SPRINGER TEXTS IN STATISTICS

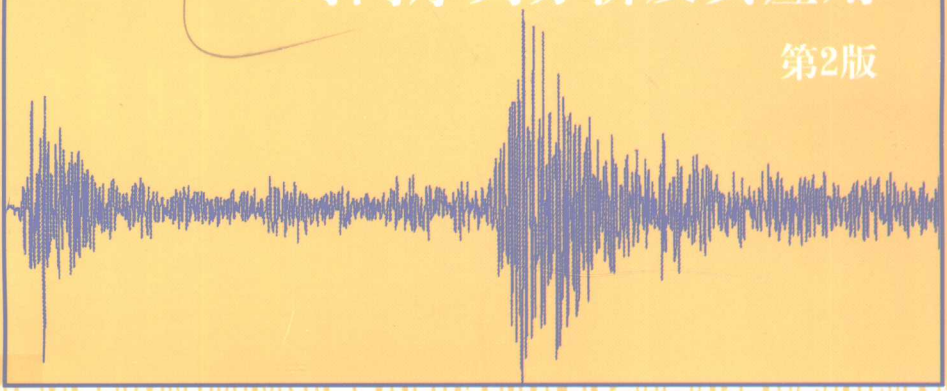
Time Series Analysis and Its Applications

With R Examples

SECOND EDITION

时间序列分析及其应用

第2版



Robert H. Shumway
David S. Stoffer

Springer

世界图书出版公司
www.wpcbj.com.cn

Robert H. Shumway
David S. Stoffer

Time Series Analysis and Its Applications

With R Examples

Second Edition

Springer

图书在版编目（CIP）数据

时间序列分析及其应用：第2版 = Time Series Analysis and Its Applications With R Examples 2nd ed. ：英文/（美）罗伯特沙姆韦（Shumway，R. H.）著. —北京：世界图书出版公司北京公司，2009.5

ISBN 978-7-5100-0438-4

Ⅰ. 时… Ⅱ. 罗… Ⅲ. 时间序列分析—教材—英文

Ⅳ. 0211.61

中国版本图书馆 CIP 数据核字（2009）第055586号

书　　名：	Time Series Analysis and Its Applications With R Examples 2nd ed.
作　　者：	Robert H. Shumway & David S. Stoffer
中 译 名：	时间序列分析及其应用 第2版
责任编辑：	高蓉　刘慧
出 版 者：	世界图书出版公司北京公司
印 刷 者：	三河国英印务有限公司
发　　行：	世界图书出版公司北京公司（北京朝内大街137号　100010）
联系电话：	010-64021602，010-64015659
电子信箱：	kjb@ wpcbj. com. cn
开　　本：	24开
印　　张：	25
版　　次：	2009年05月
版权登记：	图字：01-2009-2047
书　　号：	978-7-5100-0438-4/O · 653　　**定　　价：** 69.00元

世界图书出版公司北京公司已获得 Springer 授权在中国大陆独家重印发行

Robert H. Shumway
Department of Statistics
University of California, Davis
Davis, CA 95616
USA
rshumway@ucdavis.edu
or
shumway@wald.ucdavis.edu

David S. Stoffer
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260
USA
stoffer@pitt.edu

Library of Congress Control Number: 2005935284

ISBN-13: 978-0-387-29317-2

springer.com

To my wife, Ruth, for her support and joie de vivre, and to the memory of my thesis adviser, Solomon Kullback.
R.H.S.

To my family, who constantly remind me what is important.
D.S.S.

Preface to the Second Edition

The second edition marks a substantial change to the first edition. Perhaps the most significant change is the introduction of examples based on the freeware R package. The package, which runs on most operating systems, can be downloaded from The Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/` or any one of its mirrors. Readers who have experience with the S-PLUS® package will have no problem working with R. For novices, R installs some help manuals, and CRAN supplies links to contributed tutorials such as *R for Beginners*. In our examples, we assume the reader has downloaded and installed R and has downloaded the necessary data files. The data files can be downloaded from the website for the text, `http://www.stat.pitt.edu/stoffer/tsa2/` or any one of its mirrors. We will also provide additional code and other information of interest on the text's website. Most of the material that would be given in an introductory course on time series analysis has associated R code. Although examples are given in R, the material is not R-dependent. In courses we have given using a preliminary version of the new edition of the text, students were allowed to use any package of preference. Although most students used R (or S-PLUS), a number of them completed the course successfully using other programs such as ASTSA, MATLAB®, SAS®, and SPSS®.

Another substantial change from the first edition is that the material has been divided into smaller chapters. The introductory material is now contained in the first two chapters. The first chapter discusses the characteristics of time series, introducing the fundamental concepts of time plot, models for dependent data, auto- and cross-correlation, and their estimation. The second chapter provides a background in regression techniques for time series data. This chapter also includes the related topics of smoothing and exploratory data analysis for preprocessing nonstationary series.

In the first edition, we covered ARIMA and other special topics in the time domain in one chapter. In this edition, univariate ARIMA modeling is presented in its own chapter, Chapter 3. The material on additional time domain topics has been expanded, and moved to its own chapter, Chapter 5. The additional topics include long memory models, GARCH processes, threshold models, regression with autocorrelated errors, lagged regression, transfer function modeling, and multivariate ARMAX models. In this edition, we have removed the discussion on reduced rank models and contemporaneous models from the multivariate ARMAX section. The coverage of GARCH models has been considerably expanded in this edition. The coverage of long memory models has been consolidated, presenting time domain and frequency domain approaches in the same section. For this reason, the chapter is presented after the chapter on spectral analysis.

The chapter on spectral analysis and filtering, Chapter 4, has been expanded to include various types of spectral estimators. In particular, kernel based estimators and spectral window estimators have been included in the dis-

cussion. The chapter now includes a section on wavelets that was in another chapter in the first edition. The reader will also notice a change in notation from the previous edition.

In the first edition, topics were supplemented by theoretical sections at the end of the chapter. In this edition, we have put the theoretical topics in appendices at the end of the text. In particular, Appendix A can be used to supplement the material in the first chapter; it covers some fundamental topics in large sample theory for dependent data. The material in Appendix B includes theoretical material that expands the presentation of time domain topics, and this appendix may be used to supplement the coverage of the chapter on time series regression and the chapter on ARIMA models. Finally, Appendix C contains a theoretical basis for spectral analysis.

The remaining two chapters on state-space and dynamic linear models, Chapter 6, and on additional statistical methods in the frequency domain, Chapter 7, are comparable to the their first edition counterparts. We do mention that the section on multivariate ARMAX, which used to be in the state-space chapter, has been moved to Chapter 5. We have also removed spectral domain canonical correlation analysis and the discussion on wavelets (now in Chapter 4) that were previously in Chapter 7. The material on stochastic volatility models, now in Chapter 6, has been expanded. R programs for some Chapter 6 examples are available on the website for the text; these programs include code for the Kalman filter and smoother, maximum likelihood estimation, the EM algorithm, and fitting stochastic volatility models.

In the previous edition, we set off important definitions by highlighting phrases corresponding to the definition. We believe this practice made it difficult for readers to find important information. In this edition, we have set off definitions as numbered definitions that are presented in italics with the concept being defined in bold letters.

We thank John Kimmel, Executive Editor, Statistics, for his guidance in the preparation and production of this edition of the text. We are particularly grateful to Don Percival and Mike Keim at the University of Washington, for numerous suggestions that led to substantial improvement to the presentation. We also thank the many students and other readers who took the time to mention typographical errors and other corrections to the first edition. In particular, we appreciate the efforts of Jeongeun Kim, Sangdae Han, and Mark Gamalo at the University of Pittsburgh, and Joshua Kerr and Bo Zhou at the University of California, for providing comments on portions of the draft of this edition. Finally, we acknowledge the support of the National Science Foundation.

Robert H. Shumway
Davis, CA
David S. Stoffer
Pittsburgh, PA
August 2005

Preface to the First Edition

The goals of this book are to develop an appreciation for the richness and versatility of modern time series analysis as a tool for analyzing data, and still maintain a commitment to theoretical integrity, as exemplified by the seminal works of Brillinger (1981) and Hannan (1970) and the texts by Brockwell and Davis (1991) and Fuller (1995). The advent of more powerful computing, especially in the last three years, has provided both real data and new software that can take one considerably beyond the fitting of simple time domain models, such as have been elegantly described in the landmark work of Box and Jenkins (see Box et al., 1994). This book is designed to be useful as a text for courses in time series on several different levels and as a reference work for practitioners facing the analysis of time-correlated data in the physical, biological, and social sciences.

We believe the book will be useful as a text at both the undergraduate and graduate levels. An undergraduate course can be accessible to students with a background in regression analysis and might include Sections 1.1-1.8, 2.1-2.9, and 3.1-3.8. Similar courses have been taught at the University of California (Berkeley and Davis) in the past using the earlier book on applied time series analysis by Shumway (1988). Such a course is taken by undergraduate students in mathematics, economics, and statistics and attracts graduate students from the agricultural, biological, and environmental sciences. At the master's degree level, it can be useful to students in mathematics, environmental science, economics, statistics, and engineering by adding Sections 1.9, 2.10-2.14, 3.9, 3.10, 4.1-4.5, to those proposed above. Finally, a two-semester upper-level graduate course for mathematics, statistics and engineering graduate students can be crafted by adding selected theoretical sections from the last sections of Chapters 1, 2, and 3 for mathematics and statistics students and some advanced applications from Chapters 4 and 5. For the upper-level graduate course, we should mention that we are striving for a less rigorous level of coverage than that which is attained by Brockwell and Davis (1991), the classic entry at this level.

A useful feature of the presentation is the inclusion of data illustrating the richness of potential applications to medicine and in the biological, physical, and social sciences. We include data analysis in both the text examples and in the problem sets. All data sets are posted on the World Wide Web at the following URLs: `http://www.stat.ucdavis.edu/~shumway/tsa.html` and `http://www.stat.pitt.edu/~stoffer/tsa.html`, making them easily accessible to students and general researchers. In addition, an exploratory data analysis program written by McQuarrie and Shumway (1994) can be downloaded (as Freeware) from these websites to provide easy access to all of the techniques required for courses through the master's level.

Advances in modern computing have made multivariate techniques in the time and frequency domain, anticipated by the theoretical developments in Brillinger (1981) and Hannan (1970), routinely accessible using higher level

languages, such as MATLAB and S-PLUS. Extremely large data sets driven by periodic phenomena, such as the functional magnetic resonance imaging series or the earthquake and explosion data, can now be handled using extensions to time series of classical methods, like multivariate regression, analysis of variance, principal components, factor analysis, and discriminant or cluster analysis. Chapters 4 and 5 illustrate some of the immense potential that methods have for analyzing high-dimensional data sets.

The many practical data sets are the results of collaborations with research workers in the medical, physical, and biological sciences. Some deserve special mention as a result of the pervasive use we have made of them in the text. The predominance of applications in seismology and geophysics is joint work of the first author with Dr. Robert R. Blandford of the Center for Monitoring Research and Dr. Zoltan Der of Ensco, Inc. We have also made extensive use of the El Niño and Recruitment series contributed by Dr. Roy Mendelssohn of the National Marine Fisheries Service. In addition, Professor Nancy Day of the University of Pittsburgh provided the data used in Chapter 4 in a longitudinal analysis of the effects of prenatal smoking on growth, as well as some of the categorical sleep-state data posted on the World Wide Web. A large magnetic imaging data set that was developed during joint research on pain perception with Dr. Elizabeth Disbrow of the University of San Francisco Medical Center forms the basis for illustrating a number of multivariate techniques in Chapter 5. We are especially indebted to Professor Allan D.R. McQuarrie of the University of North Dakota, who incorporated subroutines in Shumway (1988) into ASTSA for Windows.

Finally, we are grateful to John Kimmel, Executive Editor, Statistics, for his patience, enthusiasm, and encouragement in guiding the preparation and production of this book. Three anonymous reviewers made numerous helpful comments, and Dr. Rahman Azari and Dr. Mitchell Watnik of the University of California, Davis, Division of Statistics, read portions of the draft. Any remaining errors are solely our responsibility.

Robert H. Shumway
Davis, CA
David S. Stoffer
Pittsburgh, PA
August 1999

Contents

Chapter 1

Characteristics of Time Series

1.1 Introduction

The analysis of experimental data that have been observed at different points in time leads to new and unique problems in statistical modeling and inference. The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods traditionally dependent on the assumption that these adjacent observations are independent and identically distributed. The systematic approach by which one goes about answering the mathematical and statistical questions posed by these time correlations is commonly referred to as time series analysis.

The impact of time series analysis on scientific applications can be partially documented by producing an abbreviated listing of the diverse fields in which important time series problems may arise. For example, many familiar time series occur in the field of economics, where we are continually exposed to daily stock market quotations or monthly unemployment figures. Social scientists follow populations series, such as birthrates or school enrollments. An epidemiologist might be interested in the number of influenza cases observed over some time period. In medicine, blood pressure measurements traced over time could be useful for evaluating drugs used in treating hypertension. Functional magnetic resonance imaging of brain-wave time series patterns might be used to study how the brain reacts to certain stimuli under various experimental conditions.

Many of the most intensive and sophisticated applications of time series methods have been to problems in the physical and environmental sciences. This fact accounts for the basic engineering flavor permeating the language of

time series analysis. One of the earliest recorded series is the monthly sunspot numbers studied by Schuster (1906). More modern investigations may center on whether a warming is present in global temperature measurements or whether levels of pollution may influence daily mortality in Los Angeles. The modeling of speech series is an important problem related to the efficient transmission of voice recordings. Common features in a time series characteristic known as the power spectrum are used to help computers recognize and translate speech. Geophysical time series such those produced by yearly depositions of various kinds can provide long-range proxies for temperature and rainfall. Seismic recordings can aid in mapping fault lines or in distinguishing between earthquakes and nuclear explosions.

The above series are only examples of experimental databases that can be used to illustrate the process by which classical statistical methodology can be applied in the correlated time series framework. In our view, the first step in any time series investigation always involves careful scrutiny of the recorded data plotted over time. This scrutiny often suggests the method of analysis as well as statistics that will be of use in summarizing the information in the data. Before looking more closely at the particular statistical methods, it is appropriate to mention that two separate, but not necessarily mutually exclusive, approaches to time series analysis exist, commonly identified as the time domain approach and the frequency domain approach.

The time domain approach is generally motivated by the presumption that correlation between adjacent points in time is best explained in terms of a dependence of the current value on past values. The time domain approach focuses on modeling some future value of a time series as a parametric function of the current and past values. In this scenario, we begin with linear regressions of the present value of a time series on its own past values and on the past values of other series. This modeling leads one to use the results of the time domain approach as a forecasting tool and is particularly popular with economists for this reason.

One approach, advocated in the landmark work of Box and Jenkins (1970; see also Box et al., 1994), develops a systematic class of models called autoregressive integrated moving average (ARIMA) models to handle time-correlated modeling and forecasting. The approach includes a provision for treating more than one input series through multivariate ARIMA or through transfer function modeling. The defining feature of these models is that they are multiplicative models, meaning that the observed data are assumed to result from products of factors involving differential or difference equation operators responding to a white noise input.

A more recent approach to the same problem uses additive models more familiar to statisticians. In this approach, the observed data are assumed to result from sums of series, each with a specified time series structure; for example, in economics, assume a series is generated as the sum of trend, a seasonal effect, and error. The state-space model that results is then treated by making judicious use of the celebrated Kalman filters and smoothers, developed origi-

nally for estimation and control in space applications. Two relatively complete presentations from this point of view are in Harvey (1991) and Kitagawa and Gersch (1996). Time series regression is introduced in Chapter 2, and ARIMA and related time domain models are studied in Chapter 3, with the emphasis on classical, statistical, univariate linear regression. Special topics on time domain analysis are covered in Chapter 5; these topics include modern treatments of, for example, time series with long memory and GARCH models for the analysis of volatility. The state-space model, Kalman filtering and smoothing, and related topics are developed in Chapter 5.

Conversely, the frequency domain approach assumes the primary characteristics of interest in time series analyses relate to periodic or systematic sinusoidal variations found naturally in most data. These periodic variations are often caused by biological, physical, or environmental phenomena of interest. A series of periodic shocks may influence certain areas of the brain; wind may affect vibrations on an airplane wing; sea surface temperatures caused by El Niño oscillations may affect the number of fish in the ocean. The study of periodicity extends to economics and social sciences, where one may be interested in yearly periodicities in such series as monthly unemployment or monthly birth rates.

In spectral analysis, the partition of the various kinds of periodic variation in a time series is accomplished by evaluating separately the variance associated with each periodicity of interest. This variance profile over frequency is called the power spectrum. In our view, no schism divides time domain and frequency domain methodology, although cliques are often formed that advocate primarily one or the other of the approaches to analyzing data. In many cases, the two approaches may produce similar answers for long series, but the comparative performance over short samples is better done in the time domain. In some cases, the frequency domain formulation simply provides a convenient means for carrying out what is conceptually a time domain calculation. Hopefully, this book will demonstrate that the best path to analyzing many data sets is to use the two approaches in a complementary fashion. Expositions emphasizing primarily the frequency domain approach can be found in Bloomfield (1976), Priestley (1981), or Jenkins and Watts (1968). On a more advanced level, Hannan (1970), Brillinger (1981), Brockwell and Davis (1991), and Fuller (1995) are available as theoretical sources. Our coverage of the frequency domain is given in Chapters 4 and 7.

The objective of this book is to provide a unified and reasonably complete exposition of statistical methods used in time series analysis, giving serious consideration to both the time and frequency domain approaches. Because a myriad of possible methods for analyzing any particular experimental series can exist, we have integrated real data from a number of subject fields into the exposition and have suggested methods for analyzing these data.

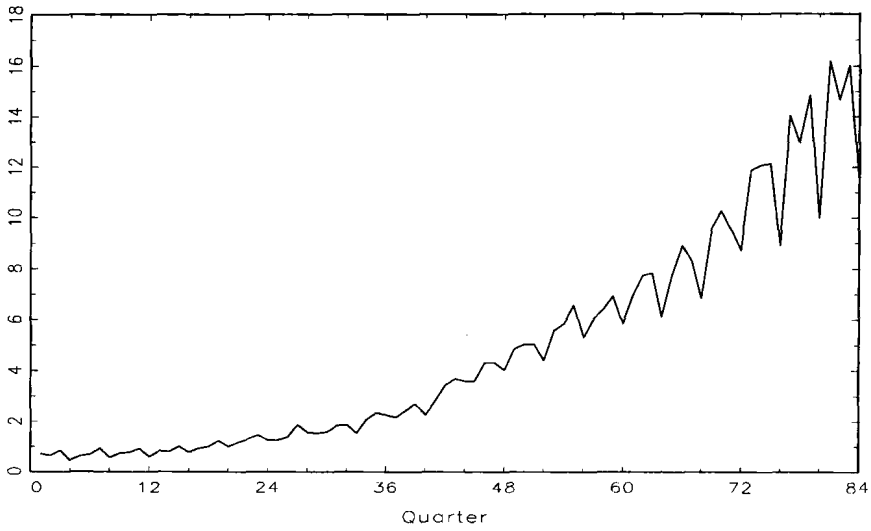


Figure 1.1 Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

1.2 The Nature of Time Series Data

Some of the problems and questions of interest to the prospective time series analyst can best be exposed by considering real experimental data taken from different subject areas. The following cases illustrate some of the common kinds of experimental time series data as well as some of the statistical questions that might be asked about such data.

Example 1.1 Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share for the U.S. company Johnson & Johnson, furnished by Professor Paul Griffin (personal communication) of the Graduate School of Management, University of California, Davis. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modeling such series begins by observing the primary patterns in the time history. In this case, note the gradually increasing underlying trend and the rather regular variation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 2 (see Problem 2.1) using regression techniques, and in Chapter 6, §6.5, using structural equation modeling.

To plot the data using the R statistical package, suppose you saved the data as `jj.dat` in the directory `mydata`. Then use the following steps to read in the data and plot the time series (the > below are prompts, you

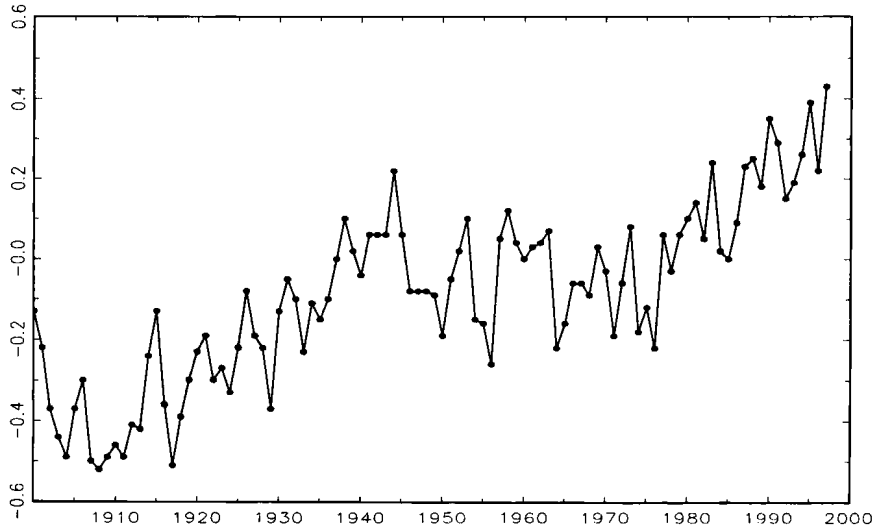


Figure 1.2 Yearly average global temperature deviations (1900–1997) in degrees centigrade.

do not type them):

```
> jj = scan("/mydata/jj.dat")   # yes forward slash
> jj=ts(jj,start=1960, frequency=4)
> plot(jj,  ylab="Quarterly Earnings per Share")
```

You can replace `scan` with `read.table` in this example.

Example 1.2 Global Warming

Consider a global temperature series record, discussed in Jones (1994) and Parker et al. (1994, 1995). The data in Figure 1.2 are a combination of land-air average temperature anomalies (from 1961-1990 average), measured in degrees centigrade, for the years 1900-1997. We note an apparent upward trend in the series that has been used as an argument for the global warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970. The question of interest for global warming proponents and opponents is whether the overall trend is natural or whether it is caused by some human-induced interface. Problem 2.8 examines 634 years of glacial sediment data that might be taken as a long-term temperature proxy. Such percentage changes in temperature do not seem to be unusual over a time period of 100 years. Again, the question of trend is of more interest than particular periodicities.