



Zeitreihenanalyse

Eine Einführung für Sozialwissenschaftler
und Historiker

Von

Prof. Dr. Helmut Thome

Martin-Luther-Universität Halle-Wittenberg

R. Oldenbourg Verlag München Wien

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

© 2005 Oldenbourg Wissenschaftsverlag GmbH
Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
www.oldenbourg.de

Das Werk einschließlich aller Abbildungen ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: Druckhaus „Thomas Müntzer“ GmbH, Bad Langensalza

ISBN 3-486-57871-5

Vorwort

Das vorliegende Lehrbuch hat eine lange Entstehungsgeschichte. Seine didaktische Konzeption hat sich seit Mitte der achtziger Jahre in Lehrveranstaltungen an verschiedenen Orten mit unterschiedlichen Teilnehmerkreisen entwickelt: an Instituten der Technischen und der Freien Universität Berlin sowie der Martin-Luther-Universität in Halle, einem "Frühjahrsseminar" des Zentralarchivs für empirische Sozialforschung an der Universität zu Köln, mehreren "Herbstseminaren" am Zentrum für historische Sozialforschung (ebenfalls Köln), einem Sommerkurs im Rahmen der *Essex Summer School in Social Science Data Analysis and Collection* an der University of Essex, England.

Ein erstes Vorlesungsskript entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Zentralarchiv für empirische Sozialforschung/Zentrum für historische Sozialforschung in Köln. Die dort gegebenen Arbeitsbedingungen und Kooperationsmöglichkeiten haben das Vorhaben erheblich gefördert; dafür möchte ich dem Geschäftsführer des Instituts, Ekkehard Mochmann, und den dortigen Kolleginnen und Kollegen herzlich danken, vor allem Priv.-Doz. Dr. Rainer Metz für einen anregenden und bis heute fortgesetzten Gedankenaustausch über verschiedene Varianten und Anwendungsmöglichkeiten zeitreihenanalytischer Methoden. Mein besonderer Dank gilt meinem früheren Mitarbeiter, Dr. Thomas Rahlf (heute Bonn), ohne dessen praktische Unterstützung und kritische Begleitung dieses Buch kaum zustande gekommen wäre. Zu den Kapitelabschnitten 2.4 und 3.4.4 hat er eigenständige Textentwürfe ausgearbeitet; weitere Teile des Manuskripts haben von seinen Korrektur- und Verbesserungsvorschlägen an früheren Fassungen profitiert. - Danken möchte ich auch meinen heutigen Mitarbeitern Christoph Birkel und Matthias Henze, die mir in der Endphase des Projekts mit der kritischen Durchsicht des gesamten Manuskripts sehr geholfen haben. Für technische Unterstützung bei der Anfertigung zahlreicher Abbildungen danke ich außerdem Tobias Jaeck. Ein herzliches Dankeschön geht auch an Dorothea Gries, die mit nicht versiegender Geduld immer wieder neu überarbeitete Manuskriptfassungen erstellt hat.

Schließlich möchte ich im voraus schon allen Leserinnen und Lesern danken, die sich der Mühe unterziehen, mich auf Fehler hinzuweisen und sonstige Verbesserungen vorzuschlagen. Sie können dazu die E-Mail-Adresse helmut.thome@soziologie.uni-halle.de benutzen.

Die Konzeption dieses Lehrbuchs wird im ersten Kapitel erläutert. Die in den Beispielanalysen verwendeten Daten sind größtenteils *online* über folgende Adresse verfügbar: www.soziologie.uni-halle.de/thome/Lehrbuch-Zeitreihenanalyse/. Außerdem kann über diese Adresse auf ein einfach zu handhabendes Analyseprogramm (TISPA) zugegriffen werden, das mir sein Autor, Karl Schableger (Linz), dankenswerterweise für diesen Zweck zugänglich gemacht hat. Es enthält Simulations- und Analyseroutinen, die bei praktischen Anwendungen sehr hilfreich sein können, aber in manchen der gängigen Programmpakete (wie SPSS) nicht angeboten werden. In den entsprechenden Kapitelabschnitten wird verschiedentlich darauf hingewiesen.

Zum Schluß noch eine Bemerkung zu Orthographie und Grammatik: Angesichts des Wirrwarrs, den eine laufende, aber offenkundig nicht abschließbare Debatte zur Rechtschreibreform geschaffen hat, habe ich mich entschieden, vorläufig an den "alten" Rechtschreibregeln festzuhalten. Auf die weitgehend aufgehobene Trennung von grammatikalischem Genus und natürlichem Geschlecht reagiere ich dagegen mit einer gewissen Unentschiedenheit. Ich benutze gelegentlich die wenig ansprechende technische Lösungsform (schreibe z. B. „Soziolog/nnen“), setze in anderen Fällen Femininum und Maskulinum nebeneinander oder verwende nur eine der beiden Formen. Leserinnen und Leser, denen diese Wankelmütigkeit nicht gefällt, bitte ich um Nachsicht.

Halle

Helmut Thome

Inhaltsverzeichnis

1.	Einleitung	
1.1	Zur Konzeption dieses Lehrbuches	1
1.2	Zum heuristischen Nutzen der Zeitreihenanalyse	7
2.	Beschreibende Analyse von Zeitreihen	14
2.1	Empirische Kennzahlen: Mittelwert, Streuung, Autokorrelation	14
2.2	Charakterisierung des Zusammenhangs zweier Zeitreihen: die Kreuzkorrelationsfunktion	19
2.3	Komponentenzerlegung	23
2.3.1	Allgemeines	23
2.3.2	Konzepterläuterung: Trend, Zyklus, Saison und irreguläre Komponente	24
2.3.3	Additive und multiplikative Komponentenmodelle	26
2.3.3.1	Box/Cox- Transformationen	28
2.3.4	Bestimmung der Trendkomponente	33
2.3.4.1	Globale Modelle	33
2.3.4.2	Lokale Trendmodelle: Gleitmittelverfahren	37
2.3.4.2.1	Ein Analysebeispiel (Arbeitslosenzahlen)	39
2.3.5	Bestimmung der Saisonkomponente mit Hilfe des Phasendurchschnittsverfahren	41
2.3.6	Bestimmung der Restkomponente und Rücktransformation der Daten	45
2.3.7	Das Gleitmittelverfahren als sukzessive Regressionsanalyse	49
2.3.8	Gleitmittelverfahren und Filterbegriff	50
2.3.9	Differenzenbildung als Methode der Trendbereinigung	52
2.4	Analyse deterministischer Zyklen: Regressionsansatz und Periodogrammanalyse	55
3.	Statistische Analyse einzelner Zeitreihen: Univariate Box/Jenkins-Modelle	
3.1	Einleitende Bemerkungen	69
3.2	Zum Begriff stochastischer Prozesse	71
3.3	„Stationarität“ als allgemeine Modellvoraussetzung	78
3.4	Autoregressive Prozesse (AR-Modelle)	80
3.4.1	AR-Modelle 1. Ordnung	80
3.4.2	AR-Modelle 2. Ordnung	87
3.4.3	AR-Modelle beliebiger Ordnung	89
3.4.4	Autoregressive Modelle zur Darstellung quasi-zyklischer Prozesse	92
3.5	Die partielle Autokorrelationsfunktion als Instrument der Modellidentifikation	99
3.6	Modellidentifikation in der Praxis (zwei Beispiele)	103

3.7	Moving-Average-Prozesse (MA-Modelle)	110
3.8	Mischprozesse: ARMA-Modelle	118
3.9	Zusätzliche Instrumente der Modellidentifikation	121
3.9.1	Die erweiterte Autokorrelationsfunktion	121
3.9.2	Die Vektor-Autokorrelation	126
3.10	Hinweise zur Schätzung der Modellparameter	131
3.11	Modellevaluierung („Diagnose“)	134
3.12	ARIMA-Modelle für nicht-stationäre Zeitreihen	142
3.12.1	Einfache Random-Walk-Prozesse	143
3.12.2	Random-Walk-Prozesse mit Drift	146
3.12.3	Allgemeine ARIMA-Modelle	148
3.12.4	Einheitswurzel-(Unit-Root)-Tests	155
3.13	Einbau saisonaler Komponenten: SARIMA-Modelle	162
4.	Interventionsanalyse nach Box und Tiao	170
4.1	Ereignis- und Wirkungstypen	171
4.2	Modelle für Ereigniswirkungen	176
4.3	Identifikation des Noise-Modells	182
4.4	Zwei Beispielanalysen	183
4.5	Interventionsanalyse mit nicht-stationären Zeitreihen	187
4.6	Zur Verwendung von Designmatrizen	197
4.7	Interventionen als Quasi-Experimente	202
5.	Dynamische Regressionanalyse	205
5.1	Einleitung	205
5.1.1	Das Problem autokorrelierter Störgrößen	206
5.1.2	Statische versus dynamische Analyseformen	211
5.2	Dynamische Regressionsmodelle 1. Ordnung	217
5.2.1	Die allgemeine Modellstruktur	217
5.2.2	Substanztheoretische Begründungen für geometrische Lag-Modelle	224
5.2.2.1	Das Modell korrigierter Erwartungen	224
5.2.2.2	Das Modell partieller Anpassung	225
5.2.3	Geometrische Lag-Modelle mit zwei Input-Variablen	226
5.3	Autokorrelatives Fehlermodell vs. dynamische Spezifikation des strukturellen Zusammenhangs der Variablen	230
5.4	Hinweise zu Test- und Schätzverfahren	233
5.5	Box/Jenkins-Transferfunktionsmodell	238
5.5.1	Die allgemeine Form der Transferfunktionsmodelle	238
5.5.2	Identifikation von Transferfunktionsmodellen	245
5.5.2.1	Prewhitening	253
5.5.2.2	Test auf eventuelle Rückkopplungseffekte	255
5.5.2.3	Ein Beispiel	255

5.5.3	Die LTF-Methode	261
5.5.4	Evaluierung von Transferfunktionsmodellen	267
5.6	Kointegrierte Prozesse und Fehlerkorrekturmodelle	271
5.6.1	Das Konzept der Kointegration	272
5.6.2	Fehlerkorrekturmodelle	276
5.6.3	Das Problem der Scheinregression	279
6.	Der Prognoseansatz von Box und Jenkins	282
6.1	Univariate Prognose	282
6.1.1	Prognose mit AR(1)-Modellen	284
6.1.2	Prognose mit MA-Modellen	287
6.1.3	Prognose mit ARMA(p,q)- Modellen	288
6.1.4	Prognose mit ARIMA-Modellen	290
6.1.5	Saisonale Modelle	294
6.1.6	Schätzung von Konfidenzintervallen	295
6.1.7	Ein Beispiel	299
6.2	Prognose mit Transferfunktionsmodellen	303
7.	Multivariate Analyse: Vektor-ARMA-Modelle	307
7.1	Grundelemente der VARMA-Modelle	307
7.2	Instrumente der Modellidentifikation	313
7.3	Zur Schätzung von Vektor-Modellen	318
7.4	Modellevaluation	320
7.5	Ein weiteres Anwendungsbeispiel aus der Literatur	321
 Anhang		
A1	Regressionsansatz zur Bestimmung der Gewichte für gleitende Durchschnitte	324
A2	Verschiebeoperator und Differenzenfilter	326
A3	Darstellung deterministischer Zyklen mit Hilfe trigonometrischer Funktionen	330
A4	Autoregressive Prozesse und Differenzengleichungen	336
A5	Zur Bestimmung der Impulsantwortfunktion mittels Koeffizientenvergleich	349
A6	5% Signifikanzpunkte für Durbin-Watson-Test	352
A7	Hinweise zu Software-Paketen	354
Verzeichnis der Abkürzungen		357
Verzeichnis der Abbildungen		359
Verzeichnis der Tabellen		364
Literatur		366
Sachregister		377

Kapitel 1

Einleitung

1.1 Zur Konzeption dieses Lehrbuches

„Zeitreihenanalyse“ ist ein Sammelbegriff für eine Gruppe von Modellen und Verfahren, die eingesetzt werden, um einen bestimmten Typus von Daten zu analysieren: eine Serie von zeitlich geordneten Meßergebnissen, die aus Erhebungen stammen, die relativ häufig (mindestens dreißig- oder vierzigmal)¹ in gleichbleibendem Abstand am gleichen „Objekt“ zur gleichen Merkmalsdimension vorgenommen worden sind. Formal läßt sich eine Zeitreihe als eine geordnete Folge $Z_t, t \in T = \{1, 2, \dots, n\}$ numerischer Größen definieren, die als Ausprägungen einer bestimmten Variablen beobachtet worden sind. Dabei ist von einer endlichen, diskreten Menge von äquidistanten² Meßzeitpunkten auszugehen, die mit $t = 1, 2, \dots, n$ durchnummeriert sind. Die einzelnen „Fälle“ einer Zeitreihe sind nicht wie in der Querschnittanalyse durch die verschiedenen Erhebungseinheiten (z. B. Personen), sondern durch die Meßzeitpunkte konstituiert.

Wichtige Quellen für die Konstruktion von Zeitreihen sind (a) die amtliche Statistik (siehe die Beispiele in Abb. 1.1.1 und 1.1.2), (b) die Inhalts- bzw. Dokumentenanalyse (siehe Abb. 1.1.3), (c) kumulierte Umfragedaten (siehe Abb. 1.1.4).

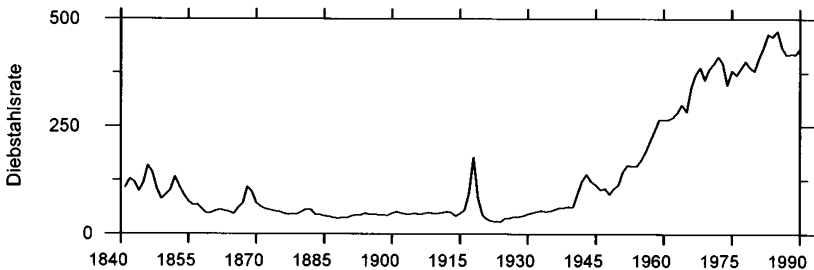


Abb. 1.1.1 Rate der wegen Diebstahl Verurteilten (pro 100.000 Einwohner) in Schweden, 1841 bis 1990 (Datenquelle: Statistics Sweden, s. von Hofer 1991)

Zeitreihenanalyse ist somit eine spezielle Form der Längsschnittanalyse. Es interessieren die kurz- und langfristigen Veränderungen im Niveau und in der Schwankungsbreite bestimmter Phänomene, die sich dabei abzeichnenden Strukturbrüche und Kontinuitäten, das Gewicht spezifischer Einflußgrößen und der dynamische Zusammenhang zwischen Zeitreihen verschiedener Indikatoren.

¹ Die Mindestgröße hängt einerseits von bestimmten Strukturmerkmalen der Daten, andererseits von formalen Eigenschaften der spezifischen Modelle ab, die bei der Analyse eingesetzt werden. Auch hundert und mehr Meßzeitpunkte können nötig sein, um die Parameter bestimmter Modelle sinnvoll schätzen zu können.

² Zeitreihenanalytische Verfahren für nicht gleichabständige Meßzeitpunkte werden in Parzen (1984) behandelt.

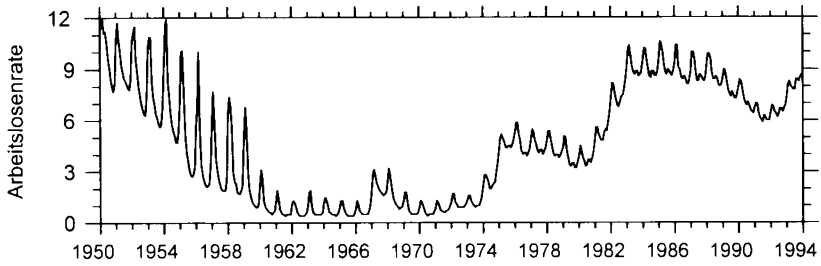


Abb. 1.1.2 Arbeitslosenrate in der Bundesrepublik Deutschland, Jan. 1950 bis Dez. 1993 (Datenquelle: Zentralarchiv für empirische Sozialforschung Köln, ZA Nr. 0800)

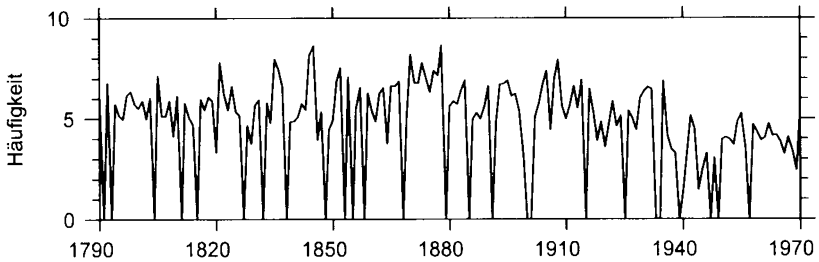


Abb. 1.1.3 Häufigkeit der Kategorie „Enlightenment Total“ in den britischen Thronreden, 1790 bis 1970 (Datenquelle: R. Weber, s. Namenwirth/Weber 1987)

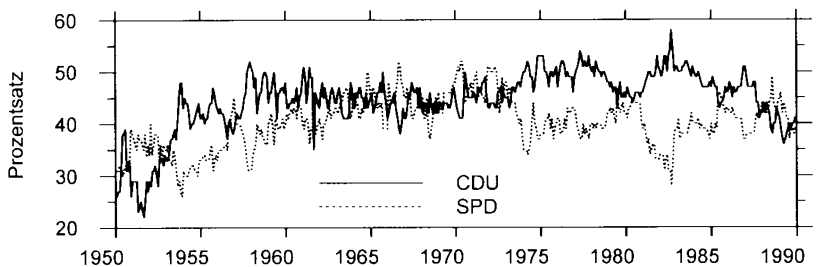


Abb. 1.1.4 Prozentanteile potentieller CDU- und SPD-Wähler, Jan. 1950 bis Dez. 1989 (Datenquelle: Zentralarchiv für empirische Sozialforschung Köln, ZA Nr. 0800)

Andere wichtige Formen sind die *Panelanalyse* und die *Verlaufs- oder Ereignisdatenanalyse (Event History Analysis)*. Bei der Panelanalyse werden Daten untersucht, die bei einer Vielzahl von Objekten wiederholt (aber nicht unbedingt in gleichen zeitlichen Abständen) erhoben worden sind. Solche Erhebungen werden bspw. in der Wahlsoziologie durchgeführt, um festzustellen, wie sich *individuelle* Einstellungen, nicht nur die durchschnittliche Orientierung einer Bevölkerung (bzw. Bevölkerungsgruppe) geändert haben (siehe z. B. Kühnel 1993). Das wahrscheinlich bekannteste „Panel“ in der Bundesrepublik ist das „sozio-ökonomische Panel“ (s. Hujer/Rendtel/Wagner 1997). Eine Kombination von Längsschnitt- und

Querschnittvariation liegt auch dem Design der *pooled time-series-cross-section analysis* (auch kürzer: *pooled time-series analysis*) zugrunde, wobei die terminologische Abgrenzung zur Panel-Analyse in der Literatur nicht einheitlich getroffen wird. Überwiegend verbindet man mit der *pooled analysis* eine Situation, in der relativ kurze Zeitreihen bei einer überschaubaren Menge von Objekten (meist Regionen oder Organisationen) in gleicher Weise erhoben worden sind. So z. B. untersuchte R. Gartner (1990) für 18 Länder die jährlichen Homizidraten von 1950-1980 in Abhängigkeit von ökonomischen, sozialen und kulturellen Einflußgrößen. Fajnzylber et al. (2002) liefern ein anderes Beispiel, in dem fünf Fünf-Jahres-Durchschnitte der Raubüberfälle in 34 Staaten analysiert werden. Eine Einführung in diese Methode bietet Sayrs (1989); s. auch Stimson (1985), Beck/Katz (1995). Bei der *Event History Analysis* (siehe z. B. Andreß 1992; Blossfeld/Hamerle/Mayer 1986) liegt wiederum eine andere Datenstruktur vor. Typischerweise werden bei einer Vielzahl von Untersuchungsobjekten (Personen, Organisationen etc.) innerhalb einer bestimmten Periode Ereignissequenzen registriert, die „Zustandswechsel“ markieren; bei einzelnen Personen bspw. den Wechsel von der Arbeitslosigkeit in ein Beschäftigungsverhältnis, bei Paaren die Geburt von Kindern oder der Bruch von Beziehungen, bei Organisationen der Übergang zu einer neuen Expansionsstufe oder zur Auflösung. Für den Wechsel von einem Zustand in einen anderen können Übergangsraten berechnet (bzw. Wahrscheinlichkeiten geschätzt) werden, die sodann in ihrer Abhängigkeit von bestimmten äußeren Einflußfaktoren (wie Lebensalter, Bildungsniveau, Mitgliederstruktur) sowie von der Dauer des Verweilens in einem vorangegangenen Zustand untersucht werden können.

Bei den längsschnittanalytischen Methoden spielen zeitliche Veränderungen, ihr Ausmaß und ihre Geschwindigkeit eine große Rolle, nicht nur weil sie historisch interessant sind, sondern weil sich kausale Zusammenhänge im allgemeinen sicherer identifizieren lassen, wenn Beobachtungen für einen längeren Zeitraum vorliegen. Dennoch dürfte die Mehrzahl sozialwissenschaftlicher Untersuchungen (außerhalb der Wirtschaftswissenschaft) immer noch in Form von Querschnittanalysen vorgenommen werden, also auf der Basis von Daten, die bei einer Vielzahl von Untersuchungsobjekten nur ein einziges Mal (zum gleichen Zeitpunkt) erhoben werden. Allerdings haben insbesondere die Panel- und die Verlaufsdatenanalyse in jüngerer Zeit erheblich an Gewicht gewonnen. Daß die statistische Zeitreihenanalyse in der soziologischen Forschungspraxis immer noch relativ selten eingesetzt wird, hat verschiedene Gründe. Der vielleicht wichtigste liegt darin, daß „Zeitreihen“ im oben angegebenen Sinne immer noch Mangelware sind. Dies beginnt sich allerdings zu ändern. Die sozialgeschichtliche Forschung, die Umfrageforschung und die amtliche Statistik produzieren mehr und mehr Daten, die sich zu langen Zeitreihen zusammenfassen lassen. In der US-amerikanischen Politikwissenschaft gehört die Zeitreihenanalyse bereits zu den Standardverfahren der empirischen Forschung.³ Zahlreiche Publikationen und Institutionen präsentieren „Soziale Indikatoren“ in Form von Zeitreihen, die in unterschiedlichen

³ Siehe die jüngeren Jahrgänge in *American Political Science Review*, zahlreiche Beiträge in den Jahrbüchern von *Political Analysis* (oder der Vorgänger-Zeitschrift *Political Methodology*) oder Editionen wie die von H. Norpoth et al. (1991).

Bereichen gesellschaftlicher Wirklichkeit langfristige Entwicklungsprozesse abbilden, so z. B. ZUMA.⁴ Leider werden dabei häufig Informationen durch zu hohe zeitliche Aggregation verschenkt; wöchentlich oder monatlich vorliegende Daten werden z. B. zu Jahresdaten gemittelt und nur in dieser Form publiziert. Eine bessere Kenntnis zeitreihenanalytischer Verfahren hätte dem mancherorts entgegenwirken können: Verfahrenkenntnisse und die Erhebung adäquater Daten, die sich mit diesen Verfahren analysieren lassen, fördern sich wechselseitig; wenn dagegen das Analysepotential bestimmter Verfahren nicht bekannt ist, kümmert man sich auch nicht um den Aufbau dazu passender, adäquater Datenstrukturen.

Die mathematischen und statistischen Grundlagen der Zeitreihenanalyse sind relativ komplex; auch dies dürfte ein Grund für ihre zögerliche Rezeption unter Soziologen und Historikern sein. Aber für die Zeitreihenanalyse gilt das gleiche wie für andere „fortgeschrittene“ statistische Verfahren, derer man sich in den Geschichts- und Sozialwissenschaften bedient: sie lassen sich für spezifische Anwendergruppen unter Berücksichtigung der jeweils gegebenen Voraussetzungen und Analyseziele auf unterschiedlichen Niveaus darstellen. Ein „Statistik“-Lehrbuch, das (wie das vorliegende) für Sozialwissenschaftler und Historiker geschrieben wird, kann die mathematisch-statistischen Grundlagen (etwa im Bereich der Wahrscheinlichkeitstheorie) allenfalls streifen; das entsprechende Fach-Vokabular nur begrenzt heranziehen. Die für den Mathematiker oder die Statistikerin selbstverständliche Einheit von formaler Genauigkeit und Verständlichkeit löst sich dabei bis zu einem gewissen Grade auf; der diskursiv-sprachliche Aufwand steigt; mathematische Ableitungen und Beweisführungen werden nur eingesetzt, wo sie für den Nicht-Mathematiker leicht nachvollziehbar sind. Wir hoffen, einen für die Adressaten dieses Lehrbuchs akzeptablen Kompromiß zwischen den nicht ganz kompatiblen Erfordernissen der Verständlichkeit und der formalen Genauigkeit sowie der Vielfalt der behandelten Themen und der Ausführlichkeit ihrer Darstellung gefunden zu haben. Das vorliegende Lehrbuch setzt somit beim Leser, bei der Leserin keine mathematischen Kenntnisse voraus, die über das durchschnittliche Abiturwissen hinausgehen.⁵ Für das Verständnis zeitreihenanalytischer Modelle ist es hilfreich, mit elementaren Formen der Differenzgleichungen vertraut zu sein. Da entsprechende Kenntnisse im Schulunterricht nicht immer erworben werden, bieten wir hierzu in Anhang 4 eine knappe Einführung. Auch andere Anhänge dienen dazu, Schulkenntnisse aufzufrischen oder in einfacher Weise zu ergänzen. Im Bereich der „Statistik“ setzen wir lediglich Grundkenntnisse der Inferenzstatistik und der (multiplen) Regressionsanalyse voraus, wie sie üblicherweise in einem sozialwissenschaftlichen Methoden-Curriculum an den Universitäten (oder - für Historiker - bspw. in den „Herbstseminaren“ des Zentrums für Historische Sozialforschung) vermittelt werden (s. Kühnel/Krebs 2004; Thome 1990).

Wie schon angedeutet, stehen für die Zeitreihenanalyse verschiedene Modelle und Verfahrensweisen zur Verfügung. Wir halten es für sinnvoll, in einem

⁴ Zentrum für Umfragen, Methoden und Analysen, Mannheim. Internet Adresse: <http://www.gesis.org/Dauerbeobachtung/Sozialindikatoren/index.htm>

⁵ Wer tiefer in die Materie einsteigen will, dem sei das Werk von Schlittgen/Streitberg (2001) empfohlen.

Einführungstext nicht möglichst viele Modelle und „Ansätze“ darzustellen, sondern einige auszuwählen und diese dann relativ ausführlich zu erörtern. Die Zeitreihenanalyse ist in ihrer ersten Stufe eher „deskriptiv“ angelegt. Dabei stützt sie sich auf graphische Darstellungen und einfache Modelle der Unterscheidung (und Verknüpfung) von „Trend“ und „Zyklen“, von „glatten“ und „irregulären“ Komponenten (s. Kap. 2). In dieser Phase wird aber nicht nur die Anwendung komplexerer (stochastischer) Modelle vorbereitet; vielmehr können die dort gewonnenen Einsichten auch in anderer Weise heuristisch genutzt werden. Insbesondere helfen sie dem Soziologen bei der (häufig notwendigen) Historisierung seines Gegenstandes; sie machen die immer wieder angemahnte „historische Perspektive“ zu einem realisierbaren Vorhaben und lassen die beiden Disziplinen näher aneinander rücken. Dies verdeutlichen wir einleitend im folgenden Kapitelabschn. 1.2.

Üblicherweise unterscheidet man in der Zeitreihenanalyse „Analysen im Frequenzbereich“ von „Analysen im Zeitbereich“. Das erste Etikett bezieht sich auf Verfahren, mit denen die Zeitreihen als Kompositum verschiedener Schwingungen mit unterschiedlichen Frequenzen (darstellbar mit Hilfe trigonometrischer Funktionen) modelliert werden. „Periodogramm-“ und „Spektralanalyse“ sind die bekanntesten Varianten aus dieser Gruppe. Wir geben hier nur eine knappe Einführung in die Periodogrammanalyse (Kap. 2.4), die allerdings so gestaltet wird, daß sie den Zugang zur Literatur über die Spektralanalyse erleichtern dürfte.⁶ Im übrigen beschränken wir uns auf Analysen im „Zeitbereich“, eine etwas wunderliche Bezeichnung, die wohl andeuten soll, daß die zeitlich diskrete Abfolge der beobachteten Werte nicht in einem Aggregat von Schwingungskomponenten aufgelöst wird. Da SozialwissenschaftlerInnen vor allem daran interessiert sind, strukturelle Zusammenhänge zwischen zwei oder mehreren Variablen(gruppen) zu untersuchen, stehen verschiedene Formen der „dynamischen“ Regressionsanalyse im Mittelpunkt unseres Lehrbuches (Kap. 5). Anders als Querschnittdaten bieten Zeitreihen im Prinzip die Möglichkeit, nicht nur Gleichgewichtsbeziehungen zu modellieren („statische“ Analyse), sondern den zeitlichen Verlauf der Wirkung verschiedener Einflußgrößen zu rekonstruieren („dynamische“ Analyse). Dazu ist es allerdings nötig, zunächst die einzelnen Zeitreihen univariat zu analysieren. In der historischen Entwicklung der Zeitreihenanalyse (s. Rahlf 1998) sind hierzu unterschiedliche Modelle (die sog. *autoregressiven* und die *moving average* Modelle) vorgeschlagen worden, für die Box/Jenkins (1976) schließlich einen vereinheitlichenden Bezugsrahmen geschaffen haben (s. Kap. 3). Die „Box/Jenkins-(ARIMA)-Modelle“ bieten zudem den Vorteil, daß in sie auch externe Einflußgrößen im Sinne der Regressionsanalyse eingebaut werden können (*Transferfunktionsmodelle*, Kap. 5.5). In Gestalt der *vektoreautoregressiven Modelle* (s. Kap. 7) sind sie außerdem zu multivariaten Analyseformen erweiterbar, die ohne die Spezifikation einer einzelnen abhängigen Variablen auskommen und auf diese Weise Rückkopplungsprozesse darstellbar machen. Die „ökonometrische“ Tradition hypothestetestender Regressionsanalyse (s. Kap. 5.2 bis 5.4) läßt sich unseres

⁶ Weitere Hinweise zur frequenzanalytischen Betrachtungsweise werden in Abschn. 3.4.4 gegeben. Zur Einführung in die Spektralanalyse siehe z. B. Leiner (1982), Gottman (1981).

Erachtens fruchtbar verbinden mit dem „statistischen“ Ansatz der Box/Jenkins-Methodologie, die versucht, aus der Datenanalyse heraus angemessene Modelle zu identifizieren. Dies gilt sowohl für die einfache Regressionsanalyse als auch für Mehrgleichungssysteme im Sinne der *Strukturgleichungsmodelle*, die ihr Pendant in den *vektorautoregressiven Modellen* finden.

Grundlage all dieser statistischen Modelle ist das Konzept „stochastischer Prozesse“, die Idee also, daß sich die in Zeitreihen registrierten Werte als Realisationen von Zufallsprozessen begreifen lassen, die jederzeit auch (etwas) anders hätten ausfallen können (oder in Zukunft ausfallen werden). Zeitreihenanalytische „Modelle“ sind somit formalisierte Vorstellungen über nicht-deterministische Prozesse, deren Erzeugungsmechanismen in gewissen Hinsichten als konstant angenommen werden (*Stationaritätsannahmen*), deren zukünftige Realisationen im einzelnen aber unbekannt sind. Dieses grundlegende Konzept werden wir zu Beginn des Kap. 3 relativ ausführlich erörtern.

Die in diesem Einführungstext behandelten Modelle und Verfahren sind so ausgewählt und dargestellt worden, daß sie einerseits eine gute Startposition verschaffen, um weiterführende (komplexere) Modelle bzw. die entsprechende Literatur zu rezipieren, andererseits aber auch selbst schon ein beachtliches Anwendungspotential anbieten. Dazu gehört das breite Feld der Evaluationsforschung, der Analyse von Folgen ungeplanter Ereignisse oder zielgerichtet eingesetzter Maßnahmen. Die von Box und Tiao vorgeschlagenen „Interventionsmodelle“ (s. Kap. 4) sind in diesem Bereich in vielfältiger Weise einsetzbar.

Die Zeitreihen, die dem Sozialwissenschaftler oder Historiker zur Verfügung stehen, weisen nicht selten Merkmale auf, die den Anwendungsvoraussetzungen der Modelle widersprechen, die wir im vorliegenden Lehrbuch primär behandeln. Dazu gehören vor allem „Ausreißer“ und fehlende Werte sowie Strukturdynamiken, die zu nicht-konstanten Varianzen führen, u. U. auch nicht-konstante Modell-Parameter erfordern. Wir können diese Problematiken aus Platzgründen im Rahmen dieses Buches nicht behandeln. Zur Modellierung von Ausreißern und nicht konstanten Varianzen sei auf die einführenden Texte von Thome (1995b; 1994a) hingewiesen. Zu zeitlich variierenden Modell-Parametern siehe Beck (1983), Isaac/Griffin (1989), Newbold/Bos (1985).

Neben der Spektralanalyse werden in diesem Einführungstext zwei weitere, relativ komplexe „Ansätze“ zur Zeitreihenanalyse nicht explizit behandelt: Harveys „struktureller Ansatz“ (s. Harvey 1989) sowie „Zustandsraummodelle“ und „Kalman-Filter“ (s. Aoki 1987). Die hier dargestellten Modelle liefern aber wichtige Bausteine auch für diese umfassenderen Ansätze bzw. Modellklassen. Harveys Ansatz ist z. B. für diejenigen Anwender von Interesse, die die Komponentenzerlegung (die wir hier nur in ihrer „klassisch“-deskriptiven Form darstellen) in einem stochastischen Bezugsrahmen betreiben wollen. Die Zustandsraummodelle bieten u.a. eine Möglichkeit, auch in der Zeitreihenanalyse mit latenten Variablen zu arbeiten, die mit Hilfe beobachtbarer „Indikatoren“ (fehlerhaft) gemessen sind (s. Beck 1989). Der Kalman-Filter ist zudem bei der Analyse von Zeitreihen mit fehlenden Werten und nicht gleichabständigen Beobachtungen sowie bei der

Schätzung von zeitabhängigen Modell-Parametern einsetzbar (s. hierzu ebenfalls Beck 1989).

1.2 Zum heuristischen Nutzen der Zeitreihenanalyse

Die Analyse einer Zeitreihe beginnt mit der genauen Betrachtung ihrer Verlaufsform, wie sie in einer angemessenen graphischen Darstellung (s. Cleveland 1994) sichtbar wird. Auch unabhängig vom Ziel einer statistischen Modellierung können auf diesem Wege Einsichten gewonnen werden, die in verschiedenen Untersuchungskontexten zu einer gehaltvolleren Problemkonzeption verhelfen. Wir möchten das anhand zweier Beispiele veranschaulichen.

In einem seiner Bücher zum „Wertewandel“ konstatiert Klages (1985) einen Zusammenhang zwischen dem „Wertewandelsschub“ von Mitte der 60er bis Mitte der 70er Jahre und der Veränderung der Geburtenrate in der Bundesrepublik Deutschland. Vor allem auf der Basis wiederholter Repräsentativbefragungen wird ein Rückgang der sog. Pflicht- und Akzeptanzwerte und eine stärkere Verbreitung der idealistisch oder hedonistisch ausgerichteten Werte der Selbstverwirklichung registriert, während die Geburtenziffern (pro 1000 Einwohner), wie sie in amtlichen Statistiken erscheinen, im gleichen Zeitraum einen abwärts gerichteten Trend aufweisen (s. Abb. 1.2.1).

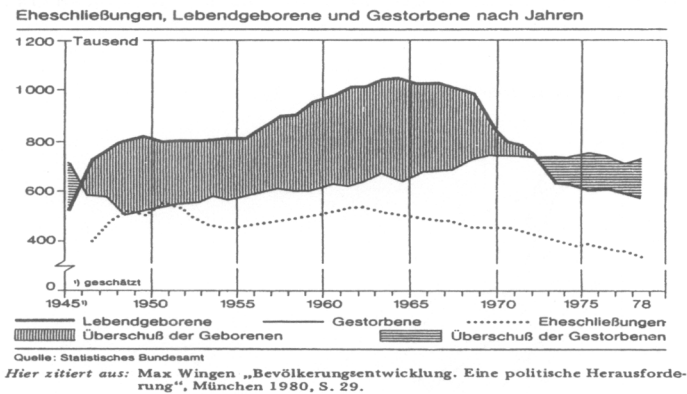


Abb. 1.2.1 Graphik zur Entwicklung der Geburtenrate (nach Klages 1985: 117)

Die zeitliche Korrespondenz der lokalen Trendverläufe wird als Indiz, wenn nicht gar als Beleg für einen strukturellen Zusammenhang angesehen; die Rückläufigkeit der Geburtenrate erscheint als Folge veränderter Wertorientierungen.⁷ Eine kurzfristige Korrespondenz im Trendverlauf zweier Variablen legt aber ganz andere Interpretationen nahe, wenn man den früheren und den weiteren Verlauf mit in den Blick nimmt. Schon der von Klages betrachtete Zeitabschnitt weist für das

⁷ Allerdings muß man Klages zugute halten, daß er seine These auch mit Individualdaten zu belegen sucht, die zeigen, daß Personen mit traditionaler Wertorientierung eher den Wunsch nach Kindern überhaupt oder nach einer großen Zahl von Kindern äußern als Personen mit „neuen“ Wertorientierungen.

Jahrzehnt zwischen Mitte der fünfziger und Mitte der sechziger Jahre einen Geburtenanstieg aus, der in das Erklärungsmuster hätte einbezogen werden müssen. Warum wird nur der Rückgang nach 1965, nicht der vorangegangene Anstieg als erklärungsbedürftig angesehen? Wenn man sich nicht von vornherein auf *Ad-hoc*-Erklärungen beschränken will, müßte man den Anstieg mit der gleichen Theorie erklären wie den Abstieg - also annehmen, daß zwischen 1955 und 1965 der Anteil der Leute mit traditioneller Wertorientierung zugenommen hat. Vermutlich resultiert dieser Anstieg aber aus einem „Baby-Boom“, der eine Generation vorher (Mitte der dreißiger Jahre) eintrat, aber erst sichtbar wird, wenn man die Zeitreihe der Geburtenrate weiter in die Vergangenheit zurückverfolgt (s. unten, Abb. 1.2.2). Wenn - aus welchen Gründen auch immer - irgendwann einmal ein Baby-Boom stattfindet, wird auch eine Generation später wieder ein gewisser Geburtenanstieg zu verzeichnen sein, weil dann der Anteil an Frauen im gebärfähigen Alter besonders hoch ist. Danach muß dann auch wieder ein Abstieg folgen. Der von Klages mit dem Wertewandel erklärte Abstieg ist sicherlich zu einem größeren Teil eine direkte Folge des vorangegangenen Anstiegs. Könnte man den irgendwie „herausrechnen“, wäre der Abstieg weniger steil und läge vielleicht sogar im Trend der ganz langfristigen Entwicklung. Wenn man die Geburtenziffern noch weiter zurückverfolgt, zeigt sich denn auch ein relativ kontinuierlicher Abstieg von etwa der Jahrhundertwende bis zum zweiten Weltkrieg in fast allen europäischen Ländern (s. Abb. 1.2.2, in der die Daten über Fünfjahresintervalle gemittelt wurden).

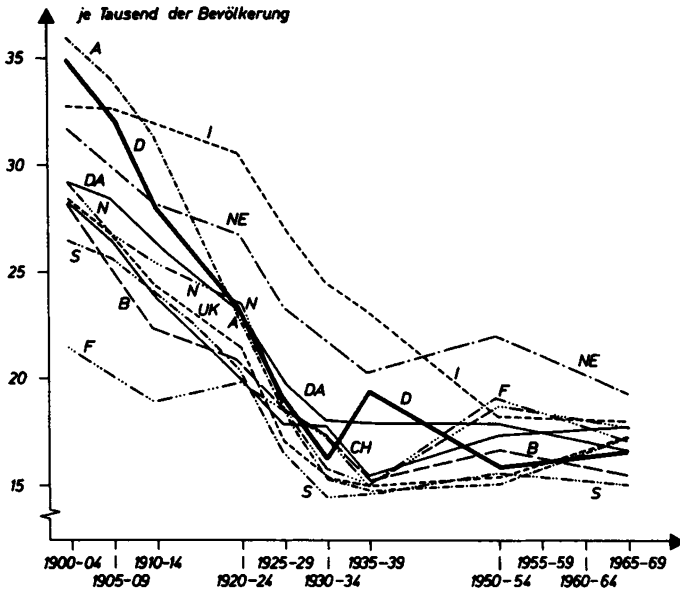


Abb. 1.2.2 Langfristige Geburtenentwicklung nach Borchardt (1982)

Dieser langfristig abfallende Trend kann ja nun nicht mehr mit einem Wertwandelsschub erklärt werden, der Mitte der 60er Jahre einsetzt. Was zunächst als „Besonderheit“ erschien, der Geburtenrückgang nach 1964, und deshalb mit einer anderen „Besonderheit“, nämlich dem Wertwandelsschub, erklärt werden sollte, erscheint nun gar nicht mehr als etwas Besonderes. Möglicherweise sind Wertewandel und Rückgang der Geburtenziffer gemeinsam in eine umfassendere Entwicklungsdynamik einbezogen, die von vielen Soziologen und Historikern bspw. unter dem Etikett gesellschaftlicher Modernisierung interpretiert wird. Darauf wollen wir hier nicht weiter eingehen (s. Hirschmann 1994; Nauck 2001). Es sollte lediglich verdeutlicht werden, wie sich durch die Erweiterung der historischen Perspektive die „Explananda“ verändern und kurzschlüssige Erklärungen vermieden werden können. Es ist oft so, daß das, was als erklärungsbedürftig erscheint, erst durch eine Vorstellung von dem langfristigen Trend sinnvoll definiert werden kann. Hierzu noch einmal ein kurzer Blick auf Abb. 1.2.2, insbesondere die fett eingezzeichnete Entwicklungslinie für Deutschland. Angenommen man findet eine plausible Erklärung für die stark abfallende Trendentwicklung von 1900 bis 1925. Die Interpretation für den folgenden Zeitabschnitt erscheint dennoch problematisch. Borchardt bemerkt hierzu: „Irgendwo scheint hier im Vergleich zum Trend eine Irregularität zu bestehen, wir wissen nur noch nicht wo. Zwei verschiedene Möglichkeiten der Erklärung gibt es: a) Man hält den Ausschlag 1930-34 für eine Abweichung vom Trend nach unten, weil in der Weltwirtschaftskrise aus wirtschaftlichen und sonstigen Gründen die Geburtenfreudigkeit besonders stark abgenommen hat. b) Man hält demgegenüber den Ausschlag 1935-39 für eine Abweichung vom Trend nach oben, weil man die NS-Bevölkerungspolitik und/oder den Sinneswandel der Bevölkerung in der Phase der NS-Erfolge im Frieden für besonders wichtig hält“ (Borchardt 1982: 122 f.). Borchardt glaubt, daß in diesem Falle der internationale Vergleich die Frage entscheidbar macht. Die Trendentwicklung verläuft in den meisten Ländern ziemlich parallel bis zur ersten Hälfte der dreißiger Jahre. Es gibt danach eher eine speziell deutsche Abweichung vom Normaltrend nach oben als zuvor eine national-spezifische Abweichung nach unten. Ähnliche Beispiele lassen sich anführen, die dann die generelle Aussage belegen: „Verschiedene Deutungen der langfristigen Entwicklung implizieren notwendig unterschiedliche Ansichten über die Eigenschaften und die Gründe der Entwicklung in den kürzeren Perioden“ (Borchardt 1982: 116). Nicht nur die Trendverläufe, auch die „irregulären“ Komponenten, d. h. die Abweichungen vom Langfristmodell können für die Analyse relevant werden, indem sie beispielsweise auf politische Gestaltungseffekte hinweisen, die nicht schon in der endogenen Prozeßdynamik enthalten sind. Borchardt diskutiert in dieser Perspektive u. a. den möglichen Einfluß des Nazi-Regimes auf die deutsche Wirtschaftsentwicklung. Die Trennung von langfristigen und kurzfristigen Komponenten ist allerdings per Augenschein nur selten mit ausreichender Sicherheit vorzunehmen, zumal wenn eine langfristige Trendkomponente noch von einer zyklischen Komponente überlagert wird. In Kap. 2 werden relativ einfache Methoden vorgestellt, mit denen dies zu bewerkstelligen ist. Zuvor möchten wir aber die heuristische Funktion der Zeitreihenanalyse noch an einem weiteren Beispiel demonstrieren.

Dabei geht es um den Anstieg der Kriminalitätsraten in der Bundesrepublik von 1964 bis 1984 und die theoretische Deutung bzw. Erklärung dieses Phänomens. Baldo Blinkert (1988) hat hierzu eine sehr interessante Abhandlung verfaßt. Der Autor sieht in der in Abb. 1.2.3 wiedergegebenen Entwicklung⁸ einen Beleg für einen exponentiellen Trendanstieg im kriminellen Verhalten.

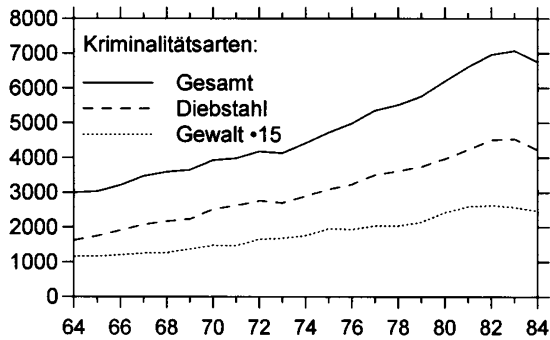


Abb. 1.2.3 Kriminalitätsraten (Häufigkeitszahlen pro 100.000 Einw.) Bundesrepublik Deutschland 1964-1984 (Quelle: ZUMA-Indikatoren)

Er deutet diesen Anstieg als Folge eines langfristigen Prozesses der Rationalisierung und Individualisierung. Diese Prozesse führen, so sein Argument, dazu, daß sich traditionale, verinnerlichte Normbindungen auflösen, daß demgemäß normkonformes Verhalten zunehmend ein Resultat nutzenorientierter Entscheidungen im Verfolgen hedonistischer Motive geworden sei. Weitere strukturelle Bedingungen - wie z. B. sinkende Aufklärungsquoten - sollen dazu führen, daß die Nutzen/Kosten-Kalkulation zunehmend ein normabweichendes, kriminelles Verhalten nahelegt. Außerdem versucht er, durch eine negative Korrelation von Wachstumsraten des Bruttosozialprodukts mit den Wachstumsraten der Kriminalitätsentwicklung eine zyklische, in diesem Falle eine zur Konjunktur antizyklische Komponente der Kriminalitätsentwicklung nachzuweisen. Darüber hinaus deutet er den exponentiellen Anstieg als Hinweis auf autokatalytische Momente, also positive Feedbackbeziehungen: je mehr kriminelle Handlungen vorliegen, desto geringer wird die Präventivwirkung des Nichtwissens, d. h., die Möglichkeiten ungesetzliches Verhalten zu erlernen, verbessern sich. Blinkerts Analyse ist insofern vorbildlich, als sie die Kriminalitätsraten als Kompositum mehrerer Komponenten ansieht, die mit unterschiedlichen Wirkungsfaktoren verknüpft sind. Er unterscheidet (a) eine langfristige Trendkomponente, die er mit den Konzepten der Individualisierung und Rationalisierung erklären möchte; (b) eine langfristige zyklische Komponente, die er auf ökonomische Konjunkturverläufe zurückführt; (c) kurzfristige Fluktuationen, die z. B. durch

⁸ Da die Abbildung aus dem Originaltext nicht reproduzierbar war, wurde eine entsprechende Abbildung mit den Daten der ZUMA-Indikatoren (s. Anm. 4) erstellt. Die Rate der Gewaltkriminalität wurde mit dem Faktor 15 multipliziert.

besondere polizeiliche Maßnahmen und bestimmte Ereigniskonstellationen hervorgerufen sein mögen; (d) positive Feedbackmechanismen, die den Prozeß aufschaukeln und durch einen exponentiellen Verlauf der Zeitreihen indiziert sein könnten.

Blinkerts Hypothesen sind interessant, aber auch hoch spekulativ. Die Interpretation greift erheblich über das vorliegende Datenmaterial hinaus. Wenn man von dem Problem einer adäquaten Operationalisierung der erklärenden Variablen einmal absieht, ist vor allem zu fragen, ob das Explanandum mit dem Kriminalitätsanstieg von 1964 bis 1984 ausreichend erfaßt ist. Wäre es noch dasselbe Explanandum, wenn in einer Periode vor 1964 oder nach 1984 eine abfallende Trendbewegung zu beobachten wäre, wenn also die aufsteigende Linie, die hier dargestellt ist, lediglich der aufsteigende Ast einer Wellenbewegung wäre? Könnte es sich hier um eine relativ kurzfristige Abweichung von einem ganz anders gestalteten langfristigen Trend handeln? Man muß sich diese Fragen insbesondere dann stellen, wenn man die relativ kurzfristig beobachtete Trendentwicklung mit einem säkularen Prozeß - Individualisierung, Rationalisierung - erklären will, der doch wohl eine viel längere Zeitspanne umfassen soll. Inzwischen liegen längere Datenreihen vor, nicht nur für Deutschland, sondern auch für andere Länder. Abb. 1.2.4 zeigt eine von Eisner (1995) vorgelegte Aggregation der Gesamtkriminalitätsraten für sechs europäische Länder.

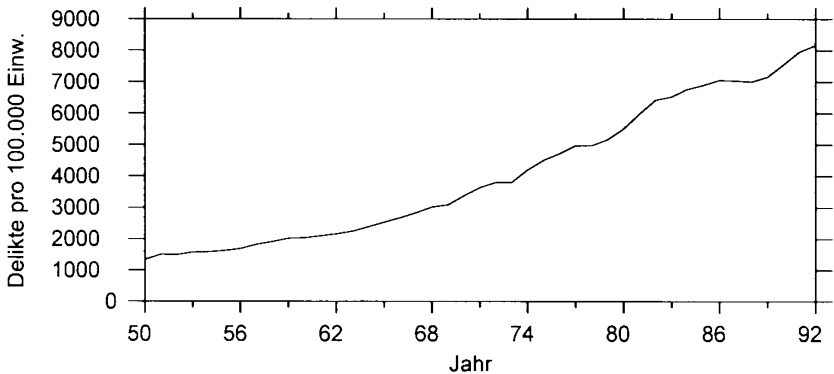


Abb. 1.2.4 Aggregierte Kriminalitätsraten für 6 Länder (Quelle: Eisner 1995)

Blinkert hat offenkundig insofern „Glück“, als sein Eindruck eines exponentiellen Wachstums der Kriminalitätsraten mit den Daten eines längeren Beobachtungszeitraums durchaus bestätigt wird - auch wenn man andere Länder mit einbezieht. Wenn man aber eine im weitesten Sinne modernisierungstheoretische Erklärung versuchen will, nimmt man in der Regel auf Prozesse Bezug, die noch weiter zurückreichen. Das heißt, die aufsteigende Linie der Kriminalitätsentwicklung müßte sich noch weiter zurückverfolgen lassen - oder man müßte begründen können, in welchem Sinne die von Blinkert bemühten Rationalisierungs- und Individualisierungsprozesse nach dem 2. Weltkrieg eine

neue Qualität oder einen entscheidenden Niveauanstieg erfahren. Jedenfalls ist es sinnvoll, die Zeitreihe der Kriminalitätsentwicklung noch früher beginnen zu lassen, bevor man den Anstieg nach 1950 summarisch der sich immer weiter voranschreitenden gesellschaftlichen Modernisierung zurechnet. Hier wird die Datenlage natürlich noch unsicherer. Für Tötungsdelikte ist sie beim internationalen Vergleich über lange Zeiträume noch am günstigsten. Eine Darstellung, die die Trendentwicklung der Tötungsdelikte seit Mitte des 19. Jahrhunderts für mehrere Länder⁹ erkennbar macht (s. Abb. 1.2.5), verdanken wir wiederum Eisner (1995).

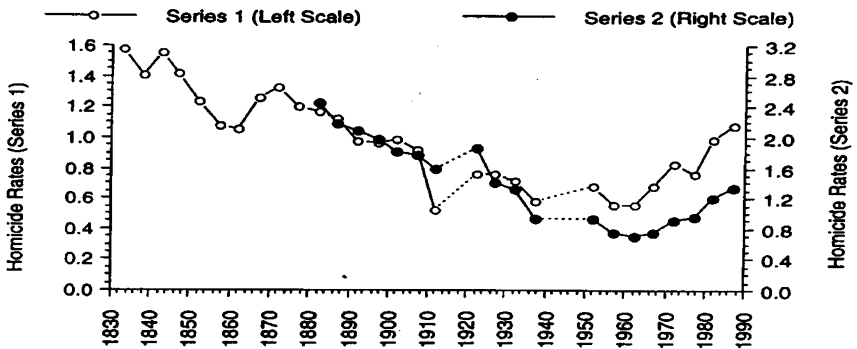


Abb. 1.2.5 Durchschnittliche Homizidraten in mehreren europäischen Ländern
(Quelle: Eisner 1995)

Jetzt wird das Explanandum doch ein bißchen komplexer, denn eine allgemeine Theorie muß ja nun beides erklären können, den abfallenden Trend spätestens seit Ende des 19. Jahrhunderts wie auch den ansteigenden Trendverlauf in der zweiten Hälfte des 20. Jahrhunderts. Wenn man weiterhin die These eines positiven Zusammenhangs zwischen Individualisierung/Rationalisierung als Element einer allgemeinen Theorie aufrechterhalten will, müßte man belegen können, daß Individualisierung/Rationalisierung zwischen ca. 1830 und 1950/60 ab- und erst danach (wieder?) zunehmen oder man müßte plausibel darlegen können, daß eine nicht-lineare, U-förmige Beziehung zwischen diesen Variablenkomplexen besteht; oder man müßte weitere Erklärungsvariablen einführen, die historisch spezifische Trendentwicklungen aufweisen.

Wir möchten auch hier die inhaltlichen Fragen nicht weiter verfolgen,¹⁰ sondern an dieser Stelle nur ein knappes methodologisches Fazit ziehen: Die Identifikation struktureller Zusammenhänge zwischen verschiedenen Komponenten sozialer Prozesse profitiert davon und ist häufig dringend darauf angewiesen, den langfristigen Verlauf der einzelnen Indikatoren genau zu betrachten. Gelingt es nicht, länger- und kurzfristige Entwicklungsdynamiken, Trend, Zyklen und irreguläre Komponenten auseinanderzuhalten, läuft man Gefahr, die kausalen Mechanismen falsch zuzuordnen,

⁹ „Serie 1“ umfaßt die Länder England/Wales, Frankreich und Schweden, „Serie 2“ zusätzlich Belgien, Italien, Niederlande, Schweiz.

¹⁰ Eisner (1997; 2002) stellt einige Erklärungsansätze vor; s. auch Thome (1995a; 2004).

bspw. eine im bisherigen Trend liegende Zunahme des Deliktiveaus ursächlich einer aktuellen, zeitlich eng begrenzten Ereigniskonstellation zuzuschreiben. Gerade aus dem Bereich der Kriminalitätsforschung ließen sich hierzu zahlreiche Beispiele nennen. Demgegenüber läßt sich folgende allgemeine Regel formulieren: Wenn sich für irgendein als erklärungsbedürftig angesehenes soziales Phänomen ein langfristiger Trendverlauf feststellen läßt, kann er nur durch Faktoren erklärt werden, deren Wirkung sich über eine *entsprechend* lange Periode erstreckt. Dies kann bedeuten, daß sich die erklärenden Faktoren selbst in einer korrespondierenden Trendbewegung entwickeln (s. die in Kap. 5.6 dargestellten Kointegrationsmodelle) oder daß sie sich zwar in einer kürzeren Zeitspanne aufbauen, aber aus angebbaren Gründen ihre Wirkung(en) erst längerfristig entfalten (s. bestimmte Varianten der in Kap. 4 dargestellten Interventionsmodelle). Dieses Prinzip paralleler Zeitdimensionalität von Explanandum und Explanans (von abhängigen und von bedingenden Größen) gilt in entsprechender Weise natürlich auch bezüglich der kurzfristigen *Abweichungen* von Trend oder Zyklus, die ebenfalls erklärungsbedürftig sein können.

Die analytische Statistik bietet ein reiches Arsenal von Modellen und Verfahrensweisen an, mit denen man Zeitreihen in verschiedene Komponenten (wie Trend, Zyklus, irreguläre Schwankungen) zerlegen kann. Ihre Anwendung stößt jedoch selbst dann auf große Schwierigkeiten, wenn man Datenlücken und die (hier nicht thematisierten) Probleme der Validitätsprüfung ebenso außer acht läßt wie die in der Sozialgeschichte häufig auftretenden „Ausreißer“ und Strukturbrüche. Denn sinnvoll einsetzbar sind diese Modelle nur mit bestimmten theoretischen Vorgaben: z. B. über die lineare oder nicht-lineare Form des Trendverlaufs oder die Länge einer Zyklus-Periode. Hier beißt sich die Katze leider wirklich in den Schwanz; denn solche Theorien (bspw. eine Theorie der langfristigen Kriminalitätsentwicklung) liegt häufig nicht vor. Vielmehr hofft man, mit Hilfe einer mehr oder weniger phantasievollen Interpretation der Datenreihen entsprechende Hypothesen allererst entwickeln zu können. Eine solche „induktiv“ gewonnene Hypothese müßte dann allerdings anhand anderer Daten auf ihre Haltbarkeit getestet werden. Das Beispiel der Wirtschaftsgeschichte, die dort geführten Diskussionen über deterministische oder stochastische Wachstumspfade, über lange und kurze Wellen der ökonomischen Entwicklung zeigen, wie schwierig ein solches Unterfangen ist, und wie unwahrscheinlich es ist, auf diesem Wege zu einem Konsens zu gelangen (s. hierzu Metz 2002). Und dort liegen immerhin ausgearbeitete Theorien ökonomischen Wachstums vor; auch die Datenlage ist wesentlich günstiger als - beispielsweise - in der historischen Kriminalitätsforschung. Aber selbst wenn die formalen Modelle der Komponentenerlegung nicht so recht zum Zuge kommen, kann die Kenntnis dieser Modelle und der mit ihnen verknüpften Analyseverfahren bei der visuellen Inspektion und Interpretation der Daten durchaus von Nutzen sein und vor Fehlinterpretationen schützen.

Kapitel 2

Beschreibende Analyse von Zeitreihen

2.1 Empirische Kennzahlen: Mittelwert, Streuung, Autokorrelation

Aus statistischen Grundkursen sind als beschreibende Kennziffern (empirische „Momente“) univariater Verteilungen vor allem das arithmetische Mittel und die Varianz (bzw. die mittlere quadratische Abweichung) bekannt. Sie lassen sich auch mit Zeitreihendaten berechnen und zur Kennzeichnung der „Lage“ und der „Streuung“ der Beobachtungswerte verwenden:

$$\text{arithm. Mittel } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i, \quad n = \text{Zahl der Beobachtungen} \quad (2.1.1)$$

$$\text{Varianz } s^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 \quad (2.1.2)$$

Diese Meßgrößen sind als summarische Beschreibung der gesamten Reihe offensichtlich nur in dem Maße aussagekräftig, wie sich deren „Systematik“ innerhalb des Beobachtungszeitraums nicht verändert. Die hiermit angesprochene „Stationarität“ einer Zeitreihe wird erst im 3. Kapitel genauer definiert. Grob gesprochen bedeutet diese Eigenschaft, daß die Kennziffern der Reihe stabil sind, daß sie allenfalls geringfügig variieren, wenn sie für unterschiedliche Abschnitte des Beobachtungszeitraums getrennt berechnet werden. Im Rahmen der deskriptiven Analyse läßt sich dies schon mit Hilfe der grafisch dargestellten Sequenz der Beobachtungswerte hinlänglich überprüfen. Bei den Arbeitslosendaten in Abb. 1.1.2 z. B. ist leicht erkennbar, daß das Jahresniveau der Zeitreihe keineswegs stabil ist; auch die saisonalen Schwankungen sind nicht konstant. Arithmetisches Mittel und Varianz, über den gesamten Zeitraum berechnet, würden diese Reihe nicht angemessen beschreiben.

In der Querschnittanalyse geht man in der Regel davon aus, daß die bei den verschiedenen Erhebungseinheiten („Fällen“) erzielten Meßergebnisse unabhängig voneinander realisiert worden sind. Bei Zeitreihen-Daten kann man dies nicht voraussetzen. In der Regel werden die sukzessiv gewonnenen Beobachtungswerte mehr oder weniger stark miteinander korrelieren, die in unmittelbarer zeitlicher Nachbarschaft registrierten Werte normalerweise stärker als die weiter auseinander liegenden Größen. Die Arbeitslosenrate für einen beliebigen Monat t , z_t , hängt mit der Rate des Vormonats, z_{t-1} , eng zusammen und diese wiederum mit der Rate ihres Vormonats, z_{t-2} , usw. Allerdings wäre es keine Überraschung, wenn die Arbeitslosenrate für den Dezember eines Jahres j mit der Dezemberrate des Jahres $j-1$ stärker korrelierte als mit der Novemberrate des Jahres j (saisonale Effekte). Man

benötigt also Maßzahlen, die die interne Abhängigkeitsstruktur einer Zeitreihe differenziert abbilden. Es liegt nahe, dafür den Korrelationskoeffizienten von Bravais-Pearson zu verwenden. In der Querschnittanalyse wird er für zwei Variablen, X und Y , bekanntlich wie folgt definiert:

$$r_{yx} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \tag{2.1.3}$$

Der Index $i = 1, 2, \dots, n$ steht für die einzelnen Fälle der untersuchten Stichprobe. Im Zähler steht die (empirische) „Kovarianz“, $cov(y, x)$, im Nenner das Produkt der Standardabweichungen der beiden Variablen, s_x und s_y . Wie kann man einen solchen Korrelationskoeffizienten berechnen, wenn nur eine einzige Zeitreihe (also nur eine Variable) vorliegt? Die Antwort heißt: kopiere die Zeitreihe $\{Z_t\}$ und verschiebe sie um eine Zeiteinheit (ein Meßintervall) gegenüber dem Original. Verschiebt man sie nach „vorne“, erhält man auf diese Weise eine zweite Zeitreihe: $\{Z_{t-1}\}$; verschiebt man sie nach hinten, ergibt sich: $\{Z_{t+1}\}$. Die Zeitreihen lassen sich dann wie in der Querschnittanalyse in einem Streudiagramm darstellen. Die Datenpaare, die dazu die Koordinatenschnittpunkte liefern, sind nun nicht (x_t, y_t) , sondern

$$(z_{1t}, z_{2t}), (z_{2t}, z_{3t}), \dots (z_{pt}, z_{t-1}), \dots (z_{n-t}, z_n),$$

wobei die Ziffern im Index die verschiedenen Meßzeitpunkte $t = 1, 2, \dots, n$ anzeigen. Offensichtlich lassen sich die gleichen Datenpaare auch mit (z_{t-1}, z_t) bezeichnen, sofern der Zeitindex mit $t = 2, \dots, n$ angegeben wird. Betrachten wir ein Beispiel: In Tab. 2.1.1 sind die ersten elf Beobachtungswerte der in Abb. 1.1.4 wiedergegebenen Reihe der CDU-Präferenzen von Januar bis November 1950 aufgelistet.

Tab. 2.1.1: Prozentwerte CDU-Anhänger, Jan.-Nov. 1950

t	1	2	3	4	5	6	7	8	9	10	11
Z_t	26	27	27	32	38	38	39	31	31	33	26

Daraus ergeben sich die folgenden zehn Datenpaare, deren Elemente jeweils um ein Zeitintervall (einen Monat) auseinander liegen:

Tab. 2.1.2: Die Datenpaare mit zeitlicher Verschiebung, s. Tab. 2.1.1

t	1	2	3	4	5	6	7	8	9	10	11
CDU (t)	26	27	27	32	38	38	39	31	31	33	26
CDU (t-1)	*	26	27	27	32	38	38	39	31	31	33

Jedes dieser Datenpaare tragen wir in ein Koordinatenkreuz ein und erhalten somit das folgende Streudiagramm.

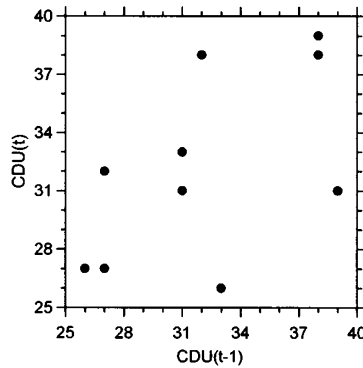


Abb. 2.1.1 Streudiagramm der Beispieldaten aus Tab. 2.1.2

Die Punktwolke streut um eine gedachte Linie, die von links unten nach rechts oben ansteigt und somit auf einen positiven Zusammenhang der beiden „Variablen“, hier Z_t und Z_{t-1} , hinweist. Man beachte, daß die Zahl der Datenpaare nicht wie in der Querschnittsanalyse n , sondern $n-1$ ist. Wenn wir diesen Zusammenhang als lineare Beziehung interpretieren wollen, können wir deren Stärke in Form des Pearsonschen Korrelationskoeffizienten analog zu Formel (2.1.3) darstellen. Der Zählerausdruck, die Auto-Kovarianz, erhält folgende Form:

$$\text{cov}(z_t, z_{t-1}) = \frac{1}{n-1} \sum_{t=2}^n (z_t - \bar{z}_t)(z_{t-1} - \bar{z}_{t-1}) \quad , \quad t=2, \dots, n \quad (2.1.4)$$

In entsprechender Weise läßt sich die Auto-Kovarianz für weiter auseinanderliegende Beobachtungen berechnen. Im folgenden sei die Zahl der Zeitintervalle, die die einzelnen Beobachtungen voneinander trennen, allgemein mit dem Buchstaben k ($k = 0, 1, 2, \dots$) angegeben. Man bezeichnet sie als „Verzögerungsspanne“ oder „Lag“. Aus Gleichung (2.1.4) wird somit:

$$\text{cov}(z_t, z_{t-k}) = \frac{1}{n-k} \sum_{k+1}^n (z_t - \bar{z}_t)(z_{t-k} - \bar{z}_{t-k}) \quad , \quad 0 < k < n \quad (2.1.5)$$

Die Zahl der Datenpaare ist $n-k$. In der Regel ermittelt man die Autokovarianzen in einem Bereich $1 \leq k \leq n/4$. In Zeitreihen mit stabilen Mittelwerten gehen die Autokovarianzen sozialwissenschaftlicher Zeitreihen mit zunehmendem k im allgemeinen rasch gegen Null. Es leuchtet unmittelbar ein, daß die Autokovarianz die interne Abhängigkeitsstruktur einer Zeitreihe nur dann gut repräsentiert, wenn dieser Zusammenhang über die gesamte Beobachtungsperiode $t = 1, 2, \dots, n$ einigermaßen konstant ist. Es hätte wenig Sinn, Kreuzprodukte über Datenpaare zu summieren, die am Anfang der Zeitreihe einen engen, gegen Ende der Zeitreihe aber nur einen schwachen Zusammenhang bilden. Dies führt zu der Forderung, die Zeitreihen sollten nicht nur mittelwert-, sondern auch *varianz/kovarianz-stationär*

sein. (Den Begriff der Stationarität werden wir eingehender erst in Kap. 3.3 erörtern.) Die schon in Gleichung (2.1.2) definierte *Varianz* ist in Gleichung (2.1.5) als Spezialfall enthalten, wenn $k = 0$ gesetzt wird. Wenn die Zeitreihen im Verhältnis zur Verzögerungsspanne relativ lang sind, werden sich die arithmetischen Mittel (wie auch die Standardabweichungen) der Originalreihe Z_t und der verschobenen Reihe Z_{t-k} kaum unterscheiden. Wir setzen deshalb in den nachfolgenden Definitionsgleichungen wie allgemein üblich sowohl für Z_t als auch für Z_{t-k} das arithmetische Mittel der gesamten Reihe ein¹¹.

Während in der Querschnittanalyse die Standardabweichungen zweier Variablen, s_x und s_y , normalerweise unterschiedlich groß sind, sind sie im Falle der Zeitreihen Z_t und Z_{t-k} als gleich anzusehen, denn es handelt sich ja um ein und dieselbe Zeitreihe (mit einem variablen Stichprobenumfang). Im Nenner des Auto-Korrelationskoeffizienten steht folglich nicht das Produkt zweier Standardabweichungen, sondern die Varianz von Z_t :

$$r_k = \frac{\frac{1}{n-k} \sum_{t=k+1}^n (z_t - \bar{z}_t)(z_{t-k} - \bar{z}_t)}{\frac{1}{n} \sum_{t=1}^n (z_t - \bar{z}_t)^2}, \quad k > 0 \quad (2.1.6)$$

Wie in der Querschnittanalyse kann er Werte zwischen -1 und +1 annehmen. Bekanntlich ist der Korrelationskoeffizient (anders als der Regressionskoeffizient) ein symmetrisches Maß, das von der Kausalrichtung ($x \rightarrow y$) oder ($y \rightarrow x$) unabhängig ist. Das gilt auch für seine Anwendung auf Zeitreihendaten. Die Stärke der Beziehung zwischen $(z_{t-k} \rightarrow z_t) \Rightarrow (z_t \rightarrow z_{t-k})$ ist die gleiche wie die Stärke der reziproken Beziehung $(z_t \rightarrow z_{t-k})$, weshalb man sich die Korrelationsrechnung für die *Lags* $k < 0$ sparen kann (anders als bei den unten dargestellten Kreuzkorrelationsfunktionen, die substantiell unterschiedliche Variablen betreffen). Will man dennoch auch für die Autokorrelation negative *Lags* zulassen, ergibt sich die Definitionsgleichung

$$r_k = \frac{\frac{1}{n-|k|} \sum_{t=1}^{n-|k|} (z_t - \bar{z}_t)(z_{t+k} - \bar{z}_t)}{\frac{1}{n} \sum_{t=1}^n (z_t - \bar{z}_t)^2}, \quad k = \dots -2, -1, 0, 1, 2, \dots \quad (2.1.7)$$

Sie schließt auch den Fall $k=0$ ein, wobei $r_0=1$ und $r_k=r_{-k}$ gilt. Betrachtet man den *Lag* k nicht als fixiert, sondern als variable Größe: $k = 1, 2, \dots, K$, lassen sich die einzelnen Auto-Korrelationskoeffizienten als Funktion von k schreiben. Abb. 2.1.2 gibt eine graphische Darstellung („Korrelogramm“) der Autokorrelationsfunktion (AKF) der ersten 40 Werte der CDU-Präferenzen aus Abb. 1.1.4 wieder.

¹¹ Häufig wird auch der Divisor nicht mit $n-k$, sondern mit n angegeben. Da die Division durch n jedoch nur im Kontext modelltheoretischer Überlegungen begründet werden kann (sie verbessert die Eigenschaften entsprechender Schätzgrößen, s. Schlittgen/Streitberg 2001: 5), bleiben wir in diesem Kapitel bei der deskriptiv plausibleren Größe $n-k$.

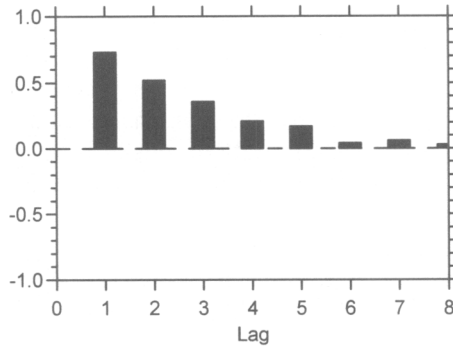


Abb. 2.1.2 Beispiel-Korrelogramm: AKF der CDU-Präferenzen, 1/1950 bis 4/1954

Wenn die untersuchte Zeitreihe saisonale oder sonstige zyklische Komponenten aufweist, nimmt ihre Autokorrelationsfunktion eine wellenförmige Gestalt an. Dies zeigt Abb. 2.1.3 für die Arbeitslosenrate der Bundesrepublik Deutschland von Januar 1959 bis Dezember 1964.

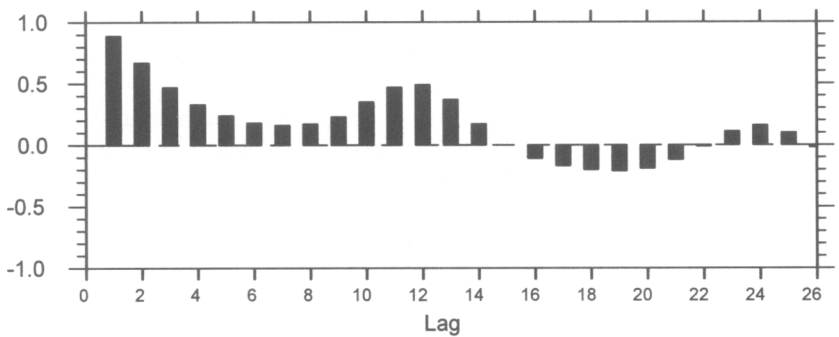


Abb. 2.1.3 AKF der Arbeitslosenrate, 1959-1964

Während dieser Teilperiode bleibt das Niveau der Arbeitslosenrate einigermaßen stabil (s. Abb. 1.1.2). Da die Zeitreihe monatliche Beobachtungswerte enthält, treten relativ hohe Korrelationskoeffizienten bei $k = 12$ und $k = 24$ auf. Bei Quartalsdaten würde die saisonale Periodizität 4 Zeit-Intervalle umfassen.

2.2 Charakterisierung des Zusammenhangs zweier Zeitreihen: die Kreuzkorrelationsfunktion

Während die Autokorrelationsfunktion (AKF) die interne Abhängigkeitsstruktur einer Zeitreihe darstellt, gibt die Kreuzkorrelationsfunktion (KKF) an, wie stark der (korrelative) Zusammenhang zwischen zwei Zeitreihen ist. Obwohl dieses Kapitel der Beschreibung einzelner Zeitreihen gewidmet ist, ist es angebracht, die KKF als empirische Kennzahl für den Zusammenhang zweier Zeitreihen, X_t und Y_t , schon jetzt darzustellen, damit die Strukturähnlichkeiten zwischen AKF und KKF deutlich werden. Wir bleiben aber weiterhin im Kontext beschreibender Analyse. Auch die KKF ist im Sinne des Produkt-Moment-Korrelations-Koeffizienten von Pearson definiert, mit ähnlichen Modifikationen wie bei der (univariaten) Autokorrelationsfunktion. Wir betrachten auch hier zunächst wieder den Ausdruck für die (Kreuz-) Kovarianz. Wenn man lediglich ermitteln will, in welchem Ausmaß die *simultan* gemessenen Werte der beiden Zeitreihen miteinander kovariieren, ergibt sich gegenüber der aus der Querschnittanalyse vertrauten Definitionsformel keine wesentliche Änderung; es wird nur der Fallindex i durch den Zeitindex t ersetzt:

$$\text{cov}(y,x) = \frac{1}{n} \sum_{t=1}^{t=n} (x_t - \bar{x})(y_t - \bar{y}) \quad (2.2.1)$$

Anders als in der Querschnittanalyse kann man in der Zeitreihenanalyse aber nicht nur die Kovarianz der simultan gemessenen Daten zweier Variablen, X_t und Y_t , ermitteln, sondern auch prüfen, wie stark Zeitreihen miteinander kovariieren, die um k Meßintervalle ($k > 0$) gegeneinander verschoben sind. Falls X_t die „führende“ Variable ist, die die Y -Reihe gleichsam hinter sich herzieht, wird aus Gleichung (2.2.1) die Definitionsformel

$$\text{cov}(y,x) = \frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x})(y_{t+k} - \bar{y}) \quad , \quad k > 0 \quad (2.2.2)$$

Den gleichen Sachverhalt können wir ausdrücken, indem wir den Zeitindex leicht modifizieren:

$$\text{cov}(y,x) = \frac{1}{n-k} \sum_{t=k+1}^n (x_{t-k} - \bar{x})(y_t - \bar{y}) \quad , \quad k > 0 \quad (2.2.3)$$

Dies ist die übliche Schreibweise für den Fall einer asymmetrischen Beziehung ($x - y$). Der zeitliche Abstand, die Verzögerungsspanne k , zwischen den beiden Variablen ist unverändert geblieben. In einem weiteren Schritt können wir den *Lag* k als variable Größe, als Laufindex behandeln. Auf diese Weise kann man die Stärke der einzelnen Kreuz-Kovarianzen als Funktion der Verzögerungsspannen $k = 0, 1, 2, \dots, K$ schreiben. Zur Berechnung der Kreuz-Kovarianzfunktion benötigt man

also nicht nur die zeitlich simultanen Datenpaare $(x_1; y_1)$, $(x_2; y_2)$, $(x_3; y_3)$ usw., sondern auch die Datenpaare $(x_{t-k}; y_t)$, die bei einem nicht fixierten Anfangszeitpunkt identisch sind mit den Paaren $(x_t; y_{t-k})$. Wir wollen das anhand zweier Zeitreihen verdeutlichen. Die erste Zeitreihe, X_t , generieren wir als eine Folge von $n = 250$ normalverteilten Zufallszahlen mit dem arithmetischen Mittel $\bar{x} = 5,0$ und der Standardabweichung $s_x = 2,0$. Die zweite Zeitreihe, Y_t , generieren wir als Funktion von X_{t-k} ($t = 6, 7, \dots, 250$; $k = 1, 2, \dots, 5$) und einem Zufallsfehler $e_t \sim N(0, 1)$ gemäß der folgenden Gleichung:

$$y_t = 0,9x_t + 0,8x_{t-1} + 0,65x_{t-2} + 0,43x_{t-3} + 0,20x_{t-4} + 0,05x_{t-5} + e_t \quad (2.2.4)$$

Die Tabelle 2.2.1 stellt für $t = 51, \dots, 250$ die jeweils ersten 15 Werte dieser beiden Zeitreihen einander gegenüber.¹²

Tab. 2.2.1: Ausdruck der ersten 15 Werte ab $t = 51$ der beiden generierten Datenreihen

t	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
x	6,9	3,0	6,5	4,1	4,7	6,2	6,0	3,1	4,6	2,9	4,7	5,8	1,2	4,0	4,6
y	20,2	14,8	17,9	14,7	14,4	16,5	17,3	13,4	14,0	10,9	12,1	14,2	8,3	10,4	11,4

Um zu ermitteln, wie stark die beiden Reihen korrelieren, wenn Y gegenüber X bspw. um $k = 2$ Intervalle nach vorne verschoben wird, sind die Datenpaare $(y = 17,9; x = 6,9)$, $(y = 14,7; x = 3,0)$ usw. zu bilden. Für $k = 7$ ergeben sich demgemäß folgende Datenpaare: $(y = 13,4; x = 6,9)$, $(y = 14,0; x = 3,0)$ usw.

Abb. 2.2.1a und 2.2.1b stellen die Streudiagramme für Datenpaare mit der Verzögerung $k=2$ und $k=7$ einander gegenüber.

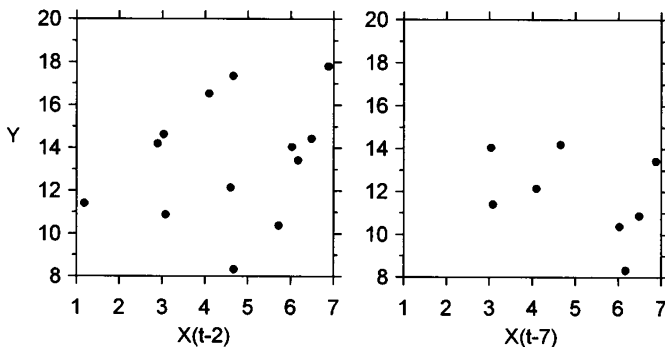


Abb. 2.2.1 a u. b: Streudiagramme aus den simulierten Zeitreihen, s. Tab.2.2.1

¹² Bei der Simulation von Modellen mit Hilfe von Computer-Algorithmen empfiehlt es sich, eine längere Datenreihe zu generieren und die Anfangswerte (bspw. wie hier: die ersten 50) zu eliminieren. Auf diese Weise vermeidet man mögliche Irregularitäten, die durch rekursive Berechnungen entstehen können.

Im ersten Diagramm (wegen $k = 2$ stehen 13 der 15 Wertepaare zur Verfügung), ist ein positiver Zusammenhang erkennbar, während im zweiten Diagramm (bei $k = 7$ bleiben 8 verwertbare „Fälle“ übrig) allenfalls ein minimaler Zusammenhang sichtbar ist. Das war auch zu erwarten, da Y_t als Funktion von X_{t-k} , $k = 1, 2, \dots, 5$, nicht aber als Funktion von $X_{t,7}$ generiert wurde. Da aber nur eine kleine Stichprobe betrachtet wird, ist mit relativ großen zufallsgenerierten Abweichungen der empirischen von den theoretisch erwartbaren Werten zu rechnen.

Anders als in unserem Beispiel mit den simulierten Zeitreihen könnte es in der Forschungspraxis unklar sein, welche der beiden Zeitreihen, X_t oder Y_t , die führende Variable (der *leading indicator*) ist. In diesem Falle wäre es sinnvoll, nicht nur die Relation $X_{t-k} \rightarrow Y_t$, sondern auch die Relation $Y_{t-k} \rightarrow X_t$ zu untersuchen. Will man beide Blickrichtungen auf einer einzigen Zeitachse darstellen, muß man negative Verzögerungsintervalle $k = -1, -2, \dots$ einführen. Aus $(X_{t-k} \rightarrow Y_t)$ wird im Falle von $k < 0$ $(X_{t-k} \rightarrow Y_t) \Rightarrow (Y_{t-k} \rightarrow X_t)$, d. h., die Reihe X folgt der Reihe Y . Abb. 2.2.2 verdeutlicht diese Darstellungsweise

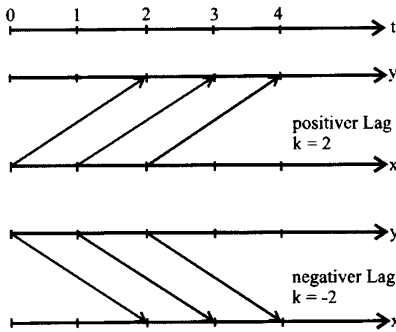


Abb. 2.2.2 Schema der Kreuzkorrelationen

Will man neben den positiven auch negative *Lags* (also reziproke Beziehungen zwischen den beiden Variablen) mit berücksichtigen, kann man dies durch eine Modifikation der Gleichung (2.2.3) wie folgt ausdrücken:

$$\text{cov}(y, x) = \frac{1}{n - |k|} \sum_{t=1}^{n-|k|} (x_t - \bar{x})(y_{t+k} - \bar{y}), \quad k = \dots, -2, -1, 0, 1, 2, \dots \quad (2.2.5)$$

Die Kreuzkorrelationsfunktion ergibt sich definitionsgemäß, indem man die Kreuzkovarianzfunktion durch das Produkt der Standardabweichungen der beteiligten Zeitreihen dividiert:

$$r_{yx}(k) = \frac{\frac{1}{n-|k|} \sum_{t=1}^{n-|k|} (x_t - \bar{x})(y_{t+k} - \bar{y})}{\frac{1}{n} \sqrt{\sum_{t=1}^n (x_t - \bar{x})^2} \frac{1}{n} \sqrt{\sum_{t=1}^n (y_t - \bar{y})^2}}, \quad k = \dots, -2, -1, 0, 1, 2, \dots \quad (2.2.6)$$

Auch für diese Korrelationskoeffizienten gilt $-1 \leq r_k \leq 1$. Im Unterschied zur Autokorrelationsfunktion verläuft die KKF für positive und negative k nicht symmetrisch. Das zeigt sich deutlich in der graphischen Darstellung der KKF, die sich für die Zeitreihen ergibt, die nach Gleichung (2.2.4) generiert wurden (Abb. 2.2.3):

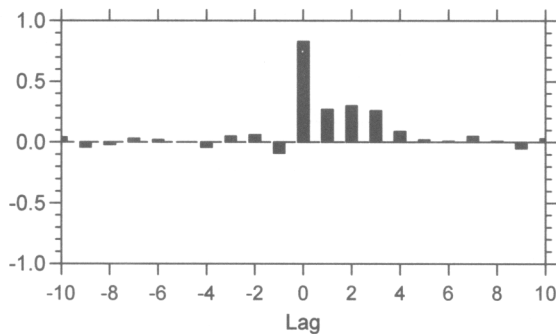


Abb. 2.2.3 KKF zu den simulierten Reihen aus Gleichung (2.2.4)

Nur auf der Seite positiver *Lags*, wenn X (wie im simulierten Modell vorgegeben) als führende Variable auftritt, zeigen sich Korrelationskoeffizienten, die deutlich von Null verschieden sind. Das abfallende Muster der vorgegebenen Regressionsgewichte wird im Muster der Kreuzkorrelationskoeffizienten ebenfalls sichtbar, aber nur „so ungefähr“. Die Korrelationskoeffizienten folgen nicht dem exponentiellen Muster der vorgegebenen Regressionskoeffizienten, die *Lag*-Spanne von 4 oder 5 Intervallen ist dagegen halbwegs erkennbar. Es liegt somit nahe, die empirische KKF als Erkennungsinstrument einzusetzen, wenn man eine eventuell bestehende strukturelle Beziehung zwischen zwei beobachteten Zeitreihen untersuchen will. Soll die KKF als Sonde für die Entdeckung kausaler Mechanismen eingesetzt werden, ist allerdings Vorsicht geboten, vor allem weil die Größe der Kreuzkorrelationskoeffizienten nicht nur von der Stärke der Kausalbeziehung zwischen den beiden Variablen, sondern auch von der Stärke der Autokorrelation der bedingenden Variablen abhängt. Hinzu kommen Verzerrungseffekte, die durch die Endlichkeit der beobachteten Reihen und den Einfluß von nicht erfaßten oder zufallsbedingten Störgrößen hervorgerufen werden. Diese Thematik geht freilich über den Kontext der beschreibenden Analyse hinaus und wird deshalb erst in Kap. 5 ausführlicher erörtert.

2.3 Komponentenzerlegung

2.3.1 Allgemeines

Wir haben im vorigen Abschnitt darauf hingewiesen, daß empirische Kennzahlen wie arithmetisches Mittel und Autokorrelationskoeffizienten nur aussagekräftig sind, wenn sie sich auf eine Zeitreihe beziehen, die „stationär“ ist, also so verläuft, daß sich für verschiedene Teilabschnitte Kennzahlen ungefähr gleicher Größe berechnen lassen. Viele Zeitreihen, die Historiker und Sozialwissenschaftler interessieren, verlaufen jedoch anders: sie weisen langfristig abfallende oder ansteigende oder wechselnde Trendrichtungen auf; und gelegentlich werden diese Trendverläufe auch noch von zyklischen Schwankungen überlagert. Zum Beispiel lassen die in Abb. 1.1.4 wiedergegebenen Parteipräferenzen wechselnde Trendverläufe erkennen. Das gilt auch für die Arbeitslosenrate (Abb. 1.1.2), die zudem von saisonalen Zyklen geprägt ist. SozialwissenschaftlerInnen verbinden mit solchen Beobachtungen Vorstellungen über eine sich langfristig entfaltende Dynamik, deren „Ursache“ in bestimmten Merkmalen eines Sozialsystems vermutet wird. Karl Marx z. B. prognostizierte, in kapitalistischen Systemen würden die Profitraten stetig fallen; Kondratieff und andere postulierten, in marktwirtschaftlichen Systemen entfalte sich das ökonomische Wachstum in „langen Wellen“; Soziologen sprechen von langfristigen Trends zunehmender funktionaler Differenzierung, zunehmender Rationalisierung, zunehmender Individualisierung usw. Die Aufdeckung der kausalen Mechanismen, die solchen Entwicklungstendenzen unterliegen, erfordert eine enge Verbindung zwischen theoretischer Reflexion und empirischer Analyse. Wir haben schon in Kapitelabschn. 1.2 anhand einiger Beispiele erläutert, wie notwendig es ist, beobachtete Zeitreihen in unterschiedliche „Komponenten“ zu zerlegen, die unterschiedliche zeitliche Dynamiken aufweisen und somit unterschiedlichen Wirkungsmechanismen zuzuordnen sind. Wie aber soll man Daten, Werte, die man beobachtet hat, in Bestandteile zerlegen, die man nicht direkt beobachten kann? Offensichtlich ist so etwas nur auf der Basis bestimmter Modellannahmen möglich, in denen die einzelnen Komponenten - wie „Trend“ und „Saison“ - formal definiert und die Art ihrer Verknüpfung festgelegt wird. Die empirische Identifikation unbeobachtbarer Komponenten ist unweigerlich mit besonderen Unsicherheiten behaftet. Ein probates Mittel, Unsicherheiten zu berücksichtigen, ist der Einbau stochastischer Elemente in die entsprechenden Modelle. In diesem Kapitel wollen wir jedoch zunächst einen traditionellen Ansatz der deskriptiven Komponentenzerlegung vorstellen, der ohne inferenzstatistische Überlegungen auskommt und auch heute noch häufig praktiziert wird, vermutlich häufiger als zufallstheoretisch fundierte Alternativen. Wir werden an verschiedenen Stellen auf diesbezügliche Unterschiede hinweisen. Der deskriptiv-deterministische Ansatz ist anschaulicher und liefert Basis-Konzepte, die auch in den stochastischen Ansätzen benötigt werden.

In den einfachsten Modellen unterscheidet man nur drei Komponenten, die man als additiv verknüpft betrachtet: Trend, Saison und Rest (bzw. „irreguläre“ Komponente). In erweiterten Modellen fügt man noch einen längerfristigen, nicht saisonal bedingten Zyklus hinzu, bspw. in der Wirtschaftsgeschichte die schon erwähnten „Langen Wellen“ der ökonomischen Entwicklung. Das Modell kann zudem dadurch komplexer gestaltet werden, daß man die Komponenten nicht additiv, sondern multiplikativ verknüpft oder gemischt additive-multiplikative Modelle konstruiert. Bevor wir uns mit der formalen Struktur der Modelle beschäftigen, sollen die Konzepte für die einzelnen Komponenten kurz erläutert werden.

2.3.2 Konzepterläuterung: Trend, Zyklus, Saison und irreguläre Komponente

Unter dem **Trend** versteht man „eine langfristige systematische Veränderung des mittleren Niveaus der Zeitreihe“ (Schlittgen/Streitberg 1989: 9). Eine präzise formale Definition des Trends seitens der Statistik liegt nicht vor. Weder ist allgemein bestimmt, auf welchen Zeitraum sich die „Langfristigkeit“ der Veränderung bezieht, noch ist klar festgelegt, welche Art von „Systematik“ ihren Verlauf kennzeichnen soll. Hier ist die statistische Analyse auf entsprechende Vorgaben der Substanzwissenschaften angewiesen, die aber oft selbst nicht in der Lage sind, präzise Definitionen zu liefern - wie z. B. die Diskussion über „lange Wellen“ in der Ökonomie gezeigt hat (s. Metz 1988; 1995). Keineswegs ist mit dem „Trend“ in jedem Fall eine kontinuierliche Zu- oder Abnahme der Werte gemeint, die linear zur Zeitachse verläuft. Die Begriffsbestimmung ist innerhalb der Statistik mit gutem Grund so vage gehalten, weil nur am konkreten Gegenstand und im Hinblick auf einen bestimmten Beobachtungszeitraum zu entscheiden ist, was als eine langfristige Entwicklungstendenz betrachtet werden soll. Was im Ablauf von zehn Jahren wie ein linearer Trend aussieht, mag sich, über ein Jahrhundert betrachtet, als Teil der ansteigenden oder abfallenden Linie einer langfristigen zyklischen Schwankung erweisen. Wie auch immer der Trend in einer konkreten Analyse festgelegt wird, formal ist er als eine Funktion der Zeit oder als gewichteter Durchschnitt mehrerer Beobachtungswerte innerhalb eines bestimmten Zeitraums zu betrachten (s. unten). Das einfachste Modell eines linear verlaufenden Trends läßt sich in Form einer Gleichung schreiben, in der der Zeitindex t als unabhängige Variable auftritt.

$$M_t = a + bt \quad , \quad t = 0, 1, 2, \dots \quad (2.3.1)$$

Das Kürzel M_t für den zeitabhängigen Trendwert erinnert daran, daß es sich beim Trend um einen sich verändernden (zeitbedingten) Mittelwert handelt; der Buchstabe a bezeichnet den Ordinatenabschnitt, den Trendwert zum Zeitpunkt $t = 0$; der Buchstabe b (Steigungskoeffizient) gibt den Trendanstieg pro Zeiteinheit an.

Diese und andere Formen des Trendverlaufs werden ausführlich in nachfolgenden Abschnitten behandelt.

Mit dem Trendbegriff ist die Vorstellung verbunden, daß die „Zeit“ als Variable kausale Größen vertritt, die nicht explizit erfaßt wurden (weil sie nicht genau bekannt sind oder weil für sie keine Daten vorliegen). So können z. B. nicht erfaßte demographische Verschiebungen den Trend sowohl in der Beitragsentwicklung bei der gesetzlichen Altersversorgung als auch in den modalen politischen Einstellungen mit bestimmen. In den Sozialwissenschaften wird auch häufig das Lebensalter als Substitut für nicht aufgeklärte Kausalfaktoren verwendet.

Die längerfristige **zyklische Komponente** ist, wie schon oben angedeutet, sowohl konzeptuell als auch empirisch oft nur schwer vom (nicht-linearen) Trend zu unterscheiden. Hierbei spielt die Länge der Beobachtungsperiode eine besonders wichtige Rolle. Zyklen sind dadurch definiert, daß sich ein bestimmtes Verlaufsmuster periodisch wiederholt. Je häufiger das innerhalb der Beobachtungsperiode geschieht, um so eher werden sie identifizierbar. Formal läßt sich ein deterministischer Zyklus, C_t , wie folgt definieren:

$$C_{t+p} = C_t, \quad t=1,2,\dots \quad (2.3.2)$$

wobei „p“ die Periode angibt, also die Zahl der Beobachtungen, die angehäuft werden müssen, bevor sich ein bestimmter Wert wiederholt. Ihrem Begriff nach sind Zyklen nicht auf wellenförmige Verlaufsmuster beschränkt, sondern können auch eine gezackte Figur (wie ein Sägeblatt zum Beispiel) darstellen. Entscheidend ist nur die periodische Wiederholung des Musters. Gelegentlich faßt man sie gemeinsam mit dem Trend unter dem Terminus *Glatte Komponente* zusammen.

Langfristige Zyklen spielen vor allem in den Wirtschaftswissenschaften und der Wirtschaftsgeschichte eine wichtige Rolle. In der allgemeinen Geschichtswissenschaft und in der Soziologie sind zwar ebenfalls zahlreiche „Zyklentheorien“ entwickelt worden (s. z. B. Sorokin 1927; Namenwirth/Weber 1987; Bornschiefer 1988), mathematisierte Zyklen-Modelle sind in diesem Bereich aber selten (s. Thome 1994b; Rahlf 1996a). Formale Analysen über zyklische Verlaufsformen bedienen sich vor allem der Spektralanalyse und der Theorie linearer Filter (s. Schlittgen/Streitberg 2001; Metz/Stier 1991). Dabei handelt es sich um mathematisch anspruchsvolle Methoden, deren Darstellung den Rahmen dieses Einführungstextes sprengen würde. Wir hoffen aber, wie in der Einleitung bereits erwähnt, mit der Periodogrammanalyse in Abschn. 2.4 den Zugang zu diesen Methoden zu erleichtern. Außerdem werden wir in Kapitel 3 zeigen, wie stochastische Zyklen (also zyklische Bewegungen mit nicht-konstanter Periode und veränderbarer Amplitude) im Rahmen der Box/Jenkins-Methode modelliert werden können. Im Rahmen der klassischen Komponentenmodelle werden wir Zyklen nur in der speziellen Form der saisonalen Schwankungen behandeln.

Auch hinter der **Saison** als jahreszeitlicher Schwankungskomponente verbergen sich nicht explizierte Kausalfaktoren wie Temperaturunterschiede, biologische Abläufe, regelmäßige Feiertage und Ferienzeiten, Budget-Terminierungen. Sie sind dadurch charakterisiert, daß sie im Rhythmus der Jahreszeiten variieren, wobei sich

diese Variationen - bei der hier praktizierten deterministischen Betrachtungsweise - unverändert jedes Jahr wiederholen. Dadurch führen sie ein zyklisches Verlaufsmuster in die Zeitreihe ein. Die Arbeitslosenzahlen liefern hierzu ein besonders prominentes Beispiel. Die Periode P in Gleichung (2.3.2) beträgt dabei genau ein Jahr; bei Monatsdaten gilt also $P = 12$, bei Vierteljahresdaten $P = 4$.

Unter der **Restkomponente** (Residualkomponente) faßt man die nicht aus der Trend-, Zyklus- oder Saisonentwicklung erklärbaren Varianzanteile einer Zeitreihe zusammen. In der Literatur werden sie auch häufig als „Störungen“ oder als „irreguläre“ Komponente bezeichnet, der man den Status einer Zufallsvariablen zuspricht (ohne sie unbedingt als solche zu modellieren). Mit dem Terminus der Störgröße wird suggeriert, daß es sich hierbei um eine Komponente handeln soll, die im Gegensatz zu Trend und Saison nicht „substanzwissenschaftlich“ erklärbar ist. Folglich wünschen sich die Analytiker Störgrößen, die über alle Meßzeitpunkte unabhängig voneinander sind („weißes Rauschen“), also keine Autokovarianzen aufweisen. In diesem Falle nämlich kann die Komponentenzerlegung in Form einer Regressionsanalyse durchgeführt werden, bei der die für die Trend- und Saisonkomponente ermittelten Größen als Parameterschätzer mit bestimmten statistischen Optimalitätseigenschaften aufgefaßt werden können. Solange wir uns nur im Kontext der beschreibenden Analyse bewegen, brauchen wir hinsichtlich der Residualkomponente allerdings keine speziellen Verteilungsannahmen einzuführen. Ökonometriker führen oft noch ein weiteres Element ein, die sog. **Konjunkturkomponente**, die eine mehrjährige, nicht unbedingt regelmäßige Schwankung darstellt. Gelegentlich wird die Konjunkturkomponente mit dem Trend zur „glatten“ Komponente oder mit der Saison zur „zyklischen“ Komponente zusammengefaßt. Wir lassen sie im folgenden außer Betracht bzw. subsumieren sie unter die Trendkomponente.

2.3.3 Additive und multiplikative Komponentenmodelle

In der Praxis geht man meistens davon aus, daß sich die Komponenten der Zeitreihe additiv zusammenfügen:

$$\begin{aligned} \text{Zeitreihe} &= \text{Trend} + \text{Saison} + \text{Rest} \\ Z_t &= M_t + S_t + R_t \end{aligned} \quad (2.3.3)$$

In vielen Fällen ist es jedoch realistischer, davon auszugehen, daß die saisonale Schwankung und die Residuen mit dem Trend wachsen, was formal durch eine multiplikative Verknüpfung der Komponenten dargestellt werden kann. (Es sei daran erinnert, daß in der Regressionsanalyse Interaktionen zwischen Variablen mit Hilfe multiplikativer Terme dargestellt werden können.)

$$Z_t = M_t \cdot S_t \cdot R_t \quad (2.3.4)$$

Solche Gleichungen lassen sich durch Logarithmieren in eine additive Form überführen:

$$\log Z_t = \log M_t + \log S_t + \log R_t \quad (2.3.5)$$

Vorstellbar ist allerdings auch, daß weder ein rein additives noch ein rein multiplikatives Modell angemessen sind. So könnten z. B. Trend und Saison miteinander „interagieren“, die Störgrößen jedoch unabhängig von diesen Komponenten bleiben:

$$Z_t = M_t \cdot S_t + R_t \quad (2.3.6)$$

Derartige Gleichungen lassen sich nicht durch Logarithmieren in eine additive Form überführen. Hier können sog. Box/Cox-Transformationen (Box/Cox 1964) weiterhelfen, die wir in Abschnitt 2.3.3.1 erläutern. Zur Bestimmung der Komponenten von „Mischmodellen“ bieten sich auch Iterationsverfahren an, die wir hier aber nicht behandeln wollen (s. z. B. Durbin/Murphy 1975).

Die Analyse geht nun im Prinzip so vor sich, daß aus den (evtl. transformierten) Beobachtungswerten die einzelnen Komponenten sukzessive isoliert und eliminiert werden. Im ersten Schritt wird für jeden Zeitpunkt die Trendkomponente bestimmt und vom jeweiligen Beobachtungswert abgezogen. Im zweiten Schritt wird auf der Basis der trendbereinigten Daten die Saisonkomponente bestimmt und ebenfalls subtrahiert, so daß die Restkomponente übrig bleibt. Dieses subtraktive Vorgehen setzt voraus, daß die einzelnen Komponenten additiv verknüpft sind, was - wie bereits angedeutet - u. U. erst nach einer Transformation der Daten der Fall ist. Bei logarithmierten oder auf andere Weise transformierten Daten müßten die Ergebnisse möglichst auf die untransformierten Größen „zurückgerechnet“ werden (s. unten). Die Residuen können schließlich auf Unabhängigkeit und andere statistisch relevante Eigenschaften getestet werden, wie das aus der allgemeinen Regressionsanalyse bekannt ist (s. Schlittgen/Streitberg 2001: 17 ff.).

Das stufenweise Vorgehen bedeutet, daß jeder nachfolgende Schritt von der Qualität des vorangegangenen abhängig ist. Dabei kann es passieren, daß die betreffende Komponente nicht vollständig oder auf eine Weise erfaßt wird, die artifizielle Strukturelemente in die verbleibende(n) Komponente(n) einschleust. Schlittgen/Streitberg (2001: 12) weisen darauf hin, daß „das Komponentenmodell unbestimmt (ist), solange nicht die einzelnen Komponenten präziser durch Modelle spezifiziert werden. Verschiedene Modellansätze können dabei sehr unterschiedliche Ergebnisse liefern, so daß bei der Interpretation veröffentlichter Reihenkomponenten (etwa saisonbereinigter Arbeitslosenziffern) stets nach dem jeweils benutzten Verfahren zu fragen ist.“

Bevor wir die einzelnen Schritte der Komponentenzzerlegung darstellen, wollen wir im nächsten Abschnitt ein Grobverfahren erläutern, mit dem festgestellt werden kann, ob ein additives oder eher ein multiplikatives oder gemischtes Modell angemessen ist.

2.3.3.1 Box/Cox-Transformationen

Der Ausdruck „multiplikatives Modell“ ist aus der Regressionsanalyse mit Querschnittdaten bekannt. Multiplikative Ausdrücke in der Regressionsgleichung repräsentieren eine vermutete „Interaktion“ zwischen unabhängigen Variablen: die Annahme, daß die Wirkung eines Regressors X_1 auf die abhängige Variable davon abhängig ist, welcher Wert eines anderen Regressors X_2 gerade realisiert ist (s. Jaccard et al. 1990; Thome 1991). Übertragen auf das Komponentenmodell der Zeitreihenanalyse bedeutet das: Ein nicht-additives Modell ist angezeigt, wenn die Größe der saisonalen oder/und der Restkomponente davon abhängig ist, welches mittlere Niveau, welcher Trendwert gegeben ist. Da sich die Größen von Saison und Restkomponente unmittelbar als stärkere oder schwächere Streuung zeigen, heißt das: ein nicht-additives Modell ist anzunehmen, wenn die Streuung mit dem Trend der Reihe kovariiert. Ob dies in einem nennenswerten Maße der Fall ist, läßt sich anhand eines einfachen Streudiagramms, des sog. *spread-versus-level-plot* (SLP), feststellen. Dazu wird die Beobachtungsperiode in K möglichst gleich lange Teilabschnitte eingeteilt.¹³ Für jeden Zeitabschnitt $k = 1, 2, \dots, K$ werden ein Streuungsmaß (Spannweite, Standardabweichung oder Interquartilsabstand) und das arithmetische Mittel (oder auch der Median) berechnet.¹⁴ Diese Kennzahlen werden sodann in ein Koordinatenkreuz eingetragen (das Streuungsmaß auf der Ordinate, der Mittelwert auf der Abszisse).

Als Beispiel kann die Reihe der monatlichen Arbeitslosendaten von Jan. 1970 bis Dez. 1982 (Abb. 2.3.1) dienen.

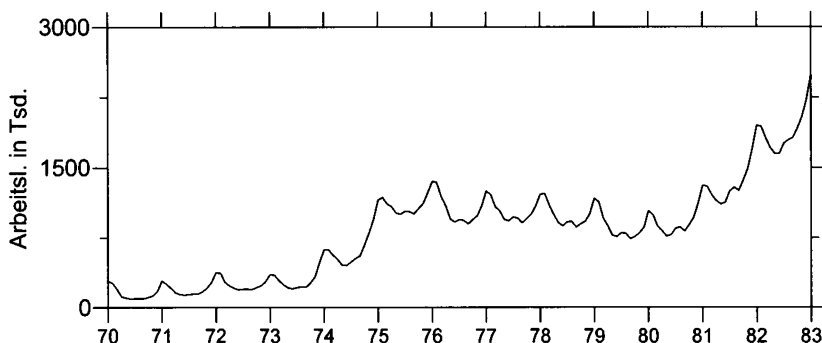


Abb. 2.3.1 Arbeitslose in Tausend von Jan. 70 bis Dez. 82 (Quelle: s. Abb. 1.1.2)

¹³ Bei Saisonalität wird die Länge der Teilabschnitte durch die saisonale Periode bestimmt, ansonsten gilt als Daumenregel eine Abschnittlänge von etwa 10 Meßzeitpunkten (s. Mohr 1980: 168).

¹⁴ Das Programmpaket SPSS bietet unter dem Kommando EXAMINE die Möglichkeit, einen „spread-versus-level-plot“ zu erhalten, in dem auf der Ordinate die jeweiligen Interquartilsabstände und auf der Abszisse die Medianwerte abgetragen werden. Beide Größen lassen sich auch logarithmieren, die Prozedur liefert dann nicht nur das Streudiagramm, sondern auch einen Schätzer für den Transformationsexponenten λ (s. unten Anm. 17).

Sie zeigt einen (nicht-linearen) ansteigenden Trendverlauf; außerdem dehnen sich die saisonalen Schwankungen parallel dazu aus. Dies legt eher ein multiplikatives als ein additives Komponentenmodell nahe, aber auch ein Mischmodell ist nicht auszuschließen. Wir wollen dies mit Hilfe eines Streudiagramms (*SLP*) näher untersuchen. Dazu teilen wir die Beobachtungsperiode in $K=13$ aufeinanderfolgende Zeitabschnitte ein, die jeweils die 12 Monate eines Jahres umfassen. Für jeden Zeitabschnitt werden die Standardabweichung (*STD*) und das arithmetische Mittel der Arbeitslosenzahlen berechnet und in den *SLP* eingetragen (s. Abb. 2.3.2). Die Koordinatenschnittpunkte für die einzelnen Jahresabschnitte sind durch Buchstaben markiert, deren alphabetische Ordnung der Abfolge der Jahre von 1970 bis 1982 entspricht.

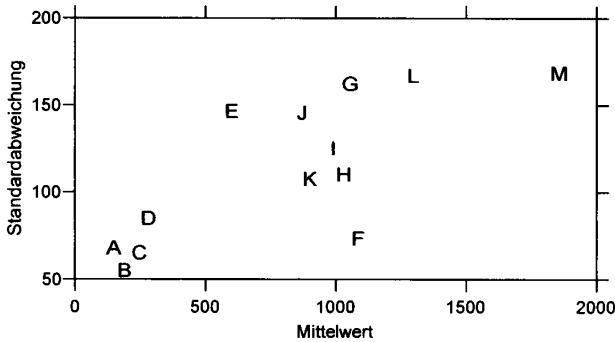


Abb. 2.3.2 *SLP* der Arbeitslosendaten

Allgemein gilt: ein rein additives Modell ist nur dann angemessen, wenn die Punkte des Streudiagramms am ehesten durch eine Gerade zu repräsentieren sind, die parallel zur Abszisse verläuft. Steigen dagegen die Ordinatenwerte gemäß einer Geraden von links unten nach rechts oben an, bedeutet dies, daß die Dispersion der Reihe (hier die saisonalen Schwankungen der Arbeitslosenzahlen) mit dem mittleren Niveau der Zeitreihe proportional zunimmt. In diesem Falle wäre eine multiplikative Verknüpfung zu wählen, die durch Logarithmieren der Werte in eine lineare überführt werden könnte. In Abb. 2.3.2 ist nicht deutlich erkennbar, ob die Steigungstendenz überhaupt geradlinig verläuft oder ob die Punkte mit zunehmendem Mittelwert nach unten von der Geraden abweichen.¹⁵ Letzteres würde ein Mischmodell anzeigen. Wir probieren zunächst ein rein multiplikatives Modell aus, logarithmieren also die Arbeitslosendaten (s. Abb. 2.3.3) und konstruieren ein neues *SLP* für die logarithmierten Daten (s. Abb. 2.3.4).

In Abb. 2.3.3 ist sofort erkennbar, daß nun die Streuung mit dem steigenden Niveau der Reihe abnimmt. Demgemäß ergibt sich im *SLP* (s. Abb. 2.3.4) eine negative Beziehung: Die Punkte streuen um eine Gerade, die von links oben nach rechts

¹⁵ Eine formale Regressionsanalyse zur Ermittlung des Steigungskoeffizienten oder zwecks Linearitätstests führt man im allgemeinen nicht durch, da die Schätzergebnisse angesichts der geringen Fallzahl und der Ausreißerphänomene zu instabil wären (s. aber unten, Anm. 17).

unten abfällt. Die Log-Transformation hat also nicht das gewünschte Ergebnis (Stabilität der Streuung) gebracht, denn dann müßten die Punkte um eine Parallele zur X-Achse variieren. Es scheint also ein Mischmodell vorzuliegen. Folglich müssen zusätzliche Möglichkeiten der Datentransformation erkundet werden.

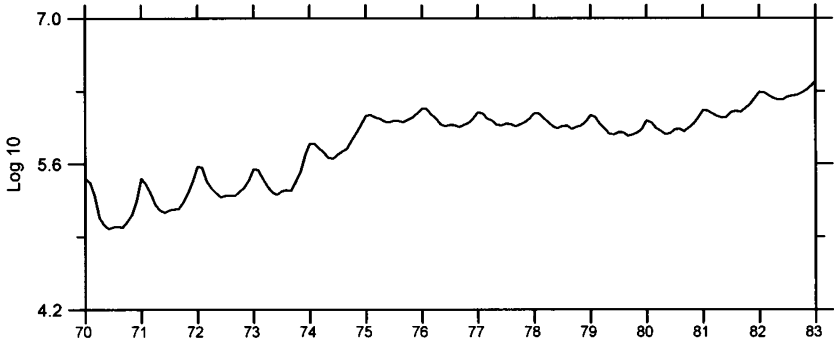


Abb. 2.3.3 *Logarithmierte Arbeitslosendaten*

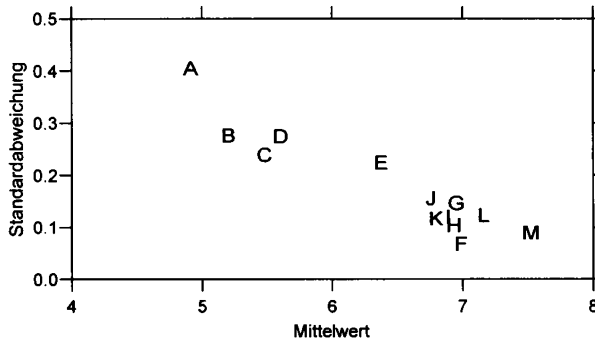


Abb. 2.3.4 *SLP der logarithmierten Arbeitslosendaten*

Box/Cox (1964) haben eine Klasse von Transformationen vorgeschlagen, die hierfür in Frage kommen.¹⁶

¹⁶ Die Box/Cox-Transformationen haben einen breiten Anwendungsbereich auch außerhalb der Zeitreihenanalyse. Box/Cox (1964) haben diese Potenzfunktion nicht spezifisch zur Varianzstabilisierung in der univariaten Zeitreihenanalyse vorgeschlagen, sondern zur Transformation heteroskedastischer Response- bzw. Fehlervariablen, um drei Zwecke gleichzeitig zu erreichen: eine lineare Beziehung zwischen Prädiktor- und Kriteriumsvariable sowie normalverteilte und homoskedastische Fehlerkomponenten. Box/Jenkins (1976) - siehe auch Jenkins (1979) - haben diesen Ansatz dann in die Zeitreihenanalyse übernommen. Zu Extensionen und Modifikationen der Transformationsformel siehe Cook/Weinberg (1982: 60 f.) sowie Emerson/Stoto (1983: 98 - 104). Zu ihrer Verwendung innerhalb der Regressionsanalyse mit Querschnittsdaten siehe z. B. Greene (1993: 329 ff.). - Dr. Herbert Stahl, seinerzeit Dozent am Institut für quantitative Methoden der Technischen Universität Berlin, hat uns als erster auf die Möglichkeit hingewiesen, Box/Cox-Transformationen auch als Instrument zur Identifikation multiplikativer oder gemischter Modelle der Komponentenzersetzung einzusetzen.

$$Z_i^* = \begin{cases} \frac{1}{\lambda}(Z_i^\lambda - 1) & , \text{für } \lambda \neq 0 \\ \ln Z_i & , \text{für } \lambda = 0 \end{cases} \quad (2.3.7)$$

Der natürliche Logarithmus von Z_i ist ein Grenzfall der Box/Cox-Transformation, der sich aus dem Grenzwert $\lim (Z_i^\lambda - 1)/\lambda = \ln Z_i$ für $\lambda \rightarrow 0$ ergibt. Im allgemeinen leistet eine rechnerisch einfachere Potenztransformation die gleichen Dienste (s. Box/Cox 1964: 214):

$$Z_i^* = \begin{cases} Z_i^\lambda & , \text{für } \lambda \neq 0 \\ \ln Z_i & , \text{für } \lambda = 0 \end{cases} \quad (2.3.8)$$

Box/Cox (1964) stellen eine formale Methode zur Schätzung von λ vor (s. auch Mohr 1980; 1983).¹⁷ Verschiedene Programmpakete (auch SPSS) bieten eine statistische Testmöglichkeit an, mit der dasjenige *Lambda* ermittelt werden kann, das die Varianz gruppierter Daten am ehesten homogenisiert. Wieso die näherungsweise Bestimmung von λ mit Hilfe des *SLP* im allgemeinen funktioniert, zeigen Schlittgen/Streitberg (2001: 102f.). Die rein multiplikativen Modelle ($\lambda = 0$) und die rein additiven Modelle ($\lambda = 1$) haben wir bereits besprochen. Abb. 2.3.5 zeigt darüber hinaus in schematischer Weise die Gestalt der Streudiagramme für diverse Mischmodelle, von denen das in der Wurzeltransformation (Fall 3 in Abb. 2.3.5) implizierte Modell in der Praxis am gebräuchlichsten ist.

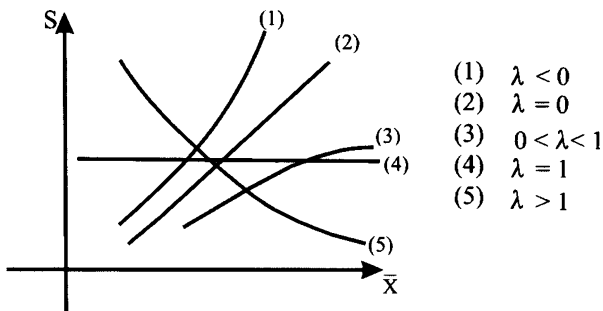


Abb. 2.3.5 Typische SLP-Muster nach Mohr (1980)

Wenn nach einer ersten Transformation die Punkte im *SLP* um eine negative Gerade bzw. Kurve streuen (wie in unserem Beispiel nach der Logarithmierung), muß *Lambda* erhöht (hier also größer Null) werden. Wenn die Punkte des Streudiagramms nach einer ersten Transformation um eine positive Gerade streuen, muß λ verringert werden. Der Versuch mit $\lambda = 0,5$ (Wurzeltransformation) bringt

¹⁷ Falls das Streudiagramm (*Spread-Level Plot*) keine markanten Ausreißer aufweist, läßt sich der Transformationsexponent wie folgt schätzen: Man regrediere linear die logarithmierten Streuungsmaße auf die logarithmierten Mittelwerte. Die Differenz (1 - Steigungskoeffizient) ist der gesuchte Schätzer (s. Mohr 1983: 186; Thome 1994a: 84 ff.; dort auch weitere Literaturhinweise).