X_{t-p}$. Andererseits stellt die rechte Seite $C + \Theta E_t$ eine Linearkombination der Innovationen $E_t, \ldots, E_{t-q}$ dar. Betrachtet man verschiedene Zeitpunkte t so sieht man, dass die rechte Seite ein sogenannter gleitender Durchschnitt (Moving Average) und die linke Seite quasi auf sich selbst zurückzuführen bzw. aus sich selbst entstehend ist (Auto Regressive). Zu jedem Zeitpunkt t gilt, dass eine Linearkombination aus der aktuellen Beobachtung X_t und den p vorangegangen Beobachtungen $X_{t-1}, \ldots, X_{t-p}$ stets übereinstimmt mit der Linearkombination der aktuellen Innovation E_t und den q vorangegangenen In-

novationen $E_{t-1}, \ldots, E_{t-q}$. Die praktische Interpretation des Modells ergibt sich, wenn die Gleichung nach X_t umgestellt wird.

Das folgende ARMA[1,0]-Modell modelliere die jährlich gemessene durchschnittliche Konzentration von PCB 28 in Weidelgras.

$$X_t - 0.92X_{t-1} = E_t$$

Schreibt man dieses Modell in der Form

$$X_t = 0.92X_{t-1} + E_t \ ,$$

so ergibt sich eine natürliche Interpretation: die Konzentration von PCB 28 in Weidelgras X_t hängt ab von der Konzentration im Vorjahr, sprich X_{t-1}, und dem über das Jahr hinzugekommenen neuen Eintrag von PCB 28 (also der Innovation) E_t. Weiterhin besagt das Modell, dass 8% der Konzentration des Vorjahres abgebaut wurden, denn X_{t-1} geht nicht vollständig in X_t ein, sondern mit Faktor 0.92. Anders gesagt: ein großer Teil des PCB Eintrags vom Vorjahr ist im Folgejahr noch messbar. Diese Modellierung ist insofern sinnvoll, als dass PCB persistente Stoffe sind, d.h. sie unterliegen nur sehr schwachen Abbauprozessen in der Natur. ARMA Modelle können somit eingesetzt werden, um die zeitliche Entwicklung eines Merkmals oder einer Variable zu beschreiben, wenn der aktuelle Zustand X_t abhängt von vorangegangenen Zuständen $X_{t-1}, \ldots, X_{t-p}$ und Innovationen $E_t, E_{t-1}, \ldots, E_{t-q}$.

3.5.1.1.1 Beispiel im multivariaten Fall Soll die Entwicklung mehrerer Variablen, wie z.B. PCB 28 und PCB 52, simultan modelliert werden, empfiehlt

sich die Verwendung eines VARMA Modells. Zwar ist es möglich beide Größen separat von einander univariat in ARMA Modellen darzustellen, jedoch finden dadurch eventuelle Wechselwirkungen keine Berücksichtigung. Die bivariate Zeitreihe $Y_1, \ldots, Y_n$ mit $Y_t = (Y_{1t}, Y_{2t})'$ folge einem VARMA Modell.

$$\begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix} = \begin{pmatrix} E_{1t} \\ E_{2t} \end{pmatrix} + \begin{pmatrix} 0.2 & 0.4 \\ 0.1 & 0.4 \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-2} \end{pmatrix} + \begin{pmatrix} 0.25 & 0.3 \\ 0.1 & 0.2 \end{pmatrix} \begin{pmatrix} E_{1,t-1} \\ E_{2,t-2} \end{pmatrix}$$

Obige Gleichung lässt sich analog zum univariaten Fall umstellen und liefert dann mit

$$\Phi(B) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - B \cdot \begin{pmatrix} -0.2 & -0.4 \\ -0.1 & -0.4 \end{pmatrix} \quad \text{und}$$

$$\Theta(B) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - B \cdot \begin{pmatrix} -0.25 & -0.3 \\ -0.1 & -0.2 \end{pmatrix}$$

die Darstellung $\Phi(B)Y_t = \Theta(B)E_t$.

□

Falls die Matrixpolynome $\Phi(B)$ und $\Theta(B)$ invertierbar sind, ergeben sich die sogenannten VMA$[\infty]$ und VAR$[\infty]$ Darstellungen der Zeitreihe X. Im Fall $k = 1$

handelt es sich dann analog zum ARMA Modell um MA und AR Darstellungen.

$$VAR: \ \Pi(B)X_t = C_\Pi + E_t, \quad \Pi(B) = \Theta(B)^{-1}\Phi(B) = I - \sum_{i=1}^{\infty} \Pi_i B^i,$$

$$C_\Pi = \Theta(B)^{-1}C$$

$$VMA: \ X_t = C_\Psi + \Psi(B)E_t, \quad \Phi(1)C_\Psi = C, \quad \Psi(B) = \Phi(B)^{-1}\Theta(B),$$

$$\Psi(B) = I + \sum_{i=1}^{\infty} \Psi_i B^i$$

Es ist bekannt, dass die Projektion $x_t = v'X_t$ $(t = 1, \ldots, n)$ $(v \neq 0)$ eines VARMA$[p, q]$ Prozesses einem ARMA$[p', q']$ Prozess mit $p' \leq kp$ und $q' \leq (k-1)p + q$ folgt [vgl. Lüt93]. Folglich besitzt die Zeitreihe $v'X_1, \ldots, v'X_n$ eine eindimensionale ARMA Darstellung und im Falle der Invertierbarkeit der zugehörigen Koeffizientenpolynome auch eine AR$[\infty]$ und eine MA$[\infty]$ Darstellung.

$$\text{ARMA}: \ \phi(B)x_t = c + \theta(B)e_t$$

$$\text{AR}: \ \pi(B)x_t = c_\pi + e_t, \ \ c_\pi = \theta(1)^{-1}c,$$

$$\pi(B) = \theta(B)^{-1}\phi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$$

$$\text{MA}: \ x_t = c_\psi + \psi(B)e_t, \ \ c_\psi = \phi(1)^{-1}c,$$

$$\psi(B) = \phi(B)^{-1}\theta(B) = 1 + \sum_{i=1}^{\infty} \psi_i B^i$$

Dabei ist e_t ein univariater white noise Prozess mit Varianz σ_e^2 und Erwartungswert 0.

3.5.1.2 Inversion eines Matrixpolynoms: von der VAR zur VMA-Darstellung

Die Frage nach der Invertierbarkeit eines beliebigen Matrixpolynoms $\kappa(B)$ ist von besonderer Bedeutung, da nur in einem solchen Fall die entsprechende VMA[∞]-Darstellung aus der VAR[p]-Darstellung der Zeitreihe gewonnen werden kann, indem $\kappa(B) = \Pi(B)$ invertiert wird. Im univariaten Fall kann die Inverse von $\kappa(B)$ genau dann bestimmt werden, wenn alle Nullstellen des Polynoms $\kappa(z)$ für $z \in \mathbb{C}$ außerhalb des Einheitskreises liegen [vgl. SS01, S.113 ff.].

3.5.1.2.1 Beispiel Sei $\kappa(B) = 1 - \frac{1}{6}B - \frac{1}{3}B^2$. Man ersetze in dieser Gleichung B durch z (die sog. z-Transformierte, [vgl. SS01, S.113]). Im zweiten Schritt werden die Nullstellen des Polynoms $\kappa(z) = 1 - \frac{1}{6}z - \frac{1}{3}z^2$ bestimmt. Diese sind nach der Lösungsformel für quadratische Funktionen gegeben durch $z_1 = 1.5$ und $z_2 = -2$. Damit gilt offenbar die folgende Gleichung.

$$\kappa(z) = 1 - \frac{1}{6}z - \frac{1}{3}z^2 = \frac{1}{-1.5 \cdot 2}(z - 1.5)(z + 2).$$

Die Inverse ergibt sich rein formal durch Partialbruchzerlegung von $1/\kappa(z)$. Nähere Informationen zu derartigen Darstellungen finden sich im Abschnitt 3.5.1.3. Die praktische Umsetzung in diesem Beispiel wird im Folgenden illustriert. Sie ist auf den allgemeinen Fall $\kappa(B) = \Pi(B)$ übertragbar, sofern Π gewisse Regularitätsbedingungen erfüllt, die im weiteren Verlauf dieses Abschnittes erläutert werden.

$$
\begin{aligned}
\kappa(z)^{-1} &= \left(\frac{1}{-1.5 \cdot 2}(z - 1.5)(z + 2) \right)^{-1} \\
&= \frac{1}{\frac{1}{-3}(z - 1.5)(z + 2)} \\
&= \frac{1}{\frac{1}{-3}(-1.5)\left(1 - \frac{1}{1.5}z\right) 2 \left(1 - \frac{1}{-2}z\right)} \\
&= \frac{1}{\left(1 - \frac{2}{3}z\right)\left(1 - \left(-\frac{1}{2}\right)z\right)} \\
&= \left(\sum_{j=0}^{\infty} \left(\frac{2}{3}\right)^j z^j \right) \left(\sum_{j=0}^{\infty} \left(-\frac{1}{2}\right)^j z^j \right) \\
&= \sum_{j=0}^{\infty} \sum_{l=0}^{\infty} \left(-\frac{1}{2}\right)^j \left(\frac{2}{3}\right)^l z^{j+l}
\end{aligned}
$$

Die Resubstitution von B für z liefert die inverse Matrix zu $\kappa(B)$, d.h. $\kappa(B)^{-1}$ [vgl. SS01, S.113 ff.]. Das allgemeine Vorgehen bei der Bildung der Inversen von $\kappa(B)$ ist vollkommen analog. Man ersetze in $\kappa(B)$ zunächst B durch z und bestimme alle Nullstellen des Polynoms. Diese seien bezeichnet mit $z_1, \ldots, z_p$ und müssen nicht paarweise verschieden sein. Anschließend wird das Polynom in seiner Linearfaktorzerlegung $c \cdot (z - z_1) \cdot \ldots \cdot (z - z_p)$ notiert und aus jedem Linearfaktor die entsprechende Nullstelle ausgeklammert, sprich man wählt die Darstellung

$$
c \cdot (z - z_1) \cdot \ldots \cdot (z - z_p) = c \cdot \left(\prod_{i=1}^{p} (-z_i) \right) \cdot (1 - z/z_1) \cdot \ldots \cdot (1 - z/z_p) \ .
$$

Die Inverse von $\kappa(B)$ ergibt sich dann aus der Darstellung

$$\begin{aligned}
&\left(c\prod_{i=1}^{p}(-z_i)\cdot\left(1-\frac{z}{z_1}\right)\cdot\ldots\cdot\left(1-\frac{z}{z_p}\right)\right)^{-1}\\
=&\frac{1}{c\cdot\prod_{i=1}^{p}(-z_i)}\frac{1}{(1-z/z_1)}\cdot\ldots\cdot\frac{1}{(1-z/z_p)}\\
=&C\cdot\left(\sum_{j=1}^{\infty}\left(\frac{1}{z_1}\right)^j z^j\right)\cdot\ldots\cdot\left(\sum_{j=1}^{\infty}\left(\frac{1}{z_p}\right)^j z^j\right)
\end{aligned}$$

Dabei ist $C=\frac{1}{c\cdot\prod_{i=1}^{p}(-z_i)}$ und alle Nullstellen $z_1,\ldots,z_p$ müssen die Bedingung $|z_i|>1$ erfüllen $(i=1,\ldots,p)$ [vgl. SS01, S.114 f.]. Möchte man beispielsweise aus der AR-Darstellung einer reellen Zeitreihe die MA[∞]-Darstellung bestimmen, so wäre $\kappa(B)=\Pi(B)$ und $\kappa(B)^{-1}=\Psi(B)$.

□

Um $\kappa(z)^{-1}$ bilden zu können, ist es notwendig, dass $\kappa(z)$ nur komplexe Nullstellen $z_1,\ldots,z_p\in\mathbb{C}$ mit $|z_i|>1$ besitzt [vgl. SS01, S.113 ff.]. Nach Schlittgen und Streitberg lässt sich dann $\kappa(B)^{-1}$ bestimmen. Ein Problem in der Berechnung von $\kappa(z)^{-1}$ ergibt sich, wenn mindestens eine der Nullstellen z_i echt komplex ist, d.h. wenn sie einen Imaginärteil ungleich Null besitzt. Ohne Beschränkung der Allgemeinheit sei z_1 echt komplex, d.h. $Im(z)\neq 0$. Damit ist $\kappa(z)^{-1}$ ein komplexes Polynom und folglich ist die MA[∞]-Darstellung der Zeitreihe ebenfalls komplex, d.h. die reelle Zeitreihe $X_1,\ldots,X_n$ besitzt sowohl eine relle AR[p]- als auch eine komplexe MA[∞]-Darstellung. Mit einer derartigen Zeitreihe sind keine weiteren ausreißeranalytischen Untersuchungen möglich. In diesem Fall versagt obiges Verfahren also. Es ist folglich für univariate Zeitreihen nur anwendbar, wenn alle Nullstellen z_i des charakteristischen Polynoms $\kappa(z)$ die folgenden Bedingungen

erfüllen:

$$(i)\ |z_i| > 1\ \forall\ i = 1, \ldots, p$$
$$(ii)\ Im(z_i) = 0\ \forall\ i = 1, \ldots, p$$

Komplizierter wird es, wenn es sich um multivariate Zeitreihen $X_1, \ldots, X_T$ mit $X_t = (X_{t1}, \ldots, X_{tk})'$ für $t = 1, \ldots, T$ handelt. In diesem Fall ist $\kappa(B) = \Pi(B)$ ein Matrixpolynom. Nach Lütkepohl existiert $\Pi(B)^{-1}$ genau dann, wenn gilt:

$$det(\Pi(z)) \neq 0\ \ \forall\ |z| \leq 1 \tag{3.10}$$

wobei det die Determinante bezeichnet [vgl. Lüt05, S.22]. Um $\Pi(B)^{-1}$ zu bestimmen, wähle man zunächst eine allgemeine Darstellung der Art

$$\Pi(B)^{-1} = \Phi_0 + \Phi_1 B + \Phi_2 B^2 + \ldots$$

mit noch unbekannten Koeffizientenmatrizen $\Phi_0, \Phi_1, \Phi_2, \ldots$. Nach Lütkepohl folgt aus der Bedingung 3.10 die absolute Summierbarkeit der Φ_i mit $\lim_{i\to\infty} \Phi_i = 0_{(k,k)}$, wobei $0_{(k,k)}$ die $k \times k$ Matrix ist, deren sämtliche Einträge Null sind[vgl. Lüt05, S.22 ff.]. Aus der Gleichung

$$\begin{aligned} I &= \Pi(B)^{-1}\Pi(B) \\ &= (\Phi_0 + \Phi_1 B + \Phi_2 B^2 + \ldots)(I - \Pi_1 B - \Pi_2 B^2 - \ldots - \Pi_p B^p) \\ &= \Phi_0 + (\Phi_1 - \Phi_0\Pi_1)B + (\Phi_2 - \Phi_1\Pi_1 - \Phi_0\Pi_2)B^2 + \ldots \end{aligned}$$

ergeben sich mittels Koeffizientenvergleich die folgenden Gleichungen [vgl. Lüt05, S.22 f.]

$$\begin{aligned} I &= \Phi_0 \\ 0 &= \Phi_1 - \Phi_0\Pi_1 \\ 0 &= \Phi_2 - \Phi_1\Pi_1 - \Phi_0\Pi_2 \\ &\vdots \\ 0 &= \Phi_i - \sum_{j=1}^{i} \Phi_{i-j}\Pi_j \\ &\vdots \end{aligned} \tag{3.11}$$

mit $\Pi_j = 0$ für $j > p$. Demzufolge können die Φ_i aus Gleichung (3.11) bestimmt werden [vgl. Lüt05, S.23]:

$$\begin{aligned} \Phi_0 &= I \\ \Phi_i &= \sum_{j=1}^{i} \Phi_{i-j}\Pi_j \ , \ \ i = 1, 2, \ldots \end{aligned}$$

$\Pi(B)^{-1}$ existiert genau dann, wenn $det(\Pi(z)) \neq 0 \ \ \forall \, |z| \leq 1$ gilt. Dies ist aber abhängig von der konkreten Besetzung der Matrizen $\Pi_1, \ldots, \Pi_p$. Daher werden an dieser Stelle einige Konventionen getroffen, um die vorliegenden Daten in ihrer Fülle auswerten zu können. Für die in dieser Arbeit untersuchten Datensätze gilt:

1. Die multivariaten Zeitreihen besitzen eine Beobachtungszahl von maximal $n = 20$ Messpunkten.

2. Für derart kurze Zeitreihen scheint es wenig sinnvoll VAR[p]-Modelle mit

$p > 1$ anzupassen.

3. Die Bestimmung von $\Psi(B) = \Pi(B)^{-1}$ (und damit die Bestimmung der MA[∞]-Darstellung) ist praktisch mit einem sehr hohen Aufwand verbunden, wenn sie denn überhaupt gelingt.

4. Additive, transiente und innovative Ausreißer werden unter Verwendung der VAR[1]-Darstellung ermittelt (zu den verschiedenen Ausreißertypen vgl. Abschnitt 3.5.1.3). Die VMA[∞]-Darstellung der Zeitreihen ist nur notwendig, um die Zeitreihe vom Effekt innovativer Ausreißer zu bereinigen bzw. um zu überprüfen, ob es sich bei gefundenen additiven und transienten Ausreißern nicht um fehlspezifizierte innovative Ausreißer handelt.

5. Um zu überprüfen, ob die gefundenen additiven und transienten Ausreißer tatsächlich additiver bzw. transienter Art sind, muss ausgeschlossen werden, dass es sich bei ihnen in Wahrheit um fehlspezifizierte innovative Ausreißer handelt.

6. Die Zeitreihen werden vom Effekt des jeweils gefundenen Ausreißers bereinigt. Aufgrund des kurzen Beobachtungszeitraumes ist die zu erwartende Auswirkung einer Fehlspezifikation von innovativen Ausreißern als transiente oder additive Ausreißer vernachlässigbar. Man überlege sich Folgendes:

   1.Fall: Es wird ein additiver Ausreißer erkannt, obwohl in Wahrheit ein innovativer Ausreißer vorliegt. Dies kann nur geschehen, wenn der Effekt des innovativen Ausreißers schnell abklingt, denn sonst würde er entweder tatsächlich als innovativer Ausreißer oder aber fälschlicher Weise als transi-

enter Ausreißer erkannt werden. Die Nachhaltigkeit der Wirkung des innovativen Ausreißers ist aufgrund der Kürze der vorliegenden Zeitreihen nur schwer zu beurteilen. Das rasche Abklingen des Effektes und der kurze Beobachtungshorizont legen nahe, dass der Effekt des innovativen Ausreißers näherungsweise als additiver Ausreißer geschätzt werden kann.

2.Fall: Es wird ein transienter Ausreißer erkannt, obwohl in Wahrheit ein innovativer Ausreißer vorliegt. Dies kann nur geschehen, wenn der Effekt des innovativen Ausreißers mittelfristig ($n = 10$ Beobachtungen) mit näherungsweise konstanter Rate abklingt. Da für die Zeitreihen nur ein kurzer Beobachtungshorizont von maximal 20 Jahren vorliegt, kann der Effekt des innovativen Ausreißers nicht langfristig geschätzt werden. Weiter ist es nicht das Ziel dieser Arbeit Prognosen zu entwickeln, wofür die korrekte Spezifikation des Ausreißers als innovativ notwendig wäre. Eine mittelfristige Approximation durch einen transienten Ausreißer wird somit als hinreichend genau erachtet.

7. Auf eine genauere Untersuchung, ob es sich bei additiven und transienten Ausreißern eventuell um fehlspezifizierte innovative Ausreißer handelt, wird aus obigen Gründen in dieser Arbeit verzichtet.

8. Der programmiertechnische Aufwand zur Bestimmung der Nullstellen des charakteristischen Polynoms $\kappa(z) = \Pi(z)$ ist enorm groß, da **R** kein Computeralgebrasystem ist und somit auch nicht in der Lage, symbolisch zu arbeiten. Allein die Überprüfung der Bedingung $det(\kappa(z)) \neq 0 \ \forall \ |z| \leq 1$ gestaltet sich für die hier betrachteten Zeitreihen sehr aufwändig.

9. Werden innovative Ausreißer in der Zeitreihe detektiert, so besteht zum

einen die Möglichkeit, sie als transiente Ausreißer zu deklarieren und ihren Effekt dadurch näherungsweise aus der Zeitreihe zu entfernen. Zum anderen ist es möglich den Effekt innovativer Ausreißer nicht zu schätzen und diese in der Zeitreihe zu belassen. Dem Anwender stehen durch die, im Zusammenhang mit dieser Arbeit implementierten Algorithmen, beide Möglichkeiten zur Verfügung. Anhand der Daten wird untersucht, ob eine Deklaration innovativer Ausreißer als transient tatsächlich praktikabel ist.

3.5.1.3 Integration von Ausreißern in das VARMA Modell

Sei die Zeitreihe $Y = (Y_1, \ldots, Y_n)$ mit $Y_t = (Y_{1t}, \ldots, Y_{kt})'$ gegeben. Enthält Y Ausreißer, so können diese unter Verwendung einer unbeobachteten und nicht mit Ausreißern kontaminierten Zeitreihe $X = (X_1, \ldots, X_n)$ modelliert werden.

$$Y_t = X_t + \alpha(B) w I_t^{(h)}$$

Der Effekt eines Ausreißer zum Zeitpunkt h wird beschrieben durch w. $I_t^{(h)}$ ist eine Dummyvariable, für die $I_h^{(h)} = 1$ und $I_t^{(h)} = 0$ für $t \neq h$ gilt. Sie modelliert den Zeitpunkt des Auftretens eines Ausreißers. Die Koeffizientenmatrix $\alpha(B)$ bestimmt die Art des Ausreißers und der Vektor w die konkrete Ausprägung in jeder der k Dimensionen. X folge einem VARMA Modell. Da X eine VARMA Dar-

stellung besitzt, gilt dies auch für Y.

$$\begin{aligned} X_t + \alpha(B) w I_t^{(h)} = Y_t &= C_\Psi + \Psi(B) A_t \qquad (3.12) \\ \Leftrightarrow\ A_t &= \underbrace{\Pi(B) X_t - C_\Pi}_{=E_t} + \underbrace{\Pi(B)\alpha(B)}_{=:\Gamma(B)} w I_t^{(h)} \\ \Leftrightarrow\ A_t &= E_t + \Gamma(B) w I_t^{(h)} \ \text{ mit } \ \Gamma(B) = I - \sum_{i=1}^{\infty} \Gamma_i B^i \end{aligned}$$

Für $t < h$ gilt folglich $A_t = E_t$. A_t und E_t werden auch Innovationen der Zeitreihen Y und X genannt. Zu den hier verwendeten Bezeichnungen wird auf Abschnitt 3.5.1.1 verwiesen. Je nach Ausreißertyp ist w deterministisch oder stochastisch.

Galeano et al. unterscheiden in stationären Zeitreihen vier verschiedene Ausreißertypen [vgl. GPT06, S.655].

(i) Multivariate Level Shift (MLS bzw. „L“)

(ii) Multivariate Additive Outlier (MAO bzw. „A“)

(iii) Multivariate Innovation Outlier (MIO bzw. „I“)

(iv) Multivariate Transitory Change (MTC bzw. „T“)

Additive Ausreißer, bzw. Ausreißer vom Typ „A“, beeinflussen lediglich den Messwert zum Zeitpunkt ihres Auftretens. Sie werden z.B. durch Messfehler generiert. Dahingegen beeinflussen Ausreißer vom Typ „L“ alle folgenden Beobachtungen, indem die Zeitreihe in ihrem mittleren Niveau (level) verschoben wird, daher die Bezeichnung „level shift“. Ein innovativer Ausreißer, oder auch Ausrei-

ßer vom Typ „I“, stellt eine interne Änderung in der Struktur der Zeitreihe dar und kann, je nach Modell bzw. Struktur der Zeitreihe, Einfluss auf die darauf folgenden Zeitpunkte haben. Das Auftreten eines Ausreißers vom Typ „T“ ist charakterisiert durch seine initialisierende Wirkung zur Änderung des Prozesses. Diese Änderung nimmt mit einer konstanten Rate in den darauf folgenden Beobachtungen ab. Es handelt sich also um eine vorübergehende („transitory“) Änderung. Für Ausreißer des Typs i=A,T,L ist der Effekt w folglich deterministisch, für innovative Ausreißer hingegen ist der Effekt w zufällig.

In nicht stationären Zeitreihen werden darüber hinaus Ausreißer eines fünften Typs betrachtet. Diese beschreiben Trendänderungen, auch Multivariate Ramp Shift (MRS oder kurz Ausreißer vom Typ „R“) genannt.

Die Art eines Ausreißers wird durch die Koeffizientenmatrix $\alpha(B)$ bestimmt. Diese hat für jeden Typ eine andere Gestalt [vgl. Tsa88, S.3 f.].

- Sei w vom Typ „A“, dann besitzt $\alpha(B)$ die Gestalt der $k \times k$ Einheitsmatrix I.
- Sei w vom Typ „T“, dann besitzt $\alpha(B)$ die Gestalt $(I - \delta IB)^{-1}$ mit $0 < \delta < 1$.
- Sei w vom Typ „I“, dann besitzt $\alpha(B)$ die Gestalt $\Psi(B)$.
- Sei w vom Typ „L“, dann besitzt $\alpha(B)$ die Gestalt $(I - IB)^{-1}$.

Im univariaten Fall ersetze man in den obigen Punkten die Einheitsmatrix I durch die reelle 1.

Die konkreten Darstellungen von $(I-\delta IB)^{-1}$ und $(I-IB)^{-1}$ sind essentiell, da sie in der Modellierung des Ausreißertyps und damit zur konkreten Schätzung des Effektes des Ausreißers benötigt werden. Um sie zu gewinnen,ist es zunächst sinnvoll im univariaten Fall eine Darstellung von $(1-\delta B)^{-1}$ in Form eines Polynoms zu finden, da $\delta BX_t = \delta X_{t-1}$ bestimmt werden kann, nicht aber $(1/(1-\delta B))X_t$. Man bedenke zu jeder Zeit, dass B der sogenannte Backshift Operator ist. Mit diesem Operator ist keine Division im üblichen Sinne möglich.

Wie in Abschnitt 3.5.1.2 erläutert, muss für die Nullstelle z_0 des charakteristischen Polynoms $\kappa(z) = (1-\delta z)$ gelten: $|z_0| > 1$. Dies ist offenbar genau dann erfüllt, wenn $|\delta| < 1$ gilt, denn $z_0 = 1/\delta$. Für $|\delta| < 1$ ist die Koeffizientenfolge $(\delta^j)_{j=1,2,...}$ absolut summierbar. Nach Schlittgen und Streitberg ist die Inverse von $\kappa(B) = 1-\delta B$ für $|\delta| < 1$ gegeben durch [vgl. SS01, S.113]

$$\kappa(B)^{-1} = \sum_{j=0}^{\infty} \delta^j B^j \quad .$$

Die Unendlichkleit der Summation ist in der praktischen Anwendung nicht weiter problematisch, da alle Schätzgleichungen nur eine endliche Anzahl dieser Summanden verwenden, wie sich noch zeigen wird. So ist nun eine, für die Praxis brauchbare, Darstellung von $(1-\delta B)^{-1}$ gefunden.

Im multivariaten Fall besteht die Aufgabe darin $(I-\delta IB)^{-1}$ zu bestimmen. I ist $k \times k$ Einheitsmatrix. Das charakteristische Polynom $\kappa(z) = I - \delta Iz$ hat die

Gestalt

$$\kappa(z) = \begin{pmatrix} 1-\delta z & 0 & \cdots & 0 \\ 0 & 1-\delta z & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1-\delta z \end{pmatrix}$$

und damit ist die Determinante $det(\kappa(z)) = (1-\delta z)^k$. Offenbar gilt für die k-fache Nullstelle $z_0 = 1/\delta$ die Bedingung $|z_0| > 1$ genau dann, wenn $|\delta| < 1$ gilt. Die Inverse $(I - \delta IB)^{-1}$ lässt sich nun einfach angeben.

$$\begin{aligned}(I-\delta IB)^{-1} &= \begin{pmatrix} 1-\delta B & 0 & \cdots & 0 & 0 \\ 0 & 1-\delta B & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & 1-\delta B & 0 \\ 0 & \cdots & & 0 & 1-\delta B \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \sum_{j=0}^{\infty}\delta^j B^j & 0 & \cdots & 0 & 0 \\ 0 & \sum_{j=0}^{\infty}\delta^j B^j & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \sum_{j=0}^{\infty}\delta^j B^j & 0 \\ 0 & \cdots & & 0 & \sum_{j=0}^{\infty}\delta^j B^j \end{pmatrix}\end{aligned}$$

Des Weiteren verbleibt die Aufgabe zunächst im univariaten Fall $(1-B)^{-1}$ zu bestimmen. Nach Schlittgen und Streitberg existiert die Inverse von $\kappa(B) = (1-B)$ für $|B| < 1$ [vgl. SS01, S.114]. Sie ist gegeben durch

$$(1-B)^{-1} = 1 + B + B^2 + \dots \quad .$$

Auch hier gilt, dass für alle praktischen Anwendungen lediglich endlich viele Summanden der Gleichung verwendet werden. Augenscheinlich ist jedoch der Fakt, dass die Nullstelle z_0 des charakteristischen Polynoms $\kappa(z) = 1 - z$ nicht die Bedingung $|z_0| > 1$ erfüllt, denn es gilt $|z_0| = 1$. Die Reihe

$$(1 - z)^{-1} = \sum_{u=1}^{\infty} z^u$$

konvergiert damit nur für $|z| < 1$, wobei die Koeffizientenfolge $1, 1^2, 1^3, \ldots$ nicht absolut summierbar ist [vgl. SS01, S.114 ff.].

Im multivariaten Fall ergibt sich $(I - IB)^{-1}$ analog. Das charakteristische Polynom

$$\kappa(z) = \begin{pmatrix} 1-z & 0 & \cdots & 0 \\ 0 & 1-z & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1-z \end{pmatrix}$$

besitzt die Determinante $det(\kappa(z)) = (1 - z)^k$, welche die k-fache Nullstelle $z_0 = 1$ hat. Weitere Ausführungen würden denen im univariaten Fall gleichen. Daher wird an dieser Stelle auf sie verzichtet und sofort die Inverse $(I - IB)^{-1}$ angegeben.

$$(I - IB)^{-1} = \begin{pmatrix} 1-B & 0 & \cdots & 0 & 0 \\ 0 & 1-B & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & 1-B & 0 \\ 0 & \cdots & & 0 & 1-B \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \sum_{j=0}^{\infty} B^j & 0 & \cdots & 0 & 0 \\ 0 & \sum_{j=0}^{\infty} B^j & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \sum_{j=0}^{\infty} B^j & 0 \\ 0 & \cdots & & 0 & \sum_{j=0}^{\infty} B^j \end{pmatrix}$$

3.5.2 Schätzung eines Zeitreihenmodells

Um Zeitreihenmodelle anhand von Daten schätzen zu können, wurden in der statistischen Theorie vielfältige Verfahren entwickelt. Sowohl Kleinste-Quadrate Ansätze ([GR06], [Fox02], [Nie05]), als auch Maximum Likelihood Methoden stehen dem Anwender zur Verfügung ([Sow92], [Ziv05], [Bar], [Mau95], [GR06]), um die AR oder ARMA Darstellung einer Zeitreihe zu schätzen. Alternativ kann die Momentenmethode nach Gallant und Tauchen [GT96] verwendet werden. Die vorgeschlagenen Algorithmen sind besonders für den Maximum Likelihood Ansatz oft sehr rechenintensiv und benötigen eine entsprechende Computerleistung. Um diesen Nachteil auszugleichen, entwickelten beispielsweise McLeod und Zhang einen weniger rechenintensiven Algorithmus ([MZ08], [Chu01]).

3.5.2.1 Kleinste-Quadrate-Schätzung eines Zeitreihenmodells

Ein einfacher praktischer Ansatz zur Schätzung von Zeitreihenmodellen ist gegeben durch die gewöhnliche Kleinste-Quadrate-Schätzung, auch als OLS (Ordinary Least Squares) bezeichnet. Die theoretischen Hintergründe und Beweise sind bei Davidson nachzulesen [vgl. Dav01]. Das Verfahren zur Schätzung der Koeffizienten wird von Fuller anschaulich dargestellt [vgl. Ful96, S.407] und an dieser Stelle kurz erläutert. Aus der asymptotischen Verteilung des Parameterschätzers ergeben sich kanonisch die üblichen α-Signifikanztests in Analogie zum Fall eines herkömmlichen linearen Regressionsmodells mit unabhängigen exogenen Variablen. Die OLS Schätzung von Zeitreihenmodellen unterscheidet sich in der konkreten Vorgehensweise nicht vom klassischen Verfahren der linearen Regression.

Sei $(y_t)_{t=1,\ldots,T}$ eine univariate Zeitreihe. Diese werde modelliert durch die Gleichung

$$y_t = x_t'\beta + \varepsilon_t \tag{3.13}$$

mit $x_t \in \mathbb{R}^p,\ \beta \in \mathbb{R}^{p+1}$ und ε als Störgröße des Modells. Die Variable x_t kann sowohl Einträge y_{t-i} ($i \in \{1, \ldots, t-1\}$), als auch rein deterministische Variablen enthalten. Ein AR[2] Prozess hat beispielsweise die Darstellung

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$$

und lässt sich mit $x_t = (1, y_{t-1}, y_{t-2})' \in \mathbb{R}^3$ und $\beta = (\beta_0, \beta_1, \beta_2)' \in \mathbb{R}^3$ in der Form $y_t = x_t'\beta + \varepsilon_t$ schreiben.

Sei das Modell $y_t = x_t'\beta + \varepsilon_t$ gegeben mit $x_t = (1, y_{t-1}, \ldots, y_{t-p})$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ für $t = 1, \ldots, T$. Es liegen folglich T Zeitpunkte vor, zu denen Daten bekannt sind. Diese seien bezeichnet mit $\dot{y_1}, \ldots, \dot{y_T}$. Da die rechte Seite von Gleichung (3.13) die Zufallsvariable ε_t enthält, ist auch die linke Seite y_t eine Zufallsvariable. Weiterhin beinhalten die Vektoren $x_T, \ldots, x_{p+1}$ die Zufallsvariablen $y_{t-1}, \ldots, y_{t-p}$ für $t = p+1, \ldots, T$. Damit sind diese ebenfalls (vektorwertige) Zufallsvariablen. Um die Notation zu vereinfachen, werden analog zu Davidson folgende Bezeichnungen eingeführt [vgl. Dav01, S.4]:

$$Y = \begin{pmatrix} y_T \\ \vdots \\ y_{T-p+1} \end{pmatrix}, \quad X = \begin{pmatrix} x_T \\ \vdots \\ x_{p+1} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_{p+1} \\ \vdots \\ \varepsilon T \end{pmatrix} .$$

Damit lässt sich das Zeitreihenmodell in der Form

$$Y = X\beta + \varepsilon$$

schreiben. Die explizite Darstellung der Matrix X ist gegeben durch

$$X = \begin{pmatrix} 1 & y_{T-1} & y_{T-2} & \cdots & y_{T-p} \\ 1 & y_{T-2} & y_{T-3} & \cdots & y_{T-p-1} \\ \vdots & & \vdots & \ddots & \vdots \\ 1 & y_p & y_{p-1} & \cdots & y_1 \end{pmatrix} .$$

Indem die Daten $\dot{y_1}, \ldots, \dot{y_T}$ in die Darstellungen für Y und X eingesetzt werden, erhält man analog $\dot{Y}$ und $\dot{X}$. Nach Davidson besitzt der unbekannte Parameter-

vektor β die folgende Darstellung [vgl. Dav01, S.12]:

$$\beta = E(x_t x_t')^{-1} E(x_t y_t) \ . \tag{3.14}$$

Schätzt man die Erwartungswerte in Formel (3.14) aus den gegebenen Daten, so erhält man einen Schätzer für β mittels

$$\hat{\beta} = (\dot{X}'\dot{X})^{-1}\dot{X}'\dot{Y} \ ,$$

sofern $det(\dot{X}'\dot{X}) \neq 0$. Nach Davidson gilt $X'X = \sum_{t=p+1}^{T} x_t x_t'$ und $X'Y = \sum_{t=p+1}^{T} x_t Y_t$ [vgl. Dav01, S.12]. Die asymptotische Konsistenz des Schätzers wird gesichert durch Nielsen [vgl. Nie05, S.3]. Um eine sinnvolle OLS Schätzung des Parametervektors β zu erhalten, bedarf es einiger Regularitätsbedingungen.

- Unter der Annahme, dass $(y_t, x_t')'$ eine gemeinsame stationäre Wahrscheinlichkeitsverteilung besitzen und dass weiterhin $(y_t, x_t')'$ und $(y_{t+k}, x_{t+k}')'$ für $k \to \infty$ asymptotisch stochastisch unabhängig sind, übertragen sich im Wesentlichen die Resultate der OLS Methode für lineare Regressionsmodelle auf die OLS Methode für Zeitreihen.

  Die Forderung der Stationarität impliziert, dass sich die Zeitreihe in ihrer Zukunft genauso verhält wie in ihrer Vergangenheit. Praktisch folgt daraus, dass die OLS Schätzung $\hat{\beta}$ des Parametervektors β asymptotisch erwartungstreu ist, sprich $E(\hat{\beta}) \to \beta$ für $T \to \infty$ [vgl. Nie05, S.3].

- Fordert man weiterhin, dass die Störgröße ε_t die Bedingungen

$$(i)\ E(\varepsilon_t x_t) = 0$$

$$(ii)\ E(\varepsilon_t^2 | x_t) = \sigma^2 \text{ (Homoskedaszität)}$$

$$(iii)\ E(\varepsilon_t \varepsilon_s | x_t, x_s) = 0 \ \forall\, t \neq s \text{ (keine Autokorrelation)}$$

  erfüllt, so ergibt sich die asymptotische Normalverteilung des Schätzers $\hat{\beta}$ [vgl. Nie05, S.8].

$$\sqrt{T}\left(\hat{\beta} - \beta\right) \to \mathcal{N}\left(0, \sigma^2 E(x_t x_t')^{-1}\right) \tag{3.15}$$

  Dies folgt aus dem Zentralen Grenzwertsatz.

Mit Hilfe des obigen Resultates lassen sich asymptotische α-Signifikanztests konstruieren. Ersetzt man in Formel (3.15) σ^2 und $V = E(x_t x_t')$ durch geeignete Schätzer $\hat{\sigma}^2$ und

$$\hat{V} := \widehat{E(x_t x_t')} = \frac{1}{T-p} \sum_{t=p+1}^{T} \dot{x}_t \dot{x}_t' \ ,$$

so ergibt sich nach Nielsen die folgende asymptotische Eigenschaft [vgl. Nie05, S.8]:

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \hat{\sigma}^2 V^{-1}\right).$$

Sei $n = T - p$ die Anzahl der $\dot{x}_t$, die zur Berechnung von $\hat{V}$ verwendet werden und $b = p + 1$. Die Zufallsvektoren $x_{p+1}, \ldots, x_T$ entstammen einer stetigen multivariaten Verteilung. Nach Hartlap et al. beträgt die Wahrscheinlichkeit dafür, dass $x_{p+1}, \ldots, x_T$ linear abhängig sind Null, bzw. äquivalent formuliert sind $x_{p+1}, \ldots, x_T$ fast sicher linear unabhängig. Weiter zeigen Hartlap et al., dass die

geschätzte Kovarianzmatrix $\hat{V}$ singulär ist, falls $b > n - 1$ gilt [vgl. HSS08, S.2]. Folglich existiert $\hat{V}^{-1}$ fast sicher für $n \geq b + 1$. Ein konsistenter Schätzer für σ^2 ergibt sich aus den Residuen der OLS Schätzung des Modells. Diese seien bezeichnet mit $\hat{\varepsilon}_t = y_t - x_t\hat{\beta}$, $t = p + 1, \ldots, T$. Dann ist

$$\hat{\sigma}^2 = \frac{1}{n - 2p - 1} \sum_{t=p+1}^{T} \hat{\varepsilon}_t^2$$

ein konsistenter Schätzer für die unbekannte Varianz σ^2 der Störgröße ε. Die theoretischen Details und Hintergründe zu den Regularitätsbedingungen finden sich bei Davidson [vgl. Dav01] und bei Fuller [vgl. Ful96, S.407ff.].

Falls Annahme (iii) verletzt ist, finden sich alternative Methoden zur konsistenten Schätzung der Autokorrelation sowohl bei Stock und Watson [vgl. SW03, S.504-507], als auch bei Verbeek [vgl. Ver04, Abschnitt 4.10.2].

Aus der Tatsache der asymptotischen Normalverteilung des OLS Schätzers $\hat{\beta}$ ergeben sich die Standard t-Tests und F-Tests als asymptotische Prüfmethode der Koeffizienten auf Signifikanz.

3.5.2.1.1 Bemerkung Eine multivariate Zeitreihe lässt sich mit der OLS Methode schätzen, indem jede ihrer Komponenten separat mit der OLS Methode geschätzt wird. Diese Behauptung wird hier kurz bewiesen.

Sei eine multivariate Zeitreihe

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{k1} \end{pmatrix}, \ldots, \begin{pmatrix} y_{1T} \\ \vdots \\ y_{kT} \end{pmatrix}$$

gegeben. Die Funktion $f = (f_1, \ldots, f_k)'$ soll anhand der Daten mit der OLS Methode geschätzt werden. Dazu wird der folgende Ansatz verwendet.

$$\sum_{t=1}^{T} \|y_t - f(y_t)\|_2^2 \ \rightarrow \ \min$$

Dabei sei $\| \ . \ \|_2$ die euklidische Norm, sprich für $x \in \mathbb{R}^k$ gilt

$$\|x\|_2 = \|(x_1, \ldots, x_k)\|_2 = \sqrt{x_1^2 + \ldots + x_k^2} \ .$$

Es gilt:

$$\begin{aligned} \sum_{t=1}^{T} \|y_t - f(y_t)\|_2^2 &= \sum_{t=1}^{T} \left((y_{1t} - f_1(y_t))^2 + \ldots + (y_{kt} - f_k(y_t))^2 \right) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{k} (y_{it} - f_i(y_t))^2 \\ &= \sum_{i=1}^{k} \sum_{t=1}^{T} (y_{it} - f_i(y_t))^2 \ . \end{aligned}$$

Sei $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_k)'$ eine OLS Schätzung von $f = (f_1 \ldots, f_k)'$. $\hat{f}$ wurde durch die OLS Methode gerade so bestimmt, dass der Ausdruck

$$\sum_{t=1}^{T} \|y_t - \hat{f}(y_t)\|_2^2$$

minimal wurde. Aus obiger Gleichungskette folgt, dass dann auch

$$\sum_{i=1}^{k}\sum_{t=1}^{T}(y_{it}-\hat{f}_i(y_t))^2$$

minimal ist. Die letzte Doppelsumme enthält durch das Quadrieren nur Summanden größer oder gleich Null. Folglich ist sie genau dann minimal, wenn alle Summanden minimal gewählt wurden. Demnach minimiert $\hat{f}_i$ die Summe

$$\sum_{t=1}^{T}(y_{it}-f_i(y_t))^2$$

für alle $i = 1, \ldots, k$. Dies wiederum ist genau dann der Fall, wenn $\hat{f}_i$ eine OLS Schätzung der i-ten Komponente ist.

Wählt man umgekehrt für die Komponenten $i = 1, \ldots, k$ OLS Schätzungen $\hat{f}_1, \ldots, \hat{f}_k$, so ergibt sich analog, dass $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_k)'$ eine OLS Schätzung des multivariaten Modells ist. Man lese zu diesem Zweck lediglich die Gleichungskette rückwärts und beachte, dass die Summe positiver reller Zahlen genau dann minimal wird, wenn alle Summanden minimal gewählt werden.

□

3.5.3 Robuste LMS Schätzung eines Zeitreihenmodells nach Rousseeuw und Leroy

Statistische Ausreißer können die OLS Schätzung eines AR[p]-Zeitreihenmodells verzerren, wie beispielsweise Rousseeuw und Leroy zeigten. Sie schlagen daher die Verwendung eines robusten Ansatzes vor, die sogenannte LMS Methode

(Least Median of Squares) um Zeitreihenmodelle zu schätzen [vgl. RL87, S.273 ff.]. Hier wird die notwendige Notation zusammengestellt um die LMS Methode, wie sie in Abschnitt 3.1.3 vorgestellt wird, anwenden zu können.

Sei $y_1, \ldots, y_n$ eine univariate Zeitreihe, die einem AR[p] Modell folge.

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i, \;\; i = 1, \ldots, n$$

Dabei gelte $x_{i1} = y_{i-1}, \ldots, x_{ip} = y_{i-p}$. Unter Verwendung dieser Notation kann die LMS Methode aus Abschnitt 3.1.3 direkt angewendet werden, um das Zeitreihenmodell robust zu schätzen. Weitere Details zur robusten Modellschätzung von Zeitreihen finden sich bei Rousseeuw und Leroy [vgl. RL87, S.273 ff.].

3.5.4 Trends in Zeitreihen

Sowohl univariate als auch multivariate Zeitreihen zeigen oft eine systematische Änderung des mittleren Niveaus. Dies wird allgemein mit dem Begriff Trend bezeichnet, wobei deterministische (z.B. polynomiale) oder stochastische Trends unterschieden werden. Einen Überblick geben Maddala und Kim [vgl. MK98]. Mit Trends behaftete Zeitreihen erfüllen nicht die Stationaritätsbedingungen. Die in den folgenden Kapiteln vorgestellten ausreißeranalytischen Methoden setzen die Stationarität der zu untersuchenden Zeitreihen jedoch voraus. Folglich ist es wichtig, Zeitreihen vor der Anwendung dieser Verfahren von jeglichen Trends zu bereinigen. Nur so erhält man konsistente Ergebnisse.

Zur Illustration des Sachverhaltes ein kleines Beispiel. Angenommen eine univariate Zeitreihe $(X_t)_{t=1,\ldots,n}$ weist einen deterministischen Trend auf. Die Zeitrei-

he sei wie folgt modelliert.

$$X_t = a + bt + E_t$$

Dabei seien die Zufallsvariablen $(E_t)_{t=1,\ldots,n}$ stochastisch unabhängig und identisch verteilt mit Erwartungswert $E(E_t) = 0$ und Varianz $Var(E_t) = \sigma^2$ für alle $t = 1, \ldots, n$. Man nennt $(E_t)_{t=1,\ldots,n}$ auch Weißes Rauschen (White Noise). Demnach ist die Zeitreihe $(X_t)_{t=1,\ldots,n}$ die Summe aus einer linearen Trendfunktion mit Anstieg b und einem Weißen Rauschen. Die Zufallskomponente der Zeitreihe ist allein durch $(E_t)_{t=1,\ldots,n}$ gegeben, wohingegen $a+bt$ deterministisch ist. Weder ist die lineare Funktion $f(t) = a + bt$ durch Ausreißer kontaminiert, noch ist es das Weiße Rauschen. Wendet man nun ausreißeranalytische Methoden auf eine Realisation $(x_t)_{t=1,\ldots,n}$ der Zeitreihe an, ohne zuvor den deterministischen Trend zu eliminieren, so wäre es möglich, dass x_1 und/oder x_n oder gar mehr Datenpunkte als Ausreißer identifiziert werden, wenn diese nur weit genug vom mittleren Niveau der Zeitreihe entfernt sind. Die hier vorgestellten ausreißeranalytischen Methoden setzen die Stationarität der Zeitreihe voraus und damit insbesondere, dass ihr mittleres Niveau konstant bleibt (sog. Mittelwertstationarität). Liegt ein Trend vor, so ändert sich das mittlere Niveau der Zeitreihe. Die Ergebnisse der Ausreißeranalytik werden durch den vorhandenen Trend unbrauchbar.

Verschiedene Ansätze zur Trendschätzung wurden unter anderem von Bianchi et al. [BBH99, vgl.], Meade [Mea00, vgl.], Busse et al. [BJK94, vgl.], Visser [Vis04, vgl.], Pollock [Pol01, vgl.], Bhattacharyya [Bha67, vgl.], El-Shaarawi [ES95, vgl.], Huber und Ronchetti [HR09, vgl.] sowie Rousseeuw und Leroy [vgl. RL87] diskutiert. Letztere verwenden die LMS Methode, um eine Trendgerade

mittels linearer Regression zu schätzen. Diese Methodik wird hier verwendet, um deterministische lineare Trends zu schätzen. Wie die LMS Methode konkret arbeitet, wird in Abschnitt 3.1.3 erläutert.

3.5.4.1 Deterministische Trends

Deterministische lineare Trends werden im Zeitreihenmodell durch eine lineare Abhängigkeit des mittleren Niveaus von der Zeit dargestellt. Allgemein liegt polynomialen Trends eine polynomiale Abhängigkeit des mittleren Niveaus von der Zeit zugrunde. Die Schwankungen der Zeitreihe um den deterministischen Trend werden als VARMA Prozess modelliert.

$$X_t = \alpha + \beta t + c_t, \;\; \Phi(B)c_t = \Theta(B)u_t, \;\; u_t \sim \text{i.i.d. Zufallsvariablen}$$

X_t ist somit im Wesentlichen ein VARMA Prozess c_t, zu dem eine deterministische Funktion addiert wird, wodurch sich das mittlere Niveau der Zeitreihe in Form von $\alpha + \beta t$ linear in der Zeit ändert. Die Koeffizientenpolynome $\Phi(B)$ und $\Theta(B)$ seien invertierbar und c_t stationär. Allgemein setzt man für polynomiale Trends statt $\alpha + \beta t$ ein Polynom $p(t)$ an. Das langfristige Verhalten der Zeitreihe entspricht im Wesentlichen dem Trendverlauf, sprich $p(t)$ und der Vorhersagefehler bei langfristigem Prozessverlauf entspricht gerade c_t, dazu siehe auch [NP82, S.141-142]. Weitere Hintergründe zu deterministischen Trends werden von Lütkepohl diskutiert [vgl. Lüt05, S.299 ff.]. Dort wird empfohlen multivariate deterministische Trends mittels Bildung erster Differenzen zu entfernen, d.h. aus der Zeitreihe $(X_t)_{t=1,\ldots,n}$ wird die Zeitreihe $(\Delta X_t)_{t=1,\ldots,n-1} = (X_{t+1} - X_t)_{t=1,\ldots,n-1}$ gebildet. Diese ist trendbereinigt und kann für die weiteren statistischen Analysen

verwendet werden. Welchen Einfluss die Bildung erster Differenzen auf Ausreißer hat, wird in Abschnitt 3.5.6.5 erläutert.

3.5.4.2 Stochastische Trends

Sei $(X_t)_{t=1,\ldots,T}$ ein VARMA Prozess mit $X_t = (X_{t1}, \ldots, X_{tp})'$ und $(Y_t)_{t=1,\ldots,T-1}$ der Prozess der ersten Differenzen von $(X_t)_{t=1,\ldots,T}$, sprich $Y_t = X_{t+1} - X_t$. Wenn $(Y_t)_{t=1,\ldots,T-1}$ stationär ist, so nennt man $(X_t)_{t=1,\ldots,T}$ differenzstationär von der Ordnung 1 oder kurz differenzstationär.

Folgt eine multivariate differenzstationäre Zeitreihe $(X_t)_{t=1,\ldots,T}$ beispielsweise einem AR[1]-Modell mit der Darstellung

$$X_t - X_{t-1} = E_t \text{ bzw. } (I - IB)X_t = E_t \ ,$$

so gilt offenbar $\Pi(B) = I - IB$ und das charakteristische Polynom $\kappa(z) := det(\Pi(z)) = det(I - Iz)$ besitzt die Nullstelle $z_0 = 1$, wobei det die Determinante bezeichnet. Man sagt auch „die Zeitreihe besitzt eine Einheitswurzel (unit root)“. Falls $det(\Pi(z))$ d Nullstellen $z_1 = \ldots = z_d = 1$ besitzt ($d \in \mathbb{N}$) und für alle weiteren Nullstellen z_i gilt $|z_i| > 1$, so nennt man $(X_t)_{t=1,\ldots,T}$ integriert von der Ordnung d, kurz $I(d)$ [vgl. Lüt05, S.243]. Weitere Details und Hintergründe finden sich bei Lütkepohl [vgl. Lüt05, S.238 ff.] und Lütkepohl [Lüt93].

Interessant ist nun die Frage, wie stochastische Trends eliminiert werden können und welchen Einfluss derartige Methoden auf die multivariate Struktur der Zeitreihe besitzen. Es ist möglich durch sukzessive Bildung erster Differenzen eine mit stochastischen Trends behaftete Zeitreihe in eine stationäre Zeitreihe zu überführen, siehe Lütkepohl [vgl. Lüt05, S.243] und Galeano et al. [vgl. GPT06,

S.660 f.]. Wie Lütkepohl weiter zeigt, ist es nicht immer zwingend notwendig exakt d-mal erste Differenzen zu bilden um Stationarität zu erreichen, wenn die Zeitreihe $(X_t)_{t=1,\dots,T}$ integriert von der Ordnung d ist. Es kann genügen d^*-mal erste Differenzen zu bilden ($d^* < d$, $d^*, d \in \mathbb{N}$). Die Bildung erster Differenzen kann jedoch die multivariate Struktur einer Zeitreihe verändern. Dies ist häufig dann der Fall, wenn die Zeitreihe cointegriert ist [vgl. Lüt05, S.244]. Die Cointegration einer Zeitreihe $(X_t)_{t=1,\dots,T}$ mit $X_t = (X_{t1}, \dots, X_{tp})'$ besagt, dass ein $\beta \in \mathbb{R}^p \backslash \{0_p\}$ existiert, so dass

$$\beta' X_t = z_t$$

stationär ist. Das Prinzip der Cointegration wird in Abschnitt 3.5.4.3 näher erläutert. Es ist somit sicherzustellen, dass keine Cointegration vorliegt, wenn erste Differenzen gebildet werden sollen [vgl. Lüt05, S.244].

Weiterhin stellt sich die Frage, wie sich das Modell mit stochastischem Trend vom Modell mit deterministischem Trend unterscheidet. Ist es überhaupt notwendig diese beiden Modelle zu betrachten? Der Unterschied lässt sich leicht illustrieren, wenn eine univariate Zeitreihe $(X_t)_{t=1\dots,T}$ mit $d = 1$ gewählt wird. Dann besitzt das Modell mit stochastischem Trend die Darstellung

$$(1 - B)X_t = \beta + \xi(B)^{-1}\zeta(B)u_t \text{ bzw. } X_t = X_{t-1} + \beta + \xi(B)^{-1}\zeta(B)u_t \ ,$$

wobei $\xi(B)$ und $\zeta(B)$ Koeffizientenpolynome sind und B ist der Backshift Operator. Der Zustand des Prozesses zum Zeitpunkt t_1 nach einem fest gewählten Zeitpunkt $t_0 < t_1$, ergibt sich aus $X_{t_1} = X_{t_0} + \beta(t_1 - t_0) + \sum_{t=t_0+1}^{t_1} \xi(B)^{-1}\zeta(B)u_t$. Man erkennt leicht, dass mit wachsendem Zeithorizont ($t_1 \to \infty$) der Vorhersa-

gefehler unbeschränkt wächst.

$$\lim_{t_1\to\infty} Var\left(\sum_{t=t_0+1}^{t_1} u_t\right) = \lim_{t_1\to\infty} \sum_{t=t_0+1}^{t_1} Var(u_t) = \lim_{t_1\to\infty} \sum_{t=t_0+1}^{t_1} \sigma_u^2$$
$$= \lim_{t_1\to\infty} (t_1 - t_0)\sigma_u^2 = \infty$$

3.5.4.3 Cointegration

Cointegration beschreibt eine spezielle Art der Abhängigkeit zwischen den Komponenten multivariater Zeitreihen. Um dies besser zu verstehen, stelle man sich eine bivariate Zeitreihe vor. Angenommen die beiden Komponenten der Zeitreihe befänden sich zu jedem Zeitpunkt $t = 1, \ldots, n$ annähernd in einer Art Gleichgewicht. Derartige Muster werden häufig in ökonomischen Zeitreihen, z.B. zwischen dem verfügbaren Einkommen und der Konsumrate, beobachtet. In der Umwelt wäre ein Beispiel die Anzahl der Wölfe in einer Region und parallel dazu die Anzahl der Beutetiere. Beide Variablen befinden sich stets in einem annähernden Gleichgewicht: je mehr Beutetiere es gibt, desto mehr Wölfe gibt es, denn ein ausreichendes Nahrungsangebot sichert das Überleben der Nachkommen. Nimmt die Zahl der Wölfe überhand, so gibt es nicht mehr genügend Beutetiere und die Zahl der Wölfe dezimiert sich wieder. Es herrscht eine Art Gleichgewicht. Dieses Gleichgewicht lässt sich mathematisch modellieren, z.B. durch die Bildung einer Linearkombination der beiden Komponenten „Wölfe“ und „Beutetiere“.

$$\beta_1 \cdot Wölfe + \beta_2 \cdot Beutetiere = 0, \;\; (\beta_1, \beta_2) \in \mathbb{R}^2 \backslash \{0_2\}$$

Allgemein sei $(X_t)_{t=1,\ldots,n}$ mit $X_t = (X_{t1}, \ldots, X_{tp})'$ eine multivariate Zeitreihe und $\beta \in \mathbb{R}^p \backslash \{0_p\}$, dann ist die Gleichgewichtsbeziehung definiert durch

$$\beta' X_t = 0 \ .$$

Nun wird es in der Realität nicht so sein, dass dieses Gleichgewicht immer exakt eingehalten wird. Vielmehr beobachtet man die zufälligen Abweichungen vom Gleichgewicht. Diese Überlegung führt zum stochastischen Modell

$$\beta' X_t = z_t \ , \tag{3.16}$$

wobei z_t eine univariate stationäre Zeitreihe sei. Die Abweichungen vom Gleichgewicht werden somit als stationäre Zeitreihe modelliert [vgl. Lüt05, S.244 f]. Das Konzept der Cointegration geht ursprünglich zurück auf Granger [vgl. Gra81, GE83] und Granger und Engle [vgl. EG87]. Johansen entwickelte Methoden, um einen Cointegrationsvektor $\beta \in \mathbb{R}^p$ zu schätzen und anschließend den nach ihm benannten Johansen-Test, um das Testproblem

$$H_0 : \ \beta = 0_p \ \text{ gegen } \ H_1 : \ \beta \neq 0$$

zu untersuchen [vgl. Joh88, Joh91]. Eine weiterführende Theorie von der Cointegration hin zur Codependence wird von Trenkler und Weber entwickelt [vgl. TW10].

Wichtig anzumerken ist, dass im Modell 3.16 die Komponenten der multivariaten Zeitreihe $(X_t)_{t=1,\ldots,n}$ Trends aufweisen dürfen. Prinzipiell wäre es somit notwendig eine trendbehaftete multivariate Zeitreihe auf Cointegration zu über-

prüfen [vgl. Lüt05]. Dies ist für die hier untersuchten Daten jedoch vernachlässigbar. Der Beobachtungszeitraum der Daten beträgt ca. 20 Jahre. Das Modell der Cointegration beschreibt eine gemeinsame langfristige Entwicklung der Komponenten multivariater Zeitreihen [vgl. DGM99] [vgl. Lüt05, S.235]. Die vorliegenden Daten besitzen somit keinen langfristigen Beobachtungshorizont. Weiterhin kann kein zwingendes sachlogisches Gleichgewicht zwischen den Komponenten der Zeitreihen abgeleitet werden. Zwar deuten Daten aus Bodenproben darauf hin, dass beispielsweise die Konzentrationen von PCB 28 und PCB 52 (sog. niedrigchlorierte PCB) oft geringer zu sein scheinen als die entsprechenden Konzentrationen von PCB 101 und PCB 138 (sog. hochchlorierte PCB), jedoch kann nicht davon ausgegangen werden, dass sich die Verhältnisse der Konzentrationen tatsächlich in einer Art Gleichgewicht im Sinne eines stationären Prozesses befinden [vgl. JM06, S.143 Abbildung 3-62 und S.122 Abb. 3-46] [vgl. Umw11]. Es kann somit weder sachlogisch, noch aus einem großen Beobachtungshorizont der Daten heraus auf Cointegration geschlossen werden. Die Entfernung multivariater Trends ohne Berücksichtigung von Cointegration wird daher in dieser Arbeit als hinreichend genau erachtet.

3.5.5 Test auf Trends in Zeitreihen

In den 1970er Jahren entwickelten Dickey und Fuller ein statistisches Verfahren, um Zeitreihen auf die Präsenz von stochastischen Trends zu testen [vgl. DF79]. Nelson und Plosser [vgl. NP82] analysierten 14 makroökonomische Zeitreihen und konnten die Nullhypothese, dass ein stochastischer Trend vorliegt, nur in einem Fall signifikant ablehnen. Der Test von Dickey und Fuller wurde von Perron

weiterentwickelt [vgl. Per89]. Er integrierte mittels Dummyvariablen die einmalige Änderung des mittleren Niveaus und die einmalige Änderung des Trends im Zeitreihenmodell. Damit war es ihm nunmehr möglich für 11 der 14 Zeitreihen Perrons die Nullhypothese der Existenz eines stochastischen Trends signifikant abzulehnen. DeJong und Whiteman zeigten, dass der Dickey-Fuller Test eine geringe Power aufweist, wenn ein autoregressives Modell mit Einheitswurzeln nahe 1 vorliegt [vgl. DW91]. Darüber hinaus untersuchten Diebold und Rudebusch die Eigenschaften des Dickey-Fuller Tests hinsichtlich der statistischen Power [vgl. DR91, S.158 ff.]. Sie kamen zu dem Schluss, dass bei geringen Stichprobengrößen ($n < 50$, [vgl. DR91]) eine geringe Power gegenüber Alternativhypothesen, wie die Existenz fraktionell integrierter Störgrößen mit langem Gedächtnis, vorliegt. Gomez und Ventosa-Santaulària entwickelten einen Test, um die Nullhypothese der Existenz eines stochastischen Trends gegen die Alternativhypothese der Existenz sowohl eines stochastischen als auch eines deterministischen Trends zu überprüfen [vgl. GZVS10]. Kurz nachdem Saïd E. Saïd. einen Test veröffentlichte, um die Existenz eines stochastischen Trends in Gegenwart eines deterministischen Trends nachzuweisen [vgl. Sa1], schlugen Schmidt und Phillips die Verwendung eines LM-Tests zur Überprüfung der Nullhypothese der Differenzstationarität gegen die Alternativhypothese der Existenz eines deterministischen linearen Trends vor [vgl. SP92]. Der von Kwiatkowski et al. vorgestellte KPSS Test ist eine Modifikation des LM-Tests und dient dazu, die Nullhypothese der Existenz eines deterministischen Trends gegen die Alternativhypothese der Differenzstationarität zu testen [vgl. KPSS92]. Giraitis et al. entwickelten einen Test zur Überprüfung der Nullhypothese der Stationarität gegen die Alternativhypothese, dass ein stochastischer oder deterministischer Trend vorliegt [vgl. GLP06]. Das

Problem einer möglichen Trendänderung wurde von Vogelsang untersucht. Dieser schlug verschiedene Teststatistiken vor, um Änderungen in der Trendfunktion einer Zeitreihe zu detektieren [vgl. Vog97]. Die Änderung des Trendverhaltens einer Zeitreihe wird von Galeano et al. als sogenannter „ramp shift“ bezeichnet und stellt eine besondere Art von statistischem Ausreißer, ähnlich einer Änderung des mittleren Niveaus einer Zeitreihe (sog. „level shift“), dar [vgl. GPT06, S.660f.].

Wie genau weist man in einer gegebenen Zeitreihe nach, dass diese trendbehaftet ist und welche Art Trend vorliegt? In der statistischen Literatur wird empfohlen mehrere Tests auf Stationarität gegen die verschiedenen Alternativhypothesen „linearer Trend“ und „stochastischer Trend“ durchzuführen. Dieses multiple Testen wirft das Problem der $\alpha-$Fehlerinflation auf. Verwendet man zu viele Tests, läuft man durch eine schrittweise Absenkung des Signifikanzniveaus α schnell Gefahr keine signifikanten Aussagen mehr treffen zu können. Auch ist die Reihenfolge der Tests entscheidend. An dieser Stelle wird eine andere Vorgehensweise propagiert. Anstatt mehrere Tests gegen die Nullhypothese der Stationarität der Zeitreihe durchzuführen, betrachte man zunächst das folgende Testproblem.

H_0 : „Die Zeitreihe ist stationär.“

H_1 : „Die Zeitreihe weist entweder einen linearen deterministischen Trend auf, oder sie ist differenzstationär.“

Ist das Ergebnis signifikant, so liegt ein Trend vor. Man teste anschließend, welche der Alternativen (linearer Trend oder stochastischer Trend oder eventuell beides)

zutrifft. Die folgenden beiden Testprobleme sind dazu geeignet.

H_{01} : „Die Zeitreihe weist einen linearen deterministischen Trend auf."

H_{11} : „Die Zeitreihe ist differenzstationär."

Wird H_{01} abgelehnt, so liegt ein stochastischer Trend vor und ein deterministischer Trend kann ausgeschlossen werden. Wird H_{01} nicht abgelehnt, so liegt ein deterministischer Trend vor und möglicherweise zusätzlich ein stochastischer Trend. Besteht ein dringender Verdacht, dass ein stochastischer Trend vorliegt, so betrachte weiterhin das folgende Testproblem.

H_{02} : „Die Zeitreihe ist differenzstationär."

H_{12} : „Die Zeitreihe weist einen linearen deterministischen Trend auf."

Mittels dieser Prozedur werden alle Möglichkeiten gegeneinander getestet und der Testaufwand so gering wie möglich gehalten. Das Problem der $\alpha-$Inflation ist in einem praktikablen Rahmen abwendbar, da lediglich drei statistische Tests angewendet werden.

In dem Datenpool, der dieser Arbeit zugrunde liegt, treten stets nur uni- und multivariate Zeitreihen mit ca. 10-20 Beobachtungen auf. Daten mit einem derart kurzen Zeithorizont von maximal 20 Beobachtungen lassen es kaum zu, dass eventuell höhergradigere polynomiale Trends erkannt werden. Es gilt die Frage zu beantworten, ob die Entwicklung der Zeitreihe abseits des geschätzten linearen Trends zufällig oder systematisch ist? Aufgrund des kurzen Beobachtungszeitrau-

mes ist diese Frage nicht zufriedenstellend zu beantworten. Andererseits ist der Beobachtungszeitraum so kurz, dass die lineare Approximation eines unbekannten polynomialen Trends hinreichend genau ist, da trendbehaftete Zeitreihe lediglich stationarisiert werden sollen. Prognosen anhand der Trends werden hingegen nicht angestrebt.

3.5.5.1 Test auf die Existenz eines Trends nach Giraitis et al.

Giraitis et al. [vgl. GLP06] entwickelten ein statistisches Verfahren zum Test der Hypothesen

H_0 : „Die Zeitreihe ist stationär."

H_1 : „Die Zeitreihe weist entweder einen linearen deterministischen Trend auf, oder sie ist differenzstationär."

Unter der Nullhypothese folgt die Zeitreihe Y_t dem Modell

$$Y_t = \beta_0 + \xi_t,$$

wobei ξ_t ein stationärer Prozess mit $E(\xi_t) = 0$ und Autokovarianzfunktion $\gamma(\tau)$ ist. An den Prozess ξ_t werden allgemeine Bedingungen gestellt. Er kann ein kurzes, ein langes oder sogar ein negatives Gedächtnis aufweisen. Die zugehörige Autokovarianzfunktion γ muss in jedem dieser drei Fälle gewisse Regularitätsbe-

dingungen erfüllen.

$$\text{(A)} \sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| < \infty$$

$$\text{(B)} \; \gamma(\tau) \sim c\tau^{2d-1}, \;\; c > 0, \;\; 0 < d < 0.5$$

$$\text{(C)} \; \gamma(\tau) \sim c\tau^{2d-1}, \;\; \sum_{\tau=-\infty}^{\infty} \gamma(\tau) = 0, \;\; -0.5 < d < 0$$

Eigenschaft (A) besagt, dass der Prozess ξ_t ein kurzes Gedächtnis besitzt. Eigenschaft (B) besagt, dass ξ_t ein langes Gedächtnis besitzt und das die Autokovarianzfunktion proportional zu $c\tau^{2d-1}$ ist. Unter der Eigenschaft (C) besitzt der Prozess ξ_t ein negatives Gedächtnis. Dabei muss sich die Summe aller Autokovarianzen zu Null aufheben und die Autokovarianzfunktion muss proportional zu $c\tau^{2d-1}$ sein. Weiterhin habe der Prozess ξ_t eine Darstellung der Art

$$\xi_t = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j} \;\; .$$

Die $(\varepsilon_j)_{j\in\mathbb{Z}}$ seien stochastisch unabhängige Zufallsvariablen mit $E(\varepsilon_j) = 0$, $Var(\varepsilon_j) = 1$ und $E(\varepsilon_j^4) < \infty$. An die Gewichte a_j werden ähnliche Regularitätsbedingungen gestellt, wie an die Autokovarianzfunktion $\gamma(\tau)$.

$$\text{(A')} \sum_{j=-\infty}^{\infty} |a_j| < \infty$$

$$\text{(B')} \; a_j \sim cj^{d-1}, \;\; c \neq 0, \;\; 0 < d < 0.5$$

$$\text{(C')} \; a_j \sim cj^{d-1}, \;\; \sum_{j=-\infty}^{\infty} a_j = 0, \;\; -0.5 < d < 0$$

Wenn Bedingung (A') gilt, so gilt auch Bedingung (A). Wenn Bedingung (B') gilt, so ist Bedingung (B) erfüllt. Analog folgt aus der Gültigkeit von Bedingung (C') die Gültigkeit von Bedingung (C). Die Forderungen an die Gewichte a_j sind damit rein formaler Natur um sicher zu stellen, dass alle Regularitätsbedingungen für die Autokovarianzfunktion erfüllt sind.

Unter der Alternativhypothese hat die Zeitreihe Y_t entweder eine Darstellung mittels eines deterministischen Trends

$$Y_t = \beta_0 + g_n(t) + \xi_t$$

oder einer Einheitswurzel

$$Y_t - Y_{t-1} = \beta_0 + g_n(t) + \xi_t \ .$$

Die Funktion $g_n(t)$ modelliert in diesem Zusammenhang den deterministischen Trend. Sie kann sowohl einen polynomialen Trend $g_n(t) = ct^\beta$ ($c \neq 0,\ \beta > -0.5$), eine nichtparametrische Regression der Form $g_n(t) = g(t/n)$ ($t = 1, \ldots, n$) als auch einen Strukturbruch modellieren.

Die Teststatistik $T_n(\hat{d})$ nach Giraitis ist gegeben durch Formel (3.17) in Abschnitt 3.5.5.1.1. Kritische Werte $c_\alpha(d)$ sind in Tabelle 3.3 gegeben. Diese können auch numerisch berechnet werden. Giraitis et al. stellen zu diesem Zweck für $\alpha = 0.05$ und $\alpha = 0.1$ Polynome zur Approximation der kritischen Werte für

nicht in Tabelle 3.3 gegebene d zur Verfügung.

$$c_{0.05}(d) = -1.98d^5 + 0.73d^4 - 0.05d^3 + 0.63d^2 - 0.66d + 0.19$$

$$c_{0.10}(d) = -0.27d^5 + 0.48d^4 - 0.55d^3 + 0.66d^2 - 0.53d + 0.14$$

d	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
$c_{0.05}(d)$	0.597	0.457	0.349	0.262	0.19	0.13	0.083	0.049	0.022
$c_{0.10}(d)$	0.524	0.393	0.292	0.213	0.153	0.105	0.067	0.039	0.016

Tabelle 3.3: Kritische Werte für die Teststatistik $T_n(\hat{d})$ [vgl. GLP06]

Die Nullhypothese wird abgelehnt, falls $T_n(\hat{d}) > c_\alpha(\hat{d})$ gilt.

Es verbleibt zu klären, wie sich die jeweiligen Faktoren der Teststatistik $T_n(\hat{d})$ errechnen. Zu diesem Zweck wird im Folgenden ein Algorithmus angegeben.

3.5.5.1.1 Algorithmus zur Bestimmung der Teststatistik $T_n(\hat{d})$.

- Setze $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- Für $j = 0, \ldots, n-1$ berechne $\hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j}(Y_i - \bar{Y})(Y_{i+j} - \bar{Y})$.
- Wähle ein ganzzahliges $q = q(n)$, so dass $\lim_{n\to\infty} q(n) = \infty$ und $\lim_{n\to\infty} \frac{q(n)}{n} = 0$ gilt, z.B. wähle q als ganzzahligen Anteil von $\sqrt{n}$ oder $\ln(n)$.
- Berechne $\hat{s}^2_{n,q}$ nach der Formel $\hat{s}^2_{n,q} = \frac{1}{q}\sum_{i=1}^{q}\sum_{j=1}^{q}\hat{\gamma}_{i-j}$.
- Setze $S^*_t = \sum_{j=1}^{t}\left(Y_j - \bar{Y}\right)$.

- Berechne V_n nach der Formel $V_n = \frac{1}{n^2}\left[\sum_{t=1}^{n}\left(S_t^*\right)^2 - \frac{1}{n}\left(\sum_{t=1}^{n} S_t^*\right)^2\right]$.
- Schätze d nach 3.5.5.1.2 [vgl. GLP06, Kün87, Rob95].
- Berechne die Teststatistik

$$T_n(\hat{d}) = \left(\frac{q}{n}\right)^{2\hat{d}} \frac{V_n}{\hat{s}_{n,q}}^2 \quad . \tag{3.17}$$

3.5.5.1.2 Die Whittle Prozedur zur Schätzung von d Um den Parameter d konsistent Schätzen zu können, empfehlen Giraitis et al. die Whittle Prozedur [GLP06, vgl.]. Dabei ist d so zu bestimmen, dass die Funktion

$$\begin{aligned} U_n(d) &= \log\left(\frac{1}{m}\sum_{j=1}^{m}\lambda_j^{2d} I_n(\lambda_j)\right) - \frac{2d}{m}\sum_{j=1}^{m}\log\lambda_j \\ &= \log\left(\frac{1}{m}\sum_{j=1}^{m} j^{2d} I_n(\lambda_j)\right) - \frac{2d}{m}\sum_{j=1}^{m}\log j \end{aligned}$$

minimiert wird. Hier ist $I_n(\lambda) := \frac{1}{2\pi n}\left|\sum_{j=1}^{n} e^{ij\lambda} Y_j\right|$ das Periodogramm und $\lambda_j = \frac{2\pi j}{n}$ $(j = 1, \ldots, \left[\frac{n-1}{2}\right])$ und $[\frac{n-1}{2}]$ der ganzzahlige Teil von $\frac{n-1}{2}$. Das Periodogramm ist leicht durch den Einsatz von Software zu bestimmen. In **R** geschieht dies am einfachsten über die Funktion `spectrum`. Der Parameter m in obiger Formel ist die Bandbreite, die bei der Schätzung des Periodogramms verwendet wird. Sie wird zusammen mit dem Periodogramm in der Funktion `spectrum` ausgegeben.

3.5.5.2 Test auf linearen deterministischen Trend gegen Differenzstationarität nach Kwiatkowski et al.

Das von Kwiatkowski et al. entwickelte Verfahren wird häufig als KPSS Test bezeichnet. Es handelt sich um eine Weiterentwicklung des von Nabeya und Tanaka propagierten LM-Tests [vgl. NT88] und untersucht für eine trendbehaftete Zeitreihe das folgende Testproblem [vgl. KPSS92].

H_{01} : „Die Zeitreihe weist einen linearen Trend auf."

H_{11} : „Die Zeitreihe ist differenzstationär."

Im Folgenden werden die Teststatistik des Verfahrens, kritische Werte und ein Algorithmus zur konkreten Umsetzung formuliert. Gegeben sei die trendbehaftete Zeitreihe $Y_1, \ldots, Y_T$. Ob eine Zeitreihe einen Trend aufweist, muss vorher mit dem Verfahren von Giraitis et al. überprüft werden [vgl. GLP06]. Das grundlegend unterstellte Zeitreihenmodell sei

$$Y_t = \xi t + r_t + \varepsilon_t,\ \ r_t = r_{t-1} + u_t,\ \ (u_t)_{t=1,\ldots,T} \text{ u.i.v.},\ E(u_t) = 0,\ Var(u_t) = \sigma_u^2\ .$$

Die Nullhypothese H_{01} ist äquivalent zur Nullhypothese H'_{01} : $\sigma_u^2 = 0$. An die Störgrößen ε_t werden allgemeine Bedingungen gestellt. Nabeyas und Tanakas Forderung der stochastischen Unabhängigkeit und identischen Verteilung der Störgrößen ε_t [vgl. NT88] ist einerseits sehr restriktiv und andererseits in der Praxis selten erfüllt. Kwiatkowski et al. zeigen, dass die Teststatistik von Nabeya und Tanaka sowohl unter den allgemeineren Bedingungen von Phillips und Perron [vgl. PP88, S.336], als auch unter Phillips' und Solos [vgl. PS89, Theoreme

3.3 und 3.14] Anforderungen an die Störgrößen ε_t eingesetzt werden kann [vgl. KPSS92, S.164ff]. Dabei sind die Bedingungen von Phillips und Solo von großer Bedeutung, da sie für die Störgrößen ε_t alle Arten von ARMA Prozessen zulassen, unabhängig davon, ob die Innovationen des Prozesses homogen oder heterogen sind. Die von Kwiatkowski et al. verwendete Modifikation des LM-Tests von Nabeya und Tanaka ist ein asymptotischer Test. Die kritischen Werte sind Tabelle 3.4 zu entnehmen.

3.5.5.2.1 Algorithmus zum Verfahren von Kwiatkowski et al.

1. Bestimme eine OLS-Schätzung des Modells $Y_t = \beta_0 + \beta_1 t$, $t = 1, \ldots, T$. Die Parameter des Modells sind β_0 und β_1. Ihre OLS-Schätzungen seien bezeichnet mit $\hat{\beta}_0$ und $\hat{\beta}_1$.

2. Bestimme die Residuen e_t des Modells aus $e_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 t$, $t = 1, \ldots, T$.

3. (a) Wenn davon auszugehen ist, dass die Störgrößen ε_t i.i.d. sind, dann setze $\hat{\sigma}_\varepsilon^2 = \frac{1}{T}\sum_{t=1}^{T} e_t$ und gehe zu Schritt 4, sonst gehe zu Schritt 3.b).

   (b) Bestimme analog zu Kwiatkowski et al. l_k als den ganzzahligen Anteil von $k[T/100]^{1/4}$ für $k = 1, \ldots, 12$ [vgl. KPSS92, S.169f.]. Setze anschließend $l = l_k$ für ein $k \in \{1, \ldots, 12\}$ und gehe zu Schritt 4.

   Bemerkung: Je größer l gewählt wird, desto kleiner ist die statistische Power des Tests. Andererseits gilt: je größer l gewählt wird, desto mehr Werte gehen in die Schätzung von $s^2(l)$ in Schritt 4 ein, was diese verlässlicher macht. Dem Beispiel von Kwiatkowski et al. folgend

wird empfohlen, ein $l \in \{l_1, \ldots, l_{12}\}$ zu wählen, das möglichst mittig in der Menge angesiedelt ist.

4. Bestimme $s^2(l)$ nach der Formel

$$s^2(l) = \frac{1}{T}\sum_{i=1}^{T} e_t^2 + \frac{2}{T}\sum_{r=1}^{l} w(r,l) \sum_{t=r+1}^{T} e_t e_{t-r} \ .$$

Die Funktion $w(r,l)$ ist das sogenannte Bartlett-Fenster mit $w(r,l) = 1 - r/(l+1)$

5. Bestimme die Folge der Partialsummen der Residuen.

$$S_t = \sum_{i=1}^{t} e_i, \ \ t = 1, \ldots, T$$

6. Bestimme den Wert der Teststatistik.

$$\hat{\eta}_\tau = \frac{1}{T^2}\sum_{t=1}^{T} S_t^2 / s^2(l)$$

7. Vergleiche $\hat{\eta}_\tau$ mit dem kritischen Wert der Teststatistik η_α zum Signifikanzniveau α. Gilt $\hat{\eta}_\tau > \eta_\alpha$, so liegt ein signifikantes statistisches Testergebnis vor und die Nullhypothese wird abgelehnt.

α	0.1	0.05	0.025	0.01
η_α	0.119	0.146	0.176	0.216

Tabelle 3.4: Kritische Werte für die Teststatistik $\hat{\eta}_\tau$ [vgl. KPSS92, S.166]

3.5.5.3 Test auf Differenzstationarität gegen linearen deterministischen Trend nach Schmidt und Phillips

Schmidt und Phillips entwickelten einen LM-Test, um für eine gegebene Zeitreihe das statistische Testproblem

H_{02} : „Die Zeitreihe ist differenzstationär."

H_{12} : „Die Zeitreihe weist einen linearen deterministischen Trend auf."

zu bearbeiten [SP92]. Das zugrunde liegende Modell lautet

$$y_t = \psi + \xi t + X_t, \;\; X_t = \beta X_{t-1} + \varepsilon_t \;\; . \tag{3.18}$$

Die Forderungen an die Innovationen bzw. Störgrößen ε_t in Modell 3.18 entsprechen denen von Phillips und Perron [vgl. PP88, S.336]. Zwar schränken diese den zulässigen Grad an Heterogenität und Autokorrelation in den ε_t ein, sie sind aber weithin allgemein gehalten und lassen verschiedene Arten von Störgrößen für das Modell zu. Unter den Bedingungen von Phillips und Perron ist der Test von Schmidt und Phillips asymptotisch konsistent [vgl. SP92, S.263ff]. Eine Erweiterung der Alternativhypothese auf höhergradige polynomiale Trends ist möglich [vgl. SP92, S.267ff]. Ein Algorithmus zur Bestimmung der Teststatistik wird angegeben. Die kritischen Werte der Teststatistik sind Tabelle 3.5 zu entnehmen.

3.5.5.3.1 Algorithmus zum Verfahren von Schmidt und Phillips

1. Setze $\tilde{\xi} = \frac{y_T - y_1}{T-1}$ und $\tilde{\psi} = y_1 - \tilde{\xi}$.
2. Für $t = 1, \ldots, T$ bestimme $\tilde{S}_t = y_t - \tilde{\psi} - \tilde{\xi} t \;\; (\tilde{S}_1 = \tilde{S}_T = 0)$.

3. Bestimme eine OLS Schätzung des Modells

$$\Delta y_t = \phi_0 + \phi \tilde{S}_{t-1} + \nu_t, \;\; (\nu_t)_{t=2,\ldots,T} \text{ i.i.d. Störgrößen, } \Delta y_t = y_t - y_{t-1} \;\; .$$

Die OLS Schätzung von ϕ sei bezeichnet mit $\tilde{\phi}$.

4. Setze $\tilde{\rho} = T\tilde{\phi}$.

5. Bestimme eine OLS Schätzung des Modells

$$y_t = \alpha + \beta y_{t-1} + \delta t + \varepsilon_t, \;\; (\varepsilon_t)_{t=1,\ldots,T} \text{ i.i.d. Störgrößen} \;\; .$$

Die Modellparameter sind α, β und δ. Ihre OLS Schätzungen seien bezeichnet mit $\hat{\alpha}, \hat{\beta}$ und $\hat{\delta}$. Bestimme aus diesen die zugehörigen Modellresiduen $\hat{\varepsilon}_t = y_t - \hat{\alpha} - \hat{\beta} y_{t-1} - \hat{\delta} t$ $(t = 2, \ldots, T)$.

6. Bestimme $s^2 = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t^2$.

7. Wähle ein l, so dass $l \to \infty$ wenn $T \to \infty$. Schmidt und Phillips empfehlen $l \in O(T^{\frac{1}{2}})$. Man wähle z.B. l als ganzzahligen Anteil von $\sqrt{T}$. Im Anschluss bestimme

$$s^2(l) = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t^2 + \frac{2}{T} \sum_{s=1}^{l} \sum_{t=s+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-s} \;\; .$$

8. Setze $\hat{\omega}^2 = \frac{s^2}{s^2(l)}$.

9. Bestimme den Wert der Teststatistik $Z(\rho) = \frac{\tilde{\rho}}{\hat{\omega}^2}$.

T	0.01	0.025	0.05	0.1	0.2	0.3	0.4	0.5
25	-20.4	-17.9	-15.7	-13.4	-10.9	-9.27	-8.02	-6.98
50	-22.8	-19.6	-17	-14.3	-11.4	-9.62	-8.26	-7.12
100	-23.8	-20.4	-17.5	-14.6	-11.6	-9.76	-8.34	-7.17
200	-24.8	-20.9	-17.9	-14.9	-11.8	-9.86	-8.42	-7.23
500	-25.3	-21.3	-18.1	-15	-11.8	-9.78	-8.32	-7.15
1000	-25.3	-21.3	-18.1	-15	-11.8	-9.78	-8.32	-7.15
2000	-25.2	-21.2	-18.1	-15	-11.8	-9.85	8.39	-7.19

T	0.6	0.7	0.8	0.9	0.95	0.975	0.99
25	-6.05	-5.21	-4.35	-3.43	-2.89	-2.51	-2.16
50	-6.14	-5.24	-4.35	-3.37	-2.78	-2.39	-2.03
100	-6.17	-5.23	-4.32	-3.35	-2.75	-2.32	-1.92
200	-6.2	-5.22	-4.3	-3.31	-2.72	-2.3	-1.89
500	-6.14	-5.19	-4.29	-3.31	-2.7	-2.28	-1.9
1000	-6.14	-5.19	-4.29	-3.31	-2.7	-2.28	-1.9
2000	-6.15	-5.21	-4.3	-3.32	-2.68	-2.27	-1.87

Tabelle 3.5: Kritische Werte für die Teststatistik $Z(\rho)$ [vgl. SP92, S.264]

3.5.6 Entfernung von Trends aus Zeitreihen

Real beobachtete Zeitreihen, seien sie uni- oder multivariat, sind häufig nicht stationär. Sie zeigen beispielsweise einen deutlichen linearen oder polynomialen Trend. Da die vorgestellten Methoden allesamt für stationäre Zeitreihen entwickelt wurden, muss eine nicht stationäre Zeitreihe trendbereinigt werden. Die Notwendigkeit der Identifikation und Bereinigung von Trends wurde bereits zu Beginn des Abschnittes 3.5.4 anhand eines einfachen Beispiels erläutert. Hier wird nun algorithmisch dargelegt, wie eine Trendbereinigung für uni- und multivariate Zeitreihen vorgenommen werden kann und welchen Einfluss die Trendbereinigung auf statistische Ausreißer hat.

Bevor Trends aus Zeitreihen entfernt werden können, ist es zunächst notwendig herauszufinden, um welche Art von Trend es sich handelt. Liegt ein linearer deterministischer Trend vor, oder ist die Zeitreihe differenzstationär? Diese Frage muss für multivariate Zeitreihen in jeder Komponente geklärt sein. Eine Zeitreihe $Y_1, \ldots, Y_T$ ist genau dann differenzstationär, wenn ihre ersten Differenzen $Y_2 - Y_1, Y_3 - Y_2, \ldots, Y_T - Y_{T-1}$ eine stationäre Zeitreihe bilden.

Sei eine multivariate Zeitreihe gegeben.

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{k1} \end{pmatrix}, \begin{pmatrix} Y_{12} \\ \vdots \\ Y_{k2} \end{pmatrix} \ldots, \begin{pmatrix} Y_{1T} \\ \vdots \\ Y_{kT} \end{pmatrix}$$

Diese weise in eventuell mehreren Komponenten einen deterministischen linearen Trend auf. Weiterhin kann die multivariate Zeitreihe in mindestens einer Komponente differenzstationär sein. Beispielsweise könnte die Zeitreihe in der zweiten Komponente einen deterministischen linearen Trend aufweisen, sprich in der Reihe $Y_{21}, \ldots, Y_{2T}$, und die Reihen $Y_{51}, \ldots, Y_{5T}$ und $Y_{61}, \ldots, Y_{6T}$ könnten differenzstationär sein. In den restlichen Komponenten sei die Zeitreihe stationär.

Um die uni- und multivariate Ausreißeranalytik anwenden zu können, muss die Stationarität der zu untersuchenden Zeitreihe gewährleistet sein. Zu diesem Zweck müssen alle vorhandenen Trends eliminiert wurden. Geschieht dies nicht, ist die Voraussetzung der Stationarität (sog. Mittelwertstationarität) nicht erfüllt. Die Anwendung der in den Abschnitten 3.6 und 3.7 vorgestellten Ausreißeranalytik liefert dann keine verlässlichen Ergebnisse.

3.5.6.0.2 Beispiel Zur Illustration ist in Tabelle 3.6 ein Beispieldatensatz, bestehend aus einer multivariaten Zeitreihe mit Dimension $k = 2$ und $T = 20$ Beobachtungen, gegeben.

Beob.nr.	Merkmal 1	Merkmal 2
1	2.55	2.43
2	1.87	1.47
3	1.68	2.32
4	0.74	1.55
5	-0.39	1.16
6	-0.18	2.83
7	0.66	2.86
8	1.24	5.13
9	1.23	5.14
10	-1.02	3.27
11	-1.39	5.89
12	0.26	5.95
13	1.14	7.44
14	-1.02	5.04
15	0.02	6.94
16	1.51	6.75
17	0.58	8.31
18	1.54	9.32
19	0.55	8.55
20	-1.35	11.52

Tabelle 3.6: Bivariate Zeitreihe mit Trend in der zweiten Komponente

Die graphische Darstellung der Zeitreihe in Abbildung 3.10 zeigt deutlich einen Trend für die zweite Komponente. Die erste Komponente weist in der Graphik keinen Trend auf. Der Einsatz statistischer Tests ist jedoch zwingend notwendig, um eventuelle Trends zu detektieren oder auszuschließen. Welche statistischen Tests zur Identifikation von Trends zur Verfügung stehen, wird in Abschnitt 3.5.5 erläutert.

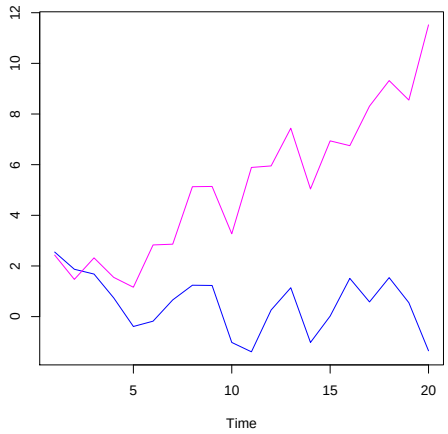

Abbildung 3.10: Bivariate Zeitreihe mit Trend in der zweiten Komponente

3.5.6.1 Algorithmus zur Elimination von Trends in univariaten Zeitreihen

1. Wende das Verfahren aus Abschnitt 3.5.5.1 an. Wurde ein Trend detektiert, so gehe zu Schritt 2, ansonsten endet der Algorithmus.

2. Überprüfe mit dem Verfahren aus Abschnitt 3.5.5.3.1, ob ein signifikanter deterministischer linearer Trend vorliegt. Gehe zu Schritt 3.

3. Liegt ein deterministischer linearer Trend vor, so eliminiere diesen wie in Abschnitt 3.5.6.3 beschrieben. Gehe zu Schritt 4.

4. Überprüfe wie in Abschnitt 3.5.5.2.1 beschrieben, ob die Zeitreihe differenzstationär ist. Gehe zu Schritt 5.

5. Wurde Differenzstationarität festgestellt, so bilde die ersten Differenzen der Zeitreihe um Stationarität herzustellen, wie in Abschnitt 3.5.6.4 beschrieben.

6. Wurde weder ein deterministischer linearer Trend noch Differenzstationarität festgestellt, kann nicht entschieden werden welche Art linearer Trend vorliegt. In diesem Fall wird konventionell festgelegt, dass der in Schritt 1 gefundene Trend deterministisch ist und entsprechend Abschnitt 3.5.6.3 entfernt wird. Diese Festlegung wird vorrangig aus praktischen Gründen getroffen. Analysiert man die Zeitreihe der ersten Differenzen auf statistische Ausreißer, so ist der Effekt der Differenzbildung auf Ausreißer zu berücksichtigen, wie in Abschnitt 3.5.6.5 dargelegt wird. Die Bildung erster Differenzen kompliziert die statistische Ausreißeranalytik. Wenn aus Schritt 1 lediglich bekannt ist, dass ein Trend vorliegt, aber nicht genau bekannt ist, um welche Art Trend es sich handelt (deterministisch oder stochastisch), so besitzt sowohl die Annahme „es handelt sich um einen deterministischen Trend“, als auch die Annahme ës handelt sich um einen stochastischen Trend‘“ ihre Berechtigung. Weiter zu berücksichtigen ist die Tatsache, dass die Zeitreihen nicht zu Prognosezwecken extrapoliert werden sollen. Vielmehr wird ein Trend innerhalb des Beobachtungszeitraumes detektiert und dessen lokale Entfernung angestrebt. Die Auswirkung der Entfernung mittels Schätzung einer Trendgerade auf die weitere Entwicklung der Zeitreihe ist hier nicht von Interesse (keine Prognose von Daten). Daher wird die Annahme „es handelt sich um einen deterministischen Trend“ als praktikabel und zu ausreißeranalytischen Zwecken als wissenschaftlich vertretbar erachtet.

3.5.6.2 Algorithmus zur Elimination von Trends in multivariaten Zeitreihen

1. Wende das Verfahren aus Abschnitt 3.5.5.1 an. Wurde ein Trend detektiert, so gehe zu Schritt 2, ansonsten endet der Algorithmus.

2. Bilde die ersten Differenzen der Zeitreihe um Stationarität herzustellen, wie in Abschnitt 3.5.6.4 beschrieben.

3.5.6.3 Elimination von linearen deterministischen Trends

Das hier vorgestellte Verfahren zur Elimination deterministischer linearer Trends darf nur auf univariate Zeitreihen angewendet werden, da deterministische Trends mittels Subtraktion entfernt werden. Dies ist nur für univariate Zeitreihen zulässig, da für multivariate Zeitreihen die komponentenweise Schätzung deterministischer Trends in Verbindung mit deren Subtraktion von der Zeitreihe, eventuelle Wechselwirkungen zwischen den Komponenten (z.B. durch Cointegration, siehe Abschnitt 3.5.4.3) nicht berücksichtigen würde. Die Zeitreihe sei bezeichnet mit $(z_t)_{t=1,\ldots,T}$.

1. Setze $y_t := z_t$ für $t = 1, \ldots, T$.

2. Bestimme eine LMS Schätzung der Parameter des Modells

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \ .$$

   Dazu verwende die LMS Methode aus Abschnitt 3.1.3 [vgl. RL87, S.24]. Die LMS Schätzungen der Parameter β_0 und β_1 seien bezeichnet mit $\hat{\beta}_0$ und $\hat{\beta}_1$.

3. Entferne den geschätzten Trend mittels Subtraktion.

$$z_t := y_t - \hat{\beta}_0 - \hat{\beta}_1 t$$

3.5.6.4 Elimination von Trends bei Differenzstationarität

Differenzstationarität wird eliminiert, indem man die sogenannten ersten Differenzen bildet. Dabei werden je zwei benachbarte multivariate Beobachtungen Y_t und Y_{t+1} durch ihre Differenz ersetzt, wodurch sich eine neue Zeitreihe Z_t ergibt.

$$Z_t := Y_{t+1} - Y_t, \;\; t = 1, \ldots, n-1$$

Die Differenzenbildung dezimiert die Anzahl der Messwerte um 1, da z.B. für $T = 20$ Beobachtungen nur die 19 Differenzen $Y_2 - Y_1, Y_3 - Y_2, \ldots, Y_{20} - Y_{19}$ berechnet werden können. Beispiel 3.5.6.0.2 einer bivariaten trendbehafteten Zeitreihe ist auf Seite 135 fortgeführt. Dort werden die ersten Differenzen gebildet, um die Wirkungsweise zu illustrieren. Die zugehörigen Daten sind in Tabelle 3.7 gegeben. Wie deren Darstellung in Abbildung 3.11 zeigt, weist die Zeitreihe keinen Trend mehr in der zweiten Komponente auf.

Beob.nr.	ΔMerkmal 1	ΔMerkmal 2
1	-0.68	-0.96
2	-0.19	0.85
3	-0.94	-0.77
4	-1.13	-0.39
5	0.21	1.67
6	0.84	0.03
7	0.58	2.27
8	-0.01	0.01
9	-2.25	-1.87
10	-0.37	2.62
11	1.65	0.06
12	0.88	1.49
13	-2.16	-2.40
14	1.04	1.90
15	1.49	-0.19
16	-0.93	1.56
17	0.96	1.01
18	-0.99	-0.77
19	-1.90	2.97

Tabelle 3.7: Bivariate Zeitreihe ohne Trend in der zweiten Komponente nach Bildung erster Differenzen

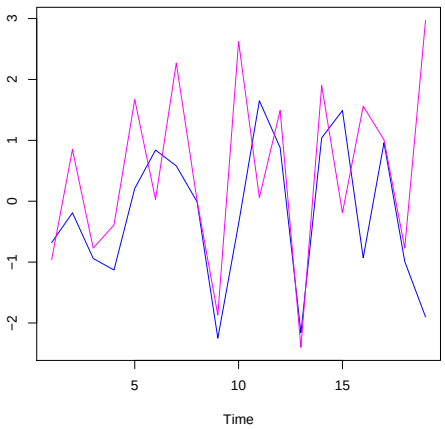

Abbildung 3.11: Bivariate Zeitreihe ohne Trend nach Bildung der ersten Differenzen

3.5.6.5 Wirkung der Differenzbildung auf Ausreißer

Wenn eine (multivariate) Zeitreihe einen Trend in einer oder in mehreren Komponenten aufweist, kann es vorkommen, dass sich dieser Trend ab einem bestimmten Zeitpunkt t_0 ändert. Beispielsweise kann eine bivariate Zeitreihe in der ersten Komponente vollkommen trendfrei sein, aber in der zweiten Komponente zunächst einen wachsenden Trend und ab dem Zeitpunkt t_0 einen fallenden Trend aufweisen. Dazu betrachte das folgende Beispiel.

Beob.nr.	Merkmal 1	Merkmal 2
1	2.55	2.43
2	1.87	1.47
3	1.68	2.32
4	0.74	1.55
5	-0.39	1.16
6	-0.18	2.83
7	0.66	2.86
8	1.24	5.13
9	1.23	5.14
10	-1.02	3.27
11	-1.39	5.56
12	0.26	6.41
13	1.14	3.36
14	-1.02	2.90
15	0.02	2.97
16	1.51	1.50
17	0.58	1.57
18	1.54	-0.49
19	0.55	1.95
20	-1.35	-0.64

Tabelle 3.8: Bivariate Zeitreihe mit Trendänderung in der zweiten Komponente

Die graphische Darstellung zeigt deutlich eine Änderung des Trends in der zweiten Komponente. Dies ist ebenfalls eine Art Ausreißer, genannt „multivariate

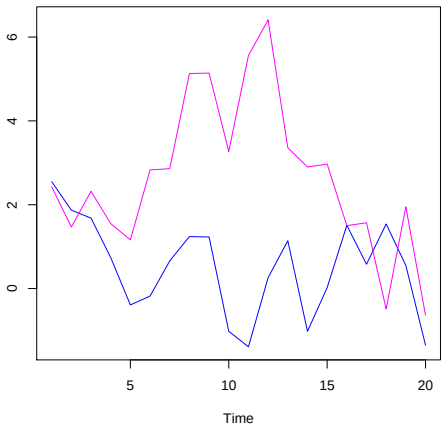

Abbildung 3.12: Bivariate Zeitreihe mit Trendänderung

ramp shift“ oder kurz ein Ausreißer vom Typ „R“. Eine Bildung der ersten Differenzen deutet an, dass nun anstelle einer Trendänderung ein level shift auftritt.

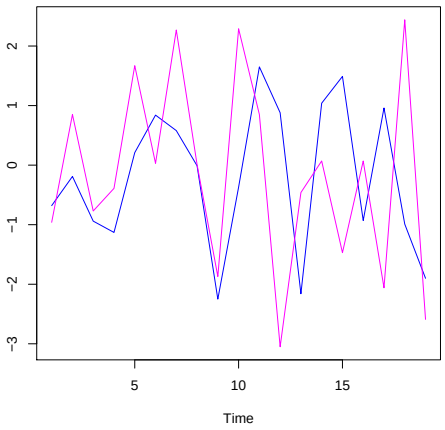

Abbildung 3.13: Bivariate Zeitreihe nach Differenzbildung: Trendänderung wird zu level shift

Die Bildung erster Differenzen kann offenbar Einfluss auf Ausreißer in Zeitreihen haben. Dadurch könnte sich sowohl die Art des Ausreißers, als auch seine Ausprägung, sprich sein Effekt, ändern. Die Wirkungsweise der Differenzbildung auf Ausreißer in Zeitreihen wurde von Galeano et al. [vgl. GPT06, S.661] hinreichend diskutiert und die Ergebnisse der Vollständigkeit halber im Folgenden zusammengestellt.

- Ein innovativer Ausreißer wird durch Bildung der ersten Differenzen wieder in einen innovativen Ausreißer umgewandelt.
- Ein additiver Ausreißer wird durch Bildung der ersten Differenzen in zwei aufeinander folgende, additive Ausreißer gleicher Größenordnung und unterschiedlichen Vorzeichens umgewandelt.
- Ein level shift wird durch Bildung der ersten Differenzen in einen additiven Ausreißer gleicher Größenordnung umgewandelt.
- Ein transienter Ausreißer wird durch Bildung der ersten Differenzen wieder in einen transienten Ausreißer umgewandelt.
- Ein Änderung des linearen Trends wird durch Bildung der ersten Differenzen in einen level shift gleicher Größenordnung umgewandelt.

3.5.7 Algorithmus zur Ausreißeranalytik bei nichtstationären Zeitreihen

Sei $Y_1, \ldots, Y_n$ eine multivariate Zeitreihe, die in mindestens einer Komponente einen linearen Trend aufweist, aber keine Trends höherer Ordnung besitzt.

Die graphische Darstellung aller Komponenten gibt einen Hinweis auf eventuell vorhandene Trends, Trendänderungen und -ordnungen. Wie man auf Trends in Zeitreihen testen kann wird in Abschnitt 3.5.5 vorgestellt. Das folgende Verfahren wird von Galeano et al. empfohlen [vgl. GPT06, S.661].

1. Bilde die ersten Differenzen $Z_t = Y_{t+1} - Y_t$. Dann besitzt Z_t keine Trends mehr.

2. Bestimme alle level shifts in der Zeitreihe Z_t, wie im Abschnitt 3.7.3 beschrieben und schätze die zugehörigen Effekte. Die so bestimmten level shifts in Z_t entsprechen Trendänderungen, sprich Ausreißern vom Typ „R“, in der Zeitreihe Y_t. Seien $w_1, \ldots, w_{r_R}$ die Effekte der entdeckten level shifts in der Zeitreihe Z_t. Diese entsprechen exakt den Effekten der jeweiligen Trendänderung in Y_t.

3. Entferne die Effekte $w_1, \ldots, w_{r_R}$ der Trendänderungen aus Y_t durch Subtraktion.
$$Y_t^* := Y_t - \sum_{i=1}^{r_R} w_i R_t^{(h)}$$
Dabei ist $R_t^{(h)} = (I - IB)^{-1} S_t^{(h)}$ mit $S_t^{(h)} = 0$ für $t < h$ und $S_t^{(h)} = 1$ für $t \geq h$. Einfach formuliert: man wähle nacheinander jeden gefundenen level shift w_i aus und subtrahiere seine Auswirkung auf die ursprüngliche Zeitreihe Y_t. Diese beträgt gerade $w_i R_t^{(h)}$.

4. Es liegt nun die Zeitreihe Y_t^* ohne Trendänderung vor. Bilde die ersten Differenzen $Z_t^* = Y_{t+1}^* - Y_t^*$. Die dadurch erzeugte Zeitreihe Z_t^* ist stationär, sprich sie enthält keine Trends und level shifts. Sie kann mit den bekannten Methoden zur Ausreißeranalytik untersucht werden.

Angenommen man hätte die Zeitreihe Z_t^* auf alle restlichen Arten von Ausreißern untersucht und die zugehörigen Effekte geschätzt. Im letzten Schritt werden die Effekte dieser Ausreißer nun umgerechnet in die korrespondierenden Effekte der Zeitreihe Y_t^* und anschließend durch Subtraktion aus ihr entfernt. Die Umrechnung geschieht wie folgt.

- Wurde in Z_t^* ein Ausreißer vom Typ „A“ mit Effekt w_A gefunden, so entspricht dies einem Ausreißer vom Typ „L“ mit gleich großem Effekt in der Zeitreihe Y_t^*. Der Effekt wird genau wie im Fall stationärer Zeitreihen aus Y_t^* entfernt.

- Wurden in Z_t^* zwei aufeinander folgende additive Ausreißer mit betragsmäßig gleich großen Effekten w_0, w_1, die sich lediglich über ihr Vorzeichen unterscheiden (d.h. $w_0 = -w_1$), gefunden, so entspricht dies einem additiven Ausreißer der Größe w_1 in der Zeitreihe Y_t^*. Dieser wird wie im Fall stationärer Zeitreihen aus Y_t^* entfernt.

- Wurde in Z_t^* ein Ausreißer vom Typ „T“ mit Effekt w_0 gefunden, so entspricht dies einem Ausreißer vom Typ „T“ mit Effekt w_0 in Y_t^*. Dieser wird wie im Fall stationärer Zeitreihen aus Y_t^* entfernt.

- Wurde in Z_t^* ein innovativer Ausreißer mit Effekt w_0 gefunden, so entspricht dies einem innovativen Ausreißer gleicher Größe in der Zeitreihe Y_t^*. Dieser wird wie im Fall stationärer Zeitreihen aus Y_t^* entfernt.

Das obige Verfahren ist auf multivariate Zeitreihen, deren Komponenten Trends beliebigen polynomialen Grades aufweisen, anwendbar, indem mittels sukzessiver Bildung erster Differenzen die Trends eliminiert werden, bis nur noch höchs-

tens lineare Trends vorliegen. Die sukzessive Rückwärtselimination von Trendänderungen, sprich Ausreißern vom Typ „R“, führt dann zu dem gewünschten Ergebnis.

3.5.8 Fehlende Werte in Zeitreihen Kompositionsdaten

Fehlende Werte in Zeitreihen stellen ein bedeutendes Problem dar, da durch sie eine Modellschätzung erschwert wird oder fehlinterpretiert werden kann. Angenommen es lägen 9 Jahresdaten der durchschnittlichen Gehalte von PCB 28 in Brassen (Abramis brama) vor, wobei die vierte Beobachtung keine wirkliche Beobachtung, sondern ein fehlender Wert wäre: $y_1, \ldots, y_9$ und y_4 ist fehlender Wert. Die Zeitreihe der ilr transformierten Daten sei $x_1, \ldots, x_9$, wobei x_4 ein fehlender Wert ist. Die Zeitreihe $(x_t)_{t=1,\ldots,9}$ folge einem AR[1]-Modell, d.h. $x_{t+1} = \beta x_t + \varepsilon_{t+1}$. Es ist nicht möglich x_4 einfach aus der Zeitreihe zu entfernen und an die daraus resultierende Zeitreihe mit 8 Messwerten ein AR[1]-Modell anzupassen. Aus dem AR[1]-Modell folgt, dass $x_4 = \beta x_3 + \varepsilon_4$ und $x_5 = \beta x_4 + \varepsilon_5$, d.h. der fehlende Wert x_4 beeinflusst im Modell die direkten Nachbarn x_3 und x_5. Würde man x_4 aus der Zeitreihe $x_1, \ldots, x_9$ entfernen und die Existenz des fehlenden Wertes gänzlich ignorieren, so erhielte man eine neue Zeitreihe $z_1, \ldots, z_8$ mit $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_3$, $z_4 = x_5$, $z_5 = x_6$, $z_6 = x_7$, $z_7 = x_8$ und $z_8 = x_9$. Unterstellt man für die Zeitreihe $z_1, \ldots, z_8$ ein AR[1]-Modell, so lautet dieses $z_{t+1} = \beta z_t + \varepsilon_{t+1}$. Aus diesem Modell folgt die Gleichung $z_4 = \beta z_3 + \varepsilon_4$. Diese ist äquivalent zu der Gleichung $x_5 = \beta x_3 + \varepsilon_5$, welche nicht das wahre AR[1]-Modell der Originaldaten $x_1, \ldots, x_9$ wiedergibt.

3.5.8.1 Schätzung fehlender Werte in Zeitreihen

Es existieren verschiedene Ansätze, um das Problem fehlender Werte zu handhaben.

1. Schätze fehlende Werte robust über einen Regressionsansatz. Um fehlende Werte in Zeitreihen zu ersetzen, ist es notwendig die Struktur der Daten so genau wie möglich zu erfassen. Folgt die Zeitreihe einem VAR[p]-Modell, so besteht ein praktikabler Ansatz darin, dieses Modell komponentenweise mit der least squares Methode aus den zur Verfügung stehenden Gleichungen, in denen kein fehlender Wert auftritt, zu schätzen. Die fehlenden Werte können dann durch das Modell prognostiziert werden [vgl. Sch01, S.183ff]. Ist die Zeitreihe durch Ausreißer kontaminiert, so empfiehlt sich eine robuste Schätzung des VAR[p]-Modells zu verwenden, z.B. über die LMS Methode (vgl. auch Abschnitt 3.5.3). Die zur Schätzung des Modells zur Verfügung stehenden Gleichungen sind dadurch charakterisiert, dass in ihnen kein fehlender Wert auftritt. Gleichungen, in denen fehlende Werte auftreten, können nicht zur Schätzung des Modells verwendet werden.

2. Ignoriere fehlende Werte und betrachte statt dessen die Zeitreihe der verbleibenden Beobachtungen. Diese werden dann als zeitlich äquidistant angenommen.

3. Lasse fehlende Beobachtungen aus und verwende Modelle, die nichtäquidistante Beobachtungszeitpunkte berücksichtigen.

Der Anteil fehlender Werte in den hier analysierten Daten beträgt in Einzelfällen bis zu 35%. Darunter sind Zeitreihen die anfänglich Beobachtungen im Ab-

stand von zwei Jahren aufweisen und ab einem bestimmten Zeitpunkt jährliche Daten. Beispielsweise enthält die Zeitreihe FICHTE(TRIEB) (Picea abies) Daten zu den Jahren 1985, 1987, 1989, 1991, 1993, 1995, 1997, 1999, 2000, 2001, 2002, 2003, 2004. Eine hier analysierte Zeitreihe zu BRASSEN enthält Daten für die Jahre 1991, 1992, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006 (alle Messprogramme sind in Anhang A aufgelistet). In dieser Arbeit wird jede Zeitreihe die fehlende Werte aufweist, zunächst mit der Methodik des Ansatzes 1 und danach mit Ansatz 2 ausgewertet. Die Ergebnisse werden anschließend verglichen. Ansatz 3 stellt einen hohen Aufwand dar. Einfache VARMA-Modelle können nicht mehr verwendet werden. Die gesamte hier vorgestellte Ausreißeranalytik uni- und multivariater Zeitreihen nach Tsay et al. [vgl. TPP98], Tsay [vgl. Tsa88] und Galeano et al. [vgl. GPT06] basiert auf der Schätzung von VAR- und VARMA-Modellen und ist nicht ohne weiteres auf Modelle für nichtäquidistante Daten übertragbar. Daher wird hier der Kompromiss angestrebt, die Auswertungen nach den Ansätzen 1 und 2 durchzuführen.

3.6 Ausreißer in univariaten Zeitreihen

Für dieses und alle folgenden Kapitel wird, wenn nicht näher spezifiziert, stets angenommen, dass eine stationäre Zeitreihe vorliegt, die sowohl eine ARMA, als auch eine AR und MA Darstellung besitzt, vgl. Abschnitt 3.5.1.1. Details zur Integration von Ausreißern in Zeitreihenmodelle finden sich in Abschnitt 3.5.1.3.

3.6.1 Schätzung des Effektes eines Ausreißers in univariaten Zeitreihen

Um die, in den folgenden Abschnitten vorgestellten Algorithmen nutzen zu können, ist es notwendig den Effekt eines Ausreißers auf eine Zeitreihe bestimmen bzw. schätzen zu können. Nur dadurch ist es möglich, diesen aus den Daten zu entfernen.

Sei $Y_1, \ldots, Y_n$ eine univariate Zeitreihe, welche möglicherweise Ausreißer enthält. Das zugehörige Modell lautet dann $Y_t = X_t + w\alpha(B)I_t^{(h)}$, mit $\Phi(B)X_t = C + \Theta(B)E_t$. Analog zum allgemeinen (k-dimensionalen) Fall, ergeben sich hier folgende Darstellungen für $\alpha(B)$ [vgl. Tsa88, S.3 f.].

- Sei w vom Typ „A“, dann gilt $\alpha(B) = 1$.
- Sei w vom Typ „T“, dann gilt $\alpha(B) = (1 - \delta B)^{-1}$ mit $0 < \delta < 1$.
- Sei w vom Typ „I“, dann gilt $\alpha(B) = \Psi(B) = \Phi(B)^{-1}\Theta(B)$.
- Sei w vom Typ „L“, dann gilt $\alpha(B) = (1 - B)^{-1}$.

3.6.1.1 Algorithmus zur Bestimmung des Effektes von Ausreißern

Die hier angegebenen Formeln zur Schätzung des Effektes eines Ausreißers in univariaten Zeitreihen finden sich bei Tsay [vgl. Tsa88, S.4]. Galeano et al. geben eine Verallgemeinerung für multivariate Zeitreihen [vgl. GPT06, S.655].

Die folgenden Schritte sind unter der Nullhypothese

„H_0 : die Zeitreihe $Y_1, \ldots, Y_n$ ist nicht mit Ausreißern kontaminiert“

zulässig. Unter H_0 gilt, dass die beobachtete Zeitreihe $Y_1, \ldots, Y_n$ mit der Zeitreihe $X_1, \ldots, X_n$ übereinstimmt und damit auch die Innovationen beider Reihen, also $A_t = E_t$. Folglich lässt sich eine einfache ARMA Darstellung $\Phi(B)Y_t = C + \Theta(B)A_t$ und eine einfache AR Darstellung $\Pi(B)Y_t = C_\Pi + A_t$ schätzen. Dies geschieht z.B. in **R** mit den Funktionen `arima` und `ar`.

1. Bestimme eine Schätzung der AR Darstellung der Zeitreihe $Y_1, \ldots, Y_t$, d.h. eine Schätzung der Koeffizienten Π_i in $\Pi(B)Y_t = C_\Pi + A_t$.

2. Bestimme die Residuen des Modells. Diese seien bezeichnet mit $\hat{\Pi}_i$. Sie ergeben sich aus $y_t = \hat{\Pi}(B)Y_t$. Bestimme die Varianz der zugehörigen Residuen. Sie liefert unter der Nullhypothese eine konsistente Schätzung der Varianz der Innovationen, $\sigma_a^2 = \sigma_E^2$. Das Softwarepaket **R** liefert zu diesem Zweck verschiedene, bereits implementierte Verfahren.

3. Sei h ein bekannter Zeitpunkt, zu dem ein Ausreißer auftritt. Wie man diesen Zeitpunkt findet, wird in den kommenden Abschnitten dargestellt. Tsay zeigte, wie die Effekte von Ausreißern in univariaten Zeitreihen geschätzt werden können [vgl. Tsa88, S.4]. Berechne für alle Zeitpunkte h den Effekt des dort detektierten Ausreißers.

   - Für einen Ausreißer vom Typ „A“ setze $\rho_{A,h}^2 = (1 + \hat{\Pi}_1^2 + \ldots + \hat{\Pi}_{n-h}^2)^{-1}$, dann ist

   $$\hat{w}_{A,h} = \rho_{A,h}^2 \left(y_h - \sum_{i=1}^{n-h} \hat{\Pi}_i y_{h+i} \right)$$

   eine Schätzung des Effektes, falls ein additiver Ausreißer zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch

$Var(\hat{w}_{A,h}) = \rho^2_{A,h}\sigma^2_a$.

- Für einen Ausreißer vom Typ „T“ setze $\beta(B) = (1-\delta B)^{-1}\hat{\Pi}(B) = \beta_0 - \beta_1 B - \beta_2 B^2 \ldots$ und zusätzlich $\rho^2_{T,h} = (1+\beta_1^2+\ldots+\beta^2_{n-h})^{-1}$, dann ist

$$\hat{w}_{T,h} = \rho^2_{T,h}\left(y_h - \sum_{i=1}^{n-h}\beta_i y_{h+i}\right)$$

eine Schätzung des Effektes, falls ein Ausreißer vom Typ „T“ zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch $Var(\hat{w}_{T,h}) = \rho^2_{T,h}\sigma^2_a$.

- Für einen Ausreißer vom Typ „L“ setze $\eta(B) = (1-B)^{-1}\hat{\Pi}(B) = \eta_0 - \eta_1 B - \ldots$ und zusätzlich $\rho^2_{L,h} = (1+\eta_1^2+\ldots+\eta^2_{n-h})^{-1}$, dann ist

$$\hat{w}_{L,h} = \rho^2_{L,h}\left(y_h - \sum_{i=1}^{n-h}\eta_i y_{h+i}\right)$$

eine Schätzung des Effektes, falls ein level shift zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch $Var(\hat{w}_{L,h}) = \rho^2_{L,h}\sigma^2_a$.

- Für einen Ausreißer vom Typ „I“ setze

$$\hat{w}_{I,h} = y_h$$

als Schätzung des Effektes, falls ein innovativer Ausreißer zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch $Var(\hat{w}_{I,h}) = \sigma^2_a$.

Die Teststatistiken (sog. Likelihood Ratio (LR) Statistiken) für Ausreißer des

Typs i=A,T,I sind im folgenden angegeben [vgl. Tsa88, S.5].

- Sei w ein Ausreißer vom Typ „A“ zur Zeit h, so ist die zugehörige LR-Teststatistik $\lambda_{A,h} = \frac{\hat{w}_{A,h}}{\rho_{A,h}\sigma_a}$.
- Sei w ein Ausreißer vom Typ „T“ zur Zeit h, so ist die zugehörige LR-Teststatistik $\lambda_{T,h} = \frac{\hat{w}_{T,h}}{\rho_{T,h}\sigma_a}$.
- Sei w ein Ausreißer vom Typ „I“ zur Zeit h, so ist die zugehörige LR-Teststatistik $\lambda_{I,h} = \frac{\hat{w}_{I,h}}{\sigma_a}$.

Tsay zeigte, dass obige LR-Statistiken unter der Nullhypothese „H_0 : die Zeitreihe ist nicht mit Ausreißern kontaminiert“ und der Annahme bekannter Modellparameter und eines bekannten Zeitpunktes h normal-$(0, 1)$-verteilt sind. Die Modellparameter sind im Allgemeinen unbekannt. Werden diese konsistent aus den Daten geschätzt, sind die angegebenen Teststatistiken asymptotisch normal-$(0, 1)$-verteilt. Signifikanz zum Niveau α liegt folglich vor, wenn $\lambda_{i,h} \notin [z_{\alpha/2}, z_{1-\alpha/2}]$, wobei $z_{\alpha/2}$ und $z_{1-\alpha/2}$ das $(\alpha/2)$, bzw. $(1 - \alpha/2)$-Quantil der Standardnormalverteilung sind und i=A,T,I. Die Alternativhypothese lautet dann jeweils „H_1 : es liegt ein Ausreißer vom Typ i zur Zeit h vor“ [vgl. Tsa88, S.5].

3.6.1.1.1 Beispiel Dieses Beispiel wird unter Verwendung der Software **R** illustriert, um dem Anwender die praktischen Methoden der Datenauswertung zu veranschaulichen. Gegeben sei die in **R** simulierte univariate Zeitreihe Y_t, $(t = 1, \ldots, 100)$ mit den Werten aus Tabelle 3.9. Das praktische Vorgehen ist leichter zu verstehen für simulierte Zeitreihen mit einem Beobachtungshorizont $n \geq 50$, da hier bereits visuell gut erkennbar ist, zu welchem Zeitpunkt und in welcher Größenordnung ein level shift vorliegt. Die in dieser Arbeit ausgewerteten realen

0.466	0.538	0.100	0.608	0.447	0.444	0.373	0.521	0.989	0.914
0.476	0.462	0.393	0.727	0.642	0.536	0.353	0.609	1.151	0.819
0.367	0.532	0.364	0.635	0.528	0.355	0.566	0.568	0.906	0.817
0.431	0.516	0.478	0.609	0.480	0.446	0.580	0.571	1.018	0.647
0.493	0.532	0.501	0.537	0.636	0.382	0.439	0.525	0.865	0.737
0.510	0.620	0.387	0.595	0.425	0.367	0.490	0.309	0.850	0.675
0.475	0.534	0.606	0.617	0.748	0.369	0.498	0.337	0.778	0.680
0.510	0.432	0.540	0.478	0.545	0.312	0.405	0.411	0.808	0.977
0.598	0.437	0.539	0.391	0.514	0.447	0.498	0.383	0.825	1.001
0.746	0.192	0.507	0.422	0.528	0.307	0.403	0.888	0.686	1.117

Tabelle 3.9: In R Simulierte Zeitreihe mit level shift

Zeitreihen weisen einen wesentlich kürzeren Beobachtungshorizont auf ($n \leq 20$). Zu rein illustrativen Zwecken soll diese simulierte Zeitreihe genügen.

Es sei bekannt, dass die Zeitreihe mit einem level shift der Größe 0.5 zum Zeitpunkt $t = 81$ kontaminiert ist, d.h. alle Messwerte Y_t mit $t = 81, \ldots, 100$ sind um 0.5 zu hoch. Die Länge der Zeitreihe ist $n = 100$. In der Praxis ist der exakte Zeitpunkt der Kontamination einer Zeitreihe mit einem Ausreißer im Allgemeinen unbekannt. Ausnahmen bekannter Zeitpunkte wären z.B. die eines Hochwassers oder eines Börsencrashs. Zu diesem Zweck wird in Abschnitt 3.6.2.1 ein Algorithmus zum Auffinden dieser Zeitpunkte angegeben.

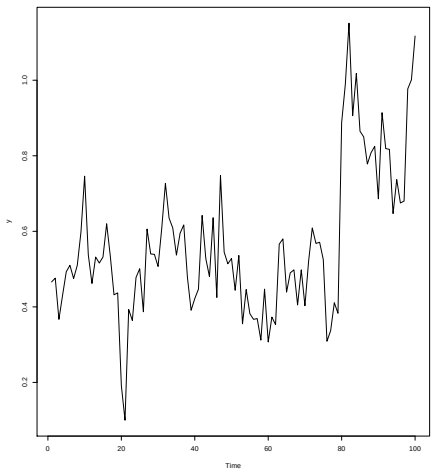

Abbildung 3.14: Zeitreihe mit level shift

Zunächst wird ein AR Modell an die Zeitreihe angepasst. Dies geschieht in **R** mit Hilfe der Funktion `ar`. **R** schätzt ein AR[2] Modell mit den Koeffizienten $\hat{\Pi}_1 = 0.6275709$ und $\hat{\Pi}_2 = 0.1602863$. Zur Bestimmung des Effektes des level shift muss $\eta(B)$ bestimmt werden. Dieses ergibt sich aus der Bestimmungsgleichung $\eta(B) = (1-B)^{-1}\Pi(B)$ unter Berücksichtigung der Darstellung $(1-B)^{-1} = (1 + B + B^2 + B^3 + \ldots)$ und $\Pi(B) = (1 - \Pi_1 B - \Pi_2 B^2)$.

$$
\begin{aligned}
&(1 - \Pi_1 B - \Pi_2 B^2)(1 + B + B^2 + B^3 + \ldots) \\
=&1 + B + B^2 + B^3 + B^4 \ldots \\
&-\Pi_1 B - \Pi_1 B^2 - \Pi_1 B^3 - \Pi_1 B^4 - \ldots \\
&-\Pi_2 B^2 - \Pi_2 B^3 - \Pi_2 B^4 - \ldots \\
=&1 - (\Pi_1 - 1)B - \sum_{l=2}^{\infty}(\Pi_1 + \Pi_2 - 1)B^l
\end{aligned}
$$

Damit sind $\hat{\eta}_1 = \hat{\Pi}_1 - 1 = -0.3724291$ und $\hat{\eta}_l = \hat{\Pi}_1 + \hat{\Pi}_2 - 1 = -0.2121428$

für $l \geq 2$. Man braucht für die Berechnung von $\rho^2_{L,h}$ aber nur $\eta_1, \ldots, \eta_{n-h}$, sprich in diesem Fall die Schätzungen $\hat{\eta}_1, \ldots, \hat{\eta}_{20}$. Setzt man diese in die zugehörige Formel ein, so ergibt sich $\hat{\rho}^2_{L,h} = 0.51314$. Die Schätzung der Varianz der Innovationen beträgt $\hat{\sigma}^2_a = 0.01350145$. Es wurde zum Zeitpunkt $t = 81$ der Effekt des level shifts geschätzt mit 0.5802699 und einem Standardfehler der Schätzung in Höhe von 0.09300751. Wie man den Effekt eines Ausreißers entfernt, wird in den folgenden Abschnitten erläutert.

Zur statistischen Prüfung von gefundenen level shifts auf Signifikanz, empfiehlt sich die im Abschnitt 3.6.2 vorgestellte C-Statistik. Dort ist beschrieben, wie univariate Zeitreihen iterativ nach level shifts durchsucht und diese auf Signifikanz geprüft werden können. Die Abgrenzung zur Prüfung von Ausreißern des Typs i=A,T,I auf Signifikanz wird hier bewusst vorgenommen, da es sich nach Galeano et al. allgemein empfiehlt, eine Zeitreihe, sei sie uni- oder multivariat, stets zuerst von level shifts zu bereinigen, bevor die Prüfung auf Existenz anderer Ausreißer durchgeführt wird [vgl. GPT06, S.659].

3.6.2 Bereinigung univariater Zeitreihen von level shifts

Balke zeigte, dass eine Methode zur Ausreißererkennung, welche die initiale Schätzung eines ARMA Modells benötigt, level shifts fälschlicher Weise als innovative Ausreißer identifizieren kann. Dazu siehe auch [BF94] und [Bal93]. Galeano et al. empfehlen, eine Zeitreihe vor Untersuchung auf Ausreißer vom Typ i=A,T,I zunächst von Ausreißern des Typs „L“ zu bereinigen [vgl. GPT06, S.659]. Zur Identifizierung von Ausreißern des Typs „A“, „T“ und „I“ findet die Likelihood

Ratio Statistic (LRS) Verwendung [vgl. Tsa88, S.5]. Um die Zeitreihen von level shifts, sprich Ausreißern des Typs „L“, zu bereinigen, wird eine Teststatistik, welche sich von der LRS unterscheidet, eingesetzt. Diese besitzt nach Galeano et al. [vgl. GPT06, S.659] entscheidende Vorteile, da sie zum einen keine Informationen über die Parameter p und q des initialen ARMA Modells benötigt und zum anderen auch eine höhere Power, als die korrespondierende LRS Statistik, aufweist.

Die Teststatistik des Verfahrens zum Auffinden von level shifts in einer eindimensionalen Zeitreihe $Y_1, \ldots, Y_n$ ist gegeben durch C mit [vgl. GPT06, S.659]

$$C := \max_{t=1,\ldots,n} C_t \, , \;\; C_t := \frac{t}{\sqrt{n}\psi(1)\sigma_e} \left(\frac{1}{t} \sum_{i=1}^{t} Y_i - \bar{Y} \right) \, , \; t = 1, \ldots, n \qquad (3.19)$$

$$\widehat{\psi(1)\sigma_e} := \left[\widehat{\gamma(0)} + 2 \sum_{i=1}^{K} \left(1 - \frac{|i|}{K} \right) \widehat{\gamma(i)} \right]^{1/2} ,$$

wobei in der praktischen Anwendung $\widehat{\psi(1)\sigma_e}$ den Wert von $\psi(1)\sigma_e$ konsistent schätzt und $\widehat{\gamma(i)} = cov(Y_t, Y_{t-i})$. Der Parameter K muss die Bedingungen $K \rightarrow \infty$ und $\lim_{n\rightarrow\infty} K/n = 0$ erfüllen, z.B. wähle K als den ganzzahligen Anteil von $\sqrt{n}$.

Bai bewies, dass C für $n \rightarrow \infty$ in Verteilung gegen das Supremum einer Brown'schen Brücke auf [0,1] konvergiert [vgl. Bai97]. Unter der Nullhypothese „H_0 : es gibt keinen level shift, sprich keinen Ausreißer vom Typ „L“ in den

Daten“, ist $\max_{t=1,\ldots,n} C_t$ dann asymptotisch mit Verteilungsfunktion

$$F(x) = 1 + 2\sum_{i=1}^{\infty} (-1)^i \exp\left(-2i^2x^2\right), \;\; x > 0$$

verteilt. Die kritischen Werte für die Untersuchung einer univariaten Zeitreihe auf level shifts lassen sich somit einfach als Quantile der durch F induzierten Wahrscheinlichkeitsverteilung bestimmen. Sei t der zu C_t korrespondierende Zeitpunkt. Zum Zeitpunkt $t + 1$ liegt genau dann ein Ausreißer vom Typ „L“ vor, wenn

$$\min\{F(C_t), 1 - F(C_t)\} \leq \alpha$$

gilt.

3.6.2.1 Algorithmus zur Bereinigung univariater Zeitreihen von level shifts

Dieser Algorithmus ist angelehnt an den Algorithmus von Galeano et al. zur Entfernung von level shifts aus multivariaten Zeitreihen und wurde leicht modifiziert [vgl. GPT06, S.659].

Lege für die Zeitreihe $Y_1, \ldots, Y_n$ einen Mindestabstand H zwischen zwei Zeitpunkten von level shifts fest, um eine Überbestimmung von level shifts zu vermeiden (z.B. H=10).

1. Setze $t_1 = 1$ und $t_2 = n$ und definiere einen Vektor $flag$ der Länge n, der nur Nullen enthält, sprich $flag = (0, \ldots, 0)'$. Der i-te Eintrag von $flag$ wird bezeichnet mit $flag_i$.

$$D = \max_{\{t_1,\ldots,t_2\} \cap \{t \mid flag_t = 0\}} |C_t| \tag{3.20}$$

Als nächstes bestimme das Argument, sprich den Index t, für den D maximal wird.

$$t_{\max} = \arg \max_{\{t_1,\ldots,t_2\} \cap \{t \mid flag_t = 0\}} |C_t| \tag{3.21}$$

Der kritische Wert zum Signifikanzniveau α sei D_α. Falls $D > D_\alpha$ ist, liegt möglicherweise ein level shift zum Zeitpunkt $t = t_{\max} + 1$ vor. Setze $flag_{t_{\max}} = 1$. Falls $D < D_\alpha$ gilt, stoppt der Algorithmus.

2. (a) Setze $t_2 = t_{\max}$ in Schritt 1 und bestimme D, sowie $t_{\max}$ wie in Gleichung (3.20) und Gleichung (3.21). Falls $D > D_\alpha$ und $t_2 - t_{\max} > H$, setze $t_2 = t_{\max}$ und $flag_{t_{\max}} = 1$. Wiederhole Schritt 2.a) bis entweder $D < D_\alpha$ oder $t_2 - t_{\max} \leq H$ gilt. Setze nun $t_{first} = t_2$, wobei t_2 der letzte Zeitpunkt ist, der die Ungleichungen $D > D_\alpha$ und $t_2 - t_{\max} > H$ erfüllt. Der Zeitpunkt $t = t_{first} + 1$ ist der des ersten möglichen level shifts.

   (b) Setze $t_1 = t_{\max}$ und $t_2 = n$ in Schritt 1. Berechne neue Werte nach Gleichung (3.20) und Gleichung (3.21). Wenn $D > D_\alpha$ und $t_{\max} - t_1 > H$, setze $t_1 = t_{\max}$ und $flag_{t_{\max}} = 1$. Wiederhole Schritt 2.b) solange, bis entweder $D < D_\alpha$ oder $t_{\max} - t_1 \leq H$. Setze $t_{last} = t_1$, wobei t_1 der letzte Zeitpunkt ist, der die Ungleichungen $D > D_\alpha$ und $t_{\max} - t_1 > H$ erfüllt. Der Zeitpunkt $t = t_{last} + 1$ ist der des letzten möglichen level shifts.

   (c) Falls $t_{last} - t_{first} < H$, stoppt der Algorithmus. Ansonsten wiederhole die Schritte 2.a) und 2.b) mit $t_1 = t_{first}$ und $t_2 = t_{last}$ bis keine wei-

teren möglichen Zeitpunkte für level shifts berechnet werden können. Gehe zu Schritt 3.

3. Definiere den Vektor $h = (h_0, \ldots, h_{r+1})$ mit $h_0 = 1$, $h_{r+1} = n$ und den aufsteigend geordneten Zeitpunkten der möglichen level shifts $h_1 < \cdots < h_r$, die in Schritt 1 und 2 bestimmt wurden. Diese korrespondieren zu den Stellen im Vektor $flag$, an denen eine 1 steht. Berechne für jedes zeitliche Teilintervall (h_i, h_{i+2}) den Wert $D = \max_{t_1 \leq t \leq t_2} |C_t|$ mit $t_1 = h_i$ und $t_2 = h_{i+2}$ und prüfe die statistische Signifikanz, sprich: gilt $D > D_\alpha$? Falls dies nicht gilt, entferne h_{i+1} aus dem Vektor h. Wiederhole diesen Schritt solange, bis nur noch signifikante D vorliegen. Entferne $h_0 = 1$ und $h_{r+1} = n$ aus h. Setze $h_i := h_i + 1$, $i = 1, \ldots, r$.

4. Seien $\{h_1, \ldots, h_r\}$ die Zeitpunkte der level shifts. Im Folgenden werden deren Auswirkungen aus der Zeitreihe entfernt. Die Effekte $w_1, \ldots, w_r$ der level shifts werden, wie im Abschnitt 3.6.1 beschrieben, geschätzt und anschließend das AR Modell

$$(I - \Pi_1 B - \ldots - \Pi_{\hat{p}} B^{\hat{p}}) y_t^* = a_t^* \tag{3.22}$$

für alle $0 \leq \hat{p} \leq p_{\max}$ mit $y_t^* = y_t - \sum_{i=1}^{r} w_i S_t^{(h_i)}$ geschätzt, sprich es wird unter Verwendung statistischer Software ein AR Modell an die, von den level shifts $w_1, \ldots, w_r$ bereinigte Zeitreihe y_t^* angepasst. Es gilt dabei $S_t^{(h_i)} = 0$ für $t < h_i$ und $S_t^{(h_i)} = 1$ für $t \geq h_i$. Anschließend wählt man das

$\hat{p}$, welches die folgende Bedingung erfüllt, aus.

$$\hat{p} = \arg \max_{0 \leq p \leq p_{\max}} AIC(p) = \arg \max_{0 \leq p \leq p_{\max}} \left\{ \log \hat{\Sigma}_p + 2\frac{k^2 p}{n} \right\}$$

Dabei gilt jeweils $\hat{\Sigma}_p = \frac{1}{n-2p-1} \sum_{t=p+1}^{n} a_t^{*2}$ und a_t^* sind die zu y_t^* korrespondierenden Innovationen. Anders gesagt: zunächst werden die Effekte der level shifts geschätzt und von der ursprünglichen Zeitreihe $y_1, \ldots, y_n$ subtrahiert. Man erhält dadurch die modifizierte Zeitreihe $y_1^*, \ldots, y_n^*$. Anschließend wird mit Hilfe einer Statistiksoftware (z.B. **R**) ein AR[$\hat{p}$] Modell an die Zeitreihe angepasst. Das Entscheidungskriterium über die Anpassungsgüte des Modells ist das AIC und $p_{\max}$ ein vorher selbst festgelegter Parameter, um die maximale Ordnung des anzupassenden Modells kontrollieren zu können. Falls nichtsignifikante level shifts im Modell auftreten, ist iterativ der am wenigsten signifikante level shift aus 3.22 zu entfernen und das Modell neu zu schätzen. Diese schrittweise Prozedur wird wiederholt, bis alle level shifts signifikant sind.

Der kritische Wert D_α ergibt sich in obigem Algorithmus als Quantil der durch F induzierten Wahrscheinlichkeitsverteilung. Das gesamte Verfahren kann nun auf die Zeitreihe y_t^* angewendet werden, sprich man setze $y_t := y_t^*$ und wiederhole die Schritte 1-4 des Algorithmus zur Bereinigung univariater Zeitreihen von level shifts solange, bis keine signifikanten level shifts mehr gefunden. Erst danach gilt die ursprüngliche Zeitreihe endgültig von Ausreißern des Typs „L“ bereinigt.

3.6.2.1.1 Ein praktisches Beispiel zur Umsetzung des Verfahrens Dieses Beispiel wird unter Verwendung der Software **R** illustriert, um dem Anwender

die praktischen Methoden der Datenauswertung zu veranschaulichen. Gegeben sei die univariate Zeitreihe Y_t, $(t = 1, \ldots, 100)$ mit den Werten aus Tabelle 3.9 von Seite 148. Es sei bekannt, dass die Zeitreihe mit einem level shift der Größe 0.5 zum Zeitpunkt $t = 81$ kontaminiert wurde, d.h. alle Messwerte Y_t mit $t = 81, \ldots, 100$ sind um 0.5 zu hoch. Die Länge der Zeitreihe ist $n = 100$. Das Verfahren zum Auffinden des Auftretenszeitpunktes des Ausreißers und die Entfernung des geschätzten Effektes werden hier vorgestellt. Die Schätzung des Effektes auf 0.5802699 wurde bereits im datengleichen Beispiel des Abschnittes 3.6.1 veranschaulicht und übersteigt den tatsächlichen, in der Praxis unbekannten, Effekt 0.5 des level shifts gering. Im letzten Schritt wird der Effekt des level shifts aus der univariaten Zeitreihe entfernt. Dazu subtrahiert man den geschätzten Effekt (hier 0.5802699) ab dem Auftretenszeitpunkt von jedem einzelnen Wert der Zeitreihe, also $y_t^* = y_t$ für $t = 1, \ldots, 80$ und $y_t^* = y_t - 0.5802699$ für $t = 81, \ldots, 100$. Die vom level shift bereinigte Zeitreihe ist in Abbildung 3.15 dargestellt.

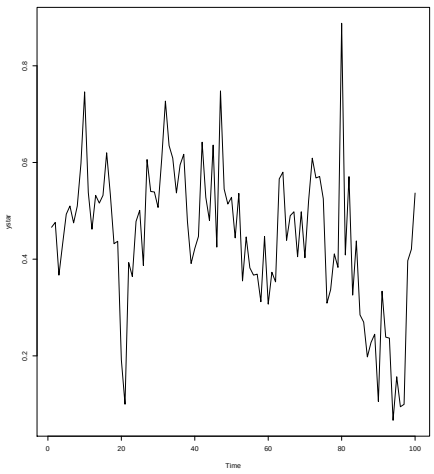

Abbildung 3.15: Von level shift bereinigte, univariate Zeitreihe

3.6.3 Bestimmung von Ausreißern des Typs „A“, „T“ und „I“ in univariaten Zeitreihen

Sei $Y_1, \ldots, Y_t$ eine univariate, stationäre und von Ausreißern des Typs „L“ bereinigte, Zeitreihe. Die Untersuchung auf Ausreißer vom Typ i=A,T,I erfolgt unter Verwendung der Likelihood Ratio Statistics (LRS) nach Tsay [vgl. Tsa88, S.5]. Diese gliedert sich in eine äußere und eine innere Schleife, wodurch Ausreißer iterativ entfernt werden. Die dort verwendete und noch zu spezifizierende Teststatistik λ folgt unter der Nullhypothese „H_0 : die Zeitreihe ist nicht mit Ausreißern kontaminiert“ keiner bekannten Verteilung. Die kritischen Werte wurden daher in statistischen Simulationsstudien ermittelt. Tsay empfiehlt, diese auf $C = 4$, $C = 3.5$ oder $C = 3$ festzulegen, je nach gewünschtem Signifikanzniveau ([vgl. Tsa88, S.8]). Dabei gilt: je größer C, desto kleiner das Signifikanzniveau α. Die Teststatistiken für jeden der drei Ausreißertypen i=A,T,I werden im Folgenden noch einmal angegeben [vgl. Tsa88, S.5 f.].

$$\lambda_{A,h} = \frac{\hat{w}_{A,h}}{\rho_{A,h}\sigma_a} \ , \ \ \lambda_{T,h} = \frac{\hat{w}_{T,h}}{\rho_{T,h}\sigma_a} \ , \ \ \lambda_{I,h} = \frac{\hat{w}_{I,h}}{\sigma_a}$$

Dann ergeben sich die Teststatistiken $\lambda_{i,\max}$ für i=A,T,I aus

$$\lambda_{i,\max} = \max\{|\lambda_{i,t}|, \ 1 \leq t \leq n\} \ .$$

Diese besitzen eine unbekannte Verteilung, da sie jeweils das Maximum abhängiger Zufallsgrößen darstellen. Die finale Teststatistik λ ergibt sich nun als Maxi-

mum der $\lambda_{i,\max}$, sprich

$$\lambda = \max\{\lambda_{A,\max}, \lambda_{T,\max}, \lambda_{I,\max}\} \ .$$

Eine genaue Betrachtung der Aussagen von Tsay ergibt, dass die kritischen Werte $C = 4$, $C = 3.5$ und $C = 3$ für die Teststatistik

$$\lambda = \max\{\lambda_{A,\max}, \lambda_{T,\max}, \lambda_{I,\max}, \lambda_{L,\max}\}$$

angegeben wurden ([vgl. Tsa88, S.7]). In dieser Arbeit wird λ unabhängig von $\lambda_{L,\max}$ definiert. Diese Festlegung erhöht aber nicht die kritischen Werte $C \in \{3, 3.5, 4\}$. Das Signifikanzniveau wird weiterhin eingehalten. Um dies einzusehen, betrachte man für eine beliebige Konstante $C \in \mathbb{R}$ und reelle Zufallsvariablen T, U, V, W den folgenden Zusammenhang.

$$\begin{aligned}
&\max\{T, U, V\} \leq \max\{T, U, V, W\} \\
\text{und } &[\max\{T, U, V\} \geq C] \ \Rightarrow \ [\max\{T, U, V, W\} \geq C] \\
\Rightarrow &\{\max\{T, U, V\} \geq C\} \subseteq \{\max\{T, U, V, W\} \geq C\} \\
\Rightarrow &P(\max\{T, U, V\} \geq C) \leq P(\max\{T, U, V, W\} \geq C) \ .
\end{aligned}$$

Es ist folglich wichtig, sich der Tatsache bewusst zu sein, dass die kritischen Werte ($C = 4$, $C = 3.5$ und $C = 3$) für λ hier strenger als notwendig ausfallen, um das Signifikanzniveau einzuhalten.

3.6.3.1 Algorithmus zur Bereinigung univariater Zeitreihen von Ausreißern des Typs i=A,T,I

Sei $Y_1, \ldots, Y_n$ eine univariate, stationäre Zeitreihe, welche möglicherweise Ausreißer enthält. Das zugehörige Modell lautet dann $Y_t = X_t + w\alpha(B)I_t^{(h)}$, mit $\Phi(B)X_t = C + \Theta(B)E_t$. Dabei ist $(X_t)_{t=1,\ldots,n}$ ein zugrunde liegender, unbeobachteter univariater ARMA Prozess, der keine Ausreißer enthält. Analog zum allgemeinen (k-dimensionalen) Fall ergeben sich hier folgende Darstellungen für $\alpha(B)$ [vgl. GPT06, Tsa88].

- Sei w vom Typ „A“, dann gilt $\alpha(B) = 1$.
- Sei w vom Typ „T“, dann gilt $\alpha(B) = (1 - \delta B)^{-1}$ mit $0 < \delta < 1$.
- Sei w vom Typ „I“, dann gilt $\alpha(B) = \Psi(B) = \Phi(B)^{-1}\Theta(B) = \Pi(B)^{-1}$.

Der hier vorgestellte Algorithmus ist angelehnt an den von Tsay veröffentlichten Algorithmus [vgl. Tsa88, S.7 f.], verzichtet jedoch auf die dort vorgeschlagene Erkennung von level shifts. Dem Beispiel von Galeano et al. folgend, werden level shifts im Vorfeld der weiteren Ausreißeranalyse aus den Daten entfernt [vgl. GPT06, S.659].

Die folgenden Schritte sind unter der Nullhypothese

„H_0 : die Zeitreihe $Y_1, \ldots, Y_n$ ist nicht mit Ausreißern kontaminiert“

zulässig. Unter H_0 gilt, dass die beobachtete Zeitreihe $Y_1, \ldots, Y_n$ mit der unbeobachteten Zeitreihe $X_1, \ldots, X_n$ aus dem Ausreißermodell übereinstimmt und damit auch die Innovationen beider Reihen, also $A_t = E_t$. Folglich lässt sich eine

einfache ARMA Darstellung $\Phi(B)Y_t = C + \Theta(B)A_t$ und eine einfache AR Darstellung $\Pi(B)Y_t = C_\Pi + A_t$ schätzen. Dies geschieht z.B. in **R** mit den Funktionen `arima` und `ar`.

1. Wähle einen kritischen Wert C und bestimme eine Schätzung der AR Darstellung der Zeitreihe $Y_1, \ldots, Y_t$, sprich eine Schätzung der Koeffizienten Π_i in $\Pi(B)Y_t = C_\Pi + A_t$. Diese seien bezeichnet mit $\hat{\Pi}_i$, die Residuen des geschätzten Modells seien mit y_t bezeichnet.

   (a) Unter Verwendung der Ergebnisse aus Schritt 1 bestimme die Varianz der zugehörigen Residuen. Sie liefert eine konsistente Schätzung der Varianz der Innovationen, $\sigma_a^2 = \sigma_E^2$. Das Softwarepaket **R** liefert zu diesem Zweck verschiedene, bereits implementierte Verfahren.

   (b) Unter Verwendung der Residuen und der geschätzten Varianz der Innovationen aus Schritt 1a), sowie den geschätzten Parametern $\hat{\Pi}_i$ der AR Darstellung, berechne nun für jeden Zeitpunkt $t = 1, \ldots, n$ die Teststatistiken $\lambda_{i,t}$ (i=A,T,I) und anschließend bestimme $\lambda_{A,\max}$, $\lambda_{T,\max}$ und $\lambda_{I,\max}$, sowie die zugehörigen Zeitpunkte h_A, h_T und h_I.

   (c) Setze $\lambda = \max\{\lambda_{A,\max}, \lambda_{T,\max}, \lambda_{I,\max}\}$. Falls $\lambda < C$, liegt keine signifikante Beobachtung vor. In diesem Fall gehe zu Schritt 2. Falls jedoch $\lambda \geq C$ gilt, liegt eine signifikante Beobachtung vor. Dann bestimme als nächstes das zu λ korrespondierende $\lambda_{i,\max}$. Dieses charakterisiert die Art des Ausreißers. Gilt z.B. $\lambda = \lambda_{A,\max}$, so handelt es sich um einen Ausreißer vom Typ A. Der zugehörige Zeitpunkt h_A gibt an, wann der Ausreißer auftritt. Der Effekt w des zu λ korrespondierenden Ausreißers wurde bereits bei der Berechnung der zugehörigen

Teststatistik $\lambda_{i,\max}$ bestimmt!

(d) Entferne den Effekt des Ausreißers durch Subtraktion von der Zeitreihe Y_t. Dabei muss die Art des Ausreißers beachtet werden. Allgemein entfernt man den Effekt w eines Ausreißers unter Berücksichtigung der zugehörigen Struktur $\alpha(B)$. Im Fall eines additiven Ausreißers gilt $\alpha(B) = 1$, daher ist $w\alpha(B)I_t^{(h_A)} = w$ für $t = h_A$ und $w\alpha(B)I_t^{(h_A)} = 0$ für $t \neq h_A$. Subtrahiert man nun für $t = 1, \ldots, n$ die Effekte von der Zeitreihe, ergibt sich die neue Zeitreihe $Y_t^* = Y_t - w\alpha(B)I_t^{(h_A)}$. Dabei wurde nur zum Zeitpunkt h_A etwas verändert, sprich es gilt $Y_t^* = Y_t$ für $t \neq h_A$ und $Y_t^* = Y_t - w$ für $t = h_A$. Bei innovativen und transienten Ausreißern verwendet man analog die bereits bekannten Strukturen $\alpha(B) = \Psi(B) = \Phi(B)^{-1}\Theta(B)$ und $\alpha(B) = (1 - \delta B)^{-1} = \sum_{j=0}^{\infty} \delta^j B^j$ mit $0 < \delta < 1$. Die Darstellung $\alpha(B) = \Psi(B)$ ergibt sich aus der Bildung von $\hat{\Pi}(B)^{-1}$, d.h. die geschätzte AR Darstellung der Zeitreihe $Y_1, \ldots, Y_n$ wird invertiert, sodass die MA Darstellung vorliegt. Dazu ist die Gleichung $\hat{\Pi}(B) \cdot \hat{\Pi}(B)^{-1} = 1$ zu lösen. Hat man $\hat{\Pi}(B)^{-1}$ bestimmt, so gilt $\alpha(B) = \hat{\Pi}(B)^{-1}$. Gehe zu Schritt 1 mit $Y_t := Y_t^*$.

2. Wenn in der inneren Iteration kein signifikanter Effekt gefunden wurde, beende die Analyse der Zeitreihe auf Ausreißer, ansonsten setze $Y_t := Y_t^*$ und gehe zu Schritt 1.

Die jeweiligen Effekte w, die zur Berechnung von $\lambda_{i,t}$ ($t = 1, \ldots, n$) und damit auch für $\lambda_{i,\max}$ (i=A,T,I) und λ notwendig sind, bestimmt man, wie im Abschnitt 3.6.1 beschrieben. Die dazu notwendigen Formeln werden hier noch einmal kurz

rekapituliert [vgl. Tsa88, S.6].

- Für einen Ausreißer vom Typ „A“ setze $\rho^2_{A,h} = (1 + \hat{\Pi}^2_1 + \ldots + \hat{\Pi}^2_{n-h})^{-1}$, dann ist

$$\hat{w}_{A,h} = \rho^2_{A,h} \left(y_h - \sum_{i=1}^{n-h} \hat{\Pi}_i y_{h+i} \right)$$

  eine Schätzung des Effektes, falls ein additiver Ausreißer zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch $Var(\hat{w}_{A,h}) = \rho^2_{A,h}\sigma^2_a$.

- Für einen Ausreißer vom Typ „T“ setze $\beta(B) = (1 - \delta B)^{-1}\hat{\Pi}(B) = \beta_0 - \beta_1 B - \beta_2 B^2 \ldots$ und zusätzlich $\rho^2_{T,h} = (1 + \beta^2_1 + \ldots + \beta^2_{n-h})^{-1}$, dann ist

$$\hat{w}_{T,h} = \rho^2_{T,h} \left(y_h - \sum_{i=1}^{n-h} \beta_i y_{h+i} \right)$$

  eine Schätzung des Effektes, falls ein Ausreißer vom Typ „T“ zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch $Var(\hat{w}_{T,h}) = \rho^2_{T,h}\sigma^2_a$.

- Für einen Ausreißer vom Typ „I“ setze

$$\hat{w}_{I,h} = y_h$$

  als Schätzung des Effektes, falls ein innovativer Ausreißer zum Zeitpunkt h vorliegt. Der Standardfehler kann geschätzt werden durch $Var(\hat{w}_{I,h}) = \sigma^2_a$.

3.7 Ausreißer in multivariaten Zeitreihen

Die in diesem Kapitel vorgestellten Methoden zur Ausreißeranalyse sind Adaptionen der univariaten Verfahren. Um sie anwenden zu können, muss eine multivariate Zeitreihe zunächst auf den eindimensionalen Raum projiziert werden. Dies geschieht aber nicht nur einmal. Die Zeitreihe wird mehrmals in unterschiedliche Richtungen projiziert, so dass viele neue eindimensionale Zeitreihen entstehen. Diese werden anschließend simultan mit univariaten Verfahren untersucht. Für dieses und alle folgenden Kapitel wird, wenn nicht näher spezifiziert, stets angenommen, dass eine stationäre Zeitreihe vorliegt, die sowohl eine VARMA, als auch eine VAR und VMA Darstellung besitzt, vgl. Abschnitt 3.5.1.1.

3.7.1 Projektion einer multivariaten Zeitreihe

Die Projektion multivariater Zeitreihen ist ein mächtiges Werkzeug, welches viele Vorteile gegenüber rein mehrdimensionalen Methoden aufweist. Durch die Projektion multivariater Zeitreihen werden univariate Zeitreihen erzeugt. Diese „Reduktion der Dimension“ vereinfacht die gesamte Analyse, da die Daten überschaubarer werden und kein multivariates Modell angepasst werden muss. Darüber hinaus weisen die, auf projizierte Zeitreihen angewandten, statistischen Tests eine höhere Power auf. Zwar gilt dies zunächst theoretisch nur, wenn die Modellparameter bekannt sind, aber Galeano et al. [GPT06, vgl.] zeigten mit Monte Carlo Simulation, dass dies auch Bestand hat, wenn die Modellparameter aus den Daten geschätzt werden.

3.7.1.1 Wirkungsweise der Projektion

Die Projektion einer multivariaten Zeitreihe und damit auch der in der Zeitreihe enthaltenen multivariaten Ausreißer, liefert eine univariate Zeitreihe mit univariaten Ausreißern. Es ist daher wichtig zu wissen, welche Wirkung die Projektion beispielsweise auf einen multivariaten innovativen Ausreißer hat. Galeano et al. haben die Wirkungsweise der Projektion untersucht und die Ergebnisse werden an dieser Stelle noch einmal kurz zusammengefasst [vgl. GPT06, S.655 f.].

- Falls w ein multivariater Ausreißer vom Typ „A“ zum Zeitpunkt h ist, so besitzt $y_t = v'Y_t$ einen univariaten Ausreißer vom Typ „A“ der Größe $\beta = v'w$ zum Zeitpunkt h, sofern $v'w \neq 0$.

- Falls w ein multivariater Ausreißer vom Typ „L“ zum Zeitpunkt h ist, so besitzt $y_t = v'Y_t$ einen univariaten Ausreißer vom Typ „L“ der Größe $\beta = v'w$ zum Zeitpunkt h, sofern $v'w \neq 0$.

- Falls w ein multivariater Ausreißer vom Typ „T“ zum Zeitpunkt h ist, so besitzt $y_t = v'Y_t$ einen univariaten Ausreißer vom Typ „T“ der Größe $\beta = v'w$ zum Zeitpunkt h, sofern $v'w \neq 0$.

- Falls w ein multivariater Ausreißer vom Typ „I“ zum Zeitpunkt h ist, so ergibt sich in der projizierten Zeitreihe $y_t = v'Y_t$ eine Folge von univariaten Ausreißern der Größe $v'w, v'\Psi_1 w, \ldots, v'\Psi_{n-h} w$, beginnend zum Zeitpunkt $t = h$. Unter der Annahme, dass h nicht nahe n ist und wegen $\Psi_j \to 0$ für $j \to \infty$, konvergiert die Größe der Ausreißer $v'\Psi_j w$ gegen Null. Es können nun noch spezielle Fälle eintreten.

- Falls $v'\Psi_i w = \psi_i v'w,\ i = 1, \ldots, n-h$ gilt, dann besitzt y_t einen innovativen Ausreißer der Größe $v'w$ zum Zeitpunkt $t = h$.
- Falls $v'\Psi_i w = 0,\ i = 1, \ldots, n-h$ gilt, dann besitzt y_t einen additiven Ausreißer der Größe $v'w$ zum Zeitpunkt $t = h$.
- Falls $v'\Psi_i w = v'w,\ i = 1, \ldots, n-h$ gilt, dann besitzt y_t einen level shift der Größe $v'w$ zum Zeitpunkt $t = h$.

Die ψ_i entstammen dabei der MA Darstellung der durch v projizierten Zeitreihe $y_t = v'Y_t$, sprich $y_t = c_\psi + \psi(B)a_t$ mit $\psi(B) = 1 + \sum_{i=1}^{\infty} \psi_i B^i$.

3.7.1.2 Zusammenhang zwischen Projektionsrichtungen und Ausreißern

Zentral in diesem Abschnitt ist die Frage nach den „richtigen“ Projektionsrichtungen v und wie man diese generiert. Hier werden die wichtigsten theoretischen Grundlagen und Ergebnisse aus Galeano et al. [vgl. GPT06] zusammengestellt.

Definiere für die multivariate Zeitreihe $Y_1, \ldots, Y_n$ zunächst die deterministische Folge $(R_t)_{t=1\ldots,n}$ mit

$$R_t := \alpha(B) w I_t^{(h)} \ .$$

Es sei an dieser Stelle daran erinnert, dass Y_t folgende Darstellung, wie im Abschnitt 3.5.1.3 erläutert wurde, besitzt.

$$Y_t = X_t + \alpha(B) w I_t^{(h)}$$

Dann gilt offenbar $Y_t = X_t + R_t$. Die Projektion von Y_t in Richtung eines beliebigen Vektors v ist gegeben durch $y_t = v'Y_t = v'X_t + v'R_t = x_t + r_t$. Es seien $\bar{R} = \frac{1}{n}\sum_{t=1}^{n} R_t$ und $\tilde{R}_t := R_t - \bar{R}$, analog dazu $\bar{Y}$ und $\tilde{Y}_t$, sowie $\bar{A}$ und $\tilde{A}_t$ aus

Formel (3.12). Weiter seien die folgenden Kovarianzschätzungen definiert.

$$\Sigma_R := \frac{1}{n}\sum_{t=1}^{n} \tilde{R}\tilde{R}'$$
$$\Sigma_Y := \frac{1}{n}\sum_{t=1}^{n} \tilde{Y}\tilde{Y}' = I + \Sigma_R = I + \beta_i w w'$$
$$\Sigma_A := \frac{1}{n}\sum_{t=1}^{n} \tilde{A}\tilde{A}'$$

Dabei ergeben sich die β_i für i=A,T,L aus den folgenden Gleichungen.

$$\beta_A = \frac{n-1}{n^2}$$
$$\beta_L = \frac{n-h+1}{n}\left(\frac{h-1}{n}\right)$$
$$\beta_T = \frac{1}{n}\left[\left(\frac{1-\delta^{2(n-h+1)}}{1-\delta^2}\right) - \frac{1}{n}\left(\frac{1-\delta^{n-h+1}}{1-\delta}\right)^2\right]$$

In der praktischen Umsetzung werden die Innovationen A_t durch die aus der Modellschätzung resultierenden Residuen ersetzt.

In der weiteren Analyse genügt es, sich ohne Beschränkung der Allgemeinheit auf Projektionsrichtungen v zu beschränken, die $v'\Sigma_Y v = 1$ erfüllen. Die Richtung, in der ein multivariater Ausreißer w zu finden ist, errechnet sich, indem zunächst der Kurtosiskoeffizient $\gamma_y(v)$ der projizierten Zeitreihe $y_t = v'Y_t$ maximiert oder minimiert wird. Dieser hängt selbstverständlich von der Projektionsrichtung v ab und ist unter Berücksichtigung der Nebenbedingung $v'\Sigma_Y v = 1$

gegeben durch

$$\gamma_y(v) = 3 - 3\left(v'\Sigma_R v\right)^2 + \nu_r(v) \text{ mit } \nu_r(v) = \frac{1}{n}\sum_{t=1}^{n} \tilde{r}_t^4 \ .$$

Die maximierenden und minimierenden Richtungen v von $\gamma_y(v)$ ergeben sich als Eigenvektoren der Matrix $\sum_{t=1}^{n} \beta_t(v)\tilde{R}_t\tilde{R}_t'$ mit den zugehörigen Eigenwerten $\mu(v) = n\left(v'\Sigma_R v\right)^2(\gamma_r(v) - 3)$. Es stellt $\gamma_r(v)$ in diesem Zusammenhang den Kurtosiskoeffizienten der Reihe r_t dar. Es gilt weiter

$$\beta_t(v) = \left(v'\tilde{R}_t\right)^2 - 3\left(v'\Sigma_R v\right) - \frac{\mu(v)}{n} \ .$$

Die den Kurtosiskoeffizienten $\gamma_y(v)$ maximierenden bzw. minimierenden Eigenvektoren korrespondieren zum maximalen bzw. minimalen Eigenwert $\mu(v)$. Je nach Ausreißertyp gilt nun folgendes.

- Für einen Ausreißervektor w vom Typ „A“ wird der Kurtosiskoeffizient $\gamma_y(v)$ maximal, wenn die Projektionsrichtung v ein Vielfaches von w ist und $\gamma_y(v)$ wird minimal, wenn v orthogonal zu w ist.

- Für einen Ausreißervektor w vom Typ „T“ wird der Kurtosiskoeffizient $\gamma_y(v)$ maximal oder minimal, wenn die Projektionsrichtung v ein Vielfaches von w ist und $\gamma_y(v)$ wird minimal oder maximal, wenn v orthogonal zu w ist.

- Für einen Ausreißervektor w vom Typ „L“ wird der Kurtosiskoeffizient $\gamma_y(v)$ minimal, wenn die Projektionsrichtung v ein Vielfaches von w ist

und $\gamma_y(v)$ wird maximal, wenn v orthogonal zu w ist, falls gilt

$$h \in \left(1+\frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right)n, 1+\frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)n\right) \quad .$$

Des Weiteren gilt für einen Ausreißervektor w vom Typ „L“, dass der Kurtosiskoeffizient $\gamma_y(v)$ maximal wird, wenn die Projektionsrichtung v ein Vielfaches von w ist und $\gamma_y(v)$ wird minimal, wenn v orthogonal zu w ist, falls gilt

$$h \notin \left(1+\frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right)n, 1+\frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)n\right) \quad .$$

Der Parameter h gibt den Zeitpunkt des Auftretens eines level shifts, sprich eines Ausreißers vom Typ „L“ an. Er tritt in der Darstellung der Zeitreihe $Y_t = X_t + \alpha(B)wI_t^{(h)}$ auf.

- Für einen Ausreißer w vom Typ „I“ gilt, dass, analog zu der bisherigen Vorgehensweise, der Kurtosiskoeffizient $\gamma_a(v)$ der projizierten Innovationen $a_t = v'A_t$ maximal wird, falls v ein Vielfaches von w ist und der Kurtosiskoeffizient wird minimal, wenn v orthogonal zu w ist.

Ein sinnvolles Verfahren, anhand dessen alle oben genannten möglichen Kombinationen untersucht werden können, ist die Maximierung und Minimierung des Kurtosiskoeffizienten $\gamma_y(v)$. Im Anschluss daran wird jeweils zum maximierenden und minimierenden Vektor v eine Orthogonalbasis des $\mathbb{R}^k$ gebildet, sprich jeweils $k-1$ zum maximierenden und minimierenden v senkrechte Vektoren. Somit werden bei Vorliegen eines Ausreißers alle Möglichkeiten der Orthogonalität oder Kollinearität von v und w berücksichtigt.

3.7.1.3 Algorithmus zum Auffinden von Projektionsrichtungen

Dieser Algorithmus stammt von Galeano et al. [vgl. GPT06, S.]. Die zentrale Aufgabe besteht in der Maximierung bzw. Minimierung der Gleichung (3.23). Wie dies im Detail durchgeführt werden kann, wird in Abschnitt 3.7.1.4 erläutert.

1. Setze $m = 1$, $Z_t^{(m)} = Y_t$ und $\tilde{Z}_t^{(m)} = Z_t^{(m)} - \bar{Z}_t^{(m)}$.

2. Setze $\Sigma_Z^{(m)} = \frac{1}{n} \sum_{t=1}^{n} \tilde{Z}_t^{(m)} \tilde{Z}_t^{(m)'}$ und finde das maximierende Argument

$$v_m = \arg \max_{v_m' \Sigma_Z^{(m)} v_m = 1} \frac{1}{n} \sum_{t=1}^{n} \left(v_m' \tilde{Z}_t^{(m)} \right)^4 \tag{3.23}$$

3. Falls $m = k$ beende die Prozedur. Ansonsten definiere

$$Z_t^{(m+1)} = \left(I - v_m v_m' \Sigma_Z^{(m)} \right) Z_t^{(m)},$$

   setze $m = m + 1$ und gehe zu Schritt 2.

4. Wiederhole die Schritte 1-3 mit dem Unterschied, dass in Gleichung (3.23) minimiert, statt maximiert wird. Dies liefert, analog zur Vorgehensweise bei der Maximierung des Kurtosiskoeffizienten 3.23, k orthogonale Vektoren.

Obiger Algorithmus generiert folglich insgesamt $2k$ Vektoren, wobei jeweils die ersten k und die letzten k orthogonal sind. Das Verfahren ist affin äquivariant, d.h. unter einer affinen Transformation der Daten liefert die Methode äquivalente Projektionsrichtungen. Die Maximierung bzw. Minimierung der Gleichung (3.23) stellt hier die eigentliche Aufgabe dar. Galeano et al. [vgl. GPT06, S.659] schlagen die Verwendung eines modifizierten Newton Verfahrens mittels linearer Approxi-

mation der Optimalitätsbedingungen $\nabla\gamma_z(v) - 2\lambda\Sigma_Z^{(m)}v = 0$ und $v'\Sigma_Z^{(m)}v - 1 = 0$ vor, wobei $\nabla\gamma_z$ den Gradienten von γ_z darstellt.

3.7.1.4 Verfahren zur Bestimmung der Eigenwerte und Eigenvektoren

Alternativ zum von Galeano et al. [vgl. GPT06, S.659] vorgeschlagenen modifizierten Newton-Verfahren lässt sich das Optimierungsproblem (3.23) durch die von Pena und Prieto diskutierte Alternative einer iterativen Bestimmung von Eigenwerten noch zu definierender Matrizen lösen [vgl. PP01a, S.1438]. Dazu wird im Folgenden das Optimierungsproblem nochmals formuliert, die zugehörigen Bedingungen erster Ordnung kombiniert und ein Algorithmus angegeben, um iterativ die gesuchte Lösung zu bestimmen.

$$\max \frac{1}{n}\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4$$

$$\text{so dass } v'\Sigma_Z^{(m)}v - 1 = 0$$

Zunächst ist festzustellen, dass der Faktor $\frac{1}{n}$ auf die Lösung des Optimierungsproblemes keinerlei Einfluss hat. Er muss daher in den folgenden Betrachtungen nicht weiter berücksichtigt werden. Als nächstes wird die Lagrangesche Funktion aufgestellt.

$$L(v_m) := \sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4 - \lambda\left(v_m'\Sigma_Z^{(m)}v_m - 1\right)$$

Die Optimalitätsbedingungen ergeben sich durch Bildung des Gradienten.

$$4\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^3\tilde{Z}_t^{(m)} - 2\lambda\Sigma_Z^{(m)}v_m = 0$$

$$\text{so dass } v_m'\Sigma_Z^{(m)}v_m = 1$$

Weiter wird die erste Gleichung von links mit v_m' multipliziert und nach λ aufgelöst.

$$4\sum_{t=1}^{n}\underbrace{\left(v_m'\tilde{Z}_t^{(m)}\right)^3 v_m'\tilde{Z}_t^{(m)}}_{=\left(v_m'\tilde{Z}_t^{(m)}\right)^4} - 2\lambda\underbrace{v_m'\Sigma_Z^{(m)}v_m}_{=1} = 0$$

$$\Leftrightarrow \lambda = 2\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4$$

Diese letzte Gleichung wird in die Optimalitätsbedingungen eingesetzt.

$$4\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^3\tilde{Z}_t^{(m)} - 2\left(2\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4\right)\Sigma_Z^{(m)}v_m = 0$$

$$\Leftrightarrow 4\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^3\tilde{Z}_t^{(m)} = 4\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4\Sigma_Z^{(m)}v_m$$

Die linke Seite wird unter Berücksichtigung folgender Gleichheit modifiziert.

$$\begin{aligned}\left(v_m'\tilde{Z}_t^{(m)}\right)^3\tilde{Z}_t^{(m)} &= \left(v_m'\tilde{Z}_t^{(m)}\right)^2\underbrace{\left(v_m'\tilde{Z}_t^{(m)}\right)}_{\in\mathbb{R}}\tilde{Z}_t^{(m)}\\ &= \left(v_m'\tilde{Z}_t^{(m)}\right)^2\tilde{Z}_t^{(m)}\underbrace{\left(v_m'\tilde{Z}_t^{(m)}\right)}_{=Z_t^{(m)\prime}v_m}\\ &= \left(v_m'\tilde{Z}_t^{(m)}\right)^2\tilde{Z}_t^{(m)}\tilde{Z}_t^{(m)\prime}v_m\end{aligned}$$

Damit ergibt sich

$$4\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^2\tilde{Z}_t^{(m)}\tilde{Z}_t^{(m)'}v_m = \underbrace{4\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4}_{\in\mathbb{R}}\Sigma_Z^{(m)}v_m \ .$$

Unter der Annahme, dass die inverse Matrix $\left(\Sigma_Z^{(m)}\right)^{-1}$ existiert, ist es möglich beide Seiten mit $\left(\Sigma_Z^{(m)}\right)^{-1}$ zu multiplizieren. Dies liefert die finale Darstellung.

$$\left[\sum_{t=1}^{n}\left(\Sigma_Z^{(m)}\right)^{-1}\left(v_m'\tilde{Z}_t^{(m)}\right)^2\tilde{Z}_t^{(m)}\tilde{Z}_t^{(m)'}\right]v_m = \left[\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4\right]v_m$$

Es ist folglich v_m als Eigenvektor der Matrix

$$Q(v_m) = \left[\sum_{t=1}^{n}\left(\Sigma_Z^{(m)}\right)^{-1}\left(v_m'\tilde{Z}_t^{(m)}\right)^2\tilde{Z}_t^{(m)}\tilde{Z}_t^{(m)'}\right]$$

zum Eigenwert

$$\left[\sum_{t=1}^{n}\left(v_m'\tilde{Z}_t^{(m)}\right)^4\right]$$

zu bestimmen. Ein iteratives Verfahren zur Bestimmung des gesuchten Vektors v_m wird im Folgenden beschrieben. Die inverse Matrix $\left(\Sigma_Z^{(m)}\right)^{-1}$ existiert genau dann, wenn $\Sigma_Z^{(m)}$ regulär ist, d.h. wenn $det(\Sigma_Z^{(m)}) \neq 0$ gilt.

3.7.1.5 Algorithmus zur Bestimmung von v_m

Dieser Algorithmus gleicht im Wesentlichen dem von Pena und Prieto vorgeschlagenen [vgl. PP01a, S.1438].

1. Wähle ein w_0, welches der Nebenbedingung $w_0'\Sigma_Z^{(m)}w_0 = 1$ genügt.

2. Setze $l = 1$.

3. Bestimme die Eigenvektoren der Matrix $Q(w_{l-1})$. Die seien bezeichnet mit $w_l^1, \ldots, w_l^{k_l}$.

4. Normiere die gefundenen Eigenvektoren $w_l^1, \ldots, w_l^{k_l}$ in dem Sinne, dass sie der Bedingung $w_l^{j'} \Sigma_Z^{(m)} w_l^j = 1$ genügen ($j = 1, \ldots, k_l$).

5. Wähle den Eigenvektor w_l^j, der die Zielfunktion maximiert.

6. Falls $\|w_l - w_{l-1}\| < \varepsilon$ setze $v_m := w_l$ und der Algorithmus stoppt. Ansonsten setze $l := l + 1$ und gehe zu Schritt 3.

Um die Normierung in Schritt 4 durchzuführen, bestimme für $j = 1, \ldots, k_l$ die Werte $u_m^j := w_m^{j'} \Sigma_Z^{(m)} w_m^j$. Anschließend definiere $w_m^j := \frac{1}{\sqrt{u_m^j}} w_m^j$. Dann genügen $w_m^1, \ldots, w_m^{k_l}$ der Nebenbedingung. Die Existenz der reellen Wurzel $\sqrt{w_m^i}$ ist gesichert, da $\Sigma_Z^{(m)}$ als Kovarianzmatrix positiv definit ist, d.h. $v' \Sigma_Z^{(m)} v > 0$ für alle v. Das Minimierungsproblem wird vollkommen analog behandelt.

3.7.2 Schätzung des Effektes eines Ausreißers in multivariaten Zeitreihen

Der theoretische Ansatz besagt, dass, bei bekannten Modellparametern, der Effekt $w_{i,h}$ eines Ausreißers vom Typ i=I,A,L,T zum Zeitpunkt h und unter der Nullhypothese „H_0 : Die multivariate Zeitreihe ist nicht mit Ausreißern kontaminiert“ nach folgender Formel für das Zeitreihenmodell in VAR Darstellung $\Pi(B) Y_t = A_t$ mit $\Gamma(B) = \Pi(B)\alpha(B) = I - \sum_{i=1}^{\infty} \Gamma_i B^i$ berechnet werden kann

[vgl. GPT06, S.655].

$$w_{i,h} = -\left(\sum_{j=0}^{n-h} \Gamma_j' \Sigma^{-1} \Gamma_j\right)^{-1} \left(\sum_{j=0}^{n-h} \Gamma_j' \Sigma^{-1} A_{h+j}\right), \ \Gamma_0 = -I$$

Die zugehörige Kovarianzmatrix hat die folgende Gestalt.

$$\Sigma_{i,h} = \left(\sum_{j=0}^{n-h} \Gamma_j' \Sigma^{-1} \Gamma_j\right)^{-1}$$

In der Praxis schätzt man zunächst mit Hilfe statistischer Software wie **R** unter Verwendung der Funktion `ar` die multivariate AR Darstellung der Zeitreihe $Y_1, \ldots, Y_n$. Die durch die Software automatisch generierten Schätzungen der Koeffizientenmatrizen $\Pi_1, \ldots, \Pi_p$ seien bezeichnet mit $\hat{\Pi}_1, \ldots, \hat{\Pi}_p$. Die Schätzung der Kovarianzmatrix lässt sich ebenfalls mit programmtechnischer Unterstützung bestimmen, bzw. wird in der Funktion `ar` automatisch mit bestimmt. Damit lassen sich die Koeffizienten des Matrixpolynoms einfach berechnen, durch $\Gamma(B) = \hat{\Pi}(B)\alpha(B)$.

Unter H_0 lässt sich die multivariate Teststatistik $J_{i,h}$ gewinnen, mit deren Hilfe man einen eventuellen Ausreißer auf Signifikanz prüfen kann.

$$J_{i,h} = w_{i,h}' \Sigma_{i,h}^{-1} w_{i,h}$$

Diese ist χ^2-verteilt mit k Freiheitsgraden und Nichtzentralitätsparameter $\eta_i = w' \Sigma_{i,h}^{-1} w$, wenn w ein Ausreißer ist [vgl. GPT06, S.655]. Unter der Nullhypothese $H_0 : \ w = 0$, sprich „$H_0 : w$ ist kein Ausreißer“, bekanntem Koeffizientenpolynom $\Pi(B)$ und normalverteilten Innovationen A_t folgt $J_{i,h}$ einer zentralen

χ^2-Verteilung mit k Freiheitsgraden [vgl. GPT06, S.655], [vgl. TPP98, S.7].

In der Praxis sind die Koeffizientenmatrizen (Π und Γ) und die Modellparameter (Σ) unbekannt. Daher werden sie unter der Nullhypothese „Die Zeitreihe ist nicht mit multivariaten Ausreißern kontaminiert“ durch konsistente Schätzungen ersetzt. Dabei korrespondiert Σ, als Kovarianzmatrix der vektorwertigen Innovationen A_t, zur Varianz σ_a^2 der Innovationen im univariaten Modell. Diese lassen sich am einfachsten mit statistischer Software, wie z.B. **R** bestimmen. Auf die explizite Angabe von iterativen Bestimmungsgleichungen wird an dieser Stelle verzichtet. Die Schätzung der Koeffizientenmatrizen mittels einer Software ist für praktische Zwecke ausreichend. Auf eine weiterführende Theorie soll an dieser Stelle verzichtet werden. Details finden sich z.B. bei Lütkepohl [vgl. Lüt05, S.69 ff.].

Ein mutmaßlicher Ausreißer $w_{i,h}$ vom Typ i zur Zeit h ist genau dann statistisch signifikant, wenn die Teststatistik $J_{i,h}$ den kritischen Wert $\chi^2_{k,\alpha}$ der χ^2-Verteilung mit k Freiheitsgraden auf dem Signifikanzniveau α überschreitet.

Dieses Verfahren ist allgemein für $k \geq 1$ einsetzbar und deckt somit den uni- und den multivariaten Fall ab.

Das allgemeine Vorgehen bei der Untersuchung von Zeitreihen gliedert sich in vier zentrale Schritte.

1. Falls eine multivariate Zeitreihe der Dimension k vorliegt, werden die optimalen $2k$ Projektionsrichtungen bestimmt, um univariate Zeitreihen zu er-

zeugen.

2. In den univariaten Zeitreihen wird nach Ausreißern gesucht.

3. Der Effekt aller gefundenen Ausreißer wird unter Verwendung eines approximativen multivariaten Modells aus den Daten entfernt.

4. Die Schritte 1-3 werden solange wiederholt, bis keine Ausreißer mehr gefunden werden.

3.7.3 Bereinigung multivariater Zeitreihen von level shifts

Im Folgenden wird ein Algorithmus zur iterativen Bestimmung von Zeitpunkten, an denen ein level shift in multivariaten Zeitreihen auftritt, angegeben. Bei diesem Verfahren werden die Daten auch von den Effekten der level shifts bereinigt. Die kritischen Werte der zugehörigen Teststatistik D_L sind in Tabelle 3.10 gegeben. Es seien $y_{1,i}, \ldots, y_{n,i}$ mit $i = 1, \ldots, 2k$ die $2k$ Projektionen der k-dimensionalen, multivariaten Zeitreihe $Y_1, \ldots, Y_n$ auf die reelle Achse, sprich $y_{t,i} = v_i' Y_t$, $t = 1, \ldots, n$. Das Verfahren zur Bestimmung der Projektionsrichtungen v_i ist im Abschnitt 3.7.1.3 beschrieben.

3.7.3.1 Algorithmus zur Bereinigung multivariater Zeitreihen von level shifts

Das hier vorgestellte Verfahren wurde von Galeano et al. entwickelt [vgl. GPT06, S.659]. Es arbeitet ähnlich dem für univariate Zeitreihen, mit dem Unterschied, dass durch das Projektionsverfahren mehrere eindimensionale Zeitreihen vorliegen. Diese werden simultan untersucht.

Seien $y_{t,i}$ die projizierten Zeitreihen ($i = 1, \ldots, 2k$, für $k > 1$).

1. Setze $t_1 = 1$ und $t_2 = n$.

$$D_L = \max_{1 \le i \le 2k} \max_{t_1 \le t \le t_2} \left| C_t^i \right| \tag{3.24}$$

Dabei korrespondiert C_t^i zu C_t in Gleichung (3.19) mit der i-ten Projektion $(y_{t,i})_{t=1,\ldots,n}$. Als nächstes bestimme die Argumente, sprich Indizes i und t, für die D_L maximal wird. Lege zunächst für die Zeitreihe einen Mindestabstand H zwischen zwei Zeitpunkten von level shifts fest, um eine Überbestimmung von level shifts zu vermeiden (z.B. H=10).

$$(i_{\max}, t_{\max}) = \arg \max_{1 \le i \le 2k} \arg \max_{t_1 \le t \le t_2} \left| C_t^i \right| \tag{3.25}$$

Der kritische Wert zum Signifikanzniveau α sei $D_{L,\alpha}$. Falls $D_L > D_{L,\alpha}$ ist, dann liegt möglicherweise ein level shift zum Zeitpunkt $t = t_{\max} + 1$ vor. Falls $D_L < D_{L,\alpha}$ gilt, stoppt der Algorithmus.

2. (a) Setze $t_2 = t_{\max}$ in Schritt 1 und bestimme D_L, sowie $(i_{\max}, t_{\max})$ wie in Gleichung (3.24) und Gleichung (3.25). Falls $D_L > D_{L,\alpha}$ und $t_2 - t_{\max} > H$, dann setze $t_2 = t_{\max}$ und wiederhole Schritt 2.a) bis entweder $D_L < D_{L,\alpha}$ oder $t_2 - t_{\max} \le H$ gilt. Setze nun $t_{first} = t_2$, wobei t_2 der letzte Zeitpunkt ist, der die Ungleichungen $D_L > D_{L,\alpha}$ und $t_2 - t_{\max} > H$ erfüllt. Der Zeitpunkt $t = t_{first} + 1$ ist der, des ersten möglichen level shifts.

(b) Setze $t_1 = t_{\max}$ und $t_2 = n$ in Schritt 1. Berechne neue Werte nach Gleichung (3.24) und Gleichung (3.25). Wenn $D_L > D_{L,\alpha}$ und $t_{\max} - t_1 > H$, setze $t_1 = t_{\max}$ und wiederhole Schritt 2.b) solange, bis entweder $D_L < D_{L,\alpha}$ oder $t_{\max} - t_1 \leq H$. Setze $t_{last} = t_1$, wobei t_1 der letzte Zeitpunkt ist, der die Ungleichungen $D_L > D_{L,\alpha}$ und $t_{\max} - t_1 > H$ erfüllt. Der Zeitpunkt $t = t_{last} + 1$ ist der, des letzten möglichen level shifts.

(c) Falls $t_{last} - t_{first} < H$, stoppt der Algorithmus. Ansonsten wiederhole die Schritte 2.a) und 2.b) mit $t_1 = t_{first}$ und $t_2 = t_{last}$ bis keine weiteren möglichen Zeitpunkte für level shifts berechnet werden können. Dann gehe zu Schritt 3.

3. Definiere den Vektor $h^L = (h_0^L, \ldots, h_{r_L+1})$ mit $h_0^L = 1$, $h_{r_L+1}^L = n$ und den aufsteigend geordneten Zeitpunkten der möglichen level shifts $h_1^L < \cdots < h_{r_L}^L$, die in Schritt 2 bestimmt wurden. Berechne für jedes zeitliche Teilintervall (h_i^L, h_{i+2}^L) den Wert D_L wie in Gleichung (3.24) mit $t_1 = h_i^L$ und $t_2 = h_{i+2}^L$ und prüfe die statistische Signifikanz, sprich: gilt $D_L > D_{L,\alpha}$? Falls dies nicht gilt, entferne h_{i+1}^L aus dem Vektor h^L. Wiederhole diesen Schritt solange, bis nur noch signifikante D_L vorliegen. Entferne $h_0^L = 1$ und $h_{r_L+1}^L = n$ aus h^L. Aus den Zeitpunkten $h^L = (h_1^L, \ldots, h_{r_L}^L)$ errechnen sich die Zeitpunkte der level shifts nur durch $h_i^L + 1$, $i = 1, \ldots, r_L$.

4. Seien $\{h_1^L, \ldots, h_{r_L}^L\}$ die Zeitpunkte der level shifts. Es werden nun die Auswirkungen der level shifts aus der Zeitreihe entfernt. Dazu wird zunächst das Modell

$$(I - \Pi_1 B - \ldots - \Pi_{\hat{p}}) B^{\hat{p}} Y_t^* = A_t^* \tag{3.26}$$

für alle $0 \leq \hat{p} \leq p_{\max}$ mit $Y_t^* = Y_t - \sum_{i=1}^{r_L} w_i S_t^{\left(h_i^L\right)}$ geschätzt. Anschließend wählt man das $\hat{p}$, welches die folgende Bedingung erfüllt, aus.

$$\hat{p} = \arg \max_{0 \leq p \leq p_{\max}} AIC(p) = \arg \max_{0 \leq p \leq p_{\max}} \left\{ \log \det(\hat{\Sigma}_p) + 2\frac{k^2 p}{n} \right\}$$

AIC bezeichnet das Akaike Information Criterion. Dabei gilt jeweils $\hat{\Sigma}_p = \frac{1}{n-2p-1} \sum_{t=p+1}^{n} A_t^* A_t^{*'}$ und A_t^* sind die zu Y_t^* korrespondierenden Innovationen. Anders gesagt: mit Hilfe einer Statistiksoftware (z.B. **R**) wird ein vektorwertiges AR[$\hat{p}$] Modell angepasst. Das Optimalitätskriterium ist AIC. Falls nichtsignifikante level shifts auftreten, ist iterativ der am wenigsten signifikante aus Gleichung (3.26) zu entfernen und das Modell neu zu schätzen. Diese schrittweise Prozedur wird wiederholt, bis alle level shifts signifikant sind.

		$\alpha = 0.05$			$\alpha = 0.01$		
n	k	$\Lambda_I, \Lambda_A, \Lambda_T$	Λ_L	D_L	$\Lambda_I, \Lambda_A, \Lambda_T$	Λ_L	D_L
50	2	3.5	2.9	1.26	3.8	3.2	1.33
	3	3.8	3.0	1.28	4.0	3.4	1.35
	5	4.0	3.3	1.31	4.2	3.7	1.36
	10	4.6	3.7	1.34	4.9	4.2	1.39
100	2	3.7	3.1	1.34	4.1	3.4	1.43
	3	3.9	3.2	1.36	4.2	3.5	1.45
	5	4.2	3.4	1.44	4.6	3.9	1.51
	10	5.0	3.8	1.49	5.3	4.2	1.53
200	2	3.9	3.2	1.40	4.2	3.6	1.50
	3	4.1	3.3	1.42	4.5	3.7	1.55
	5	4.4	3.6	1.54	4.8	3.9	1.67
	10	5.2	3.8	1.62	5.6	4.3	1.74
300	2	4.0	3.3	1.43	4.3	3.7	1.56
	3	4.2	3.4	1.47	4.6	3.8	1.60
	5	4.6	3.6	1.54	5.0	4.1	1.70
	10	5.3	3.9	1.67	5.7	4.3	1.81
400	2	4.1	3.4	1.44	4.5	3.8	1.58
	3	4.3	3.5	1.48	4.7	3.9	1.62
	5	4.6	3.6	1.57	5.1	4.2	1.73
	10	5.4	3.9	1.68	5.8	4.3	1.82
500	2	4.2	3.4	1.45	4.5	3.8	1.59
	3	4.3	3.5	1.49	4.8	3.9	1.63
	5	4.7	3.7	1.59	5.1	4.1	1.78
	10	5.5	4.0	1.69	5.9	4.4	1.84

Tabelle 3.10: Kritische Werte zur Bestimmung von Ausreißern nach Galeano et al. [vgl. GPT06, S.664]

3.7.4 Bestimmung von Ausreißern des Typs „A“, „T“ und „I“ in multivariaten Zeitreihen

Sei $Y_1^*, \ldots, Y_n^*$ eine von level shifts bereinigt, multivariate Zeitreihe, welche möglicherweise Ausreißer enthält. Das zugehörige Modell lautet dann $Y_t^* = X_t + w\alpha(B)I_t^{(h)}$, mit $\Phi(B)X_t = C + \Theta(B)E_t$. Dabei ist $(X_t)_{t=1,\ldots,n}$ ein zugrunde liegender, unbeobachteter VARMA Prozess, der keine Ausreißer enthält. Für $\alpha(B)$ ergeben sich folgende Darstellungen [vgl. GPT06, S.654 f.].

- Sei w vom Typ „A“, dann gilt $\alpha(B) = I$ (I ist die k-dimensionale Einheitsmatrix).
- Sei w vom Typ „T“, dann gilt $\alpha(B) = (I - \delta IB)^{-1}$ mit $0 < \delta < 1$.
- Sei w vom Typ „I“, dann gilt $\alpha(B) = \Psi(B) = \Phi(B)^{-1}\Theta(B) = \Pi(B)^{-1}$.

Die unbeobachtete Zeitreihe $X_1, \ldots, X_n$ besitze eine VARMA und eine VAR Darstellung, vgl. Abschnitt 3.5.1.1 und Galeano et al. [GPT06]. Die folgenden Schritte sind unter der Nullhypothese

„H_0 : die Zeitreihe $Y_1^*, \ldots, Y_n^*$ ist nicht mit Ausreißern kontaminiert“

zulässig. Unter H_0 gilt, dass die beobachtete Zeitreihe $Y_1^*, \ldots, Y_n^*$ mit der unbeobachteten, theoretischen Zeitreihe $X_1, \ldots, X_n$ aus dem Ausreißermodell, übereinstimmt und damit auch die Innovationen beider Reihen, also $A_t^* = E_t$. Folglich existiert eine VARMA Darstellung $\Phi(B)Y_t^* = C + \Theta(B)A_t^*$ und eine VAR Darstellung $\Pi(B)Y_t^* = C_\Pi + A_t^*$ der Zeitreihe $Y_1^*, \ldots, Y_n^*$.

Wie im Abschnitt 3.7.1.3 erläutert, werden zunächst die Projektionsrichtungen $v_1, \ldots, v_{2k}$ bestimmt. Diese werden anschließend verwendet, um die $2k$ Projektionen $y_{t,i} = v_i' Y_t^*$ $(i = 1, \ldots, 2k)$ der Zeitreihe , sowie die $2k$ Projektionen der zugehörigen Innovationen $a_{t,i} = v_i' A_t^*$ zu berechnen. Auf diese neu generierten Reihen wird der im Folgenden angegebene Algorithmus angewendet. Additive und transiente Ausreißer findet man durch die Untersuchung der $y_{t,i}$-Reihen, innovative Ausreißer hingegen durch die Untersuchung der $a_{t,i}$-Reihen.

3.7.4.1 Algorithmus zur Bestimmung von Ausreißern des Typs i=A,T,I in multivariaten Zeitreihen

Der hier angegebene Algorithmus stammt von Galeano et al. [vgl. GPT06, S.660]. Man wähle ein Signifikanzniveau α, z.B. $\alpha = 0.05$.

1. Schätze für jede der $2k$ univariaten Zeitreihen $y_{t,i}$ die AR Darstellung, z.B. in **R** und berechne die Teststatistiken $\lambda_{A,t}^i$ und $\lambda_{T,t}^i$. Weiterhin schätze für jede der $2k$ univariaten Zeitreihen $a_{t,i}$ die AR Darstellung und berechne die Teststatistik $\lambda_{I,t}^i$. Wie dies genau geschieht, wird im Abschnitt 3.7.2 vorgestellt. Im Anschluss berechne daraus die folgenden Maximumstatistiken.

$$\Lambda_A = \max_{1 \le i \le 2k} \max_{1 \le t \le n} |\lambda_{A,t}^i|$$

$$\Lambda_T = \max_{1 \le i \le 2k} \max_{1 \le t \le n} |\lambda_{T,t}^i|$$

$$\Lambda_I = \max_{1 \le i \le 2k} \max_{1 \le t \le n} |\lambda_{I,t}^i|$$

2. Seien $\Lambda_{A,\alpha}, \Lambda_{T,\alpha}, \Lambda_{I,\alpha}$ die kritischen Werte der Teststatistiken zum Signifi-

kanzniveau α. Sie sind in Tabelle 3.10 zu finden. Wenn $\Lambda_i < \Lambda_{i,\alpha}$ für alle i=A,T,I gilt, so wurde kein Ausreißer gefunden und der Algorithmus stoppt. Falls $\Lambda_i > \Lambda_{i,\alpha}$ für genau ein i gilt (z.B. $\Lambda_A > \Lambda_{A,\alpha}$ und $\Lambda_T < \Lambda_{T,\alpha}$, $\Lambda_I < \Lambda_{I,\alpha}$), dann liegt ein Ausreißer des zugehörigen Typs i zum korrespondierenden Zeitpunkt vor (in diesem Beispiel also ein Ausreißer vom Typ „A“). Entferne den Effekt des Ausreißers aus der multivariaten Zeitreihe $Y_1^*, \ldots, Y_n^*$. Falls $\Lambda_i > \Lambda_{i,\alpha}$ für mehr als ein i gilt (z.B. $\Lambda_A > \Lambda_{A,\alpha}$, $\Lambda_T > \Lambda_{T,\alpha}$ und $\Lambda_I < \Lambda_{I,\alpha}$), so wähle den signifikantesten Ausreißer (dieser korrespondiert zu der Teststatistik, die den signifikantesten Wert aufweist) und entferne den zugehörigen Effekt des Ausreißers aus der multivariaten Zeitreihe $Y_1^*, \ldots, Y_n^*$. Wiederhole die Schritte 1 und 2 für die, von dem Ausreißer bereinigte Reihe.

3. Seien $\{h_1^A, \ldots, h_{r_A}^A\}$ die Zeitpunkte, an denen in Schritt 2 additive Ausreißer entdeckt und aus der Zeitreihe entfernt wurden, analog dazu die Zeitpunkte zu transienten Ausreißern $\{h_1^T, \ldots, h_{r_T}^T\}$ und innovativen Ausreißern $\{h_1^I, \ldots, h_{r_I}^I\}$. Die zugehörigen Effekte $\{w_1^A, \ldots, w_{r_A}^A\}$, $\{w_1^T, \ldots, w_{r_T}^T\}$ und $\{w_1^I, \ldots, w_{r_I}^I\}$ der gefundenen Ausreißer werden, wie im Abschnitt 3.7.2 erläutert, geschätzt. Anschließend werden sie gemeinsam aus der ursprünglichen multivariaten Reihe $Y_1^*, \ldots, Y_n^*$ durch Subtraktion entfernt und ein neues AR Modell für die, von allen Ausreißern bereinigte, Zeitreihe ge-

schätzt.

$$Y_t^{**} = Y_t^* - \sum_{i=1}^{r_A} w_i^A I_t^{(h_{i_A}^A)} - \sum_{i=1}^{r_T} w_i^T (1-\delta B)^{-1} I_t^{(h_{i_T}^T)}$$
$$A_t^{**} = A_t^* - \sum_{i=1}^{r_I} w_i I_t^{h_{iI}^I}$$

Das zu schätzende Modell ist dann

$$\hat{\Pi}(B) Y_t^{**} = A_t^{**} \quad .$$

Falls nichtsignifikante Ausreißereffekte auftreten, ist iterativ der am wenigsten signifikante zu entfernen und das Modell neu zu schätzen. Diese schrittweise Prozedur wird wiederholt, bis alle Ausreißereffekte signifikant sind.

Die Schätzung des multivariaten AR Modells geschieht z.B. in **R** mit den Funktionen `arima` und `ar`.

Kapitel 4

Auswertung und Diskussion spezieller Datensätze

In diesem Kapitel werden zunächst exemplarisch einzelne Datensätze detailliert ausgewertet, um die Vorgehensweise bei der Identifikation und Bereinigung der Datensätze von Ausreißern in dieser Arbeit zu illustrieren. Die gefundenen Ausreißer und der Vergleich der angewandten Methoden werden direkt für diese Datensätze diskutiert. Da sich die Vorgehensweisen, Feststellungen und Argumentationen zu großen Teilen wiederholen, werden ein univariater und zwei multivariate Querschnittsdatensätze analysiert. Zusätzlich wird eine multivariate Zeitreihe auf zwei verschiedene Arten auf Ausreißer untersucht und detailliert diskutiert. Die hier vorgestellten Querschnittsdatensätze sind repräsentativ für die Gesamtheit der analysierten Daten. Die Auswertung der multivariaten Zeitreihe ist insofern speziell, als dass es sich dabei um die einzige Zeitreihe handelt, in der die Effekte innovativer Ausreißer dadurch entfernt werden können, dass sie als transiente Ausreißer behandelt werden (vgl. Abschnitt 3.5.1.2).

Die Datensätze werden anonymisiert, um mögliche Rückschlüsse auf das betreffende Bundesland oder die betreffende Stadt auszuschließen. Eine Liste der für diese Arbeit verwendeten Messprogramme befindet sich in Anhang A.

Zunächst sollen uni- und multivariate Querschnittsdatensätze untersucht werden. Anschließend wird eine multivariate Zeitreihe auf statistische Ausreißer untersucht und von den entsprechenden Effekten bereinigt. Die Schätzungen der natürlichen Variabilität werden angegeben und diskutiert. Weiterhin sollen die Kongenerenprofile der Querschnittsdatensätze visualisiert werden. Für Zeitreihen sind Kongenerenprofile nur sinnvoll, wenn sie aus den Jahresquerschnittsdaten erstellt werden.

Nach Abschluss der exemplarischen Auswertungen werden die Ergebnisse aller weiteren Auswertungen tabellarisch zusammengestellt und in Abschnitt 5 diskutiert. Dabei steht neben dem Kontaminationsanteil und der Kontaminationswahrscheinlichkeit auch der Einfluss statistischer Ausreißer auf die Schätzung der natürlichen Variabilität im Fokus. Aus den bereinigten Datensätzen werden, sofern möglich, Kongenerenprofile erstellt.

4.1 Querschnittsdaten

4.1.1 Univariate Querschnittsdaten

4.1.1.1 dl PCB

In diesem Abschnitt wird ein Datensatz zu Aalen (Anguilla anguilla) aus dem Jahr 2004 auf statistische Ausreißer untersucht. Der Datensatz beinhaltet $n = 124$ Messwerte zu PCB 118. Daten zu anderen PCB-Kongeneren liegen nicht vor. Die

Konzentrationen sind angegeben in $\mu g/kg$ Frischgewicht. Der Datensatz weist keine fehlenden Werte auf. Die statistische Analyse gestaltet sich folgendermaßen. Zunächst werden die Daten durch Entfernung der Einheit standardisiert.

$$x\frac{\mu g}{kg} = x\frac{10^{-9}kg}{kg} = x \cdot 10^{-9}$$

Anschließend muss der Rest jeder Probe datentechnisch erfasst werden.

$$Rest = 10^9\frac{\mu g}{kg} - x\frac{\mu g}{kg} = 1 - x \cdot 10^{-9}$$

Es liegen nun $n = 124$ Datenpunkte der Gestalt

$$\begin{pmatrix} x \cdot 10^{-9} \\ 1 - x \cdot 10^{-9} \end{pmatrix}$$

vor, d.h. es handelt sich um Kompositionsdaten mit $D = 2$. Zunächst werden alle als Null angegebenen Konzentrationen durch ihre halbe Bestimmungsgrenze ersetzt und der „Rest“ entsprechend angepasst, wie in Abschnitt 3.1.5 beschrieben. Dies betrifft 4 Werte. Für einen dieser Werte ist die Bestimmungsgrenze ebenfalls mit Null angegeben. Er kann nicht ersetzt werden und muss von den weiteren Untersuchungen ausgeschlossen werden, da die ilr Transformierte dieses Wertes nicht berechnet werden kann. Es liegen nunmehr $n = 123$ Wertepaare vor. Diese Datenpunkte werden einer ilr Transformation unterzogen. Der Scatterplot der transformierten Daten ist in Abbildung 4.1 gegeben.

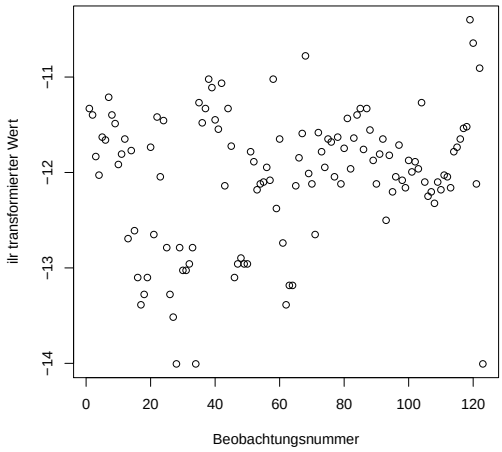

Abbildung 4.1: Scatterplot der ilr transformierten PCB118 Daten zu Aalen im Jahr 2004

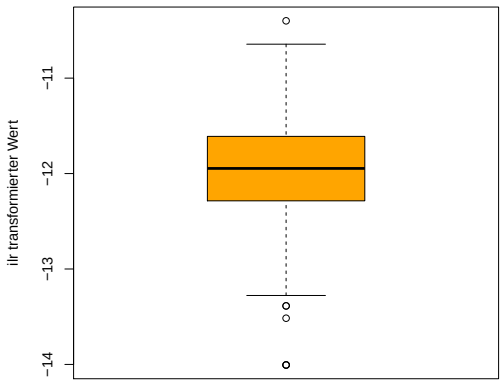

Abbildung 4.2: Boxplot der ilr transformierten PCB118 Daten zu Aalen im Jahr 2004

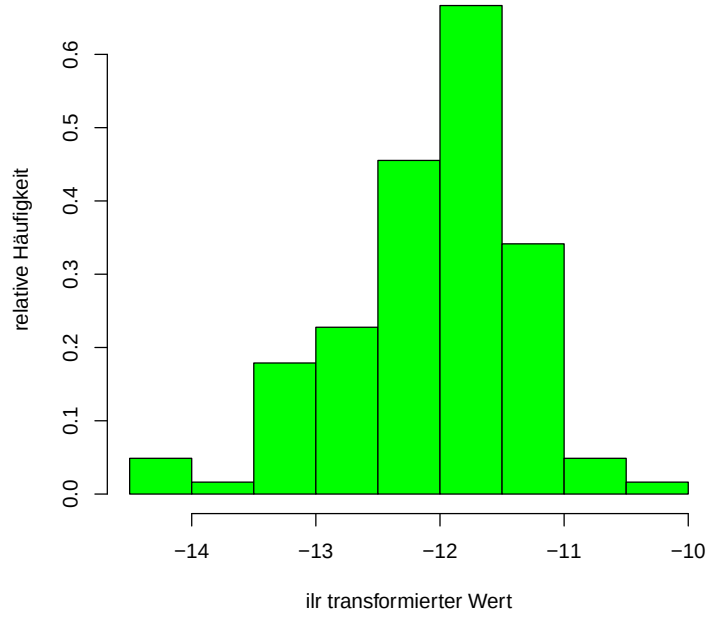

Abbildung 4.3: Histogramm der ilr transformierten PCB118 Daten zu Aalen im Jahr 2004

Die natürliche Variabilität der Daten beträgt 0.4922888.

Als globales Signifikanzniveau wird $\alpha = 0.05$ gewählt und der Hampel-Test zur Identifikation von Ausreißern eingesetzt. Dieser detektiert bei dreifacher Iteration insgesamt 11 statistische Ausreißer in den ilr transformierten Daten. Die natürliche Variabilität der bereinigten Daten beträgt 0.2953863. Sie entspricht ca. 60% der natürlichen Variabilität aller Daten bzw. einer Abnahme der natürlichen Variabilität um ca. 40%. Der Kontaminationsanteil durch Ausreißer in den Daten beträgt ca. 9%.

Betrachtet man die Ausreißer, so ist auffällig, dass alle 3 ursprünglichen Nulldaten als Ausreißer erkannt wurden. Dies kann bedeuten, dass die Ersetzungs-

methode für Nulldaten aus Abschnitt 3.1.5 ungeeignet ist. Diese Ausreißer können aber auch derart interpretiert werden, dass solch niedrige Konzentrationen von PCB 118 ungewöhnlich sind. Insgesamt wurden neben den 3 Nulldaten noch 7 weitere niedrige Konzentrationen als Ausreißer erkannt. Lediglich eine hohe Konzentration von PCB 118 wurde durch das Verfahren als Ausreißer erkannt. Insgesamt ergibt sich, dass sehr niedrige Konzentrationen (die teilweise unterhalb der Bestimmungsgrenze liegen) als statistische Ausreißer erkannt werden, hohe Konzentrationen hingegen nur in einem einzigen Fall. Wie lässt sich dies erklären?

Bedenkt man, dass der Informationsgehalt der Daten in ihren Verhältnissen zueinander liegt und nicht in den Differenzen (vgl. Abschnitt 3.1.4), so scheinen die gefundenen Ausreißer plausibel, denn: liegt eine niedrige Konzentration x nahe Null vor, so wird der Quotient

$$\frac{x}{Rest}$$

ebenfalls nahe Null sein. Null ist der kleinste mögliche Wert und ∞ der größte mögliche Wert, den obiger Quotient annehmen kann. In diesem Sinne sind also sehr große Quotienten oder Quotienten nahe Null als Randwerte zu sehen. Derartige Quotienten ergeben sich lediglich für Konzentrationen x nahe 0 $\mu g/kg$ oder nahe $1 = 10^9$ $\mu g/kg$. Betrachtet man weiterhin die ilr Transformation

$$ilr(x) = \sqrt{\frac{1}{2}} \ln \frac{x}{1-x} \ ,$$

so gilt offenbar $\lim_{x \searrow 0} ilr(x) = -\infty$ und $\lim_{x \nearrow 1} ilr(x) = \infty$.

Abbildung 4.4 zeigt die ilr Transformation für $D = 2$. Die vorliegenden Da-

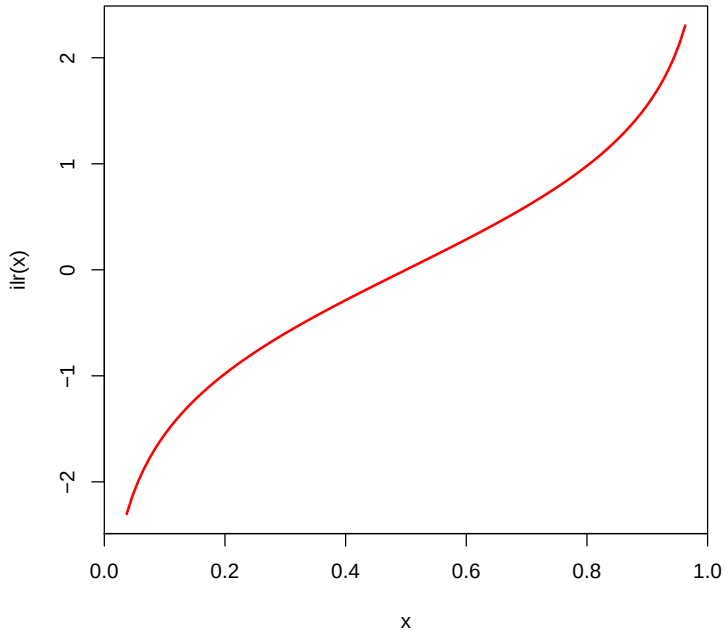

Abbildung 4.4: Die ilr Transformation für $D = 2$

ten konzentrieren sich um einen Median von $46\mu g/kg$. Ihr Maximalwert beträgt $410\mu g/kg$ und der kleinste Wert ungleich Null beträgt $5\mu g/kg$. Aus der Divergenz von $ilr(x)$ für $x \searrow 0$ folgt: je näher eine Konzentration x an Null gelegen ist, desto betragsmäßig größer ist ihr ilr transformierter Wert.

Zwar sind sämtliche gemessenen Konzentrationen nahe Null (der Maximalwert ist 410 $\mu g/kg = 0.000000410$), jedoch gilt weiter: je näher eine Konzentration x zu Null ist, desto größer ist der Anstieg der Funktion $ilr(x)$. Anders formuliert: eine marginale Verringerung der Konzentration x bewirkt eine große Änderung von $ilr(x)$ in Richtung $-\infty$. Dies ist leicht einzusehen, wenn die erste Ableitung

$ilr'(x) = \frac{\partial ilr(x)}{\partial x}$ betrachtet wird.

$$ilr'(x) = \sqrt{\frac{1}{2}} \frac{1}{x(1-x)}$$

Offenbar gilt $\lim_{x \searrow 0} ilr'(x) = \infty$. In Abbildung 4.5 ist $ilr'(x)$ dargestellt.

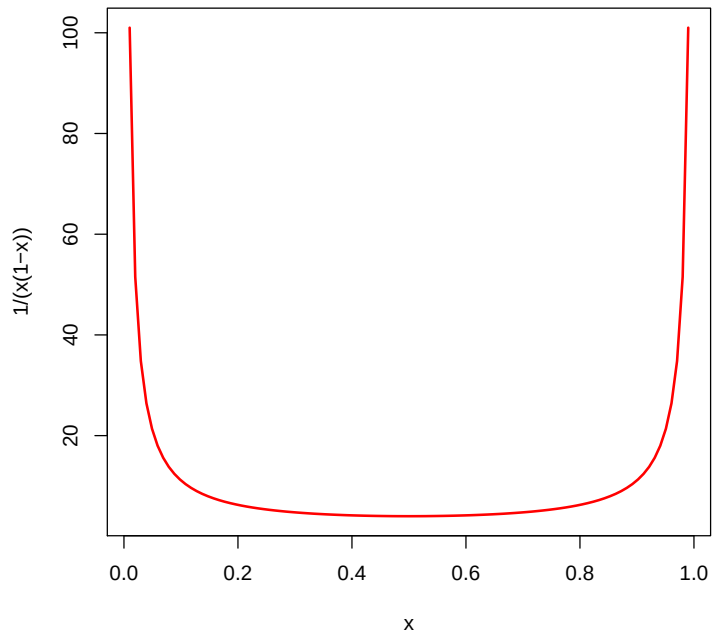

Abbildung 4.5: Die erste Ableitung der ilr Transformation für $D = 2$

Aus rein mathematischer Sicht scheint es somit durchaus plausibel, wenn niedrige Konzentrationen als statistische Ausreißer erkannt werden. Aus rein sachlogischer Sicht ist es wichtig dies kritisch zu hinterfragen. Spiegeln die gefundenen Ausreißer die natürliche Streubreite der in der Natur vorkommenden Konzentrationen von PCB 118 in Aalen wieder, oder gibt es noch weitere (möglicherweise unbekannte) Einflussfaktoren? Lebten die Aale mit niedrigen PCB 118 Konzentra-

tionen in einer wenig belasteten Stelle des Gewässers? Worauf ist die hohe Konzentration von $410 \mu g/kg$ in einem einzigen Aal zurückzuführen? Wurde dieser Aal ein- oder mehrmalig einer höheren Konzentration von PCB 118 ausgesetzt, oder ist die hohe Konzentration von PCB 118 individuell bedingt? Denkbar wäre auch, dass der entsprechende Aal durch Zuwanderung aus einem stärker durch PCB 118 belasteten Gebiet stammt und somit nicht die eigentliche Aalpopulation des Probenahmestandortes repräsentiert. Eine weitere Möglichkeit der Erklärung besteht darin, dass es sich bei der hohen Konzentration möglicherweise um einen Mess- oder Notationsfehler handelt. Da diese Fragen nicht hinreichend genau geklärt werden können, ist es stets notwendig, alle weiteren statistischen Analysen sowohl unter Einbezug der Ausreißer, als auch unter deren Ausschluss durchzuführen. An dieser Stelle werden die durch den Hampel-Test detektierten Ausreißer als solche betrachtet.

4.1.2 Multivariate Querschnittsdaten

4.1.2.1 dl PCB

In diesem Abschnitt wird ein Datensatz zu Brassen (Abramis brama) aus dem Jahr 2003 auf statistische Ausreißer untersucht. Die Proben wurden deutschlandweit genommen. Der Datensatz beinhaltet $n = 20$ Messwerte zu den 12 dl PCB (vgl. Abschnitt 2.3). Die Konzentrationen sind angegeben in $\mu g/kg$ Frischgewicht. Es liegen demnach $n = 20$ multivariate Datenpunkte der Gestalt $x = (x_1, \ldots, x_{12})'$ vor. Die statistische Analyse gestaltet sich folgendermaßen. Zunächst werden die Daten durch Entfernung der Einheit standardisiert und der „Rest“ berechnet, wie in Abschnitt 3.1.4 beschrieben. Es liegen somit Kompositionsdaten der Gestalt

$x = (x_1, \ldots, x_{12}, Rest)'$ (d.h. $D = 13$) vor. Fehlende Werte gibt es nicht im Datensatz. In zwei Beobachtungen existieren Nulldaten. Diese werden wie im Abschnitt 3.1.5 ersetzt und die ilr Transformation angewendet. Die Schätzung der natürlichen Variabilität beträgt 4.693253. Die Scatterplotmatrix der ilr transformierten Daten in Abbildung 4.6 vermittelt einen ersten Eindruck und gibt Hinweise, ob die Daten möglicherweise durch Ausreißer kontaminiert sind. Die Darstellung der Scatterplotmatrix wird erschwert durch die hohe Dimension der Daten, wodurch sich eine 12×12 Matrix ergibt. Die einzelnen Scatterplots werden in der Folge sehr klein dargestellt. An dieser Stelle empfiehlt es sich daher ausdrücklich, die visuelle Analyse des Datensatzes computergestützt am Bildschirm und unter Verwendung einer Zoomfunktion durchzuführen. Die Scatterplots in Zeile 4 deuten bei genauer Betrachtung auf mögliche Ausreißer hin. Eine Analyse der ilr transformierten Daten auf statistische Ausreißer mittels der robusten Mahalanobis-Distanz kann nicht durchgeführt werden, da die Anzahl der Daten zu gering ist. Um statistische Ausreißer unter Verwendung der robusten Mahalanobis-Distanz in d-dimensionalen Daten detektieren zu können, benötigt die von Filzmoser und Varmuze implementierte **R**-Routine `Moutlier` mindestens $n = 2d$ Datenpunkte (vgl. Abschnitt 3.4) [vgl. FV10]. Hier gilt $d = 12$ und $n = 20$.n durchzuführen. Die Scatterplots in Zeile 4 deuten bei genauer Betrachtung auf mögliche Ausreißer hin. Eine Analyse der ilr transformierten Daten auf statistische Ausreißer mittels der robusten Mahalanobis-Distanz kann nicht durchgeführt werden, da die Anzahl der Daten zu gering ist. Um statistische Ausreißer unter Verwendung der robusten Mahalanobis-Distanz in d-dimensionalen Daten detektieren zu können, benötigt die von Filzmoser und Varmuze implementierte **R**-Routine `Moutlier` mindestens $n = 2d$ Datenpunkte (vgl. Abschnitt 3.4) [vgl. FV10]. Hier gilt $d = 12$



Abbildung 4.6: Scatterplotmatrix der ilr transformierten Daten für dl PCB Brassen 2003

und $n = 20$.

Das Projektionsverfahren nach Pena und Prieto wird eingesetzt um statistische Ausreißer zu finden, vgl. Abschnitt 3.4.1. Es wurden sechs Ausreißer erkannt und die Schätzung der natürlichen Variabilität für die bereinigten Daten beträgt 2.763794. Sie entspricht damit ca. 59% der natürlichen Variabilität aller Daten bzw. einer Verringerung der natürlichen Variabilität um ca. 41%.

Zwei der sechs Ausreißer sind Beobachtungen, die ursprünglich Nulldaten enthielten. Die Erkennung dieser Beobachtungen als multivariate Ausreißer kann ein Hinweis darauf sein, dass die Ersetzungsmethode aus Abschnitt 3.1.5 möglicherweise unbrauchbar ist. Die restlichen vier, als Ausreißer erkannte, Datenpunkte enthielten ursprünglich keine Nulldaten.

Ohne Kenntnis weiterer sachlogischer Faktoren kann nicht einfach entschieden werden, ob die gefundenen Ausreißer tatsächlich als solche zu betrachten sind, gerade unter Berücksichtigung der Tatsache, dass sie in einer eher kleinen Stichprobe von $n = 20$ Datenpunkten der Dimension $d = 12$ detektiert wurden. Im Folgenden werden daher die Probenahmestandorte betrachtet, um feststellen zu können, ob sich die Ausreißer ein oder zwei Gebietstypen zuordnen lassen. Unterschieden werden dabei ländliche Räume, Kleinstädte, Mittelstädte, Großstädte, Ballungsräume und Landkreise/Regierungsbezirke. Der Datenanonymität halber werden diese hier jedoch nicht explizit genannt. Die Ausreißer entfallen wie folgt auf die Standortkategorien.

Aus den Standorten kann nicht abgeleitet werden, welchem Standorttyp die statistischen Ausreißer vornehmlich zuzuordnen sind, da Landkreise und Regierungsbezirke zu ungenaue Standortangaben darstellen. Eine weitere Möglichkeit besteht darin, die Einträge in den einzelnen Ausreißervektoren mit dem Balken-

Standorttyp	Anzahl Ausreißer
ländlicher Raum	1
Kleinstadt	0
Mittelstadt	0
Großstadt	1
Ballungsraum	0
Landkreis oder Regierungsbezirk	4

Tabelle 4.1: Ausreißer in dl PCB Brassen, nach Probenahmestandorten

diagramm der Mittelwerte oder dem Balkendiagramm der Mediane der $d = 12$ Komponenten zu vergleichen. Balkendiagramme werden in dieser Arbeit auch als Kongenerenprofile oder kurz Profile bezeichnet.

Wie man sieht, sind sich alle Balkendiagramme in Abbildung 4.7 augenscheinlich sehr ähnlich. Zusätzlich zu diesen Mittelwertprofilen und Medianprofilen können die Balkendiagramme der sechs Ausreißer angegeben werden.

Vergleicht man die Profile der Ausreißervektoren mit den Profilen der Mittelwerte bzw. der Mediane, so lassen sich möglicherweise Hinweise darauf finden, warum die gefundenen Ausreißer als solche erkannt wurden. Da sich die Mittelwert- und Medianprofile stark ähneln, wird zum Vergleich der bereinigten Daten mit den potentiellen Ausreißern das Mittelwertprofil herangezogen.

Beispielsweise entspricht der erste multivariate Ausreißer (siehe „Profil des ersten Ausreißers“ in Abbildung 4.8) in 11 von den 12 Balken in etwa dem Profil der Mittelwerte aus Abbildung 4.7. Balken 3 jedoch zeigt einen negativen Wert an, wohingegen der Mittelwert der bereinigten Daten in der dritten Komponente deutlich positiv ist. Im Profil des zweiten Ausreißers sind augenscheinlich keine Auffälligkeiten zu erkennen. Das Balkendiagramm des dritten Ausreißers zeigt Unterschiede zum Mittelwertprofil in den Balken 3, 4 und 10. Insgesamt zeigt das

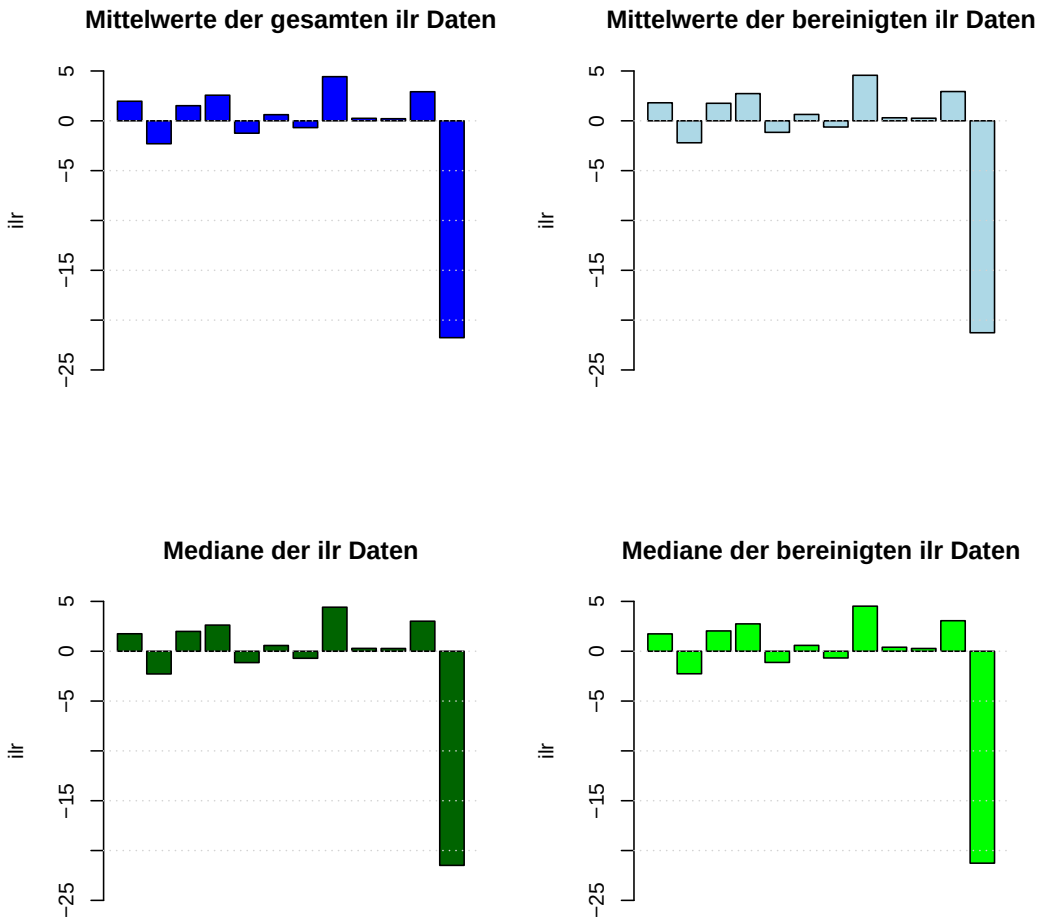

Abbildung 4.7: Balkendiagramme der Mittelwerte und Mediane der ilr transformierten Daten für dl PCB Brassen 2003

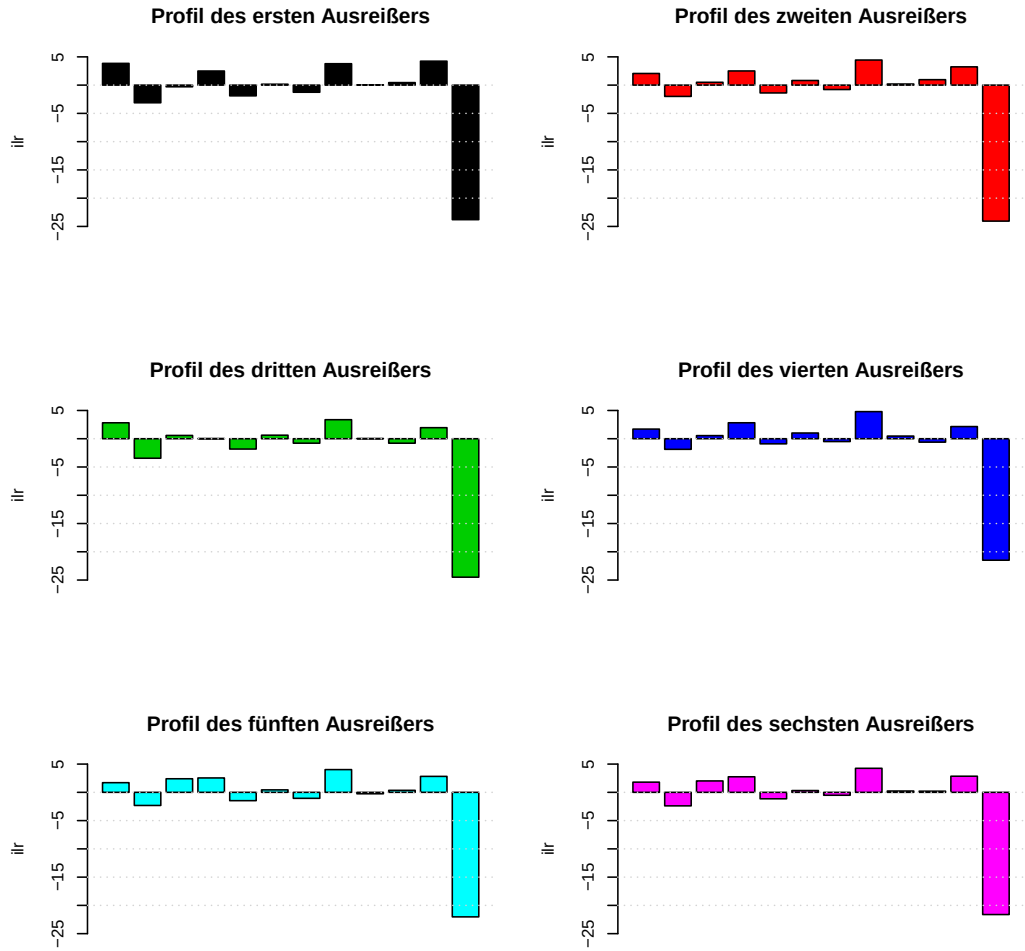

Abbildung 4.8: Balkendiagramme der sechs Ausreißer der ilr transformierten Daten für dl PCB Brassen 2003

Profil des dritten Ausreißers, dass 6 der 12 Einträge nahe Null sind, was im Vergleich zum Mittelwertprofil der bereinigten Daten auffällig scheint. Der zehnte Balken des vierten Ausreißers ist im Vergleich zum Mittelwertprofil negativ. Das Profil des fünften und sechsten Ausreißers ist im Vergleich zum Mittelwertprofil augenscheinlich nicht auffällig.

Es ist an dieser Stelle zu überlegen, ob die augenscheinlich unauffälligen Ausreißer wieder in den Datensatz integriert werden, oder ob sie weiterhin als Ausreißer angesehen werden. Zwar sind der zweite, fünfte und sechste Ausreißer augenscheinlich nicht auffällig, jedoch kann die Kombination ihrer Einträge im Vergleich zu den restlichen Messwerten ein Grund für die Deklaration als Ausreißer nach Pena und Prieto sein. Multivariate Ausreißer müssen nicht immer in einer oder mehreren Komponenten auffällig sein, wie bereits in Abschnitt 3.1.10 dargelegt wurde (vgl. Abbildung 3.7 und 3.8).

Die gefundenen Ausreißer wurden unter Verwendung rein statistischer Methoden in einer Stichprobe vom Umfang $n = 20$ Beobachtungen detektiert. Die Stichprobengröße ist vergleichsweise gering, berücksichtigt man die Dimension der Daten von $d = 12$. Sachlogische Gründe, die gefundenen Ausreißer auch tatsächlich als solche zu betrachten, können nicht aus den Daten abgeleitet werden. Die Balkendiagramme können Hinweise darauf geben, warum ein Datenpunkt als Ausreißer detektiert wurde, jedoch sind sie nicht in jedem Fall hilfreich. Werden statistische Methoden angewendet um Ausreißer in Daten zu detektieren, so ist stets die Zahl der vorhandenen Datenpunkte und die Dimension der Daten zu berücksichtigen. Für die hier untersuchte Stichprobe liegen nur wenige Datenpunkte im Vergleich zur Dimension vor. Alle ausreißeranalytischen Methoden sollten daher mit Bedacht angewendet und die Ergebnisse entsprechend kritisch hinterfragt

werden. Die wenigen Daten lassen nur begrenzt Rückschlüsse auf die zugrunde liegende natürliche Variabilität zu. In der Folge können Messwerte fälschlicherweise als statistische Ausreißer interpretiert werden, obgleich sie lediglich die natürliche Variabilität der Daten widerspiegeln. Aufgrund der vergleichsweise hohen Anzahl von sechs Ausreißern in $n = 20$ Datenpunkten der Dimension $d = 12$ ist davon auszugehen, dass möglicherweise nicht alle gefundenen Ausreißer auch tatsächlich als solche zu interpretieren sind. Vielmehr legt dies nahe, dass die natürliche Variabilität durch die geringe Größe der Stichprobe unterschätzt wird und dadurch Messwerte zu Unrecht als Ausreißer klassifiziert werden. Daher ist es zwingend notwendig eventuell folgende statistische Auswertungen sowohl anhand des vollständigen Datensatzes, als auch mit den bereinigten Daten durchzuführen und die Ergebnisse entsprechend zu interpretieren.

4.1.2.2 PCDD

In diesem Abschnitt werden Grünkohlproben (Brassica oleracea var. sabellica) aus dem Jahr 2003 auf statistische Ausreißer untersucht. Sie wurden in einem Bundesland an verschiedenen Standorten genommen. Die Konzentrationen sind angegeben in ng/kg Frischgewicht. Es liegen insgesamt $n = 16$ multivariate Datenpunkte der Gestalt $x = (x_1, \ldots, x_7)'$ vor. Zunächst werden die Daten durch Entfernung der Einheit standardisiert.

$$1\frac{ng}{kg} = 1\frac{10^{-9}g}{10^3 g} = 10^{-12}$$

Anschließend wird der „Rest“ berechnet, wie in Abschnitt 3.1.4 beschrieben. Demnach liegen Kompositionsdaten mit $D = 8$ vor. Fehlende Werte gibt es nicht

im Datensatz. Eine Beobachtung enthält für die sieben Kongenere ausschließlich Nulldaten, d.h. der entsprechende Datenpunkt hat die Gestalt $(0, 0, 0, 0, 0, 0, 0, 1)$, wobei 1 den „Rest“ angibt. Zudem wurden keine Bestimmungsgrenzen angegeben. Dieser Datenpunkt wird von den weiteren Analysen ausgeschlossen, da eine ilr Transformation für ihn nicht durchführbar ist. Die ilr Transformation wird auf die restlichen $n = 15$ Beobachtungen angewendet und die natürliche Variabilität auf 3.234171 geschätzt.

Die Scatterplotmatrix in Abbildung 4.9 vermittelt einen ersten Eindruck von der Struktur der ilr transformierten Daten.

Die Scatterplots in den Zeilen 1, 6 und 7 deuten auf Ausreißer hin. Eine Analyse der ilr transformierten Daten auf statistische Ausreißer mittels der robusten Mahalanobis-Distanz liefert vier Ausreißer. Dabei wurde analog zu Filzmoser und Hron das Signifikanzniveau $\alpha = 0.025$ gewählt. Werden diese von den Daten ausgeschlossen, so beträgt die Schätzung der natürlichen Variabilität 3.296219 und damit ca. 102% der natürlichen Variabilität der Gesamtdaten bzw. einen Zuwachs von ca. 2%. Wie ist dies zu erklären? Die natürliche Variabilität wird berechnet als Spur der geschätzten Kovarianzmatrix. Sie entspricht somit der Summe der geschätzten Varianzen der einzelnen Komponenten. Zur Illustration der Plausibilität einer Vergrößerung der geschätzten natürlichen Variabilität wird an dieser Stelle ein Beispiel gegeben. Angenommen es wurde ein multivariater Ausreißer als solcher identifiziert, weil er in seiner ersten Komponente extrem vom mittleren Niveau der Daten abweicht. In seinen anderen Komponenten sei er jedoch sehr nah am Mittelwert der Daten gelegen. Dann führt der Ausschluss dieses Ausreißers bei der Berechnung der natürlichen Variabilität unter Umständen zu folgendem Effekt: die geschätzte Varianz der ersten Komponente wird verringert und

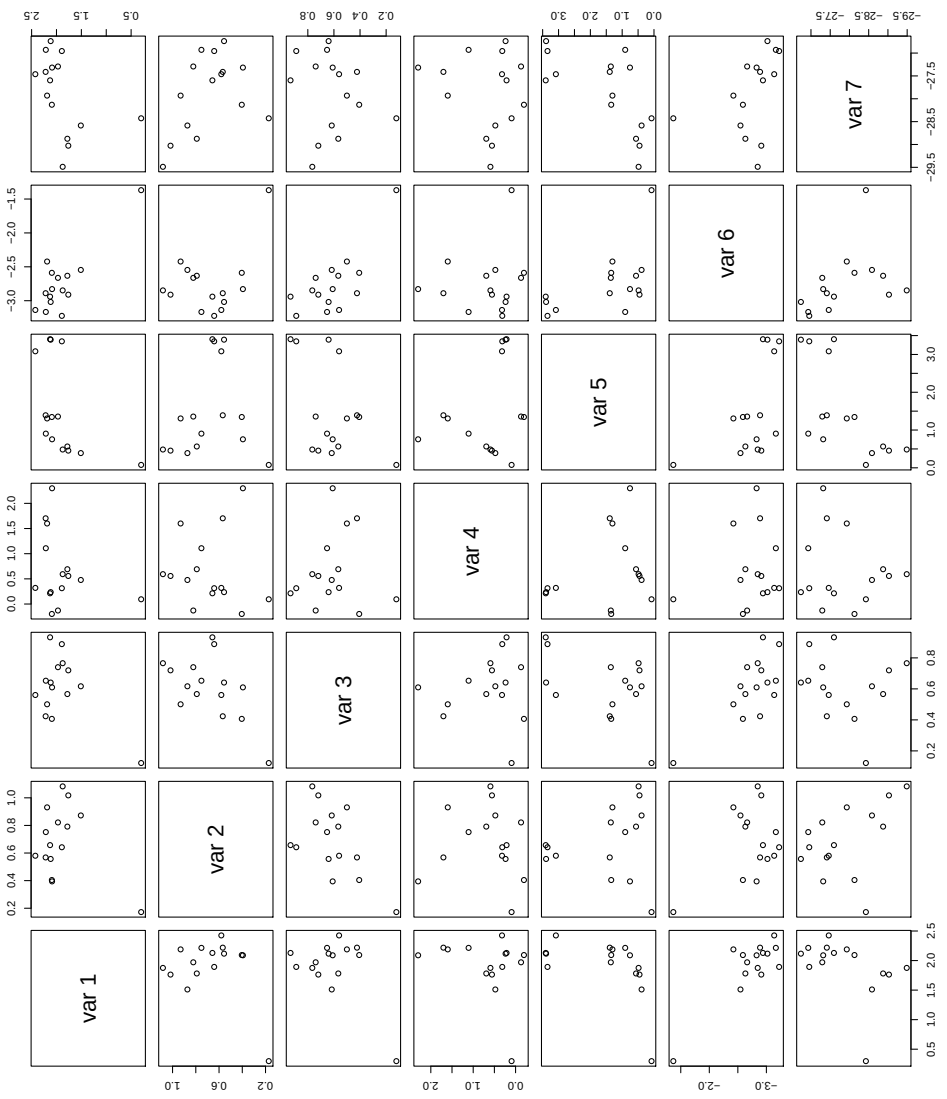

Abbildung 4.9: Scatterplotmatrix der ilr transformierten Daten für PCDD Grünkohl 2003

die Varianz der restlichen Komponenten wird vergrößert. Dazu betrachte man den folgenden Beispieldatensatz.

Beob.nr.	Komponente 1	Komponente 2
1	0.1	3
2	0.8	1
3	0.9	2
4	1	4
5	1.1	5

Tabelle 4.2: Beispiel zur Vergrößerung der Schätzung der natürlichen Variabilität durch den Ausschluss von Ausreißern

Die zugehörige geschätzte Kovarianzmatrix hat die Gestalt

$$\begin{pmatrix} 0.157 & 0.55 \\ 0.55 & 2.5 \end{pmatrix} .$$

Die Schätzung der natürlichen Variabilität ergibt sich zu 0.157+2.5=2.657. Entfernt man die erste Zeile als Ausreißer, so ergibt sich eine neue Schätzung der Kovarianzmatrix.

$$\begin{pmatrix} 1/60 & 7/30 \\ 7/30 & 10/3 \end{pmatrix}$$

Die daraus resultierende Schätzung der natürlichen Variabilität 1/60+10/3=3.35 ist größer als die natürliche Variabilität des unbereinigten Datensatzes. Wie man an diesem einfachen Beispiel sieht, kann der Ausschluss statistischer Ausreißer durchaus zu einer höheren Schätzung der natürlichen Variabilität führen. Dieses Phänomen tritt insbesondere dann auf, wenn statistische Ausreißer in einer Komponente mit geringer geschätzter Varianz detektiert werden, sie aber in Komponenten mit größerer geschätzter Varianz nicht ausreißerverdächtig sind.

Die Ausreißer im Datensatz zu Grünkohl lassen sich keinem eindeutigen Gebietstyp zuordnen. Sie treten sowohl in ländlichen Regionen, als auch in Klein- und Mittelstädten auf, vgl. Tabelle 4.3.

Standorttyp	Anzahl Ausreißer
ländlicher Raum	2
Kleinstadt	1
Mittelstadt	1
Großstadt	0
Ballungsraum	0
Landkreis oder Regierungsbezirk	0

Tabelle 4.3: Ausreißer in PCDD Grünkohl, nach Probenahmestandorten

Der Datensatz enthält zu großen Teilen ländliche Regionen, Kleinstädte und Mittelstädte als Probenahmestandorte. Lediglich zwei Großstädte werden als Probenahmestandort angegeben. Die Ausreißer können somit nicht auf Standorttypen eingeschränkt werden. Weitere Details zur Datenerhebung und etwaigen Besonderheiten können den Daten nicht entnommen werden.

Um die gefundenen Ausreißer weiter zu untersuchen, werden im Folgenden ihre Profile mit denen der Mittelwerte und Mediane der ilr transformierten Komponenten verglichen.

Optisch auffällig in Abbildung 4.11 ist, dass die ersten sechs Balken im Profil des ersten Ausreißers deutlich kürzer sind als die korrespondierenden Balken im Mittelwertprofil der bereinigten Daten. Weiterhin zeigt das Profil des vierten Ausreißers einen negativen vierten Balken, jedoch kann nicht ohne Weiteres beurteilt werden, ob dieser Abstand zum korrespondierenden Balken im Mittelwertprofil ausreicht diesen Datenpunkt als Ausreißer zu deklarieren. Für die restlichen Ausreißer können keine optischen Auffälligkeiten festgestellt werden.

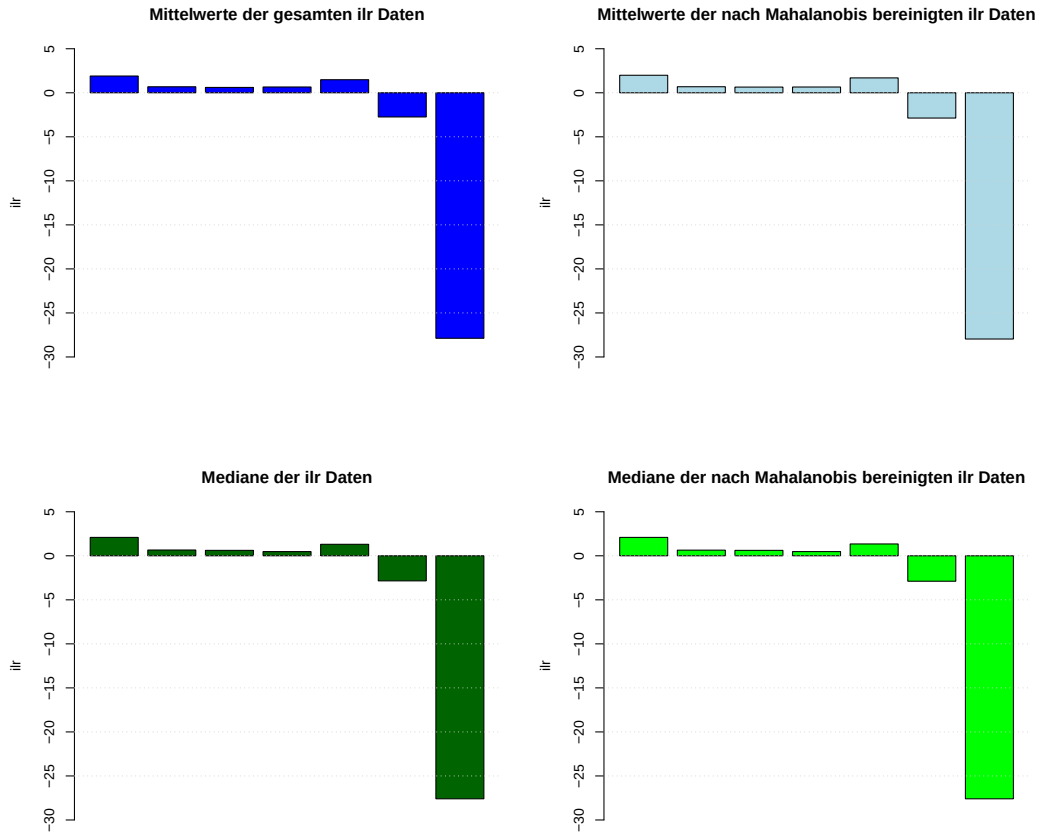

Abbildung 4.10: Balkendiagramme der Mittelwerte und Mediane der ilr transformierten Daten für PCDD Grünkohl 2003

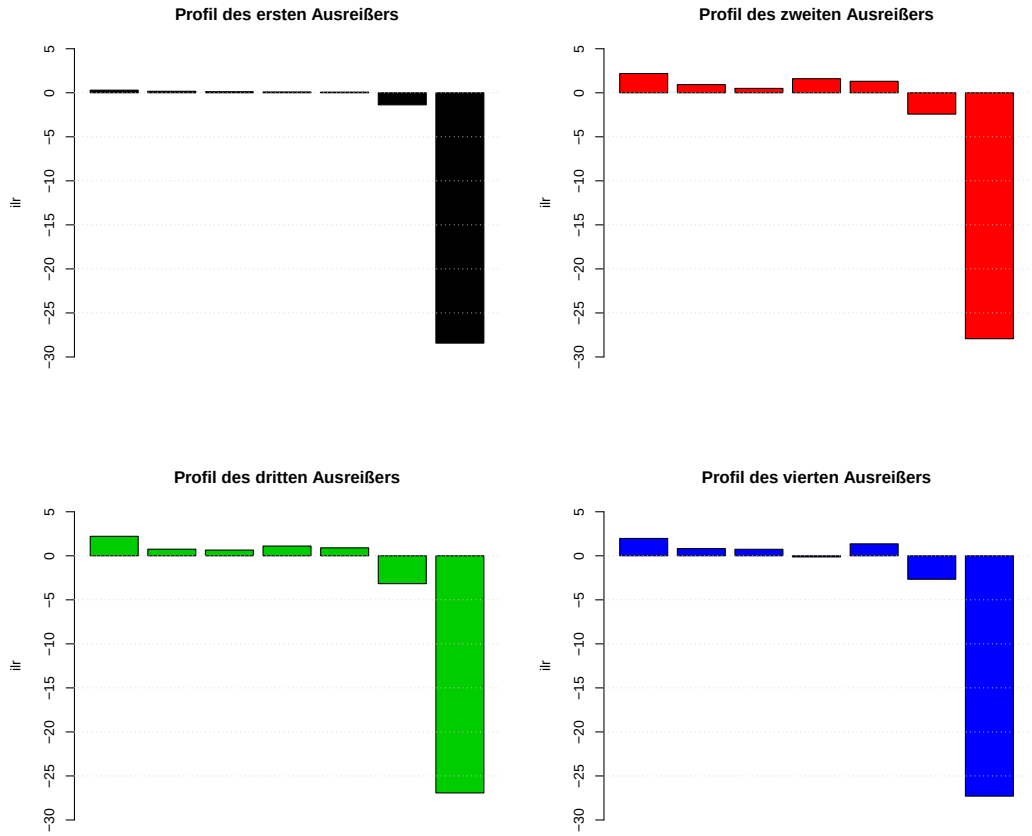

Abbildung 4.11: Balkendiagramme der vier Ausreißer der ilr transformierten Daten für PCDD Grünkohl 2003 nach Mahalanobis

Es ist zu beachten, dass die Methode der robusten Mahalanobis-Distanz zum Auffinden statistischer Ausreißer hier auf eine Stichprobe der Größe $n = 15$ angewendet wurde, wobei die Dimension der Daten $d = 7$ beträgt und $k = 4$ Ausreißer detektiert wurden. Das Verhältnis zwischen n und d entspricht circa 2:1. Die Forderung $n \geq 2d$ ist damit knapp erfüllt (vgl. Abschnitt 3.4). Im Vergleich zur Anzahl detektierter Ausreißer und der Dimension d ist die Stichprobengröße jedoch eher gering: $2d + k > n$. Die gefundenen Ausreißer sind kritisch zu hinterfragen. Auch der Einfluss dieser Datenpunkte auf die Schätzung der natürlichen Variabilität in Höhe von 2% ist vergleichsweise gering. Es ist daher zu überlegen, ob einige oder möglicherweise alle der detektierten Ausreißer eventuell fälschlich als solche vom Verfahren erkannt wurden.

Im Weiteren Verlauf werden die Daten dem Projektionsverfahren von Pena und Prieto aus Abschnitt 3.4.1 zum Signifikanzniveau $\alpha = 0.01$ unterzogen. Die Wahl des Signifikanzniveaus folgt dabei dem Vorgehen von Pena und Prieto [vgl. PP01b, S.290]. Dieses liefert lediglich einen Ausreißer, dessen Profil in Abbildung 4.13 angegeben ist. Das Profil der Mittelwerte und Mediane der nach Pena und Prieto bereinigten Daten ist in Abbildung 4.12 gegeben.

Der vierte Balken des Ausreißerprofils ist im Vergleich zum bereinigten Mittelwertprofil auffällig. Weitere optische Auffälligkeiten können nicht festgestellt werden. Die Schätzung der natürlichen Variabilität im bereinigten Datensatz nach Pena und Prieto ergibt sich zu 3.181172. Sie entspricht ca. 98% der natürlichen Variabilität der gesamten Daten bzw. einer Verringerung dieser um ca. 2%. Der mit dem Verfahren von Pena und Prieto gefundene Ausreißer wurde mit der robusten Mahalanobis-Distanz nicht detektiert. Die Überprüfung des Probenahmestandortes ergab, dass es sich um einen Datenpunkt aus einer ländlichen Region

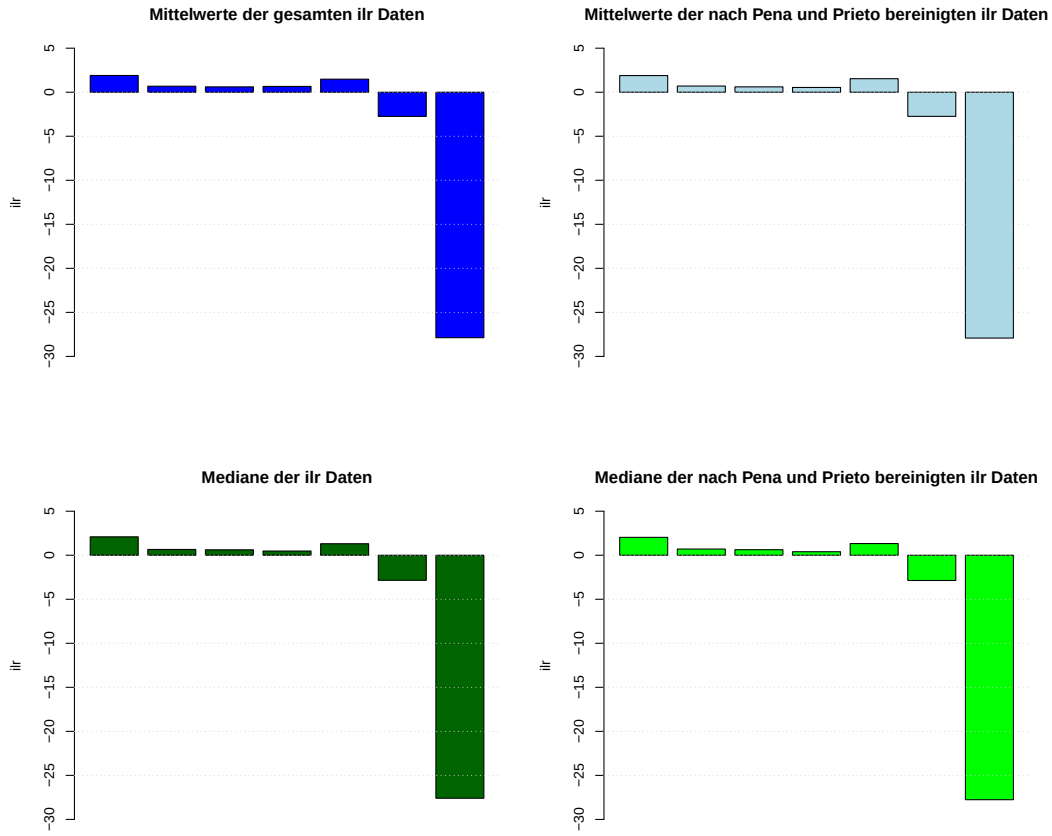

Abbildung 4.12: Balkendiagramme der Mittelwerte und Mediane der ilr transformierten Daten für PCDD Grünkohl 2003 im Vergleich zu den Mittelwerten und Medianen der ausreißerbereinigten Daten nach Pena und Prieto

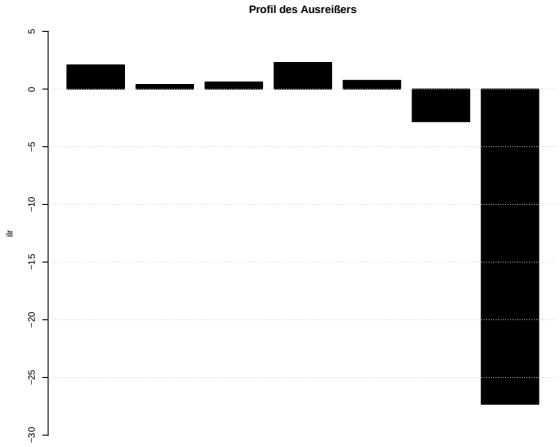

Abbildung 4.13: Balkendiagramm des Ausreißers der ilr transformierten Daten für PCDD Grünkohl 2003 nach Pena und Prieto

handelt. Da ca. 50% der Daten aus ländlichen Regionen und Kleinstädten stammen, gibt der Probenahmestandort des Ausreißers keinen Hinweis auf sachlogische Gründe diesen Punkt als Ausreißer zu identifizieren.

Je nach verwendeter Methode, ergeben sich folglich andere Ausreißer. Beide Methoden sind praktikabel. Es lässt sich an dieser Stelle nicht sagen, ob eines der beiden Verfahren eventuell konservativer ist, da die gefundenen Ausreißer keine Schnittmenge besitzen. Eine weiterführende Diskussion dieser Frage folgt in Abschnitt 5.

Auffällig ist, dass das Verfahren nach Pena und Prieto keinen der Ausreißer detektiert, die unter Verwendung der robusten Mahalanobis-Distanz erkannt wurden. Zwar kann nicht geklärt werden, ob diese Tatsache gegen den Einsatz eines der beiden Verfahren spricht, jedoch zeigen die unterschiedlichen Ergebnisse, dass die gefundenen Ausreißer kritisch hinterfragt werden müssen, bevor eine endgültige Entscheidung getroffen wird, ob ein Datenpunkt als Ausreißer zu betrachten ist.

4.2 Multivariate Zeitreihen

In diesem Abschnitt wird exemplarisch die Analyse einer multivariaten Zeitreihe auf Kontaminationen durch statistische Ausreißer anhand des Projektionsverfahrens nach Galeano et al. vorgestellt und die Ergebnisse diskutiert (vgl. Kapitel 3.7).

4.2.1 PCB Leitkongenere

Der hier untersuchte Datensatz enthält deutschlandweit genommene Proben zu Brassenfilets (Abramis brama). Es handelt sich um Daten aus den Jahren 1988 bis 2006. Für das Jahr 1989 liegen keine Daten vor. Die Anzahl der Datenpunkte variiert über die Jahre. Da es sich um Daten zu den sechs PCB Leitkongeneren handelt, liegen folglich multivariate Datenpunkte der Gestalt $x = (x_1, \ldots, x_6)'$ vor. Die Konzentrationen sind angegeben in $\mu g/kg$ Frischgewicht.

Bei den Daten handelt es sich um Jahresquerschnittsdaten. Dementsprechend wird zunächst jedes Jahr separat statistisch analysiert. Dazu werden die Daten durch Entfernung der Einheit standardisiert und der „Rest“ berechnet, wie in Abschnitt 3.1.4 beschrieben. Es liegen somit für jedes Jahr Kompositionsdaten mit $D = 7$ vor. Im Jahr 1994 treten drei fehlende Werte für PCB 28 auf, im Jahr 1995 ein fehlender Wert für PCB 28 und im Jahr 1996 zwei fehlende Werte für PCB 28. Bis auf das Jahr 1991 enthalten alle weiteren Jahre Nulldaten. Diese werden wie in Abschnitt 3.1.5 ersetzt. Anschließend unterteilt sich die Analyse in zwei Varianten.

- Variante 1: Die fehlenden Werte in den Jahresquerschnittsdaten werden, wie in Abschnitt 3.1.6 beschrieben, geschätzt.

- Variante 2: Datenpunkte, die fehlende Werte enthalten, werden von den weiteren Analysen ausgeschlossen.

Für alle in dieser Arbeit analysierten Datensätze wurden beide Varianten der Auswertung durchgeführt. Der Kompaktheit halber sei hier lediglich die Vorgehensweise für Variante 1 illustriert. Es werden die ilr transformierten Daten berechnet und anschließend alle Jahresdatensätze, die mindestens $n = 12$ Messwerttupel aufweisen, mit dem Verfahren nach Pena und Prieto aus Abschnitt 3.4.1 auf statistische Ausreißer untersucht. Die Zahl der gefundenen Ausreißer für jedes Jahr ist in Tabelle 4.4 gegeben.

Jahr	Anzahl Ausreißer (Variante 1)	Anzahl Ausreißer (Variante 2)
1988	$(n < 12)$	$(n < 12)$
1989	$(n < 12)$	$(n < 12)$
1990	$(n < 12)$	$(n < 12)$
1991	$(n < 12)$	$(n < 12)$
1992	$(n < 12)$	$(n < 12)$
1993	$(n < 12)$	$(n < 12)$
1994	$(n < 12)$	$(n < 12)$
1995	1	0
1996	1	3
1997	1	1
1998	1	1
1999	1	1
2000	1	1
2001	1	1
2002	0	0
2003	1	1
2004	0	0
2005	1	1
2006	1	1

Tabelle 4.4: Ausreißer in Daten zu PCB Leitkongeneren in Brassenfilets

Die Vorgehensweise und Argumentation bei der Analyse der ilr transformier-

ten Jahresdatensätze auf statistische Ausreißer gleicht für beide Varianten dem Vorgehen in den Abschnitten 4.1.1.1 und 4.1.2.1 und wird daher an dieser Stelle nicht explizit angegeben. Die als statistische Ausreißer detektierten Datenpunkte werden von den weiteren Analysen ausgeschlossen.

Nach Abschluss der statistischen Ausreißeranalyse der Querschnittsdatensätze werden für jedes Jahr die Mittelwertvektoren der bereinigten Daten gebildet. Berücksichtigt man an dieser Stelle, dass für das Jahr 1989 keine Daten vorliegen, so ergeben sich, wie in Abschnitt 3.5.8.1 diskutiert, zwei weitere Varianten der Auswertung.

- Variante A: Schätze fehlende Jahresmittelwertvektoren, wie in Abschnitt 3.5.8.1 beschrieben unter Verwendung der LMS-Methode (vgl. Abschnitt 3.5.3).

- Variante B: Entferne alle Jahre, zu denen keine Messwerte gegeben sind. Betrachte die verbliebenen Jahresmittelwertvektoren als Zeitreihe mit äquidistanten Beobachtungen.

Es ergeben sich insgesamt vier Auswertevarianten: 1A, 1B, 2A und 2B. Die Variante 1B wird hier exemplarisch vorgestellt und diskutiert. Alle weiteren Auswertungen ähneln sich in Vorgehensweise und Argumentation der hier vorgestellten. In Abbildung 4.14 ist die Zeitreihe der Mittelwerte aus Variante 1B dargestellt. Sie besteht aus $T = 18$ Mittelwertvektoren der Dimension $d = 6$. Um die Verfahren aus Abschnitt 3.7 anwenden zu können, muss die Zeitreihe trendbereinigt vorliegen. Die Verfahren aus Abschnitt 3.5.5.1 detektieren sowohl einen stochastischen, als auch einen deterministischen Trend in der fünften Komponente. Daher werden im nächsten Schritt die ersten Differenzen der Zeitreihe gebildet und

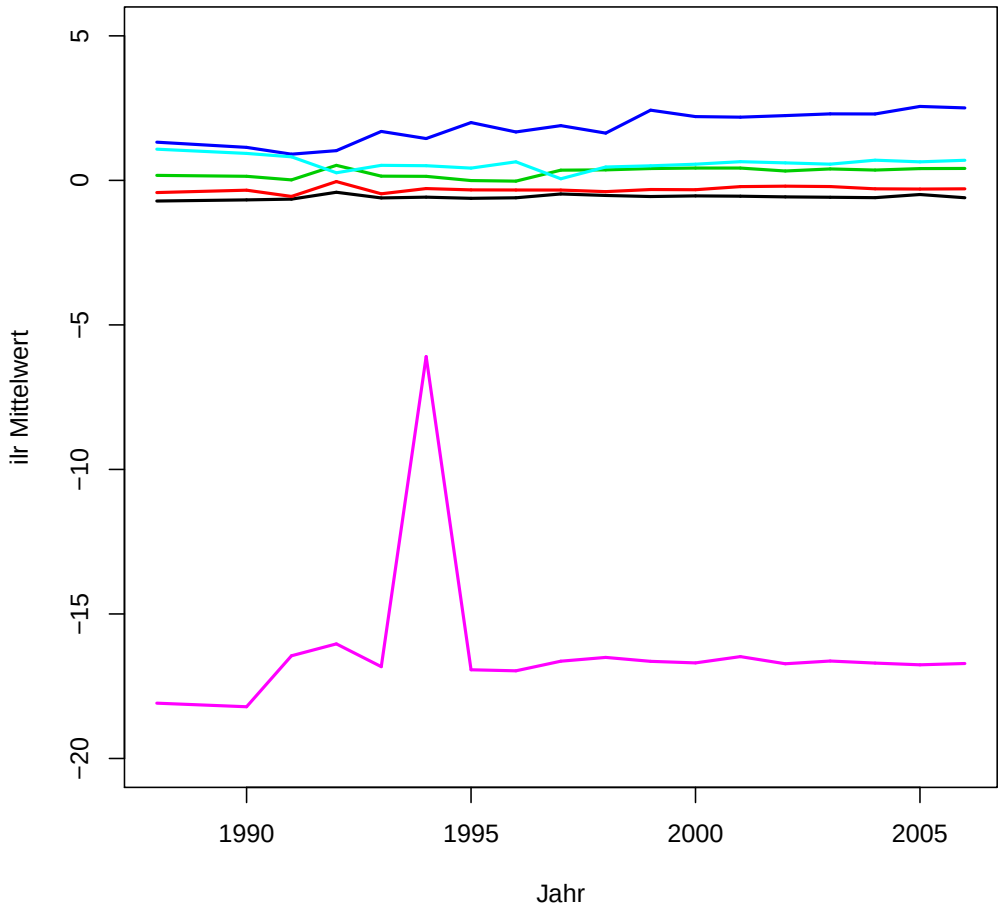

Abbildung 4.14: Mittelwerte der ilr transformierten Daten nach Variante 1B

eine Mittelwertbereinigung vorgenommen, siehe Abbildung 4.15. Für die trend-

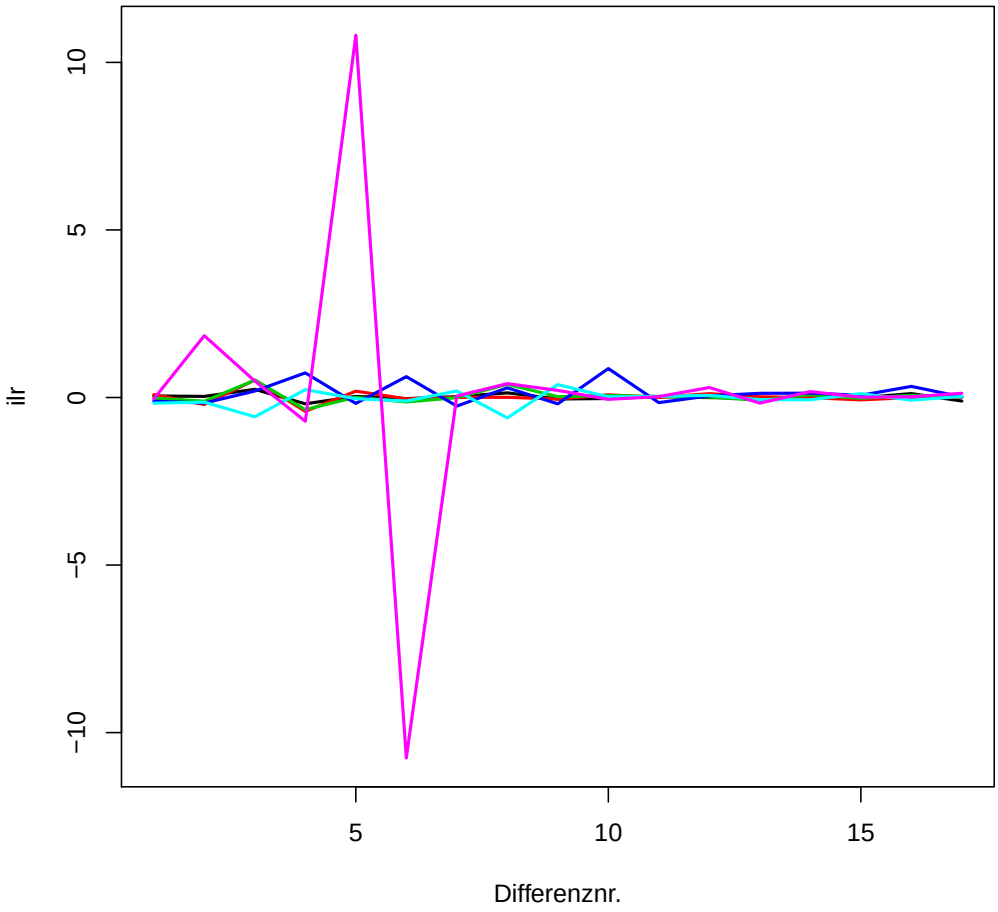

Abbildung 4.15: Mittelwertbereinigte erste Differenzen der Mittelwerte der ilr transformierten Daten nach Variante 1B

bereinigte Zeitreihe kann eine Schätzung der natürlichen Variabilität angegeben werden. Dazu wird ein $VAR[1]$-Modell an die Daten angepasst, wie in Abschnitt 3.5.2.1 beschrieben. Die natürliche Variabilität ergibt sich als Spur der geschätzten Kovarianzmatrix der Modellresiduen. Sie beträgt 11.59905. Anschließend erfolgt die statistische Analyse auf Ausreißer mittels der Verfahren aus Abschnitt 3.7.

Zum Zeitpunkt $t = 6$ wird ein innovativer Ausreißer detektiert. Dieser kann, wie in Abschnitt 3.5.1.2 beschrieben, nicht ohne Weiteres entfernt werden. Um die Zeitreihe der ersten Differenzen jedoch zumindest näherungsweise vom Effekt des innovativen Ausreißers zu bereinigen, wird die Prozedur zur Ausreißererkennung und Bereinigung der Zeitreihe modifiziert. Alle durch den Algorithmus aus Abschnitt 3.7.4.1 gefundenen innovativen Ausreißer werden während des iterativen Prozesses sofort als transiente Ausreißer deklariert. Ihr Effekt wird durch den, des entsprechenden innovativen Ausreißers geschätzt. Die Schätzung des Effektes eines deklarierten transienten Ausreißers durch den Effekt des entsprechenden innovativen Ausreißers ist notwendig. Es ist nicht sinnvoll den Effekt des deklarierten transienten Ausreißers durch den tatsächlichen transienten Effekt zu besagtem Zeitpunkt zu schätzen.

Begründung: Angenommen man täte dies dennoch. Da zu besagtem Zeitpunkt ein innovativer Ausreißer und kein signifikanter transienter Ausreißer gefunden wurde folgt, dass der Effekt des deklarierten transienten Ausreißers nahe Null geschätzt würde. Der zu entfernende Effekt ist jedoch der eines signifikanten innovativen Ausreißers, d.h. dieser ist signifikant verschieden von Null. Der Effekt des innovativen Ausreißers würde folglich durch den Effekt des transienten (nicht signifikanten) Ausreißers unterschätzt. In vorliegendem Datensatz ist zu beobachten, dass diese Vorgehensweise in einer programmtechnischen Endlosschleife mündet.

Durch die Deklaration des innovativen Ausreißers als transienten Ausreißer, werden durch den Algorithmus zwei weitere innovative Ausreißer zu den Zeitpunkten $t = 5$ und $t = 7$ erkannt. Diese wurden ebenfalls als transiente Ausreißer deklariert. Die von den Ausreißern bereinigte Zeitreihe der ersten Differenzen ist

in Abbildung 4.16 gegeben. Die Schätzung der natürlichen Variabilität beträgt 0.564466. Sie entspricht ca. 4.9% der ursprünglichen natürlichen Variabilität von 11.59905 bzw. einer Abnahme der natürlichen Variabilität von 95.1%.

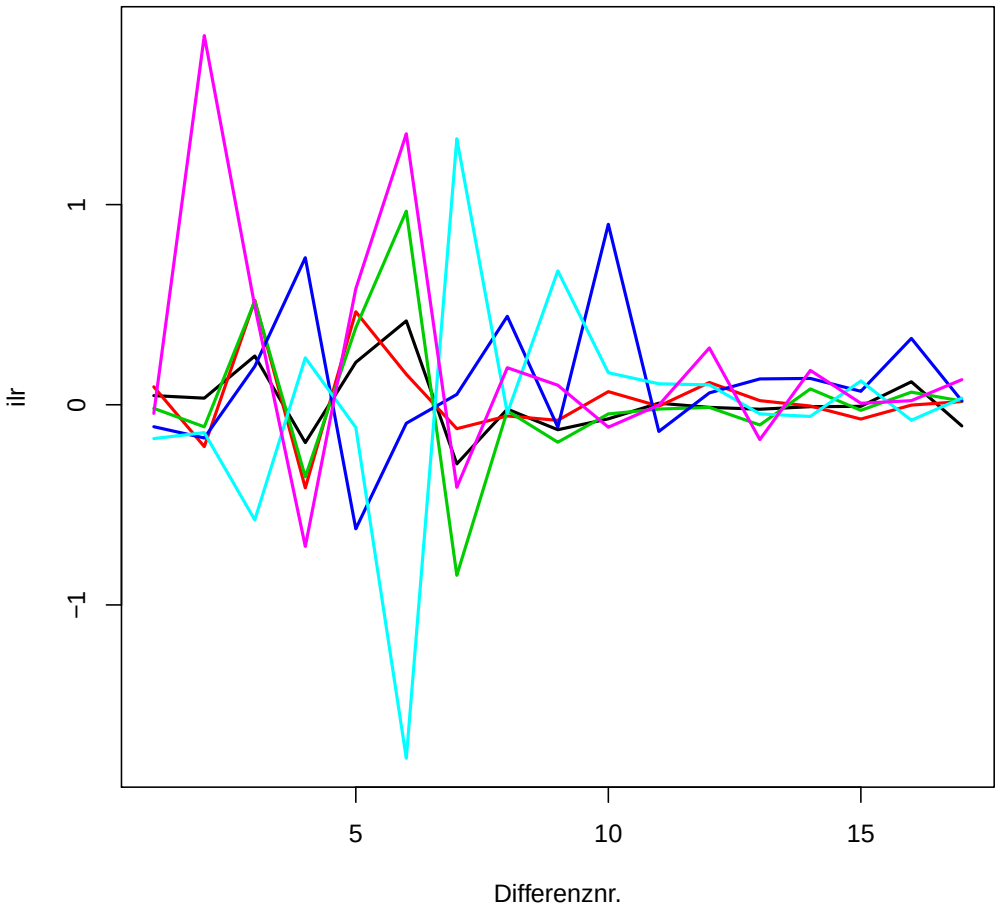

Abbildung 4.16: Von Ausreißern bereinigte Zeitreihe der ersten Differenzen der Mittelwerte der ilr transformierten Daten nach Variante 1B

Inwiefern sind diese Ergebnisse sinnvoll? Berücksichtigt man, dass der fehlende Mittelwertvektor des Jahres 1989 ausgelassen wurde in Variante 1B, so folgt, dass die Bildung der Differenzen teilweise fehlerbehaftet ist: die erste Differenz

ergibt sich aus den Mittelwertvektoren der Jahre 1990 und 1988. Alle weiteren Differenzen sind korrekt errechnet. Diese fehlerhafte Berechnung der ersten Differenz beeinflusst in der Folge die Schätzung des AR[1]-Modells der Differenzen und liefert verzerrte Modellresiduen. Diese werden in der statistischen Ausreißeranalyse untersucht. Folglich kann es zu Fehlspezifikationen von Ausreißern kommen. Um dies zu überprüfen, wurden die Daten zum Vergleich mittels Variante 1A ausgewertet. Die Schätzung des fehlenden Mittelwertvektors der ilr transformierten Daten für das Jahr 1989 erfolgte, wie in Abschnitt 3.5.8.1 beschrieben. Der Test nach Giraitis et al. (vgl. Abschnitt 3.5.5.1) zeigt keine Trends an, so dass die ersten Differenzen nicht gebildet werden. Die Zeitreihe der mittelwertbereinigten ilr transformierten Mittelwertvektoren nach Variante 1A ist in Abbildung 4.17 gegeben.

Die Bereinigung der Zeitreihe der Mittelwerte aus Variante 1A erfolgt wieder über den Ansatz, innovative Ausreißer als transiente zu deklarieren. Dies liefert die in Abbildung 4.18 bereinigte Zeitreihe der Mittelwertvektoren zu Variante 1A. Eine Analyse der Daten auf die Präsenz statistischer Ausreißer detektiert zwei benachbarte innovative Ausreißer zu den Zeitpunkten $t = 7$ und $t = 8$, bzw. für die Jahre 1994 und 1995. Nach Abschnitt 3.5.6.5 gilt, dass ein innovativer Ausreißer bei der Bildung erster Differenzen wieder einen innovativen Ausreißer liefert. Bezeichnet man die ersten Differenzen mit $\Delta_1, \ldots, \Delta_{18}$ mit $\Delta_i = x_{i+1} - x_i$, wobei $x_1, \ldots, x_{19}$ die Mittelwertvektoren seien, dann ist leicht zu sehen, dass gilt

$$\Delta_6 = x_7 - x_6, \;\; \Delta_7 = x_8 - x_7 \;\text{ und }\; \Delta_8 = x_9 - x_8,$$

d.h. die innovativen Ausreißer der nicht differenzierten Zeitreihe tauchen in der

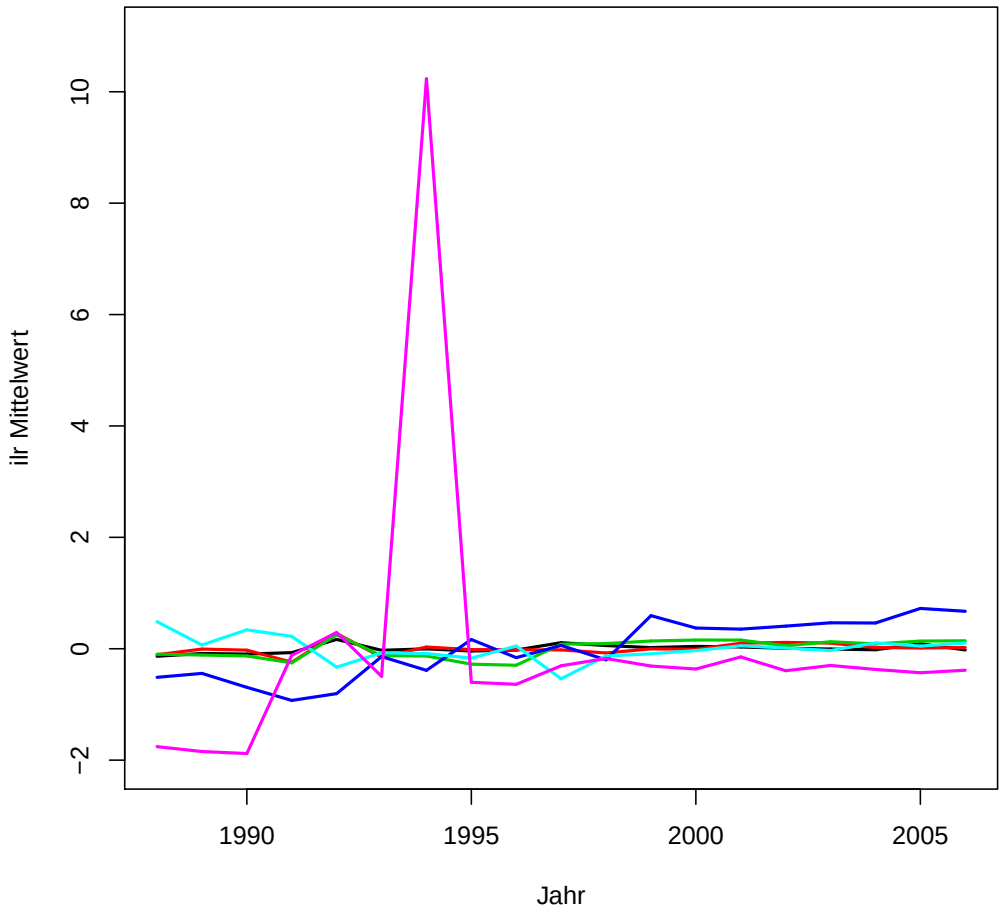

Abbildung 4.17: Zeitreihe der Mittelwerte der ilr transformierten Daten nach Variante 1A

Zeitreihe der Differenzen zu den Zeitpunkten $t = 6$, $t = 7$ und $t = 8$ auf. In Variante 1B wurden die innovativen Ausreißer zu den Zeitpunkten $t = 5$, $t = 6$ und $t = 7$ detektiert. Berücksichtigt man, dass in Variante 1B ein Datenpunkt fehlt (für das Jahr 1989 liegen keine Daten vor, daher wurde es aus der Zeitreihe entfernt), so sieht man, dass die Zeitpunkte, zu denen innovative Ausreißer in den ersten Differenzen von Variante 1A und 1B auftreten, identisch sind. Beide Ansätze liefern folglich die gleichen statistischen Ausreißer. Auffällig ist, dass der Test auf Trends nach Giraitis et al. (vgl. Abschnitt 3.5.5.1) in Variante 1A keinen Trend detektierte. Zur Erinnerung wird hier noch einmal das Testproblem angegeben.

H_0 : „Die Zeitreihe ist stationär."

H_1 : „Die Zeitreihe weist entweder einen linearen deterministischen Trend auf, oder sie ist differenzstationär."

Der Test nach Giraitis lehnt die Nullhypothese nicht ab. Daraus folgt aber keinesfalls, dass diese korrekt sein muss. Eine Nichtablehnung der Nullhypothese kann auf mehrere Gründe zurückzuführen sein, wie z.B.:

1. Die Nullhypothese ist tatsächlich wahr.
2. Die Nullhypothese ist nicht wahr, jedoch ist die Power des Tests zu gering, um dies zu erkennen.
3. Der Beobachtungszeitraum der Daten ist zu kurz, um eine langfristige Änderung des mittleren Niveaus der Zeitreihe tatsächlich erkennen zu können.

Hier wurde angenommen, dass die Zeitreihe tatsächliche keinen Trend aufweist.

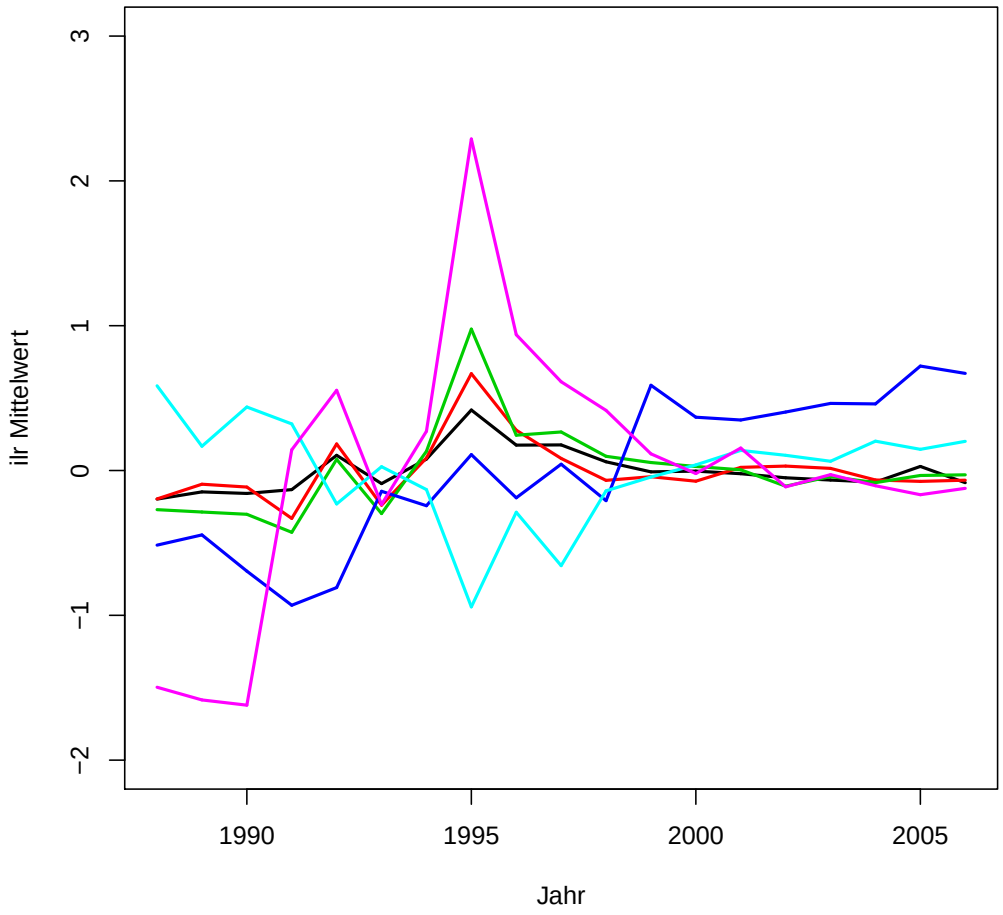

Abbildung 4.18: Von Ausreißern bereinigte Zeitreihe der Mittelwerte der ilr transformierten Daten nach Variante 1A

Ohne Kenntnis der statistischen Power des Tests kann aus einer Nichtablehnung der Nullhypothese nicht auf deren Korrektheit geschlossen werden. Andererseits kann der Trend in den Daten zu Variante 1B fälschlicher Weise detektiert worden sein. Man bedenke stets, dass die Daten in beiden Varianten sich dadurch unterscheiden, dass in Variante 1A der fehlende Vektor des Jahres 1989 geschätzt wurde, in Variante 1B wurde der Zeitpunkt komplett von den weiteren Analysen ausgeschlossen.

In der Diskussion der innovativen Ausreißer ist zu berücksichtigen, dass Variante 1 gewählt wurde, d.h. fehlende Werte in den Jahresquerschnittsdaten werden aus den verbliebenen Jahresdaten geschätzt. In den Jahren 1994 und 1995 treten fehlende Werte auf. Für das Jahr 1994 liegen $n = 7$ Datenpunkte vor, wobei vier von ihnen fehlende Werte für PCB 28 aufweisen. Es ist demnach möglich, dass die Schätzung der fehlenden Werte aufgrund der wenigen vollständigen Daten für das Jahr 1994 dazu geführt hat, dass die drei fehlenden Werte für PCB 28 schlecht geschätzt wurden. Dies verzerrt in der Folge den Mittelwertvektor der entsprechenden ilr transformierten Werte. Die Profile der Originaldaten sind in Abbildung 4.19 und 4.20 gegeben. Bis auf die fehlenden Werte, sind die Profile optisch ähnlich.

Schätzt man die fehlenden Werte und wendet die ilr Transformation auf die Daten an, zeigen sich deutliche Unterschiede in den Profilen der transformierten Daten, wie in Abbildung 4.21 und 4.22 zu erkennen ist.

Durch die Schätzungen der fehlenden Werte wird das Mittelwertprofil des Jahres 1994 stark verzerrt, wie in Abbildung 4.27 zu sehen ist. Für das Jahr 1995 liegen insgesamt $n = 13$ Datenpunkte vor. Es tritt ein fehlender Wert für PCB 28 auf. Optisch ist zwischen den Profilen der ilr transformierten vollständigen Daten-

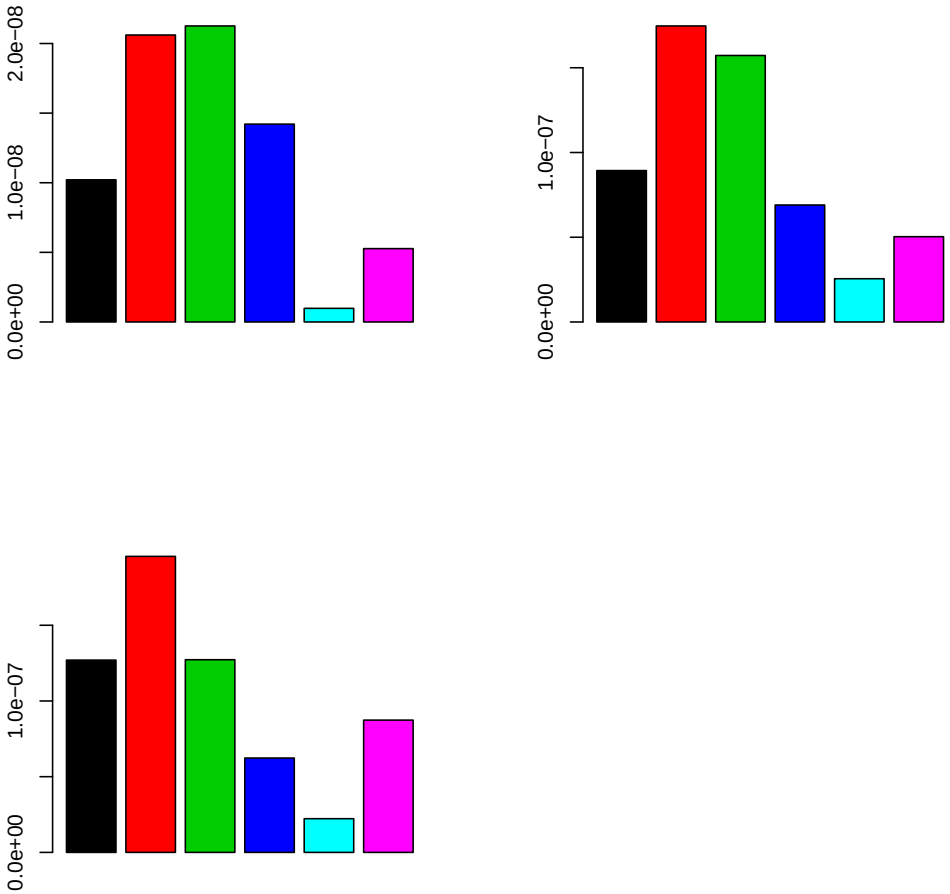

Abbildung 4.19: Profil der Daten von 1994, die keine fehlenden Werte aufweisen

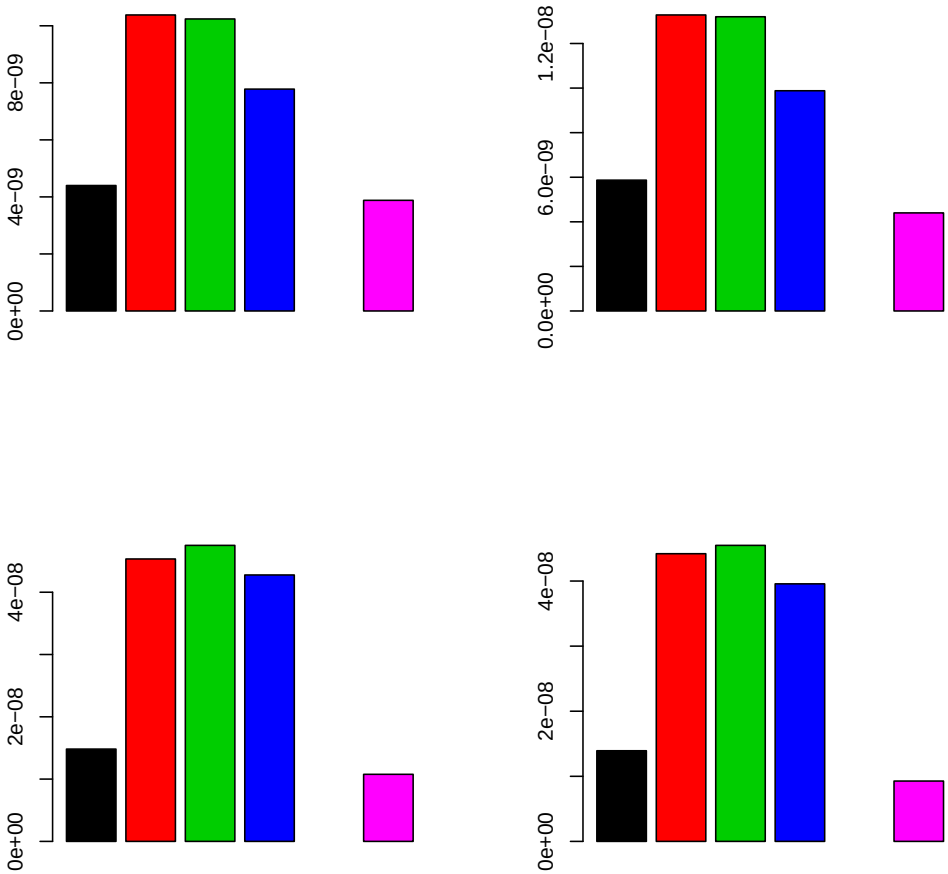

Abbildung 4.20: Profil der Daten von 1994, die fehlende Werte aufweisen

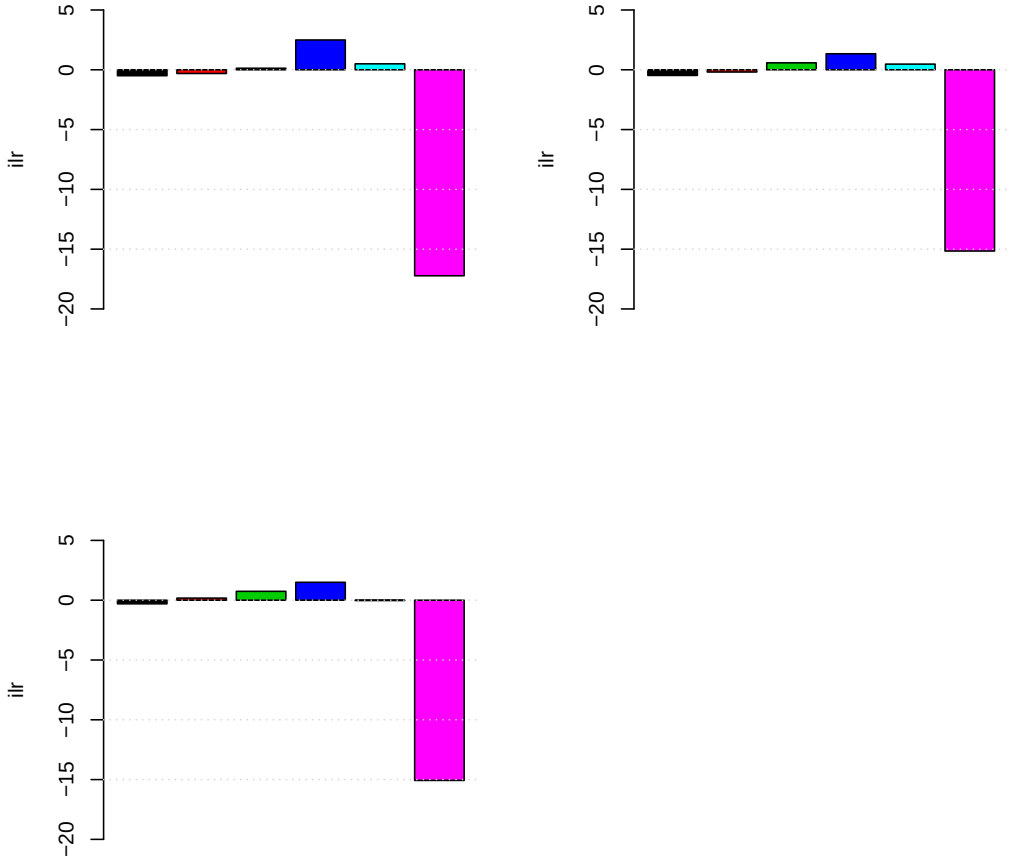

Abbildung 4.21: Profil der ilr transformierten Daten von 1994, die keine fehlenden Werte aufweisen

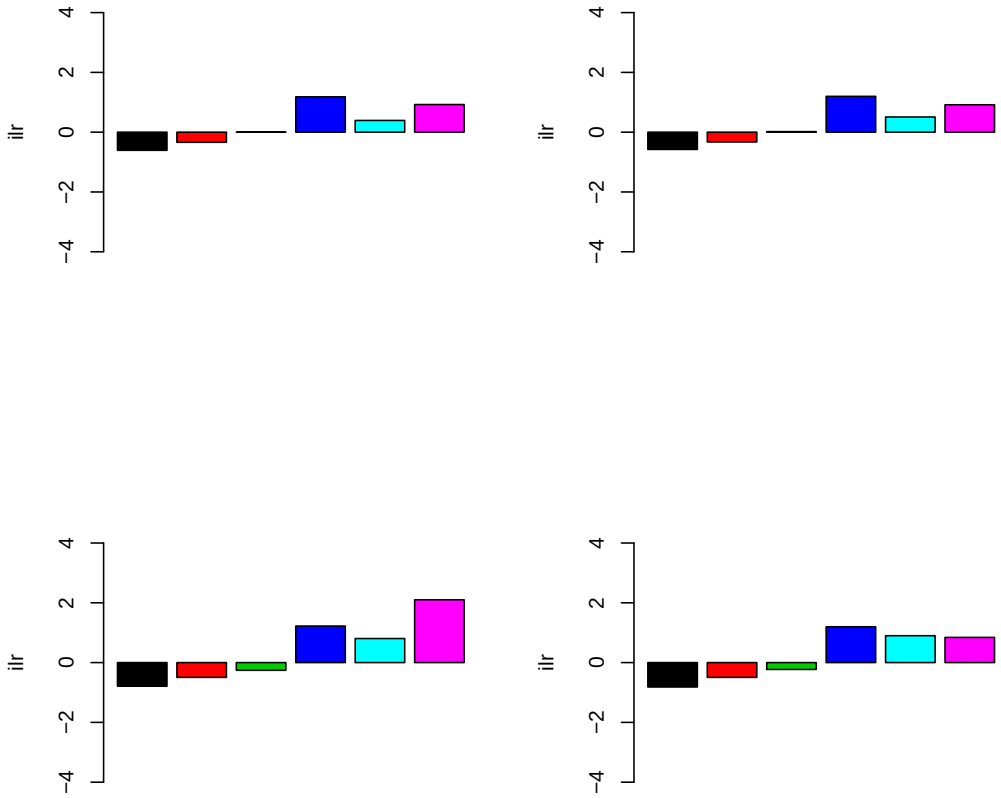

Abbildung 4.22: Profil der ilr transformierten Daten von 1994, deren fehlende Werte geschätzt wurden

punkte und dem ilr transformierten Datenpunkt mit geschätztem fehlenden Wert kein Unterschied sichtbar, wie man in den Abbildung 4.23-4.25 erkennen kann.

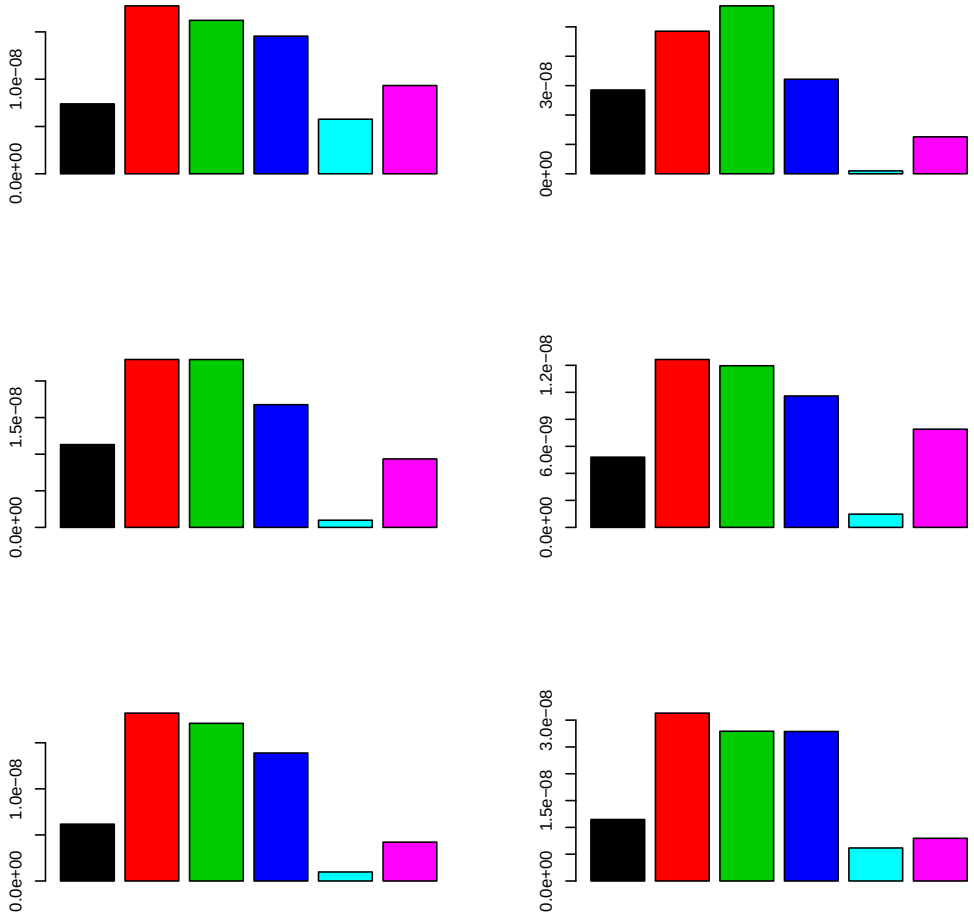

Abbildung 4.23: Profil der ilr transformierten Daten von 1995, die keine fehlenden Werte aufweisen, Teil 1

Es besteht daher die Möglichkeit, dass die Ausreißer erst durch die Ersetzungsmethode aus Abschnitt 3.1.6 in die Daten gelangten.

Wichtig ist hier die Erkenntnis, dass beide Varianten die gleichen Ausreißertypen zu denselben Zeitpunkten detektieren. Gibt es neben den statistischen Grün-

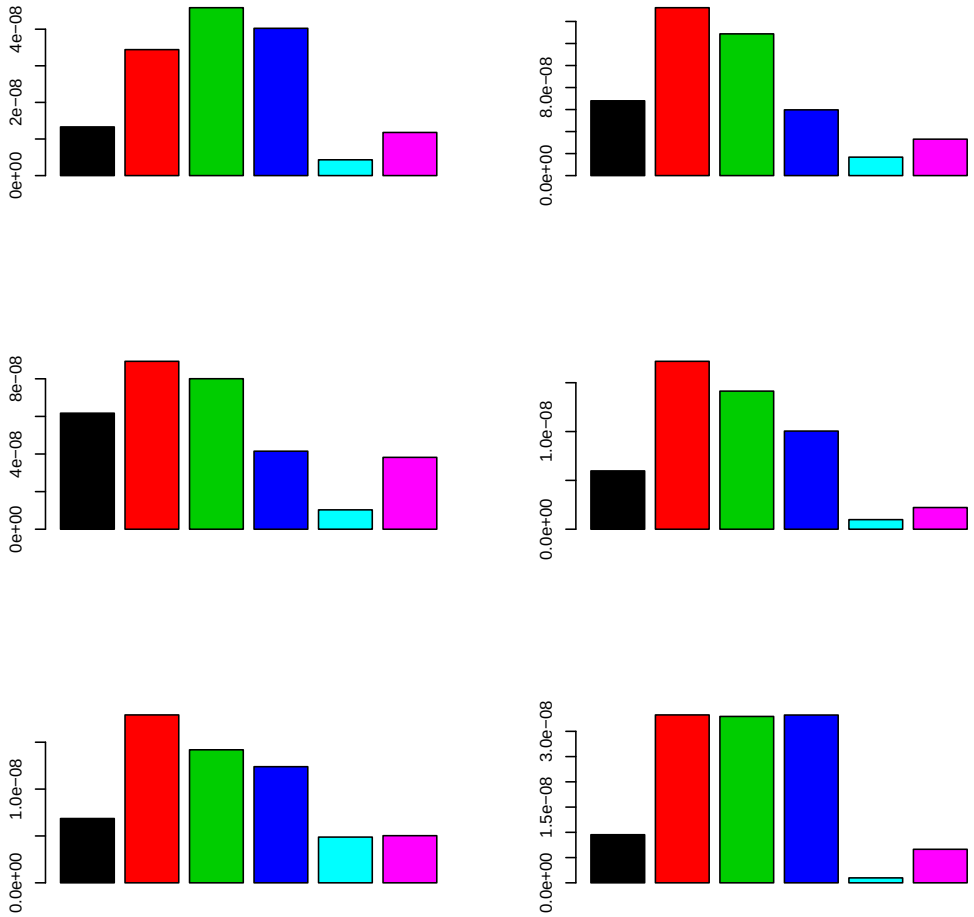

Abbildung 4.24: Profil der ilr transformierten Daten von 1995, die keine fehlenden Werte aufweisen, Teil 2

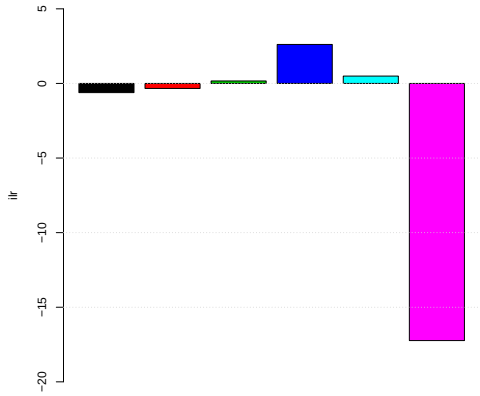

Abbildung 4.25: Profil des ilr transformierten Datenpunkte von 1995, dessen fehlender Wert geschätzt wurde

den auch sachlogische Hinweise, dass diese Daten tatsächlich Ausreißer sind? Um diese Frage zu beantworten, ist es notwendig sich die Ausgangsdaten detailliert anzuschauen. Aufgrund der Datenanonymität, werden hier keine Städtenamen angegeben. Die Anzahl der Proben je Ort und Jahr sind in Tabelle 4.5 geben.geben.

Jahr/Stadt	Ort 1	Ort 2	Ort 3	Ort 4	Ort 5	Ort 6	Ort 7
1988							
1989							
1990							
1991		1			1		
1992	1	1	1	1	1		
1993	1	1	1	1	1	1	1
1994	1	1	1	1	1	1	1
1995	1	1	1	1	1	1	1

1996	1	1	1	1	1	1	1
1997	1	1	1	1	1	1	1
1998	1	1	1	1	1	1	1
1999	1	1	1	1	1	1	1
2000	1	1	1	1	1	1	1
2001	1	1	1	1	1	1	1
2002	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1
2004	1	1	1	1	1	1	1
2005	1	1	1	1	1	1	1
2006	1	1	1	1	1	1	1

Stadt/Jahr	Ort 8	Ort 9	Ort 10	Ort 11	Ort 12	Ort 13	Ort 14
1988	1						
1989							
1990	1						
1991							
1992							
1993							
1994		1	1	1	1	1	1
1995		1	1	1	1	1	1
1996		1	1	1	1	1	1
1997	1	1	1	1	1	1	1
1998		1	1	1	1	1	1

1999	1	1	1	1	1	1	1
2000		1	1	1	1	1	1
2001	1	1	1	1	1	1	1
2002		1	1	1	1	1	1
2003	1	1	1	1	1	1	1
2004		1	1	1	1	1	1
2005	1	1	1	1	1	1	1
2006		1	1	1	1	1	1

Tabelle 4.5: Anzahl der Proben zu Brassenfilets je Stadt und Jahr

Aus Tabelle 4.5 kann man entnehmen, dass bis einschließlich 1993 jährlich insgesamt höchstens fünf Proben zur Verfügung stehen. Im Jahr 1994 wächst die Anzahl der Proben auf sieben, ab 1995 sind es jährlich stets mindestens 13, maximal 14 Proben. Ein Vergleich der Probenahmestandorte der Jahre 1988 bis 1993 zeigt, dass diese vorwiegend aus ländlichen Regionen und Kleinstädten stammen, vgl. Tabelle 4.6. Im Vergleich zum Zeitraum 1988-1993 steigt ab dem Jahr 1994 die Zahl der Proben aus Mittel- und Großstädten stark an. Waren es anfangs noch Null bis maximal zwei Proben, so sind es 1994 bereits drei Proben und ab dem Jahr 1995 stets sieben Proben. Der Anteil von Proben aus ländlichen Regionen und Kleinstädten ist ab dem Jahr 1995 in etwa so groß, wie der Anteil an Proben aus Mittel- und Großstädten. Dies könnte ein Indikator für die in den Jahren 1994 und 1995 gefundenen innovativen Ausreißer sein.

Jahr/Gebietstyp	ländlicher Raum	Kleinstadt	Mittelstadt	Großstadt
1988	1			
1989				
1990	1			
1991	2			
1992	1	1		1
1993	2	1	1	1
1994	2	2	1	2
1995	4	2	4	3
1996	4	2	4	3
1997	5	2	4	3
1998	4	2	4	3
1999	5	2	4	3
2000	4	2	4	3
2001	5	2	4	3
2002	4	2	4	3
2003	5	2	4	3
2004	4	2	4	3
2005	5	2	4	3
2006	4	2	4	3

Tabelle 4.6: Anzahl der Proben zu Brassenfilets je Gebietstyp und Jahr

Ein weiterer Ansatz zur Untersuchung der innovativen Ausreißer besteht in der Betrachtung der jährlichen Profile der ausreißerbereinigten gemittelten ilr Daten. Diese sind in den Abbildungen 4.26-4.31 gegeben. Optisch auffällig ist das Profil des Jahres 1994. Dessen sechster Balken ist bedeutend kürzer als in allen anderen Profilen. Dies kann ein Hinweis auf einen möglichen Ausreißer sein. Das Jahr 1995 zeigt im Vergleich zu den Profilen der Vorjahre die stärkste Ausprägung des vierten Balkens, während alle anderen Mittelwerte sich kaum unterscheiden. Aus der graphischen Analyse erhält man den Hinweis, dass die Beobachtung des Jahres 1994 eventuell durch einen Ausreißer kontaminiert wurde. Dies bestärken die Ergebnisse der rein statistischen Analyse. Das Profil des Jahres 1995 wirkt gesamt betrachtet eher unauffällig und gibt keinen optischen Hinweis auf eine eventuelle Kontamination durch Ausreißer.

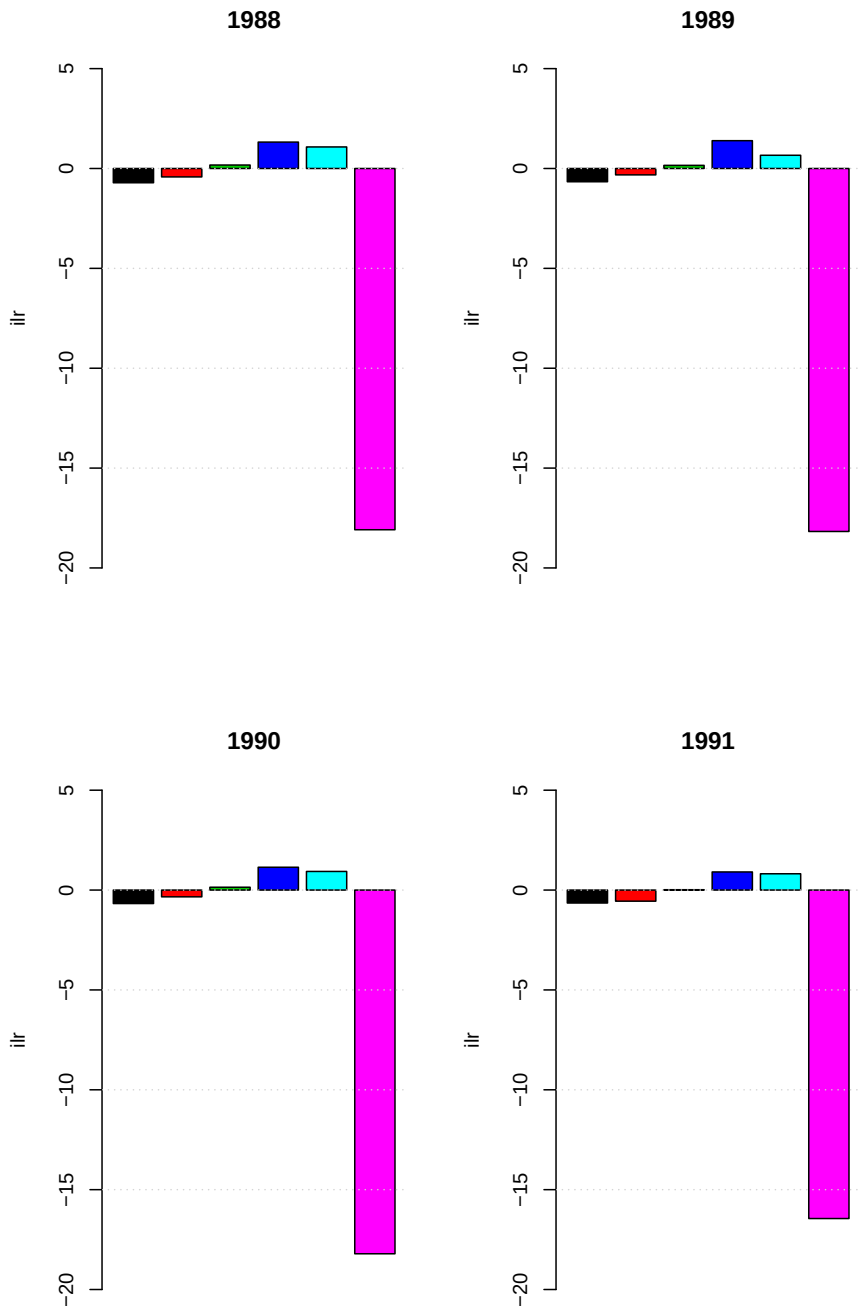

Abbildung 4.26: Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 1

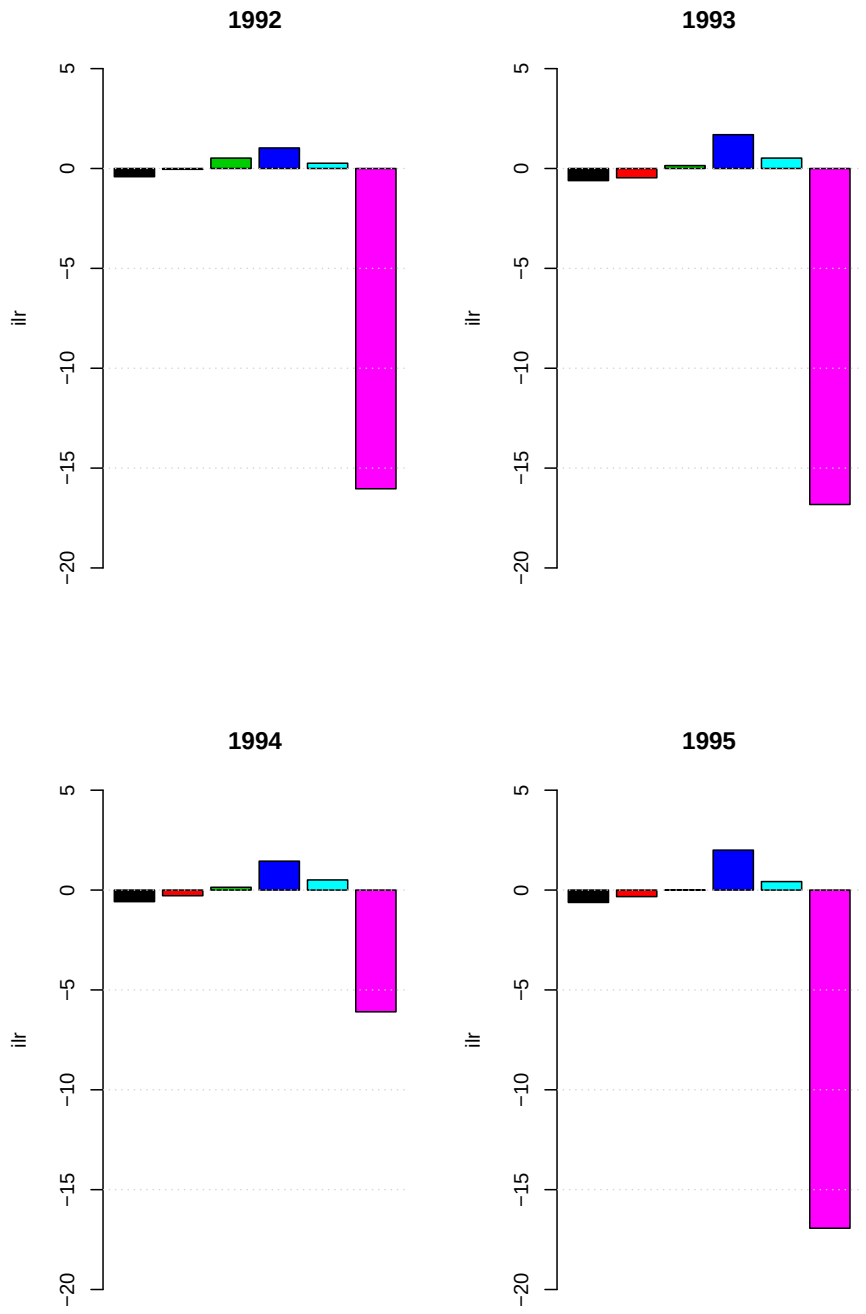

Abbildung 4.27: Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 2

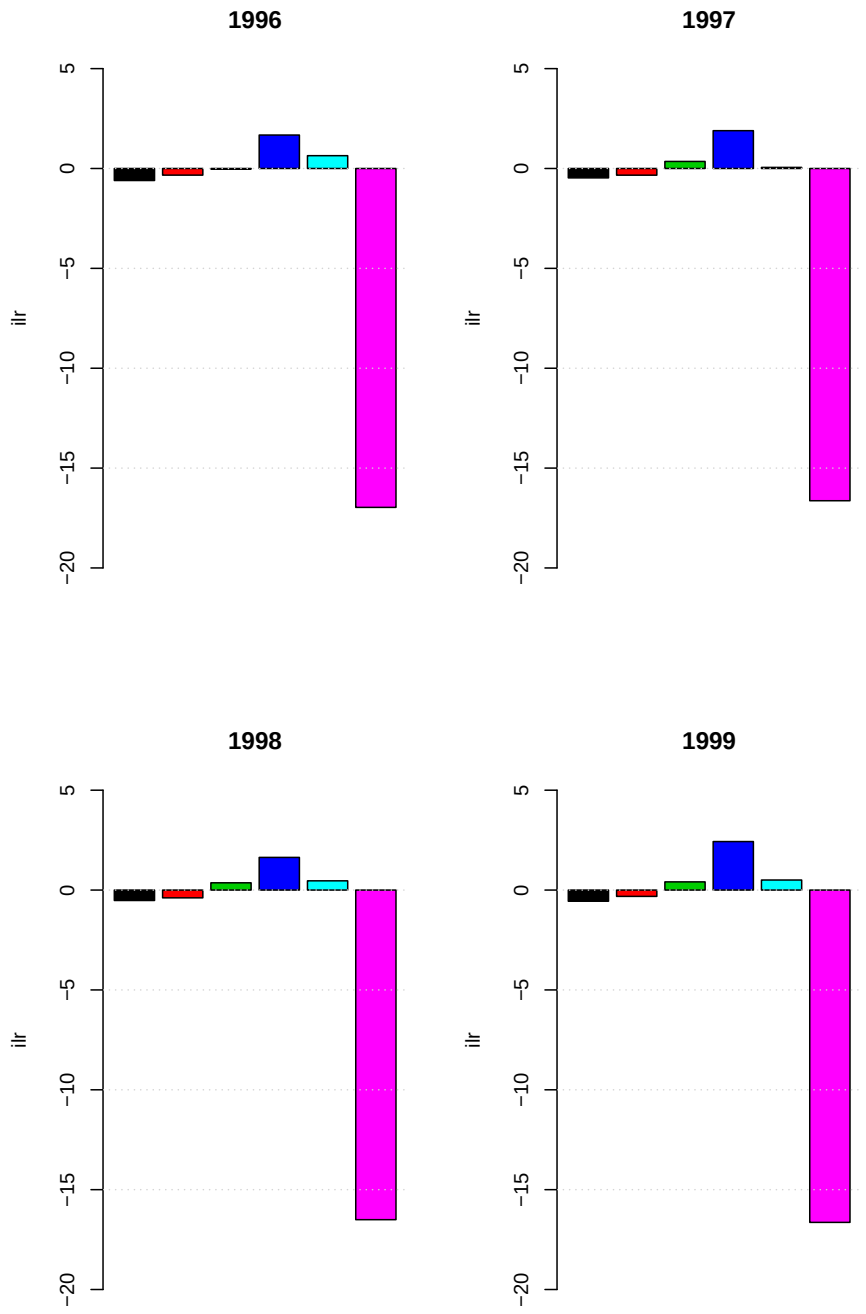

Abbildung 4.28: Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 3

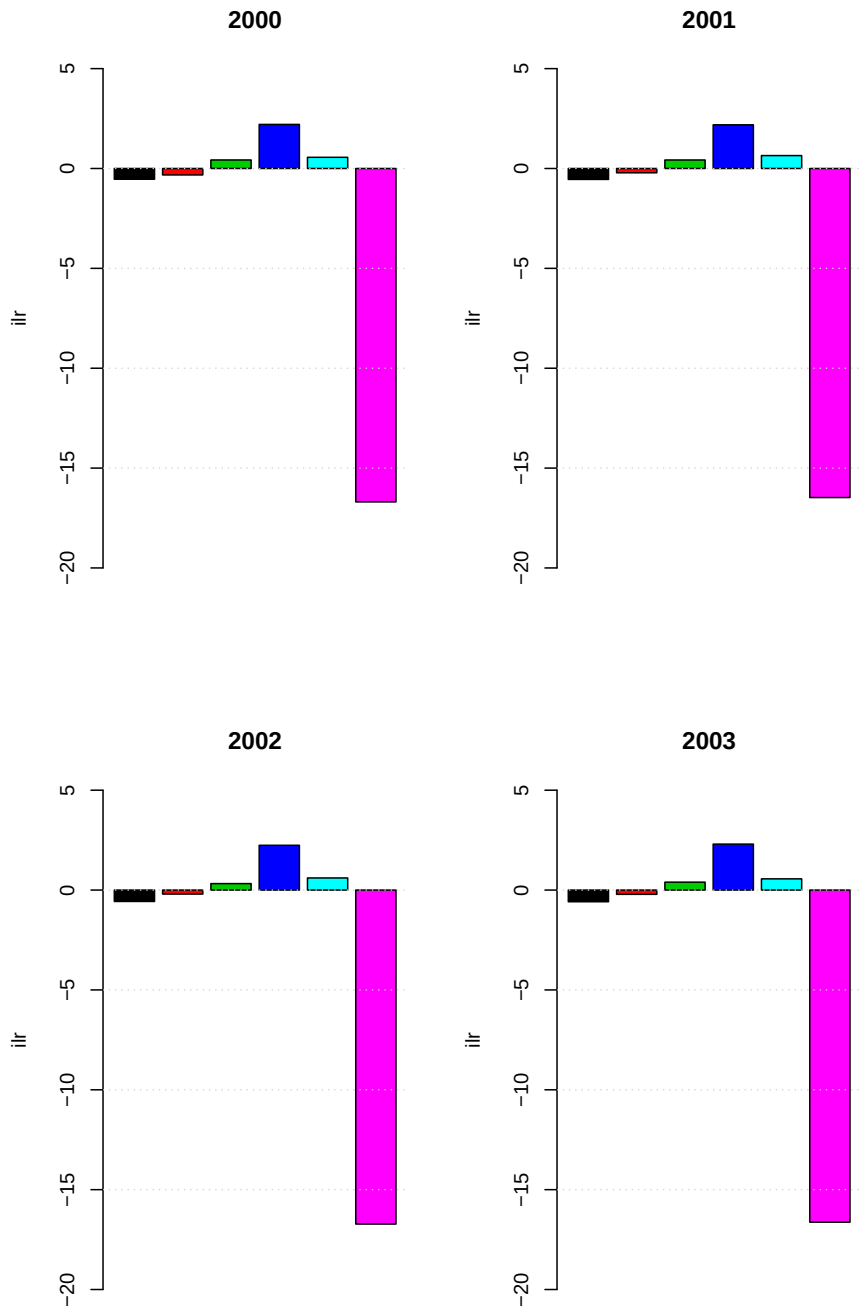

Abbildung 4.29: Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 4

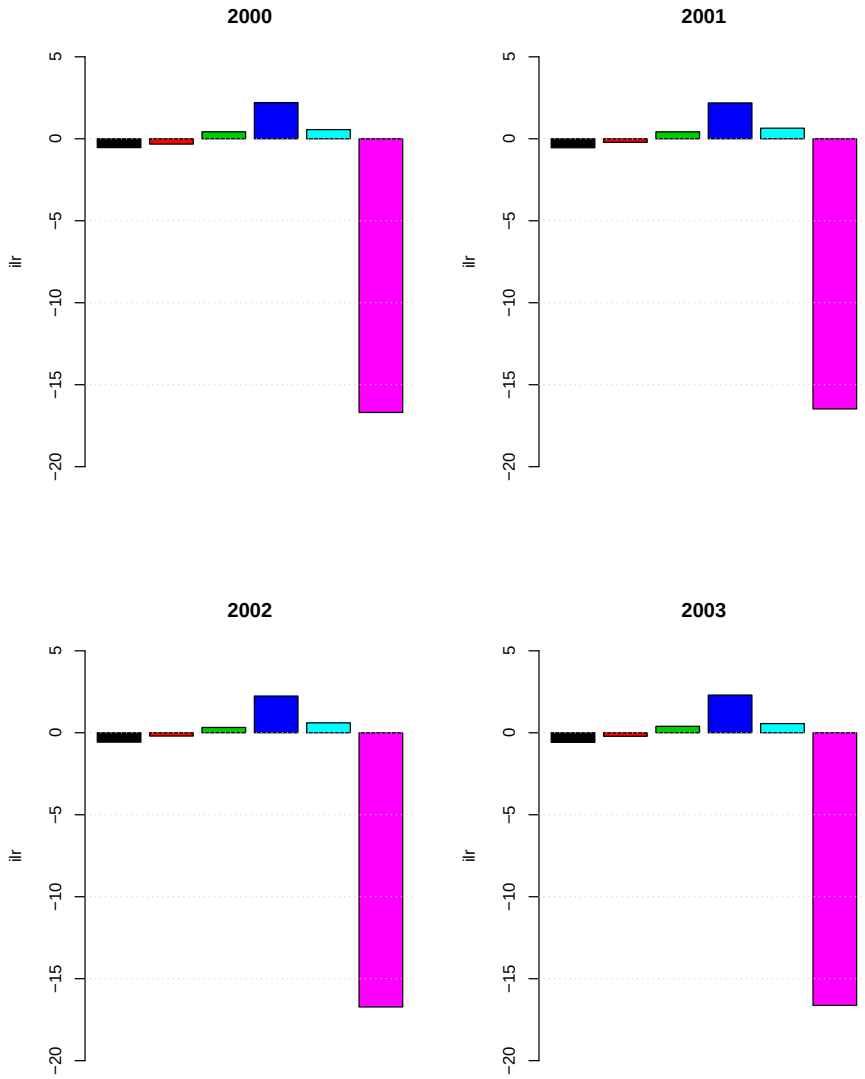

Abbildung 4.30: Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 5

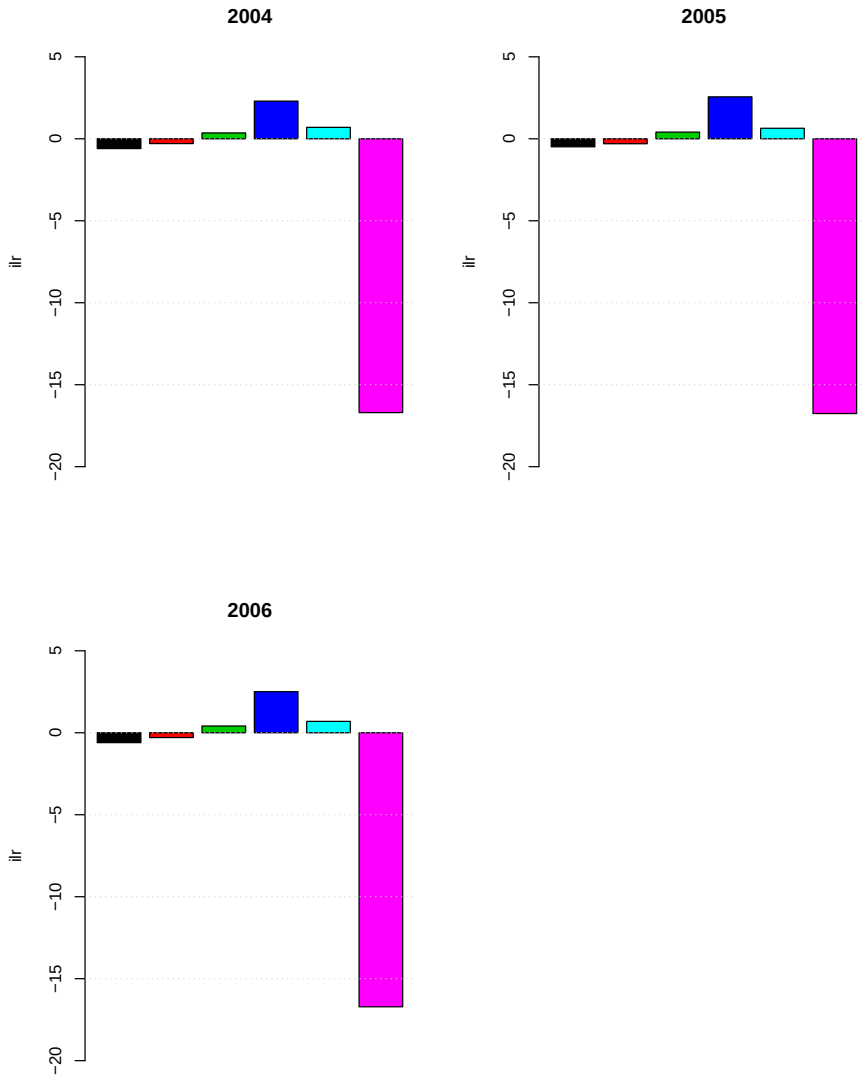

Abbildung 4.31: Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 6

Kapitel 5

Allgemeine Auswertung und Diskussion

In diesem Abschnitt werden zunächst die Ergebnisse der Auswertungen sowohl für uni- und multivariate Querschnittsdaten, als auch für multivariate Zeitreihen vorgestellt. Probleme bei der Datenauswertung und Möglichkeiten diese zu beheben sollen erläutert werden. Sowohl der Kontaminationsanteil und die Kontaminationswahrscheinlichkeit durch statistische Ausreißer (vgl. Abschnitt 3.1.9), als auch deren Einfluss auf die Schätzung der natürlichen Variabilität sind Gegenstand dieser Arbeit. Daher werden diese beiden zentralen Aspekte gesondert diskutiert (vgl. Abschnitt 3.1.7). Der Einfluss der Methode zum Auffinden statistischer Ausreißer in multivariaten Querschnittsdaten wird beleuchtet. In den Anhängen D-W sind Kongenerenprofile ausreißerbereinigter Daten nach dem Verfahren von Pena und Prieto gegeben.

5.1 Univariate Querschnittsdaten

Es wurden 18 Jahresquerschnittsdatensätze zu dl PCB 118 in verschiedenen Biota analysiert, darunter Aale (Anguilla anguilla), Barben (Barbus barbus), Hechte (Esox lucius), Zander (Sander lucioperca) und Brassen (Abramis brama). Da teilweise aus den Daten nicht ersichtlich war, in welchem Gewebe die PCB Konzentrationen in den Brassen gemessen wurden (z.B. in der Muskulatur oder in der Leber), wurden die Auswertungen gesondert durchgeführt. In der Datenbank waren die Informationen zum analysierten Teil der Brassen durch die Ausdrücke „Brassen“ und „Brassen-Filet“ gegeben. Es wurden daher zwei verschiedene Datensatztypen erstellt: „Brassen“ und „Brassen-Filet“. Die Zahl der Datensätze, der vorhandenen Proben und die Zahl der mit dem Hampel-Test gefundenen statistischen Ausreißer sind in Tabelle B.1 von Anhang B gegeben.

Insgesamt $l = 7$ Datensätze waren durch statistische Ausreißer kontaminiert, was eine geschätzte Kontaminationswahrscheinlichkeit von $\hat{p}_p = 0.4118$ und ein geschätztes Konfidenzintervall der Gestalt $(0.1844; 0.6708)$ liefert. Die Gesamtzahl der Proben beträgt $n = 606$, wobei $m = 29$ Datenpunkte als Ausreißer mit dem Ausreißertest nach Hampel detektiert wurden. Die Schätzung des Kontaminationsanteils beträgt $\hat{p}_{prop} = 0.04785$. Das zugehörige Konfidenzintervall lautet $(0.03228; 0.068)$.

Die Schätzungen der natürlichen Variabilität für die Gesamtdaten und die bereinigten Daten sind in Tabelle B.17 gegeben. Tabelle 5.1 zeigt, dass der Median der Änderungen der Schätzungen der natürlichen Variabilität durch Ausschluss statistischer Ausreißer 0% beträgt.

Min	25%-Perz.	Median	75%-Perz.	Max
-53.92	-10.96	0	0	0

Tabelle 5.1: Kenngrößen der Änderungen der Schätzungen der natürlichen Variabilität für univariate Querschnittsdaten nach Ausschluss statistischer Ausreißer in %

5.2 Multivariate Querschnittsdaten

In diesem Abschnitt werden allgemeine Ergebnisse der Auswertung von multivariaten Querschnittsdatensätzen zusammengestellt und diskutiert. Darüber hinaus werden allgemeine Schätzungen der Kontaminationswahrscheinlichkeit p_p und des Kontaminationsanteils p_{prop} angegeben. Die Ausreißeranalytik anhand der robusten Mahalanobis-Distanz wurde, dem Beispiel von Filzmoser und Hron folgend, zum Signifikanzniveau $\alpha = 0.025$ durchgeführt [vgl. FH08, S.236]. Die statistische Ausreißeranalytik nach Pena und Prieto wurde zum Signifikanzniveau $\alpha = 0.01$ durchgeführt, wie im Algorithmus aus Abschnitt 3.4.1.1 beschrieben [s.a. PP01b, S.290]. Sämtliche Auswertungen zur Kontaminationswahrscheinlichkeit p_p und zum Kontaminationsanteil p_{prop} wurden zum Signifikanzniveau $\alpha = 0.05$ vorgenommen. Für die Auswertung anhand des Projektionsverfahrens fand die Implementierung des Matlab Codes von Pena und Prieto in **R** Verwendung [Pri00].

Die Untersuchung multivariater Querschnittsdaten auf statistische Ausreißer erfolgte sowohl unter Verwendung der robusten Mahalanobis-Distanz, als auch mit Hilfe des Projektionsverfahrens nach Pena und Prieto. Zunächst werden hier die Ergebnisse vorgestellt und anschließend untersucht, ob eines der beiden Verfahren

möglicherweise konservativer arbeitet in dem Sinne, dass weniger Datenpunkte als Ausreißer deklariert werden.

Die Untersuchungen der multivariaten Querschnittsdatensätze auf statistische Ausreißer nach Pena und Prieto und anhand der robusten Mahalanobis-Distanz sind in den Tabellen B.3-B.14 zusammengefasst. In Tabelle 5.2 ist der Einfluss statistischer Ausreißer auf die Schätzung der natürlichen Variabilität dargestellt. Diese kann, wie bereits in Abschnitt 4.1.2.2 diskutiert, durch den Ausschluss statistischer Ausreißer sowohl vergrößert, als auch verkleinert werden. Die Änderungen der natürlichen Variabilität sind in % angegeben. Details zu Änderungen der natürlichen Variabilität durch die Bereinigung der einzelnen Datensätze von statistischen Ausreißern anhand der robusten Mahalanobis-Distanz und anhand des Projektionsverfahrens nach Pena und Prieto sind in den Tabellen B.18 und B.19 gegeben.

	Min	25%-Perz.	Median	75%-Perz.	Max
Mahalanobis	-96.62	-38.57	-26.48	-6.07	195.09
Pena-Prieto	-65.96	-31.71	-18.17	-4.93	21.03

Tabelle 5.2: Kenngrößen der Änderungen der Schätzungen der natürlichen Variabilität in multivariaten Querschnittsdaten nach Ausschluss statistischer Ausreißer in %

Interessant ist die Beantwortung der Frage, ob die Schätzung der natürlichen Variabilität nach Ausschluß statistischer Ausreißer sich eher vergrößert oder verkleinert. Zu diesem Zweck wird in den Tabellen B.18 und B.19 eine Vergrößerung der Schätzung kodiert mit 1 und eine Verkleinerung kodiert mit 0. Enthält ein Datensatz keine Ausreißer, so ändert sich die Schätzung der natürlichen Variabilität nicht. Dies wird mit dem Wert 9999 kodiert. Die Ergebnisse für die Ausreißerana-

lytik anhand des Verfahrens nach Pena und Prieto sind in Tabelle B.16 gegeben, die Ergebnisse der Auswertungen anhand der robusten Mahalanobis-Distanz sind in Tabelle B.15 zusammengefasst.

Im Folgenden soll der Einfluss statistischer Ausreißer auf die Schätzung der natürlichen Variabilität untersucht werden. Daher werden ausschließlich die durch statistische Ausreißer kontaminierten Datensätze betrachtet. Durch die Kodierung liegen Beobachtungen der Art $x_1, \ldots, x_m \in \{0, 1\}$ vor. Diese werden aufgefasst als Realisationen von stochastisch unabhängigen und identisch binomial-$(1, p_g)$-verteilten Zufallsvariablen $X_1, \ldots, X_m$. Eine Schätzung des Modellparameters p_g ist gegeben durch $\hat{p}_g = \bar{x}$. Um zu überprüfen, ob eine Vergrößerung der Schätzung der natürlichen Variabilität so wahrscheinlich ist wie eine Verkleinerung, wird nach Formel 3.7 ein Konfidenzintervall für p_g berechnet (vgl. Abschnitt 3.1.9). Das Testproblem lautet wie folgt:

$$H_0 : \; p_g = 0.5 \;\text{ gegen }\; H_1 : \; p_g \neq 0.5 \; . \tag{5.1}$$

Zunächst werden die Ergebnisse der Analysen anhand der robusten Mahalanobis-Distanz diskutiert. Aufgrund der Dualität von Konfidenzintervallen und statistischen Tests [vgl. HRS10, S.49], ist es möglich die Testentscheidung anhand des Konfidenzintervalles zu treffen. Die Nullhypothese in Testproblem 5.1 wird abgelehnt, da $0.5 \notin (0.0751; 0.2610)$ gilt. Untersucht man weiterhin das Testproblem

$$H_0 : \; p_g \geq 0.5 \;\text{ gegen }\; H_1 : \; p_g < 0.5 \; . \tag{5.2}$$

zum Signifikanzniveau $\alpha = 0.025$, wird die entsprechende Nullhypothese eben-

falls signifikant abgelehnt, da das zugehörige Konfidenzintervall $(0, 0.2610)$ den Wert 0.5 nicht überdeckt. Dies zeigt, dass sich die Schätzung der natürlichen Variabilität eher verringert, als vergrößert, wenn statistische Ausreißer eliminiert werden. Eine analoge Vorgehensweise für die nach Pena und Prieto kontaminierten Datensätze liefert in Testproblem 5.1 das Konfidenzintervall $(0.0788; 0.2569)$ zum Signifikanzniveau $\alpha = 0.05$ und in Testproblem 5.2 das Konfidenzintervall $(0; 0.2569)$ zum Signifikanzniveau $\alpha = 0.025$ für den unbekannten Parameter p_g. Auch hier zeigt sich, dass der Ausschluss statistischer Ausreißer eher zu einer Verringerung der Schätzung der natürlichen Variabilität führt, als zu einer Vergrößerung. Mit dem Verfahren nach Pena und Prieto konnten mehr Datensätze ausgewertet werden als mit der robusten Mahalanobis-Distanz. Die robuste Mahalanobis-Distanz unter Verwendung des MCD-Schätzers kann für d-dimensionale Daten nur dann eingesetzt werden, wenn mindestens $2d$ Beobachtungen vorliegen (vgl. Abschnitt 3.4), wohingegen das Verfahren nach Pena und Prieto gelegentlich auch Analysen zulässt, wenn weniger als $2d$ Datenpunkte vorliegen. Insgesamt konnten 66 Datensätze mit Hilfe der robusten Mahalanobis-Distanz auf statistische Ausreißer untersucht werden. Von diesen waren alle durch statistische Ausreißer kontaminiert. Die Gesamtzahl aller Proben in den 66 Datensätzen beträgt 1564, wovon 430 als statistische Ausreißer deklariert wurden. Die Kontaminationswahrscheinlichkeit wird auf $\hat{p}_{p,Maha} = 1$ geschätzt und das zugehörige Konfidenzintervall hat die Gestalt $(0.9456; 1)$. Die Schätzung des Kontaminationsanteils beträgt $\hat{p}_{prop,Maha} = 0.2749$. Das zugehörige Konfidenzintervall lautet $(0.2529; 0.2978)$.

Nach Filzmoser und Hron sind Projektionsverfahren besonders nützlich, um eine verlässliche statistische Ausreißeranalytik in multivariaten Datensätzen von

vergleichsweise geringem Umfang einzusetzen [vgl. FH08, S.236]. Insgesamt konnten mit dem Projektionsverfahren von Pena und Prieto 86 multivariate Datensätze ausgewertet werden. Von diesen waren 72 durch statistische Ausreißer kontaminiert. In den insgesamt 1852 Proben detektierte das Verfahren 389 Ausreißer. Die Kontaminationswahrscheinlichkeit wird geschätzt auf $\hat{p}_{p,Pena} = 0.8372$. Das zugehörige Konfidenzintervall hat die Gestalt $(0.7420; 0.9080)$. Die Schätzung des Kontaminationsanteils beträgt $\hat{p}_{prop,Pena} = 0.2100$ und das zugehörige Konfidenzintervall ist $(0.1917; 0.2293)$.

Um zu untersuchen, ob beide Verfahren signifikant verschiedene Kontaminationswahrscheinlichkeiten und Kontaminationsanteile liefern, wird der exakte Test von Fisher eingesetzt [vgl. Bas94, S.249 ff.]. Das Signifikanzniveau beträgt $\alpha = 0.05$. Die Testprobleme lauten wie folgt:

$$H_0 : \; p_{p,Pena} = p_{p,Maha} \;\text{ gegen }\; H_1 : \; p_{p,Pena} \neq p_{p,Maha} \tag{5.3}$$

$$H_0 : \; p_{prop,Pena} = p_{prop,Maha} \;\text{ gegen }\; H_1 : \; p_{prop,Pena} \neq p_{prop,Maha} \;. \tag{5.4}$$

Der p-Wert in Testproblem 5.3 beträgt 0.48. Die Nullhypothese kann nicht abgelehnt werden. Für Testproblem 5.4 ergibt sich ein p-Wert von 0.0005995. Die Nullhypothese in Testproblem 5.4 wird daher abgelehnt. Das Verfahren nach Pena und Prieto liefert somit einen anderen Kontaminationsanteil als die Ausreißeranalytik unter Verwendung der robusten Mahalanobis-Distanz. In einem weiteren Schritt wird das Testproblem

$$H_0 : \; p_{prop,Pena} \geq p_{prop,Maha} \;\text{ gegen }\; H_1 : \; p_{prop,Pena} < p_{prop,Maha} \tag{5.5}$$

geprüft. Der p-Wert beträgt 0.0003154. Da dieser kleiner als das Signifikanzniveau $\alpha = 0.025$ ist, wird die Nullhypothese in Testproblem 5.5 abgelehnt. Daraus folgt, dass der Kontaminationsanteil unter Verwendung der robusten Mahalanobis-Distanz größer als bei Einsatz des Projektionsverfahrens nach Pena und Prieto ist. Für die hier untersuchten Datensätze kann gefolgert werden, dass die Methode von Pena und Prieto häufig konservativer arbeitet und weniger statistische Ausreißer in den Datensätzen detektiert. Dies ist möglicherweise auf die unterschiedlichen Signifikanzniveaus von 0.01 und 0.025 zurückzuführen. Eine genauere Betrachtung zeigt jedoch, dass die Mengen der anhand beider Verfahren gefundenen Ausreißer je nach Datensatz sowohl disjunkt sein, als auch nichtleere Schnittmengen besitzen können. Wären die unterschiedlichen Signifikanzniveaus tatsächlich der Grund für die häufig geringere Zahl detektierter statistischer Ausreißer nach der Methode von Pena und Prieto, so wäre die Menge der Ausreißer eine Teilmenge der Ausreißer, die unter Verwendung der robusten Mahalanobis-Distanz detektiert wurden. Dies kann jedoch nicht beobachtet werden. Daher muss die Formulierung „Die Methode von Pena und Prieto arbeitet häufig konservativer.“ genauer gefasst werden zu „Die Methode von Pena und Prieto arbeitet im Vergleich zur robusten Mahalanobis-Distanz konservativer, was die Anzahl der gefundenen Ausreißer betrifft, wobei beide Methoden unterschiedliche Ausreißer detektieren können“. Diese Aussagen gelten zunächst lediglich für die hier untersuchten Typen von Datensätzen. Eine Erweiterung der Aussagen auf andere Datenmengen ist nicht ohne Weiteres zulässig.

Für die hier untersuchten Querschnittsdatensätze wurden Kongenerenprofile erstellt, nachdem die Daten von statistischen Ausreißern bereinigt wurden. Zur Bereinigung der Datensätze wurde das Projektionsverfahren von Pena und Prieto ein-

gesetzt, da es zum einen aus statistischer Sicht weniger Datenpunkte als Ausreißer klassifiziert, als die robuste Mahalanobis-Distanz, zum anderen konnten mehr Datensätze auf statistische Ausreißer untersucht und bereinigt werden. Die Kongenerenprofile wurden stets aus Jahresdaten erstellt. Eine Aggregation der Daten über mehrere Jahre erfolgte nicht, da die Daten, wie sich im Abschnitt zu multivariaten Zeitreihen zeigen wird, oft trendbehaftet sind. Unter Berücksichtigung der Tatsache, dass die Informationen eines Datenpunktes in den Verhältnissen seiner einzelnen Einträge zu finden sind (vgl. Abschnitt 3.1.4), werden Kongenerenprofile verzerrt, wenn Trends unberücksichtigt bleiben. Man überlege sich Folgendes: gegeben sei ein Kompositionsdatenpunkt $x_i = (0.1, 0.9)'$ im Jahre i. Angenommen es liegt ein Trend vor, der besagt, dass der Anteil von Stoff 1 pro Jahr um 0.05 steigt. Dann läge für das Jahr $i + 1$ der Datenpunkt $x_{i+1} = (0.15, 0.85)$ vor. Das Verhältnis der Einträge in Jahr i beträgt $0.1/0.9 \approx 0.11$, das Verhältnis der Einträge in Jahr $i + 1$ ergibt sich zu $0.15/0.85 \approx 0.18$. Die Bildung des Mittelwertvektors nach Aggregation beider Jahresdaten (unter Verwendung der ilr Transformation) liefert $\bar{x} = (0.1228286, 0.8771714)'$. Das Verhältnis der Einträge errechnet sich zu $0.1228286/0.8771714 \approx 0.14$. Offenbar verzerren Trends die Verhältnisse der Daten, wenn diese aggregiert und Mittelwertvektoren gebildet werden. Die Erstellung der Kongenerenprofile erfolgte daher jahresspezifisch. Sie sind in den Abbildungen D.1-W.4 gegeben.

5.3 Stichprobenumfang und Dimension der Daten

Für die hier untersuchten multivariaten Querschnittsdaten entsprach der Stichprobenumfang n in den meisten Fällen circa dem doppelten der Dimension d der

Daten. Ausreißer sind daher stets kritisch zu hinterfragen, denn aufgrund des vergleichsweise geringen Stichprobenumfangs spiegeln sie möglicherweise lediglich die natürliche Variabilität wieder. Alle folgenden Auswertungen sollten daher stets sowohl unter Berücksichtigung aller Daten, als auch unter Ausschluss von Ausreißern durchgeführt werden. An dieser Stelle wird empfohlen, neben dem Verhältnis von Stichprobengröße und Dimension auch die Anzahl gefundener Ausreißer q zu betrachten. Selbst wenn $n \geq 2d$ gilt, so kann eine große Anzahl gefundener Ausreißer (z.B. wenn $n < 2d + q$ gilt) in der Stichprobe ein Indikator für die natürliche Variabilität der Daten sein.

Bei den untersuchten Proben handelt es sich um Monitoringdaten, welche oft eine höhere Variabilität als Labordaten aufweisen. Aus diesem Grund sollten die gefundenen Ausreißer kritisch hinterfragt werden. Wenn das Verhältnis zwischen Stichprobengröße und Dimension nahe 2:1 ist, können sachlogische Hinweise wie z.B. Probenahmestandorte, als zusätzliche Informationen ausgewertet werden und dazu dienen, Messwerte als Ausreißer zu klassifizieren oder wieder in den Datenbestand aufzunehmen.

Eine rein statistische Analyse ohne Berücksichtigung der begleitenden Informationen wie z.B. die Jahreszeit zu der die Probenahme stattfand, den Standort der Probenahme, die Eigenschaften der untersuchten Spezies (z.B. Standorttreue von Brassen) und mögliche Umweltkontaminanten in der Umgebung des Probenahmestandortes (z.B. Kläranlagen) ist nicht empfehlenswert. Liegen vergleichsweise wenig Daten vor, so kann eine differenzierte Sicht auf alle Aspekte der Probenahme selbst, Aufschluss darüber geben, ob ein Datenpunkt tatsächlich als Ausreißer zu bewerten ist oder nicht.

5.4 Probleme mit der Mahalanobis-Distanz

Die Auswertung der Daten anhand der robusten Mahalanobis-Distanz war nicht immer möglich. Zum einen betraf dies Datensätze $x_1, \ldots, x_n \in \mathbb{R}^p$ mit $n < 2p$, da die Funktion `Moutlier` im **R**-Paket `chemometrics` für den Fall $n \geq 2p$ implementiert ist [FV10], da sie den MCD Schätzer verwendet (vgl. Abschnitt 3.3). Zum anderen traten Probleme der Art auf, dass zu viele Datenpunkte in einer Hyperfläche des $\mathbb{R}^p$ angesiedelt waren. Dieser Fall konnte für folgende Datensätze zu Weidelgras beobachtet werden: PCDD im Jahr 2001 (18 von 23 Datenpunkten in einer Hyperebene), PCDF im Jahr 2001 (20 von 23 Datenpunkten in einer Hyperebene) und PCDF im Jahr 2002 (17 von 21 Datenpunkten in einer Hyperebene). Warum dieser Umstand die Bestimmung der robusten Mahalanobis-Distanz unter Verwendung des MCD Schätzers aus mathematischer Sicht unmöglich macht, wurde bereits in Abschnitt 3.3 erläutert.

Eine Möglichkeit dennoch eine statistische Ausreißeranalytik durchzuführen bestünde darin, eine Koordinatentransformation für die Daten vorzunehmen. Diese Vorgehensweise ist sinnvoll, wenn der vollständige Datensatz Teilmenge einer Hyperebene ist. In einem solchen Fall wäre es möglich, die Daten derart in einem anderen Koordinatensystem darzustellen, dass alle Punkte z.B. als p-ten Eintrag die Null besäßen, d.h. die p-dimensionale Darstellung der Daten wäre redundant [vgl. Fis09, Abbildungen S.213]. Eine Reduktion der Dimension auf $p - 1$ durch Entfernung der p-ten Komponente wäre möglich und die robuste Mahalanobis-Distanz unter Verwendung des MCD Schätzers könnte für die so gewonnen Daten bestimmt werden. Solange jedoch mindestens ein Punkt des Datensatzes nicht in der Hyperebene enthalten ist, würde es auch nach Koordinatentransformation kei-

ne redundanten Komponenten geben.

5.5 Kongenerenprofile

In den Anhängen D-W sind ausreißerbereinigte Kongenerenprofile zur Belastung verschiedener Biota durch PCB und Dioxine gegeben. Die Bereinigung von statistischen Ausreißern erfolgte nach dem Projektionsverfahren von Pena und Prieto (vgl. Abschnitt 3.4.1). Profile von Datensätzen, die nicht mit der Methode analysiert und bereinigt werden konnten, sind darin nicht enthalten. Für 78 der insgesamt 86 untersuchten Datensätze sind die entsprechenden Kongenerenprofile angegeben. Dabei wiesen lediglich zwei Datensätze fehlende Werte auf. Bei diesen handelt es sich um Daten zu PCB Leitkongeneren in Brassen-Filets aus den Jahren 1995 und 1996 mit einem bzw. zwei fehlenden Werten zu PCB 28. In diesen Fällen wurden die fehlenden Werte geschätzt und die entsprechenden Kongenerenprofile im Subtext gekennzeichnet.

Da die Kongenerenprofile aus Jahresdaten generiert wurden, können eventuelle Trends in der Entwicklung der Belastungssituation des entsprechenden Biotatyps (z.B. Weidelgras) visuell untersucht werden, indem die Entwicklung der Profile über die Zeit betrachtet wird. Auffällig ist, dass die Profile eines Biotatyps sich über die Zeit hinweg visuell stark ähneln, was die optischen Verhältnisse zwischen den durchschnittlichen Konzentrationen der einzelnen PCB- bzw. Dioxinkongenere betrifft. Lediglich die absoluten Balkengrößen ändern sich und suggerieren teilweise die Präsenz eines Trends (vgl. z.B. Abbildungen J.1-J.6).

Die Profile der PCB Leitkongenere in Grünkohl und Weidelgras wurden für die Jahre 2000-2003 erstellt (vgl. Anhänge L und M). Jeweils das Balkendiagramm

des Jahres 2000 für Grünkohl und für Weidelgras unterscheiden sich optisch von denen der Jahre 2001-2003. Für Grünkohl und Weidelgras liegt keine Zeitreihe für PCB Leitkongenere mit ausreichend großem Beobachtungshorizont vor, so dass eine entsprechende Ausreißeranalytik mittels der Projektionsmethode von Galeano et al. (vgl. Abschnitt 3.7) nicht möglich ist. Es ist zu hinterfragen, inwiefern die Kongenerenprofile des Jahres 2000 repräsentativ für die Belastung von Grünkohl oder Weidelgras durch PCB sind, da diese sich optisch von denen der restlichen Jahre unterscheiden. Die Untersuchung der Rohdaten in Bezug auf Probenahmestandorte zeigt keine Auffälligkeiten. Keiner der Datensätze weist fehlende Werte auf, so dass die Verzerrungen der Profile für das Jahr 2000 im Vergleich zu den Profilen der Jahre 2001-2003 nicht auf den Einfluss der Ersetzungsmethode zur Schätzung fehlender Werte zurückzuführen ist. Eine sachlogische Erklärung für die optisch auffälligen Profile des Jahres 2000 kann aus den Rohdaten nicht abgeleitet werden. Ohne zusätzliche Informationen zum Datensatz können keine weiteren Folgerungen abgeleitet werden.

5.6 Multivariate Zeitreihen

In diesem Abschnitt werden die allgemeinen Ergebnisse der statistischen Ausreißeranalytik unter Verwendung des Verfahrens von Galeano et al. für multivariate Zeitreihen (vgl. Abschnitt 3.7) vorgestellt und diskutiert. Zusammenfassungen der Ergebnisse zu den verschiedenen Datentypen PCB Leitkongenere, dl PCB, PCDD und PCDF sind in den Tabellen C.1-C.10 gegeben. Je nach Datenlage waren mehrere Analyseansätze notwendig. Die Zeitreihen bestehen aus ausreißerbereinigten Jahresmittelwertvektoren, die unter Verwendung der ilr Transformation

bestimmt wurden (vgl. Abschnitt 3.1.4). Die Ausreißeranalytik der ilr transformierten Querschnittsdaten erfolgte mit dem Projektionsverfahren nach Pena und Prieto (vgl. Abschnitt 3.4.1). Da die Querschnittsdaten teilweise fehlende Werte aufwiesen, wurden die ausreißerbereinigten ilr-Mittelwertvektoren auf zweierlei Art bestimmt:

- Variante 1: Die fehlenden Werte in den Jahresquerschnittsdaten werden, wie in Abschnitt 3.1.6 beschrieben, geschätzt.
- Variante 2: Datenpunkte die fehlende Werte enthalten, werden von den weiteren Analysen ausgeschlossen.

Darüber hinaus lagen für manche Zeitreihen nicht zu allen Jahren Querschnittsdaten vor, so dass entsprechende ilr-Mittelwertvektoren nicht gebildet werden konnten. Aus dieser Problematik ergaben sich wiederum zwei Möglichkeiten die Daten auszuwerten:

- Variante A: Schätze fehlende Jahresmittelwerte, wie in Abschnitt 3.5.8.1 beschrieben unter Verwendung der LMS-Methode (vgl. Abschnitt 3.5.3).
- Variante B: Entferne alle Jahre, zu denen keine Messwerte gegeben sind. Betrachte die verbliebenen Jahresmittelwerte als Zeitreihe mit äquidistanten Beobachtungen.

An dieser Stelle sei vorab erwähnt, dass lediglich innovative Ausreißer (d.h. Ausreißer vom Typ I) in den multivariaten Zeitreihen detektiert wurden. Die entsprechenden Ergebnisse sind für die einzelnen Datentypen PCB Leitkongenere, dl PCB, PCDD und PCDF je nach Auswertemethode in Tabelle C.11 zusammengefasst. Die Auswertemethoden sind kodiert entsprechend obiger Varianten: 1, 2,

A, B, A1, A2, B1, B2 und 0, wobei 0 bedeutet, dass eine Zeitreihe ohne fehlende Jahresmittelwertvektoren und ohne fehlende Jahresquerschnittsdaten vorliegt. Tabelle C.11 zeigt, für welchen Datentyp zu welcher Auswertevariante wie viele Ausreißer vom Typ I gefunden wurden.

Bevor die Zeitreihen der ilr-Jahresmittelwertvektoren statistisch auf Ausreißer untersucht werden konnten, mussten diese trendbereinigt werden. Die Untersuchung, ob ein linearer deterministischer oder stochastischer Trend vorlag, erfolgte mit dem Test von Giraitis et al. zum Signifikanzniveau $\alpha = 0.05$ (vgl. Abschnitt 3.5.5.1). Sofern Trends vorlagen, wurden diese durch Bildung erster Differenzen eliminiert. Dabei ist zu hinterfragen, ob die Beobachtung von Trends in derart kurzen Zeitreihen (Beobachtungshorizont $6 \leq T \leq 19$) überhaupt sinnvoll ist. Der Test nach Giraitis et al. überprüft die Nullhypothese der Stationarität gegen die Alternativhypothese, dass ein linearer deterministischer oder stochastischer Trend vorliegt. Aufgrund des kurzen Beobachtungshorizontes ist es denkbar, dass eine Änderung des mittleren Niveaus der Zeitreihe in einer oder mehreren Komponenten als Trend fehlinterpretiert wird. Möglicherweise handelt es sich lediglich um eine kurz- bis mittelfristige Änderung des mittleren Niveaus, die sich bei einem längeren Beobachtungshorizont egalisieren würde. Eine sachlogische Begründung, warum die detektierten Trends tatsächlich als solche zu verstehen sind, ergibt sich aus der umweltpolitischen Debatte. So ist der Eintrag von PBT-Stoffen (persistent, bioakkumulierend, toxisch), zu denen PCB und Dioxine zählen, in die Umwelt nach REACh untersagt [vgl. REA06]. Die Umsetzung der Richtlinie führt dazu, dass die in der Umwelt gemessenen Konzentrationen von PCB und Dioxinen, insbesondere in Biota, sich ändern: vorhandene Einträge werden langsam abgebaut und das Hinzukommen neuer Einträge ist begrenzt. In der Folge ist zu

erwarten, dass die Belastung von Biota durch PCB und Dioxine sich langfristig ändert. In Tabelle 5.3 ist nach Datentypen aufgeschlüsselt, wie viele Zeitreihen von ilr-Jahresmittelwerten auf Trends untersucht wurden und wie viele davon als trendbehaftet deklariert wurden. Wie in Abschnitt 3.5.6.5 festgestellt, kann sich

	PCB Leitkongenere	dl PCB	PCDD	PCDF
Anz. Zeitreihen	12	12	8	8
trendbehaftet	6	8	2	7

Tabelle 5.3: Zahl der mit linearen Trends behafteten multivariaten Zeitreihen

die Bildung erster Differenzen auf statistische Ausreißer in Zeitreihen auswirken. Jedoch zeigt Abschnitt 3.5.6.5 auch, dass Ausreißer durch Bildung der ersten Differenzen nicht verschwinden. Vielmehr ist es möglich in den trendbereinigten Zeitreihen statistische Ausreißer zu detektieren und ihre entsprechenden Effekte zu schätzen. Diese Informationen können anschließend genutzt werden, um die Originaldaten von den Effekten der korrespondierenden statistischen Ausreißer zu bereinigen. Daraus folgt, dass eine Bildung erster Differenzen die Ausreißererkennung durch das Projektionsverfahren von Galeano et al. nicht beeinträchtigt, falls ein Trend fälschlicher Weise detektiert wurde.

Wie Tabelle C.11 zeigt, wurden in 40 Zeitreihen insgesamt neun innovative Ausreißer detektiert, die sich auf sieben multivariate Zeitreihen verteilen. Die Kontaminationswahrscheinlichkeit wird demnach geschätzt auf $\hat{p}_p = 0.175$, das nach Formel (3.7) bestimmte 95%-Konfidenzintervall hat die Gestalt $(0.0734; 0.3278)$. Eine Schätzung des Kontaminationsanteils ist für Zeitreihen nicht sinnvoll, da zu einem Zeitpunkt mehrere Ausreißer auftreten können. So ist es rein theoretisch möglich, dass eine Zeitreihe durch mehr statistische Ausreißer kontaminiert ist,

als sie Beobachtungen aufweist.

Es wurde untersucht, ob die innovativen Ausreißer möglicherweise durch die Methode zur Schätzung fehlender Werte in multivariaten Zeitreihen generiert wurden. In der Tat kann beobachtet werden, dass vier der neun innovativen Ausreißer zu Zeitpunkten, deren ilr-Jahresmittelwerte geschätzt wurden, auftreten. Es ist zu hinterfragen, ob die Schätzmethode ungeeignet ist, bzw. ob ihr Versagen auf gewisse Ursachen zurückzuführen ist. Betrachtet man die entsprechenden Zeitreihen, so ist auffällig, dass diese nur maximal eine Folge von vier aufeinander folgenden ilr-Jahresmittelwerten besitzen. Die restlichen Jahresmittelwerte sind getrennt durch fehlende Werte. Die Schätzung der fehlenden ilr-Jahresmittelwerte unter Verwendung der LMS-Methode (vgl. Abschnitt 3.5.8.1 und 3.5.3) kann daher durch die geringe Zahl der zur Verfügung stehenden Datenpunkte ungeeignet sein. Andererseits muss berücksichtigt werden, dass diese Beobachtungen möglicherweise durch statistische Ausreißer kontaminiert sind, d.h. es sollte eine robuste Schätzmethode verwendet werden. Ein praktikabler Ansatz besteht möglicherweise in der Wahl einer anderen robusten Schätzmethode, wie z.B. LTS (least trimmed squares) [vgl. RL87, S.15]. Weiter konnte festgestellt werden, dass Ausreißer möglicherweise durch die Schätzung fehlender Werte in den ursprünglichen Querschnittsdaten generiert wurden und dadurch die Schätzung des ilr-Jahresmittelwertvektors verzerrten (vgl. Abschnitt 4.2.1). Die Methoden zur Ersetzung fehlender Werte und deren Ergebnisse sollten stets hinterfragt werden, um in den Analysen nicht selbst Ausreißer zu generieren.

In Abschnitt 3.5.1.2 wurde vorgeschlagen, die Effekte der innovativen Ausreißer zu entfernen, indem diese wie transiente Ausreißer behandelt werden, um die Zeitreihe dementsprechend zu bereinigen. In der Praxis stellt sich dieser Ansatz

als sehr selten durchführbar heraus. Die Methode konnte nur für den in Abschnitt 4.2.1 ausgewerteten Datensatz angewendet werden. Die restlichen kontaminierten Zeitreihen ließen sich nicht mit der Methode bereinigen. Daher war es nur für eine Zeitreihe möglich, den Einfluss innovativer Ausreißer auf die Schätzung der natürlichen Variabilität zu beobachten (vgl. Abschnitt 4.2.1). Dies mag darin begründet liegen, dass stets AR[1]-Modelle für die Zeitreihen unterstellt wurden, wie in Abschnitt 3.5.1.2 erörtert. Möglicherweise müssen auch AR[p]-Modelle mit $p > 1$ für derart kurze Zeitreihen in Betracht gezogen werden. Weiterhin konnte festgestellt werden, dass es zwingend notwendig ist die MA-Darstellung zu berechnen, auch wenn dies mit einem großen programmiertechnischen Aufwand verbunden sein sollte. Eine Bereinigung multivariater Zeitreihen von den Effekten innovativer Ausreißer ist scheinbar nur selten durch den hier gewählten Ansatz möglich.

Warum wurden in 40 Zeitreihen lediglich 9 statistische Ausreißer gefunden? Es ist möglich, dass durch die nicht entfernten innovativen Ausreißer weitere Ausreißer in den kontaminierten Zeitreihen maskiert werden (vgl. Masking, Abschnitt 3.2.1). Wie in der Analyse der multivariaten Zeitreihe aus Abschnitt 4.2.1 gezeigt, wurden weitere (innovative) Ausreißer detektiert, nachdem die Zeitreihe vom Effekt des ersten gefundenen Ausreißers vom Typ I bereinigt wurde. Dies erklärt aber nicht, warum in 33 multivariaten Zeitreihen kein weiterer statistischer Ausreißer gefunden wurde. Dies kann z.B. darauf zurückzuführen sein, dass die Zeitreihen zu wenige Beobachtungen besitzen, im Vergleich zur Zahl ihrer Komponenten. In diesem Fall lägen zu wenige Informationen vor, um Ausreißer identifizieren zu können. In der Tat weisen 26 der 40 analysierten Zeitreihen $T < 2p$ ilr-Jahresmittelwertvektoren auf.

Eine weitere Erklärung, warum lediglich 9 Ausreißer in 40 Zeitreihen detektiert wurden, ist möglicherweise durch die Tatsache gegeben, dass die Zeitreihen aus ilr-Jahresmittelwertvektoren bestehen, welche wiederum mit Hilfe von ausreißerbereinigten Querschnittsdatensätzen berechnet wurden. Ein Jahresmittelwertvektor kann nur dann als Ausreißer in einer Zeitreihe erkannt werden, wenn sich z.B. die Zusammensetzung der Querschnittsdaten für das entsprechende Jahr von der Zusammensetzung der anderen Jahresquerschnittsdaten unterscheidet, z.B. indem mehr Proben an stärker durch PCB und Dioxine belasteten Standorten generiert wurden als in den Vorjahren (z.B. durch Standortwechsel für die Probenahme) oder durch ungewöhnliche Kontaminationseinträge an den bisherigen Standorten, z.B. durch Überschwemmungen. Die nicht mit Ausreißern kontaminierten Datensätze weisen bzgl. der Probenahmestandorte über die Jahre ein homogenes Profil auf. Besondere Belastungssituationen, wie z.B. Überschwemmungen, sind in den Datensätzen nicht dokumentiert. Es existieren demnach sachlogische Gründe, welche die Annahme bestärken, dass die nach dem Projektionsverfahren von Galeano et al. analysierten und für nicht kontaminiert befundenen Zeitreihen, tatsächlich keine Ausreißer beinhalten.

Anhand der von Aureißern bereinigten Zeitreihendaten lässt sich die Entwicklung der Belastungssituation von Biota durch PCB, Dioxine und dioxinähnliche PCB über einen längeren Zeitraum erfassen. Bereinigte Daten besitzen somit eine enorme Bedeutung für die Erfolgskontrolle der Wirksamkeit umweltpolitischer Maßnahmen, wie z.B. REACh [vgl. REA06].

Von Ausreißern bereinigte Querschnittsdaten können dazu verwendet werden, den Status der Belastung von Biota durch Schadstoffe wie PCB zu verschiedenen Zeitpunkten zu bewerten und verlässliche Kongenerenmuster zu erstellen. Diese

wiederum dienen als Grundlage für weiter führende Analysen, z.B. bei der Suche nach Quellen für erhöhte Konzentrationen von PCB in Tierfutter.

Wie Tabelle 5.1 und 5.2 zeigen, führt die Elimination statistischer Ausreißer tendenziell zu einer Verringerung der Schätzung der natürlichen Variabilität.

Bei der statistischen Analyse von Konzentrationsdaten ist stets darauf zu achten, dass statistische Standardmethoden erst nach entsprechender Transformation der Daten angewendet werden dürfen, denn der Informationsgehalt von Kompositionsdaten ist durch deren Verhältnisse charakterisiert.

Während der statistischen Analysen konnte festgestellt werden, dass die Ersetzung von fehlenden Werten durchaus eine Quelle für Kontaminationen mit statistischen Ausreißern sein kann (vgl. Abschnitt 4.2.1). Aus diesem Grunde ist es notwendig, Datenpunkte in denen fehlende Werte geschätzt wurden, stets kritisch zu hinterfragen, z.B. durch eine visuelle Kontrolle und Tests auf statistische Ausreißer. Treten fehlende Werte auf, sollte die statistische Analyse der Daten sowohl unter Ausschluss der entsprechenden Datenpunkte, als auch unter Verwendung eines Schätzverfahrens für fehlende Werte durchgeführt werden. Auf diese Weise werden vorhandene Informationen in den unvollständigen Datenpunkten genutzt um Ausreißer zu detektieren. Diese Ergebnisse können anschließend durch einen Abgleich mit den entsprechenden Resultaten, die unter Ausschluss von unvollständigen Datenpunkten generiert wurden, auf Plausibilität geprüft werden.

Die Datenerhebung sollte stets vollständig dokumentiert werden, insbesondere ist die Angabe der Bestimmungsgrenzen für die verschiedenen Kongenere unabdingbar um die Daten anschließend sinnvoll aus- und bewerten zu können. So ist beispielsweise die Information, die durch einen Messwert in Höhe von 0 $\mu g/kg$ bei einer Bestimmungsgrenze von 0.1 $\mu g/kg$ gegeben wird eine andere, als bei ei-

ner Bestimmungsgrenze von 0.01 $\mu g/kg$. In beiden Fällen ist der Messwert mit 0 $\mu g/kg$ angegeben, jedoch bedeutet dies im ersten Fall, dass die wahre (unbekannte) Konzentration maximal 0.1 $\mu g/kg$ beträgt, im zweiten Fall kann sie höchstens 0.01 $\mu g/kg$ betragen. Darüber hinaus sind Verfahren zur Ersetzung von Nulldaten nur dann anwendbar, wenn die Bestimmungsgrenzen bekannt sind. Ist dies nicht der Fall, kann die ilr Transformation für die entsprechenden Datenpunkte nicht durchgeführt werden. Sie müssen von den weiteren Analysen ausgeschlossen werden, was das Gesamtbild verzerrt. Die Angabe der Bestimmungsgrenzen bei der Datenerhebung ist somit bedeutend für die anschließende Datenanalyse.

Kapitel 6

Zusammenfassung

Die Untersuchung von Umweltdaten auf statistische Ausreißer bedarf spezieller mathematischer Techniken, wie die ilr Transformation und Methoden zur Ersetzung von Nulldaten (vgl. Abschnitte 3.1.4 und 3.1.5). Erst nachdem die Daten entsprechend aufbereitet wurden, können statistische Methoden zur Ausreißererkennung eingesetzt werden. Diese Arbeit befasste sich mit uni- und multivariaten Querschnittsdaten, sowie multivariaten Zeitreihen, bestehend aus PCB- und Dioxinkonzentrationen in Biota. Die statistischen Analysen umfassten sowohl Daten zu Pflanzen wie Grünkohl und Weidelgras, als auch zu wildlebenden Tieren, wie Brassen und Aalen. Sofern es möglich war, wurden aus mehreren Jahresquerschnittsdatensätzen entsprechende Zeitreihen der Jahresmittelwertvektoren gebildet, um mögliche Auffälligkeiten in der zeitlichen Entwicklung der Belastungssituation von Biota durch PCB und Dioxine erkennen zu können.

Die Analyse der Querschnittsdaten erfolgte mit zwei verschiedenen Verfahren: der robusten Mahalanobis-Distanz und mit dem Projektionsverfahren nach Pena und Prieto (vgl. Abschnitt 3.3). Für beide Verfahren wurden die entsprechenden

Kontaminationswahrscheinlichkeiten und Kontaminationsanteile geschätzt (vgl. Abschnitt 3.1.9 und 5.2). Anschließend erfolgte der Vergleich beider Verfahren anhand des exakten Tests nach Fisher (vgl. Abschnitt 3.1.8). Dadurch konnte festgestellt werden, dass unter Verwendung der robusten Mahalanobis-Distanz mehr statistische Ausreißer detektiert wurden als mit dem Projektionsverfahren. Die robuste Mahalanobis-Distanz war seltener einsetzbar als das Projektionsverfahren nach Pena und Prieto, was darin begründet lag, dass der MCD Schätzer der Kovarianzmatrix teilweise nicht bestimmt werden konnte (vgl. Abschnitt 5.4). Für das Verfahren nach Pena und Prieto ergab sich ein geschätzter Kontaminationsanteil von $\hat{p}_{prop,Pena} = 0.2100$. Der Kontaminationsanteil unter Verwendung der robusten Mahalanobis-Distanz wurde auf $\hat{p}_{prop,Maha} = 0.2749$ geschätzt. Weiterhin wurde der exakte Test von Fisher eingesetzt, um die Kontaminationswahrscheinlichkeit für beide Verfahren miteinander zu vergleichen. Die Nullhypothese, dass gleiche Kontaminationswahrscheinlichkeiten vorliegen, konnte nicht abgelehnt werden. Die Schätzungen der Kontaminationswahrscheinlichkeiten betragen $\hat{p}_{p,Maha} = 1$ und $\hat{p}_{p,Pena} = 0.8372$.

Der Einfluss statistischer Ausreißer auf die Schätzung der natürlichen Variabilität in multivariaten Querschnittsdaten wurde untersucht und in den Tabellen 5.2, B.18 und B.19 zusammengestellt. Der Median der Änderungen der Schätzungen der natürlichen Variabilität durch Ausschluss von Daten, die mittels der robusten Mahalanobis-Distanz als Ausreißer deklariert wurden, beträgt -26.48%. Wurden Daten mit dem Projektionsverfahren von Pena und Prieto als Ausreißer erkannt, so betrug der Median der Änderungen der Schätzungen der natürlichen Variabilität -18.17%.

Die Schätzung des Kontaminationsanteils und der Kontaminationswahrschein-

lichkeit für multivariate Querschnittsdaten beruht auf der statistischen Analyse vieler Datensätze, deren Stichprobengröße n und Dimension d im Verhältnis von circa 2:1 stehen. Es ist daher unbedingt darauf zu achten, dass die gefundenen Ausreißer in ihrer Bedeutung nicht überbewertet werden. Aufgrund der oft vergleichsweise geringen Stichprobenumfänge, wurden einige Datenpunkte möglicherweise fälschlich als statistische Ausreißer detektiert. Dennoch liefern die hier ausgewerteten Datensätze wertvolle Hinweise auf die tatsächliche (unbekannte) Kontaminationswahrscheinlichkeit und den Kontaminationsanteil.

Die Analyse multivariater Zeitreihen auf statistische Ausreißer erfolgte mit dem Projektionsverfahren nach Galeano et al. (vgl. Abschnitt 3.7). Die Auswertung wurde häufig auf mehrere Arten durchgeführt. Entsprechend der Datenlage mussten fehlende Werte in Jahresquerschnittsdaten geschätzt oder die entsprechenden Datenpunkte vor der Bestimmung des Jahresmittelwertvektors vollständig entfernt werden. Falls Zeitpunkte auftauchten, zu denen kein Jahresmittelwertvektor vorlag, erfolgte die Auswertung der betreffenden Zeitreihen auf zwei verschiedene Arten: die fehlenden Jahresmittelwertvektoren wurden geschätzt, oder aber die Zeitpunkte mit fehlenden Jahresmittelwertvektoren wurden ausgelassen und die Zeitreihe der verbliebenen Werte als Reihe mit äquidistanten Beobachtungen betrachtet. Es ergaben sich folglich mehrere Möglichkeiten einen Datensatz auszuwerten.

Bevor die Zeitreihen mit dem Verfahren von Galeano et al. auf statistische Ausreißer untersucht werden konnten, mussten diese noch trendbereinigt werden. Um lineare stochastische oder deterministische Trends zu detektieren wurde der Test von Giraitis et al. eingesetzt. Die Elimination von Trends erfolgte dann über die Bildung erster Differenzen. Insgesamt wiesen 23 der 40 Zeitreihen Trends auf

(vgl. Tabelle 5.3).

Im Resultat konnten durch diese Vorgehensweise 40 Zeitreihen ausgewertet werden. Die Zahl gefundener Ausreißer beträgt neun, diese verteilen sich auf sieben Zeitreihen. Einen Überblick gibt Tabelle C.11. Bei allen neun Ausreißern handelt es sich um innovative Ausreißer. Um eine multivariate Zeitreihe von den Effekten innovativer Ausreißer zu bereinigen, muss die MA-Darstellung der Zeitreihe bekannt sein, bzw. aus der Schätzung der AR-Darstellung generiert werden (vgl. Abschnitt 3.5.1.2). Da dies für multivariate Zeitreihen sehr aufwändig ist, wurde in dieser Arbeit darauf verzichtet und ersatzweise der Versuch unternommen, die Zeitreihen derart zu bereinigen, dass innovative Ausreißer wie transiente Ausreißer behandelt wurden. Dieser Ansatz stellte sich als nicht praktikabel heraus, da er in lediglich einem von sieben Fällen durchführbar war. Der Einfluss statistischer Ausreißer auf die Schätzung der natürlichen Variabilität konnte somit nicht untersucht werden. Die Kontaminationswahrscheinlichkeit wird auf $\hat{p}_p = 0.175$ geschätzt.

Abbildungsverzeichnis

2.1 Allgemeine Strukturformel für PCDD/F 10

2.2 Allgemeine Strukturformel für PCB 13

2.3 Strukturformel der 12 dl PCB 14

3.1 Scatterplot eindimensionaler ilr transformierter Daten 57

3.2 Boxplot eindimensionaler ilr transformierter Daten 58

3.3 Boxplot eindimensionaler ilr transformierter Daten mit 3-fachem Interquartilabstand 58

3.4 Boxplots und Scatterplot zweidimensionaler Messdaten 60

3.5 Scatterplot dreidimensionaler Messdaten 61

3.6 Boxplot für jedes der drei Merkmale 61

3.7 Boxplots zu Aalen aus dem Jahr 2002 63

3.8 Scatterplot zu Aalen aus dem Jahr 2002 63

3.9 Scatterplot der skalierten Merkmale des Datensatzes D_2 64

3.10 Bivariate Zeitreihe mit Trend in der zweiten Komponente 131

3.11 Bivariate Zeitreihe ohne Trend nach Bildung der ersten Differenzen 135

3.12 Bivariate Zeitreihe mit Trendänderung 137

3.13 Bivariate Zeitreihe nach Differenzbildung: Trendänderung wird zu level shift 137
3.14 Zeitreihe mit level shift 149
3.15 Von level shift bereinigte, univariate Zeitreihe 156

4.1 Scatterplot der ilr transformierten PCB118 Daten zu Aalen im Jahr 2004 188
4.2 Boxplot der ilr transformierten PCB118 Daten zu Aalen im Jahr 2004 188
4.3 Histogramm der ilr transformierten PCB118 Daten zu Aalen im Jahr 2004 189
4.4 Die ilr Transformation für $D = 2$ 191
4.5 Die erste Ableitung der ilr Transformation für $D = 2$ 192
4.6 Scatterplotmatrix der ilr transformierten Daten für dl PCB Brassen 2003 195
4.7 Balkendiagramme der Mittelwerte und Mediane der ilr transformierten Daten für dl PCB Brassen 2003 198
4.8 Balkendiagramme der sechs Ausreißer der ilr transformierten Daten für dl PCB Brassen 2003 199
4.9 Scatterplotmatrix der ilr transformierten Daten für PCDD Grünkohl 2003 203
4.10 Balkendiagramme der Mittelwerte und Mediane der ilr transformierten Daten für PCDD Grünkohl 2003 im Vergleich zu den Mittelwerten und Medianen der ausreißerbereinigten Daten nach Mahalanobis 206

4.11 Balkendiagramme der vier Ausreißer der ilr transformierten Daten für PCDD Grünkohl 2003 nach Mahalanobis 207
4.12 Balkendiagramme der Mittelwerte und Mediane der ilr transformierten Daten für PCDD Grünkohl 2003 im Vergleich zu den Mittelwerten und Medianen der ausreißerbereinigten Daten nach Pena und Prieto 209
4.13 Balkendiagramm des Ausreißers der ilr transformierten Daten für PCDD Grünkohl 2003 nach Pena und Prieto 210
4.14 Mittelwerte der ilr transformierten Daten nach Variante 1B 214
4.15 Mittelwertbereinigte erste Differenzen der Mittelwerte der ilr transformierten Daten nach Variante 1B 215
4.16 Von Ausreißern bereinigte Zeitreihe der ersten Differenzen der Mittelwerte der ilr transformierten Daten nach Variante 1B 217
4.17 Zeitreihe der Mittelwerte der ilr transformierten Daten nach Variante 1A 219
4.18 Von Ausreißern bereinigte Zeitreihe der Mittelwerte der ilr transformierten Daten nach Variante 1A 221
4.19 Profil der Daten von 1994, die keine fehlenden Werte aufweisen . 223
4.20 Profil der Daten von 1994, die fehlende Werte aufweisen 224
4.21 Profil der ilr transformierten Daten von 1994, die keine fehlenden Werte aufweisen 225
4.22 Profil der ilr transformierten Daten von 1994, deren fehlende Werte geschätzt wurden 226
4.23 Profil der ilr transformierten Daten von 1995, die keine fehlenden Werte aufweisen, Teil 1 227

4.24 Profil der ilr transformierten Daten von 1995, die keine fehlenden Werte aufweisen, Teil 2 228
4.25 Profil des ilr transformierten Datenpunktes von 1995, dessen fehlender Wert geschätzt wurde 229
4.26 Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 1 234
4.27 Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 2 235
4.28 Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 3 236
4.29 Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 4 237
4.30 Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 5 238
4.31 Profile der Mittelwerte der ausreißerbereinigten ilr transformierten Jahresdaten nach Variante 1A, Teil 6 239

D.1 Kongenerenprofil zu dl PCB in Aalen 2003, n=14 348

E.1 Kongenerenprofil zu dl PCB in Brassen 2003, n=14 349

F.1 Kongenerenprofil zu dl PCB in Grünkohl 2003, n=14 350

G.1 Kongenerenprofil zu dl PCB in Weidelgras 2003, n=24 351

H.1 Kongenerenprofil zu PCB Leitkongeneren in Eiern von Wanderfalken 1990, n=10 353

I.1 Kongenerenprofil zu Leitkongeneren in Aalen 2002, n=105 355
I.2 Kongenerenprofil zu PCB Leitkongeneren in Aalen 2003, n=144 . 355
I.3 Kongenerenprofil zu PCB Leitkongeneren in Aalen 2004, n=123 . 355

J.1 Kongenerenprofil zu PCB Leitkongeneren in Brassen 1995, n=13 . 357
J.2 Kongenerenprofil zu PCB Leitkongeneren in Brassen 1998, n=11 . 357
J.3 Kongenerenprofil zu PCB Leitkongeneren in Brassen 2000, n=12 . 357
J.4 Kongenerenprofil zu PCB Leitkongeneren in Brassen 2001, n=12 . 357
J.5 Kongenerenprofil zu PCB Leitkongeneren in Brassen 2002, n=13 . 358
J.6 Kongenerenprofil zu PCB Leitkongeneren in Brassen 2003, n=12 . 358

K.1 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1995, n=13 360
K.2 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1996, n=13 360
K.3 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1997, n=13 360
K.4 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1998, n=13 360
K.5 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1999, n=13 361
K.6 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2000, n=13 361
K.7 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2001, n=13 361

K.8 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2002, n=13 361
K.9 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2003, n=13 362
K.10 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2004, n=13 362
K.11 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2005, n=13 362
K.12 Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2006, n=13 362

L.1 Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2000, n=18 364
L.2 Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2001, n=18 364
L.3 Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2002, n=16 364
L.4 Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2003, n=16 364

M.1 Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2000, n=27 366
M.2 Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2001, n=25 366
M.3 Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2002, n=22 366
M.4 Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2003, n=24 366

N.1 Kongenerenprofil zu PCDD in Aalen 2003, n=18 367

O.1 Kongenerenprofil zu PCDD in Brassen 1995, n=13 369
O.2 Kongenerenprofil zu PCDD in Brassen 1998, n=11 369
O.3 Kongenerenprofil zu PCDD in Brassen 2000, n=12 369
O.4 Kongenerenprofil zu PCDD in Brassen 2001, n=12 369
O.5 Kongenerenprofil zu PCDD in Brassen 2002, n=15 370
O.6 Kongenerenprofil zu PCDD in Brassen 2003, n=20 370

P.1 Kongenerenprofil zu PCDD in Miemsuscheln 1994, n=13 371

Q.1 Kongenerenprofil zu PCDD in Fichtennadeln 1995, n=20 373
Q.2 Kongenerenprofil zu PCDD in Fichtennadeln 1996, n=20 373
Q.3 Kongenerenprofil zu PCDD in Fichtennadeln 1997, n=21 373
Q.4 Kongenerenprofil zu PCDD in Fichtennadeln 1998, n=20 373
Q.5 Kongenerenprofil zu PCDD in Fichtennadeln 1999, n=29 374
Q.6 Kongenerenprofil zu PCDD in Fichtennadeln 2000, n=27 374
Q.7 Kongenerenprofil zu PCDD in Fichtennadeln 2001, n=26 374
Q.8 Kongenerenprofil zu PCDD in Fichtennadeln 2002, n=21 374

R.1 Kongenerenprofil zu PCDD in Grünkohl (Frischgewicht) 1995, n=16 376
R.2 Kongenerenprofil zu PCDD in Grünkohl (Frischgewicht) 1996, n=21 376
R.3 Kongenerenprofil zu PCDD in Grünkohl (Frischgewicht) 1997, n=14 376
R.4 Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2001, n=18 377

R.5 Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2002, n=12 377
R.6 Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2003, n=15 377

S.1 Kongenerenprofil zu PCDD in Weidelgras 2002, n=22 378
S.2 Kongenerenprofil zu PCDD in Weidelgras 2003, n=24 378

T.1 Kongenerenprofil zu PCDF in Brassen 2002, n=15 379
T.2 Kongenerenprofil zu PCDF in Brassen 2003, n=20 379

U.1 Kongenerenprofil zu PCDF in Fichtennadeln 1995, n=20 381
U.2 Kongenerenprofil zu PCDF in Fichtennadeln 1996, n=20 381
U.3 Kongenerenprofil zu PCDF in Fichtennadeln 1997, n=21 381
U.4 Kongenerenprofil zu PCDF in Fichtennadeln 1998, n=20 381
U.5 Kongenerenprofil zu PCDF in Fichtennadeln 1999, n=29 382
U.6 Kongenerenprofil zu PCDF in Fichtennadeln 2000, n=27 382
U.7 Kongenerenprofil zu PCDF in Fichtennadeln 2001, n=26 382
U.8 Kongenerenprofil zu PCDF in Fichtennadeln 2002, n=21 382

V.1 Kongenerenprofil zu PCDF in Grünkohl (Frischgewicht) 1995, n=16 384
V.2 Kongenerenprofil zu PCDF in Grünkohl (Frischgewicht) 1996, n=21 384
V.3 Kongenerenprofil zu PCDF in Grünkohl (Frischgewicht) 1997, n=14 384
V.4 Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2000, n=18 385
V.5 Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2001, n=18 385

V.6 Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2003, n=15 385

W.1 Kongenerenprofil zu PCDF in Weidelgras 2000, n=26 387

W.2 Kongenerenprofil zu PCDF in Weidelgras 2001, n=23 387

W.3 Kongenerenprofil zu PCDF in Weidelgras 2002, n=21 387

W.4 Kongenerenprofil zu PCDF in Weidelgras 2003, n=24 387

Tabellenverzeichnis

3.1 Beispiel zur Simes-Hochberg Prozedur 19

3.2 Kritische Werte des zweiseitigen Hampel Tests 68

3.3 Kritische Werte für die Teststatistik $T_n(\hat{d})$ [vgl. GLP06] 121

3.4 Kritische Werte für die Teststatistik $\hat{\eta}_\tau$ [vgl. KPSS92, S.166] . . . 125

3.5 Kritische Werte für die Teststatistik $Z(\rho)$ [vgl. SP92, S.264] . . . 128

3.6 Bivariate Zeitreihe mit Trend in der zweiten Komponente 130

3.7 Bivariate Zeitreihe ohne Trend in der zweiten Komponente nach Bildung erster Differenzen 135

3.8 Bivariate Zeitreihe mit Trendänderung in der zweiten Komponente 136

3.9 In R Simulierte Zeitreihe mit level shift 148

3.10 Kritische Werte zur Bestimmung von Ausreißern nach Galeano et al. 180

4.1 Ausreißer in dl PCB Brassen, nach Probenahmestandorten 197

4.2 Beispiel zur Vergrößerung der Schätzung der natürlichen Variabilität durch den Ausschluss von Ausreißern 204

4.3 Ausreißer in PCDD Grünkohl, nach Probenahmestandorten 205

4.4 Ausreißer in Daten zu PCB Leitkongeneren in Brassenfilets 212

4.5 Anzahl der Proben zu Brassenfilets je Stadt und Jahr 231

4.6 Anzahl der Proben zu Brassenfilets je Gebietstyp und Jahr 232

5.1 Kenngrößen der Änderungen der Schätzungen der natürlichen Variabilität für univariate Querschnittsdaten nach Ausschluss statistischer Ausreißer in % 242

5.2 Kenngrößen der Änderungen der Schätzungen der natürlichen Variabilität in multivariaten Querschnittsdaten nach Ausschluss statistischer Ausreißer in % 243

5.3 Zahl der mit linearen Trends behafteten multivariaten Zeitreihen . 255

B.17 Schätzungen der natürlichen Variabilität der gesamten univariaten Querschnittsdaten im Vergleich zur Schätzung der natürlichen Variabilität nach Ausschluß der Ausreißer, die anhand des Hampel-Tests detektiert wurden; Indikatorspalte gibt an, wie sich die Schätzung der natürlichen Variabilität nach Ausschluss statistischer Ausreißer ändert: 0=Verringerung, 1=Vergrößerung, 9999=keine Änderung, da keine Ausreißer detektiert wurden 310

B.18 Schätzungen der natürlichen Variabilität der Gesamtdaten im Vergleich zur Schätzung der natürlichen Variabilität nach Ausschluß der Ausreißer, die anhand der robusten Mahalanobis-Distanz detektiert wurden; Indikatorspalte gibt an, wie sich die Schätzung der natürlichen Variabilität nach Ausschluss statistischer Ausreißer ändert: 0=Verringerung, 1=Vergrößerung, 9999=keine Änderung, da keine Ausreißer detektiert wurden 314

B.19 Schätzungen der natürlichen Variabilität der Gesamtdaten im Vergleich zur Schätzung der natürlichen Variabilität nach Ausschluß der Ausreißer, die anhand des Verfahrens von Pena und Prieto detektiert wurden; Indikatorspalte gibt an, wie sich die Schätzung der natürlichen Variabilität nach Ausschluss statistischer Ausreißer ändert: 0=Verringerung, 1=Vergrößerung, 9999=keine Änderung, da keine Ausreißer detektiert wurden 319

B.1 Zusammenfassung der Ergebenisse für dl PCB 118 Querschnittsdaten mit Ausreißerbestimmung nach Hampel; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6); Biota Aal (Anguilla anguilla), Barbe (Barbus barbus), Hecht (Esox lucius), Zander (Sander lucioperca), Brassen (Abramis brama) 320

B.2 Zusammenfassung der Ergebnisse der Analysen für univariate Querschnittsdaten auf statistische Ausreißer nach Hampel 321

B.3 Zusammenfassung der Ergebenisse für dl PCB Querschnittsdaten mit Ausreißerbestimmung nach dem Verfahren von Pena und Prieto; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6); Biota Fichte-Trieb (Picea abies), Grünkohl (Brassica oleracea var. sabellica), Weidelgras (Lolium perenne), Wanderfalke (Falco peregrinus) . . 322

B.4 Zusammenfassung der Ergebenisse für dl PCB Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 323

B.5 Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 324

B.6 Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; Proben mit fehlenden Daten wurden ausgeschlossen; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 325

B.7 Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 326

B.8 Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; Proben mit fehlenden Daten wurden ausgeschlossen; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) . . . 327

B.9 Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) . . 328

B.10 Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; Proben mit fehlenden Daten wurden ausgeschlossen; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 329

B.11 Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6); Biota Miesmuschel: lat. Mytilus 330

B.12 Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; Proben mit fehlenden Daten wurden ausgeschlossen; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 331

B.13 Zusammenfassung der Ergebenisse für PCDF Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; fehlende Werte traten nicht auf; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 332

B.14 Zusammenfassung der Ergebenisse für PCDF Querschnittsdaten mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; fehlende Werte traten nicht auf; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6) 333

B.15 Zusammenfassung der Ergebnisse der Analysen von multivariaten Querschnittsdaten auf statistische Ausreißer anhand der robusten Mahalanobis-Distanz; fehlende Werte wurden geschätzt 334

B.16 Zusammenfassung der Ergebnisse der Analysen von multivariaten Querschnittsdaten auf statistische Ausreißer anhand des Verfahrens von Pena und Prieto; fehlende Werte wurden geschätzt . . 335

C.1 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von dl PCB; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte traten keine fehlenden Werte auf; n.b.=nicht bestimmbar . . 337

C.2 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von dl PCB; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte traten keine fehlenden Werte auf; n.b.=nicht bestimmbar . . 338

C.3 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, diese wurden geschätzt 339

C.4 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, die entsprechenden Proben wurden entfernt 340

C.5 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, diese wurden geschätzt 341

C.6 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, die entsprechenden Proben wurden entfernt 342

C.7 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDD; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf 343

C.8 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDD; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf 344

C.9 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDF; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf 345

C.10 Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDF; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf; n.b.=nicht bestimmbar 346

C.11 Zusammenfassung der Ergebnisse, wieviele Ausreißer des Typs I nach welcher Auswertevariante gefunden wurden 347

Literaturverzeichnis

[AE05] AITCHISON, J. ; EGOZCUE, J. J.: Compositional Data Analysis: Where Are We and Where Should We Be Heading? In: *Mathematical Geology* 37 (2005), S. 829–850

[AGRH07] ANTUNES, Paulo ; GIL, Odete ; REIS-HENRIQUES, Maria A.: Influence of fish size in bioaccumulation of PCBs in seabass from semi-intensive aquaculture and natural environment. In: *Science of the Total Environment* 377 (2007), S. 36–44

[Ait86] AITCHISON, J.: *The statistical analysis of compositional data*. Chapman and Hall, 1986

[Ait03a] AITCHISON, J.: *The statistical Analysis of Compositional Data*. Neudruck der 1986 erschienen Ausgabe mit Zusatzmaterial. Blackburn Press, 2003

[Ait03b] AITCHISON, John ; DEPARTMENT OF STATISTICS UNIVERSITY OF GLASGOW (Hrsg.): *A Concise Guide to Compositional Data Analysis*. Department of Statistics University of Glasgow, 2003

[Bai97] BAI, Jushan: Estimating Multiple Breaks One at a Time. In: *Econometric Theory* 13 (1997), S. 315–352

[Bal93] BALKE, Nathan S.: Detecting Level Shifts in Time Series. In: *Journal of Business and Economic Statistics* 11 (1993), S. 81–92

[Bar] BARIGOZZI, Matteo: *ARIMA estimation: theory and applications*. – Matteo Barigozzi is a post-doctoral researcher at the European Center for Advanced Research in Economics and Statistics (ECARES) - Université Libre de Bruxelles

[Bas94] BASLER, Herbert: *Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistischen Methodenlehre*. Physica-Verlag, 1994

[BBH99] BIANCHI, Marco ; BOYLE, Martin ; HOLLINGSWORTH, Déirdre: A comparison of methods for trend estimation. In: *Applied Economic Letters* 6 (1999), S. 103–109

[BDR] BHAUMIK, Amitabha ; DEY, K. ; RAVISHANKER, Nalini: *A Dynamic Linear Model Approach for Compositional Time Series Analysis*

[BF94] BALKE, Nathan S. ; FOMBY, Thomas B.: Large Shocks, Small Shocks, and Economic Fluctuations: Outliers in Macroeconomic Time Series. In: *Journal of Applied Economics* 9 (1994), S. 181–200

[BF04] BERNHOLT, Thorsten ; FISCHER, Paul: The Complexity of Computing the MCD-Estimator / Universität Dortmund, Lehrstuhl In-

formatik 2 and Danish Technical University Kongens Lyngby, IMM. 2004. – Forschungsbericht

[BG99] BAUER, Marcus ; GATHER, Ursula: *The Identification of Multiple Outliers in Online Monitoring Data*. 1999

[BG00] BECKER, Claudia ; GATHER, Ursula: The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. In: *Computional Statistics & Data Analysis* 36 (2000), S. 119–127

[Bha67] BHATTACHARYYA, G. K.: *Robust Estimates of Linear Trend in Multivariate Time Series*. 1967

[BJK94] BUSSE, Ralf ; JESKE, Roland ; KRÄMER, Walter: Efficiency of least-squares-estimation of polynomial trend when residuals are autocorrelated. In: *Econometric Letters* 45 (1994), S. 267–271

[BL94] BARNETT, V. ; LEWIS, T.: *Outliers in Statistical Data*. Chichester : John Wiley & Sons, 1994

[BMP77] BREIMAN, Leo ; MEISEL, William ; PURCELL, Edward: Variable Kernel Estimates of Multivariate Densities. In: *Technometrics* 19 (1977), Nr. 2, S. 135–144

[BMW99] BRANDT, Patrick T. ; MONROE, Burt L. ; WILLIAMS, John T.: *Time Series Models for Compositional Data*. Workshop in Political Theory and Policy Analysis, Juli 1999

[Bos08] BOSCH, Siegfried: *Lineare Algebra*. Springer Verlag, 2008

[BPG96] BARCELÓ, Caries ; PAWLOWSKY, Vera ; GRUNSKY, Eric: Some Aspects of Transformations of Compositional Data and the Identification of Outliers. In: *Mathematical Geology* 28 (1996), Nr. 4, S. 501–518

[BPG05] BUCCIANTI, A. ; PAWLOWSKY-GLAHN, V.: New perspectives on water chemistry and compositional data analysis. In: *Math Geol* 37 (2005), S. 703–727

[BS92] BACON-SHONE, J.: Ranking Methods for Compositional Data. In: *Applied Statistics* 41 (1992), S. 533–537

[BSA09] ; EPA (Veranst.): *EPA - BSAF data set*. `http://www.epa.gov/med/Prods_Pubs/bsaf.htm`. Version: 2009

[BTDB06] BOOGAART, K. G. V. ; TOLOSANA-DELGADO, R. ; BREN, M.: Concept for handling with zeros and missing values in compositional data. In: *Proceedings of IAMG'06 - The XI annual conference of the International Association for Mathematical Geology* University of Liege, Belgium, 2006. – CD-ROM

[BZ80] BALLSCHMITER, K. ; ZELL, M.: Analysis of Polychlorinated Biphenyls (PCB) by Glass Capillary Gas Chromatography. In: *Fresenius Zeitschrift für Analytische Chemie* 302 (1980), S. 20–31

[Cad88] CADE, Tom J.: *Peregrine falcon populations*. Peregrine Fund, 1988

[Cam80] CAMPBELL, N. A.: Robust Procedures in Multivariate Analysis

I: Robust Covariance Estimation. In: *Applied Statistics* 29 (1980), Nr. 3, S. 231–237

[Chu01] CHUMACERO, Rómulo A.: Estimating ARMA Models Efficiently. In: *Studies in Nonlinear Dynamics and Econometrics* 5 (2001), S. 103–114

[Dav01] DAVIDSON, James: *Econometric Theory*. Blackwell Publishing, 2001

[DF79] DICKEY, D. A. ; FULLER, W. A.: Distributions of the estimators for autoregressive time series with a unit root. In: *Journal of the American Statistical Association* 74 (1979), S. 221–254

[DG89] DAVIES, L. ; GATHER, U.: *The identification of multiple outliers*. Technical Report 89/1, 1989

[DGM99] DOLADO, Juan J. ; GONZALO, Jesús ; MARMOL, Francesc: *Cointegration*. 1999

[DK11] DÜLL, R. ; KUTZELNIGG, H.: *Taschenlexikon der Pflanzen Deutschlands*. Quelle & Meyer-Verlag, 2011

[DKS+97] DETTWILER, Johannes ; KARLAGANIS, Georg ; STUDER, Christoph ; JOSS, Stefan ; STETTLER, Anton ; CHAMBAZ, Daniel: Umweltgefährdende Stoffe - Dioxine und Furane / Bundesamt für Umwelt, Wald und Landschaft BUWAL. 1997. – Forschungsbericht

[Don82] DONOHO, David L.: *Breakdown properties of multivariate location estimators*. Qualifying paper, 1982

[DR91] DIEBOLD, F. X. ; RUDEBUSCH, G. D.: On the Power of Dickey-Fuller tests against fractional alternatives. In: *Econometric Letters* 35 (1991), S. 155–160

[DS09] DIETRICH, E. ; SCHULZE, A.: *Statistische Verfahren zur Maschinen- und Prozessqualifikation*. München : Carl Hanser Verlag, 2009

[DW91] DEJONG, D. N. ; WHITEMAN, C. H.: Reconsidering 'Trends and random walks in macroeconomic time series'. In: *Journal of Monetary Economics* 28 (1991), S. 221–254

[EG87] ENGLE, R. F. ; GRANGER, C. W. J.: Co-integration and error correction: Representation, estimation and testing. In: *Econometrica* 55 (1987), S. 251–276

[EHM+06] EDGAR, P. J. ; HURSTHOUSE, A. S. ; MATTHEWS, J. E. ; DAVIES, I.M. ; HILLIER, S.: Sediment influence on congener-specific PCB bioaccumulation by Mytilus edulis: a case study from an intertidal hot spot, Clyde Estuary, UK. In: *Jornal of Environmental Monitoring* 8 (2006), S. 887–896

[ES95] EL-SHAARAWI, A. H.: Trend detection and estimation with environmental applications. In: *Mathematics and Computers in Simulation* 39 (1995), S. 441–447

[ET93] EFRON, Bradley ; TIBSHIRANI, Robert J.: *An Introduction to the Bootstrap*. Chapmall & Hall CRC, 1993

[FC05] FRY, Tim R. L. ; CHONG, Derek: A tale of two logits, compositional data analysis and zero observations. In: MATEU-FIGUERAS, G. (Hrsg.) ; BARCELÓ-VIDAL, C. (Hrsg.) ; University of Girona, Spain (Veranst.): *Proceedings of CODAWORK'05, The Second Compositional Data Analysis Workshop* University of Girona, Spain, 2005

[FFM00] FRY, J.M. ; FRY, T.R.L. ; MCLAREN, K.R.: Compositional Data Analysis and Zeros in Micro Data. In: *Applied Economics* 32 (2000), S. 953–959

[FGR05] FILZMOSER, P. ; GARRETT, R. G. ; REIMANN, C.: Multivariate outlier detection in exploration geochemistry. In: *Computional Geosciences* 31 (2005), S. 579–587

[FH05] FRERICHS, Heike ; HANN, Carmen: Coplanare polychlorierte Biphenyle (PCB) in den Oberböden Hamburgs. In: *Umweltchem Ökotox* 17 (2005), S. 201–207

[FH08] FILZMOSER, Peter ; HRON, Karel: Outlier Detection for Compositional Data Using Robust Methods. In: *Mathematical Geosciences* 40 (2008), S. 233–248

[FH09] FILZMOSER, Peter ; HRON, Karel: Correlation analysis for compositional data. In: *Mathematical Geosciences* 41 (2009), S. 905–919

[FHR09] FILZMOSER, Peter ; HRON, Karel ; REIMANN, Clemens: Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. In: *Science of the Total Environment* 407 (2009), S. 6100–6108

[Fis09] FISCHER, Gerd: *Lineare Algebra*. Vieweg+Teubner, 2009

[Fox02] FOX, John: *Time-Series Regression and Generalized Least Squares Appendix to An R and S-PLUS Companion to Applied Regression*, January 2002

[Ful96] FULLER, Wayne A.: *Introduction to Statistical Time Series*. Wiley and Sons, 1996

[FV10] FILZMOSER, P. ; VARMUZA, K.: *chemometrics: Multivariate Statistical Analysis in Chemometrics*. Vienna, Austria, 2010. `http://CRAN.R-project.org/package=chemometrics`

[FW94] FANG, K.T. ; WANG, Y.: *Number Theoretic Methods in Statistics*. London : Chapman and Hall, 1994

[GB10] GROH, Stefanie ; BORHO, Werner: *dl-PCB in den Böden von Baden-Württemberg*. `www.lubw.baden-wuerttemberg.de`. Version: 2010

[GE83] GRANGER, C. W. J. ; ENGLE, R. F.: *Dynamic model specification with equilibrium constraints*. 1983

[GK72] GNANADESIKAN, R. ; KETTENRING, J. R.: Robust estimates,

residuals, and outlier detection with multiresponse data. In: *Biometrics* 28 (1972), S. 81–124

[GK88] GATHER, U. ; KALE, B. K.: Maximum Likelihood Estimation in the Presence of Outliers. In: *Commun. Statist. -Theory Meth.* 17 (1988), S. 3767–3784

[GKP04] GATHER, Ursula ; KUHNT, Sonja ; PAWLITSCHKO, Jörg: *Concepts of Outlyingness for Various Data Structures*. 2004

[GLGL05] GERNANDT, David S. ; LOPEZ, Gretel G. ; GARCIA, Sol O. ; LISTON, Aaron: Phylogeny and classification of Pinus. In: *TAXON* 54 (2005), S. 29–42

[GLP06] GIRAITIS, Liudas ; LEIPUS, Remigijus ; PHILIPPE, Anne: The test for stationarity versus trends and unit roots for a wide class of dependent errors. In: *Econometric Theory* 22:6 (2006), S. 989–1029

[Gos92] GOSLING, Elizabeth: *The Mussel Mytilus: ecology, physiology, genetics, and culture*. Elsevier, 1992

[GPT06] GALEANO, Pedro ; PEÑA, Daniel ; TSAY, Ruey S.: Outlier detection in multivariate time series via projection pursuit. In: *Journal of the American Statistical Association* 101 (2006), S. 654–669

[GR06] GIURCANEANU, Ciprian D. ; RISSANEN, Jorma: Estimation of AR and ARMA models by stochastic complexity. In: *IMS Lec-*

ture Notes-Monograph Series, Time Series and Related Topics 52 (2006), S. 48–59

[Gra81] GRANGER, C. J.: Some properties of time series data and their use in econometric model specification. In: *Journal of Econometrics* 16 (1981), S. 121–130

[GRI11] ; Germplasm Resources Information Network (Veranst.): *GRIN Taxonomy for Plants Online Abfrage*. `www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?319629`. Version: 2011

[GT96] GALLANT, R. ; TAUCHEN, G.: Which moments to match? In: *Econometric Theory* 12 (1996), S. 657–681

[GZVS10] GÓMEZ-ZALVÍDAR, Manuel ; VENTOSA-SANTAULARIA, Daniel: Testing for a Deterministic Trend when there is Evidence of Unit-Root / Universidad de Guanajuato, Department of Economics and Statistics. Version: 2010. `http://econpapers.repec.org/RePEc:gua:wpaper:em200801`. 2010 (EM200801). – Department of Economics and Finance Working Papers

[Hal87] HALL, Peter: ESTIMATION OF INTEGRATED SQUARED DENSITY DERIVATIVES. In: *Statistics and Probability Letters* 6 (1987), S. 109–115

[Hal90a] HALL, Peter: ON LAW OF THE LOGARITHM FOR DENSITY ESTIMATORS. In: *Statistics and Probability Letters* 9 (1990), S. 237–240

[Hal90b] HALL, Peter: On the Relative Performance of Bootstrap and Edgeworth Approximations of a Distribution Function. In: *Journal of Multivariate Analysis* 35 (1990), S. 108–129

[Hal90c] HALL, Peter: Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. In: *Journal of Multivariate Analysis* 32 (1990), S. 177–203

[Hal92] HALL, P.: *The Bootstrap and Edgeworth Expansion*. New York : Springer Verlag, 1992

[Haw80] HAWKINS, D. M.: *Identification of Outliers*. Chapman and Hall, 1980

[HE89] HARTUNG, J. ; ELPELT, B.: *Multivariate Statistik*. 3. Aufl. München, Wien : Oldenbourg, 1989

[HEK05] HARTUNG, Joachim ; ELPELT, Bärbel ; KLÖSENER, Karl-Heinz: *Statistik*. Oldenbourg, 2005

[HH04] HEIBERGER, Richard M. ; HOLLAND, Burt: *Statistical analysis and data display: an intermediate course with examples in S-PLUS, R, and SAS*. Springer Verlag, 2004

[HK10] HRON, Karel ; KUBÁCEK, Lubomír: *Statistical properties of the total variation estimator for compositional data*. 2010. – Veröffentlichung in Metrika

[HR05] HARDIN, J. ; ROCKE, D. M.: The distribution of robust distances.

In: *Journal of Computional Graphics and Statistics* 14 (2005), S. 928–946

[HR09] HUBER, Peter J. ; RONCHETTI, Elvezio M.: *Robust Statistics*. John Wiley & Sons, 2009

[HRRS05] HAMPEL, Frank R. ; ROCNCHETTI, Elvezio M. ; ROUSSEEUW, Peter J. ; STAHEL, Werner A.: *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2005

[HRS10] HELD, Leonard ; RUFIBACH, Kaspar ; SEIFERT, Burkhardt ; UNIVERSITÄT ZÜRICH (Hrsg.): *Einführung in die Biostatistik*. Universität Zürich, 2010

[HSS08] HARTLAP, J. ; SIMON, P. ; SCHNEIDER, P.: Why your model parameter confidences might be too optimistic - unbiased estimation of the inverse covariance matrix. In: *Astronomy & Astrophysics* 6170 (2008), S. 1 – 6

[HSW03] HARTMANN, Edmund ; SCHNEIDER, Jörg ; WENZEL, Andrea: Schadstoffe in klärschlammgedüngten Ackerböden Baden-Württembergs / Landesanstalt für Umweltschutz Baden-Württemberg. Version: 2003. `http://www.lfu.baden-wuerttemberg.de`. Karlsruhe, 2003. – Forschungsbericht

[HT99] HALL, Peter ; TURLACH, Berwin A.: Reducing bias in curve estimation by use of weights. In: *Computional Statistics and Data Analysis* 30 (1999), S. 67–86

[HTF10a] HRON, Karel ; TEMPL, Matthias ; FILZMOSER, Peter: Imputation of missing values for compositional data using classical and robust methods. In: *Computional Statistics and Data Analysis* 54 (2010), Nr. 12, S. 3095–3107

[HTF10b] HRON, Karel ; TEMPL, Matthias ; FILZMOSER, Peter: *Robust Imputation of Missing Values in Compositional Data Using the R-Package robCompositions*, 2010. `http://en.scientificcommons.org/55854283`

[HW88] HALL, Peter ; WAND, Matthew P.: Minimizing L_1 Distance in Nonparametric Density Estimation. In: *Journal of Multivariate Analysis* 26 (1988), S. 59–88

[HW95] HALL, Peter ; WOLFF, Rodney C. L.: Estimators of integrals of powers of density derivatives. In: *Statistics and Probability Letters* 24 (1995), S. 105–110

[HW96] HALL, Peter ; WOOD, Andrew T. A.: Approximations to distributions of statistics used for testing hypothesis about the number of modes of a population. In: *Journal of Statistical Planning and Inference* 55 (1996), S. 299–317

[JM06] JOAS, Anke ; MÜLLER, Elisabeth: Abschlussbericht: Qualitätssicherung und Erweiterung des Datenbestandes der Dioxin-Datenbank des Bundes und der Länder zu Dioxinen, dioxinähnlichen nicht dioxinähnlichen PCB einschließlich der Auswertung

und Bewertung der Daten / BiPRO GmbH München im Auftrag des Umweltbundesamtes. 2006. – Forschungsbericht

[Joh88] JOHANSEN, S.: Statistical Analysis of Cointegration Vectors. In: *Journal of Economic Dynamics and Control* 12 (1988), S. 231–254

[Joh91] JOHANSEN, S.: Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. In: *Econometrica* 59 (1991), S. 1551–1580

[KEKZ09] KOZLOV, Mikhail ; ELENAZVEREVA ; KOZLOV, Michail V. ; ZVEREV, Vitali: *Impacts of Point Polluters on Terrestrial Biota: Comparative Analysis of 18 Contaminated Areas*. Springer Verlag, 2009

[Kün87] KÜNSCH, H.: Statistical aspects of self-similar processes. In: PROHOROV, Yu. A. (Hrsg.) ; SAZONOV, V. V. (Hrsg.): *Proceedings of the 1st World Congress of the Bernoulli Society* Bd. 1, 1987, S. 67–74

[KPSS92] KWIATKOWSKI, Denis ; PHILLIPS, Peter C. B. ; SCHMIDT, Peter ; SHIN, Yongcheol: Testing the null hypothesis of stationarity against the alternative of a unit-root How sure are we that economic time series have a unit root? In: *Journal of Econometrics* 54 (1992), S. 159–178

[Kre05] KRENGEL, Ulrich: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Vieweg, 2005

[Kru69] KRUSKAL, J. B. ; MILTON, R. C. (Hrsg.) ; NELDER, J. A. (Hrsg.): *Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transfomation which optimizes a new index of condensation*. New York : Academic Press, 1969. – 427–440 S.

[KT95] KRAUSS, Peter ; TAURO, Jeanette: *Mögliche Gefährdung des Grundwassers durch PCB sowie Dioxine und Furane im Boden*. Karlsruhe, 1995

[Kur06] KURUWITA, Chinthaka: *A Bayesian approach for bandwidth selection in kernel density estimationwith with censored data*. Clemson University, 2006

[103]

[LJ03] LIN, S.M. ; JOHNSON, K.F.: *Methods of microarray data analysis III: papers from CAMDA '02*. Norwell, Massachusetts : Kluwer Academic Publishers, 2003

[Lüt93] LÜTKEPOHL, H.: *Introduction to Multiple Time Series Analysis*. New York: Springer-Verlag, 1993

[Lüt05] LÜTKEPOHL, Helmut: *New introduction to multiple time series analysis*. Springer Verlag, 2005

[Mah36] *On the generalised distance in statistics*. Bd. *12, 49-55*. National Institute of Science of India, 1936

[Mau95] MAURICIO, José A.: Exact Maximum Likelihod Estimation of Stationary Vector ARMA Models. In: *Journal of the American Statistical Association* 90 (1995), S. 282–291

[MB03] MAINDONALD, John H. ; BRAUN, John: *Data analysis and graphics usind R: an example-based approach*. Cambridge Univ. Press, 2003

[Mea00] MEADE, Nidel: A note on the Robust Trend and ARARMA methodologies used in the M3 Competition. In: *International Journal of Forecasting* 16 (2000), S. 517–519

[MFBVPG03] MARTÍN-FERNÁNDEZ, J. A. ; BARCELÓ-VIDAL, C. ; PAWLOWSKY-GLAHN, V.: Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. In: *Mathematical Geology* 35 (2003), Nr. 3, S. 253–278

[MK98] MADDALA, G. S. ; KIM, I.-M.: *Unit Roots, Cointegration and Structural Changes*. Cambridge Univ. Press, 1998

[MM05] MOORE, David S. ; MCCABE, George P.: *Introduction to the practice of statistics, 5th edition*. W.H. Freeman and Co., 2005

[MMY06] MARONNA, R. ; MARTIN, R. D. ; YOHAI, V. J.: *Robust statistics: theory and methods*. Wiley, New York, 2006

[MVM10] MILLER, F. P. ; VANDOME, A. F. ; MCBREWSTER, J.: *Kernel Density Estimation*. Alphascript Publishing, 2010

[MZ01] MARONNA, R. ; ZAMAR, R.: Robust estimates of location and dispersion for high-dimensional data sets. In: *Technometrics* 44 (2001), S. 307–317

[MZ08] MCLEOD, A. I. ; ZHANG, Y.: Faster ARMA maximum likelihood estimation. In: *Computional Statistics and Data Analysis* 52 (2008), S. 2166–2176

[Nel06] NELSON, Joseph S.: *Fishes of the wold*. John Wiley & Sons, 2006

[Nie05] NIELSEN, Heino B. ; UNIVERSITY OF COPENHAGEN (Hrsg.): *Linear Regression with Time Series Data*. University of Copenhagen, September 2005. – scriptum Econometrics 2

[NP82] NELSON, Charles R. ; PLOSSER, Charles I.: Trends and Random Walks in Macroeconomic Time Series. In: *Journal of Monetary Economics* 10 (1982), S. 139–163

[NT88] NABEYA, S. ; TANAKA, K.: Asymptotic theory of a test for the constancy of regression coefficients against the random walk alternative. In: *Annals of Statistics* 16 (1988), S. 218–235

[PBL+09] PALM, Wolf-Ulrich ; BOLTEN, Mareike ; LUTZENBERGER, Alexa ; STAMER, Franziska ; RUCK, Wolfgang: Literaturstudie zur Ursachenanalyse der Kontamination von Grundfutter durch PCB unter besonderer Berücksichtigung dioxinähnlicher PCB / Leuphana Universität Lüneburg - Fakultät III, Institut für Ökologie und Umweltchemie (Bereich Umweltchemie). 2009. – Forschungsbericht

[Per89] PERRON, P.: The great cash, the oil price shock, and the unit root hypothesis. In: *Econometrica* 57 (1989), S. 1361–1401

[PFF00] PAN, J.-X. ; FUNG, W.-K. ; FANG, K.-T.: Multiple outlier detection in multivariate data using projection pursuit techniques. In: *Journal of Statistical Planning and Inference* 1 (2000), S. 153–167

[PGE02] PAWLOWSKY-GLAHN, Vera ; EGOZCUE, Juan J.: BLU Estimators and Compositional Data. In: *Mathematical Geology* 34 (2002), S. 259–274

[PGETD08] PAWLOWSKY-GLAHN, V. ; EGOZCUE, JJ ; TOLOSANO-DELGADO, R.: *Lecture notes on compositional data analysis*. `http://hdl.handle.net/10256/297`. Version: 2008

[Pol01] POLLOCK, D. S. G.: Methodology for trend estimation. In: *Economic Modelling* 18 (2001), S. 75–96

[PP88] PHILLIPS, P. C. B. ; PERRON, P.: Testing for a unit root in time series regression. In: *Biometrika* 75 (1988), S. 335–346

[PP01a] PENA, Daniel ; PRIETO, Francisco J.: Cluster Identification Using Projections. In: *Journal of the American Statistical Association* 96 (2001), S. 456

[PP01b] PENA, Daniel ; PRIETO, Francisco J.: Multivariate Outlier Detection and Robust Covariance Matrix Estimation. In: *Technometrics* 43 (2001), Nr. 3, S. 286–310

[Pri00] PRIETO, Francisco J.: *MATLAB code Outlier Detection via Pena and Prieto*. 2000

[PS89] PHILLIPS, P. C. B. ; SOLO, V.: *Asymptotics for linear processes*. 1989

[Ram99] RAMBOLD, A.: *Ausgewählte Verfahren zur Identifikation von Ausreißern und einflußreichen Beobachtungen in multivariaten Daten und Verfahren*. München : Herbert Utz Verlag, 1999

[REA06] *REACh Regulation (EC) 1907/2006*. : *REACh Regulation (EC) 1907/2006*, 2006

[Rüg02] RÜGER, Bernhard: *Test- und Schätztheorie, Band II: Statistische Tests*. Oldenbourg, 2002

[Ric00] RICHARDSON, David M.: *Ecology and biogeography*. Cambridge Univ. Press, 2000

[Rip11] RIPLEY, Brian: *Package 'MASS'*, 2011

[RL87] ROUSSEEUW, P.J. ; LEROY, A.: *Robust Regression and Outlier Detection*. Wiley, New York, 1987

[Rob95] ROBINSON, P. M.: Gaussian semiparametric estimation of long range dependence. In: *Annals of Statistics* 23 (1995), S. 387–505

[Roy53] ROY, S. N.: On a heuristic method of test construction and its use in a multivariate analysis. In: *Annals of Mathematical Statistics* 24 (1953), S. 220–238

[RZ90] ROUSSEEUW, P. J. ; ZOMERAN, B. C. V.: Unmasking multivariate outliers and leverage points. In: *Journal of the American Statistical Association* 85 (1990), S. 633–651

[Sa1] SAÏD, Saïd E.: Unit-roots test for time-series data with a linear time trend. In: *Journal of Econometrics* 47 (1991), S. 285–303

[Sac09] SACHS, L.: *Angewandte Statistik*. Berlin : Springer Verlag, 2009

[SB98] SMITH, T. M. F. ; BRUNSDON, Teresa M.: The Time Series Analysis of Compositional Data. In: *Journal of Official Statistics* 14 (1998), Nr. 3, S. 237–253

[Sch01] SCHLITTGEN, Rainer: *Angewandte Zeitreihenanalyse*. München, Wien : Oldenbourg, 2001

[Sil81] SILVERMAN, B. W.: Using kernel estimates to investigate multimodality. In: *Journal of the Royal Statistical Society: Series B* 43 (1981), S. 97–99

[SJK+05] STACHEL, Burkhard ; JANTZEN, Eckard ; KNOTH, Wilhelm ; KRÜGER, Frank ; LEPOM, Peter ; OETKEN, Matthias ; REINCKE, Heinrich ; SAWAL, Georg ; SCHWARTZ, René ; UHLIG8, Steffen: The Elbe Flood in August 2002 - Organic Contaminants in Sediment Samples Taken After the Flood Event. In: *Journal of Environmental Science and Health* A40 (2005), S. 265–287

[Sow92] SOWELL, Fallaw: Maximum likelihood estimation of stationary

univariate fractionally integrated time series models. In: *Journal of Econometrics* 53 (1992), S. 165–188

[SP92] SCHMIDT, Peter ; PHILLIPS, Peter C. B.: LM tests for a unit root in the resence of deterministic trends. In: *Oxford Bulletin of Econometrics and Statistics* 54 (1992), S. 257–287

[SS01] SCHLITTGEN, Rainer ; STREITBERG, Bernd H. J.: *Zeitreihenanalyse*. Oldenbourg, 2001

[SSM03] SCHLULZE, A. ; SALTHAMMER, T. ; MARUTZKY, R.: Vorkommen polychlorierter Biphenyle (PCB) in Altholz. In: *Holz als Roh- und Wekrstoff* 61 (2003), S. 23–28

[Sta81] STAHEL, W. A.: *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*, ETH Zürich, Diss., 1981

[Ste77] STERBA, Günther: *Süßwasserfische aus aller Welt*. Urania, 1977

[Str80] STREULI, H.: Fehlerhafte Interpretation und Anwendung von Ausreifler-Tests, insbesondere bei Ringversuchen zur Uberprüfung analytisch-chemischer Untersuchungsmethoden. In: *Fresenius Zeitschrift für Analytische Chemie* 303 (1980), S. 406–408

[Str04] STRECK, G.: *Einführung in die Statistik für Geoökologen und andere Naturwissenschaftler*. Bayreuth : Books on Demand GmbH, 2004

[SW03] STOCK, James H. ; WATSON, Mark W.: *Introduction to Econometrics*. Addison Wesley, 2003

[Swi70] SWITZER, P.: Numerical Classification. In: *Geostatistics* Plenum (1970), S. 31–47

[Tea10] TEAM, R Development C. ; R FOUNDATION FOR STATISTICAL COMPUTING (Hrsg.): *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2010. `http://www.R-project.org/.` – ISBN 3-900051-07-0

[TH96] TRACEY, G. A. ; HANSEN, D. J.: Use of Biota-Sediment Accumulation Factors to Assess Similarity of Nonionic Organic Chemical Exposure to Benthically-Coupled Organisms of Differing Trophic Mode. In: *Arch. Environ. Contain. Toxicol.* 30 (1996), S. 467–475

[Tha09] THAS, Olivier: *Comparing Distributions*. Springer Verlag, 2009

[THF10] TEMPL, Matthias ; HRON, Karel ; FILZMOSER, Peter ; TU WIEN (Hrsg.): *Package 'robCompositions'*. TU Wien, August 2010

[THMF05] THIÓ-HENESTROSA, S. ; MARTÍN-FERNÁNDEZ, J. A.: Dealing with Compositional Data: The Freeware CoDaPack. In: *Mathematical Geology* 37 (2005), Nr. 7, S. 773–793

[TPP98] TSAY, Ruey S. ; PENA, Daniel ; PANKRATZ, Alan E.: Outliers in Multivariate Time Series. In: *Statistics and Econometric Series Working Paper 98-96* 42 (1998), S. 1–20

[Tsa88] TSAY, Ruey S.: Outliers, Level Shifts, and Variance Changes in Time Series. In: *Journal of Forecasting,* 7 (1988), S. 1–20

[TW10] TRENKLER, Carsten ; WEBER, Enzo: *Codependence and Cointegration*. 2010. – University of Regensburg Working Papers in Business, Economics and Management Information Systems

[UBA99] *Ermittlung atmosphärischer Stoffeinträge in den Boden, Nutzung neuer Sammel- und Nachweisverfahren Verbundverfahren Ergebnisse 1998*. `http://www.uvm.baden-wuerttemberg.de/lfuandwww.umweltbundesamt.de`. Version: 1999

[UBA11] UBA: *Chemikalienpolitik und Schadstoffe, REACH*. `http://www.umweltbundesamt.de/chemikalien/dioxine.htm#1`. Version: 2011

[Umw08] UMWELTBUNDESAMT: *Fact Sheet Polychlorierte Dibenzodioxine und Furane (PCDD/F)*. `http://www.umweltbundesamt.at/fileadmin/site/umweltthemen/gesundheit/fact_sheets/Fact_Sheet_Dioxine.pdf`. Version: 2008

[Umw11] ; Umweltbundesamt, Dessau-Roßlau (Veranst.): *Daten zur Umwelt*. `http://www.umweltbundesamt-daten-zur-umwelt.de/umweltdaten/public/theme.do?nodeIdent=2379`. Version: 2011

[Ver04] VERBEEK, Marno: *A Guide to Modern Econometrics*. 2. Wiley and Sons, 2004

[VF09] VARMUZA, K. ; FILZMOSER, P.: *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, 2009

[Vis04] VISSER, Hans: Estimation and detection of flexible trends. In: *Atmospheric Environment* 38 (2004), S. 4135–4145

[VJ06] VEREIN JORDSAND, Deutscher Verein zum Schutz der V. Reichsbund für Vogelschutz V. Reichsbund für Vogelschutz: *Die Vogelwelt, Bände 2006-2007*. Duncker & Humblot, 2006

[VN05] VOLLAND, Gerhard ; NEUWIRTH, Adolf: Dioxinähnliche polychlorierte Biphenyle (PCB) und polychlorierte Dioxine/Furane (PCDD/F) im Innenraum - Modelluntersuchungen zum Einfluss PCB-belasteter Bauteile und Baustoffe auf die Innenraumluft in Abhängigkeit von baulichen Einrichtungen und klimatischen Randbedingungen / Materialprüfungsanstalt Universität Stuttgart; MPA - Otto-Graf-Institut. 2005. – Forschungsbericht

[Vog97] VOGELSANG, Timothy J.: Wald-Type Tests for Detecting Breaks in the Trend Function of a Dynamic Time Series. In: *Econometric Theory* 13 (1997), S. 818–849

[Wat86] WATSON, Mark W.: Univariate Detrending Methods with Stochastic Trends. In: *Journal of Monetary Economics* 18 (1986), S. 49–75

[Whi58] WHITTLE: On the smoothing of probability density functions. In: *Journal of the Royal Statistical Society: Series B* 20 (1958), S. 334–343

[Wil10] WILKER, Holger: *Weibull-Statistik in der Praxis, 2. Auflage*. Books on Demand GmbH, 2010

[Wol95] WOLF, D. ; UMWELTMINISTERIUM BADEN-WÜRTEMBERG (Hrsg.): *Handbuch Boden - Dioxine in Böden Baden-Württembergs*. Landesanstalt für Umweltschutz Baden-Württemberg. Karlsruhe: Umweltministerium Baden-Würtemberg, 1995

[Ziv05] ZIVOT, Eric ; UNIVERSITY OF WASHINGTON (Hrsg.): *Estimation of ARMA Models*. University of Washington, April 2005

Anhang A

Messprogramme

1. Untersuchung von Böden und Gemüse auf PCDD/PCDF
2. Transferuntersuchung Raum Eisenhüttenstadt
3. F&E Biomonitoring Elbe
4. Untersuchung der Belastung von Böden, Miesmuscheln und Sedimenten aus dem Bereich Wilhelmshaven
5. Untersuchungen von Biota, Emissionen, Depositionen und Immissionen des Landes NRW
6. Bioindikation luftgetragener Dioxine und Furane mit Fichtennadeln
7. Biomonitoring (Weidelgras, Grünkohl) an 6 Standorten
8. Allergene und kanzerogene Luftschadstoffe (früher: Chloraromaten-Dioxin-Messprogramm Thüringen)

9. Meßprogramm zur Untersuchung der Dioxinbelastung in der Stadtgemeinde Bremen
10. Untersuchungsprogramm zu Grünkohl
11. PCDD/Fs, dioxinähnliche PCB und Marker-PCB in Eiern von Wanderfalken aus Deutschland
12. Untersuchung von Fischen aus bayerischen Gewässern - Fischmonitoring-Programm
13. Umweltprobenbank des Bundes-Fische
14. Bestimmung von PCDD/PCDF, dl PCB und Ind. PCB in Fichten- und Kieferntrieben aus der UPB des Bundes
15. Umweltprobenbank des Bundes - Forschungsprojekt Möweneier
16. Umweltprobenbank des Bundes - Forschungsprojekt Aalmuttermuskulatur
17. Fischuntersuchung
18. Umweltprobenbank des Bundes - Routineprogramm Möweneier
19. Umweltprobenbank des Bundes - Routineprogramm Brassenmuskulatur
20. Fischuntersuchungen im Dortmunder Hafengebiet

Anhang B

Auswertungstabellen zu Querschnittsdaten

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Aal 2002	0.42	0.39	0	105	24
Aal 2003	0.62	0.49	0	144	47
Aal 2004	0.49	0.30	0	123	34
Barbe 2003	0.37	0.37	9999	10	4
Hecht 2003-2004	0.80	0.80	9999	14	3
Zander 2002-2004	0.27	0.27	9999	19	3
Brasse 1995	0.20	0.20	9999	13	3
Brasse 1998	0.34	0.34	9999	11	3
Brasse 2000	0.36	0.36	9999	12	3
Brasse 2001	0.92	0.56	0	12	4
Brasse 2002	0.52	0.52	9999	15	3

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Brasse 2003	1.23	0.77	0	20	3
Brasse-Filet 1999	0.42	0.19	0	14	3
Brasse-Filet 2000	0.41	0.41	9999	13	3
Brasse-Filet 2001	0.86	0.86	9999	14	3
Brasse-Filet 2002	0.49	0.49	9999	13	3
Brasse-Filet 2003	0.71	0.71	9999	14	3
Brasse-Filet 2004	0.43	0.43	9999	13	3
Brasse-Filet 2005	0.67	0.67	9999	14	3
Brasse-Filet 2006	0.45	0.45	9999	13	3

Tabelle B.17: Schätzungen der natürlichen Variabilität der gesamten univariaten Querschnittsdaten im Vergleich zur Schätzung der natürlichen Variabilität nach Ausschluß der Ausreißer, die anhand des Hampel-Tests detektiert wurden; Indikatorspalte gibt an, wie sich die Schätzung der natürlichen Variabilität nach Ausschluss statistischer Ausreißer ändert: 0=Verringerung, 1=Vergrößerung, 9999=keine Änderung, da keine Ausreißer detektiert wurden

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
PCB Leitkongenere					
Aal 2002	1.87	1.28	0	105	24

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Aal 2003	2.00	1.12	0	144	47
Aal 2004	2.05	1.51	0	123	34
Hecht 2003+2004	2.24	1.62	0	14	4
Brasse 1995	0.69	0.76	1	13	3
Brasse 2000	0.81	0.58	0	12	3
Brasse 2001	2.09	1.74	0	12	3
Brasse 2002	1.43	1.34	0	13	3
Brasse 2003	3.12	3.49	1	12	3
Brasse-Filet 1991-1994	1.59	0.83	0	14	4
Brasse-Filet 1995	1.29	1.08	0	13	3
Brasse-Filet 1996	1.27	0.91	0	13	3
Brasse-Filet 1997	1.78	1.06	0	13	3
Brasse-Filet 1998	0.88	0.94	1	13	3
Brasse-Filet 1999	1.35	1.01	0	13	3
Brasse-Filet 2000	1.16	1.26	1	13	3
Brasse-Filet 2001	1.96	1.27	0	13	3
Brasse-Filet 2002	1.17	1.23	1	13	3
Brasse-Filet 2003	1.60	0.98	0	13	3
Brasse-Filet 2004	0.93	0.97	1	13	3
Brasse-Filet 2005	1.63	1.56	0	13	3
Brasse-Filet 2006	0.90	0.84	0	13	3
Fichte-Trieb 2000-2004	0.37	0.36	0	23	5
Grünkohl TS 2000	0.96	0.68	0	18	6

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Grünkohl TS 2001	0.35	0.26	0	18	6
Grünkohl TS 2002	0.31	0.17	0	16	5
Grünkohl TS 2003	0.57	0.14	0	16	5
Weidelgras 2000	1.48	0.61	0	27	8
Weidelgras 2001	0.69	0.36	0	25	9
Weidelgras 2002	0.34	0.24	0	22	8
Weidelgras 2003	0.56	0.53	0	24	9
Wanderfalke-Ei 2000-2003	1.88	1.86	0	19	4
		dl PCB			
Weidelgras 2003	2.46	2.29	0	24	6
		PCDD			
Aal 2003	0.99	0.62	0	18	5
Brasse-Muskulatur 1994	57.17	1.93	0	14	3
Brasse 2002	2.44	1.16	0	15	4
Brasse 2003	5.60	3.60	0	20	6
Fichte-Nadel 1995	1.09	0.79	0	20	6
Fichte-Nadel 1996	0.65	0.42	0	20	6
Fichte-Nadel 1997	0.55	0.46	0	21	7
Fichte-Nadel 1998	0.43	0.19	0	20	6
Fichte-Nadel 1999	2.27	2.03	0	29	9
Fichte-Nadel 2000	3.63	1.79	0	27	9

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Fichte-Nadel 2001	2.16	1.33	0	26	9
Fichte-Nadel 2002	1.86	1.46	0	21	7
Grünkohl FG 1995	2.80	2.51	0	16	4
Grünkohl FG 1996	1.79	0.87	0	21	7
Grünkohl FG 1997	1.49	1.40	0	14	3
Grünkohl TS 2001	2.68	1.99	0	18	5
Grünkohl TS 2003	3.23	3.30	1	15	4
Weidelgras 2000	1.12	3.30	1	26	4
Weidelgras 2002	6.70	3.07	0	22	6
Weidelgras 2003	2.42	1.79	0	24	6
		PCDF			
Brasse 2003	7.99	5.68	0	20	5
Fichte-Nadel 1995	2.18	2.21	1	20	5
Fichte-Nadel 1996	1.52	0.89	0	20	5
Fichte-Nadel 1997	0.76	0.80	1	21	5
Fichte-Nadel 1998	0.51	0.32	0	20	5
Fichte-Nadel 1999	2.61	1.98	0	29	9
Fichte-Nadel 2000	3.35	2.44	0	27	8
Fichte-Nadel 2001	2.78	1.59	0	26	8
Fichte-Nadel 2002	2.05	2.00	0	21	5
Grünkohl FG 1995	2.54	1.50	0	16	4
Grünkohl FG 1996	1.34	1.05	0	21	5

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Weidelgras 2000	1.56	1.41	0	26	8
Weidelgras 2003	2.79	2.49	0	24	7

Tabelle B.18: Schätzungen der natürlichen Variabilität der Gesamtdaten im Vergleich zur Schätzung der natürlichen Variabilität nach Ausschluß der Ausreißer, die anhand der robusten Mahalanobis-Distanz detektiert wurden; Indikatorspalte gibt an, wie sich die Schätzung der natürlichen Variabilität nach Ausschluss statistischer Ausreißer ändert: 0=Verringerung, 1=Vergrößerung, 9999=keine Änderung, da keine Ausreißer detektiert wurden

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
PCB Leitkongenere					
Aal 2002	1.86675341	1.27475701	0	105	27
Aal 2003	1.99874751	1.67054524	0	144	38
Aal 2004	2.04661205	1.82228935	0	123	30
Hecht	2.2444918	1.59958041	0	14	2
Brasse 1995	0.68567403	0.68567403	9999	13	0
Brasse 1998	0.92438031	0.64270064	0	11	3
Brasse 2000	0.8128143	0.8128143	9999	12	0
Brasse 2001	2.09131571	2.09131571	9999	12	0

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Brasse 2002	1.42971495	1.36707539	0	13	1
Brasse 2003	3.12231279	3.12231279	9999	12	0
Brasse-Filet 1991-1994	1.59395543	0.99104107	0	14	2
Brasse-Filet 1995	1.28518093	1.15906807	0	13	1
Brasse-Filet 1996	1.2706392	1.10763557	0	13	1
Brasse-Filet 1997	1.78058499	1.52882962	0	13	1
Brasse-Filet 1998	0.8833306	0.89629242	1	13	1
Brasse-Filet 1999	1.35042311	0.95321972	0	13	1
Brasse-Filet 2000	1.15623607	1.21565911	1	13	1
Brasse-Filet 2001	1.95925891	2.06779749	1	13	1
Brasse-Filet 2002	1.16874638	1.16874638	9999	13	0
Brasse-Filet 2003	1.59656807	0.87873061	0	13	1
Brasse-Filet 2004	0.93130906	0.93130906	9999	13	0
Brasse-Filet 2005	1.62999166	1.14515544	0	13	1
Brasse-Filet 2006	0.89705848	0.92417835	1	13	1
Fichte-Trieb 2000-2004	0.36949966	0.33710115	0	23	7
Grünkohl TS 2000	0.96388179	0.65832418	0	18	6
Grünkohl TS 2001	0.35481733	0.26080792	0	18	6
Grünkohl TS 2002	0.31358452	0.17901415	0	16	4
Grünkohl TS 2003	0.57037897	0.3747403	0	16	3
Weidelgras 2000	1.48166621	1.01169233	0	27	5
Weidelgras 2001	0.69090247	0.35519218	0	25	9
Weidelgras 2002	0.34212953	0.2739654	0	22	7

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Weidelgras 2003	0.56323484	0.52747913	0	24	9
Wanderfalke Ei 1990	0.86859789	0.7271283	0	10	2
Wanderfalke Ei 2000-2003	1.87533749	1.94019588	1	19	6
		dl PCB			
Fichte-Trieb 2000-2004	3.52365778	3.33936996	0	23	9
Grünkohl TS 2003	4.70392142	4.91893336	1	16	2
Weidelgras 2003	2.46384664	2.46384664	9999	24	0
Aal 2003	2.52768755	2.41182518	0	20	6
Brasse 2003	4.69325284	2.76379358	0	20	6
Wanderfalke Ei 2000-2003	4.56325305	5.52288344	1	20	6
		PCDD			
Aal 2003	0.98884965	0.69418463	0	18	3
Brasse-Muskulatur 1994	57.1678719	57.1678719	9999	14	0
Brasse 1995	2.59417152	1.8976813	0	13	4
Brasse 1998	2.25562521	1.22768933	0	11	2
Brasse 2000	1.94426943	1.42289718	0	12	3
Brasse 2001	3.10744559	2.06394107	0	12	3
Brasse 2002	2.43875443	2.53831588	1	15	1
Brasse 2003	5.59593695	4.13605374	0	20	6
Miesmuschel 1994	0.26393929	0.28596289	1	13	4
Fichte-Nadel 1995	1.09448181	0.76445371	0	20	5

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Fichte-Nadel 1996	0.64508585	0.61257163	0	20	6
Fichte-Nadel 1997	0.54677884	0.49971222	0	21	6
Fichte-Nadel 1998	0.43186383	0.14699899	0	20	6
Fichte-Nadel 1999	2.27145737	2.00422316	0	29	10
Fichte-Nadel 2000	3.63499232	1.99117202	0	27	10
Fichte-Nadel 2001	2.16116889	1.54911753	0	26	9
Fichte-Nadel 2002	1.86443581	1.5935662	0	21	7
Grünkohl FG 1995	2.79915375	2.74062659	0	16	2
Grünkohl FG 1996	1.78920543	1.04142856	0	21	7
Grünkohl FG 1997	1.48873777	1.48873777	9999	14	0
Grünkohl TS 2001	2.67641237	2.29554473	0	18	4
Grünkohl TS 2002	1.89963785	1.84064862	0	12	2
Grünkohl TS 2003	3.23417146	3.18117171	0	15	1
Weidelgras 2002	6.69821854	3.01957833	0	22	7
Weidelgras 2003	2.42471021	1.76390136	0	24	8
Wanderfalke Ei 2003	1.07485322	0.92235228	0	13	4
		PCDF			
Brasse 2002	4.59160115	4.55209016	0	15	3
Brasse 2003	7.98551461	7.98551461	9999	20	0
Fichte-Nadel 1995	2.18455954	2.18455954	9999	20	0
Fichte-Nadel 1996	1.51920983	1.51920983	9999	20	0
Fichte-Nadel 1997	0.76217888	0.77301889	1	21	1

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.
Fichte-Nadel 1998	0.51459733	0.51459733	9999	20	0
Fichte-Nadel 1999	2.60886676	1.74423473	0	29	9
Fichte-Nadel 2000	3.34747884	2.98891108	0	27	7
Fichte-Nadel 2001	2.78389232	1.89968634	0	26	5
Fichte-Nadel 2002	2.04756221	2.03791391	0	21	1
Grünkohl FG 1995	2.53932377	2.53932377	9999	16	0
Grünkohl FG 1996	1.34014655	0.81942745	0	21	5
Grünkohl FG 1997	1.67509018	1.33483535	0	14	4
Grünkohl TS 2000	4.5435027	2.77420447	0	18	5
Grünkohl TS 2001	4.34227133	2.85420918	0	18	6
Grünkohl TS 2003	3.63698882	3.85558787	1	15	3
Weidelgras 2000	1.56269821	1.14573357	0	26	6
Weidelgras 2001	2.77430509	2.58978897	0	23	3
Weidelgras 2002	2.79538383	2.55568574	0	21	1
Weidelgras 2003	2.79170847	2.10639635	0	24	4

Biota	σ	σ_{clean}	Indikator	# Proben	# Ausr.

Tabelle B.19: Schätzungen der natürlichen Variabilität der Gesamtdaten im Vergleich zur Schätzung der natürlichen Variabilität nach Ausschluß der Ausreißer, die anhand des Verfahrens von Pena und Prieto detektiert wurden; Indikatorspalte gibt an, wie sich die Schätzung der natürlichen Variabilität nach Ausschluss statistischer Ausreißer ändert: 0=Verringerung, 1=Vergrößerung, 9999=keine Änderung, da keine Ausreißer detektiert wurden

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$hat p_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Aal	3	3	1.0000	0.2924	n.b.	372	18	0.0484	0.0289	0.0754
Barbe	1	0	0.0000	n.b.	0.9750	10	0	0.0000	n.b.	0.3085
Hecht	1	0	0.0000	n.b.	0.9750	14	0	0.0000	n.b.	0.2316
Zander	1	1	1.0000	0.0250	n.b.	19	7	0.3684	0.1629	0.6164
Brassen	6	3	0.5000	0.1181	0.8819	83	3	0.0361	0.0075	0.1020
Brassen-Filet	6	1	0.1667	0.0042	0.6412	108	1	0.0093	0.0002	0.0505

Tabelle B.1: Zusammenfassung der Ergebenisse für dl PCB 118 Querschnittsdaten mit Ausreißerbestimmung nach Hampel; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6); Biota Aal (Anguilla anguilla), Barbe (Barbus barbus), Hecht (Esox lucius), Zander (Sander lucioperca), Brassen (Abramis brama)

Datensatz	# Datensätze	# kontaminierte Datensätze	# Proben (B)	# Ausreißer
Gesamt	20	7	606	29
Aal	3	3	372	18
Barbe	1	0	10	0
Hecht	1	0	14	0
Zander	1	1	19	7
Brasse	6	2	83	3
Brasse-Filet	6	1	108	1

Tabelle B.2: Zusammenfassung der Ergebnisse der Analysen für univariate Querschnittsdaten auf statistische Ausreißer nach Hampel

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$hat p_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Trieb	1	1	1	0.025	n.b.	23	9	0.3913	0.1971	0.6146
Grünkohl	1	1	1	0.025	n.b.	16	2	0.125	0.0155	0.3835
Weidelgras	1	0	0	n.b.	0.975	24	0	0	n.b.	0.1425
Aal	1	1	1	0.025	n.b.	20	6	0.3	0.1189	0.5428
Brasse	1	1	1	0.025	n.b.	20	6	0.3	0.1189	0.5428
Wanderfalke	1	1	1	0.025	n.b.	20	6	0.3	0.1189	0.5428

Tabelle B.3: Zusammenfassung der Ergebenisse für dl PCB Querschnittsdaten mit Ausreißerbestimmung nach dem Verfahren von Pena und Prieto; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6); Biota Fichte-Trieb (Picea abies), Grünkohl (Brassica oleracea var. sabellica), Weidelgras (Lolium perenne), Wanderfalke (Falco peregrinus)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\, p_p$	$CI^{up}\, p_p$	n	# Ausr.	$hat p_c$	$CI_{low}\, p_c$	$CI^{up}\, p_c$
Fichte-Trieb	1	0	0	n.b.	0.975	23	0	0	n.b.	0.1482
Grünkohl	1	0	0	n.b.	0.975	16	0	0	n.b.	0.2059
Weidelgras	1	1	1	0.025	n.b.	24	6	0.25	0.0977	0.4671
Aal	1	n<2p	n.b.	n.b.	n.b.	20	n<2p	n.b.	n.b.	n.b.
Brasse	1	n<2p	n.b.	n.b.	n.b.	20	n<2p	n.b.	n.b.	n.b.
Wanderfalke	1	n<2p	n.b.	n.b.	n.b.	20	n<2p	n.b.	n.b.	n.b.

Tabelle B.4: Zusammenfassung der Ergebenisse für dl PCB Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Trieb	1	1	1.0000	0.0250	n.b.	23	5	0.2174	0.0746	0.4370
Grünkohl	4	4	1.0000	0.3976	n.b.	68	22	0.3235	0.2151	0.4479
Weidelgras	4	4	1.0000	0.3976	n.b.	98	34	0.3469	0.2536	0.4498
Aal	3	3	1.0000	0.2924	n.b.	372	105	0.2823	0.2371	0.3310
Hecht	1	1	1.0000	0.0250	n.b.	14	4	0.2857	0.0839	0.5810
Brasse	6	5	0.8333	0.3588	0.9958	73	15	0.2055	0.1198	0.3162
Brasse-Filet	13	11	0.8462	0.5455	0.9808	170	40	0.2353	0.1737	0.3063
Wanderfalke	1	1	1.0000	0.0250	n.b.	19	4	0.2105	0.0605	0.4557

Tabelle B.5: Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\, p_p$	$CI^{up}\, p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\, p_c$	$CI^{up}\, p_c$
Brasse-Filet	13	10	0.7692	0.4619	0.9496	164	33	0.2012	0.1427	0.2708

Tabelle B.6: Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; Proben mit fehlenden Daten wurden ausgeschlossen; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Trieb	1	1	1.0000	0.0250	n.b.	23	7	0.3043	0.1321	0.5292
Grünkohl	4	4	1.0000	0.3976	n.b.	68	19	0.2794	0.1773	0.4015
Weidelgras	4	4	1.0000	0.3976	n.b.	98	30	0.3061	0.2170	0.4074
Aal	3	3	1.0000	0.2924	n.b.	372	95	0.2554	0.2118	0.3029
Hecht	1	1	1.0000	0.0250	n.b.	14	2	0.1429	0.0178	0.4281
Brasse	6	2	0.3333	0.0433	0.7772	73	4	0.0548	0.0151	0.1344
Brasse-Filet	13	11	0.8462	0.5455	0.9808	170	12	0.0706	0.0370	0.1201
Wanderfalke	2	2	1.0000	0.1581	n.b.	29	8	0.2759	0.1273	0.4724

Tabelle B.7: Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Brasse-Filet	13	10	0.7692	0.4619	0.9496	164	14	0.0854	0.0475	0.1391

Tabelle B.8: Zusammenfassung der Ergebenisse für Querschnittsdaten der PCB Leitkongenere mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; Proben mit fehlenden Daten wurden ausgeschlossen; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Nadel	8	8	1.0000	0.6306	n.b.	184	59	0.3207	0.2539	0.3933
Grünkohl frisch	3	3	1.0000	0.2924	n.b.	51	14	0.2745	0.1589	0.4174
Grünkohl trocken	2	2	1.0000	0.1581	n.b.	33	9	0.2727	0.1330	0.4552
Weidelgras	3	3	1.0000	0.2924	n.b.	72	16	0.2222	0.1327	0.3356
Aal	1	1	1.0000	0.0250	n.b.	18	5	0.2778	0.0969	0.5348
Brasse - Muskulatur	1	1	1.0000	0.0250	n.b.	14	3	0.2143	0.0466	0.5080
Brasse	2	2	1.0000	0.1581	n.b.	35	10	0.2857	0.1464	0.4630

Tabelle B.9: Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Nadel	8	8	1.0000	0.6306	n.b.	184	59	0.3207	0.2539	0.3933
Grünkohl frisch	3	3	1.0000	0.2924	n.b.	51	14	0.2745	0.1589	0.4174
Grünkohl trocken	2	2	1.0000	0.1581	n.b.	33	9	0.2727	0.1330	0.4552
Weidelgras	3	3	1.0000	0.2924	n.b.	72	16	0.2222	0.1327	0.3356
Aal	1	1	1.0000	0.0250	n.b.	18	5	0.2778	0.0969	0.5348
Brasse - Muskulatur	1	1	1.0000	0.0250	n.b.	14	3	0.2143	0.0466	0.5080
Brasse	2	2	1.0000	0.1581	n.b.	35	10	0.2857	0.1464	0.4630

Tabelle B.10: Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; Proben mit fehlenden Daten wurden ausgeschlossen; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Nadel	8	8	1.0000	0.6306	n.b.	184	59	0.3207	0.2539	0.3933
Grünkohl frisch	3	2	0.6667	0.0943	0.9916	51	9	0.1765	0.0840	0.3087
Grünkohl trocken	3	3	1.0000	0.2924	n.b.	45	7	0.1556	0.0649	0.2946
Weidelgras	2	2	1.0000	0.1581	n.b.	46	15	0.3261	0.1953	0.4802
Aal	1	1	1.0000	0.0250	n.b.	18	3	0.1667	0.0358	0.4142
Brasse - Muskulatur	1	0	0.0000	n.b.	0.9750	14	0	0.0000	n.b.	0.2316
Brasse	6	6	1.0000	0.5407	n.b.	83	23	0.2771	0.1845	0.3862
Miesmuscheln	1	1	1.0000	0.0250	n.b.	13	4	0.3077	0.0909	0.6143
Wanderfalke - Ei	1	1	1.0000	0.0250	n.b.	13	4	0.3077	0.0909	0.6143

Tabelle B.11: Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; wenn Proben nicht vollständig waren, wurden fehlende Daten geschätzt; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6); Biota Miesmuschel: lat. Mytilus

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Brasse - Muskulatur	1	1	1.0000	0.0250	n.b.	13	4	0.3077	0.0909	0.6143

Tabelle B.12: Zusammenfassung der Ergebenisse für PCDD Querschnittsdaten mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; Proben mit fehlenden Daten wurden ausgeschlossen; n.b.=nicht bestimmbar; CI_{low} p_p=Untergrenze des Konfidenzintervalles für p_p; CI^{up} p_p=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Nadel	8	5	0.6250	0.2449	0.9148	184	50	0.2717	0.2089	0.3421
Grünkohl frisch	2	2	1.0000	0.1581	n.b.	37	9	0.2432	0.1177	0.4120
Weidelgras	2	2	1.0000	0.1581	n.b.	50	15	0.3000	0.1786	0.4461
Brasse	2	1	0.5000	0.0126	0.9874	35	5	0.1429	0.0481	0.3026

Tabelle B.13: Zusammenfassung der Ergebenisse für PCDF Querschnittsdaten mit Ausreißerbestimmung unter Verwendung der robusten Mahalanobis-Distanz; fehlende Werte traten nicht auf; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Biota	# DS	# kont. DS.	$\hat{p}_p$	$CI_{low}\ p_p$	$CI^{up}\ p_p$	n	# Ausr.	$\hat{p}_c$	$CI_{low}\ p_c$	$CI^{up}\ p_c$
Fichte-Nadel	8	5	0.6250	0.2449	0.9148	184	23	0.1250	0.0809	0.1816
Grünkohl frisch	3	2	0.6667	0.0943	0.9916	51	9	0.1765	0.0840	0.3087
Grünkohl trocken	3	3	1.0000	0.2924	n.b.	51	14	0.2745	0.1589	0.4174
Weidelgras	4	4	1.0000	0.3976	n.b.	94	14	0.1489	0.0839	0.2372
Brasse	2	1	0.5000	0.0126	0.9874	35	3	0.0857	0.0180	0.2306

Tabelle B.14: Zusammenfassung der Ergebenisse für PCDF Querschnittsdaten mit Ausreißerbestimmung unter Verwendung des Verfahrens von Pena und Prieto; fehlende Werte traten nicht auf; n.b.=nicht bestimmbar; $CI_{low}\ p_p$=Untergrenze des Konfidenzintervalles für p_p; $CI^{up}\ p_p$=Obergrenze des Konfidenzintervalles für p_p; Konfidenzintervalle nach Formel (3.6)

Datensatz	# Datensätze	# kontaminierte Datensätze	# Proben	# Ausreißer
Gesamt	66	66	1564	430
Wildlebende Pflanzen	38	38	840	239
Grünkohl TS	6	6	117	31
Grünkohl FG	5	5	88	23
Weidelgras	10	10	244	71
Fichte-Trieb	1	1	23	5
Fichte-Nadel	16	16	368	109
Wildlebende Tiere	28	28	724	191
Aal	4	4	390	110
Hecht	1	1	14	4
Brasse	8	8	117	30
Brasse-Filet	13	13	170	40
Brasse-Muskulatur	1	1	14	3
Wanderfalke	1	1	19	4

Tabelle B.15: Zusammenfassung der Ergebnisse der Analysen von multivariaten Querschnittsdaten auf statistische Ausreißer anhand der robusten Mahalanobis-Distanz; fehlende Werte wurden geschätzt

Datensatz	# Datensätze	# kontaminierte Datensätze	# Proben	# Ausreißer
Gesamt	86	72	1852	389
Wildlebende Pflanzen	46	40	958	217
Grünkohl TS	11	11	180	42
Grünkohl FG	6	4	102	18
Weidelgras	11	10	262	59
Fichte-Trieb	2	2	46	16
Fichte-Nadel	16	13	368	82
Wildlebende Tiere	40	32	894	172
Aal	5	5	410	104
Hecht	1	1	14	2
Brasse	15	10	211	32
Brasse-Filet	13	11	170	12
Brasse-Muskulatur	1	0	14	0
Miesmuscheln	1	1	13	4
Wanderfalke	4	4	62	18

Tabelle B.16: Zusammenfassung der Ergebnisse der Analysen von multivariaten Querschnittsdaten auf statistische Ausreißer anhand des Verfahrens von Pena und Prieto; fehlende Werte wurden geschätzt

Anhang C

Auswertungstabellen zu Zeitreihendaten

		# Ausr.(imp)				
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T
Fichte-Trieb Deutschland	12	-	-	-	-	-
Fichte-Trieb Stadt B	11	-	-	1	-	n.b.
Silbermöwe-Ei (Larus argentatus) Deutschland	8	-	-	-	-	-
Kiefer-Trieb (Pinus) Stadt B	9	-	-	1	-	n.b.
Silbermöwe-Ei Bundesland A	9	-	-	-	-	-
Silbermöwe-Ei Bundesland B	9	-	-	-	-	-

Tabelle C.1: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von dl PCB; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte traten keine fehlenden Werte auf; n.b.=nicht bestimmbar

		# Ausr.(imp)					#Ausr. zu t_{imp}
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T	
Fichte-Trieb Deutschland	19	-	-	-	-	-	-
Fichte-Trieb Stadt B	19	-	-	-	-	-	
Silbermöwe-Ei Deutschland	14	-	-	-	-	-	
Kiefer-Trieb Stadt B	13	-	-	1	-	n.b.	0
Silbermöwe-Ei Bundesland A	15	-	-	2	-	n.b.	2
Silbermöwe-Ei Bundesland B	15	-	-	-	-	-	

Tabelle C.2: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von dl PCB; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte traten keine fehlenden Werte auf; n.b.=nicht bestimmbar

		# Ausr.(imp)					
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T
Brassen-Filet Deutschland	17	-	-	2	-	3
Brassen-Filet Stadt A	14	-	-	-	-	-
Fichte-Trieb Deutschland	12	-	-	-	-	-
Fichte-Trieb Stadt B	11	-	-	-	-	-
Kiefer-Trieb Deutschland	9	-	-	-	-	-
Silbermöwe-Ei Deutschland	17	-	-	-	-	-

Tabelle C.3: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, diese wurden geschätzt

		# Ausr. (del)				
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T
Brassen-Filet Deutschland	17	-	-	-	-	-

Tabelle C.4: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, die entsprechenden Proben wurden entfernt

		# Ausr.(imp)				#Ausr. zu t_{imp}
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T
Brassen-Filet Deutschland	19	-	-	-	-	-
Brassen-Filet Stadt A	16	-	-	-	-	-
Fichte-Trieb Deutschland	19	-	-	-	-	-
Fichte-Trieb Stadt B	19	-	-	-	-	-
Kiefer-Trieb Deutschland	13	-	-	-	-	-
Silbermöwe-Ei Deutschland	19	-	-	-	-	-

Tabelle C.5: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, diese wurden geschätzt

Zeitreihe PCB Leitkongenere		# Ausr. (del)				#Ausr. zu t_{imp}
	T	Typ A	Typ T	Typ I	Typ L	
Brassen-Filet Deutschland	19	-	-	-	-	-

Tabelle C.6: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten der PCB Leitkongenere; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten fehlende Werte auf, die entsprechenden Proben wurden entfernt

		# Ausr.(imp)				
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T
Brassen Deutschland	6	-	-	-	-	-
Brassen Stadt C	6	-	-	-	-	-
Brassen Stadt D	6	-	-	-	-	-
Fichte-Nadel	7	-	-	-	-	-

Tabelle C.7: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDD; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf

		# Ausr.(imp)					#Ausr. zu t_{imp}
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T	
Brassen Deutschland	10	-	-	-	-	-	-
Brassen Stadt C	10	-	-	-	-	-	-
Brassen Stadt D	10	-	-	-	-	-	-
Fichte-Nadel	7	-	-	-	-	-	-

Tabelle C.8: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDD; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf

		# Ausr.(imp)				
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T
Brassen Deutschland	17	-	-	-	-	-
Brassen Stadt C	14	-	-	-	-	-
Brassen Stadt D	12	-	-	-	-	-
Fichte-Nadel	11	-	-	-	-	-

Tabelle C.9: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDF; fehlende Jahresmittelwerte wurden gelöscht; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf

		# Ausr.(imp)					#Ausr. zu t_{imp}
	T	Typ A	Typ T	Typ I	Typ L	Typ I $\rightarrow$ Typ T	
Brassen Deutschland	17	-	-	-	-	-	-
Brassen Stadt C	14	-	-	1	-	n.b.	1
Brassen Stadt D	12	-	-	1	-	n.b.	1
Fichte-Nadel	11	-	-	-	-	-	-

Tabelle C.10: Zusammenfassung der Ergebnisse der Analysen von multivariaten Zeitreihen zu Jahresmittelwerten von PCDF; fehlende Jahresmittelwerte wurden geschätzt; bei der Bestimmung der Jahresmittelwerte aus den Querschnittsdaten traten keine fehlenden Werte auf; n.b.=nicht bestimmbar

Anzahl gefundener Ausreißer vom Typ I													
	0				1				2				
Typ	Leit	dl PCB	PCDD	PCDF	Leit	dl PCB	PCDD	PCDF	Leit	dl PCB	PCDD	PCDF
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
A	5	4	3	1	-	1	-	2	-	1	-	-
B	5	4	3	3	-	2	-	-	-	-	-	-
A1	1	-	-	-	-	-	-	-	-	-	-	-
A2	1	-	-	-	-	-	-	-	-	-	-	-
B1	-	-	-	-	-	-	-	-	1	-	-	-
B2	1	-	-	-	-	-	-	-	-	-	-	-
0	-	-	1	1	-	-	-	-	-	-	-	-

Tabelle C.11: Zusammenfassung der Ergebnisse, wieviele Ausreißer des Typs I nach welcher Auswertevariante gefunden wurden

Anhang D

Kongenerenprofile dl PCB Aal

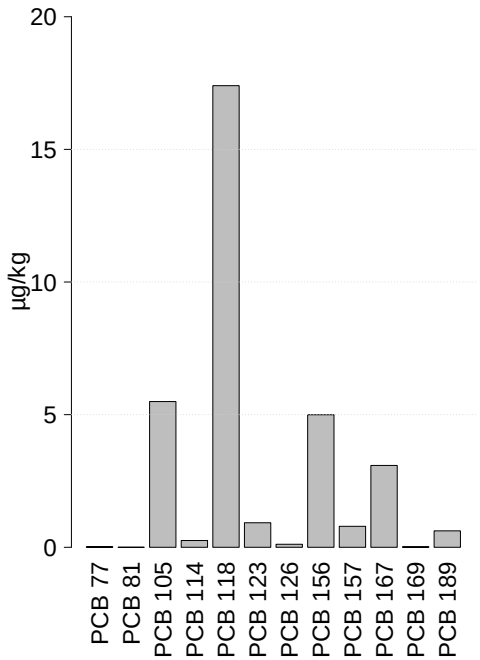

Abbildung D.1: Kongenerenprofil zu dl PCB in Aalen 2003, n=14

Anhang E

Kongenerenprofile dl PCB Brasse

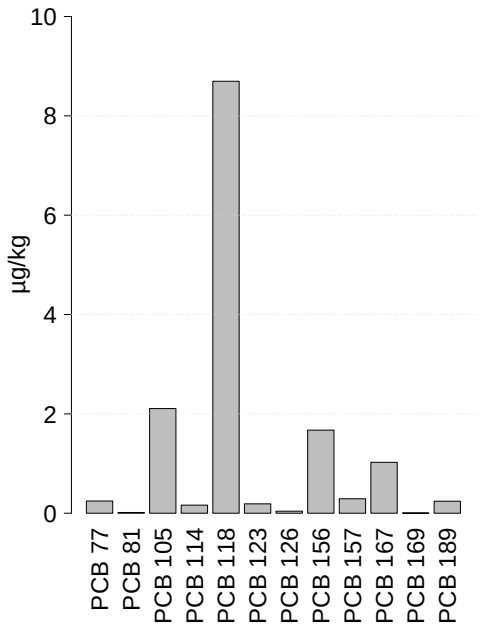

Abbildung E.1: Kongenerenprofil zu dl PCB in Brassen 2003, n=14

Anhang F

Kongenerenprofile dl PCB Grünkohl

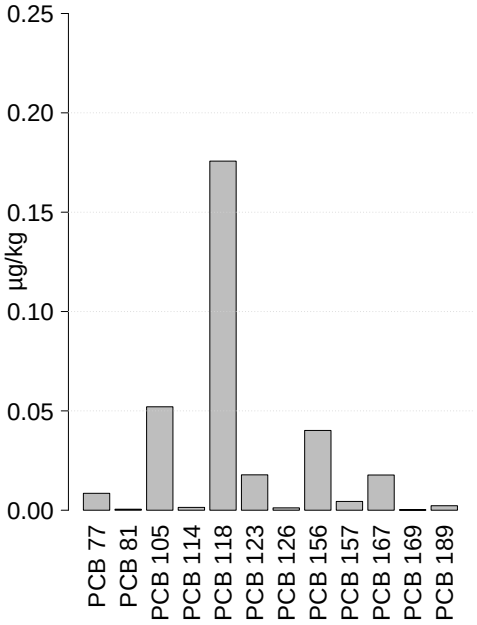

Abbildung F.1: Kongenerenprofil zu dl PCB in Grünkohl 2003, n=14

Anhang G

Kongenerenprofile dl PCB

Weidelgras

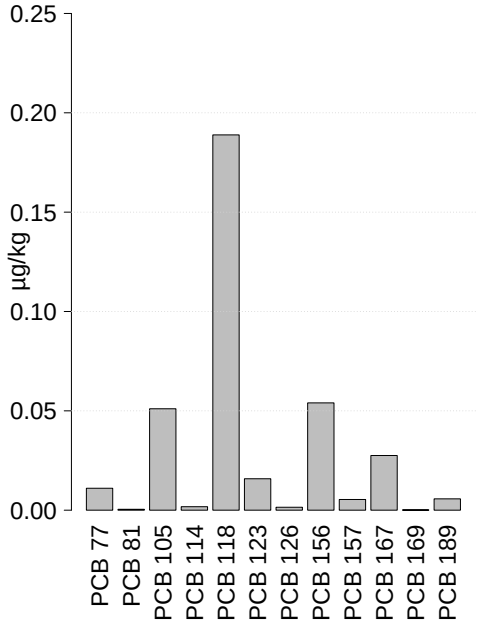

Abbildung G.1: Kongenerenprofil zu dl PCB in Weidelgras 2003, n=24

Anhang H

Kongenereprofile PCB

Leitkongenere Ei des Wanderfalken

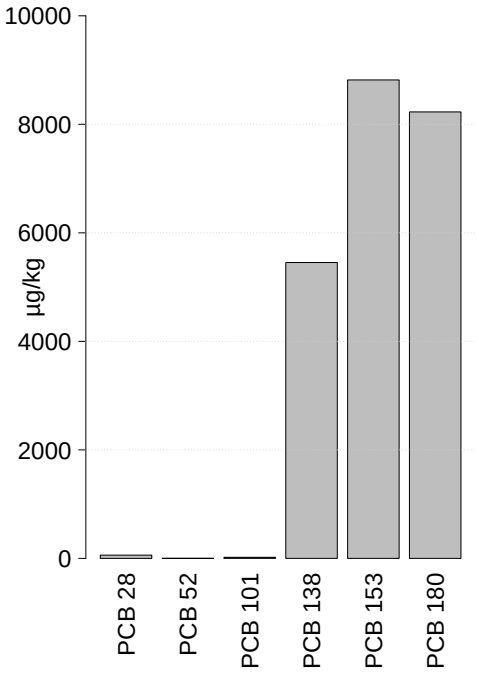

Abbildung H.1: Kongenerenprofil zu PCB Leitkongeneren in Eiern von Wanderfalken 1990, n=10

Anhang I

Kongenereprofile PCB

Leitkongenere Aal

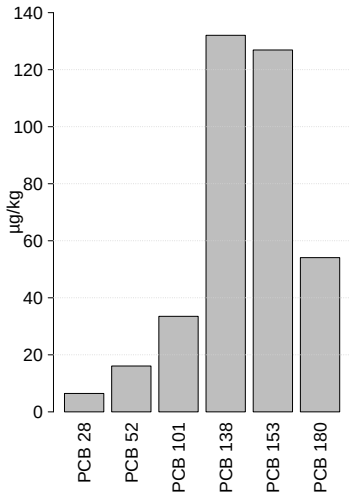

Abbildung I.1: Kongenerenprofil zu PCB Leitkongeneren in Aalen 2002, n=105

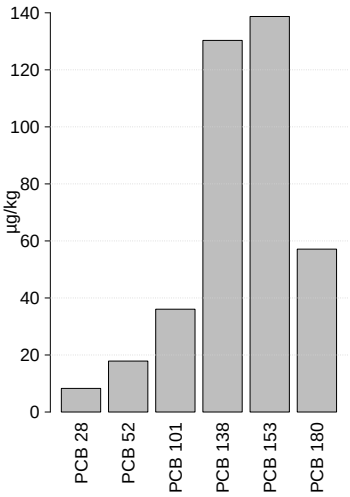

Abbildung I.2: Kongenerenprofil zu PCB Leitkongeneren in Aalen 2003, n=144

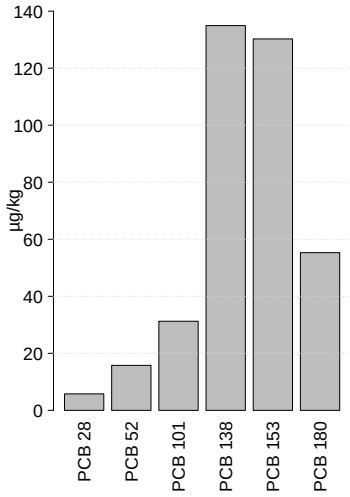

Abbildung I.3: Kongenerenprofil zu PCB Leitkongeneren in Aalen 2004, n=123

Anhang J

Kongenereprofile PCB

Leitkongenere Brasse

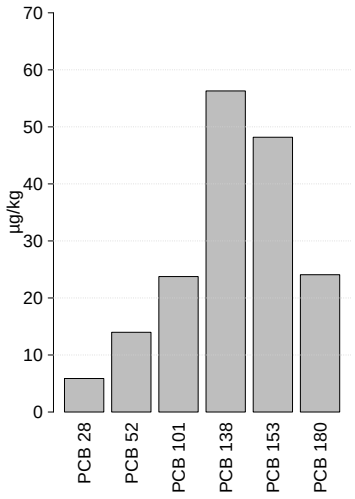

Abbildung J.1: Kongenerenprofil zu PCB Leitkongeneren in Brassen 1995, n=13

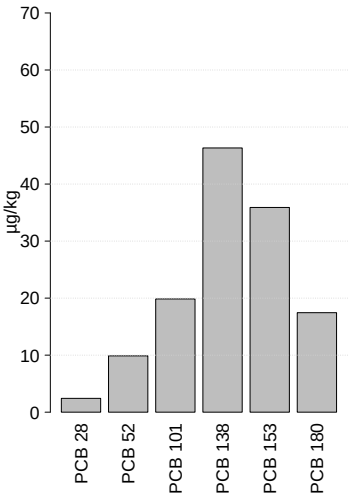

Abbildung J.2: Kongenerenprofil zu PCB Leitkongeneren in Brassen 1998, n=11

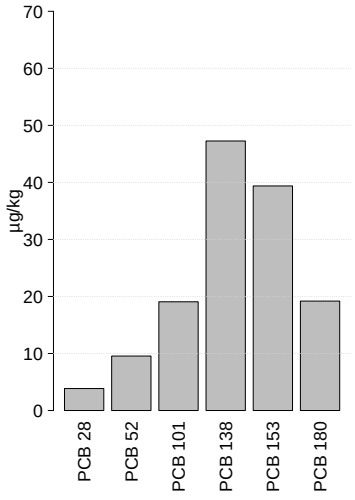

Abbildung J.3: Kongenerenprofil zu PCB Leitkongeneren in Brassen 2000, n=12

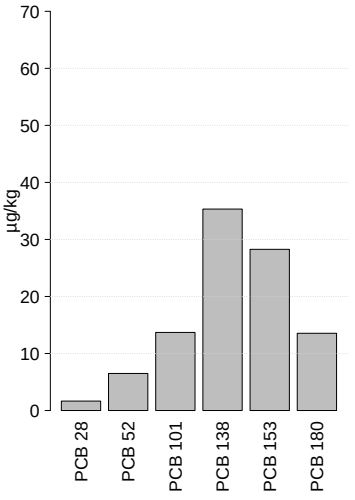

Abbildung J.4: Kongenerenprofil zu PCB Leitkongeneren in Brassen 2001, n=12

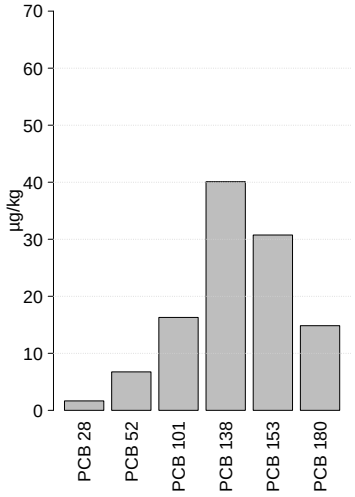

Abbildung J.5: Kongenerenprofil zu PCB Leitkongeneren in Brassen 2002, n=13

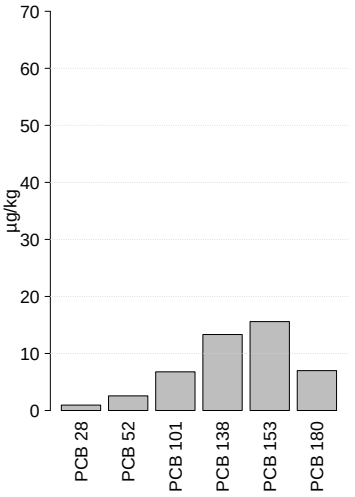

Abbildung J.6: Kongenerenprofil zu PCB Leitkongeneren in Brassen 2003, n=12

Anhang K

Kongenereprofile PCB

Leitkongenere Brassen-Filets

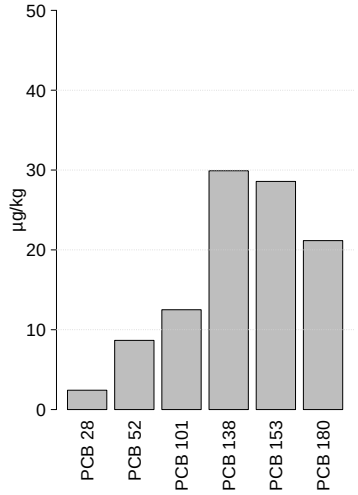

Abbildung K.1: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1995, n=13

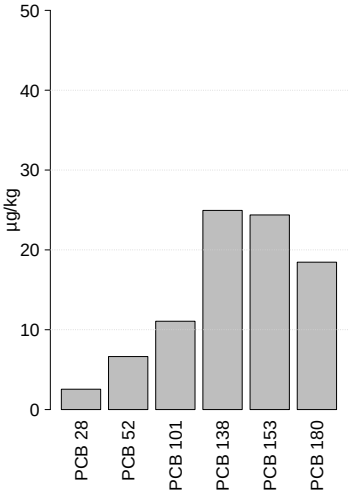

Abbildung K.2: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1996, n=13

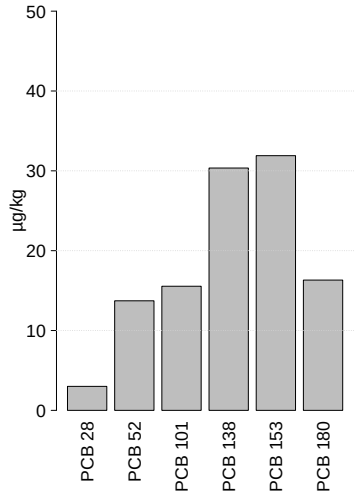

Abbildung K.3: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1997, n=13

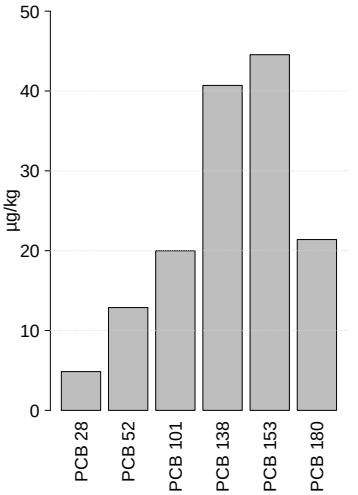

Abbildung K.4: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1998, n=13

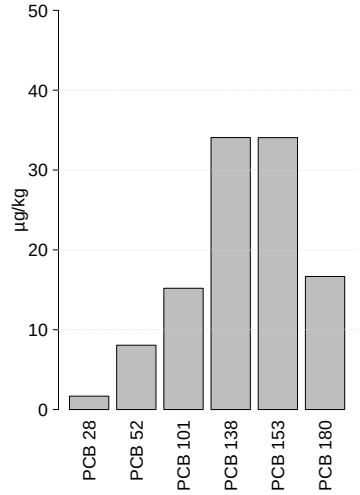

Abbildung K.5: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 1999, n=13

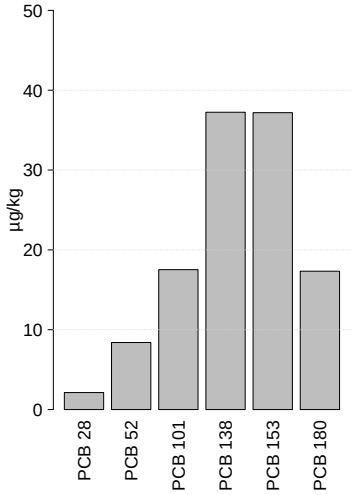

Abbildung K.6: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2000, n=13

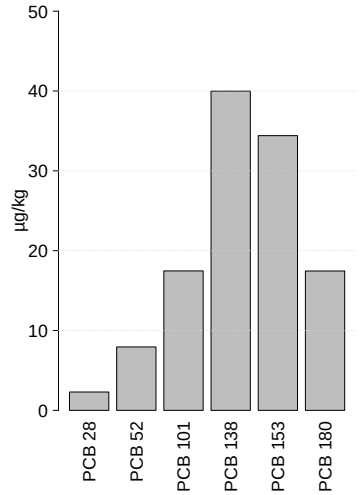

Abbildung K.7: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2001, n=13

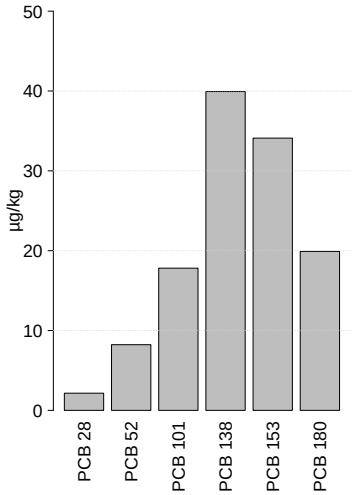

Abbildung K.8: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2002, n=13

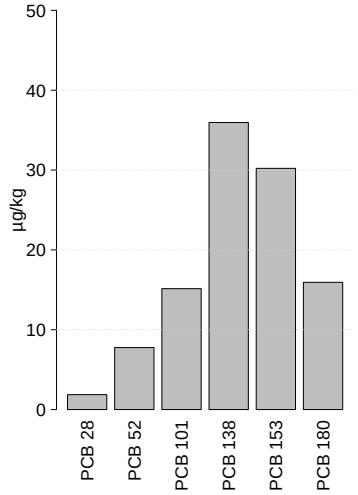

Abbildung K.9: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2003, n=13

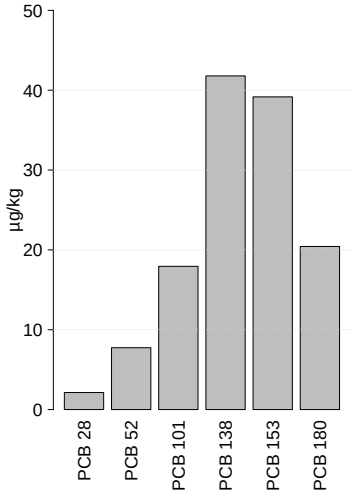

Abbildung K.10: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2004, n=13

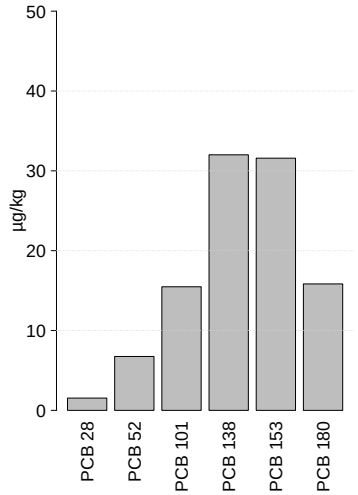

Abbildung K.11: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2005, n=13

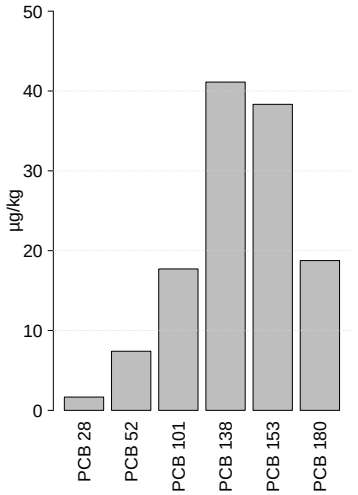

Abbildung K.12: Kongenerenprofil zu PCB Leitkongeneren in Brassen-Filets 2006, n=13

Anhang L

Kongenereprofile PCB

Leitkongenere Grünkohl

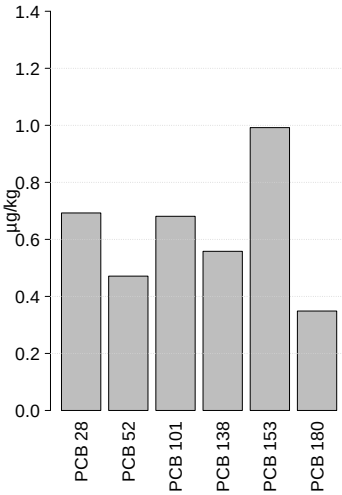

Abbildung L.1: Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2000, n=18

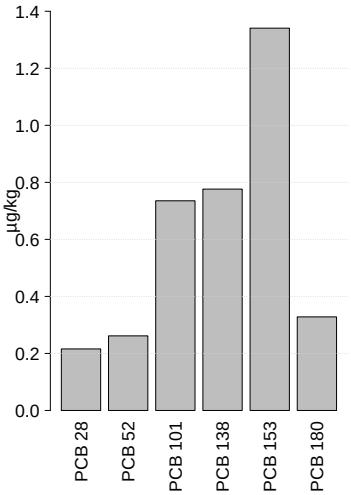

Abbildung L.2: Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2001, n=18

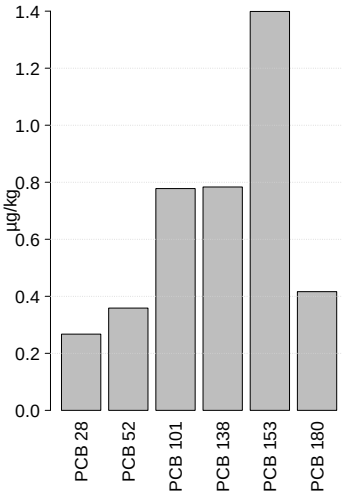

Abbildung L.3: Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2002, n=16

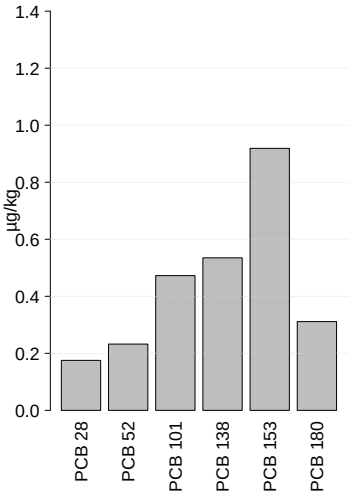

Abbildung L.4: Kongenerenprofil zu PCB Leitkongeneren in Grünkohl (Trockensubstanz) 2003, n=16

Anhang M

Kongenereprofile PCB

Leitkongenere Weidelgras

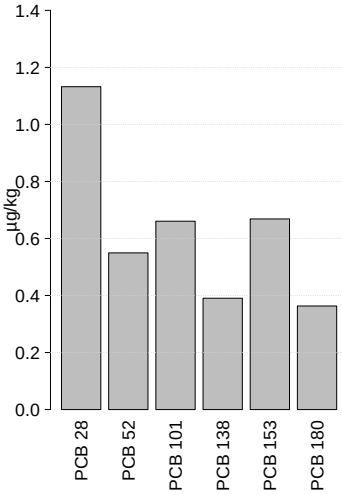

Abbildung M.1: Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2000, n=27

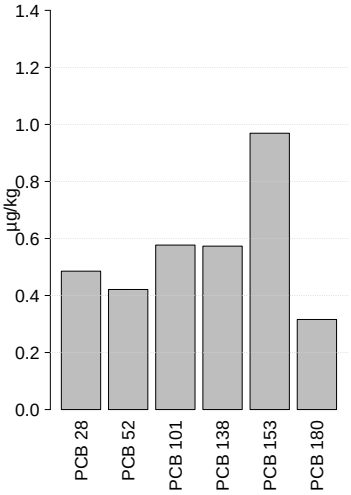

Abbildung M.2: Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2001, n=25

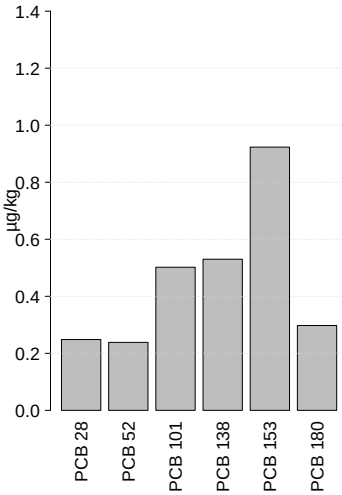

Abbildung M.3: Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2002, n=22

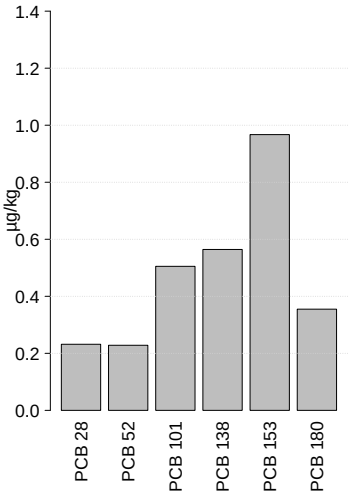

Abbildung M.4: Kongenerenprofil zu PCB Leitkongeneren in Weidelgras 2003, n=24

Anhang N

Kongenerenprofile PCDD Aal

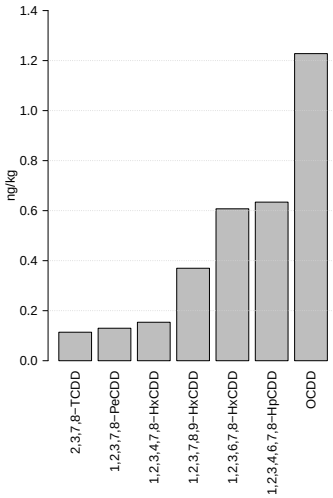

Abbildung N.1: Kongenerenprofil zu PCDD in Aalen 2003, n=18

Anhang O

Kongenerenprofile PCDD Brasse

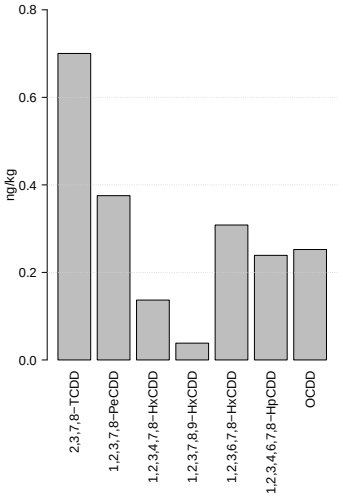

Abbildung O.1: Kongenerenprofil zu PCDD in Brassen 1995, n=13
figure

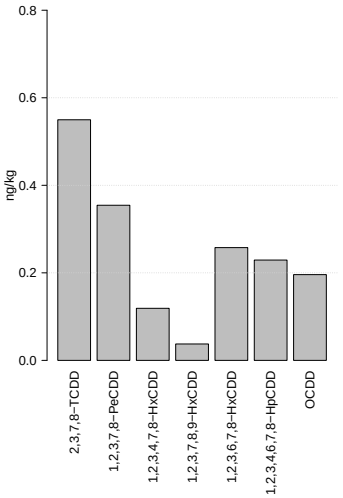

Abbildung O.2: Kongenerenprofil zu PCDD in Brassen 1998, n=11

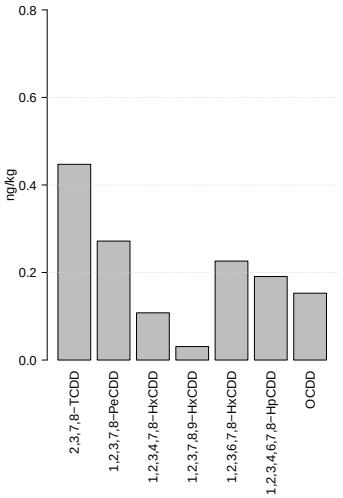

Abbildung O.3: Kongenerenprofil zu PCDD in Brassen 2000, n=12

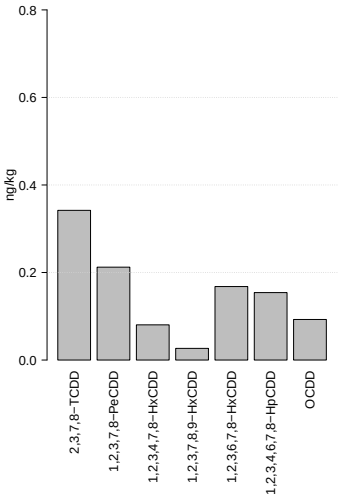

Abbildung O.4: Kongenerenprofil zu PCDD in Brassen 2001, n=12

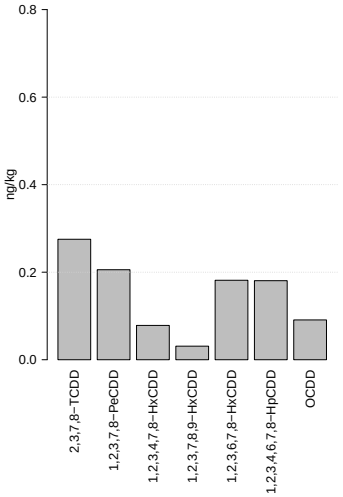

Abbildung O.5: Kongenerenprofil zu PCDD in Brassen 2002, n=15

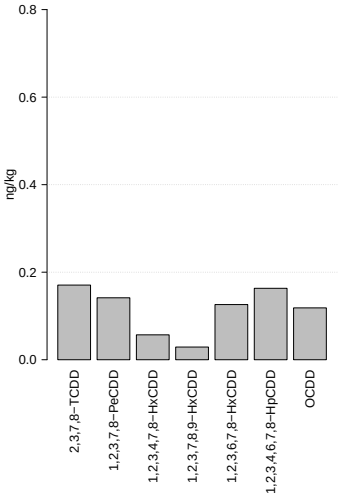

Abbildung O.6: Kongenerenprofil zu PCDD in Brassen 2003, n=20

Anhang P

Kongenerenprofile PCDD

Miesmuschel

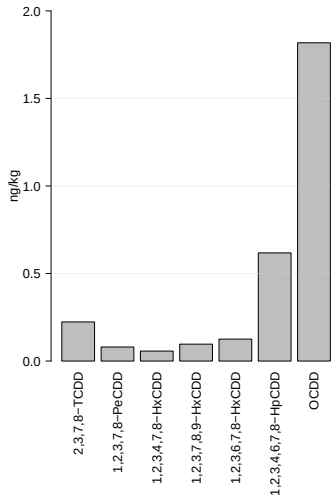

Abbildung P.1: Kongenerenprofil zu PCDD in Miesmuscheln 1994, n=13

Anhang Q

Kongenerenprofile PCDD

Fichte-Nadel

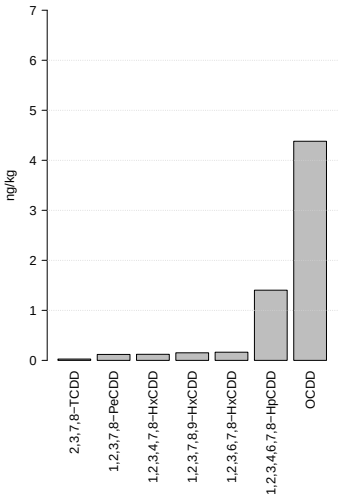

Abbildung Q.1: Kongenerenprofil zu PCDD in Fichtennadeln 1995, n=20

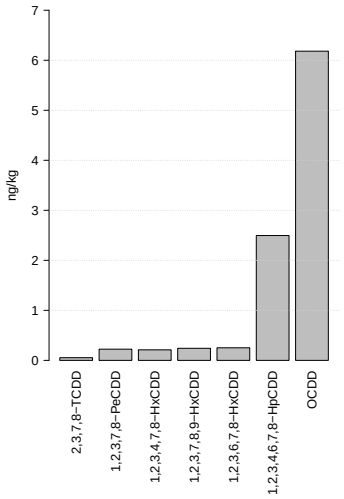

Abbildung Q.2: Kongenerenprofil zu PCDD in Fichtennadeln 1996, n=20

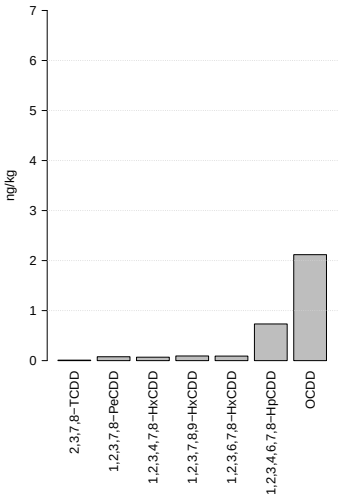

Abbildung Q.3: Kongenerenprofil zu PCDD in Fichtennadeln 1997, n=21

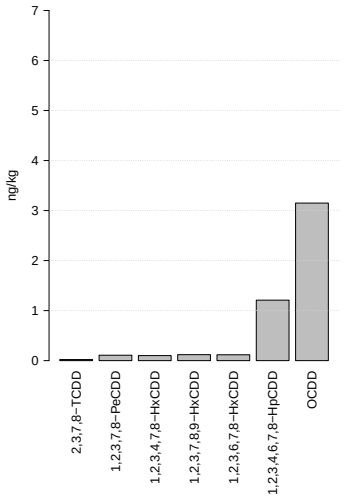

Abbildung Q.4: Kongenerenprofil zu PCDD in Fichtennadeln 1998, n=20

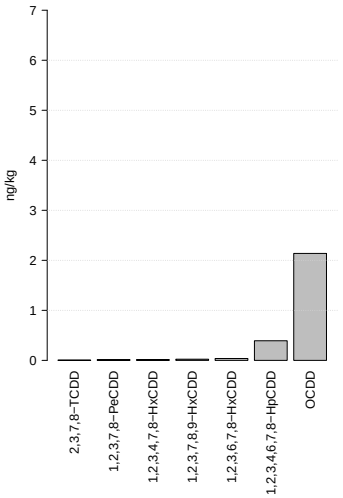

Abbildung Q.5: Kongenerenprofil zu PCDD in Fichtennadeln 1999, n=29

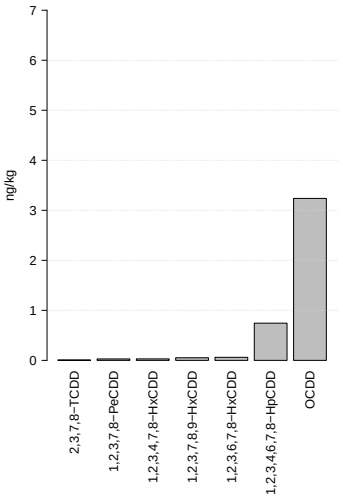

Abbildung Q.6: Kongenerenprofil zu PCDD in Fichtennadeln 2000, n=27

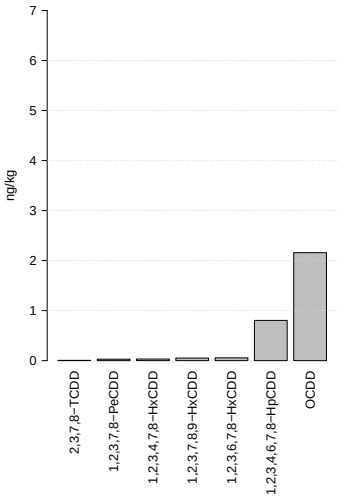

Abbildung Q.7: Kongenerenprofil zu PCDD in Fichtennadeln 2001, n=26

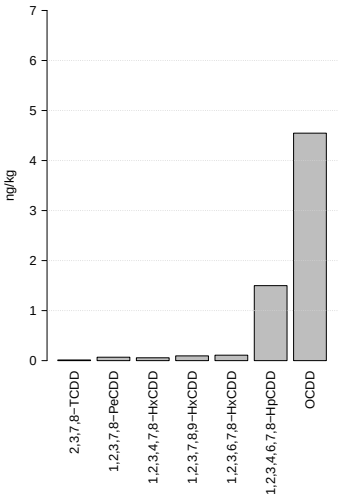

Abbildung Q.8: Kongenerenprofil zu PCDD in Fichtennadeln 2002, n=21

Anhang R

Kongenerenprofile PCDD Grünkohl

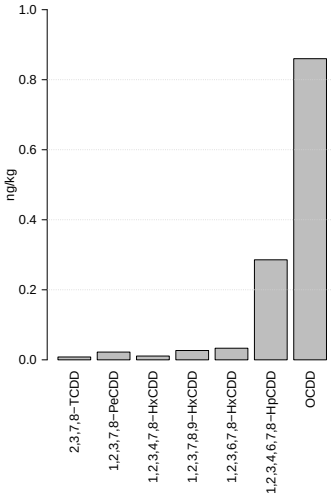

Abbildung R.1: Kongenerenprofil zu PCDD in Grünkohl (Frischgewicht) 1995, n=16

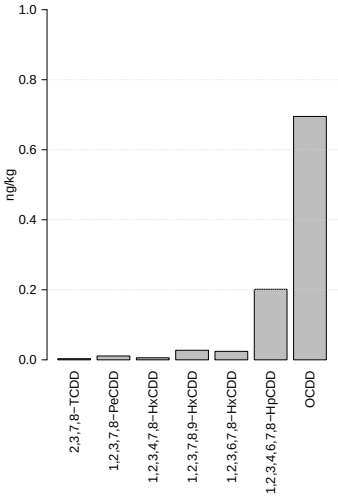

Abbildung R.2: Kongenerenprofil zu PCDD in Grünkohl (Frischgewicht) 1996, n=21

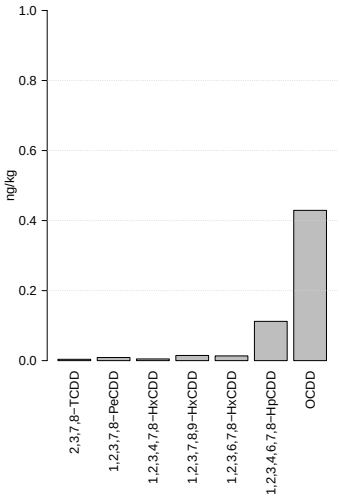

Abbildung R.3: Kongenerenprofil zu PCDD in Grünkohl (Frischgewicht) 1997, n=14

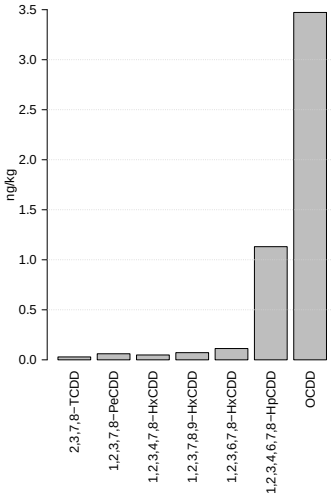

Abbildung R.4: Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2001, n=18

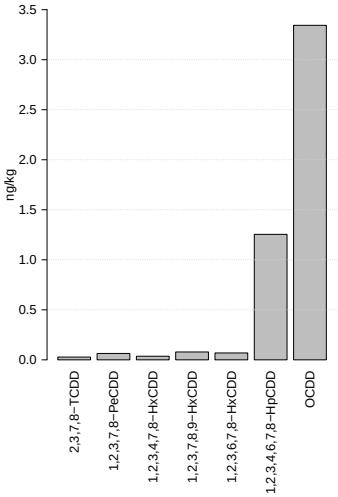

Abbildung R.5: Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2002, n=12

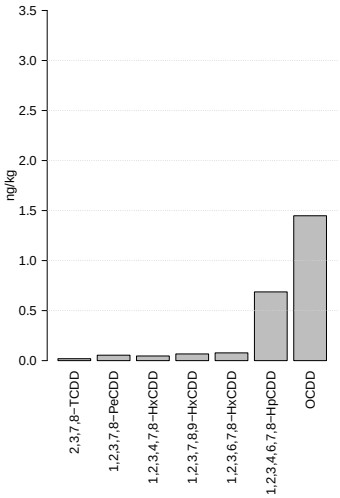

Abbildung R.6: Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2003, n=15

Anhang S

Kongenerenprofile PCDD

Weidelgras

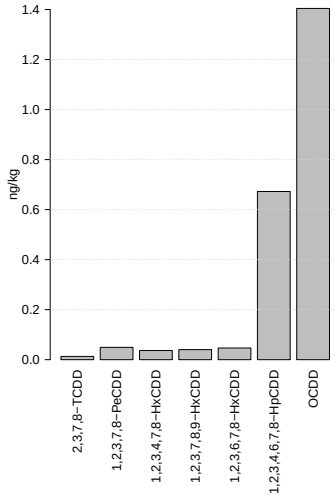

Abbildung S.1: Kongenerenprofil zu PCDD in Weidelgras 2002, n=22

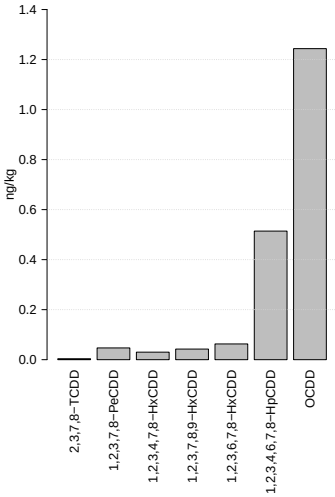

Abbildung S.2: Kongenerenprofil zu PCDD in Weidelgras 2003, n=24

Anhang T

Kongenerenprofile PCDF Brasse

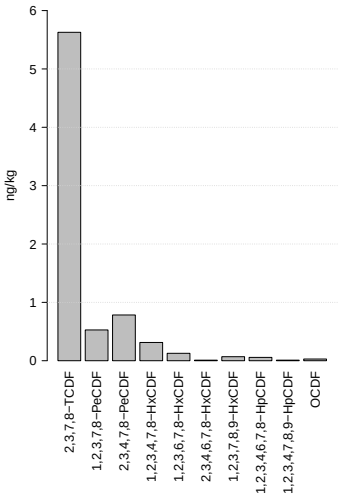

Abbildung T.1: Kongenerenprofil zu PCDF in Brassen 2002, n=15

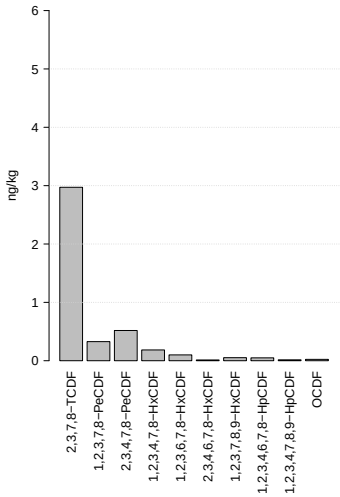

Abbildung T.2: Kongenerenprofil zu PCDF in Brassen 2003, n=20

Anhang U

Kongenerenprofile PCDF

Fichte-Nadel

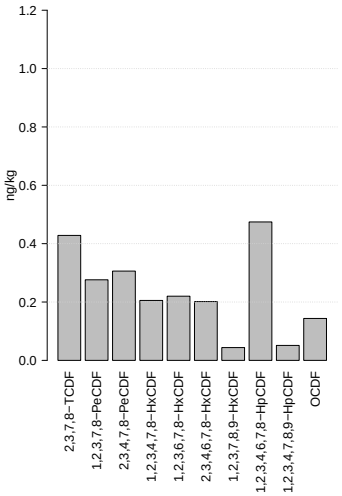

Abbildung U.1: Kongenerenprofil zu PCDF in Fichtennadeln 1995, n=20

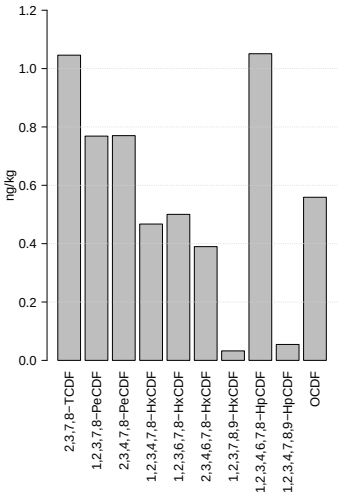

Abbildung U.2: Kongenerenprofil zu PCDF in Fichtennadeln 1996, n=20

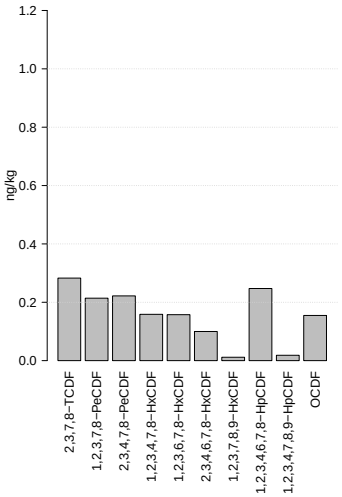

Abbildung U.3: Kongenerenprofil zu PCDF in Fichtennadeln 1997, n=21

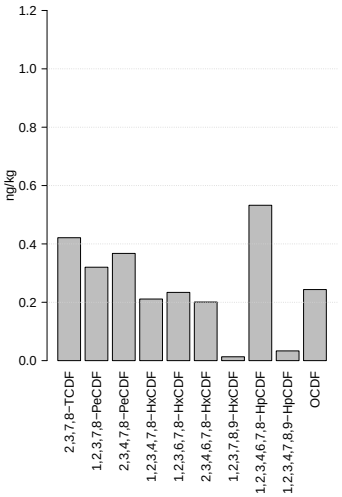

Abbildung U.4: Kongenerenprofil zu PCDF in Fichtennadeln 1998, n=20

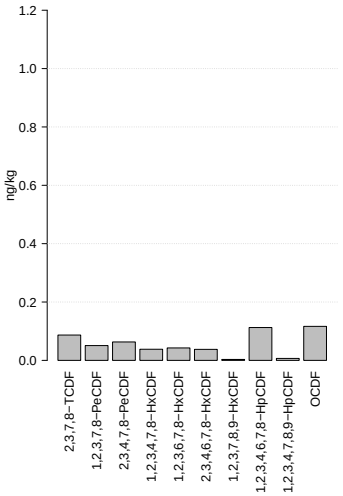

Abbildung U.5: Kongenerenprofil zu PCDF in Fichtennadeln 1999, n=29

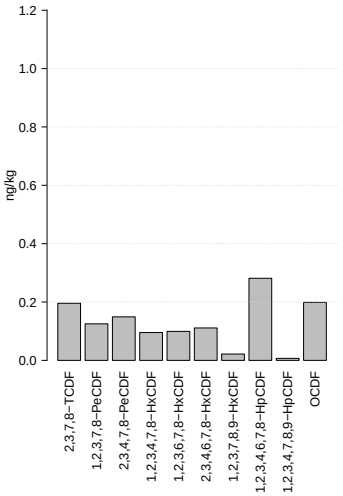

Abbildung U.6: Kongenerenprofil zu PCDF in Fichtennadeln 2000, n=27

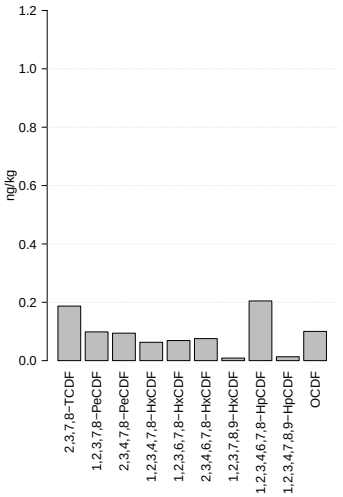

Abbildung U.7: Kongenerenprofil zu PCDF in Fichtennadeln 2001, n=26

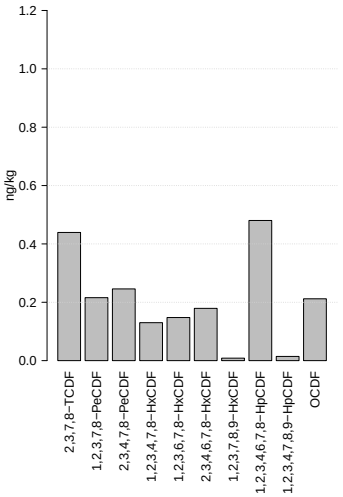

Abbildung U.8: Kongenerenprofil zu PCDF in Fichtennadeln 2002, n=21

Anhang V

Kongenerenprofile PCDF Grünkohl

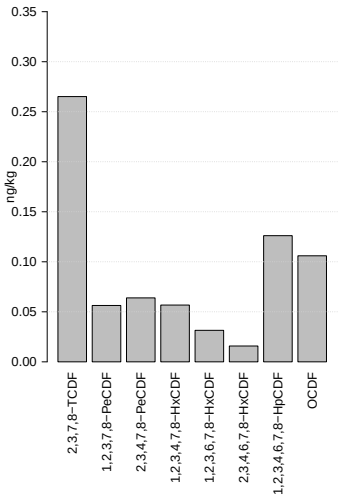

Abbildung V.1: Kongenerenprofil zu PCDF in Grünkohl (Frischgewicht) 1995, n=16

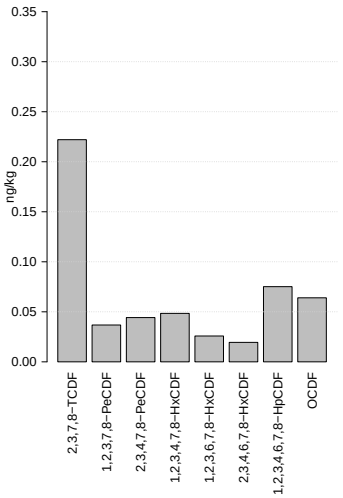

Abbildung V.2: Kongenerenprofil zu PCDF in Grünkohl (Frischgewicht) 1996, n=21

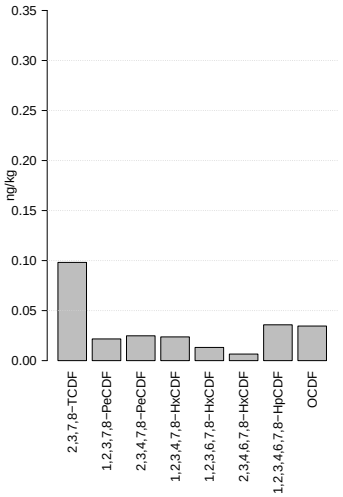

Abbildung V.3: Kongenerenprofil zu PCDF in Grünkohl (Frischgewicht) 1997, n=14

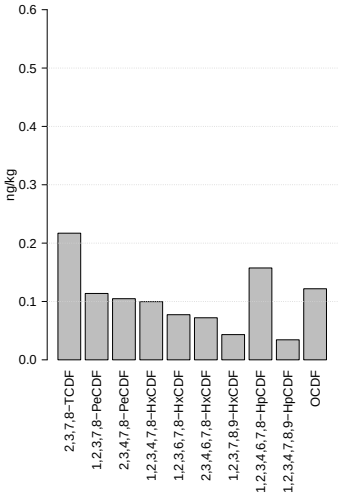

Abbildung V.4: Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2000, n=18

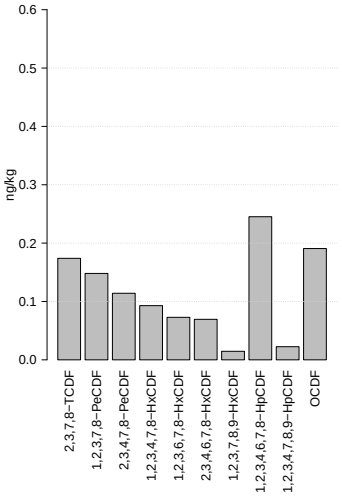

Abbildung V.5: Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2001, n=18

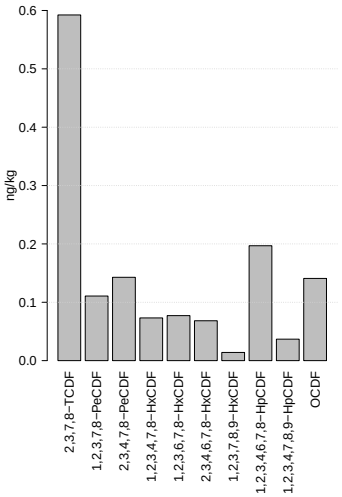

Abbildung V.6: Kongenerenprofil zu PCDD in Grünkohl (Trockensubstanz) 2003, n=15

Anhang W

Kongenerenprofile PCDF

Weidelgras

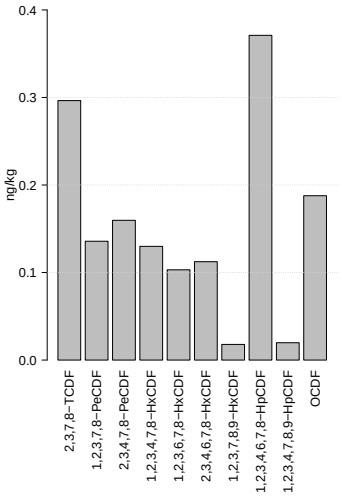

Abbildung W.1: Kongenerenprofil zu PCDF in Weidelgras 2000, n=26

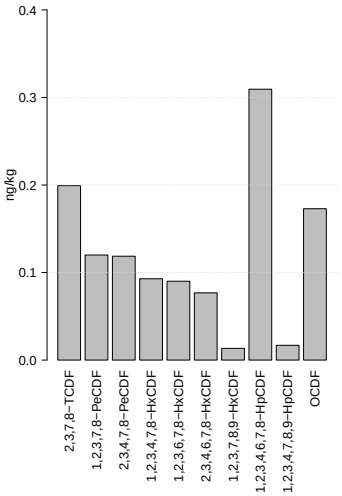

Abbildung W.2: Kongenerenprofil zu PCDF in Weidelgras 2001, n=23

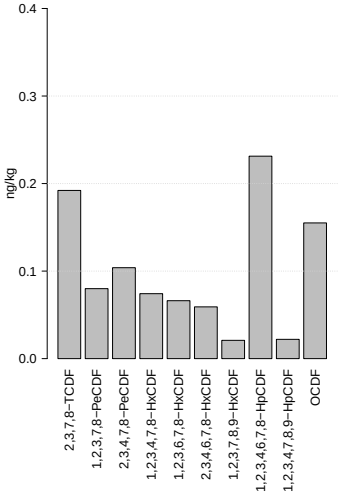

Abbildung W.3: Kongenerenprofil zu PCDF in Weidelgras 2002, n=21

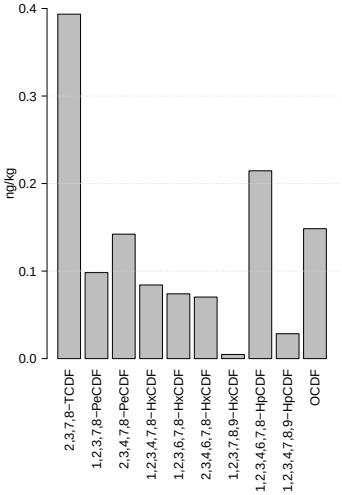

Abbildung W.4: Kongenerenprofil zu PCDF in Weidelgras 2003, n=24