WILEY SERIES IN PROBABILITY AND STATISTICS

FIFTH EDITION

Time Series Analysis

Forecasting and Control

George E. P. Box • Gwilym M. Jenkins
Gregory C. Reinsel • Greta M. Ljung

WILEY

TIME SERIES ANALYSIS

A complete list of the titles in this series appears at the end of this volume.

TIME SERIES ANALYSIS

Forecasting and Control

Fifth Edition

GEORGE E. P. BOX
GWILYM M. JENKINS
GREGORY C. REINSEL
GRETA M. LJUNG

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Box, George E. P.
Time series analysis : forecasting and control. – Fifth edition / George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-67502-1 (cloth : alk. paper) 1. Time-series analysis. 2. Prediction theory. 3. Transfer functions. 4. Feedback control systems–Mathematical models. I. Jenkins, Gwilym M. II. Reinsel, Gregory C. III. Ljung, Greta M., 1941- IV. Title.
QA280.B67 2016
519.5'5–dc23

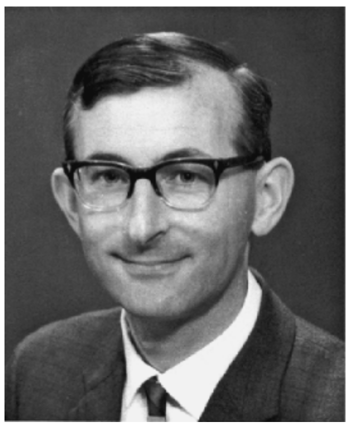
2015015492

Printed in the United States of America

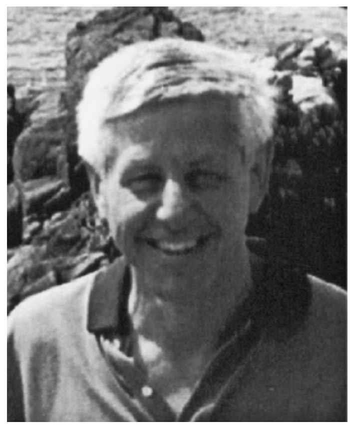
10 9 8 7 6 5 4 3 2 1

To the memory of

George E. P. Box



Gwilym M. Jenkins



Gregory C. Reinsel

CONTENTS

PREFACE TO THE FIFTH EDITION

This book describes statistical models and methods for analyzing discrete time series and presents important applications of the methodology. The models considered include the class of autoregressive integrated moving average (ARIMA) models and various extensions of these models. The properties of the models are examined and statistical methods for model specification, parameter estimation, and model checking are presented. Applications to forecasting nonseasonal as well as seasonal time series are described. Extensions of the methodology to transfer function modeling of dynamic relationships between two or more time series, modeling the effects of intervention events, multivariate time series modeling, and process control are discussed. Topics such as state-space and structural modeling, nonlinear models, long-memory models, and conditionally heteroscedastic models are also covered. The goal has been to provide a text that is practical and of value to both academicians and practitioners.

The first edition of this book appeared in 1970 and around that time there was a great upsurge in research on time series analysis and forecasting. This generated a large influx of new ideas, modifications, and improvements by many authors. For example, several new research directions began to emerge in econometrics around that time, leading to what is now known as time series econometrics. Many of these developments were reflected in the fourth edition of this book and have been further elaborated upon in this new edition.

The main goals of preparing a new edition have been to expand and update earlier material, incorporate new literature, enhance and update numerical illustrations through the use of R, and increase the number of exercises in the book. Some of the chapters in the previous edition have been reorganized. For example, Chapter 14 on multivariate time series analysis has been reorganized and expanded, placing more emphasis on vector autoregressive (VAR) models. The VAR models are by far the most widely used multivariate time series models in applied work. This edition provides an expanded treatment of these models that includes software demonstrations.

Chapter 10 has also been expanded and updated. This chapter covers selected topics in time series analysis that either extend or supplement material discussed in earlier chapters.

This includes unit roots testing, modeling of conditional heteroscedasticity, nonlinear models, and long memory models. A section of unit root testing that appeared in Chapter 7 of the previous edition has been expanded and moved to Section 10.1 in this edition. Section 10.2 deals with autoregressive conditionally heteroscedastic models, such as the ARCH and GARCH models. These models focus on the variability in a time series and are useful for modeling the volatility or variability in economic and financial series, in particular. The treatment of the ARCH and GARCH models has been expanded and several extensions have been added.

Elsewhere in the text, the exposition has been enhanced by revising, modifying, and omitting text as appropriate. Several tables have either been edited or replaced by graphs to make the presentation more effective. The number of exercises has been increased throughout the text and they now appear at the end of each chapter.

A further enhancement to this edition is the use of the statistical software R for model building and forecasting. The R package is available as a free download from the R Project for Statistical Computing at www.r-project.org. A brief description of the software is given in Appendix A1.1 of Chapter 1. Graphs generated using R now appear in many of the chapters along with R code that will help the reader reconstruct the graphs. The software is also used for numerical illustration in many of the examples in the text.

The fourth edition of this book was published by Wiley in 2008. Plans for a new edition began during the fall of 2012. I was deeply honored when George Box asked me to help him with this update. George was my Ph.D. advisor at the University of Wisconsin-Madison and remained a dear friend to me over the years as he did to all his students. Sadly, he was rather ill when the plans for this new edition were finalized towards the end of 2012. He did not have a chance to see the project completed as he passed away in March of 2013. I am deeply grateful for the opportunity to work with him and for the confidence he showed in assigning me this task. The book is dedicated to his memory and to the memory of his distinguished co-authors Gwilym Jenkins and Gregory Reinsel. Their contributions were many and they are all missed.

I also want to express my gratitude to several friends and colleagues in the time series community who have read the manuscript and provided helpful comments and suggestions. These include Ruey Tsay, William Wei, Sung Ahn, and Raja Velu who have read Chapter 14 on multivariate time series analysis, and David Dickey, Johannes Ledolter, Timo Teräsvirta, and Niels Haldrup who have read Chapter 10 on special topics. Their constructive comments and suggestions are much appreciated. Assistance and support from Paul Lindholm in Finland is also gratefully acknowledged. The use of R in this edition includes packages developed for existing books on time series analysis such as Cryer and Chan (2010), Shumway and Stoffer (2011), and Tsay (2014). We commend these authors for making their code and datasets available for public use through the R Project.

Research for the original version of this book was supported by the Air Force Office of Scientific Research and by the British Science Research Council. Research incorporated in the third edition was partially supported by the Alfred P. Sloan Foundation and by the National Aeronautics and Space Administration. Permission to reprint selected tables from *Biometrika Tables for Statisticians*, Vol. 1, edited by E. S. Pearson and H. O. Hartley is also acknowledged. On behalf of my co-authors, I would like to thank George Tiao, David Mayne, David Pierce, Granville Tunnicliffe Wilson, Donald Watts, John Hampton, Elaine Hodkinson, Patricia Blant, Dean Wichern, David Bacon, Paul Newbold, Hiro Kanemasu, Larry Haugh, John MacGregor, Bovas Abraham, Johannes Ledolter, Gina Chen, Raja Velu, Sung Ahn, Michael Wincek, Carole Leigh, Mary Esser, Sandy Reinsel, and

Meg Jenkins, for their help, in many different ways, in preparing the earlier editions. A very special thanks is extended to Claire Box for her long-time help and support.

The guidance and editorial support of Jon Gurstelle and Sari Friedman at Wiley is gratefully acknowledged. We also thank Stephen Quigley for his help in setting up the project, and Katrina Maceda and Shikha Pahuja for their help with the production.

Finally, I want to express my gratitude to my husband Bert Beander for his encouragement and support during the preparation of this revision.

GRETA M. LJUNG

Lexington, MA
May 2015

PREFACE TO THE FOURTH EDITION

It may be of interest to briefly recount how this book came to be written. Gwilym Jenkins and I first became friends in the late 1950s. We were intrigued by an idea that a chemical reactor could be designed that optimized itself automatically and could follow a moving maximum. We both believed that many advances in statistical theory came about as a result of interaction with researchers who were working on real scientific problems. Helping to design and build such a reactor would present an opportunity to further demonstrate this concept.

When Gwilym Jenkins came to visit Madison for a year, we discussed the idea with the famous chemical engineer Olaf Hougen, then in his eighties. He was enthusiastic and suggested that we form a small team in a joint project to build such a system. The National Science Foundation later supported this project. It took 3 years, but suffice it to say, that after many experiments, several setbacks, and some successes the reactor was built and it worked.

As expected, this investigation taught us a lot. In particular, we acquired proficiency in the manipulation of difference equations that were needed to characterize the dynamics of the system. It also gave us a better understanding of nonstationary time series required for realistic modeling of system noise. This was a happy time. We were doing what we most enjoyed doing: interacting with experimenters in the evolution of ideas and the solution of real problems, with real apparatus and real data.

Later there was fallout in other contexts, for example, advances in time series analysis, in forecasting for business and economics, and also developments in statistical process control (SPC) using some notions learned from the engineers.

Originally Gwilym came for a year. After that I spent each summer with him in England at his home in Lancaster. For the rest of the year, we corresponded using small reel-to-reel tape recorders. We wrote a number of technical reports and published some papers but eventually realized we needed a book. The first two editions of this book were written during a period in which Gwilym was, with extraordinary courage, fighting a debilitating illness to which he succumbed sometime after the book had been completed.

Later Gregory Reinsel, who had profound knowledge of the subject, helped to complete the third edition. Also in this fourth edition, produced after his untimely death, the new material is almost entirely his. In addition to a complete revision and updating, this fourth edition resulted in two new chapters: Chapter 10 on nonlinear and long memory models and Chapter 12 on multivariate time series.

This book should be regarded as a tribute to Gwilym and Gregory.

I was especially blessed to work with two such gifted colleagues.

GEORGE E. P. BOX

Madison, Wisconsin
March 2008

PREFACE TO THE THIRD EDITION

This book is concerned with the building of stochastic (statistical) models for time series and their use in important areas of application. This includes the topics of forecasting, model specification, estimation, and checking, transfer function modeling of dynamic relationships, modeling the effects of intervention events, and process control. Coincident with the first publication of *Time Series Analysis: Forecasting and Control,* there was a great upsurge in research in these topics. Thus, while the fundamental principles of the kind of time series analysis presented in that edition have remained the same, there has been a great influx of new ideas, modifications, and improvements provided by many authors.

The earlier editions of this book were written during a period in which Gwilym Jenkins was, with extraordinary courage, fighting a slowly debilitating illness. In the present revision, dedicated to his memory, we have preserved the general structure of the original book while revising, modifying, and omitting text where appropriate. In particular, Chapter 7 on estimation of ARMA models has been considerably modified. In addition, we have introduced entirely new sections on some important topics that have evolved since the first edition. These include presentations on various more recently developed methods for model specification, such as canonical correlation analysis and the use of model selection criteria, results on testing for unit root nonstationarity in ARIMA processes, the state-space representation of ARMA models and its use for likelihood estimation and forecasting, score tests for model checking, structural components, and deterministic components in time series models and their estimation based on regression-time series model methods. A new chapter (12) has been developed on the important topic of *intervention and outlier* analysis, reflecting the substantial interest and research in this topic since the earlier editions.

Over the last few years, the new emphasis on industrial quality improvement has strongly focused attention on the role of control both in process *monitoring* and in process *adjustment.* The control section of this book has, therefore, been completely rewritten to serve as an introduction to these important topics and to provide a better understanding of their relationship.

The objective of this book is to provide practical techniques that will be available to most of the wide audience who could benefit from their use. While we have tried to remove the inadequacies of earlier editions, we have not attempted to produce here a rigorous mathematical treatment of the subject.

We wish to acknowledge our indebtedness to Meg (Margaret) Jenkins and to our wives, Claire and Sandy, for their continuing support and assistance throughout the long period of preparation of this revision.

Research on which the original book was based was supported by the Air Force Office of Scientific Research and by the British Science Research Council. Research incorporated in the third edition was partially supported by the Alfred P. Sloan Foundation and by the National Aeronautics and Space Administration. We are grateful to Professor E. S. Pearson and the Biometrika Trustees for permission to reprint condensed and adapted forms of Tables 1, 8, and 12 of *Biometrika Tables for Statisticians*, Vol. 1, edited by E. S. Pearson and H. O. Hartley, to Dr. Casimer Stralkowski for permission to reproduce and adapt three figures from his doctoral thesis, and to George Tiao, David Mayne, Emanuel Parzen, David Pierce, Granville Wilson, Donald Watts, John Hampton, Elaine Hodkinson, Patricia Blant, Dean Wichern, David Bacon, Paul Newbold, Hiro Kanemasu, Larry Haugh, John MacGregor, Bovas Abraham, Gina Chen, Johannes Ledolter, Greta Ljung, Carole Leigh, Mary Esser, and Meg Jenkins for their help, in many different ways, in preparing the earlier editions.

GEORGE BOX AND GREGORY REINSEL

1

INTRODUCTION

A *time series* is a sequence of observations taken sequentially in time. Many sets of data appear as time series: a monthly sequence of the quantity of goods shipped from a factory, a weekly series of the number of road accidents, daily rainfall amounts, hourly observations made on the yield of a chemical process, and so on. Examples of time series abound in such fields as economics, business, engineering, the natural sciences (especially geophysics and meteorology), and the social sciences. Examples of data of the kind that we will be concerned with are displayed as time series plots in Figures 2.1 and 4.1. An intrinsic feature of a time series is that, typically, adjacent observations are *dependent*. The nature of this dependence among observations of a time series is of considerable practical interest. *Time series analysis* is concerned with techniques for the analysis of this dependence. This requires the development of stochastic and dynamic models for time series data and the use of such models in important areas of application.

In the subsequent chapters of this book, we present methods for building, identifying, fitting, and checking models for time series and dynamic systems. The methods discussed are appropriate for discrete (sampled-data) systems, where observation of the system occurs at equally spaced intervals of time.

We illustrate the use of these time series and dynamic models in five important areas of application:

1. The *forecasting* of future values of a time series from current and past values.
2. The determination of the *transfer function* of a system subject to inertia—the determination of a dynamic input–output model that can show the effect on the output of a system of any given series of inputs.
3. The use of indicator input variables in transfer function models to represent and assess the effects of unusual *intervention* events on the behavior of a time series.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

4. The examination of interrelationships among several related time series variables of interest and determination of appropriate *multivariate* dynamic models to represent these joint relationships among the variables over time.
5. The design of simple *control schemes* by means of which potential deviations of the system output from a desired target may, so far as possible, be compensated by adjustment of the input series values.

1.1 FIVE IMPORTANT PRACTICAL PROBLEMS

1.1.1 Forecasting Time Series

The use at time t of available observations from a time series to forecast its value at some future time $t + l$ can provide a basis for (1) economic and business planning, (2) production planning, (3) inventory and production control, and (4) control and optimization of industrial processes. As originally described by Holt et al. (1963), Brown (1962), and the Imperial Chemical Industries (ICI) monograph on short term forecasting (Coutie, 1964), forecasts are usually needed over a period known as the *lead time*, which varies with each problem. For example, the lead time in the inventory control problem was defined by Harrison (1965) as a period that begins when an order to replenish stock is placed with the factory and lasts until the order is delivered into stock.

We will assume that observations are available at *discrete*, equispaced intervals of time. For example, in a sales forecasting problem, the sales z_t in the current month t and the sales $z_{t-1}, z_{t-2}, z_{t-3}, \ldots$ in previous months might be used to forecast sales for lead times $l = 1, 2, 3, \ldots, 12$ months ahead. Denote by $\hat{z}_t(l)$ the forecast made at *origin* t of the sales z_{t+l} at some future time $t + l$, that is, at *lead time* l. The function $\hat{z}_t(l)$, which provides the forecasts at origin t for all future lead times, based on the available information from the current and previous values $z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots$ through time t, will be called the *forecast function* at origin t. Our objective is to obtain a forecast function such that the mean square of the deviations $z_{t+l} - \hat{z}_t(l)$ between the actual and forecasted values is as small as possible *for each lead time* l.

In addition to calculating the best forecasts, it is also necessary to specify their accuracy, so that, for example, the risks associated with decisions based upon the forecasts may be calculated. The accuracy of the forecasts may be expressed by calculating *probability limits* on either side of each forecast. These limits may be calculated for any convenient set of probabilities, for example, 50 and 95%. They are such that the realized value of the time series, when it eventually occurs, will be included within these limits with the stated probability. To illustrate, Figure 1.1 shows the last 20 values of a time series culminating at time t. Also shown are forecasts made from origin t for lead times $l = 1, 2, \ldots, 13$, together with the 50% probability limits.

Methods for obtaining forecasts and estimating probability limits are discussed in detail in Chapter 5. These forecasting methods are developed based on the assumption that the time series z_t follows a *stochastic* model of known form. Consequently, in Chapters 3 and 4 a useful class of such time series models that might be appropriate to represent the behavior of a series z_t, called autoregressive integrated moving average (ARIMA) models, are introduced and many of their properties are studied. Subsequently, in Chapters 6, 7, and 8 the practical matter of how these models may be developed for actual time series data is explored, and the methods are described through the three-stage procedure of tentative