Diplomarbeit
zur Erlangung
des akademischen Grades Diplomingenieur
an der Fakultät für Wirtschaftswissenschaften und Informatik
der Universität Klagenfurt

Erkennen von Strukturbrüchen in Data Warehouses mit Data Mining-Techniken

eingereicht von

Dieter Mitsche

Betreuer:
O. Univ. Prof. DI Dr. Johann Eder
Institut für Informatik-Systeme
Forschungsgruppe Betriebliche Informations- und Kommunikationssysteme

Klagenfurt, im Feber 2003

Ich erkläre ehrenwörtlich, dass ich die vorliegende Schrift verfasst und alle ihr vorausgehenden oder sie begleitenden Arbeiten durchgeführt habe. Die in der Schrift verwendete Literatur sowie das Ausmaß der mir im gesamten Arbeitsvorgang gewährten Unterstützung sind ausnahmslos angegeben.

Die Schrift ist noch keiner anderen Prüfungsbehörde vorgelegt worden.

Klagenfurt, am 7. Feber 2003

Zusammenfassung

Ein gravierendes Problem von Analysen in Data Warehouses stellen Veränderungen in den Strukturen, die den Stammdaten (dimension members) zugrundeliegen, sowie Veränderungen in Berechnungsformeln oder Maßeinheiten von Kennzahlen dar. In diesem Fall spiegeln OLAP-Analysen Trends in den Werten wider, die zu falschen Schlussfolgerungen und inadäquaten Maßnahmen führen können.

In dieser Arbeit wird versucht, diese Veränderungen in den Strukturen von Data Warehouses mit Hilfe von Data Mining-Techniken zu identifizieren; verschiedene einsetzbare Methoden werden zunächst vorgestellt und auf kleinen Datenbeständen mit wenigen dimension members und bekannten Strukturbrüchen erprobt, um die prinzipielle Funktionsfähigkeit der Methoden zu überprüfen. Anschließend werden Probleme beleuchtet, die auftreten, wenn die Verfahren auf realen Data Warehouses angewendet werden: einerseits muss wegen der Größe von realen Datenbeständen der Aspekt der Laufzeit- und Speicherkomplexität untersucht werden, andererseits müssen die Fehlerraten erster und zweiter Ordnung analysiert werden.

In der Arbeit wird ein schrittweiser Ansatz vorgeschlagen, der zuerst die Daten auf Veränderungen in der Kennzahlendimension überprüft und anschließend Brüche in den Strukturdimensionen untersucht, wozu die Werte im Data Warehouse nach einer Strukturdimension aggregiert werden. Sollten bei der Analyse nach einer Strukturdimension keine Auffälligkeiten auftreten, werden die Werte nach zwei Strukturdimensionen aggregiert analysiert, fällt auch hier nichts auf, werden drei Strukturdimensionen gemeinsam analysiert, etc. Die abschließenden Experimente an einem realen Data Warehouse zeigten auf, dass die Methoden durchwegs gut skalieren und die Fehlerraten erster und zweiter Ordnung stark von der Qualität und Volatilität des Datenmaterials abhängen.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1

Einleitung

In diesem einführenden Kapitel werden einerseits grundlegende Begriffe, die für die weitere Arbeit von Bedeutung sind, definiert, andererseits wird auch die Motivation der Behandlung dieser Thematik erläutert.

1.1 Grundlagen von Data Mining und Data Warehouses

Data Mining ist definiert als der automatisierte Prozess der Extraktion neuen und nützlichen Wissens, das in großen Datenbeständen versteckt ist [Ka02]. Die Bedeutung des Data Mining nimmt aufgrund der steigenden Mengen an elektronisch verfügbaren Daten und dank der kontinuierlich wachsenden Rechenleistung von Computern ständig zu; der Einsatz reicht von der klassischen Analyse des Einkaufsverhaltens in Supermärkten und der Kreditwürdigkeitsprüfung eines Darlehenswerbers bis hin zu wissenschaftlichen Anwendungen in der Bildverarbeitung, Meteorologie, Biologie und Medizin, ohne hier einen Anspruch auf Vollständigkeit der Aufzählung zu erheben [Ka02].

Data Warehouses[1] sind meist materialisierte Views über strukturierte oder semi-strukturierte Datenbestände aus unterschiedlichen Datenquellen [EK02], die in aller Regel aggregierte Daten enthalten. Im Gegensatz zu relationalen Datenbanken steht in Data Warehouses die systematische und effiziente Suche bzw. Analyse von Daten nach speziellen Kriterien im Vordergrund [Ze02]. Data Warehouses enthalten eine Menge von Dimensionen; die eigentlichen Daten des Data Warehouses liegen am Schnittpunkt der Ausprägungen der Dimensionen (den *dimension members* bzw. *Dimensionsdaten*); ein Datenschema bzw. kurz Schema bezeichnet alle Dimensionen mit ihren dimension members (sämtliche Metadaten des Data Warehouses) zu einem bestimmten Zeitpunkt. Üblicherweise umfasst die Menge aller Dimensionen eine Zeitdimension,

[1] Der von [Ze02] verwendete deutsche Begriff 'Datenlager' hat sich nicht durchgesetzt.

die die Granularität der Gültigkeitszeit eines Wertes spezifiziert, eine Kennzahlendimension, die die zu analysierenden Kennzahlen beinhaltet, sowie mehrere Strukturdimensionen, die festlegen, welche Kriterien zur Dateneinteilung herangezogen werden [EK02].

Data Mining hat aufgrund der Aggregiertheit der Daten sowie der Multidimensionalität in Data Warehouses mit besonderer Sorgfalt zu erfolgen - einerseits können auf Primärdaten beliebte, klassifizierende Algorithmen nicht auf verdichteten Daten angewendet werden (Assoziationsregeln, neuronale Netzwerke, etc.), andererseits birgt die Multidimensionalität die Gefahr, Veränderungen in einer Dimension auf andere Dimensionen überzuwälzen. Dennoch bieten sich eine Reihe von Data Mining-Techniken an, um Brüche in den Kennzahlen- und Strukturdimensionen aufzudecken.

Die Arbeit ist so aufgebaut, dass nach der nun folgenden Beschreibung der Problemstellung der Thematik in Kapitel 2 zunächst verschiedene Typen von Strukturbrüchen vorgestellt werden, darauf aufbauend werden mögliche einsetzbare Data-Mining-Methoden in Kapitel 3 an einfachen Beispielen illustriert; danach zeigt Kapitel 4 Grenzen der Verfahren auf, die teilweise überwunden werden können - es wird auch gezeigt, wie die zunächst anhand von zwei Dimensionen skizzierten Methoden auf Werten, die über n Dimensionen referenziert werden, appliziert werden können. Schließlich sind in Kapitel 5 die Ergebnisse der Anwendung der beschriebenen Verfahren an einem realen Data Warehouse festgehalten.

1.2 Problemstellung und Motivation

Zum gegenwärtigen Zeitpunkt sehen sich Data Warehouses und Online Analytical Processing (OLAP)-Systeme mit dem Problem, Schema- und Instanzänderungen zu erkennen sowie mit ihnen umzugehen, konfrontiert. Da diese Systeme jedoch gerade zu Analysezwecken konzipiert wurden, ist es von fundamentaler Bedeutung, Analysen über mehrere Schemaversionen durchführen zu können. Sollte es zwischen aufeinanderfolgenden Schemaversionen zu einer oder auch mehreren Strukturevolutionen gekommen sein, so sollte vor der eigentlichen Interpretation der Daten eine automatisierte Überprüfung indizieren, dass die Strukturen verändert wurden [EK02].

Als einleitende Überlegung kann folgendes Beispiel dienen: bis zum Jahr 1993 existierte für den Studiengang Informatik an der Universität Klagenfurt lediglich ein Institut für Informatik. Im darauffolgenden Jahr (Jahr 1994) wurde das Institut für Informatik (IFI) in drei verschiedene Institute (Institut für Wirtschaftsinformatik und Anwendungssysteme (IWAS), Institut für Informatiksysteme (ISYS), Institut für Informationstechnologie (ITEC)) aufgespalten. Will man im Jahr 1994 eine Analyse über die Entwicklung der Mitarbeiterzahlen oder über ein sonstiges im Data Warehouse gespeichertes Feature der Kennzahlendimension durchführen, so wird man für die Zeit vor 1994 überraschenderweise vom aktuellen Jahr deutlich abweichende Werte er-

halten, die auf ein Diskontinuum in der Struktur schließen lassen. Obgleich in diesem Fall jedem Beobachter aufgrund der Simplizität die Vermutung einer internen Umgliederung sofort in den Sinn kommen könnte (ein sich in Kenntnis der wahren Sachlage befindlicher User würde beispielsweise die Zahl der Mitarbeiter des IWAS vor 1994 als *Zahl der Mitarbeiter des IFI * 0.33* errechnen), wäre es auch bereits hier eine gedankliche Unterstützung, wenn ein automatisierter Mechanismus einen leisen Hinweis auf eine Umstrukturierung geben könnte.

Da im Falle von Strukturmodifikationen die Koeffizienten der Transformationsvektoren[2] bzw. Transformationsmatrizen nur manuell unter Zuhilfenahme domainspezifischen Wissens geschätzt bzw. experimentell erprobt werden können [EK02], sollte bei der Analyse von Daten über mehrere Chronone[3] zumindest eine automatische Erkennung solcher Änderungen möglich sein, um den Durchführenden der Analyse vor unsorgfältiger Interpretation der Ergebnisse zu warnen. Erkennungen derartiger Änderungen sollten Anlass sein, Nachforschungen bei Domain-Experten über die tatsächlichen Ursachen und Auswirkungen von Evolutionen einzelner Dimensionsdaten anzustellen, um fundiertere Aussagen über die Ursachen bestimmter Trends treffen zu können. Zentraler Gegenstand der vorliegenden Arbeit ist es daher, Analysemethoden zu untersuchen, die automatisiert Strukturbrüche aufdecken können. Sämtliche Techniken werden zunächst an einem sehr kleinen, konstruierten Beispiel vorgestellt, um in weiterer Folge an einem größeren Exempel die Eignung, Effizienz und Effektivität der Methoden für die Praxis zu überprüfen.

[2] Die Koeffizienten der Transformationsvektoren geben an, zu welchen Teilen alte Strukturen in neue Strukturen einfließen. Zwischen zwei aufeinanderfolgenden Strukturversionen SV_1 und SV_2 werden für jede Strukturdimension D_i und jede Kennzahl F Transformationsmatrizen $T_{SV_1,SV_2,D_i,F}$ erstellt, wobei $T(d_i, d_j)$ den Gewichtungsfaktor der Abbildung der Kennzahl F des dimension members d_i in Strukturversion SV_1 auf F des dimension members d_j in Strukturversion SV_2 repräsentiert [EK02].

[3] [Je98] definiert ein Chronon als jenes kleinste, nicht mehr weiter zerlegbare Zeitintervall eines Data Warehouses, in dessen Granularität Schemaveränderungen auftreten können.

2

Arten von Strukturbrüchen

In diesem Abschnitt sollen mögliche Strukturbrüche und deren Ursachen im Detail beschrieben werden. Auf komplizierte Strukturevolutionen - ein dimension member ist in mehreren verschiedenen Dimensionen enthalten - wird nicht eingegangen, da von einem 'optimalen' Design des Data Warehouses ausgegangen wird, in dem sämtliche dimension members aller Dimensionen zueinander orthogonal sind[1].

Im folgenden werden Strukturbrüche in der Kennzahlen- sowie in den Strukturdimensionen dargestellt, wobei stets ein Data Warehouse mit Verkaufsstatistikdaten verwendet wird, welches die Entwicklung der Kennzahl 'Gewinn' anhand der Strukturdimensionen Land und Produktgruppe im Jahresverlauf darstellt.

2.1 Strukturbrüche in der Kennzahlendimension

In der Kennzahlendimension werden folgende zwei Strukturbrüche unterschieden:

- Veränderung der Berechnungsformel einer Kennzahl
- Veränderung der Maßeinheit einer (mehrerer) Kennzahl(en)

2.1.1 Veränderung der Berechnungsformel einer Kennzahl

Ein Strukturbruch in der Kennzahlendimension liegt unter anderem vor, wenn sich die Rechenregel einer Kennzahl im Zeitablauf verändert.

[1] Beispielsweise werden die Umsätze verschiedener Produkte nicht nach den Strukturdimensionen Produktionsland und Verkaufsland analysiert - in beiden Dimensionen wären die Dimensionsinstanzen identisch -, sondern es wird für die Strukturdimension Land (und Produkte) der Umsatz nach den Kennzahlen 'produziert in' sowie 'verkauft in' ausgewertet.

Ein einfaches Beispiel: im Verkaufsstatistik-Data Warehouse wird in $Jahr_t$ erkannt, dass der Gewinn sämtlicher Produktgruppen sämtlicher Länder nicht mehr 20% des Umsatzes, sondern nur mehr 10% dessen beträgt. Tabelle 2.1 zeigt auf, dass der Gewinn sämtlicher Kombinationen von Produktgruppe und Land in $Jahr_t$ deutlich unter jenem der Vorjahre liegt.

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	**106**
$Land_1$	PG_B	155	156	158	159	**81**
$Land_1$	PG_C	689	692	694	701	**353**
$Land_2$	PG_A	16	17	18	19	**10**
$Land_2$	PG_B	13	14	15	16	**9**
$Land_2$	PG_C	86	87	88	89	**46**
$Land_3$	PG_A	353	356	363	366	**184**
$Land_3$	PG_B	310	311	315	318	**160**
$Land_3$	PG_C	989	993	995	997	**500**

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.1: Änderung der Berechnungsvorschrift des Gewinns

Falls die Veränderung deutlich über der natürlichen Varianz der Werte liegt, so ist im Normalfall eine automatisierte Aufdeckung des Strukturbruchs möglich, wie in Kapitel 3 gezeigt wird.

2.1.2 Veränderung der Maßeinheit einer (mehrerer) Kennzahl(en)

Ein ähnlicher Strukturbruch liegt bei der Veränderung der Maßeinheit der Kennzahl vor: bis zu einem bestimmten Zeitpunkt wurde die Kennzahl in der Maßeinheit m_1 angegeben, danach in einer von m_1 verschiedenen Einheit m_2.

Zur Illustration dient wieder das Verkaufsstatistik-Data Warehouse: der Gewinn wird ab einem bestimmten Chronon t nicht mehr in ATS, sondern in EURO angegeben. Im Gegensatz zum vorangegangenen Kapitel hat sich somit von $Jahr_{t-1}$ auf $Jahr_t$ keine Kalkulationsvorschrift modifiziert - der Gewinn beträgt in beiden Jahren 20% des Umsatzes -, es wurde jedoch die Währungseinheit verändert.

In Kapitel 3 werden Methoden vorgestellt, um beide Typen von Strukturbrüchen in der Kennzahlendimension (Veränderung der Berechnungsformel einer Kennzahl sowie Veränderung der Maßeinheit einer Kennzahl) erkennen zu können. Während es in diesem Kapitel belanglos ist, ob die Kennzahlen additiv, semi-additiv oder nicht-additiv sind[2], da keine Hierarchien vor-

[2]Additive Kennzahlen können entlang jeder Dimensionshierarchie summiert werden (z.B. Gewinn, Umsatz, Kosten, etc.), semi-additive Kennzahlen sind zwar ebenso numerische Werte, sie können aber nicht einfach durch Summation aggregiert werden (z.B. der aktuelle Lagerstand einer Abteilung kann nicht durch einfache Summation der relativen Füllhöhen der Lager von Unterabteilungen berechnet werden; die Werte sind aber dennoch

Gew	ATS/EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	**15**
$Land_1$	PG_B	155	156	158	159	**12**
$Land_1$	PG_C	689	692	694	701	**51**
$Land_2$	PG_A	16	17	18	19	**1**
$Land_2$	PG_B	13	14	15	16	**1**
$Land_2$	PG_C	86	87	88	89	**7**
$Land_3$	PG_A	353	356	363	366	**27**
$Land_3$	PG_B	310	311	315	318	**23**
$Land_3$	PG_C	989	993	995	997	**73**

Legende: PG=Produktgruppe, Gew=Gewinn in Schilling bzw. EURO

Tabelle 2.2: Änderung der Währungseinheit des Gewinns

herrschen, wird in Kapitel 4.4 und in Kapitel 5.2.2 von additiven Kennzahlen ausgegangen, da die Werte entlang einer Strukturdimension gruppiert werden.

2.2 Strukturbrüche in den Strukturdimensionen

Auch in den Strukturdimensionen kann es im Laufe der Zeit zu Veränderungen der einzelnen dimension members kommen, die nur zum Teil von Laderoutinen eines Data Warehouses erkannt werden können. Folgende Strukturbrüche werden dabei näher betrachtet:

- Aufnahme eines neuen dimension members (NEW)
- Löschung eines bisherigen dimension members (DELETE)
- Veränderung der Bedeutung eines dimension members (UPDATE)
- Änderungen in der Zugehörigkeitshierarchie einer Strukturdimension (MOVE)
- Verschmelzung zweier (mehrerer) dimension members (MERGE)
- Aufspaltung eines dimension members (SPLIT)

numerisch, sodass durch eine bestimmte Durchschnittsfunktion auch der Lagerstand der übergeordneten Ebene ermittelt werden kann.), nicht-additive Kennzahlen hingegen können überhaupt nicht entlang der Dimensionshierarchie aggregiert werden (z.B. der Zufriedenheitsgrad der Mitarbeiter in der Abteilung kann nicht durch eine Durchschnittsberechnung aller Unterabteilungen errechnet werden, da in den möglichen Ausprägungen des Zufriedenheitsgrads maximal eine Reihenfolge, aber keine Abstände zwischen einzelnen Werten sinnvoll ermittelt werden können.) [Ki97].

2.2.1 Aufnahme eines neuen dimension members

Ein Strukturbruch in der Strukturdimension herrscht unter anderem dann vor, wenn zu einem bestimmten Zeitpunkt t eine zusätzliche, bislang noch nicht vorhandene Instanz einer Strukturdimension aufgenommen wird.

Das naheliegendste Beispiel im Verkaufsstatistik-Data Warehouse ist die Erweiterung des Produktprogrammes des Unternehmens und somit die Aufnahme einer neuen Produktgruppe $Produktgruppe_D$ in $Jahr_t$ in das Data Warehouse, wie Tabelle 2.3 aufzeigt.

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	211
$Land_1$	PG_B	155	156	158	159	161
$Land_1$	PG_C	689	692	694	701	705
$Land_2$	PG_A	16	17	18	19	20
$Land_2$	PG_B	13	14	15	16	17
$Land_2$	PG_C	86	87	88	89	92
$Land_3$	PG_A	353	356	363	366	368
$Land_3$	PG_B	310	311	315	318	319
$Land_3$	PG_C	989	993	995	997	999
$Land_1$	PG_D	-	-	-	-	**47**
$Land_2$	PG_D	-	-	-	-	**11**
$Land_3$	PG_D	-	-	-	-	**92**

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.3: Aufnahme einer zusätzlichen Produktgruppe

Da allerdings während des Ladens der Daten aus operativen Vorsystemen in das Data Warehouse der Versuch, Werte nicht vorhandener dimension members einzutragen, sofort als Fehler erkannt wird, bedarf es zur Erkennung dieses Strukturbruchs keiner weiteren Data Mining-Techniken, und dieser Bruch wird in den folgenden Kapiteln daher nicht mehr weiter behandelt.

2.2.2 Löschung eines bisherigen dimension members

Komplementär zum im vorangegangenen Abschnitt dargestellten Hinzufügen eines neuen dimension members ist das Löschen eines bisherigen dimension members.

Im Data Warehouse über Verkaufsstatistiken wird eine Produktgruppe wegen Veränderungen des Verbraucherverhaltens nun nicht mehr angeboten, die ehemals existierende Produktgruppe PG_C wird aus dem Produktionsprogramm eliminiert, wie Tabelle 2.4 dokumentiert.

Auch hier fällt allerdings bereits während der Laderoutine auf, dass die Daten nicht das korrekte Format haben - in der Strukturdimension *Produktgruppe* bleiben Werte für Produktgruppe PG_C aus, eine gesonderte Analyse dieses Strukturbruchs ist im folgenden daher auch

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	211
$Land_1$	PG_B	155	156	158	159	161
$Land_1$	PG_C	689	692	694	701	-
$Land_2$	PG_A	16	17	18	19	20
$Land_2$	PG_B	13	14	15	16	17
$Land_2$	PG_C	86	87	88	89	-
$Land_3$	PG_A	353	356	363	366	368
$Land_3$	PG_B	310	311	315	318	319
$Land_3$	PG_C	989	993	995	997	-

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.4: Elimination von Produktgruppe PG_C aus dem Produktionsprogramm

nicht mehr notwendig.

2.2.3 Veränderung der Bedeutung eines dimension members

Ein schwierig zu erkennender Strukturbruch liegt dann vor, wenn ein bisheriger dimension member unter der gleichen Bezeichnung in einer späteren Schemaversion eine andere Bedeutung hat. In diesem Fall können Laderoutinen keinen Fehler erkennen, da das Datenformat der verlangten Struktur exakt entspricht.

Im Data Warehouse über Verkaufsstatistiken liegt ein Bruch dieser Art beispielsweise vor, wenn ab $Jahr_t$ $Land_1$ eine andere Bedeutung hat; wegen eines Handelsembargos dürfen in das ursprüngliche Land keinerlei Waren mehr exportiert werden, andererseits wurden jedoch in einem Land, das bislang seine inländischen Erzeugnisse durch rigorose Importrestriktionen vor ausländischen Produkten geschützt hat, vollständige Handelsliberalisierungsmaßnahmen erlassen - $Land_1$ ist ab $Jahr_t$ der Schlüssel des neuen Landes. Tabelle 2.5 zeigt auf, dass in sämtlichen Produktgruppen von $Land_1$ der Gewinn in $Jahr_t$ um nahezu das Vierfache gegenüber vorangegangenen Jahren angestiegen ist; in den anderen Ländern hingegen setzte sich der mehr oder weniger stationäre Trend der Vorjahre fort.

In Kapitel 3 wird anhand verschiedener Methoden versucht werden, diesen Strukturbruch sowie die im folgenden aufgeführten Strukturbrüche in Strukturdimensionen aufzudecken; es muss aber festgehalten werden, dass nicht immer ein Aufdecken eines solchen Bruchs möglich sein wird - gibt es in dem neuen Land aus Tabelle 2.5 beispielsweise ähnlich viele Einwohner wie im ursprünglichen Land und ist die Nachfrage nach Produkten von ähnlicher Gestalt, so könnten die Unterschiede in den einzelnen Werten möglicherweise zu gering sein, um sie als Strukturbrüche klassifizieren zu können.

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	**808**
$Land_1$	PG_B	155	156	158	159	**612**
$Land_1$	PG_C	689	692	694	701	**2798**
$Land_2$	PG_A	16	17	18	19	20
$Land_2$	PG_B	13	14	15	16	17
$Land_2$	PG_C	86	87	88	89	92
$Land_3$	PG_A	353	356	363	366	368
$Land_3$	PG_B	310	311	315	318	319
$Land_3$	PG_C	989	993	995	997	999

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.5: Änderung der Bedeutung des Schlüssels eines Landes

2.2.4 Änderungen in der Zugehörigkeitshierarchie einer Strukturdimension

Ein Strukturbruch, der ebenfalls nicht automatisch während des Ladens der Daten von operativen Vorsystemen in das Data Warehouse erkannt wird, ist die Änderung in der Zugehörigkeitshierarchie einer Strukturdimension. Operative Vorsysteme liefern oft nicht die unterste, detaillierteste Stufe der Daten weiter, sondern bereits eine teilweise aggregierte Version; falls in der Zuordnung der untersten Stufe zur darüberliegenden Aggregationsstufe eine Änderung gemacht wird, ist dies für die auf den aggregierten Daten operierende Laderoutine nicht erkennbar.

Im Verkaufsstatistik-Data Warehouse wäre ein solcher 'MOVE' in der Hierarchie beispielsweise dann vorhanden, wenn Produktgruppe PG_A ab dem Zeitpunkt t nicht mehr die Produkte P_{A1}, P_{A2} und P_{A3}, sondern die Produkte P_{C1}, P_{A2} und P_{A3} umfasst (vice versa die Änderung in Produktgruppe PG_C), damit in Produktgruppe PG_C alle umsatz- und gewinnträchtigen Produkte zusammengefasst sind, wogegen Produktgruppe PG_A alle Produkte mit geringem Umsatz und geringem Rohertrag erfassen soll. Tabelle 2.6 illustriert, dass in sämtlichen Ländern die Produktgruppe PG_C in $Jahr_t$ bezüglich des Gewinns deutlich zugelegt hat, wogegen PG_A in $Jahr_t$ überall deutliche Gewinneinbußen zu verzeichnen hat.

2.2.5 Verschmelzung zweier (mehrerer) dimension members

Ein Strukturbruch, der nur zum Teil bei der Überspielung der Daten aus operativen Vorsystemen erkannt wird, ist die Vereinigung zweier dimension members. Es wird zwar erkannt, dass eine vormals mit Datenwerten befüllte Dimensionsinstanz nun fehlende Werte aufweist, es wird allerdings nicht erkannt, dass eine andere Dimensionsinstanz nun nicht mehr die ursprünglichen Werte, sondern die Summe der Werte der beiden vereinigten dimension members enthält.

Im Verkaufsstatistik-Data Warehouse kann es beispielsweise vorkommen, dass $Land_1$ und $Land_3$ in $Jahr_t$ vereinigt werden; die Kennzahlenwerte des vereinigten Landes sind ab $Jahr_t$

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	**111**
$Land_1$	PG_B	155	156	158	159	161
$Land_1$	PG_C	689	692	694	701	**805**
$Land_2$	PG_A	16	17	18	19	**10**
$Land_2$	PG_B	13	14	15	16	17
$Land_2$	PG_C	86	87	88	89	**102**
$Land_3$	PG_A	353	356	363	366	**198**
$Land_3$	PG_B	310	311	315	318	319
$Land_3$	PG_C	989	993	995	997	**1169**

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.6: Änderung der Zugehörigkeit von Produkten zu Produktgruppen

gesammelt in $Land_1$ ausgewiesen, in $Land_3$ fehlen ab $Jahr_t$ die Werte, wie Tabelle 2.7 darstellt.

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	**579**
$Land_1$	PG_B	155	156	158	159	**480**
$Land_1$	PG_C	689	692	694	701	**1704**
$Land_2$	PG_A	16	17	18	19	20
$Land_2$	PG_B	13	14	15	16	17
$Land_2$	PG_C	86	87	88	89	92
$Land_3$	PG_A	353	356	363	366	**-**
$Land_3$	PG_B	310	311	315	318	**-**
$Land_3$	PG_C	989	993	995	997	**-**

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.7: Vereinigung zweier Länder

Sind an der Vereinigung mehr als zwei Dimensionsinstanzen beteiligt, so müssten in mehreren Ländern ab dem Vereinigungszeitpunkt Werte fehlen, was bei der Überspielung der Daten aus den Vorsystemen allerdings sofort auffallen würde. Es gäbe also wieder nur ein Land, dessen Diskontinuität der Werte nicht automatisch erkennt werden würde.

2.2.6 Aufspaltung eines dimension members

Ähnlich wie mit der Vereinigung verhält es sich mit der Aufspaltung eines dimension members in zwei (mehrere) dimension members: Laderoutinen bemerken zwar, dass eine bislang nicht vorhandene Dimensionsinstanz mit Werten befüllt werden soll, es wird aber nicht erkannt, dass die Werte des ursprünglichen dimension members um die nun im neu hinzugefügten dimension member ausgewiesenen Werte reduziert wurden.

Das Pendantbeispiel zur Vereinigung zweier Länder ist somit die Aufspaltung eines Landes in

zwei Länder. Tabelle 2.8 exemplifiziert die Auswirkungen der Aufspaltung von $Land_3$ in $Jahr_t$ in $Land_3$ und $Land_4$ - die Gewinne in $Land_3$ gehen unverhältnismäßig stark zurück, wogegen in den anderen Ländern der in den Vorjahren zu verzeichnende nahezu stationäre Trend der Entwicklung der Gewinne anhält; andererseits werden aber nun erstmalig Werte für $Land_4$ ausgewiesen.

Gew	EUR	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	206	208	211
$Land_1$	PG_B	155	156	158	159	161
$Land_1$	PG_C	689	692	694	701	705
$Land_2$	PG_A	16	17	18	19	20
$Land_2$	PG_B	13	14	15	16	17
$Land_2$	PG_C	86	87	88	89	92
$Land_3$	PG_A	353	356	363	366	**266**
$Land_3$	PG_B	310	311	315	318	**238**
$Land_3$	PG_C	989	993	995	997	**721**
$Land_4$	PG_A	-	-	-	-	**102**
$Land_4$	PG_B	-	-	-	-	**81**
$Land_4$	PG_C	-	-	-	-	**278**

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 2.8: Aufspaltung eines Landes in zwei Länder

Spaltet sich ein Land in mehr als zwei Länder auf, so ändert sich auch hier aus Data Mining-Perspektive die Problemstellung nicht: sämtliche neu hinzugekommenen Staaten waren bislang nicht vorhanden, es wird beim Füllen des Data Warehouses sogleich erkannt, dass Werte nicht vorhandener dimension members eingetragen werden. Lediglich die Tatsache, dass das ursprünglich vereinigte Land nun wesentlich kleiner ist, wird nicht automatisch erfasst.

2.2.7 Kombinationen verschiedenster Ursachen

Die bislang aufgeführten Strukturbrüche müssen nicht voneinander vollständig isoliert auftreten, sie können in beliebiger Kombination zusammenfallen. Dabei können auch Strukturbrüche, die bislang wenigstens teilweise als solche erkannt wurden, möglicherweise nicht mehr erkannt werden:

- Gleichzeitig wird ein dimension member gelöscht, und ein neuer kommt hinzu, welcher die Bezeichnung des gelöschten dimension members unverändert übernimmt - zwei Strukturbrüche, die getrennt voneinander sofort aufgefallen wären, können nun nicht mehr automatisch identifiziert werden.

- Gleichzeitig werden zwei dimension members miteinander vereinigt, eine andere Dimensionsinstanz spaltet sich jedoch in zwei dimension members auf, wobei der aus der Spaltung neu hinzugekommene dimension member die Bezeichnung des durch die Vereinigung aufgelösten dimension members exakt übernimmt. Wären die beiden Strukturbrüche getrennt aufgetreten, so wären wenigstens vorherige bzw. nunmehrige Lücken in den Daten durch die Aufspaltung bzw. Vereinigung erkannt worden.

- Die Bezeichnung eines dimension members bleibt zwischen zwei Schemaversionen gleich, obwohl sich die Bedeutung dieses ändert, der neue dimension member wird in der Aggregationshierarchie anders zugeordnet. Diese beiden Brüche wären getrennt voneinander durch Laderoutinen nicht erkannt worden, und sie fallen auch nicht auf, wenn sie in Kombination miteinander auftreten.

Dies ist natürlich nur eine kleine Auswahl der zahlreichen Möglichkeiten von Kombinationen verschiedenster Strukturbrüche - sie zeigt allerdings auf, dass bei gleichzeitigem Auftreten mehrerer Brüche niemals Information über einzelne Brüche dazugewonnen werden kann, vielmals jedoch zumindest teilweise Information verloren geht.

3

Aufdecken von Strukturbrüchen durch Data Mining-Techniken

In diesem Abschnitt sollen verschiedene Data Mining-Techniken vorgestellt werden, die Strukturbrüche in Data Warehouses aufdecken können; jede Methode wird dabei an Daten (z.T. aus obigen Tabellen) erprobt, an denen die Vorzüge der Methode gut zum Ausdruck kommen. Im einzelnen werden folgende Methoden skizziert:

- Einfache Abweichungsmatrizen
- Bivariate Kreuzkorrelation
- Lineare Regression
- Autokorrelation
- Autoregression
- Differenzialgleichungen
- Eigen- und Singulärwertzerlegung
- Trigonometrische Transformationen
- Diskrete Wavelet-Transformationen

3.1 Einfache Abweichungsmatrizen

Die in diesem Kapitel vorgestellte Methode stellt keine Data Mining-Technik im engeren Sinn dar, sie ist allerdings sehr gut geeignet, grobe Strukturbrüche einer Datenmatrix $A \in \mathbb{R}^{m \times n}$ auf

einfache Weise in $O(m + n)$ aufzudecken. Diese Methode berechnet für die zu untersuchende Datenmatrix vier Tabellen mit den absoluten und relativen Differenzen, die erste Aufschlüsse über Schemaänderungen liefern können:

- Matrix mit absoluten Differenzbeträgen der Werte aller Strukturkombinationen zwischen aufeinanderfolgenden Chrononen (Matrix M_1 in den Tabellen unten)
- Matrix mit prozentualen Änderungen der Werte aller Strukturkombinationen zwischen aufeinanderfolgenden Chrononen (Matrix M_2)
- Matrix mit den prozentuellen Anteilen aller Strukturkombinationen für jedes Chronon (Matrix M_3)[1]
- Matrix mit den Veränderungen der Anteile aller Strukturkominationen zwischen aufeinanderfolgenden Chrononen (Matrix M_4)

Die resultierenden Matrizen sind mit Ausnahme der dritten Matrix, die in den Dimensionen mit der Inputmatrix übereinstimmt, somit Matrizen $\in \mathbb{R}^{m \times n-1}$. Tritt ein Strukturbruch auf, der sämtliche Strukturkombinationen umfasst (beispielsweise Veränderung der Berechnungsvorschrift einer Kennzahl oder Änderung der Maßeinheit einer Kennzahl), so sollten vor allem die ersten beiden Matrizen grobe Abweichungen aufweisen, bei einem Strukturbruch, der nur einen dimension member oder nur einen Teil der dimension members betrifft (beispielsweise ein Update oder Merge von Strukturdaten), kommen die Unterschiede am deutlichsten in der dritten und vierten Matrix zum Ausdruck.

Untersucht man die Werte in Tabelle 2.2 (Änderung der Währungseinheit von ATS auf EURO) mit dieser Methode[2], so erhält man die in den Tabellen 3.1 bis 3.4 angeführten Ergebnisse.

Es zeigt sich bei Betrachtung der vier Matrizen die für eine Änderung der Maßeinheit typische Situation: die Matrizen M_1 und M_2 weisen deutliche Ausreißer in der letzten Spalte auf; M_2 betont zudem, dass die Veränderung bei allen Werten von ungefähr gleichem Ausmaß war (zwischen 90% und 95%). Die Matrizen M_3 und M_4 können keine besonderen Ausreißer feststellen - die Anteile einzelner Werte am Gesamtkuchen sind in allen Jahren nahezu unverändert geblieben.

Anders präsentiert sich die Situation bei einer Veränderung, die nur einen Wert bzw. einen Teil der Werte betrifft. Dazu wird das in Tabelle 2.5 dargestellte 'UPDATE' eines dimension

[1] In großen Data Warehouses wird diese Matrix alleine aufgrund der geringen Anteile einzelner Werte kaum Aussagekraft haben. Sie ist aber wichtig, da sie als Vorstufe zur Berechnung von Matrix M_4 dient.

[2] Die Methode der Abweichungsmatrix wird wie alle anderen Methoden in diesem Kapitel so verwendet, als hätte die Datenmatrix nur zwei Dimensionen (eine Struktur- und eine Zeitdimension), obwohl insgesamt drei Dimensionen vorliegen und hier eine andere Vorgehensweise zu wählen ist (siehe Kapitel 4.3 weiter unten). In diesem Kapitel sollen jedoch primär die Methoden illustriert werden, die Brüche aufdecken können; da die Daten dieses Abschnitts mit 'präparierten' Brüchen versehen sind, ist die zweidimensionale Betrachtungsweise möglich.

	AbsDiff	$Jahr_{t-4;t-3}$	$Jahr_{t-3;t-2}$	$Jahr_{t-2;t-1}$	$Jahr_{t-1;t}$
$Land_1$	PG_A	3	1	2	**-193**
$Land_1$	PG_B	1	2	1	**-147**
$Land_1$	PG_C	3	2	7	**-650**
$Land_2$	PG_A	1	1	1	**-18**
$Land_2$	PG_B	1	1	1	**-15**
$Land_2$	PG_C	1	1	1	**-82**
$Land_3$	PG_A	3	7	3	**-339**
$Land_3$	PG_B	1	4	3	**-295**
$Land_3$	PG_C	4	2	2	**-924**

Legende: PG=Produktgruppe, AbsDiff=Absolutbetrag der Differenzen der Gewinne, $Jahr_{s;t}$=Vergleich von $Jahr_s$ mit $Jahr_t$

Tabelle 3.1: Matrix M_1 bei Veränderung der Maßeinheit einer Kennzahl

	RelDiff	$Jahr_{t-4;t-3}$	$Jahr_{t-3;t-2}$	$Jahr_{t-2;t-1}$	$Jahr_{t-1;t}$
$Land_1$	PG_A	1.49%	0.49%	0.97%	**-92.79%**
$Land_1$	PG_B	0.65%	1.28%	0.63%	**-92.45%**
$Land_1$	PG_C	0.44%	0.29%	1.01%	**-92.72%**
$Land_2$	PG_A	6.25%	5.88%	5.56%	**-94.74%**
$Land_2$	PG_B	7.69%	7.14%	6.67%	**-93.75%**
$Land_2$	PG_C	1.16%	1.15%	1.14%	**-92.13%**
$Land_3$	PG_A	0.85%	1.97%	0.83%	**-92.62%**
$Land_3$	PG_B	0.32%	1.29%	0.95%	**-92.77%**
$Land_3$	PG_C	0.4%	0.2%	0.2%	**-92.68%**

Legende: PG=Produktgruppe, RelDiff=Veränderung der Gewinne in %, $Jahr_{s;t}$=Vergleich von $Jahr_s$ mit $Jahr_t$, prozentuelle Veränderung auf Basis von $Jahr_s$

Tabelle 3.2: Matrix M_2 bei Veränderung der Maßeinheit einer Kennzahl

	Anteile	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	7.18%	7.24%	7.22%	7.24%	7.14%
$Land_1$	PG_B	5.51%	5.51%	5.54%	5.53%	5.71%
$Land_1$	PG_C	24.49%	24.44%	24.33%	24.40%	24.29%
$Land_2$	PG_A	0.57%	0.60%	0.63%	0.66%	0.48%
$Land_2$	PG_B	0.46%	0.49%	0.53%	0.56%	0.48%
$Land_2$	PG_C	3.06%	3.07%	3.09%	3.10%	3.33%
$Land_3$	PG_A	12.55%	12.58%	12.73%	12.74%	12.86%
$Land_3$	PG_B	11.02%	10.99%	11.04%	11.07%	10.95%
$Land_3$	PG_C	35.16%	35.08%	34.89%	34.70%	34.76%

Legende: PG=Produktgruppe, Anteile=Anteile einzelner Kombinationen von dimension members am Gesamtkuchen eines Jahres

Tabelle 3.3: Matrix M_3 bei Veränderung der Maßeinheit einer Kennzahl

	AnteileDiff	$Jahr_{t-4;t-3}$	$Jahr_{t-3;t-2}$	$Jahr_{t-2;t-1}$	$Jahr_{t-1;t}$
$Land_1$	PG_A	0.06%	-0.02%	0.02%	-0.10%
$Land_1$	PG_B	0.00%	0.03%	-0.01%	0.18%
$Land_1$	PG_C	-0.05%	-0.11%	0.07%	-0.11%
$Land_2$	PG_A	0.03%	0.03%	0.03%	-0.19%
$Land_2$	PG_B	0.03%	0.03%	0.03%	-0.08%
$Land_2$	PG_C	0.02%	0.01%	0.01%	0.24%
$Land_3$	PG_A	0.03%	0.15%	0.01%	0.12%
$Land_3$	PG_B	-0.03%	0.06%	0.02%	-0.12%
$Land_3$	PG_C	-0.08%	-0.19%	-0.19%	0.06%

Legende: PG=Produktgruppe, AnteileDiff=Absolute Veränderung der Anteile einzelner Kombinationen von dimension members am Gesamtkuchen, $Jahr_{s;t}$=Vergleich der Anteile zwischen $Jahr_s$ und $Jahr_t$

Tabelle 3.4: Matrix M_4 bei Veränderung der Maßeinheit einer Kennzahl

members herangezogen, die resultierenden vier Matrizen sind in den Tabellen 3.5 bis 3.8 aufgeführt. In diesem Fall liefern alle Ergebnismatrizen Aufschlüsse über Ausreißer: in den Ma-

	AbsDiff	$Jahr_{t-4;t-3}$	$Jahr_{t-3;t-2}$	$Jahr_{t-2;t-1}$	$Jahr_{t-1;t}$
$Land_1$	PG_A	3	1	2	**600**
$Land_1$	PG_B	1	2	1	**453**
$Land_1$	PG_C	3	2	7	**2097**
$Land_2$	PG_A	1	1	1	1
$Land_2$	PG_B	1	1	1	1
$Land_2$	PG_C	1	1	1	3
$Land_3$	PG_A	3	7	3	2
$Land_3$	PG_B	1	4	3	1
$Land_3$	PG_C	4	2	2	2

Legende: PG=Produktgruppe, AbsDiff=Absolutbetrag der Differenzen der Gewinne, $Jahr_{s;t}$=Vergleich von $Jahr_s$ mit $Jahr_t$

Tabelle 3.5: Matrix M_1 bei Veränderung der Bedeutung eines dimension members

trizen M_1 und M_2 fallen in den ersten drei Zeilen in der letzten Kolumne die Werte so deutlich aus der Reihe, dass zweifelsohne auf einen Strukturbruch geschlossen werden kann. Untersucht man die Matrix M_3, so sieht man, dass sich die Anteile der ersten drei Zeilen nahezu verdoppelt haben. M_4 streicht zusätzlich heraus, dass alle anderen Zeilen relative Anteile am Gesamtkuchen an die Strukturkombinationen aus den ersten drei Zeilen verloren haben - in der letzten Kolumne sind lediglich die Werte der geänderten dimension members positiv, alle anderen sind negativ.

Diese Methode eignet sich aus Performance- und Simplizitätsgründen als 'erstes grobes Sieb', um extreme Ausreißer identifizieren zu können. Feinere Unterschiede sollten jedoch eher mit mathematisch fundierteren, in den folgenden Kapiteln präsentierten Verfahren analysiert werden.

	RelDiff	$Jahr_{t-4;t-3}$	$Jahr_{t-3;t-2}$	$Jahr_{t-2;t-1}$	$Jahr_{t-1;t}$
$Land_1$	PG_A	1.49%	0.49%	0.97%	**288.46%**
$Land_1$	PG_B	0.65%	1.28%	0.63%	**284.91%**
$Land_1$	PG_C	0.44%	0.29%	1.01%	**299.14%**
$Land_2$	PG_A	6.25%	5.88%	5.56%	5.26%
$Land_2$	PG_B	7.69%	7.14%	6.67%	6.25%
$Land_2$	PG_C	1.16%	1.15%	1.14%	3.37%
$Land_3$	PG_A	0.85%	1.97%	0.83%	0.55%
$Land_3$	PG_B	0.32%	1.29%	0.95%	0.31%
$Land_3$	PG_C	0.40%	0.20%	0.20%	0.20%

Legende: PG=Produktgruppe, RelDiff=Veränderung der Gewinne in %, $Jahr_{s;t}$=Vergleich von $Jahr_s$ mit $Jahr_t$, prozentuelle Veränderung auf Basis von $Jahr_s$

Tabelle 3.6: Matrix M_2 bei Veränderung der Bedeutung eines dimension members

	Anteile	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	7.18%	7.24%	7.22%	7.24%	**13.39%**
$Land_1$	PG_B	5.51%	5.51%	5.54%	5.53%	**10.14%**
$Land_1$	PG_C	24.49%	24.44%	24.33%	24.40%	**46.38%**
$Land_2$	PG_A	0.57%	0.60%	0.63%	0.66%	**0.33%**
$Land_2$	PG_B	0.46%	0.49%	0.53%	0.56%	**0.28%**
$Land_2$	PG_C	3.06%	3.07%	3.09%	3.10%	**1.52%**
$Land_3$	PG_A	12.55%	12.58%	12.73%	12.74%	**6.10%**
$Land_3$	PG_B	11.02%	10.99%	11.04%	11.07%	**5.29%**
$Land_3$	PG_C	35.16%	35.08%	34.89%	34.70%	**16.56%**

Legende: PG=Produktgruppe, Anteile=Anteile einzelner Kombinationen von dimension members am Gesamtkuchen eines Jahres

Tabelle 3.7: Matrix M_3 bei Veränderung der Bedeutung eines dimension members

3.2 Bivariate Kreuzkorrelation

Die bivariate Kreuzkorrelation misst die Stärke der Abhängigkeit zweier Wertereihen im Zeitverlauf. Die am häufigsten verwendete Metrik dieser Methode ist der *Pearson'sche Korrelationskoeffizient* r, der wie folgt definiert ist (vgl. u.a. [KP99]):

$$r = \frac{(\frac{1}{N-1})\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{(\frac{1}{N-1})\sum_{i=1}^{N}(x_i-\mu_x)^2}\sqrt{(\frac{1}{N-1})\sum_{i=1}^{N}(y_i-\mu_y)^2}} = \frac{E[(x_t-\mu_x)(y_t-\mu_y)]}{\sqrt{E[(x_t-\mu_x)^2]E[(y_t-\mu_y)^2]}}, \quad (3.1)$$

wobei x_i und y_i die Werte der Vektoren $\vec{x} = (x_1, x_2, \ldots, x_N)^T$ bzw. $\vec{y} = (y_1, y_2, \ldots, y_N)^T$ zum Zeitpunkt i signalisieren; die arithmetischen Mittel der Werte der Vektoren sind durch μ_x respektive μ_y dargestellt. Der zweite Teil der Gleichung streicht heraus, dass der Korrelationskoeffizient

	AnteileDiff	$Jahr_{t-4;t-3}$	$Jahr_{t-3;t-2}$	$Jahr_{t-2;t-1}$	$Jahr_{t-1;t}$
$Land_1$	PG_A	0.06%	-0.02%	0.02%	**6.15%**
$Land_1$	PG_B	0.00%	0.03%	-0.01%	**4.61%**
$Land_1$	PG_C	-0.05%	-0.11%	0.07%	**21.98%**
$Land_2$	PG_A	0.03%	0.03%	0.03%	**-0.33%**
$Land_2$	PG_B	0.03%	0.03%	0.03%	**-0.28%**
$Land_2$	PG_C	0.02%	0.01%	0.01%	**-1.57%**
$Land_3$	PG_A	0.03%	0.15%	0.01%	**-6.64%**
$Land_3$	PG_B	-0.03%	0.06%	0.02%	**-5.78%**
$Land_3$	PG_C	-0.08%	-0.19%	-0.19%	**-18.14%**

Legende: PG=Produktgruppe, AnteileDiff=Absolute Veränderung der Anteile einzelner Kombinationen von dimension members am Gesamtkuchen, $Jahr_{s;t}$=Vergleich der Anteile zwischen $Jahr_s$ und $Jahr_t$

Tabelle 3.8: Matrix M_4 bei Veränderung der Bedeutung eines dimension members

dem erwarteten Produkt der Abweichungen der beiden Variablen von ihrem jeweiligen Mittelwert entspricht; eine Veränderung der Skalierung der Variablen hat somit keine Auswirkung auf den Korrelationskoeffizienten. Der Wertebereich von r in (3.1) liegt im Intervall [-1;1], wobei 1 maximale gleichläufige Abhängigkeit der Variablen, -1 maximale gegenläufige Abhängigkeit der Variablen - bei einer Erhöhung der Werte von x fallen die Werte von y im selben Ausmaß - und 0 absolute Unkorreliertheit, absolute Unabhängigkeit der beiden Variablen dokumentiert. Die Werte innerhalb der Intervallgrenzen sind sorgfältig zu interpretieren, gute Statistikprogramme geben im Normalfall aber Auskunft darüber, ob auf 95% bzw. 99% Signifikanzniveau (Sig. 2-tailed (zweiseitig) in folgenden Abbildungen) eine auffällige Korrelation der Werte der beiden Variablen gegeben ist. Um die Brücke zur Terminologie der vorherigen Kapitel zu schlagen, kann x_i auch als die Ausprägung einer bestimmten Kennzahl einer konkreten Kombination von dimension members bzw. eines konkreten dimension members x in der Periode i verstanden werden.

Die Methode der Kreuzkorrelation eignet sich vor allem zum Aufdecken von Strukturbrüchen, die nur einen Teil der dimension members betreffen; aus diesem Grund soll die Methode zunächst zur Erkennung der Änderung eines dimension members an den Daten aus Tabelle 2.5 erprobt werden. Aus Abbildung 3.1 wird ersichtlich, dass die Kreuzkorrelationskoeffizienten, an denen $Land_1$ mit einem anderen Land beteiligt ist, wesentlich geringer sind als die anderen. Da alle Korrelationskoeffizienten von $Land_1$ mit entweder $Land_2$ oder $Land_3$ unter dem geforderten Signifikanzniveau von 95%, alle anderen Korrelationskoeffizienten jedoch teilweise deutlich über diesem Level liegen, ist dies ein deutliches Indiz für einen Strukturbruch bei $Land_1$.

Die Kreuzkorrelationsanalyse zeigt auch bei Änderungen in der Zugehörigkeitshierarchie (Werte aus Tabelle 2.6) Auffälligkeiten, wie Abbildung 3.2 darstellt: Gewinne gleicher Produkt-

Correlations

		L1PGA	L1PGB	L1PGC	L2PGA	L2PGB	L2PGC	L3PGA	L3PGB	L3PGC
L1PGA	Pearson Correlation	1	1,000**	1,000**	,713	,713	,878	,595	,615	,645
	Sig. (2-tailed)	,	,000	,000	,177	,177	,050	,290	,270	,239
	N	5	5	5	5	5	5	5	5	5
L1PGB	Pearson Correlation	1,000**	1	1,000**	,713	,713	,878	,595	,615	,645
	Sig. (2-tailed)	,000	,	,000	,177	,177	,050	,290	,269	,240
	N	5	5	5	5	5	5	5	5	5
L1PGC	Pearson Correlation	1,000**	1,000**	1	,710	,710	,876	,592	,613	,643
	Sig. (2-tailed)	,000	,000	,	,179	,179	,051	,293	,272	,242
	N	5	5	5	5	5	5	5	5	5
L2PGA	Pearson Correlation	,713	,713	,710	1	1,000**	,962**	,979**	,979**	,986**
	Sig. (2-tailed)	,177	,177	,179	,	,	,009	,004	,004	,002
	N	5	5	5	5	5	5	5	5	5
L2PGB	Pearson Correlation	,713	,713	,710	1,000**	1	,962**	,979**	,979**	,986**
	Sig. (2-tailed)	,177	,177	,179	,	,	,009	,004	,004	,002
	N	5	5	5	5	5	5	5	5	5
L2PGC	Pearson Correlation	,878	,878	,876	,962**	,962**	1	,901*	,909*	,926*
	Sig. (2-tailed)	,050	,050	,051	,009	,009	,	,037	,032	,024
	N	5	5	5	5	5	5	5	5	5
L3PGA	Pearson Correlation	,595	,595	,592	,979**	,979**	,901*	1	,992**	,970**
	Sig. (2-tailed)	,290	,290	,293	,004	,004	,037	,	,001	,006
	N	5	5	5	5	5	5	5	5	5
L3PGB	Pearson Correlation	,615	,615	,613	,979**	,979**	,909*	,992**	1	,953*
	Sig. (2-tailed)	,270	,269	,272	,004	,004	,032	,001	,	,012
	N	5	5	5	5	5	5	5	5	5
L3PGC	Pearson Correlation	,645	,645	,643	,986**	,986**	,926*	,970**	,953*	1
	Sig. (2-tailed)	,239	,240	,242	,002	,002	,024	,006	,012	,
	N	5	5	5	5	5	5	5	5	5

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Legende: L_iPG_j=Kombination von Land i und Produktgruppe j; in jeder Zelle drei Werte:
1.) Pearson'scher Korrelationskoeffizient zwischen den beteiligten Land-/Produktgruppen-Kombinationen; 2.) Sig. (2-tailed)=Signifikanzniveau (zweiseitig); 3.) N=Anzahl der Werte je Land-/Produktgruppenkombination
Korrelationskoeffizienten < 0.9 sind grau unterlegt

Abbildung 3.1: Kreuzkorrelationsanalyse bei Änderung eines dimension members

gruppen in unterschiedlichen Ländern sind hoch positiv korreliert, zwischen den Produktgruppen PG_A und PG_C herrschen hingegen stets signifikant negative Korrelationen - die Gewinne von Produktgruppe PG_A gingen durch die Aggregationsänderungen in $Jahr_t$ stark zurück, während die Gewinne von PG_C in diesem Jahr unverhältnismäßig stark anstiegen. Produktgruppe PG_B, die von sämtlichen Änderungen verschont geblieben war, weist demzufolge mit keiner der beiden anderen Produktgruppen weder positiv noch negativ signifikante Korrelationen auf. Bei einem größeren Datenbestand mit 1000 Produktgruppen wären die Korrelationskoeffizienten zwischen PG_B und allen anderen Produktgruppen mit Ausnahme von PG_A und PG_C zumeist signifikant hoch, sodass man sofort erkennen würde, welche Produktgruppen einen Strukturbruch aufweisen

Correlations

		L1PGA	L1PGB	L1PGC	L2PGA	L2PGB	L2PGC	L3PGA	L3PGB	L3PGC
L1PGA	Pearson Correlation	1	-,716	-,990**	,963**	-,671	-,976**	1,000**	-,571	-,996**
	Sig. (2-tailed)	,	,174	,001	,008	,215	,005	,000	,315	,000
	N	5	5	5	5	5	5	5	5	5
L1PGB	Pearson Correlation	-,716	1	,801	-,503	,993**	,850	-,700	,975**	,773
	Sig. (2-tailed)	,174	,	,103	,387	,001	,068	,188	,005	,126
	N	5	5	5	5	5	5	5	5	5
L1PGC	Pearson Correlation	-,990**	,801	1	-,918*	,765	,996**	-,988**	,674	,998**
	Sig. (2-tailed)	,001	,103	,	,028	,132	,000	,002	,212	,000
	N	5	5	5	5	5	5	5	5	5
L2PGA	Pearson Correlation	,963**	-,503	-,918*	1	-,447	-,881*	,969**	-,333	-,936*
	Sig. (2-tailed)	,008	,387	,028	,	,450	,048	,007	,584	,019
	N	5	5	5	5	5	5	5	5	5
L2PGB	Pearson Correlation	-,671	,993**	,765	-,447	1	,817	-,655	,979**	,733
	Sig. (2-tailed)	,215	,001	,132	,450	,	,091	,230	,004	,159
	N	5	5	5	5	5	5	5	5	5
L2PGC	Pearson Correlation	-,976**	,850	,996**	-,881*	,817	1	-,971**	,732	,991**
	Sig. (2-tailed)	,005	,068	,000	,048	,091	,	,006	,160	,001
	N	5	5	5	5	5	5	5	5	5
L3PGA	Pearson Correlation	1,000**	-,700	-,988**	,969**	-,655	-,971**	1	-,551	-,994**
	Sig. (2-tailed)	,000	,188	,002	,007	,230	,006	,	,336	,001
	N	5	5	5	5	5	5	5	5	5
L3PGB	Pearson Correlation	-,571	,975**	,674	-,333	,979**	,732	-,551	1	,636
	Sig. (2-tailed)	,315	,005	,212	,584	,004	,160	,336	,	,248
	N	5	5	5	5	5	5	5	5	5
L3PGC	Pearson Correlation	-,996**	,773	,998**	-,936*	,733	,991**	-,994**	,636	1
	Sig. (2-tailed)	,000	,126	,000	,019	,159	,001	,001	,248	,
	N	5	5	5	5	5	5	5	5	5

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Legende: L_iPG_j=Kombination von Land i und Produktgruppe j; in jeder Zelle drei Werte:
1.) Pearson'scher Korrelationskoeffizient zwischen den beteiligten Land-/Produktgruppen-Kombinationen; 2.) Sig. (2-tailed)=Signifikanzniveau (zweiseitig); 3.) N=Anzahl der Werte je Land-/Produktgruppenkombination
Positiv signifikante Korrelationen $>$ 0.875 sind hellgrau unterlegt, negativ signifikante Korrelationen $<$ -0.875 dunkelgrau

Abbildung 3.2: Kreuzkorrelationsanalyse bei Änderung der Zugehörigkeitshierarchie

und welche gleichgeblieben sind; allerdings lassen die extrem gegenläufigen Korrelationskoeffizienten zwischen den Produktgruppen PG_A und PG_C bereits bei drei Produktgruppen vermuten, dass letztere beiden an Strukturbrüchen beteiligt gewesen sein mussten.

3.3 Lineare Regression

Regression ist im Allgemeinen als Modellierung einer abhängigen Variable mittels einer oder mehrerer unabhängiger Variablen definiert [We85]. Ein bestimmter Datenvektor eines Data

Warehouses wird somit erklärt durch die Entwicklung der Werte eines anderen Datenvektors (bzw. mehrerer anderer Datenvektoren), wobei die hier analysierten Zusammenhänge stets als linear angenommen werden; es gilt bei der Interpretation des Ergebnisses allerdings zu beachten, dass die Abbildung eines Vektors als Funktion anderer Vektoren nicht Kausalität impliziert [Os90].

Im einfachsten Fall wird eine abhängige Variable Y durch eine andere Variable X zu erklären versucht, i.e.:

$$\hat{Y_i} = A + BX_i, \quad i = 1, \ldots, N \tag{3.2}$$

wobei $\hat{Y_i}$ den geschätzten Wert für Y zum Zeitpunkt i signalisieren soll; A und B sind reelle Zahlen, N repräsentiert die Anzahl der Komponenten der jeweiligen Vektoren. Der tatsächliche Wert Y_i ist demnach

$$Y_i = \hat{Y_i} + \epsilon_i = A + BX_i + \epsilon_i, \tag{3.3}$$

wobei ϵ_i als Fehlprognose ('Errorterm') zum Zeitpunkt t interpretiert wird. Ziel einer optimalen Regressionsgerade ist es, die Summe aller ϵ_i kleinstmöglich zu machen; in der klassischen Regressionsanalyse wird dabei versucht, ϵ^2 zu minimieren (least squares approach)[3]. Formal gilt es daher, folgende Gleichung zu minimieren:

$$F(A, B) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (Y_i - A - BX_i)^2 \tag{3.4}$$

Um die optimalen Parameter A und B zu erhalten, werden die jeweiligen Ableitungen nach A bzw. B null gesetzt, wobei $\bar{X}$ das arithmetische Mittel des Vektors X ($\bar{X} := N^{-1} \sum_{i=1}^{N} x_i$) und

[3]Es wäre natürlich auch möglich, den Absolutbetrag von ϵ zu minimieren, dieses ist aber mathematisch aufgrund der nötigen Fallunterscheidungen schwieriger abzuleiten. Will man jedoch bewusst einzelne extreme Ausreißer nicht zu stark bestrafen, so eignet sich dieser Ansatz besser als die Minimierung der quadrierten Fehlerterme ϵ^2. Eine einfache Minimierung der Summe aller ϵ_i hingegen ist niemals zielführend, da positive Differenzen negative aufheben könnten.

$\bar{Y}$ den Mittelwert des Vektors Y repräsentieren:

$$\begin{aligned}
\frac{\partial F(A,B)}{\partial A} &= \sum_{i=1}^{N} 2(Y_i - A - BX_i)(-1) \stackrel{!}{=} 0 \\
&\quad \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} A - B\sum_{i=1}^{N} X_i = 0 \\
&\quad N\bar{Y} - NA - BN\bar{X} = 0 \\
&\quad A = \bar{Y} - B\bar{X} \\
\frac{\partial F(A,B)}{\partial B} &= \sum_{i=1}^{N} 2(Y_i - A - BX_i)(-X_i) \stackrel{!}{=} 0 \\
&\quad \sum_{i=1}^{N} BX_i^2 = \sum_{i=1}^{N}(X_iY_i - AX_i) \\
&\quad \sum_{i=1}^{N} BX_i^2 = \sum_{i=1}^{N}(X_iY_i - \bar{Y}X_i + B\bar{X}X_i) \\
&\quad B\sum_{i=1}^{N}(X_i^2 - \bar{X}X_i) = \sum_{i=1}^{N} X_i(Y_i - \bar{Y}) \\
&\quad B = \frac{\sum_{i=1}^{N} X_i(Y_i-\bar{Y})}{\sum_{i=1}^{N} X_i(X_i-\bar{X})} = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{N}(X_i-\bar{X})^2},
\end{aligned} \tag{3.5}$$

womit die enge Verwandtschaft zur in Kapitel 3.2 dargestellten Korrelationsanalyse erkennbar wird.

Die Methode lässt sich verallgemeinern für mehrere abhängige Variablen:

$$Y_i = \hat{Y}_i + \epsilon_i = A + B_1X_{i1} + B_2X_{i2} + \ldots + B_NX_{iN} + \epsilon_i.$$

Die Ableitungen sind jedoch nun nicht mehr so einfach zu berechnen, im Falle von zwei abhängigen Variablen liefern die partiellen Ableitungen nach A, B_1 und B_2 bereits folgende umfangreiche Formeln, wobei zur Vereinfachung der Notation $X_i^* := X_i - \bar{X}$ und $Y_i^* := Y_i - \bar{Y}$ eingeführt werden:

$$\begin{aligned}
A &= \bar{Y} - B_1\bar{X}_1 - B_2\bar{X}_2 \\
B_1 &= \frac{\sum_{i=1}^{N} X_1^*Y^* \sum_{i=1}^{N}(X_2^*)^2 - \sum_{i=1}^{N} X_2^*Y^* \sum_{i=1}^{N} X_1^*X_2^*}{\sum_{i=1}^{N}(X_1^*)^2 \sum_{i=1}^{N}(X_2^*)^2 - (\sum_{i=1}^{N} X_1^*X_2^*)^2} \\
B_2 &= \frac{\sum_{i=1}^{N} X_2^*Y^* \sum_{i=1}^{N}(X_1^*)^2 - \sum_{i=1}^{N} X_1^*Y^* \sum_{i=1}^{N} X_1^*X_2^*}{\sum_{i=1}^{N}(X_1^*)^2 \sum_{i=1}^{N}(X_2^*)^2 - (\sum_{i=1}^{N} X_1^*X_2^*)^2}
\end{aligned} \tag{3.6}$$

Als Metrik zur Evaluation der Prognosegenauigkeit eines Regressionsmodells werden in der Literatur u.a. folgende Kennzahlen hervorgehoben [We85], [Os90]:

- R^2-Wert: der R^2-Wert eines Regressionsmodells ist definiert als

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{N} \hat{\epsilon_i}^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{3.7}$$

  und sagt über die Erklärungskraft des Modells aus, indem die Fehlprognosen ins Verhältnis zur Varianz in den Daten gesetzt werden. Der Hauptnachteil dieser Kennzahl liegt darin, dass durch Aufnahme zusätzlicher unabhängiger Variablen der R^2-Wert beliebig bis zum Maximum eins erhöht werden kann. Die nächste Kennzahl behebt diesen Fehler.

- Angepasster R^2-Wert: in dieser Kennzahl wird die Berechnung des ursprünglichen R^2-Wertes aus (3.7) um eine Bestrafung für eine höhere Anzahl an unabhängigen Variablen im Modell modifiziert:

$$\bar{R}^2 = 1 - \frac{N-1}{N-p}(1 - R^2), \tag{3.8}$$

  wobei p die Anzahl der unabhängigen Variablen signalisiert. $\bar{R}^2$ ist somit niedriger als R^2, wenn mehr als eine unabhängige Variable in des Modell einfließt.

- F-Verhältnis: diese Kennzahl berücksichtigt ebenso wie der angepasste R^2-Wert die Länge der Datenvektoren sowie die Anzahl der unabhängigen Variablen:

$$F = \frac{(\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2)/p}{(\sum_{i=1}^{N} \hat{\epsilon_i}^2)/(N-p-1)}, \tag{3.9}$$

  wobei p wie oben für die Anzahl der Modellparameter steht. Je höher das F-Verhältnis ist, desto höher ist die statistische Signifikanz des Modells.

Die Regressionsmethode kann am besten an den Daten aus Tabelle 2.5 illustriert werden: es soll dabei ein lineares Modell mit einer unabhängigen Variable verwendet werden, sodass jede Strukturkombination aus Land und Produktgruppe durch jeweils eine andere Strukturkombination erklärt wird. In Tabelle 3.9 ist das Ergebnis der R^2- und F-Werte für die Strukturkombinationen $Land_1PG_A$, $Land_2PG_A$ und $Land_3PG_A$ aufgeführt.

Abhängige Var. ⇒	$Land_1PG_A$		$Land_2PG_A$		$Land_3PG_A$	
Unbhängige Var. ⇓	R^2	F	R^2	F	R^2	F
$Land_1PG_A$	-	-	0.508	3.096	0.354	1.642
$Land_1PG_B$	1	**469 793**	0.508	3.094	0.354	1.643
$Land_1PG_C$	1	**192 704**	0.505	3.055	0.351	1.62
$Land_2PG_A$	0.508	3.096	-	-	0.959	70.588
$Land_2PG_B$	0.508	3.096	1	∞	0.959	70.588
$Land_2PG_C$	0.771	10.089	0.925	36.750	0.812	12.996
$Land_3PG_A$	0.354	1.642	0.959	70.588	-	-
$Land_3PG_B$	0.378	1.826	0.959	69.444	0.983	174.509
$Land_3PG_C$	0.417	2.143	0.973	108	0.941	47.935

Legende: PG=Produktgruppe

Tabelle 3.9: R^2- und F-Werte bei linearer Regression der Daten aus Tabelle 2.5 mit einer abhängigen Variable

Es zeigt sich ganz deutlich, dass sich die Strukturkombination $Land_1PG_A$ gänzlich durch $Land_1PG_B$ oder $Land_1PG_C$ erklären lässt - der R^2-Wert ist in beiden Fällen 1, der F-Wert mit Werten im sechsstelligen Bereich extrem hoch. Die Erklärungskraft mit Hilfe einer anderen

Strukturkombination allerdings ist wesentlich geringer - der R^2-Wert übersteigt in keinem Fall 0.8; $Land_2PG_A$ und $Land_3PG_A$ hingegen lassen sich sehr gut durch alle anderen Kombinationen, an denen $Land_1$ nicht beteiligt ist, erklären: in jedem Fall liegt der R^2-Wert über 0.8, bis auf einen Fall auch immer über 0.9. Die R^2-Werte in den ersten drei Zeilen (in Verbindung mit $Land_1$) bewegen sich hingegen lediglich zwischen 0.35 und 0.51 - in diesen Zeilen muss somit ein Strukturbruch stattgefunden haben. Die Analyse deckt auch auf, dass die Veränderung bei $Land_1$ in allen Produktgruppenkombinationen gleich war - der Strukturbruch muss somit im Bereich von $Land_1$ und nicht im Bereich der Produktgruppen liegen.

3.4 Autokorrelation

Autokorrelation bezieht sich in Analogie zur Kreuzkorrelation auf die Ähnlichkeit der Werte eines Datenvektors $\vec{x} = (x_1, x_2, \ldots, x_N)^T$ mit seinen eigenen vergangenen Werten[4]. Der Autokorrelationskoeffizient berechnet sich analog zum Pearson'schen Korrelationskoeffizienten (3.1) aus dem vorangegangenen Kapitel, wobei $1/(N-1)$ im Zähler und $1/N$ im Nenner für N genügend groß weggelassen werden können. Die Similarität einer Zeitreihe mit sich selbst um τ Zeiteinheiten verschoben ist daher wie folgt definiert:

$$\rho(\tau) = \frac{\sum_{i=1}^{N-\tau}(x_i - \mu_x)(x_{i+\tau} - \mu_x)}{\sum_{i=1}^{N-\tau}(x_i - \mu_x)^2}, \tag{3.10}$$

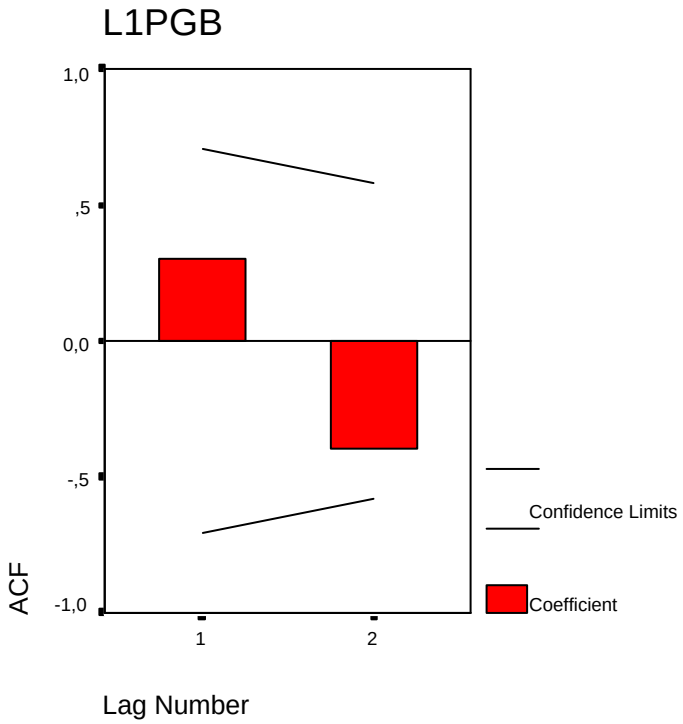
wobei x_i den Wert des Datenvektors $\vec{x}$ zum Zeitpunkt i darstellt, μ_x das arithmetische Mittel der Werte repräsentiert. Die Visualisierung der Autokorrelationsfunktion erfolgt mit Hilfe des Korrelogramms, in dem die Werte der Autokorrelationskoeffizienten in Abhängigkeit des Lags τ in ein Diagramm aufgetragen werden [Wü02].

Zur Erkennung von Strukturbrüchen können einerseits die Differenzen der Autokorrelationskoeffizienten aus der ersten Zeitreihe mit den Werten von x_{t-1} bis $x_{t+\tau-1}$ und der zweiten Zeitreihe mit den Werten von x_t bis $x_{t+\tau}$ herangezogen werden, andererseits liefert auch der Einfluss des Wertes an der Stelle $\tau = 1$ Ergebnisse über die Kontinuität bzw. Diskontinuität in den Daten: bei einem kontinuierlichen Wachstumstrend der Werte wie im Falle der Gewinne von Produktgruppen in obigem Data Warehouse sollte dieser Wert auf alle Fälle hoch positiv sein. Darüber hinaus dient das Korrelogramm noch als wichtiges Analyseinstrument im Rahmen der Autoregression (siehe folgender Abschnitt) und den darauf aufbauenden Kennzahlen.

Wendet man die Autokorrelationsanalyse auf die Daten der Veränderung der Berechnungsvorschrift einer Kennzahl aus Tabelle 2.1 an, so sollten vor allem die Autokorrelationswerte der Zeitperioden $Jahr_{t-4}$ bis $Jahr_{t-1}$ mit jenen der Jahre $Jahr_{t-3}$ bis $Jahr_t$ verglichen wer-

[4] autos (griechisch)=selbst

den. Das Korrelogramm der Autokorrelationswerte der Land-/Produktgruppenkombinationen von $Jahr_{t-4}$ bis $Jahr_{t-1}$ ist exemplifiziert an der Kombination $Land_1PG_B$ in Abbildung 3.3 dargestellt. Der bei kontinuierlich steigenden Gewinnen erwartete positive Einfluss des lediglich

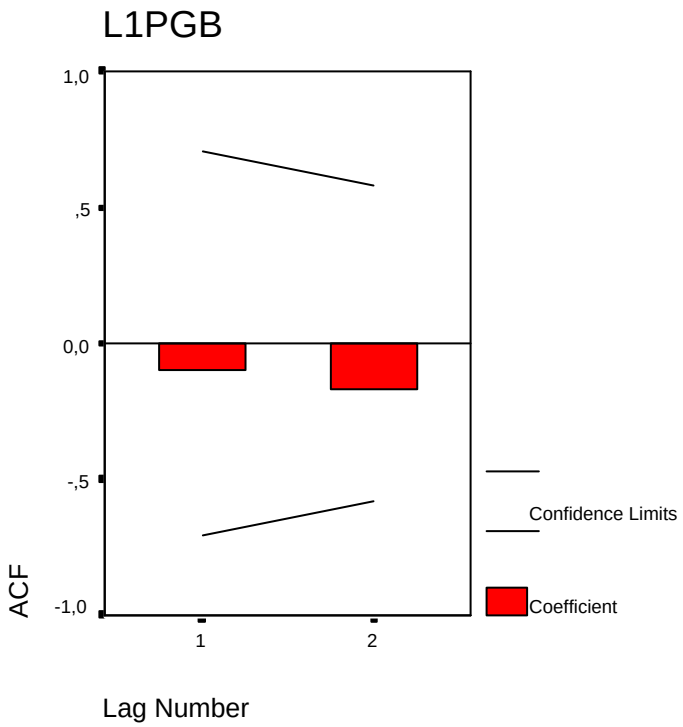


Legende: $L1PGB$=Strukturkombination Land 1 und Produktgruppe B;
ACF=Autokorrelationskoeffizienten zu Lags 1 und 2; Confidence limits zeigen die Grenzen an, ab deren Überschreitung der ACF-Werte die Hypothese der Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegt werden kann

Abbildung 3.3: Korrelogramm vor der Veränderung der Berechnungsvorschrift einer Kennzahl

um eine Zeiteinheit verschobenen Wertes ist gegeben, der Autokorrelationskoeffizient ist 0.3; der um zwei Zeiteinheiten zurückliegende Wert hat einen Wert von -0.4, was auf ein moderates Wachstum der Gewinne schließen lässt: der positive Einfluss des Wertes an der Stelle $\tau = 1$ muss gedämpft werden. Wie aus der Abbildung weiters hervorgeht, liegen die Werte deutlich unter der 95%-Signifikanzgrenze - dies lässt folgern, dass nachfolgende Werte mit vorherigen Werten der untersuchten Zeitreihe nicht signifikant korreliert sind; falls die Koeffizienten außerhalb dieser Grenzen liegen würden, könnte man die Hypothese einer Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegen. Diese Widerlegung ist hier aber vor allem deshalb nicht möglich, weil die analysierten Zeitreihen zu kurz sind, um ein derart hohes Signifikanzniveau erreichen zu können.

Betreibt man nun die gleiche Analyse für den Zeitabschnitt von $Jahr_{t-3}$ bis $Jahr_t$, so ergibt sich das in Abbildung 3.4 dargestellte Szenario: aufgrund der drastischen Kürzung der Gewinne



Legende: $L1PGB$=Strukturkombination Land 1 und Produktgruppe B;
ACF=Autokorrelationskoeffizienten zu Lags 1 und 2; Confidence limits zeigen die Grenzen an, ab deren Überschreitung der ACF-Werte die Hypothese der Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegt werden kann

Abbildung 3.4: Korrelogramm nach der Veränderung der Berechnungsvorschrift einer Kennzahl

sind jetzt die Autokorrelationswerte für $\tau = 1$ ebenso wie für $\tau = 2$ negativ, was einem Wachstumstrend definitiv widerspricht - dies ist als klarer Hinweis (und durch die Differenz zur obigen Abbildung als noch klarerer Hinweis) auf einen Strukturbruch zu werten.

3.5 Autoregression

Bei der Untersuchung von Datenvektoren zu Data Mining-Zwecken ist es nützlich, die beobachteten Wertereihen $\vec{x} = (x_1, x_2, \ldots, x_N)^T$ als eine spezielle Realisierung eines stochastischen Prozesses aufzufassen. Jeder Wert ist demnach als Zusammenspiel von einerseits Zufallsvariablen und andererseits vergangenen Beobachtungen zu interpretieren, deren Einflüsse speziellen

Gesetzen der Wahrscheinlichkeitstheorie (z.B. Normalverteilung) gehorchen.

In den folgenden Abschnitten werden nun die grundlegenden Autoregressionsmodelle vorgestellt, um dann im nächsten Schritt für konkrete Datenvektoren das richtige Modell sowie die richtigen Parameter des gewählten Modells auswählen zu können; es wird dabei stets von (zumindest schwach)[5] stationären Prozessen ausgegangen, d.h. von Prozessen, für die bezüglich Erwartungswert, Varianz und Autokovarianz gilt (vgl. hier und im folgenden u.a. [Wü02], [Ha93]):

$$E(x_1) = E(x_2) = \ldots E(x_N) = \mu_x, \tag{3.11}$$

$$Var(x_i) = \sigma^2 = \gamma(0) = E[(x_i - \mu_x)^2], \;\; i = 1, \ldots, N \tag{3.12}$$

sowie

$$Cov(x_i, x_{i-\tau}) = \gamma(\tau) = E[(x_i - \mu_x)(x_{i-\tau} - \mu_x)], \;\; i = \tau + 1, \ldots, N; \tag{3.13}$$

die Notation $\gamma(0)$ bzw. $\gamma(\tau)$ ist angegeben, da sich der in (3.10) definierte und auch hier benötigte Autokorrelationskoeffizient $\rho(\tau)$ einer Datenreihe mit sich selbst um τ Zeiteinheiten verschoben somit auch darstellen lässt als

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}, \;\; \tau = 0, \pm 1, \pm 2, \ldots \tag{3.14}$$

Grundlage der im folgenden dargestellten Autoregressionsmodelle ist das Dekompositionstheorem von *Wold*, das besagt, dass sich jeder stationäre, nicht-deterministische Prozess als Linearkombination von unkorrelierten Zufallsvariablen darstellen lässt:

$$x_t - \mu_x = u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + \ldots = \sum_{j=0}^{\infty} \psi_j u_{t-j}, \psi_0 = 1, \tag{3.15}$$

wobei u_t unabhängige Zufallsvariablen mit $\mu_u = 0$ und $Var(u) = \sigma^2 < \infty$ (sonst keine Konvergenz) sind und ψ_j als Gewichte interpretiert werden; aufgrund der Unkorreliertheit der einzelnen Zufallsvariablen gilt auch: $Cov(u_i, u_{i-k}) = 0$, falls $k \neq 0$.

[5] Starke Stationärität impliziert, dass die gemeinsame Wahrscheinlichkeitsverteilung einer Menge von r Werten t_1,t_2,...,t_r die gleiche ist wie die Zusammenhangsverteilung zu den Zeitpunkten $t_{1+\tau}$,$t_{2+\tau}$,...,$t_{r+\tau}$ für ein beliebiges τ. Starke Stationärität zieht somit automatisch schwache Stationärität nach sich; die Umkehrung gilt im Allgemeinen jedoch nicht.

3.5.1 AR(p)-Prozesse

Setzt man in (3.15) die Parameter $\psi_j = \phi^j$, so erhält man (unter der vereinfachenden Annahme, dass $\mu_x = 0$) den einfachsten autoregressiven Prozess, den AR(1)-Prozess:

$$\begin{aligned} x_t &= u_t + \phi u_{t-1} + \phi^2 u_{t-2} + \dots \\ &= \phi x_{t-1} + u_t, \quad t = 1, \dots, N \end{aligned} \tag{3.16}$$

wobei man die zweite Zeile von (3.16) erhält, indem man von der ersten Zeile das ϕx_{t-1}-fache subtrahiert. Der Wert x_t geht somit einerseits auf seinen vorherigen Wert x_{t-1} multipliziert mit dem Parameter ϕ zurück ($\Rightarrow$ Regression), andererseits erklärt sich x_t auch noch durch eine Zufallsvariable u_t, die die Stochastizität des Prozesses hervorhebt [Wü02]. Obwohl der erste Wert erst bei $t = 1$ beobachtet wird, wird der Prozess so betrachtet, als hätte er bereits vor langer Zeit (zum Zeitpunkt J) begonnen. Ersetzt man in (3.16) sukzessive x_{t-1} durch $\phi x_{t-2} + u_{t-1}$, x_{t-2} durch $\phi x_{t-3} + u_{t-2}$, etc., so erhält man:

$$\begin{aligned} x_t &= \textstyle\sum_{j=0}^{\infty} \phi^j u_{t-j} \\ &= \textstyle\sum_{j=0}^{J-1} \phi^j u_{t-j} + \phi^J x_{t-J} \end{aligned} \tag{3.17}$$

Der Erwartungswert μ_x dieses Prozesses ist 0, weil die u_t iid[6] sind. Für die Varianz $Var(x_t)$ bei $|\phi| < 1$ hingegen gilt:

$$\begin{aligned} \gamma(0) &= E(x_t^2) = E(\textstyle\sum_{j=0}^{\infty} \phi^j u_{t-j})^2 \\ &= \textstyle\sum_{j=0}^{\infty} \phi^{2j} E(u_{t-j}^2) \\ &= \sigma^2 \textstyle\sum_{j=0}^{\infty} \phi^{2j} \\ &= \sigma^2/(1-\phi^2) \end{aligned} \tag{3.18}$$

Die zweite Zeile von (3.18) nützt die Linearität des Erwartungswertes aus, die dritte Zeile setzt für die Varianz der u_t σ^2 ein, und die vierte Zeile baut auf der Voraussetzung auf, dass $|\phi| < 1$ ist, womit die Reihe $\sum_{j=0}^{\infty} \phi^{2j}$ gegen $\frac{1}{1-\phi^2}$ konvergiert. Ist $|\phi| \geq 1$ (nur in Ausnahmefällen), dann hängt der Erwartungswert des Prozesses primär vom Startwert des Prozesses vor langer Zeit ab. Die Autokovarianz des AR(1)-Prozesses erhält man, indem man in der zweiten Zeile von (3.17) $J = \tau$ setzt:

$$\gamma(\tau) = E(x_t x_{t-\tau}) = E\left[\left(\sum_{j=0}^{\tau-1} \phi^j u_{t-j} + \phi^\tau x_{t-\tau}\right) x_{t-\tau}\right]$$

[6] iid=independent and identically distributed = voneinander unabhängig und gleichverteilt

Da alle $u_t, \ldots, u_{t-\tau+1}$ den Wert von $x_{t-\tau}$ gemäß der Annahmen des AR(1)-Prozesses nicht beeinflussen, ist der Erwartungswert der Summe von $\sum_{j=0}^{\tau-1} \phi^j u_{t-j} = 0$, und somit gilt:

$$\gamma(\tau) = \phi^\tau E(x_{t-\tau}^2) = \phi^\tau \gamma(0) \tag{3.19}$$

Der AR(1)-Prozess kann verallgemeinert werden auf den AR(p)-Prozess:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + u_t, \quad t = 1, \ldots, N \tag{3.20}$$

wobei Varianz und Autokovarianz ähnlich wie im AR(1)-Prozess in (3.18) und (3.19) hergeleitet werden können [Ha93].

3.5.2 MA(q)-Prozesse

In (3.15) kann alternativ zum AR(p)-Prozess auch $\psi_1 = \theta$ und $\psi_j = 0$ für $j > 1$ gewählt werden, womit man den einfachsten 'Moving Average'-Prozess, den MA(1)-Prozess, erhält:

$$x_t = u_t + \theta u_{t-1}, \quad t = 1, \ldots, N \tag{3.21}$$

Zur Vereinfachung der Analyse ist es hilfreich, den MA(q)-Prozess als autoregressiven Prozess auszudrücken; setzt man im konkreten Fall des MA(1)-Prozesses für u_{t-1} (wegen $x_{t-1} = u_{t-1} + \theta u_{t-2}$) $x_{t-1} - \theta u_{t-2}$ ein, für u_{t-2} ebenso $x_{t-2} - \theta u_{t-3}$, etc., so erhält man:

$$\begin{aligned} x_t &= u_t + \theta(x_{t-1} - \theta(x_{t-2} - \theta(x_{t-3} - \ldots))) \\ &= u_t + \theta^1 x_{t-1} - \theta^2 x_{t-2} + \theta^3 x_{t-3} \ldots \\ &= -\textstyle\sum_{j=1}^{\infty} (-\theta)^j x_{t-j} + u_t \end{aligned} \tag{3.22}$$

Der Erwartungswert des MA(1)-Prozesses aus (3.21) ist wegen $\mu_{u_i} = 0$ auch hier 0, die Varianz entspricht (wiederum unter Berücksichtigung der Unkorreliertheit von u_t und u_{t-1})

$$\begin{aligned} \gamma(0) &= E(x_t^2) = E[(u_t + \theta u_{t-1})(u_t + \theta u_{t-1})] \\ &= E(u_t^2) + \theta^2 E(u_{t-1}^2) + 2\theta E(u_t u_{t-1}) \\ &= \sigma^2 + \theta^2 \sigma^2 \\ &= (1 + \theta^2)\sigma^2 \end{aligned} \tag{3.23}$$

Die Autokovarianz für $\tau = 1$ ist ähnlich definiert als

$$\begin{aligned}\gamma(1) &= E(x_t x_{t-1}) = E[(u_t + \theta u_{t-1})(u_{t-1} + \theta u_{t-2})] \\ &= E(u_t u_{t-1}) + \theta E(u_{t-1})^2 + \theta E(u_t u_{t-2}) + \theta^2 E(u_{t-1} u_{t-2}) \\ &= \theta E(u_{t-1}^2) \\ &= \theta \sigma^2\end{aligned} \tag{3.24}$$

Ist $\tau > 1$, so ergibt sich in der Ausmultiplizierung des Polynoms $E(x_t x_{t-\tau})$ kein Produkt gleicher u_t mehr, somit gilt für den MA(1)-Prozess:

$$\gamma(\tau) = 0, \quad \tau = 2, 3, \ldots \tag{3.25}$$

Auch der MA(1)-Prozess kann generalisiert werden auf Prozesse der Ordnung q, i.e.:

$$x_t \sim MA(q) \Leftrightarrow x_t = u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q}, \;\; t = 1, \ldots, N, \tag{3.26}$$

wobei Varianz und Autokovarianzen wieder in ähnlicher Weise zu (3.23) und (3.24) hergeleitet werden können.

3.5.3 ARMA(p,q)-Prozesse

Die beiden Prozesse der vorangegangenen Abschnitte können zusammengefasst werden zur generelleren Klasse der ARMA(p,q)-Modelle. Der einfachste stochastische Prozess dieser Kategorie ist der ARMA(1,1)-Prozess, der definiert ist als

$$x_t = \phi x_{t-1} + u_t + \theta u_{t-1} \tag{3.27}$$

Der Erwartungswert μ_x für das ARMA(1,1)-Modell ist auch hier wieder gemäß Annahme 0, für die Varianz und Autokovarianz multipliziert man zunächst (3.27) mit $x_{t-\tau}$, somit gilt für die Erwartungswerte unter Ausnützung ihrer Linearitätseigenschaft:

$$E(x_t x_{t-\tau}) = \phi E(x_{t-1} x_{t-\tau}) + E(u_t x_{t-\tau}) + \theta E(u_{t-1} x_{t-\tau}) \tag{3.28}$$

Aus $\gamma(\tau) = E(x_t x_{t-\tau})$ folgt:

$$\gamma(\tau) = \phi \gamma(\tau - 1) + E(u_t x_{t-\tau}) + \theta E(u_{t-1} x_{t-\tau})$$

Für $\tau \geq 2$ sind die beiden letzten Terme jeweils 0, da x_t lediglich von u_t, u_{t-1}, …, etc. abhängt, nicht aber von $u_{t+1}, \ldots, u_{t+\tau}$. Ist $\tau = 1$ (Autokovarianz zu Lag 1), so ist der vorletzte Term von

(3.28) 0, für den Erwartungswert des letzten Terms aus (3.28) hingegen gilt wegen (3.27):

$$\begin{aligned} E(u_{t-1}x_{t-1}) &= E[u_{t-1}(\phi x_{t-2} + u_{t-1} + \theta u_{t-2})] \\ &= \phi E(u_{t-1}x_{t-2}) + E(u_{t-1}^2) + \theta E(u_{t-1}u_{t-2}) \\ &= \sigma^2, \end{aligned}$$

wobei die zweite Zeile wiederum die Linearität des Erwartungswertes ausnützt; die dritte Zeile baut auf der Voraussetzung auf, dass x_t von $u_{t+1}, \ldots, u_{t+\tau}$ unabhängig ist. Ist $\tau = 0$ (=Varianz), so ist der Erwartungswert des vorletzten Terms von (3.28) mit den obigen Umformungen σ^2, für den Erwartungswert des letzten Terms gilt wegen (3.27):

$$\begin{aligned} E(u_{t-1}x_t) &= E[u_{t-1}(\phi x_{t-1} + u_t + \theta u_{t-1})] \\ &= \phi E(u_{t-1}x_{t-1}) + E(u_{t-1}u_t) + \theta E(u_{t-1}^2) \\ &= \phi\sigma^2 + \theta\sigma^2, \end{aligned}$$

wobei die dritte Zeile das vorher berechnete Resultat $E(u_{t-1}x_{t-1}) = \sigma^2$ sowie die Unkorreliertheit von u_{t-1} und u_t ausnützt. Setzt man die errechneten Erwartungswerte in die Formel von (3.27) ein, so erhält man:

$$\begin{aligned} \gamma(0) &= \phi\gamma(1) + \sigma^2 + \theta\phi\sigma^2 + \theta^2\sigma^2 \\ \gamma(1) &= \phi\gamma(0) + \theta\sigma^2 \\ \gamma(\tau) &= \phi\gamma(\tau - 1), \quad \tau = 2, 3, \ldots \end{aligned} \tag{3.29}$$

Setzt man die zweite Gleichung von (3.29) in die erste ein, so ergibt sich:

$$\gamma(0) = \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2}\sigma^2 \tag{3.30}$$

und für die Autokovarianz für $\tau = 1$:

$$\gamma(1) = \frac{(1 + \phi\theta)(\phi + \theta)}{1 - \phi^2}\sigma^2 \tag{3.31}$$

Auch hier kann der ARMA(1,1)-Prozess verallgemeinert werden zu einem ARMA(p,q)-Prozess

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q}, \tag{3.32}$$

wobei Varianz und Autokovarianzen wieder ähnlich (wenn auch mit nun komplizierteren Umformungen) hergeleitet werden können.

3.5.4 ARIMA(p,d,q)-Prozesse

Gelten für eine Datenreihe die in (3.11), (3.12) und (3.13) geforderten Stationäritätsbedingungen nicht, so muss vor Anwendung des ARMA(p,q)-Modells zunächst eine Differenzoperation durchgeführt werden. Dazu führt man die Hilfsvariable y_t ein, die die Veränderung von x_t misst, i.e.:

$$y_t = (1 - B)^d x_t, \tag{3.33}$$

wobei B für den Backshiftoperator steht, i.e.: $Bx_t = x_{t-1}, B^2x_t = x_{t-2}, \ldots, B^dx_t = x_{t-d}$. Für $d = 1, 2, 3, \ldots$ in (3.33) erhält man folgende modifizierte Wertereihe:

$$\begin{aligned} y_t &= (1-B)x_t = x_t - x_{t-1}, && d = 1, \\ &= (1-B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}, && d = 2, \\ &= (1-B)^3 x_t = (1 - 3B + 3B^2 - B^3)x_t = x_t - 3x_{t-1} + 3x_{t-2} - x_{t-3}, && d = 3, \\ &\ldots \end{aligned}$$

Ein Prozess ist somit ein ARIMA(p,d,q)-Prozess[7] ($d \geq 0$), wenn die transformierte Wertereihe mit $y_t = (1 - B)^d x_t$ ein gewöhnlicher ARMA(p,q)-Prozess ist.

Die Frage, welche Differenzordnung d auf die Daten angewendet werden muss, hängt vollständig von den Differenzen in den Werten ab: Tabelle 3.10 zeigt beispielhaft Daten auf, für die die Differenzordnung $d = 1$ absolut ausreichend ist, die Werte der zweiten Spalte können durchaus als stationär betrachtet werden - die Wahl einer höheren Differenzordnung ($d = 2$ in der dritten Spalte) trägt nicht mehr zur Verbesserung der Stationärität bei; zudem muss als weiterer Nachteil höherer Differenzordnungen stets berücksichtigt werden, dass bei Differenzordnung d und ursprünglicher Anzahl von Werten N nur mehr $N - d$ Datenwerte zur Verfügung stehen, die restlichen fallen der Differenzoperation zum Opfer. Bei superlinearem Anstieg der

x_t	$(1-B)x_t$	$(1-B)^2x_t$
20		
25	5	
28	3	-2
31	3	0
35	4	1
37	2	-2
41	4	2

Legende: B=Backshiftoperator ($B^dx_t = x_{t-d}$)

Tabelle 3.10: Stationärität bereits bei Differenzordnung $d = 1$

[7] ARIMA=Autoregression Integrated Moving Average

Werte x_t reicht jedoch die Differenzordnung $d = 1$ nicht aus; Tabelle 3.11 zeigt eine Datenreihe auf, für die erst bei Differenzordnung $d = 2$ Stationärität erreicht wird. Falls die Werte

x_t	$(1-B)x_t$	$(1-B)^2x_t$
9		
16	7	
25	9	2
36	11	2
49	13	2
64	15	2
81	17	2

Legende: B=Backshiftoperator ($B^d x_t = x_{t-d}$)

Tabelle 3.11: Stationärität erst bei Differenzordnung $d = 2$

jedoch nicht nur superlinear ansteigen, sondern sich von einer Zeiteinheit auf die andere verdoppeln, verdreifachen, etc., so ist eine vorherige Transformation der Daten notwendig. Betrachtet man beispielsweise die Daten in Tabelle 3.12, so erkennt man, dass die Differenzordnung beliebig erhöht werden kann, ohne jedoch Stationärität der Daten zu erreichen. In diesem Fall

x_t	$(1-B)x_t$	$(1-B)^2x_t$	$(1-B)^3x_t$
20			
40	20		
80	40	20	
160	80	40	20
320	160	80	40
640	320	160	80
1280	640	320	160

Legende: B=Backshiftoperator ($B^d x_t = x_{t-d}$)

Tabelle 3.12: Kontinuierliche Verdopplung der Werte - auch höhere Differenzordnungen bringen keine Stationärität

ist es notwendig, die Daten bereits vor der Analyse mittels eines ARIMA(p,d,q)-Modells zu transformieren: als Möglichkeit bietet sich hier die Box-Cox-Transformation [BD02] an, die die ursprünglichen Daten transformiert:

$$f_\lambda(x_t) = \begin{cases} \lambda^{-1}(x_t^\lambda - 1), & x_t \geq 0, \lambda > 0, \\ \ln x_t, & x_t > 0, \lambda = 0. \end{cases} \tag{3.34}$$

Für Data Mining-Belange kann $\lambda = 0$ gesetzt werden [BD02], womit die logarithmierten Werte der Datenreihe untersucht werden; bei Werten ≤ 0 muss zunächst zu jedem Wert der Absolut-

betrag des kleinsten Wertes+1[8] hinzugezählt werden. Wendet man die Box-Cox-Transformation auf die Daten aus Tabelle 3.12 an, so erkennt man, dass nun die Differenzoperationen Stationärität erzeugen, wie in Tabelle 3.13[9] aufgeführt ist.

$\ln x_t$	$(1-B)\ln x_t$
3	
3.69	0.69
4.38	0.69
5.08	0.69
5.77	0.69
6.46	0.69
7.15	0.69

Legende: B=Backshiftoperator ($B^d x_t = x_{t-d}$)

Tabelle 3.13: Differenzoperationen auf den transformierten Daten - Stationärität bereits bei Differenzordnung $d = 1$

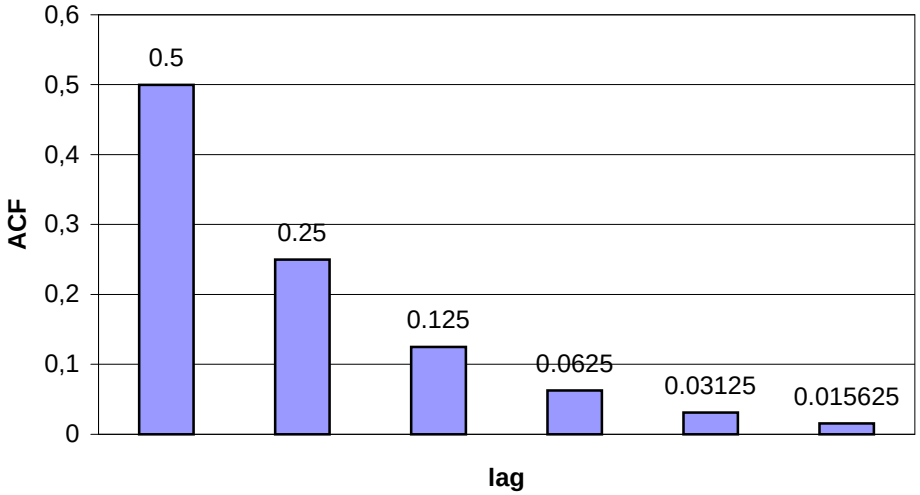
3.5.5 Identifikation des korrekten Modells und Schätzung der Parameter

Hat man eine bzw. mehrere konkrete Datenreihen zur Verfügung, so gilt es zunächst zu analysieren, welches der oben genannten AR(p)-, MA(q)- bzw. ARMA(p,q)-Modelle anzuwenden ist bzw. ob vorher noch Transformationen und Differenzoperationen durchzuführen sind. Diese Identifikation erfolgt zumeist mit Hilfe der Visualisierung der Autokorrelationsfunktion $\rho(\tau)$ in einem Korrelogramm. Da der Autokorrelationskoeffizient an der Stelle τ (Lag τ) der Division der Autokovarianz zu Lag τ durch die Varianz der Datenreihe entspricht, können aus den in den obigen Kapiteln hergeleiteten Formeln für die Varianz und die Autokovarianz(en) folgende Eigenschaften des Autokorrelationskoeffizienten abgeleitet werden:

- Die Autokorrelationsfunktion eines AR(1)-Prozesses ergibt sich gemäß (3.18) und (3.19) als
$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \frac{\phi^\lambda \sigma^2}{1-\phi^2}\frac{1-\phi^2}{\phi^2} = \phi^\lambda.$$
Zeigt sich somit in einem Korrelogramm einer Datenreihe eine Korrelation, die mit zunehmendem τ exponentiell zurückgeht, aber nicht sofort abbricht, dann ist grundsätzlich ein AR-Prozess heranzuziehen. Wenn die Korrelationswerte stets positiv sind, so ist ein AR(1)-Prozess mit positivem Parameter ϕ ($0 < \phi < 1$) zu wählen, exemplifiziert ist das Korrelogramm für $\phi = 0.5$ in Abbildung 3.5 dargestellt. Wenn die Korrelationswerte ab-

[8] Somit ist auch der kleinste Wert positiv, und der Logarithmus kann von jedem Wert gezogen werden.

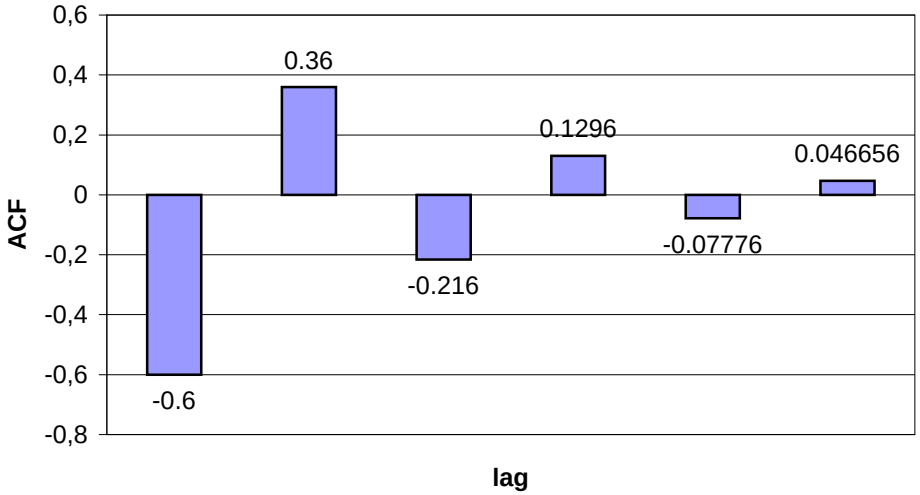
[9] Die in der ersten Spalte von Tabelle 3.13 angegebenen Werte enthalten Rundungsfehler, die erste Differenzoperation wurde aber auf den korrekten Werten durchgeführt, wodurch auch die Differenz zwischen der dritten und vierten Zeile 0.69 und nicht 0.7 ist.



Legende: ACF=Autokorrelationskoeffizient für Lag 1 bis 6

Abbildung 3.5: Korrelogramm eines AR(1)-Prozesses mit $\phi = 0.5$

wechselnd negativ und positiv sind, so entspricht dies einem AR(1)-Prozess mit negativem ϕ $(-1 < \phi < 0)$; beispielhaft zeigt Abbildung 3.6 das Korrelogramm eines AR(1)-Prozesses mit $\phi = -0.6$. Sind die Korrelationen positiv und negativ, aber nicht alternierend nega-



Legende: ACF=Autokorrelationskoeffizient für Lag 1 bis 6

Abbildung 3.6: Korrelogramm eines AR(1)-Prozesses mit $\phi = -0.6$

tiv und positiv, so ist ein AR-Prozess höherer Ordnung heranzuziehen, dessen Parameter

$\phi_1, \ldots, \phi_p$ negativ und positiv sind, ebenso sind Prozesse höherer Ordnung p heranzuziehen, wenn sich die Autokorrelationen durch Differenzengleichungen p-ter Ordnung darstellen lassen.

- Die Autokorrelationsfunktion eines MA(1)-Prozesses ermittelt sich analog aus (3.23), (3.24) und (3.25):

$$\begin{aligned} \rho(1) &= \frac{\gamma(\tau)}{\gamma(0)} = \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{1+\theta^2}, \\ \rho(\tau) &= 0, \quad \tau = 2, 3, \ldots \end{aligned}$$

  Zeigt sich in einem Korrelogramm eines Datenvektors eine Korrelation, die abrupt (nicht jedoch wegen der geringen Länge des Vektors) zu $\tau = q$ endet, so ist ein MA(q)-Prozess zu selektieren. Der Parameter θ ist positiv zu wählen, wenn die Werte stärker als eine Zufallsreihe von Werten korrelieren, und negativ, wenn die Differenzen aufeinanderfolgender Werte von $\vec{x}$ eher abwechselnd positiv und negativ sind.

- Die Autokorrelationsfunktion eines ARMA(1,1)-Prozesses berechnet sich gemäß (3.29), (3.30) und (3.31) als:

$$\begin{aligned} \rho(1) &= \frac{\gamma(\tau)}{\gamma(0)} = \frac{(1+\phi\theta)(\phi+\theta)\sigma^2}{1-\phi^2} \frac{1-\phi^2}{(1+\theta^2+2\phi\theta)\sigma^2} = \frac{(1+\phi\theta)(\phi+\theta)}{(1+\theta^2+2\phi\theta)} \\ \rho(\tau) &= \frac{\phi\gamma(\tau-1)}{\gamma(0)} = \phi\rho(\tau-1), \quad \tau = 2, 3, \ldots \end{aligned}$$

  Das idealtypische Korrelogramm eines ARMA(1,1)-Modells entspricht somit im wesentlichen jenem eines AR(1)-Modells; daher sind auch hier bei stets positiven Korrelationen positive Werte für θ und ϕ heranzuziehen, bei stets negativen Korrelationen ist $\theta < 0$ zu wählen, und bei alternierend positiven und negativen Korrelationen ist ϕ negativ zu setzen.

Ist das korrekte Modell nun gefunden, so gilt es, die Parameter $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ so zu wählen, dass der Fehler zwischen den prognostizierten Daten und den tatsächlichen Daten minimal ist. Dies erfolgt mit der Maximum Likelihood-Methode: der Likelihood einer Parametermenge $\Theta = \{\sigma_u^2, \mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q\}$ ist definiert als:

$$L(\Theta) = p(\vec{x}|\Theta) = \prod_{n=1}^{N} p(x_n|\Theta); \tag{3.35}$$

der Likelihood einer Menge von Parametern kann daher interpretiert werden als die Wahrscheinlichkeit, dass die Daten $x_1, \ldots, x_N$ von Θ generiert worden sind. Der Likelihood soll maximiert werden, wobei aus rechentechnischen Gründen in praxi der negative Log-Likelihood minimiert

wird[10]:

$$-\ln L(\Theta) = -\sum_{n=1}^{N} \ln p(x_n|\Theta). \tag{3.36}$$

Die einzelnen Werte von Θ erhält man, indem die Ableitungen des Log-Likelihoods nach dem jeweiligen Parameter null gesetzt werden. Können die Daten als normalverteilt betrachtet werden, so gilt:

$$\begin{array}{rcl} -\ln L(\Theta) & = & -\sum_{n=1}^{N} \ln p(x_n|\mu,\sigma^2) \\ & = & -\sum_{n=1}^{N} (\ln \frac{1}{\sqrt{2\Pi}\sigma} - \frac{1}{2}\frac{(x_n-\mu)^2}{\sigma^2}), \end{array} \tag{3.37}$$

womit sich bei Ableiten nach μ und σ^2 die bekannten Formeln ergeben:

$$\begin{array}{rcl} \mu & = & \frac{1}{N}\sum_{n=1}^{N} x_n \\ \sigma^2 & = & \frac{1}{N}\sum_{n=1}^{N} (x_n-\mu)^2. \end{array}$$

Hat man zudem einen AR(1)-Prozess als passend herausgewählt, so gilt wegen

$$x_t = \phi x_{t-1} + u_t$$

unter Vernachlässigung der Zufallsvariablen u_t[11]:

$$-\ln L(\Theta) = -\sum_{n=2}^{N} (\ln \frac{1}{\sqrt{2\Pi}\sigma} - \frac{1}{2}\frac{(\phi x_{n-1}-\mu)^2}{\sigma^2}). \tag{3.38}$$

Beim Ableiten von (3.38) nach ϕ ergibt sich

$$\begin{array}{rcl} \frac{\partial L}{\partial \phi}\left[-\sum_{n=2}^{N} (\ln \frac{1}{\sqrt{2\Pi}\sigma} - \frac{1}{2}\frac{(\phi x_{n-1}-\mu)^2}{\sigma^2})\right] & = & \\ -\sum_{n=2}^{N} -\frac{1}{2}\frac{2(x_{n-1}\phi-\mu)x_{n-1}}{\sigma^2} & = & \\ \sum_{n=2}^{N} \frac{x_{n-1}^2\phi-\mu x_{n-1}}{\sigma^2} & \stackrel{!}{=} & 0 \\ \sum_{n=2}^{N} x_{n-1}^2\phi & = & \sum_{n=2}^{N} \mu x_{n-1} \\ \phi\sum_{n=2}^{N} x_{n-1}^2 & = & \mu\sum_{n=2}^{N} x_{n-1} \\ \phi & = & \frac{\mu\sum_{n=2}^{N} x_{n-1}}{\sum_{n=2}^{N} x_{n-1}^2}. \end{array} \tag{3.39}$$

Bei Modellen höherer Ordnung wird genauso vorgegangen; allerdings muss dabei beachtet werden, dass zur Ableitung eines Prozesses mit t Parametern ein lineares Gleichungssystem mit t Gleichungen gelöst werden muss.

[10] Durch das Ziehen der Logarithmen wird aus dem Produkt- ein Summenzeichen, womit sich die Differenzierungsoperationen wesentlich vereinfachen.

[11] Die Summe läuft nun von zwei bis N, da der erste Wert nicht durch einen Vorgänger erklärt werden kann und somit aus der Berechnung herausfällt.

3.5.6 Einsetzbarkeit von Autoregression im Data Mining

Bei der Eignung der aus den vorangegangenen Abschnitten gewonnenen Erkenntnisse über Autoregression zu Data Mining-Zwecken gilt es zu beachten, dass bei der Analyse mehrerer Wertereihen möglicherweise nicht alle Wertereihen den gleichen ARIMA(p,d,q)-Prozess $(p, d, q \geq 0)$ als optimal ansehen. Um die Vergleichbarkeit gewährleisten zu können, muss aber für alle Datenreihen ein und dasselbe Modell herangezogen werden; im Zweifelsfall ist eher ein Prozess niedrigerer Ordnung zu wählen, um kein 'Overfitting' einzelner Datenreihen zu betreiben, auch unter Berücksichtigung der Performance erscheinen einfachere Modelle als vorteilhaft. Im Normalfall ist darüber hinaus ein AR(p)-Prozess einem MA(q)-Prozess vorzuziehen, da beim Vergleich mehrerer Datenreihen selten abrupte Brüche der Autokorrelationen bei allen Datenreihen auftreten sollten[12].

Als Distanzfunktion bietet sich die Möglichkeit an, die Koeffizienten sämtlicher Parameter ϕ und θ des gewählten Modells für jede Datenreihe mittels der oben aufgeführten Maximum-Likelihood-Methode zu bestimmen und dann die Gesamtdifferenz zwischen allen Koeffizienten als Distanzmetrik heranzuziehen, i.e.:

$$dist(\vec{x}, \vec{y}) = \sum_{i=1}^{p}(abs(\phi_{ix} - \phi_{iy})) + \sum_{j=1}^{q}(abs(\theta_{jx} - \theta_{jy})),$$

wobei ϕ_{ix} und θ_{ix} als i-ter Parameter ϕ bzw. j-ter Parameter θ des untersuchten Datenvektors $\vec{x}$ zu interpretieren sind. Dies kann allerdings einerseits aufgrund der bei Modellen höherer Ordnung zu lösenden LGS zu den oben erwähnten Performanceproblemen führen, andererseits sind die exakten Werte von ϕ und θ aus Data Mining-Perspektive sekundär; es ist daher sinnvoll, die einzelnen Parameter ϕ und θ mit Hilfe der Korrelogramme (unter Zuhilfenahme der im vorigen Abschnitt erläuterten Spezifika von Parametern von Autokorrelationsfunktionen) zu schätzen und auf eine exakte Ermittlung zu verzichten. Es existiert dennoch eine breite Palette von Distanzmetriken, die auch bei geschätzten Modellparametern ϕ und θ eingesetzt werden können, von denen drei bedeutende im folgenden vorgestellt werden.

Eine weit verbreitete Kennzahl zur Erklärung eines Datenvektors durch ein bestimmtes ARMA(p,q)-Modell ist das *AIC*, das *Akaike Information Criterion* [Wü02], das wie folgt definiert ist:

$$AIC(p, q) = -2L(\Theta|\vec{x}) + \frac{2(p+q)}{N} = \log \hat{\sigma}^2 + \frac{2(p+q)}{N}, \tag{3.40}$$

[12] Wäre dies der Fall, so hätte man allerdings Indizien für einen Strukturbruch, der sämtliche Strukturkombinationen erfassen würde.

wobei $\hat{\sigma}^2$ ist die geschätzte Varianz der Residualwerte u_t darstellt, die sich berechnet als

$$E(x_t - \hat{x_t}) = \frac{1}{N - max(p,q)} \sum_{i=max(p,q)+1}^{N} (x_i - \hat{x_i})^2; \tag{3.41}$$

$\hat{x_i}$ sind dabei die aufgrund der Prämissen des Modells prognostizierten Werte. Wird somit beispielsweise ein AR(p)-Prozess als Basis herangezogen, so ist $\hat{x_t} = \sum_{i=1}^{p} \phi_i x_{t-i}$. Der zweite Term von (3.40) addiert auf höhere Modelle einen 'Penalty': je höher die Anzahl der eingesetzten Parameter ϕ und θ ist, umso flexibler sind die Modelle, und umso genauer können die Basisdaten approximiert werden. Da damit jedoch ein Performanceverlust verbunden ist sowie möglicherweise Overfitting betrieben wird, erhöht der zweite Term den AIC-Wert komplexer Modelle. Während in der klassischen Autoregressionsanalyse bei Untersuchung dieser Kennzahl das Minimieren des AIC-Wertes oberste Priorität besitzt, sind aus Data Mining-Sicht primär die Differenzen der AIC-Werte einzelner Datenreihen von Bedeutung.

In der Literatur (vgl. [Ha93]) wird auch häufig das BIC (Schwarz-Kriterium) herangezogen, das Modelle höherer Ordnung mit geringeren Strafen belegt:

$$BIC(p,q) = -2L(\Theta|\vec{x}) + \frac{(p+q)}{N \log N} = \log \hat{\sigma}^2 + \frac{(p+q)}{N \log N}. \tag{3.42}$$

Wenn AR(p)-Prozesse miteinander verglichen werden, kann auch der *Final Prediction Error* (FPE) als Distanzmetrik herangezogen werden. Der FPE eines AR(p)-Prozesses ist definiert als

$$FPE(p) = \hat{\sigma}^2 \frac{n+p}{n-p}. \tag{3.43}$$

Bei Data-Mining-Analysen gilt es prinzipiell, bei Vergleichen verschiedener Datenvektoren auch die durchschnittliche Größe der Werte der einzelnen Vektoren zu beachten - aus diesem Grund werden am folgenden Beispiel zwei Varianten obiger Kennzahlen erprobt: die erste Variante berechnet die Kennzahlen ohne Berücksichtigung der Größen der Werte der Datenreihen klassisch nach obigen Formeln, in der zweiten Variante fließt in die Berechnung des AIC, des BIC und des FPE nicht das in (3.41) ermittelte $\hat{\sigma}^2$ ein, sondern es wird vorher noch die Quadratwurzel von $\hat{\sigma}^2$ gezogen und diese durch das arithmetische Mittel der Wertereihe dividiert[13];

[13]Da vorher Werte quadriert werden, muss vor der Division durch das arithmetische Mittel dieser Quadrierungseffekt durch das Wurzelziehen wieder wettgemacht werden. Es wäre nicht zielführend, zur Vermeidung des Wurzelziehens lediglich die Absolutbeträge der Differenzen von x_i und $\hat{x_i}$ (und nicht deren Quadrate) in die Berechnung von $\hat{\sigma}^2$ einfließen zu lassen, da in diesem Falle Datenreihen mit insgesamt gleicher Differenz vom Mittelwert die gleiche Beurteilung erhielten, obwohl eine Datenreihe mit bei allen (vielen) Werten vorhandenen relativ kleinen Varianzen wesentlich besser durch das Modell repräsentiert bzw. approximiert wäre als eine andere Datenreihe, die einige Ausreißer mit großer Varianz und einige Werte mit minimaler bis keiner Varianz hätte; beispielsweise wären dann bei Vorhersage $\hat{x_i} = \mu_x$ Datenreihen $d_1 = (6,6,6,1,11,1,11,6,6,6)$ und $d_2 = (4,8,4,8,4,8,4,8,4,8)$ gleichgestellt, obwohl d_2 klar besser dem Modell entspricht.

der gesamte Programmcode von Variante I (mit den entsprechenden Ergänzungskommentaren im Falle von Variante II) ist in Abbildung A.1 des Anhanges aufgeführt.

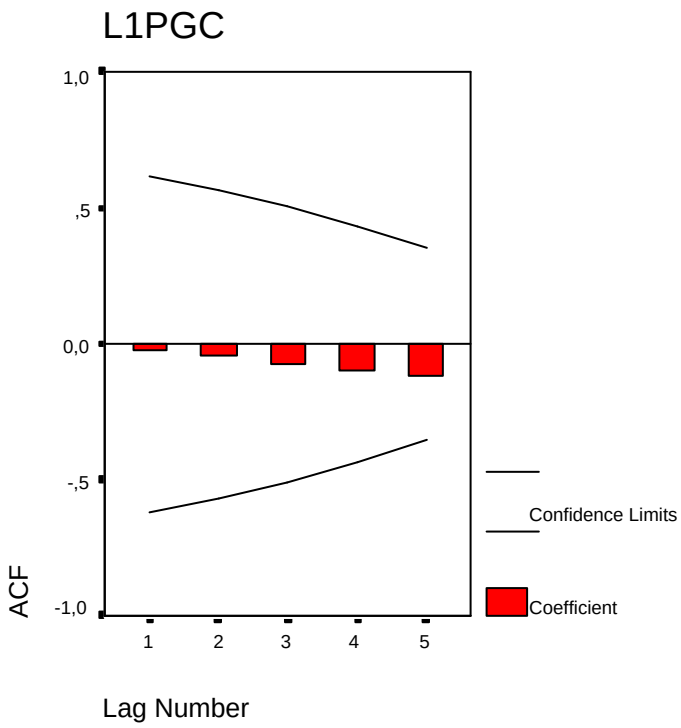
Die Methode der Autoregression wird anhand eines UPDATE-Strukturbruchs in Tabelle 3.14 vorgestellt, da diese Werte als stationär betrachtet werden können und somit nicht vorher Differenzoperationen angewendet werden müssen. Zunächst werden die Korrelogramme inspiziert,

Gew	EUR	$Jahr_{t-6}$	$Jahr_{t-5}$	$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	202	205	203	206	208	204	**808**
$Land_1$	PG_B	155	156	158	157	155	159	**612**
$Land_1$	PG_C	689	692	690	686	690	692	**2798**
$Land_2$	PG_A	16	17	18	17	19	18	20
$Land_2$	PG_B	13	14	15	13	16	17	14
$Land_2$	PG_C	86	87	88	85	89	88	86
$Land_3$	PG_A	353	356	350	363	354	357	353
$Land_3$	PG_B	310	307	312	308	312	311	310
$Land_3$	PG_C	989	993	991	990	995	987	989

Legende: PG=Produktgruppe, Gew=Gewinn in EURO

Tabelle 3.14: Änderung der Bedeutung des Schlüssels eines Landes

um mit visueller Unterstützung ein korrektes Modell auswählen zu können. Untersucht man beispielsweise lediglich die in den Abbildungen 3.7, 3.8 und 3.9 dargestellten Korrelogramme der Strukturkombinationen, an denen Produktgruppe PG_C beteiligt ist, so fällt es auf den ersten Blick schwer, Gemeinsamkeiten zu entdecken. Allerdings erkennt man bei genauerer Betrachtung, dass die Autokorrelationen von $Land_2$ und $Land_3$ relativ ähnlich sind - die Autokorrelationskoeffizienten sind bis auf den Wert zu $\tau = 3$ stets negativ, und sie nehmen insgesamt betragsmäßig eher ab; im Falle von $Land_1$ sind die Autokorrelationen jedoch stets negativ, und die Beträge der Werte nehmen mit zunehmendem Lag τ zu. Da kein abruptes Ende der Korrelationen zu Lag q festgestellt werden kann, sollte prinzipiell ein AR(p)-Prozess gewählt werden. Es existieren bei den untersuchten Korrelogrammen positive und negative Korrelationen, jedoch nicht in alternierender Weise, womit ein AR(1)-Modell nicht gerechtfertigt ist. Zieht man einen AR(3)-Prozess mit $\phi_1 = 0.6$, $\phi_2 = 0.6$ und $\phi_3 = -0.2$ als Referenzmodell heran, so ergeben sich beim Vergleich aller Strukturkombinationen in den beiden erwähnten Varianten (klassische Kennzahlenberechnung und adaptierte Berechnung) über alle Jahre die in den Tabellen 3.15 (klassische Variante) und 3.16 (adaptierte Variante) aufgeführten Ergebnisse. Es zeigt sich, dass die Berechnung der Kennzahlen auf klassische Weise die Ausreißer deutlicher zum Ausdruck bringt: die *Final Prediction Errors* der Strukturkombinationen mit Beteiligung von $Land_1$ sind um Zehnerpotenzen höher als die der anderen Kombinationen; auch die Werte der beiden anderen Kriterien sind für $Land_1$ höher - hier fallen nur die Unterschiede in Absolutbeträgen nicht so deutlich aus,



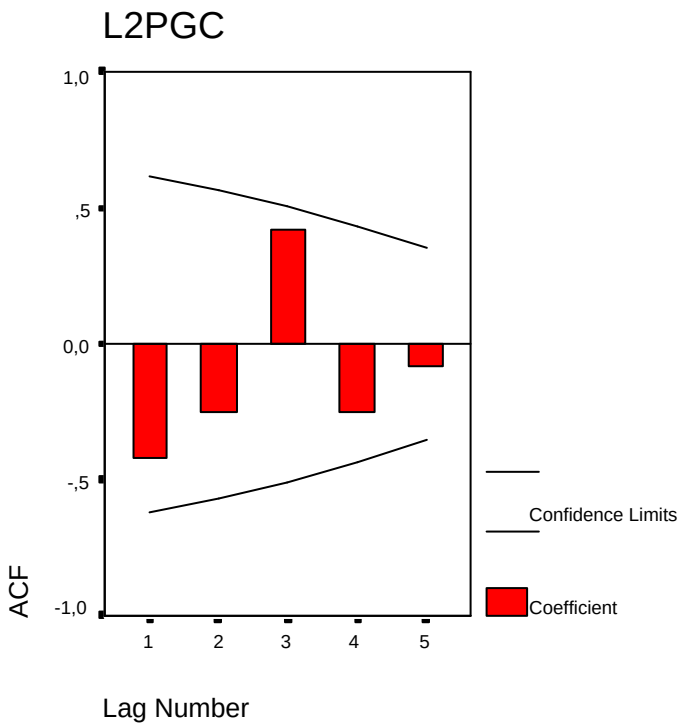
Legende: $L1PGC$=Strukturkombination Land 1 und Produktgruppe C;
ACF=Autokorrelationskoeffizienten zu Lags 1 bis 5; Confidence limits zeigen die Grenzen an, ab deren Überschreitung der ACF-Werte die Hypothese der Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegt werden kann

Abbildung 3.7: Korrelogramm der Strukturkombination $Land_1PG_C$ aus Tabelle 3.14

$AR(3)$	AIC_{cl}	BIC_{cl}	FPE_{cl}
$Land_1PG_A$	**10**	**10**	**226 060**
$Land_1PG_B$	**10**	**10**	**129 450**
$Land_1PG_C$	**14.8**	**14.1**	**2 770 700**
$Land_2PG_A$	0.84	0.2	2.48
$Land_2PG_B$	2.43	1.79	12.05
$Land_2PG_C$	2.77	2.12	16.9
$Land_3PG_A$	4.44	3.8	89.9
$Land_3PG_B$	1.69	1.05	5.7
$Land_3PG_C$	3.84	3.2	49.37

Legende: PG=Produktgruppe

Tabelle 3.15: Kennzahlenberechnung von AIC, BIC und FPE auf klassische Weise über alle Jahre



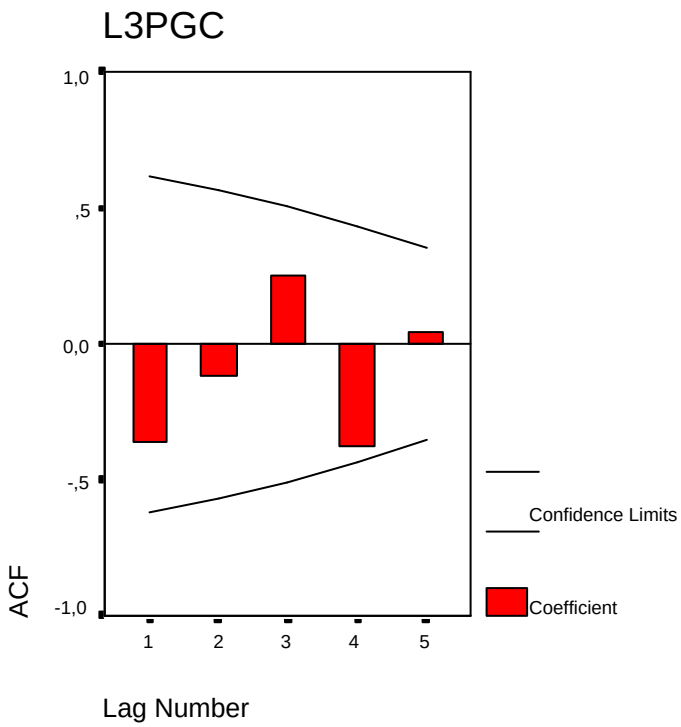
Legende: $L2PGC$=Strukturkombination Land 2 und Produktgruppe C;
ACF=Autokorrelationskoeffizienten zu Lags 1 bis 5; Confidence limits zeigen die Grenzen an, ab deren Überschreitung der ACF-Werte die Hypothese der Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegt werden kann

Abbildung 3.8: Korrelogramm der Strukturkombination $Land_2PG_C$ aus Tabelle 3.14

$AR(3)$	AIC_{ad}	BIC_{ad}	FPE_{ad}
$Land_1PG_A$	**0.89**	**0.23**	**2.58**
$Land_1PG_B$	**0.88**	**0.24**	**2.56**
$Land_1PG_C$	**0.91**	**0.28**	**2.65**
$Land_2PG_A$	-1.77	-2.41	0.18
$Land_2PG_B$	-0.83	-1.46	0.46
$Land_2PG_C$	-2.63	-3.27	0.001
$Land_3PG_A$	-3.28	-3.91	0.04
$Land_3PG_B$	-4.1	-4.75	0.01
$Land_3PG_C$	-4.6	-5.25	0.01

Legende: PG=Produktgruppe

Tabelle 3.16: Adaptierte Kennzahlenberechnung von AIC, BIC und FPE über alle Jahre



Legende: $L3PGC$=Strukturkombination Land 3 und Produktgruppe C;
ACF=Autokorrelationskoeffizienten zu Lags 1 bis 5; Confidence limits zeigen die Grenzen an, ab deren Überschreitung der ACF-Werte die Hypothese der Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegt werden kann

Abbildung 3.9: Korrelogramm der Strukturkombination $Land_3PG_C$ aus Tabelle 3.14

da die logarithmierten Varianzen $\hat{\sigma^2}$ in die Berechnung einfließen. Skaliert man hingegen die $\hat{\sigma^2}$-Werte, wie in der zweiten Tabelle dargestellt, so erscheinen die Unterschiede nicht mehr so groß: die FPE-Werte der ersten drei Zeilen betragen 'nur' mehr rund das Fünffache des maximalen FPE der anderen Zeilen, die Absolutbeträge der Differenzen des AIC und des BIC sind noch geringer; aufgrund des vorher erfolgten Logarithmierens von $\hat{\sigma^2}$ haben aber auch geringere Absolutdifferenzen relativ hohe Aussagekraft - alleine bei Betrachtung des Vorzeichens erkennt man, dass die ersten drei Zeilen Ausreißer sein müssten[14].

Untersucht man sämtliche Werte von Strukturkombinationen aus Tabelle 3.14 nur im Zeitintervall $[Jahr_{t-6}; Jahr_{t-1}]$ wiederum mit dem gleichen AR(3)-Modell mit $\phi_1 = \phi_2 = 0.6$ und $\phi_3 = -0.2$, so ergeben sich in den beiden Varianten die in den Tabellen 3.17 und 3.18 dargestell-

[14] Adaptiert man die zweite Variante dahingehend, dass man die Logarithmen von $\hat{\sigma^2}$ nicht zieht, dann sind die Absolutbeträge der Differenzen des AIC und des BIC nicht größer; durch die nun überall positiven Vorzeichen wird die Aussagekraft der Ergebnisse sogar noch reduziert.

ten Resultate: vergleicht man nun die Ergebnisse der beiden Berechnungsvarianten, so fallen

$AR(3)$	AIC_{cl}	BIC_{cl}	FPE_{cl}
$Land_1PG_A$	3.2785	2.5576	29.2869
$Land_1PG_B$	2.7463	2.0254	17.2005
$Land_1PG_C$	3.6764	2.9555	43.5978
$Land_2PG_A$	0.4976	-0.2233	1.8153
$Land_2PG_B$	2.9128	2.1918	20.3157
$Land_2PG_C$	3.0103	2.2894	22.3974
$Land_3PG_A$	**4.5926**	**3.8716**	**108.9811**
$Land_3PG_B$	1.5722	0.8513	5.3167
$Land_3PG_C$	4.1850	3.4641	72.5020

Legende: PG=Produktgruppe

Tabelle 3.17: Kennzahlenberechnung von AIC, BIC und FPE auf klassische Weise (Analyseintervall: $Jahr_{t-6}$ bis $Jahr_{t-1}$)

$AR(3)$	AIC_{ad}	BIC_{ad}	FPE_{ad}
$Land_1PG_A$	-3.2605	-3.9814	0.0423
$Land_1PG_B$	-3.0633	-3.7842	0.0516
$Land_1PG_C$	-4.0533	-4.7742	0.0192
$Land_2PG_A$	-1.8341	-2.5551	0.1763
$Land_2PG_B$	**-0.8954**	**-1.6163**	**0.4508**
$Land_2PG_C$	-2.9046	-3.6255	0.0604
$Land_3PG_A$	-2.9905	-3.7114	0.0555
$Land_3PG_B$	-4.1558	-4.8768	0.0173
$Land_3PG_C$	-4.3411	-5.0621	0.0144

Legende: PG=Produktgruppe

Tabelle 3.18: Adaptierte Kennzahlenberechnung von AIC, BIC und FPE (Analyseintervall: $Jahr_{t-6}$ bis $Jahr_{t-1}$)

in beiden Tabellen die Werte der ersten drei Zeilen überhaupt nicht mehr als Ausreißer auf - der Strukturbruch muss somit zwischen $Jahr_{t-1}$ und $Jahr_t$ stattgefunden haben. Interessant ist aber, dass bei Betrachtung der ersten Tabelle die Strukturkombination $Land_3PG_A$ mit einem FPE von über 108 und auch Maximalwerten in den beiden anderen Kategorien AIC und BIC als Ausreißer interpretiert werden könnte; da in der zweiten Tabelle diese Kombination überhaupt nicht mehr auffällt, sind die hohen Werte aus der ersten Tabelle nicht aufgrund eines Strukturbruchs entstanden, sondern eher durch die relativ größeren Werte bedingt gewesen. In der zweiten Tabelle könnte eher $Land_2PG_B$ als Ausreißer erkannt werden, da diese Strukturkombination in allen Kategorien mit relativ deutlichem Abstand die größten Werte ausweist - dies lässt sich wiederum mit den betragsmäßig kleinen Werten von $Land_2PG_B$ erklären; ein

Strukturbruch hat auch hier nicht stattgefunden. Als conclusio dieses Kapitels kann daher festgehalten werden, dass Differenzen in einzelnen Kennzahlen sehr sorgfältig interpretiert werden müssen, da durch die Wahl der Berechnungsformeln die Resultate beachtlich verändert werden können. Bei der Identifikation eines Strukturbruchs gilt es daher, einerseits die Ergebnisse verschiedener Berechnungsmethoden von Kennzahlen miteinander zu vergleichen und andererseits erst ab Differenzen in bedeutenden Größenordnungen Schlüsse auf Strukturveränderungen zu ziehen.

3.6 Differenzialgleichungen

Eine weitere Möglichkeit zur Aufdeckung von Strukturbrüchen bietet die Modellierung einer Datenreihe durch eine Differenzialgleichung; ähnlich wie bei der Methode der Autoregression ist auch in diesem Fall die Wahl des korrekten Differenzialgleichungsmodells von größter Bedeutung. Als Metrik der Erkennung von Strukturbrüchen kann die Differenz zwischen dem von der Funktion, die die Lösung der Differenzialgleichung darstellt, prognostizierten Wert und dem tatsächlichen Wert in einem beliebigen Chronon t herangezogen werden. Bevor das Verfahren an einem Beispiel vorgestellt wird, seien einige grundlegende Definitionen, die im folgenden verwendet werden, erwähnt (vgl. u.a. [Di00]):

Definition 3.6.1 *Eine Differenzialgleichung heißt gewöhnlich, wenn lediglich die Funktionen und Ableitungen einer Variable auftreten*[15]*.*

Definition 3.6.2 *Eine gewöhnliche Differenzialgleichung ist n-ter Ordnung, wenn n die höchste vorkommende Ableitung der Funktion ist. In allgemeiner Form ist diese Differenzialgleichung dargestellt als*

$$x^{(n)} = f(t, x, x', x'', \ldots, x^{(n-1)}). \tag{3.44}$$

Definition 3.6.3 *Eine gewöhnliche Differenzialgleichung n-ter Ordnung gemäß (3.44), für die die zusätzlichen Bedingungen*

$$x(t_0) = x_0, \;\; x'(t_0) = x'_0, \ldots, \; x^{(n-1)}(t_0) = x_0^{(n-1)} \tag{3.45}$$

gelten, wobei t_0, x_0, …, $x_0^{(n-1)}$ vorgegebene Zahlen sind, heißt Anfangswertproblem von (3.44).

Definition 3.6.4 *Eine Lösung der Differenzialgleichung aus (3.44) ist eine Funktion φ, die auf*

[15] Bei der hier verwendeten Modellierung von Datenreihen werden nur gewöhnliche Differenzialgleichungen verwendet.

einem offenen Intervall definiert ist, mindestens n-fach ableitbar ist und die Gleichung

$$\varphi^{(n)}(t) = f(t, \varphi(t), \varphi'(t), \ldots, \varphi^{(n-1)}(t)) \tag{3.46}$$

erfüllt. Eine Lösung des Anfangswertproblems aus (3.45) muss zudem den Bedingungen $\varphi(t_0) = x_0, \varphi'(t_0) = x_0', \ldots, \varphi^{(n-1)}(t_0) = x_0^{(n-1)}$ genügen.

Definition 3.6.5 *Eine gewöhnliche Differenzialgleichung erster Ordnung heißt separabel, wenn sie in der Form*

$$x'(t) = g(t)\, h(x),\ h(x) \neq 0, \tag{3.47}$$

darstellbar ist.

In diesem Abschnitt wird zunächst ein sehr einfaches Modell herangezogen, um die Werte aus Tabelle 2.5 abzubilden. Basierend auf dem Hintergrundwissen, dass das Gewinnwachstum jeder Produktgruppe in jedem Land jährlich rund 1% betragen sollte, kann jeder Datenvektor mit einer gewöhnlichen Differenzialgleichung erster Ordnung folgendermaßen modelliert werden:

$$x'(t) = 0.01\ x(t). \tag{3.48}$$

Da diese Gleichung separabel ist ($g(t) = 0.01$, $h(x) = x(t)$[16]), erhält man ihre allgemeine Lösung durch direkte Integration:

$$\begin{aligned} \tfrac{x'}{x} &= 0.01, \\ \int \tfrac{x'}{x}\, dt &= \int 0.01\, dt \end{aligned}$$

Wegen $\int x' dt = \int dx$ gilt somit:

$$\begin{aligned} \int \tfrac{dx}{x} &= \int 0.01\, dt \\ \ln\ |x| &= 0.01t + c \\ x &= e^{0.01t} k, \end{aligned}$$

wobei c die Kombination der Integrationskonstanten auf der linken und rechten Seite der Gleichung darstellt; $k = e^c$ für $x > 0$ und $k = e^{-c}$ für $x < 0$; $k \neq 0$. Für jeden Datenvektor kann durch den Wert des ersten aufgezeichneten Chronons in $Jahr_{t-4}$ (Anfangsbedingung $x(0) = Value(Jahr_{t-4})$) die spezielle Lösung für k gefunden werden: exemplifiziert gilt in der Strukturkombination $Land_1 PG_A$ wegen $x(0) = 202$ somit:

$$x(0) = e^{0.02*0} k = 202,$$

[16]Im folgenden wird die Abhängigkeit der Funktion x von t aus Einfachheitsgründen weggelassen.

und die spezielle Lösung dieses Problems lautet daher

$$x(t) = 202e^{0.02t}.$$

Zum Erkennen von Strukturbrüchen eignet sich der Absolutbetrag der Differenz zwischen dem mittels der Differenzialgleichung prognostizierten Wert und dem tatsächlichen Wert des Datenvektors i zum Zeitpunkt t skaliert durch den jeweiligen Mittelwert des Vektors. Die zu den Daten aus Tabelle 2.5 gehörige Differenzmatrix ist in Tabelle 3.19 abgebildet. Bei Betrachtung

ScaledDiff$\times 10^3$		$Jahr_{t-4}$	$Jahr_{t-3}$	$Jahr_{t-2}$	$Jahr_{t-1}$	$Jahr_t$
$Land_1$	PG_A	0	2.98	0.25	0.47	**1834.7**
$Land_1$	PG_B	0	2.25	0.53	2.91	**1817.2**
$Land_1$	PG_C	0	3.52	8	8.06	**1866.6**
$Land_2$	PG_A	0	46.62	93.15	139.6	185.95
$Land_2$	PG_B	0	57.96	115.83	173.6	231.3
$Land_2$	PG_C	0	1.53	2.97	4.3	28.17
$Land_3$	PG_A	0	1.52	7.94	6.23	1.64
$Land_3$	PG_B	0	6.72	4.01	4.58	11.6
$Land_3$	PG_C	0	5.97	14.06	22.24	30.53

Legende: PG=Produktgruppe, ScaledDiff=Absolutbetrag der Differenz der Gewinne skaliert durch den jeweiligen Mittelwert der Land-/Produktgruppen-Kombination

Tabelle 3.19: Differenzen zwischen dem mittels der Differenzialgleichung prognostizierten und dem tatsächlichen Wert zum Zeitpunkt t

des Ergebnisses fallen die Ausreißer in der letzten Kolumne der ersten drei Zeilen auf - die Werte liegen um knapp eine Zehnerpotenz höher als das Maximum der anderen Werte; der offensichtliche Strukturbruch kann somit auch mit dieser Methode erkannt werden.

Die Mächtigkeit der Methode der Modellierung einer Datenreihe mittels einer Differenzialgleichung kommt noch stärker zum Ausdruck, wenn man das Vorwissen hat, dass die generelle Entwicklung von Wertereihen nicht ein stetiger Aufwärts-, Abwärts- oder Stagnationstrend war, sondern aus stückweise stetigen Trends besteht; bevor an einem konkreten Beispiel eine solche stückweise stetige Funktion erläutert wird, sind zunächst zusätzliche Definitionen notwendig, um die entsprechende Differenzialgleichung lösen zu können.

Definition 3.6.6 *Eine Funktion f heißt stetig auf einem Intervall I, wenn sie für alle Punkte $a \in I$ stetig ist, i.e. wenn gilt:*

$$\lim_{x \to a} f(x) = f(a), \quad \forall a \in I.$$

Definition 3.6.7 *Die Laplace-Transformation $\mathscr{L}[f]$ einer zumindest stückweise stetigen Funktion*

f ist definiert als:

$$\mathscr{L}[f] = F(s) = \int_0^\infty e^{-st} f(t)dt,$$

wobei F(s) für alle s, für die das Integral $\int_0^\infty e^{-st} f(t)$ konvergiert, definiert ist; i.e. für alle s, für die $lim_{A\to\infty} \int_0^A e^{-st} f(t)dt$ existiert und endlich ist.

Beispiel 3.6.1 *Die Laplace-Transformation $\mathscr{L}[f]$ der konstanten Funktion $f(t) = 1,\ t \geq 0$, ist*

$$\mathscr{L}[1] = F(s) = \int_0^\infty e^{-st}dt = \lim_{A\to\infty}(-\frac{e^{-st}}{s}\ |_0^A) = 0 - (-\frac{1}{s}) = \frac{1}{s}$$

und ist definiert für $s > 0$, da ein negativer Wert von s einen positiven Wert im Exponenten von e nach sich ziehen würde; $\lim_{A\to\infty}(-\frac{e^{-st}}{s}\ |_0^A)$ wäre somit unendlich; bei $s = 0$ ist $\frac{1}{s}$ nicht definiert.

Die Laplace-Transformation besitzt zahlreiche wichtige Eigenschaften, die für das Lösen von Differenzialgleichungen von Bedeutung sind [Di00]:

Definition 3.6.8 *[Linearität] Seien f und g zwei Funktionen mit Laplace-Transformationen $\mathscr{L}[f] = F(s)$ und $\mathscr{L}[g] = G(s)$, und seien c_1 und c_2 zwei reelle Konstanten, so gilt:*

$$\mathscr{L}[c_1 f + c_2 g] = c_1\mathscr{L}[f] + c_2\mathscr{L}[g]$$

Definition 3.6.9 *[Ableitungstheorem] Sei f eine zumindest stückweise stetige, differenzierbare Funktion mit der Laplace-Transformation $\mathscr{L}[f] = F(s)$, und sei f' zumindest stückweise stetig, dann gilt:*

$$\mathscr{L}[f'] = s\mathscr{L}[f] - f(0).$$

Sofern auch höhere Ableitungen existieren und stetig sind, lässt sich die Formel für höhere Ableitungen verallgemeinern:

$$\mathscr{L}[f^{(n)}] = s^n\mathscr{L}[f] - s^{n-1}f(0) - s^{n-2}f'(0) - \ldots sf^{(n-2)}f(0) - f^{(n-1)}(0)$$

Definition 3.6.10 *[Invertierbarkeit] Sei $\mathscr{L}[f] = F(s)$ die Laplace-Transformation von f, so existiert auch die Inverse $\mathscr{L}^{-1}[F]$, die ebenso linear ist.*

Die einfachste stückweise stetige Funktion, durch die sich sämtliche andere stückweise stetigen Funktionen ausdrücken lassen, die *Heaviside*-Funktion u_c, ist definiert als:

$$u_c(t) \quad = \quad \begin{cases} 0, & t < c \\ 1, & t \geq c, \end{cases} \qquad c \geq 0. \tag{3.49}$$

Eine beliebige stückweise stetige Funktion

$$f(t) \quad = \quad \begin{cases} a_1, & t \in [0, c_1) \\ a_2, & t \in [c_1, c_2) \\ a_3, & t \in [c_2, c_3), \\ \vdots & \end{cases}$$

kann demnach geschrieben werden als

$$f(t) = a_1 + (a_2 - a_1)u_{c1}(t) + (a_3 - a_2)u_{c2}(t) + \dots . \tag{3.50}$$

Die Laplace-Transformationen der wichtigsten elementaren Funktionen sind in Tabelle 3.20 dargestellt.

$f(t)$	$\mathscr{L}[f] = F(s)$	$Domain(f)$
1	$\frac{1}{s}$	$s > 0$
e^{at}	$\frac{1}{s-a}$	$s > a$
$\sin at$	$\frac{a}{s^2+a^2}$	$s > 0$
$\cos at$	$\frac{s}{s^2+a^2}$	$s > 0$
t^n, $n \in \mathbb{N}^+$	$\frac{n!}{s^{n+1}}$	$s > 0$
$u_c(t)$	$\frac{e^{-cs}}{s}$	$s > 0$

Legende: $f(t)$=elementare Funktion, $\mathscr{L}[f] = F(s)$=Laplace-Transformation der elementaren Funktion, $Domain(f)$=Definitionsbereich der Laplace-Transformation

Tabelle 3.20: Laplace-Transformationen wichtiger elementarer Funktionen

Ein weiteres Theorem im Zusammenhang mit der $Heaviside$-Funktion ist das folgende:

Definition 3.6.11 (Verschiebbarkeit) *Sei $f(t)$ eine Funktion mit der Laplace-Transformation $\mathscr{L}[f] = F(s)$ für $s > a$. Dann gilt:*

$$\mathscr{L}[e^{ct} f(t)] = F(s - c), \dots, für \;\; s > a + c$$

Für Laplace-Transformationen von Funktionen in Kombination mit der $Heaviside$-Funktion $u_c(t)$ gilt [Di00]:

$$\mathscr{L}[u_c(t) f(t - c)] = e^{-cs} F(s). \tag{3.51}$$

Ein Anfangswertproblem mit einer stückweise stetigen Differenzialgleichung kann mittels der Laplace-Transformation in folgenden Schritten gelöst werden:

1. Anwenden der Laplace-Transformation auf beiden Seiten der Differenzialgleichung - aus der Differenzialgleichung wird somit eine algebraische Gleichung.

2. Einsetzen der Anfangsbedingung und Lösen der aus obigem Schritt erhaltenen Gleichung.

3. Anwenden der inversen Laplace-Transformation zum Rücktransferieren der algebraischen Gleichungslösung in eine Lösung der Differenzialgleichung.

Das Verfahren soll an den Daten aus Tabelle 3.21 erprobt werden: durchschnittlich kann davon

Gewinn	$Jahr_0$	$Jahr_1$	$Jahr_2$	$Jahr_3$	$Jahr_4$	$Jahr_5$	$Jahr_6$	$Jahr_7$	$Jahr_8$
PG_A	100	105	107	100	95	93	86	91	96
PG_B	200	201	203	194	186	177	189	202	225
PG_C	300	306	310	292	275	259	282	302	331
PG_D	400	409	419	408	401	376	403	430	462
PG_E	500	503	505	488	466	450	485	532	574

Legende: PG=Produktgruppe

Tabelle 3.21: Entwicklung der Gewinne einzelner Produktgruppen

ausgegangen werden, dass die Produktgewinne zwischen $Jahr_0$ und $Jahr_8$ jährlich um 2% steigen, wobei allerdings der Faktor des allgemeinen Wirtschaftswachstums berücksichtigt werden muss - im Zeitraum $Jahr_0 - Jahr_2$ stagnierte dieses, von $Jahr_3$ bis $Jahr_5$ gab es einen jährlichen Rückgang um 5%, von $Jahr_6$ bis $Jahr_8$ hingegen wuchs die Wirtschaft um 5%. Die allgemeine Differenzialgleichung für diese Daten lautet daher:

$$x' = 0.02x + g, \tag{3.52}$$

und g ist definiert als

$$g = \begin{cases} 0, & t \in [0,2], \\ -0.05\mu, & t \in (2,5], \\ 0.05\mu, & t \in (5,\infty), \end{cases}$$

wobei μ den Mittelwert des jeweiligen Datenvektors repräsentiert[17].

Im speziellen Fall der Produktgruppe PG_A ist $\mu = 97$, und somit lautet die Differenzialgleichung dieser Produktgruppe

$$x' = 0.02x + g, \;\; g = \begin{cases} 0, & t \in [0,2], \\ -0.05 * 97, & t \in (2,5], \\ 0.05 * 97, & t \in (5,\infty). \end{cases}$$

[17]Gemäß (3.50) müssten die Intervalle von g in diesem Fall $[0,3), [3,6), [6,\infty)$ anstelle von $[0,2], (2,5], (5,\infty)$ lauten. Es wird aber davon ausgegangen, dass die jeweiligen Trends nicht erst ganz kurz vor den Jahren drei und sechs aufhören, sondern sofort nach den Jahren zwei und fünf zu Ende sind - aus diesem Grund wurde diese Modellierung gewählt.

Die Gleichung kann gemäß (3.50) geschrieben werden als

$$x' = 0.02x - 4.85u_2(t) + 9.7u_5(t), \;\; x(0) = 100.$$

Unter Zuhilfenahme des Ableitungstheorems von oben wird nun die Laplace-Transformation $\mathscr{L}[f] = X(s)$ angewandt:

$$\begin{array}{rcl} sX(s) - x(0) & = & 0.02X(s) - \frac{4.85e^{-2s}}{s} + \frac{9.7e^{-5s}}{s} \\ (s - 0.02)X(s) & = & -\frac{4.85e^{-2s}}{s} + \frac{9.7e^{-5s}}{s} + x(0) \end{array}$$

Setzt man die Anfangsbedingung $x(0) = 100$ ein, so kann die nunmehr algebraische Gleichung gelöst werden:

$$\begin{array}{rcl} (s - 0.02)X(s) & = & -\frac{4.85e^{-2s}}{s} + \frac{9.7e^{-5s}}{s} + 100 \\ X(s) & = & \frac{-4.85e^{-2s}+9.7e^{-5s}+100s}{s(s-0.02)} \\ X(s) & = & \frac{-4.85e^{-2s}}{s(s-0.02)} + \frac{9.7e^{-5s}}{s(s-0.02)} + \frac{100}{s-0.02} \end{array}$$

Nun wird die inverse Laplace-Transformation unter Ausnützung ihrer Linearitätseigenschaft, der Verschiebbarkeitseigenschaft der *Heaviside*-Funktion gemäß (3.51), der Partialbruchzerlegung $\frac{a}{s(s-k)} = \frac{a}{k(s-k)} - \frac{a}{sk}$, $a, k \in \mathbb{R}$, $k \neq 0$, unter Zuhilfenahme von Tabelle 3.20 durchgeführt:

$$\begin{array}{rcl} x(t) & = & \mathscr{L}^{-1}[\frac{-4.85e^{-2s}}{0.02(s-0.02)}] + \mathscr{L}^{-1}[\frac{4.85e^{-2s}}{0.02s}] + \\ & + & \mathscr{L}^{-1}[\frac{9.7e^{-5s}}{0.02(s-0.02)}] - \mathscr{L}^{-1}[\frac{9.7e^{-5s}}{0.02s}] + 100\mathscr{L}^{-1}[\frac{1}{s-0.02}] \\ x(t) & = & \mathscr{L}^{-1}[\frac{-242.5e^{-2s}}{s-0.02}] + \mathscr{L}^{-1}[\frac{242.5e^{-2s}}{s}] + \\ & + & \mathscr{L}^{-1}[\frac{485e^{-5s}}{s-0.02}] - \mathscr{L}^{-1}[\frac{485e^{-5s}}{s}] + 100\mathscr{L}^{-1}[\frac{1}{s-0.02}] \\ x(t) & = & -242.5\mathscr{L}^{-1}[\frac{e^{-2s}}{s-0.02}] + 242.5\mathscr{L}^{-1}[\frac{e^{-2s}}{s}] + \\ & + & 485\mathscr{L}^{-1}[\frac{e^{-5s}}{s-0.02}] - 485\mathscr{L}^{-1}[\frac{e^{-5s}}{s}] + 100\mathscr{L}^{-1}[\frac{1}{s-0.02}] \\ x(t) & = & -242.5u_2(t)e^{0.02(t-2)} + 242.5u_2(t) + 485u_5(t)e^{0.02(t-5)} - 485u_5(t) + 100e^{0.02t} \\ x(t) & = & 100e^{0.02t} + (242.5 - 242.5e^{0.02t-0.04})u_2(t) + (485e^{0.02t-0.1} - 485)u_5(t) \end{array}$$

Die letzte Zeile kann gemäß (3.50) umgeschrieben werden in:

$$x(t) = \begin{cases} 100e^{0.02t}, & t \in [0, 2], \\ 100e^{0.02t} - 242.5e^{0.02t-0.04} + 242.5, & t \in (2, 5], \\ 100e^{0.02t} + 485e^{0.02t-0.10} - 242.5e^{0.02t-0.04} - 242.5, & t \in (5, \infty). \end{cases}$$

Analog werden die Koeffizienten der Differenzialgleichungen der anderen Produktgruppen aus Tabelle 3.21 errechnet; als Differenzmetrik zur Erkennung von Strukturbrüchen wird wie oben

$$\frac{1}{\mu} \ |\Delta(Istwert, Modellwert)|$$

herangezogen. Das Ergebnis ist in Tabelle 3.22 dargestellt: es fällt auf, dass in Produktgruppe

$\frac{\lvert\Delta\rvert}{\mu} * 10^3$	$Jahr_0$	$Jahr_1$	$Jahr_2$	$Jahr_3$	$Jahr_4$	$Jahr_5$	$Jahr_6$	$Jahr_7$	$Jahr_8$
PG_A	0	30.72	30.09	13.25	35.38	26.00	**168.56**	**188.83**	**210.55**
PG_B	0	15.40	26.14	42.52	53.24	68.43	77.65	83.22	39.58
PG_C	0	0.20	7.60	39.43	67.29	91.16	83.32	87.06	61.75
PG_D	0	2.23	6.49	9.89	23.59	5.77	9.29	14.21	8.41
PG_E	0	14.19	30.79	35.28	49.16	50.44	50.18	27.34	15.93

Legende: PG=Produktgruppe, $\frac{\lvert\Delta\rvert}{\mu}$=Absolutbetrag der Differenz zwischen tatsächlichem und prognostiziertem Wert skaliert durch den Mittelwert der jeweiligen Produktgruppe

Tabelle 3.22: Skalierte Differenzen zwischen dem tatsächlichen und dem prognostizierten Wert der Daten aus Tabelle 3.21

PG_A die relativen Fehlprognosen für $Jahr_6$, $Jahr_7$ und $Jahr_8$ deutlich höher sind als in den anderen Produktgruppen - während alle anderen Produktgruppen Werte von teilweise deutlich unter 100 aufzuweisen haben, hat sich bei PG_A offensichtlich in $Jahr_6$ ein Strukturbruch ereignet, dessen Auswirkungen in $Jahr_7$ und $Jahr_8$ noch verstärkt zum Ausdruck kommen. Tatsächlich ist im Falle von PG_A in $Jahr_6$ entgegen dem allgemeinen Wachstumstrend von 5% ein Rückgang von 5% vorgefallen - in den nächsten Jahren sind die Werte zwar wieder in normalem Ausmaß im Vergleich zum Vorjahr gestiegen, doch aufgrund der 'schlechteren Ausgangslage' aus $Jahr_6$ sind die Differenzen zwischen dem tatsächlichen und dem prognostizierten Wert noch immer bedeutend höher als in den anderen Produktgruppen.

Zusammenfassend kann daher festgehalten werden, dass die Methode der Modellierung einer Datenreihe mittels Differenzialgleichungen eine sehr mächtige, vielseitig einsetzbare Vorgangsweise ist; der Nachteil ist darin zu sehen, dass es eines allgemeinen Vorwissens bedarf, um ein für eine bestimmte Zeitreihe adäquates Modell aufstellen zu können.

3.7 Eigen- und Singulärwertzerlegung

In diesem Abschnitt wird zunächst die Eigenwertzerlegung definiert, da sich die Singulärwertzerlegung unmittelbar aus dieser ableiten lässt. Für die Analyse von Datenbeständen in Data Warehouses ist jedoch primär die Singulärwertzerlegung von Bedeutung, da lediglich in den seltensten Fällen die Datenmatrizen quadratisch sind.

3.7.1 Eigenwertzerlegung

Die Zahl $\lambda \in \mathbb{E}$ ($\mathbb{E} := \mathbb{R}$ oder $\mathbb{C}$) heißt Eigenwert einer quadratischen Matrix $A \in \mathbb{R}^{n \times n}$ und der Vektor $\vec{x}$ heißt Eigenvektor von A, falls gilt:

$$A\vec{x} = \lambda\vec{x}. \tag{3.53}$$

Die Eigenwerte der Matrix können beispielsweise berechnet werden als Nullstellen des charakteristischen Polynoms $\det(A - \lambda I)$ der Matrix A, wobei I die Identitätsmatrix in $\mathbb{R}^{n \times n}$ mit 1ern auf der Hauptdiagonale und Nullwerten in allen anderen Zeilen-Spalten-Kombinationen repräsentiert, und $\det(A-\lambda I)$ ein Polynom n-ten Grades ist und dementsprechend n Nullstellen besitzt; die Nullstellen müssen nicht alle verschieden sein, und sie können auch komplex sein[18]. Die zu den Nullstellen bzw. Eigenwerten zugehörigen Eigenvektoren sind zueinander orthogonal (linear unabhängig), sämtliche Eigenvektoren zusammen bilden die *Eigenbasis* der Matrix.

Sei k der Rang der Matrix A, dann lässt sich A darstellen als:

$$A = U\Lambda U^T, \tag{3.54}$$

wobei Λ eine $\mathbb{E}^{k \times k}$-Diagonalmatrix mit den Eigenwerten auf der Diagonalen und U eine $\mathbb{E}^{n \times k}$-spaltenorthogonale Matrix ist (es gilt daher: $U^T U = I$), deren Spalten die Eigenvektoren von A repräsentieren [Gu02].

3.7.2 Singulärwertzerlegung

Die Singulärwertzerlegung verallgemeinert die Eigenwertzerlegung auf Rechtecksmatrizen: sei A eine beliebige $\mathbb{R}^{m \times n}$-Matrix mit Rang k. Für die stets quadratischen und symmetrischen[19] $\mathbb{R}^{n \times n}$- bzw. $\mathbb{R}^{m \times m}$-Matrizen $A^T A$ und AA^T gilt

$$A^T A = (USV^T)^T(USV^T) = VSU^T USV^T = VS^2V^T \tag{3.55}$$

und

$$AA^T = (USV^T)(USV^T)^T = USV^T VSU^T = US^2U^T, \tag{3.56}$$

da $(AB)^T = B^T A^T$ ist; demnach sind die Spalten von U Eigenvektoren von AA^T und die Spalten von V sind Eigenvektoren von $A^T A$. Die Matrix A kann daher zerlegt werden in

$$A = USV^T = s_1(u_1 v_1^T) + s_2(u_2 v_2^T) + \ldots s_k(u_k v_k^T), \tag{3.57}$$

[18] Eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ hat reelle Eigenwerte. Ist der Eigenwert einer reellen Matrix komplex, so ist auch der zugehörige Eigenvektor komplex.

[19] Falls die Matrix A komplex ist, so ist $A^T A$ bzw. AA^T Hermitesch (komplex-konjugiert).

wobei $s_1 \geq s_2 \geq \ldots \geq s_k > s_{k+1} = \ldots = s_{min(m,n)} = 0$ ist, i.e. die Singulärwerte werden absteigend geordnet. Für Data Mining-Anwendungen von Bedeutung ist vor allem der folgende Satz: bei Weglassen der Summanden mit den kleinsten Singulärwerten in obiger Formel ergibt sich eine optimale Approximation A_p für A durch Auswahl der p größten Singulärwerte [Gu02], i.e.

$$\| A - A_p \|_2 = \min \| A - B \|_2, \forall p = 1, \ldots, k, \tag{3.58}$$

wobei B eine beliebige Matrix mit gleicher Dimensionalität wie A ist und Rang p besitzt und $\| A \|_2$ die Spektralnorm der Matrix A darstellt, i.e.:

$$\| A \|_2 = \sup_{x \neq 0} \frac{\| A\vec{x} \|_2}{\| \vec{x} \|_2}. \tag{3.59}$$

Aus der Perspektive des Data Mining ist dieser Punkt besonders wichtig, da mit dem Weglassen kleinerer Singulärwerte ein nicht unwesentlicher Performancegewinn verbunden ist. Als allgemeine Richtlinie kann gelten, dass bereits die zwei bis drei größten Singulärwerte das Ergebnis hinreichend genau approximieren, sodass alle weiteren im Normalfall vernachlässigt werden können.

Bei der Analyse von zwei Datenmatrizen auf mögliche Strukturbrüche mittels der Methode der Singulärwertzerlegung muss sowohl die Distanz der Singulärwerte als auch die Distanz der normierten Singulärvektoren in Betracht gezogen werden; da jedoch die Differenzbeträge der Spalten von U und V bei zeitlich direkt aufeinanderfolgenden Datenmatrizen zumeist gering sind - ein Großteil der Werte ist in beiden Datenbeständen identisch, die Hauptrichtungen der Singulärvektoren werden sich im Normalfall nicht gravierend verändern - , sollten eher die Differenzen der Singulärwerte gemessen werden. Allerdings gilt es zu beachten, dass geringe Differenzen in den Singulärwerten noch lange nicht geringe Unterschiede der Datenquellen bedeuten: die Singulärwerte geben nur die Länge (Stärke) der Vektoren in ihrer jeweiligen Richtung an, sie sagen aber nichts über die Richtung der Vektoren aus. Zwei absolut unähnliche Matrizen mit orthogonalen Singulärvektoren können die gleichen Singulärwerte haben - es muss daher auch die Differenz der Singulärvektoren mitberücksichtigt werden[20]. Ab welcher Differenzgröße der Singulärwerte bzw. Singulärvektoren ein Strukturbruch vorherrscht, kann in absoluten Zahlen nicht beurteilt werden: der Differenzbetrag muss jedenfalls signifikant größer sein als die Differenzbeträge vorheriger Datenbestände. Der Programmcode zur Ermittlung der Distanzen zwischen Singulärvektoren und -werten (bzw. Eigenvektoren und -werten) ist im Anhang in Abbildung A.2 dargestellt, Abbildung A.3 zeigt ihren Aufruf aus der konkreten Methode der

[20] Als Hauptkriterium gilt demnach die Differenz der Singulärwerte; der Abstand der normierten Singulärvektoren dient als 'Kontrollinstrument', um nicht von geringen Singulärwertdifferenzen sofort auf ähnliche Matrizen zu schließen.

Singulärwertzerlegung.

Die Singulärwertzerlegung eignet sich durch die Berücksichtigung der gesamten Datenmatrix demzufolge primär für Strukturbrüche, von denen sämtliche dimension members betroffen sind; sie eignet sich daher vorzüglich zur Aufdeckung von Änderungen der Berechnungsvorschrift einer Kennzahl sowie Änderungen der Maßeinheit einer Kennzahl.

Um die Methode der Singulärwertzerlegung im Zeitvergleich sinnvoll anwenden zu können, wird für diese Methode eine etwas größere Datenmatrix verwendet, die in Tabelle 3.23 aufgeführt ist. Bei diesen willkürlich gewählten Werten handelt es sich wiederum um Gewinne von Produktgruppen in Ländern, die jahrelang stets in ATS ausgewiesen worden, in $Jahr_9$ jedoch nur mehr in EURO aufgeführt sind. Zunächst soll der Vergleich von jeweils vier aufeinanderfolgenden

Gew	$Land_1$	$Land_1$	$Land_2$	$Land_2$
ATS/EUR	PG_A	PG_B	PG_A	PG_B
$Jahr_1$	100	200	300	400
$Jahr_2$	102	203	304	404
$Jahr_3$	104	204	305	405
$Jahr_4$	105	205	306	411
$Jahr_5$	106	207	310	413
$Jahr_6$	108	209	311	414
$Jahr_7$	110	211	315	417
$Jahr_8$	113	215	318	418
$Jahr_9$	9	16	23	30

Legende: PG=Produktgruppe, Gew=Gewinn in ATS bzw. EURO

Tabelle 3.23: Beispiel der Gewinnentwicklung zweier Produktgruppen in zwei Ländern

Jahren auf einer auf allen Singulärwerten beruhenden Differenzmaßzahl beruhen: die Differenzen zwischen den einzelnen Jahren in den Singulärwerten und in den Singulärvektoren sind in Tabelle 3.24 aufgeführt ($Jahr_{14-25}$ bedeutet beispielsweise, dass die Datenmatrix von $Jahr_1$ bis $Jahr_4$ mit der Datenmatrix von $Jahr_2$ bis $Jahr_5$ verglichen wird; das in Abbildung A.3 im Anhang aufgeführte Programm, das die Distanz in den Singulärwerten sowie die Distanz in den Singulärvektoren misst, erhält als erstes Argument die Zeilen eins bis vier aus Tabelle 3.23 und als zweites Argument die Zeilen zwei bis fünf[21]). Alle Werte der Tabelle wurden dabei mit 1000 multipliziert, um die Unterschiede in den Daten leichter zu erkennen[22]. Bei der Betrachtung der

[21] Das Programm wird dabei mit den transformierten Matrizen aufgerufen, da in Tabelle 3.23 die Zeiteinheit ausnahmsweise entlang der Ordinate läuft.

[22] Bei allen verwendeten Verfahren spielen die Absolutbeträge eine unwesentliche Rolle. Aussagekräftig sind vor allem die Relationen zwischen den Metriken der einzelnen Datenreihen - deshalb werden in Tabelle 3.24 sowie in nachfolgenden Tabellen immer wieder Skalierungen der Werte vorgenommen, um die Unterschiede besser zu veranschaulichen.

dist * 10^3	$SVec$	$SVal$
$Jahr_{14-25}$	124	0.9
$Jahr_{25-36}$	559.1	0.7
$Jahr_{36-47}$	856.8	1.3
$Jahr_{47-58}$	682.3	2.2
$Jahr_{58-69}$	500.3	**138.6**

Legende: dist=euklidische Distanz zwischen Singulärwerten ($SVal$) und Singulärvektoren ($SVec$) zwischen Matrizen mit jeweils vier Jahren (z.B.: $Jahr_{14-25}$ =Vergleich der Matrix der Jahre 1 bis 4 mit der Matrix der Jahre 2 bis 5)

Tabelle 3.24: Summierte euklidische Differenzen zwischen aufeinanderfolgenden Singulärvektoren und Singulärwerten

Distanzen der Singulärwerte fällt sofort der Ausreißer der letzten Zeile aus Tabelle 3.24 auf: alle anderen Differenzen liegen mit den skalierten Werten zwischen 0.7 und 2.2 relativ gleichauf, in der letzten Zeile beträgt die Differenz hingegen mehr als 138 - ein klares Indiz für einen Strukturbruch; in Anbetracht der Tatsache, dass in der Spalte der Singulärwerte der Wert der letzten Zeile um das 63fache größer ist als die Werte der anderen Zeilen, kann die geringere Distanz der Singulärvektoren in der letzten Zeile im Vergleich zur dritten oder vierten Zeile vernachlässigt werden.

Das Ergebnis ist bereits gleich aussagekräftig, wenn man lediglich die zwei größten Singulärwerte jeder Matrix in die Berechnung aufnimmt, wie Tabelle 3.25 aufzeigt (auch hier wurden alle Werte mit 1000 multipliziert). Auch in diesem Fall[23] erkennt man, dass der Betrag

dist * 10^3	$SVec$	$SVal$
$Jahr_{14-25}$	988.7	1.1
$Jahr_{25-36}$	354	1.4
$Jahr_{36-47}$	1009.7	2.4
$Jahr_{47-58}$	1409.1	4.4
$Jahr_{58-69}$	999.5	**277.1**

Legende: dist=euklidische Distanz zwischen Singulärwerten ($SVal$) und Singulärvektoren ($SVec$) zwischen Matrizen mit jeweils vier Jahren (z.B.: $Jahr_{14-25}$ =Vergleich der Matrix der Jahre 1 bis 4 mit der Matrix der Jahre 2 bis 5)

Tabelle 3.25: Summierte euklidische Differenzen zwischen den jeweils zwei größten aufeinanderfolgenden Singulärwerten und dazugehörigen Singulärvektoren

der Differenz der Singulärwerte zwischen den Perioden $Jahr_{58-69}$ um nahezu das 63fache über den anderen Differenzbeträgen liegt - die anderen, hier vernachlässigten Singulärwerte tragen

[23] Bei Verwendung von nur k Singulärwerten wird das in Abbildung A.3 aufgeführte Programm so modifiziert, dass als drittes Argument die Anzahl der zu berücksichtigenden Singulärwerte - in diesem Fall zwei - angegeben wird; die Singulärwertzerlegung erfolgt dann nicht mehr mit dem Aufruf *svd(Matrix)*, sondern mit *svds(Matrix, k)*.

betragsmäßig nur mehr sehr wenig zum Ergebnis bei[24]. Wogegen es bei Optimierungsproblemen stets eine Erschwerung des Rechnens ist, wenn Matrizen derart schlecht konditioniert sind, i.e. die Konditionszahl $= \frac{\lambda_{max}}{\lambda_{min}} = \frac{\sqrt{s_{max}}}{\sqrt{s_{min}}}$ sehr hoch ist, zum Aufdecken von Strukturbrüchen wie in diesem Fall ist dies jedoch kein Nachteil - der dritte und vierte Singulärwert wirkt sich in den Differenzen von Tabelle 3.25 gering aus[25].

3.7.3 Principal Component Analysis

Principal Component Analysis (PCA) - auch genannt Karhunen-Loeve-Transformation - kann als eine Erweiterung der im vorigen Kapitel präsentierten klassischen Eigenwertzerlegung betrachtet werden. Der prinzipielle Gedanke dieser Analysemethode liegt darin, eine Anzahl von möglicherweise korrelierten Werten in eine kleinere Zahl von unkorrelierten Variablen zu transformieren, die die Varianz der Daten möglichst gut erklärt. Während die klassische Eigenwertzerlegung die gesamte Datenmatrix untersucht, ist die PCA primär zur Aufdeckung solcher Strukturbrüche, die nur einen (oder wenige) Wert(e) betreffen, geeignet [Ho02].

Bei der PCA wird so vorgegangen, dass zunächst von der zu analysierenden Wertereihe $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ der Mittelwert $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ errechnet wird. Dann wird die Kovarianzmatrix $C_x \in \mathbb{R}^{n\times n}$ aufgestellt, wobei $(C_x)_{ij} = (x_i - \mu_x)(x_j - \mu_x)$ ist; die Elemente auf der Hauptdiagonale der Kovarianzmatrix entsprechen somit den individuellen Varianzen $(x_i - \mu_x)^2$ der Komponente x_i. Von der Kovarianzmatrix[26] werden die Eigenvektoren mit ihren korrespondierenden Eigenwerten errechnet, i.e. jene Vektoren e_i, für die gilt:

$$C_x e_i = \lambda_i e_i,\ i = 1, \ldots, n. \tag{3.60}$$

Sei A nun die Matrix, die die Eigenvektoren der Kovarianzmatrix als Zeilenvektoren besitzt; dann kann der ursprüngliche Datenvektor $\vec{x}$ geschrieben werden als

$$\vec{y} = A(\vec{x} - \vec{\mu_x}), \tag{3.61}$$

wobei $\vec{\mu_x}$ ebenso wie in nachfolgender Gleichung als $\mathbb{R}^n$-Vektor mit dem Wert μ_x in jeder Dimension betrachtet werden muss. Da die Eigenvektoren einer Matrix stets zueinander orthogonal sind

[24] Die Differenzen in den Singulärvektoren sollten bei der Auswahl von den k größten Singulärwerten sehr behutsam interpretiert werden, da die MATLAB-Funktion $svds(M, k)$ die Vorzeichen der Singulärvektoren indeterministisch verändert, wodurch die Differenzen unterschiedlich sein können. Zu Analysezwecken sollte man daher in aller Regel zunächst die Differenzen in den Singulärwerten beachten.

[25] Die Tatsache, dass die Distanzwerte der λ_i bei der Approximation mit nur zwei Singulärwerten nahezu doppelt so groß sind, ist auf die Normierung der Distanz mittels Division durch die Anzahl von Singulärwerten zurückzuführen.

[26] Die Kovarianzmatrix ist per definitionem quadratisch und symmetrisch - folglich ist die Eigenwertzerlegung aus (3.54) anwendbar.

und somit $A^T = A^{-1}(A^T A = I)$ ist, kann der ursprüngliche Datenvektor $\vec{x}$ geschrieben werden als

$$\vec{x} = A^T \vec{y} + \vec{\mu_x}. \tag{3.62}$$

Diese Schreibweise von $\vec{x}$ als Linearkombination von orthogonalen Basisvektoren entspricht der Zerlegung von $\vec{x}$ in seine prinzipiellen Komponenten (= PCA) [Ho02].

Zur Aufdeckung von Strukturbrüchen können auch bei dieser Methode einerseits die in Abbildung A.2 dargestellten Differenzen in den Eigenvektoren und Eigenwerten herangezogen werden, andererseits deuten auch grobe Abweichungen im Mittelwert μ_x in Relation zur Größe des Mittelwertes sowie zur Länge des analysierten Vektors auf Veränderungen in der Struktur hin - die Vorgehensweise der PCA ist in Abbildung A.4 des Anhangs aufgeführt.

Die Vorzüge der Methode sollen illustriert werden am Beispiel der Daten aus Tabelle 2.5: $Land_1$ erhielt in $Jahr_t$ eine andere Bedeutung[27], sodass nur hier beim Vergleich der Zeitreihen $Jahr_{t-4}$ bis $Jahr_{t-1}$ zu $Jahr_{t-3}$ bis $Jahr_t$ signifikante Auffälligkeiten auftreten dürften. Das erwartete Ergebnis bestätigt sich bei Durchlauf des Programms, wie Tabelle 3.26 zeigt: die

dist	$EigVal$	$EigVec$	$Mean * 10$
$Land_1PG_A$	**190.24**	0.6432	**1.062**
$Land_2PG_A$	0.0039	0	0.135
$Land_3PG_A$	0.0188	0.6515	0.026

Legende: dist=euklidische Distanz zwischen Eigenwerten ($EigVal$), Eigenvektoren ($EigVec$) zwischen den Matrizen $Jahr_{t-4}$ bis $Jahr_{t-1}$ und $Jahr_{t-3}$ bis $Jahr_t$, Mean=Absolutveränderung des Mittelwertes zwischen $Jahr_{t-3}$ und $Jahr_t$ dividiert durch den größeren der beiden Werte und die Länge der Vektoren

Tabelle 3.26: Differenzen in Eigenwerten, -vektoren und Mittelwerten mit Werten aus Tabelle 2.5

Differenzen in den Eigenwerten sowie im Erwartungswert sind in der ersten Zeile um vier Zehnerpotenzen (Eigenwerte) und um nahezu den Faktor acht (Erwartungswert) größer als in den beiden anderen Zeilen - bei relativen Differenzen in diesen Größenordnungen muss im Normalfall ein Strukturbruch in der Strukturkombination $Land_1PG_A$ stattgefunden haben.

Ebenso wie bei der Methode der klassischen Eigen- und Singulärwertzerlegung bietet sich auch im Rahmen der PCA die Möglichkeit an, nur die k größten Eigenwerte (vgl. (3.58)) für die Berechnung heranzuziehen. Dazu werden von der Kovarianzmatrix A lediglich die k größten Zeilen selektiert, mit Hilfe der Transformation

$$\vec{y} = A_k(\vec{x} - \vec{\mu_x}) \tag{3.63}$$

[27] Da die Produktgruppen in allen Ländern gleichgeblieben sind, wird der Vergleich der Länder exemplarisch nur an Produktgruppe PG_A durchgeführt.

und der Rücktransformation kann $\vec{x}$ somit geschrieben werden als

$$\vec{x} = A_k^T \vec{y} + \vec{\mu_x}, \tag{3.64}$$

wobei $\vec{\mu_x}$ nun ein $\mathbb{R}^k$-Vektor mit den Mittelwerten μ_x als Elementen ist. Lässt man beispielsweise nur die zwei größten Eigenwerte mit ihren korrespondierenden Eigenvektoren in die Berechnung miteinfließen[28], so ergibt sich bei Anwendung der Methode wiederum auf die gleichen Daten aus Tabelle 2.5 das in Tabelle 3.27 dargestellte Bild: der Ausreißer in der ersten Zeile sticht wieder heraus - die Differenz in den Eigenwerten ist in der Strukturkombination $Land_1PG_A$ auch hier um mehr als das 10 000fache größer als in den anderen Strukturkombinationen[29]; diese deutlichen Größenunterschiede in den Eigenwertdistanzen zwischen $Land_1$ und den anderen Zeilen sollten ein genügend großes Indiz für einen Strukturbruch bei $Land_1$ in $Jahr_t$ bedeuten.

dist	$EigVal$	$EigVec$	$Mean * 10$
$Land_1PG_A$	**380.48**	0.78	**1.062**
$Land_2PG_A$	0.0077	1.29	0.135
$Land_3PG_A$	0.0377	1.23	0.026

Legende: dist=euklidische Distanz zwischen Eigenwerten ($EigVal$), Eigenvektoren ($EigVec$) zwischen den Matrizen $Jahr_{t-4}$ bis $Jahr_{t-1}$ und $Jahr_{t-3}$ bis $Jahr_t$, Mean=Absolutveränderung des Mittelwertes zwischen $Jahr_{t-3}$ und $Jahr_t$ dividiert durch den größeren der beiden Werte und die Anzahl der gewählten Eigenwerte (in diesem Fall: Anzahl=2)

Tabelle 3.27: Differenzen in Eigenwerten, -vektoren und Mittelwerten mit Werten aus Tabelle 2.5 unter Berücksichtigung nur zweier Eigenwerte und -vektoren

3.8 Trigonometrische Transformationen

3.8.1 Diskrete Fourier-Transformation

Das Prinzip der Fourier-Transformation basiert auf der Überlegung, dass sich eine beliebige Funktion $f(t)$ als Summe (bei diskreten Werten) bzw. als Integral (bei kontinuierlichen Werten wie z.B. Signalen) von Sinus- und Cosinus-Funktionen mit ganzzahligen Frequenzen darstellen lässt. Für Data Mining-Zwecke ist jedoch nur der diskrete Fall von Bedeutung, da Daten diskrete

[28] Ähnlich wie im vorangegangenen Kapitel wird auch hier das Programm um ein drittes Argument erweitert, in welchem die Anzahl der zu berücksichtigenden größten Eigenwerte und -vektoren übergeben wird - die Distanzberechnung basiert somit nur auf diesen k Eigenwerten und zugehörigen Eigenvektoren.

[29] Die Distanzen in den Eigenwerten sind wie bei der Singulärwertzerlegung auch in diesem Fall bei Berücksichtigung nur zweier Eigenwerte nahezu doppelt so groß, da die Werte nun nur mehr durch zwei statt durch vier dividiert werden und der dritte und vierte Eigenwert kaum noch Auswirkungen hat. Die Distanzen in den Eigenvektoren sind auch in diesem Fall sehr vorsichtig zu interpretieren - auch hier werden von MATLAB die Vorzeichen beliebig vertauscht, sodass sich die Differenzen der Eigenvektoren von Durchlauf zu Durchlauf stark unterscheiden können.

Objekte sind; im folgenden wird daher nur die diskrete Fourier-Transformation (DFT) vorgestellt [At89].

Die DFT einer bestimmten Wertereihe von eindimensionalen Daten $\vec{f} = (f_0, f_1, \ldots, f_{N-1})^T$ ist definiert als:

$$F(n) = \sum_{k=0}^{N-1} f(k) e^{-i2\pi kn/N}, \;\; n = 0, \ldots, N-1, \;\; i = \sqrt{-1}, \tag{3.65}$$

wobei N die Periodizität der Funktion darstellt - es wird somit von einer repetitiven Funktion ausgegangen, sodass $f(k) = f(k+N)\ \forall k$ bzw. $E(x) = E(x+N)\ \forall x$ gilt. Aus dieser Definition lässt sich bereits ableiten, dass diese Analyse nur bei Datenreihen sinnvoll ist, die relativ stationär sind und nicht markante Aufwärts- bzw. Abwärtstrends im Zeitverlauf beinhalten; wegen der Beziehung

$$e^{ix} = \cos x + i \sin x$$

sind in obiger Formel sowohl Sinus- als auch Cosinus-Funktionen erfasst.

Aus der Perspektive des Data Mining sind primär die Differenzen in den Amplituden der Frequenzen[30] der ermittelten Sinusoide bzw. Cosinusoide von Bedeutung. Die Amplituden erhält man, indem man die obige Formel zur Berechnung der Fourier-Transformations-Koeffizienten leicht modifiziert: man berechnet zunächst

$$d_l = \frac{1}{N} \sum_{k=0}^{N-1} f(k) e^{-i2\pi kl/N}, \;\; l = 0, \ldots, \lfloor \frac{N}{2} \rfloor, \;\; i = \sqrt{-1}, \tag{3.66}$$

dann werden die dazugehörigen Amplituden A_l ermittelt als:

$$A_l = 2\sqrt{d_l d_l^*} = 2|d_l| = 2\sqrt{d_l d_{-l}}, \tag{3.67}$$

wobei d_l^* das komplex-konjugierte Pendant zu d_l darstellt, i.e.: wird d_l dargestellt als $a + bi$, wobei a den Realteil $\Re(d_l)$ und b den Imaginärteil $\Im(d_l)$ signifiziert, so ist $d_l^* = a - bi$. Ab welcher Differenz der Amplituden ein Strukturbruch vorliegt, kann auch hier nicht generell beantwortet werden, die Differenz der Werte eines Ausreißers muss allerdings bedeutend größer als die der anderen Werte sein. Als Distanzfunktion dieser wie auch der beiden folgenden Wavelet-Methoden wird die in Abbildung A.5 im Anhang dargestellte Funktion verwendet, die einerseits die maximale Amplitudendifferenz der beiden Wertereihen und andererseits die Gesamtdifferenz aller Amplituden berechnet, wobei die Fourier-Koeffizienten jedes Datenvektors vor der Differenzbe-

[30] Die Fourier-Transformation bildet vom Zeitraum auf den Frequenzraum ab - demnach kann im Ergebnis von Amplituden gesprochen werden.

rechnung mit dem jeweiligen Mittelwert des Vektors skaliert werden (die Gesamtdifferenz wird noch zusätzlich durch die Länge der Vektoren dividiert). Der DFT-spezifische Programmcode, der diese Distanzfunktion aufruft, ist in Abbildung A.6 des Anhanges angegeben[31].

Die Methode soll zunächst an den Daten aus Tabelle 2.5 erprobt werden, in denen es in $Jahr_t$ zu einer Veränderung der Bedeutung von $Land_1$ kam. Wie auch bei der PCA-Methode werden hier die Zeitreihen von $Jahr_{t-4}$ bis $Jahr_{t-1}$ und $Jahr_{t-3}$ bis $Jahr_t$ miteinander verglichen. Das Ergebnis der Berechnung ist in Tabelle 3.28 dargestellt: ganz deutlich fällt auf, dass sowohl die

Land	PG	MaxDiff*100	SumDiff*100
$Land_1$	PG_A	**332.62**	**83.16**
$Land_1$	PG_B	**329.35**	**82.34**
$Land_1$	PG_C	**341.6**	**85.4**
$Land_2$	PG_A	1.75	0.44
$Land_2$	PG_B	2.52	0.63
$Land_2$	PG_C	3.58	0.9
$Land_3$	PG_A	1.71	0.43
$Land_3$	PG_B	0.38	0.1
$Land_3$	PG_C	0.32	0.08

Legende: PG=Produktgruppe, MaxDiff=maximale DFT-Amplitudendifferenz, SumDiff=summierte DFT-Amplitudendifferenz

Tabelle 3.28: Maximale und summierte DFT-Amplitudendifferenzen im Vergleich $Jahr_{t-4}$ bis $Jahr_{t-1}$ zu $Jahr_{t-3}$ bis $Jahr_t$

maximale Differenz als auch die summierte Differenz von Strukturdimensionskombinationen, an denen $Land_1$ beteiligt ist, um ein Vielfaches höher ist als in den anderen Ländern. Während die Werte der mit 100 multiplizierten maximalen Differenz von $Land_2$ und $Land_3$ nicht über 3.58 hinausgehen, sind die Werte der maximalen Differenz in allen Kombinationen mit $Land_1$ stets über 320; die Werte der insgesamten Differenz sind ebenso in allen Kombinationen mit $Land_1$ um mehr als das 91fache größer als in den anderen Zeilen - dieses Ergebnis ist ein eindeutiges Indiz für eine Veränderung des dimension members $Land_1$.

3.8.2 Diskrete Cosinus-Transformation

Die diskrete Cosinus-Transformation (DCT) bildet ebenso wie die DFT die Inputdaten im Zeitraum auf den Frequenzraum ab; sie verwendet jedoch nur Cosinus-Funktionen, wodurch sämtliche Multiplikationen im Rahmen ihrer Berechnung[32] reelle Resultate liefern. Der Vorteil dieses

[31] Der Faktor $\frac{1}{N}$ wird bei der Berechnung der Amplituden weggelassen, da dieser die Ordnung sowie die relativen Abstände der einzelnen Elemente nicht verändert; die Absolutwerte der Amplituden sind aus Data-Mining-Perspektive eher unbedeutend.

[32] Betrachtet man die Gleichung $e^{ix} = \cos x + i \sin x$, so erkennt man, dass durch den Sinus der komplexe Teil hinzugefügt wird.

Verfahrens ist somit eine geringere Rechenzeit für die notwendigen Multiplikationen, wodurch die DCT in dieser Arbeit auch zur Analyse gesamter Datenmatrizen (nicht nur einzelner Datenvektoren wie die DFT) eingesetzt wird.

Die DCT einer Inputmatrix A ist definiert als:

$$B(k_1, k_2) = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} 4\, A(i,j) \cos[\frac{\pi\ k_1}{2N_1}(2i+1)] \cos[\frac{\pi\ k_2}{2N_2}(2j+1)], \tag{3.68}$$

wobei $A(i,j)$ den Wert der Matrix A in der i-ten Zeile und j-ten Spalte angibt und $B(k_1, k_2)$ den DCT-Koeffizienten in Zeile k_1 und Spalte k_2 indiziert; N_1 und N_2 entsprechen der Anzahl der Zeilen bzw. Spalten der Inputmatrix A. Eine allgemein gültige Formel zur Wahl der Anzahl der Zeilen und Spalten der Koeffizientenmatrix B kann nicht gegeben werden - der Tradeoff zwischen Approximationsgenauigkeit der Inputdaten und Performance muss fallspezifisch abgewogen werden [DC02].

Für Data-Mining-Belange sind primär die relativen Unterschiede der Differenzen zwischen einzelnen Analyseintervallen von Bedeutung - die Absolutbeträge der Differenzen an sich sind eher nicht als Interpretationsgrundlage heranzuziehen. Im Gegensatz zur DFT und der Wavelet-Transformation im folgenden Kapitel wird bei dieser Methode die maximale Differenz nicht berechnet, da diese zu stark von den Zeilen- und Spaltenindizes k_1 und k_2 der Ergebnismatrix B beeinflusst wäre, es wird lediglich die summierte euklidische Differenz über alle Koeffizienten (evtl. skaliert mit dem Mittelwert der Inputdaten) berechnet. Eine Skalierung der Differenz mit der Anzahl der Spalten und Zeilen der Ergebnismatrix B wird nicht vorgenommen, da davon ausgegangen wird, dass diese bei verschiedenen Zeitvergleichen der selben Datenquelle stets die gleichen Dimensionen hat. Der entsprechende Sourcecode ist im Anhang in Abbildung A.7 dargestellt[33].

Die Aussagekraft der DCT kann am besten am Beispiel der Veränderung der Berechnungsvorschrift des Gewinns an den Daten aus Tabelle 2.1 überprüft werden: vergleicht man jeweils zwei aufeinanderfolgende Jahre[34] sämtlicher Produktgruppenkombinationen bei einer 3×3-Ergebnismatrix B anhand ihrer skalierten und unskalierten Differenzen, so erhält man das in Tabelle 3.29 aufgeführte Resultat: es fällt auf, dass die skalierten Differenzen den Ausreißer wesentlich schlechter aufdecken als die unskalierten Werte - im zweiten Fall ist der Wert der letzten

[33] Da die Fourier-Transformation bei einer Inputmatrix der Größe $2^x \times 2^y$ einen schnelleren Algorithmus mit Laufzeit $O(n \log n)$ verwendet, ist dieser selbst geschriebene Algorithmus nur in allen anderen Fällen schneller (dort gilt: $O(n^2)$ der DCT vs. $O(n^3)$ der klassischen Fourier-Transfomation).

[34] Im Normalfall sollten nicht nur zwei aufeinanderfolgende Jahre, sondern mehrere Chronone, die in einem Sliding Window um eine Zeiteinheit weiterverschoben werden, analysiert werden. Da die Beispieltabelle jedoch nur Daten von fünf Jahren enthält, wurde die erste Variante gewählt, um zumindest vier verschiedene Differenzwerte zu erhalten.

YearComb	ScaledDiff	UnscaledDiff
$Jahr_{t-4;t-3}$	0.0834	32.3807
$Jahr_{t-3;t-2}$	0.0824	54.2999
$Jahr_{t-2;t-1}$	0.1149	40.9486
$Jahr_{t-1;t}$	**0.1736**	**4622.7**

Legende: ScaledDiff=Differenzen der DCT-Koeffizienten skaliert mit dem Durchschnitt der Mittelwerte der jeweiligen Jahre, UnscaledDiff=unskalierte Dfifferenzen der DCT-Koeffizienten, YearComb=Jahre, die miteinander verglichen werden (z.B.: $Jahr_{t-4;t-3}$: $Jahr_{t-4}$ wird mit $Jahr_{t-3}$ verglichen)

Tabelle 3.29: Differenzen der diskreten Cosinus-Transformation für die Werte aus Tabelle 2.1 von $Jahr_{t-4}$ bis $Jahr_t$

Zeile um rund zwei Zehnerpotenzen höher als die Werte der anderen Zeilen, während im ersten Fall nur geringe Unterschiede zwischen den einzelnen Zeilen festzustellen sind.

3.9 Diskrete Wavelet-Transformation

Eine andere Möglichkeit der Transformation vom Zeit- in den Frequenzraum stellt die Wavelet-Transformation dar. Der größte Vorteil dieser Methode im Vergleich zur DFT liegt darin, dass die Koeffizienten der Wavelet-Transformation auch Aussagekraft über die Zeitlokalität der Frequenzen in den ursprünglichen Daten enthalten. Während bei der Fourier-Transformation ein Signal, welches in einer Periode t_0 bis t_1 zunächst durch eine hohe Frequenz und eine geringe Amplitude charakterisiert ist und ab dem Zeitpunkt $\frac{t_0+t_1}{2}$ mit niedriger Frequenz und hoher Amplitude auftritt, die gleiche Fourier-Transformation[35] besitzt wie das umgekehrte Signal, das zunächst durch eine niedrige Frequenz und in der zweiten Periodenhälfte durch eine hohe Frequenz gekennzeichnet ist, so unterscheidet die Wavelet-Transformation die beiden Signale deutlich voneinander. Auch in diesem Kapitel wird nur die diskrete Wavelet-Transformation (DWT) vorgestellt, da nur sie für die Identifikation von Strukturbrüchen in Data Warehouses von Bedeutung ist (vgl. hier und im folgenden [Vi99]).

Sei $\psi(x) \in L_2(\mathbb{R})$ eine quadrat-integrierbare Funktion, i.e. $\int_{\mathbb{R}} |\psi(x)|^2 dx < \infty$, so ist die Wavelet-Transformation $\psi_{a,b}(x)$ mit $a \in \mathbb{R} \setminus \{0\}$, $b \in \mathbb{R}$ allgemein definiert als

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}} \psi(\frac{x-b}{a}), \tag{3.69}$$

[35] Wenn man die Fourier-Transformation modifiziert, i.e. die Fourier-Koeffizienten für kurze Zeitintervalle jeweils separat berechnet, kann dieser nachteilige Effekt allerdings minimiert werden. Für Data-Mining-Zwecke ist dies jedoch nicht von weiterer Bedeutung, da das zu analysierende Intervall ohnehin minimale Länge - nur zwei aufeinanderfolgende Chronone - aufweisen kann.

wobei der Faktor $\frac{1}{\sqrt{|a|}}$ zur Normalisierung der Koeffizienten dient; die Division durch a ist als Skalierung des Wavelets, die Subtraktion von b als Verschiebung (Shift) zu betrachten, sodass die Wavelet-Funktion unabhängig von a und b ist. Die Wavelet-Funktion muss zudem der Bedingung

$$C_\psi = \int_{\mathbb{R}} \frac{|\Psi(w)|^2}{|w|} dw \ < \infty \tag{3.70}$$

genügen, wobei $\Psi(w)$ die Fourier-Transformation von $\psi(x)$ kennzeichnet. Somit stehen als mögliche Wavelet-Funktionen all jene Funktionen zur Verfügung, die dieses Kriterium erfüllen. Im folgenden werden zwei verschiedene Wavelets anhand konkreter Data Mining-Beispiele vorgestellt.

3.9.1 Haar-Wavelet

Das Haar-Wavelet war das erste, einfachste Wavelet, das bereits anno 1910 von Alfred Haar definiert wurde. Die Haar-Basis besteht zunächst aus der Skalierungsfunktion (dem Haar Vater-Wavelet) $\phi(t)$:

$$\phi(t) = \begin{cases} 1 & t \in [0,1] \\ 0 & t \notin [0.1]. \end{cases} \tag{3.71}$$

Damit das zu skalierende Intervall immer kleiner gewählt werden kann, um die ursprüngliche Funktion besser approximieren zu können, definiert man:

$$\phi_i^j(t) = \phi(2^j t - i), \ j = 0, 1, \ldots \quad i = 0, 1, \ldots 2^j - 1, \tag{3.72}$$

wobei intuitiv i als die Phasenverschiebung und j als die Intervallgröße (=Frequenz) interpretiert werden kann. Darüber hinaus besteht die Haar-Basis noch aus der eigentlichen Haar Wavelet-Funktion (dem Haar Mutter-Wavelet) $\psi(t)$:

$$\psi(t) = \begin{cases} 1 & t \in [0, \frac{1}{2}] \\ -1 & t \in [\frac{1}{2}, 1] \\ 0 & t \notin [0,1]. \end{cases} \tag{3.73}$$

Auch hier definiert man

$$\psi_i^j(t) = \psi(2^j t - i), \ j = 0, 1, \ldots \quad i = 0, 1, \ldots 2^j - 1. \tag{3.74}$$

Die beiden Funktionen (3.71) und (3.73) spannen die Vektorräume V^j und W^j auf:

$$V^j = span\{\phi_i^j\}, i = 0, 1, \ldots 2^j - 1, \tag{3.75}$$

$$W^j = span\{\psi_i^j\}, i = 0, 1, \ldots 2^j - 1, \tag{3.76}$$

wobei $span\{a_1, \ldots, a_n\}$ den Unterraum bezeichnet, der von der Menge aller Linearkombinationen von $a_1, \ldots, a_n$ erzeugt wird, i.e.: $span\{a_1, \ldots, a_n\} := \sum_{k=1}^{n} \alpha_k a_k;\ \alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Die Zusammenhänge zwischen den beiden Vektorräumen (3.75) und (3.76) sind die Grundlage für die Ermittlung der Koeffizienten der Wavelet-Transformation: es gilt

$$V^{j-1} \subseteq V^j, \quad W^{j-1} \subseteq W^j \tag{3.77}$$

sowie

$$V^j = V^{j-1} \oplus W^{j-1} = V^{j-2} \oplus W^{j-2} \oplus W^{j-1} = \ldots = V^0 \oplus W^0 \oplus W^1 \oplus \ldots \oplus W^{j-1}, \tag{3.78}$$

wobei $\oplus$ die Vereinigung von Vektorräumen bedeutet. Auf die zu transformierende Datenreihe mit $N = 2^k$ Werten wird iterativ bis zu V^0 die Expansion der Koeffizienten in die Vektorräume angewandt (*Multiresolution Analysis*)[Yp02]; es kommt somit zu einer Folge von Hoch- und Tiefpassfilteroperationen, um die Amplituden der hohen und niedrigen Frequenzen der Wavelet-Transformation zu berechnen. Der Sourcecode zur Ermittlung der Koeffizienten der Wavelet-Transformation ist in Abbildung A.8 im Anhang aufgeführt, es wird beim Vergleich der Koeffizienten zweier Datenvektoren wiederum die im vorigen Kapitel in Abbildung A.5 (vgl. Anhang) vorgestellte Distanzfunktion verwendet.

Die Transformation mit dem Haar-Wavelet soll ebenso wie die DFT am Beispiel der Daten aus Tabelle 2.5 illustriert werden, das Ergebnis ist in Tabelle 3.30 aufgeführt.

Land	PG	MaxDiff*100	SumDiff*100
$Land_1$	PG_A	**118.24**	**50.54**
$Land_1$	PG_B	**117.64**	**50.01**
$Land_1$	PG_C	**120.7**	**51.60**
$Land_2$	PG_A	0.62	0.26
$Land_2$	PG_B	0.89	0.38
$Land_2$	PG_C	1.58	0.67
$Land_3$	PG_A	0.77	0.42
$Land_3$	PG_B	0.67	0.32
$Land_3$	PG_C	0.14	0.06

Legende: PG=Produktgruppe, MaxDiff=maximale DWT-Amplitudendifferenz, SumDiff=summierte DWT-Amplitudendifferenz

Tabelle 3.30: Maximale und summierte DWT-Amplitudendifferenzen im Vergleich $Jahr_{t-4}$ bis $Jahr_{t-1}$ zu $Jahr_{t-3}$ bis $Jahr_t$

Es zeigt sich auch bei der DWT ganz deutlich, dass sowohl die maximale als auch die gesamte

Differenz bei allen Strukturkombinationen, an denen $Land_1$ beteiligt ist, um mehr als das 74fache sowohl hinsichtlich der maximalen Differenz als auch hinsichtlich der gesamten Differenz höher ist als in allen anderen Kombinationen - ähnlich wie bei der DFT kann auch hier auf einen Strukturbruch in $Land_1$ geschlossen werden.

3.9.2 DWT als Folge linearer Transformationen

Eine zweite Möglichkeit der diskreten Wavelet-Transformation besteht in der Ermittlung der Amplituden der einzelnen Frequenzen mittels iterativer linearer Transformationen. Dazu benötigt man Transformationsmatrizen H_k und G_k der Größe $2^{J-k} \times 2^{J-k+1}$ für $k = 1, \ldots, J-k$, wobei 2^J die Länge der Inputvektoren ist und N eine passend gewählte Konstante ist. Die Transformationsmatrizen enthalten als Werte die Wavelet-Filter $h_s,\ s = 0, \ldots 2^N - 1$ (Details zu N und h_s unten), in der ersten Zeile der H_k-Matrizen sind die Werte h_1 in der 1.Spalte, h_2 in der 2.Spalte, h_3 in der 3.Spalte, ..., und h_0 in der letzten Spalte positioniert, alle anderen Spalten enthalten Nullwerte. In jeder weiteren Zeile werden die h_s um zwei Spalten nach rechts (*modulo* der Spaltenanzahl 2^{J-k+1}) verschoben. Für $J = 3$, $N = 2$ und $k = 1$ ergibt sich beispielsweise folgende H_1-Matrix:

$$H_1 = \begin{bmatrix} h_1 & h_2 & h_3 & 0 & 0 & 0 & 0 & h_0 \\ 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 \\ h_3 & 0 & 0 & 0 & 0 & h_0 & h_1 & h_2 \end{bmatrix}.$$

In Analogie zu H_k ist die Matrix G_k definiert, die an den entsprechenden Stellen die Werte von h_s in H_k durch $(-1)^s\, h_{N+1-s}$ ersetzt, die zur obigen H_1-Matrix korrespondierende G_1-Matrix ist daher:

$$G_1 = \begin{bmatrix} -h_2 & h_1 & -h_0 & 0 & 0 & 0 & 0 & h_3 \\ 0 & h_3 & -h_2 & h_1 & -h_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & h_3 & -h_2 & h_1 & -h_0 & 0 \\ -h_0 & 0 & 0 & 0 & 0 & h_3 & -h_2 & h_1 \end{bmatrix}.$$

Die Ermittlung der Koeffizienten der Wavelet-Transformation für einen Input-Datenvektor $\vec{y}$ mit Länge $N = 2^J$ ist somit eine Transformation aus maximal J Schritten: die Transformation nach dem 1. Schritt ist $W_1\,\vec{y} = \begin{bmatrix} H_1 \\ G_1 \end{bmatrix} \vec{y}$, die Transformation nach dem 2. Schritt $W_2\,\vec{y} = \begin{bmatrix} \begin{bmatrix} H_2 \\ G_2 \end{bmatrix} (H_1\,\vec{y}) \\ G_1\,\vec{y} \end{bmatrix}$, nach dem 3. Schritt $W_3\,\vec{y} = \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} H_3 \\ G_3 \end{bmatrix} H_2(H_1\,\vec{y}) \\ G_2(H_1\,\vec{y}) \end{bmatrix} \\ G_1\,\vec{y} \end{bmatrix}$, etc. Die

Matrix $\begin{bmatrix} H_k \\ G_k \end{bmatrix}$ ist eine basis-verändernde Matrix im 2^{J-k+1}-dimensionalen Raum, sie ist somit orthogonal, i.e. $[H_k^T G_k^T] \begin{bmatrix} H_k \\ G_k \end{bmatrix} = I_{2^{J-k}}$ [Vi99].

Bislang wurde N als eine geeignet zu wählende Konstante definiert: Standardwahl für N bei kontinuierlicher Wavelet-Transformation ist die Anzahl der verschwindenden Momente der zu analysierenden Funktion $\Psi(x)$, i.e. jene größte natürliche Zahl, für die gilt:

$$\int_{-\infty}^{+\infty} x^n \Psi(x) dx = 0, \quad n = 0, 1, \ldots, N-1; \tag{3.79}$$

Wavelets werden auf Basis verschwindender Momente definiert, da sie somit orthogonal zu Polynomen niedrigen Grades sind und nicht-oszillierende Funktionen in komprimierter Form darstellen. Bei kontinuierlichen Funktionen kann die Anzahl verschwindender Momente mittels einer Taylorreihen-Entwicklung der Funktion[36] gefunden werden, für Data Mining-Zwecke kann als mögliche Lösung $N = 2$ vorgeschlagen werden [Vi99]. $N = 2$ ist eine Folge des Einsatzes der Matrix G mit ihren vier Filtern, denn bei Schreibweise der Koeffizienten von G als Vektor $\vec{G}$=(h_3, $-h_2$, h_1, $-h_0$) gilt:

$$\sum_{i=0}^{4} G(i) i^m = 0, \quad m = 0, \ldots, N-1,$$

wobei $G(i)$ das i-te Element des Vektors $\vec{G}$ für $i = 0, \ldots, 3$ signalisiert[37]. Als Werte für die Filter h_s sind die Filter für Daubechies-Wavelets für das jeweils gewählte N heranzuziehen, in Tabelle 3.31 sind die h_s-Filter für $N = 2$ aufgelistet[38]. Der Programmcode zur Ermittlung der Koeffizienten der diskreten Wavelet-Transformation mittels linearer Transformationen unter Anwendung von Daubechies-Filtern für $N = 2$ ist in Abbildung A.9 im Anhang dargestellt; nachfolgende Abbildung A.10 zeigt auf, dass beim Vergleich der ermittelten Transformations-Koeffizienten auch in diesem Fall wieder auf die generische Distanzfunktion aus Abbildung A.5 zurückgegriffen wird.

Ebenso wie die Transformation mit dem Haar-Wavelet sowie die diskrete Fourier-

[36] Eine kontinuierliche Funktion $f(x)$ hat N verschwindende Momente, wenn

$$\frac{\partial^n f(x)}{\partial x^n} \mid_{x=0} = 0, \quad n = 0, 1, \ldots, N-1.$$

[37] Da somit vier Filter zum Einsatz kommen, erfolgt die Transformation nur bis zu jener Matrix H_k, die noch vier Spalten hat, um alle Filter verarbeiten zu können - die Transformation kann somit schon nach weniger als J Schritten beendet sein.

[38] Daubechies-Wavelets sind im Gegensatz zu Haar nicht ein einzelnes konkretes Wavelet, es handelt sich hierbei um Klassen von jeweils zueinander orthogonalen Skalierungs- und Waveletfunktionen, die jeweils einen spezifischen Grad N an Krümmungssanftheit aufweisen.

s	Daub2
0	$\frac{1+\sqrt{3}}{4\sqrt{2}}$
1	$\frac{3+\sqrt{3}}{4\sqrt{2}}$
2	$\frac{3-\sqrt{3}}{4\sqrt{2}}$
3	$\frac{1-\sqrt{3}}{4\sqrt{2}}$

Legende: Daub2=Filter für Daubechies-Wavelet bei $N = 2$, s=Index der h_s-Filter, s=0 ... 3

Tabelle 3.31: Filter für Daubechies-Wavelet bei $N = 2$

Land	PG	MaxDiff*100	SumDiff*100
$Land_1$	PG_A	**139.35**	**51.42**
$Land_1$	PG_B	**138.66**	**50.8**
$Land_1$	PG_C	**142.84**	**52.35**
$Land_2$	PG_A	0.6	0.3
$Land_2$	PG_B	0.86	0.43
$Land_2$	PG_C	1.84	0.68
$Land_3$	PG_A	0.97	0.43
$Land_3$	PG_B	0.49	0.41
$Land_3$	PG_C	0.098	0.085

Legende: PG=Produktgruppe, MaxDiff=maximale DWT-Amplitudendifferenz, SumDiff=summierte DWT-Amplitudendifferenz

Tabelle 3.32: Maximale und summierte Amplitudendifferenzen im Vergleich $Jahr_{t-4}$ bis $Jahr_{t-1}$ zu $Jahr_{t-3}$ bis $Jahr_t$ mit Daubechies-Filtern

Transformation wird auch diese Methode wieder an den Daten aus Tabelle 2.5 erprobt - Tabelle 3.32 stellt das Resultat dar. Es fällt bei der Analyse ähnlich wie bei der Wavelet-Transformation mit dem Haar-Wavelet und der diskreten Fourier-Transformation auf, dass $Land_1$ ein deutlicher Ausreißer ist - die Differenzbeträge (sowohl maximale als auch gesamte Differenz) sind in allen Zeilen, an denen $Land_1$ beteiligt ist, um rund das 75fache höher als in den anderen Zeilen; auch diese Methode kann somit auf den 'präparierten' Daten einen Strukturbruch des dimension members $Land_1$ aufzeigen - die relativen Größenunterschiede zwischen den einzelnen Werten entsprechen dabei jenen der Haar-Wavelet-Transformation.

Zusammenfassend kann daher gefolgert werden, dass zahlreiche mathematisch fundierte Methoden existieren, um Strukturbrüche auf 'präparierten Daten' aufzeigen zu können. Im folgenden wird untersucht, welche bislang nicht erwähnten Schwierigkeiten bei der Identifikation von Brüchen in praxi auftreten können, und welche Ansätze möglich sind, um diese zu überwinden.

4

(Überwindbare) Grenzen der untersuchten Methoden

In diesem Kapitel sollen Grenzen des Data Mining sowie Ansätze, diese zu überwinden, beleuchtet werden.

- Es darf durch die im vorigen Kapitel dargestellten Ergebnisse verschiedener Methoden nicht der Eindruck entstehen, sämtliche Strukturbrüche erkennen zu können: einerseits spiegelt sich nicht jeder Strukturbruch in einer massiven Veränderung der Werte wider, und andererseits kann es zu massiven Veränderungen kommen, ohne dass die dahinterliegende Struktur modifiziert wurde (Abschnitt 4.1).

- Es ist bedeutend leichter, einen Strukturbruch aufzudecken, wenn man bereits vor Durchführung der Analysen eine 'Vermutung' besitzt, in welchem Zeitraum bzw. bei welchem dimension member ein solcher Bruch vorgefallen sein könnte (ebenso Abschnitt 4.1).

- Der Rechenzeitkomplexität der Methoden kommt große Bedeutung zu, da sich die Datenvolumina in der Realität verwendeter Data Warehouses im Gigabyte-Bereich bewegen (Abschnitt 4.2).

Weiters wird in diesem Kapitel auch erläutert, wie die vorgestellten Methoden sinnvoll auf Daten, die über n Dimensionen referenziert werden, angewendet werden können (Abschnitt 4.3). Den Abschluss bildet die Conclusio aller vorherigen Erkenntnisse, die ein schrittweises Vorgehen zur Erkennung von Strukturbrüchen nahe legt (Abschnitt 4.4).

4.1 Schwierig zu erkennende Strukturbrüche

Dieser Abschnitt behandelt eine Datenmatrix, in der ein einzelner dimension member von einer Diskontinuität in der Struktur betroffen war: untersucht man die jährlichen prozentualen Wachstumsraten des Bruttonationalprodukts von 1988 bis 1998 [Un00] der wirtschaftlich hochentwickelten Länder der Erde (vgl. Tabelle 4.1[1]), so ist bekannt, dass lediglich in Deutschland in dieser Zeit eine Veränderung der Landesgrenzen stattgefunden hat.

BIPW	*1988*	*1989*	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*
CAN	4.9	2.5	-0.2	-1.8	0.8	2.2	4.1	2.3	1.5	3.6	3.25
FRA	4.5	4.3	2.5	0.8	1.2	-1.3	2.8	2.1	1.5	2.2	3.25
GER	3.7	3.3	**4.7**	**1.2**	2.2	-1.1	2.9	1.9	1.4	2.4	2.75
ITA	3.9	2.9	2.2	1.1	0.6	-1.2	2.2	2.9	0.7	1.3	2
JPN	6.2	4.8	5.1	3.8	1.0	0.3	0.6	1.4	3.5	0.8	1.25
GBR	5.0	2.2	0.4	-2.0	-0.5	2.1	4.3	2.7	2.4	3.5	2.5
USA	3.8	3.4	1.3	-1.0	2.7	2.3	3.5	2.0	2.4	3.7	2.5
AUS	3.8	4.2	1.2	-1.3	2.6	3.9	5.4	4.0	3.6	3.1	3.75
AUT	4.1	3.8	4.3	2.8	2.1	0.4	3.1	1.8	1.2	1.8	2.5
BEL	4.9	3.6	3.3	2.3	1.8	-1.2	2.3	2.0	1.5	2.3	2.75
DEN	1.2	0.6	1.4	1.3	0.2	1.5	4.3	2.6	2.4	2.5	3
FIN	4.9	5.7	0.0	-7.1	-3.6	-1.2	4.4	4.3	3.7	4.6	4
GRE	4.5	3.5	-0.6	3.5	0.4	-0.9	1.5	2.0	1.8	3.3	3.5
ISL	-0.1	0.3	1.3	1.1	-3.4	1.0	3.7	1.0	5.2	1.0	0.5
IRE	4.3	6.1	7.8	1.9	3.9	3.1	7.0	10.5	7.7	8.0	7
MAL	8.4	8.2	6.3	6.3	4.7	4.5	4.0	9.0	4.2	3.7	4.25
NED	2.6	4.7	4.1	2.3	2.1	0.3	2.6	2.3	3.5	3.2	3.5
NZL	2.3	-0.6	0.3	-2.3	0.6	5.1	5.5	2.7	2.1	1.5	3.5
NOR	-0.1	0.9	2.0	3.0	3.4	2.7	5.5	3.6	5.3	3.4	4.25
POR	4.0	4.9	4.1	2.1	4.2	7.8	1.9	2.0	3.0	3.3	3.25
ESP	5.1	4.8	3.7	2.3	0.7	-1.2	2.1	2.8	2.2	3.2	3.5
SWE	2.7	2.4	1.4	-1.7	-1.4	-2.2	2.7	3.0	1.1	2.3	3
SUI	3.1	4.4	3.7	-0.8	-0.1	-0.5	0.5	0.8	-0.2	0.7	1.75

Legende: BIPW=jährliches prozentuelles Wachstum des Bruttoinlandsprodukts

Tabelle 4.1: Jährliche prozentuale Wachstumsraten des BIP der hochentwickelten Länder der Erde von 1988 bis 1998 [Un00]

Ziel der Data Mining-Methoden sollte es sein, die Diskontinuität der Struktur Deutschlands soweit aufzudecken, dass auch ein Benutzer der Daten, der nicht das Hintergrundwissen der Vereinigung Deutschlands besitzt, durch auffällige Kennzahlen vor unvorsichtiger Interpretation der Daten gewarnt wäre.

[1]Die Zahlen für 1998 sind Schätzungen, jene für 1997 sind vorläufige Werte. Die Werte Deutschlands enthalten ab 1991 die Zahlen der neuen Bundesländer, der Wert Belgiens beinhaltet auch die Daten von Luxemburg.

Abweichungsmatrizen

Einen ersten Versuch stellt die Berechnung der Abweichungsmatrizen aus Kapitel 3.1 dar: bei einer Veränderung lediglich eines dimension members könnte dies in den Matrizen M_1 und M_2 auffallen; da es sich hier um prozentuelle Werte handelt, ist eine Berechnung von M_3 und M_4 nicht sinnvoll. Das Ergebnis der Matrizen M_1 und M_2 ist in den Tabellen 4.2 und 4.3 dargestellt. Die Abweichungsmatrizen, die im vorigen Kapitel die Strukturbrüche klar identifiziert hatten,

AD	89/88	90/89	91/90	92/91	93/92	94/93	95/94	96/95	97/96	98/97
CAN	-2.4	-2.7	-1.6	2.6	1.4	1.9	-1.8	-0.8	2.1	-0.35
FRA	-0.2	-1.8	-1.7	0.4	-2.5	4.1	-0.7	-0.6	0.7	1.05
GER	-0.4	1.4	*-3.5*	1	-3.3	4	-1	-0.5	1	0.35
ITA	-1	-0.7	-1.1	-0.5	-1.8	3.4	0.7	-2.2	0.6	0.7
JPN	-1.4	0.3	-1.3	-2.8	-0.7	0.3	0.8	2.1	-2.7	0.45
GBR	-2.8	-1.8	-2.4	1.5	2.6	2.2	-1.6	-0.3	1.1	-1
USA	-0.4	-2.1	-2.3	3.7	-0.4	1.2	-1.5	0.4	1.3	-1.2
AUS	0.4	-3	-2.5	3.9	1.3	1.5	-1.4	-0.4	-0.5	0.65
AUT	-0.3	0.5	-1.5	-0.7	-1.7	2.7	-1.3	-0.6	0.6	0.7
BEL	-1.3	-0.3	-1	-0.5	-3	3.5	-0.3	-0.5	0.8	0.45
DEN	-0.6	0.8	-0.1	-1.1	1.3	2.8	-1.7	-0.2	0.1	0.5
FIN	0.8	-5.7	**-7.1**	3.5	2.4	5.6	-0.1	-0.6	0.9	-0.6
GRE	-1	-4.1	**4.1**	-3.1	-1.3	2.4	0.5	-0.2	1.5	0.2
ISL	0.4	1	-0.2	-4.5	4.4	2.7	-2.7	4.2	-4.2	-0.5
IRE	1.8	1.7	**-5.9**	2	-0.8	3.9	3.5	-2.8	0.3	-1
MAL	-0.2	-1.9	0	-1.6	-0.2	-0.5	5	-4.8	-0.5	0.55
NED	2.1	-0.6	-1.8	-0.2	-1.8	2.3	-0.3	1.2	-0.3	0.3
NZL	-2.9	0.9	-2.6	2.9	4.5	0.4	-2.8	-0.6	-0.6	2
NOR	1	1.1	1	0.4	-0.7	2.8	-1.9	1.7	-1.9	0.85
POR	0.9	-0.8	-2	2.1	3.6	**-5.9**	0.1	1	0.3	-0.05
ESP	-0.3	-1.1	-1.4	-1.6	-1.9	3.3	0.7	-0.6	1	0.3
SWE	-0.3	-1	-3.1	0.3	-0.8	4.9	0.3	-1.9	1.2	0.7
SUI	1.3	-0.7	**-4.5**	0.7	-0.4	1	0.3	-1	0.9	1.05

Legende: AD=Differenz der jährlichen prozentuellen Wachstumsraten des Bruttoinlandsprodukts

Tabelle 4.2: Abweichungsmatrix M_1 der jährlichen BIP-Wachstumsraten

liefern in diesem Fall keinen Hinweis auf einen Bruch bei Deutschland mehr: die Matrix M_1 in Tabelle 4.2 beinhaltet als betragsmäßig höchsten Wert die Differenz Finnlands zwischen 1990 und 1991 - die Grenzen dieses Landes haben sich in diesem Zeitraum bekanntermaßen aber nicht verändert; die betragsmäßig zweitgrößten Werte finden sich für Portugal zwischen 1993 und 1994 und Irland zwischen 1990 und 1991 - auch diese können nicht durch territorische bzw. politische Veränderungen erklärt werden; da die Zeitperiode 1990-1991 die zwei höchsten

RD	89/88	90/89	91/90	92/91	93/92	94/93	95/94	96/95	97/96	98/97
CAN	-49%	-108%	**800%**	-144%	175%	86%	-44%	-35%	140%	-10%
FRA	-4%	-42%	-68%	50%	-208%	-315%	-25%	-29%	47%	48%
GER	-11%	42%	*-74%*	83%	-150%	-364%	-34%	-26%	71%	15%
ITA	-26%	-24%	-50%	-45%	-300%	-283%	32%	-76%	86%	54%
JPN	-23%	6%	-25%	-74%	-70%	100%	133%	150%	-77%	56%
GBR	-56%	-82%	-600%	-75%	-520%	105%	-37%	-11%	46%	-29%
USA	-11%	-62%	-177%	-370%	-15%	52%	-43%	20%	54%	-32%
AUS	11%	-71%	-208%	-300%	50%	38%	-26%	-10%	-14%	21%
AUT	-7%	13%	-35%	-25%	-81%	675%	-42%	-33%	50%	39%
BEL	-27%	-8%	-30%	-22%	-167%	-292%	-13%	-25%	53%	20%
DEN	-50%	133%	-7%	-85%	650%	187%	-40%	-8%	4%	20%
FIN	16%	-100%	**DIV/0!**	-49%	-67%	-467%	-2%	-14%	24%	-13%
GRE	-22%	-117%	-683%	-89%	-325%	-267%	33%	-10%	83%	6%
ISL	-400%	333%	-15%	-409%	-129%	270%	-73%	420%	-81%	-50%
IRE	42%	28%	-76%	105%	-21%	126%	50%	-27%	4%	-13%
MAL	-2%	-23%	0%	-25%	-4%	-11%	125%	-53%	-12%	15%
NED	81%	-13%	-44%	-9%	-86%	767%	-12%	52%	-9%	9%
NZL	-126%	-150%	**-867%**	-126%	750%	8%	-51%	-22%	-29%	133%
NOR	**-1000%**	122%	50%	13%	-21%	104%	-35%	47%	-36%	25%
POR	23%	-16%	-49%	100%	86%	-76%	5%	50%	10%	-2%
ESP	-6%	-23%	-38%	-70%	-271%	-275%	33%	-21%	45%	9%
SWE	-11%	-42%	-221%	-18%	57%	-223%	11%	-63%	109%	30%
SUI	42%	-16%	-122%	-88%	400%	-200%	60%	-125%	-450%	150%

Legende: RD=Veränderung der prozentuellen Wachstumsraten des Bruttoinlandsprodukts in %

Tabelle 4.3: Abweichungsmatrix M_2 der jährlichen BIP-Wachstumsraten

Veränderungen hat, die jeweils negativ sind, könnte auf eine weltweite Rezession in diesem Zeitraum geschlossen werden. Untersucht man die anderen Werte dieser Kolumne, so erkennt man lediglich zwei positive Werte, wodurch diese These zumindest nicht zurückgewiesen werden kann. Das ursprüngliche Ziel, den Strukturbruch Deutschlands aufzudecken, kann hier aber nicht erreicht werden. Sucht man den höchsten positiven Wert in der Tabelle (Finnland zwischen 1993 und 1994), so kann auch hier lediglich in Analogie zur obigen These auf einen weltweiten konjunkturellen Aufschwung geschlossen werden, da in diesem Zeitintervall nur zwei negative Werte existieren.

Auch bei Betrachtung der Matrix M_2 fällt Deutschland nicht auf: neben dem fehlenden Wert Finnlands von 1990 bis 1991 fällt der Wert Norwegens zwischen 1988 und 1989 auf: -1000%! Bei näherer Untersuchung der Daten erkennt man aber, dass dieser hohe Betrag nur darauf zurückzuführen ist, dass die BIP-Wachstumsrate Norwegens im Jahr 1998 sehr nahe bei null liegt - eine in absoluten Zahlen geringe Differenz wird dadurch mit hohen prozentualen

Veränderungen ausgewiesen (gleiches gilt auch für die beiden betragsmäßig nächstgrößten Werte - Neuseeland und Kanada jeweils 1990 bis 1991); M_2 ist daher bei der Datenanalyse nicht sehr hilfreich.

Bivariate Kreuzkorrelation

Eine andere im vorigen Kapitel erläuterte Methode zum Aufdecken von Strukturbrüchen, die bivariate Kreuzkorrelation, bringt ebenfalls nicht das gewünschte Ergebnis (vgl. Abbildung 4.1, in der aus Platzgründen lediglich die Korrelationskoeffizienten der ersten neun Länder aus Tabelle 4.1 aufgelistet sind): die in Abbildung 4.1 dargestellten Zusammenhänge zwischen BIP-

Correlations

		CAN	FRA	GER	ITA	JPN	GBR	USA	AUS	AUT
CAN	Pearson Correlation	1	,493	,176	,387	-,161	,955**	,890**	,851**	,033
	Sig. (2-tailed)	,	,123	,606	,239	,636	,000	,000	,001	,924
	N	11	11	11	11	11	11	11	11	11
FRA	Pearson Correlation	,493	1	,846**	,918**	,533	,479	,470	,299	,773**
	Sig. (2-tailed)	,123	,	,001	,000	,091	,136	,144	,371	,005
	N	11	11	11	11	11	11	11	11	11
GER	Pearson Correlation	,176	,846**	1	,800**	,548	,192	,257	,020	,880**
	Sig. (2-tailed)	,606	,001	,	,003	,081	,572	,445	,953	,000
	N	11	11	11	11	11	11	11	11	11
ITA	Pearson Correlation	,387	,918**	,800**	1	,560	,420	,277	,188	,779**
	Sig. (2-tailed)	,239	,000	,003	,	,073	,198	,409	,581	,005
	N	11	11	11	11	11	11	11	11	11
JPN	Pearson Correlation	-,161	,533	,548	,560	1	-,052	-,139	-,322	,718*
	Sig. (2-tailed)	,636	,091	,081	,073	,	,880	,684	,334	,013
	N	11	11	11	11	11	11	11	11	11
GBR	Pearson Correlation	,955**	,479	,192	,420	-,052	1	,822**	,831**	,042
	Sig. (2-tailed)	,000	,136	,572	,198	,880	,	,002	,002	,902
	N	11	11	11	11	11	11	11	11	11
USA	Pearson Correlation	,890**	,470	,257	,277	-,139	,822**	1	,855**	,039
	Sig. (2-tailed)	,000	,144	,445	,409	,684	,002	,	,001	,910
	N	11	11	11	11	11	11	11	11	11
AUS	Pearson Correlation	,851**	,299	,020	,188	-,322	,831**	,855**	1	-,161
	Sig. (2-tailed)	,001	,371	,953	,581	,334	,002	,001	,	,635
	N	11	11	11	11	11	11	11	11	11
AUT	Pearson Correlation	,033	,773**	,880**	,779**	,718*	,042	,039	-,161	1
	Sig. (2-tailed)	,924	,005	,000	,005	,013	,902	,910	,635	,
	N	11	11	11	11	11	11	11	11	11

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Legende: in jeder Zelle drei Werte:
1.) Pearson'scher Korrelationskoeffizient zwischen den beteiligten Ländern; 2.) Sig. (2-tailed)=Signifikanzniveau (zweiseitig); 3.) N=Anzahl der Werte je Land
Signifikante Korrelationskoeffizienten (99% Signifikanzniveau) sind dunkelgrau unterlegt

Abbildung 4.1: Kreuzkorrelationsanalyse der jährlichen BIP-Wachstumsraten

Wachstumsraten von Ländern können dahingehend interpretiert werden, dass zwischen den

meisten Staaten klar positive Korrelationen herrschen - in Abbildung 4.1 weisen alle Länder mit Ausnahme Japans mit jeweils drei anderen Ländern signifikant hohe Korrelationskoeffizienten (auf 99% Signifikanzniveau) aus; der auffallende Wert dieser Tabelle ist demnach Japan. Nimmt man in die Analyse allerdings sämtliche 23 Länder aus Tabelle 4.1 mit auf, so fallen als mögliche Ausreißer drei signifikant niedrige (95% Signifikanzniveau) negative Korrelationskoeffizienten auf: -0.678 zwischen Norwegen und Japan, -0.678 zwischen Norwegen und Malta sowie -0.656 zwischen der Schweiz und Malta - die aus Data Mining-Sicht wichtige Aussagekraft dieser Analyse ist darin zu sehen, dass vor Vergleichen der jährlichen BIP-Wachstumsraten mit Malta und Norwegen, und mit Abstrichen auch vor Vergleichen mit der Schweiz und Japan, Vorsicht angebracht ist, da es möglicherweise eine Strukturveränderung in diesen Ländern gegeben hat. Deutschland weist hochsignifikante positive Korrelationskoeffizienten (99% Sig.) mit Frankreich, Italien, Österreich, Belgien, den Niederlanden, Spanien und der Schweiz auf, es existieren jedoch in Verbindung mit keinem anderen Land signifikante negative Korrelationswerte - Deutschland entspricht somit bei dieser Analyse dem Durchschnitt, es kann kein Strukturbruch Deutschlands diagnostiziert werden.

Lineare Regression

Die Methode der linearen Regression wird in diesem Fall mit sieben unabhängigen Variablen durchgeführt, wobei für jedes Land die in Tabelle 4.1 nach dem jeweiligen Land aufgeführten sieben Länder (*modulo* 23) als Vorhersage dienen (beispielsweise sind die Vorhersageländer für Portugal somit Spanien, Schweden, die Schweiz, Kanada, Frankreich, Deutschland und Italien). Tabelle 4.4 zeigt das Ergebnis der R^2-, $\bar{R}^2$- und F-Werte auf: Island und Griechenland weisen mit relativ deutlichem Abstand die geringsten Werte auf, i.e., sie lassen sich am schlechtesten durch andere Länder erklären, Ausreißer am anderen Ende der Skala mit sehr hohen Erklärungswerten der Modelle sind die Niederlande und Portugal. Der deutsche Wert hingegen befindet sich im Mittelfeld aller Länder - es kann auch in diesem Fall kein Strukturbruch erkannt werden.

Autokorrelation

Betrachtet man die Autokorrelationskoeffizienten der BIP-Wachstumsraten der Länder, so kann zu Data Mining-Zwecken primär die Stärke des Koeffizienten zu Lag $\tau = 1$ herangezogen werden; die Korrelogramme einiger Länder sind in Abbildung 4.2 am Ende dieses Kapitels dargestellt. Bei der Analyse zeigt sich, dass in nahezu allen Ländern der Autokorrelationskoeffizient $\rho(1)$ einen positiven Wert annimmt: lediglich in Portugal und Deutschland sind diese Werte mit -0.043 bzw. -0.015 negativ; somit können diese beiden Länder als Ausreißer identifiziert werden. In Portugal hat es während der Zeitperiode von 1988 bis 1998 keine territorische Veränderung

Land	R^2	$\bar{R}^2$	F
CAN	0.975	0.918	16.946
FRA	0.931	0.771	5.8
GER	*0.874*	*0.58*	*2.969*
ITA	0.967	0.89	12.602
JPN	0.863	2.542	2.689
GBR	0.874	0.579	2.965
USA	0.854	0.513	2.502
AUS	0.923	0.742	5.112
AUT	0.937	0.789	6.329
BEL	0.815	0.385	1.894
DEN	0.939	0.798	6.633
FIN	0.934	0.781	6.096
GRE	**0.55**	**-0.502**	**0.523**
ISL	**0.429**	**-0.902**	**0.323**
IRE	0.881	0.602	3.163
MAL	0.764	0.212	1.385
NED	**0.993**	**0.967**	**63.029**
NZL	0.903	0.675	3.969
NOR	0.92	0.735	4.956
POR	**0.988**	**0.961**	**36.185**
ESP	0.952	0.839	8.422
SWE	0.97	0.9	13.849
SUI	0.754	0.181	1.315

Tabelle 4.4: R^2-, $\bar{R}^2$- und F-Werte der jährlichen BIP-Wachstumsraten der einzelnen Länder

gegeben; die Unterschiede zu den anderen Staaten könnten sich jedoch daraus ergeben haben, dass Portugal noch einen relativen wirtschaftlichen Aufholbedarf hat(te). Der zweite Ausreißer, Deutschland, ist hingegen tatsächlich einer strukturellen Veränderung unterlegen gewesen - diese Analyse konnte somit Deutschland als 'Sonderfall' identifizieren.

Autoregression

Untersucht man die 23 verschiedenen Zeitreihen der jährlichen BIP-Wachstumsraten aus Tabelle 4.1 mit der Methode der Autoregression, so kann beispielsweise ein $AR(5)$-Modell als Basis herangezogen werden - betrachtet man die Korrelogramme aus obigem Abschnitt, so erkennt man positive wie negative Korrelationen, was gegen ein $AR(1)$-Modell spricht. Ein MA(q)-Anteil erscheint nicht unbedingt als notwendig, da einige der Autokorrelationskoeffizienten in den Korrelogrammen in Abbildung 4.2 am Ende dieses Kapitels ($\rho(1)$ im Falle Japans und $\rho(3)$ im Falle Großbritanniens) so groß sind, dass die Hypothese einer Zufallsreihe von Werten widerlegt

werden kann; auch auf vorherige Differenzoperationen vor Anwendung des AR(5)-Modells kann verzichtet werden, da die Zeitreihen durchaus als stationär betrachtet werden können. Eine eindeutige Ermittlung der Parameter ϕ macht bei mehreren verschiedenen Datenvektoren keinen Sinn, es werden daher nach Betrachtung der Korrelogramme die Werte $\phi_1 = 0.6$, $\phi_2 = -0.5$, $\phi_3 = 0.4$, $\phi_4 = 0.3$ und $\phi_5 = 0.2$ gewählt. Das Ergebnis der klassischen Berechnungsvariante der Kennzahlen AIC, BIC und FPE (keine Skalierung der Werte mit dem Mittelwert der Daten[2]) für die jeweiligen Länder ist in Tabelle 4.5 dargestellt - es kann dabei ein relativ eindeutiger

Land	AIC	BIC	FPE
CAN	2.1382	1.4187	9.1152
FRA	2.7407	2.0212	16.6513
GER	*2.9916*	*2.272*	*21.3979*
ITA	1.7564	1.0369	6.2224
JPN	2.7515	2.032	16.8316
GBR	2.3428	1.6233	11.1848
USA	2.1238	1.4043	8.9848
AUS	1.9887	1.2692	7.8493
AUT	1.9403	1.2208	7.4786
BEL	2.2409	1.5214	10.1009
DEN	1.2656	0.546	3.8087
FIN	**3.9998**	**3.2802**	**58.6448**
GRE	1.7366	1.017	6.1
ISL	**3.3919**	**2.6723**	**31.9315**
IRE	2.8381	2.1186	18.3541
MAL	2.9881	2.2686	21.3247
NED	2.0008	1.2812	7.9447
NZL	2.3584	1.6389	11.3608
NOR	2.7792	2.0597	17.3044
POR	2.535	1.8155	13.5547
ESP	2.5319	1.8124	13.5128
SWE	2.9765	2.2569	21.0773
SUI	2.4891	1.7695	12.9462

Tabelle 4.5: AIC, BIC und FPE der jährlichen BIP-Wachstumsraten der einzelnen Länder

Ausreißer identifiziert werden: Finnland. Der Wert dieses Staates liegt im Kriterium FPE um nahezu das Doppelte über allen anderen Werten, in den anderen beiden Kriterien sind die Unterschiede aufgrund der vorherigen Logarithmierungen der Fehlprognosen in absoluten Zahlen geringer. Auch der zweithöchste Wert, Island, liegt relativ deutlich über all den anderen Wer-

[2]Da die Daten aus Tabelle 4.1 ohnehin prozentuale Werte darstellen, würde eine nochmalige Skalierung keinen Sinn haben.

ten, obwohl auch hier ebenso wie bei Finnland keine politische bzw. territorische Veränderung stattgefunden hat. Deutschland liegt zwar auf dem dritten Platz, die Abstände zu den nachfolgenden Staaten sind aber zu gering, um hier noch einen Ausreißer erkennen zu können - gerade bei Autoregressionsanalysen sollten aufgrund der Einbindung von Zufallsvariablen u_t geringe Differenzen mit größter Vorsicht interpretiert werden.

Differenzialgleichungen

Die BIP-Wachstumsdaten aus Tabelle 4.1 können mit einer einfachen Funktionsgleichung modelliert werden, in denen die Ableitungen nicht zu berücksichtigen sind: es wird angenommen, dass in den Jahren 1988 und 1989 das Wirtschaftswachstum um 1% über dem Durchschnitt des jeweiligen Landes von 1988 bis 1998 lag, von 1990 bis 1993 das Wachstum 2.5% unter dem Mittelwert aller betrachteten Jahre sowie von 1994 bis einschließlich 1998 wiederum um 1.6% über dem Durchschnitt. Formal ist der prognostizierte Wert $x(t)$, $t = 0 \dots 10$, ($t_0 = 1988,\ t_{10} = 1998$) somit definiert als

$$x(t) = \begin{cases} \mu + 1, & t \in [0, 2), \\ \mu - 2.5, & t \in [2, 6), \\ \mu + 1.6, & t \in [6, 10]. \end{cases} \tag{4.1}$$

Da es sich um prozentuelle Werte handelt, sind die in Tabelle 4.6 dargestellten Absolutbeträge der Differenzen zwischen dem mittels der Differenzialgleichung prognostizierten Wert und dem tatsächlichen Wert zum jeweiligen Zeitpunkt t unskaliert dargestellt: die größten Differenzen treten mit 6.61 im Jahr 1993 im Falle von Portugal und mit 6.39 anno 1991 in Finnland auf - das eigentliche Ziel, einen Strukturbruch in Deutschland aufzudecken, kann allerdings nicht erreicht werden.

Principal Component Analysis

Die Methode der PCA kann auf zweifache Weise eingesetzt werden: einerseits können die Datenvektoren der einzelnen Länder miteinander verglichen werden, andererseits können auch verschiedene Zeitintervalle des gleichen Landes gegenübergestellt werden[3]. Tabelle 4.7 vergleicht zunächst verschiedene Länder untereinander, wobei aus Platzgründen nur die Differenzen zwischen acht Ländern aufgelistet sind[4]: betrachtet man die Differenzen der Eigenwerte und Eigenvektoren der Kovarianzmatrizen der in Tabelle 4.7 dargestellten Länder, so kann festgestellt

[3] Es wird dabei jeweils die Variante der PCA verwendet, die alle Eigenwerte berechnet, um nicht durch die indeterministischen Vorzeichenwechsel von MATLAB falsche Eigenvektorendifferenzen zu erhalten.

[4] Da es sich um prozentuale Werte handelt, sind lediglich die Differenz der Eigenvektoren der Kovarianzmatrizen in der rechten, oberen Dreiecksmatrix sowie die Differenz der Eigenwerte in der linken, unteren Dreiecksmatrix der Tabelle aufgelistet; die Differenz der Mittelwerte ist nicht angegeben.

Land/Diff	*1988*	*1989*	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*
CAN	1.8	0.6	0.2	1.4	1.2	2.6	0.4	1.4	2.2	0.1	0.45
FRA	1.33	1.13	2.83	1.13	1.53	0.97	0.97	1.67	2.27	1.57	0.52
GER	0.4	0	*4.9*	*1.4*	2.4	0.9	1	2	2.5	1.5	1.15
ITA	1.21	0.21	3.01	1.91	1.41	0.39	1.09	0.39	2.59	1.99	1.29
JPN	2.59	1.19	4.99	3.69	0.89	0.19	3.61	2.81	0.71	3.41	2.96
GBR	1.95	0.85	0.85	1.55	0.05	2.55	0.65	0.95	1.25	0.15	1.15
USA	0.38	0.02	1.38	0.92	2.78	2.38	0.52	2.02	1.62	0.32	1.52
AUS	0.31	0.09	0.59	1.91	1.99	3.29	0.69	0.71	1.11	1.61	0.96
AUT	0.56	0.26	4.26	2.76	2.06	0.36	1.04	2.34	2.94	2.34	1.64
BEL	1.58	0.28	3.48	2.48	1.98	1.02	1.62	1.92	2.42	1.62	1.17
DEN	1.71	2.31	1.99	1.89	0.79	2.09	0.79	0.91	1.11	1.01	0.51
FIN	2.11	2.91	0.71	**6.39**	2.89	0.49	1.01	0.91	0.31	1.21	0.61
GRE	1.45	0.45	0.15	3.95	0.85	0.45	2.15	1.65	1.85	0.35	0.15
ISL	2.15	1.75	2.75	2.55	1.95	2.45	1.05	1.65	2.55	1.65	2.15
IRE	2.82	1.02	4.18	1.72	0.28	0.52	0.72	2.78	0.02	0.28	0.72
MAL	1.62	1.42	3.02	3.02	1.42	1.22	3.38	1.62	3.18	3.68	3.13
NED	1.24	0.86	3.76	1.96	1.76	0.04	1.84	2.14	0.94	1.24	0.94
NZL	0.58	3.48	0.92	1.68	1.22	5.72	2.02	0.78	1.38	1.98	0.02
NOR	4.19	3.19	1.41	2.41	2.81	2.11	0.81	1.09	0.61	1.29	0.44
POR	0.69	0.21	2.91	0.91	3.01	**6.61**	3.39	3.29	2.29	1.99	2.04
ESP	1.45	1.15	3.55	2.15	0.55	1.35	2.15	1.45	2.05	1.05	0.75
SWE	0.49	0.19	2.69	0.41	0.11	0.91	0.11	0.19	1.71	0.51	0.19
SUI	0.89	2.19	4.99	0.49	1.19	0.79	2.31	2.01	3.01	2.11	1.06

Legende: Diff=Differenzen zwischen Prognose- und Ist-Wert

Tabelle 4.6: Differenzen zwischen dem prognostizierten und tatsächlichen Wert des Wirtschaftswachstums

ΔEVal\ΔEVec	CAN	FRA	GER	ITA	JPN	GBR	USA	AUS
CAN	-	0.4367	0.3866	0.4415	0.4167	0.438	0.4221	0.419
FRA	0.5189	-	0.3839	0.3748	0.4702	0.4341	0.452	0.4004
GER	*0.7399*	*0.2211*	-	0.4196	0.4388	0.4398	0.4301	0.4153
ITA	0.6033	0.0844	*0.1367*	-	0.4539	0.3969	0.4671	0.4221
JPN	0.0959	0.423	*0.6441*	0.5074	-	0.3732	0.4215	0.4132
GBR	0.2213	0.7402	*0.9613*	0.8246	0.3172	-	0.3988	0.408
USA	0.9385	0.4196	*0.1986*	0.3353	0.8427	**1.1598**	-	0.4082
AUS	0.705	0.1861	*0.035*	0.1017	0.6091	0.9263	0.2336	-

Legende:Δ EVal=Eigenwertdifferenz,ΔEVec=Eigenvektorendifferenz der prozentuellen BIP-Wachstumsraten

Tabelle 4.7: Eigenwert- und Eigenvektorendifferenzen zwischen Kovarianzmatrizen der jährlichen BIP-Wachstumsraten ausgewählter Länder

werden, dass diese generell relativ gering sind; als mögliche Ausreißer dieses Datenausschnitts könnten eventuell Großbritannien oder die USA erkannt werden, deren Eigenwerte zueinander um mehr als eins differieren - die Unterschiede sind aber auch hier zu gering, um von einem Strukturbruch sprechen zu können. Vergleicht man jedoch sämtliche Länder miteinander (hier nicht dargestellt), dann kann ein eindeutiger Ausreißer identifiziert werden: Finnland. Die maximale Eigenwertdifferenz wird mit 8.4289 zwischen Finnland und den Niederlanden gemessen, die zweitgrößte Differenz findet sich zwischen Finnland und Österreich mit 8.3259, und auch der Vergleich Finnlands mit anderen Ländern hinsichtlich der Eigenwertdifferenz liefert Werte über acht (8.2595 mit Malta, 8.2072 mit Dänemark, 8.1718 mit Portugal, 8.1591 mit den USA, 8.0209 mit Norwegen). Die größte Eigenwertdifferenz, an der nicht Finnland beteiligt ist, liegt mit 3.5514 (zwischen den Niederlanden und Island) um mehr als das Doppelte unter den größten finnischen Werten und noch immer deutlich unter der kleinsten Differenz, an der auch Finnland beteiligt ist (4.8775 zwischen Finnland und Island) - diese Analyse würde somit deutlich vor einer voreiligen Interpretation bei Vergleichen der BIP-Wachstumsraten eines Landes mit Finnland zwischen 1988 und 1998 warnen.

In der zweiten Variante wird die Principal Component Analysis verwendet, um die BIP-Wachstumsraten jedes Landes zwischen 1988 und 1993 mit jenen zwischen 1993 und 1998 zu vergleichen; Tabelle 4.8 zeigt das Ergebnis. Bei Betrachtung der Eigenwertdifferenzen fallen als klare Ausreißer Finnland und Island auf: die Differenzen dieser beiden Länder liegen um mehr als das Vierfache über den Differenzen der anderen Länder. Auch der Wert Schwedens ist relativ deutlich höher als die Werte der anderen Länder, sodass auch dieser Staat als kleiner Ausreißer identifiziert werden könnte - in all diesen drei Ländern sind die Differenzen aber nicht auf territoriale oder bedeutende politische Veränderungen zurückzuführen, sie lassen sich lediglich durch unterschiedliche Wirtschaftswachstumsphasen erklären. Deutschland fügt sich mit Durchschnittswerten in die Eigenwert-Rangliste ein; ein Strukturbruch ist auch mittels der PCA nicht zu erkennen. Die Eigenvektorendifferenzen sind in allen Ländern so gering, dass sie nicht weiter berücksichtigt werden müssen.

Hat man jedoch bereits vor Durchführung der Analysen eine Vermutung, dass zwischen 1990 und 1991 möglicherweise die Staatsgrenzen eines Landes verändert wurden, so kann man versuchen, das Intervall 1988-1995 einmal mit dem Intervall 1990-1997 und einmal mit dem Intervall 1991-1998 bezüglich der Eigenwertdifferenzen[5] zu vergleichen, um anschließend die relativen Unterschiede der beiden Eigenwertdifferenzen zu berechnen. Tabelle 4.9 zeigt das Ergebnis: Finnland und Island fallen nun nicht mehr als Ausreißer auf; die größten Werte stehen für Australien, Deutschland, Österreich und Schweden zu Buche - in diesen Ländern könnte es

[5]Die Eigenvektoren waren bereits oben so ähnlich, dass sie in Tabelle 4.9 vernachlässigt wurden.

Land	ΔEVal	ΔEVec
CAN	2.9604	0.5694
FRA	0.8271	0.4918
GER	*0.4619*	0.5545
ITA	0.4042	0.4144
JPN	0.4863	0.5521
GBR	3.9747	0.6251
USA	1.0711	0.5382
AUS	1.4449	0.5574
AUT	0.2181	0.6157
BEL	0.4154	0.6593
DEN	0.0444	0.5509
FIN	**94.3512**	0.553
GRE	2.9203	0.6324
ISL	**76.2866**	0.6286
IRE	0.1707	0.4581
MAL	0.3203	0.5215
NED	0.2833	0.4771
NZL	5.3838	0.4642
NOR	0.5148	0.5766
POR	0.4739	0.6393
ESP	0.8206	0.5265
SWE	**18.3333**	0.4796
SUI	1.7683	0.5509

Legende:ΔEVal=Eigenwertdifferenz, ΔEVec=Eigenvektorendifferenz der prozentuellen BIP-Wachstumsraten

Tabelle 4.8: Eigenwert- und Eigenvektorendifferenzen der BIP-Wachstumsraten eines Landes mit sich selbst zwischen den Intervallen 1988-1993 und 1993-1998

demnach zwischen 1990 und 1991 zu Strukturbrüchen gekommen sein[6].

Diskrete Fourier-Transformation

Die diskrete Fourier-Transformation kann ebenso wie die PCA im vorigen Abschnitt auf zwei verschiedene Weisen angewendet werden; hier wird lediglich das Ergebnis der erste Variante (Berechnung der Differenzen zwischen verschiedenen Ländern) gezeigt, da die zweite Variante anschließend anhand der Haar-Wavelet-Transformation erprobt wird. Bei Betrachtung von Tabelle 4.10 - aus Platzgründen sind auch hier nur die Vergleichswerte zwischen acht Ländern

[6]Die reinen prozentuellen Differenzen sollten allerdings nicht überbewertet werden, da Werte nahe bei null leicht starken prozentuellen Schwankungen unterworfen sind.

Land	ΔEW_{I1}	ΔEW_{I2}	$\max(\frac{\Delta EW_{I1}}{\Delta EW_{I2}}, \frac{\Delta EW_{I2}}{\Delta EW_{I1}})$
CAN	0.1566	0.7254	463.11%
FRA	0.4844	0.3832	126.4%
GER	0.0413	0.3276	**792.63%**
ITA	0.0875	0.1005	114.87%
JPN	0.2636	0.6682	253.47%
GBR	0.2972	0.7247	243.81%
USA	0.0421	0.1401	332.94%
AUS	0.0117	0.2466	**2103.44%**
AUT	0.0317	0.2056	**648.35%**
BEL	0.3095	0.3773	121.92%
DEN	0.2392	0.2833	118.48%
FIN	5.7421	4.8514	118.36%
GRE	0.315	0.8783	278.85%
ISL	1.5849	1.2047	131.57%
IRE	0.0325	0.0522	160.78%
MAL	0.0073	0.0268	368.05%
NED	0.1422	0.2067	145.34%
NZL	0.8478	1.4360	169.38%
NOR	0.6316	0.7481	118.45%
POR	0.0516	0.0693	134.23%
ESP	0.4341	0.4623	106.49%
SWE	0.895	0.1719	**520.59%**
SUI	0.7754	0.3823	202.84%

Legende:ΔEW_{I1}, ΔEW_{I2} = Eigenwertdifferenzen der prozentuellen BIP-Wachstumsraten im Intervallvergleich $I1$ (1988-1995 mit 1990-1997) bzw. im Intervallvergleich $I2$ (1988-1995 mit 1992-1998), $\max(\frac{\Delta EW_{I1}}{\Delta EW_{I2}}, \frac{\Delta EW_{I2}}{\Delta EW_{I1}})$=prozentuelle Veränderung der Eigenwertdifferenzen (größere EW-Differenz dividiert durch kleinere EW-Differenz)

Tabelle 4.9: Eigenwertdifferenzen der jährlichen BIP-Wachstumsraten eines Landes mit sich selbst zwischen den Zeitintervallen 1988-1995 und 1990-1997 und zwischen den Zeitintervallen 1988-1995 und 1991-1998 sowie prozentuale Veränderungen der Eigenwertdifferenz

dargestellt[7] - fällt die Diskontinuität der deutschen Struktur kaum auf, auch wenn die maximale Differenz der Fourier-Koeffizienten zwischen Deutschland und Kanada die zweitgrößte ist; anhand des in der Tabelle gezeigten Ausschnitts könnten am ehesten noch die USA und Großbritannien als Ausreißer interpretiert werden, da diese beiden Länder in der Gegenüberstellung sowohl die höchste maximale Differenz als auch die höchste Gesamtdifferenz ausweisen - die Unterschiede zu den anderen Ländern sind jedoch zu gering, um hier auf Strukturbrüche

[7]Bei längeren Datenvektoren wäre es vorteilhaft, Vektoren mit 2^n Werten heranzuziehen, um die Laufzeit des Algorithmus zu verbessern.

MaxDiff\SumDiff	CAN	FRA	GER	ITA	JPN	GBR	USA	AUS
CAN	-	0.6711	0.9176	0.8319	0.7515	0.2492	**1.0641**	0.8621
FRA	4.6487	-	0.4455	0.5992	0.2561	0.7176	0.592	0.5636
GER	**6.0915**	*2.1937*	-	0.7381	0.6873	0.9798	0.3713	0.3947
ITA	4.1059	3.0026	*3.4233*	-	0.5957	0.8785	0.6916	0.7008
JPN	4.218	0.9967	*2.8601*	2.0059	-	0.7742	0.6668	0.7414
GBR	1.6267	4.2547	*5.6975*	3.4715	3.8241	-	**1.1263**	0.9244
USA	4.8685	3.98	*1.7863*	2.8024	4.6464	**6.142**	-	0.3524
AUS	3.7806	3.2453	*2.3759*	3.0242	3.9117	5.4074	1.1614	-

Legende: MaxDiff=Maximale Differenz der DFT-Koeffizienten, SumDiff=insgesamte Differenz der DFT-Koeffizienten

Tabelle 4.10: Maximale und insgesamte Differenzen der DFT-Koeffizienten der jährlichen BIP-Wachstumsraten ausgewählter Länder

schließen zu können. Betrachtet man sämtliche (hier nicht dargestellte) Länderkombinationen, so erkennt man, dass sowohl die größte maximale Differenz als auch die größte Gesamtdifferenz der DFT-Koeffizienten zwischen Finnland und Malta mit 25.5184 bzw. 4.5486 gemessen wird - somit könnten diese beiden Länder als Ausreißer identifiziert werden, auch wenn die Abstände zu den nächstgrößeren Differenzen gering sind.

Diskrete Wavelet-Transformation

Die im vorigen Kapitel vorgestellte Methode der diskreten Wavelet-Transformation mit dem Haar-Wavelet wird hier eingesetzt, um die Differenzen in den BIP-Wachstumsraten jedes Landes mit sich selbst zwischen den Zeiträumen 1988-1995 und 1990-1997 sowie zwischen 1988-1995 und 1991-1998 zu erkennen - die beiden Methoden benötigen Datenvektoren der Länge 2^n, deshalb wurden jeweils acht Jahre zusammengefasst. Das Ergebnis in den ersten beiden Spalten von Tabelle 4.11 zeigt, dass Finnland mit Spitzenwerten in der maximalen und insgesamten Differenz in beiden Vergleichsintervallen auch von dieser Methode klar als Ausreißer erkannt wird. Die zweitgrößte Werte (Schweden im Intervallvergleich $I1$, die Schweiz im Intervallvergleich $I2$) liegen zum einen weit hinter Finnland und zum anderen nur relativ knapp vor den nächsthöheren Werten, sodass gemäß dieser Analyse lediglich in Finnland ein Strukturbruch stattgefunden hat.

Hat man zu Beginn der Analyse eine zusätzliche 'Vorausahnung', dass es zwischen 1990 und 1991 zu einer Veränderung eines Landes gekommen sein könnte, so macht es Sinn, die maximalen und insgesamten Differenzen aus den verschiedenen Vergleichsperioden gegenüberzustellen. Vergleicht man nun die prozentualen Unterschiede der maximalen Differenzen und der gesamten Differenzen (siehe Tabelle 4.11), so erkennt man in beiden Spalten relativ deutliche Ausreißer: in der Kolumne der maximalen Differenzen ist der Wert Deutschlands mit 201.81% mit relativ

Land	$Max\Delta_{I1}$	$Sum\Delta_{I1}$	$Max\Delta_{I2}$	$Sum\Delta_{I2}$	$\max(\frac{Max\Delta_{I1}}{Max\Delta_{I2}}, \frac{Max\Delta_{I2}}{Max\Delta_{I1}})$	$\max(\frac{Sum\Delta_{I1}}{Sum\Delta_{I2}}, \frac{Sum\Delta_{I2}}{Sum\Delta_{I1}})$
CAN	4.1405	1.4392	4.3713	1.1521	105.5722%	124.9121%
FRA	2.5161	0.838	2.4726	1.0184	101.7612%	121.5225%
GER	*1.3432*	*0.7462*	*2.7108*	*0.8355*	**201.8134%**	111.9608%
ITA	2.8269	0.9814	2.4297	0.9924	116.3512%	101.1121%
JPN	1.4804	0.6274	2.3033	0.8343	155.5895%	132.9635%
GBR	3.4711	1.5726	4.8522	1.4712	139.7879%	106.8916%
USA	2.6458	0.722	2.4394	0.6833	108.459%	105.6625%
AUS	2.5347	0.7854	2.6388	0.7949	104.1075%	101.2098%
AUT	0.9086	0.4459	1.3516	0.4243	148.7656%	105.0987%
BEL	1.7454	0.6913	2.0587	0.6912	117.9506%	100.0048%
DEN	2.0816	0.7069	1.4754	0.6356	141.0855%	111.2153%
FIN	**15.8749**	**6.8068**	**15.6775**	**5.3782**	101.2592%	126.5613%
GRE	3.2017	1.2761	2.7501	0.7258	116.4208%	**175.8183%**
ISL	6.4809	2.7878	7.6573	2.5726	118.1525%	108.3671%
IRE	1.086	0.5449	1.2024	0.5391	110.7187%	101.0609%
MAL	0.7735	0.307	0.8123	0.3459	105.0232%	112.6983%
NED	1.2502	0.4337	1.6048	0.5987	128.3635%	138.0559%
NZL	3.0753	1.6617	3.7193	1.7922	120.9382%	107.857%
NOR	0.6478	0.336	1.089	0.4986	168.1198%	148.3955%
POR	1.3832	0.7177	1.2702	0.5295	108.8948%	135.548%
ESP	2.4616	0.7409	3.0165	0.7945	122.5442%	107.2425%
SWE	8.1778	2.8462	6.6177	3.0507	123.5744%	107.1864%
SUI	6.8713	1.9585	7.6681	3.1123	111.5968%	158.9078%

Legende:$Max\Delta_{I1}$, $Sum\Delta_{I1}$, $Max\Delta_{I2}$, $Sum\Delta_{I2}$ = Maximale Differenz sowie summierte Differenz der DWT-Koeffizienten im Intervallvergleich $I1$ (1988-1995 mit 1990-1997) bzw. im Intervallvergleich $I2$ (1988-1995 mit 1991-1998), $\max(\frac{Max\Delta_{I1}}{Max\Delta_{I2}}, \frac{Max\Delta_{I2}}{Max\Delta_{I1}})$, $\max(\frac{Sum\Delta_{I1}}{Sum\Delta_{I2}}, \frac{Sum\Delta_{I2}}{Sum\Delta_{I1}})$ =prozentuelle Veränderung der maximalen bzw. summierten Differenzen der DWT-Koeffizienten (größere Differenz dividiert durch kleinere Differenz)

Tabelle 4.11: Maximale und insgesamte Differenzen der DWT-Koeffizienten (Haar-Wavelet) der jährlichen BIP-Wachstumsraten eines Landes mit sich selbst zwischen den Zeitintervallen 1988-1995 und 1990-1997 und zwischen den Zeitintervallen 1988-1995 und 1991-1998 sowie prozentuale Veränderungen der maximalen und gesamten Differenz

deutlichem Abstand am größten, in der Kolumne der Gesamtdifferenzen der Wert Griechenlands mit 175.82% - in diesen beiden Ländern könnte es somit gemäß der Analyse zu einem Strukturbruch zwischen 1990 und 1991 gekommen sein.

4.2 Skalierbarkeit der untersuchten Methoden

In diesem Abschnitt wird die Performanz der untersuchten Methoden bei großen Datenmengen untersucht. Es wird hierzu mittels Zufallszahlen (zur Generierung der Daten vergleiche Abbildung A.11 im Anhang) ein Data Warehouse mit einer Kennzahl, einer Struktur-[8] und einer Zeitdimension simuliert, das in Matrizenform dargestellt wird: in der i-ten Zeile in der j-ten Spalte steht somit der Wert der Kennzahl des i-ten dimension members dieser Strukturdimension im j-ten Chronon; die Randomgenerator-Funktion in Abbildung A.11 wird somit mit einer mit Nullwerten initialisierten $N^3 \times N$-Matrix aufgerufen[9] und liefert eine mit Zufallszahlen (normalverteilt um den Mittelwert 200) gefüllte Datenmatrix zurück, die allerdings in den Zeilen 25-30 (im konkreten Fall somit die Werte der dimension members $SDim1_{25}$, ...,$SDim1_{30}$) in den Zeitperioden 8-10 deutliche Ausreißer enthält; die Daten sind somit in diesem Abschnitt wie zu Beginn zur Aufdeckung von Ausreißern 'präpariert', sodass jede Methode diese bei richtiger Einstellung der Schwellwerte, bei deren Überschreitung der zugehörige dimension member als brüchig klassifiziert wird, entdecken sollte (in Abbildung A.12 im Anhang ist beispielhaft für alle Methoden der Aufruf der Fourier-Transformation mit Schwellwerten für die maximale und insgesamte Distanz zweier Vektoren angegeben).

Primäres Ziel ist es jedoch, die benötigte Rechenzeit festzustellen. Im konkreten Fall der 1000×10-Matrix liefern alle untersuchten Methoden das gewünschte Ergebnis, die dazu benötigten Rechenzeiten (Pentium III, 866 MHz mit 128 MB RAM) sind im Anschluss an die Anmerkungen zu den speziellen Einstellungen jeder Methode in Tabelle 4.12 dargestellt.

Anmerkungen zu den verwendeten Verfahren:

- Es werden lediglich die Abweichungsmatrizen M_1 und M_2 berechnet, da aufgrund der großen Datenmengen eine Verschiebung der Anteile einzelner Vektoren zu minimal ist, um in den Matrizen M_3 bzw. M_4 aufzufallen. Zwei aufeinanderfolgende Spalten einer Zeile der Datenmatrix werden dann ausgegeben, wenn der relative Unterschied der Werte mehr als 60% beträgt.

- Die bivariate Kreuzkorrelation interpretiert das Verhältnis zweier Wertereihen als auffällig, wenn deren Korrelationskoeffizient größer als 0.98 oder kleiner als -0.98 ist.

- Um den Laufzeitvergleich mit anderen Verfahren nicht zu verzerren, werden die Autokorrelationswerte nicht mehr in einem Korrelogramm visualisiert, sondern ebenso wie die Werte

[8] Die Besonderheiten beim Auftreten mehrerer Strukturdimensionen sind in Kapitel 4.3 dargestellt.

[9] Die Matrix wird bereits vorher initialisiert, da durch die frühzeitige Allokation des benötigten Speicherplatzes ein bedeutender Performancegewinn entsteht. Es ist auch unabhängig von der Eingabegröße der Matrix schneller, jeden Wert der Matrix direkt durch einen Aufruf der *randn*-Funktion zu berechnen, als im vorhinein einen Vektor mit einer bestimmten Anzahl n an bereits generierten Zufallszahlen zu übergeben und der Matrix anschließend den entsprechenden Wert des Zufallszahlenvektors (z.B. Summe von Zeilen- und Spaltenindizes *modulo* n) zuzuweisen.

der anderen Methoden in textueller Form ausgegeben; die Ausgabe eines Datenvektors erfolgt bei Überschreitung der Grenzwerte $\rho(1) = 0.7$ und $\rho(2) = 0.3$.

- Die lineare Regression wird aus Performanzgründen nur mit einer abhängigen Variable (zufällig gewählt) durchgeführt. Das Ergebnis ist auffällig, wenn der R^2-Wert eines Modells größer als 0.93 ist.

- Dem Autoregressionsverfahren liegt ein ARMA(3,2)-Prozess mit den Parametern $\phi_1 = \phi_2 = 0.6$, $\phi_3 = -0.2$ sowie $\theta_1 = 0.4$ und $\theta_2 = -0.4$ zugrunde. Der Grenzwert zur Erkennung eines Ausreißers liegt bei $AIC = 9$.

- Die Singulärwertzerlegung sieht zwei aufeinanderfolgende Intervalle der gesamten Matrix als 'ausreißerverdächtig' an, wenn die Singulärwertdifferenz der beiden Matrizen größer als 0.06 ist. Um die Berechnung zu beschleunigen, werden lediglich die beiden größten Singulärwerte berechnet. Ebenso wie bei der diskreten Cosinus-Transformation werden jeweils vier Zeitintervalle in einer Matrix analysiert.

- Die Ergebnismatrix der diskreten Cosinus-Transformation ist eine 4×4-Matrix, auf der dann die Differenzen einzelner Zeitintervalle (nicht skaliert mit dem Mittelwert) berechnet werden. Zwei aufeinanderfolgende Zeitintervalle werden ausgegeben, wenn ihre Differenz in der DCT-Matrix größer als 8000 ist.

- Die maximalen bzw. insgesamten Differenzen zwischen aufeinanderfolgenden Vektoren müssen in der Analyse mit der Fourier-Transformation größer als 1.75 bzw. größer als 0.35 sein, um die entsprechenden Zeitintervalle des jeweiligen Vektors als Ausreißer zu interpretieren; es werden bei der DFT ebenso wie bei den Wavelet-Transformationen und der Principal Component Analysis jeweils vier Zeitperioden gemeinsam betrachtet.

- Im Rahmen der Principal Component Analysis wird ein Strukturbruch eines Vektors vermutet, wenn die Eigenwertdifferenz zweier aufeinanderfolgender Kovarianzmatrizen dieses Vektors größer als 20 ist.

- Im Rahmen der Analyse mittels Haar- und Daubechies-Wavelets werden Zeitintervalle einzelner Datenvektoren als auffällig interpretiert, wenn die maximale Distanz zum nachfolgenden Zeitintervall größer als 0.5 und die insgesamte Distanz größer als 0.25 ist.

- Die Methode der Modellierung mittels Differenzialgleichungen wird mit der einfachen Funktionsgleichung $x(t) = 200$, welches für alle dimension members und alle Zeitintervalle gilt, durchgeführt, da kein weiteres Vorwissen über die Entwicklung der Wertereihen

vorhanden ist[10]. Wie bei der klassischen Methode der Abweichungsmatrix wird ein Wert dann als Strukturbruch ausgegeben, wenn der Absolutbetrag der relativen Differenz des Wertes zu 200 größer als 60% ist.

Methode	Zeit
Abweichungsmatrix	0.16
Differenzialgleichungen	0.16
Autokorrelation	0.55
DFT	1.27
Autoregression	1.38
Haar-Wavelet	2.47
Singulärwertzerlegung	4.67
PCA	6.76
Daubechies-Wavelet	23.57
DCT	29.61
Kreuzkorrelation	**181.92**
Lineare Regression	**504.98**

Legende: Rechenzeiten der Methoden in Sekunden

Tabelle 4.12: Rechenzeiten der untersuchten Methoden auf einer $10^3 \times 10$-Matrix

Es zeigt sich, dass die Laufzeit der Methoden stark unterschiedlich ist: die Berechnung der klassischen Abweichungsmatrizen M_1 und M_2 benötigt ebenso wie die Berechnung der Abweichungsmatrix zwischen dem tatsächlichen und dem prognostizierten Wert 200 lediglich 0.16 Sekunden, und dennoch werden in beiden Varianten die Differenzen in den Zeilen 25-30 in den Spalten acht bis zehn als Ausreißer aufgedeckt - diese beiden Methoden sollen demnach zuerst eingesetzt werden. Die lineare Regression am anderen Ende der Rechenzeitskala ist hingegen erst nach über 500 Sekunden mit der Berechnung fertig, sie identifiziert dabei auch die Zeilen 25-30 als Ausreißer.

Analysiert man im folgenden die Laufzeitkomplexitäten der Methoden, dann werden die Unterschiede in den Berechnungszeiten aufgeklärt: seien D die Anzahl an dimension members in der Strukturdimension[11] und C die Anzahl der Chronone mit $C \ll D$, dann gilt für die Komplexitätsordnungen der einzelnen Methoden:

- Die Methode der einfachen Abweichungsmatrix sowie die Abweichungsmatrizen mit den absoluten und relativen Differenzen zwischen dem von der Differenzialgleichung prognostizierten und dem tatsächlichen Wert besuchen alle Werte genau einmal $\rightarrow O(DC) = O(D)$.

[10] Auch wenn $x(t) = 200$ eine reine Funktionsgleichung und keine Differenzialgleichung darstellt, da keine Ableitungen vorkommen, so wird zur Erhaltung der Konsistenz der Terminologie auch im weiteren von der Methode der Differenzialgleichungen gesprochen.

[11] Die Thematik der Laufzeitkomplexitäten bei Erweiterung der Methoden auf n Strukturdimensionen wird in Kapitel 4.3 nochmals kurz angerissen.

- Singulärwertzerlegung und DCT[12] arbeiten jeweils auf $D \times I$-Matrizen (I=Intervallgröße, zumeist 4 oder 8), von denen es maximal C gibt. Die Laufzeitkomplexität der DCT ist somit

$$O(DIk_1k_2C) = O(Dk_1k_2) = O(D),$$

wobei k_1 und k_2 die Anzahl der Zeilen bzw. Spalten der Ergebnismatrix angeben; da diese Zahlen im Allgemeinen ebenso wie I und C jedoch sehr klein im Verhältnis zu D sind, kann die Komplexitätsordnung auch als $O(D)$ angesehen werden.

Für die Singulärwertzerlegung gilt: eine komplette Singulärwertzerlegung einer $n \times n$-Matrix benötigt O(n^3) Operationen, bei einer $D \times I$-Matrix (mit $I \ll D$), von denen es maximal C gibt, ist die Laufzeitkomplexität somit

$$O(D^3C) = O(D^3).$$

Die approximierte Variante der Singulärwertzerlegung, die lediglich die k größten Singulärwerte berechnet, ist jedoch bedeutend schneller als die klassische Methode; hier gestaltet sich eine exakte Laufzeitanalyse allerdings als schwierig, da MATLAB nicht preis gibt, mit welchem Verfahren die beiden größten Singulärwerte gefunden werden - aus der Tatsache, dass dieses Verfahren bei der Erprobung auf dem Real-World Data Warehouse (siehe Kapitel 5.2.1) ungefähr die gleiche Laufzeit hat wie die DCT, könnte jedoch der Schluss gezogen werden, dass die Berechnung möglicherweise in $O(D^2)$ oder sogar in nahezu linearer Zeit $O(D^{1+\epsilon})$, $\epsilon > 0$, erfolgt.

- Autokorrelation, Autoregression, Principal Component Analysis, die diskrete Fourier-Transformation sowie die beiden Wavelet-Transformationen betrachten jeden dimension member über bestimmte Zeitintervalle getrennt, d.h. sie analysieren nur C Werte zu einem bestimmten Zeitpunkt. Ob die Berechnung eines Datenvektors (mit C Werten) nun linear [$O(C)$], logarithmisch-linear [$O(C \log C)$] oder höher polynomial [$O(C)^2$ bzw. $O(C)^3$] mit der Anzahl der Werte in Verbindung steht, spielt für die gesamte Matrix nur eine geringe Rolle: es müssen D solcher Vektoren analysiert werden, insgesamt ergibt dies beispielsweise bei quadratischer Rechenzeit in Abhängigkeit eines Vektors

$$O(C^2D) = O(D),$$

wobei die Gleichung aus der Annahme $C \ll D$ folgt.

[12]Die diskrete Cosinus-Transformation könnte ebenso wie die DFT auf einzelnen Vektoren operieren; da in diesem Fall der Performancevorteil wesentlich geringer ist als bei Betrachtung ganzer Matrizen, wird sie in dieser Arbeit nur für gesamte Datenmatrizen eingesetzt.

- Die Verfahren der bivariaten Kreuzkorrelation und der linearen Regression vergleichen alle D dimension members mit allen anderen dimension members über alle C Zeitintervalle hinweg, die Komplexität ist somit

$$O(D^2C) = O(D^2).$$

  Werden die bivariate Kreuzkorrelation und die lineare Regression jedoch so eingesetzt, dass nicht dimension members untereinander, sondern jeder dimension member lediglich mit der Summe aller Werte (bzw. mit einem speziellen dimension member) verglichen wird, so ist deren Laufzeit

$$O(D),$$

  da alle Daten im Data Warehouse (mit Ausnahme der Daten des speziellen dimension members) nur einmal besucht werden; dieses Verfahren wird weiter unten in Kapitel 5.2.3 verwendet, die nun folgenden Rechenzeitanalysen gehen allerdings noch von ersterer Variante aus.

Die Rechenzeiten der unterschiedlichen Methoden auf $N^3 \times N$-Matrizen für N=10, 20, 30,...,70 sind in Tabelle 4.14 dargestellt[13] - auch bei Vergrößerung der Matrizen bestätigt sich die obige These, dass die klassische Abweichungsmatrix und die Abweichungsmatrix mit den Unterschieden zwischen dem von einem Modell (einer Differenzialgleichung) prognostizierten Wert und dem tatsächlichen Wert die schnellsten Verfahren sind; diese beiden Methoden sollen demnach zuerst eingesetzt werden. Davor zeigt Tabelle 4.13 die benötigen Zeiten zum Generieren[14], Laden und Speichern der $N^3 \times N$-Matrizen auf. Bei Betrachtung der Ergebnisse muss

Datenmatrix	$10^3 \times 10$	$20^3 \times 20$	$30^3 \times 30$	$40^3 \times 40$	$50^3 \times 50$	$60^3 \times 60$	$70^3 \times 70$
Generierung	0.11	1.98	9.94	32.51	152.75	349.55	778.02
Speichern	0.11	1.27	6.86	20.82	52.67	104.69	827.61
Laden	0.16	2.26	11.09	38.67	85.69	182.79	853.93

Legende: Generierungs-, Speicherungs- und Ladezeiten in Sekunden

Tabelle 4.13: Generierungs-, Speicherungs- und Ladezeiten in Abhängigkeit der Größe der Datenmatrix

berücksichtigt werden, dass die Methoden der Auto- und Kreuzkorrelation, der linearen Regression und der Autoregression, der Singulärwertzerlegung und der DCT mit den Einstellungen der $10^3 \times 10$-Matrix nicht mehr die gewünschten Ausreißer identifizieren - da die Zeitintervalle der

[13] Tabelle 4.14 ist unvollständig, da die Methoden, die bereits bei kleinen Datenmatrizen sich als langsam erwiesen hatten, nicht mehr auf den größeren Tabellen erprobt wurden.

[14] Ein Versuch der Allokation einer $80^3 \times 80$=512000×80-Matrix hat eine 'OUT OF MEMORY'-Exception zur Folge (Pentium III, 866 MHz mit 128 MB RAM).

Meth \ Matrix	$10^3 \times 10$	$20^3 \times 20$	$30^3 \times 30$	$40^3 \times 40$	$50^3 \times 50$	$60^3 \times 60$	$70^3 \times 70$
Abw.matrix	0.16	2.52	23.34	67.72	206.47	436.05	1046.5
Diff.gleichung	0.16	1.92	16.04	56.47	71.46	408.26	897.76
Autokorrelation	1.05	7.58	46.3	121.49	343.95	681.68	1673.7
DFT	1.15	52.51	334.88		3659.8		12350
Autoregression	1.26	19.78	110.23	346.3			3491.4
Haar-Wavelet	2.47	88.17					20536
Daubechies-W.	23.57	975.2					
PCA	6.76	336.47					
SVD	4.67	146.16					
DCT	29.61	657.85					
Korrelation	**181.92**						
Regression	**504.98**						

Legende: Meth=Methode; Rechenzeiten in Sekunden

Tabelle 4.14: Rechenzeiten der untersuchten Methoden in Abhängigkeit der Größe der Datenmatrix

einzelnen dimension members nun länger sind, haben drei 'verfälschte' Werte (die Spalten 8-10 in den Zeilen 25-30) nun nicht mehr einen so großen Einfluss auf die jeweiligen Kennzahlen; will man die Ausreißer trotzdem erkennen, müssen die Schwellwerte, bei deren Überschreitung ein Strukturbruch signalisiert wird, verändert werden.

4.3 Erkennung von Strukturbrüchen in n Dimensionen

In den vorangegangenen Abschnitten wurde zu Zwecken der Vereinfachung stets von zwei Dimensionen - einer Strukturdimension und einer Zeitdimension - ausgegangen. Im Allgemeinen besitzen Data Warehouses jedoch mehrere (5-20) Strukturdimensionen; auch das spezielle Real-World-Data Warehouse des nachfolgenden Kapitels weist vier Strukturdimensionen aus. Strukturbrüche können nun in all diesen Dimensionen auftreten, wobei in der Mehrzahl aller Fälle lediglich dimension members einer einzelnen Dimension betroffen sein sollten. Betrachtet man beispielsweise das in Tabelle 4.15 dargestellte Data Warehouse, das die Werte dreier Strukturdimensionen SD_{1j}, SD_{2j} und SD_{3j} mit je zwei dimension members SD_{i1} und SD_{i2} in drei Jahren enthält, so fallen auf den ersten Blick viele Ausreißer in der letzten Kolumne ($Jahr_3$) auf - die Werte aller Strukturkombinationen dieses Jahres unterscheiden sich vom Vorjahreswert um rund 50%; auf dieser Betrachtungsebene von kombinierten Strukturdimensionen würden somit mit allen Methoden alle dimension members als brüchig klassifiziert werden, obwohl lediglich in einer Strukturdimension ein Strukturbruch stattgefunden hat. Um der Ursache der Veränderung auf den Grund zu gehen, werden alle dimension members jeder Strukturdimension

SD_1	SD_2	SD_3	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	SD_{21}	SD_{31}	50	55	30
SD_{11}	SD_{21}	SD_{32}	100	110	60
SD_{11}	SD_{22}	SD_{31}	150	165	90
SD_{11}	SD_{22}	SD_{32}	200	220	120
SD_{12}	SD_{21}	SD_{31}	50	55	80
SD_{12}	SD_{21}	SD_{32}	100	110	150
SD_{12}	SD_{22}	SD_{31}	150	165	225
SD_{12}	SD_{22}	SD_{32}	200	220	295

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.15: Data Warehouse mit einem Strukturbruch in einer Dimension (1a)

getrennt analysiert[15], i.e. die Werte eines speziellen dimension members werden jeweils nach seiner zugehörigen Strukturdimension aggregiert; Tabelle 4.16 zeigt das Resultat der Daten aus Tabelle 4.15. Hierauf kann beispielsweise die einfache Abweichungsmatrix M_2 für zwei Dimen-

SD	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	500	550	300
SD_{12}	500	550	750
SD_{21}	300	330	320
SD_{22}	700	770	730
SD_{31}	400	440	425
SD_{32}	600	660	625

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.16: Data Warehouse mit einem Strukturbruch in einer Dimension (1b)

$\Delta(\%)$	$Jahr_{12}$	$Jahr_{23}$
SD_{11}	10%	**45.5%**
SD_{12}	10%	**36.36%**
SD_{21}	10%	3.03%
SD_{22}	10%	5.19%
SD_{31}	10%	3.4%
SD_{32}	10%	5.3%

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i, $\Delta(\%)$=Differenz der Werte in %, $Jahr_{mn}$=Vergleich von Jahr m mit Jahr n, prozentuelle Veränderung auf Basis von Jahr m

Tabelle 4.17: Data Warehouse mit einem Strukturbruch in einer Dimension (1c)

[15]Dieses Verfahren hätte korrekterweise bereits in Kapitel 3 angewendet werden sollen - dort waren allerdings die Brüche so offensichtlich, dass sie von den Methoden auch bei Betrachtung von Strukturdimensionskombinationen identifiziert werden konnten.

sionen angewendet werden (siehe Tabelle 4.17), und es wird deutlich, dass die Unterschiede auf Veränderungen in Strukturdimension SD_1 zurückzuführen sind - die Werte der beiden dimension members dieser Strukturdimension haben sich im Vergleich zum Vorjahr um über 45% bzw. über 36% verändert, wogegen die Veränderungen sämtlicher dimension members anderer Strukturdimensionen nur rund 5% betragen haben.

Ein häufig in Data Warehouses anzutreffender Fall ist auch der, dass bei der getrennten Analyse einzelner Strukturdimensionen alle (bzw. mehrere) Strukturdimensionen große Brüche aufweisen, die jedoch nur auf Brüche einer einzigen Strukturdimension zurückzuführen sind. Liegt ein Data Warehouse mit zwei Strukturdimensionen mit den in Tabelle 4.18 dargestellten Daten vor, so fällt zunächst auf, dass alle Werte verschiedener Strukturdimensionskombinationen

SD_1	SD_2	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	SD_{21}	100	500	200
SD_{11}	SD_{22}	60	12	30
SD_{11}	SD_{23}	40	8	20
SD_{12}	SD_{21}	200	1000	400
SD_{12}	SD_{22}	120	24	60
SD_{12}	SD_{23}	80	16	40
SD_{13}	SD_{21}	300	1500	600
SD_{13}	SD_{22}	180	36	90
SD_{13}	SD_{23}	120	24	60

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.18: Data Warehouse mit einem Strukturbruch in einer Dimension (2a)

in gleichem Maße brechen. Analysiert man dies auf Ebene einzelner Dimensionen, so erhält man das in Tabelle 4.19 präsentierte Resultat: alle beteiligten dimension members in beiden Struktur-

SD	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	200	520	250
SD_{12}	400	1040	500
SD_{13}	600	1560	750
SD_{21}	600	3000	1200
SD_{22}	360	72	180
SD_{23}	240	48	120

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.19: Data Warehouse mit einem Strukturbruch in einer Dimension (2b)

dimensionen weisen Brüche auf; untersucht man die einzelnen Dimensionen jedoch nicht mit der Abweichungsmatrix M_2, sondern mit der Abweichungsmatrix M_4 (prozentuelle Veränderungen der Anteile einzelner dimension members am Gesamtkuchen der jeweiligen Strukturdimension

im Vergleich zum Vorchronon)[16], so wird klar, dass der Bruch in der Strukturdimension SD_2 stattgefunden hat, wie Tabelle 4.20 illustriert: während die prozentuellen Veränderungen inner-

$M_4(\%)$	$Jahr_{12}$	$Jahr_{23}$
SD_{11}	0%	0%
SD_{12}	0%	0%
SD_{13}	0%	0%
SD_{21}	**46.2%**	**-16.2%**
SD_{22}	**-27.7%**	**9.7%**
SD_{23}	**-18.5%**	**6.5%**

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i, $M_4(\%)$=absolute Veränderung der Anteile einzelner dimension members am Gesamtkuchen zwischen einzelnen Jahren, $Jahr_{mn}$=Vergleich der Anteile einzelner dimension members zwischen Jahr m und Jahr n

Tabelle 4.20: Data Warehouse mit einem Strukturbruch in einer Dimension (2c)

halb der Strukturdimension SD_1 stets null waren, haben sich die Anteile der einzelnen dimension members in SD_2 zwischen den verschiedenen Jahren deutlich verschoben - diese Verschiebung hat sich somit bei getrennter Betrachtung aller Dimensionen auch auf andere Strukturdimensionen ausgewirkt. In großen Data Warehouses wird die Abweichungsmatrix in aller Regel jedoch nicht so klar zum Ausdruck bringen, welche Strukturdimensionen sich verändert haben und welche gleich geblieben sind; in diesem Fall ist es äußerst hilfreich, einen dimension member in jeder Strukturdimension zu kennen, der mit großer Sicherheit von Strukturbrüchen verschont geblieben ist - sämtliche Analysen wie PCA, DFT, DWT, Autokorrelation, etc. können bei der Analyse anderer dimension members als Vergleichsmaßstab den 'nicht-brüchigen' dimension member heranziehen[17].

Ein Strukturbruch kann aber auch in mehreren Strukturdimensionen gleichzeitig auftreten: trifft man beispielsweise auf ein Data Warehouse mit den Daten aus Tabelle 4.21, so fallen ähnlich wie oben gravierende Ausreißer in der letzten Spalte ($Jahr_3$) auf; in allen Dimensionskombinationen weichen die Werte um mehr als 50% vom Vorjahreswert ab. Betrachtet man nun wieder die Werte für jede Strukturdimension getrennt, so erhält man das in Tabelle 4.22 aufgeführte Ergebnis: kein einziger Strukturbruch kann nun erkannt werden, die Werte sämtlicher dimension members unterscheiden sich vom Vorjahr um weniger als 5%. Da die ursprünglichen Daten jedoch enorme Brüche in $Jahr_3$ aufweisen, muss ein kombinierter Strukturbruch mehrerer Dimensionen stattgefunden haben. Folglich werden im nächsten Schritt alle möglichen Kombi-

[16]Ein ähnliches Resultat könnte man auch mit einer Kreuzkorrelations- oder einer Regressionsanalyse erhalten, aus Performanzgründen wird jedoch die Berechnung der Abweichungsmatrix bevorzugt.

[17]Gelingt es nicht, einen solchen dimension member in jeder Dimension herauszuforschen, so kann auch die Entwicklung der Summe aller Absolutbeträge zwischen zwei Intervallen als Trendvorgabe herangezogen werden.

SD_1	SD_2	SD_3	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	SD_{21}	SD_{31}	50	52.5	15
SD_{11}	SD_{21}	SD_{32}	50	52.5	15
SD_{11}	SD_{22}	SD_{31}	50	52.5	85
SD_{11}	SD_{22}	SD_{32}	50	52.5	85
SD_{12}	SD_{21}	SD_{31}	50	52.5	90
SD_{12}	SD_{21}	SD_{32}	50	52.5	90
SD_{12}	SD_{22}	SD_{31}	50	52.5	15
SD_{12}	SD_{22}	SD_{32}	50	52.5	15

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.21: Data Warehouse mit einem kombinierten Strukturbruch in zwei Dimensionen (1)

SD	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	200	210	200
SD_{12}	200	210	210
SD_{21}	200	210	210
SD_{22}	200	210	200
SD_{31}	200	210	205
SD_{32}	200	210	205

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.22: Data Warehouse mit einem kombinierten Strukturbruch in zwei Dimensionen (2)

nationen von jeweils zwei dimension members in verschiedenen Strukturdimensionen analysiert: Tabelle 4.23 zeigt das Resultat auf. Unterzieht man die Daten von Tabelle 4.23 wiederum der Abweichungsmatrix M_2 (Tabelle 4.24), so erkennt man sofort, dass ein kombinierter Strukturbruch zwischen den dimension members aus Strukturdimension SD_1 und SD_2 vorgefallen ist - die Abweichungen in diesen Kombinationen betragen teilweise über 70%, wogegen alle anderen Abweichungen anderer Dimensionskombinationen unter 5% liegen.

Die Analyse funktioniert auch, wenn mehr als zwei Strukturdimensionen beteiligt sind; um einen Vorgeschmack auf die realen Daten aus dem nächsten Abschnitt zu liefern, wird hier ein Data Warehouse mit vier Strukturdimensionen mit je zwei dimension members und einer Zeitdimension dargestellt (siehe Tabelle 4.25). In diesem Data Warehouse sind drei Brüche versteckt: zwischen $Jahr_1$ und $Jahr_2$ schrumpft der Wert von dimension member SD_{21} auf 20% des ursprünglichen Wertes (UPDATE, MOVE oder Teilresultat eines SPLIT), zwischen $Jahr_2$ und $Jahr_3$ tauschen SD_{31} und SD_{32} die Strukturen (UPDATE oder MOVE), und zwischen $Jahr_3$ und $Jahr_4$ gibt es eine Veränderung in der Strukturdimension SD_4: der dimension member SD_{41} verliert 70% seines Wertes an SD_{42} (MOVE). Ohne das Hintergrundwissen dieser Brüche

SD_1	SD_2	$Jahr_1$	$Jahr_2$	$Jahr_3$
SD_{11}	SD_{21}	100	105	30
SD_{11}	SD_{22}	100	105	170
SD_{12}	SD_{21}	100	105	180
SD_{12}	SD_{22}	100	105	30
SD_{11}	SD_{31}	100	105	100
SD_{11}	SD_{32}	100	105	100
SD_{12}	SD_{31}	100	105	105
SD_{12}	SD_{32}	100	105	105
SD_{21}	SD_{31}	100	105	105
SD_{21}	SD_{32}	100	105	105
SD_{22}	SD_{31}	100	105	100
SD_{22}	SD_{32}	100	105	100

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.23: Data Warehouse mit einem kombinierten Strukturbruch in zwei Dimensionen (3)

$\Delta(\%)$		$Jahr_{12}$	$Jahr_{23}$
SD_{11}	SD_{21}	5%	**71.43%**
SD_{11}	SD_{22}	5%	**61.9%**
SD_{12}	SD_{21}	5%	**71.43%**
SD_{12}	SD_{22}	5%	**71.43%**
SD_{11}	SD_{31}	5%	4.76%
SD_{11}	SD_{32}	5%	4.76%
SD_{12}	SD_{31}	5%	0%
SD_{12}	SD_{32}	5%	0%
SD_{21}	SD_{31}	5%	0%
SD_{21}	SD_{32}	5%	0%
SD_{22}	SD_{31}	5%	4.76%
SD_{22}	SD_{32}	5%	4.76%

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i, $\Delta(\%)$=Veränderung der Werte in %, $Jahr_{mn}$=Vergleich von Jahr m mit Jahr n, prozentuelle Veränderung auf Basis von Jahr m

Tabelle 4.24: Data Warehouse mit einem kombinierten Strukturbruch in zwei Dimensionen (4)

sieht man in Tabelle 4.25 zunächst lediglich, dass die Werte große Diskontinuitäten aufweisen. Untersucht man demnach jede der vier Strukturdimensionen getrennt, so erhält man die in Tabelle 4.26 dargestellten Werte; berechnet man nun die Abweichungsmatrix M_4 für die Daten aus Tabelle 4.26, so werden die Ursachen der 'brüchigen' Daten aufgedeckt, wie Tabelle 4.27

SD_1	SD_2	SD_3	SD_4	$Jahr_1$	$Jahr_2$	$Jahr_3$	$Jahr_4$
SD_{11}	SD_{21}	SD_{31}	SD_{41}	100	20	60	18
SD_{11}	SD_{21}	SD_{31}	SD_{42}	200	40	80	122
SD_{11}	SD_{21}	SD_{32}	SD_{41}	300	60	20	6
SD_{11}	SD_{21}	SD_{32}	SD_{42}	400	80	40	54
SD_{11}	SD_{22}	SD_{31}	SD_{41}	500	500	700	210
SD_{11}	SD_{22}	SD_{31}	SD_{42}	600	600	800	1290
SD_{11}	SD_{22}	SD_{32}	SD_{41}	700	700	500	150
SD_{11}	SD_{22}	SD_{32}	SD_{42}	800	800	600	950
SD_{12}	SD_{21}	SD_{31}	SD_{41}	900	180	220	66
SD_{12}	SD_{21}	SD_{31}	SD_{42}	1000	200	240	394
SD_{12}	SD_{21}	SD_{32}	SD_{41}	1100	220	180	54
SD_{12}	SD_{21}	SD_{32}	SD_{42}	1200	240	200	326
SD_{12}	SD_{22}	SD_{31}	SD_{41}	1300	1300	1500	450
SD_{12}	SD_{22}	SD_{31}	SD_{42}	1400	1400	1600	2650
SD_{12}	SD_{22}	SD_{32}	SD_{41}	1500	1500	1300	390
SD_{12}	SD_{22}	SD_{32}	SD_{42}	1600	1600	1400	2310

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.25: Brüche in einem Data Warehouse mit vier Strukturdimensionen (1)

SD	$Jahr_1$	$Jahr_2$	$Jahr_3$	$Jahr_4$
SD_{11}	3600	2800	2800	2800
SD_{12}	10000	6640	6640	6640
SD_{21}	5200	1040	1040	1040
SD_{22}	8400	8400	8400	8400
SD_{31}	6000	4240	5200	5200
SD_{32}	7600	5200	4240	4240
SD_{41}	6400	4480	4480	1344
SD_{42}	7200	4960	4960	8096

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i

Tabelle 4.26: Brüche in einem Data Warehouse mit vier Strukturdimensionen (2)

unterstreicht: es fällt auf, dass in $Jahr_2$ im Vergleich zum Vorjahr die Unterschiede in Strukturdimension SD_2 am größten sind[18] - mit über 27% liegen die Veränderungen deutlich über jenen der anderen Dimensionen; es muss in diesem Zeitraum somit eine Veränderung der Strukturen in SD_2 stattgefunden haben. Zwischen $Jahr_2$ und $Jahr_3$ verändern sich die Anteile lediglich in

[18]Bei einer Vielzahl an dimension members wird im Normalfall nur ein dimension member der jeweiligen Dimension einen Bruch aufweisen - da hier jedoch nur zwei dimension members dargestellt sind, wiegen die positiven Veränderungen des einen dimension members stets die negativen des anderen auf. Bei mehreren dimension members würden sich die gegenteiligen Auswirkungen in den nicht veränderten dimension members auf die Vielzahl der members verteilen, sodass nur mehr ein Bruch erkannt werden würde.

$M_4(\%)$	$Jahr_{12}$	$Jahr_{23}$	$Jahr_{34}$
SD_{11}	3.19%	0%	0%
SD_{12}	-3.19%	0%	0%
SD_{21}	**-27.22%**	0%	0%
SD_{22}	**27.22%**	0%	0%
SD_{31}	0.8%	**10.17%**	0%
SD_{32}	-0.8%	**-10.17%**	0%
SD_{41}	0.4%	0%	**-33.22%**
SD_{42}	-0.4%	0%	**33.22%**

Legende: SD=Strukturdimension, SD_{ij}=j-ter dimension member in Strukturdimension i, $\Delta(\%)$=absolute Veränderung der Anteile einzelner dimension members am Gesamtkuchen zwischen einzelnen Jahren, $Jahr_{mn}$=Vergleich der Anteile einzelner dimension members zwischen Jahr m und Jahr n

Tabelle 4.27: Brüche in einem Data Warehouse mit vier Strukturdimensionen (3)

Strukturdimension SD_3, und dort mit über 10% relativ drastisch - somit kann in diesem Intervall nur in SD_3 eine Veränderung vorgefallen sein. Ähnliches gilt für den Vergleich von $Jahr_3$ mit $Jahr_4$: auch hier haben sich nur die Anteile in SD_4 verschoben, und auch hier mit über 33% so dramatisch, dass sich hier ein Strukturbruch eingeschlichen hat[19].

Das Verfahren lässt sich in einem Data Warehouse mit n Strukturdimensionen zur Erkennung von Brüchen in bis zu n Dimensionen verallgemeinern, indem im nächsten Schritt des Algorithmus stets Strukturkombinationen analysiert werden, deren Anzahl an Strukturdimensionen um eins über der Anzahl aus dem vorhergehenden Schritt liegt. Aus den folgenden zwei Gründen wird dennoch bei der Erkennung von Strukturbrüchen mit Werten, die nach einer Strukturdimension gruppiert sind, begonnen:

1. In der überwiegenden Anzahl der Strukturbrüche in den Strukturdimensionen wird lediglich eine Strukturdimension (i.e. ein, mehrere oder alle dimension members dieser Dimension) von dieser Diskontinuität betroffen sein. Hat man bereits auf dieser Ebene zahlreiche Brüche entdeckt, so wird man in der Regel nicht mehr kombinierte Brüche betrachten.

2. Der Aspekt der Performance spricht für die primäre Untersuchung von Brüchen in ein-

[19]Dieses Ergebnis zeigt auf, dass ein Bruch in einer Dimension, bei dem die Strukturen lediglich zwischen einzelnen dimension members verschoben werden, summiert jedoch keine Veränderungen auftreten, keine Auswirkungen auf andere Dimensionen hat. Ändern sich hingegen dimension members einer Strukturdimension so, dass die neue Summe größer oder kleiner als die alte Summe ist, so ergibt dies auch Veränderungen in anderen Dimensionen, die allerdings wesentlich geringer ausfallen als in der tatsächlich veränderten Strukturdimension. In realen Data Warehouses wird eine Veränderung der Anteile von null in einer Strukturdimension aber zumeist nicht festzustellen sein, da die Werte auch ohne Strukturveränderungen in geringem Ausmaße schwanken - somit kann die Analyse lediglich Auskunft darüber geben, dass ein Strukturbruch existiert; sie kann im Normalfall jedoch nicht sagen, welcher Art dieser war.

zelnen Strukturdimensionen: seien D_1, D_2, ..., D_n die Anzahl der Elemente in der i-ten Strukturdimension ($i = 1 \ldots n$; geordnet, sodass $D_1 \geq D_2 \geq \ldots \geq D_n$), so müssen im ersten Schritt nur $D_1 + D_2 + \ldots + D_n = O(D_1)$ verschiedene Werte analysiert werden, auf der zweiten Ebene bereits $D_1D_2 + \ldots D_1D_n + \ldots + D_{n-1}D_n = O(D_1^2)$, auf der i-ten Ebene somit $O(D_1^i)$, $i = 1, \ldots, n$.

4.4 Conclusio der vorherigen Abschnitte: schrittweises Vorgehen

Als Konsequenz der unterschiedlichen Laufzeitkomplexitäten der untersuchten Methoden sowie deren Einsetzbarkeit auf n Dimensionen kann folgende schrittweise Vorgehensart vorgeschlagen werden:

1. Zunächst wird die gesamte Datenmatrix auf eventuelle Strukturbrüche in der Kennzahlendimension untersucht (Veränderungen der Berechungsvorschrift einer Kennzahl, Änderung der Maßeinheit einer Kennzahl). Bei performanzkritischen Applikationen eignet sich dafür in erster Linie eine Abweichungsmatrix, die die Differenz der Summen aller Absolutbeträge (oder euklidische Distanz) zwischen den beiden Chrononen berechnet; ist die Performanz weniger kritisch, können auch eine Singulärwertzerlegung oder diskrete Cosinus-Transformation der Matrizen durchgeführt werden. Treten bereits hier Brüche auf, so muss versucht werden, ihre Ursachen zu finden, um sie bereinigen zu können; können diese Brüche nicht eliminiert werden, so bleibt nur noch die Möglichkeit, einzelne 'brüchige' Chronone aus der Analyse herauszunehmen[20].

2. Im nächsten Schritt wird jede Strukturdimension getrennt analysiert; die Abweichungsmatrizen M_1, M_2 und vor allem M_4 sind hier (wie oben beschrieben) einzusetzen.

3. Kann die Datenmatrix sinnvoll mit einer Differenzialgleichung (bzw. mit einer einfachen Funktionsgleichung wie im Falle der Zufallszahlen von Kapitel 4.2) modelliert werden, so lohnt es sich, die Abweichungsmatrizen mit den absoluten und relativen Differenzen (M_1 und M_2) zwischen dem vom Modell prognostizierten und dem tatsächlichen Wert jedes dimension members zu jedem Zeitpunkt zu berechnen - die Rechenzeit entspricht jener der Berechnung der einfachen Abweichungsmatrizen.

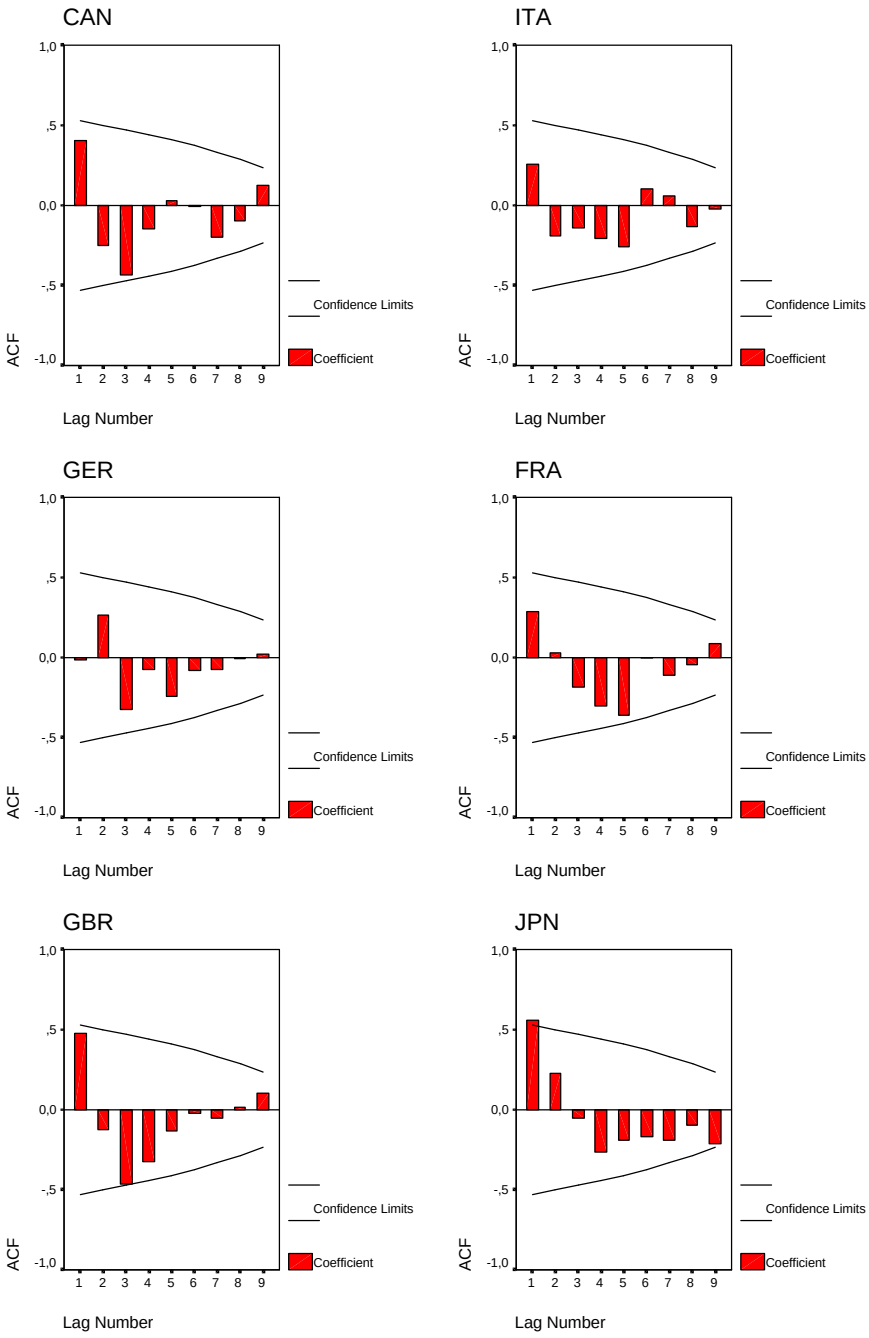
4. Kennt man in jeder Strukturdimension einen dimension member, der im Untersuchungszeitraum definitiv keinem Strukturbruch unterlegen ist, so können Autokorrelation, Auto-

[20] Nimmt man die Brüche auf dieser Ebene nicht heraus, kann man in nachfolgenden Schritten keine weiteren Brüche mehr sinnvoll entdecken.

regression, DFT, Wavelet-Transformationen, PCA, Kreuzkorrelation und lineare Regression so eingesetzt werden, dass die Ergebnisse (bzw. Koeffizienten) der in obigen Schritten aufgedeckten brüchigen dimension members mit denen des definitiv nicht veränderten dimension members verglichen werden[21]. Auch wenn die Performanz auf dieser Stufe nicht mehr so kritisch sein sollte, empfiehlt es sich, zunächst die Methode der Autokorrelation zu versuchen; will man Wavelet-Transformationen verwenden, ist aus Performanzgründen das Haar-Wavelet vor der DWT als Sequenz linearer Transformationen einzusetzen.

5. Können bei der Analyse einzelner Strukturdimensionen keine Strukturbrüche erkannt werden, und vermutet man dennoch Brüche, so gilt es, Kombinationen von Strukturdimensionen zu untersuchen, deren Anzahl an Strukturdimensionen um eins größer ist als die Anzahl im aktuellen Schritt, und es können wiederum schrittweise die verschiedenen Methoden aus 2.), 3.) und 4.) eingesetzt werden.

[21] Kann ein solcher unveränderter dimension member nicht definiert werden, so kann der Koeffizientenvergleich der brüchigen dimension members auch mit der Gesamtentwicklung der Werte (Summe aller Absolutbeträge aller Werte) erfolgen, sofern der Anteil der Werte in brüchigen dimension members im Verhältnis zur Gesamtsumme aller Werte gering ist - in Kapitel 5 wird diese Vergleichsmethode benutzt.



Legende: ACF=Autokorrelationskoeffizienten zu Lags 1 bis 9; Confidence limits zeigen die Grenzen an, ab deren Überschreitung der ACF-Werte die Hypothese der Zufallsverteiltheit der Werte mit 95% Sicherheit widerlegt werden kann

Abbildung 4.2: Korrelogramme der jährlichen BIP-Wachstumsraten ausgewählter Länder

5

Erprobung der Methoden auf einem realen Data Warehouse

In diesem Kapitel wird die oben präsentierte Vorgehensweise (Abschnitt 4.4) zur Erkennung von Strukturbrüchen an einem realen Data Warehouse erprobt. Zunächst wird in Abschnitt 5.1 erläutert, welche möglicherweise problematischen Transformationen der Daten notwendig sind, um diese in das Eingabeformat des Algorithmus zu bringen, anschließend sind in Abschnitt 5.2 die Resultate der Anwendung dieses Verfahrens dokumentiert. Kapitel 5.3 zeigt daraufhin einerseits allgemein mögliche Interpretationen von Strukturbrüchen auf, andererseits weist es auf Probleme hin, die speziell in Verbindung mit dem verwendeten Data Warehouse stehen.

5.1 Datenquelle und notwendige Vortransformationen

Die bislang auf präparierten bzw. randomisierten Daten erprobten Methoden zur Erkennung von Strukturbrüchen sollen nun auf Realitätsdaten eines Krankenhaus-Data Warehouses angewendet werden; das Data Warehouse beinhaltet die Kennzahl der Kosten aufgegliedert nach vier aus Datenschutzgründen verschlüsselten Strukturdimensionen und zwei Zeitdimensionen (Jahr und Monat); das ursprüngliche Format der Daten

$$[SD_1,\ Jahr,\ Monat,\ SD_2,\ SD_3,\ SD_4]\ Wert$$

muss dazu in das in obigen Abbildungen stets verwendete tabellarische Format

$$[SD_1,\ SD_2,\ SD_3,\ SD_4]\ Wert_{Jahr_1\ Monat_1},\ Wert_{J_1\ M_2},\ldots,Wert_{J_1\ M_{12}},\ldots,Wert_{Jahr_n\ Monat_m}$$

($n = \#Jahre$, $m = \#Monate\ in\ Jahr_n$) gebracht werden, wobei sich diese Transformation als relativ kompliziert gestaltet: die Monatsnamen müssen in numerische Werte umgewandelt werden (Jänner=1, ..., Dezember=12), alle fehlenden Werte (im Textfile als '#MI' gekennzeichnet) müssen eliminiert werden. Für das Laden der Daten bietet MATLAB zwar ähnlich wie C einen sehr einfachen *fread()* bzw. einen *textread()*-Befehl an, der ein nach Wunsch formatiertes Einlesen der Daten erlaubt, die Performanz dieses Befehls ist jedoch außerordentlich schlecht - auch nach rund zwei Tagen Laufzeit sind die rund 38 MB großen Daten (rund 650 000 Zeilen) nicht eingelesen. Somit werden die Daten in eine relationale Datenbank importiert, um dort mit entsprechenden SQL-Operationen die notwendigen Joins durchführen zu können. Die fertigen Tabellen mit den Werten der Kennzahl 'Kosten' und den Strukturdimensionen (nach Elimination von nicht in allen Chrononen vorhandenen dimension members bleiben in Strukturdimension SD_1 523 dimension members, in SD_2 8, in SD_3 2 und in SD_4 385 dimension members übrig; es gibt 10656 Kombinationen von dimension members) werden nun wieder zurück in ein textuelles Format gebracht - aufgrund der nun wesentlich geringeren Größe und der Trennung in eine Matrix mit den (numerischen) Werten und eine zweite Matrix mit den Bezeichnungen der jeweiligen dimension members können die Daten nun problemlos in MATLAB geladen werden.

5.2 Ergebnisse der Methoden

5.2.1 Analyse des gesamten Datenbestandes

Im folgenden werden die Resultate der Methoden präsentiert; es wird dabei die im vorangegangenen Kapitel vorgeschlagene mehrstufige Methode angewendet, wobei die Strukturdimensionen zunächst getrennt betrachtet werden. Vor diesem Schritt wird allerdings im ersten Schritt überprüft, ob zwischen zwei aufeinanderfolgenden Chrononen möglicherweise Strukturbrüche erkannt werden können, die die Kennzahlendimension und somit alle dimension members betreffen: dazu werden in der ursprünglichen Datenstruktur $(SD_1, \ldots, SD_4, Val(Jahr_1 Monat_1), \ldots, Val(Jahr_n Monat_m))$[1] zwischen allen aufeinanderfolgenden Chrononen die euklidischen Distanzen[2] der Differenzen der Werte aller Strukturkombinationen zwischen diesen beiden Intervallen aufsummiert[3]; die jeweiligen Quadratwurzeln der resultierenden Differenzen sind in Tabelle 5.1 in absteigender Ordnung dargestellt.

- Laufzeit der Berechnung: 3.13 sec[4].

[1] Im untersuchten Data Warehouse ist n = drei und m = sieben → es gibt insgesamt 31 Chronone.

[2] Es wird in gewissem Sinne eine insgesamte Abweichungsmatrix über alle Werte zwischen je zwei Zeitintervallen berechnet.

[3] Die summierten Differenzen an sich (ohne Quadrierung bzw. Betrag) können nicht herangezogen werden, da sonst negative Werte positive ausgleichen könnten.

[4] Diese Methode wurde wie alle anderen folgenden auf einem Pentium III-Prozessor mit 866 MHz und 128 MB

$Monate$	$\Delta * 10^7$
12-13	5.1247
11-12	3.3383
7-8	1.3002
1-2	1.2313
2-3	1.0609
$24-25$	0.8618
6-7	0.736
8-9	0.7207
5-6	0.6992
4-5	0.6904
$23-24$	0.6889
3-4	0.6401
9-10	0.5868
10-11	0.5401
$26-27$	0.1737
$19-20$	0.1484
$20-21$	0.1467
$27-28$	0.1389
$28-29$	0.1290
$30-31$	0.1196
$13-14$	0.0994
$25-26$	0.0913
$15-16$	0.0910
$18-19$	0.0901
$17-18$	0.0852
$22-23$	0.0823
$14-15$	0.0789
$16-17$	0.0724
$21-22$	0.0693
$29-30$	0.0671

Legende:Δ=Summe der euklidischen Distanzen aller dimension members zwischen jeweils zwei Monaten

Tabelle 5.1: Erste Abweichungsanalyse auf dem Real-World-Data Warehouse

Bei Betrachtung der Ergebnisse fallen zwei Dinge auf: einerseits sind die Differenzen zwischen den Monaten zwölf und 13 sowie zwischen elf und zwölf außergewöhnlich hoch, andererseits befinden sich unter den differenzmäßig größten 14 Zeilen alle Monatsvergleiche des ersten Jahres (nebst zwei weiteren); nach der 14. Zeile ist ein abrupter Rückgang der Differenz zu den weiteren Zeilen zu verzeichnen. Nachforschungen bei Domain-Experten machen klar: die Werte der ersten zwölf

RAM durchgeführt.

Monate waren in ATS, die Werte danach in EURO angegeben. Sämtliche Werte der ersten zwölf Monate werden daher korrigiert, sie werden durch den Umrechungsfaktor der beiden Währungen (13.7603) dividiert. Die Abweichungsanalyse wird nun nochmals angewandt, um eventuell einen weiteren Bruch identifizieren zu können (siehe Tabelle 5.2)[5]. Betrachtet man nun Tabelle 5.2,

$Monate$	$\Delta * 10^6$
24-25	8.6182
23-24	6.8894
12-13	3.7099
11-12	2.4261
$26-27$	1.7369
$19-20$	1.4837
$20-21$	1.4673
$27-28$	1.3887
$28-29$	1.2896
$30-31$	1.1962
$13-14$	0.9941
$7-8$	0.9449
$25-26$	0.9132
$15-16$	0.9096
$18-19$	0.9007
$1-2$	0.8949
$17-18$	0.8518
$22-23$	0.8226
$14-15$	0.789
$2-3$	0.771
$16-17$	0.7243
$21-22$	0.6926
$29-30$	0.6707
$6-7$	0.5349
$8-9$	0.5238
$5-6$	0.5081
$4-5$	0.5018
$3-4$	0.4652
$9-10$	0.4265
$10-11$	0.3925

Legende:Δ=Summe der euklidischen Distanzen aller dimension members zwischen jeweils zwei Monaten

Tabelle 5.2: Zweite Abweichungsanalyse auf dem Real-World-Data Warehouse

[5]Die Laufzeit des Verfahrens beträgt hier und in der dritten Iteration dieses ersten Schrittes wie oben 3.13 Sekunden.

so sind die meisten Differenzen zwischen Monaten des ersten Jahres im Hinterfeld anzutreffen. Auffällig sind nun aber hohe Differenzen zwischen den Monaten 24-25, 23-24, 12-13 und 11-12: durch nochmaliges Nachfragen bei Domain-Experten wird dies dadurch erklärt, dass die gesamte interne Leistungsverrechnung aller Kostenstellen im Dezember eines jeweiligen Jahres berücksichtigt wird; alle anderen Monate beinhalten Werte vor dieser Umlage. Um auf einen komplizierten Umlageschlüssel, der diese Verzerrung kompensieren kann, zu verzichten, werden die Dezemberwerte aller untersuchten Jahre aus der weiteren Analyse ausgeschlossen (i.e. der Jänner des aktuellen Jahres wird mit dem November des Vorjahres verglichen). Analysiert man die nochmals adaptierten Daten erneut mit derselben Rechenmethode, gelangt man zu dem in Tabelle 5.3 aufgeführten Ergebnis; es können nun keine Ausreißer mehr identifiziert werden.

Die Ergebnisse der dreimaligen Anwendung der Singulärwertzerlegung mit zwei Singulärwerten (mit dem Wissen über die Brüche aus den Abweichungsmatrizen) sind anschließend zu Zwecken der Gegenüberstellung mit den Resultaten der Abweichungsmatrix in Tabelle 5.4 aufgelistet. Es kommt dabei klar heraus, dass die Singulärwertzerlegung sehr gut abrupte Brüche zwischen einzelnen Chrononen (plötzlicher Währungswechsel) entdecken kann, hingegen jedoch nicht gut zur Identifikation schleichender Brüche (andere Währung über mehrere Monate hinweg) geeignet ist - mit der Singulärwertzerlegung hätte man zwar erkennen können, dass zwischen Monat 12 und Monat 13 die Währung wechselte (dies zeigt sich durch hohe Differenzen in allen Intervallen, in denen Monat 12 und Monat 13 vorkommen), man hätte aber aufgrund der Skalierung der Distanzen nicht sofort daraus schließen können, dass in den Monaten 1 bis 12 eine andere Währung als danach vorherrschte.

- Laufzeit der Singulärwertzerlegung in der 1. Iteration (zwei Singulärwerte): 383.98 sec.
- Laufzeit in der 2. Iteration (zwei Singulärwerte): 382.56 sec.
- Laufzeit in der 3. Iteration (zwei Singulärwerte): 354.49 sec.

Die Ergebnisse der Anwendung der diskreten Cosinus-Transformation[6] mit einer 2×2-Koeffizientenmatrix (wiederum mit dem Wissen der Brüche aus den Abweichungsmatrizen) sind ebenso primär zu Vergleichszwecken in Tabelle 5.5 dargestellt. Die DCT deckt im Gegensatz zur Singulärwertzerlegung in der ersten Iteration zwar auf, dass in den Monaten 1 bis 12 aufgrund der viel höheren Differenzen möglicherweise eine andere Währung vorherrschte, sie kann aber kurze, abrupte Brüche (Probleme mit der internen Leistungsverrechnung in den Monaten 12 und 24) nicht erkennen - bereits in der zweiten Iteration (nach Vereinheitlichung der unterschiedlichen Währungen) fallen keine Werte mehr auf.

[6]Die DCT wird dabei in der Variante, die keine Skalierung der DCT-Koeffizienten mit dem Mittelwert der jeweiligen Matrizen (vgl. dazu Kapitel 3.8.2) vorsieht, durchgeführt.

$Monate$	$\Delta * 10^6$
$22-23$	1.8792
$24-25$	1.7369
$11-12$	1.4939
$18-19$	1.4837
$19-20$	1.4673
$25-26$	1.3887
$26-27$	1.2896
$28-29$	1.1962
$12-13$	0.9941
$7-8$	0.9449
$23-24$	0.9132
$14-15$	0.9096
$17-18$	0.9007
$1-2$	0.8949
$16-17$	0.8518
$21-22$	0.8226
$13-14$	0.789
$2-3$	0.771
$15-16$	0.7243
$20-21$	0.6926
$27-28$	0.6707
$6-7$	0.5349
$8-9$	0.5238
$5-6$	0.5081
$4-5$	0.5018
$3-4$	0.4652
$9-10$	0.4265
$10-11$	0.3925

Legende:Δ=Summe der euklidischen Distanzen aller dimension members zwischen jeweils zwei Monaten

Tabelle 5.3: Dritte Abweichungsanalyse auf dem Real-World-Data Warehouse

- Laufzeit der DCT (2×2-Koeffizientenmatrix) in der 1. Iteration: 352.51 sec.
- Laufzeit in der 2. Iteration (2×2-Koeffizientenmatrix): 350.37 sec.
- Laufzeit in der 3. Iteration (2×2-Koeffizientenmatrix): 323.13 sec.

Eine Kombination von DCT und Singulärwertzerlegung könnte zwar alle Strukturbrüche aufdecken, wenn man aber die wesentlich höheren Laufzeiten von DCT und Singulärwertzerlegung im Vergleich zu der einfachen Abweichungsmatrix heranzieht, so ist letzterem Verfahren bei gleich guten Resultaten aus Performanzgründen eindeutig der Vorzug zu geben. Da in der

$Monate$	$\Delta(SV)DW_1$	$\Delta(SV)DW_2$	$\Delta(SV)DW_3$
$1-4; 2-5$	67.4	67.4	67.4
$2-5; 3-6$	139.8	139.8	139.8
$3-6; 4-7$	79.4	79.4	79.4
$4-7; 5-8$	164.4	164.4	164.4
$5-8; 6-9$	21.3	21.3	21.3
$6-9; 7-10$	41.3	41.3	48.33
$7-10; 8-11$	93.1	93.1	93.1
$8-11; 9-12$	**662.6**	**662.6**	59.9
$9-12; 10-13$	**566.9**	76.5	72.3
$10-13; 11-14$	**774.7**	42.8	48.9
$11-14; 12-15$	**1286.9**	125	107.3
$12-15; 13-16$	**3237.0**	**589.5**	-
$13-16; 14-17$	110.2	110.2	110.2
$14-17; 15-18$	239.3	239.3	239.3
$15-18; 16-19$	25.6	25.6	25.6
$16-19; 17-20$	256.9	256.9	256.9
$17-20; 18-21$	115.9	115.9	115.9
$18-21; 19-22$	70.4	70.4	70.4
$19-22; 20-23$	174.8	174.8	174.8
$20-23; 21-24$	**2034.6**	**2034.6**	**531.7**
$21-24; 22-25$	136.8	136.8	156.6
$22-25; 23-26$	47.9	47.9	198.4
$23-26; 24-27$	65.2	65.2	164
$24-27; 25-28$	**2320.1**	**2320.1**	-
$25-28; 26-29$	250.4	250.4	250.4
$26-29; 27-30$	134.2	134.2	134.2
$27-30; 28-31$	234.0	234.0	234.0

Legende: $\Delta(SV)DW_i$=Differenz der Singulärwerte auf Data Warehouse i, i=1...3

Tabelle 5.4: Singulärwertzerlegung auf dem ursprünglichen Data Warehouse (DW_1), nach EURO/ATS-Korrektur (DW_2) und nach ILV-Korrektur (DW_3)

dritten Iteration mittels der Abweichungsmatrix, der Singulärwertzerlegung und der DCT (bei dieser bereits in der zweiten Iteration) keine Ausreißer mehr festgestellt werden können, werden die auf diese Weise korrigierten Daten als Input für die Methoden auf der nächsten Ebene verwendet.

5.2.2 Getrennte Analyse der Strukturdimensionen

Zunächst werden auf jeder der vier Strukturdimensionen klassische Abweichungsmatrizen berechnet: ein Strukturbruch in einer Dimension liegt dann vor, wenn der Absolutunterschied in

$Monate$	$\Delta(coeffs)DW_1 * 10^6$	$\Delta(coeffs)DW_2 * 10^6$	$\Delta(coeffs)DW_3 * 10^6$
$1-4; 2-5$	**111.7**	8.12	8.12
$2-5; 3-6$	**158.63**	11.52	11.52
$3-6; 4-7$	**215.45**	15.66	15.66
$4-7; 5-8$	**210.2**	15.28	15.28
$5-8; 6-9$	**292.41**	21.25	21.25
$6-9; 7-10$	**235.2**	17.09	17.09
$7-10; 8-11$	**407.81**	29.64	29.64
$8-11; 9-12$	**150.55**	10.94	6.99
$9-12; 10-13$	**138.4**	11.5	11.76
$10-13; 11-14$	**123.41**	9.02	10.7
$11-14; 12-15$	**102.12**	4.95	12.07
$12-15; 13-16$	**113.7**	19.01	-
$13-16; 14-17$	3.65	3.65	3.65
$14-17; 15-18$	16.04	16.04	16.04
$15-18; 16-19$	2.6	2.6	2.6
$16-19; 17-20$	4.29	4.29	4.29
$17-20; 18-21$	11.04	11.04	11.04
$18-21; 19-22$	9.33	9.33	9.33
$19-22; 20-23$	11.41	11.41	11.41
$20-23; 21-24$	31.01	31.01	8.88
$21-24; 22-25$	4.82	4.82	7.98
$22-25; 23-26$	41.61	41.61	15.11
$23-26; 24-27$	23.85	23.85	7.81
$24-27; 25-28$	20.11	20.11	-
$25-28; 26-29$	18.62	18.62	18.62
$26-29; 27-30$	6.19	6.19	6.19
$27-30; 28-31$	10.19	10.19	10.19

Legende: $\Delta(coeffs)DW_1$=Differenz der DCT-Koeffizienten auf Data Warehouse i, i=1...3

Tabelle 5.5: Diskrete Cosinus-Transformation auf dem ursprünglichen Data Warehouse (DW_1), nach EURO/ATS-Korrektur (DW_2) und nach ILV-Korrektur (DW_3)

zwei aufeinanderfolgenden Werten eines dimension members (M_1) mehr als 50 000 beträgt, der relative Unterschied zwischen zwei Werten mehr als 80% ausmacht (M_2, bezogen auf den jeweils größeren Absolutbetrag[7] der beiden Vergleichswerte), und die Anteilsveränderung des dimension members zwischen zwei Jahren mehr als 1% absolut und mehr als 85% relativ bezogen auf den größeren Absolutbetrag der beiden Werte beträgt (M_4, Beispiel: steigt der Anteil eines dimension members von 1% auf 7%, so ist die Absolutveränderung 6% und die relative Veränderung

[7]Damit Strukturbrüche mit einem hohen positiven Wert und einem hohen negativen Wert in zwei aufeinanderfolgenden Monaten dennoch als solche erkannt werden, werden verschiedene Vorzeichen extra berücksichtigt - hier gilt lediglich die Einschränkung, dass die Absolutdifferenz der beiden Werte mehr als 75 000 betragen muss - dies entspricht in etwa der kleinsten Absolutdifferenz eines mit obigen Kriterien aufgedeckten Strukturbruchs.

85.7%)[8]. Zudem werden jene dimension members, in denen mehr als fünf der nunmehr 29 Werte null bzw. fehlend sind, aus der Berechnung herausgenommen.

Wendet man dieses Verfahren auf die Strukturdimension SD_1 an, in der bekannt ist, dass lediglich ein dimension member (A729103292) im Vergleichszeitraum einem Strukturbruch unterlegen war, so erhält man 25 dimension members (allerdings ohne diesen einen), die als brüchig kategorisiert werden. Tabelle 5.6 zeigt die gefundenen dimension members mit den jeweiligen an einem Strukturbruch beteiligten Chrononen und Werten auf. Die in Tabelle 5.6 aufgezeigten Werte schwanken so stark, dass es sich hierbei kaum um natürliche Streuungen handeln kann; da es jedoch definitiv nicht Strukturbrüche sind, könnte es sich nur um Eingabefehler handeln - zur Interpretation der Ergebnisse vgl. Kapitel 5.3. Untersucht man mit gleicher Wahl der Parameter die Dimensionen SD_2 und SD_4[9], in denen Strukturbrüche nicht bekannt sind[10], so erhält man das in Tabelle 5.7 aufgeführte Ergebnis - auch in den Strukturdimensionen SD_2 und SD_4 werden einige dimension members (drei bzw. 20) als brüchig klassifiziert.

- Laufzeit der Methode auf Strukturdimension SD_1: 1.21 sec.
- Laufzeiten auf Strukturdimensionen SD_2 und SD_3: 0 sec.
- Laufzeit auf Strukturdimension SD_4: 0.6 sec.

Würde man die Strukturbruchsanalyse nun auch noch kombiniert für die Strukturdimensionen SD_1 und SD_4 durchführen, obwohl es auf unterer Ebene bereits zahlreiche Brüche gegeben hat (was somit der vorgeschlagenen Vorgangsweise widerspricht), so wäre das Resultat bei gleicher Einstellung der Parameter wie oben nicht das gleiche, wie das in Tabelle 5.8 dargestellte Resultat unterlegt. Auffallend bei der Betrachtung von Tabelle 5.8 ist, dass der dimension member $F223683619$ (Strukturdimension SD_4) in Kombination mit fünf verschiedenen dimension members aus Strukturdimension SD_1 als Strukturbruch vorkommt, obwohl er bei getrennter Betrachtung der Dimensionen (vgl. Tabelle 5.7) nicht ein einziges Mal als Bruch auftritt - man könnte somit meinen, hier gäbe es einen Bruch, der von obiger Analyse nicht aufgedeckt wurde. Bei näherer Betrachtung der Daten fällt allerdings auf, dass dieser dimension member den mit Abstand größten Wert in allen Monaten ausweist und daher nur deshalb so häufig in Tabelle 5.8 erscheint, weil er die Schwellwerte zur Elimination von kleinen Werten 'leicht' überbieten

[8] Ist allerdings nur einer der absoluten oder relativen Werte in den beiden miteinander verglichenen Monaten null bzw. fehlend, so wird kein Bruch festgestellt.

[9] In Strukturdimension SD_3 kann kein Bruch festgestellt werden - es existieren dort allerdings auch nur zwei dimension members.

[10] Selbst von Domain-Experten kann nicht mehr nachvollzogen werden, welche dimension members in Strukturdimension SD_4 im Analysezeitraum Strukturbrüche zu verzeichnen hatten, da die Regeln der Zubuchung von Transaktionen zu den einzelnen dimension members dieser Strukturdimension häufig verändert werden (vgl. Kapitel 5.3). In Strukturdimension SD_2 ist hingegen bekannt, dass keine Strukturbrüche in diesem Zeitintervall vorgefallen sind.

dim member	*Intervall*	$Wert_1$	$Wert_2$	Anz
A102638053	25-26	-1846.4	-167 230	2
A127906421	7-8	803 600	78.92	2
A135105873	26-27	591 770	-350 370	8
A175546859	17-18	-76 800	19 459	2
A176492395	21-22	141 450	-223.53	3
A233252045	16-17	170 450	15 830	1
A315347857	16-17	50 660	360 830	2
A346226871	4-5	21 898	219 220	1
A378566270	17-18	-74 805	18 692	2
A443164712	25-26	134.21	159 190	7
A468143681	13-14	109 110	5131.3	1
A502408343	23-24	134 930	17 950	2
A502626747	1-2	6568.2	830 940	1
A546705177	19-20	790.94	164 590	2
A549964299	25-26	3311.2	136 300	1
A633897121	12-13	4236.9	329 750	5
A768474977	27-28	1508	142 160	1
A776617093	1-2	78 825	3104.5	1
A799350957	24-25	179.89	125 360	1
A841768422	21-22	302 930	-304 460	3
A855748746	26-27	-278 570	296 630	3
A864725209	22-23	2690.7	143 180	3
A882270984	3-4	17 868	175 070	1
A913796677	22-23	195.63	200 190	5
A971709069	21-22	192 640	2954.3	3

Legende: *dim member*=dimension member, Intervall=die beiden Monate mit der größten Differenz (zum Beweis deren Werte $Wert_1$ und $Wert_2$), Anz=Anzahl der als brüchig klassifizierten Differenzen zweier Monate dieses speziellen dimension members (hier sind nur die Werte der betragsmäßig größten Differenz jedes members dargestellt)

Tabelle 5.6: Erkannte Strukturbrüche in Strukturdimension SD_1 mit der Methode der Abweichungsmatrix

kann; eine kombinierte Strukturdimensionsanalyse ist demnach - wie bereits oben erwähnt - nicht zielführend, wenn bereits Brüche bei getrennter Betrachtung der Dimensionen auftraten.

5.2.3 Vergleichsanalysen einzelner dimension members

Um die im vorigen Schritt aufgedeckten Strukturbrüche einer Plausibilitätskontrolle zu unterziehen, wird nun eine Auswahl der in Kapitel 3 vorgestellten Analysen auf Ebene einzelner Datenvektoren angewendet. Diese Verfahren sollen vor allem dazu dienen, jene oben identifizierten Strukturbrüche, die einmaliger Art sind und möglicherweise in die Kategorie 'Eingabefehler'

dim member	*Intervall*	$Wert_1$	$Wert_2$	Anz
D160866068	1-2	44 764	727 040	1
D380554747	1-2	5387.5	765 090	2
D430391667	11-12	-1 870 800	-146 350	2
F108498872	7-8	890 120	4370.2	7
F112308097	18-19	-414 850	109 990	3
F112312887	1-2	9626.7	122 340	1
F139701034	21-22	336 270	-268 010	3
F196028405	1-2	110 040	7813.7	1
F239909113	24-25	29 187	271 940	1
F270414432	1-2	4411.8	117 520	1
F366256661	13-14	2660.3	228 510	3
F378476541	22-23	276.81	749 870	5
F386807789	1-2	8956.5	430 040	5
F429797213	4-5	-46.39	280 990	8
F535841867	28-29	-32 524	-201 940	1
F657245554	17-18	-322 040	-34 409	6
F701638710	1-2	5704.6	873 210	1
F836945175	12-13	3120.9	119 900	4
F851172809	22-23	211 627	12 955	2
F920147523	25-26	-14 799	-168 850	3
F938244641	13-14	8342.2	230 110	2
F971043799	11-12	103 190	478.44	1
F979663437	4-5	17 606	206 930	2

Legende: *dim member*=dimension member, Intervall=die beiden Monate mit der größten Differenz (zum Beweis deren Werte $Wert_1$ und $Wert_2$), Anz=Anzahl der als brüchig klassifizierten Differenzen zweier Monate dieses speziellen dimension members (hier sind nur die Werte der betragsmäßig größten Differenz jedes members dargestellt)

Tabelle 5.7: Erkannte Strukturbrüche in den Strukturdimensionen SD_2 und SD_4 mit der Methode der Abweichungsmatrix

einzuordnen sind, herauszufiltern. Als Vergleichsmaßstab wird dabei in allen nun folgenden Analysen die Entwicklung der Summe aller Absolutbeträge aller Vektoren in der jeweiligen Strukturdimension herangezogen.

Diskrete Fourier-Transformation

Die Methode der DFT wird verwendet, um jeweils ein Intervall von acht aufeinanderfolgenden Monaten der Absolutbeträge der Werte[11] eines der in den Tabellen 5.6 und 5.7 als Strukturbruch

[11]Es werden bei der DFT sowie im folgenden bei den Wavelet-Transformationen und der PCA stets die Absolutbeträge der Werte der dimension members herangezogen, da die Anzahl der dimension members mit abwechselnd positiven und negativen Werten sehr gering ist - somit wird das Vorzeichen vernachlässigt.

SD_1 *dim member*	SD_4 *dim member*	Intervall	$Wert_1$	$Wert_2$	Anz
*A*102638053	*F*920147523	25-26	-1846.4	-166 990	2
*A*123023807	*F*112308097	25-26	-112 900	15 173	1
*A*127906421	*F*108498872	7-8	803 600	78.92	2
*A*135105873	*F*112308097	18-19	-336 750	200 160	3
*A*135105873	**F223683619**	25-26	8090.4	1 059 500	2
*A*159866326	*F*300483476	1-2	10 878	155 650	1
*A*233252045	*F*444326175	16-17	167 410	8858.5	2
*A*276450736	*F*851172809	21-22	316 420	51 447	1
*A*315347857	*F*851172809	22-23	142 830	75.79	2
*A*346226871	*F*979663437	4-5	17 606	206 930	2
*A*468143681	*F*386807789	13-14	109 080	5131.3	1
*A*502408343	*F*112312887	11-12	10 298	93 029	1
*A*549964299	*F*366256661	25-26	3311.2	136 300	1
*A*633897121	*F*112312887	24-25	24 750	367 180	4
*A*633897121	*F*377698827	3-4	132 550	1042.1	1
*A*662149231	**F223683619**	24-25	8350.9	134 980	2
*A*667148201	**F223683619**	14-15	143 240	20 445	1
*A*855491625	**F223683619**	28-29	21 547	137 260	1
*A*855748746	**F223683619**	26-27	-290 150	291 070	3
*A*864725209	*F*231632116	22-23	519.69	124 770	2
*A*882270984	*F*851172809	11-12	141 520	163.88	2

Legende: SD_i=i-te Strukturdimension, *dim member*=dimension member, Intervall=die beiden Monate mit der größten Differenz (zum Beweis deren Werte $Wert_1$ und $Wert_2$), Anz=Anzahl der als brüchig klassifizierten Differenzen zweier Monate dieses speziellen dimension members (hier sind nur die Werte der betragsmäßig größten Differenz jedes members dargestellt)

Tabelle 5.8: Erkannte Strukturbrüche bei Kombination der Strukturdimensionen SD_1 und SD_4 mit der Methode der Abweichungsmatrix

identifizierten dimension members, in dem kein Wert null ist[12], mit der Gesamtentwicklung der summierten Absolutbeträge der Werte aller dimension members im Analysezeitraum zu vergleichen. Bei Schwellwerten von 11 für die maximale Distanz und 2.3 für die längennormierte summierte Distanz der nunmehr skalierten Fourier-Koeffizienten[13] werden von den bereits in den

[12]Diese Einschränkung gilt auch für die nachfolgenden Methoden der Principal Component Analysis und der Haar-Wavelet-Transformation.

[13]Die Wahl der Schwellwerte sollte im Normalfall auf Basis einer maximum-margin-Methode basieren, i.e. die Schwellwerte sollten so angesetzt werden, dass der Abstand zwischen den als brüchig und nicht-brüchig klassifizierten Elementen maximal ist; durch geeignete Gewichtung verschiedener Faktoren (maximale und insgesamte Distanz in diesem Fall) wäre dies auch bei mehr als einem Parameter möglich. Der Nachteil dieses Verfahrens ist aber darin zu sehen, dass die Anzahl der erkannten Strukturbrüche von DFT, PCA, Haar-Wavelet, etc. stark schwanken würde. Aus diesem Grund wurden die Schwellwerte der Analysen dieses Abschnitts so gesetzt, dass zwischen neun und fünfzehn dimension members als brüchig erkannt werden (Ausnahmen sind die bivariate Kreuzkorrelation und die lineare Regression, wo 24 bzw. 28 Brüche erkannt werden).

Tabellen 5.6 und 5.7 identifizierten Strukturbrüchen die in Tabelle 5.9 dargestellten Elemente auch im Rahmen der DFT als brüchig klassifiziert.

- Laufzeit der DFT auf Strukturdimension SD_1: 0.33 sec.
- Laufzeit auf Strukturdimension SD_2: 0.06 sec.
- Laufzeit auf Strukturdimension SD_4: 0.22 sec.

SD	*dim member*
SD_1	**A443164712**
SD_1	**A546705177**
SD_1	**A633897121**
SD_1	**A768474977**
SD_1	**A776617093**
SD_1	**A841768422**
SD_1	**A855748746**
SD_1	**A864725209**
SD_1	**A913796677**
SD_4	**F378476541**
SD_4	**F971043799**

Legende: *dim member*=dimension member, SD_i=i-te Strukturdimension

Tabelle 5.9: Im Rahmen der DFT erkannte obige Strukturbrüche

Principal Component Analysis

Die PCA wird in ähnlicher Weise wie die DFT verwendet - auch hier wird die Entwicklung der Absolutbeträge der Werte der im vorherigen Schritt als brüchig kategorisierten dimension members mit der Gesamtentwicklung der summierten Absolutbeträge der Werte aller dimension members verglichen[14]. Setzt man die Grenzwerte dieses Verfahrens zur Identifikation von Strukturbrüchen auf 500 000 (euklidische Distanz der Eigenwerte) bzw. 0.5 (normierte euklidische Distanz der Eigenvektoren), so werden von den im ersten Schritt als brüchig erkannten dimension members (Tabellen 5.6 und 5.7) die in Tabelle 5.10 aufgeführten dimension members auch hier wiederum als Ausreißer klassifiziert.

- Laufzeit der PCA auf Strukturdimension SD_1: 1.7 sec.

[14]Von der in Kapitel 3.7.3 vorgestellten Möglichkeit, die Veränderung der Mittelwerte eines dimension members zwischen verschiedenen Perioden zu berechnen, wird hier nicht Gebrauch gemacht. Die euklidischen Distanzen von Eigenwerten und Eigenvektoren werden zudem in diesem Fall auf Basis aller Eigenwerte (nicht nur auf Basis der k größten Eigenwerte) berechnet.

- Laufzeit auf Strukturdimension SD_2: 0.22 sec.
- Laufzeit auf Strukturdimension SD_4: 1.76 sec.

Dim	*dim member*
SD_1	**A175546859**
SD_1	**A502626747**
SD_1	**A776617093**
SD_1	*A*864725209
SD_2	**D160866068**
SD_2	**D380554747**
SD_4	**F139701034**
SD_4	**F196028405**
SD_4	**F378476541**
SD_4	**F535841867**
SD_4	**F657245554**
SD_4	**F701638710**
SD_4	**F851172809**

Legende: *dim member*=dimension member, SD_i=i-te Strukturdimension

Tabelle 5.10: Im Rahmen der PCA erkannte obige Strukturbrüche

Haar-Wavelet-Transformation

Die Methode der Transformation mittels dem Haar-Wavelet identifiziert beim Vergleich der Absolutbeträge der Werte einzelner dimension mebers mit der Gesamtentwicklung der Summe der Absolutbeträge aller dimension members (wie oben PCA und DFT) bei Schwellwerten der maximalen Distanz der Koeffizienten von 2.3 und jener der skalierten summierten Distanz von 0.95 die in Tabelle 5.11 aufgelisteten dimension members als Strukturbrüche.

- Laufzeit der Haar-Wavelet-Transformation auf Strukturdimension SD_1: 0.49 sec.
- Laufzeit auf Strukturdimension SD_2: 0.06 sec.
- Laufzeit auf Strukturdimension SD_4: 0.44 sec.

Die Resultate der DWT als Folge linearer Transformationen mit Daubechies-Filtern sind hier nicht aufgeführt, zum Vergleich nur ihre Laufzeiten (bei MaxThreshold=3.2 und SumThreshold=1.2 werden 9 dimension members als brüchig erkannt):

- Laufzeit der der DWT als Folge linearer Transformationen mit Daubechies-Filtern auf Strukturdimension SD_1: 5.66 sec.

Dim	*dim member*
SD_1	**A378566270**
SD_1	**A443164712**
SD_1	**A546705177**
SD_1	**A633897121**
SD_1	**A768474977**
SD_1	**A776617093**
SD_1	**A841768422**
SD_1	**A855748746**
SD_1	**A864725209**
SD_1	**A913796677**
SD_4	**F196028405**
SD_4	**F378476541**

Legende: *dim member*=dimension member, SD_i=i-te Strukturdimension

Tabelle 5.11: Im Rahmen der Haar-Wavelet-Transformation erkannte obige Strukturbrüche

- Laufzeit auf Strukturdimension SD_2: 0.72 sec.
- Laufzeit auf Strukturdimension SD_4: 4.18 sec.

Autokorrelation

Die Autokorrelationsmethode muss im Vergleich zu Kapitel 3.4 und Kapitel 4.2 leicht angepasst werden - als Schwellwert zur Erkennung eines Strukturbruchs wird nicht wie oben eine konstante Zahl herangezogen, sondern die Differenz zwischen den Werten eines einzelnen dimension members und den Werten des Vektors mit den summierten Beträgen aller dimension members. In Tabelle 5.12 sind jene dimension members aufgeführt, deren Autokorrelationskoeffizienten $\rho(1)$ und $\rho(2)$ um mehr als 0.4 bzw. 0.15 von jenen des Vektors mit den summierten Absolutbeträgen aller Werte differieren[15].

- Laufzeit der Autokorrelation auf Strukturdimension SD_1: 0.06 sec.
- Laufzeit auf Strukturdimension SD_2: 0 sec.
- Laufzeit auf Strukturdimension SD_4: 0.06 sec.

Die Resultate der Autoregression sind hier nicht aufgeführt, zum Vergleich hier ihre Laufzeiten (bei Anwendung eines AR(3)-Modells[16] mit ϕ_1=0.6, ϕ_2=0.6, ϕ_3=-0.2 sowie einem *AIC*-

[15]Dimension members mit mehr als fünf Nullwerten werden aus der Berechnung hier ebenso ausgeklammert wie unten bei der Berechnung der Kreuzkorrelationskoeffizienten.

[16]Da kein Vorwissen über ein Ansteigen bzw. Sinken der Werte im Laufe der Monate vorhanden ist, wird von stationären Daten ausgegangen, sodass keine vorherigen Transformationen durchgeführt werden.

Dim	*dim member*
SD_1	**A135105873**
SD_1	**A378566270**
SD_1	**A546705177**
SD_1	**A776617093**
SD_1	**A841768422**
SD_1	**A855748746**
SD_1	**A971709069**
SD_4	**F139701034**
SD_4	**F196028405**
SD_4	**F239909113**
SD_4	**F429797213**
SD_4	**F657245554**

Legende: *dim member*=dimension member, SD_i=i-te Strukturdimension

Tabelle 5.12: Im Rahmen der Autokorrelation erkannte obige Strukturbrüche

Threshold[17] von 7 werden 15 dimension members als brüchig identifiziert):

- Laufzeit der Autoregression auf Strukturdimension SD_1: 0.11 sec.
- Laufzeit auf Strukturdimension SD_2: 0.06 sec.
- Laufzeit auf Strukturdimension SD_4: 0.05 sec.

Bivariate Kreuzkorrelation

Ähnlich wie die vorangegangenen Methoden wird auch die bivariate Kreuzkorrelation so verwendet, dass die Gleichläufigkeit der Wertentwicklung der oben als brüchig identifizierten dimension members mit der Entwicklung der summierten Absolutbeträge der Werte verglichen wird - somit ist auch die Laufzeit dieser Methode linear in der Anzahl der Eingabedaten. Sieht man einen geringeren Korrelationskoeffizient als 0.15 als Bruch an[18], so werden die in Tabelle 5.13 aufgelisteten dimension members als brüchig erkannt.

- Laufzeit der Kreuzkorrelation im Vergleich zur Gesamtentwicklung auf Strukturdimension SD_1: 0.05 sec.
- Laufzeit auf Strukturdimension SD_2: 0 sec.

[17] Die Rechenzeiten wurden mit der klassischen Berechnungsvariante ermittelt (zur Terminologie vgl. Kapitel 3.5), die Laufzeiten der adaptierten Variante sind jedoch identisch.

[18] Bei dieser Analyse werden zwar deutlich mehr dimension members als in den vorangegangenen Analysen aufgedeckt; der Schwellwert ist mit 0.15 allerdings schon so nahe bei null, dass die erkannten dimension members durchaus als brüchig bezeichnet werden können.

- Laufzeit auf Strukturdimension SD_4: 0 sec.

Dim	*dim member*
SD_1	**A102638053**
SD_1	**A135105873**
SD_1	**A175546859**
SD_1	**A315347857**
SD_1	**A378566270**
SD_1	**A468143681**
SD_1	**A502408343**
SD_1	**A546705177**
SD_1	**A633897121**
SD_1	**A768474977**
SD_1	**A776617093**
SD_1	**A841768422**
SD_1	**A855748746**
SD_1	**A864725209**
SD_1	**A913796677**
SD_2	**D430391667**
SD_4	**F112308097**
SD_4	**F139701034**
SD_4	**F196028405**
SD_4	**F378476541**
SD_4	**F386807789**
SD_4	**F535841867**
SD_4	**F657245554**
SD_4	**F920147523**

Legende: *dim member*=dimension member, SD_i=i-te Strukturdimension

Tabelle 5.13: Im Rahmen der bivariaten Kreuzkorrelation erkannte obige Strukturbrüche

Auch hier soll lediglich ein Laufzeitvergleich mit der linearen Regression (einfachstes Modell Y=A+B*X, wobei X die Entwicklung der summierten Absolutbeträge aller Werte und Y den jeweiligen in obigem Schritt als brüchig identifizierten dimension member repräsentiert; eine Unterschreitung des R^2-Wertes von 0.15 gilt als Strukturbruch, und 28 dimension members werden dabei als Brüche erkannt) angestellt werden, ohne dass die Resultate der linearen Regression aufgeführt sind:

- Laufzeit der linearen Regression auf Strukturdimension SD_1: 0.05 sec.
- Laufzeit auf Strukturdimension SD_2: 0 sec.
- Laufzeit auf Strukturdimension SD_4: 0 sec.

5.3 Gesamtinterpretation der verschiedenen Methoden

Interpretiert man die Ergebnisse der Methoden sehr eng, dann sind nur jene dimension members als Strukturbrüche zu klassifizieren, die in allen Methoden als solche erkannt worden - in diesem Fall bleibt allerdings kein einziger dimension member übrig. Interpretiert man die Ergebnisse hingegen sehr weit, so können alle im Rahmen der Abweichungsmatrix identifizierten Brüche zur weiteren Analyse, ob es sich hierbei um tatsächliche Brüche oder um Scheinbrüche (Eingabefehler, tatsächliche große Schwankungen, ...) handelt, freigegeben werden - in diesem Fall hätte man die weiteren Methoden aber gar nicht anwenden sollen. Diese eindimensionale Interpretation ist jedoch sehr zu hinterfragen: einerseits sind die Schwellwerte der Methoden unterschiedlich 'streng' eingestellt, andererseits sollte auch die Absolutanzahl der aufgedeckten Brüche im Rahmen der Abweichungsmatrix berücksichtigt werden. Eine mögliche Lösung wäre es, die einzelnen Methoden sowie die Häufigkeit der Brüche in der Abweichungsmatrix mit Gewichten zu versehen.

Wie gut das Ergebnis der Methoden auf den vorliegenden Daten ist, sollte von Domain-Experten beurteilt werden. Als eine mögliche Metrik bietet sich die Präzision (*Precision*) p des Systems an, die definiert ist als

$$p = \frac{|F_Q \cap R_Q|}{|F_Q|},$$

wobei $|F_Q|$ die Kardinalität der Menge der vom System gelieferten dimension members und $|R_Q|$ die Anzahl der tatsächlichen Strukturbrüche signalisiert; die Präzision drückt somit aus, wie viele der erkannten Strukturbrüche tatsächlich als solche zu klassifizieren sind [Wb02][19].

In diesem Fall ist die Ermittlung der Präzision jedoch problematisch: die Präzision des Systems kann nicht eruiert werden, da Strukturbrüche in Strukturdimension SD_4 auch von Domain-Experten nicht mehr nachvollzogen werden können. Daher sind bei der Interpretation der Ergebnisse dieses Data Warehouses folgende Punkte zu beachten:

- Die Zubuchungsregeln von Transaktionen zu einzelnen dimension members in Strukturdimension SD_4 können von jedem Mitarbeiter selbständig verändert werden - da diese Änderungen zumeist nicht dokumentiert werden, ist die Sinnhaftigkeit einer Analyse von Daten speziell entlang Strukturdimension SD_4 in Frage zu stellen.

- Sowohl bei Betrachtung der Werte auf der Ebene von Strukturdimensionskombinationen

[19] Die zweite klassische Evaluationskennzahl, der *Recall* r, definiert als

$$r = \frac{|F_Q \cap R_Q|}{|R_Q|},$$

die angibt, welcher Prozentsatz der tatsächlichen Strukturbrüche auch wirklich erkannt wurde, kann im Normalfall nicht berechnet werden, da die Gesamtanzahl der tatsächlichen Brüche nicht bekannt ist.

als auch bei Betrachtung der Werte gruppiert entlang einer Strukturdimension sind starke Schwankungen festzustellen - dies könnte auf natürliche Schwankungen der Werte zurückzuführen sein (z.B. in manchen Monaten werden hohe Einmalinvestitionen getätigt), es könnte aber auch in fehler- bzw. lückenhaften Dateneingaben begründet sein.

- Die stark schwankenden Werte führen dazu, die Schwellwerte zur Identifikation von Strukturbrüchen relativ scharf anzusetzen, um nicht zu viele dimension members als brüchig zu klassifizieren - tatsächliche Strukturbrüche, die möglicherweise etwas weniger krass zum Ausdruck kommen, könnten daher verborgen bleiben; der einzige dimension member, von dem bekannt ist, dass sich seine Struktur im Analysezeitraum verändert hat (dimension member $A729103292$ in Strukturdimension SD_1), wird im Rahmen der Analysen nicht als brüchig erkannt.

Die mangelhafte Datenqualität dieses Data Warehouses lässt wenig Aufschlüsse über die Mächtigkeit des vorgeschlagenen Ansatzes zu: es kann lediglich gefolgert werden, dass alle untersuchten Methoden hinsichtlich der Laufzeit sehr gut abschneiden - mit Ausnahme der Singulärwertzerlegung und der DCT, die etwas mehr als sechs Minuten benötigen, sind alle Verfahren unter sechs Sekunden fertig. Um die Qualität der Ergebnisse der jeweiligen Verfahren beurteilen zu können, wäre die Erprobung des Ansatzes an einem anderen realen Data Warehouse erforderlich.

6

Resümee

In dieser Arbeit wurden Methoden zur Identifikation von Strukturbrüchen vorgestellt, die prinzipiell als sehr leistungsfähig erscheinen: ihre Laufzeitkomplexitäten sind zumeist niedrig (in vielen Fällen linear in Abhängigkeit der Eingabegröße), und sie lassen sich auf n-dimensional referenzierte Daten im Allgemeinen gut erweitern.

Wie die Erprobung der Methoden auf einem realen Data Warehouse gezeigt hat, hängt die Güte der Erkennung von Strukturbrüchen zum einen von der Datenqualität und zum anderen von der natürlichen Volatilität der Werte ab: fehlerhaft bzw. lückenhaft aufgezeichnete Werte sowie starke Schwankungen der Werte können zu hohen Fehlerraten erster und zweiter Ordnung führen. Darüber hinaus ist dem korrekten, situationsspezifischen Tuning der Parameter der verschiedenen Methoden große Bedeutung zu schenken - zu straff eingestellte Parameter ziehen eine Erhöhung der Fehlerrate erster Art nach sich, zu lasch eingestellte Schwellwerte führen zu einer Erhöhung der Fehlerrate zweiter Art.

Anhang A

Quellcode verwendeter Funktionen

```
% computes AIC, BIC and FPE for a data vector
% INPUT: data, coefficients of AR and MA (possibly 0)
% OUTPUT: AIC, BIC, FPE (if q=0)
% At least one of p and q must be ~= 0!!
function f=armacriteria(xvec,phivec,thetavec)
plen=length(phivec);
qlen=length(thetavec);
N=length(xvec);
xdach=zeros(1,N-plen);
% compute iid random numbers mean 0 and var 1
uivec=zeros(1,N);
for i=1:N
   uivec(1,i)=(rand-0.5)*2;
end
% prediction of xt starts at value rlen+1
rlen=max(plen,qlen);
if (rlen~=0)
  % predict estimates for xj based on AR
  if (plen ~= 0)
    for j=(rlen+1):N
       for k=1:plen
          xdach(1,j-rlen)=xdach(1,j-rlen)+phivec(1,k)*xvec(1,j-k);
       end
    end
  end
  % predict estimates for xj based on MA
  if (qlen ~= 0)
    for j=(rlen+1):N
       for k=1:qlen
          xdach(1,j-rlen)=xdach(1,j-rlen)+thetavec(1,k)*uivec(1,j-k);
       end
    end
  end
  % add ut for each xt-estimator
  for j=(rlen+1):N
     xdach(1,j-rlen)=xdach(1,j-rlen)+uivec(1,j);
  end
end   %if rlen(~= 0)
% calc delta(xvec,xdach)
sigmahoch2=0;
for j=(rlen+1):N
   sigmahoch2=sigmahoch2+(xdach(1,j-rlen)-xvec(1,j))^2;
end
sigmahoch2=sigmahoch2/(N-rlen);  % scaling by length
% 2nd variant: include those 2 lines
%sigmahoch2=sqrt(sigmahoch2); % take root so that scaling by avg size makes sense
%sigmahoch2=sigmahoch2/mean(xvec); % scaling by avg size of values
% compute AIC, BIC and FPE with natural logarithm
aic=log(sigmahoch2)+((2*(plen+qlen))/N);
bic=log(sigmahoch2)+((plen+qlen)/(N*log(N)));
fpe=NaN; % if q>0 then fpe is not computed
if (qlen == 0) % then also compute fpe
   fpe=sigmahoch2*((N+plen)/(N-plen));
end
f=[aic,bic,fpe];
```

Abbildung A.1: Berechnung von Kennzahlen im Rahmen der Autoregression

```
% distance function for matrices
% computes scaled distances in eigen(singular-)values & -vectors
% result=2*1 vector, 1st val=valuedist, 2ndval=vectordist
% comparison of eigenvalues and eigenvectors in my eyes makes sense since
% we compare a data vector (matrix) with itself (just slided ahead one period)
% => main directions of vectors and order of directions should remain the same
% same reason holds for comparing maxdiff and sumdiff in DFT and DWT
function f=value2vectordist(valmat1, valmat2,vecmat1, vecmat2, mean1, mean2)
N=size(valmat1);
% scaling of eigenvalues with mean of vector (matrix)
% diff in eigenvalues
valdistance=0;

%take min of both dimensions of matrix (if nonsquare singular value matrices)
dim=min(N(1),N(2));
for i=1:dim
  valmat1(i,i)=valmat1(i,i)/mean1;
  valmat2(i,i)=valmat2(i,i)/mean2;
end
for i=1:dim
  dimdistance=(valmat1(i,i)-valmat2(i,i))^2;
  valdistance=valdistance+dimdistance;
end
% output 1: root of valdistance/length
%valdistance=sqrt(valdistance)/(dim*avgval);
valdistance=sqrt(valdistance)/dim;

% diff in eigenvectors
vecdistance=0;
M=size(vecmat1);  %M(1)=#rows, M(2)=#cols
for j=1:M(2)
  % eigenvectors have to be normed
  vecmat1(:,j)=vecmat1(:,j)/(norm(vecmat1(:,j)));
  vecmat2(:,j)=vecmat2(:,j)/(norm(vecmat2(:,j)));
  for i=1:M(1)
    dimdistance=(vecmat1(i,j)-vecmat2(i,j))^2;
    vecdistance=vecdistance+dimdistance;
  end
end

% output 2: root of vecdistance/length
vecdistance=sqrt(vecdistance)/M(2);
result=zeros(2,1);
result(1,1)=valdistance;
result(2,1)=vecdistance;
f=result;
```

Abbildung A.2: 'Euklidische' Distanzfunktion der Eigenwerte und Eigenvektoren (bzw. Singulärwerte und Singulärvektoren) zweier Matrizen

```
% INPUT: two rectangular matrices of same size
% OUTPUT: normed difference of singularvalues and eigenvectors
%
function f=svddistance(matrix1,matrix2)
% singular value decomposition
% [U, S, V]=svd(M)
% cols of matrix U are eigenvectors of MM'
% values on maindiagonal of S are singularvalues of M
[u1, s1, v1]=svd(matrix1);
[u2,s2,v2]=svd(matrix2);
% no error checks here, assumed that matrix sizes are the same
% get the avgval of both matrices
avg1=mean(mean(matrix1));
avg2=mean(mean(matrix2));
valvectdist=value2vectordist(s1,s2,u1,u2,avg1,avg2);
f=valvectdist;
```

Abbildung A.3: Klassische Singulärwertzerlegung

```
% PCA for Data Mining
% INPUT: 2 (consecutive) vectors of values over time
% OUTPUT: difference in eigenvectors, eigenvalues
% OUTPUT also difference in mean (if wanted) => omit 1 line below (see comment)
%
function f=pcaanalysis(vector1,vector2)
% calc old and new mean
% assumed that sizes are equal
N=length(vector1);
mean1=mean(vector1);
mean2=mean(vector2);
% calc covmatrices
covmatrix1=zeros(N,N);
covmatrix2=zeros(N,N);
for i=1:N
  for j=1:N
    covmatrix1(i,j)=(vector1(i)-mean1)*(vector1(j)-mean1);
    covmatrix2(i,j)=(vector2(i)-mean2)*(vector2(j)-mean2);
  end
end
% calc eigenvectors & -values
% in cols of V are eigenvectors, in D eigenvalues
[V1, D1]=eig(covmatrix1);
[V2,D2]=eig(covmatrix2);
%diff in means in relation to bigger of the 2 means
biggermean=0;
if (mean1 > mean2)
  biggermean=mean1;
else
  biggermean=mean2;
end
meandiff=sqrt((mean1-mean2)^2);

%omit semicolon at end if you want to use mean-difference
meandiff=meandiff/(biggermean*N);
eigvalvectdist=value2vectordist(D1,D2,V1,V2,mean1,mean2);
f=eigvalvectdist;
```

Abbildung A.4: Principal Component Analysis (PCA)

```
%distance function used by DFT and Wavelet-Transforms
function f=generaldistance2(vector1, vector2, coeff1, coeff2)
mean1=mean(vector1);
mean2=mean(vector2);
coeff1=coeff1/mean1;
coeff2=coeff2/mean2;
sumdist=0;
maxdist=0;
N=length(vector1);
for i=1:N
  currdist=abs(coeff1(1,i)-coeff2(1,i));
  if (currdist > maxdist)
    maxdist=currdist; end
  sumdist=sumdist+currdist;
end
sumdist=sumdist/N;
result=zeros(2,1);
result(1,1)=maxdist;
result(2,1)=sumdist;
f=result;
```

Abbildung A.5: Generische Distanzermittlungsfunktion für Koeffizienten unterschiedlicher Transformationen

```
% Fourier-Transform for Data Mining
% INPUT: 2 (consecutive) vectors of values over time
% OUTPUT: difference in max-amplitude and summed difference
%
function f=fourieranalysis(vector1,vector2)

% Let Matlab compute fft
% length of vectors should be power of two - then FFT O(n log n)
% else rather O(n³)
fouriercoeffs1=fft(vector1);
fouriercoeffs2=fft(vector2);
N=length(vector1);

% scaling for calculation of amplitudes is omitted since no change in order
% fouriercoeffs therefore not divided by N (both vectors should always have same length)

% if N odd => floor is used
numfreqs=floor(N/2);
% 2nd half of freq1 and freq2 remains zero - just for calling distance f.
freq1=zeros(1,N);
freq2=zeros(1,N);
% the prime of a number is its complex conjugate
for i=1:numfreqs
  freq1(1,i)=2*sqrt(fouriercoeffs1(1,i)*fouriercoeffs1(1,i)');
  freq2(1,i)=2*sqrt(fouriercoeffs2(1,i)*fouriercoeffs2(1,i)');
end
maxsumdist=generaldistance2(vector1,vector2,freq1,freq2);
f=maxsumdist;
```

Abbildung A.6: Diskrete Fourier-Transformation

```
% computes DCT of 2 matrices and differences between 'em
% INPUT: 2 matrices of same size (2 consecutive time periods)
% INPUT: num rows and num cols of coefficients matrices
% OUTPUT:  differences between coefficients
% in this case: dct-coefficients unscaled by mean
function f=dctcompare(matrix1, matrix2, numrows, numcols)
% no size checks, assumed that they are the same
rowlength=size(matrix1,1);
collength=size(matrix1,2);
mat1mean=mean(mean(matrix1));
mat2mean=mean(mean(matrix2));
dct1=0; dct2=0;
% initialize DCTcoefficient-matrices
DCTcoeff1=zeros(numrows,numcols);
DCTcoeff2=zeros(numrows,numcols);
% perform DCT-transform
for i=1:numrows
  for j=1:numcols
    for k=1: rowlength
      for l= 1: collength
        dct1=4*matrix1(k,l)*cos(((pi*i)/(2*rowlength))*(2*k+1))*cos(((pi*j)/(2*collength))*(2*l+1));
        DCTcoeff1(i,j)=DCTcoeff1(i,j)+dct1;
        dct2=4*matrix2(k,l)*cos(((pi*i)/(2*rowlength))*(2*k+1))*cos(((pi*j)/(2*collength))*(2*l+1));
        DCTcoeff2(i,j)=DCTcoeff2(i,j)+dct2;
      end
    end
  end
end

% scale all dct-coefficients by mean - in some cases better OMITTED
%for i=1:numrows
 %  for j=1:numcols
  %    DCTcoeff1( i,j)=DCTcoeff1(i,j)/mat1mean;
   %   DCTcoeff2(i,j)=DCTcoeff2(i,j)/mat2mean;
   %end
%end

% measure Euclidean distance
sumdist=0; currdist=0;
for i=1:numrows
  for j=1:numcols
    currdist=(DCTcoeff1(i,j)-DCTcoeff2(i,j))^2;
    sumdist=sumdist+currdist;
  end
end
% scale by num of rows and cols of resultmat
f=sqrt(sumdist);
```

Abbildung A.7: Distanzermittlung mit der DCT

```
% HAAR-WAVELET FOR DATA MINING
% Input: data vectors,their length must be a power of 2
% equal lengths of both vectors assumed
% Output: max-coefficientdifference and overall-distance
function f=wavelethaar(vector1, vector2)
% scaling factor of Haar wavelet
scaling=sqrt(2);
tempcoeffs1=vector1;
tempcoeffs2=vector2;
N=length(vector1);
currentcoeffs1=zeros(1,N);
currentcoeffs2=zeros(1,N);
currentN=N;

% one loop is one expansion step in current vector space
% expand input-coefficients until V(0) is reached
while (currentN > 1)
 i=1;
coeffpos=1;
 % taking averages => low-pass filtering
 while (i < currentN)
  currentcoeffs1(1, coeffpos)=(tempcoeffs1(1,i)+tempcoeffs1(1,i+1))/scaling;
  currentcoeffs2(1, coeffpos)=(tempcoeffs2(1,i)+tempcoeffs2(1,i+1))/scaling;
   i=i+2;
  coeffpos=coeffpos+1;
 end
i=1;
 % taking differences => high-pass filtering
while (i < currentN)
  currentcoeffs1(1, coeffpos)=(tempcoeffs1(1,i)-tempcoeffs1(1, i+1))/scaling;
  currentcoeffs2(1, coeffpos)=(tempcoeffs2(1,i)-tempcoeffs2(1, i+1))/scaling;
  i=i+2;
  coeffpos=coeffpos+1;
end

% expanded coefficients are input for next expansion step
tempcoeffs1=currentcoeffs1;
tempcoeffs2=currentcoeffs2;
% high frequences (differenced values) remain => half N
currentN=currentN/2;
end
maxsumdist=generaldistance2(vector1, vector2, tempcoeffs1, tempcoeffs2);
f=maxsumdist;
```

Abbildung A.8: Diskrete Wavelet-Transformation mit dem Haar-Wavelet

```
% DAUBECHIES-WAVELET-FILTER (N=2)
% DWT with LINEAR TRANSFORMATIONS
% INPUT: data vector, their length must be a power of 2
% OUTPUT:  wavelet-coefficients
function f=waveletdaubechies(vector)
N=length(vector);
waveletcoefficients=zeros(N,1);
% FILTER COEFFICIENTS FOR transf. matrices
h0=0.4829629131445342;
h1=0.836516303737808;
h2=0.2241438680420134;
h3=-0.1294095225512604;
numrows=N/2;
numcols=N;
Hmat=zeros(numrows, numcols); Gmat=zeros(numrows, numcols);
tempvec=zeros(N,1);
tempvec=vector';
currN=N;
% the main loop : 4 filters => 4 cols needed at least
while (currN >= 4)  % we have 4 filters
   pos=0;
   for i=1:(currN/2)
      % position of h0..h3
      pos=mod(pos,currN);
      pos0=pos; pos1=pos+1; pos2=pos+2; pos3=pos+3;
      % make modulo corrections
         if (pos==0)
          pos0=currN;  end
         if ((mod(pos+2,currN))==0)
          pos3=1; end
         if ((mod(pos+1,currN))==0)
          pos2=1; pos3=2; end
      % do positioning
      Hmat(i, pos0)=h0; Gmat(i,pos0)=h3;
      Hmat(i,pos1)=h1;  Gmat(i,pos1)=-h2;
      Hmat(i,pos2)=h2;  Gmat(i,pos2)=h1;
      Hmat(i,pos3)=h3;  Gmat(i,pos3)=-h0;
      pos=pos+2; %in next row pos is right-shifted by 2
   end      %for
   tmpmat=Hmat*tempvec;
   tmpmat2=Gmat*tempvec;
   tempvec=zeros(currN/2, 1);
   for i=1:(currN/2)
     waveletcoefficients(i,1)=tmpmat(i,1);
      tempvec(i,1)=waveletcoefficients(i,1);
   end
   for i=(currN/2 +1) : currN
     waveletcoefficients(i,1)=tmpmat2(i-currN/2,1);
   end
   % half matrix dimensions at every step
   currN=currN/2;
   Hmat=zeros(currN/2, currN);  Gmat=zeros(currN/2, currN);
end   %while
f=waveletcoefficients;
```

Abbildung A.9: Diskrete Wavelet-Transformation mittels linearer Transformationen unter Anwendung von Daubechies-Filtern für $N = 2$

```
% compares 2 vectors by Daubechies-Wavelet
% output: max and overall difference
%
function f=daubechiescompare(vector1, vector2)
coeffs1=waveletdaubechies(vector1);
coeffs2=waveletdaubechies(vector2);
% coeffs have to be transposed
maxsumdist=generaldistance2(vector1,vector2,coeffs1',coeffs2');
f=maxsumdist;
```

Abbildung A.10: Ermittlung der Distanzen der Koeffizienten der DWT mit Daubechies-Filtern für $N = 2$

```
%  random generator with noise
% computes normally distributed numbers for
%  given matrix (allocation in advance saves time)
% return random matrix
function f=randomgenerator(matrix)
nsq=size(matrix,1);
timechronons=size(matrix,2);
for i=1:nsq
for j=1:timechronons
matrix(i,j)=(randn)*10+200; %~170-230
end
end
% Include some noise in time periods 8-10
for i=25:30
for j=8:10
 matrix(i,j)=i+j;
end
end
f=matrix;
```

Abbildung A.11: Generierung von Zufallszahlen mit 'Fehlern'

```
%compare DFT-coefficients of dimension members with coefficients of sum of all members
%sumvec could also be a specified certainly unchanged dimension member
% Input rowindices is (double-cleaned) result of previous step,
%Input matrix contains only these data, but no information about the indices
%parametrize: MaxThreshold, SumThreshold
function f=all2fourierdifferences(matrix,sumvec, rowindices)
tic;
M=size(matrix,1);
N=size(matrix,2);
MaxThreshold=11;
SumThreshold=2.3;
numberbreaks=0;
resultmat=zeros(1,0);
for i=1:M
  intervsize=8; % length of vector (size of interval)
  dist=0; % dist from t=1
  %diff=1; % time difference between consecutive vectors
  while  ( ( dist+intervsize) <= N)
          zeroval=0;
  for j=1:intervsize
    if (matrix(i,j+dist)==0) zeroval=1; end
  end
  if (zeroval==0)
    % use absolute value of dimension members - in some cases perhaps better omitted
    result=fourieranalysis(abs(matrix(i,(1+dist):(intervsize+dist))),sumvec(1,(1+dist):(intervsize+dist)));
    if ((result(1,1) > MaxThreshold) &(result(2,1) > SumThreshold))
      numberbreaks=numberbreaks+1;
      resultmat(numberbreaks,1)=rowindices(i,1);
    end  %if
  end %zeroval
  dist=dist+1;
  end %while
end %for i
toc
f=resultmat;
```

Abbildung A.12: Aufruf der Fourier-Transformation mit bestimmten Schwellwerten für die Ausgabe auffälliger Zeilen

Literaturverzeichnis

[At89] Atkinson, K.E., An Introduction to Numerical Analysis, 2nd edition, John Wiley, New York 1989

[BD02] Brockwell, P.J., Davis, R.A., Introduction to Time Series and Forecasting, Springer Verlag, New York 2002

[DC02] The Discrete Cosine Transform, Download `http://www.ece.purdue.edu/~ace/jpeg-tut/jpgdct1.html`, August 2002

[Di00] Diacu, F., An Introduction to Differential Equations - Order and Chaos, W. H. Freeman, New York 2000

[EK02] Eder, J., Koncilia, C., Morzy, T., The COMET Metamodel for Temporal Data Warehouses, Proceedings of the 14th International Conference on Advanced Information Systems Engineering, CAISE 2002, Toronto

[Gu02] Gutknecht, M., Lineare Algebra für Informatiker, Vorlesungsskript, ETH Zürich 2002

[Ha93] Harvey, C.A., Time Series Models, Harvester Wheatsheaf, New York 1993

[Ho02] Hollmen, J., Principal Component Analysis, Download `http://www.cis.hut.fi/~jhollmen/dippa/node30.html`, August 2002

[Je98] Jensen, C.S. (Hrsg.), Dyreson, C. E. (Hrsg.), A consensus Glossary of Temporal Database Concepts - Feb. 1998 Version, Springer-Verlag, Berlin 1998

[Ka02] Karypis, G., Data Mining, in IEEE: Computing in Science and Engineering, July-August 2002

[Ki97] Kimball, R., Features for Query Tools, Download `http://www.dbmsmag.com/9702d05.html`, Jänner 2003

[KP99] Kröpfl, B., Peschek, W., Schneider, E., Schönlieb, A., Angewandte Statistik - eine Einführung für Wirtschaftswissenschaftler, Carl Hanser Verlag, München, Wien 1999

[Os90] Ostrom, C.W., Time Series Analysis, Regression Techniques, 2nd edition, Sage Publications, Beverly Hills, London 1990

[Un00] United Nations 2000, International Financial Statistics, Download `http://www.un.org/esa/analysis/annextab.pdf`, August 2002

[Vi99] Vidakovic, B., Statistical Modeling by Wavelets, John Wiley, New York 1999

[Wb02] Weber, R., Multimedia Retrieval, Vorlesungsskript, ETH Zürich 2002

[We85] Weisberg, S., Applied Linear Regression, John Wiley, New York 1985

[Wü02] Würtz, D., ARMA Modelling: Basic Concepts of Linear Processes, in: S-PLUS for Financial Engineers, Vorlesungsskript zu 'Theoretische Physik', ETH Zürich 2002

[Yp02] Yonghong, P., Wavelets for Time-Series Data Mining, Download `http://cs.bris.ac.uk/~peng/`, August 2002

[Ze02] Zehnder, C. A., Gestaltung grosser Informationssysteme, Vorlesungsskript, ETH Zürich 2002