palgrave
macmillan

The Foundations of Modern Time Series Analysis

Terence C. Mills

The Foundations of Modern Time Series Analysis

Palgrave Advanced Texts in Econometrics series.

Palgrave Advanced Texts in Econometrics is a series that provides coverage of econometric techniques, applications and perspectives at an advanced research level. It will include research monographs that bring current research to a wide audience; perspectives on econometric themes that develop a long term view of key methodological advances; textbook style presentations of advanced teaching and research topics. An over-riding theme of this series is clear presentation and accessibility through excellence in exposition, so that it will appeal not only to econometricians, but also to professional economists and, particularly, to Ph.D students and MSc students undertaking dissertations. The texts will include developments in theoretical and applied econometrics across a wide range of topics and areas including time series analysis, panel data methods, spatial econometrics and financial econometrics.

Palgrave Advanced Texts in Econometrics
Series Standing Order ISBN 978–0–230–34818–9

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and the ISBN quoted above.

Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke, Hampshire RG21 6XS, England.

The Foundations of Modern Time Series Analysis

Terence C. Mills
Professor of Applied Statistics and Econometrics, Department of Economics, Loughborough University, UK

palgrave
macmillan

First published 2011 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN 978–0–230–29018–1 hardback
ISBN 0–230–29018–3 hardback

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

10 9 8 7 6 5 4 3 2 1
20 19 18 17 16 15 14 13 12 11

Printed and bound in Great Britain by
CPI Antony Rowe, Chippenham and Eastbourne

Contents

List of Tables

List of Figures

1
Prolegomenon: A Personal Perspective and an Explanation of the Structure of the Book

Time series analysis: a personal perspective

1.1 My interest in time series analysis began around 1977, soon after I had been appointed to a lectureship in econometrics in the School of Economic Studies at the University of Leeds. I had earlier been subjected to a rather haphazard training in econometrics and statistics, both as an undergraduate at Essex and as a postgraduate at Warwick, so that when I entered academia I was basically self-taught in these subjects. This was an undoubted advantage in that my enthusiasm for them remained undiminished but it was accompanied by a major drawback: I was simply unacquainted with large areas of econometric and statistical theory. As an example of this haphazard background, in my final undergraduate year in 1973 I attended a course on the construction of continuous time economic models given by Peter Phillips, now the extremely distinguished Sterling Professor of Econometrics and Statistics at Yale but then in his first academic appointment, while my econometrics course consisted of being taught the yet to be examined thesis of a temporary lecturer who subsequently left academia, never to return, at the end of that academic year!

I was made painfully aware of these lacunae in my education by the arrival of Brendan McCabe – one of the finest theoretical time series analysts of our generation – to a lectureship at Leeds just three months after my own appointment, and then to what seemed at the time to be a flood of papers by Denis Sargan, David Hendry and Grayham Mizon outlining a new approach to time series econometrics (see, for example, Davidson et al., 1978; Hendry and Mizon, 1978; Sargan, 1980). The serial appearance of these papers meant that I had continually to rethink my doctoral thesis for the University of Warwick on modelling the UK demand for money function, for which the time limit for submission was rapidly approaching!

Hendry (1977) had a particularly major impact on my research and this led me – and not before time, many would say – to George Box and Gwilym Jenkins' classic book (Box and Jenkins, 1970) which, it is not too fanciful to say, altered my academic career completely! Throughout my academic 'training' in departments of economics I had never been comfortable with either economic theory or the traditional econometric approach of estimation conditional on a given theory, preferring to take an unashamedly empirical approach to econometric modelling (I had, in fact, been offered a grant to take the MSc in Operational Research at Lancaster in 1973, where Gwilym Jenkins was then based, as well as one to take the MA in Economics at Warwick, opting for the latter on the grounds that, as a Londoner, Lancaster was much too far 'up north' – an interesting decision given that I subsequently spent almost twenty years at the universities of Leeds and Hull!)

The model-building philosophy expounded by Box and Jenkins was therefore intellectually very congenial to me and I embraced it with enthusiasm. Assimilating all these ideas, along with the then extremely popular approach of Granger–Sims causality testing (Granger, 1969; Sims, 1972), enabled me to successfully complete my thesis in 1979 and to get my first publications under my belt.

My time series education was extended further during a part-time stint in the Bank of England's Monetary Policy Group during the early 1980s, where a chance encounter with Peter Burman, then Head of Statistical Techniques, enabled me to become acquainted with unobserved component models and signal extraction techniques (Burman, 1980; Mills, 1982a, 1982b). I was now up and running and a few years later *Time Series Techniques for Economists* (Mills, 1990) was published, which, to my continued surprise, remains in print over twenty years later.

1.2 I have always been interested in the historical development of econometrics and statistics, no doubt in part a consequence of my long collaborations and friendships with economic historians, notably Nick Crafts and Forrest Capie. My early forays into the subject were restricted to the introductions to Edward Elgar collections on economic and financial market forecasting and on the modelling of trends and cycles (Mills, 1999, 2002a, 2002b), but later articles (Mills, 2009, 2011) consolidated my interest and led directly to the writing of this book.

Scope of the study

1.3 The early, essentially descriptive, history of time series analysis has been covered in detail by Klein (1997). I therefore quickly decided that my starting point would be the formal development of the concept of correlation and the first statistical analyses of meteorological and economic time series, which took place during the last decade of the nineteenth century. My end point was chosen rather more subjectively, but it became clear that the publication of Box and

Jenkins' book in 1970 marked, in retrospect, a watershed in the development of the subject, as it synthesized much of the analysis that had been carried out up to that point and, as a consequence, acted as a catalyst for the explosion of research that has subsequently been undertaken over the last 40 years. The choice of 1970 also resonated from a personal perspective, as it was the year in which I entered higher education, where I have remained ever since!

Style and structure of the book

1.4 Natural reference points to the development of time series analysis in the first half of the twentieth century are Udny Yule and Maurice Kendall's *An Introduction to the Theory of Statistics* (Yule and Kendall, 14th edition, 1950) and Kendall's *Advanced Theory of Statistics* (Kendall, 1946). As well as being hugely impressed by the general excellence of these texts, I was also taken by the format of subheading and section number used in them. I have adopted this format here, both to pay homage to these two British greats of the subject and also because of the ease with which it allows cross-referencing, an essential part of a study such as this. Thus a cross-reference to section **y** of Chapter **x** will be denoted §**x.y** in subsequent chapters.

On reading many of the early papers on time series, particularly those in *Biometrika* and the *Journal of the Royal Statistical Society*, I was immediately struck by their discursive prose style and, it must be said, by the length of the articles, which facilitated such discursiveness (no doubt this was helped by the relatively small number of active time series analysts writing at the time, the lack of a peer review process – not necessarily a bad thing under the circumstances – and the fact that authors were also often the editors of the journals!) I have thus taken the opportunity of quoting at length from these seminal contributions as it is my opinion that being able to read the original descriptions, arguments and, quite frankly, the prejudices and hobby horses of the major protagonists, adds much to our understanding of the development of the subject and, indeed, to the overall gaiety of these contributions. Indeed, the contrast between these papers and the terseness of many current journal articles is quite striking.

I have also provided, in various endnotes, short 'pen pictures' of some of the major figures in time series to provide background colour to the analysis being developed. Of course, biographies exist for several characters and references to these are given in the notes.

1.5 The book contains 16 chapters, including this. Chapter 2 introduces the early work of Yule on regression and correlation and of Hooker on the concept of trend. Chapter 3 is devoted to periodogram analysis and focuses on the applications of this technique made by Schuster and Beveridge to sunspots and wheat prices, respectively. Early concerns with detrending are the focus of Chapter 4, which examines the variate differencing method of Student and Pearson and its critique by Yule and Persons. By this time, the early 1920s, formal statistical models of time series had begun to be developed and Chapter 5 concentrates on the 'first generation' of these models proposed by Yule, Slutzky and Working, with the analyses of Yule and Walker on periodicities in sunspots and air pressure being a consequence of superposed fluctuations forming the material of Chapter 6.

During the 1930s the probabilistic theory of time series began to be developed, first by Russian mathematicians and then by the Swede Herman Wold: Chapter 7 is devoted to his 1938 monograph *A Study in the Analysis of Stationary Time Series*, which laid the foundations for subsequent theoretical research in the subject. Chapter 8 covers various extensions to the autoregressive class of models, in particular the oscillatory models of Kendall. Hard on the heels of Wold, the 1940s saw major research activity, by an increasing number of statisticians, on developing a theory of statistical inference for stationary time series. This is developed in Chapter 9, which then goes on to discuss various proposals for estimating autoregressive, moving average and mixed processes, culminating in the univariate modeling methodology that was developed by Box and Jenkins during the 1960s.

Of course, analysts since the beginning of the twentieth century had been confronted with time series that were not stationary but which contained trends, hence the need for methods such as variate differencing. A parallel literature had also developed, primarily in the actuarial profession, of detrending by 'graduation' – the taking of successive moving averages. Chapter 10 begins by linking this literature to the more conventional detrending method of fitting local polynomial trends. It then goes on to consider other methods of eliminating trend movements, most notably by differencing, which led on to the concept of an integrated process and the associated ARIMA model. Forecasting time series with local trends became of increasing concern during the 1950s in a variety of disciplines and this is the subject of Chapter 11, which looks at both exponential smoothing techniques and the 'full blown' theory of forecasting ARIMA models whilst also examining the links between them.

Up to this point, the development has been focused almost exclusively on methods for analysing time series individually, but during the 1950s the modelling of several series together began to attract attention. Chapter 12 thus develops the transfer function approach of Box and Jenkins, in which an 'input' affects an 'output', and also the more general framework of multiple time series analysis, which allows feedback between various series.

Chapter 13 focuses on the modern extension of the periodogram, spectral analysis, while Chapter 14 discusses the various techniques that have been developed to deal with seasonal patterns in time series, both to adjust the data for such fluctuations and to explicitly model the observed seasonality.

Chapter 15 examines four sub-themes that developed between the late 1950s and 1970, namely inference concerning nonstationarity, the use of model selection criteria, state space models, the Kalman filter and recursive estimation, and nonlinearity in time series. Finally, Chapter 16 links the emerging themes from the previous chapters to the huge explosion of research undertaken over the last forty years since 1970 and offers some thoughts as to where the subject is likely to go from the position it finds itself in at the start of the second decade of the twenty-first century.

2
Yule and Hooker and the Concepts of Correlation and Trend

Yule on regression and correlation

2.1 The foundations of modern time series analysis began to be laid in the late nineteenth century and were made possible by the invention of regression and the related concept of the correlation coefficient. By the final years of the century the method of correlation had made its impact felt primarily in biology, through the work of Francis Galton on heredity (Galton, 1888, 1890) and of Karl Pearson on evolution (Pearson 1896; Pearson and Filon, 1898).[1] Correlation had also been used by Edgeworth (1893, 1894) to investigate social phenomena and by G. Udny Yule in the field of economic statistics, particularly to examine the relationship between welfare and poverty (Yule, 1895, 1896).[2] This led Yule (1897a, 1897b) to provide a full development of the theory of correlation which, unusually from a modern perspective – but, as we shall see, importantly for time series analysis –, was based on the related idea of a regression between two variables X and Y.[3] It also did not rely on the assumption that the two variables were jointly normally distributed, which was central to the formal development of correlation in Edgeworth (1892) and Pearson (1896). This was an important generalization, for Yule was quick to appreciate that much of the data appearing in the biological and social sciences were anything but normally distributed, typically being highly skewed.

2.2 Yule's development is worth setting out in some detail. Let $x = X - \overline{X}$ and $y = Y - \overline{Y}$ denote the deviations of the variables from their respective means. Suppose that x takes on k distinct values and that a particular value, $x_i, i = 1, 2, \ldots, k,$ is associated with n_i values of $y, y_{ij}, j = 1, 2, \ldots, n_i, (n_1 + \cdots + n_k = n)$. These n_i pairs of values (y_{ij}, x_i) are called a 'y-array', from which can be defined the array mean[4]

$$\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$$

and array variance

$$\sigma_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$

Let $y = bx$ be the regression line that is, in some sense, the best linear representation of the relationship between the k pairs $(\overline{y}_i, x_i)$ and define

$$d_i = \overline{y}_i - bx_i$$

to be the distance from the ith array mean to the regression line. For the ith y-array,

$$\sum_{j=1}^{n_j}(y_{ij} - bx_i)^2 = n_i\sigma_i^2 + n_i d_i^2$$

and summing over all k arrays gives

$$\sum_{i=1}^{k} n_i d_i^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij} - bx_i)^2 - \sum_{i=1}^{k} n_i\sigma_i^2 \qquad (2.1)$$

Yule chose his best-fitting regression line to be the one that minimizes the left-hand side of (2.1). Because the second term on the right-hand side of (2.1) does not depend on b, this minimization is equivalent to choosing b to minimize

$$\sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij} - bx_i)^2$$

This, of course, is the standard method of least squares and leads to

$$b = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_j} y_{ij}x_i}{\sum_{i=1}^{n_i} n_i \sum_{i=1}^{k_i} x_i^2} \qquad (2.2)$$

Yule referred to b as the regression coefficient or, somewhat confusingly from today's perspective, as simply the regression. Redesignating the individual pairs of observations as (y_p, x_p), $p = 1, 2, \ldots, n$, then allows (2.2) to be written in the familiar form[5]

$$b = \frac{\sum_{p=1}^{n} y_p x_p}{\sum_{p=1}^{n} x_p^2} \qquad (2.3)$$

Yule then defined

$$\sum_{p=1}^{n} x_p^2 = n\sigma_x^2; \qquad \sum_{p=1}^{n} y_p^2 = n\sigma_y^2; \qquad \sum_{p=1}^{n} y_p x_p = nr\sigma_y\sigma_x$$

where σ_x and σ_y are the standard deviations of x and y.[6] The formula (2.3) can then be expressed as

$$b = r\frac{\sigma_y}{\sigma_x} \tag{2.4}$$

and what Yule termed the *characteristic relation* between y and x becomes

$$y = r\frac{\sigma_y}{\sigma_x}x$$

By analogous reasoning, the characteristic relation between x and y is

$$x = r\frac{\sigma_x}{\sigma_y}y = b'y$$

How is the new variable $r = \sqrt{bb'}$, the geometric mean of the two regressions, to be interpreted? Note first that the two characteristic relations can be written as

$$\frac{y}{\sigma_y} = r\frac{x}{\sigma_x} \qquad \frac{x}{\sigma_x} = r\frac{y}{\sigma_y}$$

prompting Yule to state that

> *if we measure x and y each in terms of its own standard deviation, r becomes at once the regression of x on y and the regression of y on x, these two regressions being then identical.* (Yule, 1897b, page 820: italics in original)

Using (2.4), and dropping subscripts and limits of summation for notational convenience, obtains

$$\sum (y - bx)^2 = \sum \left(y - r\frac{\sigma_y}{\sigma_x}x\right)^2 = n\sigma_y^2(1 - r^2) \tag{2.5}$$

and, analogously,

$$\sum (x - b'y)^2 = n\sigma_x^2(1 - r^2) \tag{2.6}$$

$$\sum \left(\frac{y}{\sigma_y} - \frac{x}{\sigma_x}\right)^2 = \sum \left(\frac{x}{\sigma_x} - \frac{y}{\sigma_y}\right)^2 = n(1 - r^2)$$

All these quantities, being sums of squares, must necessarily be positive, so that r cannot be numerically greater than unity (i.e., $|r| \leq 1$). If $r = \pm 1$, all these quantities become zero, but

$$\sum \left(\frac{y}{\sigma_y} \pm \frac{x}{\sigma_x} \right)^2 = 0$$

requires that

$$\frac{y_p}{\sigma_y} \pm \frac{x_p}{\sigma_x} = 0 \quad p = 1, 2, \ldots, n$$

or

$$\frac{y_1}{x_1} = \frac{y_2}{x_2} = \cdots = \frac{y_p}{x_p} = \pm \frac{\sigma_y}{\sigma_x}$$

the sign of the last term being the sign of r. Hence, '*when the value of r is unity, all pairs of deviations bear the same ratio to one another, or the values of the two variables are related by a simple linear law*' (*ibid.*, page 821: italics in original). In other words, the distribution of the scatter of Y and X values has collapsed into a distribution along a straight line. The greater the value of $|r|$, the more closely this result holds, and hence r is termed the *coefficient of correlation*. Yule took great care to contrast the interpretation of $|r| = 1$ – that of perfect correlation – with its 'polar' opposite:

> ... $r = 0$ does not *in general* imply that the variables are strictly independent in the sense that the chance of getting a pair of deviations is equal to the product of the chances of getting either separately. The condition $r = 0$ is necessary but is not sufficient. (*ibid.*, page 821: italics in original)

Yule was clearly aware that the linear regression model underlying the calculation of r was just an assumption: 'if the [true] regression be very far from linear some caution must evidently be used in employing r to compare two different distributions' (*ibid.*, page 821: see also the discussion on pages 816–17).

Yule then noted that the quantities in (2.5) and (2.6), $\sigma_y\sqrt{1-r^2}$ and $\sigma_x\sqrt{1-r^2}$, were the standard errors made in estimating y and x from their respective characteristics, i.e., regressions, and regarded $\sqrt{1-r^2}$ as such an important quantity that he provided a table (Table I of the Appendix) of its values for r incrementing in hundredths.

2.3 After a detailed numerical example reworking the pauperism and welfare relief data of Yule (1896), he then extended the regression framework to three variables, now denoted X_1, X_2 and X_3, with mean deviations x_1, x_2 and x_3, and with

$$\sum x_i^2 = n\sigma_i^2, \quad i = 1, 2, 3; \quad \sum x_i x_j = n r_{ij} \sigma_i \sigma_j, \quad i \neq j$$

The characteristic relation, or regression,

$$x_1 = b_{12}x_2 + b_{13}x_3 \tag{2.7}$$

is then fitted by solving the following normal equations for b_{12} and b_{13}:

$$\sum x_1x_2 = b_{12}\sum x_2^2 + b_{13}\sum x_2x_3$$

$$\sum x_1x_3 = b_{12}\sum x_2x_3 + b_{13}\sum x_3^2$$

As these can be written

$$r_{12}\sigma_1 = b_{12}\sigma_2 + b_{13}r_{23}\sigma_3$$

$$r_{13}\sigma_1 = b_{12}r_{12}\sigma_2 + b_{13}\sigma_3$$

the solutions are

$$\begin{aligned} b_{12} &= \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\frac{\sigma_1}{\sigma_2} \\ b_{13} &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\frac{\sigma_1}{\sigma_3} \end{aligned} \tag{2.8}$$

There are, of course, two further characteristic relations expressing x_2 and x_3, respectively, in terms of the remaining pair of variables: '(t)he value of any b in terms of the r's can be written down from the expressions [2.8] by simply interchanging the suffixes. Thus b_{23} could be written down by simply writing 2 for 1 and 3 for 2 in the expression for b_{12}' (Yule, 1897b, page 832). Yule then defined

$$v = x_1 - (b_{12}x_2 + b_{13}x_3)$$

to be the 'error made in estimating x_1 from relation [2.7] or a deviation of x_1 from the value $(b_{12}x_2 + b_{13}x_3)$', remarking that the 'relation [2.7] has been so formed that

$$\sum v^2 = \sum (x_1 - (b_{12}x_2 + b_{13}x_3))^2$$

is the least possible' (*ibid.*, page 832). Using the solutions (2.8), this sum of squared errors can be written as

$$\sum v^2 = n\sigma_1^2\left(1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}\right) = n\sigma_1^2(1 - R_1^2)$$

where $\sigma_1\sqrt{1 - R_1^2}$ is the standard error made in estimating x_1 from the regression (2.7) and R_1 is the coefficient of correlation between x_1 and (x_2, x_3), which Yule suggested might be termed a 'coefficient of double correlation'.[7] Yule termed the quantities b_{12}, b_{13}, etc., the *net* or *partial* regression coefficients, and the quantity

$$r_{12.3} = \sqrt{b_{12}b_{21}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)}\sqrt{(1 - r_{23}^2)}} \tag{2.9}$$

the net, or presumably partial, correlation coefficient.[8] It retains the chief properties of the ordinary correlation coefficient in that $r_{12.3}$ will be zero if both partial regression coefficients are zero, it is a symmetric function of the variables ($r_{12.3} = r_{21.3}$), and $|r_{12.3}| \leq 1$. The definition (2.9) has some interesting implications. Since

$$(r_{12} - r_{13}r_{23})^2 \leq (1 - r_{13}^2)(1 - r_{23}^2)$$

r_{12} must lie between the limits

$$r_{13}r_{23} \pm \sqrt{1 + r_{13}^2 r_{23}^2 - r_{13}^2 - r_{23}^2}$$

By providing a table of special cases, Yule (*ibid.*, page 834) showed that it was perfectly possible for both r_{13} and r_{23} to be positive, yet for r_{12} to be negative or zero: indeed, only when r_{13} and r_{23} both exceed $\sqrt{0.5} = 0.707$ will $0 < r_{12} \leq 1$.

2.4 Yule then considered the conditions under which the standard error of the regression of x_1 on x_2 and x_3, $\sigma_1\sqrt{1 - R_1^2}$, would be smaller than the standard error of the regression of x_1 on just x_2, $\sigma_1\sqrt{1 - r_{12}^2}$. This is equivalent to finding the conditions that guarantee $R_1^2 > r_{12}^2$. The necessary condition is easily shown to be $(r_{13} - r_{12}r_{23})^2 > 0$. But, from (2.9), $r_{13} - r_{12}r_{23}$ is the numerator of $r_{13.2}$, so that $R_1^2 > r_{12}^2$ as long as $r_{13.2}$ is non-zero. For example, if $r_{12} = \pm 0.8$, $r_{23} = 0.5$ and $r_{13} = 0.4$, then $r_{13.2} = 0$ and, although x_3 is reasonably positively correlated with x_1, it turns out to be of no assistance in estimating x_1. Conversely, if $r_{13} = 0$ it cannot be concluded that x_3 is of no use (i.e., $r_{13.2} = 0$) unless $r_{12} = 0$ as well.

2.5 The remainder of Yule (1897b) extended the analysis to four variables and then considered the cases of two and three variable correlation when the variables are jointly normally distributed. In the latter development, Yule introduced a result on the probable error of the correlation coefficient that was contained in the then unpublished Pearson and Filon (1898), being

$$0.674489\frac{1 - r^2}{\sqrt{n}} \tag{2.10}$$

The constant 0.674889 is the 0.25 value of the standard normal distribution, so that this formula provides the approximate bounds for a 50% confidence interval for the 'true' value of the correlation coefficient: $(1-r^2)/\sqrt{n}$ can therefore be seen to be the standard error of the correlation coefficient from a normal population.

Hooker and the concept of trend

2.6 By the turn of the twentieth century, applications of the theory of correlation were becoming more popular, particularly using economic and social data (in addition to the references given in §**2.1**, see also Yule, 1899, and Hooker, 1901a).[9] Hooker (1901b), in examining the correlation between the marriage rate and trade (taken to be the value of exports per capita) over the period 1857–1899, raised an important difficulty with correlation analysis when applied to time series data.[10] If the movements in the two series that are being correlated are produced by a combination of slow secular movements and more rapid, say year to year, changes, then the latter may be highly correlated while the former may be unrelated, so that the overall correlation between the two series may turn out to be small. This is exactly what appeared to be happening with the marriage rate and trade, which over the period exhibited declining and increasing secular movements, respectively, thus producing a calculated correlation of just 0.18 with a probable error (see equation (2.10)) of 0.09. Arguably, what was really of interest was the correlation between the minor oscillations – the short-run movements in the series – and to counteract this, Hooker proposed the following strategy, which is worth quoting in full.

> What I wish to suggest . . . is an elementary method of eliminating the general movement in the particular case of phenomena exhibiting similar regular periodic movements, so as to enable us to correlate the oscillations.
>
> To correlate the oscillations of two curves, I propose that all deviations should be reckoned, not from the average of the whole period, but from the instantaneous average at the moment. The curve or line representing the successive instantaneous averages I propose to call the *trend*. Any point on the trend will be represented by the average of all observations in the period of which that moment is the central point; e.g., if a curve shows a period of p years, the instantaneous average in any year is the average of the p years of which that particular year is the middle. By working out this instantaneous average for consecutive observations, we obtain the trend in the curve; i.e., the direction in which the variable is really moving when the oscillations are disregarded. And by replacing the deviations from the average in the formula $r=\sum x_1x_2/n\sigma_1\sigma_2$ by the deviations from this trend, we shall obtain a measure of the correlation of the oscillations of two curves exhibiting similar regular fluctuations. (Hooker 1901b, page 486)

Thus, not only did Hooker introduce for the first time the notion of a trend, but he also proposed detrending by a moving average.[11] Choosing $p=9$ on the grounds 'that a trade maximum occurs, on an average, approximately every ninth year' (*ibid.*, page 487), this strategy produced a correlation of 0.80 (probable error 0.04) between the detrended marriage rate and trade, leading Hooker to conclude 'that while there is no connection between the *general* movements of the two curves, there is a close correspondence between the oscillations' (*ibid.*, page 487, italics in original). Hooker then asked the following question

> does the marriage-rate respond immediately to general prosperity? In other words, will not a maximum in the marriage rate occur some time *after* a maximum in the trade curve; and ought we not therefore to correlate the marriage-rate with the trade in the previous year? (*ibid.*, page 487, italics in original)

To answer this, Hooker also calculated correlations between the marriage rate and trade lagged by one year and by half a year (taken as the average of the current and previous year's trade) and led by one year and by half a year. The maximum correlation was 0.86 when trade was lagged by half a year, allowing Hooker to 'conclude that, on the average of the thirty-five years, the marriage-rate follows the exports at an interval of half a year' (*ibid.*, page 488). This would therefore represent the, admittedly rudimentary, first appearance of what would come to be known as a cross-correlation function (see §**12.13**).

2.7 Hooker repeated the analysis using various other measures of trade and also broke the sample into two, enabling him to contrast the correlations and the lead/lags across the two sub-samples. As Hendry and Morgan (1995, page 11) remark, 'Hooker's paper demonstrates the new level of technology brought in from the biometricians and the new skills of inference needed for such techniques, as well as their remarkable range of application. In modern parlance, he explicitly considers non-stationarity due to both stochastic trends and regime shifts as well as deterministic trends, cross serial correlations and lead-lag determination, and issues of model selection when there are multiple correlated causes so that the empirical model has to be discovered from the data'. Hooker had thus taken the analysis of time series data to a much higher plane than ever before.[12]

2.8 Hooker's core example of the correlation between the marriage rate and per capita trade can be recreated using data from Mitchell (1998). Table 2.1 presents the marriage rate for the UK, the trend in the marriage rate, calculated as a 9-year centred moving average, and the detrended marriage rate, which Hooker called the oscillations in the rate, along with similar calculations for total UK trade per capita, for the period 1857 to 1899. The correlation between

Table 2.1 Marriage rate and trade per capita in the UK, 1857–1899

	Marriage rate	Nine-year moving average	Detrended marriage rate	Trade per capita	Nine-year moving average	Detrended trade per capita
1857	16.19	–	–	15.00	–	–
1858	15.60	–	–	13.56	–	–
1859	16.59	–	–	14.69	–	–
1860	16.67	–	–	16.38	–	–
1861	15.90	16.39	–0.49	15.98	16.83	–0.85
1862	15.73	16.49	–0.76	16.66	17.57	–0.91
1863	16.47	16.55	–0.08	18.77	18.30	0.47
1864	17.18	16.46	0.72	20.26	18.97	1.29
1865	17.15	16.34	0.81	20.15	19.46	0.69
1866	17.13	16.33	0.80	21.69	20.03	1.66
1867	16.16	16.41	–0.25	20.11	20.79	–0.68
1868	15.74	16.48	–0.74	20.74	21.51	–0.77
1869	15.58	16.49	–0.91	20.76	22.09	–1.33
1870	15.87	16.45	–0.58	21.17	22.58	–1.41
1871	16.39	16.38	0.01	23.51	22.81	0.70
1872	17.10	16.39	0.71	25.25	23.09	2.16
1873	17.33	16.37	0.96	25.40	23.32	2.08
1874	16.77	16.30	0.47	24.56	23.40	1.16
1875	16.46	16.12	0.34	23.77	23.39	0.38
1876	16.31	15.93	0.38	22.64	23.41	–0.77
1877	15.54	15.69	–0.15	22.83	23.19	–0.36
1878	14.97	15.47	–0.50	21.46	23.02	–1.56
1879	14.20	15.31	–1.11	21.07	22.97	–1.90
1880	14.69	15.13	–0.44	23.68	22.81	0.87
1881	14.95	14.91	0.04	23.30	22.59	0.71
1882	15.32	14.74	0.58	23.89	22.24	1.65
1883	15.33	14.66	0.67	24.09	22.11	1.98
1884	14.91	14.66	0.25	22.31	22.15	0.16
1885	14.33	14.67	–0.34	20.66	22.06	–1.40
1886	14.00	14.71	–0.71	19.71	22.02	–2.31
1887	14.19	14.71	–0.52	20.26	21.86	–1.60
1888	14.20	14.70	–0.50	21.42	21.56	–0.14
1889	14.79	14.66	0.13	22.95	21.31	1.64
1890	15.28	14.71	0.57	22.89	21.23	1.66
1891	15.39	14.80	0.59	22.46	21.29	1.17
1892	15.24	14.95	0.29	21.34	21.38	–0.04
1893	14.52	15.13	–0.61	20.13	21.34	–1.21
1894	14.79	15.27	–0.47	19.90	21.16	–1.26
1895	14.82	15.38	–0.56	20.28	21.08	–0.80
1896	15.52	–	–	21.06	–	–
1897	15.81	–	–	21.01	–	–
1898	16.03	–	–	21.33	–	–
1899	16.32	–	–	22.19	–	–

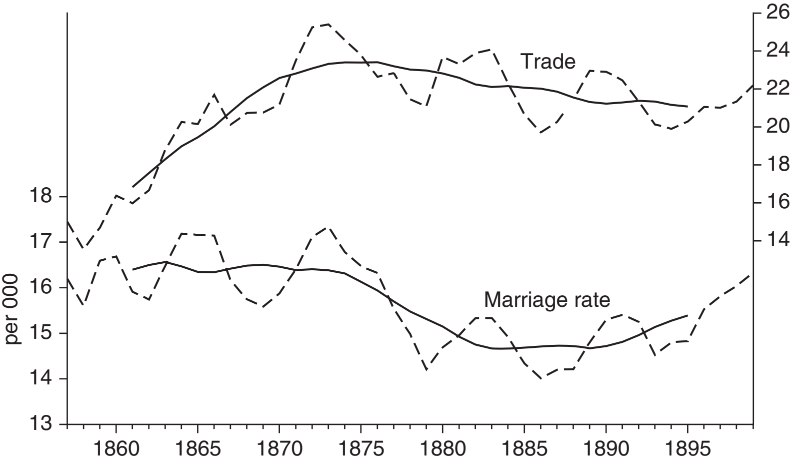


Figure 2.1 Marriage rate and trade per capita in the UK, 1857–1899, with nine year centred moving average superimposed

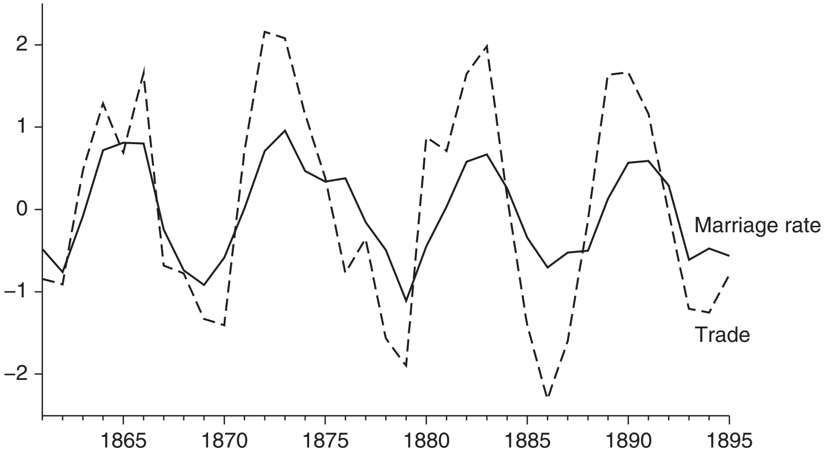


Figure 2.2 Detrended marriage rate and trade per capita in the UK, 1861–1895

the marriage rate and trade per capita is −0.001, so that the two observed series are uncorrelated, but the detrended 'oscillations' have a correlation of 0.85 with a standard error, calculated from (2.10), of 0.03.

The data are plotted in Figures 2.1 and 2.2 and these very different correlations are obviously borne out from the plots. Table 2.2 reports the 'cross-correlations', where $r(k)$ denotes the correlation between the current detrended marriage rate and detrended trade lagged k years (with negative k implying a lead). Following Hooker, $k = 1/2$ denotes the correlation using the average of the current and

Table 2.2 Correlations between the detrended marriage rate and lags of detrended trade per capita

Lag k	2	$1\frac{1}{2}$	1	$\frac{1}{2}$	0	$-\frac{1}{2}$	−1	$-1\frac{1}{2}$	−2
				(a) 35 years: 1861–1895					
$r(k)$	0.28	0.59	0.76	0.91	0.85	0.68	0.37	0.10	−0.23
				(b) 15 years: 1861–1875					
$r(k)$	0.11	0.45	0.68	0.91	0.93	0.80	0.52	0.26	−0.07
				(c) 20 years: 1876–1895					
$r(k)$	0.36	0.68	0.82	0.92	0.80	0.59	0.24	−0.09	−0.39

lagged one year detrended trade and other non-integer k are defined in analogous fashion.[13] As in Hooker, the sample is split at 1876 and cross-correlations computed for the full and the two sub-samples.

These findings replicate those of Hooker in that the lag looks to have increased across the samples, at least in terms of the value of k that produces the maximum cross-correlation: this is found to be $k=1/2$ for the later sub-sample but $k=0$ for the earlier one.

2.9 Hooker (1905) returned to the issue of 'detrending', but now attacked it by suggesting a method that 'consists simply in calculating the correlation coefficients of the differences between successive values of two variables' (page 697). Thus, given two time series (Y_t, X_t), $t=1,2,\ldots,T$, then, rather than calculating the usual correlation coefficient from the mean deviations $y_t = Y_t - \overline{Y}$, $x_t = X_t - \overline{X}$,

$$r_{xy} = \frac{\sum x_t y_t}{n\sigma_x\sigma_y} \quad \sigma_x^2 = \frac{\sum x_t^2}{n} \quad \sigma_y^2 = \frac{\sum y_t^2}{n} \tag{2.11}$$

the correlation coefficient between the *successive differences*, $\Delta x_t = x_t - x_{t-1}$, $\Delta y_t = y_t - y_{t-1}$, $t=2,\ldots,T$, is calculated.[14] Noting that the sample means of these differences can be written as $\overline{\Delta x} = (x_T - x_1)/T$ and $\overline{\Delta y} = (y_T - y_1)/T$, this correlation is given by

$$r_{\Delta x \Delta y} = \frac{\sum (\Delta x_t - \overline{\Delta x})(\Delta y_t - \overline{\Delta y})}{T\sigma_{\Delta x}\sigma_{\Delta y}}$$

$$\sigma_{\Delta x}^2 = \frac{\sum (\Delta x_t - \overline{\Delta x})^2}{T} \quad \sigma_{\Delta y}^2 = \frac{\sum (\Delta y_t - \overline{\Delta y})^2}{T}$$

This correlation coefficient was applied to the daily changes of the corn prices analysed in Hooker (1901a), finding that the absolute sizes of the correlation coefficients so obtained were considerably smaller than (often less than half the size of) the corresponding correlation coefficients calculated from the levels. The conclusion drawn by Hooker from this analysis seems particularly prescient when viewed from a modern perspective:

> in examining the relationship between two series of observations extending over a considerable period of time, correlation of absolute values (deviations from the arithmetic mean) is the most suitable test of 'secular' interdependence, and may also be the best guide when the observations tend to deviate from an average that may be regarded as constant. Correlation of the deviations from an instantaneous average (or *trend*) may be adopted to test the similarity of more or less marked periodic influences. Correlation of the difference between successive values will probably prove most useful in cases where the similarity of the shorter rapid changes (with no apparent periodicity) are the subject of investigation, or where the normal level of one or both series does not remain constant. It may even, in certain cases, be desirable to combine the two methods, and to correlate the deviations from the mean in the one series with the successive changes of the other. (Hooker, 1905, page 703: italics in original)

Hooker was thus clearly aware of the distinction between what are now called integrated (here $I(0)$ and $I(1)$) processes and of the difficulties inherent in modelling the relationships between series of different orders of integration (see the discussion in §**16.20**). Almost contemporaneously, Cave-Browne-Cave (1905) was considering both the correlation between daily changes in barometric heights between two meteorological stations and the correlation between successive daily barometric heights at the two stations themselves, thus providing the first example of calculating serial correlations. Thus by the early years of the twentieth century, the first hesitant steps along the path of modern time series analysis were clearly being taken, although formalization of these methods would have to wait another twenty years.

3 Schuster, Beveridge and Periodogram Analysis

Periodogram analysis

3.1 Around the time that Hooker and Yule were developing correlation and detrending techniques for economic time series, the physicist Sir Arthur Schuster was investigating periodicities in series such as earthquake frequency and sunspot numbers using a technique that became to be known as *periodogram analysis* (see Schuster, 1897, 1898, 1906).[1] Periodogram analysis is based on the technique of harmonic analysis and the use of Fourier series, which we outline using the classic approach taken in Whittaker and Robinson (1924) and Davis (1941).[2]

By a harmonic we mean a function of the form

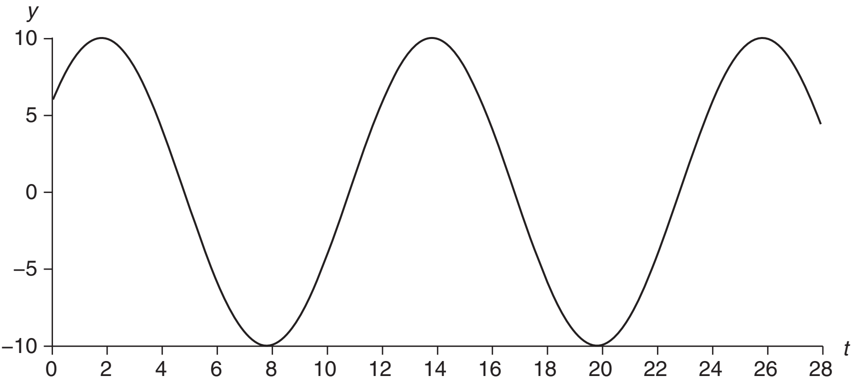
$$\begin{aligned} y &= A\cos\frac{2\pi t}{n} + B\sin\frac{2\pi t}{n} = \rho\cos\alpha\cos\frac{2\pi t}{n} + \rho\sin\alpha\sin\frac{2\pi t}{n} \\ &= \rho\left(\cos\left(\frac{2\pi t}{n} - \alpha\right)\right) \qquad (3.1) \\ &= \sqrt{A^2+B^2}\cos\left(\frac{2\pi t}{n} - \alpha\right) \end{aligned}$$

where we use the trigonometric identity $\cos(\beta-\alpha) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$ and note that the definitions $A = \rho\cos\alpha$ and $B = \rho\sin\alpha$ imply that $\tan\alpha = B/A$ and $\rho^2 = A^2 + B^2$ since $\cos^2\alpha + \sin^2\alpha = 1$.

In (3.1) n is the *period* of the harmonic, its reciprocal, $1/n$, is the *frequency*, and $\rho = \sqrt{A^2+B^2}$ is the *amplitude*; $\alpha =$ arctan B/A is the *phase angle*, whose effect is to delay by $n\alpha/2\pi$ time periods the peak of the cosine function, which would otherwise occur at $t = 0, n, 2n, \ldots$. A plot of the harmonic function

$$\begin{aligned} y &= 6\cos\frac{2\pi t}{12} + 8\sin\frac{2\pi t}{12} = 10\cos\left(\frac{2\pi t}{12} - \alpha\right) \\ \alpha &= \arctan 1.3333 = 53.13^\circ = 0.2952\pi \end{aligned}$$

is shown below, where the amplitude and the period of the cycle are clearly 10 and 12 respectively, with the phase angle of 53.13° inducing a *phase shift* of $12 \times 0.2952/2 = 1.77$ time periods.



3.2 By writing $\omega = 2\pi/n$ as the frequency measured in radians, a series of the form

$$y_t = \tfrac{1}{2}A_0 + \sum_{j=1}^{\infty} A_j \cos j\omega t + \sum_{j=1}^{\infty} B_j \sin j\omega t$$

may be defined, which is known as a (*trigonometrical*) *Fourier series*. To obtain the coefficients A_j and B_j in terms of the observed series y_t, we can make use of the orthogonality conditions which prevail amongst the harmonic components, which lead to[3]

$$\begin{aligned} A_0 &= \frac{2}{n}\int_0^n y_t\,dt \\ A_j &= \frac{2}{n}\int_0^n y_t \cos j\omega t\,dt \qquad j > 0 \\ B_j &= \frac{2}{n}\int_0^n y_t \sin j\omega t\,dt \qquad j > 0 \end{aligned}$$

Suppose now that y_t is to be approximated by the first J harmonics of a Fourier series:

$$y_t = \tfrac{1}{2}A_0 + \sum_{j=1}^{J} A_j \cos j\omega t + \sum_{j=1}^{J} B_j \sin j\omega t + e_{J,t} = y_{J,t} + e_{J,t}$$

The integral of the square of the residual, $e_{J,t}$, is

$$I = \frac{2}{n}\int_0^n e_{J,t}^2\,dt = \frac{2}{n}\int_0^n (y_t - y_{J,t})^2\,dt = \frac{2}{n}\int_0^n (y_t^2 - 2y_t\,y_{J,t} + y_{J,t}^2)\,dt$$

Noting the well-known integrals

$$\int_0^n \sin p\omega t \sin r\omega t \, dt = \int_0^n \cos p\omega t \cos r\omega t \, dt = \int_0^n \sin p\omega t \cos r\omega t \, dt = 0, \quad p \neq r$$

$$\frac{2}{n}\int_0^n \sin^2 p\omega t \, dt = \frac{2}{n}\int_0^n \cos^2 p\omega t \, dt = 1$$

we obtain

$$I = \tfrac{2}{n}\int_0^n y_t^2 \, dt - \left(\tfrac{1}{2}A_0^2 + \rho_1^2 + \rho_2^2 + \cdots + \rho_J^2\right) \geq 0 \quad \rho_j^2 = A_j^2 + B_j^2$$

which implies the *Bessel inequality*

$$\tfrac{1}{2}A_0^2 + \rho_1^2 + \rho_2^2 + \cdots + \rho_J^2 \leq \tfrac{2}{n}\int_0^n y_t^2$$

with equality holding if $J = \infty$. Now, from the definition of A_0, it is clear that the mean of y_t is $\mu_y = \frac{1}{2}A_0$. Hence, the variance, σ_y^2, of y_t is

$$\sigma_y^2 = \tfrac{1}{2}\int_0^n (y_t^2 - \mu_y^2) \, dt = \tfrac{1}{2}\sum_{j=1}^{\infty} \rho_j^2$$

The variance of the residual, σ_e^2, is similarly given by

$$\sigma_e^2 = \tfrac{1}{2}I = \tfrac{1}{2}(\rho_{J+1}^2 + \rho_{J+2}^2 + \cdots)$$

which can thus be made smaller than any preassigned number by choosing J large enough.

3.3 The set ρ_j, $j = 1, 2, \ldots, n/2$, is referred to as the *periodogram*. As an example of the construction of a periodogram, consider the harmonic function $y_t = A\sin(\kappa t + \beta)$, for which the Fourier coefficients are given by

$$A_j = A\sin\beta\left(\frac{\sin\pi(\kappa/\omega - j)}{\pi(\kappa/\omega - j)} + \frac{\sin\pi(\kappa/\omega + j)}{\pi(\kappa/\omega + j)}\right)$$

and

$$B_j = A\cos\beta\left(\frac{\sin\pi(\kappa/\omega - j)}{\pi(\kappa/\omega - j)} - \frac{\sin\pi(\kappa/\omega + j)}{\pi(\kappa/\omega + j)}\right)$$

so that

$$\begin{aligned}\rho_j^2 &= \frac{A^2}{\pi^2}\left(\frac{\sin^2\pi(\kappa/\omega+j)}{(\kappa/\omega+j)^2}+\frac{\sin^2\pi(\kappa/\omega-j)}{(\kappa/\omega-j)^2}-2\cos 2\beta\frac{\sin\pi(\kappa/\omega+j)\sin\pi(\kappa/\omega-j)}{\kappa^2/\omega^2-j^2}\right)\\ &= \frac{A^2}{\pi^2}\left(\frac{\sin^2\pi(\kappa/\omega+j)}{(\kappa/\omega+j)^2}+\frac{\sin^2\pi(\kappa/\omega-j)}{(\kappa/\omega-j)^2}-2\cos 2\beta\frac{\sin\pi(\kappa/\omega+j)\sin\pi(\kappa/\omega-j)}{(\kappa/\omega+j)(\kappa/\omega-j)}\right)\end{aligned} \tag{3.2}$$

Since the function $\sin\pi(\kappa/\omega-j)/(\kappa/\omega-j)$ has a maximum value of 2π as $j\rightarrow\kappa/\omega$, the expression (3.2) has a limiting value, as $j\rightarrow\kappa/\omega$, of

$$\begin{aligned}\rho_j^2 &= \frac{A^2}{\pi^2}\left(\frac{\sin^2 2\pi\kappa/\omega}{4(\kappa/\omega)^2}+4\pi^2-\frac{2\pi\cos 2\beta\sin 2\pi\kappa/\omega}{\kappa/\omega}\right)\\ &= A^2+A^2\left(\frac{\sin^2 2\pi\kappa/\omega}{4\pi^2(\kappa/\omega)^2}-\frac{2\cos 2\beta\sin 2\pi\kappa/\omega}{\pi\kappa/\omega}\right)\end{aligned}$$

The second term in this expression will be small compared to the first so that ρ_j^2 will have a maximum in the neighbourhood of $j=\kappa/\omega$. This is the fundamental idea underlying the use of the periodogram in the discovery of hidden periodicities.

The dominating term of (3.2) is $\sin^2\pi(\kappa/\omega-j)/(\kappa/\omega-j)^2$, so that ρ_j^2 will also have minima in the neighbourhood of the value of j which makes this term zero. Such zero values are obtained from the equation $\kappa/\omega-j=m$, where m is an integer. An equivalent expression for κ is $\kappa=2\pi/p$, where p is the 'true' period of the cycle. The above equation then becomes $n=p(j+m)$.

3.4 As an application of this approach, consider the function

$$y_t = 100\sin\left(\frac{2\pi t}{43}+\frac{\pi}{4}\right)$$

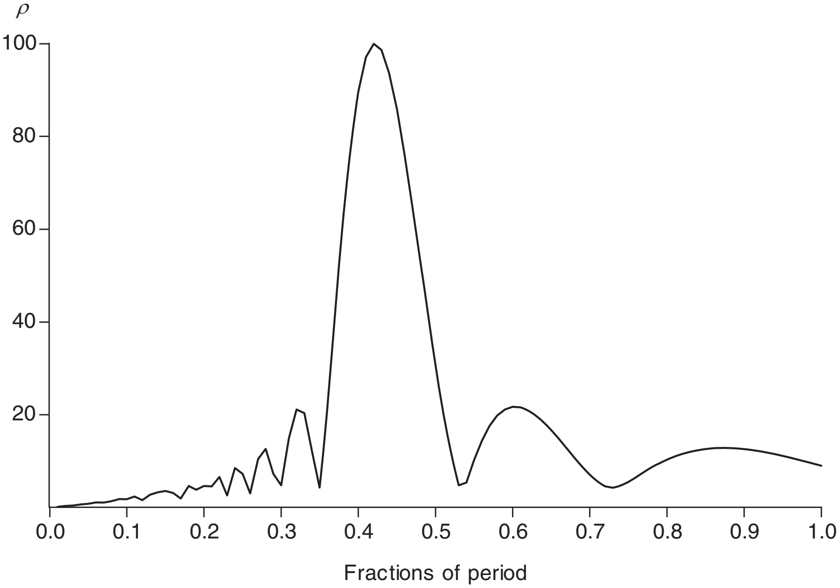
Using (3.2) and noting that $\cos 2\beta=\cos\pi/2=0$, we have

$$\begin{aligned}\rho_j^2 &= \left(\frac{100}{\pi}\right)^2\left(\frac{\sin^2\pi\left(\frac{2\pi}{43\omega}+j\right)}{\left(\frac{2\pi}{43\omega}+j\right)^2}+\frac{\sin^2\pi\left(\frac{2\pi}{43\omega}-j\right)}{\left(\frac{2\pi}{43\omega}-j\right)^2}\right)\\ &= \left(\frac{100}{\pi}\right)^2\left(\frac{\sin^2\pi\left(\frac{n}{43}+j\right)}{\left(\frac{n}{43}+j\right)^2}+\frac{\sin^2\pi\left(\frac{n}{43}-j\right)}{\left(\frac{n}{43}-j\right)^2}\right)\end{aligned}$$

If we set $n = 204$ and define $x = 2/j$ to be the *Fourier sequence* 2.00, 1.00, 0.6667, 0.50, 0.40, 0.333, ..., we have

$$\rho_x^2 = \left(\frac{100}{2\pi}\right)^2 \left(\frac{\sin^2 2\pi\left(\frac{102}{43} + \frac{1}{x}\right)}{\left(\frac{102}{43} + \frac{1}{x}\right)^2} + \frac{\sin^2 2\pi\left(\frac{102}{43} - \frac{1}{x}\right)}{\left(\frac{102}{43} - \frac{1}{x}\right)^2} \right)$$

The plot of ρ_x against x for $0 < x < 1$ is shown below and clearly reveals the existence of a period at $x = 43/102 = 0.4216$. However, minor peaks are found on either side of the major peak. This is a characteristic of periodograms and such 'shadows' should not be interpreted as being evidence of other periodicities, which clearly do not exist here.



The minimum points can be found by recalling that, when $m = 1$, $j = (n - p)/p$, thus implying that, since $x = 2/j$, the 'greater' minima is given by $x_1 = 2p/(n - p) = 86/(204 - 43) = 0.5342$. Similarly, the 'smaller' minima, obtained when $m = -1$, is given by $x_2 = 2p/(n + p) = 86/(204 + 43) = 0.3482$: the interval (x_2, x_1) may be termed the 'interference band'. The width of this band is thus $\Delta x = x_1 - x_2 = 4p^2/(n - p)(n + p) = 0.1860$ and these can all clearly be seen from the periodogram.

If the period was unknown, but we knew x_1 and x_2, and hence Δx, we could estimate p as

$$p = n\sqrt{\frac{\Delta x}{4+\Delta x}}$$

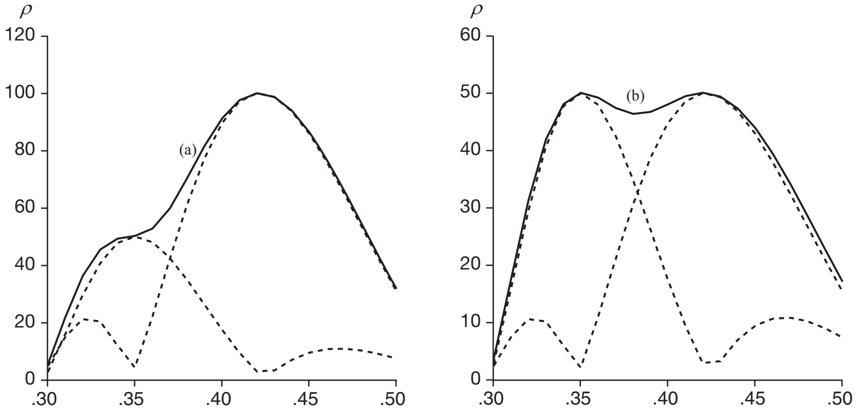
and this recovers $p = 43$.

3.5 Consider now the periodograms of the following functions (again with $n = 204$)

$$\text{(a)}\ y_t = 50\sin\left(\frac{2\pi t}{35.7}+\frac{\pi}{4}\right)+100\sin\left(\frac{2\pi t}{43}+\frac{\pi}{4}\right)$$

$$\text{(b)}\ y_t = 50\sin\left(\frac{2\pi t}{35.7}+\frac{\pi}{4}\right)+50\sin\left(\frac{2\pi t}{43}+\frac{\pi}{4}\right)$$

The first component in each function has an interference band stretching from $x_2 = 0.2979$ to $x_1 = 0.4242$ and this will seriously overlap with the interference band of the second component, which, as we have seen, extends from 0.3482 to 0.5342. Consequently, the periodograms of functions (a) and (b), which are shown with their components in the figure below, have peaks that are much too broad to have been derived from a single harmonic and thus reveal the importance of checking the theoretical width of any peak suspected to have arisen from a single harmonic.



Calculating the periodogram

3.6 With this framework in mind, the approach taken by Schuster (1906) to examine the periodicity of sunspots may be set out in the following way.

Suppose we have T observations available on the variable y: $y_1, y_2, \ldots, y_T$. These are arranged in m rows of p observations, where m and p are such that $mp \leq T \leq (m+1)p$:

$$\begin{array}{cccccc}
y_1 & y_2 & y_3 & y_4 & \cdots & y_p \\
y_{p+1} & y_{p+2} & y_{p+3} & y_{p+4} & \cdots & y_{2p} \\
y_{2p+1} & y_{2p+2} & y_{2p+3} & y_{2p+4} & \cdots & y_{3p} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
y_{(m-1)p+1} & y_{(m-1)p+2} & y_{(m-1)p+3} & y_{(m-1)p+4} & \cdots & y_{mp}
\end{array}$$

For a 'trial period' $P = p/s$, the amplitude of the periodogram is given by $\rho_P = \sqrt{A_P^2 + B_P^2}$, where

$$\begin{aligned}
A_P &= \frac{2}{pm}\sum_{j=1}^{p} M_j \cos\frac{2\pi j}{P} = \frac{2}{pm}\sum_{j=1}^{p} M_j \cos\frac{2s\pi j}{p} \\
B_P &= \frac{2}{pm}\sum_{j=1}^{p} M_j \sin\frac{2\pi j}{P} = \frac{2}{pm}\sum_{j=1}^{p} M_j \sin\frac{2s\pi j}{p} \\
M_j &= \sum_{i=1}^{m} y_{j+(i-1)p}
\end{aligned}$$

3.7 Figure 3.1 shows the annual mean number of sunspots for the years 1700 to 2007.[4] This is clearly a series with a pronounced, but certainly not deterministic, periodicity. The calculated periodogram is plotted in Figure 3.2 and shows that a maximum amplitude occurs at a period of $p = 11.1$ years, consistent with the known behaviour of the sunspot cycle and also consistent with Schuster's analysis of the shorter sample from 1749 to 1901.[5]

The further important period of around ten years was also found by Schuster, and this led him to split his sample into two subsamples of 150 years and to calculate the periodograms for both. Figure 3.3 repeats Schuster's subsample calculations and also shows the periodogram for a third subsample running from 1901 to 2007.

The features observed by Schuster are again revealed clearly. During the interval 1750 to 1825 there are peaks in the periodogram at approximately 9 and 14 years, while during the years 1826 to 1900 there is a pronounced single peak between 11 and 11.5 years. The peak for the 'post-Schuster' observations is around 10.5 years, suggesting that during the twentieth century the periodicity of the sunspot cycle may have declined slightly.[6]

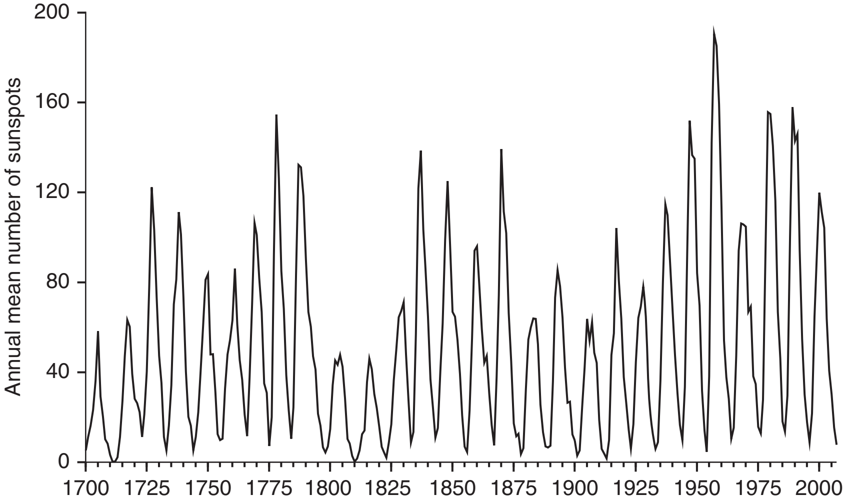


Figure 3.1 Annual mean number of sunspots, 1700–2007

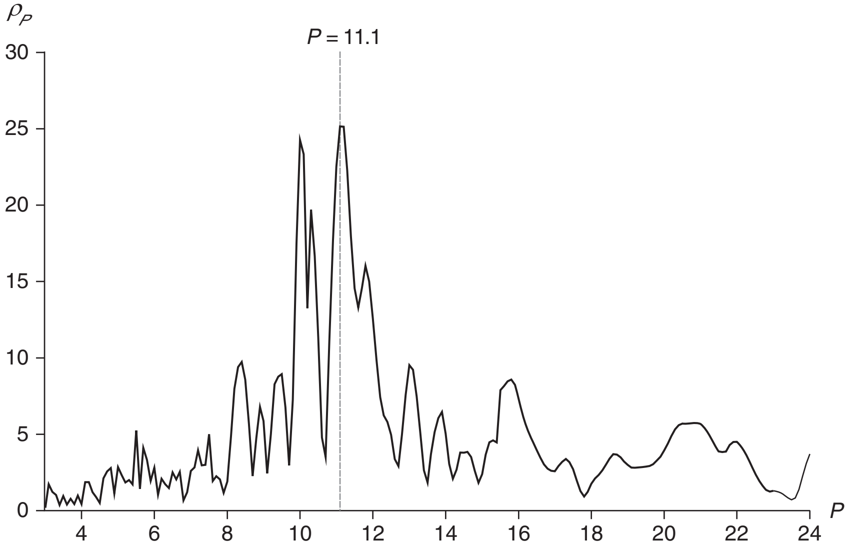


Figure 3.2 Periodogram of sunspot activity, 1700–2007

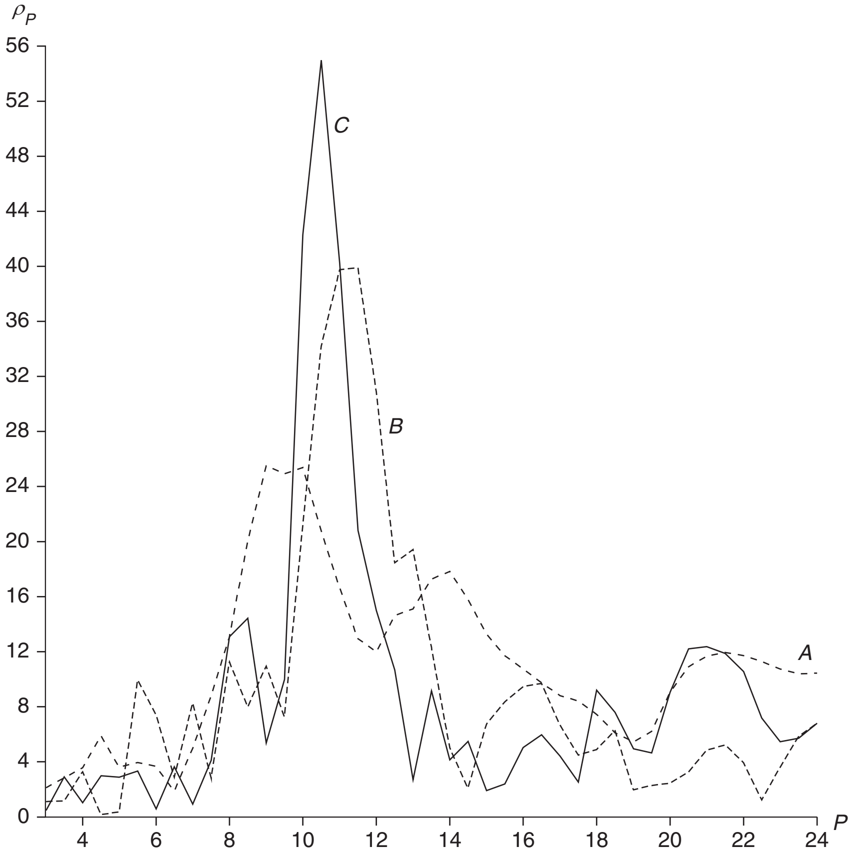
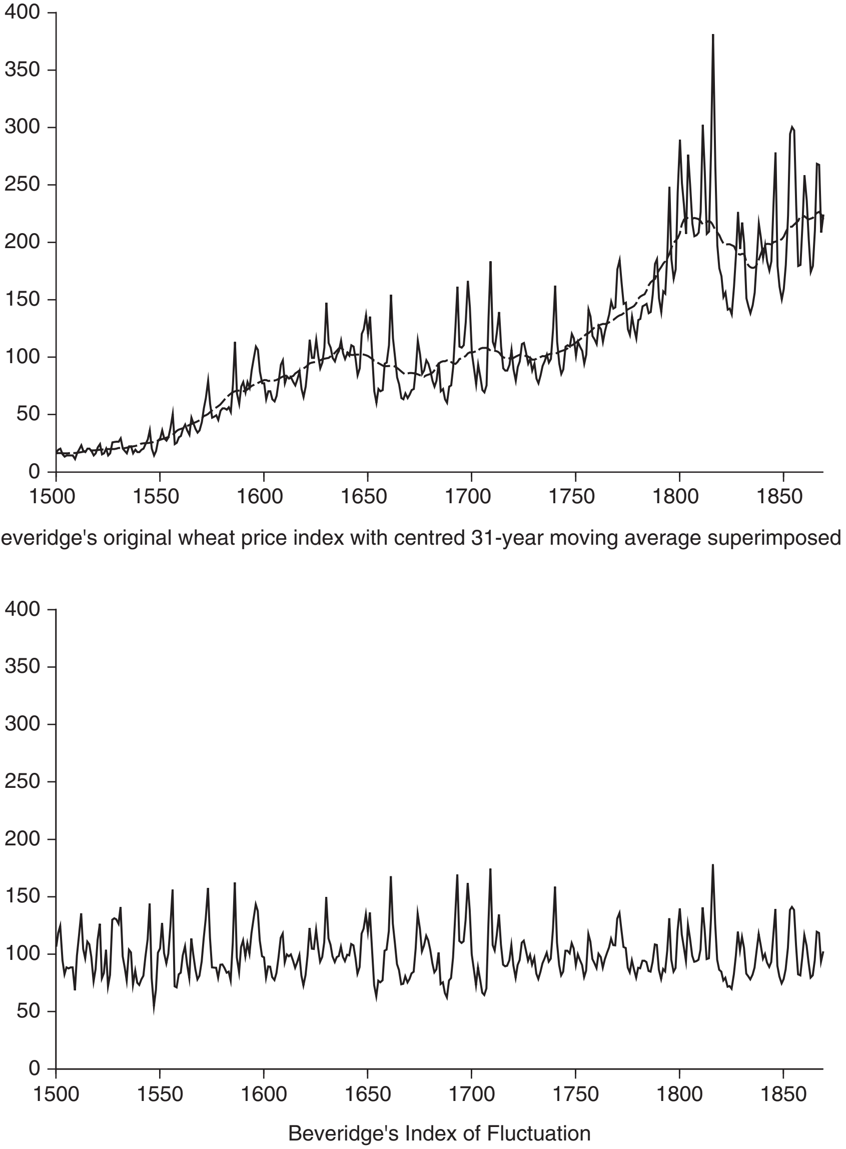


Figure 3.3 Periodograms of sunspot activity: *A*: 1750–1825; *B*: 1826–1900; *C*: 1901–2007

3.8 It was another 15 years before the next serious attempt at constructing a periodogram. This was made by (the then) Sir William Beveridge (1921, 1922) in his investigation of cycles in European wheat prices.[7] Figure 3.4 plots Beveridge's Western and Central Europe wheat price index for 1500 to 1869, as reported in Beveridge (1921, Appendix). Unlike the sunspot activity series, this price index has no clear periodicity but a pronounced secular trend. To eradicate this trend, Beveridge divided the series by a centred 31-year moving average, i.e., if the price index is denoted y_t, the detrended index, which Beveridge terms the 'Index of Fluctuation', is defined as

$$x_t = \frac{y_t}{\frac{1}{31}\sum_{j=-15}^{15} y_{t-j}} \tag{3.3}$$



Beveridge's original wheat price index with centred 31-year moving average superimposed

Beveridge's Index of Fluctuation

Figure 3.4 Beveridge's wheat price index and Index of Fluctuation, 1500–1869

The moving average is shown superimposed on the wheat index in Figure 3.4 and the Index of Fluctuation is also plotted, from which it is clear that the detrending has been successful.

The periodogram of the Index of Fluctuation is shown in Figure 3.5.[8] A peak at approximately 15 years is observed, and this is the cycle that was emphasized

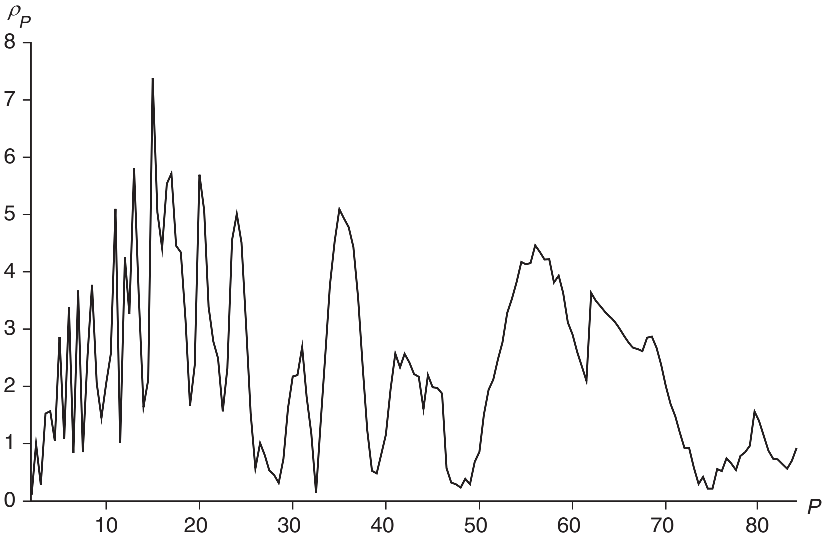


Figure 3.5 Periodogram of the Index of Fluctuation

by Beveridge, although he regarded it as resulting from combinations of shorter cycles. Many other local peaks are observed and, in particular, there appear to be longer cycles having periods of approximately 35, 54 and 68 years, which Beveridge attributed to meteorological cycles. The 35-year cycle is known as the Brűckner cycle in temperature, rainfall and barometric pressure, the 54-year cycle corresponds to one found in English rainfall and wind direction, while the 68-year cycle is close to a cycle observed in air pressure.

The interpretation of the 15-year cycle as a combination of shorter cycles caused a good deal of disquiet to the discussants of Beveridge (1922) because, as Yule pointed out, combining several cyclical components could not produce a 'composite' component with a longer cycle. Beveridge's response (page 473) appears somewhat obfuscatory and, it is fair to say, many discussants seemed unconvinced by the justification for many of the cycles that Beveridge claimed to have found, perhaps anticipating the criticisms of periodogram analysis that were to be made over the next decade or so by a variety of researchers, Yule included, and these are discussed in Chapters 5 and 13.

3.9 The 'detrending' equation (3.3) may be written in the general multiplicative form $y_t = a_t x_t$, where a_t is the 'trend function' multiplying the trend free series x_t to give the observed series y_t. The approach to detrending taken by Hooker (1901b) that was discussed in §§**2.6–2.8** takes the general additive form $y_t = a_t + x_t$. Half a century after Beveridge, the two approaches were shown to

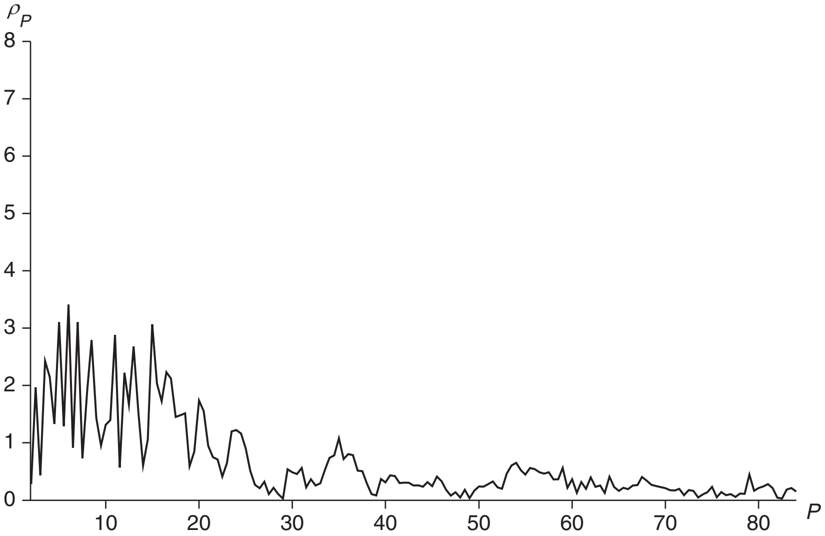


Figure 3.6 Periodogram of the first difference of the Index of Fluctuation

be approximately identical by Granger and Hughes (1971). As was discussed in §**2.9**, Hooker (1905) later considered first differencing as a detrending method. Figure 3.6 plots the periodogram for $x_t = y_t - y_{t-1}$, and shows that, in comparison to the Beveridge method of detrending, those cycles having long periods have been downgraded, with greater emphasis being placed on shorter cycles, so much so that the 15-year cycle is no longer dominant. The impact of different detrending procedures on the periodogram would not be worked out for some years but, when it was, provided further ammunition for detractors of the technique.

4

Detrending and the Variate Differencing Method: Student, Pearson and Their Critics

'Student' and the variate differencing method

4.1 The differencing approach to detrending time series proposed by Hooker (1905) and Cave-Brown-Cave (1905) (§**2.9**) was reconsidered some years later by 'Student' (1914) in rather more formal fashion.[1] Student began by assuming that y_t and x_t were *randomly distributed in time and space*, by which he meant that, in modern terminology, $E(y_t y_{t-i})$, $E(x_t x_{t-i})$ and $E(y_t x_{t-i})$, $i \neq 0$, were all zero if it was assumed that both variables had zero mean. If the correlation between y_t and x_t was denoted $r_{yx} = E(y_t x_t)/\sigma_y \sigma_x$, where $\sigma_y^2 = E(y_t^2)$ and $\sigma_x^2 = E(x_t^2)$, Student showed that the correlation between the dth differences of x and y was the same value. To show this result using modern notation, define these dth differences as

$$\Delta^d y_t = (y_t - y_{t-1})^d \quad \Delta^d x_t = (x_t - x_{t-1})^d$$

Consider first $d = 1$. Then

$$\sigma_{\Delta y}^2 = E(\Delta y_t^2) = E(y_t^2 - 2y_t y_{t-1} + y_{t-1}^2) = 2\sigma_y^2 \tag{4.1}$$

$$\sigma_{\Delta x}^2 = 2\sigma_x^2$$

$$E(\Delta y_t \Delta x_t) = E(y_t x_t + y_{t-1} x_{t-1} - y_t x_{t-1} - y_{t-1} x_t) = 2r_{yx}\sigma_y \sigma_x$$

and

$$r_{\Delta y \Delta x} = \frac{E(\Delta y_t \Delta x_t)}{\sigma_{\Delta y}\sigma_{\Delta x}} = r_{yx}$$

Thus, proceeding successively, we have

$$r_{\Delta^d y \Delta^d x} = r_{\Delta^{d-1} y \Delta^{d-1} x} = \cdots = r_{yx}$$

Student then assumed that y_t and x_t were given by polynomials in time:

$$y_t = Y_t + \sum_{j=1}^{d} \beta_j t^j \quad x_t = X_t + \sum_{j=1}^{d} \gamma_j t^j$$

where $E(Y_t Y_{t-i})$, $E(X_t X_{t-i})$ and $E(Y_t X_{t-i})$, $i \neq 0$, are all zero. Since a polynomial of order d,

$$T_t^{(d)} = \sum_{j=1}^{d} \beta_j t^j$$

becomes, on differencing d times,

$$\Delta^d T_t^{(d)} = d!\beta_d$$

we have

$$\Delta^d x_t = \Delta^d X_t + d!\beta_d, \quad \Delta^d y_t = \Delta^d Y_t + d!\gamma_d,$$

so that $\Delta^d x_t$ and $\Delta^d y_t$ are independent of time. Thus

$$r_{\Delta^d y \Delta^d x} = r_{\Delta^d Y \Delta^d X} = r_{YX}$$

and

$$r_{\Delta^{d+1} y \Delta^{d+1} x} = r_{\Delta^d y \Delta^d x}$$

leading Student to the conclusion that

> if we wish to eliminate variability due to position in time or space and to determine whether there is any correlation between the residual variations, all that has to be done is to correlate the 1st, 2nd, 3rd...*d*th differences between successive values of our variable with the 1st, 2nd, 3rd...*d*th differences between successive values of the other variable. When the correlation between the two *d*th differences is equal to that between the two $(d + 1)$th differences, this value gives the correlation required. (Student, 1914, page 180)

4.2 Student's paper, which contained only a rudimentary empirical example, was swiftly followed by several further contributions in *Biometrika* by Anderson (1914; in German), Cave and Pearson (1914), Elderton and Pearson (1915) and Ritchie-Scott (1915).[2] This led Cave and Pearson, in what was the first serious empirical application of the technique, to remark that the

> method appears to be one of very great importance, and like many new methods it has developed in a co-operative manner, which is a good reason

> for not entitling it by the name of any single contributor. We prefer to term it the *Variate Difference Correlation Method.* (Cave and Pearson, 1914, page 341; italics in original)

4.3 Equation (4.1) is easily generalized. Since

$$\Delta^d Y_t = (Y_t - Y_{t-1})^d = Y_t - {}_dC_1 Y_{t-1} + {}_dC_2 Y_{t-2} - \cdots + (-1)^d\, {}_dC_d Y_{t-d}$$

where

$${}_dC_j = \frac{d!}{(d-j)!j!}$$

is the standard combinatorial formula, then

$$\begin{aligned}\sigma^2_{\Delta^d Y} &= E(\Delta^d Y_t)^2 = E(Y_t^2) + {}_dC_1^2 E(Y_{t-1}^2) + \cdots + {}_dC_d^2 E(Y_{t-d}^2)\\ &= \sigma_Y^2({}_dC_0^2 + {}_dC_1^2 + \cdots + {}_dC_d^2)\end{aligned}$$

Since (see Anderson, 1914)

$${}_dC_0^2 + {}_dC_1^2 + \cdots + {}_dC_d^2 = {}_{2d}C_d$$

the variance of the dth difference of Y is

$$\sigma^2_{\Delta^d Y} = {}_{2d}C_d \sigma_Y^2 = \frac{2d!}{d!d!}\sigma_Y^2$$

Anderson (1914, page 278) then derived the variance of $r_{\Delta^d y \Delta^d x}$ as the expression

$$\sigma^2(r_{\Delta^d y \Delta^d x}) = \frac{(1-r_{yx}^2)^2}{(T-d)}\left(1 + \sum_{j=1}^{d} \frac{2(T-d-j)}{(T-d)}\left(\frac{d!d!}{(d-j)!(d+j)!}\right)^2\right) \qquad (4.2)$$

where T is the number of observations available.[3] Thus, for $d = 0$

$$\sigma^2(r_{yx}) = \frac{(1-r_{yx}^2)^2}{T}$$

and, consequently,

$$\sigma^2_{\Delta Y} = 2\sigma_Y^2; \quad \sigma^2(r_{\Delta y \Delta x}) = \frac{(1-r_{yx}^2)^2}{T-1}\frac{3T-4}{2(T-1)}$$

$$\sigma^2_{\Delta^2 Y} = 6\sigma_Y^2; \quad \sigma^2(r_{\Delta^2 y \Delta^2 x}) = \frac{(1-r_{yx}^2)^2}{T-2}\frac{35T-88}{18(T-2)}$$

$$\sigma^2_{\Delta^3 Y} = 20\sigma_Y^2; \quad \sigma^2(r_{\Delta^3 y \Delta^3 x}) = \frac{(1-r_{yx}^2)^2}{T-3}\frac{231T-843}{100(T-3)}$$

and so on.

4.4 Cave and Pearson noted that the ratio of the variances of the successive differences of Y had a simple form:

$$\frac{\sigma^2_{\Delta^d Y}}{\sigma^2_{\Delta^{d-1} Y}} = \frac{{}_{2d}C_d}{{}_{2(d-1)}C_{d-1}} = \frac{2d!}{d!d!}\frac{(d-1)!(d-1)!}{(2d-2)!} = \frac{2d(2d-1)}{d^2} = 4 - \frac{2}{d}$$

They therefore suggested focusing on this ratio as it avoids the need to estimate σ^2_Y, which can only practically be found from $\sigma^2_{\Delta^d y}$ after that value has become equal to $\sigma^2_{\Delta^d Y}$, i.e., after 'steadiness has set in' (*ibid.*, page 346).

4.5 The issue of estimating the successive variances was a major concern of Cave and Pearson. Anderson (1914) had implicitly assumed that the sample of available observations, T, was large enough so that he could take as approximately true that, for example,

$$\frac{1}{T-d}\sum_{t=1}^{T-d} X_t = \frac{1}{T}\sum_{t=1}^{T} X_t \quad \text{and} \quad \frac{1}{T-d}\sum_{t=1}^{T-d} X_t^2 = \frac{1}{T}\sum_{t=1}^{T} X_t^2$$

Cave and Pearson were acutely aware of the difficulties that this, and other features of data collection, posed for practical applications of the method:

> Now such relations will undoubtedly be very approximately true, if the X's are random variates uncorrelated to each other, and provided d is small compared with T. These conditions seem amply satisfied when we proceed to fourth or fifth differences in barometric pressures, taken, say, over ten or twelve years; the addition of four or five daily pressures will hardly affect sensibly either the mean or the standard deviation. But such extensive data, while not only involving a great deal of labour in the difference work are not those which, perhaps, most frequently demand the attention of the statistician, whether he be economist, sociologist or a student of scientific agriculture. In such cases it not infrequently happens that the available data only provide a range of 20 to, perhaps, at most 50 years; and we need to discover whether there is a true relationship between our variates, apart from a continuous change in both due to the time factor. At present accurate statistics of annual trade or revenue, or satisfactory annual demographic data hardly extend at most beyond a period of 50 years. Very often – under even approximately like methods of record – we shall hardly have more than twenty years' trustworthy returns. Not only has the method of record been changed, but the conditions of transit and trade may have been immensely modified and in a manner which we could not suppose to be even approximately represented by a continuous function of time. (*ibid.*, page 342)

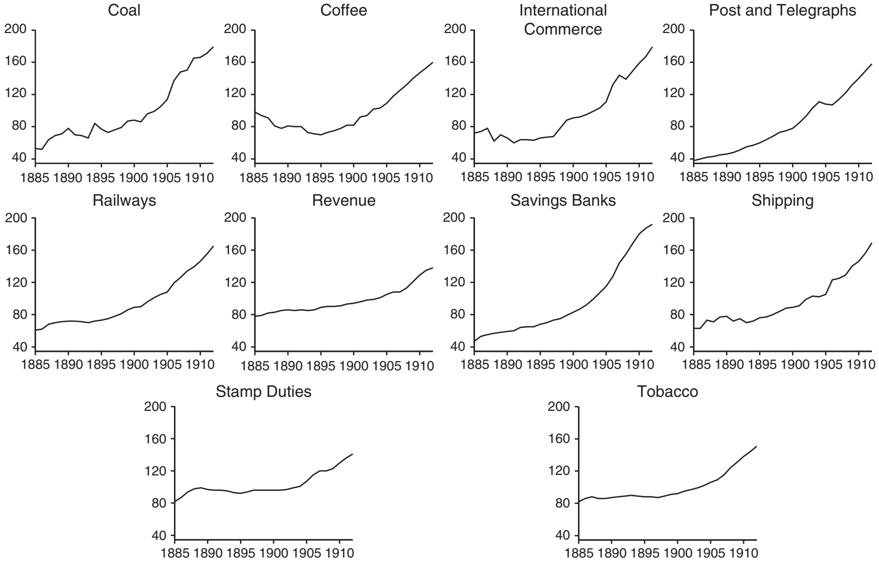


Figure 4.1 Italian economic indices, 1885–1912

4.6 Cave and Pearson therefore decided

> to illustrate the theory of the variate difference correlation method in its present stage of development on a *short* series of economic data, in order to test what approximation there is in such a short series to stability, and further how Dr Anderson's values for the successive standard deviations apply to such cases. (*ibid.*, page 342; italics in original)

To this end they used indices of ten Italian economic sectors for the years 1885 to 1912.[4] Figure 4.1 shows the set of indices and a variety of trending behaviours may be observed in their evolution. Because of these pronounced and varied trends, the high correlations between the indices shown in Table 4.1 may be viewed with some suspicion:

> the correlations ... are very high solely because the individual indices are variates increasing one and all as continuous functions of time. ... For example, the correlation between the indices for tobacco and savings is .984; are we to interpret this to signify, that, if there are large savings this means that much will be spent on tobacco? Or is this high correlation simply in whole or part spurious, merely indicating that both savings and consumption of tobacco increased markedly with the time? (*ibid.*, page 344)

Table 4.1 Correlation coefficients for Italian economic indices. Probable errors are all less than 0.03 as calculated using equation (2.10)

	Coal	Coffee	Commerce	Post and Telegraph	Railways	Revenue	Savings Banks	Shipping	Stamps	Tobacco
Coal	1									
Coffee	0.921	1								
Internat' Comm	0.970	0.969	1							
Post and Telegraph	0.957	0.884	0.949	1						
Railways	0.989	0.938	0.982	0.979	1					
Revenue	0.974	0.918	0.962	0.971	0.989	1				
Savings	0.990	0.940	0.980	0.972	0.996	0.988	1			
Shipping	0.986	0.937	0.981	0.970	0.997	0.990	0.990	1		
Stamps	0.965	0.913	0.941	0.898	0.961	0.981	0.963	0.967	1	
Tobacco	0.968	0.955	0.967	0.935	0.979	0.983	0.984	0.978	0.973	1

Table 4.2 Values of the ratio $\sigma^2_{\Delta^d y}/\sigma^2_{\Delta^{d-1} y}$ and their approach to $4-(2/d)$

d	1	2	3	4	5	6
$4-(2/d)$	2	3	3.333	3.5	3.6	3.667
Coal	0.035	2.074	3.075	3.379	3.580	3.682
Coffee	0.036	0.843	3.307	3.619	3.701	3.791
International Commerce	0.038	1.720	3.032	3.213	3.104	2.881
Post and Telegraph	0.009	0.799	1.959	2.597	3.010	3.208
Railways	0.012	0.708	2.816	3.128	3.449	3.711
Revenue	0.019	0.763	2.124	2.747	3.020	3.164
Savings Banks	0.010	0.350	2.214	3.106	3.275	3.455
Shipping	0.031	1.834	3.093	3.174	3.189	3.195
Stamp Duties	0.040	0.585	1.660	2.008	2.328	2.499
Tobacco	0.022	0.352	2.213	3.025	3.117	3.101
Mean	0.025	1.003	2.549	3.000	3.177	3.269

Cave and Pearson thus computed differences of the indices up to $d=6$ and assessed how the ratio $\sigma^2_{\Delta^d y}/\sigma^2_{\Delta^{d-1} y}$ approached $4-(2/d)$ as d increased. These calculations are reported in Table 4.2. The ratios do not 'approach steadiness' until $d=3$, after which

> there is increasing approach to agreement in the observed and theoretical values, but this approach is slow, and we believe that there is greater steadiness than is really indicated by this test. The source of this apparent unsteadiness lies we think in the relative largeness of d compared with T (i.e. at a maximum 6 as compared with 28), rather than in our not having taken sufficiently high differences. (*ibid.*, page 347)

4.7 Cave and Pearson then computed correlation coefficients for all pairs of indices at each level of differencing. We shall content ourselves with reporting the correlations (± probable errors) between tobacco and savings for $d=0, 1, \ldots, 6$, the $d=0$ correlation being the focus of concern in the second quotation in §**4.6**:

d	Correlation
0	0.984 ± 0.005
1	0.766 ± 0.065
2	-0.044 ± 0.182
3	-0.327 ± 0.181
4	-0.380 ± 0.188
5	-0.402 ± 0.197
6	-0.432 ± 0.204

It is clear that the large positive correlation between tobacco and savings at $d = 0$ does appear to be spurious: by $d = 3$ the correlation is negative and by $d = 6$ it is probably significantly so. Interestingly, Cave and Pearson found that, by this latter degree of differencing, tobacco was also significantly negatively correlated with coal (-0.514 ± 0.184), but was insignificantly correlated with all other indices, leading them to the view

> that the consumption of tobacco can hardly be considered as a measure of general prosperity; it appears to be the greatest when trade conditions are unfavourable, and in particular when savings are least and manufacturing conditions as measured by the importation of coal are slack. The result suggests the pipe of the unemployed at the street corner, rather than the increased expenditure of the fully occupied artisan. (*ibid.*, page 352)

Indeed, Cave and Pearson become quite effusive about the findings of the variate difference method, and their enthusiasm is well worth presenting as an extended quotation.

> If we turn ... to the actual correlations of the indices themselves, we find in every case an arid and scarcely undulating waste of high correlation. No one can obtain any nourishment whatever from the statement that the *Tobacco Index* is correlated with the ... *Savings Bank Index* to the extent of .984! The organic relationship between these variates is wholly obscured by the continuous increase ... of them with time. But when we proceed to sixth differences and see that the consumption of tobacco ... is associated substantially but *negatively* with savings, we seem to touch realities, and realities of some worth.... [T]here can be small doubt that to proceed from the actual correlation of such indices to the correlations of their higher differences gives the feeling of clearing away the sand of the desert, and reaching all the ordered arrangements of an excavated town below; the slight undulations of the waste above are all really fallacious, and enable us to appreciate nothing of the actual topography of the city.
>
> The method is at present in its infancy, but it gives hope of greater results than almost any recent development of statistics, for there has been no source more fruitful of fallacious statistical argument than the common influence of the time factor. One sees at once how the method may be applied to growth problems in man and in lower forms of life with a view to measuring common extraneous influences, to a whole variety of economic and medical problems obscured by the influences of the national growth factor, and to a great range of questions in social affairs where contemporaneous change of the community in innumerable factors has been interpreted as a causative nexus, or society assumed to be at least an organic whole; the flowers in a

> meadow would undoubtedly exhibit highly correlated development, but it is not a measure of mutual correlation, and the development of various social factors has to be freed from the time effect before we can really appreciate their organic relationships. (*ibid.*, pages 353–4)

4.8 Nevertheless, Cave and Pearson were well aware that the sample of Italian indices data was comparatively small and that this led to some uncertainties and difficulties with the method:

> In the present paper we have dealt only with very sparse 'populations' (only 28 values of the variates), but this has enabled us to consider not only a very large number of correlations, but to see the practical influence of terminal conditions on our theory. This may we think be summed up in the statement that the Andersonian formulae for the standard deviations will hardly in many practical cases be more than very roughly approximated before the size of the population becomes too small to make the deductions reliable. Further in most cases our difference correlations have hardly even with the sixth differences reached a steady state. ... From an examination of the actual numerical working of the correlations, it appear to us that the terminal values are in the case of these short series of very great importance. It is further clear that the theory as given by 'Student' depends upon certain equalities which are not fulfilled in short series. (*ibid.*, page 354)

The arguments being made in this passage were formulated in a footnote (*ibid.*, page 355) as follows. Allowing for X, and hence ΔX, to have non-zero means, the latter estimated as

$$\overline{\Delta X} = \frac{1}{T-1}\sum_{t=2}^{T}(X_t - X_{t-1}) = \frac{X_T - X_1}{T-1},$$

Cave and Pearson estimated the variance of ΔX as

$$\hat{\sigma}^2_{\Delta X} = \frac{1}{T-1}\sum_{t=2}^{T}(X_t - X_{t-1})^2 - (\overline{\Delta X})^2 = \frac{1}{T-1}\left(\sum_{t=2}^{T}(X_t^2 + X_{t-1}^2)\right) - (\overline{\Delta X})^2 \quad (4.3)$$

on the assumption that

$$\sum_{t=2}^{T} X_t X_{t-1} = 0$$

Equivalently, (4.3) may be written as

$$\hat{\sigma}^2_{\Delta X} = \frac{1}{T-1}\left(2\sum_{t=1}^{T} X_t^2 - (X_1^2 + X_T^2)\right) - (\overline{\Delta X})^2$$

On defining $\hat{\sigma}_X^2 = \sum_{t=1}^{T} X_t^2/T$, this becomes

$$\begin{aligned}\hat{\sigma}_{\Delta X}^2 &= 2\hat{\sigma}_X^2 + \frac{2}{T-1}(\hat{\sigma}_X^2 - \tfrac{1}{2}(X_1^2 + X_T^2)) - (\overline{\Delta X})^2 \\ &= 2\hat{\sigma}_X^2 + \frac{2}{T-1}\left(\hat{\sigma}_X^2 - \tfrac{1}{2}\left(X_1^2 + X_T^2 + \frac{1}{T-1}(X_T - X_1)^2\right)\right)\end{aligned}$$

Since, on average,

$$X_1^2 + X_T^2 \approx (X_T^2 - X_1^2)^2 \approx 2\hat{\sigma}_X^2$$

it follows that

$$\hat{\sigma}_{\Delta X}^2 = 2\hat{\sigma}_X^2\left(1 + \frac{1}{(T-1)^2}\right)$$

Although for large T, the Anderson formula $\sigma_{\Delta X}^2 = 2\sigma_X^2$ is clearly recovered, for small T there could obviously be discrepancies, and particularly so for the more complex formulae holding for the higher-order differences.

4.9 Elderton and Pearson (1915) provided a further application of the variate difference method, this time analysing death rates in the first five years of life for the period 1850 to 1912. Variate differencing was used to annihilate any trend movement in infant mortality, which they attributed to environmental factors, and they contrasted this approach favourably with the partial correlation analysis that Pearson (1912) had previously employed, concluding that

> for both sexes, a heavy deathrate in one year of life means a markedly lower deathrate in the same group in the following year of life, and that this extends to a lessened degree to the year following that, but is not by the present method easy to trace further. It is difficult to believe that this important fact can be due to any other source than the influence of natural selection, i.e., a heavy mortality leaves behind it a stronger population. (Elderton and Pearson, 1915, page 506)

Persons and Yule's critiques and Pearson's response

4.10 Although Pearson and his co-workers were concerned about the impact that short series may have on the variate difference method (see §**4.8**), in general they were extremely enthusiastic about using the technique to attack problems across a wide range of areas that suffered from the 'time-correlation' problem:

> there is small doubt that it is the most important contribution to the apparatus of statistical research which has been made for a number of years past.

> Its field of application to physical problems alone seems inexhaustible. We are no longer limited to the method of partial correlation, nor compelled to seek for factors which rendered constant will remove the changing influence of environment. (*ibid.*, page 489)

Nevertheless, it was not long before two major critiques of the method were published – Persons (1917) and Yule (1921) – and these prompted a detailed rejoinder from Pearson and Elderton (1923). The points of disagreement were essentially fourfold:[5]

(i) Persons and Yule, writing primarily from an economic perspective, naturally saw a time series as made up of a number of components, the isolation and analysis of which was of fundamental concern to them; Pearson and Elderton, from a predominantly medical/geneticist viewpoint, were more concerned with eradicating common time varying factors so that attention could be focused on 'organic' relationships.

(ii) Persons, in particular, but also Yule to a lesser extent, emphasized the underlying assumptions of Student and Anderson (see §§**4.1–4.3**), arguing that these were too restrictive in many applications; Pearson and Elderton responded by claiming that these assumptions were not critical and could be, and indeed had been, generalized.

(iii) Yule was more exercised by the potential for short-period fluctuations to confound the effects of differencing, with Pearson and Elderton counter-claiming that this possibility had been over-emphasized by Yule.

(iv) Persons felt that polynomials in time were too limited to offer an adequate representation of the underlying trend for many time series and that other possibilities should be investigated; Pearson and Elderton, although first appearing to dispute this, effectively conceded this point when they reanalysed the infant mortality data (§**4.9**) using more flexible 'smooths' taken from the actuarial literature (see Chapter 10 for related methods).

As these disagreements lie at the heart of many recurring issues in time series analysis, we shall discuss each of them in some detail.

4.11 Persons (1917) was one of the first statisticians to explicitly consider the decomposition of a time series into unobserved components.

> The items of annual time series of economic data may be conceived to be constituted of the following elements or component parts:
>
> First, the *secular trend* or growth element due to the increase of population and development of industry;

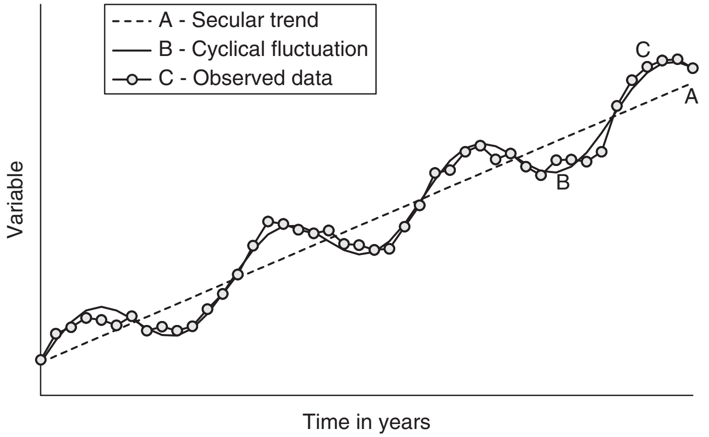


Figure 4.2 Idealistic representation of a time series as the sum of trend, cyclical and irregular components

> Second, *cyclical fluctuations*, extending over a number of years and having a greater or less degree of periodicity, due to the alternating periods of business prosperity and depression;
>
> Third, *irregular fluctuations* from year to year due to the influence of accidental or, at any rate, unpredictable events such as inventions, striking changes in fashion, or war. (Persons, 1917, page 619; italics in original)

In the spirit of Persons (1917, Figure 1), an idealistic representation of a time series containing these three components is shown as Figure 4.2. Of course, as emphasized by Persons, the secular trend may be other than a straight line and the cyclical fluctuations could be more complicated than the simple sine curve shown here.

Persons viewed the secular trend, or 'normal growth element', as increasing or decreasing regularly according to some principle and, consequently, did not think that it should 'fit' the cyclical or irregular fluctuations of the data. Suppose that the trend is linear, $\beta_0 + \beta_1 t$, so that the cycle (plus irregular) is $x_t - \beta_0 - \beta_1 t$. The first difference of the cycle is thus $\Delta x_t - \beta_1$, which differs from the first difference of the observed series x_t by just a constant. Hence any correlations calculated from the first differences of the cycles of two 'linearly detrended' series will be identical to the correlations calculated from the first differences of the observed series themselves.

Persons then derived a further interesting result. Consider the deviation of x_t from a three-year moving average centred on x_t: $x_t - (x_{t+1} + x_t + x_{t-1})/3$. This is equivalent to $-\frac{1}{3}(x_{t+1} - 2x_t + x_{t-1}) = -\frac{1}{3}\Delta^2 x_{t+1}$, so that the correlation between the second differences of two series will be identical to the correlation between the deviations from three-year moving averages. These considerations led Persons to the view that

> [i]n general, significant coefficients of correlation for the raw figures of two series indicate the similarity of the growth elements of the two series, *if large growth elements exist*. The existence or non-existence of such elements is readily determined graphically or by fitting a simple function to the data.
>
> Significant coefficients of correlation for first differences indicate that the cyclical fluctuations synchronize, *if there be cyclical fluctuations*. Evidence of such cycles may be secured by plotting the deviations from the assumed linear trend.
>
> Significant coefficients of correlation for second and in some cases higher differences indicate, in general, that the irregular fluctuations synchronize. Coefficients for higher differences of short series contain a large spurious element which increases with the order of the difference. This element is due to the tendency of the items to alternate in sign. (*ibid.*, page 622: italics in original)

Yule (1921) took a similar stance, averring that

> [t]he essential difficulty of the time correlation problem is the difficulty of isolating for study different components in the total movement of each variable: the slow secular movement, probably non-periodic in character or, if periodic, with a very long period; the oscillations of some ten years' duration, more or less, corresponding to the wave in trade; the rapid movements from year to year which give the appearance of irregularity to the curve in a statistical chart and which may in fact be irregular or may possess a quasi-periodicity of some two years duration; the seasonal movements within the year, and so on. (page 501)

He then contrasted this approach with what he took to be the variate difference perspective.

> [A]nd if 'Student' desires to remove from his figures secular movements, periodic movements, uniform movements, and accelerated movements – well, the reader is left wondering with what sort of movements he *does* desire to deal. (*ibid.*, page 502; italics in original)

> ['Student'] desires to find the correlation between x and y when every component in each of the variables is eliminated which can well be called a function of the time, and nothing is left but residuals such that the residual of a given year is uncorrelated with those that precede or that follow it. (*ibid.*, page 503)

Yule left the reader in no doubt as to which position he preferred.

> But which view of the problem is correct? Do we want to isolate oscillations of different durations, two years, ten years, or whatever it may be, or nothing but these random residuals? Personally I cannot hesitate for a moment as to the answer. The only residuals which it is easy to conceive as being totally uncorrelated with one another in the manner supposed are errors of observation, errors due to the 'rounding off' of index numbers and the like, fluctuations of sampling, and analogous variations. And an error of observation or fluctuation of sampling in x would normally be uncorrelated with an error of observation or fluctuation in y, so that if the generalized variate-difference method did finally isolate nothing but residuals of the kind supposed I should expect it in general to lead to nothing but correlations that were zero within the limits of sampling. ... [T]he problem is not to isolate random residuals but oscillations of different durations, and unless the generalized method can be given some meaning in terms of oscillations it is not easy to see what purpose it can serve. (*ibid.*, page 504)

In their response to these views, Pearson and Elderton were typically combative:

> Now we know that if we correlate the falling phthisis deathrate with the falling birthrate we shall have a correlation of the order 0.9. But no one is likely to believe there is an *organic* relationship between the two of this order, – any more than one believes that the correlation between the cancer deathrate and the increasing expenditure on apples per head of the population, the value of which is 0.89, is a true organic relationship, i.e. is due to one or more common factors in the two variates. Such high correlations as arise from common growth or decline with time, when interpreted as causal or semicausal relationships, are in our opinion perfectly idle, indeed are only apt to be mischievous, and we shall reach nothing, or less than nothing – knighthoods, – by the investigation of them.
>
> But when we take the *apparently* random deviations from the secular trend, it does seem a perfectly legitimate problem to ask: is there any relationship between them?
>
> If the deviations of two variates from their secular trends be X and Y, we want to discover their correlation r_{xy}. All are agreed, we think, as to the desirability of finding this correlation, – including even Mr. Yule, although he apparently confesses that he cannot find any source for such correlation except in common periodic terms. Now the real problem before us is this: Having by means of a high order parabola or an adequate smooth got rid of the secular trend, will the variate difference method give us r_{xy} or what does it give us?

> ... [I]n our opinion the deviations X and Y from the secular trends of the two variates, such as occur in vital and economic statistics, are dependent on factors which are obviously non-periodic in character. They are summed up in sanitation, legislation, new routes and methods of transport, over- and under-production, new methods of agriculture, wars, famines, transfer of population and thousands of other factors which make up civilized human life. We might define them as 'historical factors', history takes place in time, but its events are not mathematical functions of the time, still less periodic functions, whatever folk experience may whisper about history repeating itself.
>
> It is only legitimate to call the effects of these historical factors random fluctuations, if that term is used in a special sense as Mr Yule appears to use it, i.e. for everything which is not due to a periodic variation. Such 'random fluctuations' are by no means as Mr Yule would seem to suggest due only to errors of observation or the deviations of random sampling; they are due to non-periodic causes which may affect both the ... variates or may not. The question is to what extent have [X] and [Y] common causes behind their fluctuations, apart from growth with time. We think this is a perfectly legitimate question to ask, and that in asking it we are not open to the insinuation, contained in the term 'random disturbances', of asking whether pure chance fluctuations are or are not correlated. (Pearson and Elderton, 1923, pages 282–3)[6]

This extended quote makes it clear that Pearson and Elderton conceived of a time series as being decomposed into just two components, a 'catchall' component comprising the secular trend and periodic fluctuations, modelled by a polynomial in time, and a random component – in modern time series parlance they work within a trend stationary specification à la Nelson and Plosser (1982): see §**16.2**. Persons and Yule, on the other hand, preferred to decompose the series into its secular trend, which will typically be a simple linear function of time or something similar, and cyclical (periodic) and irregular components, thus giving rise to an unobserved components formulation. The elimination of the signal by differencing shows the variate differencing procedure to be an early forerunner of the Box–Jenkins approach to modelling nonstationarity (see §**10.16–10.19**), and is indeed mentioned by them (Box and Jenkins, 1970, page 89), although they state that the motivation and objectives of the procedure were quite different from their own differencing approach. In contrast, the Persons–Yule unobserved component formulation is what would now be referred to as a *structural model* (Nerlove, Grether and Carvalho, 1979, Harvey, 1989). From a bivariate perspective, using variate differencing prior to correlating a pair of time series was a forerunner of the 'prewhitening' approach (see Pierce, 1977), while the Persons–Yule idea of correlating the components has its descendants in Mills (1982b) and Watson (1986).

4.12 Persons was concerned about the underlying assumptions made by Student (cf. §**4.1**).

> It is my contention that these assumptions are such that as cannot be retained in applying the method to the most common types of problems. For instance the pairing of items of two time series is made possible by the position of those items in time either because they occur in the same time interval (concurrent) or in definitely related intervals (lag). Our problem may be, and usually is, not only to determine the correlation but to find what pairings give the maximum correlation. In such case the assumption that only one pairing is significant vitiates the conclusion at the outset. The writers on the variate difference correlation method all assume that 'the true r_{XY}' is for pairs concurrent in time. (Persons, 1917, page 604)

Persons thus claimed that the variate difference approach would necessarily rule out the results obtained by Hooker when looking at the relationship between trade and the marriage rate that was discussed in §§**2.6–2.9**.[7]

After showing how oscillatory movements in a time series cannot be removed by differencing through using artificially constructed observations, Persons then investigated a set of 21 American economic series observed over the period 1879 to 1913. Working within the framework of §**4.11**, Persons assumed that the secular trend was linear, fitted a straight line by least squares or by the method of moments, and designated the 'deviations of the raw figures from the lines of secular trend' as 'cycles'.[8] Nine of these cycles were found to 'synchronize' and these were combined into a 'business barometer'.[9] This barometer was then cross-correlated, for lags running for three years in either direction, with several of the series using up to sixth differences. Persons remarked that, while the correlation coefficient for concurrent items held fairly steady for all differences, the correlations showed a marked tendency to alternate in sign as 'successive degrees of lag are taken in either direction', which led him to ask

> (w)hat is the explanation of the observed steadiness, and of the alternation of sign of coefficients for various degrees of lag? 'Student' believes that the steadiness is due to the random distribution, with respect to time, of the differences. The alternation in sign is a phenomenon not noticed, or if noticed not considered, by the writers on the subject. (*ibid.*, page 609)

Persons then showed that

> (i)f consecutive items of a series alternate in sign the first and higher differences will also alternate in sign and the resulting items will increase

> numerically as the order of the difference increases. A succession of like signs may persist with the first and higher differences but the number resulting will be smaller numerically than those resulting where the sign alternates. Where the variate difference method is applied to two short series we may, therefore, expect the terms alternating in size to be of dominating influence upon the coefficient of correlation. Also when a lag is taken in either direction the coefficients will tend to alternate in sign. (*ibid.*, pages 609–610)

Persons examined the phenomenon of alternating signs in the cross-correlations of the various differences of his series, finding a 'marked tendency to alternation in sign' for second and higher differences. He then investigated what would happen for two random series. In an early example of a simulation experiment, Persons constructed two series of 35 random observations and then calculated cross-correlation coefficients for up to eighth differences.[10]

Table 4.3 recreates Persons' simulation and, for comparison, also shows the results for a second, much larger, sample of 1,000 random observations. For the smaller sample of just 35 observations (that available for the data being analysed by Persons), there is indeed a persistent alternation in the sign of the cross-correlations as lags are taken in either direction from zero. This is because the two series being correlated have a tendency to alternate in sign on differencing so that, as further differences are taken, the observations that alternate in sign become dominating ones: lagging either series in either direction then brings a different set of signs into correspondence. Moreover, as predicted by Persons, the absolute magnitudes of the correlations increase, although they at best only become marginally significant by the eighth difference. For the much larger sample, such patches of alternating signs are no longer able to dominate and all cross-correlations at all differences remain close to zero.

Persons then made an analytical investigation of the conditions under which cross-correlations at a particular lag would remain 'steady' as successive differences were taken. The Student assumptions were found to be sufficient but not necessary, for such steadiness would also occur if certain conditions, termed by Persons the 'balancing conditions' and involving the product sums $\sum x_t x_{t-1}$, $\sum y_t y_{t-1}$ and $\sum x_t y_{t+1} + \sum x_t y_{t-1}$, were to be approximately satisfied. Indeed, Persons (1917, page 615) claimed that, for time series of the length that he was considering, 'the conditions ... are apt to occur and be the cause of any stability of the coefficients of correlation between multiple differences', backing this up with computational evidence that this had, in fact, actually happened within his data set.

Yule (1921) focused attention on the correlation induced into a series by differencing. Using the set-up of §**4.1**, it can be seen that

$$E(\Delta y_t \Delta y_{t-1}) = E(y_t y_{t-1} - y_t y_{t-2} - y_{t-1}^2 + y_{t-1} y_{t-2}) = -\sigma_y^2$$

Table 4.3 Cross-correlation coefficients between two random series and their differences

d	$r(-2)$	$r(-1)$	$r(0)$	$r(1)$	$r(2)$
0	−0.04 (0.17)	−0.07 (0.17)	+0.13 (0.17)	−0.04 (0.17)	+0.11 (0.17)
1	+0.02 (0.21)	−0.08 (0.21)	+0.17 (0.20)	−0.14 (0.20)	+0.22 (0.20)
2	+0.10 (0.24)	−0.09 (0.24)	+0.18 (0.23)	−0.21 (0.23)	+0.30 (0.22)
3	+0.15 (0.26)	−0.15 (0.26)	+0.20 (0.26)	−0.28 (0.25)	+0.36 (0.23)
4	+0.20 (0.28)	−0.22 (0.27)	+0.26 (0.27)	−0.34 (0.25)	+0.40 (0.24)
5	+0.23 (0.29)	−0.28 (0.28)	+0.32 (0.28)	−0.39 (0.27)	+0.40 (0.28)
6	+0.24 (0.31)	−0.31 (0.29)	+0.38 (0.28)	−0.41 (0.27)	+0.39 (0.28)
7	+0.24 (0.32)	−0.33 (0.31)	+0.42 (0.28)	−0.42 (0.28)	+0.38 (0.29)
8	+0.25 (0.34)	−0.35 (0.32)	+0.45 (0.29)	−0.43 (0.29)	+0.37 (0.24)
0	−0.02 (0.03)	−0.03 (0.03)	−0.01 (0.03)	−0.01 (0.03)	+0.03 (0.03)
1	+0.00 (0.04)	−0.01 (0.04)	+0.00 (0.04)	−0.02 (0.04)	+0.04 (0.04)
2	+0.01 (0.04)	−0.01 (0.04)	+0.01 (0.04)	−0.02 (0.04)	+0.03 (0.04)
3	+0.01 (0.05)	−0.01 (0.05)	+0.02 (0.05)	−0.03 (0.05)	+0.03 (0.05)
4	+0.01 (0.05)	−0.01 (0.05)	+0.02 (0.05)	−0.03 (0.05)	+0.02 (0.05)
5	+0.01 (0.05)	−0.01 (0.05)	+0.02 (0.05)	−0.03 (0.05)	+0.02 (0.05)
6	+0.01 (0.06)	−0.02 (0.06)	+0.02 (0.06)	−0.03 (0.06)	+0.01 (0.06)
7	+0.01 (0.06)	−0.02 (0.06)	+0.02 (0.06)	−0.02 (0.06)	+0.01 (0.06)
8	+0.01 (0.06)	−0.02 (0.06)	+0.02 (0.06)	−0.02 (0.06)	+0.01 (0.06)

Top panel: $T = 35$; bottom panel: $T = 1000$. $r(k)$ denotes the correlation between Y_t and X_{t-k}. Figures in parentheses are standard errors computed using equation (4.2).

If the correlation between $\Delta^d y_t$ and $\Delta^d y_{t-k}$ is denoted ${}_d r_y(k)$, then clearly

$$ {}_1 r_y(1) = \frac{E(\Delta y_t \Delta y_{t-1})}{\sqrt{E(\Delta y_t^2)E(\Delta y_{t-1}^2)}} = \frac{-\sigma_y^2}{2\sigma_y^2} = -\tfrac{1}{2} $$

so that the adjacent differences are (negatively) correlated even though the original series is random. Note, though, that this correlation does not extend any further than adjacent observations, for

$$E(\Delta y_t \Delta y_{t-2}) = E(y_t y_{t-2} - y_t y_{t-3} - y_{t-1} y_{t-2} + y_{t-1} y_{t-3}) = 0$$

implying that ${}_1r_y(2) = 0$ and, by extension, ${}_1r_y(k) = 0$ for $k > 1$. Having shown this, Yule then generalized these results to dth differences:

$${}_dr_y(1) = -\frac{d}{d+1}, \quad {}_dr_y(2) = \frac{d(d-1)}{(d+1)(d+2)}, \quad \cdots$$

$${}_dr_y(k) = (-1)^k \frac{d(d-1)\cdots(d-k+1)}{(d+1)(d+2)\cdots(d+k)} = (-1)^k \frac{d!d!}{(d-k)!(d+k)!} \quad k \leq d$$

with ${}_dr_y(k) = 0$ for $k > d$.[11] Hence dth differences of a random series y have nonzero correlations between observations up to d intervals apart, with these correlations declining and alternating in sign, being negative for odd d:

> The correlations start with a high negative value between adjacent terms, and the values slowly die away with alternating signs. Differencing a random series tends therefore to produce a series in which the successive terms are alternately positive and negative. (Yule, 1921, page 521)

Yule thus echoed Persons' findings of alternating signs of correlations but these are now between the lagged differences of a random series as well as between the differences of two random series.

Pearson and Elderton's response to this critique was to claim that Elderton and Pearson (1915, section (8)) had, in fact, investigated cross-correlations using their mortality data, but agreed that the non-randomness of x and y was a valid criticism:

> . . . there seems to us two main criticisms of the Variate Difference correlation method: . . . that we cannot assume that X is solely correlated with a single Y, and that the series of X's is not intercorrelated, nor the series of Y's. We think that this is a valid criticism which has to be met, either by showing that there are many cases in which the causes which produce the X's and Y's do not last over more than one interval, or else by enlarging our method and supposing that correlations between the X's and Y's of the above character really exist. (Pearson and Elderton, 1923, page 286)

They then proceeded to generalize the variate difference method when the original series were 'intercorrelated', utilizing results provided by Egon Pearson (see note 9).[12] Thus we extend our notation to denote the correlation between x_t and x_{t+k} as $r_x(k)$ and the correlation between x_t and y_{t+k} as $r_{xy}(k)$, so that the correlation between x_t and y_{t-k} is $r_{xy}(-k)$, etc. (where, implicitly, $r_x(k) \equiv {}_0r_x(k)$ and so on). Asymmetry is allowed, so that $r_x(k)$ is not necessarily equal to $r_x(-k)$, and so on. The key result (Pearson and Elderton, 1923, page 294, equation (xviii)) is

$$\frac{E(\Delta^d x_t \Delta^d y_{t+k})}{\sigma_x \sigma_y} = \frac{(2d)!}{d!d!}\phi(d, r_{xy}, k) \tag{4.4}$$

where, for a generic sequence of correlations $r(j)$, $j = 0, \pm1, \pm2, \ldots, \pm d$,

$$\phi(d, r, k) = \sum_{j=-d}^{d} (-1)^{|j|} \frac{d!d!}{(d-j)!(d+j)!} r(j+k)$$

From (4.4), it follows that (cf. §**4.4**)

$$\frac{E(\Delta^d x_t \Delta^d y_{t+k})}{E(\Delta^{d-1} x_t \Delta^{d-1} y_{t+k})} = \left(4 - \frac{2}{d}\right) \frac{\phi(d, r_{xy}, k)}{\phi(d-1, r_{xy}, k)} \tag{4.5}$$

and, on defining $\phi(d, r, k) = r(0)\,\phi(d, r', k)$, where $r' = r/r(k)$, the correlation between $\Delta^d x_t$ and $\Delta^d y_{t+k}$ can be expressed as

$${}_d r_{xy}(k) = \frac{E(\Delta^d x_t \Delta^d y_{t+k})}{\sqrt{E(\Delta^d x_t)^2 E(\Delta^d y_{t+k})^2}} = r_{xy}(k) \frac{\phi(d, r'_{xy}, k)}{\sqrt{\phi(d, r_x, 0)\,\phi(d, r_y, k)}} \tag{4.6}$$

If $r'_{xy}(k) = r_x(k) = r_y(k) = 0$ for all $k \neq 0$, then ${}_d r_{xy}(0) = r_{xy}(0)$ and ${}_d r_{xy}(k) = 0$, which recovers the original Student result. This result is also recovered if, more generally, $r'_{xy}(k) = r_x(k) = r_y(k)$, i.e., if the correlations die out at the same rate, which implies that $r_{xy}(k) = r_{xy}(0) \times r_x(k)$

Suppose now that the correlations decay at different rates, for example as

$$r_{xy}(k) = r_{xy}(0)\varepsilon_{xy}^k \quad r_x(k) = \varepsilon_x^k \quad r_y(k) = \varepsilon_y^k$$

We then have

$$\phi(d, r, k) = 2\varphi(d, \varepsilon, k) - 1$$

where

$$\varphi(d, \varepsilon, k) = \sum_{j=0}^{d} \frac{d!d!}{(d-j)!(d+j)!} \varepsilon^j$$

Equation (4.5) then becomes

$$\frac{E(\Delta^d x_t \Delta^d y_{t+k})}{E(\Delta^{d-1} x_t \Delta^{d-1} y_{t+k})} = \left(4 - \frac{2}{d}\right) \frac{2\varphi(d, r_{xy}, k) - 1}{2\varphi(d-1, r_{xy}, k) - 1}$$

and similarly

$$\frac{E(\Delta^d x_t)^2}{E(\Delta^{d-1} x_t)^2} = \left(4 - \frac{2}{d}\right) \frac{2\varphi(d, r_x, 0) - 1}{2\varphi(d-1, r_x, 0) - 1}$$

Thus (4.6) is

$${}_d r_{xy}(k) = r_{xy}(k) \frac{2\varphi(d, r'_{xy}, k) - 1}{\sqrt{(2\varphi(d, r_x, 0) - 1)(2\varphi(d, r_y, k) - 1)}}$$

If both ε_x and ε_y are larger than ε_{xy}, so that the intercorrelations within the series decay slower than the cross-correlations, which is plausible in many cases, then ${}_d r_{xy}(k) > r_{xy}(k)$. On the other hand, if ε_x and ε_y are negative, as would be the case if there were two-period oscillations, then the sign of this inequality would be reversed, leading Pearson and Elderton (1923, page 298) to conclude that '(w)e think the theory of the variate difference method thus generalized will meet such criticisms as those of Persons, which are valid. It may be harder to meet those of Mr Yule', to which we now turn.

4.13 Yule (1921) considered taking differences of the periodic function

$$y_t = \rho \sin\left(2\pi \frac{t+\alpha}{n}\right) = \rho \sin\left(\frac{2\pi t}{n} + \frac{2\pi\alpha}{n}\right) \tag{4.7}$$

where (cf. §§**3.1–3.5**) ρ is the amplitude of the sine wave, n is the period, and α is the phase, whose effect is to advance the peak of the sine function by $n\alpha/2\pi$ periods. The first difference of interval h of (4.7) is

$$\begin{aligned} \Delta y_{t+h} &= \rho\left(\sin\left(2\pi\frac{t+\alpha+h}{n}\right) - \sin\left(2\pi\frac{t+\alpha}{n}\right)\right) \\ &= 2\rho \sin\left(\pi\frac{h}{n}\right)\cos\left(2\pi\frac{t+\alpha+0.5h}{n}\right) \\ &= 2\rho \sin\left(\pi\frac{h}{n}\right)\sin\left(2\pi\frac{t+\alpha+0.5h+0.25n}{n}\right) \end{aligned} \tag{4.8}$$

The second equality in (4.8) uses the trigonometric identity $2\cos A \sin B = \sin(A+B) - \sin(A-B)$, with $A = 2\pi(t+\alpha+0.5h)/n$ and $B = \pi h/n$, while the third equality uses $\sin(A + 0.5\pi) = \cos A$.

Thus the first difference of y is given by a sine wave of the same period as the original function but with the phase shifted by the amount $0.5h + 0.25n$

and the amplitude multiplied by the factor $2\sin(\pi h/n)$. The second difference will therefore be derived from the first difference by multiplying the amplitude by the same factor and shifting the phase by the same amount, and so on for successive differences.

Yule focused attention on the factor $2\sin(\pi h/n)$, since whether this is greater or less than unity will determine if successive differences will continually diverge (because of an increasing amplitude) or will converge with the amplitude getting smaller and smaller. It is clear that the factor will exceed unity if $n/6 < h < 5n/6$, so that if h lies in this interval, differencing will emphasize periodic fluctuations rather than eliminate them. Equivalently, this interval can be written as $6h/5 < n < 6h$, so that if $h = 1$ year, a period of between 1.2 years and 6 years will produce a diverging amplitude.

Since $2\sin(\pi h/n)$ reaches a maximum of 2 at $h = n/2$ then, for $h = 1$, a period of $n = 2$ produces the greatest increase in amplitude. For example, by taking sixth differences the amplitude will be multiplied $2^6 = 64$-fold, leading Yule to conclude that

> (t)he effect then of differencing the values of a function which is given by a series of harmonic terms is not gradually to extinguish all the terms, but selectively to emphasize the term with a period of 2 intervals; terms with a period between 2 and 6 intervals, or between 2 and 1.2 intervals have their amplitude increased, but not so largely; terms with a period between 1 and 1.2 intervals, or greater than 6 intervals, are reduced in amplitude. Further, every term is altered in phase, by an amount depending on its period. Correlations between high differences will accordingly *tend* to give the correlations between component oscillations of very short period – predominantly of a two-yearly period, in so far as such oscillations exist in the original observations, even though they may not be the most conspicuous or characteristic oscillations. (Yule, 1921, page 509; italics in original)

In responding to Yule, Pearson and Elderton adopted a slightly different framework to show that, if there was a single harmonic in y_t with a period n, then

$$\frac{E(\Delta^d y_t)^2}{E(\Delta^{d-1} y_t)^2} = 4\left(\sin\left(\frac{\pi h}{n}\right)\right)^2$$

and

$$\sigma_y^2 = \frac{E(\Delta^d y_t)^2}{2^{2d}\left(\sin\left(\frac{\pi h}{n}\right)\right)^{2d}}$$

These equations were then used to determine the period if such a single harmonic existed and to examine how far a given harmonic would account for fluctuations in the original data (after any secular trend had been removed, of course).

If two series, y_t and x_t, had common periodic terms but with different phases α and β, then Pearson and Elderton showed that

$$\frac{E(\Delta^d x_t \Delta^d y_t)}{E(\Delta^{d-1} x_t \Delta^{d-1} y_t)} = \frac{E(\Delta^d x_t)^2}{E(\Delta^{d-1} x_t)^2} = \frac{E(\Delta^d y_t)^2}{E(\Delta^{d-1} y_t)^2} = 4\left(\sin\left(\frac{\pi h}{2n}\right)\right)^2 \tag{4.9}$$

and

$${}_d r_{xy}(0) = \cos(\alpha - \beta) \tag{4.10}$$

leading them to conclude that

> we know that on 'Student's' hypothesis, the ratios of the differences given in [4.9] tend to the value 4. They would only tend to 4 in the case of a single periodic term if the period were twice the fundamental unit of time, i.e. ... for $n = 2h$. After the secular terms have been eliminated, the ratios in [4.9] and the correlation in [4.10] should always be the same in the case of a single periodic term. But in the case of correlated fluctuations, while the correlation in [4.10] would be constant, the ratio of differences – of course on 'Student's' hypothesis – would be $4 - 2/d$ and thus only *tend* to a constant value = 4. On the other hand if $r_{xy,k}$, $r_{x,k}$ and $r_{y,k}$ be not a series of zero correlations (except, of course for $k = 0$) then the ratios in [4.9] will tend more slowly to 4, because they depend on quantities like $d/(d+1)$ being equal to $(d-1)/d$ practically, as d becomes large. (Pearson and Elderton, 1923, page 300; italics in original; notation altered for consistency)

4.14 As discussed in §**4.11**, a particular interest of Persons was the correlation between the cyclical components of two series. He took great pains, however, to emphasize how such a correlation was dependent upon the method used to eliminate the secular trend:

> judgment concerning the correlation of cyclical fluctuations of two series must be preceded by elimination of the secular trend. The choice of a function to represent the secular trend, indeed the choice of the method of eliminating the trend, whether by curve fitting or otherwise, these are questions fundamental to the process. (Persons, 1917, pages 623–4)

Persons then assessed the importance of the choice of trend function by analysing two of the series used to illustrate the variance difference method by

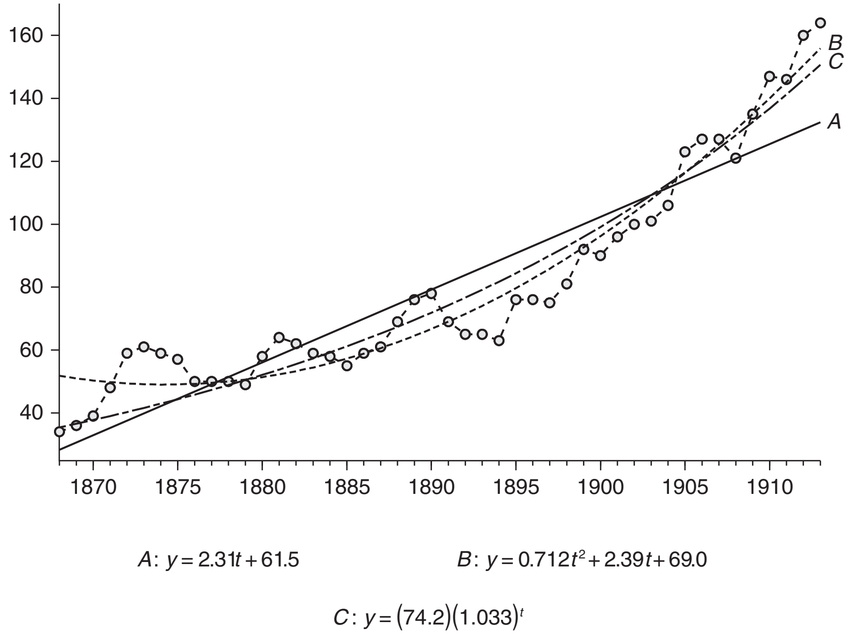


Figure 4.3 London bank clearings (in millions of pounds), 1868–1913, with straight line (*A*), parabola (*B*) and compound interest curve (*C*) fitted to data

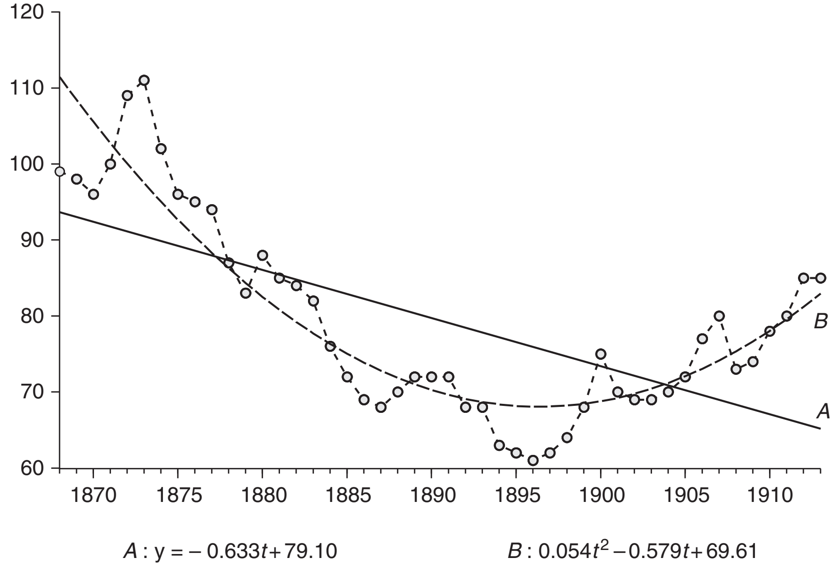


Figure 4.4 Sauerbeck's index numbers of wholesale prices, 1868–1913, with straight line (*A*) and parabola (*B*) fitted to data

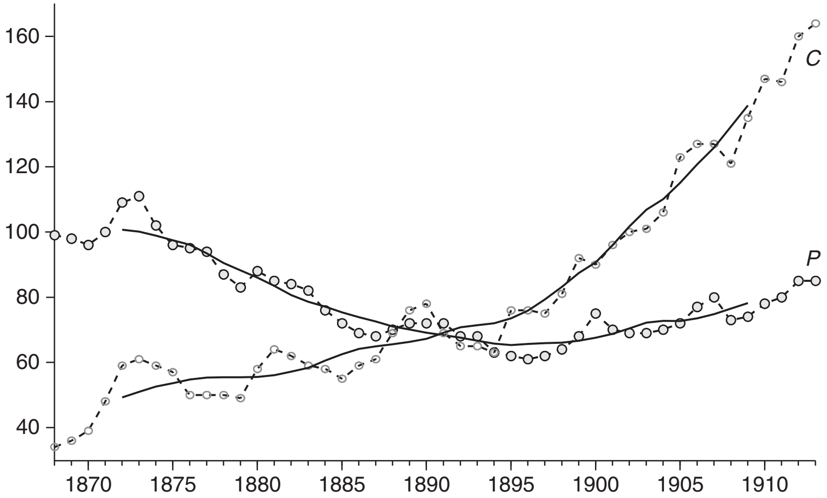


Figure 4.5 Sauerbeck's price indices (*P*) and London clearings (*C*), 1868–1913, with their respective nine-year moving averages, 1872–1909

Student (1914), London bank clearings and Sauerbeck's price index, providing the data in Persons (1917, Table IX). Because this was the first serious attempt at decomposing time series we discuss it in some detail. Figures 4.3 and 4.4 recreate Persons' figures 2 and 3, showing various trends fitted to the two series, while Figure 4.5 recreates his figure 4, in which the two series are plotted with nine-year moving averages superimposed.[13]

Figures 4.6–4.9 (Persons' figures 5–8) show the deviations from the various secular trends and hence may be regarded as alternative estimates of the cyclical fluctuations. Whatever the method used to compute these fluctuations, Persons (1917, page 625) felt able to conclude that 'fluctuations of clearings show a tendency to precede or forecast the fluctuations in prices', before going on to emphasize that '(t)he main question upon which we wish to get light is, however, the effect of the various methods of eliminating the secular trend upon the coefficients of correlation between corresponding deviations' (*ibid.*, page 629). Table 4.4 repeats Persons' calculations that underlie his Table X. Although there are some minor numerical differences in the correlations, Table 4.4 supports Persons' original conclusions.

> The coefficients of correlation for the raw figures [−.36, −.31 and −.28] show that the secular trends of prices and clearings are in opposite directions. The coefficients for the first differences of the raw figures and of all the deviations

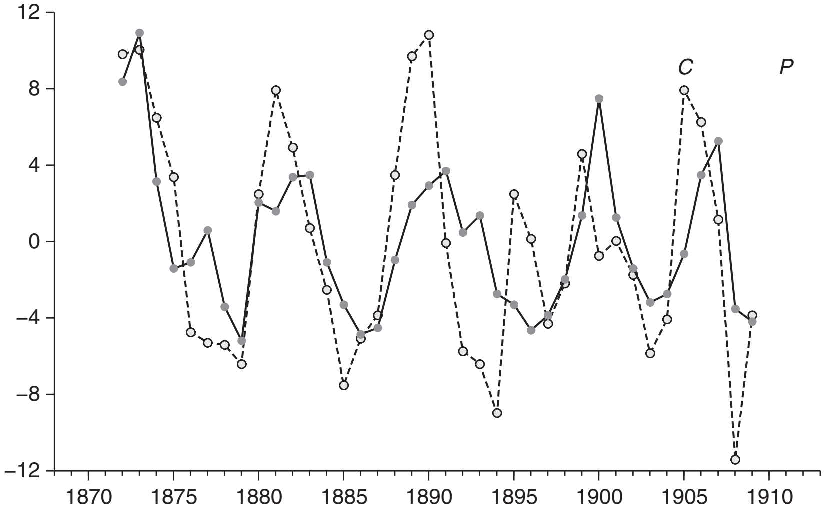


Figure 4.6 Deviations of London clearings (*C*) and Sauerbeck's prices (*P*) from their respective nine-year moving average secular trends, 1872–1909

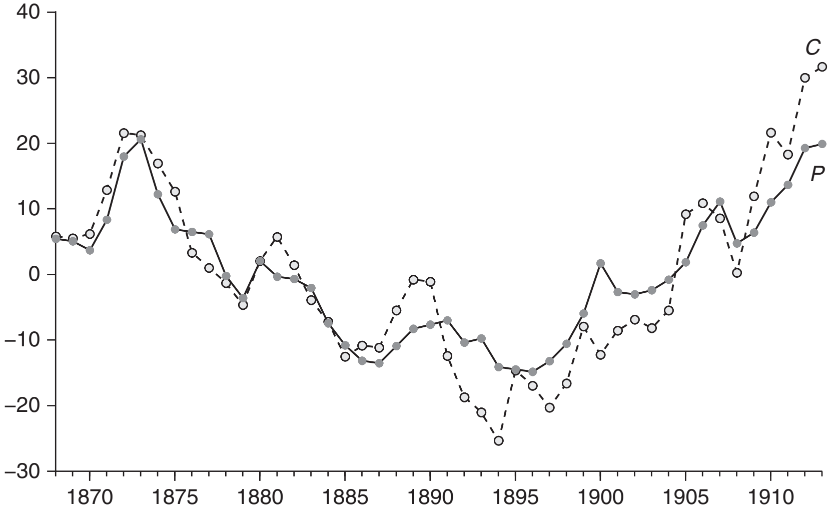


Figure 4.7 Deviations of London clearings (*C*) and Sauerbeck's prices (*P*) from their respective linear secular trends, 1868–1913

indicate an appreciable positive correlation for concurrent items [$r(0)$] and for prices one-year lag [$r(1)$], with the coefficient [$r(0)$] larger. The coefficients for second and higher differences of the raw figures, and deviations as well, decrease as the order of difference increases; the coefficients for one-year

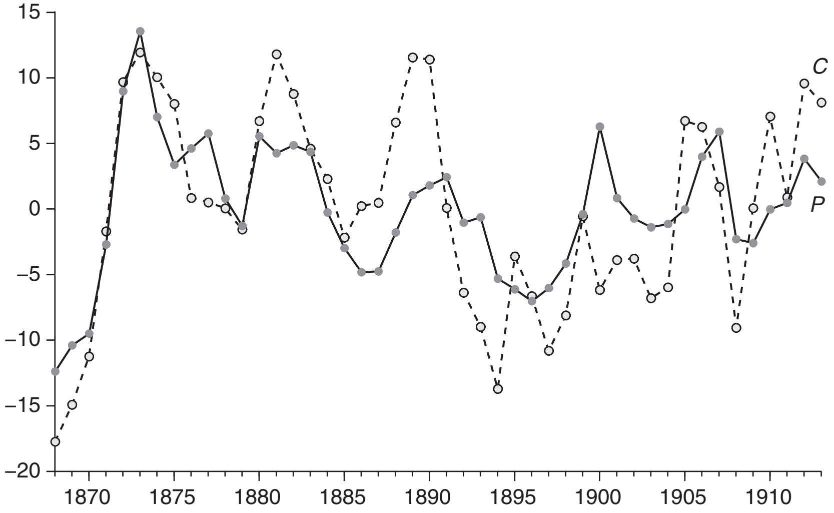


Figure 4.8 Deviations of London clearings (*C*) and Sauerbeck's prices (*P*) from their respective parabolic secular trends, 1868–1913

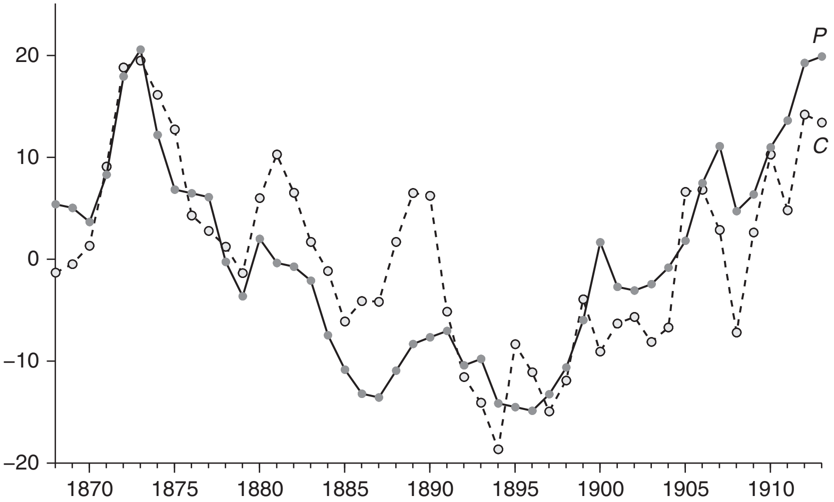


Figure 4.9 Deviations of London clearings from trend as compound interest curve (*C*) and Sauerbeck's prices from linear trend (*P*), 1868–1913

lag of prices decreasing more rapidly than for concurrent items. These facts indicate that the maximum correlation of business cycles (including irregular fluctuations) is for clearings preceding prices by less than half a year, say, four months. (*ibid.*, page 630)

Table 4.4 Coefficients of correlation between Sauerbeck's price indices and London clearings, 1868–1913

d	*r*(−2)	*r*(−1)	*r*(0)	*r*(1)	*r*(2)
			A		
0	−0.42	−0.36	−0.31	−0.25	−0.22
1	−0.04	+0.10	+0.52	+0.46	+0.17
2	−0.05	−0.19	+0.31	+0.14	−0.00
3	+0.02	−0.23	+0.25	+0.02	−0.02
4	+0.04	−0.22	+0.23	−0.03	−0.02
5	+0.05	−0.19	+0.21	−0.05	−0.01
6	+0.08	−0.17	+0.18	−0.07	−0.00
			B		
0	−0.22	+0.20	+0.64	+0.64	+0.25
1	−0.15	−0.04	+0.40	+0.35	+0.16
2	−0.02	−0.13	+0.28	+0.07	+0.08
3	+0.06	−0.08	+0.18	+0.01	+0.11
			C		
0	+0.68	+0.81	+0.92	+0.84	+0.69
1	−0.04	+0.10	+0.52	+0.46	+0.17
2	−0.05	−0.19	+0.31	+0.14	−0.00
3	+0.02	−0.23	+0.25	+0.02	−0.02
			D		
0	+0.14	+0.45	+0.75	+0.65	+0.33
1	−0.09	+0.04	+0.49	+0.42	+0.15
2	−0.05	−0.19	+0.31	+0.14	−0.00
3	+0.02	−0.23	+0.25	+0.02	−0.02
			E		
0	+0.59	+0.69	+0.78	+0.69	+0.50
1	−0.08	+0.05	+0.48	+0.42	+0.13
2	−0.05	−0.19	+0.31	+0.14	−0.00
3	+0.02	−0.23	+0.25	+0.02	−0.02

A: Raw figures and their differences
B: Deviations from nine-year moving average and differences
C: Deviations from straight line and differences
D: Deviations from parabola and differences
E: Deviations from compound interest law for clearings and straight line for prices and differences.

Table 4.5 Coefficients of correlation between Sauerbeck's price index and London clearings from their respective linear secular trends for the two periods 1868–1896 and 1897–1913 together with coefficients for lag differences

d	$r(-1)$	$r(0)$	$r(1)$	$r(2)$	$r(3)$
			1868–1896		
0	+0.36	+0.71	+0.63	+0.32	−0.06
1	+0.19	+0.57	+0.38	+0.13	−0.08
			1897–1913		
0	−0.21	+0.48	+0.68	+0.19	−0.31
1	−0.36	+0.31	+0.46	+0.06	−0.34

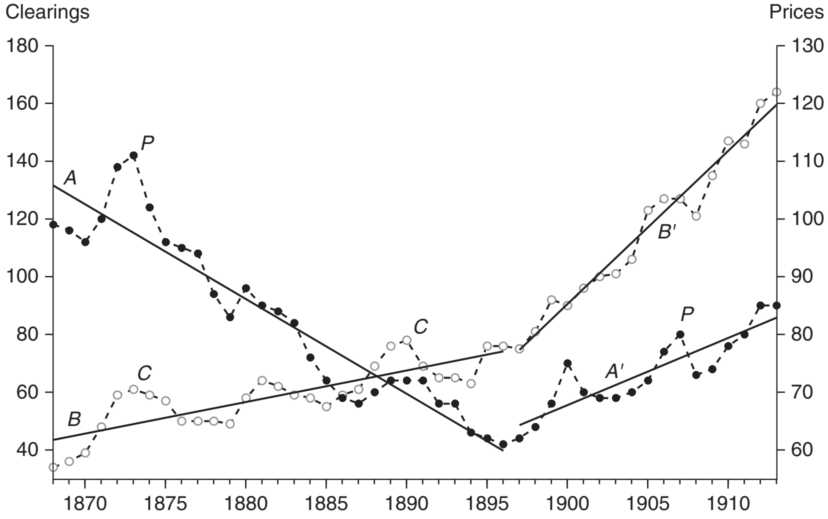
$r(k) \equiv r_{CP}(k)$ denotes the correlation between clearings at time t and prices at time $t+k$

Persons then considered how the various ways of removing the secular trend had performed.

> The coefficients of correlation for the deviations all agree in locating the maximum, and therefore the lag of prices, at less than a year. The actual maximum found was for concurrent items, except for deviations from the nine-year averages which gives a maximum at one year lag of prices. Since our judgment is based upon the relative values of the coefficients for various degrees of lag, rather than upon their absolute values, the type of secular trend chosen does not appear to have great significance. Curve-fitting, however, does appear to be preferable to the taking of moving averages because, first, all the items may be used in determining the correlation and, second, the coefficients for deviations and first differences disagree in their location of the maximum when deviations from the moving average are taken. (*ibid.*, page 630–631)

Persons then went on to consider fitting linear trends to two subperiods of the data; 1868 to 1896 and 1897 to 1913 (see Table 4.5). These trends are shown superimposed on the series in Figure 4.10 and the deviations from trend are shown in Figure 4.11 (cf. Persons Figures 9 and 10), leading Persons to the conclusion that

> (d)ivision of the data into two sections throws new light onto the problem. Clearings and prices fluctuated concurrently during the first period, but prices lagged behind clearings by a year during the period 1896–1913. Perhaps increased speculation has changed the character of clearings during the second period. (*ibid.*, page 632)



Period	Prices	Clearings
1868–1896	*A* $y = -1.64t + 68$	*B* $y = 1.09t + 69$
1897–1913	*A′* $y = +1.16t + 57$	*B′* $y = 5.31t + 43$

Figure 4.10 Sauerbeck's price index (*P*) and London clearings (*C*), 1868–1913, with two straight lines fitted to both series, 1868–1896 and 1897–1913, respectively

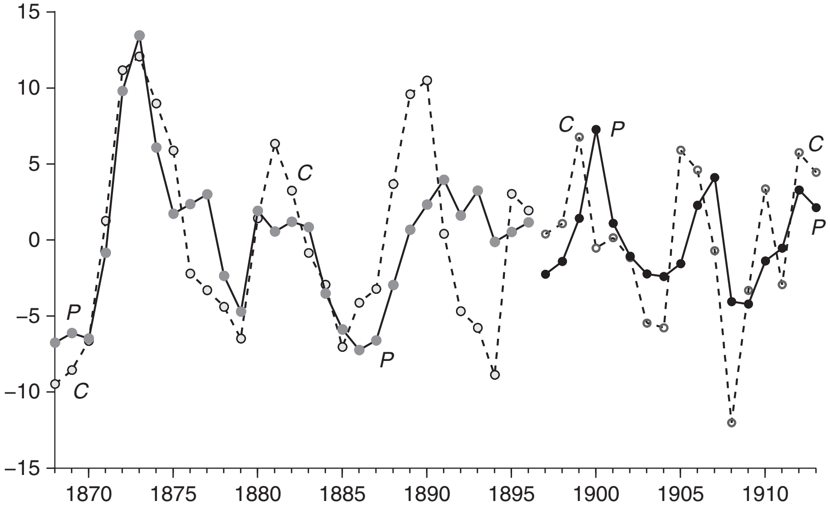


Figure 4.11 Deviations of Sauerbeck's prices (*P*) and London clearings (*C*) from their respective linear trends for the two periods 1868–1896 and 1897–1913

After a second example in a similar vein examining wholesale prices and pig iron production in the US, Persons felt able to make the following recommendations.

> The variate difference correlation method has been invented to eliminate spurious correlation due to position of items in time and space.
>
> The method involves the assumption that the taking of multiple differences leads to series of random variates. In practice for short series this assumption is not fulfilled.
>
> Coefficients for higher differences of short series tend to alternate in sign and to conceal rather than to reveal the nature of the correlation between the series being tested.
>
> Stability of coefficients for higher differences appears to have little significance for short series, and perhaps for long series as well. The assumption that the series correlated are made up of variates 'randomly distributed in time,' if fulfilled, will lead to stable coefficients for successive differences. However, though this condition is *sufficient* for stability it is not *necessary*.
>
> In testing economic time series for correspondence of their cyclical fluctuations, especially in determining the relative position of the cycles upon the assumption that there are cycles, the correlation coefficients between deviations from a linear secular trend together with coefficients for first differences constitute a reliable basis for judgment.
>
> When the question is one of the *existence* or *non-existence* of similar cycles in two time series great care must be used in the choice of the function used to represent the secular trend and in the nature of the fit of the curve or line to the data. The method of first differences is an extremely valuable aid in investigating such a question.
>
> Coefficients of correlation between second differences may give information concerning minor oscillations as distinct from secular trend and major cycles. Even for this purpose the use of higher than second differences appears to be unreliable, especially so for short series. The coefficients of correlation between second differences are identical with those between deviations from three-year progressive averages.
>
> The method of measuring correlation between cycles of time series, that is both easy of application and reliable, is the method of *first* differences. In general, however, this method should be supplemented by curve fitting. To secure a picture of the cycles it is, of course, necessary to take deviations from a closely fitted curve.
>
> Finally, curve fitting to eliminate the secular trend of a time series should always be adapted to the problem in hand and interpretation of coefficients of correlation between time series should be made with continual reference to the fundamental data. Important light may be secured by dividing statistical

> series into more homogeneous sub-series and analyzing the latter. The nature of the data is as important as the method to be applied. Rules-of-thumb concerning method or data are apt to lead to pitfalls. (*ibid.*, pages 641–642)

This extremely thoughtful set of recommendations did, in hindsight, lay the foundations for many of the developments in time series that have taken place right up to the present day, especially when dealing with economic time series: see, in particular, chapters 10 and 16. They met with complete indifference from Pearson and Elderton, however, who simply responded by stating that 'Dr Persons' criticism is valid, although his attempt to get rid of secular trends are from our standpoint wholly inadequate, and further his criticism was hypothetical' (Pearson and Elderton, 1923, page 309). They then moved on to an extended reworking of their 1915 mortality example with little or no reference to these concerns! Nevertheless, Pearson and Elderton (1923) represented the end of Karl Pearson's involvement with the variate difference method, leaving the field open for other statisticians to become involved.

Later developments: Anderson, Tintner and Quenouille

4.15 The variate difference baton was subsequently picked up by Oskar Anderson, who extended the methodology that he had introduced in Anderson (1914) (§§**4.2**–**4.5**) to a sequence of papers (Anderson, 1923, 1926, 1927a, 1927b) and finally a book (Anderson, 1929). From §**4.3**, the variance of the random component Y_t is given by

$$\sigma_Y^2 = \frac{\sigma_{\Delta^d Y}^2}{{}_{2d}C_d}$$

so that an unbiased estimate of this variance is given by

$$\hat{V}_d = \frac{S_d}{(T-d)_{2d}C_d} \qquad S_d = \sum_{t=1}^{T-d} (\Delta^d y_t)^2$$

The variance of $\hat{V}_d$ has an extremely complicated formula, but if terms of order $(T-d)^{-2}$ are neglected (i.e., if the sample size T is large), it can be approximated by[14]

$$var\hat{V}_d = \frac{1}{T-d}\left(\mu_{Y,4} - 3\sigma_Y^4 + 2\frac{{}_{4d}C_{2d}}{({}_{2d}C_d)^2}\sigma_Y^4\right) \approx \frac{1}{T-d}(\mu_{Y,4} - 3\sigma_Y^4 + \sqrt{2d\pi}\sigma_Y^4)$$

where $\mu_{Y,4} = E(Y_t^4) = \kappa_4 + 3\sigma_Y^4$, κ_4 being the moment measure of skewness, and the approximation

$$\frac{2^{2d}}{{}_{2d}C_d} \sim \sqrt{2d\pi}$$

is used. If κ_4 exists then $\hat{V}_d/(var\hat{V}_d)^{\frac{1}{2}}$ has a limiting standard normal distribution.

Anderson also derives an even more complicated formula for the variance of the difference between $\hat{V}_d$ and $\hat{V}_{d+1}$, but provides an easier approximation when $d \geq 6$ and the observations are normally distributed:

$$var(\hat{V}_d - \hat{V}_{d+1}) = \frac{(3d+1)\sqrt{2d\pi}\hat{V}_d^2}{2(2d+1)^2(T-d-1)}$$

The obvious problem with implementing these results was the amount of computation that was involved. The evaluation of the variances was facilitated by the provision of extensive tables of the constants required in the formulae by Tintner (1940), building on the work of Anderson and also of Zaycoff (1937). Indeed, Tintner's monograph provides the definitive discussion of the variance difference approach up to 1940.

4.16 Tintner (1940) also offered an extension of the asymptotic result given above concerning the variance of the difference between the variances of two consecutive differences. An exact test can be computed which takes account of the correlation induced in an independent series through differencing by selecting subsamples of the differenced series that ensure that any correlation is eradicated. To be precise, the subsamples of $\Delta^d y_t$ and $\Delta^{d+1} y_t$ should be selected as:

$$\Delta^d y_r : r = t,\ t + (2d+3),\ t + 2(2d+3),\ \ldots,\ t + (j-1)(2d+3);$$

$$\Delta^{d+1} y_s : s = t + d + 1,\ t + d + 1 + (2d+3),\ \ldots,\ t + d + 1 + (j-1)(2d+3)$$

where t may have any integral value from 1 to $2d + 3$ inclusive, with each value giving rise to a different selection; j takes the largest possible value such that $(j-1)(2d+3) \leq T$, the length of the sample. This ensures that none of the quantities $\Delta^d y_r$ and $\Delta^{d+1} y_s$ have a y_t in common, thus ensuring their independence. If the non-random elements have been removed by taking dth differences, then $\Delta^d y_r$ and $\Delta^{d+1} y_s$ should be sequences of independent variables, each with expected value zero and with variances satisfying the ratio

$$\frac{2dC_d}{2d + 2C_{d+1}} = \frac{d+1}{2(2d+1)}$$

Tintner thus suggested using the test statistics

$$F = \frac{S_d}{S_{d+1}} \frac{2(2d+1)}{d+1} \sim F(d, d) \quad \text{or} \quad z = \frac{1}{2} \ln F \sim N(0,\ 1)$$

to provide exact significance limits.

Although providing an exact test, this sample selection method involves the sacrifice of a considerable proportion of the available data, with only $(T - d - 1)/(2d + 3)$ observations being used. Subsequently, Johnson (1948) proposed a modification of this approach which allowed a greater proportion of the data to be used.

4.17 The aim of the variate difference method was thus to establish the order of the polynomial that gave the 'best fit' to a series, in the sense of reducing the error to randomness, this being the order of differencing after which the variances $\hat{V}_d$ do not seriously change for higher differences. But on the assumption that the original series consisted of a polynomial plus a random error (as in §**4.1**), we might also enquire what is the best estimate of the error variance σ_Y^2 given a set of $\hat{V}_d$s. This question was examined by Quenouille (1953), who sought a linear function of the $\hat{V}_d$s which had minimal variance. Quenouille (1951, 1953) also extended the method to allow for cases when the errors were not assumed to be random but were intercorrelated. This, however, requires the concept of serial correlation, probably the most fundamental concept in time series analysis, to which Yule, in particular, now turned his attention.

5
Nonsense Correlations, Random Shocks and Induced Cycles: Yule, Slutzky and Working

Modern foundations

5.1 By the mid-1920s the methodological advances discussed in the previous two chapters, namely periodogram analysis and the variate differencing method, appeared to be running out of steam, with few new applications appearing and increasing concern about the underlying assumptions of the techniques. At this point, three papers appeared in quick succession which transformed the subject and laid the foundations for modern approaches to the analysis of time series. Two of the papers, by Yule (1926) and Slutzky (1927), went a long way to establishing the basis for the theoretical analysis of stationary time series and, because of the enduring importance of their contributions, are consequently subjected to detailed scrutiny in this chapter, along with a subsequent and closely related paper by Working (1934). The third paper, also by Yule (Yule, 1927), attacked periodic time series in a new way and, in turn, provided the foundations for analysing oscillatory time series: this is our focus in Chapter 6.

Yule and nonsense correlations

5.2 In his Presidential Address to the Royal Statistical Society in November 1925, Yule considered a problem that had puzzled him for many years. Since it lies at the centre of all attempts to analyse the relationships between time series, Yule's statement of the problem is worth setting out in full:

> It is fairly familiar knowledge that we sometimes obtain between quantities varying with the time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly 'significant'. As the occurrence of such 'nonsense-correlations' makes one mistrust the serious arguments that are sometimes put forward on the basis of correlations

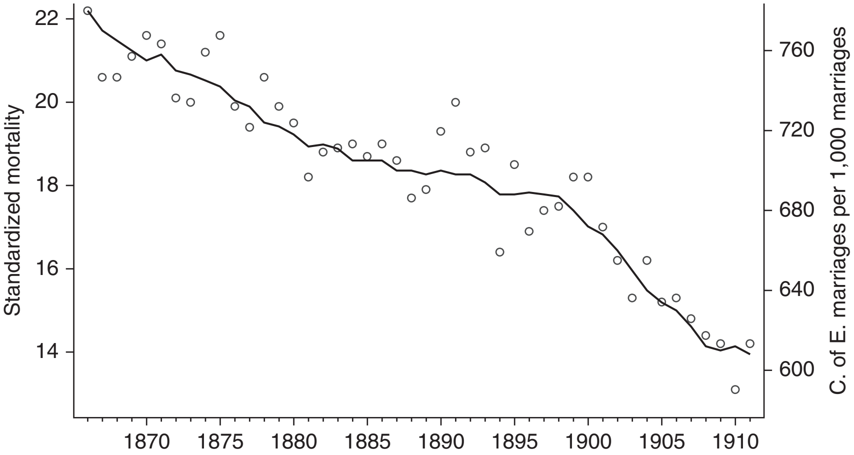


Figure 5.1 Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0.9512$. (Recreated from Yule, 1926, Fig. 1, page 3)

between time-series ... it is important to clear up the problem of how they arise and in what special cases. [Figure 5.1] gives a very good illustration. The full line shows the proportion of Church of England marriages to all marriages for the years 1866–1911 inclusive: the small circles give the standardized mortality per 1,000 persons for the same years. Evidently there is a very high correlation between the two figures for the same year: the correlation coefficient actually works out at +0.9512.

Now I suppose it is possible, given a little ingenuity and goodwill, to rationalize very nearly anything. And I can imagine some enthusiast arguing that the fall in the proportion of Church of England marriages is simply due to the Spread of Scientific Thinking since 1866, and the fall in mortality is also clearly to be ascribed to the Progress of Science: hence both variables are largely or mainly influenced by a common factor and consequently ought to be highly correlated. But most people would, I think, agree with me that the correlation is simply sheer nonsense; that it has no meaning whatever; that it is absurd to suppose that the two variables in question are in any sort of way, however indirect, causally related to one another.

And yet, if we apply the ordinary test of significance in the ordinary way, the result suggests that the correlation is certainly 'significant' – that it lies far outside the probable limits of fluctuations of sampling. The standard error of a coefficient of correlation is $(1 - r^2)/\sqrt{n}$, where n is the number of observations: that is to say, if we have the values of the two variables x and y entered in their associated pairs on cards, if we take out a random sample of n cards (small compared with the total of cards available) and work out

> the correlation, for this sample, take another sample in the same way, and so on – then the correlation coefficients for the samples will fluctuate round the correlation r for the aggregate of cards with a standard deviation $(1 - r^2)/\sqrt{n}$. For the assigned value of r, viz. 0.9512 and 46 observations, the standard error so calculated is only 0.0140, and on this basis we would judge that we could probably trust the coefficient within 2 or 3 units in the second place of decimals. But we might ask ourselves a different question, and one more germane to the present enquiry. If we took samples of 46 observations at random from a record in which the correlation for the entire aggregate was zero, would there be any appreciable chance of our getting such a correlation as 0.9512 merely by chances of sampling? In this case the standard error would be $1/\sqrt{46}$, or 0.1474, the observed correlation is 6.45 times this, and the odds would be many millions to one against such a value occurring 'by chance' – odds so great that the event may be written down as for all practical purposes impossible. On the ordinary test applied in the ordinary way we seem compelled to regard the correlation as having *some* meaning. (Yule, 1926, pages 2–4; italics in original)

Having thus restated the standard statistical argument of the day, Yule then made a crucial assertion:

> Now it has been said that to interpret such correlations as implying causation is to ignore the common influence of the time-factor. While there is a sense – a special and definite sense – in which this may perhaps be said to cover the explanation ... , to my own mind the phrase has never been intellectually satisfying. I cannot regard time *per se* as a causal factor; and the words only suggest that there is some third quantity varying with the time to which the changes in both the observed variables are due ... But what one feels about such a correlation is, not that it must be interpreted in terms of some very indirect catena of causation, but that it has no meaning at all; that in non-technical terms it is simply a fluke, and if we had or could have experience of the two variables over a much longer period of time we would not find any appreciable connection between them. But to argue like this is, in technical terms, to imply that the observed correlation *is* only a fluctuation of sampling, whatever the ordinary formula for the standard error may seem to imply: *we are arguing that the result given by the ordinary formula is not merely wrong, but very badly wrong.* (*ibid.*, page 4: italics added for emphasis)

Yule next set out the problem, as he saw it, more formally:

> When we find that a theoretical formula applied to a particular case gives results which common sense judges to be incorrect, it is generally as well

> to examine the particular assumptions from which it was deduced, and see which of them are inapplicable to the case in point. In obtaining the formula for the standard error we assume, to speak as before in terms of drawing cards from a record: (1) that we are drawing throughout from the same aggregate and not taking one sample from one aggregate, a second sample from another aggregate and so on; (2) that every card in each sample is also drawn from the same aggregate, not the first card from one batch, the second from another, and so on; (3) that the magnitude of *x* drawn on, say, the second card of the sample is quite independent of that on the first card, and so on for all other pairs in the sample; and similarly for *y*; there must be no tendency for a high value of *x* on the first card drawn to imply that the value of *x* on the second card will also probably be high; (4) in order to reduce the formula to the very simple form given, we have also to make certain assumptions as to the form of the frequency distribution in the correlation table for the aggregate from which the samples are taken. (*ibid.*, pages 4–5)

In what ways does the example chosen by Yule and shown in Figure 5.1 diverge from these basic assumptions?

> In the particular case considered and in many similar cases there are two of these assumptions – leaving aside the fourth as comparatively a minor matter – which quite obviously do not apply, namely, the related assumptions (2) and (3). Our data necessarily refer to a *continuous* series of years, and the changes in both variables are, more or less, continuous. The proportion of marriages celebrated in the Established Church falls without a break for years together; only a few plateaus and little peaks here and there interrupt the fall. The death-rate, it is true, shows much larger and more irregular fluctuations from year to year, but there is again a steady tendency to fall throughout the period; only one rate (the last) in the first half of the years chosen, 1866–88, is below the average, only five in 1889–1911 are above it. Neither series, obviously, in the least resembles a random series as required by assumption (3). (*ibid.*, page 5)

What, then, are the implications for these violations of the basic assumptions?

> But can this breach of the assumed conditions render the usual formula so wholly inapplicable as it seems to be? May it not merely imply ... some comparatively slight modification? Even if the standard error by the usual formula were doubled, this would still leave the correlation almost significant. ... [*W*]*hen the successive* x*'s and* y*'s in a sample no longer form a random series, but a series in which successive terms are closely related to one another, the usual conceptions to which we are accustomed fail totally and entirely to apply.* (*ibid.*, pages 5–6; italics added for emphasis)

This, in a nutshell, is the problem of 'nonsense correlations' that Yule intended to analyse in his Presidential Address.

5.3 Yule began his attack on the problem by considering two simple harmonic functions

$$y_t = \sin\left(2\pi\frac{t}{n}\right) \quad x_t = \sin\left(2\pi\frac{t+\alpha}{n}\right)$$

where (cf. §§**3.1–3.5**, §**4.13**) n is the period and α is the difference in phase between the two functions (the amplitude is taken as unity as its value is irrelevant to the analysis). Yule wished to compute the correlation between simultaneous values of y and x over an interval $\pm h$ around the time $t = u$, treating the observed values as continuous. Since, for example,

$$\begin{aligned}\int_{u-h}^{u+h} \sin\left(2\pi\frac{t+\alpha}{n}\right)dt &= \frac{n}{2\pi}\left(\cos\left(2\pi\frac{u+\alpha-h}{n}\right) - \cos\left(2\pi\frac{u+\alpha+h}{n}\right)\right)\\ &= \frac{n}{\pi}\sin\left(2\pi\frac{u+\alpha}{n}\right)\sin\left(2\pi\frac{h}{n}\right)\end{aligned}$$

dividing this by $2h$ will give the mean of x over the interval $u \pm h$:

$$\bar{x}(u \pm h) = \frac{n}{2\pi h}\sin\left(2\pi\frac{u+\alpha}{n}\right)\sin\left(2\pi\frac{h}{n}\right) \tag{5.1}$$

Similarly,

$$\int_{u-h}^{u+h} \sin^2\left(2\pi\frac{t+\alpha}{n}\right)dt = h - \frac{n}{4\pi}\cos\left(4\pi\frac{u+\alpha}{n}\right)\sin\left(4\pi\frac{h}{n}\right)$$

so that, on division by $2h$, we have

$$s_x^2(u \pm h) = \tfrac{1}{2} - \frac{n}{8\pi h}\cos\left(4\pi\frac{u+\alpha}{n}\right)\sin\left(4\pi\frac{h}{n}\right) - \bar{x}^2\,(u \pm h) \tag{5.2}$$

which is the variance of x over the interval $u \pm h$. In a similar vein, using

$$\int_{u-h}^{u+h} \sin\left(2\pi\frac{t}{n}\right)\sin\left(2\pi\frac{t+\alpha}{n}\right)dt = h\cos\left(2\pi\frac{\alpha}{n}\right) - \frac{n}{4\pi}\cos\left(2\pi\frac{2u+\alpha}{n}\right)\sin\left(4\pi\frac{h}{n}\right)$$

enables the covariance between y and x over the interval $u \pm h$ to be written as

$$\overline{yx}(u \pm h) = \tfrac{1}{2}\cos\left(2\pi\frac{\alpha}{n}\right) - \frac{n}{8\pi h}\cos\left(2\pi\frac{2u+\alpha}{n}\right)\sin\left(4\pi\frac{h}{n}\right)$$

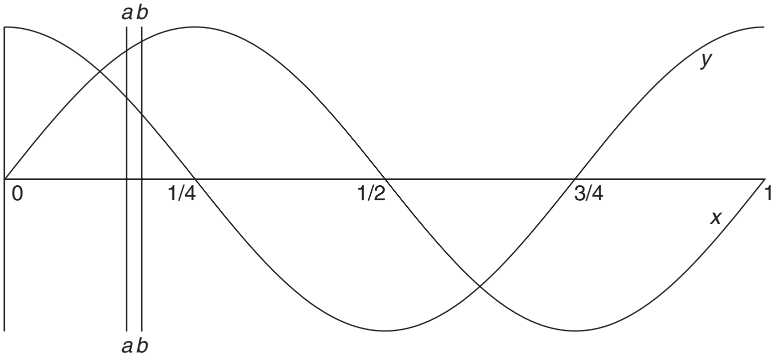


Figure 5.2 Two sine curves differing by a quarter-period in phase, and consequently uncorrelated when the correlation is taken over a whole period

The correlation between y and x over $u \pm h$ is then given by

$$r_{yx}(u \pm h) = \frac{\overline{yx}(u \pm h) - \bar{y}(u \pm h)\bar{x}(u \pm h)}{s_y(u \pm h)\, s_x(u \pm h)} \tag{5.3}$$

where $\bar{y}(u \pm h)$ and $s_y^2(u \pm h)$ are the mean and variance of y calculated in an analogous fashion to (5.1) and (5.2).

Yule focused attention on the case where the phase shift was a quarter of the period, $\alpha = n/4$. The correlation between y and x over a whole period is then obviously zero, as positive deviations from zero in y are exactly matched in frequency by negative deviations from zero in x, as in Figure 5.2. Now suppose we only observe data for a short interval of the whole period, say that enclosed between the two verticals *aa*, *bb*. This interval is so short that the segments of the two curves enclosed between *aa* and *bb* are very nearly straight lines, that for y rising and that for x falling, so that the correlation between the two variables within this interval will therefore be close to -1. Suppose further that the interval from a to b is represented by $t = u \pm h$ and we let $h \to 0$, so that the interval becomes infinitesimally short and the segments of the two curves can be taken to be strictly linear. For $u = 0, 0.25, 0.5, 0.75, 1, \ldots$ the correlation between the two curves will be zero, while for the intervals between these points the correlation will alternate between -1 and $+1$ (see Figure 5.3).

Yule then considered how this correlation 'function' varied as the length of the interval increases from $h = 0$ to $h = n/2$. When $\alpha = n/4$ we have

$$\bar{y}(u \pm h) = \frac{n}{2\pi h} \sin\left(2\pi \frac{u}{n}\right) \sin\left(2\pi \frac{h}{n}\right)$$

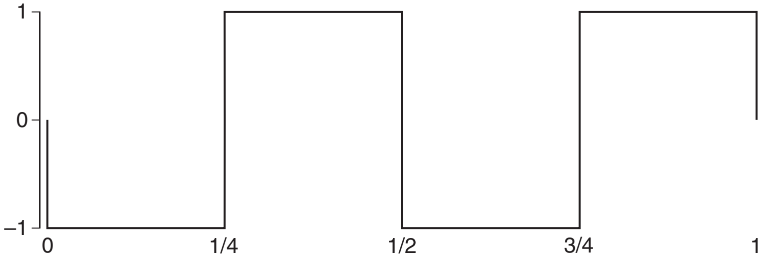


Figure 5.3 Variation of the correlation between two simultaneous intervals of the sine curves of Figure 5.2, as the centre of the interval is moved across from left to right

$$\bar{x}(u \pm h) = \frac{n}{2\pi h}\cos\left(2\pi\frac{u}{n}\right)\sin\left(2\pi\frac{h}{n}\right)$$

$$s_y^2 = \tfrac{1}{2} - \frac{n}{8\pi h}\cos\left(4\pi\frac{u}{n}\right)\sin\left(4\pi\frac{h}{n}\right) - \bar{y}^2(u \pm h)$$

$$s_x^2 = \tfrac{1}{2} + \frac{n}{8\pi h}\cos\left(4\pi\frac{u}{n}\right)\sin\left(4\pi\frac{h}{n}\right) - \bar{x}^2(u \pm h)$$

$$\overline{yx}(u \pm h) = \frac{n}{8\pi h}\sin\left(4\pi\frac{u}{n}\right)\sin\left(4\pi\frac{h}{n}\right)$$

from which the correlation as u varies for a given value of h can be calculated from (5.3). Figure 5.4 recreates Yule's Fig. 4, which shows 'correlation curves' for $2h/n = 0.1, 0.3, \ldots, 0.9$ and from which Yule concluded that

> The first effect of lengthening the interval from something infinitesimally small up to 0.1 of a period is only slightly to round off the corners of the rectangles of [Figure 5.3], and quite slightly to decrease the maximum correlation attainable; it is not until the sample-interval becomes as large as half a period, or thereabouts, that the contours of the curve round off and the maximum undergoes a rather sudden drop. (*ibid.*, page 8)

Yule then used these curves to construct the frequency distribution of the correlation coefficient for a given value of $2h/n$. These distributions are shown in Figure 5.5 and led Yule to conclude that the[1]

> answer to our question, how the distribution of isolated frequencies at +1 and −1 closes up to the distribution of an isolated clump of frequency at zero, is then that the distribution first of all becomes a U-shaped distribution, with limits not far from +1 and −1, and that these limits, at first gradually

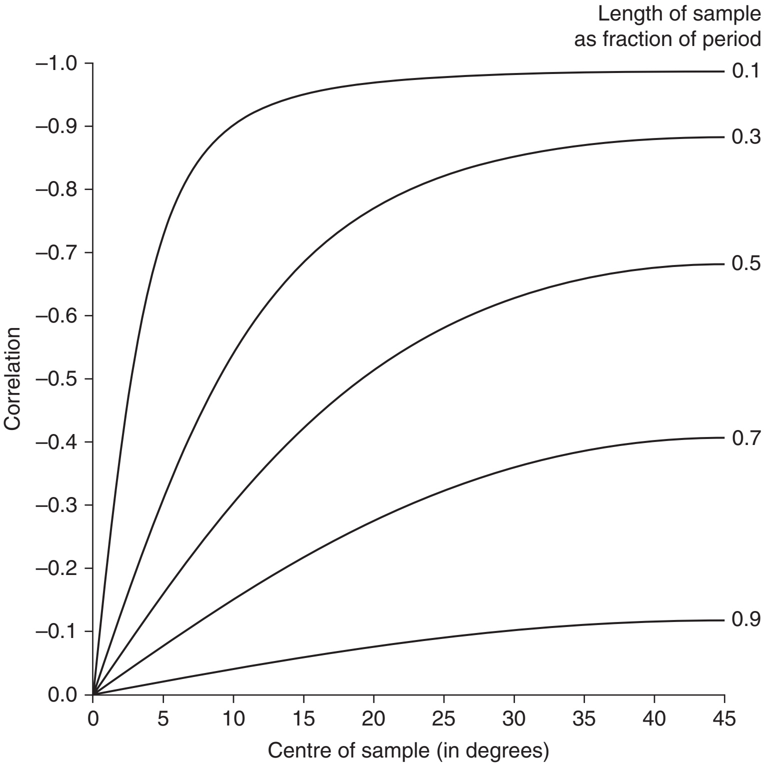


Figure 5.4 Variation of the correlation coefficient between two simultaneous finite intervals of the harmonic curves of Figure 5.2, when the length of the interval is 0.1, 0.3, . . . , 0.9 of the period, as the centre of the interval is moved across from left to right; only one-eighth of the whole period shown

and then more rapidly, close in on zero; but *the distribution always remains U-shaped, and values of the correlation as far as possible removed from the true value (zero) always remain the most frequent.*

The result is in complete contrast with what we expect in sampling under the conditions usually assumed, when successive values of either variable drawn from the sample are independent of one another. In that case the values of *r* in successive samples may differ widely, but the mode tends to coincide with the 'true' value in the aggregate from which the sample is drawn – zero in the present illustration. Here the values in the samples tend to diverge as widely as possible, in both directions, from the truth. We must evidently divest ourselves, in such a case, from all our preconceptions based on sampling under fundamentally different conditions. And evidently the result *suggests* – it cannot do more – the answer to the problem with which we

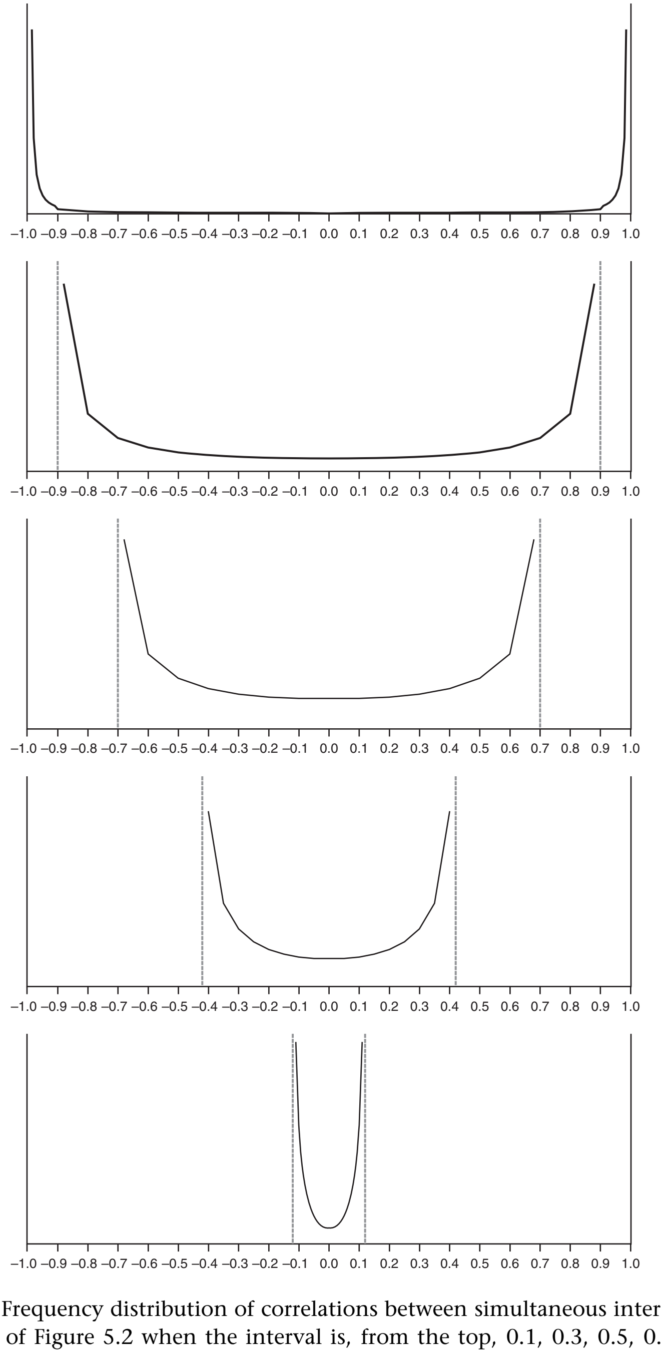


Figure 5.5 Frequency distribution of correlations between simultaneous intervals of the sine curves of Figure 5.2 when the interval is, from the top, 0.1, 0.3, 0.5, 0.7 and 0.9, respectively, of the period

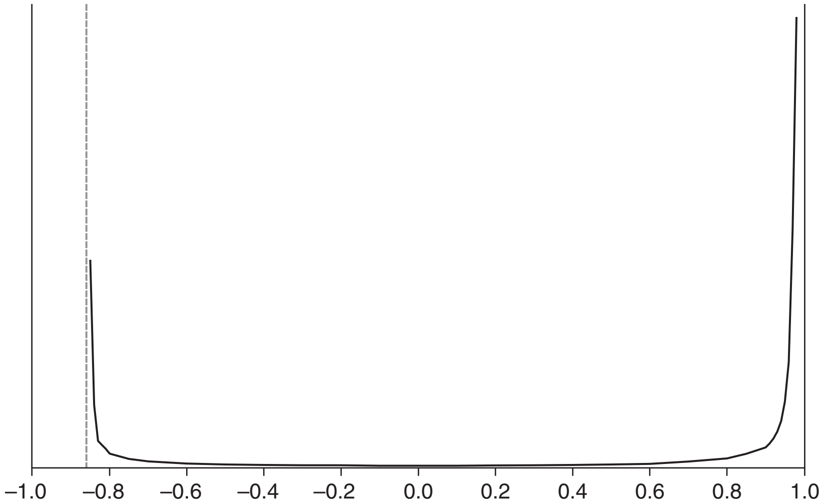


Figure 5.6 Frequency distribution of correlations between two simultaneous intervals of sine curves differing by 60° in phase (correlation over a whole period +0.5) when the length of interval is 0.2 of the period

> started. We tend – it suggests – to get 'nonsense-correlations' between time-series, *in some cases*, because *some* time series are *in some way* analogous to the harmonic series that we have taken as illustration, and our available samples must be regarded as very small samples, if not practically infinitesimal, when compared with the length required to give the true correlation. (*ibid.*, pages 10–12; italics in original)

Yule then considered the case of two sine curves for which the correlation over the whole period was not zero. Specifically, he took two curves that differed in phase by 60° (i.e., $\alpha = n/6$), so that the correlation over a whole period is 0.5, and assumed that $2h/n = 0.2$. The resulting frequency distribution is shown in Figure 5.6, and was described by Yule thus:[2]

> (i)t remains U-shaped, but has become asymmetrical. The limits are −0.85055 and +0.98221, and frequencies are much higher near the positive limit. Roundly 68 per cent of the correlations are positive, 32 per cent are negative, nearly 48 per cent exceed +0.9, only some 13 per cent are less than −0.8. We could only conjecture, in such a case, that the true correlation was positive, if we had a number of samples available, and noted that those giving a positive correlation were to those giving a negative correlation as about 2 to 1. Quite often, at about one trial in eight, a single sample might entirely mislead us by giving a high negative correlation exceeding 0.8. And, be it remembered,

> we have taken a fairly long sample, amounting to one-fifth of the period; if the complete period were something exceeding, say, 500 years, it is seldom that we would have such a sample at our disposal. (*ibid.*, pages 12–13)

5.4 The implication of the analysis in §**5.2** is that meaningless correlations between time series could arise because the series are in some way analogous to harmonic functions, leading Yule to ask

> (w)hat characteristics must two empirical series possess in order that small random samples, taken from them in the same way that we took the small samples from the sine-curves, may tend to give a U-shaped frequency-distribution for the resultant correlations? (*ibid.*, page 14)

The phenomenon is clearly related to the fact that a small segment of a sine curve, when taken at random, will usually be either rising or falling and so will tend to be highly correlated (of either sign) with other segments taken at random. It is easily seen that, if $h=2n$, then

$$\bar{x} = \bar{x}(u \pm n/2) = \bar{y} = \bar{y}(u \pm n/2) = 0$$

$$s_x^2 = s_x^2(u \pm n/2) = s_y^2 = s_y^2(u \pm n/2) = 0.5$$

$$\overline{yx} = \overline{yx}(u \pm n/2) = \tfrac{1}{2}\cos\left(2\pi\frac{\alpha}{n}\right)$$

so that

$$r_{yx} = r_{yx}(u \pm n/2) = \cos\left(2\pi\frac{\alpha}{n}\right)$$

If the whole period is $n=360$ years and the phase is taken to be $\alpha=1$ year, then $r=\cos 1^\circ=0.99985$ gives the correlation between the value of the variable in one year and the value in the next. Similarly, the correlation between the value in one year and that in the next but one year is $\cos 2^\circ=0.99939$, so that, for example, the correlation between values ten years apart is $\cos 10^\circ=0.98481$.

If, following the notation used previously in §**4.12**, we denote the correlation between x_t and x_{t+k} as $r_x(k)$, then Yule proposed that such correlations should be termed the *serial correlations* of the x series (*ibid.*, page 14). With this concept thus defined, Yule then considered answering the following question:

> will it suffice to give us a U-shaped distribution of correlations for samples from two empirical series, if the serial correlations for both of them are high, and positive at least as far as $r_x(T-1)$ where T is the number of terms in the sample? (*ibid.*, page 14: notation altered for consistency)

Yule argued that, if the first term in a sample of consecutive observations taken from a variable having positive serial correlations is considerably above the sample average, then the next few terms will probably be above the average as well, but later terms will have to be below the average to compensate, thus implying that a plot of the sample against time would tend to show a downward movement from left to right. Conversely, if the first term is below the average such a plot will show an upward movement from left to right. Different segments of two such variables would then tend to have markedly positive or negative correlations, depending on whether the two segments had movements in the same or opposite directions. 'This suggests that the frequency-distribution of correlations will be widely dispersed and possibly tend to be bimodal. But will it tend to the extreme of bimodality, a definite U-shape?' (*ibid.*, page 15).

To answer this question, Yule referred back to Figure 5.2.

> When we take a small sample out of either of the curves, such as that between the verticals *aa*, *bb* of the figure, the sample does not tend to show a more or less *indefinite* upward or downward trend; it moves upward or downward with a clear unbroken sweep. This must imply something more: if the curve is going up from year t to year $t+1$, it tends to rise further from year $t+1$ to $t+2$, which is to say, that *first differences are positively correlated with each other*, as well as the values of the variable. For the sine-curve, in fact, we know that the first differences form a curve of the same period as the original: the serial correlations for the *first differences* are therefore precisely the same as those for the values of the variable, given above. This is a very important additional property. It suggests that, for random samples from two empirical series to give a U-shaped distribution of correlations, each series should not merely exhibit positive values for the serial correlations up to $r_x(T-1)$, but their difference series should also give positive serial correlations up to the limit of the sample. (*ibid.*, page 15; italics in original, notation altered for consistency)

5.5 Yule formalized these ideas by first considering the case of a *random series*, for which all the serial correlations are zero, and utilized a well-known result that, in a sample of size T taken from such a series, the correlation between the deviations of any two terms from the sample mean is $-1/(T-1)$.[3] If the first sample value was then above the sample mean, there would be no tendency for the remaining terms to show a downward movement, as they would all have an equal, although slight, tendency to lie below the sample mean. Yule then took 60 sets of 10 random terms, obtained by drawing cards from two packs of playing cards in the following way:

> The court cards were removed from two patience packs; black cards were reckoned as positive, red cards as negative and tens as zeros, so that the

> frequency-distribution in the pack was uniform from −9 to +9, with the exception that there were two zeros. The mean of this distribution is zero, and the standard deviation is $\sqrt{28.5}$, or 5.3385. The pack was shuffled and a card drawn; thoroughly shuffled again and another card drawn, and so on. Every precaution was taken to avoid possible bias and ensure randomness. The use of a double pack helps, I think, towards this, as the complete series is repeated four times. Shuffling was very thorough after every draw; after shuffling, the pack was cut and, say, the fifth card from the cut taken as the card drawn, so as to avoid any possible tendency of the cards to cut at a black rather than a red, or a ten rather than an ace, and so on. (*ibid.*, page 30)

He then computed the deviations from the means in each sample and next separated the samples into two groups, depending on whether the first deviation was positive or negative. Taking each group separately, he then averaged the deviations of each term across the group. Since the standard deviations of all the terms are the same, and the correlation of every term with every other is −1/9, then if the mean of the first term of the positive deviation group is rescaled as 1,000, the most probable deviation of each of the other terms is −1,000/9 or −111, with a similar expectation for the probable deviations of the terms in the negative deviation group on reversing signs.

We recreate this simulation in Table 5.1 but, rather than physically repeating Yule's rather heroic sampling procedure, we utilize modern computing power and software![4] Column (3) gives the average deviations for the first deviation

Table 5.1 Deviations from the mean of the sample in samples of 10 terms from a random series, averaging separately samples in which the first deviation is positive and samples in which the first deviation is negative: average of first deviations taken as +1,000

		Experimental results		
Term (1)	Expectation (2)	First term + (3)	First term − (4)	Together (5)
1	+1,000	+1,000	+1,000	+1,000
2	−111	−379	−198	−274
3	−111	−167	−464	−340
4	−111	−131	−105	−116
5	−111	−158	+141	+15
6	−111	+173	+21	+85
7	−111	−2	−192	−112
8	−111	+99	−132	−35
9	−111	−222	−178	−197
10	−111	−213	+108	−27

positive group; column (4) the average deviations for the first deviation negative group; and column (5) for the two groups taken together. As Yule concluded,

> (t)he figures of neither [column 3], nor [column 4], nor [column 5] show any definite trend in terms 2 to 10. Selection of the first term does not bias the remainder of the sample, or give it any trend or 'tilt' either upwards or downwards; the remaining terms are still random in their order. (*ibid.*, page 16)

He then constructed a correlated series by cumulating a random series:

> Now suppose we take from a series of random terms (with mean zero) a sample of ten terms $a, b, c, d, e, f, g, h, k, l$, and form from it, by successive addition, a new series a, $a+b$, $a+b+c$ .... In this new series the terms are correlated with each other, since each term contains the term before, but the differences are random. (*ibid.*, page 16)

The mean of the sample is thus

$$a+0.9b+0.8c+0.7d+0.6e+0.5f+0.4g+0.3h+0.2k+0.1l$$

so that the deviation of the first term, a, from the mean is

$$-0.9b-0.8c-0.7d-0.6e-0.5f-0.4g-0.3h-0.2k-0.1l$$

Table 5.2 gives the deviations of the successive terms in the sample from the mean. The standard deviation of each deviation for a series of such samples is given by the square root of the sum of squares of the coefficients in the appropriate row in Table 5.2 (scaled by the standard deviation of the original random series). These are given in the rightmost column and show that the end terms in the sample are the most variable, the central terms are the least variable, and the standard deviations are symmetrical about the centre of the sample. The correlation between any pair of terms will be given by the ratio of the product sum of the coefficients associated with the two terms divided by the product of their respective standard deviations. These coefficients are shown in Table 5.3. The correlations of terms adjacent to each other at either end of the sample are high and positive, but terms at opposite ends have moderately high and negative correlations. The general effect of this arrangement of correlations, argued Yule, was to 'give the sample *as a whole a tendency* to be tilted one way or the other as the first term is above or below the average' (*ibid.*, page 18; italics in original).

If the first term in the sample is one unit above the sample mean then the expected mean deviations of the other terms are given by multiplying the appropriate correlation by the ratio of their standard deviations to the standard deviation of the first term. These mean deviations (multiplied by 1,000)

Table 5.2 Coefficients of the terms in the deviations from the mean of the sample, in a sample of 10 terms from a series with random differences *a*, *b*, *c*, . . . , *l*

Term	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Coefficient
	b	*c*	*d*	*e*	*f*	*g*	*h*	*k*	*l*	of s.d.
1	−0.9	−0.8	−0.7	−0.6	−0.5	−0.4	−0.3	−0.2	−0.1	1.688
2	+0.1	−0.8	−0.7	−0.6	−0.5	−0.4	−0.3	−0.2	−0.1	1.432
3	+0.1	+0.2	−0.7	−0.6	−0.5	−0.4	−0.3	−0.2	−0.1	1.204
4	+0.1	+0.2	+0.3	−0.6	−0.5	−0.4	−0.3	−0.2	−0.1	1.025
5	+0.1	+0.2	+0.3	+0.4	−0.5	−0.4	−0.3	−0.2	−0.1	0.922
6	+0.1	+0.2	+0.3	+0.4	+0.5	−0.4	−0.3	−0.2	−0.1	0.922
7	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	−0.3	−0.2	−0.1	1.025
8	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	−0.2	−0.1	1.204
9	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	−0.1	1.432
10	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	+0.9	1.688

Table 5.3 Coefficients between deviations from the mean of the sample, in a sample of 10 terms from a series of random differences

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	+1	+0.81	+0.57	+0.26	−0.10	−0.42	−0.61	−0.66	−0.64	−0.58
2	+0.81	+1	+0.73	+0.37	−0.04	−0.42	−0.65	−0.73	−0.71	−0.64
3	+0.57	+0.73	+1	+0.61	+0.14	−0.32	−0.61	−0.72	−0.73	−0.66
4	+0.26	+0.37	+0.61	+1	+0.48	−0.05	−0.43	−0.61	−0.65	−0.61
5	−0.10	−0.04	+0.14	+0.48	+1	+0.41	−0.05	−0.32	−0.42	−0.42
6	−0.42	−0.42	−0.32	−0.05	+0.41	+1	+0.48	+0.14	−0.04	−0.10
7	−0.61	−0.65	−0.61	−0.43	−0.05	+0.48	+1	+0.61	+0.37	+0.26
8	−0.66	−0.73	−0.72	−0.61	−0.32	+0.14	+0.61	+1	+0.73	+0.57
9	−0.64	−0.71	−0.73	−0.65	−0.42	−0.04	+0.37	+0.73	+1	+0.81
10	−0.58	−0.64	−0.66	−0.61	−0.42	−0.10	+0.26	+0.57	+0.81	+1

are shown in column (2) of Table 5.4: they show a continuous decline from +1,000 for the first term to −579 for the tenth term. The deviations from the mean of each sample constructed from accumulating each of the 60 random samples drawn earlier were then calculated and an analogous computation to that reported in Table 5.1 is shown in column (3) of Table 5.4.

As Yule noted, since the correlations and standard deviations in Table 5.2 are symmetrical, the calculations could be repeated if the samples were sorted depending on whether the *last* term was positive or negative. These calculations are shown in column (5) of Table 5.4 and the results from combining the data on which columns (3) and (5) are based are shown in column (6), leading Yule to conclude that

> (i)n marked contrast with the random series, the sample from the series with random differences shows a clear tendency to tilt one way or the other as a

Table 5.4 Deviations from the mean of the sample in samples of 10 terms from a series with random differences, averaging separately samples in which (*a*) first deviation is positive, (*b*) first deviation is −, (*c*) last deviation is +, (*d*) last deviation is −. The average of first or last deviations, respectively, called +1,000

		Experimental		Experimental results	
Term	Expectation	results *a* and *b*	Term	*c* and *d*	Together
(1)	(2)	(3)	(4)	(5)	(6)
1	+1,000	+1,000	10	+1,000	+1,000
2	+684	+738	9	+754	+746
3	+404	+436	8	+513	+474
4	+158	+283	7	+274	+278
5	−53	−30	6	+79	+25
6	−228	−184	5	−194	−189
7	−368	−346	4	−479	−412
8	−474	−621	3	−498	−559
9	−544	−655	2	−674	−664
10	−579	−621	1	−776	−698

> whole; and hence one random sample from such a series will tend to give more or less marked correlations, either positive or negative, with another. (*ibid.*, page 19)

although he did add the proviso

> it must be remembered that this *tendency* of the sample to be tilted one way or the other as a whole *is* only a tendency; it is sufficiently clearly marked to attract attention during experimental work, but by no means stringent, as is evident from the moderate values of the correlations in [Table 5.3]. (*ibid.*, page 19)

Yule finally considered a third type of series, one whose first differences were positively correlated. He investigated a special case of such a series: that obtained by cumulating a random series twice, i.e., from our original random sample of size 10, we calculate

$$\begin{array}{l} a \\ 2a + b \\ 3a + 2b + c \\ \vdots \\ 10a + 9b + 8c + 7d + 6e + 5f + 4g + 3h + 2k + l \end{array}$$

Table 5.5 Coefficients of the terms in the deviations from the mean of the sample, in a sample of 10 terms from a series of which the second differences are random

Term	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Coefficient
	b	*c*	*d*	*e*	*f*	*g*	*h*	*k*	*l*	of s.d.
1	−4.5	−4.5	−3.6	−2.8	−2.1	−1.5	−0.6	−0.3	−0.1	2.635
2	−3.5	−3.5	−3.6	−2.8	−2.1	−1.5	−0.6	−0.3	−0.1	2.311
3	−2.5	−2.5	−2.6	−2.8	−2.1	−1.5	−0.6	−0.3	−0.1	1.877
4	−1.5	−1.5	−1.6	−1.8	−2.1	−1.5	−0.6	−0.3	−0.1	1.357
5	−0.5	−0.5	−0.6	−0.8	−1.1	−1.5	−0.6	−0.3	−0.1	0.801
6	+0.5	+0.5	+0.4	+0.2	−0.1	−0.5	−0.6	−0.3	−0.1	0.492
7	+1.5	+1.5	+1.4	+1.2	+0.9	+1.5	−0.6	−0.3	−0.1	0.971
8	+2.5	+2.5	+2.4	+2.2	+1.9	+1.5	+0.4	−0.3	−0.1	1.738
9	+3.5	+3.5	+3.4	+3.2	+2.9	+2.5	+1.4	+0.7	−0.1	2.597
10	+4.5	+4.5	+4.4	+4.2	+3.9	+3.5	+2.4	+1.7	+0.6	3.513

Table 5.6 Coefficients between deviations from the mean of the sample, in a sample of 10 terms from a series of which the second differences are random

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	+1	+0.99	+0.97	+0.91	+0.71	−0.32	−0.94	−0.98	−0.96	−0.94
2	+0.99	+1	+0.99	+0.94	+0.75	−0.27	−0.94	−0.99	−0.98	−0.96
3	+0.97	+0.99	+1	+0.97	+0.82	−0.18	−0.91	−0.99	−0.99	−0.97
4	+0.91	+0.94	+0.97	+1	+0.91	+0.01	−0.84	−0.96	−0.98	−0.98
5	+0.71	+0.75	+0.82	+0.91	+1	+0.36	−0.59	−0.80	−0.87	−0.89
6	−0.32	−0.27	−0.18	+0.01	+0.36	+1	+0.51	+0.21	+0.07	−0.01
7	−0.94	−0.94	−0.91	−0.84	−0.59	+0.51	+1	+0.94	+0.87	+0.82
8	−0.98	−0.99	−0.99	−0.96	−0.90	+0.21	+0.94	+1	+0.98	+0.96
9	−0.96	−0.98	−0.99	−0.98	−0.87	+0.07	+0.87	+0.98	+1	+0.99
10	−0.93	−0.96	−0.97	−0.98	−0.89	−0.01	+0.82	+0.96	+0.99	+1

for which the mean is

$$5.5a + 4.5b + 3.6c + 2.8d + 2.1e + 1.5f + g + 0.6h + 0.3k + 0.1l$$

Analogous calculations to those reported in Tables 5.2 and 5.3 are shown as Tables 5.5 and 5.6:

> It will be seen that the standard deviations are now no longer symmetrical about the centre of the sample, the s.d. of term 10 being much larger than that of term 1; while the general arrangement of the correlations is similar to that of [Table 5.2], the correlations are much higher, and again they are not symmetrical with respect to the two ends of the sample. But the magnitude of

Table 5.7 Deviations from the mean of the sample, in samples of 10 terms from a series of which the second differences are random, averaging separately samples in which (*a*) first deviation is positive, (*b*) first deviation is −, (*c*) last deviation is +, (*d*) last deviation is −. The average of first or last deviations, respectively, called +1,000

Term (1)	Expectation (2)	Experimental results *a* and *b* (3)	Term (4)	Expectation (5)	Experimental Results *c* and *d* (6)
1	+1,000	+1,000	10	+1,000	+1,000
2	+870	+868	9	+733	+726
3	+689	+691	8	+473	+459
4	+467	+489	7	+226	+206
5	+215	−258	6	−1	−10
6	−59	−10	5	−203	−205
7	−347	−300	4	−377	−367
8	−644	−637	3	−520	−502
9	−945	−1002	2	−629	−615
10	−1247	−1357	1	−702	−692

> the correlations is now *very* high. Between terms 1 and 2 there is a correlation of 0.992, and between terms 9 and 10 a correlation of 0.991. The maximum negative correlation is that between terms [3 and 8, and is −0.990]. The tendency of the sample to 'tilt' as a whole becomes now very clearly marked, so clear that it becomes quite evident on forming even a few experimental samples in this way. (*ibid.*, page 20; italics in original)[5]

Table 5.7 reports analogous simulations to those given in Table 5.4 and, as should be expected, these show an appropriate degree of conformity, with the experimental results being very close to their expectations.

5.6 After reporting these simulations, Yule summarized their implications in a crucial insight into what are now called *integrated processes* (a term introduced by Box and Jenkins, 1970: see §§**10.17–10.18**):

> Now this argument has led us to a remarkable result, which at first sight may seem paradoxical: namely, that for the present purpose we are really only concerned with the serial correlations for the *differences* of our given series, and not with the serial correlations of those series themselves. For if we take a long but finite series of random terms and sum it, the serial correlations for the sum-series are not determinate and will vary from one such series to another: and yet all such series evidently have the same characteristics from the present standpoint. And obviously again, if we form the second-sum of a long but finite series of random terms, the serial correlations for the second-sum are not determinate and will vary from one such series to another, and

> yet all such series, from the present standpoint, have the same characteristics. If in either case we make the series indefinitely long, all the serial correlations will tend towards unity, but the samples remain just the same as they were before, so evidently we cannot be concerned with the mere magnitude of the serial correlations themselves: they are dependent on the length of the series. (*ibid.*, page 22; italics in original)

To formalize this insight, suppose that $x_1, x_2, \dots, x_T$ is a zero mean series with standard deviation σ_x for which the serial correlations are $r_x(1), r_x(2), \dots, r_x(k)$, using the notation of §**4.12**. Then, if T is assumed to be large,

$$\begin{aligned}\sum_{t=1}^{T-1}(x_{t+1}-x_t)^2 &= \sum_{t=1}^{T-1}x_{t+1}^2+\sum_{t=1}^{T-1}x_t^2-2\sum_{t=1}^{T-1}x_{t+1}x_t\\ &\approx 2\sum_{t=1}^{T-1}x_t^2-2\sum_{t=1}^{T-1}x_{t+1}x_t\\ &= 2\sum_{t=1}^{T-1}x_t^2\left(1-\frac{\sum_{t=1}^{T-1}x_{t+1}x_t}{\sum_{t=1}^{T-1}x_t^2}\right)\end{aligned}$$

or, again utilizing the notation of §**4.7**,

$$\sigma_{\Delta x}^2 = 2\sigma_x^2(1-r_x(1))$$

Similarly, and dropping summation limits to ease notation,

$$\begin{aligned}\sum(x_{t+2}-x_{t+1})(x_{t+1}-x_t) &= \sum x_{t+2}x_{t+1}+\sum x_{t+1}x_t-\sum x_{t+2}x_t-\sum x_{t+1}^2\\ &\cong 2\sum x_{t+1}x_t-\sum x_{t+2}x_t-\sum x_{t+1}^2\\ &= \sum x_{t+1}^2\left(2\frac{\sum x_{t+1}x_t}{\sum x_{t+1}^2}-\frac{\sum x_{t+2}x_t}{\sum x_{t+1}^2}-1\right)\end{aligned}$$

Denoting the serial correlations of the differences as ${}_1r_x(k)$ (as in §**4.12**), we thus have

$${}_1r_x(1)\sigma_{\Delta x}^2 = \sigma_x^2(2r_x(1)-r_x(2)-1)$$

i.e.,

$${}_1r_x(1) = \frac{2r_x(1)-r_x(2)-1}{2(1-r_x(1))}$$

Generalizing this result gives

$$ {}_1r_x(k) = \frac{2r_x(k) - r_x(k+1) - r_x(k-1)}{2(1 - r_x(1))} = -\frac{1}{2(1 - r_x(1))}\Delta^2 r_x(k+1) \tag{5.4} $$

Suppose that the differences are random, so that all the ${}_1r_x(k)$ are zero and $\Delta^2 r_x(k+1) = 0$ for all k, implying that

$$ r_x(k) = 2r_x(k-1) - r_x(k-2) $$

Successive serial correlations are then generated by the arithmetical progression

$$ \begin{aligned} r_x(2) &= 2r_x(1) - r_x(0) = 2r_x(1) - 1 \\ r_x(3) &= 2r_x(2) - r_x(1) = 3r_x(1) - 2 \\ &\vdots \\ r_x(k) &= kr_x(1) - (k-1) \end{aligned} $$

To compute these serial correlations obviously requires a value of $r_x(1)$, say $\hat{r}_x(1)$. Yule (*ibid.*, page 59) suggested determining $\hat{r}_x(1)$ by making the sum of the calculated correlations equal to the sum of the observed correlations, so that the mean error was zero. This gives

$$ \sum_{j=1}^{k} r_x(j) = \tfrac{1}{2}k(k+1)\hat{r}_x(1) - \tfrac{1}{2}k(k-1) $$

from which $\hat{r}_x(1)$ can be calculated. To implement these results, Yule generated three series with random differences, denoted A_1, B_1 and C_1, in the same fashion as in §**5.4** above, these being shown in Figure 5.8 with the underlying random series, A_0, B_0 and C_0, being shown in Figure 5.7. Formally, if the random series is denoted $u_1, u_2, \ldots, u_T$, then $x_t = u_1 + u_2 + \cdots + u_t$ is a series with random differences (in the simulations T is set at 100). Setting $k = 10$, $r_x(1)$ was computed for each series by solving

$$ 11\hat{r}_x(1) = 9 + 0.2\sum_{j=1}^{10} r_x(j) $$

producing the serial correlations shown in Table 5.8 and plotted in Figure 5.9. The fits are quite accurate but it is noticeable how the magnitudes of the serial correlations differ across the three series: $r_x(10)$ is 0.764, 0.191 and 0.697 for A_1, B_1 and C_1, respectively. Yule considered a potential difficulty arising from these

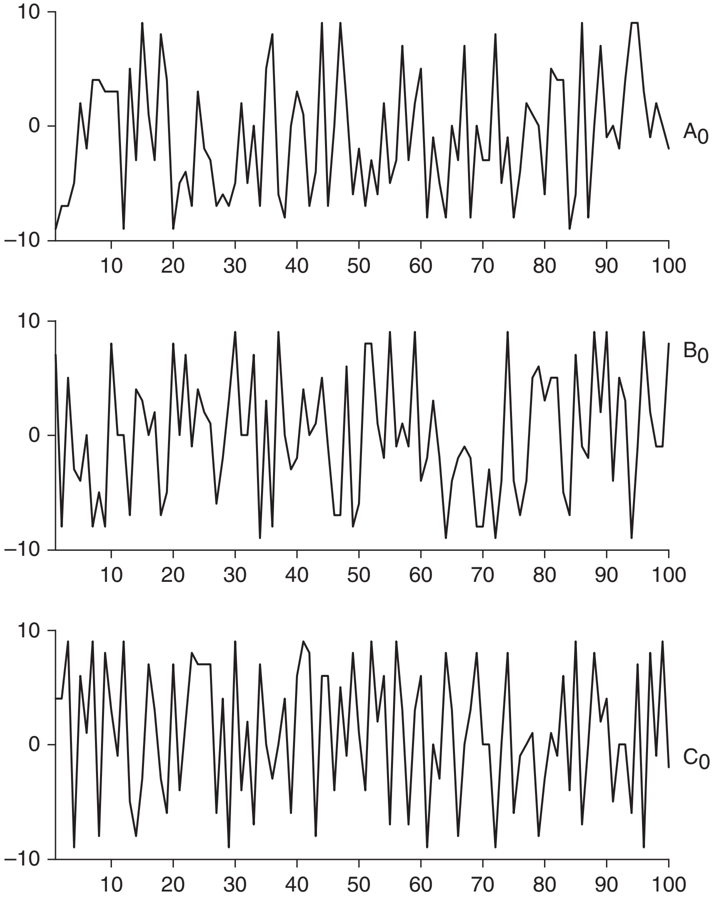


Figure 5.7 Three random series

linearly declining serial correlations: 'if the lines are continued downwards, they will lead to negative and then to impossible values of the correlation' (*ibid.*, page 60). He responded to this by emphasizing that

> we can only obtain such series as those in [Table 5.8] if the serial correlations are determined from a *finite* series, and for a finite series [$\Delta^2 r_x(k+1)=0$] will be only approximately true for moderate values of k and will cease to be valid for large values.' (*ibid.*, page 60; italics in original)

Yule next considered the case when the differences are correlated such that ${}_1r_x(k)$ is a linear function of k. This can be expressed as ${}_1r_x(k)=1-\alpha k$

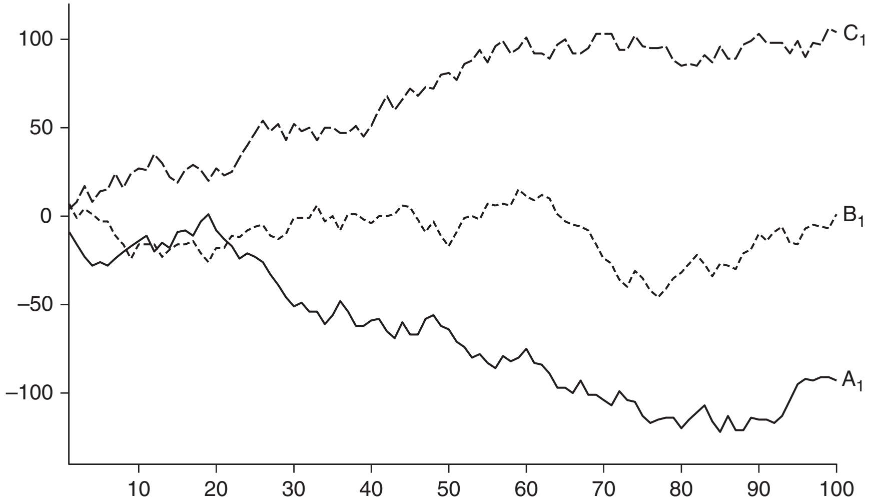


Figure 5.8 Three series with random differences (conjunct series with random differences)

Table 5.8 Comparison of serial correlations for three series with random differences, with fitted arithmetical progressions

	Series A_1		Series B_1		Series C_1	
	Observed correlation	Calculated correlation	Observed correlation	Calculated correlation	Observed correlation	Calculated correlation
1	0.975	0.978	0.909	0.920	0.954	0.967
2	0.953	0.956	0.835	0.840	0.920	0.935
3	0.935	0.934	0.766	0.760	0.894	0.902
4	0.916	0.912	0.691	0.679	0.864	0.870
5	0.897	0.890	0.594	0.599	0.834	0.837
6	0.876	0.868	0.515	0.519	0.801	0.805
7	0.853	0.846	0.458	0.439	0.780	0.772
8	0.826	0.824	0.366	0.360	0.747	0.740
9	0.796	0.802	0.268	0.279	0.720	0.707
10	0.764	0.780	0.191	0.199	0.697	0.675

since ${}_1r_x(0) = 1$. From (5.4) we then have

$$\Delta^2 r_x(k+1) = -2(1 - r_x(1))(1 - \alpha k)$$

and, since their second differences are a linear function of k, the serial correlations $r_x(k)$ must be generated by a cubic in k:

$$r_x(k) = 1 + bk + ck^2 + dk^3$$

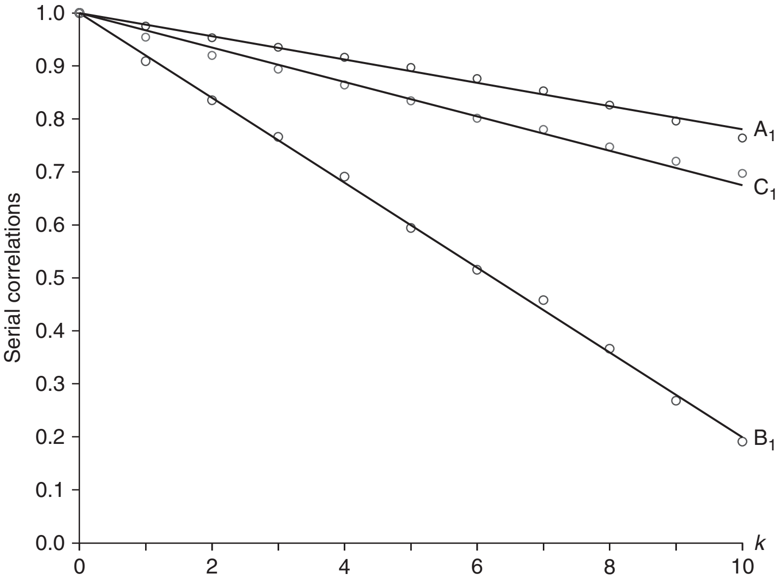


Figure 5.9 Serial correlations up to $r(10)$ for three experimental series (of 100 terms) with random differences

This implies that

$$\Delta^2 r_x(k+1) = 2(c + 3dk)$$

and, on equating coefficients, we have

$$c = -(1 - r_x(1)) \quad d = \tfrac{1}{3}\alpha(1 - r_x(1)) \quad b = -d = -\tfrac{1}{3}\alpha(1 - r_x(1))$$

Defining $m = 1 - r_x(1)$, we can thus write the cubic as

$$r_x(k) = 1 - mk^2 + \tfrac{1}{3}\alpha mk(k^2 - 1) \tag{5.5}$$

Again determining m by making the sum of the calculated correlations equal to the sum of the observed correlations yields the general equation

$$\sum_{j=1}^{k} r_x(j) = k - \hat{m}\left\{\tfrac{1}{6}k(k+1)(2k+1) + \tfrac{1}{6}\alpha k(k+1) - \tfrac{1}{12}\alpha k^2(k+1)^2\right\} \tag{5.6}$$

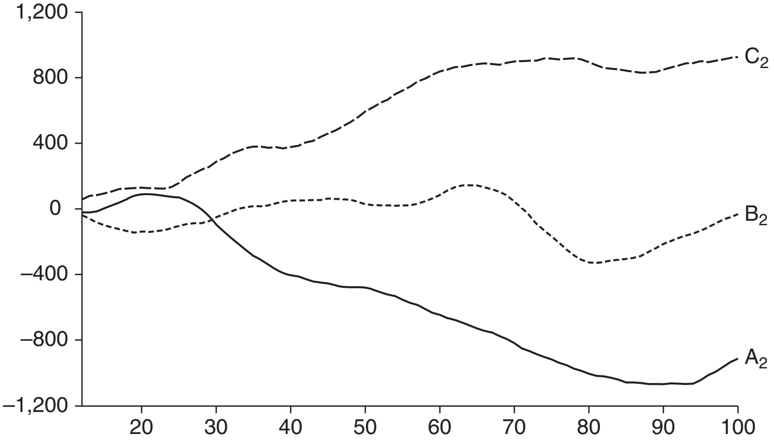


Figure 5.10 Three series with positively correlated differences (conjunct series with conjunct differences)

To utilize this result, Yule constructed a series with correlated differences by taking the random series u_t and cumulating 11-period moving sums, i.e., by calculating

$$s_t = \sum_{j=t-10}^{t} u_j, \quad x_t = \sum_{j=11}^{t} s_j = u_t + 2u_{t-1} + \cdots + 2u_{t-10} + u_{t-11}, \quad t = 11, \ldots, T$$

It is thus straightforward to show that

$$_1r_x(k) = r_s(k) = \begin{cases} 1 - (k/11) & \text{for } k = 1, \ldots, 10 \\ 0 & \text{for } k \geq 11 \end{cases}$$

Thus, setting $\alpha = \frac{1}{11}$ and $k = 10$ reduces (5.6) to

$$295\hat{m} = 10 - \sum_{j=1}^{k} r_x(j)$$

The series so generated, A$_2$, B$_2$ and C$_2$, are shown in Figure 5.10, with their observed serial correlations and the serial correlations calculated from the cubic in *k* reported in Table 5.9 and plotted in Figure 5.11. The cubic fit is fairly accurate for A$_2$ and B$_2$, but is rather poor for series C$_2$, for which the serial correlations appear to decline linearly rather than as a cubic. Again, the serial correlations differ considerably from series to series.

Table 5.9 Comparison of serial correlations for three series with correlated differences, with fitted cubic series

	Series A_1		Series B_1		Series C_1	
	Observed correlation	Calculated correlation	Observed correlation	Calculated correlation	Observed correlation	Calculated correlation
1	0.984	0.995	0.989	0.990	0.973	0.995
2	0.965	0.983	0.960	0.963	0.946	0.980
3	0.944	0.962	0.916	0.921	0.919	0.956
4	0.919	0.936	0.858	0.864	0.891	0.925
5	0.892	0.903	0.789	0.795	0.862	0.887
6	0.862	0.866	0.711	0.716	0.831	0.843
7	0.829	0.824	0.625	0.628	0.801	0.794
8	0.793	0.779	0.534	0.533	0.770	0.742
9	0.756	0.732	0.441	0.432	0.738	0.686
10	0.718	0.683	0.348	0.329	0.706	0.629

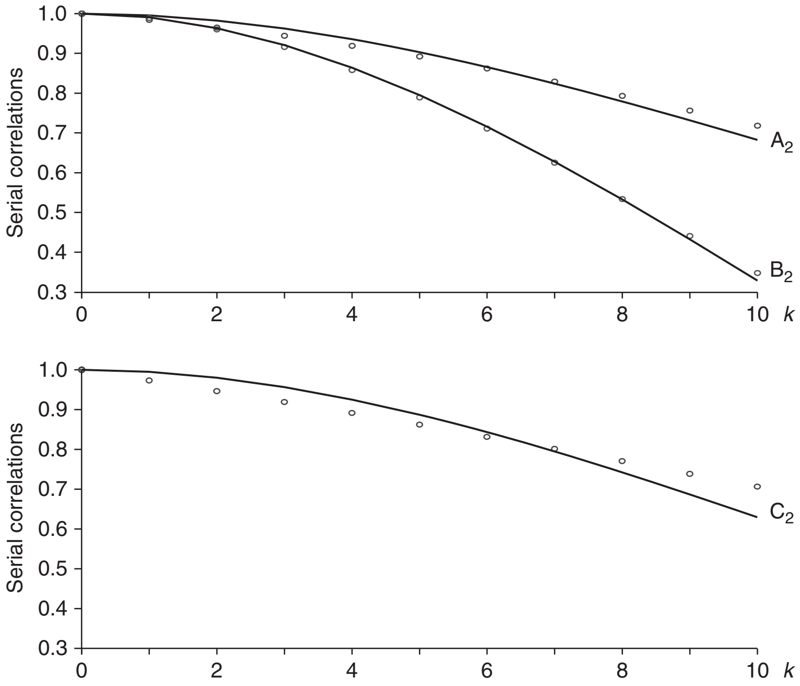


Figure 5.11 Serial correlations up to $r(10)$ for three experimental series (of 100 terms) with positively correlated (conjunct) differences

Finally, Yule briefly considered the case when the second differences of a series were random, so that the series is the 'second sum' of a random series, i.e.,

$$s_t = \sum_{j=1}^{t} u_j, \quad x_t = \sum_{j=1}^{t} s_j = tu_t + (t-1)u_{t-1} + \cdots + u_1, \quad t = 1, \ldots, T$$

In this case the first differences of x_t are the sum of a random series and therefore the serial correlations of Δx_t are given by $\Delta_1^2 r_x(k+1) = 0$, or

$$_1r_x(k) = k_1r_x(1) - (k-1) = 1 - k(1 - {}_1r_x(1)) = 1 - \alpha k$$

with $\alpha = 1 - {}_1r_x(1)$. Thus, the $r_x(k)$ are given by (5.5) and the analysis is identical to that above.

5.7 This analysis led Yule to classify time series into the following categories based on the nature of their serial correlations:

Random series: Series for which all serial correlations are zero.

Conjunct series: Series for which all serial correlations are positive. With finite series, $r(k)$ may well decrease with k and become negative at some point, in which case the series is said to be 'conjunct up to $r(k)$'.

Disjunct series: Series for which the serial correlations are all negative. Although Yule (*ibid.*, pages 62–3) provided a setup that would generate such a series, the conditions under which this might occur are extremely stringent. However, a series that is 'disjunct up to $r(1)$' is simply obtained by taking first differences of a random series, for which $r(1) = -0.5$ and all higher serial correlations are zero (see §**4.12**).

Oscillatory series: Series for which the serial correlations change sign, alternating between runs of positive and negative values.

Yule regarded these classifications as building blocks: 'clearly in the endless variety presented by facts we may expect to meet with compound series of any type, e.g., conjunct series with an oscillatory series superposed' (*ibid.*, page 26). Nevertheless, his focus continued to be on the three types of series analysed in §**5.5**: (a) *random series*, (b) *conjunct series having random differences*; and (c) *conjunct series having differences which are themselves conjunct*. In terms of these three types, the random series A_0, B_0 and C_0 shown in Figure 5.7 display 'no secular trend, and the whole movement is highly irregular. The graphs are not, to the eye at least, very unlike graphs of some annual averages in meteorological data' (*ibid.*, page 26). Figure 5.8 shows A_1, B_1 and C_1, conjunct series having random differences: 'we now get a marked 'secular movement,' with irregular oscillations superposed on it' (*ibid.*, page 26). Figure 5.10 shows A_2, B_2 and C_2, conjunct series with conjunct differences: 'the curves are smoothed out, the

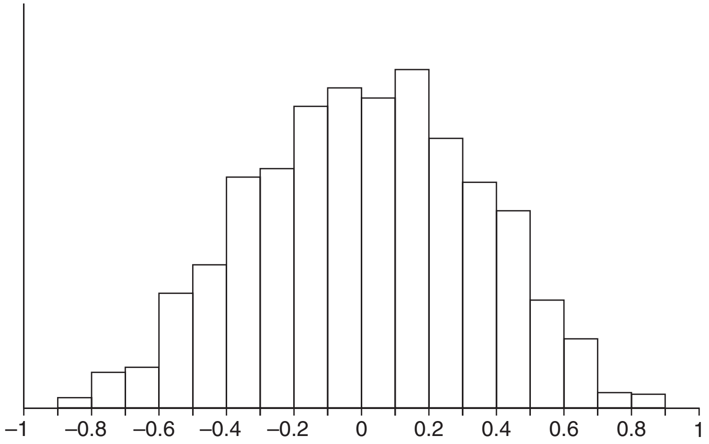


Figure 5.12 Frequency distribution of 600 correlations between samples of 10 observations from random series

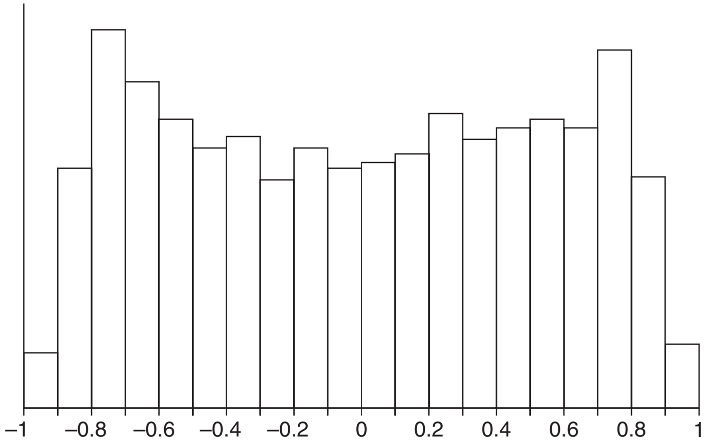


Figure 5.13 Frequency distribution of 600 correlations between samples of 10 observations from conjunct series with random differences

secular movements or long waves are conspicuous, but there are no evident oscillations of short duration' (*ibid.*, page 26).

5.8 Having considered the various 'internal' properties of the different types of time series, Yule then turned his attention to his primary aim, that of analysing the correlations between pairs of series drawn from each of the types. Using samples of size 10, he correlated 600 pairs of random series, 600 pairs of conjunct series with random differences, and 600 pairs of conjunct series with conjunct differences. These series were generated using the sampling procedure of §**5.4**.

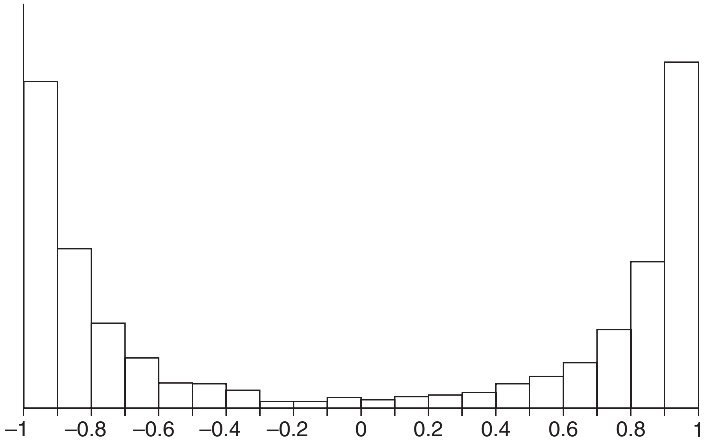


Figure 5.14 Frequency distribution of 600 correlations between samples of 10 observations from conjunct series with conjunct differences

We again recreate Yule's calculations and show in Figures 5.12–5.14 the frequency distributions of the correlations between pairs drawn from the three types of series. The three distributions are quite distinct, being approximately normal, uniform and U-shaped, respectively. The distribution of correlations between random series (Figure 5.12) matches theory: 'the distribution ... should be symmetrical about zero, and ... should approximate the normal form with the mode at zero' (*ibid.*, page 31). With regard to

> the two simple types of conjunct series, those with random differences and those with conjunct differences respectively, correlations between samples of the first type are subject to a much higher standard error than that given by the usual formula [$1/\sqrt{10} = 0.3162$], but do not tend definitely to mislead [Figure 5.13]; correlations between samples of the second type tend definitely to be 'nonsense-correlations' – correlations approaching plus or minus unity in value [Figure 5.14]. The tentative answer to the problem of my title is therefore this: that some time-series are conjunct series with conjunct differences, and that when we take samples from two such series the distribution of correlations between them is U-shaped – we tend to get high positive or high negative correlations between the samples, without any regard to the true value of the correlation between the series that would be given by long experience over an indefinitely extended time. (*ibid.*, page 39)

Yule emphasized that conjunct series with random differences (the sum of a random series with zero mean) would swing above and below the zero base line but, as the length of the series was increased, would not tend to be correlated with time (viz. Figure 5.8). The second sum of a random series, being a conjunct series with conjunct differences, would display swings above and below the base line that would be smoother, longer and of greater amplitude, but there would still be no tendency to be correlated with time as the series length was increased (viz. Figure 5.10). With this analysis, Yule was making the first tentative steps towards identifying what are now referred to as *stochastic trends* (see §**16.12**).

Interestingly, Yule ended the theoretical part of his paper with this statement:

> I give my answer to the problem as a tentative answer only, for I quite recognize that the discussion is inadequate and incomplete. The full discussion of the mathematical problem – given two series, each with specified serial correlations, required to determine the frequency distribution of correlations between samples of n consecutive observations – I must leave to more competent hands. It is quite beyond my abilities, but I hope that some mathematician will take it up. The results that he may obtain may seem to be of mere theoretical importance, for in general we only have the sample itself, which may be quite inadequate for obtaining the serial correlations. But to take such a view would, I think, be short-sighted. The work may not lead, it is unlikely to lead, to any succinct standard error, or even frequency-distribution applicable to the particular case. But only such direct attack can, it seems to me, clear up the general problem; show us what cases are particularly liable to lead to fallacious conclusions, and in what cases we must expect a dispersion of the sample-correlations greater than the normal. ... If my view is correct, that the series correlations of the difference series are the really important factor [then] the sample may be a more adequate basis for the approximate determination of the difference correlations than for the determination of the serial correlations of the series itself. (*ibid.*, page 40)

The statement is extraordinarily prescient on at least two counts. Examination of the serial correlations of the difference series underlies the famous Box and Jenkins (1970) approach to time series model building, to be discussed in Chapter 10, while the mathematical treatment of the nonsense regression problem had to wait some sixty years before a complete solution was provided by Phillips (1986): see §**16.19**.

5.9 Yule then turned his attention to applying these ideas to two time series: Beveridge's (1921, 1922) wheat price index and rainfall at Greenwich. We rework here the first application by using the price index that was subjected to periodogram analysis in §§**3.8–3.9**. Concentrating, as did Yule, on the 300-year period from 1545 to 1844, and using the index numbers themselves, rather than the smoothed Index of Fluctuation, the serial correlations up to $k = 40$ are displayed in Figure 5.15.[6]

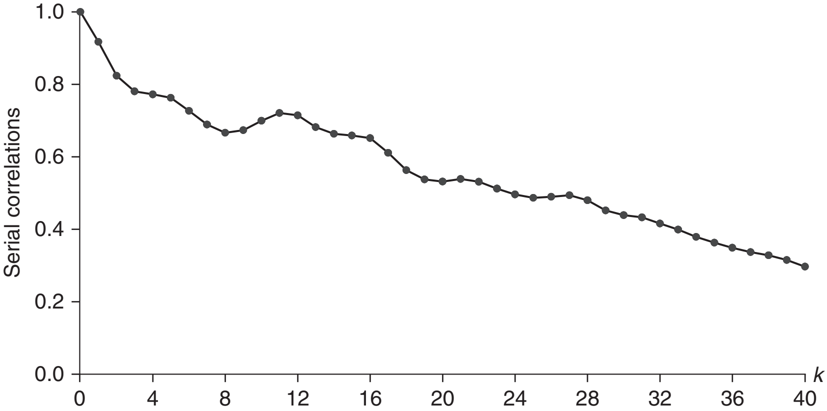


Figure 5.15 Serial correlations up to $r(40)$ for Beveridge's index numbers of wheat prices in Western Europe, 1545–1844

> The correlations are all positive, as they evidently must be in a series that sweeps up from values round about 20 or 30 in earlier years to 100, 200 and over in the later years [recall Figure 3.4]. They fall away at first with some rapidity to a minimum of [0.67] at $r(8)$; there is then a large broad hummock in the curve followed by some minor oscillations, and finally, from about $r(25)$ onwards, the curve tails away comparatively smoothly to [0.30] at $r(40)$. (*ibid.*, pages 42–43; notation altered for consistency)

Yule's next step was to compute the serial correlations of various differences of the index. By a similar reasoning to that of §**5.5**, the serial correlations of the h-step differences $x_{t+h} - x_t$, which we denote as $r_x^h(k)$ (noting that $r_x^1(k) \equiv {}_1r_x(k)$), are given by a generalization of (5.4)

$$r_x^h(k) = \frac{2r_x(k) - r_x(k+h) - r_x(k-h)}{2(1 - r_x(h))}$$

on noting that if $k < h$, $r_x(k-h) = r_x(h-k)$. The 'serial difference correlations' for various values of h are plotted in Figure 5.16. Yule then embarked on a detailed discussion of the oscillations contained in the plots of these serial correlations, which we summarize thus. The plot of the serial correlations for the first differences ($h = 1$) shows that both the peaks and troughs occur between five and six years apart, which is thus consistent with Beveridge's findings of important periodicities in this interval (see Figure 3.4). These oscillations in the serial correlations would be practically eliminated by setting the differencing interval to either 5 or 6, thus determining the next two choices for h. The two serial

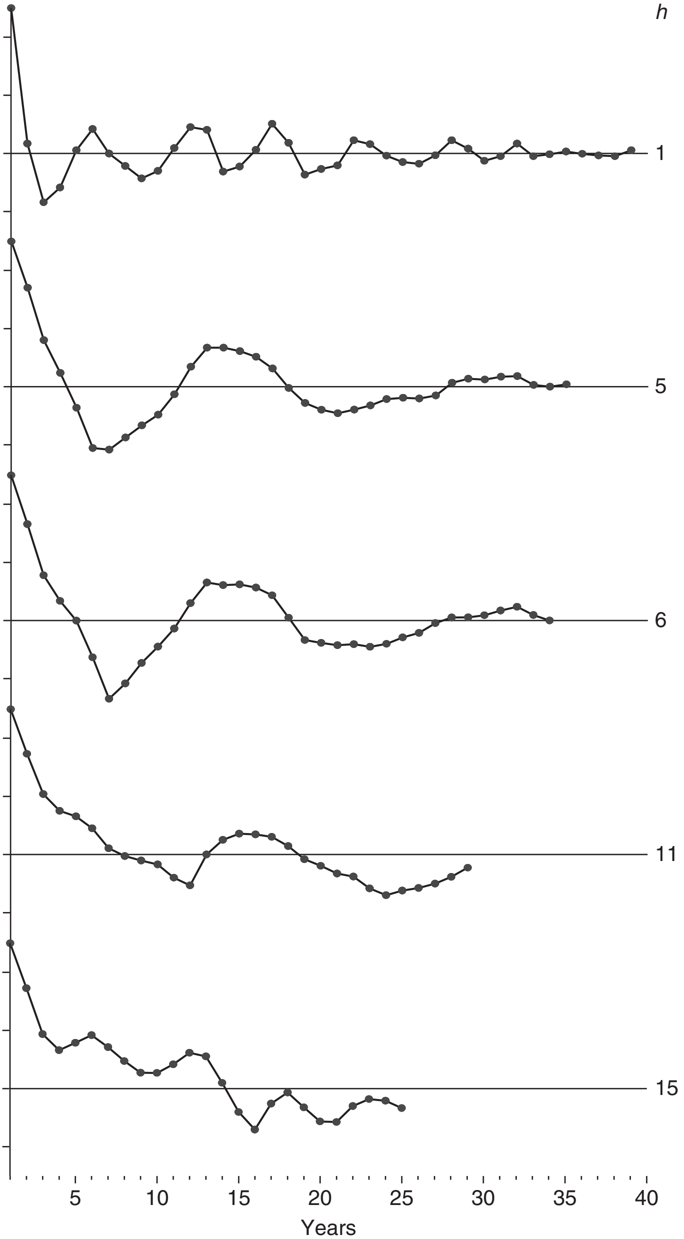


Figure 5.16 Serial difference correlations $r^h(k)$ for the index numbers of wheat prices in Western Europe; intervals for differencing $h = 1, 5, 6, 11$ and 15 years respectively

correlation plots are almost identical, having pronounced oscillations with a peak-to-peak period of around 18 years and a trough-to-trough period of about 14 years. Setting $h = 11$ shows a peak-to-peak period of around 14 years and trough-to-trough period of 12 years, these again being reasonably consistent with the earlier periodogram analysis.

Yule's final choice was $h = 15$, which produces many minor oscillations in the serial correlations. Ignoring these, he argued that, since the curve cuts the zero axis at around 13.5 years, this was consistent with the long cycle of 54 years found by Beveridge. He concluded that analyses such as this 'may suffice to suggest the interesting way in which the serial correlations can be used to bring out, at least by a rough first analysis, the predominant characteristics of a given series. In the series in question there can be no doubt about the differences being oscillatory' (*ibid.*, page 47).

Yule finally compared the curve for $h = 5$ with a compound cosine curve constructed by taking the predominant periodicities found by Beveridge (see *ibid.*, Tables XV and XVI). These are plotted together in Figure 5.17. Although there is only a rough agreement between the two plots, Yule felt that, given the circumstances, 'the agreement is, perhaps, as good as we have any right to expect' (*ibid.*, page 49).

5.10 Yule concluded his address with the following summary which, since it encapsulates what are arguably the most important concepts so far introduced for the foundations of time series analysis, is quoted in detail.

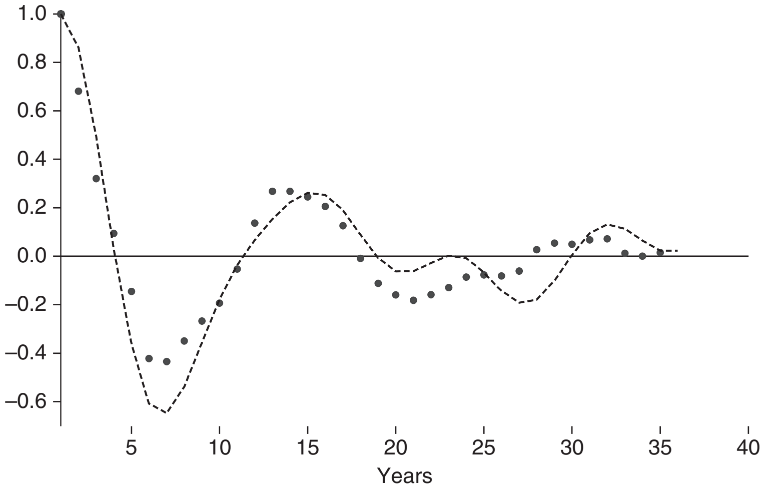


Figure 5.17 Serial difference correlations for $h = 5$ ($r^5(k)$) (dots) and a curve constructed from certain of the periodicities given by Beveridge (dashed line)

> Starting from a question that may have seemed to some silly and unnecessary, we were led to investigate the correlations between samples of two simple mathematical functions of time. It appeared that small samples ... of such functions tended to give us correlations departing as far as possible from the truth, the correlations tending to approach ±1 if the time for which we had experience was very small compared with the time necessary to give the true correlation. Asking ourselves, then, what types of statistical series might be expected to give results analogous to those given by the mathematical function considered, we were led to a classification of series by their serial correlations $r(1), r(2), r(3), \ldots, r(k)$, $r(k)$ being the correlation between terms t and $t+k$. The important matter in classification was the *form* of the function relating $r(k)$ to k, which indicated the nature of the serial correlations between *differences* of the time series. If this function is linear, the time-series has random differences; if it gives a graph concave downwards the difference correlations are positive. We concluded that it was series of the latter type (positively correlated series with positively correlated differences, or conjunct series with conjunct differences to use my suggested term) that formed the dangerous class of series, correlations between short samples tending towards unity. Experimental investigation completely confirmed this suggestion. Samples from conjunct series with random differences gave a widely dispersed distribution of correlations; samples from conjunct series with conjunct differences gave a completely U-shaped distribution, with over one-third of the correlations exceeding ±0.9. (*ibid.*, page 53)

Slutzky and the summation of random causes

5.11 In contrast to Yule, Slutzky focused on inducing serial correlations by taking moving sums of a random series, as he was particularly interested in modelling the recurring cycles that were a predominant feature of economies at the time. He began his paper with this evocative description of the evolution of economic time series.[7]

> Almost all of the phenomena of economic life, like many other processes, social, meteorological, and others, occur in sequences of rising and falling movements, like waves. Just as waves following each other on the sea do not repeat each other perfectly, so economic cycles never repeat earlier ones exactly either in duration or in amplitude. Nevertheless, in both cases, it is almost always possible to detect, even in the multitude of individual peculiarities of the phenomena, marks of certain approximate uniformities and regularities. The eye of the observer instinctively discovers on waves of a certain order other smaller waves, so that the idea of harmonic analysis, viz., that of the possibility of expressing the irregularities of the form and spacing

> of the waves by means of the summation of regular sinusoidal fluctuations, presents itself to the mind almost spontaneously. (Slutzky, 1937, page 105)

However, Slutzky was not convinced that harmonic analysis, as discussed in Chapter 3, was necessarily the appropriate way to model such fluctuations. Although unsatisfactory aspects of the fit of a periodogram might be explained by 'casual' deviations superposed on regular waves or, if there seemed to be shifts in the fit across the first and second halves of the sample, by the interference of factors that caused the original regularity to be replaced by a new one of a similar type (recall the shift in the sunspot periodogram identified in §**3.7**), Slutzky felt that empirical series were typically too short for such hypotheses to be either conclusively proved or refuted. Moreover, the typical assumption of periodogram analysis, that successive terms of a series were independent, was clearly false, leading him to ask

> is it possible that a definite structure of a connection between random fluctuations could form them into a system of more or less regular waves? ... What means of explanation ... would be left to us if we decided to give up the hypothesis of the superposition of regular waves complicated only by pure random components? ... (*T*)*he undulatory character of the processes and the appropriate regularity of the waves* are the two facts for which we shall try to find a possible source in random causes combining themselves in their common effect. (*ibid.*, pages 106–7; italics in original)

5.12 Slutzky began to answer this question by defining

> two kinds of chance series: (1) those in which the probability of the appearance, in a given place in the series, of a certain value of the variable, depends on previous or subsequent values of the variable, and (2) those in which it does not. In this way we distinguish between *coherent* and *incoherent* (or random) series. The terms of the second series are not correlated. In series in which there is correlation between terms, one of the most important characteristics is the value of the coefficient of correlation between terms, considered as a function of the distance between the terms correlated. We shall call it the *correlational function*. (*ibid.*, pages 107–8; italics in original)

The similarity with Yule's analysis is clear, with 'coherent' and 'correlational function' being used instead of 'conjunct' and 'serial correlations'. Slutzky then defined the correlational function more precisely, by limiting

> our investigation to those cases in which the distribution of probability remains constant. The coefficient of correlation, then, is exclusively

> determined by the distance between the terms and not by their place in the series. The coefficient of correlation of each member with itself ($r(0)$) will equal unity, and its coefficient of correlation ($r(k)$) with the kth member following will necessarily equal its coefficient ($r(-k)$) with the kth member preceding. (*ibid.*, page 108; notation altered to maintain consistency)

Slutzky next introduced the idea that 'any concrete instance of an experimentally obtained chance series' should be regarded 'as a *model* of empirical processes which are structurally similar to it' (*ibid.*, page 108). He then considered coherent series that were derived by *moving summations* of either another coherent series or an incoherent series. If the 'causes', $\dots, x_{t-2}, x_{t-1}, x_t, \dots$ produced the 'consequences' $\dots, y_{t-2}, y_{t-1}, y_t, \dots,$ so that each consequence was determined by the influence of a number of preceding causes, then

$$\begin{aligned} y_t &= A_0 x_t + A_1 x_{t-1} + \cdots + A_{n-1} x_{t-(n-1)} \\ y_{t-1} &= A_0 x_{t-1} + \cdots + A_{n-2} x_{t-(n-1)} + A_{n-1} x_{t-n} \\ \dots \end{aligned}$$

Each consequence y_t has one particular cause of its own, x_t, and $n-1$ causes in common with y_{t-1}. As the successive consequences possess common causes there will be a correlation between them even if the sequence of causes is random. Thus, suppose that the causes are indeed random, so that (cf. §**4.1**)

$$E(x_t) = 0 \quad E(x_t^2) = \sigma_x^2 \quad E(x_t x_s) = 0, \quad t \neq s$$

and that the consequences are given by

$$y_t = \sum_{i=0}^{n-1} A_i x_{t-i} \tag{5.7}$$

It therefore follows that

$$E(y_t) = 0$$

$$E(y_t^2) = \sigma_y^2 = \sigma_x^2 \sum_{i=0}^{n-1} A_i^2$$

$$E(y_t y_{t+k}) = \sigma_x^2 \sum_{i=0}^{(n-1)-k} A_i A_{i+k}$$

Since these expectations do not depend on t, the serial correlation $r_y(k)$ is also independent of t:

$$r_y(k) = \frac{\sum_{i=0}^{(n-1)-k} A_i A_{i+k}}{\sum_{i=0}^{n-1} A_i^2} \tag{5.8}$$

from which it immediately follows that

$$r_y(0) = 1 \quad r_y(k) = r_y(-k) \quad r_y(k) = 0, \quad k \geq n$$

When all the weights are equal ($A_0 = A_1 = \cdots = A_{n-1}$), so that (5.7) defines simple moving summation, the serial correlation coefficients will be given by

$$r(0) = 1, r(1) = r(-1) = \frac{n-1}{n}, \ldots, r(n-1) = r(-(n-1)) = \frac{1}{n} \tag{5.9}$$

with $r(k) = r(-k) = 0$ for $k \geq n$.[8]

The process of moving summation may be repeated. Consider an s-fold moving summation of x_t. This will produce, successively,

$$x_t^{(1)} = \sum_{i=0}^{n-1} a_i^{(1)} x_{t-i}; \; x_t^{(2)} = \sum_{i=0}^{n-1} a_i^{(2)} x_{t-i}^{(2)}; \; \ldots$$

$$y_t = x_t^{(s)} = \sum_{i=0}^{n-1} a_i^{(s-1)} x_{t-i}^{(s-1)} = \sum_{i=0}^{s(n-1)} A_i^{(s)} x_{t-i} \tag{5.10}$$

which is clearly of the form (5.7).

5.13 Slutzky used various derived series in his empirical work, each based on the final digits of lottery numbers.[9] Since they are effectively all derived from uniformally distributed random variables, we use here for our recreation of Slutzky's simulations a series of length 1,000 drawn from the integers $0, 1, \ldots, 9$, a short segment of this 'basic' series x_t being shown in Figure 5.18. Setting $n = 10$, $A_i = 1$ for $0 \leq i \leq 9$ and $A_i = 0$ for $i \geq 10$ in (5.7) yields the y_t series, termed Model I, shown in Figure 5.19. Model II is obtained by repeating this moving summation on y_t itself (so that 'in turn, the consequences become causes'), this being shown in Figure 5.20.

Model III is derived by using a moving sum whose weights are defined in the following way:

$$A_i = 10^4 \exp\left(\tfrac{1}{2}(0.1(i-47)^2\right)/\sqrt{2\pi}$$

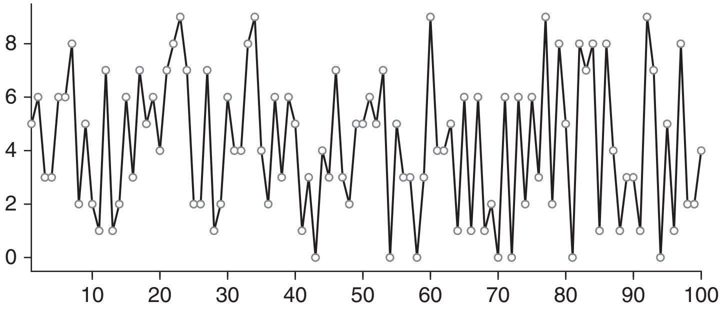


Figure 5.18 The first 100 terms from the basic series

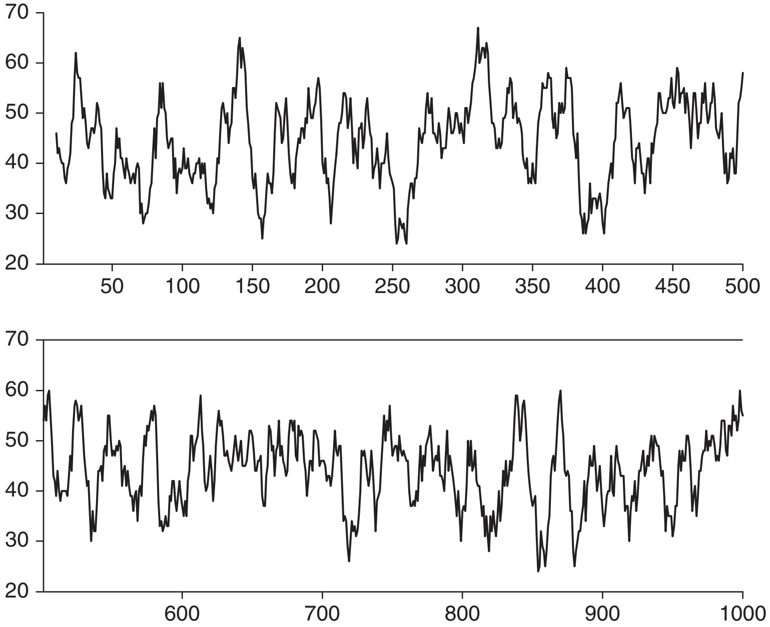


Figure 5.19 Model I constructed from the first 1,000 terms of the basic series

i.e., the weights trace out a scaled Gaussian curve. This series is shown in Figure 5.21. Finally, Model IVa is derived by taking a moving summation of order two 12 times in succession (i.e., $n=2$, $s=12$ in (5.10)), which is easily shown to imply that

$$A_i^{(12)} = {}_{12}C_i \quad i=0,\,1,\,\ldots,\,12, \quad A_i=0 \quad i>12$$

Models IVb and IVc are then obtained as the first and second differences of Model IVa. All three series are shown in Figure 5.22.

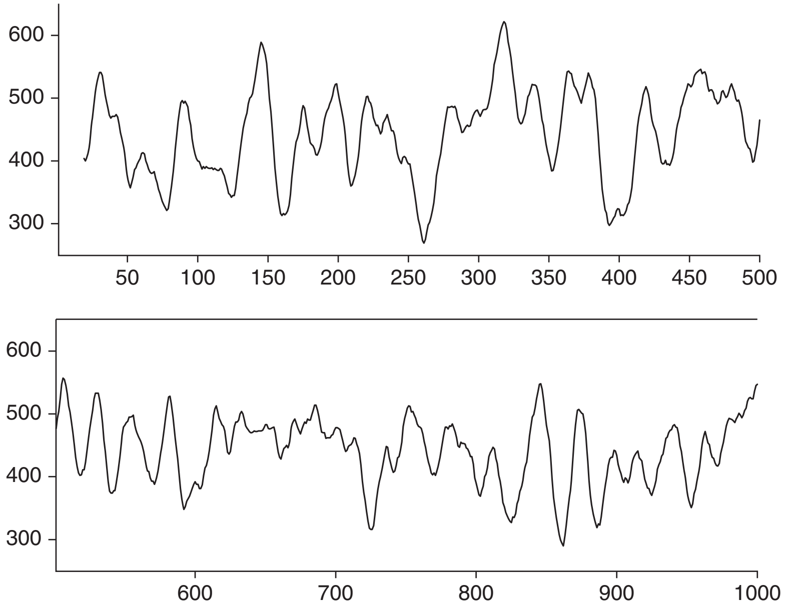


Figure 5.20 Model II constructed from the first 1,000 terms of the basic series

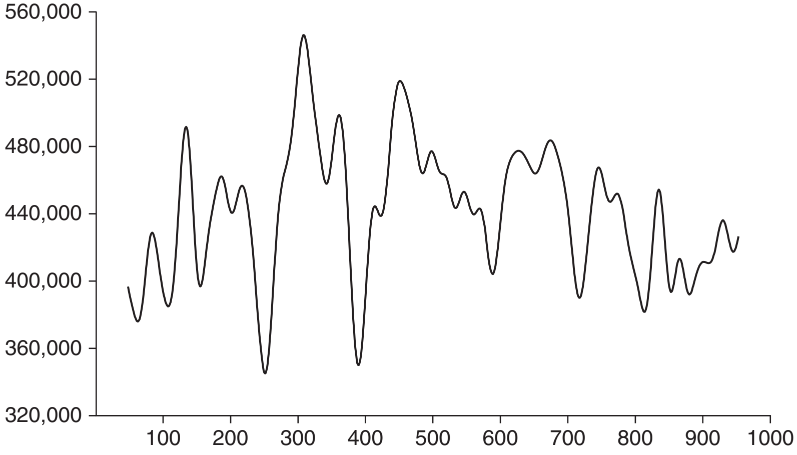


Figure 5.21 Model III constructed from the first 1,000 terms of the basic series

5.14 The serial correlations of the various series can be obtained using (5.8). For Model I, putting $n = 10$ into (5.9) gives, for $y_{I,t} = \sum_{i=0}^{10} x_{t-i}$,

$$r_I(k) = \begin{cases} 1 - (k/10) & 0 \leq k \leq 9 \\ 0 & k \geq 10 \end{cases}$$

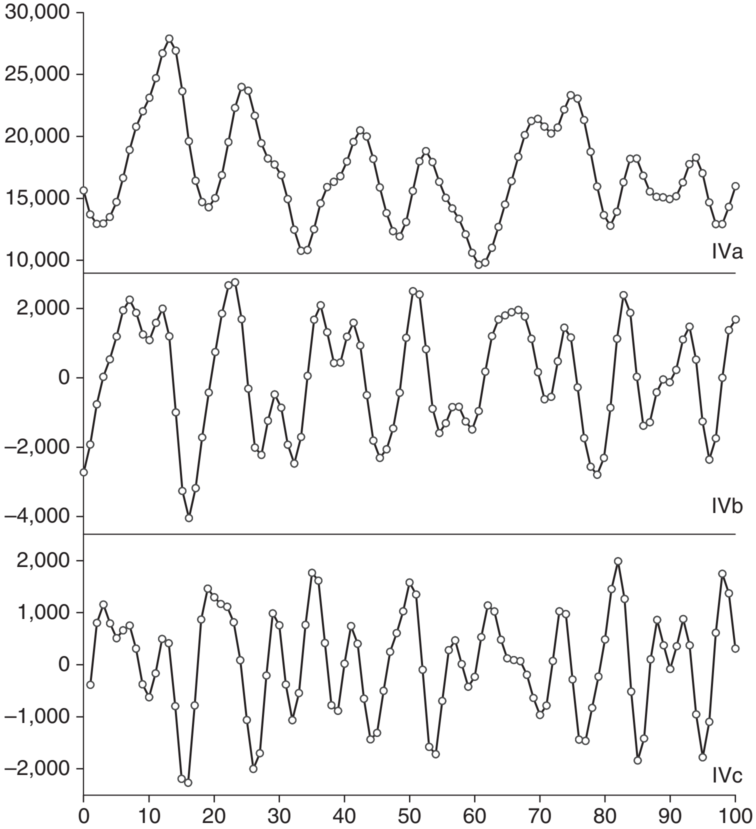


Figure 5.22 The first 100 terms of Models IVa, IVb and IVc

Using (5.10), Model II is given by $y_{II,t} = \sum_{i=0}^{18} A_i^{(2)} x_{t-i}$, where

$$A_i^{(2)} = \begin{cases} i+1 & 0 \leq i \leq 9 \\ 19-i & 10 \leq i \leq 18 \end{cases}$$

For Model IVa, inserting $A_i^{(12)} = {}_{12}C_i$ into (5.9) and noting that, from §**4.3** (and by obvious extension),

$$\sum_{i=0}^{12} {}_{12}C_i^2 = {}_{24}C_{12} \qquad \sum_{i=0}^{12} ({}_{12}C_i)({}_{12}C_{i+k}) = {}_{24}C_{12-k}$$

yields

$$r_{IVa}(k) = \frac{{}_{24}C_{12-k}}{{}_{24}C_{12}} = \frac{12!\,12!}{(12-k)!(12+k)!} = \frac{12.11\ldots(12-(k-1))}{13.14\ldots(12+k)}$$

Models IVb and IVc are given by, respectively,

$$y_{IVb,t} = \Delta y_{IVa,t} = \sum_{i=0}^{12} ({}_{12}C_i)\Delta x_{t-i}$$

and

$$y_{IVc,t} = \Delta y_{IVb,t} = \Delta^2 y_{IVa,t} = \sum_{i=0}^{12} ({}_{12}C_i)\Delta^2 x_{t-i}$$

By extending the results given in §**4.12** and using those of Anderson (1923), Slutzky (*ibid.*, pages 141–2) showed that the serial correlations of these two series were given by

$$r_{IVb}(k) = {}_1r_{IVa}(k) = \frac{\Delta^2 r_{IVa}(k-1)}{\Delta^2 r_{IVa}(-1)} \quad r_{IVc}(k) = {}_2r_{IVa}(k) = \frac{\Delta^4 r_{IVa}(k-2)}{\Delta^4 r_{IVa}(-2)} \tag{5.11}$$

Finally, Slutzky (*ibid.*, page 139) showed, by an argument too tangential to the theme being developed here to be discussed in any detail, that the serial correlations of Model III can be approximated by

$$r_{III}(k) = \exp(-k^2/400)$$

The serial correlations of the various models are plotted in Figure 5.23. Slutzky went on to show that the 'correlational function' – the set of serial correlations as a function of the lag *k* – of a derived series could either be approximated by

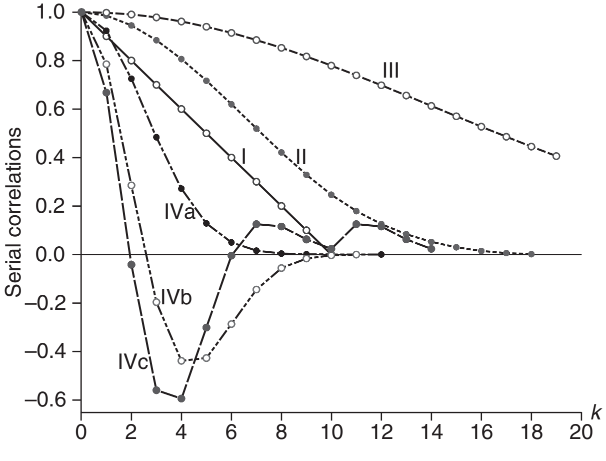


Figure 5.23 Serial correlations of Models I–IVc

a Gaussian curve (Models II–IVa) or, if the series was obtained by dth differencing, by the $2d$th differences of the ordinates of the Gaussian curve (Models IVa and IVb).

5.15 Slutzky used the serial correlation structure of the derived series shown in Figure 5.23 to argue that these models 'give an inductive proof of our first thesis, namely, *that the summation of random causes may be the source of cyclic, or undulatory processes*' (*ibid.*, page 114; italics in original). He went on:

> If a variable ... happens to remain above (or below) its general level, then in that interval it will have a temporary level about which it will almost certainly oscillate. Thus on the waves of one order there appear superimposed waves of another order.
>
> The unconnected random waves are usually called irregular zigzags. A correlation between the items of a series deprives the waves of this characteristic and introduces into their rising and falling movements an element of *graduality*. (*ibid.*, page 116; italics in original)

We can see this property operating in the series of Model I, shown in Figure 5.18, where there are 'gradual transitions from the maximum point of a wave to its minimum and vice versa, since the correlation between neighboring items of the series makes small differences between them more probable than large ones' (*ibid.*, page 116). Slutzky makes a distinction between *graduality* and *fluency*: 'we could speak about the absence of the latter property if a state of things existed where there would be an equal probability for either a rise or a fall after a rise as well as after a fall. If fluency were missing we should obtain waves covered by zigzags such as we find in Model I' (*ibid.*, page 116). Recall that Model I is defined as

$$\begin{aligned}
y_1 &= x_1 + x_2 + x_3 + \cdots + x_{10}, \\
y_2 &= x_2 + x_3 + \cdots + x_{10} + x_{11}, \\
y_3 &= x_3 + \cdots + x_{10} + x_{11} + x_{12}, \\
&\ldots
\end{aligned}$$

so that its first differences are

$$\begin{aligned}
\Delta y_1 &= y_2 - y_1 = x_{11} - x_1, \\
\Delta y_2 &= y_3 - y_2 = x_{12} - x_2, \\
&\ldots
\end{aligned}$$

Thus we see that adjacent first differences have no causes in common and hence are uncorrelated. The same will apply to pairs of differences that are further apart, except for the pairs $\Delta y_1 = x_{11} - x_1$ and $\Delta y_{11} = x_{21} - x_{11}$, etc. The series of

Table 5.10 Serial correlation coefficients for Models I–III

Model	Coefficient of correlation between:		
	Terms $r_y(1)$	First differences ${}_1r_y(1)$	Second differences ${}_2r_y(1)$
I	0.9	0	−0.5
II	0.985	0.85	0
III	0.9975	0.9925	0.9876

differences will be almost incoherent (uncorrelated) and hence the waves of y will be 'covered by chaotically irregular zigzags' (by extension of the result in §**4.12**, ${}_1r_y(10) = -\frac{1}{2}$ with all other ${}_1r_y(k)$ equal to zero).

If adjacent differences are positively correlated then a rise will tend to be followed by further rises and a fall by further falls, so that

> a steep rise will have the *tendency* to continue with the same steepness, a moderate one with the same moderateness. So small sections of a wave will tend to be straight lines; and the greater the coefficient of correlation between adjacent differences the closer the sections approximate straight lines. (*ibid.*, pages 116–17)

Serial correlations between second differences play an analogous role. The higher the correlation the less variable will be the second differences, so that a series with roughly constant second differences will tend to approximate a second-degree parabola. In Table 5.10 are shown $r_y(1)$, ${}_1r_y(1)$ and ${}_2r_y(1)$ for Models I, II and III, the latter two correlations being computed using (5.10):

> As we go from the ... basic series to Model I and then to Models II and III, we find progressive changes in their graphic appearance (see Figures [5.19, 5.20 and 5.21] respectively). These changes are produced at first by the introduction and then by the growth of graduality and of fluency in the movements of the respective chance waves. The growth of the degree of correlation between items (or between their differences) as we go from the ... basic series to Model I, etc. ... corresponds to changes in the graphic appearance of our series. (*ibid.*, page 117)

5.16 These results were taken by Slutzky as evidence that his first thesis, that of 'inducing undulatory processes of a more or less fluent character as the result of the summation of random causes', may be regarded as being 'practically proved'. He wanted, however, to go further and to demonstrate that the waves of the derived process have an approximate regularity. To do this he re-examined Model II, shown in Figure 5.20, describing its evolution thus:

> In many places there are, apparently, large waves with massive outlines as well as smaller waves lying, as it were, over them; sometimes these are detached from them, sometimes they are almost completely merged into them. (*ibid.*, page 120)

This led Slutzky to suggest that, across the set of models being considered,

> (a) careful examination of graphs of our models will disclose ... a number of places where the approximate equality of the length of the waves is readily apparent. If we had a much shorter series, such as series offered by the ordinary statistics of economic life with its small number of waves, *we should be tempted to consider the sequence as strictly periodic, that is, as composed of a few regular harmonic fluctuations complicated by some insignificant casual fluctuations.* (*ibid.*, page 120; italics added for emphasis)

By fitting sums of sinusoids to Models II and III, and observing how different sub-periods of the series were approximated by sinusoids of the same type but with different parameters, Slutzky arrived at the following tentative hypothesis.

> *The summation of random causes generates a cyclical series which tends to imitate for a number of cycles a harmonic series of a relatively small number of sine curves. After a more or less considerable number of periods every regime becomes disarranged, the transition to another regime occurring sometimes rather gradually, sometimes more or less abruptly, around certain critical points.*
>
> In addition to the tendencies towards graduality and fluency (that is towards linear and parabolic forms for small sections) we find a third tendency, namely, the tendency toward a sinusoidal form. (*ibid.*, pages 123–4; italics in original)

To investigate this tendency, consider the simple harmonic function of §**5.2**:

$$x_t = \sin\left(2\pi\frac{t+\alpha}{n}\right)$$

from which it follows that

$$\Delta^2 x_t = -4\sin^2\pi\frac{1}{n}x_{t+1} = -\theta x_{t+1} \tag{5.12}$$

The proof of this fundamental result uses standard trigonometric identities. If we define

$$A = 2\pi \frac{t+\alpha+1}{n} \quad B = 2\pi \frac{1}{n}$$

then

$$\begin{aligned}\Delta^2 x_t &= x_{t+2} - 2x_{t+1} + x_t = \sin(A+B) - 2\sin A + \sin(A-B)\\ &= 2\sin A\cos B - 2\sin A\\ &= 2\sin A(1 - 2\sin^2(B/2)) - 2\sin A\\ &= -4\sin^2 B\sin A = -\theta x_{t+1}\end{aligned}$$

on using, first, the addition theorem $\sin(A \pm B) = \sin A\cos B \pm \cos A\sin B$; second, the double-angle formula $\cos B = 1 - 2\sin^2(B/2)$; and, finally, setting $\theta = 4\sin^2(2\pi/n)$.

If there is a high correlation between $\Delta^2 x_t$ and x_{t+1} then (5.12) will be approximately true and there will exist a tendency towards a sinusoidal form in the series: the closer this correlation is to -1, the more pronounced will this tendency be. In fact, what will happen is that the sinusoidal form will appear as a sequence of 'regimes' which disrupt gradually depending on the size of the correlation: the accumulation of deviations will necessarily destroy every regime but will then create a new regime having different parameters, so that a coherent series is patched together out of a number of sinusoids whose parameters vary in an unpredictable way. By extending (5.11) the correlation between $\Delta^2 x_t$ and x_{t+1} may be denoted as

$$\tfrac{1}{2}r_x(1) = \frac{\Delta^2 r_x(-1)}{\sqrt{\Delta^4 r_x(-2)}}$$

The correlations for Models I-IVa are, respectively, $\tfrac{1}{2}r_I(1) = -0.316$, $\tfrac{1}{2}r_{II}(1) = -0.315$, $\tfrac{1}{2}r_{III}(1) = -0.578$ and $\tfrac{1}{2}r_{IVa}(1) = -0.599$, none of which are particularly close to -1. (Note that a tendency towards either a linear or a parabolic form cannot appear in a long section of a coherent series because this would disrupt its cyclic character.) Of course, (5.12) holds only for a single sinusoid and cannot be applied to a sum of sinusoids having different periods, which is the traditional setup of periodogram analysis (see §§**3.1–3.5**).

These informal ideas were expressed more precisely in Slutzky's Theorem A (*ibid.*, page 130: proof given on pages 142–4), which he termed the *Law of the Sinusoidal Limit*. This considered a series y_t with the following properties:

$$E(y_t) = 0, \quad E(y_t^2) = \sigma_y^2 = f(n)$$

$$\frac{E(y_t y_{t+k})}{E(y_t^2)} = r_y(k) = \phi(k, n)$$

where n is a parameter related to the series as a whole and $f(n)$ and $\phi(k, n)$ are independent of k. If $r_y(1)$ and $\frac{1}{2}r_y(1)$ satisfy the conditions

$$|r_y(1)| \leq c < 1 \quad \lim \tfrac{1}{2}r_y(1) = -1$$

as $n \to \infty$, then: (1) for arbitrarily small ε and η and for s arbitrarily large, there will exist a number, n_0, such that, for every $n > n_0$, the probability that the absolute deviations of a sequence of observations $y_\tau, y_{\tau+1}, \ldots, y_{\tau+s}$ from a certain sinusoid will not exceed $\varepsilon\sigma$ will be greater than $1 - \eta$; (2) the period of this sinusoid will be determined by the equation $\cos(2/L) = r_y(1)$; (3) the number of periods in the interval $(\tau, \tau + s)$ will be arbitrarily large provided s and n are taken large enough.

The practical usefulness of this theorem was provided by Theorem B, also proved in (*ibid.*, page 144–5). If x_t is a random series from which are derived

$$x_t^{(1)} = x_t + x_{t-1}, x_t^{(2)} = x_t^{(1)} + x_{t-1}^{(1)}, \ldots, x_t^{(n)} = x_t^{(n-1)} + x_{t-1}^{(n-1)}$$

and

$$y_t = \Delta^m x_t^{(n)}$$

then y_t will tend to obey the law of the sinusoidal limit provided m and n increase indefinitely and the ratio m/n is constant.

Both propositions can be generalized to the case of a series coinciding with the sum of a number of sinusoids, but the practical coincidence does not extend to the series as a whole, for the respective sinusoids of closest fit differ across intervals. This is plainly the case for the series in Theorem B, since, in the limit, y_t and y_{t+k} will be independent of each other as soon as $k > m + n + 1$, so that the phases and amplitudes of the sinusoids practically coincident with the partial series $y_t, y_{t+1}, \ldots, y_{t+s}$ and $y_{t+k+s}, y_{t+k+s+1}, \ldots, y_{t+k+2s}$, respectively, will also be independent of each other provided $k > m + n + 1$. Slutzky thus felt entitled to conclude that

> the chance functions of the type just considered appearing on the one end of the scale, and the random functions on the other, there evidently must exist all possible intermediate gradations between these extremes. The ability of the coherent chance series to simulate the periodic, or the nearly periodic, functions, seems thus to be definitely demonstrated. (*ibid.*, page 132)

5.17 Before the English translation of Slutzky's paper appeared in *Econometrica* in 1937, Kuznets (1929) had provided an introduction, a detailed interpretation, and an extension of Slutzky's ideas. In fact, economists had already become familiar with the consequences of cumulation, as shown by the extended, and extremely apposite, discussion in Bullock, Persons and Crum (1927, pages 80–4). Kuznets explained the appearance of cyclical patterns in the cumulation of a random series in the following way. Although the successive observations of a purely random series were uncorrelated, the successive observations of a cumulation or moving average of this random series, because they will contain many identical observations, will necessarily be serially correlated: the longer the moving average, the higher the serial correlation. Moreover, because there will typically be clusters of positive and negative values in the random series, their cumulation will almost certainly impart cyclical fluctuations into the moving average. This interpretation was directly in accord with Slutzky, but Kuznets offered a second source of cyclical patterns in a moving average of a random series: such patterns could be induced by an extremely large value of the random series. This would be included in several of the observations of the moving average and so would raise or depress the level of a sequence of observations, thus inducing cyclical swings. Consequently, Kuznets argued that both the shape of the distribution underlying the random series and the length of the moving average would influence the amplitude and period of the cycles so generated: a skewed, peaked distribution would be the most likely source of clear-cut cycles.

Thus the cumulation of random shocks could produce cyclical movements, but Kuznets was wary of making the inverse inference that actually observed cycles were indeed caused by cumulating random shocks.

> It has been shown that the summation of random causes yields cycles, that certain peculiarities of the distribution of these random causes and of the averaging process influences the characteristics of the cyclical oscillations obtained. But can one invert the proposition and say that, therefore, cyclical oscillations may be conceived primarily as results of summations of random causes, and that the characteristics of some of these cyclical oscillations can best be grasped as a result of the underlying random events or of the process of cumulation?
>
> Such inference, of course, cannot carry with it any certainty, since we are never certain of all the contingencies of other hypotheses to which the formations of cycles in economic data may be reduced. But the only way to establish the significance of the conclusions stated, is to proceed from it as from a significant but not exclusive hypothesis and to see how it agrees or disagrees with the other elements of the known universe which are associated with the cyclical oscillations. (Kuznets, 1929, pages 273–4)

5.18 Slutzky's law of the sinusoidal limit was subsequently extended by Romanovsky (1932, 1933), who showed that it continued to hold when the length of the summation was increased. He also obtained a necessary and sufficient condition for the limit to be a sum of several sinusoids and also relaxed the assumption that the original series had to be purely random. Moran (1949) was later able to shorten the proof of the theorem using newly developed analytical techniques (see §**8.11**) and, in Moran (1950), he went on to show that taking repeated moving averages of random series could result in several sinusoids whose number and periods depended on the weights of the moving averages.

Working and random-difference series

5.19 Interestingly, Kuznets ruled out the possibility that, at least economic, time series could be generated by simple cumulations of all past shocks:

> (w)e shall omit the straight cumulation altogether, because it is too far removed from reality. In economic life, there is no perpetual influence of an event once occurred on all the subsequent events. Rather, we must conceive of it in the nature of a moving average in the sense that after a certain time its influence is reduced to zero and the item must be omitted altogether from the averaging process. (Kuznets, 1929, page 269)

This was not the view of Holbrook Working, however, who observed just a few years later that

> time series commonly possess in many respects the characteristics of series of cumulated random numbers. The separate items in such time series are by no means random in character, but the changes between successive items tend to be largely random. . . . The fact that series commonly used as indexes of business activity closely resemble series obtainable by cumulating random numbers has given support to the theory that so called business cycles result in large degree from cumulative effects of independent random influences bearing on the business situation – some favourably, some unfavourably. (Working, 1934, page 11)

Although he attributed this view to Slutzky (1927), Working took the idea somewhat further, first introducing the term *random-difference series* to describe a series obtained by cumulating random numbers, 'since it is the first differences of the series and not the items of the series itself which are random', and then contrasting such series with the typical view of an economic time series promulgated by both economic statisticians and theorists alike:

> Economic theory has fallen far short of recognising the full implications of the resemblance of many economic time series to random-difference series; and methods of statistical analysis in general use have given these implications virtually no recognition. Economic theories and the techniques that have been employed in analysis of time series generally deal in terms of norms and of deviations therefrom. The norms may be regarded as constants (as is common in economic theory) or they may be regarded as changing progressively (represented in statistical practice by trend lines). The deviations from norms or from trends are commonly regarded as having one of three characteristics: (a) that of random deviation, each item independent of all others; (b) that of cycles, either regular in periodicity and amplitude or irregular, but in any case with a definite tendency for deviation in one direction to be followed after an interval by deviation in the other (not by accident but in consequence of a specific reaction tendency); or (c) some combination of random deviations with a cyclical tendency, or with several cyclical tendencies of differing period and amplitude. (*ibid.*, pages 11–12)

Working thus clearly took 'standard' economic practice of the time as lying firmly within a trend stationary framework (recall the earlier discussion in §**4.11**), regarding this practice as being

> inappropriate and misleading when applied to cases in which the dominant tendency is for the effects of successive events to be independent and cumulative. An outstanding characteristic of a series of this type is that its *changes* are largely random and unpredictable. Even in a purely random series (one of random deviations from a norm) in which individual values are unpredictable, *changes* are predictable with considerable accuracy ... In a series characterised by primarily random *changes*, however, absolute values of immediately subsequent items are predictable with an accuracy that for many purposes may be regarded as very satisfactory, but subsequent changes are largely unpredictable. (*ibid.*, page 12; italics in original)

What Working was getting at in this quote is the following. Suppose that x_t is the random series of §**5.12** and $y_t = \sum_{i=0}^{t-1} x_{t-i}$ is its cumulation. The correlation between x_t and its immediately subsequent change, $\Delta x_t = x_{t+1} - x_t$, is, using the notation and results of §**5.14–5.16**,

$$ {}_1^0r_x(0) = \frac{(-1)\Delta r_x(-1)}{\sqrt{(-1)\Delta^2 r_x(-1)}} = \frac{r_x(1) - r_x(0)}{\sqrt{(-1)(r_x(1) - 2r_x(0) - r_x(1))}} = -\frac{1}{\sqrt{2}} = -0.7071 $$

whereas the correlation between y_t and y_{t+1} is, by using (5.8) with $A_i = 1$ and $n = t$, $r_y(1) = (t-1)/t \approx 1$ for large t, even though $r_x(1) = 0$. Thus,

> in a purely random series (one of random deviations from a norm) conspicuous trends will be found. Such 'trends,' however, must be regarded merely as generalized descriptions of the course of the series over a certain period, not as norms, nor as bases for predicting the future course of the series over even the briefest subsequent period. (*ibid.*, page 12)

5.20 Although Working thought that few of the time series actually encountered in practice would have purely random changes, he nevertheless felt that 'the characteristic of random changes is present in such important degree in ... many ... economic series, as to deserve serious consideration'. His aim was then to present a long *standard* random-difference series that was known to possess 'those characteristics and those alone' and to this end presented a series of length 2,400 that was the cumulation of random normal variates with standard deviation 10.[10] We recreate this simulation in Figure 5.24 where, following Working, we arrange the data as a weekly series observed over 47 years (1900 to 1946). To give a clearer impression of the course of the entire series than can be obtained from the separate detailed segments shown in Figure 5.24, we plot 'annual' averages, along with 'mid-year' values, in Figure 5.25.

Working's discussion of the uses to which such an 'experimental' series could be put is worth quoting in detail, as it predates many of the subsequent discussions of the behaviour of what is nowadays referred to as an integrated process of order one (see §**10.18**).

> An important application of the random-difference series here presented will be found in visual comparisons with actual time series, and probably in comparisons of appropriate statistical constants derived from the experimental random-difference series and from the actual series, as an aid in ascertaining whether and to what extent the actual series shows the characteristics of a cumulation of random changes.... It will be apparent from brief study of the curves in [Figure 5.24] that an essentially random-difference series of only 200–300 items [for example, an individual segment in Figure 5.24], which would be regarded as a very long annual series or a fairly long monthly series in economic statistics, might very easily be taken mistakenly to be a series dominated by a true irregular cycle with superimposed random fluctuations.
>
> In the actual statistical analysis of time series found to have important or dominant random-difference characteristics, a number of applications of the accompanying 'experimental time series' readily suggest themselves. Some students of stock and commodity prices attribute great forecasting significance to certain 'formations' that appear more or less conspicuously in the

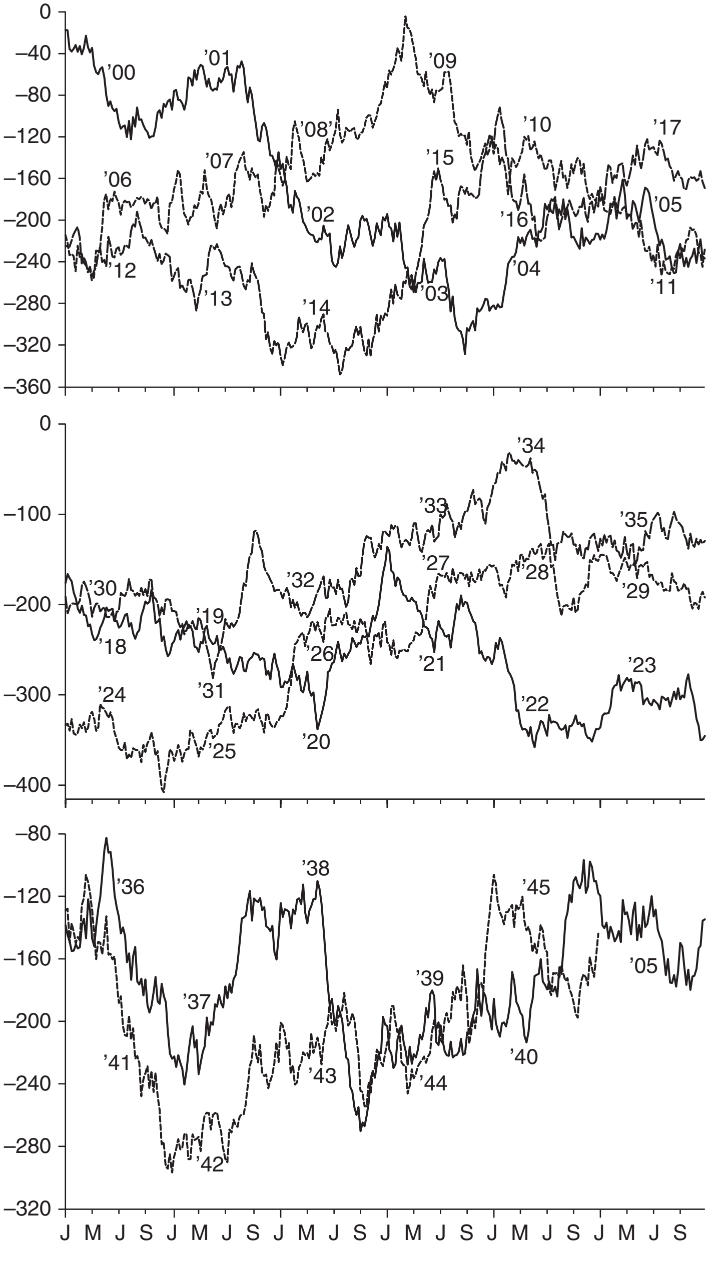


Figure 5.24 A random-difference experimental time series

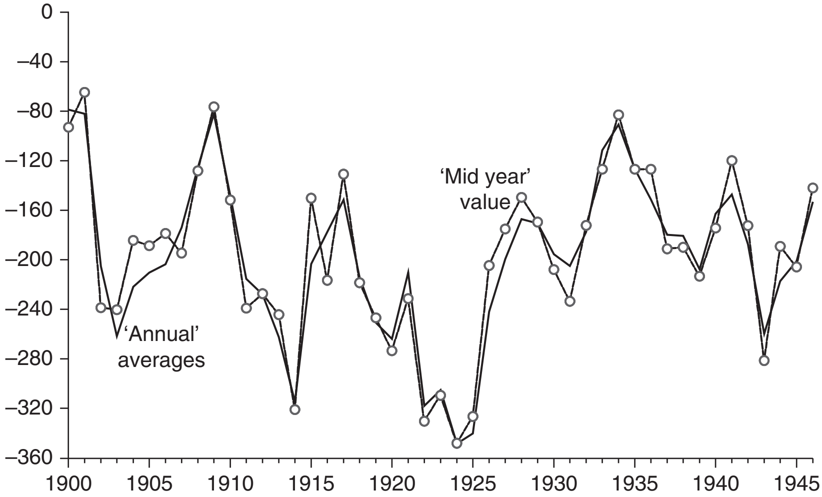


Figure 5.25 'Annual' averages and mid-points of random-difference experimental time series

> charted price data – such as 'resistance and support levels,' 'lines,' and 'head and shoulder formations.' Other students of these prices scoff at such ideas. Valuable evidence on the probable significance of such 'formations' could be obtained by ascertaining with what relative frequency they are found in a random-difference series. If they occur as frequently and as clearly in the random-difference series as in the actual series, it is to be supposed that they are without forecasting significance, for it is known that changes in a random-difference series are quite unpredictable. (*ibid.*, page 21)

This discussion of stock and commodity price formations – the foundation of chartism and technical analysis[11] – was returned to later by Kendall (1953) and Roberts (1959) and is key to understanding the behaviour of financial time series: see §§**11.2–11.3**. However, Working's concluding paragraph, and in particular the footnote that comes at the very end of his paper, is even more prescient.

> Finally it should be noted that the series here presented cannot fulfill all the requirements that may arise for an experimental random-difference series. This series represents only one possible type of class, though one chosen because of its probably superior generality. Other types, with some notably different features, may be obtained by varying in some systematic fashion the standard deviation of the population from which the drawings are made.† *I find that to the important extent that wheat prices resemble a random-difference series, they resemble most closely one that might be derived by cumulating random*

> *numbers drawn from a slightly skewed population of standard deviation varying rather systematically through time.*
>
> †It may be questioned whether a series so drawn should be regarded as strictly random. Systematic variation of the standard deviation of the population, its mean being kept at zero, would introduce no correlations among algebraic values of the numbers drawn, *but would introduce correlations among their values, signs neglected.*
>
> (*ibid.*, page 24; italics added for emphasis)

From a modern perspective, Working had clearly hit upon the idea of a martingale difference having time-varying variances (which may be interpreted as measuring volatility), so planting the seed that, some half a century later, would begin to grow dramatically into the Amazonian forests of the autoregressive conditional heteroskedastic (ARCH) model of Engle (1982) and its myriad generalizations (see §**16.11** and, for example, Gouriéroux, 1997, for a recent and detailed treatment)!

5.21 These three key papers thus introduced and analysed a number of fundamental concepts in the foundations of time series analysis, some of the most important being: (i) the formalization of the concept of serial correlation; (ii) the analysis of the conditions under which nonsense correlations between time series are likely to exist; (iii) the demonstration that cyclical fluctuations may be induced by summing a sequence of random causes, and that these random shocks will also lead to the breakdown and subsequent reformation of any cyclical pattern; and (iv) that the cumulation of random observations will produce a series that exhibits short-term trends and cycles and other familiar patterns and hence gives the appearance of apparent predictability, even though such predictability is illusory. These fundamental themes will reappear in both the theory and practice of time series analysis throughout our development.

6

Periodicities in Sunspots and Air Pressure: Yule, Walker and the Modelling of Superposed Fluctuations and Disturbances

Yule, superposed fluctuations and disturbances

6.1 At the same time as he was analysing the nonsense correlation problem, Yule was also turning his attention back to harmonic motion and, in particular, to how harmonic motion responds to external shocks. This attention led to yet another seminal paper in the foundations of time series analysis: Yule (1927). Yule's starting point was to take a simple harmonic function of time and to superpose upon it a sequence of random errors. If these errors were small, 'the only effect is to make the graph somewhat irregular, leaving the suggestion of periodicity still quite clear to the eye' (*ibid.*, page 267), and an example of this situation is shown in Figure 6.1(a). If the errors were increased in size, as in Figure 6.1(b), 'the graph becomes more irregular, the suggestion of periodicity more obscure, and we have only sufficiently to increase the "errors" to mask completely any appearance of periodicity' (*ibid.*, page 267). Nevertheless, no matter how large the errors, periodogram analysis would still be applicable and should, given a sufficient number of observations, continue to provide a close approximation to both the period and the amplitude of the underlying harmonic function. Yule referred to this setup as one of *superposed fluctuations* – 'fluctuations which do not in any way disturb the steady course of the underlying periodic function' (*ibid.*, page 268).

But Yule did not see this set-up as being the most likely hypothesis in most physical situations, leading him to suggest a delightful thought experiment, based on the following set-up of a pendulum.

> If we observe at short intervals of time the departures of a simple harmonic pendulum from its position of rest, errors of observation will cause superposed fluctuations of the kind supposed in [Figure 6.1]. But by improvement of apparatus and automatic methods of recording, let us say, errors of observation are practically eliminated. (*ibid.*, page 268)

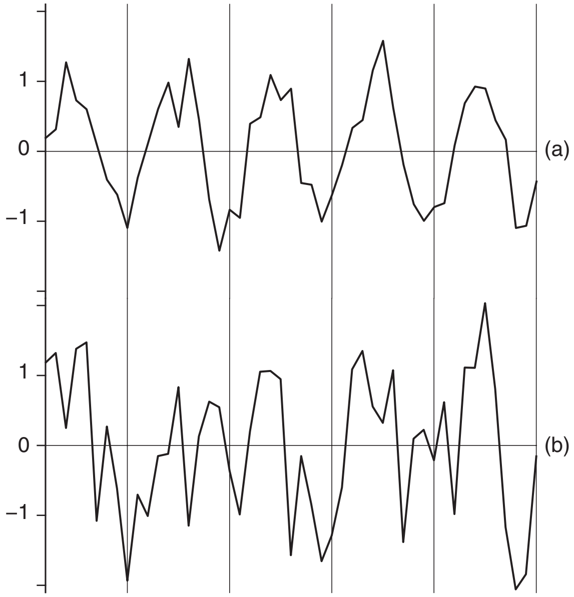


Figure 6.1 Graphs of simple harmonic functions of unit amplitude with superposed random fluctuations: (a) smaller fluctuations; (b) larger fluctuations

The recording apparatus is then left to itself, but

> unfortunately boys get into the room and start pelting the pendulum with peas, sometimes from one side and sometimes from the other. The motion is now affected, not by *superposed fluctuations* but by true *disturbances*, and the effect on the graph will be of an entirely different kind. The graph will remain surprisingly smooth, but amplitude and phase will vary continually. (*ibid.*, page 268; italics in original)

To illustrate this experiment formally, consider the simple harmonic function given by

$$x_t = \rho \sin 2\pi \frac{t}{n} \tag{6.1}$$

where (cf. §§**3.1–3.5**) ρ is the amplitude of the sine wave and n is the period. Using the result (5.12) proved in §**5.15**, (6.1) can be written as

$$\Delta^2 x_t = -4 \sin^2 \pi \frac{1}{n} = -\theta x_{t+1} \tag{6.2}$$

where

$$\theta = 4\sin^2 \pi\frac{1}{n} = 2\left(1 - \cos 2\pi\frac{1}{n}\right) = 2 - 2\cos\vartheta$$

on defining $\vartheta = 2\pi/n$. Equation (6.2) may be written equivalently as

$$x_{t+2} = (2-\theta)x_{t+1} - x_t \tag{6.3}$$

The 'errors' produced by the boys pelting the pendulum with peas leads to the inclusion of an error, ε_{t+2}, in (6.3), which we may rewrite in the more convenient form

$$x_t = (2-\theta)x_{t-1} - x_{t-2} + \varepsilon_t \tag{6.4}$$

Figure 6.2 shows a graph of x_t constructed from (6.4) by setting $n = 10$, so that $\theta = 4\sin^2 18^\circ = 4 \times 0.3090^2 = 0.382$ and thus

$$x_t = 1.618x_{t-1} - x_{t-2} + \varepsilon_t \tag{6.5}$$

Following Yule, ε_t was defined to be 1/20th of the deviation of the sum of four independent throws of a dice from the expected value of the four throws (which is 14). This defines a discrete random variable taking the values $-0.5(0.05)0.5$, with mean zero and standard deviation 0.1708. Setting $x_1 = 0$ and $x_2 = \sin 36^\circ = 0.588$, Figure 6.2 shows the simulation of (6.5) for $t = 1, \ldots, 300$, which led Yule (*ibid.*, page 269) to observe that '(i)nspection of the figure shows that there are now no abrupt variations in the graph, but the amplitude varies within wide limits, and the phase is continually shifting. Increasing the magnitude of the disturbances simply increases the amplitude: the graph remains smooth'.

6.2 Why does the simulated series in Figure 6.2 present such a smooth appearance? An undisturbed harmonic function may be regarded as the solution of the difference equation

$$\Delta^2 x_t + \theta x_{t+1} = 0 \tag{6.6}$$

If the motion is disturbed, however, we now have, say,

$$\Delta^2 x_t + \theta x_{t+1} = \phi(t) \tag{6.7}$$

where $\phi(t)$ is some 'disturbance function'. Hence we see that (6.6) is the complementary function of the solution to (6.7) and $\phi(t)$ is the particular integral.

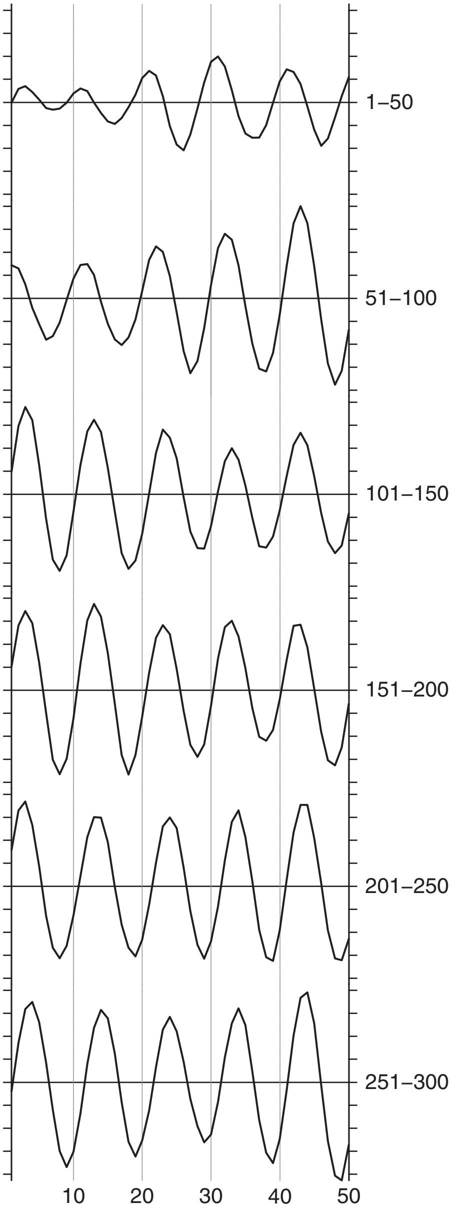


Figure 6.2 Graph of a disturbed harmonic function, equation (6.5)

The solution to (6.4), given initial values x_1 and x_2 and writing $k = 2 - \theta$, is the following series for $t > 2$

$$\begin{aligned} x_3 &= kx_2 - x_1 + \varepsilon_3 \\ x_4 &= (k^2 - 1)x_2 - kx_1 + k\varepsilon_3 + \varepsilon_4 \\ x_5 &= \{k(k^2 - 1) - k\}x_2 - (k^2 - 1)x_1 + (k^2 - 1)\varepsilon_3 + k\varepsilon_4 + \varepsilon_5 \\ x_6 &= \{(k(k^2 - 1) - k) - (k^2 - 1)\}x_2 - \{k(k^2 - 1) - k\}x_1 \\ &\quad + \{k(k^2 - 1) - k\}\varepsilon_3 + (k^2 - 1)\varepsilon_4 + k\varepsilon_5 + \varepsilon_6 \end{aligned}$$

etc.

The coefficients on the ε terms form the sequence $1, k, k^2 - 1, k(k^2 - 1) - k, \ldots$ and hence are related by an equation of the form

$$A_m = kA_{m-1} - A_{m-2}$$

where A_m is the coefficient on ε_m, $m \geq t - 3$. But this is simply an equation of the form (6.3), so that the coefficients on the ε's are therefore the terms of a sine function having the same period as the complementary function (6.6) and with initial terms 1 and k: for our simulated series, they take the values +1, +1.6180, +1.6180, +1, 0, −1, −1.6180, etc.

The first 30 terms of the simulated series, its complementary function and particular integral, and the disturbances are shown in Table 6.1.

> The series tends to be oscillatory, since, if we take adjacent terms, most of the periodic coefficients of the ε's are of the same sign, and consequently the adjacent terms are positively correlated; whereas if we take terms, say, 5 places apart, the periodic coefficients of the ε's are of opposite signs, and therefore the terms are negatively correlated. The series tends to be smooth – i.e., adjacent terms highly correlated – since adjacent terms represent simply differently weighted sums of ε's, all but one of which are the same. (*ibid.*, page 272)

Yule pointed out (in an addition to the original text) that if the initial conditions were set as $x_1 = x_2 = 0$ then there would be no true harmonic component and the series would reduce to the particular integral alone, although the graph of the series would look little different to that shown in Figure 6.2 – 'the case would correspond to that of a pendulum initially at rest, but started into movement by the disturbances' (*ibid.*, page 272).

The peak-to-peak periods range from 8.24 to 10.83 with an average of 10.03, while the trough-to-trough periods range from 8.75 to 10.85 with an average of 10.05, the true period being, of course, 10. Considerations of this type led Yule to conclude that

> (i)t is evident that the problem of determining with any precision the period of the fundamental undisturbed function from the data of such a graph as

Table 6.1 Decomposition of the first 30 terms of the simulated series used in Figure 6.2 into complementary function (simple harmonic function) and particular integral (function of the disturbances alone)

t	Observed x_t	Complementary function	Particular integral	Disturbance ε_t
1	0	0	0	0
2	+0.5878	+0.5878	0	0
3	+0.7014	+0.9511	−0.2497	−0.25
4	+0.4468	+0.9511	−0.5042	−0.10
5	+0.1216	+0.5878	−0.4662	+0.10
6	−0.2501	0	−0.2501	0
7	−0.3262	−0.5878	+0.2616	+0.20
8	−0.2778	−0.9511	+0.6733	0
9	−0.0232	−0.9511	+0.9279	+0.10
10	+0.3902	−0.5878	+0.9780	+0.15
11	+0.6046	0	+0.6046	−0.05
12	+0.4880	+0.5878	−0.0998	−0.10
13	−0.0150	+0.9511	−0.9661	−0.20
14	−0.4623	+0.9511	−1.4134	+0.05
15	−0.8330	+0.5878	−1.4208	−0.10
16	−0.9355	0	−0.9355	−0.05
17	−0.6806	−0.5878	−0.0928	0
18	−0.2158	−0.9511	+0.7353	−0.05
19	+0.3315	−0.9511	+1.2826	0
20	+1.0521	−0.5878	+1.6399	+0.30
21	+1.3709	0	+1.3709	0
22	+1.1659	+0.5878	+0.5781	0
23	+0.2856	+0.9511	−0.6655	−0.25
24	−1.0362	+0.9511	−1.9873	−0.30
25	−1.8422	+0.5878	−2.4300	+0.10
26	−2.0944	0	−2.0944	−0.15
27	−1.3966	−0.5878	−0.8088	+0.15
28	−0.2653	−0.9511	+0.6858	−0.10
29	+0.8674	−0.9511	+1.8185	−0.10
30	+1.7687	−0.5878	+2.3556	+0.10

> [Figure 6.2] is a much more difficult one than that of determining the period when we have only to deal with superposed fluctuations. It is doubtful if any method can give a result that is not subject to an unpleasantly large margin of error if our data are available for no more than, say, 10 to 15 periods. [F]rom mere inspection of [Figure 6.2] it is, I think, clear that [periodogram analysis] must give results subject to a much larger margin of error than is usually supposed – results, consequently, which must be interpreted with the greatest caution, and that if applied to data covering only a few periods it may easily give results which are apparently absurd or highly paradoxical. (*ibid.*, page 278)

6.3 Yule proposed analysing models of the type (6.4) by using least squares regression. If (6.4) is written

$$x_t = kx_{t-1} - x_{t-2} + \varepsilon_t \tag{6.8}$$

and it is assumed that the disturbances ε_t have zero mean, the regression of $x_t + x_{t-2}$ on x_{t-1} will provide an estimate of k and hence of $\cos\vartheta = k/2$, from which estimates of ϑ and the period of the harmonic may be calculated. Yule first obtained such estimates for the simulated series of Figure 6.2, having split the series into two halves of length 150. Here we provide estimates for the complete sample and for the two halves:

Complete sample of 300 terms

$$x_t = 1.62338x_{t-1} - x_{t-2}$$

$$\cos\vartheta = 0.81169; \quad \vartheta = 35.74^\circ; \quad \text{period} = 10.07$$

First 150 terms

$$x_t = 1.62897x_{t-1} - x_{t-2}$$

$$\cos\vartheta = 0.81448; \quad \vartheta = 35.46^\circ; \quad \text{period} = 10.15$$

Second 150 terms

$$x_t = 1.62026x_{t-1} - x_{t-2}$$

$$\cos\vartheta = 0.81013; \quad \vartheta = 35.89^\circ; \quad \text{period} = 10.03$$

The periods thus found are not far from those obtained in §**6.2** and the estimates of k are close to the 'true' value of 1.61803. The three regressions give values of the disturbances which have correlations of +0.998, +0.992 and +0.999 with the true disturbances: 'on the whole, I think that the result may be regarded as reasonably satisfactory' (*ibid.*, page 275).

6.4 Yule then turned his attention to annual sunspot numbers between 1749 and 1924, an updated series of which was introduced and analysed using periodogram methods in §**3.7**. Rather than just focusing on the raw numbers, Yule also constructed a 'graduated' series, defined as

$$x'_t = \frac{w_t}{3} - \frac{\Delta^2 w_{t-1}}{9}$$

where $w_t = x_{t-1} + x_t + x_{t+1}$. Some simple algebra shows that x'_t is the weighted moving average

$$x'_t = \tfrac{1}{9}(-x_{t-2} + 4x_{t-1} + 3x_t + 4x_{t+1} - x_{t+2})$$

The sunspot and the graduated numbers for the years 1700 to 2007 are shown in the top two panels of Figure 6.3. To Yule

> the upper curve in [Figure 6.3] ... suggests quite definitely to my eye that we have to deal with a graph of the type of [Figure 6.2], not of the type of [Figure 6.1], at least as regards its principal features. It is true that there are minor irregularities, which may represent superposed fluctuations, probably in part of the nature of errors of observation; for the sunspot numbers can only be taken as more or less approximate 'index numbers' to sunspot activity. But in the main the graph is wonderfully smooth, and its departures from true periodicity, which have troubled all previous analysts of the data, are precisely those found in [Figure 6.2] – great variation in amplitude and continual changes of phase. (*ibid.*, page 273)

It was to reduce the impact of superposed fluctuations that the graduated series was constructed and this aspect is discussed further below.

As both the sunspot and graduated numbers have positive means, being necessarily non-negative, a constant was included in the regression (6.8). Estimation over both the extended sample period 1700 to 2007 and the period available to Yule gave the following results (the first two observations in each period are lost due to the construction of x_{t-1} and x_{t-2}):

Sunspot Numbers, 1749–1924

s.d. of whole series = 34.75

$$x_t = 1.61979x_{t-1} - x_{t-2} + 17.06$$

$$\cos\vartheta = 0.80989; \quad \vartheta = 35.91°; \quad \text{period} = 10.02$$

s.d. of disturbances = 17.08

Sunspot Numbers, 1700–2007

s.d. of whole series = 40.43

$$x_t = 1.64572x_{t-1} - x_{t-2} + 17.74$$

$$\cos\vartheta = 0.82286; \quad \vartheta = 34.63°; \quad \text{period} = 10.40$$

s.d. of disturbances = 18.13

The regression for the shorter sample period available to Yule recovers his estimates quite closely. The results for the extended sample show that the estimate of the period has increased by almost 0.4 of a year and the variability of the series has also increased somewhat. The disturbances estimated from the extended

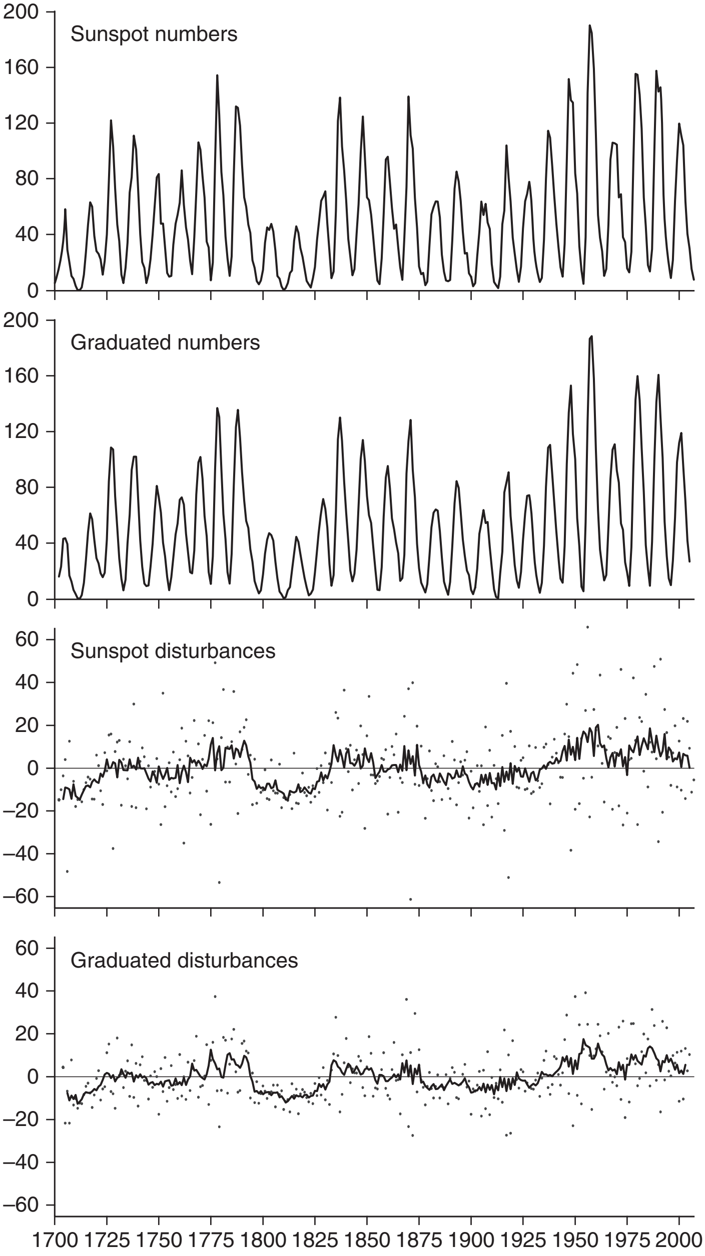


Figure 6.3 Graphs of the sunspots and graduated numbers, and of the disturbances given by equation (6.7): the lines on the disturbance graphs show quinquennial averages

sample regression are plotted in the third panel of Figure 6.3 with a quinquennial moving average superimposed. Focusing on the sample from 1749 to 1924, Yule described their behaviour thus.

> It will be seen that the disturbances are very variable, running up to over ±50 points. But the course of affairs is rather curious. From 1751 to 1792, or thereabouts, the disturbances are mainly positive and highly erratic; from 1793 to 1834 or thereabouts, when the sunspot curve was depressed, they are mainly negative and very much less scattered; from 1835 to 1875, or thereabouts, they are again mainly positive and highly erratic; and finally, from 1876 to 1915, or thereabouts, once more mainly negative and much less erratic. It looks as if the 'disturbance function' had itself a period of somewhere about 80 to 84 years, alternate intervals of 40 to 42 years being highly disturbed and relatively quiet. (*ibid.*, pages 275–6)

The additional observations now available serve only to confirm Yule's impressions. The disturbances in the first half of the eighteenth century were predominantly negative and not particularly erratic. The final interval isolated by Yule probably continued until the late 1930s, whereupon there was again an extended sequence of generally positive and highly erratic disturbances.

One problem that exercised Yule was that the estimated period for the shorter sample, here 10.02 years, was too low compared to the usual estimates of somewhat over 11 years. In Yule's opinion 'this was probably due to the presence of superposed fluctuations: as already noted, the graph of sunspot numbers suggests the presence of minor irregularities due to this cause' (*ibid.*, page 273), leading him to the view that

> if such fluctuations are present, our two variables $x_t + x_{t-2}$ and x_{t-1} are, as it were, affected by errors of observation, which would have the effect of reducing the correlation and also the regression [coefficient]. Reducing the regression [coefficient] means reducing the value of $\cos\vartheta$ – that is, increasing ϑ or reducing the apparent period. (*ibid.*, page 276)

Yule therefore re-estimated the regressions using the graduated data. Doing that here obtains the following results.

Graduated Sunspot Numbers, 1753–1920

s.d. of whole series = 34.10

$$x'_t = 1.68431x'_{t-1} - x'_{t-2} + 14.23$$

$$\cos\vartheta = 0.84216; \quad \vartheta = 32.63^\circ; \quad \text{period} = 11.03$$

s.d. of disturbances = 11.50

Graduated Sunspot Numbers, 1704–2003

$$\text{s.d. of whole series} = 39.46$$

$$x'_t = 1.69408x'_{t-1} - x'_{t-2} + 15.41$$

$$\cos\vartheta = 0.84704; \quad \vartheta = 32.11^\circ; \quad \text{period} = 11.21$$

$$\text{s.d. of disturbances} = 12.19$$

From the first regression, Yule felt able to conclude that '(t)he estimate of the period is now much closer to that usually given, and I think it may be concluded that the reason assigned for the low value obtained from the ungraduated numbers is correct' (*ibid.*, page 276). Interestingly, the period obtained from the extended sample, 11.21, turns out to be identical to the period obtained using Fourier analysis for the period 1750–1914 by Larmor and Yamaga (1917). The calculated disturbances are shown as the bottom panel of Figure 6.3: 'the scatter is greatly reduced (s.d. of disturbances [12.19] against [18.13]), but the general course of affairs is very similar to that shown from the graph for the ungraduated numbers' (*ibid.*, page 276).

Figure 6.4 shows a scatterplot of $x_t + x_{t-2}$ on x_{t-1} and its graduated counterpart and these provide scant indication of any nonlinearity in the relationships. The proportion of the variance of x_t that has been accounted for by x_{t-1} and x_{t-2} is calculated to be 76 per cent for Yule's sample and 80 per cent for the extended sample, with the graduated counterpart values being 89 per cent and 90 per cent.

6.5 Yule then extended the model (6.1) to contain two harmonics:

$$x_t = \rho_1 \sin 2\pi \frac{t}{n_1} + \rho_2 \sin 2\pi \frac{t}{n_2}$$

Writing $x_t = a + (x_t - a)$, where a is that part of x_t due to the first harmonic and $(x_t - a)$ is that part due to the second, (6.2) extends naturally to

$$\Delta^2 x_t = x_t - 2x_{t+1} + x_{t+2} = -\theta_1 a - \theta_2(x_{t+1} - a)$$

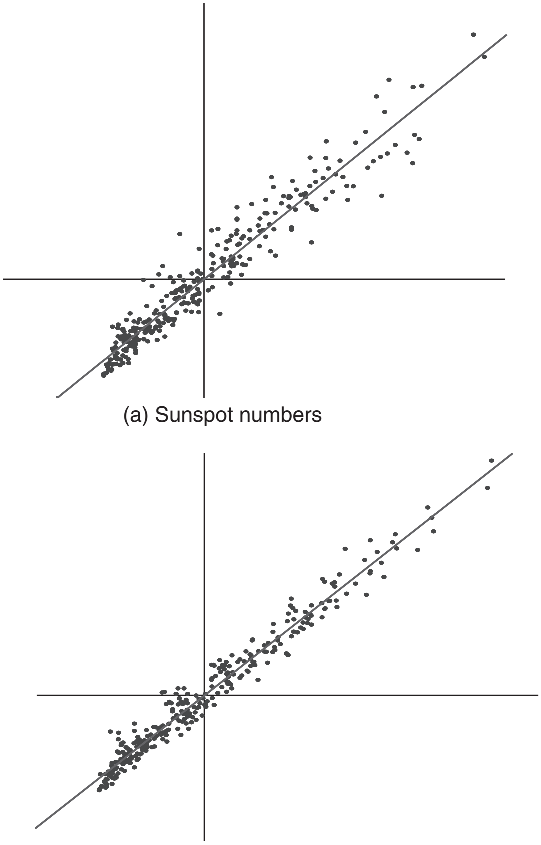
$$\Delta^4 x_t = x_t - 4x_{t+1} + 6x_{t+2} - 4x_{t+3} - x_{t+4} = \theta_1^2 a + \theta_2^2(x_{t+2} - a)$$

where

$$\theta_i = 4\sin^2 \frac{\pi}{n_i} = 2(1 - \cos\vartheta_i) \quad i = 1, 2$$

By eliminating a, this pair of equations can be reduced to (cf. (6.3))

$$x_{t+4} = (4 - \theta_1 - \theta_2)(x_{t+3} + x_{t+1}) - (6 - 2\theta_1 - 2\theta_2 + \theta_1\theta_2)x_{t+2} - x_t$$



(a) Sunspot numbers

(b) Graduated numbers

Figure 6.4 Scatterplot of $x_t + x_{t-2}$ (horizontal) on x_{t-1} (vertical)

and, if a disturbance is again appended, we can write (cf. (6.4))

$$x_t = k_1(x_{t-1} + x_{t-3}) - k_2 x_{t-2} - x_{t-4} + \varepsilon_t \tag{6.9}$$

While questioning the theoretical legitimacy of appending such an error, Yule thought that it could be justified in practice.

> If ... we nevertheless assume a relation of the form [6.9] and proceed to determine k_1 and k_2 by the method of least squares, regarding $x_t + x_{t-4}$, $x_{t-1} + x_{t-3}$ and x_{t-2} as our three variables, and forming the regression equation for the first on the last two, can this give us any useful information? I think it can. The results may afford a certain criterion as between the

> respective conceptions of the curve being affected by superposed fluctuations or by disturbances. If there are no *disturbances* in the sense in which the term here is used, the application of the suggested method is perfectly legitimate, and should bring out any secondary period that exists. To put the matter in a rather different way: *disturbances* occurring in every interval imply an element of unpredictability very rapidly increasing with the time. *Superposed fluctuations* imply an element of unpredictability which is no greater for several years than for one year. If, then, there is a secondary period in the data, and we might well expect a period of relatively small amplitude – if only a sub-multiple of the fundamental period – equation [6.9] should certainly bring out this period, *provided that we have only to do with superposed fluctuations and not disturbances.* (*ibid.*, page 279: italics in original, notation altered for consistency)

Estimates of the regression (6.9) for the various series and samples were obtained as follows:

Sunspot Numbers, 1749–1924

$$x_t = 1.15975(x_{t-1} + x_{t-3}) - 1.016367x_{t-2} - x_{t-4} + 31.21$$

$$\theta_1 = 2.56899 \quad \cos\vartheta_1 = -0.284495$$

$$\vartheta_1 = 106.53^\circ \text{ or } 253.47^\circ \quad \text{period} = 1.42 \text{ or } 3.38 \text{ years}$$

$$\theta_2 = 0.27126 \quad \cos\vartheta_1 = 0.86437$$

$$\vartheta_1 = 30.19^\circ \quad \text{period} = 11.92 \text{ years}$$

$$\text{s.d. of disturbances} = 21.97 \text{ years}$$

Graduated Sunspot Numbers, 1753–1920

$$x'_t = 1.67128(x'_{t-1} + x'_{t-3}) - 1.86233x'_{t-2} - x'_{t-4} + 23.48$$

$$\theta_1 = 2.07867 \quad \cos\vartheta_1 = -0.03933$$

$$\vartheta_1 = 92.25^\circ \text{ or } 267.75^\circ \quad \text{period} = 1.34 \text{ or } 3.90 \text{ years}$$

$$\theta_2 = 0.25005 \quad \cos\vartheta_1 = 0.87498$$

$$\vartheta_1 = 28.96^\circ \quad \text{period} = 12.43 \text{ years}$$

$$\text{s.d. of disturbances} = 17.47 \text{ years}$$

Sunspot Numbers, 1700–2007

$$x_t = 1.16854(x_{t-1} + x_{t-3}) - 1.01714x_{t-2} - x_{t-4} + 34.14$$

$$\theta_1 = 2.56648 \quad \cos\vartheta_1 = -0.28324$$

$$\vartheta_1 = 106.45^\circ \text{ or } 253.55^\circ \quad \text{period} = 1.42 \text{ or } 3.38 \text{ years}$$

$$\theta_2 = 0.26498 \quad \cos\vartheta_1 = 0.86757$$

$$\vartheta_1 = 28.93^\circ \quad \text{period} = 12.07 \text{ years}$$

$$\text{s.d. of disturbances} = 24.09 \text{ years}$$

Graduated Sunspot Numbers, 1704–2003

$$x'_t = 1.69927(x'_{t-1} + x'_{t-3}) - 1.91419x'_{t-2} - x'_{t-4} + 25.96$$

$$\theta_1 = 2.04908 \quad \cos\vartheta_1 = -0.02454$$

$$\vartheta_1 = 91.41^\circ \text{ or } 268.59^\circ \quad \text{period} = 1.34 \text{ or } 3.94 \text{ years}$$

$$\theta_2 = 0.25165 \quad \cos\vartheta_1 = 0.87418$$

$$\vartheta_1 = 29.05^\circ \quad \text{period} = 12.39 \text{ years}$$

$$\text{s.d. of disturbances} = 19.12 \text{ years}$$

Since the values of the θ's give $\cos\vartheta$ and not ϑ itself, the value of ϑ is not strictly determinate; the longer period is naturally taken as approximate to the fundamental, but the choice of the shorter period is quite uncertain. So far as the results go then, they at first sight suggest the existence of two periods, one year or more longer than the value which anyone, on a mere inspection of the graph, would be inclined to take for the fundamental, and the other much shorter. On the face of it the result looks odd, and the last figures given for the ungraduated and graduated numbers respectively show that it is really of no meaning. *The standard deviations found for the disturbances are* ... *larger than when we assumed the existence of a single period only.* ... So far from having improved matters by the assumption of a second period, we have made them very appreciably worse: we get a worse and not a better estimate of x_t when x_{t-3} and x_{t-4} are brought into account than when we confine ourselves to x_{t-1} and x_{t-2} alone. To put it moderately, there is no evidence that any secondary period exists. ... The result also bears out the assumption that it is disturbances rather than superposed fluctuations which are the main cause of the irregularity, the element of unpredictability, in the data. (*ibid.*, page 280)

Yule explained this result, which might be taken as paradoxical, in a way that is now familiar to econometricians but which demonstrated his mastery of contemporary regression analysis:

> it is simply due to the fact that we have insisted on the regression equation being of a particular form, the coefficients of x_{t-1} and x_{t-3} being identical, and the coefficient of x_{t-4} unity. The result tells us merely that, if we insist on this, such and such values of the coefficients are the best, but even so they cannot give as good a result as the equation of form [6.8] with only two terms on the right. (*ibid.*, page 280)

6.6 As a second approach, Yule considered the 'ordinary regression equation'

$$x_t = b_1 x_{t-1} - b_2 x_{t-2} \tag{6.10}$$

For x_t to have a harmonic component, the roots of the equation

$$z^2 - b_1 z + b_2 = 0$$

must be imaginary. If these roots are $\alpha \pm i\beta$ and we let

$$\alpha^2 + \beta^2 = b_2 = e^{2\lambda} \quad \text{and} \quad \tan\vartheta = \beta/\alpha$$

then the general solution of the difference equation (6.10) is of the form

$$x_t = e^{\lambda t}(A\cos\vartheta t + B\sin\vartheta t) \tag{6.11}$$

For a real physical phenomenon, λ would be expected to be either negative ($b_2 < 1$), so that the solution (6.11) would be a damped harmonic vibration, or zero ($b_2 = 1$), in which case the solution would be simple harmonic vibration.

The regression (6.10), with a disturbance term ε_t implicitly appended, was first fitted to the simulated series of Figure 6.2, producing the following results.

Complete sample of 300 terms

$$x_t = 1.6220x_{t-1} - 0.9983x_{t-2}$$

Roots: $0.8110 \pm 0.5836i$

$\tan\vartheta = 0.71959 \quad \vartheta = 35.74^\circ \quad \text{Period} = 10.07 \quad \lambda = -0.0009$

First 150 terms

$$x_t = 1.6253x_{t-1} - 0.9955x_{t-2}$$

Roots: $0.8126 \pm 0.5789i$

$\tan\vartheta = 0.712334 \quad \vartheta = 35.46^\circ \quad \text{Period} = 10.15 \quad \lambda = -0.0023$

Second 150 terms

$$x_t = 1.601x_{t-1} - 0.9999x_{t-2}$$

Roots: $0.8101 \pm 0.5863i$

$\tan\vartheta = 0.723753 \quad \vartheta = 35.89^\circ \quad \text{Period} = 10.03 \quad \lambda = -0.0001$

The periods are identical to those obtained previously, with the value of λ being very close to its true value of zero, leading Yule (*ibid.*, page 281) to conclude that 'the agreement seems quite satisfactory'!

For the various sunspot series and sample periods, the following results are obtained

Sunspot Numbers, 1749–1924

$$x_t = 1.33597x_{t-1} - 0.64986x_{t-2} + 13.94$$

Roots: $0.66798 \pm 0.45128i$

$\tan\vartheta = 0.67559 \quad \vartheta = 34.04^\circ \quad \text{Period} = 10.58 \quad \lambda = -0.21550$

s.d. of disturbances = 15.56

Graduated Sunspot Numbers, 1753–1920

$$x'_t = 1.51975x'_{t-1} - 0.80457x'_{t-2} + 12.84$$

Roots: $0.75987 \pm 0.47661i$

$\tan\vartheta = 0.62723 \quad \vartheta = 32.10^\circ \quad \text{Period} = 11.22 \quad \lambda = -0.10872$

s.d. of disturbances = 10.96

Sunspot Numbers, 1700–2007

$$x_t = 1.39078x_{t-1} - 0.69026x_{t-2} + 15.00$$

Roots: $0.69539 \pm 0.45466i$

$\tan\vartheta = 0.65382 \quad \vartheta = 33.18^\circ \quad \text{Period} = 10.85 \quad \lambda = -0.18533$

s.d. of disturbances = 16.69

Graduated Sunspot Numbers, 1704–2003

$$x'_t = 1.55218x'_{t-1} - 0.83264x'_{t-2} + 14.14$$

Roots: $0.77609 \pm 0.47992i$

$\tan\vartheta = 0.61838 \quad \vartheta = 31.73^\circ \quad \text{Period} = 11.35 \quad \lambda = -0.09158$

s.d. of disturbances = 11.69

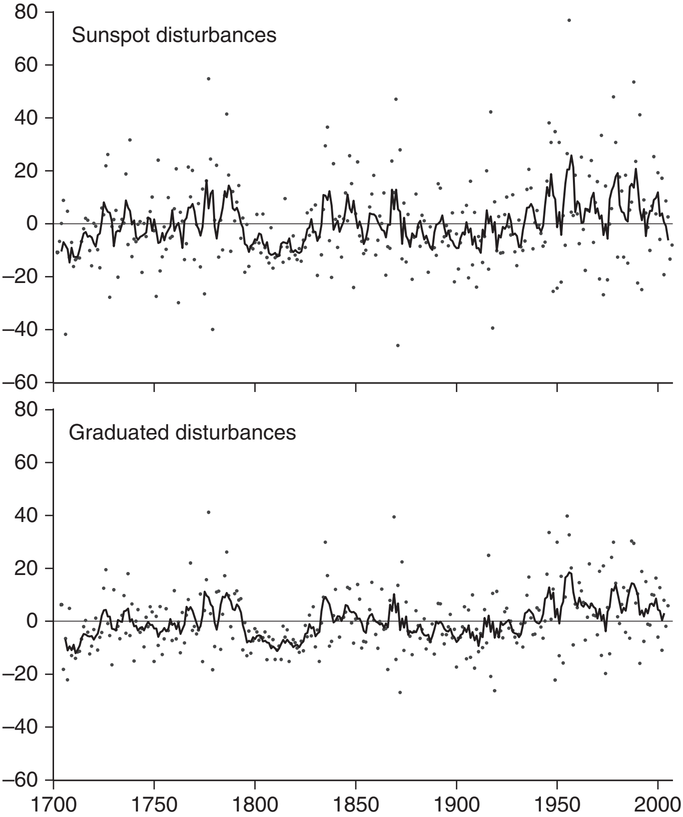


Figure 6.5 Graphs of the disturbances given by equation (6.7): the lines on the graphs show quinquennial averages

For Yule's sample, the period for the ungraduated sunspot numbers is increased when compared with the harmonic formula (10.58 to 10.02) although it is still too low, but that obtained from the graduated numbers (11.22 against 11.03) is now almost the same as that suggested by Larmor and Yamaga (1917). For the extended samples, both periods are increased to 10.85 and 11.35, respectively (against 10.40 and 11.21 from the harmonic formula).

6.7 The two disturbance series for the extended sample period are shown in Figure 6.5. Yule analysed these in the context of the alternating quiet and disturbed periods of approximately 42 years alluded to in §**6.4**. Table 6.2 extends Yule's periods both backwards and forwards in time to cover the extended sample now available. As discussed in §**6.4**, Yule found that there were alternating periods of 42 years in which the disturbances gave positive and negative mean values accompanied by high and low standard deviations respectively.

Table 6.2 Means and standard deviations of disturbances in successive periods of 42 years. (Y) corresponds to periods investigated by Yule (1927, Table II)

Period	Sunspot disturbances		Graduated disturbances	
	Mean	St. Dev.	Mean	St. Dev.
1709–1750	−2.32	12.07	−2.77	8.69
1751–1792 (Y)	2.48	18.91	2.33	11.81
1793–1834 (Y)	−7.23	7.41	−6.46	5.87
1835–1876 (Y)	2.55	17.99	2.11	12.62
1877–1918 (Y)	−4.35	13.95	−3.78	8.66
1919–1960	4.19	19.22	2.94	14.23
1961–2002	6.77	19.66	6.19	12.69

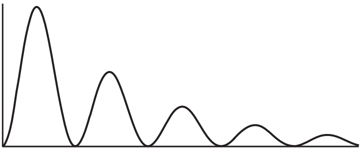


Figure 6.6 Graph of the square of a damped harmonic vibration, (6.12)

In the extended period covered in Table 6.2, this alternating pattern continues to be found from the early 1700s up until 1960, but the final period 'bucks the trend', having a positive mean accompanied by a high standard deviation, rather than a negative mean and a low standard deviation.[1,2]

6.8 Yule concluded from his examination of these disturbances that a damped vibration did explain the evolution of the sunspot numbers. However, rather than being a simple damped vibration, Yule argued that the process generating the sunspot numbers was more akin to a 'train' of squared damped harmonic vibrations superposed upon each other. The square of a damped harmonic vibration,

$$x_t = Ae^{-at}(1 - \cos \vartheta t) \tag{6.12}$$

is shown in Figure 6.6. Figure 6.7 shows a train of such functions, each with different amplitude A and each starting when the one before reaches its first minimum. This looks much more like the graph of the sunspot numbers. However, if (6.12) is regarded as the solution of a difference equation, then it is seen to imply that there must be a real root, thus giving rise to at least a third order difference equation. The difference equation (6.10) would then need extending to include x_{t-3}, in which case it becomes necessary to examine the correlation between x_t and x_{t-3} and, possibly, between x_t and more distant terms.

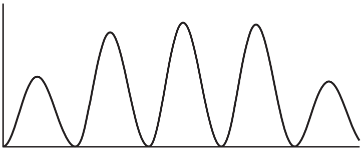


Figure 6.7 Graph of a series of superposed functions of the form of Figure 6.6, each one starting when the one before reaches its first minimum

Yule thus considered the serial correlations of the sunspot numbers up to lag five, i.e., he computed $r_x(1), \ldots, r_x(5)$ using the notation of §**5.3**. He then computed corresponding partial correlations (cf. §**2.3**), which we may denote as $r_x(1), r_x(2 \cdot 1), r_x(3 \cdot 12), r_x(4 \cdot 123)$ and $r_x(5 \cdot 1234)$, where $r_x(k \cdot 12 \ldots (k-1))$ denotes the correlation between x_t and x_{t-k} with the intervening lagged values, $x_{t-1}, \ldots, x_{t-k+1}$, held constant. Although no details are presented, presumably Yule computed the partial correlations by following the recursive scheme outlined in Yule (1907, §§**14–16**) with the assumption 'that the correlation between x_{t-1} and x_{t-2} is the same as that between x_t and x_{t-1}, and so forth – an assumption which implies corresponding equalities between partial correlations' (Yule, 1927, pages 286–7). The serial and partial correlations for the extended sample are shown in Table 6.3. The third column, labeled $1 - r^2$, uses the partial correlations in its calculation. The fourth column is computed so as to be able to use the result, taken from Yule (1907, §17), that

$$1 - R^2_{1 \cdots k} = (1 - r^2(1))(1 - r^2(2 \cdot 1)) \cdots (1 - r^2(k \cdot 1 \ldots (k-1)))$$

where $R^2_{1 \cdots k}$ is the coefficient of multiple correlation (cf. §**2.3**). $1 - R^2_{1 \cdots k}$ then measures the proportionate reduction in the variance of x_t induced by taking into account k lags of x_t.

For both the sunspot numbers and their graduations, the original conclusions of Yule continue to hold.

> It will be seen that after the first two terms all the [partial] correlations are so small that the continued product of $(1 - r^2)$ hardly falls at all. ... It seems quite clear that ... it would be an entire waste of time to take into account any terms more distant from x_t then x_{t-2} for purposes of estimation. As regards the idea suggested that the difference equation should be of the form required for such a function as [6.12], it may be noted that $r_x(3 \cdot 12)$ is of the wrong sign: a positive correlation would be required. The correlations give no evidence at all of any periodicity other than the fundamental, nor of any other exponential function. They strongly emphasise the increase of the element of predictability with the time. (Yule, 1927, page 288)

Table 6.3 Serial correlations of the sunspot numbers and the deduced partial correlations for the extended sample period 1700–2007. In the serial correlations, 1 denotes the correlation between x_t and x_{t-1}, i.e., $r(1)$, and so on. In the partial correlations, 2.1 denotes the correlation between x_t and x_{t-2} with x_{t-1} constant, i.e., $r(2 \cdot 1)$, and so on

Serial correlations		Partial correlations		$1 - r^2$	Continued product of $1 - r^2$
		Sunspot Numbers			
1	0.820	1	0.820	0.328	0.328
2	0.450	2.1	−0.677	0.542	0.178
3	0.038	3.12	−0.144	0.979	0.174
4	−0.278	4.123	0.044	0.998	0.174
5	−0.426	5.1234	0.015	0.999	0.174
		Graduated Sunspot Numbers			
1	0.845	1	0.845	0.286	0.286
2	0.480	2.1	−0.821	0.326	0.093
3	0.050	3.12	0.058	0.997	0.093
4	−0.280	4.123	0.306	0.906	0.084
5	−0.433	5.1234	0.179	0.968	0.082

6.9 After conducting some experiments that use periodogram analysis on 'disturbed' harmonic functions, which need not concern us here, Yule concluded his paper with the following observation:

> many series which have been or might be subjected to periodogram analysis may be subject to 'disturbance' in the sense in which the term is here used, and that this may possibly be the source of some rather odd results which have been reached. Disturbance will always arise if the value of the variable is affected by external circumstances and the oscillatory variation with time is wholly or partly self-determined, owing to the value of the variable at any time being a function of the immediately preceding values. Disturbance, as it seems to me, can only be excluded if either (1) the variable is quite unaffected by external circumstances, or (2) we are dealing with a forced vibration and the external circumstances producing this forced vibration are themselves undisturbed. (*ibid.*, pages 295–7)

Walker's extension of Yule's model

6.10 While the correct modelling of the impact of disturbances on harmonic functions had certainly been demonstrated by Yule, of more fundamental importance to time series analysis was the introduction of the more general models in which x_t was an unrestricted function of lagged values of itself, as in the 'ordinary regression' equation (6.10). The analysis of such models was taken further by Walker (1931), who considered the extension of (6.10) to[3]

$$x_t = g_1 x_{t-1} + g_2 x_{t-2} + \cdots + g_p x_{t-p} \tag{6.13}$$

Multiplying this equation by x_{t-p-1} and summing over all values of x_t from $t = p + 2$ to T obtains

$$\sum_{t=p+2}^{T} (x_t x_{t-p-1} - g_1 x_{t-1} x_{t-p-1} - \cdots - g_p x_{t-p} x_{t-p-1}) = 0 \tag{6.14}$$

If it is assumed that T is large, then, with σ_x being the standard deviation of the series, (6.14) becomes

$$\sigma_x^2 (r_{p+1} - g_1 r_p - g_2 r_{p-1} - \cdots - g_p r_1) = 0$$

or

$$r_{p-1} = g_1 r_p + g_2 r_{p-1} + \cdots + g_p r_1 \tag{6.15}$$

Note that we now denote the serial correlation $r_x(k)$ as r_k to maintain consistency of notation. Walker (*ibid.*, page 519; notation altered for consistency) was thus able to remark that 'the relationships between the successive x's and the successive r's are, in the limit when T is very large, identical.'

Denoting the roots of

$$z^p - g_1 z^{p-1} - \cdots - g_p = 0$$

as $h_1, h_2, \ldots, h_p$, the solutions to (6.14) and (6.15) are, respectively,

$$x_t = A_1 h_1^t + A_2 h_2^t + \cdots + A_p h_p^t \tag{6.16}$$

and

$$r_i = B_1 h_1^i + B_2 h_2^i + \cdots + B_p h_p^i \tag{6.17}$$

where the As and Bs are constants.

> Thus the r's must have the same periods as the x's. This is obvious in slightly damped simple oscillations each occupying say q of the intervals between the x's; for then $r_{q+1}, r_{q+2}, \ldots, r_{2q}$ will tend to be the same as $r_1, r_2, \ldots, r_q$; the r's will thus have a period of q intervals and will be damped if the x's are damped. (*ibid.*, page 520)

6.11 Suppose now that

$$x_t = gx_{t-1} + \varepsilon_t$$

where g is a fraction and ε_t is a zero mean disturbance. This can be written as

$$x_t - x_{t-1} = -(1-g)x_{t-1} + \varepsilon_t$$

If g is interpreted as the degree of 'persistence', then this form is seen to be equivalent to the 'damping' of a mechanical system, the diminution being proportional to the magnitude of the previous term. Since $r_0 = 1$, it is clear that $B_1 = 1$, $r_1 = h_1 = g$ and $r_i = g^i$. Denoting the standard deviation of the disturbances as σ_ε, it is then straightforward to show that

$$\sigma_\varepsilon^2 = (1 - r_1^2)\sigma_x^2$$

In the general case where there are p lags of x

$$\sigma_\varepsilon^2 = (1 - g_1 r_1 - g_2 r_2 - \cdots - g_p r_p)\sigma_x^2$$

6.12 Walker then considered the relationship between the serial correlations and the trigonometric coefficients of a (zero mean) Fourier series of length $T = 2J + 1$ (cf. §**3.2**)

$$x_t = \sum_{j=1}^{J} A_j \cos j\omega t + \sum_{j=1}^{J} B_j \sin j\omega t, \quad t = 1, \ldots, T$$

Walker showed the relationship to be

$$r_k = \sum_{j=1}^{J} f_j^2 \cos jk\omega$$

where $f_j^2 = (A_j^2 + B_j^2)/2\sigma_x^2$ is the square of the *amplitude-ratio* introduced in Walker (1925). Calling the plot of the r_k the 'correlation periodogram', Walker (1931, pages 524–5) observed that

> if the series forms an accurate cosine curve with a period of T/q or k terms, ... the property of amplitude ratios tells us that $f_q = 1$, all the other f's vanishing; and as the series repeats itself completely after k terms we shall have $r_k = 1$.
>
> Thus any period of q terms with an amplitude ratio f will produce as graph for r_k a cosine curve with maxima of f_k at $k = q, 2q, 3q, \ldots$, and equal and opposite minima half-way between.

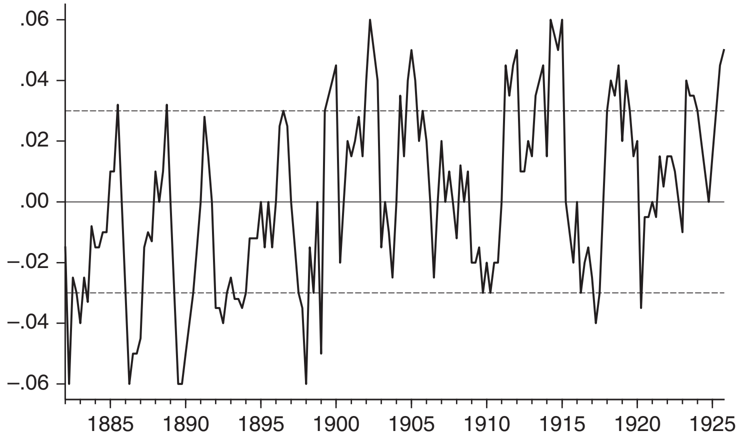


Figure 6.8 Port Darwin pressure, 1882–1925 (quarterly)

> Accordingly if there are only one or two periods and they are well-marked, inspection of the correlation periodogram will reveal them; but if there are three or four periods or they are ill marked, Fourier analysis of the r_k curve will be necessary.

Walker applied these ideas to the air pressure at Port Darwin, which 'displays surges of varying amplitude and period with irregularities superposed, suggesting that pressure in this region has a natural period of its own, based presumably on the physical relationships of world-weather, but that oscillations are modified by external disturbances' (*ibid.*, page 525). 177 quarterly pressure values from 1882 to 1926, shown in Figure 6.8, are analysed here and the periodogram of the series, covering periods from 9 to $1\frac{1}{2}$ years, is shown in Figure 6.9, panel (a). The periods having the largest amplitude ratios are 22 (0.24), $13\frac{1}{2}$ (0.29), $11\frac{2}{3}$ (0.29) and 11 (0.24) quarters (f values shown in parentheses). Walker noted that the probable value of a single f was 0.09 and the probable value of the greatest of 26 of these (the number calculated here) would be 0.20 if the f values were independent. However, from the correlation periodogram shown in Figure 6.10, the correlation between successive quarterly pressures is of the order of 0.76, so that the series is far from being independent, leading Walker to the view that we should

> naturally interpret the pressure variations in one of two ways. Either (*a*) the pressure is like a mechanical system, with persistence but without natural periods and acted on by a series of disturbances; in this case it is the periodicity of the disturbances that must be examined. Or (*b*) the pressure behaves

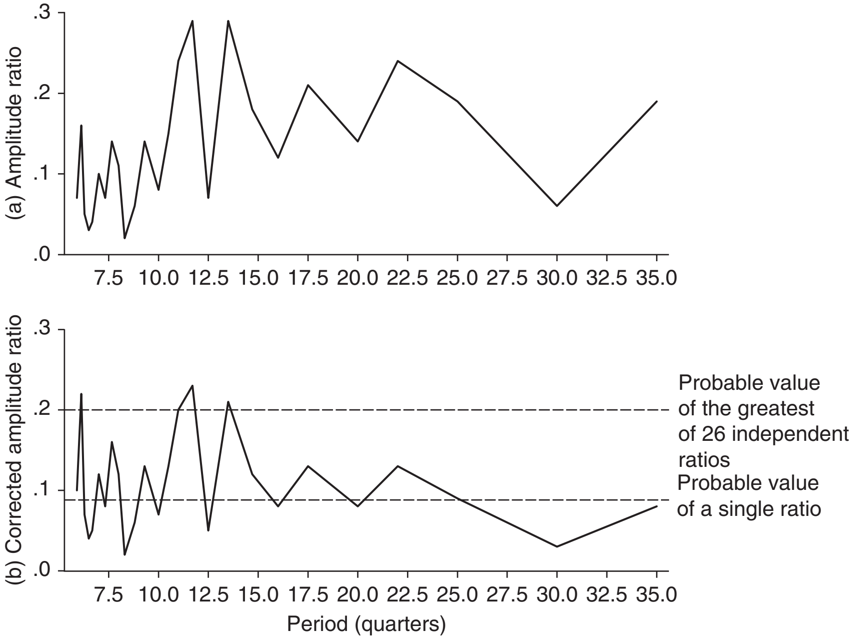


Figure 6.9 Periodograms of pressure at Port Darwin: (a) periodogram of observed series; (b) periodogram when persistent disturbances are allowed for

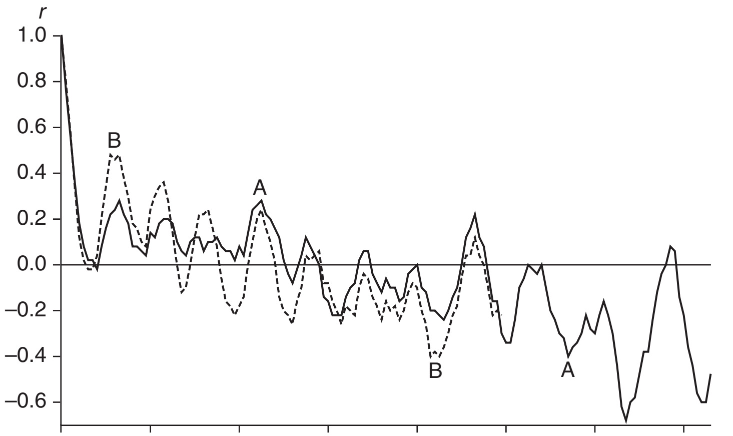


Figure 6.10 Serial correlation coefficients of Port Darwin pressure. A: coefficients calculated using all observations; B: coefficients calculated using 77 pairs of correlates

> like a mechanical system with persistence and natural periods, and then these periods interest us. (*ibid.*, page 527)

Walker showed that if there were two 'physical systems', one in which the successive values were independent and another in which persistence produced a correlation r_1 between successive terms, and if the same disturbances were imposed on the systems, then the ratio of the squares of the amplitude ratios f'_q and f_q was given by

$$\frac{f_q^2}{f_q'^2} = \frac{1 - r_1^2}{1 - 2r_1 \cos q\omega + r_1^2}$$

The amplitude ratios f' obtained using this relationship are shown in Figure 6.9, panel (b). Four ratios reach the limit of 0.20 that is expected to be produced by mere chance and these suggest natural periods of $11\frac{1}{3}$ and $13\frac{1}{2}$ quarters (2.8 and 3.4 years), with a possible third period of 6.1 quarters (1.5 years).

6.13 Walker was far from convinced that these periods were a physical reality and went on to consider the second explanation, that Port Darwin pressure has natural periods of its own, maintained by non-periodic disturbances from outside. After some experimentation with the first 40 serial correlations, he decided to compute serial correlation coefficients r_k from $k = 1$ to $k = 147$: these are shown as line A in Figure 6.10. Walker explained these results thus:

> here when k is 20 we have 157 pairs of correlates, but as k grows the number of correlates diminishes until when $k = 147$ it is only 30. A glance shows outstanding oscillations near $k = 44$, 93 and 137 with three smaller oscillations between 0 and 44, three between 44 and 93, and two between 93 and 137; the general downward slope is maintained. The obvious interpretation is that we have an oscillation with eight periods in 92 quarters, fitting well with the intermediate maxima; superposed on this there is evidence of a rise up to maxima near 46, 92 and 137 with minima in between, or of an oscillation with a period of about 46 quarters. But far from showing damping the oscillations grow with k, and the explanation seems to lie in the contrast between the number $(177 - k)$ of correlates when k is small and the number when k exceeds 100. Thus for the last 40 terms the number of values correlated averages 50, covering $12\frac{1}{2}$ years, and we have the first 12 or 13 years correlated, with different lags, with the last 12 or 13 years; as [Figure 6.8] shows, each has well-marked waves and it is obvious that there will be relative positions in which the waves correspond; so there will be big oscillations in r_k on a scale that would not arise if the number of years correlated were longer. (*ibid.*, pages 530–1)

To confirm this explanation, Walker recomputed the r_k by ‘correlating the first 77 quarters with the groups of 77 quarters which occur $1, 2, 3, \ldots, 100$ quarters later; in this way each correlation coefficient is based on 77 pairs of terms’ (*ibid.*, page 531). The resulting serial correlations are plotted as curve B in Figure 6.10 and contain the main features of curve A along with the expected reduction in amplitude due to damping. Walker suggested ignoring serial correlations for which k exceeds 100 in Figure 6.10 and concluded that the apparent damping in B was largely due to a greater amplitude in the first half of the data than in the second.

6.14 Although Yule and Walker were predominantly interested in analysing the role of disturbances on harmonic functions, with both having real physical systems in mind, their analysis led to the introduction of a number of fundamental results in time series analysis: the concept of partial correlations to go with serial correlations; the result that there will be similar solutions to the difference equations generating the observed data and the serial correlations; the graphical device of a ‘correlation periodogram’; and the introduction of a fundamental model – the ‘ordinary regression model’, to use Yule’s rather prosaic phrase, in which the current value of a series was a linear function of past values of the series. These had all been developed using intuition and the observation of actual physical systems. What was now required were formal theoretical foundations on which to place these concepts, and such foundations were not long in being put into place.

7
The Formal Modelling of Stationary Time Series: Wold and the Russians

Moving averages and autoregressions

7.1 Slutzky's modelling of the 'summation of random causes', introduced in §§**5.11–5.16**, and the 'ordinary regression equations' (6.10) and (6.13) of Yule and Walker were to become the basic models of time series analysis. One of the reasons why they have been such enduring features, apart from their obvious usefulness, may be because of their renaming as *moving averages* and *linear autoregressions*, respectively, by Herman Wold (1938, page 2), as these are terms that convey their structure with great clarity and effectiveness.[1]

Wold regarded these two models as lying within the more general scheme of *linear regression*, having the common feature that a 'random element plays a fundamental, active role', a feature which 'constitutes a distinct contrast to the scheme of *hidden periodicities* – as we shall call the hypothesis of strict periods – and makes the schemes of linear regression *a priori* plausible in several instances where the scheme of hidden periodicities has been criticized' (*ibid.*, page 3; italics in original). Not content with simply introducing new terminology, however, Wold's real desire was to provide a formal theory, based on probability concepts, within which to set these models.

> While the schemes of linear regression thus form a type of hypothesis of the greatest importance, the development of the subject is still little advanced, both as to the theory and the application of the schemes. For instance, earlier definitions concerning the scheme of autoregression are incomplete. One of the chief purposes of the present volume is to give some contributions for completion in these respects. It also aims at bringing the schemes into place in the theory of probability, thereby uniting the rather isolated results hitherto reached.
>
> In the theory of probability, the schemes of linear regression fall under the heading of the discrete stationary random process as defined by [Khinchin (1932, 1933)]. (Wold, 1938, page 29)

Stationary random processes

7.2 As is made apparent in the second paragraph of the quote above, the models intuitively developed by Slutzky, Yule and Walker are all members of the class of discrete stationary random processes. Such a distinction had not been made explicitly before Wold and, in this context, it is worth quoting the opening two paragraphs of his introduction:

> Observational series which describe phenomena changing with time may be roughly classified in two broad categories, viz. *evolutive* and *stationary*. In the former case, different sections of the time series are dissimilar in one or more respects. For instance, the sectional averages may be distinctly different, or some other structural property of the series may present variation. In the analysis of evolutive time series, absolute time plays a fundamental role, e.g. as the independent variable in a trend function, or as a fixed scale in studying the development of a phenomenon from an initial state of rest.
>
> Stationary time series are unchanging in respect to their general structure. The fluctuations up and down in such a series may seem random or show tendencies to regularity – in any case, the character of the series is, on the whole, the same in different sections. Or otherwise expressed, in the analysis of stationary series time is allotted the secondary role of a passive medium. Even without preparation, observational time series are frequently stationary. On the other hand, the deviations from trend form a type of derived time series which is often stationary. (*ibid.*, page 1)

7.3 Wold's theoretical development of stationary random processes used the concepts and techniques introduced by the Russian mathematicians Khinchin (1932, 1933, 1934) and Kolmogorov (1931, 1933).[2] In this chapter we introduce the basic concepts employed by Wold and the fundamental theorems that he proved to obtain the representations that now form the formal basis of modern time series analysis. In order to keep the development manageable, however, no attempt is made to be inclusive and proofs are not provided.

In his formal development, Wold let $\{t\}$ stand for the set of values taken by a real parameter, assumed to represent time, and let one random variable $\xi(t)$ correspond to each time point t in $\{t\}$. The corresponding set of random variables, denoted $\{\xi(t)\}$, will then be a random process if the following conditions are satisfied.

(A) If a subset of $\{t\}$, say $(t) = (t_1, \ldots, t_n)$, is arbitrarily chosen, then the variable $\xi(t_1, \ldots, t_n) = [\xi(t_1), \ldots, \xi(t_n)]$ will be well-defined. Let the distribution function of $\xi(t_1, \ldots, t_n)$ be denoted by

$$F(t_1, \ldots, t_n; u_1, \ldots, u_n) = P[\xi(t_1) \leq u_1, \ldots, \xi(t_n) \leq u_n] \qquad (7.1)$$

Here the right-hand side of (7.1) denotes the joint probability that $\xi(t_1) \leq u_1, \ldots, \xi(t_n) \leq u_n$ and may be termed the probability function: let the sets of distribution and probability functions of $\xi(t_1, \ldots, t_n)$ be denoted by $\{F\}$ and $\{P\}$ respectively.

(B) Given $(t) = (t_1, \ldots, t_n)$ and with $(i_1, \ldots, i_n)$ being an arbitrary permutation of the sequence $(1, 2, \ldots, n)$, the functions $\{F\}$ will satisfy the following relations identically in $u_1, \ldots, u_n$:

$$F(t_{i_1}, \ldots, t_{i_n}; u_{i_1}, \ldots, u_{i_n}) = F(t_1, \ldots, t_n; u_1, \ldots u_n) \tag{7.2}$$

$$F(t_1, \ldots, t_m; u_1, \ldots, u_m) = F(t_1, \ldots, t_n; u_1, \ldots, u_m, +\infty, \ldots, +\infty) \tag{7.3}$$

where $m < n$. Equations (7.2) and (7.3) imply that the probability laws governing $\{\xi(t)\}$ must not contradict themselves and they are thus referred to as the *consistency relations*.

This random (or stochastic) process thus extends the notion of a random variable to an infinite number of dimensions. The sample elements of the process $\{\xi(t)\}$, also called the *realizations* of the process, are functions of t, say $\xi_i(t)$. Keeping t fixed at, say, $t = t_1$, the set of sample values $\xi_i(t_1)$ that constitute the random variable $\xi(t_1)$ are obtained. More generally, if we keep $t_1, \ldots, t_n$ fixed, the realizations will provide the 'universe' of sample elements $[\xi_i(t_1), \ldots, \xi_i(t_n)]$ that constitute the n-dimensional random variable $[\xi(t_1), \ldots, \xi(t_n)]$.

To be able to define stationarity, arbitrary translations within the set $\{t\}$ must be considered. Assume that $\{t\}$ consists either of all real values or is formed by an unbroken sequence of equidistant values, say $\ldots, -1, 0, 1, 2, \ldots$. A random process $\{\xi(t)\}$ as defined by a set $\{F\}$ is then termed *stationary* if, for an arbitrary subset $(t) = (t_1, \ldots, t_n)$, the relation

$$F(t_1 + t, \ldots t_n + t; u_1, \ldots, u_n) = F(t_1, \ldots t_n; u_1, \ldots, u_n)$$

is identically satisfied in $u_1, \ldots, u_n$ and in t. If t is restricted to be a sequence of equidistant values then $\{\xi(t)\}$ is called a *discrete* stationary random process: if t is arbitrary then the process will be *continuous*.

7.4 Expectations derived from the distribution functions $\{F\}$ determining a stationary process $\{\xi(t)\}$ are called the *characteristics* of the process and will be independent of t, as will be the distribution functions $F(t; u)$: the function of u so obtained is then the *principal* distribution function, $F(u)$. Restricting attention to a one-dimensional stationary process, the mean μ and variance σ^2 are then defined as[3]

$$\mu = E(\xi) = \int_{-\infty}^{\infty} u dF(u) \quad \sigma^2 = E[(\xi - \mu)^2] = \int_{-\infty}^{\infty} (u - \mu)^2 dF(u)$$

If the variance is finite, the *automoments* of second order, as defined by

$$\nu_2^{(k)} = E(\xi(t) \cdot \xi(t+k)) = \int_{R_2} uv \cdot d_{u,v}F(t, t+k; u, v) = \nu_2^{(-k)}$$

will also be finite. These characteristics determine the *autocorrelation coefficients* of the stationary process $\{\xi(t)\}$

$$r_k = r_k(\xi) = (\nu_2^{(k)} - \mu^2)/\sigma^2 = r_{-k}$$

If $r_k(\xi) = 0$ for all $k \neq 0$, the process $\{\xi(t)\}$ is termed *non-autocorrelated.*

Now consider a set of random processes $\{\xi^{(1)}(t)\}, \ldots, \{\xi^{(k)}(t)\}$ with an arbitrarily chosen set of time points $(t) = (t_1, \ldots, t_n)$ and k sets of real numbers $(u^{(s)}) = (u_1^{(s)}, \ldots, u_n^{(s)}), s = 1, \ldots, k$. The processes $\{\xi^{(i)}(t)\}$ will be called *independent* if the following relation is satisfied

$$\begin{aligned}&P[\xi^{(1)}(t_1) \leq u_1^{(1)}, \ldots, \xi^{(1)}(t_n) \leq u_n^{(1)}; \ldots; \xi^{(k)}(t_1) \leq u_1^{(k)}, \ldots, \xi^{(k)}(t_n) \leq u_n^{(k)}]\\ &\quad = P[\xi^{(1)}(t_1) \leq u_1^{(1)}, \ldots, \xi^{(1)}(t_n) \leq u_n^{(1)}] \ldots P[\xi^{(k)}(t_1) \leq u_1^{(k)}, \ldots, \xi^{(k)}(t_n) \leq u_n^{(k)}]\end{aligned}$$

If it is assumed that the independent processes $\{\xi^{(i)}(t)\}$ are stationary and have finite variances $\sigma^2(\xi^{(i)})$ then the sum process

$$\{\zeta_k(t)\} = a_1\{\xi^{(1)}(t)\} + \cdots + a_k\{\xi^{(k)}(t)\}$$

is stationary with expectation, variance and autocorrelation coefficients given by

$$E\{\zeta_k\} = a_1 E\{\xi^{(1)}\} + \cdots + a_k E\{\xi^{(k)}\}$$

$$\sigma^2(\zeta_k) = a_1^2\sigma^2(\xi^{(1)}) + \cdots + a_k^2\sigma^2(\xi^{(k)}) \tag{7.4}$$

$$r_p(\zeta_k) = a_1^2 \frac{\sigma^2(\xi^{(1)})}{\sigma^2(\zeta_k)} r_p(\xi^{(1)}) + \cdots + a_k^2 \frac{\sigma^2(\xi^{(k)})}{\sigma^2(\zeta_k)} r_p(\xi^{(k)}) \tag{7.5}$$

The expressions (7.4) and (7.5) depend on the identities

$$r(\xi^{(r)}(t \pm p); \xi^{(s)}(t \pm q)) = 0 \quad p \geq 0, \ \ q \geq 0 \tag{7.6}$$

where r and s are arbitrary. If (7.6) is satisfied then $\{\xi^{(r)}\}$ and $\{\xi^{(s)}\}$ are said to be *uncorrelated*. In fact, (7.4) and (7.5) will hold if $\{\xi^{(r)}\}$ and $\{\xi^{(s)}\}$ are simply uncorrelated processes, stationary or otherwise.

Similarly, the moving average process defined by

$$\zeta(t) = a_0\xi(t) + a_1\xi(t-1) + \cdots + a_h\xi(t-h) \tag{7.7}$$

will also be stationary if $\xi(t)$ is stationary (*ibid.*, page 38).

These operations may also be applied to observed time series, so that, if $\ldots, \overline{\xi}_{t-1}, \overline{\xi}_t, \overline{\xi}_{t+1}, \ldots$ represents such a series, the counterpart to (7.7), for example, is

$$\overline{\zeta}_t = a_0\overline{\xi}_t + a_1\overline{\xi}_{t-1} + \cdots + a_h\overline{\xi}_{t-h}$$

7.5 Wold then utilized the concept of convergence in probability to state his first theorem. A sequence $\xi^{(1)}, \xi^{(2)}, \ldots$ of random variables is said to *converge in probability* to a random variable ξ if, for every $\varepsilon > 0$

$$P[|\xi^n - \xi| > \varepsilon] \to 0$$

as $n \to \infty$. A sequence of random processes $\{\xi^{(1)}(t)\}, \{\xi^{(2)}(t)\}, \ldots$ is then called *convergent in probability to a limit process* $\{\xi(t)\}$ if, for an arbitrary set $(t) = (t_1, \ldots, t_n)$, the sequence

$$\xi^{(1)}(t_1, \ldots, t_n), \xi^{(2)}(t_1, \ldots, t_n), \ldots$$

is convergent in probability to the limit variable $\xi(t_1, \ldots, t_n)$. This allows Wold (*ibid.*, page 40) to state

Theorem 1.
A necessary and sufficient condition that a sequence $\{\xi^{(1)}(t)\}, \{\xi^{(2)}(t)\}, \ldots$ of random processes be convergent in probability is that, for an arbitrary t, the sequence $\xi^{(1)}(t), \xi^{(2)}(t), \ldots$ be convergent in probability. If the sequence is convergent and if every process $\{\xi^{(n)}(t)\}$ is stationary, the limit process will be stationary.

7.6 Suppose $\xi = [\xi^{(1)}, \ldots, \xi^{[n]}]$ represents an n-dimensional random variable with distribution function $F(u_1, \ldots, u_n)$ and there exists a linear function, say

$$L[x - \mu] = a_1(x^{(1)} - \mu_1) + \cdots + a_n(x^{(n)} - \mu_n)$$

such that

$$P[L[\xi - \mu] \neq 0] = P[a_1(\xi^{(1)} - \mu_1) + \cdots + a_n(\xi^{(n)} - \mu_n) \neq 0] = 0$$

The distribution of ξ is then said to be (*linearly*) *singular* and the variables $\xi^{(i)}$ are said to be connected by the relation $L[\xi - \mu] = 0$. The singularity is of *rank h* if there exist only $n - h$ independent relations between the variables $\xi^{(i)}$, say

$$\begin{aligned} a_{1,h+1}(\xi^{(1)} - \mu_1) + \cdots + a_{n,h+1}(\xi^{(n)} - \mu_n) &= 0 \\ a_{1,n}(\xi^{(1)} - \mu_1) + \cdots + a_{n,n}(\xi^{(n)} - \mu_n) &= 0 \end{aligned}$$

Suppose that the singularity is of the form

$$\xi(t) - \mu + a_1(\xi(t-1) - \mu) + \cdots + a_h(\xi(t-h) - \mu) = 0 \tag{7.8}$$

where $h \leq n$. This is known as a *stochastical difference relation of order h*. Wold (*ibid.*, page 45) was then able to prove the following result.

Theorem 2.
Let $\{\xi(t)\}$ be a discrete stationary process with autocorrelation coefficients r_k. If $\{\xi(t)\}$ is linearly singular then it is a process of superposed harmonics. A necessary condition that $\{\xi(t)\}$ be linearly singular, say on account of the relation $L[\xi(t) - m] = 0$ given by [7.8], *is that r_k satisfies the difference equation $L[r_k] = 0$.*

On defining the *principal correlation determinants*

$$\Delta(r,n) = \begin{vmatrix} 1 & r_1 & r_2 & \dots & r_n \\ r_1 & 1 & r_1 & \dots & r_{n-1} \\ r_2 & r_1 & 1 & \dots & r_{n-2} \\ \vdots & \dots & \dots & \dots & \vdots \\ r_n & r_{n-1} & r_{n-2} & \dots & 1 \end{vmatrix} \geq 0 \tag{7.9}$$

Wold (*ibid.*, page 47) could then assert

Theorem 3.
Let $\{\xi(t)\}$ be a discrete stationary process with principal correlation determinants $\Delta(r,n)$. A necessary and sufficient condition that $\{\xi(t)\}$ be singular of rank h is that $\Delta(r,h)$ be the first vanishing determinant in the sequence $\Delta(r,1)$, $\Delta(r,2), \dots$.

Wold showed that a stationary process with finite variance which satisfies (7.8) will also satisfy the difference relation

$$\Delta^{2s}\xi(t-s) + h_1\Delta^{2s-2}\xi(t-s+1) + \cdots + h_s[\xi(t) - \mu] = 0 \tag{7.10}$$

Types of discrete stationary processes

7.7 A *purely random process* is a process such that (7.1) has the form

$$F(t_1, \dots, t_n; u_1, \dots, u_n) = F(u_1) \dots F(u_n)$$

On extending the notation to let $\{\xi(t; F)\}$ denote a purely random process defined by a distribution function $F(u)$, the following theorem holds (*ibid.*, page 48)

Theorem 4.
Let $\{\xi^{(1)}(t; F^{(1)})\}$, $\{\xi^{(2)}(t; F^{(2)})\}, \ldots$ *represent independent, purely random processes such that the infinite convolution* $F^{(1)} \otimes F^{(2)} \otimes \cdots$ *is convergent. Then the sum*

$$\{\xi^{(1)}(t; F^{(1)})\} + \{\xi^{(2)}(t; F^{(2)})\} + \cdots$$

will be convergent, and will constitute a purely random process with this convolution for its principal distribution function.

In Theorem 4, the convolution of two distribution functions $F_1(u)$ and $F_2(u)$ is given by

$$G(u) = F_1(u) \otimes F_2(u) = \int_{-\infty}^{\infty} F_1(u - x) \cdot dF_2(x)$$

7.8 From §**7.4**, a stationary process $\{\xi(t)\}$ will be obtained by taking

$$\xi(t) = b_0\eta(t) + b_1\eta(t-1) + \cdots + b_h\eta(t-h) \tag{7.11}$$

Here $\{\eta(t)\}$ represents a purely random process and $(b) = (b_0, b_1, \ldots, b_h)$ an arbitrary sequence of real numbers. Equation (7.11) defines a *process of moving averages* with $\{\eta(t)\}$ known as the *primary* process. Usually the identifying assumption is made that $b_0 = 1$. The principal distribution functions $F_\xi(u)$ and $F_\eta(u)$ are connected by

$$F_\xi(u) = F_\eta(u/b_0) \otimes F_\eta(u/b_1) \otimes \cdots \otimes F_\eta(u/b_h)$$

and, as long as the variance of $\{\eta(t)\}$, $\sigma^2(\eta)$, is finite,

$$\sigma^2(\xi) = (b_0^2 + b_1^2 + \cdots + b_h^2)\sigma^2(\eta)$$

If it is assumed that $E(\eta) = 0$ and, as $h \to \infty$, the real sequence (b) is such that $\sum_{k=0}^{\infty} b_k^2$ is convergent, then (7.11) extends to

$$b_0\eta(t) + b_1\eta(t-1) + b_2\eta(t-2) + \cdots \tag{7.12}$$

It follows from the independence of $\eta(t)$ that the variance of

$$b_n\eta(t-n) + b_{n+1}\eta(t-n-1) + \cdots + b_{n+p}\eta(t-n-p)$$

is given by

$$(b_n^2 + b_{n+1}^2 + \cdots + b_{n+p}^2)\sigma^2(\eta)$$

and thus tends to zero uniformally in p as $n \to \infty$. Accordingly, (7.12) is convergent and, from Theorem 1, the stationary process $\{\xi(t)\}$ may be defined as

$$\xi(t) = b_0\eta(t) + b_1\eta(t-1) + b_2\eta(t-2) + \cdots$$

Wold stated that this is the general formula for a *process of linear regression.*

7.9 Now let $(a) = (a_1, \ldots, a_h)$ be a set of real numbers such that $a_h \neq 0$ and for which the roots of the *characteristic equation*

$$z^h + a_1 z^{h-1} + \cdots + a_{h-1}z + a_h = 0$$

all have modulus less than 1. Let $(b) = (b_1, b_2, \ldots)$ be a sequence such that the difference equation

$$x(t) + a_1 x(t-1) + \cdots + a_h x(t-h) = 0$$

is satisfied when $x_t \equiv b_t$ and where the initial values $b_1, \ldots, b_h$ are solutions of the following system of linear equations

$$\begin{aligned} &a_1 + b_1 = 0 \\ &a_2 + a_1 b_1 + b_2 = 0 \\ &\vdots \\ &a_{h-1} + a_{h-2}b_1 + \cdots + a_1 b_{h-2} + b_{h-1} = 0 \\ &a_h + a_{h-1}b_1 + \cdots + a_1 b_{h-1} + b_h = 0 \end{aligned} \qquad (7.13)$$

The b_i are seen to be real and uniquely determined and, if $\sum_{k=1}^{\infty} b_k^2$ is convergent, a stationary process will be defined by

$$\xi(t) = \eta(t) + b_1\eta(t-1) + b_2\eta(t-2) + \cdots \qquad (7.14)$$

for purely random $\{\eta(t)\}$ with finite variance. Since $\{\xi(t)\}$ is stationary, so also will be

$$\zeta(t) = \xi(t) + a_1\xi(t-1) + \cdots + a_h\xi(t-h)$$

Wold (*ibid.*, page 54) then showed that $\{\zeta(t)\}$ and $\{\eta(t)\}$ are equivalent, so that

$$\xi(t) + a_1\xi(t-1) + \cdots + a_h\xi(t-h) = \eta(t) \qquad (7.15)$$

which implies that the variables $\xi(t)$, $\xi(t-1), \ldots, \xi(t-h)$ are connected by a 'relation of linear regression', with (7.15) then defining a *process of (linear) autoregression*.

7.10 A stationary and singular process given by

$$\xi(t) - \xi(t-h) = 0$$

is called a *periodic process* and any sample series will be strictly periodic with period h. If $\{\xi^{(1)}(t)\}, \ldots, \{\xi^{(k)}(t)\}$ are independent stationary processes then the sum $\{\xi(t)\} = \{\xi^{(1)}(t)\} + \cdots + \{\xi^{(k)}(t)\}$ will constitute a stationary process. If at least one of the processes $\{\xi^{(i)}(t)\}$ is a periodic process, or a process of superposed harmonics, then $\{\xi(t)\}$ will be called a *process of hidden periodicities*.

7.11 Wold (*ibid.*, pages 60–5) termed a variable $\xi(t, t-1, \ldots)$ connected with a stationary process $\{\xi(t)\}$ *normal* if it had the *characteristic function*

$$f(X_t, X_{t-1}, \ldots) = \exp\left(i\mu \sum_{p=0}^{\infty} X_{t-p} - \frac{\sigma^2}{2} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} r_{|p-q|} X_{t-p} X_{t-q} \right)$$

where $\mu, \sigma > 0$ and the autocorrelation coefficients r_k are real and satisfy (7.9), i.e., $\Delta(r, n) \geq 0$. This characteristic function has the normal distribution for its principal distribution function.

From Theorem 2, a singular normal process satisfying (7.10) will exist if the autocorrelation coefficients of any superposed harmonic are such that

$$\sum_{p=0}^{n} \sum_{q=0}^{n} r_{|p-q|} X_{t-p} X_{t-q} \geq 0$$

for any n and for the real sequence $(X) = (X_t, X_{t-1}, X_{t-2}, \ldots)$.

Wold used this result to show that Slutzky's Law of the Sinusoidal Limit (as discussed in §**5.16** and extended by another Russian mathematician, Romanovsky, 1932, 1933) may be verified by analogy to the properties of singular normal processes in that certain sections of stationary processes which, in the limit, satisfy singularity restrictions, will approximate superposed harmonics with the sections having the same period but varying amplitudes and phases: see §**7.20** for further details.

Autocorrelation coefficients as Fourier constants

7.12 Wold's next theorem (*ibid.*, page 66) related the autocorrelation coefficients r_k to the Fourier coefficients of a non-decreasing function.

Theorem 5.

Let $r_k(k=0,\pm1,\pm2,\dots)$ be an arbitrary sequence of constants. A necessary and sufficient condition that there exists a discrete stationary process with the r_ks for autocorrelation coefficients is that the r_ks are the FOURIER *coefficients of a non-decreasing function, say $W(x)$, such that $W(0)=0$; $W(\pi)=\pi$,*

$$r_k = \frac{1}{\pi}\int_0^\pi \cos kx \cdot dW(x)$$

The 'inversion formula' which allows $W(x)$ to be uniquely determined by the autocorrelation coefficients is

$$W(x) = x + 2\sum_{k=1}^{\infty}\frac{r_k}{k}\sin kx$$

This formula, called *the generating function* of the r_k, has a structure given by the following corollary to Theorem 5.

Corollary

Let $\{\xi(t)\}$ be a stationary process with autocorrelation coefficients r_k such that $\sum_{k=1}^{\infty}|r_k|$ is convergent. Then $W(x)$ will be absolutely continuous and will have a derivative $W'(x)$ that is bounded in modulus and given by

$$W'(x) = \sum_{k=-\infty}^{\infty} r_k \cos kx, \quad 0 \le x < \pi$$

Linear autoregression analysis of the discrete stationary process

7.13 Wold (*ibid.*, pages 75–80) showed that the variable $\xi(t)$ connected with the stationary process $\{\xi(t)\}$ may be approximated by $\xi(t-1),\dots,\xi(t-n)$, with the approximating error, termed the *residual*, being given by

$$\eta(t;n) = \xi(t) - \mu - a(1,n)\cdot(\xi(t-1)-\mu) - \cdots - a(n,n)\cdot(\xi(t-n)-\mu)$$

Here $\{\xi(t)\}$ has mean μ and principal correlation determinants $\Delta(r,n)$ given by (7.9). It is also assumed to have finite variance $\sigma^2(\xi)$ and, if $\Delta(r,n-1)\neq 0$, this variance will satisfy the inequalities

$$\sigma^2(\xi) \ge \sigma^2(\eta(n)) = \sigma^2(\xi)\frac{\Delta(r,n)}{\Delta(r,n-1)} \ge 0$$

which implies that

$$1 \ge \frac{\Delta(r,1)}{1} \ge \frac{\Delta(r,2)}{\Delta(r,1)} \ge \cdots \ge \frac{\Delta(r,n)}{\Delta(r,n-1)} \ge 0$$

From the analysis of §**7.6**, this implies that either $\Delta(r,n) > 0$ for all n, or $\Delta(r,n) > 0$ for $n < h$ and $\Delta(r,n) = 0$ for $n \geq h$, where h is the rank of linear singularity. It must therefore be the case that any stationary process belongs to one, and only one, of the following classes:

(I) The process is non-singular, and there exists a positive constant $\chi^2 \leq 1$ such that

$$\frac{\sigma^2(\eta(n))}{\sigma^2(\xi)} = \frac{\Delta(r,n)}{\Delta(r,n-1)} \to \chi^2 \leq 1 \quad \text{as } n \to \infty$$

(II) The process is singular, say of rank h.

(III) The process presents no singularity of finite rank, but

$$\frac{\sigma^2(\eta(n))}{\sigma^2(\xi)} = \frac{\Delta(r,n)}{\Delta(r,n-1)} \to 0 \quad \text{as } n \to \infty$$

in which case the process is termed *singular of infinite rank*.

This led Wold (*ibid.*, pages 81–4) to prove for case (**I**)

Theorem 6.
A residual process $\{\eta(t)\}$ obtained from a non-singular stationary process $\{\xi(t)\}$ is stationary and non-autocorrelated. The variable $\eta(t)$ is non-correlated with $\xi(t-1), \xi(t-2), \ldots$, while

$$r(\xi(t), \eta(t)) = \frac{\sigma(\eta)}{\sigma(\xi)}$$

This theorem also holds for cases (**II**) and (**III**) except that, as the residual variables $\eta(t)$ are then vanishing, their correlation properties will be indeterminate.

A canonical form of the discrete stationary process

7.14 The analysis of §**7.13** enabled Wold (*ibid.*, pages 84–9) to prove the most fundamental theorem in time series analysis, which has since become known as *Wold's Decomposition.*[4]

Theorem 7.
Denoting by $\{\xi(t)\}$ an arbitrary discrete stationary process with finite dispersion, there exists a three-dimensional stationary process $\{\psi(t), \zeta(t), \eta(t)\}$ with the following properties:

(*A*) $\{\xi(t)\} = \{\psi(t)\} + \{\zeta(t)\}$
(*B*) $\{\psi(t)\}$ *and* $\{\zeta(t)\}$ *are non-correlated.*

(*C*) $\{\psi(t)\}$ *is singular.*
(*D*) $\{\eta(t)\}$ *is non-autocorrelated, and* $E[\eta(t)]=E[\zeta(t)]=0$.
(*E*) $\{\zeta(t)\}=\{\eta(t)\}+b_1\{\eta(t-1)\}+b_2\{\eta(t-2)\}+\cdots$
where b_n *represent real numbers such that* $\sum b_n^2$ *is convergent.*

Thus, by starting from a purely random process $\{\eta(t)\}$, forming a sum process of the type $\{\zeta(t)\}=\{\eta(t)\}+b_1\{\eta(t-1)\}+b_2\{\eta(t-2)\}+\cdots$ and adding an independent process $\{\psi(t)\}$ ruled by an appropriate stationarity, an arbitrarily prescribed stationary process $\{\xi(t)\}$ is obtained (one of $\{\psi(t)\}$ and $\{\zeta(t)\}$ may be vanishing). The implications and importance of this theorem will be seen throughout later developments.

Stochastical difference equations

7.15 As already stated, Wold referred to equation (7.8) of §**7.6** as a *stochastical difference relation.* A more general representation is the linear autoregression (7.15), rewritten here as

$$\{\xi(t)\}+a_1\{\xi(t-1)\}+\cdots+a_h\{\xi(t-h)\}=\{\eta(t)\} \tag{7.16}$$

which Wold (*ibid.*, page 93) termed a *stochastical difference relation* between the processes $\{\xi(t)\}$ and $\{\eta(t)\}$. If $\{\eta(t)\}$ is known and $\{\xi(t)\}$ is unknown, (7.16) is termed a *stochastical difference equation.*

Equation (7.16) presents obvious analogies with ordinary difference equations of the form

$$x(t)+a_1x(t-1)+\cdots+a_hx(t-h)=y(t) \tag{7.17}$$

If there are no 'external influences' present, so that $y(t)=0$, the solutions to (7.17) describe how $x(t)$ develops through time from the initial values $x(t-1)=x_{t-1},\ldots,\ x(t-h)=x_{t-h}$, say, so that the expected value for $x(t)$ is $-a_1x_{t-1}-\cdots-a_hx_{t-h}$. If there is an external influence this expected value becomes $y(t)-a_1x_{t-1}-\cdots-a_hx_{t-h}$, with $y(t)$ being regarded as functional, i.e., uniquely determined at any future time point.

In contrast, the stochastical approach assumes that the external factors $\{\eta(t)\}$ are only ruled by certain probability laws: although $\{\eta(t)\}$ must follow the consistency relations (7.1) and (7.2), it could be a non-random or even a non-stationary process. If the $\{\eta(t)\}$ process is known, any sample series $(\ldots,\eta_{t-1},\eta_t,\eta_{t+1},\ldots)$ will describe an actual realization of the external influence. This will then determine the sample series $(\ldots,\xi_{t-1},\xi_t,\xi_{t+1},\ldots)$ of the process $\{\xi(t)\}$. However, typically only probabilistic knowledge of the actual path

$(\ldots,\eta_{t-1},\eta_t,\eta_{t+1},\ldots)$ will be available and so only a probability law about the behaviour of $\{\xi(t)\}$ can be reached, so that $(\ldots,\xi_{t-1},\xi_t,\xi_{t+1},\ldots)$ forms a solution to the stochastical difference equation (7.16).

The probability laws will provide information on the 'average' behaviour of $\{\xi(t)\}$ but Wold was careful to point out that it will not generally be the case that such behaviour will be identical to the solution of 'functional' difference equations for which there are no external influences.

Wold briefly contrasted the solutions of the stationary linear autoregression (7.16) with those from the process

$$\{\xi(t)\} - \{\xi(t-1)\} = \eta(t)$$

This is referred to as a *discrete homogenous process* which is *evolutive*, having solutions that are oscillatory with amplitude increasing over time.

The process (7.16) will be a stochastical difference equation if the coefficients a_i are real and if $\{\eta(t)\}$ is a discrete random process with finite variance. If $a_h \neq 0$ the equation is said to be of *order h*. If $\{\eta(t)\} = 0$ then, if (7.16) has any solutions, these will be singular and will have (7.8) with $\mu = 0$ as the relation of singularity. There will therefore be a non-vanishing stationary process that satisfies this equation if the related characteristic equation (see §**7.9**) has roots with modulus less than one. It also follows that

Theorem 8.
The series b_i defined by [7.12] *and* [7.13] *does not satisfy any difference equation of lower order than h.*

Theorem 9.
Let [7.16] *be a stochastical difference equation such that all roots of its characteristic equation are of a modulus less than unity, let $\{\eta(t)\}$ be stationary and have finite dispersion, and let the sequence b_i be given by* [7.12] *and* [7.13]. *Then*

$$\lim_{n\to\infty}[\{\eta(t)\} + b_1\{\eta(t-1)\} + b_2\{\eta(t-2)\} + \cdots + b_n\{\eta(t-n)\}]$$

will exist and forms a stationary solution of the equation.

The process of linear autoregression is thus, by construction, a solution of a stochastical difference equation such that $\{\eta(t)\}$ is stationary and all roots of the characteristic equation have modulus less than unity. Theorem 9 states that a mechanism whose intrinsic movements are damped will give rise to stationary oscillations when influenced by stationary external shocks.

7.16 Suppose that $\{\zeta(t)\}$ and $\{\eta(t)\}$ are two stationary processes such that

$$\zeta(t) = \eta(t) + b_1\eta(t-1) + b_2\eta(t-2) + \cdots \tag{7.18}$$

$$\eta(t) = \zeta(t) + a_1\zeta(t-1) + a_2\zeta(t-2) + \cdots \tag{7.19}$$

where

$\{\eta(t)\}$ is non-autocorrelated;
$\sigma^2(\eta, t) > 0$ is finite;
$E[\eta(t)] = 0$;
the sum $\sum b_k^2$ is convergent.

Substituting (7.18) into (7.19) obtains

$$\begin{aligned}\eta(t) &= \eta(t) + (a_1 + b_1)\zeta(t-1) + (a_2 + b_1a_1 + b_2)\zeta(t-2) + \cdots \\ &\quad + (a_k + b_1a_{k-1} + \cdots + b_{k-1}a_1 + b_k)\zeta(t-k) + \cdots\end{aligned}$$

so that

$$a_k + b_1a_{k-1} + \cdots + b_{k-1}a_1 + b_k = 0, \quad k = 1, 2, \ldots$$

Writing

$$K^2 = 1 + b_1^2 + b_2^2 + \cdots$$

we thus have

$$\sigma(\zeta) = K \cdot \sigma(\eta)$$

$$K \cdot r(\zeta(t+n); \eta(t)) = b_n$$

$$r_k = r_k(\zeta) = (b_k + b_1b_{k+1} + b_2b_{k+2} + \cdots)/K^2$$

Consider next rewriting (7.18) and (7.19) as

$$\zeta(t+s) = \eta(t+s) + b_1\eta(t+s-1) + b_2\eta(t+s-2) + \cdots$$

$$\zeta(t) + a_1\zeta(t-1) + a_2\zeta(t-2) + \cdots = \eta(t)$$

multiplying them together and taking expectations to obtain, for $s < 0$,

$$r_k + a_1r_{k-1} + \cdots + a_{k-1}r_1 + a_k + a_{k+1}r_1 + a_{k+2}r_2 + \cdots = 0$$

for all $k > 0$. In the same way, for $s \geq 0$,

$$(1 + a_1r_1 + a_2r_2 + \cdots)\sigma^2(\zeta) = \sigma^2(\eta)$$

and

$$r_k + a_1 r_{k+1} + a_2 r_{k+2} + \cdots = b_k / K^2 \quad k \geq 0$$

Forecasting with autoregressions

7.17 Wold (*ibid.*, pages 101–3) considered forecasting using the linear autoregressive process, taking the variable $\zeta(t+k)$ to be conditioned by the development of the processes (7.18) and (7.19) up to time point t inclusive. Thus

$$\zeta(t-k) = \zeta_{t-k}, \quad \eta(t-k) = \eta_{t-k}; \quad k = 0, 1, 2, \ldots$$

where $(\zeta_t, \zeta_{t-1}, \ldots)$ and $(\eta_t, \eta_{t-1}, \ldots)$ are observed sample series, and hence

$$\begin{aligned} \zeta_{t-k} &= \eta_{t-k} + b_1 \eta_{t-k-1} + b_2 \eta_{t-k-2} + \cdots \quad k = 0, 1, 2, \ldots \\ \eta_{t-k} &= \zeta_{t-k} + a_1 \zeta_{t-k-1} + a_2 \zeta_{t-k-2} + \cdots \quad k = 0, 1, 2, \ldots \end{aligned} \tag{7.20}$$

Since the variables $\eta(t)$ are uncorrelated, the linear forecast of $\zeta(t+k)$ made at time t, denoted $F_t[\zeta(t+k)]$, is given by (7.18) as

$$F_t[\zeta(t+k)] = b_k \eta_t + b_{k+1} \eta_{t-1} + b_{k+2} \eta_{t-2} + \cdots \quad k = 1, 2, \ldots \tag{7.21}$$

Equivalently, from (7.19) the forecast can be written, for $k = 1, 2, \ldots$, as

$$\begin{aligned} F_t[\zeta(t+k)] = &-a_1 \cdot F_t[\zeta(t+k-1)] - a_2 \cdot F_t[\zeta(t+k-2)] - \cdots \\ &-a_{k-1} \cdot F_t[\zeta(t+1)] - a_k \zeta_t - a_{k+1} \zeta_{t-1} - a_{k+2} \zeta_{t-2} - \cdots \end{aligned} \tag{7.22}$$

This form shows how successive forecasts may be calculated. It can be shown to be equivalent to (7.21) by writing every $F_t[\zeta(t+k-i)]$ in the above expression in the form implied by (7.22) and expressing every ζ_{t-i} in terms of η_{t-i} by means of (7.20). Alternatively, the forecasts may be expressed in terms of ζ_{t-i}:

$$F_t[\zeta(t+k)] = f_{k,0} \zeta_t + f_{k,1} \zeta_{t-1} + f_{k,2} \zeta_{t-2} + \cdots \tag{7.23}$$

For $k = 1$, (7.22) becomes

$$F_t[\zeta(t+1)] = -a_1 \zeta_t - a_2 \zeta_{t-1} - a_3 \zeta_{t-2} - \cdots$$

so that

$$f_{1,i} = -a_{i+1}$$

Substituting (7.23) into (7.22) obtains

$$\begin{aligned}
&f_{k,0} + a_1 f_{k-1,0} + a_2 f_{k-2,0} + \cdots + a_{k-1} f_{1,0} + a_k = 0 \\
&f_{k,1} + a_1 f_{k-1,1} + a_2 f_{k-2,1} + \cdots + a_{k-1} f_{1,1} + a_{k+1} = 0 \\
&\ldots
\end{aligned}$$

Thus, after having calculated the coefficients $f_{k-i,j}$ appearing in the forecasts $F_t[\zeta(t+k-i)]$, these relations yield the coefficients $f_{k,j}$ necessary for computing $F_t[\zeta(t+k)]$ in terms of the ζ_{t-i}s.

The relations (7.21)–(7.23) are referred to by Wold as the *forecasting formulae*. Given the sample series $(\zeta_t, \zeta_{t-1}, \zeta_{t-2}, \ldots)$ and/or $(\eta_t, \eta_{t-1}, \eta_{t-2}, \ldots)$, 'these formulae furnish the best linear forecast as to the future development of the series' (*ibid.*, page 102). 'Best' is used in the sense that the expected squared error of the forecast, which, from (7.21), is $(1 + b_1^2 + b_2^2 + \cdots + b_{k-1}^2)\sigma^2(\eta)$, is shown to be a minimum. As $k \to \infty$, this expression tends to $K^2\sigma^2(\eta) = \sigma^2(\zeta)$, so that for large values of k, the forecast $F_t[\zeta(t+k)]$ is approximately of the same efficiency as the trivial forecast $E[\zeta(t+k)] = E[\zeta(t)] = 0$.

Linear autoregressions

7.18 The linear autoregression process $\{\zeta(t)\}$ is

$$\{\zeta(t)\} + a_1\{\zeta(t-1)\} + \cdots + a_h\{\zeta(t-h)\} = \{\eta(t)\} \tag{7.24}$$

where the stationary process $\{\eta(t)\}$ is non-autocorrelated and $E[\eta(t)] = E[\zeta(t)] = 0$. From the expressions at the end of §**7.16**, we have

$$(1 + a_1 r_1 + a_2 r_2 + \cdots + a_h r_h)\sigma^2(\zeta) = \sigma^2(\eta)$$

and three groups of relations involving the autocorrelation coefficients:

$$\begin{cases}
\ldots \\
r_k + a_1 r_{k-1} + a_2 r_{k-2} + \cdots + a_{h-1} r_{k-h+1} + a_h r_{k-h} = 0 \\
\ldots \\
r_{h+1} + a_1 r_h + a_2 r_{h-1} + \cdots + a_{h-2} r_3 + a_{h-1} r_2 + a_h r_1 = 0 \\
r_h + a_1 r_{h-1} + a_2 r_{h-2} + \cdots + a_{h-2} r_2 + a_{h-1} r_1 + a_h = 0
\end{cases} \tag{7.25}$$

$$\begin{cases}
r_{h-1} + a_1 r_{h-2} + a_2 r_{h-3} + \cdots + a_{h-2} r_1 + a_{h-1} + a_h r_1 = 0 \\
\ldots \\
r_1 + a_1 + a_2 r_1 + \cdots + a_{h-2} r_{h-3} + a_{h-1} r_{h-2} + a_h r_{h-1} = 0
\end{cases} \tag{7.26}$$

$$\begin{cases} 1+a_1r_1+a_2r_2+\cdots+a_{h-2}r_{h-2}+a_{h-1}r_{h-1}+a_hr_h=1/K^2 \\ r_1+a_1r_2+a_2r_3+\cdots+a_{h-1}r_h+a_hr_{h+1}=b_1/K^2 \\ \cdots \\ r_k+a_1r_{k+1}+a_2r_{k+2}+\cdots+a_{h-1}r_{h+k-1}+a_hr_{h+k}=b_k/K^2 \\ \cdots \end{cases} \tag{7.27}$$

The group (7.25) is given in Walker (1931): see §**6.10**, equation (6.15). The r_k for $k \geq h$ satisfy a difference equation which is the same as that satisfied by the b_k sequence and both evolve as damped oscillations.

The second group (7.26) contains $h-1$ relations and involves $r_1, r_2, \ldots, r_{h-1}$, which may be obtained directly from the a_i by solving the system (7.26), and may be regarded as a corollary to

Theorem 10.
Let $\{\zeta(t)\}$ *be a process of linear autoregression of order h. Then the autocorrelation coefficients of* $\{\zeta(t)\}$ *satisfy no difference equation* (*cf.* [7.8]) *of lower order than h.*

Since $a_k=0$ for $k>h$, the forecasting formula (7.22) shows that the forecasts $F_t[\zeta(t+1)], F_t[\zeta(t+2)], \ldots, F_t[\zeta(t+k)], \ldots$ will satisfy a difference equation with respect to k, so that the forecasts will also form a damped oscillation, revealing how the series will evolve from time t if there were no external influences present at the future times $t+1, t+2, \ldots$. Consequently, $F_t[\zeta(t+k)] \to 0 = E[\zeta(t)]$ as $k \to \infty$, in agreement with the concluding remark of §**7.17**.

By referring to Theorem 5 of §**7.12**, Wold next obtained

Theorem 11 (abridged)
The generating function $W(x)$ *of the autocorrelation coefficients in a process* $\{\zeta(t)\}$ *of linear autoregression is absolutely continuous, and has a bounded derivative* $W'(x)$ *given by* $W'(x)=G(x)+G(-x)-1$, where

$$G(x)=\frac{1+(a_1+r_1)e^{ix}+\cdots+(a_{h-1}+a_{h-2}r_1+\cdots+a_1r_{h-2}+r_{h-1})e^{i(h-1)x}}{1+a_1e^{ix}+a_2e^{i2x}+\cdots+a_he^{ihx}}$$

7.19 Wold (*ibid.*, pages 110–21) analysed in detail the autoregression (7.24) when $h=2$. Denoting the roots of the associated characteristic equation (cf. §**7.9**) as p and q, we then have

$$a_1=p+q, \quad a_2=pq, \quad a_n=0 \quad \text{for } n>2; \ \ |p|<1, |q|<1$$

and thus

$$\zeta(t)-(p+q)\zeta(t-1)+pq\zeta(t-2)=\eta(t) \tag{7.28}$$

As a_1 and a_2 must be real, two possibilities exist: either I: p and q are real, or II: $p = A + iB$, $q = A - iB$, where A and B represent real numbers such that

$$A^2 + B^2 = |p^2| = |q^2| < 1$$

Assuming that $p \neq q$, the general solution of the difference equation obtained from (7.28) with $\eta(t) = 0$ is, for P_1 and P_2 arbitrary,

$$P_1 \cdot p^t + P_2 \cdot q^t \tag{7.29}$$

For case II, the solution is

$$Q_1 \cdot C^t \cos \lambda t + Q_2 \cdot C^t \sin \lambda t$$

where Q_1 and Q_2 are arbitrary and

$$C = +\sqrt{A^2 + B^2} \quad \cos \lambda = A/C, \quad 0 < \lambda < \pi$$

I. p and q are real.
Substituting the general solution (7.29) for b_1 and b_2 into the system (7.13) and solving for P_1 and P_2 obtains

$$b_k = \frac{p}{p-q} \cdot p^k + \frac{q}{q-p} \cdot q^k = \frac{p^{k+1} - q^{k+1}}{p-q} \quad k \geq 0$$

Inserting this result into the expression for K^2 yields

$$K^2 = \frac{\sigma^2(\zeta)}{\sigma^2(\eta)} = \frac{1+pq}{(1-p^2)(1-q^2)(1-pq)}$$

The system (7.26) reduces to the single equation

$$r_1 + a_1 + a_2 r_1 = 0$$

Solving for

$$r_1 = \frac{p+q}{1+pq}$$

observing that $r_0 = 1$, and equating these two coefficients to (7.29) for $t = 0$, 1 obtains

$$r_k = \frac{p(1-q^2)}{(p-q)(1+pq)} \cdot p^k + \frac{q(1-p^2)}{(q-p)(1+pq)} \cdot q^k \quad k \geq 0$$

Two special cases are worth considering. If $q=0$ then the various relations reduce to

$$\zeta(t) - p\zeta(t-1) = \eta(t)$$

$$b_k = r_k = p^k, \quad k \geq 0 \quad \sigma^2(\zeta) = \frac{\sigma^2(\eta)}{1-p^2}$$

These formulae cover the case $h=1$ and were discussed in Walker (1931). If $q=-p$, we have

$$\zeta(t) - p^2\zeta(t-2) = \eta(t)$$

$$b_{2k} = r_{2k} = p^{2k}, \quad b_{2k+1} = r_{2k+1} = 0, \quad k \geq 0 \quad \sigma^2(\zeta) = \frac{\sigma^2(\eta)}{1-p^4}$$

II. p and q are complex conjugates.

$$p = A + iB, \quad q = A - iB$$

Here we have, by a similar analysis,

$$\zeta(t) - 2A \cdot \zeta(t-1) + (A^2 + B^2) \cdot \zeta(t-2) = \eta(t)$$

$$b_k = C^k \cos k\lambda + \frac{A}{B} \cdot \sin k\lambda \tag{7.30}$$

$$r_k = C^k \cos k\lambda + \frac{A}{B} \cdot \frac{1-C^2}{1+C^2} C^k \sin k\lambda \tag{7.31}$$

$$\sigma^2(\zeta) = \frac{1+C^2}{(1-C^2)(1+C^4-2A^2+2B^2)} \cdot \sigma^2(\eta) \tag{7.32}$$

This set-up covers the case of an oscillatory mechanism whose intrinsic oscillations consist of a single damped harmonic with a frequency lying in the interval $0 < \lambda < \pi$ and a damping factor C^t. Wold then considered whether periodogram analysis would accurately uncover λ. If the roots $A \pm iB$ of the characteristic equation lie close to the periphery of the unit circle, so that the intrinsic oscillations are only slightly damped, then periodogram analysis will be able to discover the frequency of the intrinsic oscillation. The more heavily damped the intrinsic oscillation is, however, the larger will the bias be in estimating λ, with periodogram analysis overestimating the intrinsic period if this is above 4 time units and underestimating it if it is between 2 and 4 units. Wold (*ibid.*, page 117) summed up these conclusions by stating that 'the situation may be described by saying that the inference drawn from the characteristic equation of the intrinsic oscillations does not apply directly to the oscillations of the mechanism when influenced by random external factors'.

7.20 Wold used the linear autoregression of order two to reveal a connection with the law of the sinusoidal limit (see §**5.15** and §**7.11**). Let

$$L(x) = x^2 - 2Ax + 1 = 0 \quad -1 < A < 1$$

be the characteristic equation of the simple harmonic

$$P_1 \cos \lambda_1 t + P_2 \sin \lambda_2 t$$

and let $\{\zeta^{(1)}(t)\}, \{\zeta^{(2)}(t)\}, \ldots$ be a sequence of autoregressions of the form

$$\zeta^{(p)}(t) - 2A_p \cdot \zeta^{(p)}(t-1) + C_p^2 \cdot \zeta^{(p)}(t-2) = \eta^{(p)}(t)$$

where

$$\lim_{p\to\infty} A_p = A, \quad \lim_{p\to\infty} C_p = 1$$

Using (7.32),

$$\sigma^2(\zeta^{(p)}) = K_p^2 \cdot \sigma^2(\eta^{(p)}) = \frac{1 + C_p^2}{(1 - C_p^2)(1 + C_p^4 - 2A_p^2 + 2B_p^2)} \cdot \sigma^2(\eta^{(p)})$$

where $K_p^2 = 1 + (b_1^{(p)})^2 + (b_2^{(p)})^2 + \cdots$. Since $1 - C_p^2$ tends to zero as $p \to \infty$, it follows that $K_p \to \infty$ as $p \to \infty$. Since, from (7.30), $b_k^{(p)}$ is bounded,

$$\lim_{p\to\infty} \frac{b_k^{(p)}}{K_p^2} = 0$$

Thus, the systems of equations (7.25–7.27) imply that

$$\lim_{p\to\infty} L(r_k^{(p)}) = 0, \quad -\infty < k < \infty$$

which, in turn, implies that the sequence $\{\zeta^{(p)}(t)\}$ is ruled by the singularities that embody the sinusoidal limit theorem (see §**7.11**). Thus, if we approximate an arbitrary sample series $(\zeta) = (\zeta_1^{(p)}, \ldots, \zeta_n^{(p)})$ by a simple harmonic with frequency λ, say $x_p(t, \zeta)$, then, holding n fixed, it follows that, for every $\varepsilon > 0$

$$\lim_{p\to\infty} P[|x_p(1, \zeta) - \zeta_1^{(p)}| < \varepsilon, \ldots, |x_p(n, \zeta) - \zeta_n^{(p)}| < \varepsilon] = 1$$

This result generalizes to a relation $L(x) = 0$ of arbitrary order, so that the process of linear autoregression 'forms a convenient starting point for the construction of sequences covered by the sinusoidal limit theorems' (*ibid.*, page 121).

Moving average processes

7.21 The general moving average of order h is

$$\{\zeta(t)\} = \{\eta(t)\} + b_1\{\eta(t-1)\} + \cdots + b_h\{\eta(t-h)\} \tag{7.33}$$

where $\{\eta(t)\}$ is purely random or, more generally, non-autocorrelated, and the sequence $(b) = (b_1, \ldots, b_h)$ is real. We continue to assume that $\sigma^2(\eta)$ is finite and that $E(\eta_t) = 0$. Unlike the process of linear autoregression, where the sequences of autocorrelation coefficients and forecasts follow damped harmonic processes, only the first h elements of these sequences are non-zero. The variance of $\{\zeta(t)\}$ is

$$\sigma^2(\zeta) = (1 + b_1^2 + b_2^2 + \cdots + b_h^2) \cdot \sigma^2(\eta)$$

while the autocorrelations are given by

$$r_k(\zeta) = \begin{cases} (b_k + b_1 b_{k+1} + \cdots + b_h b_{h+k})/(1 + b_1^2 + \cdots + b_h^2) & \text{for } k \leq h \\ 0 & \text{for } k > h \end{cases} \tag{7.34}$$

where $k \geq 0$ and $b_0 = 1$. Specializing the forecast formula (7.21) gives

$$F_t[\zeta(t+k)] = \begin{cases} b_k \eta_t + b_{k+1}\eta_{t-1} + \cdots + b_h \eta_{t-h+k} & \text{for } 0 \leq k \leq h \\ 0 & \text{for } k > h \end{cases}$$

Given the moving average process (7.33), does a primary process of the form (7.19)

$$\eta(t) = \zeta(t) + a_1\zeta(t-1) + a_2\zeta(t-2) + \cdots$$

exist and, if so, how can the coefficients (a) be obtained? Consider the characteristic equation

$$x^h + b_1 x^{h-1} + \cdots + b_{h-1}x + b_h = 0 \tag{7.35}$$

If all the roots of (7.35) have modulus less than unity, Theorem 9 states that an infinite sequence $(a) = (a_1, a_2, \ldots)$ such that (7.19) holds is given by the system (7.13) and the difference relation of §**7.9** on replacing the a_is by the b_is. These relations constitute a difference equation of order h satisfied by the sequence (a), which forms a damped harmonic. Under these circumstances the relations of §**7.16** hold, and take the following form:

$$a_{2h+k}r_h + a_{2h+k-1}r_{h-1} + \cdots + a_{h+k+1}r_1 + a_{h+k} + a_{h+k-1}r_1 + \cdots + a_{k+1}r_{h-1} + a_k r_h = 0 \tag{7.36}$$

$$\begin{cases} a_{2h}r_h + a_{2h-1}r_{h-1} + \cdots + a_{h+1}r_1 + a_h + a_{h-1}r_1 + \cdots + a_1r_{h-1} + r_h = 0 \\ a_{2h-1}r_h + a_{2h-2}r_{h-1} + \cdots + a_hr_1 + a_{h-1} + a_{h-2}r_1 + \cdots + a_1r_{h-2} + r_{h-1} = 0 \\ \cdots \\ a_{h+1}r_h + a_hr_{h-1} + \cdots + a_2r_1 + a_1 + r_1 = 0 \end{cases} \tag{7.37}$$

$$\begin{cases} a_hr_h + a_{h-1}r_{h-1} + \cdots + a_1r_1 + 1 = 1/K^2 \\ a_{h-1}r_h + \cdots + a_1r_2 + r_1 = b_1/K^2 \\ \cdots \\ a_2r_h + a_1r_{h-1} + r_{h-2} = b_{h-2}/K^2 \\ a_1r_h + r_{h-1} = b_{h-1}/K^2 \\ r_h = b_h/K^2 \end{cases} \tag{7.38}$$

Thus, if (7.35) has no root x_k falling outside the unit circle, a set of well-defined linear operations on the moving average (7.33) will yield the primary process $\{\eta(t)\}$ given by (7.19): if $|x_k| \leq 1$ for all k, the sequence (b) and the process $\{\zeta(t)\}$ is termed *regular*.

Wold then uses the generating function $W(x)$ of §**7.12** to obtain the fundamental identity

$$\begin{aligned} &\frac{1}{K^2}(x^h + b_1x^{h-1} + \cdots + b_{h-1}x + b_n)(b_hx^h + b_{h-1}x^{h-1} + \cdots + b_1x + 1) \\ &= r_hx^{2h} + r_{h-1}x^{2h-1} + \cdots + r_1x^{h+1} + x^h + r_1x^{h-1} + \cdots + r_{h-1}x + r_h \end{aligned} \tag{7.39}$$

Since the zeros of the factor $b_hx^h + b_{h-1}x^{h-1} + \cdots + b_1x + 1$ will be x_k^{-1}, the zeros of the right-hand side of (7.39) may be denoted $x_1, x_2, \ldots, x_{2h-1}, x_{2h}$, where

$$x_k = x_{2h+1-k}^{-1}, \quad 0 < |x_1| \leq |x_2| \leq \cdots \leq |x_h| \leq 1 \leq |x_{h+1}| \leq \cdots \leq |x_{2h}|$$

It then follows that if there exists another sequence, say $(1, b_1^{(i)}, \ldots, b_h^{(i)})$, such that the associated moving average has autocorrelation coefficients coinciding with those of (7.33), then one zero of the polynomial $x^h + b_1^{(i)}x^{h-1} + \cdots + b_{h-1}^{(i)}x + b_h^{(i)}$ will equal either x_1 or x_1^{-1}, another either x_2 or x_2^{-1}, etc. There will be at most 2^h real sequences of this type, say $(b_k^{(0)}), \ldots, (b_k^{(s)})$. If $(b_k^{(0)})$ represents the regular sequence then all other sequences are non-regular.

Letting $(b_k^{(i)})$ be a group of sequences attached to the regular sequence $(b_k^{(0)})$ and writing

$$(K^{(i)})^2 = 1 + (b_i^{(1)})^2 + (b_i^{(2)})^2 + \cdots + (b_i^{(h)})^2$$

we can then define a group of moving averages as

$$\zeta^{(i)}(t) = \frac{K^{(0)}}{K^{(i)}}[\eta(t) + b_1^{(i)}\eta(t-1) + \cdots + b_h^{(i)}\eta(t-h)] \quad i = 1, \ldots, s < 2^h \tag{7.40}$$

It follows from its construction that this group will contain one, and only one, regular process and that

$$\sigma^2(\zeta^{(i)}) = \sigma^2(\zeta^{(j)}); \quad r_k^{(i)} = r_k^{(j)}, \quad k = 0, \pm 1, \pm 2, \ldots \quad i, j = 1, \ldots, s < 2^h$$

If all the roots of (7.35) lie of the periphery of the unit circle, i.e., $x_k = 1$ for all k, then the group will contain just the process $\{\zeta^{(0)}(t)\} = \{\zeta(t)\}$; otherwise the group will contain at most 2^h. If we denote by $\{\eta^{(i)}(t)\}$ the residuals of the non-regular processes, then these are given by

$$\begin{aligned}\frac{K^{(i)}}{K^{(0)}}\eta^{(i)}(t) &= \eta(t) + (a_1 + b_1^{(i)}) \cdot \eta(t-1) + (a_2 + a_1 b_1^{(i)} + b_2^{(i)}) \cdot \eta(t-2) + \cdots \\ &\quad + (a_h + a_{h-1} b_1^{(i)} + \cdots + b_h^{(i)}) \cdot \eta(t-h) \\ &\quad + (a_{h+1} + a_h b_1^{(i)} + \cdots + a_1 b_h^{(i)}) \cdot \eta(t-h-1) + \cdots \end{aligned} \tag{7.41}$$

and we have

$$\{\zeta^{(i)}(t)\} = \{\eta^{(i)}(t)\} + b_1\{\eta^{(i)}(t-1)\} + \cdots + b_h\{\eta^{(i)}(t-h)\} \tag{7.42}$$

for all processes in the group $\{\zeta^{(i)}\}$.

These ideas may be illustrated by the following examples.

Example 1. Let $h = 1$ with $b_1 = 2$. Then (7.39) is

$$0.2(x+2)(2x+1) = 0.4x^2 + x + 0.4$$

Hence $r_1 = 0.4$ and $r_k = 0$ for $k > 1$. The characteristic equation is $(x+2)(2x+1) = 0$, which gives two sequences $b^0 = (1, 0.5)$ and $b^1 = (1, 2)$. The system (7.13) gives $a_1 = -0.5$ and, in general, $a_k = -0.5^k$. Thus, for the regular process b^0,

$$\{\zeta(t)\} = \{\eta(t)\} + 0.5\{\eta(t-1)\}$$

and it follows that

$$\{\eta(t)\} = \{\zeta(t)\} - 0.5\{\zeta(t-1)\} + 0.5^2\{\zeta(t-2)\} - 0.5^3\{\zeta(t-3)\} + \cdots$$

Since $K^2 = 1.25$ and $[K^{(1)}]^2 = 5$, (7.40) gives

$$\{\zeta^{(1)}(t)\} = 0.5\{\eta(t)\} + \{\eta(t-1)\}$$

while (7.41) gives

$$\eta^{(1)}(t) = \tfrac{1}{2}\eta(t) + \tfrac{3}{4}\eta(t-1) - \tfrac{3}{8}\eta(t-2) + \tfrac{3}{16}\eta(t-3) - \cdots$$

Thus, using (7.42), we have

$$\zeta^{(1)}(t) = \{\eta^{(1)}(t)\} + 0.5\{\eta^{(1)}(t-1)\} = 0.5\{\eta(t)\} + \{\eta(t-1)\}$$

from which it can easily be verified that $\sigma^2(\eta^{(1)}(t)) = \sigma^2(\eta(t))$ and that $r_k(\eta^{(1)}) = 0$ for $k \neq 0$, i.e., $\{\eta^{(1)}(t)\}$ is non-autocorrelated.

Example 2. Here we suppose that $r_1 = \frac{1}{6}$, $r_2 = -\frac{1}{3}$ and $r_k = 0$ for $k > 2$. The fundamental identity (7.39) reads

$$\tfrac{2}{3}(x^2 + 0.5x - 0.5)(-0.5x^2 + 0.5x + 1) = -\tfrac{1}{3}x^4 + \tfrac{1}{6}x^3 + x^2 + \tfrac{1}{6}x - \tfrac{1}{3}$$

Noting that $x^2 + 0.5x - 0.5 = (x - 0.5)(x + 1)$, so that neither root lies outside the unit circle, there are two sequences, $b^0 = (1, 0.5, -0.5)$, which is regular, and $b^1 = (1, -1, -2)$. Thus the regular process is defined as

$$\{\zeta(t)\} = \{\eta(t)\} + 0.5\{\eta(t-1)\} - 0.5\{\eta(t-2)\}$$

from which we obtain the (a) sequence as

$$a_k = \tfrac{1}{3}\left(\tfrac{1}{2}\right)^k + \tfrac{2}{3}(-1)^k$$

The non-regular process is given by

$$\begin{aligned}\{\zeta^{(1)}(t)\} &= 0.5\{\eta(t)\} - 0.5\{\eta(t-1)\} - \{\eta(t-2)\} \\ &= \{\eta^{(1)}(t)\} + 0.5\{\eta^{(i)}(t-1)\} - 0.5\{\eta^{(i)}(t-2)\}\end{aligned}$$

with the non-autocorrelated residual being

$$\eta^{(1)}(t) = \tfrac{1}{2}\eta(t) - \tfrac{3}{4}\eta(t-1) - \tfrac{3}{8}\eta(t-2) - \tfrac{3}{16}\eta(t-3)$$

Example 3. Finally, suppose $r_1 = -0.5$ and $r_k = 0$ for $k > 1$. Now (7.39) reads

$$0.5(x-1)(-x+1) = 0.5x^2 + x - 0.5$$

and we conclude that there is just one sequence $b = (1, -1)$, associated with the process

$$\{\zeta(t)\} = \{\eta(t)\} - \{\eta(t-1)\}$$

Here (7.35) has just one root, which falls on the unit circle. In these circumstances, Wold (*ibid.*, pages 124–6) showed that a limit process of the form

$$\{\eta(t)\} = \lim_{i\to\infty}[\{\zeta(t)\} + a_1^{(i)}\{\zeta(t-1)\} + a_2^{(i)}\{\zeta(t-2)\} + \cdots]$$

exists where, in this case $a_k^{(i)} = (1-\varepsilon)^k$, with $0 < \varepsilon \to 0$ as $i \to \infty$. Setting $\varepsilon = 10^{-i}$, then

$$\{\eta(t)\} = \lim_{i\to\infty}[\{\zeta(t)\} + (1-10^{-i})\{\zeta(t-1)\} + (1-10^{-i})^2\{\zeta(t-2)\} + \cdots]$$

which we see approaches the process

$$\{\eta(t)\} = \{\zeta(t)\} + \{\zeta(t-1)\} + \{\zeta(t-2)\} + \cdots$$

Some applications of stationary processes

7.22 In §§**6.12–6.13** we discussed Walker's (1931) analysis of the Port Darwin air pressure data, focusing on the complete 'correlation periodogram', which Wold (*ibid.*, page 135) referred to more succinctly as the correlogram – 'for the sake of brevity in writing, the graphs of serial and autocorrelation coefficients will be termed *correlograms* (*empirical* and *hypothetical* respectively)'. Wold focused attention on Walker's preliminary analysis of the first 40 serial coefficients, which showed that the r_k, $0 \le k \le 40$, could be approximately represented by the function

$$r_k = 0.19 \cdot 0.96^k \cos \pi k/6 + 0.15 \cdot 0.98^k + 0.66 \cdot 0.71^k \qquad (7.43)$$

This function has a damped harmonic with a period of 12 quarters and satisfies the difference equation

$$r_k - 3.35r_{k-1} + 4.43r_{k-2} - 2.71r_{k-3} + 0.64r_{k-4} = 0$$

Walker then used the argument that, since (7.25) implies that

$$r_k + a_1r_{k-1} + \cdots + a_hr_{k-h} = 0, \quad k \ge h$$

then the empirical series, $\ldots, \bar{\zeta}_{t-1}, \bar{\zeta}_t, \bar{\zeta}_{t+1}, \ldots$, follows the autoregression

$$\bar{\zeta}_t + a_1\bar{\zeta}_{t-1} + \cdots + a_h\bar{\zeta}_{t-h} = \bar{\eta}_t \qquad (7.44)$$

i.e.,

$$\bar{\zeta}_t - 3.35\bar{\zeta}_{t-1} + 4.43\bar{\zeta}_{t-2} - 2.71\bar{\zeta}_{t-3} + 0.64\bar{\zeta}_{t-4} = \bar{\eta}_t \qquad (7.45)$$

Wold (*ibid.*, pages 144–5) pointed out that such an argument was, in fact, invalid, as the autocorrelation coefficients not only satisfy (7.25) but also the systems (7.26) and (7.27): in fact, the coefficients $r_1, r_2, \ldots, r_{h-1}$ will be uniquely determined by (7.26) in terms of the a_is. Thus it is not certain that the autocorrelation coefficients corresponding to (7.45) will be given by the function (7.43). Wold showed that the system (7.26) corresponding to (7.45) gives the values $r_1 = 0.93$, $r_2 = 0.72$ and $r_3 = 0.43$, rather than the values 0.75, 0.55 and 0.35 given by Walker. Wold also showed that the relationship between the variance of the disturbances and the observed series given by Walker was incorrect, so that all the parameters of Walker's model required modification.

In fact, Wold suggested that the simpler model

$$\bar{\zeta}_t - 0.73\bar{\zeta}_{t-1} = \bar{\eta}_t \tag{7.46}$$

gave a good fit to the first few serial coefficients. Since this model is an example of case I of §**7.19** with $q = 0$, it gives the sequence of correlations $r_1 = 0.73$, $r_2 = 0.73^2 = 0.53$, $r_3 = 0.73^3 = 0.39$, $r_4 = 0.73^4 = 0.28$, compared to the actual air pressure serial coefficients of 0.76, 0.56, 0.36 and 0.18 respectively.

Turning his attention to the empirical correlogram of air pressure, shown in Figure 6.10, Wold remarked that

> the serial coefficients show rather small deviations from zero in the interval $3 < k < 40$. On the other hand, the increase in amplitude for certain k-values > 40 might be due to the successive reduction in the number of correlates. Perhaps this argument is sufficient to explain also why the fluctuations are somewhat larger in that alternative variant of a correlogram given by Walker, where all serial coefficients are based on 77 pairs of correlates. As the fluctuations, furthermore, seem rather irregular and aperiodic – at least to my eye – it is doubtful whether it would be possible to improve sensibly the approach [7.46] by taking into account more distant elements $\bar{\zeta}_{t-2}, \bar{\zeta}_{t-3}$, etc. In this connexion it is rather interesting to notice that according to the general analysis there exists no process of linear autoregression having [7.43] for autocorrelation coefficients. Another reason for resting satisfied with the simple approach [7.46] is that the ordinates of the periodogram presented [in Figure 6.9] are all lying on about the same level – this periodogram does not, like that of the sunspots, suggest a scheme of linear autoregression with a tendency to periodicity. (*ibid.*, pages 145–6)

Thus Wold suggested that a simpler, first-order linear autoregression presented the best fit to the Port Darwin air pressure data: a view that would hold that any tendency to periodicity in the series was of a spurious nature.[5]

7.23 A function of the form

$$y(t) = \mu + \sum_{k=1}^{s} C_k \cos(\lambda_k + \varphi_k) = \mu + \sum_{k=1}^{s} (A_k \cos \lambda_k t + B_k \sin \lambda_k t) \tag{7.47}$$

is referred to by Wold as a *composed harmonic*. Suppose that $\zeta(t) = y(t) + \eta(t)$, where $\eta(t)$ is purely random with variance $\sigma^2(\eta)$. This is known as the *scheme of hidden periodicities*. The autocorrelations of $\zeta(t)$ are, for $k \neq 0$, given by (Wold, equation 46)

$$r_k = \frac{\sum_{i=1}^{s} C_i^2 \cos \lambda_i k}{2\sigma^2(\eta) + \sum_{i=1}^{s} C_i^2} \tag{7.48}$$

so that r_k is also a composed harmonic and there exist arbitrarily large k-values such that

$$r_k \approx r_0 = \frac{\sum_{i=1}^{s} C_i^2}{2\sigma^2(\eta) + \sum_{i=1}^{s} C_i^2}$$

The implication of this is that even if an observed time series clearly shows a cyclical character but has serial coefficients that are gradually vanishing, then the scheme of hidden periodicities is inappropriate.

In contrast to (7.48), the correlogram of a linear autoregression will form a damped harmonic while that for a moving average will cut off beyond a certain k-value. These possibilities are illustrated in Figure 7.1, in which the

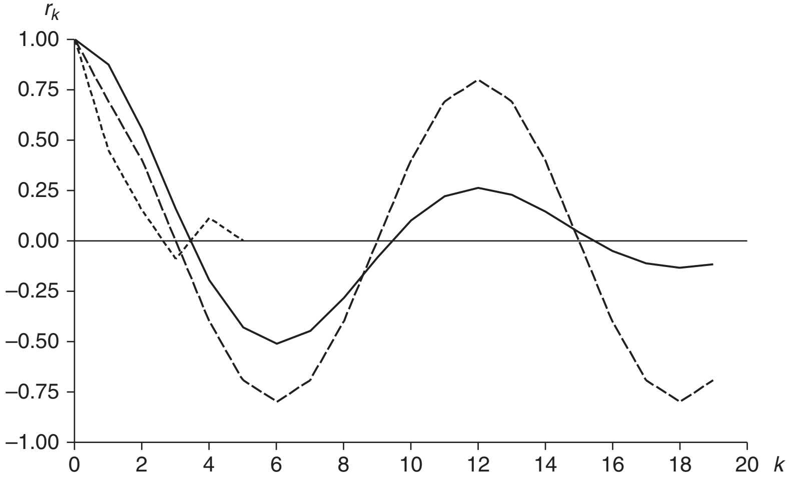


Figure 7.1 Correlograms illustrating the schemes of hidden periodicities (dashed line), linear autoregression (unbroken line), and moving average (dotted line)

correlograms are constructed in the following way. For the scheme of hidden periodicities, (7.48) was used with $s = C_1 = 1$, $\sigma^2(\eta) = 1.25$ and $\lambda_1 = \pi/6$. The linear autoregression uses (7.31) with $A = 0.8$, $B = 0.4$ and $\lambda = \pi/6$, while the moving average uses (7.34) with $b_1 = 0.7$, $b_2 = 0.4$, $b_3 = -0.3$ and $b_4 = 0.2$. Given the very different behaviour of the alternative schemes, Wold (*ibid.*, page 147) argued that 'it may be expected that we would obtain useful suggestions by inspecting the empirical correlogram when searching for an adequate scheme to be applied to an observational time series. For this reason, the construction of an empirical correlogram is taken as (a) starting point in ... applications'.

This led Wold to recommend the following approach to applying the theory developed above.

> If the empirical correlogram suggests a scheme of hidden periodicities, the next step would be to construct a periodogram for a more detailed analysis of possible periodicities in the material under investigation.
>
> Next, if the correlogram suggests a scheme of linear autoregression, our first problem is to find a scheme [7.15] such that the corresponding hypothetical correlogram will fit the empirical one. The chief difficulty is to derive suitable values for the coefficients (a) – when having arrived at a set of coefficients (a), the corresponding autocorrelation coefficients will be uniquely determined by the system [7.25–7.26], and the residuals $\overline{\eta}_t$ by the relations [7.44]. It is further a desideratum that these residuals be as small as possible. Having seen above that these problems are more intricate than emphasized in earlier studies of the graph of serial coefficients, it will be found that an empirical autoregression analysis as proposed by Yule (1927) will be useful in this connection.
>
> Finally, it may happen that the empirical correlogram will suggest a scheme of moving averages. As far as I know, the problem of fitting this scheme to observational data has not been attacked.... It will be seen that the relation [7.39] gives a starting point for attacking this problem. (*ibid.*, page 148)

7.24 Before embarking on applications, Wold took great pains to point out various limitations of the methodology. A major drawback was the lack of an inferential framework within which any results might be assessed –

> in time series analysis, significance problems are extremely intricate.... Consequently, all questions about the significance and the interpretation of the quantitative results fall outside the scope of this study, and again an explicit warning is given against attaching importance to the numerical values found for the parameters of the different models fitted to the observational data. (*ibid.*, pages 148–9)

A second question was that of identification: it will not be possible to distinguish between different schemes that give rise to the same set of autocorrelation coefficients. As an example of this, Wold considered the nonlinear function

$$\xi(t) = \eta(t) \cdot \eta(t-1)$$

in which $\eta(t)$ is, as usual, a zero mean random process. It is clear that $\xi(t)$ will be non-autocorrelated and hence indistinguishable from $\eta(t)$.

> Thus, if we have found a hypothetical scheme that fits well to an empirical correlogram, it is perfectly possible that there are other schemes which yield an equally close approximation. When it is necessary to choose between different schemes, it may happen that theoretical arguments will speak in favour of one of the schemes.... (T)he schemes of linear regression often seem plausible from theoretical viewpoints, at least to a first approximation. On the other hand, a rational choice between different schemes may be alternatively based on an examination of other structural properties of the time series than its serial coefficients. (*ibid.*, pages 149–50)

Moreover, if a process is actually generated by a nonlinear function, then restricting analysis to only linear autoregressions may lead to unduly complicated processes being arrived at.

An application of moving averages

7.25 Wold's first application was to analyse Beveridge's Index of Fluctuation, the periodogram of which was constructed in §**3.8**, in which he focused attention on the last 100 years of observations from 1770 to 1869. The correlogram for $0 \leq k \leq 15$ is shown in Figure 7.2, where it is observed that $\bar{r}_1 \approx 0.6$ and all following serial coefficients lie in the interval $-0.16 < \bar{r}_k < 0.13$, i.e., they are all rather close to zero, allowing Wold (*ibid.*, pages 151–2) to conclude that '(t)o my eye, the correlogram definitely suggests a scheme of moving averages', leading him to set out the following problem:

> A set of numbers $u_1, u_2, \ldots, u_h$ being given, does there exist a moving average [7.32] with autocorrelation coefficients r_k such that $r_k = u_k$ for $1 \leq k \leq h$? If the answer is in the affirmative, we know from [§**7.21**] that there in general will exist a finite group of moving averages with the prescribed autocorrelation coefficients, and we are also in possession of a direct method for determining the coefficients (b) of these moving averages. (*ibid.*, page 152)

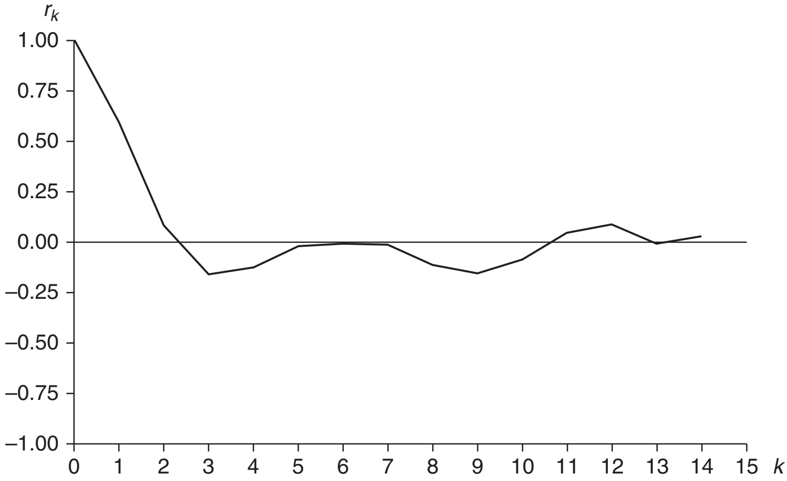


Figure 7.2 Correlogram of Beveridge's Index of Fluctuation, 1770–1869

Under these conditions (7.39) becomes

$$\begin{aligned} u(x) &= u_h x^h + u_{h-1}x^{h-1} + \cdots + u_1 x + 1 + \frac{u_1}{x} + \cdots + \frac{u_{h-1}}{x^{h-1}} + \frac{u_h}{x^h} \\ &= \frac{1}{K^2}(x^h + b_1 x^{h-1} + \cdots + b_{h-1}x + b_h)\left(b_h + \frac{b_{h-1}}{x} + \cdots + \frac{b_1}{x^{h-1}} + \frac{1}{x^h}\right) \end{aligned} \tag{7.49}$$

If x_0 is a root of $u(x)=0$ then so will be x_0^{-1}. Thus the substitution $z=x+x^{-1}$ will transform $u(x)$ to

$$v(z) = v_0 z^h + v_1 z^{h-1} + \cdots + v_{h-1}z + v_h \tag{7.50}$$

For example, with $h=3$, (7.49) becomes

$$\begin{aligned} u(x) &= u_3(x^3 + x^{-3}) + u_2(x^2 + x^{-2}) + u_1(x + x^{-1}) + 1 \\ &= u_3(z^3 - 3z) + u_2(z^2 - 2) + u_1 z + 1 \\ &= u_3 z^3 + u_2 z^2 + (u_1 - 3u_3)z + (1 - 2u_2) \end{aligned}$$

If z is a root of $v(z)=0$ then two roots of $u(x)=0$ will be obtained from the equation

$$P(x,z) = z - x - x^{-1} = x^2 - zx + 1 = 0$$

The roots of this equation are given by

$$\frac{z}{2} \pm \sqrt{\frac{z^2}{4} - 1}$$

Since the products of these roots is unity then, unless both roots have modulus unity, one of them must be situated inside, and the other outside, the unit circle.

If z is a complex root of $v(z) = 0$, then another root will be the complex conjugate of z, which we denote z^*. If $P(x, z) = 0$ has the roots x and x^{-1}, then $P(x, z^*) = 0$ will have the roots x^* and $(x^*)^{-1}$, in which case one of the real polynomials $(x - x_1)(x - x_1^*)$ and $(x - (x_1)^{-1})(x - (x_1^*)^{-1})$ must be a factor in the polynomial

$$b(x) = x^h + b_1 x^{h-1} + \cdots + b_{h-1}x + b_h$$

appearing in (7.49).

If z_0 is a real root of $v(z) = 0$, two cases need to be distinguished. If $|z_0| \geq 2$, both x_1 and x_2 must be real and this will correspond to real roots in $b(x) = 0$. If, on the other hand, $|z_0| < 2$ then x_1 and x_2 will be complex conjugates of modulus unity, and both $(x - x_1)$ and $(x - x_2)$ must be contained in $b(x)$. Since one zero of $u(x)$ corresponds to one zero of $v(z)$, this is impossible unless z_0 is a root of even multiplicity of $v(z) = 0$. The following theorem is thus obtained.

Theorem 12.

A necessary and sufficient condition that there exists a moving average [7.33] *with autocorrelation coefficients r_k equalling u_k for $1 \leq k \leq h$ is that the auxiliary polynomial $v(z)$ defined by* [7.50] *has no zero z_0 of odd multiplicity in the real interval $-2 < z_0 < 2$.*

If this condition is satisfied, the sequences (b) sought for will be given by the real polynomials $b(x)$ satisfying (7.49). There will at most be 2^h of these sequences and the polynomials $b(x)$ may be written in the form $(x - x_1)(x - x_2) \ldots (x - x_h)$, where the real or complex quantity x_i is a root of $P(x, z_i) = 0$, where the z_i, $i = 1, \ldots, h$, are the roots of $v(z) = 0$.

Thus, returning to the correlogram of the Index of Fluctuation, Wold assumed that the small deviations from zero of r_k for $k > 1$ were merely the product of chance fluctuations and asked whether there existed a moving average $\eta(t) + b_1\eta(t-1)$ with autocorrelation coefficient r_1 equalling 0.595.[6] Putting $h = 1$ and $u_1 = 0.595$ into (7.49) obtains $u(x) = 0.595x + 1 + 0.595x^{-1}$ and $v(z) = 0.595z + 1$. Since the root $-0.595^{-1} = -1.68$ of $v(z) = 0$ lies in the critical interval $-2 < z < 2$, it must be concluded from Theorem 12 that there exists no moving average with $r_1 = 0.595$ and $r_k = 0$ for $k > 1$.

For z to lie outside the critical interval $-2 < z < 2$ it must therefore be the case that $-0.5 \leq r_1 \leq 0.5$: i.e., all moving averages of the form $\eta(t) + b_1\eta(t-1)$

have $|r_1| \leq 0.5$ and there will only be one moving average of this form for which $r_1 = 0.5$, namely

$$\zeta(t) - \mu = \eta(t) + \eta(t-1)$$

Consequently, this moving average will yield the closest fit to the prescribed value of 0.595, in which case the deviations of the serial correlation coefficients shown in Figure 7.2 from the values $r_1 = 0.5$, $r_2 = r_3 = \cdots = 0$ must be ascribed to pure chance.

To obtain a better fit, higher-order moving averages must be considered. Using the first two serial correlations, $u_1 = 0.595$ and $u_2 = 0.081$, gives

$$u(x) = 0.081x^2 + 0.595x + 1 + 0.595x^{-1} + 0.081x^{-2}$$

and

$$v(z) = 0.081z^2 + 0.595z + 0.838$$

Since the roots of $v(z) = 0$ are $z_1 = -1.90$ and $z_2 = -5.45$, it follows from Theorem 12 that no moving average of order 2 exists with these autocorrelation coefficients. To remove z_1 from the critical interval, u_2 will need to be modified. The general expression for $v(z_1)$ is

$$v(z_1) = u_2 z_1^2 + u_1 z_1 + (1 - 2u_2)$$

Putting $z_1 = -2$ into $v(z_1) = 0$ along with $u_1 = r_1 = 0.595$ yields the solution $u_2 = r_1 - \frac{1}{2} = 0.095$ with corresponding function $0.095v^2 + 0.595z + 0.810 = 0$, from which $z_1 = -2.0$ and $z_2 = -4.263$.

We next solve $P(x, 2) = x^2 + 2x + 1 = 0$, which gives the double root $x = -1$, and $P(x, -4.263) = x^2 + 4.263x + 1 = 0$, which gives the real roots $x = -0.2491$ and $x = -4.0139$. It then follows that there exist two functions which satisfy the conditions

$$\begin{aligned} b_1(x) &= (x+1)(x+0.2491) = x^2 + 1.2491x + 0.2491 \\ b_2(x) &= (x+1)(x+4.0139) = x^2 + 5.0139x + 4.0139 \end{aligned}$$

The function $b_1(x)$ gives rise to the regular moving average

$$\zeta_1(t) - \mu = \eta(t) + 1.2491\eta(t-1) + 0.2491\eta(t-2)$$

while the function $b_2(x)$ yields

$$\begin{aligned} \zeta_2(t) - \mu &= \frac{K_1}{K_2}(\eta(t) + 5.0139\eta(t-1) + 4.0139\eta(t-2)) \\ &= 0.2491\eta(t) + 1.2491\eta(t-1) + \eta(t-2) \end{aligned}$$

on using $K_1^2 = 1 + 1.2491^2 + 0.2491^2$ and $K_2^2 = 1 + 5.0139^2 + 4.0139^2$. Alternatively, because of the symmetry of the two roots, $\zeta_2(t)$ may be written down immediately once $\zeta_1(t)$ has been obtained.

Wold argued, after using a second example in which u_2 was further adjusted to make the roots z_1 and z_2 coincide, that even small changes in autocorrelations would lead to substantial alterations in the values taken by the moving average coefficients.

A further example was considered in detail by Wold. The values $u_1 = 0.60$, $u_2 = 0.09$, $u_3 = -0.15$ and $u_4 = -0.10$ closely approximate the first four serial coefficients, which we estimate as 0.595, 0.081, –0.161 and –0.126 respectively. These values yield

$$-10^3 v(z) = 10z^4 + 15z^3 - 49z^2 - 105z - 62$$

and on solving $v(z) = 0$ we obtain

$$z_1 = -2.1272 \quad z_2 = 2.5103 \quad z_3, z_4 = -0.9415 \pm 0.5240i$$

From Theorem 12 there will therefore exist a group of moving averages with the prescribed correlogram, and this group will consist of eight processes. Solving $P(x, z_i) = 0$, $i = 1, \ldots, 4$, gives the following solutions

$$\begin{aligned} x_{11} &= -0.7013 & x_{12} &= -1.4259 \\ x_{21} &= 0.4966 & x_{22} &= 2.0137 \\ x_{31} &= -0.3381 - 0.6679i & x_{32} &= -0.6034 + 1.1919i \\ x_{41} &= -0.3381 + 0.6679i & x_{42} &= -0.6034 - 1.1919i \end{aligned}$$

Writing

$$B(x) = (x + 0.3381 - 0.6679i)(x + 0.3381 + 0.6679i) = x^2 + 0.6762x + 0.5604$$

the regular moving average will be obtained from

$$\begin{aligned} b(x) &= (x + 0.7013)(x - 0.4966) \cdot B(x) \\ &= x^4 + 0.8809x^3 + 0.3505x^2 - 0.1208x - 0.1952 \end{aligned}$$

i.e., as

$$\eta(t) + 0.8809\eta(t-1) + 0.3505\eta(t-2) - 0.1208\eta(t-3) - 0.1952\eta(t-4)$$

with $K^2 = 1.9515$. A second moving average with the same correlogram will be delivered by

$$\begin{aligned} b_1(x) &= (x + 1.4259)(x - 0.4966) \cdot B(x) \\ &= x^4 + 1.6055x^3 + 0.4807x^2 + 0.0420x - 0.3968 \end{aligned}$$

with $K_1^2 = 3.9679$. Multiplying $b_1(x)$ by $K/K_1 = 0.7013$ yields the second moving average

$$\eta(t) + 1.1259\eta(t-1) + 0.3371\eta(t-2) + 0.0294\eta(t-3) - 0.2783\eta(t-4)$$

Proceeding in analogous fashion, the third and fourth moving averages are obtained from

$$b_2(x) = (x + 0.7013)(x - 2.0137) \cdot B(x)$$

and

$$b_3(x) = (x + 1.4259)(x - 2.0137) \cdot B(x)$$

to yield

$$\eta(t) - 0.3159\eta(t-1) - 0.8637\eta(t-2) - 0.8395\eta(t-3) - 0.3930\eta(t-4)$$

and

$$\eta(t) + 0.0308\eta(t-1) - 0.9432\eta(t-2) - 0.7909\eta(t-3) - 0.5604\eta(t-4)$$

The four remaining moving averages correspond to the complex roots $x = -0.6034 \pm 1.1919i$ of $u(x) = 0$. Due to symmetry, these processes can be obtained directly from the four processes above by reversing the order of the coefficients. For example, the regular moving average gives

$$-0.1952\eta(t) - 0.1208\eta(t-1) + 0.3505\eta(t-2) + 0.8809\eta(t-3) + \eta(t-4)$$

7.26 If $\{\zeta(t)\}$ is a regular moving average then the primary process $\{\eta(t)\}$ will be given either by (7.19) or by its limiting counterpart of Example 3 of §**7.21** when the characteristic equation has a root with modulus unity. Thus, consider the regular moving average

$$\zeta(t) = \eta(t) + 0.8809\eta(t-1) + 0.3505\eta(t-2) - 0.1208\eta(t-3) - 0.1952\eta(t-4)$$

Using the system (7.13) obtains $a_1 = -0.8809$, $a_2 = 0.4255$, $a_3 = 0.0548$ and $a_4 = 0.1043$, after which the (a) coefficients are given by the difference relation

$$a_k = -0.8809a_{k-1} - 0.3505a_{k-2} + 0.1208a_{k-3} + 0.1952a_{k-4} \quad k > 4$$

The primary process associated with a non-regular moving average may be obtained in a similar fashion as outlined by Wold (*ibid.*, pages 160–2). The moving averages in a group will, by construction, present the same correlogram and same variance, so that the autocorrelation properties of the corresponding series $\overline{\eta}_t$ will provide no basis for deciding which of the moving averages should be preferred.

In terms of forecasting, Wold argued that the forecast for which the expected squared deviation from the future path of $\{\zeta(t)\}$ was minimized is given by

$$F_t[\zeta(t+k)] = b_h\eta_t^{(i)} + b_{k+1}\eta_{t-1}^{(i)} + \cdots + b_h\eta_{t-h+k}^{(i)}$$

where $\{\eta_t^{(i)}\}$ is the primary process constructed from the *i*th non-regular moving average in the group and the sequence (b) is that for the regular moving average. In other words, the different moving averages in a group will give rise to the same sequence of optimal forecasts and thus in general

$$F_t[\zeta(t+k)] = \overline{\mu} + b_h\overline{\eta}_t + b_{k+1}\overline{\eta}_{t-1} + \cdots + b_h\overline{\eta}_{t-h+k}$$

where $\overline{\mu}$ is the sample average of $\overline{\zeta}_t$. From this formula it is seen that forecasts beyond the next h observations reduce to the sample average of the data. The squared deviation of errors of these forecasts are given by

$$(1 + b_1^2 + b_2^2 + \cdots + b_{h-1}^2)\sigma^2(\eta)$$

so that the efficiency of the forecasts decreases gradually as the number of periods being forecasted is extended, leading Wold to the opinion that

> especially in view of economic time series, the type of forecast delivered by the scheme of moving averages seems *a priori* more realistic, seems to correspond better to what might be reasonably possible to find out from the past development. Further, considering the forecasts over a short period, the prognosis given by the scheme of moving averages is, as a rule, rather efficient. In my opinion, this is a circumstance of central importance, for often the main interest is concentrated upon the prognosis concerning the near future. (*ibid.*, page 168)

An application of linear autoregression

7.27 Wold's second major application was to consider the Swedish cost of living index between 1840 and 1913 after he had removed a trend in the data

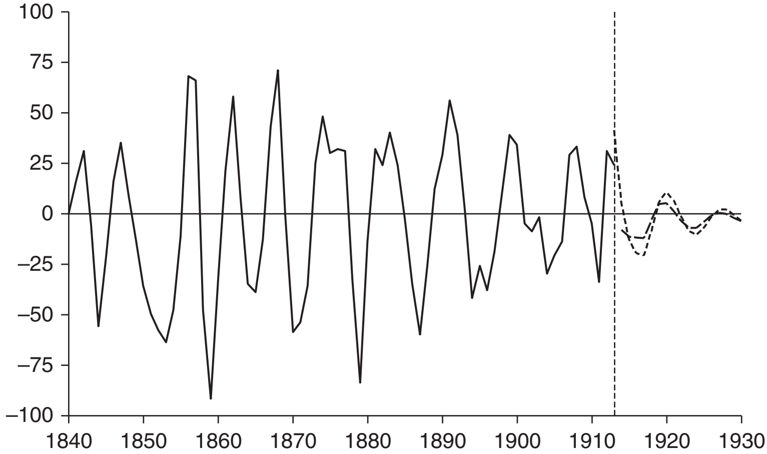


Figure 7.3 Swedish Cost of Living Index, 1840–1913, with forecasts out to 1930

to induce stationarity. This index is shown in Figure 7.3 and 'is seen to reflect clearly changes between economic expansion and contraction. A certain regularity seems to be present in the movement up and down, but the distance between two adjacent maxima is rather inconstant, varying between some 5 and 10 years' (*ibid.*, page 177). The correlogram is shown in Figure 7.4:

> The correlogram looks rather like a simple damped oscillation, say $C \cdot q^k \cdot \cos(\lambda k + \varphi)$. An inspection of the graph shows that in approximating the correlogram by such a function we would have to take the period $p = 2\pi/\lambda$ to be about 7 or 8 years, the phase φ to be approximately vanishing, and $q^7 \sim 1/2$, the latter relation corresponding to a damping of some 50% in the duration of one period. (*ibid.*, page 177)

This led Wold to consider a linear autoregression of order two, since 'this will present a correlogram forming a simple damped harmonic' (*ibid.*, page 177):

$$\zeta(t) + a_1\zeta(t-1) + a_2\zeta(t-2) = \eta(t) \tag{7.51}$$

Wold preferred such a process to a scheme of hidden periodicities since the latter model, because each harmonic component will produce an undamped harmonic in the correlogram, would require at least two superposed harmonics to adequately represent its shape. Such a scheme would therefore involve at least six parameters rather than just the two required by (7.51).

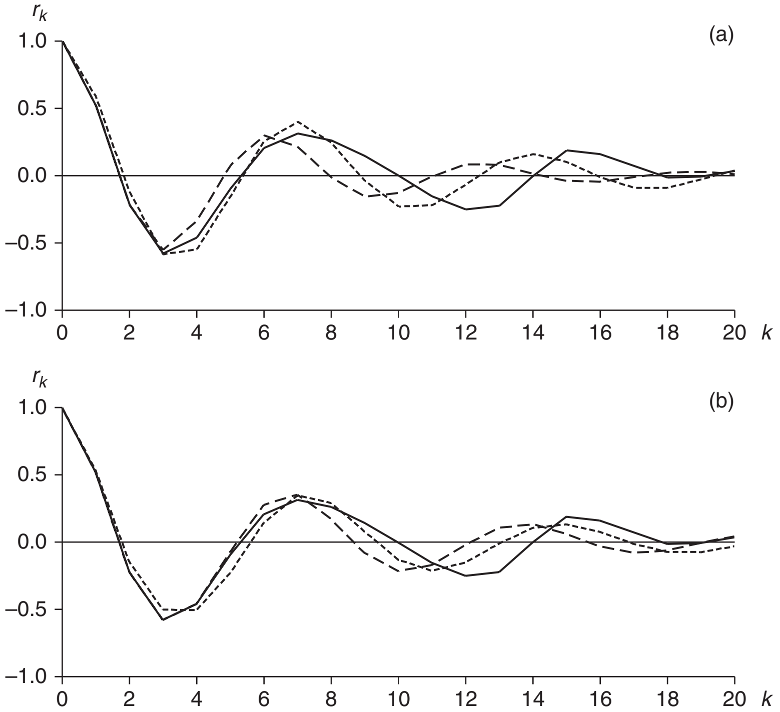


Figure 7.4 Correlogram of the cost of living index (unbroken line) with the hypothetical correlograms from equations (7.54) (dashed line) and (7.55) (dotted line) in panel (a), and from equations (7.56) (dashed line) and (7.57) (dotted line) in panel (b)

7.28 In §**7.18**, the system of equations (7.26) with coefficients $a_1, \ldots, a_h$ will deliver the autocorrelation coefficients $r_1, \ldots, r_{h-1}$ required for deriving the following coefficients $r_h, r_{h+1}, \ldots$ from the difference relations (7.25). In searching for an adequate autoregressive process of the form (7.51), the 'inverse problem' has to be confronted, i.e., that of finding a set of coefficients $a_1, \ldots, a_h$ given a set of serial coefficients. Wold thus suggested replacing the r_k with the corresponding serial coefficients $\bar{r}_k$ in the system (7.26) and the last relation in (7.25) and solving the following system of equations for a 'trial' set of (a) coefficients:

$$\begin{cases} \bar{r}_1 + a_1 + a_2\bar{r}_1 + a_3\bar{r}_2 + \cdots + a_h\bar{r}_{h-1} = 0 \\ \bar{r}_2 + a_1\bar{r}_1 + a_2 + a_3\bar{r}_1 + \cdots + a_h\bar{r}_{h-2} = 0 \\ \cdots \\ \bar{r}_h + a_1\bar{r}_{h-1} + a_2\bar{r}_{h-2} + \cdots + a_h = 0 \end{cases} \tag{7.52}$$

If the roots of the characteristic equation associated with (a) lie in the unit circle, these coefficients will define a linear autoregression of the form (7.24).

By construction, the first h autocorrelation coefficients of this process will coincide with the serial coefficients $\overline{r}_1, \ldots, \overline{r}_h$ and the subsequent coefficients can be obtained using the difference relations (7.25) and (a).

Having thus derived the correlogram of the hypothetical process defined by (a), this can then be compared with the empirical correlogram. If the fit appears satisfactory then the analysis can be carried further using (a), but if the deviations between the hypothetical and empirical correlograms are deemed to be too large, some adjustment of the (a) coefficients would then be required.

Given the h coefficients (a), the primary series $\overline{\eta}_t$ may be constructed from the observed values $\overline{\zeta}_t$ as

$$\overline{\eta}_t = \overline{\zeta}_t - \overline{\mu} + a_1(\overline{\zeta}_{t-1} - \overline{\mu}) + \cdots + a_h(\overline{\zeta}_{t-h} - \overline{\mu})$$

The method employed by Yule (1927) to obtain the (a) coefficients was least squares (see §**6.3**), which chooses (a) to minimize $\sum \overline{\eta}^2$. In fact, this approach closely approximates solving the system (7.52). In this case,

$$\sigma^2(\overline{\eta}) \approx (1 + a_1 r_1 + a_2 r_2 + \cdots + a_h r_h) \cdot \sigma^2(\overline{\zeta}) \tag{7.53}$$

where the $\approx$ sign conveys the approximation produced by having to discard the first h terms of $\overline{\zeta}_t$ for which the corresponding values of $\overline{\eta}_t$ cannot be calculated. In other words, the variance of $\overline{\eta}_t$ will approximate the hypothetical variance $\sigma^2(\eta)$, although this will not be the case if the trial set (a) is determined otherwise than by (7.52).

Consequently, (7.52) gives a set of coefficients $a_1, \ldots, a_h$ that minimize the variance of the residuals $\overline{\eta}_t$. The first h autocorrelation coefficients will coincide with the corresponding serial coefficients but there is no guarantee that the complete hypothetical correlogram will provide a good fit to the empirical correlogram throughout its whole range. In practice, a compromise needs to be met 'between the two desiderata of obtaining small residuals $\overline{\eta}_t$ and small deviations between the correlograms, and besides try to satisfy the relation $\sigma^2(\eta) \sim \sigma^2(\overline{\eta})$.

7.29 Applying this approach to the cost of living index, Wold used the values $\overline{r}_1 = 0.5216$ and $\overline{r}_2 = -0.2240$ so that, with $h = 2$, the system (7.52) becomes[7]

$$\begin{aligned} 0.5216 + a_1 + 0.5216 a_2 &= 0 \\ -0.2240 + 0.5216 a_1 + a_2 &= 0 \end{aligned}$$

with the solution $a_1 = -0.8771$ and $a_2 = 0.6815$. The roots of the characteristic equation $z^2 + a_1 z + a_2 = 0$ are $0.4385 \pm 0.6994i$ and are thus less than unity in modulus, so that the relation

$$\zeta(t) - 0.8771\zeta(t-1) + 0.6815\zeta(t-2) = \eta(t) \tag{7.54}$$

defines a process of linear autoregression. By construction, the first two autocorrelations of the process $\{\zeta(t)\}$ will be $r_1 = \overline{r}_1 = 0.5216$ and $r_2 = \overline{r}_2 = -0.2240$, with subsequent autocorrelations being obtained recursively from the difference relation $r_k - 0.8871 r_{k-1} + 0.6815 r_{k-2} = 0$, $k \geq 3$. The resulting correlogram is also shown in Figure 7.4(a), prompting Wold to argue that

> (c)omparing with the empirical correlogram, it is seen that the period in the hypothetical correlogram is too short, and that the damping is a little too heavy. ... (T)he damping factor equals $\sqrt{a_2}$, while the period is given by $p = 2\pi/\lambda$, where $\cos\lambda = -a_1/2\sqrt{a_2}$. Thus, an increase in a_2 will bring on a slighter damping. Further, reducing λ we obtain a longer period. However, ... we cannot conclude without further evidence that it will be possible to improve the fit – the coefficients a_1 and a_2 determine also the constant factor and the phase of the damped harmonic, and it might happen that an adjustment in a_1 and a_2 would cause such a change, e.g. in the phase, that the total result of the adjustment would be a poorer fit. (*ibid.*, page 180)

The hypothetical correlogram of (7.54) has a period of 6.22 years. To achieve a period close to seven years with reduced damping, Wold adjusted the coefficients to $a_1 = -1.10$ and $a_2 = 0.77$. The correlogram of the process defined by

$$\zeta(t) - 1.10\zeta(t-1) + 0.77\zeta(t-2) = \eta(t) \tag{7.55}$$

is also shown in Figure 7.4(a): ‘up to r_8 and r_9, the hypothetical correlogram seems to fit rather well. Beyond this point, the fit is less satisfactory, partly because the graph of the serial coefficients presents a slow descent to the minimum in $k \sim 12.5$, and a rapid rise to the next maximum’ (*ibid.*, page 181).

Substituting the appropriate values from the model (7.54) into (7.53) obtains $\sigma^2(\overline{\eta}) \sim \sigma^2(\eta) = 0.390\sigma^2(\zeta)$. Wold showed that (7.55) led to a larger residual variance, leading him to conclude that ‘all in all, neither of the schemes seems adequate ... It seems as if we cannot find a satisfactory approach without taking into account more distant elements $\overline{\zeta}_{t-3}$, $\overline{\zeta}_{t-4}$, etc.’ (*ibid.*, page 182). Wold thus extended the model by taking $h = 4$, arriving at the process

$$\zeta(t) - 0.8100\zeta(t-1) + 0.7452\zeta(t-2) - 0.0987\zeta(t-3) + 0.2101\zeta(t-4) = \eta(t) \tag{7.56}$$

The correlogram of this process is shown in Figure 7.4(b) and is seen to be almost identical to that from (7.55) although here, of course, $r_k = \overline{r}_k$ for $k \leq 4$. The roots of (7.56) are $0.5385 \pm 0.6814i$ and $-0.1335 \pm 0.5106i$, so that two of the roots are reasonably close to those of (7.55). Wold found that an improved fit to the empirical correlogram was obtained by adjusting the roots to $0.5888 \pm 0.6540i$

and $-0.20 \pm 0.58i$, leading to

$$\zeta(t) - 0.7776\zeta(t-1) + 0.6797\zeta(t-2) - 0.1342\zeta(t-3) + 0.2914\zeta(t-4) = \eta(t) \tag{7.57}$$

This is also shown in Figure 7.4(b) and Wold regarded the general shape as being 'rather satisfactory'. For (7.56), the relation (7.53) gives $\sigma^2(\eta) = 0.371\sigma^2(\zeta)$, which represents a slight increase in efficiency over (7.54) as compensation for introducing two further parameters.

7.30 Forecasts of the cost of living index may be calculated using equation (7.22). For a model of the type (7.57), these forecasts are built up as

$$\begin{aligned} F_t[\zeta(t+1)] &= (1 + a_1 + a_2 + a_3 + a_4)\overline{\mu} - a_1\overline{\zeta}_t - a_2\overline{\zeta}_{t-1} - a_3\overline{\zeta}_{t-2} - a_4\overline{\zeta}_{t-3} \\ F_t[\zeta(t+2)] &= (1 + a_1 + a_2 + a_3 + a_4)\overline{\mu} - a_1 F_t[\zeta(t+1)] - a_2\overline{\zeta}_t - a_3\overline{\zeta}_{t-1} - a_4\overline{\zeta}_{t-2} \end{aligned}$$

and so on. These forecasts, calculated as $F_{1912}[\zeta(1913)] = 41.4$, $F_{1912}[\zeta(1914)] = 5.2$, etc., are shown in Figure 7.3 up to 1930, i.e., for k up to 18. Also shown are the set of forecasts $F_{1913}[\zeta(1914)]$, etc., up to 1930.

> The two forecasts curves in [Figure 7.3] yield a good illustration of the prognosis situation in an approach of linear autoregression. Firstly, while a forecast $F_t[\zeta(t+k)]$ is often rather efficient for small k-values, the efficiency vanishes asymptotically as k increases. Further, as soon as we are in a position to take a new observation $\overline{\zeta}_{t+1}$ into consideration when forming the prognosis, the new forecast curve is often substantially modified; how much, will depend on the residual $\overline{\eta}_{t+1} = \overline{\zeta}_{t+1} - F_t[(t+1)]$. – Summing up, it is the short forecasts that are efficient. In this respect, we meet the same situation as in the scheme of moving averages, and the same contrast to the scheme of hidden periodicities. On the other hand, under special circumstances the oscillations in a scheme of linear regression are nearly functional, viz. nearly strictly periodic – as remarked in discussing the sinusoidal limit theorem ..., processes of hidden periodicities can be obtained as limit cases of the schemes of linear autoregression. (*ibid.*, page 187)

7.31 Wold finally considered how linear autoregressions were formed by the *complete systems* analysed in economics by Frisch (1933) and Tinbergen (1937). A simple example of a complete system is given by

$$\begin{aligned} \xi(t) &= c_1\zeta(t-1) + \eta'(t) \\ \zeta(t) &= d_0\xi(t) + d_1\xi(t-1) + \eta''(t) \end{aligned}$$

Such a system may be rewritten as

$$\zeta(t) = d_0 c_1 \zeta(t-1) + d_1 c_1 \zeta(t-2) + (d_0 + d_1)\eta'(t) + \eta''(t)$$

i.e., as (7.51) with $a_i = -d_i c_1$, $i = 0, 1$, and $\eta_t = (d_0 + d_1)\eta'(t) + \eta''(t)$.

7.32 Wold's monograph was rightly hailed as a major contribution to the foundations of time series analysis, fusing together the intuitive autoregressive and moving average models of Yule, Slutzky and Walker, developed in response to observing physical and economic phenomena, with the advances in probability theory made by the Russian mathematicians Kolmogorov and Kinchine. But there is more to Wold's contribution, for he also introduced the first formal concepts in the theory of forecasting and made suggestions as to how these models may be arrived at by examining the contrast between empirical and hypothetical correlograms.

Wold, however, was acutely aware of the limitations of his framework, most notably in the absence of an inferential framework to bring to bear on the model selection process – a subject that, unsurprisingly, would quickly engage the attention of the new breed of mathematical statisticians encouraged by the work of Wold to research in the area of time series. Wold was also concerned with two other problems that had been avoided by focusing attention just on stationary time series – the necessity for detrending an 'evolutive' series before this modelling framework could be employed and the possibility that observed time series might be generated by a nonlinear process. Again, these were to become major research agendas in subsequent developments in time series modelling.

8

Generalizations and Extensions of Stationary Autoregressive Models: From Kendall to Box and Jenkins

Oscillatory autoregressions

8.1 After being introduced by Yule and Walker and having its theoretical foundations established by Wold, the autoregressive model was further developed in a trio of papers written during the Second World War by Maurice Kendall (1943, 1944, 1945a).[1] Because of his interest in what seemed to be systematic fluctuations in a wide variety of agricultural time series, Kendall's focus was on 'oscillatory' time series generated by the second-order autoregressive process studied by Yule (1927)[2]

$$x_t + ax_{t-1} + bx_{t-2} = \varepsilon_t \tag{8.1}$$

for which the roots of the characteristic equation $z^2 + az + b = 0$ are assumed to be the complex conjugates $\alpha \pm i\beta$. The complementary function of (8.1) is then

$$p^t (A \cos \theta\, t + B \sin \theta\, t) \tag{8.2}$$

where $p = +\sqrt{b}$,

$$\theta = \tan^{-1} \frac{\beta}{\alpha} = \tan^{-1} \sqrt{\left(\frac{4b}{a^2} - 1\right)} = \cos^{-1}\left(\frac{-a}{2\sqrt{b}}\right)$$

and A and B are arbitrary constants (cf. §**6.6**). Assuming $b > 0$, $0 < p < 1$, and $4b > a^2$, the complementary function (8.2) represents a damped harmonic with a fundamental period of $2\pi/\theta$. If ξ_t is a particular value of (8.2) such that $\xi_0 = 0$ and $\xi_1 = 1$, so that $A = 0$, $B = 1/p \sin\theta$, and

$$\begin{aligned}\xi_t &= p^t B \sin \theta\, t = p^t \sin \theta\, t / p \sin \theta = p^t \sin \theta\, t / p \tan \theta \cos \theta \\ &= \frac{2}{\sqrt{(4p^2 - a^2)}} p^t \sin \theta\, t\end{aligned}$$

then a particular integral of (8.2) is $\sum_{j=0}^{\infty} \xi_j \varepsilon_{t-j+1}$ and the complete solution becomes

$$x_t = p^t (A \cos \theta\, t + B \sin \theta\, t) + \sum_{j=0}^{\infty} \xi_j \varepsilon_{t-j+1}$$

If the series was 'started up' some time ago, so that the complementary function has been damped out of existence, then this solution is just

$$x_t = \sum_{j=0}^{\infty} \xi_j \varepsilon_{t-j+1}$$

which is a moving sum of a random series with damped harmonic weights. For a long series, Kendall showed that the autocorrelations were given by

$$\rho_k = \frac{p^k \sin (k\theta + \psi)}{\sin \psi} \qquad \tan \psi = \frac{1+p^2}{1-p^2} \tan \theta$$

so that, apart from a constant factor, ρ_k is given by the product of the damping factor p^k and a harmonic term which has the fundamental period of the generating equation (8.1). Using the systems of equations in §**7.18**, it is easily shown that

$$\rho_1 = -\frac{a}{(1+b)} \qquad \rho_2 = \frac{a^2 - b(1+b)}{1+b}$$

with subsequent autocorrelations being computed using the recursion

$$\rho_k + a\rho_{k-1} + b\rho_{k-2} = 0$$

Focusing on the oscillatory characteristics of both the generated series x_t and its correlogram, Kendall pointed out that, although $\rho_0 = 1$ will always be a peak at the beginning of the correlogram, the presence of the phase angle ψ implies that the interval from $k = 0$ to the next maximum of the correlogram will not be equal to the fundamental period $2\pi/\theta = 2\pi/\cos^{-1}(-a/2\sqrt{b})$. Consequently, Kendall preferred to judge the length of the period by measuring from upcross to upcross (i.e., values of k at which the correlogram turns from negative to positive) or from trough to trough of the correlogram – if peaks are to be preferred, then the peak at $k = 0$ should not be counted. On the assumption that the ε_t are normal, Kendall (1945a, Appendix) showed that the mean distance (m.d.) between upcrosses was

$$\text{m.d. (upcrosses)} = \frac{2\pi}{\cos^{-1} \rho_1} = \frac{2\pi}{\cos^{-1}(-a/(1+b))}$$

while the mean distance between peaks was

$$\text{m.d. (peaks)} = \frac{2\pi}{\cos^{-1}\tau_1}, \quad \tau_1 = \frac{-1 + 2\rho_1 - \rho_2}{2(1-\rho_1)} = \frac{b^2 - (1+a)^2}{2(1+a+b)}$$

The relationship between the variances of the random error ε_t and the generated series x_t, denoted σ_ε^2 and σ_x^2 respectively, is easily shown to be

$$\frac{\sigma_\varepsilon^2}{\sigma_x^2} = \frac{1-b}{1+b}((1+b)^2 - a^2) \tag{8.3}$$

a result that will be found useful in §**8.8**.

8.2 Kendall illustrated these properties of an oscillatory autoregressive process by generating 480 observations from the model (8.1) with $a = -1.1$ and $b = 0.5$, i.e.,

$$x_t = 1.1x_{t-1} - 0.5x_{t-2} + \varepsilon_t \tag{8.4}$$

The error process was assumed to be an integer rectangular random variable ranging from −49 to +49. The observations on this variable, termed Series I, are listed in Kendall (1945a, Table 2) and are plotted as Figure 8.1. 'Evidently systematic movements are present although they are obscured to some extent by the random variable. The series is, in fact, highly damped, the damping factor being $\sqrt{0.5} = 0.7071$, so that we should expect the disturbance function to exercise considerable influence on the course of the series' (*ibid.*, page 105).

The frequency distributions of the peak to peak and upcross to upcross intervals are shown in Table 8.1. As $\tau_1 = 0.3$, so that $\cos^{-1}\tau_1 = 72.54°$, the expected mean-distance between peaks is $360/72.54 = 4.96$: the observed mean distance in Series I of 5.05 thus represents an excellent agreement.

The expected mean-distance between upcrosses is $2\pi/\cos^{-1}(0.7333) = 8.40$ compared to an observed value of 8.30. The fundamental period of the generating equation, however, is $2\pi/\theta = 2\pi/\cos^{-1}(-a/2\sqrt{b}) = 9.25$, which is rather longer.

Given these oscillatory properties of Series I, Kendall considered whether a standard periodogram analysis would uncover them. The periodogram calculated by Kendall is shown in Figure 8.2, the top panel for integer values of the period P up to 50, the bottom panel for a finer mesh of periods between 8 and 9. This led him to conclude that

> (t)he results are rather striking. There are about a dozen peaks, two of which, at 20 and 42, stand out as offering substantial evidence of significant periods. In fact there are periods almost everywhere except in the right place, at 8 or 9. (*ibid.*, page 106)

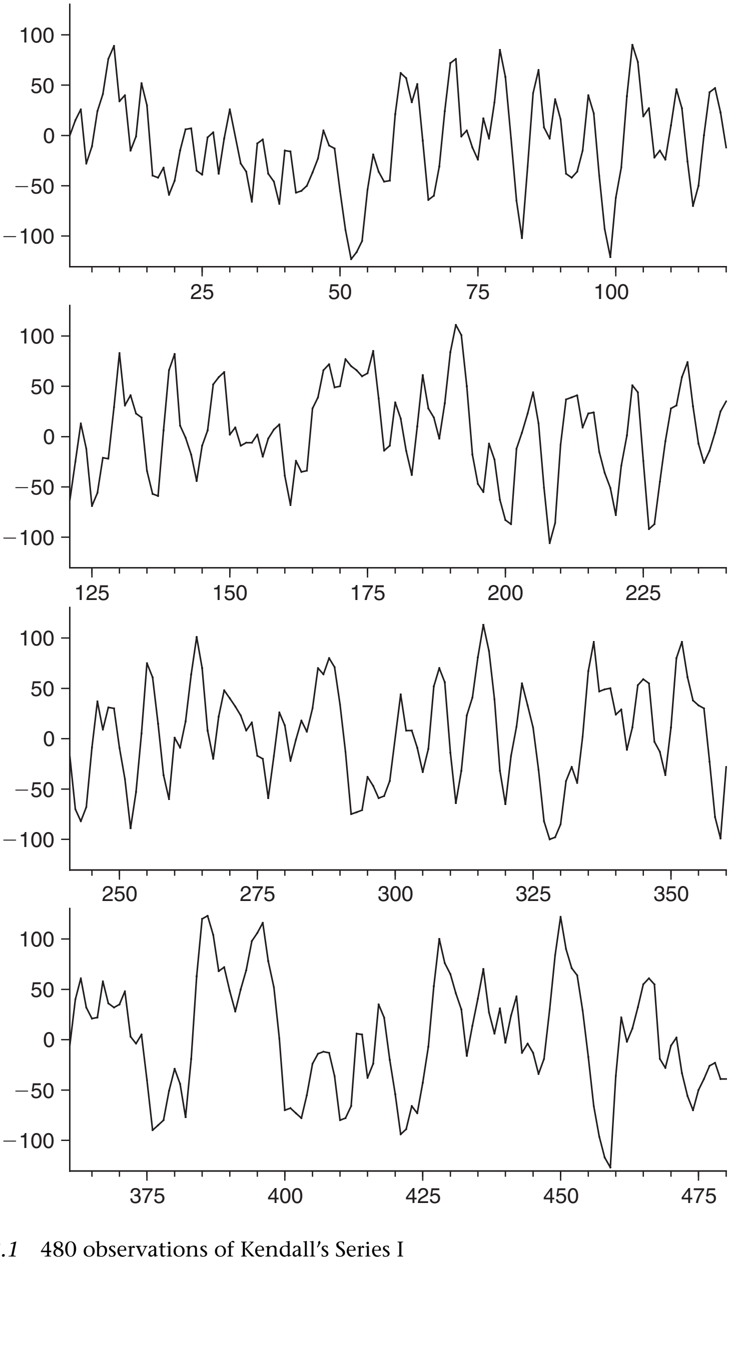


Figure 8.1 480 observations of Kendall's Series I

Table 8.1 Distribution of intervals from peak to peak and upcross to upcross for Series I

Interval (units)	Peak-to-peak Frequency	Upcross-to-upcross frequency
2	10	3
3	17	3
4	14	5
5	13	2
6	14	6
7	13	9
8	5	10
9	4	5
10	1	2
11	2	2
12	–	3
13	–	2
14	–	1
15	–	1
17	–	2
29	–	1
Total	93	57

Kendall compared 'the ambiguous and confusing picture presented by the periodogram' with the correlogram of Series I, shown in Figure 8.3.

> The damped oscillatory effect is now clearly evident, and the only doubt that would occur is that after a point the oscillations do not continue to damp out. This is due to the shortness of the series ... The average interval between troughs of the correlogram is 7.2 (or 8.0 if we ignore the doubtful ripple at 41), moderately close to the mean-distance between upcrosses (but considerably longer, one may remark, than the mean-distance between peaks).
>
> It seems undeniable that so far as this particular series is concerned the correlogram gives much better results than the periodogram. Without prior knowledge of the way in which the series was generated, we should be led by the correlogram to suspect a simple autoregressive scheme. (*ibid.*, page 110)

Indeed, using the observed serial correlations leads to the scheme

$$x_t = 1.132x_{t-1} - 0.486x_{t-2} + \varepsilon_t$$

which is a good approximation to the true generating equation (8.4).

8.3 Kendall (1943, 1944) applied these ideas to several agricultural series for England and Wales. Figure 8.4 shows the annual observations from 1871 to

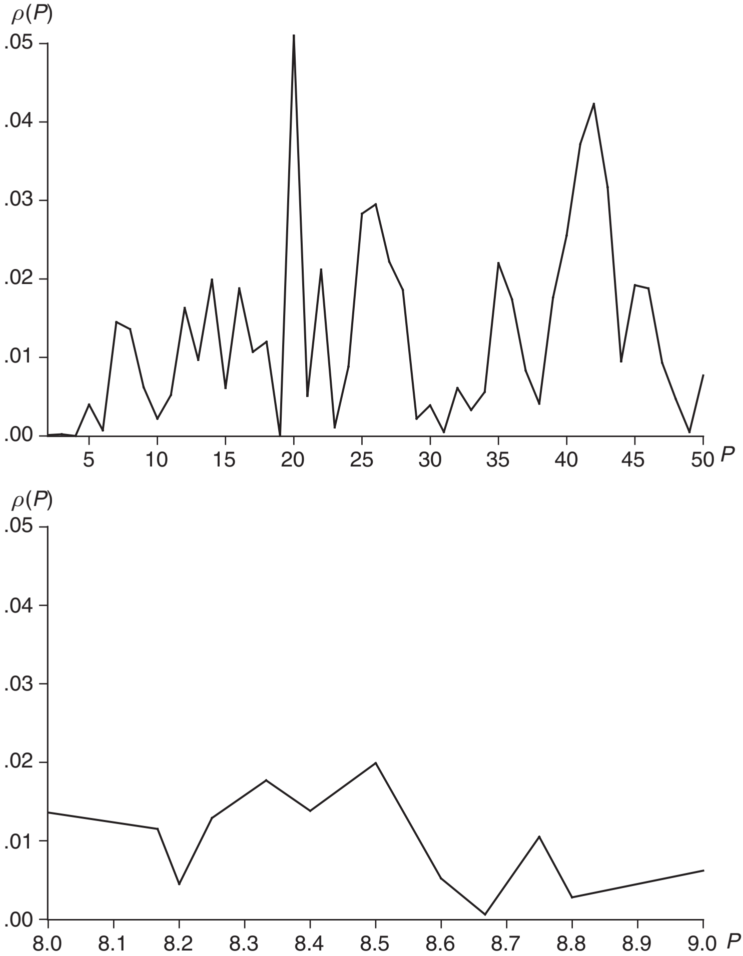


Figure 8.2 Periodogram of Series I

1934/1935 for wheat prices and sheep population taken from Kendall (1943, Table 1), while Figure 8.5 shows their correlograms.[3] Kendall concluded that both show 'real systematic fluctuations' and used, for the first time, concepts of statistical significance to affirm this conclusion.

> Owing to the comparative shortness of the series one has to safeguard against being misled by sampling effects and against seeing more in the diagrams than actually exists. No test is known for the significance of a correlogram. For any *given* serial correlation the theory of large samples may be used to

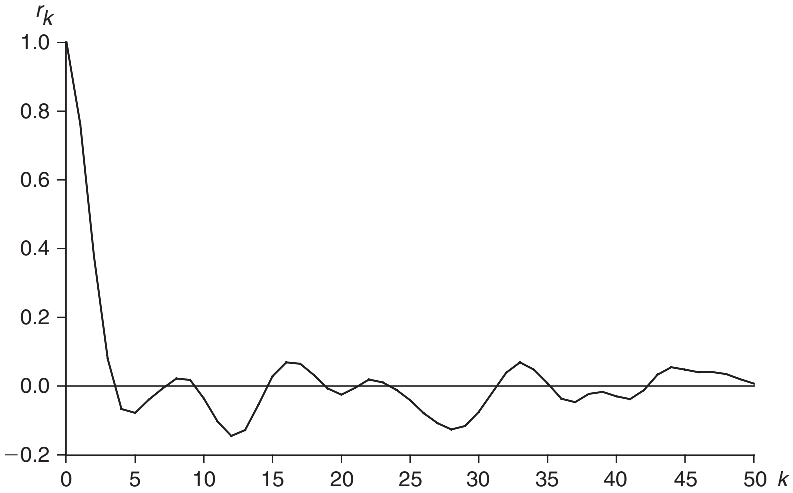


Figure 8.3 Correlogram of Series I

> show that the standard error is approximately $1/\sqrt{n}$, where n is the number of pairs entering into the correlation. To test the hypothesis that correlations are zero we should probably not make a serious misjudgment by using the standard error to obtain probabilities in the normal way – that is, by reference to the normal distribution; but it is not clear that the number of terms used in calculating these particular coefficients (*e.g.*, ... 64 for r_1, 63 for r_2 ... 35 for r_{30}) is large enough to justify the use of large sample theory. However, taking the standard error as $1/\sqrt{n}$, we see that, to the 5 per cent level of probability, a value of 0.25 would be required for r_1 before we could assume its significance, and a value of 0.33 for r_{30}.
>
> This applies for any given coefficient, but it does not help much in deciding whether the undulatory character of the whole set of serial correlations is significant of regular oscillation. However, I do not think that anyone would doubt, after looking at the correlograms ... that the undulations are not accidental.' (Kendall, 1943, pages 102–103; italics in original)[4]

Focusing attention here on the sheep population data, Kendall considered the partial correlations of the series (cf. §**6.8**), the first six being shown in Table 8.2, along with the continued product of $1-r^2$ (as in Table 6.3), concluding that 'it is clear that no appreciable gain in representation is to be obtained by taking the regression on more than two preceding terms' (*ibid.*, page 104). A similar pattern of partial correlations is found for the wheat price series, also shown in Table 8.2.

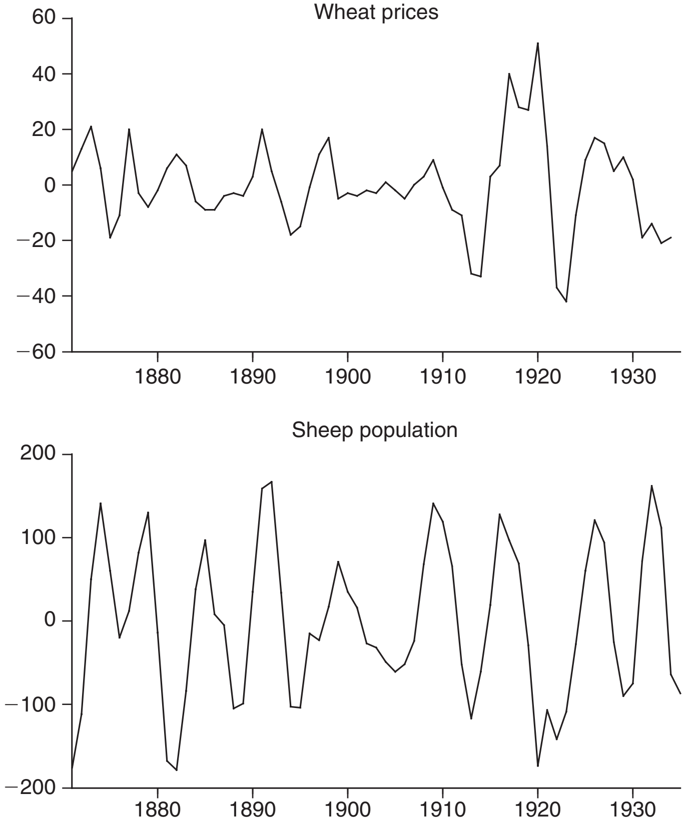


Figure 8.4 Detrended wheat prices and sheep population for England and Wales, 1871–1934/5

The autoregression implied by the correlogram of the sheep population series is

$$x_t = 1.029x_{t-1} - 0.741x_{t-2} + \varepsilon_t$$

Since

$$\tan\theta = \sqrt{\left(\frac{4b}{a^2} - 1\right)} = 1.341, \quad \theta = 53.3^\circ$$

the period is calculated as $360/53.3 = 6.8$ years. In the correlogram there are peaks at $k = 7$, 17 and 25 years (ignoring $k = 0$: see §**8.1**), giving periods of 10 and 8 years with a mean of 9, while there are troughs at $k = 3$, 13, 21 and 28, giving periods of 10, 8 and 7 with a mean of 8.3 years. ‘We therefore conclude

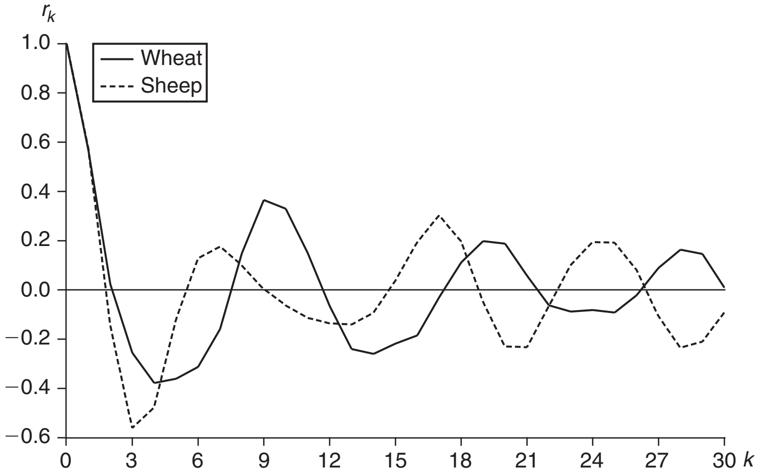


Figure 8.5 Correlograms of wheat prices and sheep population

Table 8.2 Partial correlations of the sheep population and wheat price series

Serial correlations		Partial correlations		$1 - r^2$	Continued product of $1 - r^2$
		(a) Sheep population			
1	0.575	1	0.575	0.669	0.669
2	−0.144	2.1	−0.709	0.497	0.332
3	−0.561	3.12	−0.036	0.999	0.332
4	−0.477	4.123	−0.049	0.998	0.331
5	−0.119	5.1234	−0.089	0.992	0.329
6	0.128	6.12345	−0.209	0.956	0.314
		(b) Wheat prices			
1	0.568	1	0.568	0.677	0.677
2	0.023	2.1	−0.442	0.805	0.545
3	−0.255	3.12	−0.041	0.998	0.544
4	−0.378	4.123	−0.260	0.991	0.539
5	−0.361	5.1234	−0.097	0.995	0.536
6	−0.313	6.12345	−0.271	0.927	0.497

that the real period is between 8 and 9 years, whereas that given by solving the autoregressive equation is much shorter' (*ibid.*, page 107).

Similar calculations for the wheat price series obtains

$$x_t = 0.826x_{t-1} - 0.448x_{t-2} + \varepsilon_t$$

with $\theta = 51.9°$ and a period of 6.9 years. The correlogram has peaks at $k = 9$, 19 and 28 and troughs at $k = 4$, 14 and 25, thus implying a period of around 10 years, again larger than the fundamental period implied by the autoregressive scheme.

8.4 Kendall considered whether this underestimation of the period from the autoregression could be a consequence of an additional *superposed* random element of the type discussed by Yule (1927) (see §**6.1**). If this is denoted η_t and assumed to have variance σ_η^2 and to be independent of the disturbance ε_t, then, if superposed on x_t, it will increase the variance of the observed series from σ_x^2 to $\sigma_x^2 + \sigma_\eta^2$. The autocovariances will not be affected, so that all autocorrelations (except $\rho_0 = 1$) will be reduced by the ratio

$$c = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2} \tag{8.5}$$

To illustrate this effect, Kendall constructed an autoregressive series of 65 terms as

$$u_t = 1.5u_{t-1} - 0.9u_{t-2} + \varepsilon_t \tag{8.6}$$

where the ε_t are rectangular random variables in the range $-49.5(1)\cdots + 49.5$. On to the series so derived were superposed (a) a second rectangular random variable with the same range, and (b) a further rectangular random variable with the range $-199.5(1)\ \cdots\ + 199.5$, the combined variable then being divided by 10 and rounded up to the nearest integer. These constructed series are given in Kendall (1944, Table 5) and their correlograms are shown in Figure 8.6. Kendall showed that, for infinite series, the value of c would be 0.93 for (a) and 0.45 for (b), so that the autocorrelations for the second series should be much smaller than those for the first.

> The correlograms run according to expectation. The effect of the bigger random element is to reduce the amplitude at the beginning of the series and to introduce some minor irregularities in the data, but not to effect substantially the lengths of the correlogram oscillations. (*ibid.*, page 114)

From the equations for ρ_1 and ρ_2 in §**8.1**, the coefficients a and b can be written in terms of the serial correlations r_1 and r_2 as

$$-a = \frac{r_1(1 - r_2)}{1 - r_1^2} \qquad -b = \frac{r_2 - r_1^2}{1 - r_1^2} \tag{8.7}$$

Apart from the fact that r_1 and r_2 may not be reliable estimates of ρ_1 and ρ_2 if the observed series is short, thus imparting sampling error into the estimates of

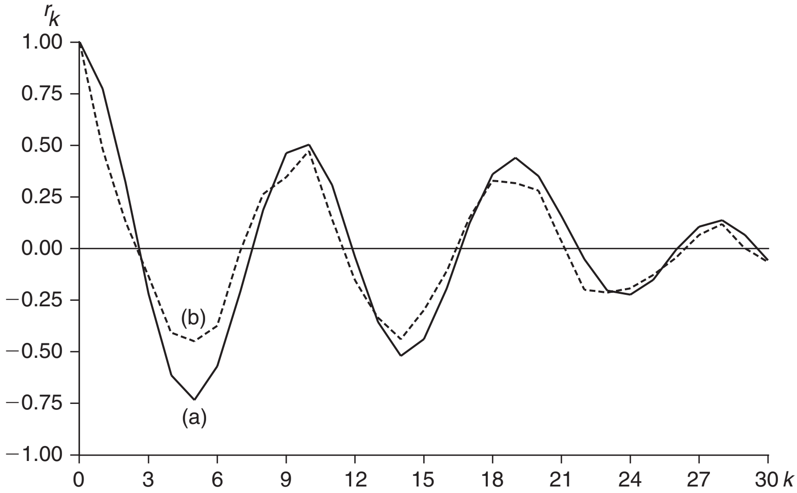


Figure 8.6 Correlograms of two artificial series with (a) a slight superposed variation, and (b) a large superposed variation

a and b, the presence of superposed variation will reduce the serial correlations by a factor c, leading to the estimates

$$-a' = \frac{cr_1(1-cr_2)}{1-cr_1^2} \qquad -b' = \frac{cr_2-cr_1^2}{1-cr_1^2}$$

The estimated fundamental period of the generating equation is then given by

$$4\cos^2\theta' = \frac{a'^2}{b'} = \frac{cr_1^2(1-cr_2)^2}{(1-c^2r_1^2)(r_2-cr_1^2)}$$

which Kendall expanded in powers of $\gamma = 1-c$ to obtain, as a first-order approximation,

$$\frac{a'^2}{b'} = \frac{a^2}{b}\left(1-\gamma\frac{(1+b(3b^2-b-a^2))}{b((1+b)^2-a^2)}\right)$$

Hence, if $3b^2-b-a^2>0$ the effect of a superposed variation (i.e., γ positive) is to make $a'^2/b' < a^2/b$ or, in other words, to result in a shortening of the observed period. The condition $3b^2-b-a^2>0$ is equivalent to

$$b > \tfrac{1}{6}\left(-1+\sqrt{(12a^2+1)}\right)$$

which is not very restrictive since, in any case, $a^2 \leq 4$ and $4b \geq a^2$. Kendall was thus led to

> the interesting conclusion that if there is any superposed random variation present, the period calculated from the observed regression equation according to formulae [8.7] will probably be too short even for long series. Yule himself found too short a period for his sunspot material and, suspecting that it was due to superposed variation, attempted to reduce that variation by graduation [§**6.4**]. The result was a longer period more in accordance with observation. It does not appear, however, that the superposed variation in his case was very big. In a number of agricultural time series which I have examined it is sometimes about half the variation of the series and the effect on the period as calculated from the serial correlations is very serious. For instance, in the cases of wheat prices and sheep population referred to above, formulae [8.7] give periods of 7.0 and 6.8 years, whereas the correlograms indicate periods of about 9.5 and 8.5 years respectively. (*ibid.*, page 116)

To demonstrate this effect, the correlogram of series (b) in Figure 8.6 has $r'_1 = 0.486$ and $r'_2 = 0.133$, thus giving, according to (8.7),

$$-a' = 0.552 \quad b' = 0.135 \quad \cos\theta' = \frac{-a'}{2\sqrt{b'}} = 0.751 \quad \theta' = 41.3^\circ$$

which corresponds to a period of about 8.7 years. In contrast, since it is known that $a = -1.5$, $b = 0.9$ and $\theta = 37.7^\circ$, the true period is 9.5 years.

This may not seem to be too large an effect, given that the first two serial correlations have been reduced from 0.78 and 0.33 to 0.49 and 0.13, respectively. Kendall argued, however, that the example served to bring out the difficulties associated with short series and the consequent unreliability of coefficients calculated from the first two serial correlations in such situations, pointing out that if $r'_2 = 0.18$ rather than 0.13 then an *increased* period of about 12 years would have been obtained and if $r'_2 = 0.20$ no solution would be possible since then $a'^2 > 4b'$ and $\cos\theta' > 1$. Both these changes in values were well within the one standard error bound of $1/\sqrt{65} = 0.12$.

Kendall also pointed out that the proportionate declines in the first two serial correlations brought about as a consequence of superposed variation were rather different, being $0.49/0.78 = 0.63$ and $0.13/0.33 = 0.40$ respectively, making it illegitimate to conclude that r_1 and r_2 were reduced by a constant proportion c. In fact, Kendall went on to show that, even in long series where it is legitimate to make this assumption, the length of the period was very sensitive to superposed variation, providing an example based on (8.6) in which a superposed variation of about 10% of the total ($c = 0.9$) shortened the period by around one year.

8.5 Kendall employed these results to investigate the oscillatory properties of the wheat price series of Figure 8.4. The correlogram shown in Figure 8.5 has upcrosses at about 7.5, 17.2 and 26.1 years, giving periods of 9.7 and 8.9 years with a mean of 9.3 years, with a similar result being obtained from the troughs in the correlogram. Calculating a' and b' by (8.7) with $r_1' = 0.5773$ and $r_2' = 0.0246$ gives

$$a' = -0.8446 \quad b' = 0.4630$$

so that

$$\cos\theta' = 0.6206 \quad \theta' = 51.63^\circ$$

with an estimated period of 6.97 years. As this is rather smaller than that calculated from the correlogram, Kendall suspected the existence of superposed variation. To estimate the variance of the superposed element η, he assumed that this was random with no periodic terms of very short period, thus enabling him to use the variate differencing method of Chapter 4. By taking up to 10th differences of the original series (i.e., before the trend was eliminated), Kendall estimated the random variance as 27.72. Since the total variance of the series is 272.8, this gives c as $1-(27.72/272.8) = 0.90$, so that $r_1 = r_1'/c = 0.5773/0.90 = 0.641$ and, similarly, $r_2 = 0.027$. From these are obtained

$$a = -1.059 \quad b = 0.652 \quad \cos\theta = 0.6551 \quad \theta = 49.07^\circ$$

giving a period of 7.34 years, which is still too short.

To produce a period of 9.3 years would require a random superposed variance of about 25 per cent, rather than 10 per cent, of the total variance and this led Kendall to question the assumption of a random superposed variation:

> we have little ground for expecting that it should be. A positive correlation between successive values of η will reduce the variance shown as random by the variance difference method and unless we have prior reason to suppose that η is random the values given by the variate difference method are likely to be too small. Unfortunately we rarely have any prior knowledge of η, but from general economic considerations one would not be surprised to find that there do exist positive correlations from one year to the next, owing to the enduring nature of some of the causes which can give rise to superposed variation. I conclude generally that discrepancies of the type here considered support the view that the period is to be determined from the correlogram, not from solution of the regression equation. (*ibid.*, pages 118–19)

Interactions and cross-correlations between time series

8.6 After mentioning extensions to higher-order and nonlinear autoregressive schemes, in his final paragraph Kendall (1944) introduced a further potential difficulty.

> A more serious problem arises if the series ε is itself not random, a state of affairs which one fears might be fairly common in economic series. To take the wheat price data once again, it would not be surprising to find that the wheat price oscillations were regenerated by a series of disturbances, part of which were attributable to variations in acreages, yields, or the prices of other crops. Such disturbances might themselves be oscillatory. For such cases the problem becomes exceedingly complicated. To discuss it at all satisfactorily one would require a long series or collateral evidence in the form of other series of a similar character. If there is a royal road in this subject it has not yet been discovered. (*ibid.*, page 119)

In fact, Kendall (1943) had already addressed the case in which the oscillations of two series could be correlated.

> When a number of products are associated or are likely to be affected together by external shocks there may appear interactions of a very complicated kind. Movements in one series may affect the disturbance function in others, and in consequence the functions may cease to be random: and even if they continue to be random, the functions for different products may be correlated. (*ibid.*, page 112)

To analyse such a situation, Kendall used the cross-correlations first introduced over forty years earlier by Hooker (§**2.6**), which were denoted $r_{xy}(k)$ using the notation introduced in §**4.12**. Suppose there are two series of the form (8.1) with solutions

$$x_t = \sum_{j=0}^{\infty} \xi_j \varepsilon_{t-j+1}$$

and

$$y_t = \sum_{j=0}^{\infty} \chi_j \zeta_{t-j+1}$$

The covariance between x_t and y_{t+k} is then given by

$$E(x_t, y_{t+k}) = \sum_{t=-\infty}^{\infty} \left(\sum_{j=0}^{\infty} \xi_j \varepsilon_{t-j+1} \right) \left(\sum_{j=0}^{\infty} \chi_j \zeta_{t-j+1} \right)$$

Kendall assumed that the disturbances were random but that $\xi_t = \mu\zeta_t$, so that an external disturbance affects both series to a similar extent but in different proportions. The covariance then reduces to

$$E(x_t, y_{t+k}) = \sum_{j=0}^{\infty} (\xi_j \chi_{j+k}) \mu^2 \sigma_\zeta^2$$

so that it and the cross-correlation $r_{xy}(k)$ will be proportional to $\sum \xi_j \chi_{j+k}$. If

$$\xi_j = A_1 p_1^j \sin \theta_1 j \quad \chi_j = A_2 p_2^j \sin \theta_2 j$$

then

$$r_{xy}(k) \propto p_2^k \sum_{j=0}^{\infty} p_1^j p_2^j \sin \theta_1 j \sin \theta_2 (j+k) \tag{8.8}$$

Thus, for $k \geq 0$, $r_{xy}(k)$ will have the appearance of a damped sinusoid because of the presence of p_2^k. For $k \leq 0$ the effect will be the same except that the damping will be according to the factor p_1^k, so that the damping is not symmetrical and thus $r_{xy}(k) \neq r_{xy}(-k)$.

8.7 Figure 8.7 shows the sheep and cow population series, while the cross-correlation function $r_{cs}(k)$, using an obvious nomenclature, is shown in Figure 8.8: Kendall referred to this as the *lag correlogram*. The series clearly show a similar pattern of oscillations, while the lag correlogram appears to be of the type arrived at above, although Kendall was careful to point out that the assumptions made

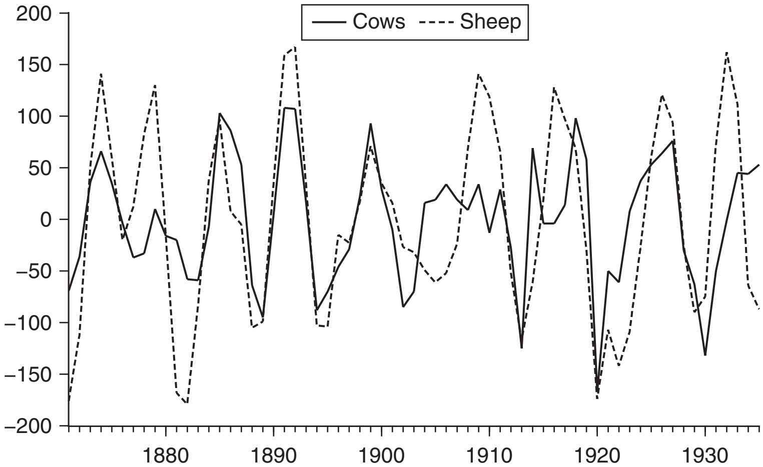


Figure 8.7 Cow and sheep populations for England and Wales, 1871–1935

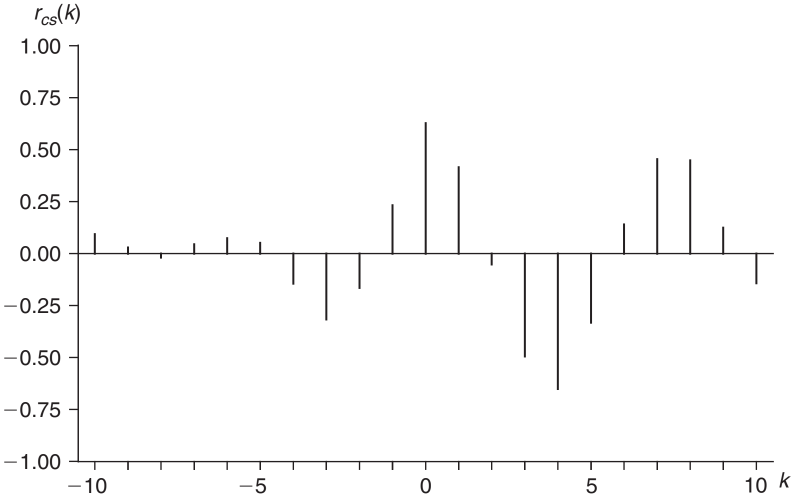


Figure 8.8 Cross-correlations between cow and sheep populations, $-10 \leq k \leq 10$

to reach (8.8) were 'rather specialized, and unlikely to be realized exactly in practice' (*ibid.*, page 113). Nevertheless, he concluded that

> the [cross-]correlations . . . reach a maximum for $k=0$, which indicates that the oscillations have some cause in common. It may be inferred that the oscillations do not take place one at the expense of the other – that is to say, an increase in cows is not accompanied by a decline in sheep. On the contrary, the two seem, on the average, to react in the same direction. This conforms to the idea that the oscillations in livestock populations are excited by disturbance functions outside the farming system. (*ibid.*, page 116)

'Internal' correlations and the lambdagram

8.8 In his final paper on time series, Yule (1945) broke away from the analysis of oscillatory processes to consider an alternative way of characterizing the properties of a time series.[5] This was based on a result in Yule and Kendall (1950, page 390) concerning the variance of the means of independent samples drawn from a time series and which focused on the behaviour of the quantity

$$\lambda_n = \frac{2}{n}((n-1)\rho_1 + (n-2)\rho_2 + \cdots + \rho_{n-1}) \tag{8.9}$$

as n increases. This can be written as

$$\lambda_n = \frac{2}{n}T_n$$

where

$$T_n = \sum_{i=1}^{n-1} S_i \quad S_i = \sum_{i=1}^{n-1} \rho_i$$

so that it is the second sum of the serial correlations scaled by the factor $2/n$. If S_m has a finite value such that m and T_m become negligible when compared to n and T_n, then the limiting value of λ_n is $2S_m$.

Yule termed λ_n the *coefficient of linkage*. If $\lambda_n = 0$ then either all of the serial correlations are zero or any positive correlations are balanced by negative correlations. Yule showed that $-1 < \lambda_n < n - 1$ and the implications of these limits are revealed when we use Yule's result that the variance of the means of independent samples of length n is $(\sigma^2/n)(1 + \lambda_n)$, where σ^2 is the variance of the series itself. The maximum value $\lambda_n = n - 1$ occurs when $\rho_i = 1$ for $i = 1, \ldots, n - 1$, so that the terms of samples of size n are *completely* linked together and the means of the successive samples have the same variance as the series itself. The minimum value $\lambda_n = -1$ is achieved when the terms in the sample are as completely negatively linked as possible (bearing in mind that not *all* pairs in a sample can have a correlation of -1) and the means of the successive samples have zero variance and hence do not vary at all. If $\lambda_n = 0$ then the terms are unlinked and the means of successive samples behave like means of random samples. Yule termed a plot of λ_n against n a *lambdagram.*

If a correlated series is formed by summing a random series in overlapping runs of k terms, i.e., as $v_t = \sum_{j=1}^{k} u_{t+j}$, then $\rho_i = (k - i)/k$, $i = 1, \ldots, k - 1$, $\rho_i = 0$, $i \geq k$, $S_n = 1/2(k - 1)$ and, in the limit, $\lambda_n = k - 1$. Thus all values of λ_n are positive and the lambdagram clearly approaches a limit, as is seen in Figure 8.9, which displays the lambdagram for $k = 5$.

Figure 8.10 displays calculated lambdagrams for a variety of series analysed by Yule and Kendall, as well as the sunspot index (n is generally set at the value chosen by Yule). They display a variety of patterns, with Kendall's agricultural series having similar lambdagrams both between themselves and with Beveridge's wheat price index. The sunspot index has a lambdagram that is generally increasing towards a maximum that appears to be in the region of 3.75, while the lambdagram of Kendall's series I looks to be declining towards a value of around 1.2. Since this latter series is generated by the oscillatory process (8.4), Kendall (1945b) analysed the implications for the lambdagram of this underlying generating process. For the process of §**8.1**, Kendall showed that the limiting value of the lambdagram for large n is

$$\lambda = \frac{-2(a + b + b^2)}{(1 + b)(1 + a + b)} \tag{8.10}$$

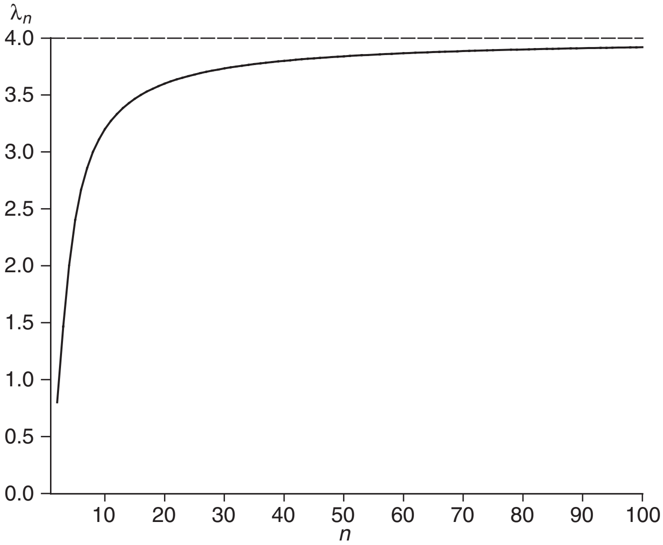


Figure 8.9 Lambdagram for a correlated series formed by summing the terms of a random series in overlapping groups of five

If $b = 1$ then it is easy to see that $\lambda = -1$, while using (8.3) and (8.7) allows λ to be written as

$$\lambda = \frac{2}{1 + a + b}(\rho_1 - b)$$

For an oscillatory process $1 + a + b = (1 - \rho_1)/(1 + b) \geq 0$ because $b > 0$ and $-1 \leq \rho_1 \leq 1$. Hence λ will be positive or negative depending on whether ρ_1 is greater than or less than b, the square of the damping factor p.

8.9 Of course, the 'true' autocorrelations are given by $\rho_0 = 1$ and $\rho_1 = -a/(1 + b)$ followed by the recursion $\rho_{i+2} = -a\rho_{i+1} - b\rho_i$. The set of autocorrelations thus generated with $a = -1.1$ and $b = 0.5$ can then be used to calculate the 'theoretical' lambdagram, which is shown with the empirical lambdagram of Series I in Figure 8.10. The limiting value from (8.10) is $\lambda = 1.167$ and by $n = 50$ both the observed and theoretical lambdagrams are consistent with this and are themselves almost identical. However

> throughout the previous course of the lambdagram the observed values are much higher than the theoretical values.
>
> It seems clear that these differences are due to the failure of the observed correlations to damp out according to theoretical explanation [cf. the discussion of §**8.2**]. If this is the correct explanation I should expect it to be equally

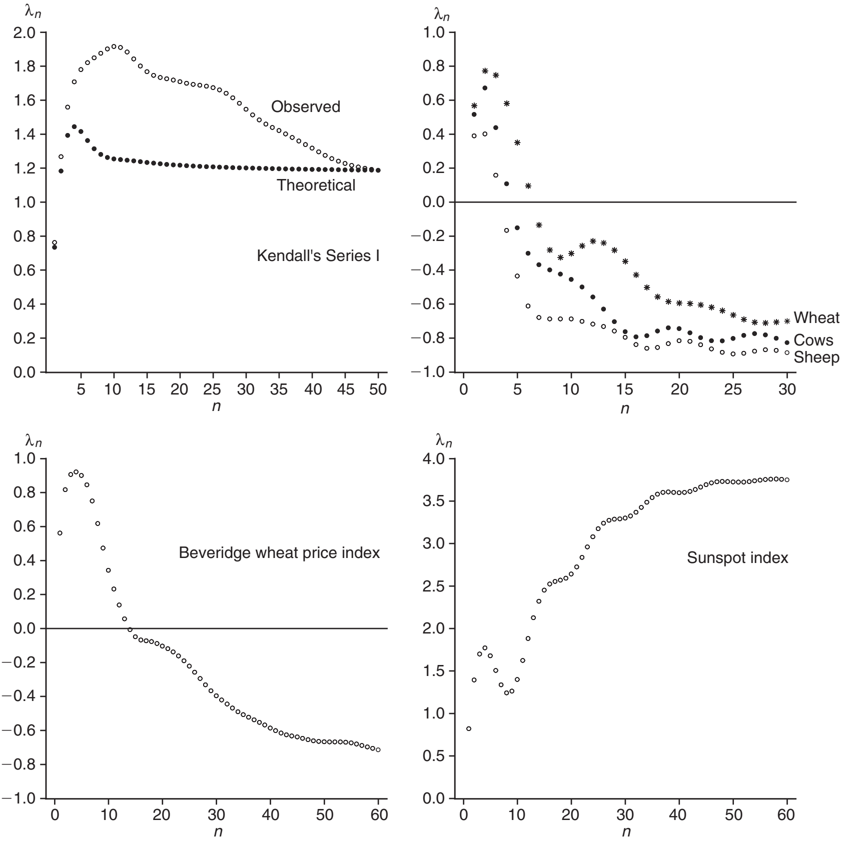


Figure 8.10 Calculated lambdagrams for a variety of time series

> possible on occasion for the observations to be systematically lower than the theoretical over parts of the range. Series I, it is to be remembered, is based on 480 terms and we are entitled to expect that for shorter series observation and theory will be less in agreement. (Kendall, 1945b, page 228)

Values for *a* and *b* for each of the other series shown in Figure 8.10 can be computed using (8.7) and the limiting values of the lambdagram calculated using (8.10). This produces λ values of −0.421, −0.394 and 0.004 for the sheep, wheat and cow series, 0.876 for the Beveridge wheat index and 0.935 for the sunspot index. From Figure 8.10 it is clear that none of these limiting values look to be very close to the values that the empirical lambdagrams look to be tending towards. While Kendall thought that short oscillatory series would give rise to serial correlations that did not damp out according to theoretical expectation, and hence empirical lambdagrams at odds with their theoretical counterparts, an alternative explanation could be that these series are not adequately represented by oscillatory processes, so that more general autoregressions are required. What is lacking here, of course, is a method for selecting the appropriate autoregression and, more generally, of assessing serial correlations for their statistical significance, a method that will be developed in Chapter 9.

Mixed autoregressive-moving average processes

8.10 The general *p*th-order autoregression

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = \varepsilon_t$$

was generalized by (A.M.) Walker (1950) to include a moving average of order q:

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q} \tag{8.11}$$

For x_t to be stationary the roots of the characteristic equation $z^p + a_1 z^{p-1} + \cdots + a_p = 0$ are required to have moduli less than unity and, for the moving average to be regular, Wold's conditions of §**7.21** need to hold. Box and Jenkins (1968, 1970) referred to a model of the form (8.11) as a *general mixed autoregressive-moving average model of order* (p, q) and gave it the acronym 'ARMA (p, q)', arguing that such models would often offer parsimonious representations of a time series in the sense that the orders of the ARMA process will be much smaller than the orders required by pure autoregressive or moving average representations: typically p and q would be 0, 1 or 2 in most applications.[6]

Box and Jenkins introduced a notation for ARMA processes that has since become standard. If x_t is allowed to have a non-zero mean μ, define the deviation as $\tilde{x}_t = x_t - \mu$ and write the ARMA (p, q) process for $\tilde{x}_t$ as

$$\tilde{x}_t - \phi_1 \tilde{x}_{t-1} - \cdots - \phi_p \tilde{x}_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{8.12}$$

where a_t is a 'white noise' series consisting of uncorrelated random normal deviates all having mean zero and variance σ_a^2. To manipulate models such as (8.12), it is convenient to define the *backshift operator* B such that $Bx_t \equiv x_{t-1}$, $B^j x_t \equiv x_{t-j}$ and $B^j \mu = \mu$. Using B, (8.12) can be written

$$\phi_p(B)\tilde{x}_t = \theta_q(B) a_t \tag{8.13}$$

where

$$\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

are polynomials in B of degree p and q and are called the autoregressive and moving average operators respectively. The stationarity and invertibility (regularity) conditions for (8.12) may then be expressed by saying that the roots of $\phi_p(B) = 0$ and $\theta_q(B) = 0$ must lie outside the unit circle.

If the model is expressed in terms of x_t rather than the deviations $\tilde{x}_t$, (8.13) becomes

$$\phi(B)x_t = \theta_0 + \theta(B)a_t$$

where

$$\theta_0 = (1 - \phi_1 - \cdots - \phi_p)\mu$$

and the subscripts to the operators have been omitted as there is here no danger of confusion.

It is easily shown (for example, Box and Jenkins, 1970, chapter 3.4) that the autocorrelations will be given by $\phi(B)\rho_k = 0$ for $k > q$, the first q autocorrelations depending on the q moving average parameters as well as the p autoregressive parameters. The p values $\rho_q, \rho_{q-1}, \ldots, \rho_{q-p+1}$ provide the necessary starting values for $\phi(B)\rho_k = 0$, $k > q$, a recursion which then entirely determines the autocorrelations at higher lags. If $q - p < 0$, the whole autocorrelation function will be dictated by $\phi(B)$ and the starting values, but if $q - p \geq 0$ there will be $q - p + 1$ initial values $\rho_q, \rho_{q-1}, \ldots, \rho_{q-p}$ which do not follow this general pattern. The partial autocorrelation function (cf. §**6.8**) will be infinite in extent, behaving eventually like the partial autocorrelation function of a pure moving average process.

An important member of this class of model is the ARMA (1,1) process

$$(1 - \phi B)\tilde{x}_t = (1 - \theta B)a_t \tag{8.14}$$

for which it can be shown that

$$\rho_1 = \frac{(1 - \phi\theta)(\phi - \theta)}{1 + \theta^2 - 2\phi\theta} \quad \rho_j = \phi^{j-1}\rho_1 \quad j > 1 \tag{8.15}$$

so that there is one initial value ρ_1 that is a function of both ϕ and θ before the remaining autocorrelations follow the geometric decay of an AR(1) process. The stationarity and invertibility conditions are $-1 < \phi < 1$ and $-1 < \theta < 1$ and, along with (8.15), these can be used to show that ρ_1 and ρ_2 must lie in the region

$$|\rho_2| < |\rho_1|$$

$$\rho_2 > \rho_1(2\rho_1 + 1) \quad \rho_1 < 0$$

$$\rho_2 > \rho_1(2\rho_1 - 1) \quad \rho_1 > 0$$

The partial autocorrelation function has a single initial value of ρ_1 and thereafter behaves like the partial autocorrelation function of a pure MA(1) process: if $\theta > 0$ it is dominated by a smoothly damped exponential which decays from a value of ρ_1, with sign determined by the sign of $(\phi - \theta)$; if $\theta < 0$ the exponential decay oscillates.

Generating functions

8.11 Wold's Theorem 5 (§**7.12**) states that a necessary and sufficient condition for a stationary process to exist is that the autocorrelations ρ_k are the coefficients of a non-decreasing function $W(\vartheta)$ such that $W(0) = 0$, $W(\pi) = \pi$ and

$$\rho_k = \frac{1}{\pi}\int_0^{\pi} \cos k\vartheta \cdot dW(\vartheta) \tag{8.16}$$

The inversion formula which allows $W(\vartheta)$ to be uniquely determined by the autocorrelation coefficients,

$$W(\vartheta) = \vartheta + 2\sum_{k=1}^{\infty} \frac{\rho_k}{k} \sin k\vartheta$$

is termed the generating function of the ρ_k by Wold. The corollary to the theorem states that if the autocorrelations are absolutely convergent then the derivative of $W(\vartheta)$ will exist and will be given by

$$W'(\vartheta) = \sum_{k=-\infty}^{\infty} \rho_k \cos k\vartheta = 1 + 2\sum_{k=1}^{\infty} \rho_k \cos k\vartheta \tag{8.17}$$

Moran (1949) referred to $W(\vartheta)$ as the integrated power spectrum of the process and $W'(\vartheta)$ as the spectral density. Following Quenouille (1947a) and Moran (1949), defining $z = e^{i\vartheta}$ enables (8.17) to be written as

$$W'(z) = \sum_{k=-\infty}^{\infty} \rho_k z^k$$

or, in terms of the autocovariances γ_k, as

$$\gamma(z) = \sum_{k=-\infty}^{\infty} \gamma_k z^k \tag{8.18}$$

This is termed the covariance-generating function (c.g.f.) and Moran (1949) showed that if the autocorrelations are absolutely convergent then the γ_k are uniquely determined and (8.18) converges for $|z| = 1$. Note that $\gamma(e^{i\vartheta}) = \gamma_0 W'(\vartheta)$.

Suppose that we now have such a process $\{x_t\}$ with autocovariances γ_k and c.g.f. $\gamma_x(z)$. A new process $\{y_t\}$ is then defined as

$$y_t = \sum_{i=0}^{\infty} \alpha_i x_{t-i}$$

where $\sum_{i=0}^{\infty} \alpha_i$ is absolutely convergent. $\{y_t\}$ will then be stationary with autocovariances

$$\begin{aligned}\gamma_{y,k} = \gamma_{y,-k} &= E(y_t y_{t+k}) \\ &= E\left\{\left(\sum_{i=0}^{\infty} \alpha_i x_{t-i}\right)\left(\sum_{i=0}^{\infty} \alpha_i x_{t-i}\right)\right\} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \alpha_i \alpha_j \gamma_{k+i-j}\end{aligned}$$

and c.g.f.

$$\begin{aligned}\gamma_y(z) &= \sum_{k=-\infty}^{\infty} \gamma_{y,k} z^k \\ &= \sum_{k=-\infty}^{\infty}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \alpha_i \alpha_j \gamma_{k+i-j} z^k \\ &= \left(\sum_{i=0}^{\infty} \alpha_i z^{-i}\right)\left(\sum_{j=0}^{\infty} \alpha_j z^j\right) \gamma_x(z) \qquad (8.19)\end{aligned}$$

which shows the effect on the c.g.f. of taking a moving average.

As an example, consider

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = \eta_t \qquad (8.20)$$

where $\{\eta_t\}$ is a random process, so that $\gamma_\eta(z) = 1$, and the characteristic equation

$$z^p + a_1 z^{p-1} + \cdots + a_p = 0$$

has all its roots outside the unit circle $|z| = 1$. Using (8.19), we then have

$$(1 + a_1 z + \cdots + a_p z^p)(1 + a_1 z^{-1} + \cdots + a_p z^{-p})\gamma_x(z) = 1 \qquad (8.21)$$

from which it follows that

$$\gamma_0 W'(\vartheta) = [(1 + a_1^2 + \cdots + a_p^2) + 2(a_1 + a_1 a_2 + \cdots + a_{p-1} a_p) \cos \vartheta \\ + \cdots + 2a_p \cos p\vartheta]^{-1}$$

This result can be extended by dropping the assumption that $\{\eta_t\}$ is random, so that we have an ARMA process. The c.g.f. $\gamma_x(z)$ can then be obtained from (8.21) by replacing the right-hand side with $\gamma_\eta(z)$. Moran (1949, 1950) used this approach to provide a much shorter proof of, and generalizations to, Slutzky's sinusoidal limit theorem of §**5.15**.

If the variances of x_t and η_t are σ_x^2 and σ_η^2, respectively, then the autocorrelation generation function (a.g.f.) is defined from the c.g.f. as

$$\rho_x(z) = \sum_{k=-\infty}^{\infty} \rho_{x,k} z^k = \frac{\sigma_\eta^2}{\sigma_x^2} \gamma_x(z) \tag{8.22}$$

and this will be found to be very useful in Chapter 9.

9
Statistical Inference, Estimation and Model Building for Stationary Time Series

The sampling theory of serial correlations

9.1 As we saw in §**8.3**, Kendall (1945a) expressed frustration at the lack of a sampling theory related to serial correlations when attempting to interpret the correlograms obtained from his experimental series.

> The significance of the correlogram is ... difficult to discuss in theoretical terms. ... (O)ur real problem is to test the significance of a set of values which are, in general, correlated. It is quite possible for a part of the correlogram to be below the significance level and yet to exhibit oscillations which are themselves significant of autoregressive effects. At the present time our judgments of the reality of oscillations in the correlogram must remain on the intuitive plane. (*ibid.*, page 103)

In his discussion of the paper from which this quote is taken, Bartlett actually took Kendall to task for not attempting any form of inference: 'it might have been useful, and probably not too intractable mathematically, to have evaluated at least the approximate theoretical standard errors for the autocorrelations' (*ibid.*, page 136). This rebuke may have been a marker for a major development in the sampling theory of serial correlations that was to be published within a year of the appearance of Kendall's paper.

Large-sample theory

9.2 This paper was Bartlett (1946), which, by extending results presented a decade earlier (Bartlett, 1935), aimed to 'amplify some suggestions I made in the discussion on [Kendall's] paper about the sampling errors of a correlogram' (Bartlett, 1946, page 27). Bartlett's focus was on large samples and so attention was concentrated on the serial correlation formula

$$r_k = \frac{\frac{1}{T}\sum_{t=1}^{T-k} x_t x_{t+k}}{\frac{1}{T}\sum_{t=1}^{T} x_t^2} = \frac{c}{\upsilon}, \tag{9.1}$$

say, regarded as an estimate of the true autocorrelation ρ_k of the zero mean, variance σ_x^2, and normally distributed series x_t, $t = 1, 2, \ldots, T$. Writing the total differential of (9.1) as $\delta r_k = \delta c/\upsilon - c\delta\upsilon/\upsilon^2$, so that

$$(\delta r_k)^2 = \frac{(\delta c)^2}{\upsilon^2} - \frac{2c\delta c\delta\upsilon}{\upsilon^3} + \frac{c^2(\delta\upsilon)^2}{\upsilon^4}$$

then, on setting $\upsilon = 1$ without loss of generality and equating $(\delta r_k)^2$ with $V(r_k)$, etc., we have

$$V(r_k) = V(c) - 2cCov(c, \upsilon) + c^2 V(\upsilon)$$

Bartlett showed that

$$V(\upsilon) = \frac{2}{T}\sum_{i=-\infty}^{\infty} \rho_i^2$$

$$V(c) = \frac{1}{T}\sum_{i=-\infty}^{\infty} (\rho_i^2 + \rho_{i+k}\rho_{i-k})$$

$$Cov(c, \upsilon) = \frac{2}{T}\sum_{i=-\infty}^{\infty} \rho_i\rho_{i+k}$$

so that the variance of r_k could be expressed as

$$V(r_k) = \frac{1}{T}\sum_{i=-\infty}^{\infty} (\rho_i^2 + \rho_{i-k}\rho_{i+k} - 4\rho_k\rho_i\rho_{i+k} + 2\rho_i^2\rho_k^2) \tag{9.2}$$

This result shows that, even for large samples with the simplifying assumption of normality, the variance of r_k depends on *all* the autocorrelations and these, of course, cannot all be estimated directly from a finite series.

9.3 Useful approximations may, however, be obtained in certain cases. If x_t is random, so that $\rho_k = 0$, $k \neq 0$, then, from (9.2), $V(r_k) = 1/T$, which is the variance of a correlation coefficient from a bivariate normal sample and was the formula employed by Kendall (1943): cf. §**8.3**. Using the fact that $\rho_{-k} = \rho_k$ then, if $\rho_i \neq 0$, $0 < i < k$, and $\rho_i = 0$, $i \geq k$, from (9.2) we have

$$V(r_k) = \frac{1}{T}\sum_{i=-(k-1)}^{k-1} \rho_i^2 = \frac{1}{T}(1 + 2\rho_1^2 + \cdots + 2\rho_{k-1}^2) \tag{9.3}$$

and it may also be shown that

$$\sigma_x^4 Cov(r_k, r_{k+j}) = \frac{1}{T}\sum_{i=-\infty}^{\infty} \rho_i \rho_{i+j} \tag{9.4}$$

If x_t is generated by an AR(1), or Markov, process, so that $\rho_k = \rho^k$, then[1]

$$V(r_k) = \frac{1}{T}\left(\frac{(1+\rho^2)(1-\rho^{2k})}{1-\rho^2} - 2k\rho^{2k}\right)$$

which, for large k, becomes

$$V(r_k) = \frac{1}{T}\sum_{i=-\infty}^{\infty} \rho^{|2i|} = \frac{1}{T}\frac{1+\rho^2}{1-\rho^2} \tag{9.5}$$

Equations (9.3) and (9.4) can also be derived directly using the generating function approach of §**8.11**, since from the a.g.f. (8.22), it can be seen that

$$(\rho(z))^2 = \sum_{j=-\infty}^{\infty} z^j \sum_{i=1}^{\infty} \rho_i \rho_{i+j}$$

will deliver the sums required in (9.3) and (9.4). For example, as the a.g.f. of a Markov process is

$$\rho(z) = -1 + \frac{1}{1-\rho z} + \frac{1}{1-\rho z^{-1}}$$

it follows that

$$\begin{aligned}(\rho(z))^2 =& 1 + (1-\rho z)^{-2} + (1-\rho z^{-1})^{-2} \\ &- 2(1-\rho z)^{-1} - 2(1-\rho z^{-1})^{-1} + 2(1-\rho z)(1-\rho z^{-1})\end{aligned}$$

The coefficient of z^0 is then

$$1 + 1 + 1 - 2 - 2 + 2(1 + \rho^2 + \rho^4 + \cdots) = \frac{1+\rho^2}{1-\rho^2}$$

thus delivering (9.5), while the coefficient of z^k is given by

$$(k+1)\rho^k - 2\rho^k + 2(\rho^k + \rho^{k+2} + \cdots) = \rho^k \left(k - 1 + \frac{2}{1-\rho^2}\right)$$

Hence, using (9.4), the covariance between r_k and r_{k+j} in a Markov scheme is

$$Cov(r_k, r_{k+j}) = \frac{1}{T}\rho^k \left(k - 1 + \frac{2}{1-\rho^2}\right)$$

and the correlation between them is then

$$\frac{\rho^k((k+1) - (k-1)\rho^2)}{1+\rho^2}$$

9.4 Bartlett (1946) used these results to analyse the correlogram of Kendall's (1944) artificial series of length $T=65$ generated as (8.4) but with the error process now an integer rectangular random variable ranging from -9.5 to $+9.5$ (cf. the process in §**8.2**). Two estimates of the correlogram and the true autocorrelations, calculated from $\rho_k = 1.1\rho_{k-1} - 0.5\rho_{k-2}$, with $\rho_0 = 1$ and $\rho_1 = 1.1/1.5 = 0.733$, are shown for k up to 30 in Figure 9.1 (two-standard error bounds under the null hypothesis that the series is random are $2/\sqrt{65} \approx 0.25$). The first estimate of the correlogram uses the large sample formula (9.1), while

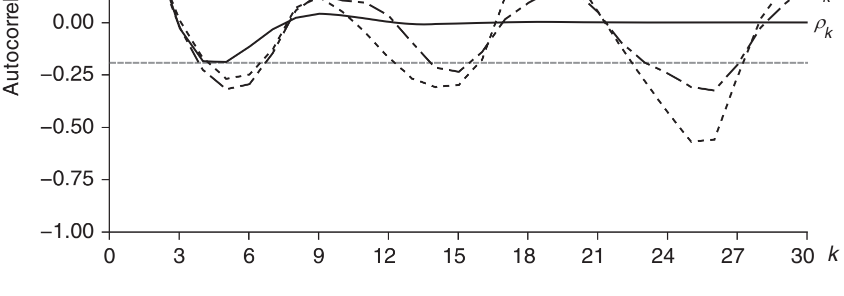


Figure 9.1 Correlogram and autocorrelations of Kendall's (1944) artificial series $x_t - 1.5x_{t-1} + 0.5x_{t-2} = u_t$

the second uses the formula employed by Kendall (1944, equation (1)):

$$r'_k = \frac{\sum_{t=1}^{T-k} x_t x_{t+k}}{\left(\sum_{t=1}^{T-k} x_t^2 \sum_{t=1}^{T-k} x_{t+k}^2\right)^{\frac{1}{2}}}$$

Neither r_k nor r'_k die down as k increases in the manner predicted by the theoretical autocorrelations ρ_k: indeed, $r'_{24} = -0.43$, $r'_{25} = -0.57$ and $r'_{26} = -0.56$ are unexpectedly large compared to their corresponding ρ_k values, which by this time are essentially zero. The 'large sample' counterparts, $r_{24} = -0.24$, $r_{25} = -0.31$ and $r_{26} = -0.33$, are somewhat smaller but still apparently far larger than they 'should' be. However, using (9.3), $V(\rho_k) \approx 2.44/T$ for $k > 10$ and so these serial correlations have standard errors of approximately 0.20, implying that, although they are quite large in magnitude, they are not significantly so (one standard error bounds of ±0.20 are also shown on Figure 9.1).

Large sample goodness of fit tests

9.5 Bartlett (1946) showed that, for linear processes, the formula (9.4) for the covariance between two observed serial correlations, r_k and r_{k+j}, in terms of the theoretical autocorrelations, could be specialized to

$$Cov(r_k, r_{k+j}) \sim \frac{1}{T} \sum_{i=-\infty}^{\infty} (\rho_i \rho_{i-j} + \rho_{i-k-j}\rho_{i+k} + 2\rho_k \rho_{k+j} \rho_i^2 - 2\rho_k \rho_i \rho_{i-k-j} - 2\rho_{k+j} \rho_i \rho_{i-k})$$

which, on defining

$$\lambda_l = \sum_{i=-\infty}^{\infty} \rho_i \rho_{i-l} = \lambda_{-l}$$

may be written as

$$Cov(r_k, r_{k+j}) \sim \frac{1}{T}(\lambda_j + \lambda_{2k+j} + 2\rho_k \rho_{k+j} \lambda_0 - 2\rho_k \lambda_{k+j} - 2\rho_{k+j} \lambda_k)$$

Suppose we have the AR(p) process (8.20). Using (8.22), its a.g.f. is

$$\rho(z) = \sum_{i=-\infty}^{\infty} \rho_i z^i = \frac{\sigma_\eta^2}{\sigma_x^2}(1 + a_1 z + \cdots + a_p z^p)^{-1}(1 + a_1 z^{-1} + \cdots + a_p z^{-p})^{-1} \tag{9.6}$$

from which it follows that

$$\rho^2(z) = \sum_{i=-\infty}^{\infty} \lambda_i z^i = \frac{\sigma_\eta^4}{\sigma_x^4}(1 + a_1 z + \cdots + a_p z^p)^{-2}(1 + a_1 z^{-1} + \cdots + a_p z^{-p})^{-2}$$

Following Quenouille (1947a, 1947b), define

$$(1 + a_1 z + \cdots + a_p z^p)^2 = \sum_{i=0}^{2p} A_i z^i \tag{9.7}$$

Using (9.6) and (9.7) it may be shown that

$$\sum_{i=0}^{2p} A_i \rho_{j-i} = 0 \quad j \geq p$$

$$\sum_{i=0}^{2p} A_i \lambda_{j-i} = 0 \quad j \geq 0$$

Now define the variables

$$R_s = \sum_{i=0}^{2p} A_i r_{s-i} \quad s > p$$

Quenouille (1947b) showed that

$$\begin{aligned} Cov(r_k, R_j) &\sim 0 && j > k \\ &\sim \frac{\sigma_\eta^4}{T\sigma_x^4} \sum_{i=0}^{2p} A_i \lambda_i && j = k \end{aligned}$$

from which it follows that, for T large, the R_s are independently and normally distributed about zero with variance

$$V(R_s) = \frac{\sigma_\eta^4}{T\sigma_x^4} \sum_{i=0}^{2p} A_i \lambda_i$$

Furthermore, the variables $r_s - \rho_s$, $s = 1, 2, \ldots, k$, are jointly distributed independently of R_s. Thus, if the equations $\rho_s = r_s$, $s = 1, 2, \ldots, p$, are used to fit an AR(p) scheme then the R_s can be used to test the adequacy of the fit.

For example, for the Markov scheme $x_t = \rho x_{t-1} + \eta_t$, for which $\rho_k = \rho^k$ and

$$\lambda_k = \rho^k \frac{1 + \rho^2}{1 - \rho^2} + k\rho^k$$

then, since $A_0 = 1$, $A_1 = -2\rho$ and $A_2 = \rho^2$, the variables $r_1 - \rho_1$ and $R_s = r_s - 2\rho r_{s-1} + \rho^2 r_{s-2}$, $s \geq 2$, will be independently and normally distributed with variances $(1 - \rho^2)/T$ and $(1 - \rho^2)^2/T$.

9.6 Quenouille (1947b) used this approach to check the adequacy of the oscillatory models fitted by Kendall (1943, 1945a) using the AR(2) scheme $x_t + ax_{t-1} + bx_{t-2} = \eta_t$. From the relationships in (8.7), the variables

$$R_s = r_{s+2} + 2ar_{s+1} + (a^2 + 2b)r_s + 2abr_{s-1} + b^2 r_{s-2}, \quad s \geq 1$$

will be distributed with mean zero and variance

$$V(R_s) = \frac{1}{T-s}\left(\frac{(1-b)((1+b)^2 - a^2)}{1+b}\right)^2$$

if the scheme is correctly specified. Quenouille focused on the variable $\chi_s^2 = R_s^2 / V(R_s)$, which will be distributed as $\chi^2(1)$, and its accumulation $\sum_{j=1}^{s} \chi_j^2$, which will be distributed as $\chi^2(s)$, if the AR(2) scheme is indeed the correct specification (for a related approach, see Whittle, 1952).

For Kendall's Series I (cf. §**8.2**), $T = 480$, $a = -1.1$ and $b = 0.5$, so that

$$R_s = r_s - 2.20r_{s-1} + 2.21r_{s-2} - 1.10r_{s-3} + 0.25r_{s-4}$$

and

$$V(R_s) = \frac{0.120}{480 - s}$$

Table 9.1 reports the values of R_s, χ_s^2 and $\sum_{j=1}^{s} \chi_j^2$ for $s \leq 30$. Although the statistics are significant at the 10% level for $s = 1$, no others reach any conventional level of significance, which should not, of course, come as any surprise given that this scheme did actually generate the data.

Applying this approach to Kendall's (1943) wheat price and sheep population series, analysed as AR(2) schemes in §**8.3**, found no evidence of misspecification for the latter series but, for the former, $\chi_2^2 = 4.254$, suggesting the possibility that a higher-order scheme might provide a better fit. Table 9.2 reports the statistics for the AR(2) scheme fitted to the annual sunspot index from 1702 to 2007 analysed in §**6.6**, for which $a = -1.39078$ and $b = 0.69026$. Seven of the first 17 χ_s^2 values are significant, so that the entire sequence of $\sum_{j=1}^{s} \chi_j^2$ are significant, implying that an AR(2) scheme for the sunspot index offers a poor fit that should be able to be improved upon.[2]

9.7 Quenouille (1947b) extended his method to deal with superposed variation of the type considered in §§**8.4–8.5**. Further extensions were provided by Bartlett and Diananda (1950) and A.M. Walker (1950, 1954), who considered the case of correlated residuals, so that the scheme under test was the ARMA process of §**8.10**. Such a test also encompasses Wold's (1949) test of a pure moving average scheme.

Table 9.1 Goodness of fit statistics obtained from fitting an AR(2) scheme to Kendall's Series I

s	R_s	χ_s^2	$\sum_{j=1}^{s} \chi_j^2$	s	R_s	χ_s^2	$\sum_{j=1}^{s} \chi_j^2$
1	0.0263	2.761	2.761	16	−0.0033	0.043	14.141
2	0.0020	0.015	2.776	17	−0.0027	0.027	14.169
3	0.0199	1.570	4.346	18	0.0064	0.155	14.324
4	−0.0062	0.151	4.497	19	0.0162	1.004	15.328
5	−0.0033	0.044	4.541	20	−0.0108	0.446	15.774
6	0.0236	2.189	6.730	21	−0.0149	0.848	16.622
7	−0.0259	2.646	9.375	22	0.0060	0.139	16.761
8	−0.0247	2.402	11.777	23	−0.0147	0.827	17.588
9	−0.0136	0.728	12.505	24	−0.0188	1.342	18.930
10	−0.0091	0.327	12.832	25	−0.0106	0.425	19.356
11	0.0049	0.095	12.927	26	−0.0189	1.351	20.706
12	0.0148	0.855	13.782	27	−0.0010	0.004	20.710
13	−0.0065	0.163	13.945	28	0.0012	0.005	20.716
14	−0.0056	0.120	14.065	29	0.0010	0.004	20.719
15	0.0029	0.033	14.098	30	0.0083	0.260	20.980

Table 9.2 Goodness of fit statistics obtained from fitting an AR(2) scheme to the sunspot index

s	R_s	χ_s^2	$\sum_{j=1}^{s} \chi_j^2$	s	R_s	χ_s^2	$\sum_{j=1}^{s} \chi_j^2$
1	−0.0249	6.646	6.646	16	0.0207	4.363	55.967
2	0.0100	1.074	7.720	17	0.0273	7.607	63.575
3	−0.0001	0.000	7.720	18	−0.0039	0.151	63.726
4	0.0319	10.806	18.526	19	0.0149	2.242	65.968
5	0.0253	6.777	25.303	20	−0.0097	0.946	66.914
6	0.0368	14.340	39.643	21	−0.0120	1.444	68.358
7	0.0294	9.090	48.733	22	−0.0049	0.237	68.596
8	−0.0083	0.724	49.456	23	0.0026	0.066	68.662
9	0.0044	0.201	49.658	24	−0.0121	1.444	70.106
10	0.0036	0.137	49.795	25	0.0053	0.279	70.384
11	0.0033	0.111	49.906	26	0.0159	2.509	72.894
12	0.0089	0.812	50.717	27	−0.0087	0.745	73.638
13	0.0062	0.391	51.108	28	0.0194	3.681	77.320
14	0.0069	0.492	51.600	29	0.0021	0.045	77.365
15	0.0007	0.005	51.605	30	0.0019	0.034	77.398

Bias in the estimation of serial correlations

9.8 Orcutt (1948) and Moran (1948) defined the sample serial correlation coefficient as in (9.1) but with $T - k$ as the divisor in the numerator and with x_t

explicitly defined to be the deviation of the observed series X_t from its sample mean $\overline{X} = \sum_{t=1}^{T} X_t/T$, i.e., as

$$r_k = \frac{\frac{1}{T-k}\sum_{t=1}^{T-k} x_t x_{t+k}}{\frac{1}{T}\sum_{t=1}^{T} x_t^2} \quad x_t = X_t - \overline{X} \tag{9.8}$$

On the assumption that x_t is independently distributed, so that $\rho_k = 0$, $k \geq 1$,

$$\begin{aligned} E(r_k) &= \frac{T}{T-k} E\left(\frac{\sum x_t x_{t+k}}{\sum x_t^2}\right) \\ &= TE\left(\frac{x_t x_{t+k}}{\sum x_t^2}\right) \\ &= \frac{1}{T-1} E\left(\frac{\sum_{t \neq j} x_t x_j}{\sum x_t^2}\right) \\ &= \frac{1}{T-1} E\left(\frac{\left(\sum x_t\right)^2 - \sum x_t^2}{\sum x_t^2}\right) \\ &= -\frac{1}{T-1} \end{aligned} \tag{9.9}$$

since $\sum x_t = 0$. Thus there is a downward bias in r_k even for a random series. If the serial correlation is defined using the *cyclic* definition

$$r_k^c = \frac{\sum_{t=1}^{T} x_t x_{t+k}}{\sum_{t=1}^{T} x_t^2} \tag{9.10}$$

where it is assumed that $x_{T+t} = x_t$, then it will also be the case that $E(r_k^c) = -(T-1)^{-1}$.[3]

Marriott and Pope (1954) and Kendall (1954) showed that, for the Markov scheme $x_t = \rho x_{t-1} + \eta_t$, for which $\rho_k = \rho^k$,

$$E(r_k) = \rho^k - \frac{1}{T-k}\left(\frac{1+\rho}{1-\rho}(1-\rho^k) + 2k\rho^k\right)$$

to terms of order T^{-1}, denoted $O(T^{-1})$, so that, for example,[4]

$$E(r_1) = \rho - \frac{1}{T-1}(1+3\rho)$$

Thus, for $T=25$ and $\rho=0.5$, $E(r_1)\approx 0.4$, and for $\rho=0.9$, $E(r_1)\approx 0.75$. If, on the other hand, we have the MA(1) scheme $x_t=\eta_t+\theta\eta_{t-1}$, so that $\rho_1=\theta/(1+\theta^2)=\rho$ and $\rho_k=0$, $k\geq 2$, we obtain, using the method of Kendall (1954),

$$E(r_1)=\rho+\frac{1}{T-1}(1+\rho)(4\rho^2-2\rho-1)$$

$$E(r_2)=-\frac{1}{T-2}(1+2\rho+2\rho^2)$$

$$E(r_k)=-\frac{1}{T-k}(1+2\rho)\quad k>2$$

Once again, the bias is always downwards (for $T=25$ and $\rho=0.5$, $E(r_1)\approx 0.44$, $E(r_2)\approx -0.11$, $E(r_3)\approx -0.09$, etc.). Quenouille (1949a) suggested that the bias of a generic serial correlation r could be reduced by computing the serial correlations for the two halves of the available sample, ${}_{(1)}r$ and ${}_{(2)}r$, and using $\tilde{r}=2r-\frac{1}{2}({}_{(1)}r+{}_{(2)}r)$, which will be unbiased to $O(T^{-2})$. For example, for $k>2$ (and assuming that T is even for simplicity)

$$E({}_{(1)}r_k)=E({}_{(2)}r_k)=-\frac{1}{T/2-k}(1+2\rho)$$

so that

$$E(\tilde{r}_k)=\frac{2k}{(T-k)(T-2k)}(1+2\rho)\sim O(T^{-2})$$

Thus with $T=25$ and $\rho=0.5$, $E(\tilde{r}_3)\approx 0.03$, which may be compared to $E(r_3)\approx -0.09$ obtained above.

9.9 Kendall (1954), as well as producing ‘cyclic’ versions of these expectations, cautioned against using such expressions when ρ was near to unity, where the distribution of r_1, for example, is so highly skewed that using expectations as a criteria for bias is itself open to question (Figures 9.4, 9.6 and 9.7 later in this chapter illustrate this aspect of the distribution). Moreover, Kendall argued that expansions of the type being used here are asymptotic and may not be accurate unless the serial correlations decline rapidly. Although he suggested that approximations using terms of $O(T^{-2})$ or $O(T^{-3})$ may not necessarily be better, various attempts at obtaining expectations and higher moments using higher-order expansions have nevertheless been made: see, for example, White (1961) and Shenton and Johnson (1965).

Partial autocorrelations

9.10 The approach taken in previous sections may be used to construct large sample tests of the partial autocorrelations (recall §**6.8**). The sample partial autocorrelations (or partial serial correlations) will now be denoted $r_{k\cdot}$, rather than $r(k \cdot 12 \ldots (k-1))$, to economize on notation. If X_t follows an AR(p) scheme then, for $k > p$, $r_{k\cdot}$ will be asymptotically normally distributed with zero mean and variance T^{-1} or, equivalently, that $\sqrt{T} \cdot r_{k\cdot}$ will be distributed as $N(0, 1)$.

As an example of the application of this result, the first five partial serial correlations of the sunspot numbers are 0.820, –0.677, –0.144, 0.044 and 0.015 (see Table 6.3). At the 5% level, a partial serial correlation in excess of $1.96/\sqrt{308} = 0.112$ (in absolute value) would reject the hypothesis that the true partial correlation was zero. The existence of $r_{3\cdot} = -0.144$ thus provides strong evidence against the series being generated by an AR(2) scheme, consistent with the finding in §**9.6**, but the insignificance of $r_{4\cdot}$ and $r_{5\cdot}$ provide no evidence against an AR(3) specification, although higher order partial serial correlations may well do so (see the further analysis of the sunspot index in §§**9.43–9.46**).

Exact moments of the null distribution of the serial correlation coefficient

9.11 Alongside the construction of approximate, large sample, goodness of fit tests, the small sample distribution theory of serial correlation also began to be developed, with the aim of obtaining, where possible, exact results. Such exactitude naturally came at a cost, usually that of assuming normality of X_t and sometimes in using the circular definition of serial correlation and/or assuming that the mean of X_t was known.

9.12 Anderson (1942) and Dixon (1944) assumed X_t to be independently and normally distributed with mean μ and variance σ^2, so that $\rho_k = 0$, $k \geq 1$. Using the cyclic definition r_1^c of the first-order serial correlation from (9.10), they obtained, as well as $E(r_1^c) = -1/(T-1)$, the variance

$$V(r_1^c) = \frac{T(T-3)}{(T-1)^2(T+1)}$$

Dixon (1944) also went on to obtain exact expressions for the higher moments:

$$E(r_1^c)^{2j-1} = \frac{-1 \cdot 3 \cdot 5 \cdots (2j-1)}{(T-1)(T+3)(T+5)\cdots(T+2j-1)}$$

$$E(r_1^c)^{2j} = \frac{1 \cdot 3 \cdot 5 \cdots (2j-1)}{(T+1)(T+3)(T+5)\cdots(T+2j-1)}$$

so that, for example,

$$E(r_1^c)^3 = \frac{-3}{(T-1)(T+3)} \quad E(r_1^c)^4 = \frac{3}{(T+1)(T+3)}$$

Moran (1948) provided analogous results for the variance of the non-circular correlation r_1 from (9.8),

$$V(r_1) = \frac{(T-2)^2}{(T-1)^3}$$

and for the third moment:

$$E(r_1)^3 = -\frac{3(T^3 - 2T^2 + 2T - 5)}{(T-1)^4(T+3)}$$

although higher-order moments were not given. Analogous results when the mean of X_t was estimated, rather than being known, were also provided by these authors.

Moran (1967a) showed that, irrespective of the distribution assumed for X_t,

$$V(r_1) \leq \frac{T(T-2)}{(T-1)^3} = \frac{(T-2)^2}{(T-1)^3} + \frac{2(T-2)}{(T-1)^3}$$

Thus the variance under normality almost attains this maximum: for example, for $T = 20$ the maximum variance is 0.0525 while the normal variance is 0.0472. Moran (1967b, 1970) showed that, for long-tailed distributions, $V(r_1)$ could be much smaller than this: from Moran (1970, Table 1) the variance could be as low as 0.0307 for a sample of $T = 20$ drawn from a gamma distribution with index $-\frac{1}{2}$, and simulation experiments by Cox (1966) showed that it could be even lower for the Cauchy distribution (although this table also shows that a uniform distribution has a variance even closer to the maximum than the normal). In general, though, a variance based on a normal assumption will work quite well except for long tailed distributions.

Distribution of the first-order serial correlation coefficient under independence

9.13 Anderson (1942) considered the distribution, under the assumption of independence and normality of X_t, of the cyclic first-order correlation coefficient r_1^c, which will now be denoted simply as r. Anderson used a result from Cochran (1934) that states that every quadratic form $\sum_{t=1}^{T}\sum_{s=1}^{T} a_{ts}X_tX_s$ is distributed as $\sum_{i=1}^{m} \zeta_i \nu_i$, where m is the rank of the matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & \dots & a_{1T} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{T1} & \dots & \dots & a_{TT} \end{bmatrix}$$

of the quadratic form, $\nu_1,\ldots,\nu_m$ are independently distributed as $\chi^2(1)$, and $\zeta_1,\ldots,\zeta_m$ are the non-zero roots of the characteristic equation of $\mathbf{A}$ (if each ζ_i appears k_i times as a root then ν_i will be distributed as $\chi^2(k_i)$). Anderson was then able to show that r can be expressed as the transformation

$$\begin{aligned} r &= \sum_{i=1}^{\frac{1}{2}(T-1)} \zeta_i\nu_i \Big/ \sum_{i=1}^{T-1} \nu_i \qquad && T \text{ odd} \\ &= \sum_{i=1}^{\frac{1}{2}(T-2)} \zeta_i\nu_i \Big/ \left(\sum_{i=1}^{T-1} \nu_i + \nu\right) && T \text{ even} \end{aligned}$$

where $\zeta_i = \cos 2\pi i/T$, $i = 1,2,\ldots,T-1$, ν_i is distributed as $\chi^2(2)$ and ν is distributed as $\chi^2(1)$.

Anderson first considered the case of $T=6$, for which the first-order serial correlation, now denoted $r(6)$, can be expressed as

$$r(6) = \frac{\zeta_1\nu_1 + \zeta_2\nu_2 - \nu}{\nu_1 + \nu_2 - \nu}$$

where

$$\zeta_1 = \cos\frac{2\pi}{6} = \frac{1}{2} \quad \zeta_2 = -\frac{1}{2}$$

The density function of the ν's is given by

$$p(\nu_1,\nu_2,\nu) = \frac{1}{4\sqrt{2\pi}}\nu^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}V_6\right)$$

on defining $V_6 = \nu_1 + \nu_2 + \nu$. The density function for $r(6)$ is then the piecewise function

$$\begin{aligned} p(r(6)) &= \frac{3}{2}\frac{(\zeta_1 - r(6))^{\frac{1}{2}}}{(\zeta_1-\zeta_2)(1+\zeta_1)^{\frac{1}{2}}} && \zeta_2 \le r(6) \le \zeta_1 \\ &= \frac{3}{2}\frac{(\zeta_1 - r(6))^{\frac{1}{2}}}{(\zeta_1-\zeta_2)(1+\zeta_1)^{\frac{1}{2}}} + \frac{3}{2}\frac{(\zeta_2 - r(6))^{\frac{1}{2}}}{(\zeta_2-\zeta_1)(1+\zeta_2)^{\frac{1}{2}}} && -1 \le r(6) \le \zeta_2 \end{aligned}$$

and the cumulative probability function has the same general form:

$$\begin{aligned} P(r(6) > r') &= \frac{(\zeta_1 - r')^{\frac{3}{2}}}{(\zeta_1-\zeta_2)(1+\zeta_1)^{\frac{1}{2}}} && \zeta_2 \le r(6) \le \zeta_1 \\ &= \frac{(\zeta_1 - r')^{\frac{3}{2}}}{(\zeta_1-\zeta_2)(1+\zeta_1)^{\frac{1}{2}}} + \frac{(\zeta_2 - r')^{\frac{3}{2}}}{(\zeta_2-\zeta_1)(1+\zeta_2)^{\frac{1}{2}}} && -1 \le r(6) \le \zeta_2 \end{aligned}$$

These results generalize to the following functions which, for values of r between the roots ζ_i, split into separate analytical expressions which are continuous although their derivatives may not necessarily be so.

$$p(r) = \frac{T-3}{2}\sum_{i=1}^{n}\left(\frac{(\zeta_i - r)^{\frac{1}{2}(T-5)}}{\prod_{j=1,i\neq j}^{\frac{1}{2}(T-1)}(\zeta_i - \zeta_j)}\right) \qquad \zeta_{n+1} \le r \le \zeta_n \quad T \text{ odd}$$

$$= \frac{T-3}{2}\sum_{i=1}^{n}\left(\frac{(\zeta_i - r)^{\frac{1}{2}(T-5)}}{\prod_{j=1,i\neq j}^{\frac{1}{2}(T-1)}(\zeta_i - \zeta_j)(1+\zeta_1)^{\frac{1}{2}}}\right) \qquad \zeta_{n+1} \le r \le \zeta_n \quad T \text{ even}$$

$$P(r > r') = \sum_{i=1}^{n}\left(\frac{(\zeta_i - r')^{\frac{1}{2}(T-3)}}{\prod_{j=1,i\neq j}^{\frac{1}{2}(T-1)}(\zeta_i - \zeta_j)}\right) \qquad \zeta_{n+1} \le r \le \zeta_n \quad T \text{ odd}$$

$$= \sum_{i=1}^{n}\left(\frac{(\zeta_i - r)^{\frac{1}{2}(T-3)}}{\prod_{j=1,i\neq j}^{\frac{1}{2}(T-1)}(\zeta_i - \zeta_j)(1+\zeta_1)^{\frac{1}{2}}}\right) \qquad \zeta_{n+1} \le r \le \zeta_n \quad T \text{ even}$$

$$n = 1, 2, \ldots, \left[\tfrac{1}{2}(T-1)\right]$$

The density functions for $T=6$ and 7 are shown in Figure 9.2, while the density function for $T=15$ is shown in Figure 9.3, which by this sample size is clearly approaching normality centred on a mean of $-1/(T-1) = -0.0714$. Also shown is a normal approximation with this mean and a variance given by the cyclic definition, $V(r(15)) = T(T-2)^3/(T-1)^2(T+1) = 0.0574$. Exact and approximate critical values were given in Dixon (1944, page 127) and show that the normal approximation is adequate for samples as small as 20 for a 5% significance level and for around 45 when using a 1% level.[5]

9.14 Anderson (1942) showed that the distribution of r_k was identical to that of r_1 when T and k are prime to each other, i.e., when T/k is an integer. In general, the distribution of r_k can be derived for any k and T by using only those distributions for which k is a factor of T and Anderson derived the distributions and accompanying critical values for $T/k = 2$, 3 and 4. For $T > 4k$ the critical values for r_1 can probably be used for $k \ge 4$.

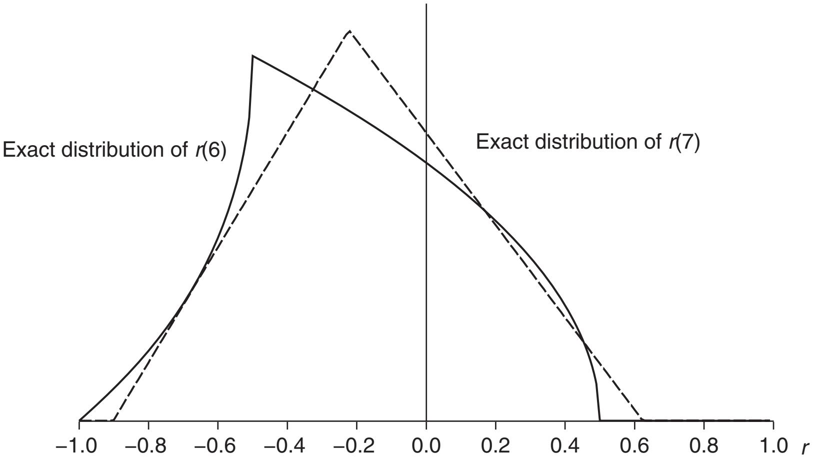


Figure 9.2 Exact distributions of the first-order serial correlation coefficient for $T = 6$ and $T = 7$

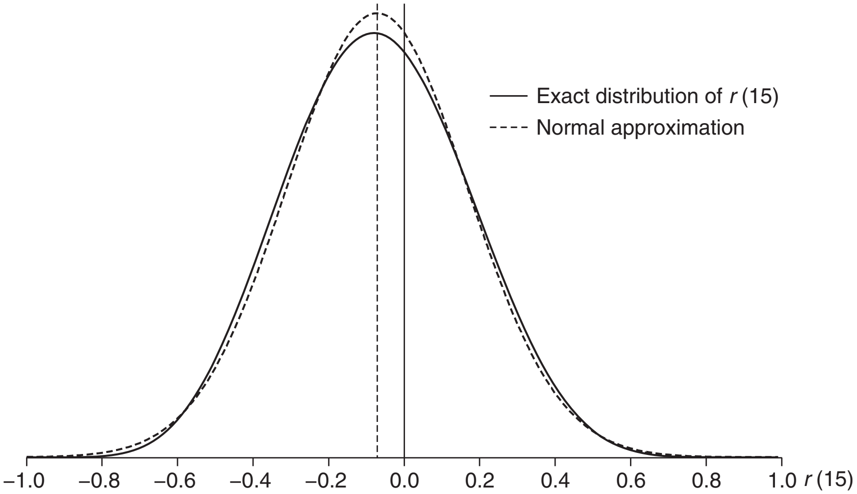


Figure 9.3 Exact distribution of the first-order serial correlation coefficient for $T = 15$ with its normal approximation

9.15 By extending the approach of Koopmans (1942), Dixon (1944) and Rubin (1945) provided an approximation to $p(r)$ which has the reasonably simple form[6]

$$\bar{p}(r) = K(1 - r^2)^{\frac{1}{2}(T-1)} = K(1 + r)^{\frac{1}{2}(T-1)}(1 - r)^{\frac{1}{2}(T-1)} \tag{9.11}$$

where

$$K = \frac{\Gamma(\frac{1}{2}(T+2))}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}(T+1))} = \frac{1}{B(\frac{T+1}{2},\frac{1}{2})}$$

with $\Gamma(a) = \int_0^\infty \exp(-y)y^{a-1}dy$ and $B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ being the standard gamma and beta functions, respectively.

Distribution of the serial correlation coefficient for a Markov process

9.16 If the true value of ρ_1 is $\rho \neq 0$, Madow (1945) extended the results of **§9.13** to

$$\begin{aligned} p(r) &= K_1 \sum_{i=1}^{n} \left(\frac{(\zeta_i - r)^{\frac{1}{2}(T-5)}}{\prod_{j=1,i\neq j}^{\frac{1}{2}(T-1)} (\zeta_i - \zeta_j)} \right) \qquad \zeta_{n+1} \leq r \leq \zeta_n \quad T \text{ odd} \\ &= K_1 \sum_{i=1}^{n} \left(\frac{(\zeta_i - r)^{\frac{1}{2}(T-5)}}{\prod_{j=1,i\neq j}^{\frac{1}{2}(T-1)} (\zeta_i - \zeta_j)(1+\zeta_1)^{\frac{1}{2}}} \right) \qquad \zeta_{n+1} \leq r \leq \zeta_n \quad T \text{ even} \end{aligned} \tag{9.12}$$

where K_1 is a function of r, ρ, σ^2 and T. Since the moments and percentiles of this distribution are difficult to obtain, Leipnik (1947) extended the Dixon–Koopmans approach to obtain

$$\overline{p}_\rho(r) = K \frac{(1-r^2)^{\frac{1}{2}(T-1)}}{(1+\rho^2-2\rho r)^{-\frac{T}{2}}} \tag{9.13}$$

which reduces to the distribution (9.11) when $\rho = 0$. This has a maximum when

$$r = r_{\max} = \frac{1}{2\rho(T-2)}((1+\rho^2)(T-1) - \sqrt{T(T-2)(1-\rho^2)^2 + (1+\rho^2)^2})$$

and it follows that $1 > |r_{\max}| > |\rho|$ and that $r_{\max} \rightarrow \rho$ asymptotically. Figure 9.4 shows the distribution (9.13) for $T = 15$ and various values of ρ, along with the envelope of $\overline{p}_\rho(r)$, $K(1-r^2)^{-\frac{1}{2}}$, this being obtained by differentiating (9.13) with respect to ρ and then eliminating this parameter. It is clearly seen that, for $|\rho|$ near 1, the distribution becomes highly concentrated about $r_{\max}$.

Leipnik (1947) showed that the mean and variance of (9.13) were given by

$$E_\rho(r) = \frac{T\rho}{T+2}$$

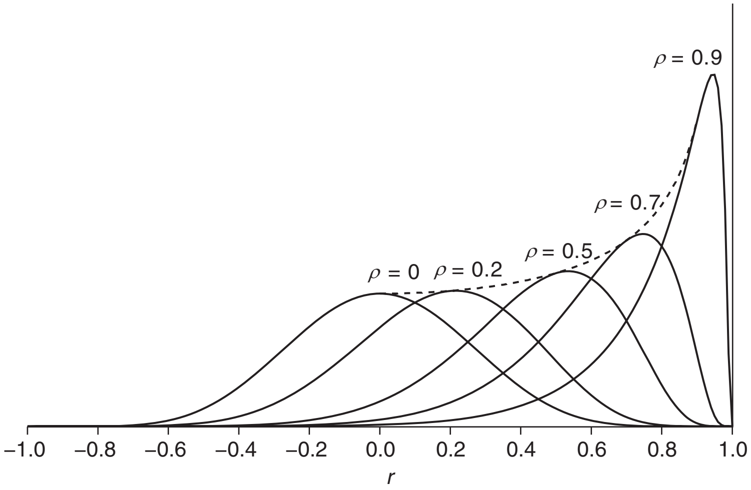


Figure 9.4 Distribution of the circular serial correlation coefficient for $T = 15$ for various values of ρ when the mean is known

and

$$V_\rho(r) = \frac{1}{T+2}\left(1 - \frac{\rho^2 T(T-2)}{(T+2)(T+4)}\right) \sim \frac{1-\rho^2}{T}$$

which confirm that r is a consistent estimate of ρ. Daniels (1956) showed that the error incurred in using the approximation (9.13) rather than the exact distribution (9.12) was negligible in the tails of the distribution when $T=20$ and $\rho=0$ but could be of concern in the upper tail of the distribution when $\rho=0.5$.

Higher-order moments of the distribution (9.13) were obtained by Jenkins (1954a) and Kendall (1957), where exact expressions are given. In general, the jth moment will be a polynomial of order j in ρ, with even-order moments containing only even powers of ρ and odd-order moments containing only odd powers of ρ.

9.17 Of course, it must be remembered that these distributions are for the cyclic definition of serial correlation when the mean of X is known (and can therefore be assumed to be zero). When the mean is unknown and has to be estimated by the sample mean, Daniels (1956) showed that an approximation to the distribution of r is

$$h(r) = \frac{K'}{(T(1-\rho)+1+\rho)} \frac{(1-r)(1-r^2)^{\frac{1}{2}T-1}}{(1-2\rho r+\rho^2)^{\frac{1}{2}(T-1)}}$$

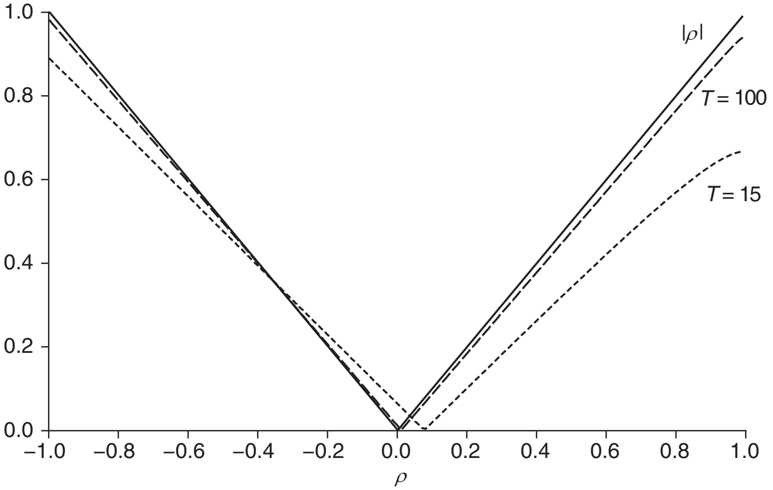


Figure 9.5 $|E(r)|$ for $T = 15$ and 100 with $|\rho|$ shown for comparison

where $K' = 1/B(\frac{T}{2}, \frac{1}{2})$, which will be accurate to $O(T^{-\frac{3}{2}})$. Explicit expressions for the first four moments were provided by Kemp (1970): for example, the mean is

$$E(r) = \frac{-1 + (T-1)\rho - \frac{(T-1)T\rho^2}{T+3}}{T(1-\rho) + 1 + \rho}$$

Figure 9.5 plots $|E(r)|$ against ρ for samples of size $T = 15$ and 100, while Figure 9.6 shows this distribution using the same values of T and ρ as Figure 9.4. It is seen that the effect of having to estimate the mean is to shift the distribution leftwards and to increase the variance and that $|E(r)| < |\rho|$ except in the region $-0.35 \leq \rho \leq 0.04$ for $T = 15$ and $-0.33 \leq \rho \leq 0.01$ for $T = 100$, with the size of the absolute bias $|\rho - E(r)|$ declining with T.

Distributions of non-circular serial correlation statistics

9.18 Watson and Durbin (1951) argued that the circular conception of the stochastic process generating X_t, as embodied in the cyclic definition (9.10) of r_1^c that has been used in §§**9.11–9.17**, was rarely plausible in practice. They thus relaxed the assumptions of circularity and known (zero) mean and considered the following statistic for testing independence, i.e., $\rho = 0$:

$$d = \frac{\sum_{i=2}^{T} (X_{i-1} - X_i)^2 - (X_n - X_{n+1})^2}{\sum_{i=1}^{T} (X_i - \overline{X})^2} \quad T = 2n$$

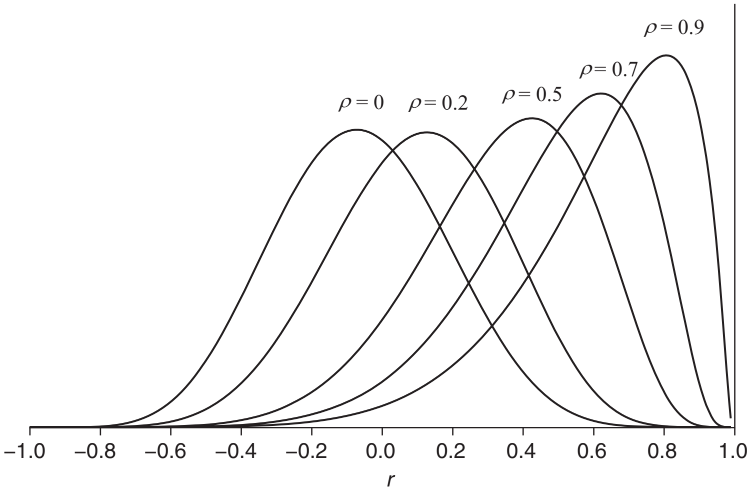


Figure 9.6 Distribution of the circular serial correlation coefficient for $T = 15$ for various values of ρ when the mean is unknown and estimated by the sample mean

where $n = T/2$ if T is even and $n = (T - 1)/2$ if T is odd. The exclusion of the central squared difference in the numerator sum is a device to give the statistic a known distribution. By extending the results of Anderson (1942) (cf. §**9.13**), Watson and Durbin (1951) showed that, for $\zeta_i = 4\sin^2(n - i)\pi/2n$, the distribution of d is

$$P(d > d') = \sum_{i=1}^{s} \frac{(\zeta_i - d')^{n-\frac{3}{2}}}{\zeta_i^{\frac{1}{2}} \prod_{j=1, j\neq i}^{n-1} (\zeta_i - \zeta_j)} \quad \zeta_{s+1} \leq d' \leq \zeta_s \quad s = 1, 2, \ldots, n - 1$$

Watson and Durbin provided 5% critical values for d for various values of T that may be used for testing independence against the alternative of positive serial correlation, $\rho > 0$. This statistic was extended by Durbin and Watson (1950, 1951, 1971) to test for first-order serial correlation in regression models, becoming probably the most recognized test statistic in econometrics (see Durbin, 1982).

9.19 Daniels (1956) investigated the distribution of ρ for a non-circular Markov process with unknown mean using the following estimator of the first-order serial correlation, which he termed the *intra-class* correlation coefficient

$$r = \frac{\sum_{t=2}^{T} x_{t-1}x_t}{\sum_{t=2}^{T-1} x_t^2 + \frac{1}{2}(x_1 + x_T)}$$

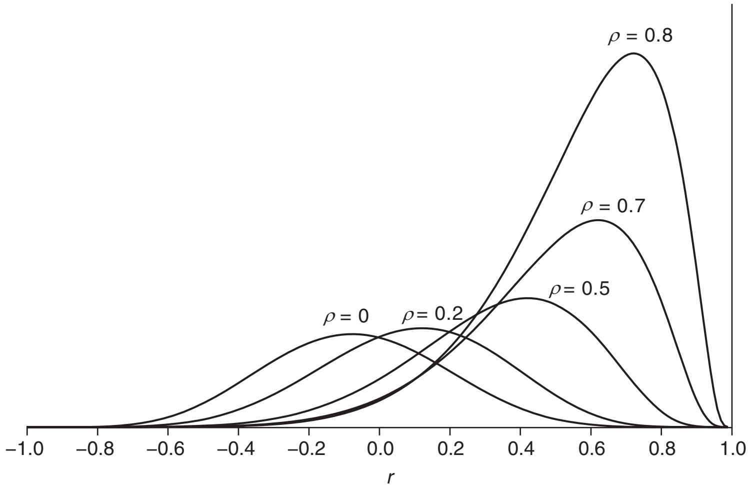


Figure 9.7 Distribution of the non-circular serial correlation coefficient for $T = 15$ for various values of ρ when the mean is unknown and estimated by the sample mean

where, for $\overline{X} = (T-1)^{-1}(\sum_{t=2}^{T-1} X_t + \frac{1}{2}(X_1 + X_T))$, $x_t = X_t - \overline{X}$. The distribution of r is then given by

$$h(r) = \frac{\Gamma(\frac{1}{2}T' + \frac{3}{2})}{2\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}T')(T'(1-\rho)+1+\rho)} \frac{(1-r)(1-r^2)^{\frac{1}{2}T'-1}}{(1-2\rho r+\rho^2)^{\frac{1}{2}(T'-1)}}$$

where $T' = T - 1 + \rho^2/(1-\rho^2)$. Figure 9.7 shows this distribution for $T = 15$ for various values of ρ: in comparison to Figures 9.4 and 9.6, relaxing the assumption of a known mean and then circularity leads to a much greater overlap between the individual distributions and, hence, less precise inference. The distribution is derived by Daniels (1956) using a saddlepoint approximation that, while accurate for small values of ρ, becomes undefined for large values, so that no curve for $\rho = 0.9$, for example, can be drawn.

Distribution of partial serial correlations

9.20 Daniels (1956) and Jenkins (1954b, 1956) extended these results to analyse the small sample distribution of the partial serial correlation, $r_{k\cdot}$, for which a large sample approximation was given in §**9.10**. For a fitted mean, the approximate density function of the circularly defined partial serial correlation is

$$Q \prod_{k\,\text{odd}} (1-r_{k\cdot})(1-r_{k\cdot}^2)^{\frac{1}{2}(T-3)} \prod_{k\,\text{even}} (1-r_{k\cdot})^2(1-r_{k\cdot}^2)^{\frac{1}{2}(T-3)}$$

where Q is a function of both T and the coefficients of the underlying autoregression. To test the hypothesis that, in the AR(p) model (8.20), $a_p = 0$, $r_{p\cdot}$ can be taken to have a distribution with density proportional to

$$(1 - r_{p\cdot})(1 - r_{p\cdot}^2)^{\frac{1}{2}(T-3)} \quad \text{when } p \text{ is odd}$$

and

$$(1 - r_{p\cdot})^2(1 - r_{p\cdot}^2)^{\frac{1}{2}(T-3)} \quad \text{when } p \text{ is even.}$$

Jenkins (1954b) showed that $E(r_{2\cdot}) \sim -2/(T-1)$ and $E(r_{2\cdot}^2) \sim 1/(T-2)$, while Daniels (1956) obtained $E(r_{2\cdot}) \sim -2/(T+1)$ and $E(r_{2\cdot}^2) \sim (T+5)/(T+1)(T+2)$, which agree with Jenkins to $O(T^{-2})$ and $O(T^{-3})$ respectively. No results for non-circularly defined partial autocorrelations have been obtained, with Kendall, Stuart and Ord (1983, page 571) remarking that '(e)xcept in the Markov case it appears that only circularly defined statistics and processes are reasonably tractable. Daniels' method can be extended to the non-circular case, but apparently nobody has yet had the stamina to embark on the labour involved'!

Estimation and inference in autoregressive models

9.21 Mann and Wald (1943) considered the estimation of the coefficients of an AR(p) process and the accompanying sampling theory of the estimators. For the model

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \tag{9.14}$$

it is assumed that ε_t is identically and independently distributed with zero mean and finite higher moments that all exist and that all the roots of the characteristic equation $z^p - \alpha_1 z^{p-1} - \cdots - \alpha_p = 0$ are less than unity in absolute value.[7] If it is further assumed that the ε_t are normally distributed with variance σ^2, Mann and Wald showed that the maximum likelihood estimators of $\alpha_1, \ldots, \alpha_p, \alpha_0$ coincide with the least squares estimators $\hat{\alpha}_1, \ldots, \hat{\alpha}_p, \hat{\alpha}_0$ and that, for large T, the AR(p) process (9.14)

> can be treated in exactly the same way as a classical regression problem where X_t is the dependent variable and $X_{t-1}, \ldots, X_{t-p}$ are the independent variables. That is to say, the estimates of the coefficients $\alpha_1, \ldots, \alpha_p, \alpha_0$, as well as the joint limiting distribution of these estimates, are the same as if [9.14] were treated as a classical regression problem. Hence, the joint limiting distribution of $\sqrt{T}(\hat{\alpha}_1 - \alpha_1), \ldots, \sqrt{T}(\hat{\alpha}_p - \alpha_p)$ and $\sqrt{T}(\hat{\alpha}_0 - \alpha_0)$ is a multivariate normal distribution with zero means and a finite covariance matrix. The covariance between $\xi_i = \sqrt{T}(\hat{\alpha}_i - \alpha_i)$ and $\xi_j = \sqrt{T}(\hat{\alpha}_j - \alpha_j)(i, j = 0, 1, \ldots, p))$ can be obtained as follows: Denote $(1/T)\sum_{t=1}^{T} X_{t-i}X_{t-j}$ by $D_{ijT}(i, j = 1, \ldots, p)$,

> $(1/T)\sum_{t=1}^{T} X_{t-i}$ by $D_{i0T} = D_{0iT}$ and let $D_{00T} = 1$. Furthermore, let $\|c_{ijT}\| = \|D_{ijT}\|^{-1}(i, j = 0, 1, \ldots, p)$ and $s^2 = (1/T)\sum_{t=1}^{T}(X_t - \hat{\alpha}_1 X_{t-1} - \cdots - \hat{\alpha}_p X_{t-p})^2$. Then the limit covariance between ξ_i and ξ_j is equal to the stochastic limit of $s^2 c_{ijT}$. Thus, for large T the covariance of ξ_i and ξ_j can be replaced by the quantity $s^2 c_{ijT}$ which can be calculated from the observations. (Mann and Wald, 1943, page 217: notation altered for consistency)

Kendall (1949) considered further the estimation of the coefficients of an autoregressive scheme. Using the second-order autoregression as an example, then, with $k = 2$, equations (7.25) become

$$r_1 + a_1 + r_1 a_2 = 0$$

$$r_2 + a_1 r_1 + a_2 = 0$$

from which we obtain (cf. (8.7))

$$a_1 = -\frac{r_1(1 - r_2)}{1 - r_1^2} \quad a_2 = \frac{r_1^2 - r_2}{1 - r_1^2}$$

which are asymptotically equivalent to the least squares estimates (i.e, they are identical if the serial correlations are estimated as $r_k = \sum_{t=1}^{T-k} x_t x_{t+k} / \sum_{t=1}^{T} x_t^2$). Kendall also considered using the complete set of equations (7.25)–(7.26), which he termed the *Yule–Walker* equations: for $k = 2$ these are

$$r_1 + a_1 + r_1 a_2 = 0$$

$$r_2 + a_1 r_1 + a_2 = 0$$

$$r_3 + a_1 r_2 + a_2 r_1 = 0$$

$$\vdots$$

The solution of the first two equations yields the least squares estimates. The least squares solution to the first m of these equations is obtained by minimizing

$$\sum_{i=1}^{m}\left(\sum_{j=0}^{2}(a_j r_{i-j})^2\right)$$

For example, using the first three Yule–Walker equations leads to the pair of equations

$$(r_1 + r_1 r_2 + r_2 r_3) + a_1(1 + r_1^2 + r_2^2) + a_2(2r_1 + r_1 r_2) = 0$$

$$(r_1^2 + r_2 + r_1 r_3) + a_1(2r_1 + r_1 r_2) + a_2(1 + 2r_1^2) = 0$$

and the solutions

$$a_1 = \frac{(2r_1 + r_1 r_2)(r_1^2 + r_2 + r_1 r_3) - (1 + 2r_1^2)(r_1 + r_1 r_2 + r_2 r_3)}{(1 + 2r_1^2)(1 + r_1^2 + r_2^2) - (2r_1 + r_1 r_2)^2}$$

$$a_2 = \frac{(2r_1 + r_1 r_2)(r_1 + r_1 r_2 + r_2 r_3) - (1 + r_1^2 + r_2^2)(r_1^2 + r_2 + r_1 r_3)}{(1 + 2r_1^2)(1 + r_1^2 + r_2^2) - (2r_1 + r_1 r_2)^2}$$

From a set of simulation experiments Kendall concluded that this approach provided no improvement over the least squares approach of solving the first two Yule–Walker equations, particularly for large values of m, and he suggested that this was because the higher-order serial correlations were so affected by sampling variability that any gain from using these additional equations was more than offset by the increase in sampling unreliability.

Kendall considered two further estimation methods. The first was a method of moments type estimator in which the first k covariances of ε_t were set to zero and the resulting expressions solved, while the second extended the approach of Quenouille (1947b) in which the expressions for R_s in §**9.6** were set to zero and solved for the autoregressive coefficients. Again, neither method proved superior to least squares, which has since become the standard method of estimating the coefficients of autoregressions.

Durbin (1960) showed that, for the pth order autoregression (9.17), the first p Yule–Walker equations

$$\begin{aligned}
r_1 + a_1 + r_1 a_2 + \cdots + r_{p-1} a_p &= 0\\
r_2 + r_1 a_1 + a_2 + \cdots + r_{p-2} a_p &= 0\\
&\vdots\\
r_p + r_{p-1} a_1 + \cdots + a_p &= 0
\end{aligned}$$

may be solved by a pivotal reduction to provide the recurrence relations

$$a_{ii} = -\frac{r_i + a_{i-1,1} r_{i-1} + a_{i-1,2} + \cdots + a_{i-1,i-1} r_1}{1 + a_{i-1,1} r_1 + \cdots + a_{i-1,i-1} r_{i-1}} \qquad i = 1, \ldots, p$$

$$a_{i,j} = a_{i-1,j} + a_{ii} a_{i-1,i-j} \qquad j = 1, \ldots, j - 1$$

using $a_{11} = -r_1$ as the starting value. The $a_{i1}, \ldots, a_{ii}$ are the coefficients of the best-fitting autoregressive model of order i, while $-a_{22}, \ldots, -a_{pp}$ are estimates of the partial correlation coefficients $r_{2\cdot}, \ldots, r_{p\cdot}$.

Small sample bias in the estimation of autoregressive models

9.22 Mann and Wald's (1943) analysis outlined in §**9.21** considered the limiting distribution of the least squares estimates of an autoregressive process, showing that they coincided with the maximum likelihood estimates. The (very) small sample properties of these estimators were considered by Hurwicz (1945), who showed that it makes a considerable difference whether the 'initial value' – the value taken by x_0 – was taken as fixed or stochastic, a distinction that loses its relevance for large sample sizes.

Focusing attention on the first-order case, consider the model

$$x_t = \alpha x_{t-1} + \varepsilon_t, \quad t = 2, 3, \dots, T$$

where ε_t is as in (9.14) but with variance $\sigma^2 = 1$ for simplicity. In the stochastic initial value case, $E(x_0) = E(\varepsilon_0) = 0$,

$$E(x_t) = 0 \quad E(x_t^2) = \frac{1}{1-\alpha^2} \quad t = 1, 2, \dots, T$$

and

$$E(x_t|x_{t-1}) = \alpha x_{t-1} \quad t = 2, 3, \dots, T$$

If the least squares estimator of α using T observations is

$$\hat{\alpha}_T = \frac{\sum_{t=2}^{T} x_t x_{t-1}}{\sum_{t=2}^{T} x_{t-1}^2}$$

then it will be said to be unbiased if $E(\hat{\alpha}_T) = \alpha$ for all α and for all T. Hurwicz (1945) thus considered the case of $T = 3$:

$$\hat{\alpha}_3 = \frac{x_1 x_2 + x_2 x_3}{x_1^2 + x_2^2}$$

and showed that

$$E(\hat{\alpha}_3) = \frac{\alpha}{2}\left(1 + \frac{1-\sqrt{1-\alpha^2}}{\alpha^2}\right)$$

By defining $\beta = \alpha^2$ and $N_T(\beta) = E(\hat{\alpha}_T)/\alpha$, the 'relative bias' for $T = 3$ is

$$N_3(\beta) = \frac{1}{2}\left(1 + \frac{1-\sqrt{1-\beta}}{\beta}\right)$$

and a plot of this function is shown in Figure 9.8. Defining

$$N_T(0) \equiv \lim_{\beta \to 0} N_T(\beta)$$

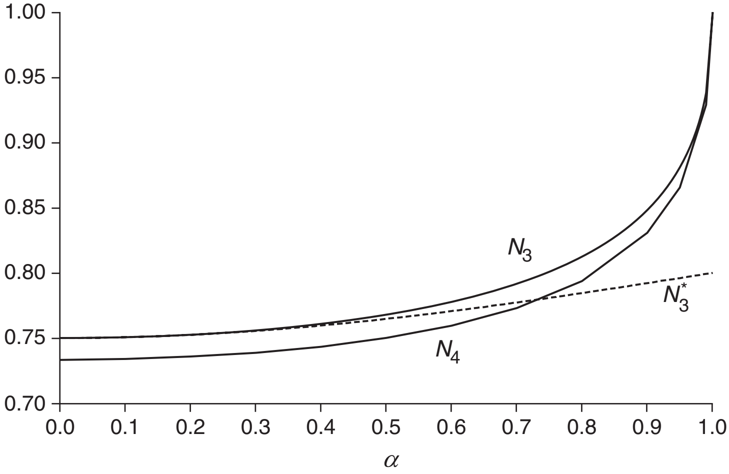


Figure 9.8 Relative bias of estimators of α for $T = 3$ and 4

then we see that $N_3(0) = 3/4$ and, although $N_3(1) = 1$, this convergence is very slow: for example, $N_3(0.94^2) = 0.875$. Thus, when the initial value is stochastic, $\hat{\alpha}_3$ is a biased estimate of α.

When x_0 is fixed, the model becomes

$$x_t = \alpha x_{t-1} + \varepsilon_t, \quad t = 1, 2, \ldots, T$$

and the least squares estimator (which is now also the maximum likelihood estimator) is

$$\hat{\alpha}_T^* = \frac{\sum_{t=1}^T x_t x_{t-1}}{\sum_{t=1}^T x_{t-1}^2}$$

Note that if $x_0 = 0$, $\hat{\alpha}_T^* = \hat{\alpha}_T$. With this particular initial value, then, on defining $N_T^*(\beta) = E(\hat{\alpha}_T^*)/\alpha$, Hurwicz showed that

$$N_3^*(\beta) = \frac{3 + \beta}{4 + \beta}$$

which is also plotted in Figure 9.8. It is seen that, although $N_3^*(0) = 3/4 = N_3(0)$, $N_3^*(1) = 4/5$ and so $\hat{\alpha}_3^*$ is biased for all values of α. Hurwicz did not treat the case when $x_0 \neq 0$, but conjectured that, for given β and T, $N_T^*(\beta) \to 1$ as x_0 becomes numerically large.

Hurwicz also derived an expression for $N_4(\beta)$, which depends upon elliptic integrals of the first kind (see Hurwicz, 1945, equations (3.37)–(3.39)), and showed that $N_4(0) = 11/15 < N_3(0)$ and $N_4(1) = 1$. This function is also plotted in Figure 9.8 using the values provided by Hurwicz (1945, Table 1) and it is seen that $N_4(\beta) < N_3(\beta)$ for all β and that the convergence of $N_4(\beta)$ to 1 is also very slow as $\beta \to 1$.

Hurwicz then derived the more general result

$$N_T(0) = N_T^*(0) = \frac{T^2 - 2T + 3}{(T-1)(T+1)}$$

and, using a third-order Maclaurin expansion, obtained the following approximation for $N_T(\beta)$:

$$\tilde{N}_T(\beta) = N_T(0) + N_T'(0)\beta + \tfrac{1}{2}N_T''(0)\beta^2 + \cdots$$

where

$$N_T'(0) = \frac{2(T^2 - 8T + 21)}{(T-1)(T+1)(T+3)(T+5)}$$

and

$$N_T''(0) = \frac{4(T^4 + 24T^3 + 98T^2 - 264T - 99)}{(T-1)(T+1)(T+3)(T+5)(T+7)(T+9)}$$

While this does not give a very good approximation for values of β near 1 (for example, $\tilde{N}_4(1) = 0.7765$ rather than 1), it is more accurate for smaller values: for $|\alpha| = 0.5$, $\tilde{N}_4(0.25) = 0.7380$ compared to the true value $N_4(0.25) = 0.7501$. Figure 9.9 plots $N_T(\beta)$ for various values of α for a wide range of T values. The function

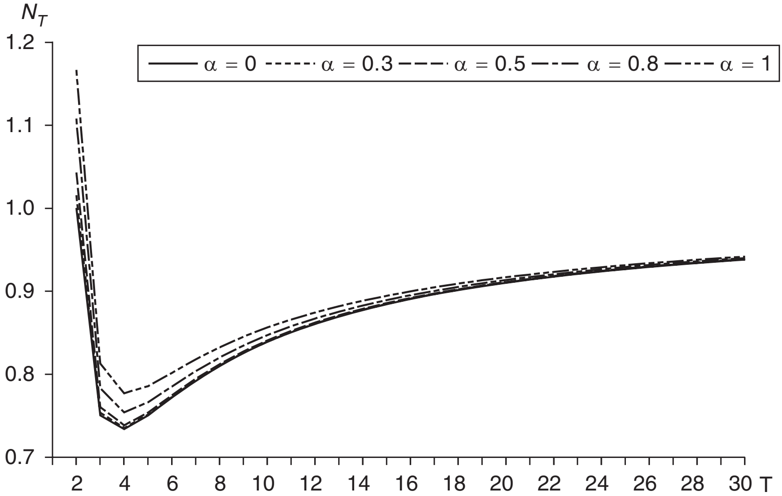


Figure 9.9 Relative bias of $\hat{\alpha}_T$ for various values of α and T

reaches a minimum at $T = 4$, after which it slowly increases, with calculations showing that $N_{500}(\beta)$ exceeds 0.996 for all values of β.

Estimation and inference in moving average models

9.23 While the estimation of autoregressive models has been shown to be the focus of great attention during the 1940s and 1950s, much less progress was being made on the estimation of moving average schemes and hence, unsurprisingly, on mixed autoregressive moving average models. Whittle (1953a, 1954) developed a large sample approach to the estimation of moving average models that, while being a complete solution, was extremely difficult to implement in practice. The search was thus on for feasible estimators that had satisfactory properties and this led to the approach proposed by Durbin (1959) and extended by Walker (1961).

9.24 We shall focus attention on estimating the parameter β in the first-order moving average model

$$x_t = \varepsilon_t + \beta\varepsilon_{t-1} \quad t = 1, 2, \ldots, T \tag{9.15}$$

where ε_t is as defined in (9.14) and it is assumed that $|\beta| < 1$, so that the moving average is 'regular' (cf. §**7.21**). A perhaps obvious estimator is to use the result that $\rho_1 = \beta/(1+\beta^2)$ (obtained by setting $k = h = 1$ in (7.34)), solve the quadratic $r_1\tilde{\beta}^2 - \tilde{\beta} + r = 0$, and use the regular solution $|\tilde{\beta}| < 1$. Whittle (1953a), however, showed that this estimator was very inefficient but his proposed adjustment was extremely complicated. Durbin (1959) thus considered the infinite autoregressive representation (cf. §**7.21**) truncated at lag p

$$x_t + \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} = \varepsilon_t$$

where $\alpha_i = (-\beta)^i$. This finite representation can be made as close as desired to the infinite autoregression by taking p sufficiently large. Durbin showed that an approximate maximum likelihood estimator of β is given by

$$\hat{\beta} = -\frac{\sum_{k=0}^{p-1} \hat{\alpha}_k \hat{\alpha}_{k+1}}{\sum_{k=0}^{p} \hat{\alpha}_k^2} \tag{9.16}$$

where the $\hat{\alpha}_k$ are the least squares estimates of the α_k (taking $\hat{\alpha}_0 = 1$). Moreover, for sufficiently large p the asymptotic variance of $\hat{\beta}$ is $T^{-1}(1-\beta^2)$, which was shown by Whittle (1953a) to be the minimum asymptotic variance of all consistent estimators under the assumption of normality of ε_t. Without the assumption of normality, the efficiency property is no longer assured.

To test the hypothesis $\beta = \beta_0$, the statistic $\sqrt{T}(\hat{\beta} - \beta_0)(1 - \beta_0^2)^{-\frac{1}{2}} \sim N(0, 1)$ may be used, while to assess the goodness-of-fit of the model (9.15) the statistic

$$T\left((1 - \beta^2)\sum_{k=0}^{p} \alpha_k^2 - 1\right) \sim \chi^2(p - 1)$$

can be employed.

The extension to higher-order moving averages is straightforward, at least in theory. For the model

$$x_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \tag{9.17}$$

assumed to be regular, the estimators $\hat{\beta}_1, \ldots, \hat{\beta}_q$ of $\beta_1, \ldots, \beta_q$ are given by the solution of the linear equation system

$$\begin{bmatrix} \sum_{k=0}^{p} \hat{\alpha}_k^2 & \sum_{k=0}^{p-1} \hat{\alpha}_k \hat{\alpha}_{k+1} & \cdots & \sum_{k=0}^{p-q+1} \hat{\alpha}_k \hat{\alpha}_{k+q-1} \\ \sum_{k=0}^{p-1} \hat{\alpha}_k \hat{\alpha}_{k+1} & \sum_{k=0}^{p} \hat{\alpha}_k^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{k=0}^{p-q+1} \hat{\alpha}_k \hat{\alpha}_{k+q-1} & \cdots & \cdots & \sum_{k=0}^{p} \hat{\alpha}_k^2 \end{bmatrix}$$

$$\times \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_q \end{bmatrix} = -\begin{bmatrix} \sum_{k=0}^{p-1} \hat{\alpha}_k \hat{\alpha}_{k+1} \\ \sum_{k=0}^{p-2} \hat{\alpha}_k \hat{\alpha}_{k+2} \\ \vdots \\ \sum_{k=0}^{p-q} \hat{\alpha}_k \hat{\alpha}_{k+q} \end{bmatrix}$$

The asymptotic variance matrix of $\hat{\beta}_1, \ldots, \hat{\beta}_q$ is $T^{-1}\mathbf{V}_q$, where

$$\mathbf{V}_q = \begin{bmatrix} 1 - \beta_q^2 & \beta_1 - \beta_{q-1}\beta_q & \beta_2 - \beta_{q-2}\beta_q & \cdots & \beta_{q-1} - \beta_1\beta_q \\ \beta_1 - \beta_{q-1}\beta_q & \begin{array}{c} 1 + \beta_1^2 \\ -\beta_{q-1}^2 - \beta_q^2 \end{array} & & & \vdots \\ \beta_2 - \beta_{q-2}\beta_q & \begin{array}{c} \beta_1 + \beta_1\beta_2 \\ -\beta_{q-2}\beta_{q-1} \\ -\beta_{q-1}\beta_q \end{array} & \ddots & & \vdots \\ \vdots & & & \begin{array}{c} 1 + \beta_1^2 \\ -\beta_{q-1}^2 - \beta_q^2 \end{array} & \beta_1 - \beta_{q-1}\beta_q \\ \beta_{q-1} - \beta_1\beta_q & \cdots & \cdots & \beta_1 - \beta_{q-1}\beta_q & 1 - \beta_q^2 \end{bmatrix}$$

Thus, for $q = 1, 2, 3$,

$$\mathbf{V}_1 = 1 - \beta_1^2 \quad \mathbf{V}_2 = \begin{bmatrix} 1 - \beta_2^2 & \beta_1 - \beta_1\beta_2 \\ \beta_1 - \beta_1\beta_2 & 1 - \beta_2^2 \end{bmatrix}$$

$$\mathbf{V}_3 = \begin{bmatrix} 1-\beta_3^2 & \beta_1-\beta_2\beta_3 & \beta_2-\beta_1\beta_3 \\ \beta_1-\beta_2\beta_3 & 1+\beta_1^2-\beta_2^2-\beta_3^2 & \beta_1-\beta_2\beta_3 \\ \beta_2-\beta_1\beta_3 & \beta_1-\beta_2\beta_3 & 1-\beta_3^2 \end{bmatrix}$$

The hypothesis $\beta_k = \beta_{0k}$, $k = 1, \ldots, q$, may be tested using the statistic

$$T\sum_{i=1}^{q}\sum_{j=1}^{q} v_{q,0}^{ij}(\hat{\beta}_i - \beta_{0i})(\hat{\beta}_j - \beta_{oj}) \sim \chi^2(q)$$

where $v_{q,0}^{ij}$ is the *ij*th element of $\mathbf{V}_q^{-1}$ evaluated at $\beta_k = \beta_{0k}$, $k = 1, \ldots, q$. The goodness-of-fit of (9.17) can be assessed using

$$T\left(\sum_{k=0}^{p}\hat{\alpha}_k^2 + \sum_{j=1}^{q}\hat{\beta}_j\sum_{i=0}^{p-j}\hat{\alpha}_i\hat{\alpha}_{i+j} - 1\right) \sim \chi^2(p-q)$$

with large values of the statistic indicating that the fit is inadequate.

9.25 Durbin (1959) examined this method by simulating twenty series of length $T = 100$ from the model (9.15) with $\beta = 0.5$ and $\varepsilon_t \sim N(0,1)$, and computing $\hat{\beta}$ from (9.16) using fitted autoregressions with $p = 5$, i.e.,

$$\hat{\beta} = -\frac{\hat{\alpha}_1 + \hat{\alpha}_1\hat{\alpha}_2 + \cdots + \hat{\alpha}_4\hat{\alpha}_5}{1 + \hat{\alpha}_1^2 + \cdots + \hat{\alpha}_5^2}$$

He also compared this estimator with the simple estimator $\tilde{\beta}$ obtained from r_1 (when the roots of $r_1\tilde{\beta}^2 - \tilde{\beta} + r = 0$ are imaginary $\tilde{\beta}$ was taken to be one: this will occur when $r_1 > 0.5$). Table 9.3 shows the results obtained by recreating Durbin's simulation, for which summary statistics are

	r_1	$\hat{\beta}$	$\hat{\beta}_C$	$\tilde{\beta}$
Mean	0.365	0.457	0.461	0.471
Std. Dev.	0.084	0.097	0.098	0.184
SE of mean	0.019	0.022	0.022	0.041

$\hat{\beta}_C$ is the corrected estimator suggested by Durbin (but not actually used by him) to mitigate the downward bias observed in $\hat{\beta}$. It is obtained by using only the first $p-1$ terms in the divisor of (9.16). The results of the simulation accord well with those presented by Durbin (1959, Table 1). The mean value of r_1 is below, but reasonably close to, the true value of ρ_1, $0.5/(1+0.5^2) = 0.4$. The variance of $\hat{\beta}$, $0.097^2 = 0.0094$, is a little larger than the theoretical variance

Table 9.3 Twenty simulations of length $T = 100$ from a first-order moving average with $\beta = 0.5$

Series	r_1	$\hat{\beta}$	$\hat{\beta}_C$	$\tilde{\beta}$	Series	r_1	$\hat{\beta}$	$\hat{\beta}_C$	$\tilde{\beta}$
1	0.502	0.601	0.603	1.000	11	0.349	0.525	0.529	0.407
2	0.346	0.428	0.434	0.401	12	0.382	0.388	0.397	0.465
3	0.389	0.417	0.423	0.478	13	0.256	0.486	0.487	0.275
4	0.423	0.445	0.445	0.553	14	0.410	0.586	0.587	0.522
5	0.481	0.486	0.488	0.756	15	0.256	0.409	0.420	0.274
6	0.171	0.254	0.255	0.176	16	0.332	0.361	0.361	0.380
7	0.384	0.434	0.442	0.469	17	0.290	0.494	0.502	0.320
8	0.250	0.300	0.300	0.268	18	0.403	0.375	0.389	0.506
9	0.430	0.445	0.452	0.571	19	0.416	0.543	0.543	0.534
10	0.393	0.520	0.520	0.485	20	0.435	0.637	0.649	0.582

$(1-0.5^2)/100 = 0.0075$, and is considerably less than that of $\tilde{\beta}$. The downward bias in $\hat{\beta}$, which is not substantial here, is mitigated a little by Durbin's correction.[8]

9.26 Walker (1961) was concerned that the truncation of the infinite autoregression to a finite order p might lead to problems in some circumstances and thus proposed an extension of Durbin's method which had the added advantage of allowing bias adjustments to be made fairly straightforwardly. Again focusing on the first-order case, this approach is based on the result that, for large T, the joint distribution of $T^{\frac{1}{2}}(r_i - \rho_i)$ has the covariance matrix (Walker, 1961, equation (11))

$$\mathbf{W}(\rho) = \begin{bmatrix} 1-3\rho^2+4\rho^4 & 2\rho(1-\rho^2) & \rho^2 & 0 & 0 & \cdots \\ 2\rho(1-\rho^2) & 1+2\rho^2 & 2\rho & \rho^2 & 0 & \cdots \\ \rho^2 & 2\rho & 1+2\rho^2 & 2\rho & \rho^2 & \cdots \\ 0 & \rho^2 & 2\rho & 1+2\rho^2 & 2\rho & \cdots \\ \vdots & & & & & \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

where we write ρ for ρ_1 as usual. The estimate of ρ is then based on the first p serial correlations $r_1, \dots, r_p$ rather than simply taking $\hat{\rho} = r_1$:

$$\hat{\rho}^{(p)} = r_1 + \sum_{s=2}^{p} \hat{c}_{1s} r_s \qquad (9.18)$$

The 'weights' $\hat{c}_{12}, \dots, \hat{c}_{1p}$ are obtained from the equations

$$\sum_{s=2}^{p} c_{1s} w_{sj}(\rho) = -w_{1j}(\rho) \quad j = 2, \dots, p \qquad (9.19)$$

Table 9.4 Twenty simulations of length $T=100$ from a first-order moving average with $\beta=0.5$

Series	$\hat{\rho}^{(5)}$	$\hat{\beta}^{(5)}$	Series	$\hat{\rho}^{(5)}$	$\hat{\beta}^{(5)}$
1	0.425	0.557	11	0.410	0.522
2	0.353	0.413	12	0.353	0.413
3	0.384	0.467	13	0.365	0.434
4	0.367	0.437	14	0.437	0.588
5	0.403	0.505	15	0.319	0.360
6	0.304	0.339	16	0.329	0.376
7	0.359	0.423	17	0.416	0.536
8	0.279	0.305	18	0.374	0.450
9	0.395	0.490	19	0.411	0.524
10	0.405	0.510	20	0.454	0.640

where $w_{ij}(\rho)$ is the ijth element of $\mathbf{W}(\rho)$, through the following procedure. Taking r_1 as an initial value for ρ and setting $p=2$ in (9.19) yields

$$\hat{c}_{12}=-\frac{w_{12}(r_1)}{w_{22}(r_1)}=-\frac{2r_1(1-r_1^2)}{1+2r_1^2}$$

so that

$$\hat{\rho}^{(2)}=r_1+\hat{c}_{12}r_2$$

Equations (9.19) can then be solved iteratively conditional on $\hat{\rho}^{(2)}$ as

$$\begin{aligned}
&\hat{c}_{12}w_{23}(\hat{\rho}^{(2)})+\hat{c}_{13}w_{33}(\hat{\rho}^{(2)})=-w_{13}(\hat{\rho}^{(2)})\\
&\hat{c}_{12}w_{24}(\hat{\rho}^{(2)})+\hat{c}_{13}w_{34}(\hat{\rho}^{(2)})+\hat{c}_{14}w_{44}(\hat{\rho}^{(2)})=-w_{14}(\hat{\rho}^{(2)})\\
&\vdots
\end{aligned}$$

Walker (1961) showed that this estimator has excellent asymptotic efficiency for p as small as 4 unless ρ is close to its maximum value of 0.5 for a first-order moving average. He also showed that the limiting distribution of $\sqrt{T}(\hat{\rho}^{(p)}-\rho)$ is $N(0,1)$ and that a bias adjusted estimator is given by

$$\hat{\rho}^{(p)}-2T^{-1}\hat{\rho}^{(p)}(2\hat{\rho}^{(p)^2}-\hat{c}_{12}\hat{\rho}^{(p)}-1)$$

Given $\hat{\rho}^{(p)}$, the estimate of β is then obtained as the regular solution to the equation $\hat{\rho}^{(p)}\hat{\beta}^{(p)^2}-\hat{\beta}^{(p)}+\hat{\rho}^{(p)}=0$.

9.27 Walker (1961) provided the extension of this procedure to the general moving average (9.17) and illustrated the method using the simulation setup of §**9.24** with p again set at 5. The results are shown in Table 9.4, from which the summary statistics of the simulation were calculated to be

	$\hat{\rho}^{(5)}$	$\hat{\beta}^{(5)}$
Mean	0.377	0.465
Std. Dev.	0.046	0.086
SE of mean	0.010	0.019

Both $\hat{\rho}^{(5)}$ and $\hat{\beta}^{(5)}$ are closer to their true values and have smaller standard errors than their Durbin counterparts and adjusting for the bias increases the mean estimates to 0.385 and 0.470 respectively. Walker, however, argued that it was by no means clear why this method appeared to suffer from less bias than Durbin's, suggesting that the improvement found both here and in his own simulations was 'probably fortuitous'.

Estimation and inference in autoregressive moving average models

9.28 Durbin (1960) and Walker (1962) extended their methods for estimating moving averages to mixed autoregressive moving average models. In a similar vein to §§**9.22–9.27**, we focus attention on the ARMA(1,1) process (cf. (8.14))

$$x_t - \phi x_{t-1} = \varepsilon_t + \theta \varepsilon_{t-1} \tag{9.20}$$

Durbin (1960) suggested fitting an autoregression of order p, as in §**9.23**, and estimating the parameters by

$$\hat{\phi} = \frac{\hat{\alpha}_1 r_2 + \hat{\alpha}_2 r_3 + \cdots + \hat{\alpha}_p r_{p+1}}{\hat{\alpha}_1 r_1 + \hat{\alpha}_2 r_2 + \cdots + \hat{\alpha}_p r_p} \tag{9.21}$$

and

$$\hat{\theta} = -\hat{\phi} + \frac{r_1 + \hat{\alpha}_1 r_2 + \cdots + \hat{\alpha}_p r_{p+1}}{1 + \hat{\alpha}_1 r_1 + \cdots + \hat{\alpha}_p r_p} \tag{9.22}$$

showing that this was the solution obtained by minimizing the sum of squared residuals from (9.20) with the ε_t replaced by the residuals from the approximating autoregression. Durbin then used these estimates as the starting values for the following iterative procedure. Given $\hat{\phi}$ and defining

$$\ell_i = \hat{\alpha}_i + \hat{\phi}\ell_{i-1} \quad \ell_0 = 1 \quad i = 1, \ldots, p$$

Durbin showed that an efficient estimator of θ was

$$\hat{\theta} = \frac{\sum_{i=0}^{p} \ell_i \ell_{i+1}}{\sum_{i=0}^{p} \ell_i^2} \tag{9.23}$$

Given $\hat{\theta}$, and now defining

$$w_t = x_t - \theta w_{t-1} = \theta w_{t-1} + \varepsilon_t, \quad w_0 = 0, \quad t = 1, \ldots, T$$

an efficient estimator of ϕ is then

$$\hat{\phi} = \frac{\sum_{t=1}^{T-1} w_t w_{t+1}}{\sum_{t=1}^{T-1} w_t^2} \tag{9.24}$$

Thus, given either (9.21) or (9.22) as an initial condition, (9.23) and (9.24) can be used iteratively. Durbin (1960) showed how this approach can readily be extended to the general ARMA(p, q) model.

9.29 Walker (1962) extended the approach of §**9.27** to the ARMA(1,1) model. Provisional estimates of $\rho = \rho_1$ and ϕ are obtained from the formulae

$$\hat{\rho}^{(1)} = r_1 - \left(\frac{\gamma_1^2}{1 + 2\gamma_1^2}\right) S_{3,1}$$

$$\hat{\phi}^{(1)}\hat{\rho}^{(1)} = r_2 - \left(\frac{2\gamma_1(1 + a_1\gamma_1)}{1 + 2\gamma_1^2}\right) S_{3,1}$$

In these formulae, $a_1 = r_2/r_1$ and $\gamma_1 = (r_1 - a_1)/(1 + a_1^2 - 2a_1 r_1)$ are estimates of ϕ and $\gamma = \theta/(1 + \theta^2) = (\rho - \phi)/(1 + \phi^2 - 2\phi\rho)$ based on the first two serial correlations r_1 and r_2, and $S_{3,1} = r_3 - 2a_1 r_2 + a_1 r_1^2$.

Next, assuming that p is set at 5 as before, coefficients $\hat{c}_{ij}$, $i = 1, 2$, $j = 3, 4, 5$, are determined from the two sets of equations

$$\begin{bmatrix} 1 + 2\hat{\gamma}^2 & 2\hat{\gamma} & \hat{\gamma}^2 \\ 2\hat{\gamma} & 1 + 2\hat{\gamma}^2 & 2\hat{\gamma} \\ \hat{\gamma}^2 & 2\hat{\gamma} & 1 + 2\hat{\gamma}^2 \end{bmatrix} \begin{bmatrix} \hat{c}_{13} \\ \hat{c}_{14} \\ \hat{c}_{15} \end{bmatrix} = - \begin{bmatrix} \hat{\gamma}^2 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 + 2\hat{\gamma}^2 & 2\hat{\gamma} & \hat{\gamma}^2 \\ 2\hat{\gamma} & 1 + 2\hat{\gamma}^2 & 2\hat{\gamma} \\ \hat{\gamma}^2 & 2\hat{\gamma} & 1 + 2\hat{\gamma}^2 \end{bmatrix} \begin{bmatrix} \hat{c}_{23} \\ \hat{c}_{24} \\ \hat{c}_{25} \end{bmatrix} = - \begin{bmatrix} 2\hat{\gamma}(1 + \hat{\phi}^{(1)}\hat{\gamma}) \\ \hat{\gamma}^2 \\ 0 \end{bmatrix}$$

where $\hat{\gamma} = (\hat{\rho}^{(1)} - \hat{\phi}^{(1)})/(1 + \hat{\phi}^{(1)^2} - 2\hat{\phi}^{(1)}\hat{\rho}^{(1)})$. These are then used to compute the final estimates

$$\hat{\rho}^{(5)} = r_1 + \hat{c}_{13}S_3 + \hat{c}_{14}S_4 + \hat{c}_{15}S_5$$

$$\hat{\phi}^{(5)}\hat{\rho}^{(5)} = r_2 + \hat{c}_{23}S_3 + \hat{c}_{24}S_4 + \hat{c}_{25}S_5$$

where $S_j = r_j - 2\hat{\phi}^{(1)} r_{j-1} + \hat{\phi}^{(1)^2} r_{j-2}$, from which an estimate of θ can be obtained as the regular root of the equation

$$(\hat{\rho}^{(5)} - \hat{\phi}^{(5)})(\theta^2 + 1) - (1 + \hat{\phi}^{(5)^2} - 2\hat{\rho}^{(5)}\hat{\phi}^{(5)})\,\theta = 0$$

9.30 Durbin (1960) provided no simulation evidence on the properties of his procedure, but Walker (1962) extended the simulations of §§**9.25–9.28** to the

Table 9.5 Twenty simulations of length $T = 100$ from a first-order autoregressive-moving average model with $\phi = 0.8$ and $\theta = 0.5$

Series	$\hat{\rho}^{(5)}$	$\hat{\phi}$	$\hat{\phi}^{(5)}$	$\hat{\theta}$	$\hat{\theta}^{(5)}$
1	0.879	0.649	0.788	0.718	0.476
2	0.834	0.603	0.745	0.650	0.315
3	0.797	0.578	0.637	0.643	0.520
4	0.785	0.578	0.622	0.582	0.493
5	0.923	0.738	0.883	0.752	0.299
6	0.774	0.575	0.547	0.610	1.000
7	0.896	0.712	0.820	0.728	0.455
8	0.893	0.762	0.798	0.699	0.634
9	0.856	0.653	0.727	0.686	0.643
10	0.929	0.774	0.880	0.760	0.421
11	0.875	0.751	0.767	0.670	0.588
12	0.850	0.673	0.727	0.696	0.542
13	0.883	0.749	0.728	0.733	1.000
14	0.858	0.714	0.675	0.688	1.000
15	0.916	0.787	0.846	0.744	0.558
16	0.923	0.691	0.898	0.758	0.168
17	0.844	0.617	0.691	0.678	0.858
18	0.889	0.717	0.822	0.708	0.354
19	0.904	0.734	0.831	0.735	0.480
20	0.853	0.666	0.715	0.684	0.717

ARMA(1,1) process (9.20) with $\phi = 0.8$ and $\theta = 0.5$. Table 9.5 repeats this simulation for both the Durbin and Walker methods. Since no suggestions were provided by Durbin as to the number of iterations to use or the convergence criteria to employ, we used ten iterations, by which time the estimates of both ϕ and θ had settled down sufficiently. We follow Walker and set $p = 5$, thus using the equations in §**9.29**. From the results presented in Table 9.5, the following summary statistics were calculated.

	$\hat{\rho}^{(5)}$	$\hat{\phi}$	$\hat{\phi}^{(5)}$	$\hat{\theta}$	$\hat{\theta}^{(5)}$
$\bar{a}$	0.8681	0.6860	0.7573	0.6960	0.5760
s	0.0465	0.0705	0.0963	0.0497	0.2443
σ	0.0349	0.0648	0.0648	0.0967	0.0967
$(\bar{a} - a)$	−0.0197	−0.1140	−0.0427	−0.1960	−0.0760
$\sqrt{20}(\bar{a} - a)/s$	−1.89	−7.23	−1.98	−17.64	1.39
$\sqrt{20}(\bar{a} - a)/\sigma$	−2.52	−7.87	−4.47	−9.06	−3.51

Here $\bar{a}$ denotes the mean of the twenty estimates of the parameter a, while s and σ are the observed and theoretical standard deviations of a (the latter taken from Walker, 1962, Table 2: note that the true value of ρ is 0.8878 from (8.15)).

It is seen that the Walker estimates are superior in terms of both smaller bias and in the ratio of the bias to the standard error. Walker also proposed a bias adjustment that reduces the bias in the estimates of ρ and ϕ but makes the bias in θ worse.

The likelihood function of an ARMA model

9.31 The advances in both computing power and numerical algorithms (see, for example, Hartley, 1961, and Marquardt, 1963) during the 1960s meant that the estimation methods outlined in the previous sections were quickly superseded by nonlinear estimation techniques based on the *likelihood principle*.[9] This approach was developed in considerable detail in Box and Jenkins (1970, chapter 7) and, because of its central importance to the analysis of time series, we review it in commensurate detail here. The general model to be estimated is the stationary and invertible ARMA(p, q) process (8.12), which may be written

$$a_t = x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} \quad t = 1, 2, \ldots, T \qquad (9.25)$$

where the notation of Box and Jenkins is adopted with $x_t = X_t - \mu$ and $a_t \sim IID(0, \sigma_a^2)$. Typically the mean μ will be replaced by the sample mean $\overline{X}$, but if desired it can be estimated along with the other parameters of (9.25), $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)'$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$ and σ_a. The observations $x_1, x_2, \ldots, x_T$ are gathered together in the vector $\mathbf{x}$, while the innovations $a_1, a_2, \ldots, a_T$ are gathered together in the vector $\mathbf{a}$.

The x's cannot be substituted immediately into (9.25) to calculate the a's because of the difficulty inherent in starting up the difference equation. However, if the p values $x_{-p+1}, \ldots, x_0$ and the q values $a_{-q+1}, \ldots, a_0$ were available, then (9.25) could be used recursively to calculate $a_1, a_2, \ldots, a_T$ conditional on this choice of starting values, and these can be gathered together in the vectors $\mathbf{x}_*$ and $\mathbf{a}_*$.

Thus, for any given choice of parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ and starting values $(\mathbf{x}_*, \mathbf{a}_*)$, we could calculate recursively a set of values $a_t(\boldsymbol{\phi}, \boldsymbol{\theta}\,|\mathbf{x}_*, \mathbf{a}_*, \mathbf{x})$, $t = 1, 2, \ldots, T$. If it is assumed that the a's are normally distributed then their joint probability distribution is

$$p(a_1, a_2, \ldots, a_T) \propto \sigma_a^{-T} \exp\left(-\left(\sum_{t=1}^{T} a_t^2/\sigma_a^2\right)\right)$$

For a particular set of data $\mathbf{x}$, the log likelihood associated with the parameter values $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a)$, *conditional* on the choice $(\mathbf{x}_*, \mathbf{a}_*)$, would then be

$$\ell_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a) = -T \ln \sigma_a - \frac{S_*(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \qquad (9.26)$$

where

$$S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{T} a_t^2(\boldsymbol{\phi}, \boldsymbol{\theta} \,|\mathbf{x}_*, \mathbf{a}_*, \mathbf{x}) \tag{9.27}$$

Since the conditional log likelihood ℓ_* involves the data only through the conditional *sum of squares function* S_* (ℓ_* being a linear function of S_* for any fixed σ_a), the maximum likelihood estimates will be the same as the least squares estimates and the behavior of the conditional likelihood can therefore be studied by examining the conditional sum of squares function.

9.32 Although the unconditional likelihood is strictly what is needed for parameter estimation, if T is reasonably large then a sufficient approximation to it is obtained by using the conditional likelihood with suitable values substituted for the elements of $\mathbf{x}_*$ and $\mathbf{a}_*$ in (9.27). One possibility is to set these elements equal to their unconditional expectations, which are zero. This approximation can be poor, however, if some of the roots of $\phi(B) = 0$ lie close to the boundary of the unit circle, using the terminology adopted in §**8.10**. In these circumstances the process is approaching nonstationarity and, as a consequence, the initial value x_1 could deviate considerably from its unconditional expectation of zero, thus introducing a large transient which would be slow to die out. An alternative is then to use (9.25) to calculate the a's *from* a_{p+1} *onwards*, setting previous a's equal to zero. Consequently, actually occurring values are used for the x's throughout the recursion, but only $T - p$ terms appear in the summation in (9.27), a loss of information which should only be slight for large T. For pure moving average models with $p = 0$, the two procedures are obviously equivalent.

9.33 The unconditional likelihood of (9.25) is given by (Box and Jenkins, 1970, chapter 7.1.4 and Appendix A7.4)

$$\ell(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a) = f(\boldsymbol{\phi}, \boldsymbol{\theta}) - T \ln \sigma_a - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{9.28}$$

where $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ is a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, and

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{T} E(a_t|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x})^2 \tag{9.29}$$

is the *unconditional sum of squares function*. Usually $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ is only important for small T and quickly becomes dominated by $S(\boldsymbol{\phi}, \boldsymbol{\theta})/2\sigma_a^2$ as T increases. Consequently, the parameter estimates obtained by minimizing the sum of squares (9.29), known as the least squares estimates, usually provide very close approximations to the maximum likelihood estimates.

Backward ARMA processes

9.34 In order to compute $S(\boldsymbol{\phi}, \boldsymbol{\theta})$, the set of conditional expectations $E(a_t | \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x})$ for $t = -\infty, \ldots, -1, 0, 1, \ldots, T$ need to be calculated. To construct an algorithm to do this, Box and Jenkins (1970, chapter 6.4) introduced the concept of a *backward* process. Consider the regular, invertible, MA(1) process

$$x_t = (1 - \theta B)\, a_t \quad |\theta| < 1 \tag{9.30}$$

From the analysis of §**7.21** (especially Example 1), this has the dual, but not invertible, representation

$$x_t = (1 - \theta^{-1} B)\, \alpha_t$$

with $\sigma_\alpha^2 = \theta^2 \sigma_a^2$. This can be written as

$$\begin{aligned} x_t &= (1 - \theta B^{-1})(-\theta^{-1} B)\, \alpha_t \\ &= ((1 - \theta^{-1} B)(-\theta B^{-1}))(-\theta^{-1} B)\, \alpha_t \\ &= ((1 - \theta^{-1} B)(-\theta B^{-1}))\, e_t \\ &= (1 - \theta B^{-1})\, e_t \end{aligned}$$

on setting $e_t = -\theta^{-1} B \alpha_t = -\alpha_{t-1}/\theta$, which has variance σ_a^2. By defining $F \equiv B^{-1}$ to be the *forward* operator, the 'backward' process

$$x_t = (1 - \theta F)\, e_t \tag{9.31}$$

is then seen to be the dual of the forward process (9.30), in which the innovation e_t is expressible as the convergent sum of current and *future* values of x:

$$e_t = x_t + \theta x_{t+1} + \theta^2 x_{t+2} + \cdots$$

An ARMA(p, q) process thus has both a forward and a backward representation:

$$\phi(B) x_t = \theta(B)\, a_t \tag{9.32}$$

$$\phi(F) x_t = \theta(F)\, e_t \tag{9.33}$$

A value x_{-h} therefore bears exactly the same probability relationship to the sequence $x_1, x_2, \ldots, x_T$ as does the value x_{T+h+1} to the sequence $x_T, x_{T-1}, \ldots, x_1$. The expected value of x_{-h} can then be obtained in exactly the same way as x_{T+h+1} but by using the backward model (9.33), a procedure termed by Box and Jenkins as 'back forecasting' (or simply 'backcasting').

Calculating the sum of squares function

9.35 The two representations can be used to generate the conditional expectations $E(a_t|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x})$, which we now denote as $[a_t]$, by taking conditional expectations of (9.33) to generate the backcasts

$$\phi(F)[x_t] = \theta(F)[e_t]$$

and then using (9.32) to generate the $[a_t]$s, from which the unconditional sum of squares can be calculated.

To illustrate the procedure, consider the following $T = 12$ successive values of x_t .

t	1	2	3	4	5	6	7	8	9	10	11	12
x_t	2.0	0.8	−0.3	−0.3	−1.9	0.3	3.2	1.6	−0.7	3.0	4.3	1.1

Suppose we wish to compute the unconditional sum of squares $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ associated with the ARMA(1,1) process

$$(1 - \phi B)x_t = (1 - \theta B)\, a_t$$

$$(1 - \phi F)x_t = (1 - \theta F)\, e_t$$

with parameter values $\phi = 0.3$ and $\theta = 0.7$. If it is assumed that backcasts are negligible beyond $t = -Q$, then the non-zero $[e_t]$s can be generated from

$$[e_t] = [x_t] - 0.3[x_{t+1}] + 0.7[e_{t+1}] \quad t = 1, 2, \dots, T - 1 = 11$$

on noting that $[e_{12}] = 0$ and $[e_t] = 0$ for $t \leq 0$. The backcasts of x_t are then generated from

$$[x_t] = 0.3[x_{t+1}] - 0.7[e_{t+1}] \quad t = -Q, -Q + 1, \dots, 0$$

With the starting value $[a_{-Q}] = [x_{-Q}]$, the successive values of $[a_t]$ are then generated from

$$[a_t] = [x_t] - 0.3[x_{t-1}] + 0.7[a_{t-1}] \quad t = -Q + 1, -Q + 2, \dots, T = 12$$

with the unconditional sum of squares being calculated as

$$S(0.3, 0.7) = \sum_{t=-Q}^{T=12} [a_t^2]$$

Table 9.6 Calculation of the $[a_t]$s from 12 values of a series assumed to be generated by the process $(1-0.3B)x_t = (1-0.7B)a_t$

t	$[a_t]$	$[x_t]$	$[e_t]$
−4	−0.008	−0.008	0
−3	−0.031	−0.028	0
−2	−0.107	−0.094	0
−1	−0.359	−0.312	0
0	−1.197	−1.039	0
1	1.474	2.0	2.342
2	1.232	0.8	0.831
3	0.322	−0.3	−0.838
4	0.016	−0.3	0.180
5	−1.799	−1.9	−0.128
6	−0.389	0.3	2.660
7	2.837	3.2	4.743
8	2.626	1.6	2.890
9	0.658	−0.7	1.542
10	3.671	3.0	4.489
11	5.970	4.3	3.970
12	3.989	1.1	0

The calculations are shown in Table 9.6 for $Q=4$, from which we obtain $S(0.3, 0.7)=89.16$. Box and Jenkins discuss how a second iteration could be carried out by using the forward model with $[a_{12}]=3.989$ to obtain $[x_{13}], [x_{14}], \ldots$ and then substituting these into the backward equation to obtain new backcasts $[x_0], [x_{-1}], \ldots$. They show that little is gained by doing this and that, in general, the procedure converges very quickly.

The two conditional sums of squares suggested as approximations in §**9.32** were (i) to start the recursion at the first available observation, setting all unknown a's and e's to zero and all the x's equal to their unconditional expectation; and (ii) to start the recursion at the pth observation using only observed values of the x's and zeros for unknown a's and e's. In the above example the unconditional expectation of x is zero and $p=1$, so that the two approximations produce

$$\sum_{t=1}^{12} (e_t | 0.3, 0.7, x_{13}=0, e_{13}=0, \mathbf{x})^2 = 101.0$$

and

$$\sum_{t=2}^{12} (e_t | 0.3, 0.7, x_{12}=1.1, e_{12}=0, \mathbf{x})^2 = 82.44$$

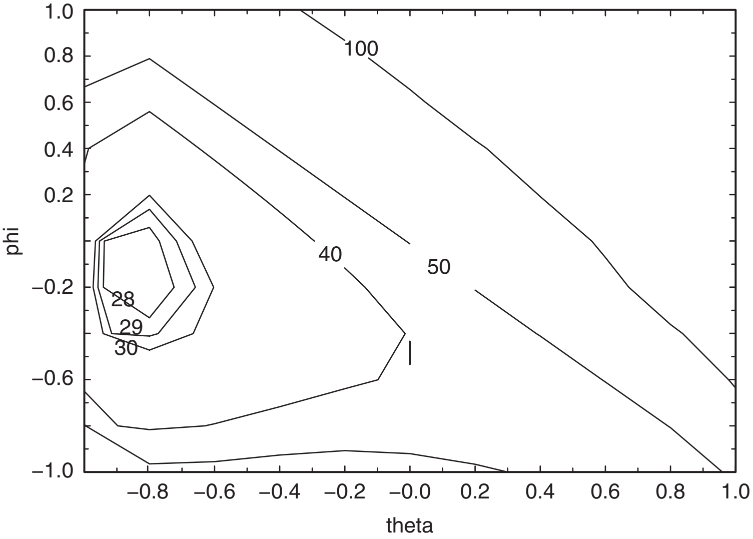


Figure 9.10 Contour plot of $S(\phi, \theta)$ calculated from the 12 values of x_t

respectively. The sum of squares using (i) is a poor approximation, although the discrepancy, which is over 10% in a series of 12 values, would be diluted if the sample was larger, since the transient introduced by the choice of starting value will die out. The approximation (ii) is much more accurate and confirms Box and Jenkins' view in §**9.32** that this is the method to employ if a conditional approximation is to be used.

9.36 Figure 9.10 presents a contour plot of $S(\phi, \theta)$ obtained by calculating the unconditional sum of squares for $\phi, \theta = -1, (0.1), 1$: the minimum is obtained at $S(0.1, -0.9) = 26.05$. While the results for such a small sample cannot be taken too seriously, the example does illustrate the usefulness of studying the complete sum of squares function and hence the likelihood function.

Box and Jenkins (1970, chapter 7.1.6) discussed alternative ways of graphically presenting sums of squares functions, and hence likelihood functions, for two and three parameters. They pointed out that the likelihood function does not merely indicate the maximum likelihood values but, according to the likelihood principle, also represents all the information contained in the data. Its overall shape can therefore be extremely informative: the existence of multiple peaks, for example, will imply that there are more than one set of values of the parameters that might explain the data, whereas the existence of a sharp ridge means that one parameter, considerably different from the maximum likelihood, could explain the data if accompanied by a value of the other parameter which deviated appropriately from its maximum value. Box and Jenkins referred to this as the *estimation situation*, which needed to be understood by examining the likelihood both graphically and analytically. For example, care needs to be taken when the maximum may be on or near a boundary, as in Figure 9.10 where the maximum likelihood estimate of θ looks to be close to -1.

Analytically, the treatment of likelihood functions has typically consisted of (i) differentiating the log likelihood and setting first derivatives to zero to obtain the maximum likelihood (ML) estimates, and (ii) deriving approximate variances and covariances of these estimates from either the second derivatives of the log likelihood or from their expected values. Mechanical application of this treatment can be problematic for two reasons: setting first derivatives to zero does not always produce maxima, and the information contained in the likelihood is only fully expressed by the ML estimates and the second derivatives of the log likelihood if the function can be adequately represented by a quadratic approximation over the region of interest.

Variances and covariances of ML estimates

9.37 Following Box and Jenkins (1970, chapter 7.1.7), we define $\boldsymbol{\beta}$ to be a vector whose $k=p+q$ elements, β_i, $i=1,\dots,k$, are the autoregressive and moving average parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, and $\boldsymbol{\xi}$ as the complete set of parameters $\boldsymbol{\beta}, \sigma_a$. The log likelihood can then be written

$$\ell(\boldsymbol{\xi}) = \ell(\boldsymbol{\beta}, \sigma_a) \cong \ell(\hat{\boldsymbol{\beta}}, \sigma_a) + \tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\ell_{ij}(\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_j)$$

where, on the assumption that a quadratic approximation is adequate, the derivatives

$$\ell_{ij} = \frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma_a)}{\partial\beta_i \partial\beta_j}$$

are constant. For large T, the quadratic approximation will be valid if $S(\boldsymbol{\beta})$ is, or if the conditional expectations in (9.29) are, approximately locally linear in the elements of $\boldsymbol{\beta}$. Under these circumstances, useful approximations to the variances and covariances of the estimates may be obtained and approximate confidence intervals constructed.

9.38 The *information matrix* for the $\boldsymbol{\beta}$ parameters is the $k \times k$ matrix defined by Whittle (1953a) as $\mathbf{I}(\boldsymbol{\beta}) = -E(\ell_{ij})$. For a given value of σ_a, the *variance-covariance matrix* $\mathbf{V}(\hat{\boldsymbol{\beta}})$ for the ML estimates $\hat{\boldsymbol{\beta}}$ is, for large T, given by the inverse of this information matrix:

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \cong -E(\ell_{ij})^{-1}$$

For example, if $k=2$,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_1, \hat{\beta}_2) & V(\hat{\beta}_2) \end{bmatrix} \cong - \begin{bmatrix} E(\ell_{11}) & E(\ell_{12}) \\ E(\ell_{12}) & E(\ell_{11}) \end{bmatrix}^{-1}$$

Now, using (9.28),

$$\ell_{ij} \cong -\frac{S_{ij}}{2\sigma_a^2} = -\frac{1}{2\sigma_a^2}\frac{\partial^2 S(\boldsymbol{\beta}|\mathbf{x})}{\partial\beta_i\partial\beta_j}$$

so that

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \cong 2\sigma_a^2 \begin{bmatrix} \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1^2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_2} \\ \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_2^2} \end{bmatrix}^{-1} = 2\sigma_a^2 \begin{bmatrix} S^{11} & S^{12} \\ S^{12} & S^{22} \end{bmatrix} \tag{9.34}$$

where $S^{ij} = S_{ij}^{-1}$. If $S(\boldsymbol{\beta})$ were exactly quadratic in $\boldsymbol{\beta}$ over the relevant region of the parameter space, then all the derivatives S_{ij} would be constant over this region. In practice the S_{ij} will vary somewhat and they are usually evaluated at or near the point $\hat{\boldsymbol{\beta}}$. Box and Jenkins showed that an estimate of σ_a^2 is provided by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\beta}})/T$ and that, for T large, $\hat{\sigma}_a^2$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated.

Confidence regions for the parameters

9.39 The square roots of the diagonal elements of (9.34) define the *standard errors* of the estimates, $\mathrm{SE}(\hat{\beta}_i)$. When several parameters are considered simultaneously, joint *confidence regions* may be constructed from the result that

$$-\sum_{i,j} E(\ell_{ij})(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) = \frac{1}{2\sigma_a^2}\sum_{i,j} S_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) < \chi_\varepsilon^2$$

defines an approximate $1-\varepsilon$ confidence region. Such a region will be bounded by the contour of the sum of squares surface for which

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}})\left(1 + \frac{\chi_\varepsilon^2}{T}\right)$$

Given the estimates $\hat{\phi}=0.1$, $\hat{\theta}=-0.9$ and $S(0.1,\ -0.9)=26.05$ obtained by minimizing $S(\phi,\theta)$ for the data in Figure 9.10, 0.95 and 0.99 confidence regions are bounded by the contours given by

$$S_{0.95}(\phi,\theta) = 26.05\left(1 + \frac{5.99}{12}\right) = 39.05$$

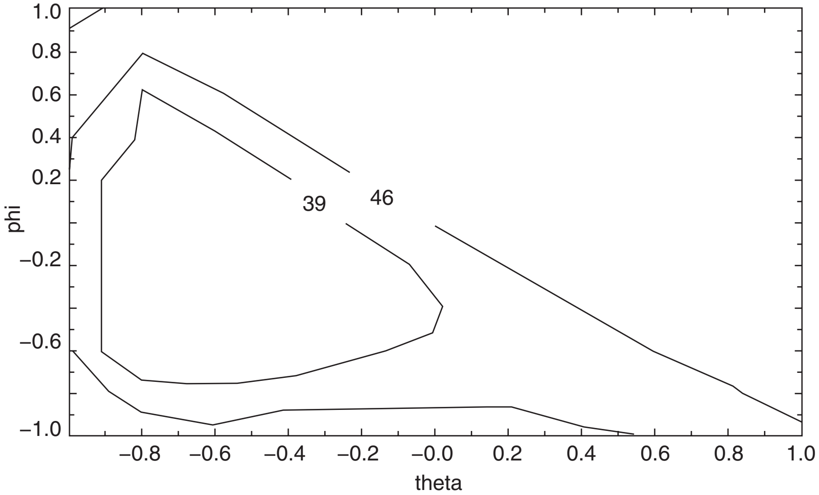


Figure 9.11 0.95 (labelled 39) and 0.99 (labelled 46) confidence regions for ϕ, θ around $(0.1, -0.9)$

$$S_{0.99}(\phi, \theta) = 26.05\left(1 + \frac{9.21}{12}\right) = 46.04$$

These regions are shown in Figure 9.11 and, unsurprisingly given the very small sample size, are rather wide.

The covariance matrix for an ARMA(1,1) model was shown by Box and Jenkins (1970, Appendix 7.5) to be

$$\mathbf{V}(\hat{\phi}, \hat{\theta}) = T^{-1}\frac{1-\phi\theta}{(\phi-\theta^2)}\begin{bmatrix}(1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2)\\(1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta)\end{bmatrix}$$

so that in the example here it is

$$\mathbf{V}(0.1, -0.9) = \begin{bmatrix}0.09802 & 0.01709\\0.01709 & 0.01882\end{bmatrix}$$

leading to the standard errors $\text{SE}(\hat{\phi}) = 0.313$ and $\text{SE}(\hat{\theta}) = 0.137$.

Nonlinear estimation

9.40 Although plotting the sum of squares function is important as it ensures that any peculiarities in the estimation situation are shown up, once we are satisfied that anomalies are unlikely, nonlinear estimation algorithms may be applied. The need for a nonlinear algorithm is seen by contrasting the autoregressive process $[a_t] = \phi(B)[x_t]$, for which

$$\frac{\partial[a_t]}{\partial\phi_i} = -[x_{t-i}] + \phi(B)\frac{\partial[x_t]}{\partial\phi_i} \tag{9.35}$$

with the moving average process $[a_t] = \theta^{-1}(B)[x_t]$, for which

$$\frac{\partial[a_t]}{\partial\theta_j} = \theta^{-2}(B)[x_{t-j}] + \theta^{-1}(B)\frac{\partial[x_t]}{\partial\theta_j} \tag{9.36}$$

In (9.35) $[x_t] = x_t$ and $\partial[x_t]/\partial\phi_i = 0$ for $t > 0$, while for $t \leq 0$ both are functions of $\boldsymbol{\phi}$, so that, except for the effect of 'starting values', $[a_t]$ is linear in $\boldsymbol{\phi}$. In contrast, $[a_t]$ is always a nonlinear function of $\boldsymbol{\theta}$ in (9.36). Nevertheless, iterative application of linear least squares may be used to estimate the parameters of any ARMA model.

The problem, as set out earlier, is to minimize $\sum_{t=1-Q}^{T}[a_t]^2$. Suppose $[a_t]$ is expanded in a Taylor series about its value corresponding to some initial set of 'guessed' parameter values $\boldsymbol{\beta}_0' = (\beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0})$:

$$[a_t] = [a_{t,0}] - \sum_{i=1}^{k}(\beta_i - \beta_{i,0})z_{i,t} \tag{9.37}$$

where

$$[a_{t,0}] = [a_t|\mathbf{x}, \boldsymbol{\beta}_0]$$

and

$$z_{i,t} = -\left.\frac{\partial[a_t]}{\partial\beta_i}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

If $\mathbf{Z}$ is the $(T+Q)\times k$ matrix containing the $z_{i,t}$ as elements, the $T+Q$ equations (9.37) may be expressed as

$$[\mathbf{a}_0] = \mathbf{Z}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + [\mathbf{a}]$$

where $[\mathbf{a}_0]$ and $[\mathbf{a}]$ are column vectors with $T+Q$ elements. The adjustments $\boldsymbol{\beta} - \boldsymbol{\beta}_0$, which minimize $S(\boldsymbol{\beta}) = [\mathbf{a}]'[\mathbf{a}]$, can now be obtained by linear least squares, regressing the $[a_0]$s onto the zs. Because the $[a_t]$s will not be exactly linear in $\boldsymbol{\beta}$, a single adjustment will not immediately produce least squares values, so that an iterative procedure, in which the adjusted values are substituted as new guesses and the process is repeated until convergence occurs, becomes necessary. The speed of convergence and, indeed, whether there is convergence at all, often depends on how good the initial guess $\boldsymbol{\beta}_0$ is to the 'true' vector $\boldsymbol{\beta}$.

While (9.35) and (9.36) allow the derivatives $z_{i,t}$ to be obtained analytically, it is often easier to obtain them numerically. Box and Jenkins (1970, chapter 7.2) outlined the methods then available to do this and also provided a suite of computer programs that enabled nonlinear estimation of ARMA models to be carried out.

Identification and initial estimates of ARMA models

9.41 Before embarking on the nonlinear estimation of an ARMA model for an observed series, the actual form of the model, i.e., the orders p and q, need to be selected and the initial estimates $\boldsymbol{\beta}_0' = (\beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0})$ need to be chosen. Box and Jenkins (1970, chapter 6) recognized that this was an essential first stage of the model building process and formalized a procedure, known as the *identification* stage, for the purposes of doing just this. The 'philosophy' behind identification is best summed up by their statement that

> identification methods are rough procedures applied to a set of data to investigate the kind of representational model which is worthy of further investigation. It should be explained that identification is necessarily inexact. It is inexact because the question of what types of models occur in practice and in what circumstances, is a property of the behavior of the physical world and cannot, therefore, be decided by purely mathematical argument. Furthermore, because at the identification stage no precise formulation of the problem is available, statistically 'inefficient' methods must necessarily be used. It is a stage at which graphical methods are particularly useful and judgment must be exercised. However, it should be borne in mind that preliminary identification commits us to nothing except to tentatively entertaining a class of models which will later be efficiently fitted and checked. (*ibid.*, page 173)

The principal tools for the identification of an ARMA process are the autocorrelation and partial autocorrelation functions: '(t)hey are used not only to help guess the form of the model, but also to obtain approximate estimates of the parameters. Such approximations are often useful at the estimation stage to provide starting values for iterative procedures employed at that stage' (*ibid.*, page 174). This involves studying the general appearance of the sample autocorrelation (or serial correlation) and partial autocorrelation functions to obtain clues about the choice of the autoregressive and moving average orders p and q. This is done by relating their appearance to the characteristic behaviour of the theoretical autocorrelation and partial autocorrelation functions for moving average, autoregressive, and mixed processes. This behaviour is succinctly summarized by Box and Jenkins, and brings together the properties of these processes that have been developed in chapters 7 and 8 (in particular §**8.10**):

> Briefly, whereas the autocorrelation function of an autoregressive process of order p tails off, its partial autocorrelation function has a cutoff after lag p. Conversely, the autocorrelation function of a moving average process of order q has a cutoff after lag q, while its partial autocorrelation function tails off. If both the autocorrelations and partial autocorrelations tail off, a mixed process is suggested. Furthermore, the autocorrelation function for a mixed process, containing a pth order autoregressive component and a qth order moving average component, is a mixture of exponentials and damped sine waves after the first $q-p$ lags. Conversely, the partial autocorrelation function for a mixed process is dominated by a mixture of exponentials and damped sine waves after the first $p-q$ lags.
>
> In general, autoregressive (moving average) behavior, as measured by the autocorrelation function, tends to mimic moving average (autoregressive) behavior as measured by the partial autocorrelation function. For example, the autocorrelation function of a first-order autoregressive process decays exponentially, while the partial autocorrelation function cuts off after the first lag. Correspondingly, for a first-order moving average process, the autocorrelation function cuts off after the first lag. The partial autocorrelation function, while not precisely exponential, is dominated by exponential terms and has the general appearance of an exponential. (*ibid.*, pages 175–6)

Particularly important for model building are the first- and second-order autoregressive and moving average processes and the simple mixed ARMA(1, 1) process. The theoretical properties of these models are summarized in Table 9.7, which has been adapted from Box and Jenkins' Table 6.1.

9.42 Comparing the behaviour of the sample and theoretical autocorrelation functions is by no means straightforward, particularly with small sample sizes. As discussed in Chapter 8 (particularly §**8.3** and §**8.9**), Kendall had been particularly concerned that moderately large sample autocorrelations could occur after the theoretical autocorrelation function had damped out, and that apparent ripples and trends could occur in the sample autocorrelation function which had no basis in the theoretical function. Box and Jenkins thus recommended caution when attempting to use the sample autocorrelation function as a tool for identification, because while 'it is usually possible to be fairly sure about broad characteristics, . . . more subtle indications may or may not represent real effects, and two or more related models may need to be entertained and investigated further at the estimation and diagnostic checking stages of model building' (*ibid.*, page 177).

Table 9.7 Behaviour of the autocorrelation and partial autocorrelation functions of various ARMA(p, q) processes. ϕ_{kk} is the kth partial autocorrelation, being the coefficient on the kth lag of an AR(k) process

ARMA order	(1,0)	(0,1)
Behaviour of ρ_k	decays exponentially	only ρ_1 nonzero
Behaviour of ϕ_{kk}	only ϕ_{11} non-zero	exponential dominates decay
Preliminary estimates from	$\phi_1 = \rho_1$	$\rho_1 = \dfrac{-\theta_1}{1+\theta_1^2}$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$
ARMA order	**(2,0)**	**(0,2)**
Behaviour of ρ_k	mixture of exponentials or damped sine wave	only ρ_1 and ρ_2 nonzero
Behaviour of ϕ_{kk}	only ϕ_{11} and ϕ_{22} nonzero	Dominated by mixture of exponentials or damped sine wave
Preliminary estimates from	$\phi_1 = \dfrac{\rho_1(1-\rho_2)}{1-\rho_1^2}$ $\phi_2 = \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2}$	$\rho_1 = \dfrac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$ $\rho_2 = \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2}$
Admissible region	$\begin{cases} -1 < \phi_2 < 1 \\ \phi_2 + \phi_1 < 1 \\ \phi_2 - \phi_1 < 1 \end{cases}$	$\begin{cases} -1 < \theta_2 < 1 \\ \theta_2 + \theta_1 < 1 \\ \theta_2 - \theta_1 < 1 \end{cases}$
ARMA order	**(1,1)**	
Behaviour of ρ_k	decays exponentially from first lag	
Behaviour of ϕ_{kk}	Dominated by exponential decay from first lag	
Preliminary estimates from	$\rho_1 = \dfrac{(1-\theta_1\phi_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1}$	$\rho_2 = \rho_1\phi_1$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$

Given the behaviour of the theoretical autocorrelation and partial autocorrelation functions, as in Table 9.7, it is also important that there are means to judge whether their sample counterparts are effectively zero after some specific lag. Box and Jenkins suggested using the Bartlett formula (9.3), with sample estimates replacing theoretical autocorrelations, to compute the standard error of r_k as

$$s(r_k) \cong T^{-\frac{1}{2}}(1 + 2r_1^2 + 2r_2^2 + \cdots + 2r_{k-1}^2)^{1/2}$$

and to use the result in §**9.10** that the standard error of the sample partial autocorrelation $r_{k\cdot}$, which we may denote $\hat{\phi}_{kk}$, is $s(\hat{\phi}_{kk}) = T^{-\frac{1}{2}}$. In both cases the ratio of the estimate to its standard error may be taken to be asymptotically standard normal.

9.43 To illustrate the identification stage of ARMA model building, we shall again use the sunspot index from 1700 to 2007, last investigated in §**9.6**, where

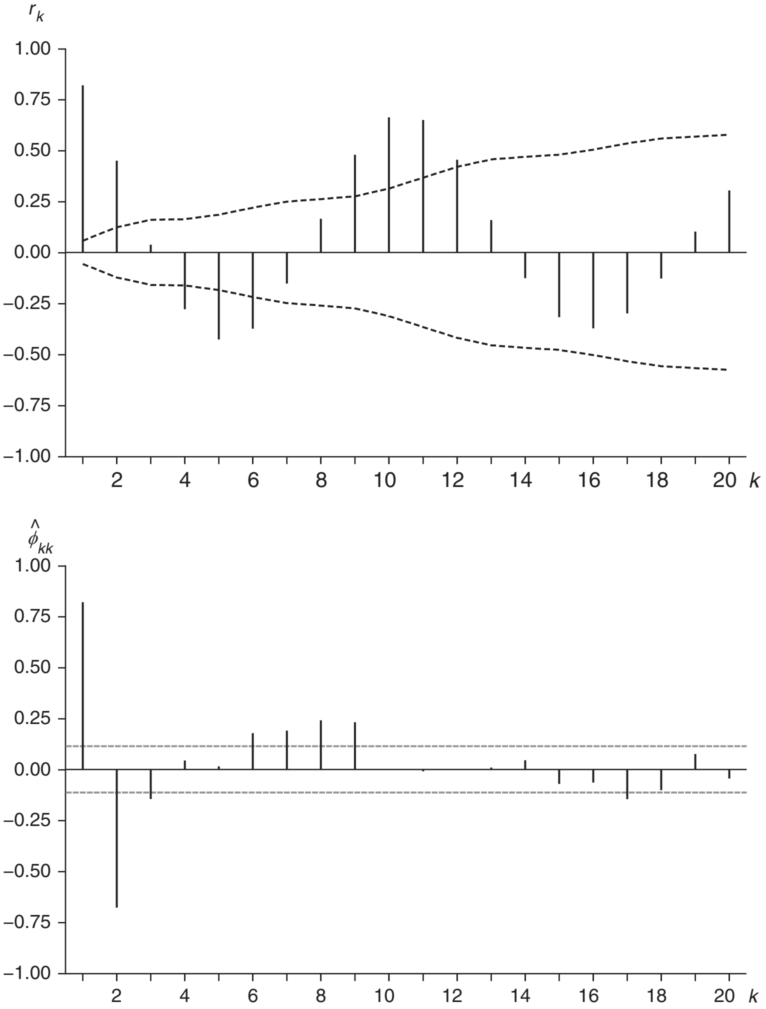


Figure 9.12 Sample autocorrelation and partial autocorrelation functions for the sunspot index with, respectively, one- and two-standard error bounds

an AR(2) fit was found to be unsatisfactory, and also Series A from Box and Jenkins (1970, page 525). The sample and partial autocorrelation functions for the sunspot index are shown in Figure 9.12. The sample autocorrelation function shows the familiar oscillatory pattern, while the sample partial autocorrelation function appears to cut off at $k = 9$ when compared to its two-standard error bounds, thus tentatively identifying an AR(9) process, as was suggested by both Craddock (1965) and Morris (1977), although a mixed model, such as an

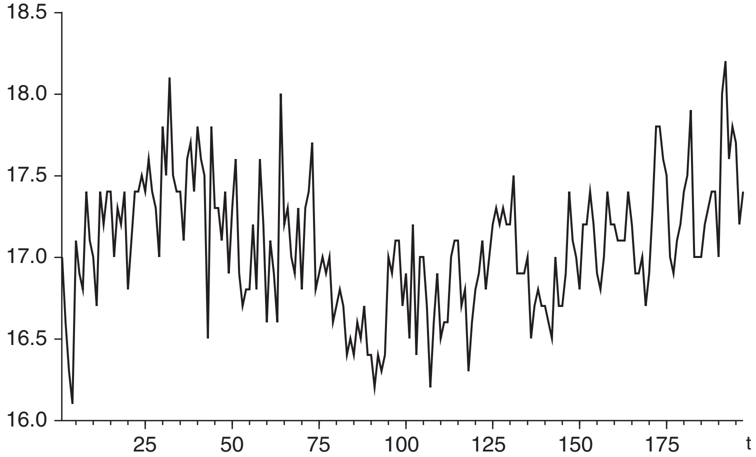


Figure 9.13 Series A from Box and Jenkins (1970): $T = 197$ two-hourly concentration readings of a chemical process

ARMA(2,1) process, might be appropriate (recall that it was found in §**9.6** that an AR(2) process offered a poor fit to this series).

Series A, consisting of 197 two-hourly concentration readings on a chemical process, is plotted as Figure 9.13 and appears to be stationary. The sample autocorrelation and partial autocorrelation functions are shown in Figure 9.14 and from these Box and Jenkins tentatively identified the series as being generated by an ARMA(1,1) process on the grounds that, from r_1 onwards, the sample autocorrelations decay roughly exponentially, albeit rather slowly. Initial estimates of ϕ and θ may be obtained by solving the expressions for ρ_1 and ρ_2 in Table 9.7 on substitution with $r_1 = 0.57$ and $r_2 = 0.50$. Chart D of Box and Jenkins (1970) may be used to read off values for these initial estimates, which Box and Jenkins report as $\hat{\phi} \approx 0.87$ and $\hat{\theta} \approx 0.48$.

Estimated models for the sunspot index and Series A

9.44 Table 9.8 reports the estimated parameters of various models fitted to the sunspot index. Initial values are not needed for the various autoregressive models as these are linear least squares fits. Initial estimates for the ARMA(2,1) model were obtained using the procedure set out in Box and Jenkins (1970, Appendix A6.2). The ARMA(2,1) model clearly gives a better fit than the AR(2), with the additional parameter θ_1 being significantly different from zero and $\hat{\sigma}_a$ being a little smaller. The innovation standard error from the ARMA(2,1) model is reduced considerably (the innovation variance being reduced by more than 17%) when the previously identified AR(9) model is fitted. Several autoregressive

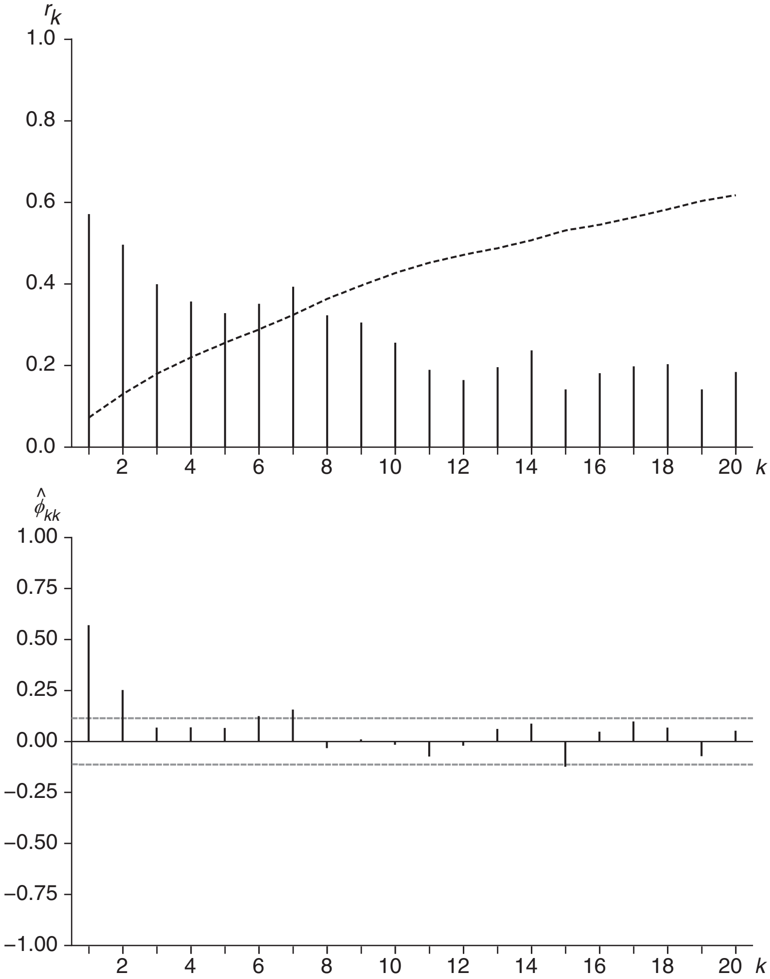


Figure 9.14 Sample autocorrelation and partial autocorrelation functions for Box and Jenkins' Series A with, respectively, one- and two-standard error bounds

coefficients are found to be insignificant, however, and so a restricted autoregression was also fitted with the coefficients $\phi_3, \ldots, \phi_8$ set to zero, which further improves the fit.

The estimated ARMA(1,1) model for Series A is

$$\underset{}{x_t -} \underset{(\pm 0.04)}{0.92}\, x_{t-1} = 1.41 + \hat{a}_t - \underset{(\pm 0.08)}{0.61}\, a_{t-1} \quad \hat{\sigma}_a^2 = 0.099$$

which accords well with the estimated model provided by Box and Jenkins (1970, Table 7.13).

Table 9.8 Alternative model estimates for the sunspot index. Standard errors are shown in parentheses. AR(9)* denotes an AR(9) model with the restrictions $\phi_3 = \cdots = \phi_8 = 0$ imposed

	AR(2)	AR(9)	AR(9)*	ARMA(2,1)
$\hat{\mu}$	50.07 (3.19)	52.77 (6.92)	53.49 (9.22)	50.10 (2.81)
$\hat{\phi}_1$	1.39 (0.04)	1.16 (0.06)	1.21 (0.04)	1.47 (0.05)
$\hat{\phi}_2$	−0.69 (0.04)	−0.40 (0.09)	−0.51 (0.04)	−0.76 (0.05)
$\hat{\phi}_3$	–	−0.16 (0.09)	–	–
$\hat{\phi}_4$	–	0.15 (0.09)	–	–
$\hat{\phi}_5$	–	−0.10 (0.09)	–	–
$\hat{\phi}_6$	–	0.02 (0.09)	–	–
$\hat{\phi}_7$	–	0.04 (0.09)	–	–
$\hat{\phi}_8$	–	−0.08 (0.09)	–	–
$\hat{\phi}_9$	–	0.25 (0.06)	0.21 (0.03)	–
$\hat{\theta}_1$	–	–	–	0.16 (0.08)
$\hat{\sigma}_a$	16.69	15.10	15.09	16.60

Diagnostic checking of fitted ARMA models

9.45 The iterative model-building procedure proposed by Box and Jenkins consists of three stages, the first two of which are identification and estimation, which have already been discussed. The third stage is that of *diagnostic checking*, that of deciding whether the fitted model is adequate. Box and Jenkins' general philosophy is that if

> there should be evidence of serious inadequacy, we shall need to know how the model should be modified in the next iterative cycle. What we are doing is only partially described by the words, 'testing goodness of fit.' We need to discover *in what way* a model is inadequate, so as to suggest appropriate modification. . . .
>
> No model form ever represents the truth absolutely. It follows that, given sufficient data, statistical tests can discredit models which could nevertheless be entirely adequate for the purpose at hand. Alternatively, tests can fail to indicate serious departures from assumptions because these tests are insensitive to the types of discrepancies that occur. The best policy is to devise the most sensitive statistical procedures possible but be prepared, for sufficient reason, to employ models which exhibit slight lack of fit. Know the facts as clearly as they can be shown – then use judgment.
>
> Clearly, diagnostic checks must be such that they *place the model in jeopardy*. That is to say, they must be sensitive to discrepancies which are likely to happen. No system of diagnostic checks can ever be comprehensive, since it

> is always possible that characteristics in the data of an unexpected kind could be overlooked. However, if diagnostic checks, which have been thoughtfully devised, are applied to a model fitted to a reasonably large body of data and fail to show serious discrepancies, then we shall rightly feel more comfortable about using that model. (*ibid.*, pages 286–7: italics in original)

9.46 One technique proposed by Box and Jenkins is that of *overfitting*: '(h)aving identified what is believed to be a correct model, we actually fit a more elaborate one. This puts the identified model in jeopardy, because the more elaborate model contains additional parameters covering feared directions of discrepancy' (*ibid.*, page 286). They emphasized that care needed to be taken as to how the model should be augmented: for example, additional autoregressive and moving average terms should not be added simultaneously as this may lead to model redundancy, as discussed in Box and Jenkins (1970, section 7.3.5).

The ARMA(1,1) model for Series A was subjected to overfitting by estimating both ARMA(2,1) and ARMA(1,2) models, producing

$$\begin{array}{llllll} x_t - & \underset{(\pm 0.15)}{1.05}\, x_{t-1} + & \underset{(\pm 0.12)}{0.11} = 1.14 + a_t - & \underset{(\pm 0.13)}{0.68}\, a_{t-1} & & \hat{\sigma}_a^2 = 0.098 \end{array}$$

$$\begin{array}{llllll} x_t - & \underset{(\pm 0.04)}{0.94}\, x_{t-1} = 1.06 + a_t - & \underset{(\pm 0.08)}{0.59}\, a_{t-1} - & \underset{(\pm 0.08)}{0.08}\, a_{t-2} & \hat{\sigma}_a^2 = 0.098 \end{array}$$

In both cases the additional parameter is insignificant, thus providing no evidence that the ARMA(1,1) model is inadequate.

Because the model is extended in a particular direction, overfitting assumes that we know what kind of discrepancies are to be feared. Box and Jenkins also considered procedures that were less dependent upon knowledge of this type, being based on the analysis of the residuals $\hat{a}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)x_t$, for if the fitted model was inadequate in some way this should be reflected in the existence of patterns and predictabilities in the $\hat{a}_t$, which should mimic white noise if the fitted model is an adequate representation of the data.

If the form of the model and the true parameter values $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ were actually known then, using the results of §**9.2**, the autocorrelations of the a's, the $r_k(a)$, would be uncorrelated and approximately normally distributed about zero with variance T^{-1}, so that the statistical significance of apparent departures of these autocorrelations from zero could be assessed. In practice, of course, the true values $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are unknown and we only have their estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$, from which the residuals $\hat{a}_t$, but not the true innovations a_t, may be calculated. Although the autocorrelations $r_k(\hat{a})$ of the residuals can yield valuable evidence concerning lack of fit and the possible nature of model inadequacy, it might be dangerous to make this assessment on the basis of a standard error of $T^{-\frac{1}{2}}$. To confirm this, Durbin (1970) showed that, for an AR(1) process with parameter ϕ, the variance of $r_1(\hat{a})$ was $\phi^2 T^{-1}$, which could be substantially less than T^{-1}. Box and Pierce (1970) derived the large sample variances and covariances of the $\hat{a}$s from any ARMA process, and showed that $T^{-\frac{1}{2}}$ should be regarded as an upper bound for the standard error of $r_k(\hat{a})$, and its use could seriously underestimate the significance of apparent departures from zero of the residual autocorrelations for small values of k, although for moderate to large values of k this standard error estimate would be accurate.

Box and Pierce (1970) also considered assessing the significance of a group of residual autocorrelations, rather than just examining the $r_k(\hat{a})$ individually. They showed that if the fitted ARMA(p, q) model was appropriate then, for the group containing the first K autocorrelations, the statistic

$$Q(K) = T\sum_{k=1}^{K} r_k^2(\hat{a})$$

was approximately distributed as $\chi^2(K - p - q)$, so that significantly large values of $Q(K)$ would indicate model inadequacy of some form.

For the residuals obtained from the ARMA(1,1) fit to Series A, $Q(20) = 23.50 \sim \chi^2(18)$, which is not significant at the 10% level and so offers no evidence against the adequacy of this model. For the AR(2) fit to the sunspot index, $Q(20) = 53.02 \sim \chi^2(18)$, which is significant at the 0.1% level and thus confirms the inadequacy of this model found previously. However, the statistic for the AR(9)* model is $Q(20) = 21.77 \sim \chi^2(17)$, which is insignificant at the 10% level (note that only three AR coefficients have actually been fitted for this restricted model, so that the degrees of freedom are $20 - 3 = 17$).

Box and Jenkins emphasized that a variety of other diagnostic checks should be performed on the residuals from a fitted ARMA model, such as examining the cumulative periodogram, and the adequacy of a model could also be assessed by looking at the stability of the parameter estimates across sub-samples of the data.

9.47 If the residuals are found to be correlated then this information can be used to identify a modified model and the three-stage model-building strategy could then be repeated. For example, suppose the residuals b_t from the model

$$\phi_0(B)x_t = \theta_0(B)b_t \tag{9.38}$$

are non-random and from their autocorrelation function the model

$$\overline{\phi}(B)b_t = \overline{\theta}(B)a_t \tag{9.39}$$

was identified. Eliminating b_t from (9.38) and (9.39) leads to the new model

$$\phi_0(B)\overline{\phi}(B)x_t = \theta_0(B)\overline{\theta}(B)a_t$$

which can now be fitted and diagnostically checked. For example, if, after fitting an ARMA(1,1) model to series A, the residuals had been found to follow an AR(1) process $(1-\overline{\phi}B)=b_t$, then the ARMA(2,1) model

$$(1-\phi B)(1-\overline{\phi}B)x_t=(1-\phi_1 B-\phi_2 B^2)x_t=(1-\theta B)a_t$$

could then be fitted in a second iteration of the modelling strategy.

9.48 The research effort over the thirty-year period from 1940 to 1970 thus produced a practical methodology of inference and estimation that enabled ARMA models to be identified, estimated and checked. Although Box and Jenkins (1970) may be regarded as a synthesis of this research programme, it was much more than that, for it also extended the analysis to nonstationary time series and to the modelling of the relationships between such series, and it is to these areas that we now turn.

10

Dealing with Nonstationarity: Detrending, Smoothing and Differencing

Early recognition of the presence of nonstationarity

10.1 As we discussed in §§**2.6–2.9**, Hooker (1901b, 1905) was the first to be concerned with the problems of dealing with time series containing trends, proposing both differencing and the use of moving averages to 'detrend' the data prior to statistical analysis.[1] Beveridge (1921, 1922) later used a variation on the moving average to eliminate a secular trend from his wheat prices before subjecting them to periodogram analysis (§§**3.8–3.9**). The variate differencing approach examined in detail in Chapter 4 explored the link between successive differencing and fitting polynomials in time to a series, with Persons (1917) explicitly considering the decomposition of an observed time series into various unobserved components, one of which was the secular trend (§**4.11**). Indeed, the identification and removal of the trend component became a preoccupation of many analysts of time series data for much of the twentieth century, even though it was conceded that even the definition of a trend posed considerable conceptual problems: as Kendall (1941, page 43) remarked

> (t)he concept of 'trend', like that of time itself, is one of those ideas which are generally understood but difficult to define with exactitude. A movement which has the evolutionary appearance of a trend over a period of thirty or forty years may in reality be one phase of an oscillatory movement of greater extent. A good deal depends on the length of the series under consideration whether we regard any particular tendency in the series as a trend, or a long-term movement, or an oscillation, or short-term movement. But in any case we require of a trend curve that it shall exhibit only the general direction of the time-series, and in practice this amounts to saying that it must be representable, at least locally, by a smooth non-periodic function such as a polynomial or a logistic curve.

In his own research on agricultural time series, discussed in §**8.3**, Kendall (1943, 1944) eliminated trends by taking nine-year moving averages prior to analysing their oscillatory movements.

In contrast to the removal of trends for the purpose of concentrating on shorter-run components such as cycles and seasonal patterns, Macaulay's (1931) focus was on the underlying trend itself, i.e., on the 'smoothed' series, as his primary interest was in examining the longer-run relationships between economic and financial time series in the absence of confounding short-run, transitory, fluctuations. While an underlying smooth function can be fitted to stationary series, particularly those having a cyclical pattern of some form (see, for example, Spencer-Smith, 1947), Macaulay (1931, pages 39–40) was primarily interested in the behaviour of trending economic time series that

> seem to be of a type somewhat analogous to … cumulated chance series. Some economic series suggest chance series which have been cumulated twice, … (since) each observation is not only highly correlated with the immediately preceding observation, but the first differences are highly correlated with the preceding first differences. The commonest type of economic time series suggest a cumulated chance series on which has been superposed another but non-cumulated chance series and a more or less regular and unchanging seasonal fluctuation.
>
> How should such series be smoothed? What sort of procedure would seem to eliminate all seasonal and erratic fluctuations leaving a reasonable picture of the cyclical fluctuations and any underlying trend?

Such a view, based on practical experience, was obviously very close to that of Working's, which was discussed in §§**5.19–5.21**. From a theoretical perspective, Wold (1938), in the opening paragraph of his treatise, also clearly made the distinction between stationary and nonstationary (or, in his term, evolutive) time series, as the quote offered in §**7.2** makes quite apparent.

These three approaches, that of transforming the data using moving averages, fitting a deterministic function of time, typically a polynomial or perhaps some other simple function, and taking successive differences of the data (or the closely related transformation of taking percentage changes), have remained the basic ways of eliminating trends, i.e., of transforming a nonstationary time series into a stationary one, up to the present day. However, the statistical implications of these approaches and the links between them took many years to discover and tease out, with several early contributions coming from a discipline yet to be encountered in this narrative, that of actuarial science.

'Graduation' by moving averages

10.2 An important task in actuarial science is to describe the actual but unknown mortality pattern of a population. To do this, the actuary calculates crude mortality rates from raw data, which usually form an irregular series, and then revises them to produce smoother estimates of mortality, a procedure termed by actuaries as 'graduation'. Various methods have been proposed to accomplish this, but one of the earliest and most enduring was the use of iterated moving averages, as originally suggested by Spencer (1904, 1907) and outlined in Whittaker and Robinson (1924, chapter XI).

Suppose that the primary series to be graduated is u_t and, as usual, define $\Delta u_t = u_{t+1} - u_t$ to be the (first-order) difference and $\Delta^d u_t$ to be the dth-order difference (cf. §**4.1**). Whittaker and Robinson introduced the notation

$$[2m+1]u_t = \sum_{j=-m}^{m} u_{t+j}$$

to denote the sum of $2m+1$ u's centred on u_t. It is then possible to find combinations of the operations Δ and [ ] which, when differences above a certain order are neglected, reproduce the series operated on, i.e.,

$$f(\Delta, [\,])u_t = u_t + \text{higher differences}$$

The function $f(\Delta, [\,])u_t$ is then taken to be the graduated value of u_t, denoted v_t,

$$v_t = f(\Delta, [\,])u_t$$

with 'the merit of this v_t depending on the circumstance that $f(\Delta, [\,])\,u_t$ involves a large number of the observed u's, whose errors to a considerable extent neutralise each other and so produce a smoothed value v_t in place of u_t' (Whittaker and Robinson, 1924, pages 288–9).

Whittaker and Robinson then introduced the *central difference* δ, defined such that $\delta^2 = u_{t+1} - 2u_t + u_{t-1}$, and showed that

$$\frac{[p]\,[q]\,[r]}{p \cdot q \cdot r}u_t = u_t + \frac{(p^2-1)+(q^2-1)+(r^2-1)}{24}\delta^2 u_t + \text{terms in } \delta^4 u_t, \delta^6 u_t, \ldots$$

Thus, if $f(\,)$ is a cubic in t, terms in fourth- and higher-order differences vanish, and the graduated value

$$v_t = \frac{[p]\,[q]\,[r]}{p \cdot q \cdot r}\left\{1 - \frac{p^2+q^2+r^2-3}{24}\delta^2\right\}u_t \qquad (10.1)$$

then reproduces the cubic perfectly.

10.3 With this result, *Spencer's 21-term moving average* is defined by substituting $p=q=5$ and $r=7$ into (10.1) to obtain

$$v_t = \frac{[5][5][7]}{5\cdot 5\cdot 7}(1-4\delta^2)\,u_t \tag{10.2}$$

If the [ ] notation is extended to *weighted* moving sums such that

$$[w_{-m},\ldots,w_0,\ldots,w_m]u_t = \sum_{j=-m}^{m} w_j u_{t+j}$$

then, since

$$(1-4\delta^2)u_t = -4u_{t+1}+9u_t-4u_{t-1} = [-4,9,4]u_t$$

the expression (10.2) can be written as the iterated moving average

$$v_t = \frac{[5][5][7][-4,9,4]}{5\cdot 5\cdot 7}u_t$$

This may then be expanded to obtain

$$\begin{aligned} v_t = {} & \tfrac{1}{350}\{60u_t + 57(u_{t-1}+u_{t+1}) + 47(u_{t-2}+u_{t+2}) + 33(u_{t-3}+u_{t+3}) \\ & + 18(u_{t-4}+u_{t+4}) + 6(u_{t-5}+u_{t+5}) - 2(u_{t-6}+u_{t+6}) - 5(u_{t-7}+u_{t+7}) \\ & - 5(u_{t-8}+u_{t+8}) - 3(u_{t-9}+u_{t+9}) - (u_{t-10}+u_{t+10})\} \end{aligned}$$

or

$$\begin{aligned} v_t = {} & 0.171u_t + 0.163(u_{t-1}+u_{t+1}) + 0.134(u_{t-2}+u_{t+2}) + 0.094(u_{t-3}+u_{t+3}) \\ & + 0.051(u_{t-4}+u_{t+4}) + 0.017(u_{t-5}+u_{t+5}) - 0.006(u_{t-6}+u_{t+6}) \\ & - 0.014(u_{t-7}+u_{t+7}) - 0.014(u_{t-8}+u_{t+8}) - 0.009(u_{t-9}+u_{t+9}) \\ & - 0.003(u_{t-10}+u_{t+10}) \end{aligned}$$

which is a symmetric moving average containing ten leads and ten lags of the primary series u_t to obtain the graduation v_t.

In a similar fashion, *Spencer's 15-term moving average* is defined with $p=q=4$ and $r=5$, leading to

$$v_t = \frac{[4][4][5]}{4\cdot 4\cdot 5}\left(1-\frac{9}{4}\delta^2\right)u_t = \frac{1}{320}[4][4][5][-9,22,-9]u_t$$

which is a symmetric moving average containing seven leads and seven lags of u_t:

$$v_t = \tfrac{1}{320}\{74u_t + 67(u_{t-1} + u_{t+1}) + 46(u_{t-2} + u_{t+2}) + 21(u_{t-3} + u_{t+3})$$
$$+ 3(u_{t-4} + u_{t+4}) - 5(u_{t-5} + u_{t+5}) - 6(u_{t-6} + u_{t+6}) - 3(u_{t-7} + u_{t+7})$$

or

$$v_t = 0.231u_t + 0.209(u_{t-1} + u_{t+1}) + 0.144(u_{t-2} + u_{t+2}) + 0.066(u_{t-3} + u_{t+3})$$
$$+ 0.009(u_{t-4} + u_{t+4}) - 0.016(u_{t-5} + u_{t+5}) - 0.019(u_{t-6} + u_{t+6})$$
$$- 0.009(u_{t-7} + u_{t+7})$$

10.4 As an example of graduation, suppose that u_t is generated by a cubic with an added random element, specifically

$$u_t = f(t) + \varepsilon_t = (t - 26) + \tfrac{1}{10}(t - 26)^2 + \tfrac{1}{100}(t - 26)^3 + \varepsilon_t$$

where the ε_t are drawn from a uniform distribution ranging from -149.5 to $+149.5$. Figure 10.1 shows the cubic curve $f(t)$, the primary series u_t, and the graduation v_t obtained using the 21-term Spencer moving average (10.2), for a segment of the observations running from $t = 20$ to 80. The graduation certainly does its job of smoothing out the fluctuations in u_t but it does not reproduce $f(t)$ exactly. Moreover, from Figure 10.2 it can be seen that the deviations from the

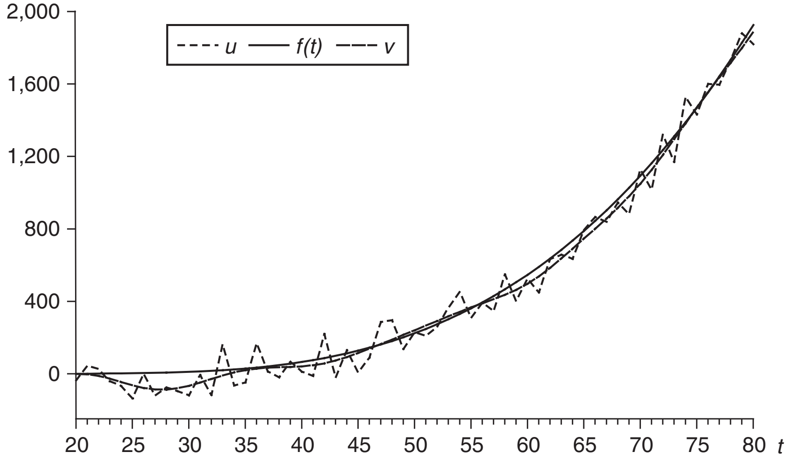


Figure 10.1 Plots of the cubic $f(t)$, the primary series u_t and the graduation v_t for $t = 20$ to 80

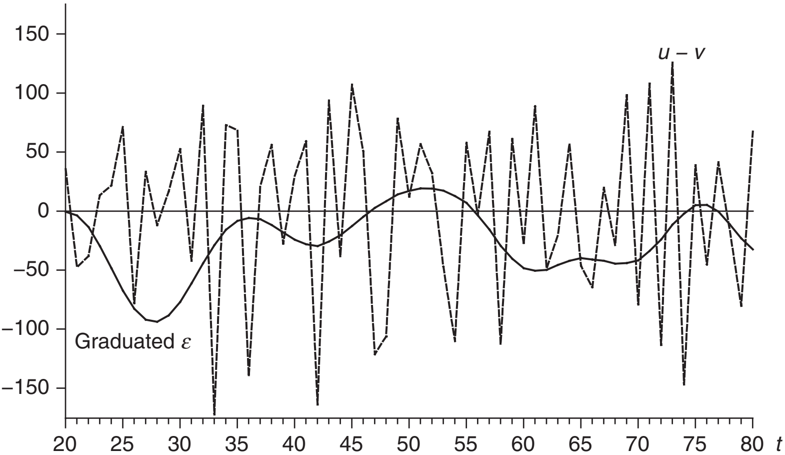


Figure 10.2 Deviations from the graduation, $u_t - v_t$, and the graduation of ε_t

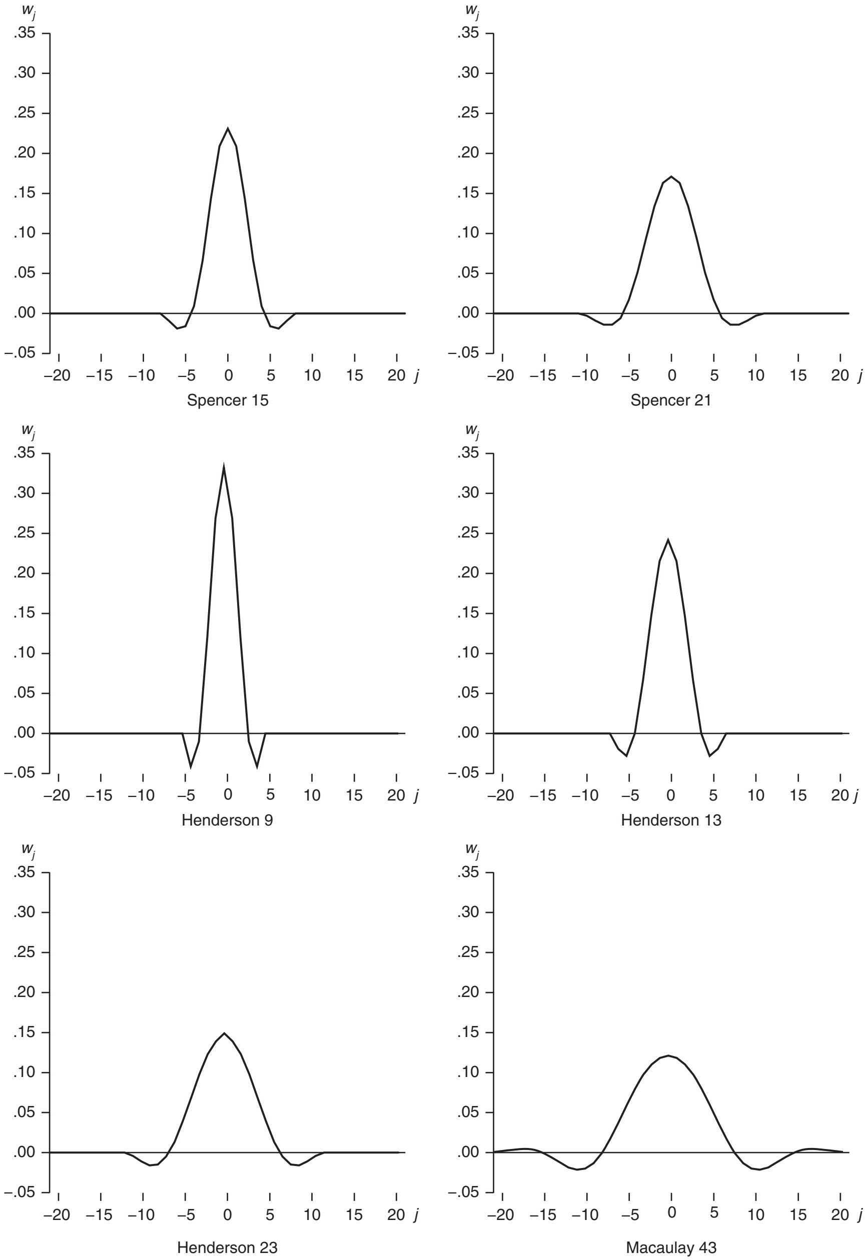
fitted graduation, $u_t - v_t$, are not random, as should be expected, but actually have a first-order sample autocorrelation of −0.52. Also shown in Figure 10.2 is the graduation of the random element ε_t itself: this graduation is certainly not random, having a negative mean with fluctuations that oscillate, and these impart the small oscillatory movement to the graduation v_t that can be seen in Figure 10.1.

The reasons for these effects, which passed unnoticed in the conventional actuarial applications of the early twentieth century, as they were focused directly on the smoothed series, became of concern when series were detrended by a moving average prior to their analysis as potentially oscillatory series during the 1940s (cf. §**8.3**). This important issue will be returned to in §§**10.9–10.13**.

10.5 Macaulay (1931) was perhaps the first to use such graduations, or 'smooths', outside of the actuarial profession, examining a wide variety of moving averages and other methods to be discussed below. These tended to be considerably more complicated than the Spencer moving averages, primarily because Macaulay was concerned with removing seasonal patterns from monthly data as well as erratic fluctuations. His favoured smooth was a 43-term symmetric moving average which closely approximated a fifth-order polynomial and was defined, using the notation introduced above, as

$$v_t = \frac{1}{9600}[5][5][8][12][7, -10, 0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0, 0, -10, 7]\, u_t$$

The set of weights $w_0, w_{\pm 1}, \ldots, w_{\pm 21}$, often referred to as the *weight function*, is shown in Figure 10.3, along with the weight functions of various other moving averages discussed in this chapter.



Spencer 15

Spencer 21

Henderson 9

Henderson 13

Henderson 23

Macaulay 43

Figure 10.3 Weight functions for Spencer, Macaulay and Henderson moving averages

Henderson–Whittaker moving averages

10.6 In a sequence of papers, Henderson (1916, 1924) and Whittaker (1923, 1924) independently considered the problem of designing a smoother (later more commonly known as a *filter*) that, as well as reproducing a cubic polynomial trend without distortion, would also satisfy certain *smoothness conditions*. The primary condition was that the filter should minimize the variance of the third differences of the smoothed series, i.e., it should minimize $Var(\Delta^3 v_t)$. In deriving this filter, the approach of Kenny and Durbin (1982, Appendix) is followed, as this is more accessible and transparent than the original derivations of Henderson and Whittaker.

Given the symmetric moving average of length $2m+1$

$$v_t = \sum_{j=-m}^{m} w_j u_{t+j}$$

with u_t represented as the sum of a cubic in t and a random error, i.e., as $u_t = f(t) + \varepsilon_t$, then the third differences of v_t can be written as

$$\Delta^3 v_t = \sum_{j=-m}^{m} \Delta^3 w_j \varepsilon_{t+j} + \sum_{j=-m}^{m} w_j \Delta^3 f(t+j)$$

Since $\sum w_j = 1$ and the third difference of a cubic is a constant, this reduces to

$$\Delta^3 v_t = \sum_{j=-m}^{m} \Delta^3 w_j \varepsilon_{t+j} + \text{constant}$$

so that, given the properties of ε_t,

$$Var(\Delta^3 v_t) = Var(\varepsilon_t) \sum_{j=-m}^{m} (\Delta^3 w_j)^2$$

From this expression it is seen that the weights which minimize the variance of the third differences of v_t will also be those which minimize the sum of squares of the third differences of the weights themselves. The requirement that v_t must itself follow a cubic, i.e., that

$$v_t = \sum_{j=-m}^{m} w_j (f(t+j) + \varepsilon_{t+j}) = f(t)$$

implies that the required weights can be obtained as the solution of the following problem: minimize

$$\sum_{j=-m}^{m} (\Delta^3 w_j)^2$$

subject to the constraints

$$\sum_{j=-m}^{m} w_j = 1 \tag{10.3}$$

and

$$\sum_{j=-m}^{m} j^k w_j = 0 \quad k = 1, 2, 3 \tag{10.4}$$

Condition (10.4) is automatically satisfied for $k = 1, 3$ if the weights are symmetric and Kenny and Durbin claim that the solution to this constrained minimization problem is given by

$$\Delta^6 w_{j-3} = \alpha + \beta j^2 \quad j = -m, \ldots, m \tag{10.5}$$

where α and β are chosen so that (10.3) and (10.4) are satisfied (it is understood that $w_j = 0$ for $|j| > m$).

If (10.5) is true then w_j has to be generated by an eighth-order polynomial in j which must be valid for $|j| \leq m + 3$. This implies that

$$w_j = ((m+1)^2 - j^2)((m+2)^2 - j^2)((m+3)^2 - j^2)(a + bj^2)$$

where a and b are determined by (10.3) and (10.4). It can then be shown that

$$w_j \propto ((m+1)^2 - j^2)((m+2)^2 - j^2)((m+3)^2 - j^2)(3(m+2)^2 - 16 - 11j^2)$$

with the constant of proportionality being chosen to ensure that (10.3) is satisfied.

The same set of weights are obtained by fitting a cubic to u_{t+j}, $j = -m, \ldots, m$, by weighted least squares and taking the value of the fitted function at t as the smoothed value v_t, i.e., by minimizing the sum of squares function

$$\sum_{j=-m}^{m} w_j (u_{t+j} - a - bj - cj^2 - dj^3)^2$$

The weight functions for $m = 4$, 6 and 11 (i.e., the Henderson 9-, 13- and 23-term moving averages) are also shown in Figure 10.3. These filters were adopted for trend estimation in the X-11 seasonal adjustment procedure by Shiskin, Young and Musgrave (1967), replacing the Spencer moving averages used in earlier versions of the US Bureau of the Census seasonal adjustment procedure: see §§**14.7–14.9** for further discussion.

10.7 Almost forty years later, Leser (1961) revisited the Whittaker–Henderson approach and extended the methodology by deriving the weights using the principle of penalized least squares, in which a linear combination of two sums of squares is minimized. The first sum of squares contains the deviations of the observations u_t from the filter v_t, the second contains the second differences of successive smoothed values $\Delta^2 v_t$, with the linear combination of the two being defined by the weights of unity and λ, i.e., for the observed sequence $u_1, u_2, \ldots, u_T$, the minimand is

$$\sum_{t=1}^{T} (u_t - v_t)^2 + \lambda \sum_{t=2}^{T-1} (\Delta^2 v_{t+1})^2 \tag{10.6}$$

The first term measures the goodness of fit of the filter, the second penalizes the departure from zero of the variance of the second differences of the filter, so that it is a measure of smoothness: hence λ is referred to as the smoothness parameter. Successive partial differentiation of (10.6) with respect to the sequence v_t leads to the first-order conditions

$$\Delta^2 v_{t+2} - 2\Delta^2 v_{t+1} + \Delta^2 v_t = \lambda(u_t - v_t)$$

so that, given T and λ, v_t is a linear function of u_t with time-varying weights:

$$v_t = \sum_{j=1}^{T} w_{t,j} u_j$$

Leser then developed an algebraic method of obtaining the coefficients $w_{t,j}$, providing a number of examples in which the solutions were obtained in, it has to be admitted, laborious and excruciating detail, which must have lessened the impact of the paper at the time! However, its historical importance lies in the fact that the method developed by Leser was exactly that proposed some two decades later by Hodrick and Prescott (1997) and which has entered into macroeconomic modelling as the H–P filter.[2]

Fitting local polynomial trends

10.8 The Spencer and Macaulay moving averages were initially popular because they could be computed essentially as a sequence of summations, which minimized the computational burden. As computational requirements became less of a concern, attention focused on the direct fitting of local polynomials. The general approach is to take the first n terms of a time series, $u_1, \ldots, u_n$ say, where n is taken to be an odd number, fit a polynomial of degree $p \leq n - 1$ to these observations, and use this polynomial to determine the 'trend' value v_t for $t = (n+1)/2$ (the choice of an odd value of n ensures that a unique 'middle' value exists at any observed time). The operation is then repeated using the terms $u_2, \dots, u_{n+1}$ to obtain the next trend value $v_{(n+3)/2}$, and the operation is repeated throughout the time series, finally obtaining, for the terms $u_{T-n+1}, \dots, u_T$, the trend value $v_{T-(n-1)/2}$.[3]

While this procedure would, on the face of it, require the continual fitting of a pth-degree polynomial by least squares, the recursive nature of the computations enabled the trend values to be calculated directly as a weighted moving average. To see this, put $n = 2m+1$ and, without loss of generality, consider the sequence of terms $u_{-m}, u_{-m+1}, \dots, u_0, \dots, u_{m-1}, u_m$. To fit a polynomial of degree p by least squares to this sequence requires solving the $p+1$ equations

$$\frac{\partial}{\partial a_j}\sum_{t=-m}^{m}(u_t - a_0 - a_1 t - \cdots - a_p t^p)^2 = 0 \quad j = 0, 1, \dots, p$$

which gives equations of the form

$$\sum t^j u_t - a_0 \sum t^j - a_1 \sum t^{j+1} - \cdots - a_p \sum t^{j+p} = 0 \quad j = 0, 1, \dots, p \qquad (10.7)$$

Since the summations in (10.7) are functions of m only, solving for a_0 yields an equation of the form

$$a_0 = c_0 + c_1 u_{-m} + c_2 u_{-m+1} + \cdots + c_{2m+1} u_m \qquad (10.8)$$

where the c's depend on m and p, but not on the u's. As $u_0 = a_0$ at $t = 0$, this value, as given by (10.8), is the value required for the polynomial and is seen to be a weighted average of the observed sequence of values, the weights being independent of which part of the series is being used. The process of fitting the polynomial trend then consists of determining the constants c and then calculating, for each consecutive sequence of $2m+1$ terms of the series, a value given by (10.8): if the sequence is $u_k, \dots, u_{2m+k}$, the calculated value will correspond to $t = m+k$.

As an example of the procedure, suppose $m = p = 3$, so that the cubic

$$u_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

is fitted to sequences of seven terms. Since the origin is $t = 0$, the summations in (10.7) are

$$\sum t^0 = 7; \quad \sum t^2 = 28; \quad \sum t^4 = 196; \quad \sum t^6 = 1588;$$

$$\sum t = \sum t^3 = \sum t^5 = \sum t^7 = 0$$

and the set of equations are

$$\begin{aligned}\sum u &= 7a_0 & &+28a_2 & \\ \sum tu &= & 28a_1 & & +196a_3 \\ \sum t^2 u &= 28a_0 & &+196a_2 & \\ \sum t^3 u &= & 196a_1 & & +1588a_3\end{aligned} \tag{10.9}$$

These may be solved to give, for a_0,

$$\begin{aligned}a_0 &= \frac{1}{21}\left(7\sum u - \sum t^2 u\right) \\ &= \frac{1}{21}(-2u_{-3} + 3u_{-2} + 6u_{-1} + 7u_0 + 6u_1 + 3u_2 - 2u_3) \\ &= \frac{1}{21}[-2, 3, 6, 7, 6, 3, -2]\end{aligned}$$

To illustrate this example, suppose the series is given by the following values

t	1	2	3	4	5	6	7	8	9	10
u_t	0	1	8	27	64	125	216	343	512	729

The trend value at $t = 4$ is then

$$\begin{aligned}a_0 &= \frac{1}{21}((-2 \times 0) + (3 \times 1) + (6 \times 8) + \cdots - (2 \times 216)) \\ &= \frac{1}{21}567 = 27\end{aligned}$$

which is, of course, equal to the actual value u_4 since a cubic is being fitted to the series $u_t = (t-1)^3$. In (10.9) it is seen that a_0 does not depend on a_3, so that the same value for a_0 would have been obtained if a quadratic rather than a cubic had been fitted. This is a general result: fitting a polynomial of odd degree p gives the same trend values as fitting a polynomial of even degree $p-1$. The implied moving averages for $p \leq 5$ and $m \leq 10$ are given in, for example, Kendall, Stuart and Ord (1983, §46.6).[4,5]

Oscillations induced by taking moving averages

10.9 Suppose an observed series y_t has a decomposition into trend, τ_t, oscillatory, γ_t, and random, ε_t, components

$$y_t = \tau_t + \gamma_t + \varepsilon_t$$

and a moving average of the form

$$Wy_t = [w_{-m}, \ldots, w_0, \ldots, w_m]y_t = \sum_{j=-m}^{m} w_j y_{t+j} \qquad \sum_{j=-m}^{m} w_j = 1$$

is applied, so that

$$W y_t = W\tau_t + W\gamma_t + W\varepsilon_t$$

As Kendall (1941) pointed out, the ideal moving average is one that reproduces the trend exactly, i.e., $W\tau_t = \tau_t$, in which case the 'detrended' series is

$$y_t - W y_t = \gamma_t + \varepsilon_t - W\gamma_t - W\varepsilon_t \tag{10.10}$$

> (T)he point to be emphasized is that the existence of the terms $W\gamma_t$ and $W\varepsilon_t$ in [10.10] may introduce oscillatory terms which were not, or annihilate oscillatory terms which were, in the original y_t. That is to say, the method of moving averages may induce into the data oscillations which are entirely spurious or may reduce or remove oscillations which are entirely genuine. (*ibid.*, page 45: notation altered for consistency)

10.10 Kendall considered first the effect on the random component of taking a moving average. Given that the typical moving average can be expressed as an iteration of simple sums (or, to be precise, averages), the results of Slutzky (1937), discussed in §§**5.11–5.16**, and those provided by Dodd (1939, 1941a, 1941b) on the effect of summing random series may be used. Thus suppose that

$$\varepsilon_t^{[2]} = W\varepsilon_t = \frac{[2m+1]}{2m+1}\varepsilon_t = \frac{1}{2m+1}\sum_{j=-m}^{m}\varepsilon_{t+j} = \frac{1}{n}\sum_{j=-m}^{m}\varepsilon_{t+j}$$

is a simple moving average of ε_t. If ε_t is random, so that consecutive values are independent, consecutive values of $\varepsilon_t^{(2)}$ will not be independent, since $\varepsilon_t^{(2)}$ and $\varepsilon_{t+k}^{(2)}$ will have $n-k$ values of ε_t in common and will thus be correlated if $n > k$. $\varepsilon_t^{(2)}$ will then be much smoother than the random series ε_t and, if further moving averages are taken, the result will be smoother still. Indeed, as Slutzky pointed out, after only a few summations the resulting series becomes very smooth, having fluctuations with varying amplitude and phase and periods concentrated around a particular modal value – just those features that are characteristic of oscillatory time series.

Dodd utilized the following useful geometrical result. Consider the two sums

$$x_t = \sum_{j=1}^{n} a_j \varepsilon_{t+j}$$

$$z_t = \sum_{j=1}^{n} b_j \varepsilon_{t+j}$$

where it is now assumed that the ε_t are normally distributed with zero mean and constant variance, V say. Treating x_t and y_t as planes, the cosine of the angle θ between them is given by

$$\cos\theta = \frac{\sum a_j b_j}{\left(\sum a_j^2 \sum b_j^2\right)^{1/2}}$$

When θ is expressed in radians, $\theta/360$ has the interpretation of being the probability that x_t and y_t are of opposite sign. Using this result, it follows that the probability of x_t and x_{t+1} changing signs is obtained from

$$\cos\theta = \frac{\sum a_j a_{j+1}}{\sum a_j^2}$$

The change of sign from negative to positive between successive values of x_t is known as an 'upcross', so that the mean distance between upcrosses is $2\pi/\theta$ (cf. §**8.1**: this, of course, is also the mean distance between 'downcrosses' – changes of sign from positive to negative). For

$$\Delta x_t = x_{t+1} - x_t = b_1\varepsilon_t + b_2\varepsilon_{t+1} + \cdots + b_n\varepsilon_{t+n-1} + b_{n+1}\varepsilon_{t+n}$$

the probability that $\Delta x_{t-1} > 0$ and $\Delta x_t < 0$, i.e., that x_t is a maximum, is then $\theta'/360$, and the mean 'peak to peak' distance between maxima is $2\pi/\theta'$, where

$$\cos\theta' = \frac{\sum b_j b_{j+1}}{\sum b_j^2}$$

$$b_1 = -a_1, \quad b_{n+1} = a_n, \quad b_j = a_{j-1} - a_j, \quad j = 2, 3, \ldots, n-1$$

Dodd considered various extensions and generalizations of these results. For example, minor oscillations, or 'ripples', may be eliminated by requiring that, for maxima, the condition $x_t > x_{t+p}$, for p arbitrarily chosen, must hold along with $\Delta x_{t-1} > 0$ and $\Delta x_t < 0$. The assumption of normal random variation may be relaxed and these results seem to be applicable to various other distributional assumptions, being used previously in §**8.1** to measure the oscillatory properties of various of Kendall's agricultural series.

10.11 The amplitude of the induced oscillations in $W\varepsilon_t$ was also considered by Kendall. Since $\varepsilon_t^{(2)}$ is the sum of n independent random variables each with variance V, it will have variance V/n. As further sums are taken, the variance of these sums becomes progressively more complicated to derive, although an expression was given in Kendall (1941, equation (11)). The general effect is clear, however:

> the variance of the series ... is reduced very considerably by the first averaging but less so by subsequent averagings, and this is what we might expect from

> the correlations between members of the series. For example, when $n = 7$, the first averaging reduced the variance by $\frac{1}{7}$, whereas the next four averagings reduce it by little more than a further $\frac{1}{2}$. (*ibid.*, page 47)

Although oscillatory movements in $W\varepsilon_t$ will thus tend to be small compared to the random fluctuations in ε_t itself if n is large, they are not necessarily negligible: as Kendall pointed out, even though a periodogram analysis of ε_t would reveal no periodicities, an analysis of $W\varepsilon_t$ may and probably would.

To reduce the effect of $W\varepsilon_t$ as much as possible, n should be made large rather than increasing the number of iterations of the moving average, i.e., the individual weights should be as small and as equal as possible. Unfortunately, this runs counter to the size of the weights that are required to eliminate the trend. As the weight functions shown in Figure 10.3 reveal, these have individual weights that are very far from being equal, being determined from several iterations of simple summations. Interestingly, Macaulay's 43-term moving average can be shown to reduce the variance of a random series to about 0.11 of its original value, roughly the same as a simple nine-term moving average.

10.12 Kendall then considered the effect of taking a moving average on the genuinely oscillatory part of the original series, i.e., on the behaviour of $W\gamma_t$. Suppose that this component follows a simple sine wave, $\gamma_t = \sum_{j=1}^{n} \sin(\alpha + j\lambda)$. Since

$$\sum_{j=1}^{n} \sin(\alpha + j\lambda) = \frac{\sin \frac{1}{2} n\lambda}{\sin \frac{1}{2}\lambda} \sin\left(\alpha + \frac{1}{2}(n-1)\lambda\right)$$

a simple moving average of n consecutive terms centred at the middle term will result in a sine series of the same period and phase as the original, but with its amplitude reduced by the factor

$$\frac{1}{n} \frac{\sin \frac{1}{2} n\lambda}{\sin \frac{1}{2}\lambda}$$

Iterating q times will reduce the amplitude by the qth power of this factor. This implies that $W\gamma_t$ will be small if n and q are both large or if $\frac{1}{2} n\lambda = 0 \,(\text{mod}\, \pi)$, i.e., if the extent of the moving average is a period of the oscillation. On the other hand, if λ and $n\lambda$ are small then the amplitude will barely be reduced at all and $\gamma_t - W\gamma_t$ will largely disappear because the moving average will partially obliterate the harmonic term in γ_t. With $n\lambda$ being small, the extent of the moving average will be short compared to the period of the harmonic. The oscillation will then be a very slow one and will be treated as part of the trend by the moving average and eliminated accordingly. The moving average will therefore emphasize the shorter oscillations at the expense of the longer ones.

If, on the other hand, the moving average is longer than the period, $W\gamma_t$ may have the original oscillation but with the sign reversed, so that the fluctuations from trend may exaggerate the true oscillations. Kendall thus concluded that

> in the study of oscillations obtained from a time-series by eliminating trend with moving averages it is desirable to safeguard against the introduction of spurious effects and the distortion of genuine effects due respectively to the random and oscillatory terms of the original series. This can best be done by extending the moving average so far as possible and by making it approximate to a multiple of any cycles which are suspected to exist. Iteration rapidly reduces the distortion of genuine oscillatory movements, but does not exert such a great effect on the spurious cycles due to random fluctuations.
>
> These considerations support the desirability of extending the moving average as far as possible; but other considerations will work in the reverse direction. The saving of arithmetic; the avoidance of sacrificing terms at the beginning and end of the series; and the nature of the weighting dictated by trend elimination itself are factors of this kind. (*ibid.*, page 49)

10.13 Given these considerations, Kendall was able to argue that, although the mean period of the oscillations induced by taking a nine-term moving average to eliminate the trend in his agricultural series was not sufficiently different from the observed mean period to dispose of the suggestion that the observed oscillations were spurious, the use of variate differencing revealed that the variance of the random component was almost certainly very much smaller than that implied by the process of moving averaging. Hence Kendall was able to conclude that detrending his series in this way did not induce spurious oscillations and that the oscillatory character of the detrended series were indeed an inherent feature of the data.[6]

Modelling deterministic trends

10.14 Economists during the early part of the twentieth century particularly favoured deterministic trend functions and, although simple linear trends were especially popular, perhaps after logarithmically transforming the series being analysed, nonlinear functions such as the logistic and Gompertz curves,

$$\begin{aligned} x_t &= \frac{k}{1+e^{-\lambda t}} \\ x_t &= \exp\left(a_0 - a_1 e^{-\lambda t}\right) \end{aligned}$$

were also employed to model various patterns of growth.

A common impression was that trends were to be removed quickly and simply so that attention could be focused on the more economically interesting cyclical fluctuations thereby revealed. The possibility that the function that was used to remove the trend could actually affect the cyclical component or, even, that the trend and cycle could be theoretically interlinked, only became apparent during the 1930s, particularly after Frickey (1934) showed that a wide range of cyclical patterns could be obtained in economic time series simply by using different functions for trend removal. Using some 23 trend functions, Frickey (1934, table 1) was able to obtain cycles for US pig iron production between 1854 and 1926 ranging from 3.3 to 45 years in average length, leading him to conclude

> first, that the average length of 'cycle' for a series – and for that matter, the whole form of the supposed cyclical picture – may exhibit great variation depending upon the kind of secular trend which has previously been fitted; second, that the discovery, about a particular trend representation which has been set up for a given economic time series, of oscillations which may conform more or less closely to a certain average length cannot in itself be taken as establishing the statistical or economic validity of such movements as cycles. (*ibid.*, pages 16–17)

10.15 Almost no attention was paid to the possibility that the fitted trends might incorporate a measure of uncertainty: as Working and Hotelling (1929, page 73) remarked, such trends were 'frequently discussed as though they represented observed facts, subject to some error in consequence of possible errors in the original data, but, when the basic data are reasonably trustworthy, to be accepted at face value. On this interpretation, trends of economic data must usually have a negligible probable error.' Working and Hotelling were not convinced by this argument and neither were they taken by the use of trends as purely descriptive devices.

> The proper use of a trend is as a representation of the course that would probably have been taken by the data if certain classes of effects appearing in the data could have been eliminated. Viewed thus, a trend may be a highly trustworthy representation of this course, or it may be, *and usually is*, subject to considerable possible error. (*ibid.*, page 73; italics added for emphasis)

Working and Hotelling were at pains to contrast the classical 'measurement error' interpretation of trend fitting in science to the situation typically found with economic data, explaining the contrast in some style.

> The theory of curve fitting is simplest where the effects which it is desired to eliminate are simple errors of observations. The data from observations on

> the course of a comet will ordinarily show the heavenly wanderer following a somewhat erratic course. No sober comet so wavers and stumbles on its path. Deviations of the observations from a smooth curve may properly be attributed to errors of observation. The curve fitted to the observations may be described as representing the probable true course of the comet.
>
> The interpretation of curves fitted to economic data requires a higher degree of abstraction. The trend of potato yields in the United States, for example, as shown in [Figure 10.4], does not represent the true course of potato yields. The original jagged curve is the best available representation of that true course. The trend gives rather a representation of the course that would have been followed had yields been affected by but one class of the factors actually affecting them, namely, those changing uniformly from year to year. If it is not reasonable to suppose that such a class of factors was actually in operation, the trend can carry no real meaning: there is no justification for fitting a trend. (*ibid.*, pages 73–4)

After providing a list of those factors that gave the trend fitted in Figure 10.4 'real meaning', Hotelling and Working proceeded to investigate the probable error of such a fitted trend.[7] Using a variety of considerations, they settled on a linear trend fitted to the means of non-overlapping pairs of successive observations for the years 1910/11 to 1914/15 and 1921/22 to 1927/28. The use of such means was a device to eliminate serial correlation in the residuals from the fitted trend, termed by Hotelling and Working as the 'method of independent groups' and employed because negative first order serial correlation was found

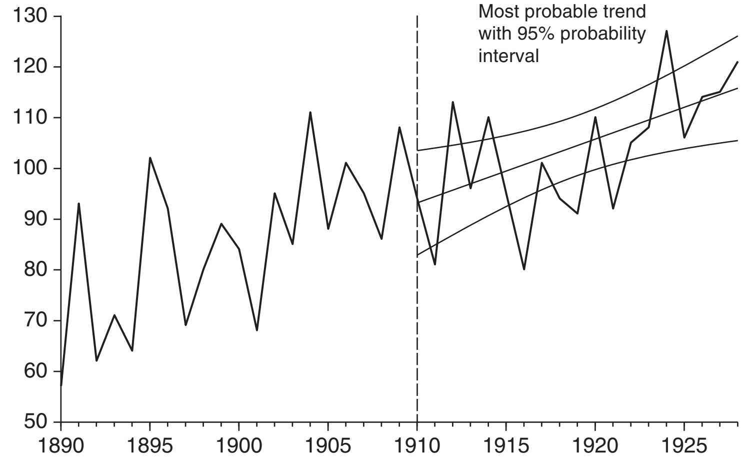


Figure 10.4 Annual average yield of potatoes in the United States, 1890–1928 and recent trend (bushels per acre)

in the residuals from a linear trend fitted to the actual potato yields. The fitted trend has intercept $\hat{\alpha} = 104.4$ with standard error $\hat{\sigma}_\alpha = 2.81$ and slope $\hat{\beta} = 1.26$ with standard error $\hat{\sigma}_\beta = 0.44$.

Working and Hotelling then showed that the standard error of the fitted trend line $\hat{\alpha} + \hat{\beta}t$ was $\sqrt{\hat{\sigma}_\alpha^2 + t^2\hat{\sigma}_\beta^2}$ (where the origin for t has been chosen so that $\bar{t} = 0$). Consequently

$$\hat{\alpha} + \hat{\beta}t \pm K\sqrt{\hat{\sigma}_\alpha^2 + t^2\hat{\sigma}_\beta^2}$$

determines a value of the series that, at time t, differs from that given by the fitted trend line by K times the standard error. The probability that this difference will be exceeded by chance is obtained by equating K with an appropriate value of Student's t-distribution with degrees of freedom given by two less than the number of observations used in the fitting. Here the degrees of freedom are 17 and, for a 5% probability, the appropriate value of the t-distribution is 2.110. Using this value for K obtains the 95% probability interval shown in Figure 10.4.

For the next year, 1929, the best estimate of the 'normal' potato yield is 117 bushels but, given this probability interval, the chances are one in 40 that the true normal yield is below 106 bushels and one in 40 that it is above 128 bushels, leading Working and Hotelling to conclude that 'it is always appropriate to consider the probable errors of a trend and that the probable errors are frequently so large as to deserve careful consideration in drawing conclusions from the data' (*ibid.*, page 84).

Differencing to remove nonstationarity

10.16 While the attention of many statisticians and economists was focused on trend fitting, either directly or via moving averages, to remove nonstationarity in time series data, Irving Fisher, the famous American economist, statistician and polemicist, returned to the idea of taking differences to transform data to stationarity (cf. the discussion of Persons analysis in §**4.14**). While investigating the relationship between US trade and price indices (denoted T and P respectively), Fisher made the observation that, when looking at a plot of the two series,

> (a)t first glance these two curves reveal no evident relationship. But this is chiefly because their relationship is not directly between the height (ordinates) of these curves but between the *slope* of the P curve (*i.e.* the rate of change of the price level) and the *height* of the T curve. (I. Fisher, 1925, page 182: italics in original)

Fisher measured the slope, P', of the P curve for a particular month 'by subtracting the index number for the *preceding* month from that for the *succeeding* month and reducing the result to a percentage of the given or intervening month' (*ibid.*, page 182, footnote 3: italics in original). Thus

$$P'_t = \frac{P_{t+1} - P_{t-1}}{P_t}$$

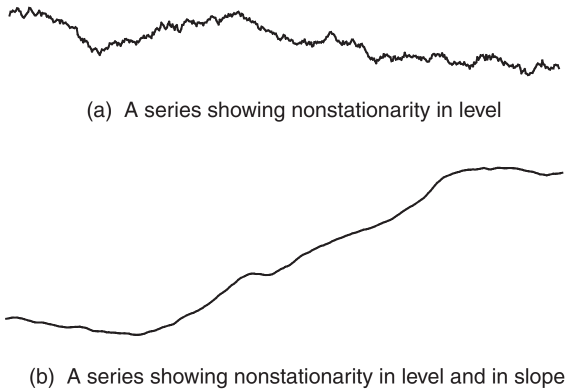
which can be expressed as

$$P'_t = \frac{\Delta P_{t+1}}{P_t} + \frac{\Delta P_t}{P_t} \approx \Delta \log P_t + \frac{P_{t-1}}{P_t} \Delta \log P_{t-1}$$

using the familiar approximation $\log(1+x) \approx x$ for x small. Thus the slope is a weighted average of the current and past growth rates. It is clearly seen in I. Fisher (1925, Chart 1) that the wandering behaviour of P is removed by this transformation, leaving P' stationary: '(w)e note at once that P' supplies an oscillating barometer without requiring any of the corrections for secular trend and seasonal variation found necessary in most "cycle" data' (*ibid.*, page 183).[8]

10.17 The use of differencing to induce stationarity was reconsidered by Box and Jenkins (1962) in the context of developing stochastic models for adaptive optimization and control, where they proposed that a nonstationary series should be differenced enough times until it appeared stationary, although in their experience second differencing had always proved adequate. This was the forerunner of the approach utilized in Box and Jenkins (1968, 1970), where they termed series that could be reduced to stationarity by differencing one or more times as being *homogeneous nonstationary*.[9]

Figure 10.5(a) shows a nonstationary series that is homogeneous in its *level*: except for a vertical translation, one part of the series looks much the same



(a) A series showing nonstationarity in level

(b) A series showing nonstationarity in level and in slope

Figure 10.5 Two kinds of homogeneous nonstationary behaviour

as any other. Such a series can be rendered stationary by differencing once, i.e., by analysing $z_t = \Delta x_t$ rather than x_t. Figure 10.5(b) shows a second type of nonstationarity of fairly common occurrence, where the series has neither a fixed level nor a fixed slope but exhibits homogeneous behaviour if differences in these characteristics are allowed for, i.e., if second differences $z_t = \Delta^2 x_t$ are considered.

Integrated processes

10.18 In general, if *d*th differences are required to render x_t stationary then the series to be analysed is $z_t = \Delta^d x_t$. This can be 'inverted' to give

$$x_t = \Delta^{-d} z_t = S^d z_t$$

where S is the infinite summation operator defined by

$$Sz_t = \sum_{j=-\infty}^{t} z_j = (1 + B + B^2 + \cdots)z_t = (1 - B)^{-1} z_t = \Delta^{-1} z_t$$

The operator $S^2 z_t$ is similarly defined as

$$S^2 z_t = (1 - B)^2 z_t = \Delta^{-2} z_t = Sz_t + Sz_{t-1} + Sz_{t-2} + \cdots = \sum_{i=-\infty}^{t} \sum_{j=-\infty}^{i} z_j$$

and so on. Thus x_t can be obtained by summing (or 'integrating') z_t d times. Recalling the terminology first introduced by Hall (1925) (cf. §**10.8**), x_t is an *integrated process* of order d.

Autoregressive-integrated-moving average processes

10.19 If the stationary *d*th differences $z_t = \Delta^d x_t$ can be represented by an ARMA(p, q) process (cf. §**8.10**, §§**9.30–9.47**),

$$\phi(B) z_t = \theta(B) a_t \tag{10.11}$$

then Box and Jenkins (1970, chapter 4) call the equivalent model for x_t itself,

$$\phi(B) \Delta^d x_t = \theta(B) a_t \tag{10.12}$$

an *autoregressive-integrated-moving average* process of orders p, d and q, succinctly given the acronym ARIMA(p, d, q). Such processes have some important and interesting properties which have led to them becoming perhaps the most widely used class of model for dealing with nonstationary processes.

Recall the simple AR(1) process $(1 - \phi B)x_t = a_t$. If $|\phi| < 1$, x_t is stationary and will therefore always revert back to its mean, here taken to be zero for simplicity. On the other hand, if $\phi > 1$ the process is said to be explosive, with x_t increasing rapidly with t. The important point is that, in both cases, the *local* behaviour of a series generated from the model is heavily dependent upon the *level* of x_t. This is in contrast to the behaviour of the series shown in Figure 10.5(a), where its local behaviour appears to be independent of its level. For an ARMA model to exhibit such behaviour, the autoregressive operator must be chosen such that

$$\phi(B)(x_t + c) = \phi(B)x_t$$

where c is any constant. Thus the autoregressive operator must satisfy $\phi(B)c = 0$, which implies that $\phi(1) = 0$, which will be satisfied if $\phi(B)$ is of the form

$$\phi(B) = \phi_1(B)(1 - B) = \phi_1(B)\Delta$$

Hence the class of processes having the desired property will be of the form

$$\phi_1(B)\Delta x_t = \theta(B)a_t$$

which, of course, is (10.12) with $d = 1$, i.e., an ARIMA$(p - 1, 1, q)$ process. The required homogeneity excludes the possibility that $z_t = \Delta x_t$ should increase explosively. This means that either $\phi_1(B)$ is a stationary autoregressive operator or $\phi_1(B) = \phi_2(B)(1 - B)$, so that $\phi_2(B)z_t = \theta(B)a_t$, where $z_t = \Delta^2 x_t$, which is the case for the series shown in Figure 10.5(b). In the latter case the same argument can be applied to the second difference and so on. Consequently, it must be the case that, for time series that are nonstationary, but nevertheless exhibit homogeneity, the autoregressive operator must be of the form shown in (10.12).

10.20 For the AR(1) process, the requirement that $\phi(1) = 0$ implies that $\phi = 1$, so that the model becomes $x_t = x_{t-1} + a_t$ or, equivalently, $\Delta x_t = a_t$. This, of course, is the famous *random* (or drunkards) *walk*, so termed in a correspondence between Karl Pearson and Lord Rayleigh in the journal *Nature* in 1905 (see Pearson and Rayleigh, 1905). Although first employed by Pearson to describe a mosquito infestation in a forest, the model was subsequently, and memorably, used to describe the optimal search strategy for finding a drunk who had been left in the middle of a field at the dead of night! The solution is to start exactly where the drunk had been placed, as that point is an unbiased estimate of the drunk's future position since he will presumably stagger along in an unpredictable and random fashion: '(t)he lesson of Lord Rayleigh's solution is that in open country the most probable place to find a drunken man who is at all capable of keeping on his feet is somewhere near his starting point' (*ibid.*, page 342).[10] If the random walk starts at time $t=0$ then

$$x_t = x_0 + \sum_{j=1}^{t} a_j$$

so that x_t is the accumulation of all past innovations. The random walk is thus equivalent to Yule's (1926) conjunct series with random differences (§§**5.6–5.7**), to Working's (1934) 'random-difference series' discussed in §§**5.16–5.17**, and to Macaulay's (1931) 'cumulated chance series' referred to in §**10.1**. Macaulay's 'chance series which has been cumulated twice' is thus an integrated series of order two, and may be thought of as a random walk with random walk innovations, since the process $\Delta x_t = b_t$ with $\Delta b_t = a_t$ can be written as $\Delta^2 x_t = a_t$. The two series shown in Figure 10.5 are generated as $\Delta x_t = a_t$ and $\Delta^2 x_t = a_t$, respectively, in both cases with a_t being a standard normal variate.

If a constant is included, the process

$$x_t = x_{t-1} + \theta_0 + a_t \tag{10.13}$$

is known as a *random walk with drift*. Figure 10.6 depicts such a process with a_t standard normal and $\theta_0 = 0.2$. It is often remarked that the evolution of many macroeconomic time series look very much like this.

If the process again starts at $t=0$, the random walk with drift can be written as

$$x_t = x_0 + t\theta_0 + \sum_{j=1}^{t} a_j$$

It therefore follows that the mean of the process will be time varying

$$\mu_t = E(x_t) = x_0 + t\theta_0$$

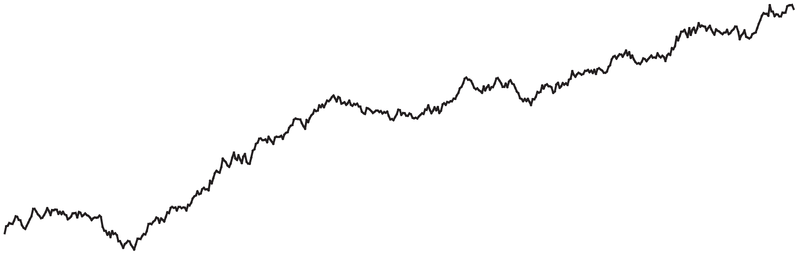


Figure 10.6 A random walk with drift

as will be the variance and all the auto-covariances

$$\gamma_{k,t} = Cov(x_t x_{t-k}) = (t-k)\sigma^2 \quad k > 0$$

where $\sigma^2 = E(a_t^2)$. Thus the autocorrelation between x_t and x_{t-k} is given by

$$\rho_{k,t} = \frac{t-k}{\sqrt{t(t-k)}} = \sqrt{\frac{t-k}{t}}$$

If t is large compared to k, all $\rho_{k,t}$ will be approximately unity. The sequence of x values will therefore be very smooth, but x_t will, of course, be nonstationary since both its mean and variance increase with t.

With a constant included, the ARIMA(p, d, q) process takes the form

$$\phi(B)\Delta^d x_t = \theta_0 + \theta(B)a_t$$

The inclusion of θ_0 has the effect of including a deterministic function of time, a polynomial of order d, into the model, but this will now be 'buried' in nonstationary noise. This should be contrasted with the traditional model of a deterministic trend, in which x_t is expressed as the sum of a polynomial and stationary noise, e.g.

$$x_t = \sum_{j=0}^{d} \beta_j t^j + b_t \quad \phi(B)b_t = \theta(B)a_t$$

This can be written as

$$\Delta^d x_t = \beta_d d! + \Delta^d b_t = \beta_d d! + \Delta^d \frac{\theta(B)}{\phi(B)} a_t$$

or

$$\phi(B)\Delta^d x_t = \phi(1)\beta_d d! + \Delta^d \theta(B)a_t$$

with the stationary nature of the noise in x_t being manifested in d roots of the moving average operator being unity.

Determining the order of differencing

10.21 The autocorrelations of an ARIMA$(p, 0, q)$ process will satisfy the difference equation $\phi(B)\rho_k = 0$ for $k > q$ (see §**8.10**). On factorizing the autoregressive operator as

$$\phi(B) = \prod_{i=1}^{p} (1 - G_i B)$$

then the solution of this difference equation for the kth autocorrelation is, assuming distinct roots,

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \quad k > q - p \tag{10.14}$$

The stationarity requirement that the roots of $\phi(B)$ must lie outside the unit circle thus implies that $|G_i| < 1$, $i = 1, \ldots, p$. From (10.14) it is clear that, in the case of a stationary process in which none of the roots lie close to the boundary of the unit circle, the autocorrelation function will quickly 'die out' for moderate and large k. However, suppose that a single real root, say G_1, approaches unity, so that $G_1 = 1 - \delta$ where δ is small and positive. Then, for k large,

$$\rho_k \approx A_1(1-\delta)^k = A_1(1 - k\delta + k^2\delta^2 - \cdots) \approx A_1(1 - \delta k)$$

and the autocorrelations will not die out quickly but will decline only slowly and approximately linearly. (A similar argument may be applied if more than one of the roots approaches unity.) This led Box and Jenkins (1970, page 175) to the conclusion that

> a tendency for the autocorrelation function not to die out quickly [can be] taken as an indication that a root close to unity may exist. The estimated autocorrelation function tends to follow the behavior of the theoretical autocorrelation function. Therefore, failure of the estimated autocorrelation function to die out rapidly might logically suggest that we should treat the underlying stochastic process as nonstationary in x_t, but possibly as stationary in Δx_t, or in some higher difference.

Box and Jenkins emphasized that the sample autocorrelations need not be high at low lags: all that is required for nonstationarity is that they do not die out rapidly. It may then be assumed that the degree of differencing necessary to achieve stationarity has been reached when the sample autocorrelations of $z_t = \Delta^d x_t$ die out fairly quickly. In practice, Box and Jenkins found that typically $d \leq 2$ and that it was usually sufficient to inspect the first twenty or so sample autocorrelations of the original series and its first and second differences.

Once d has been so chosen, the autoregressive and moving average orders p and q of $z_t = \Delta^d x_t$ can be identified using the procedure outlined in §§**9.41–9.43**.

10.22 Figures 10.7 and 10.8 show Series B and C taken from Box and Jenkins (1970, pages 526 and 528 respectively), along with plots of the sample autocorrelation functions for $d \leq 2$ and $k \leq 20$. It is clear that both series are nonstationary with the autocorrelations for $d = 0$ declining only very slowly. For Series B, which is the IBM common stock price for 369 days during 1961 and 1962, first

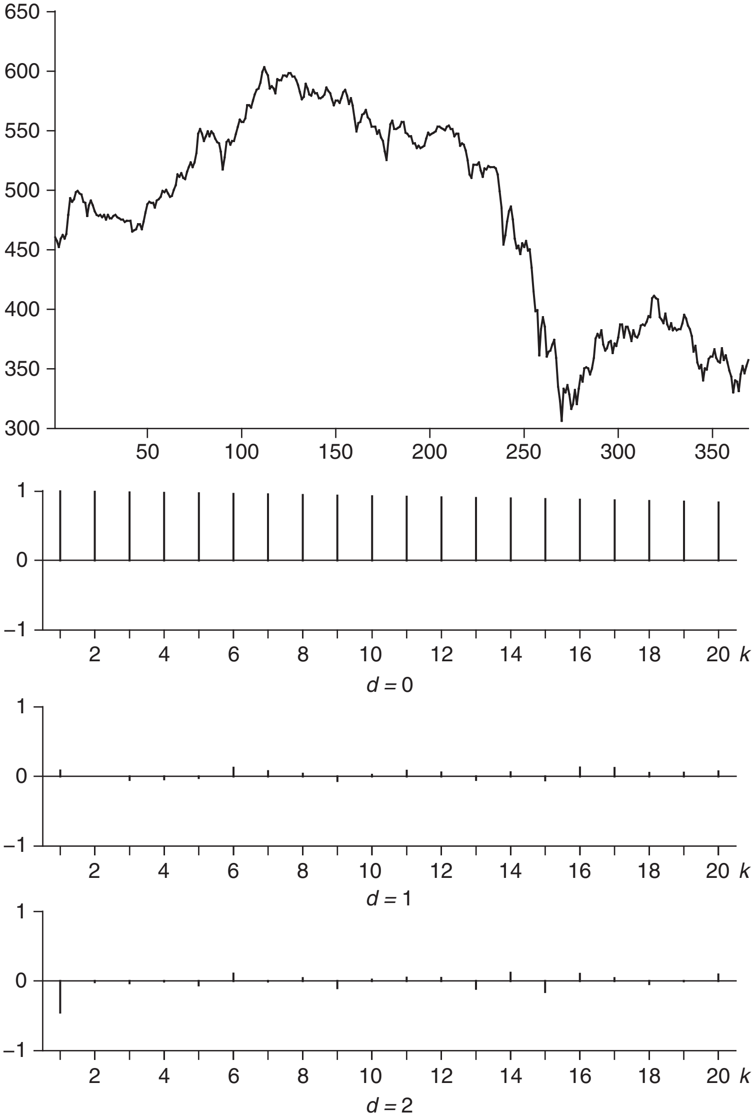


Figure 10.7 Series B from Box and Jenkins (1970); IBM common stock closing prices: daily 17 May 1961–2 November 1962

differencing is seen to induce stationarity: indeed, for $d = 1$ all sample autocorrelations are close to zero, thus implying that Δx_t is white noise or that the series itself follows a random walk, the traditional model used to model stock prices.

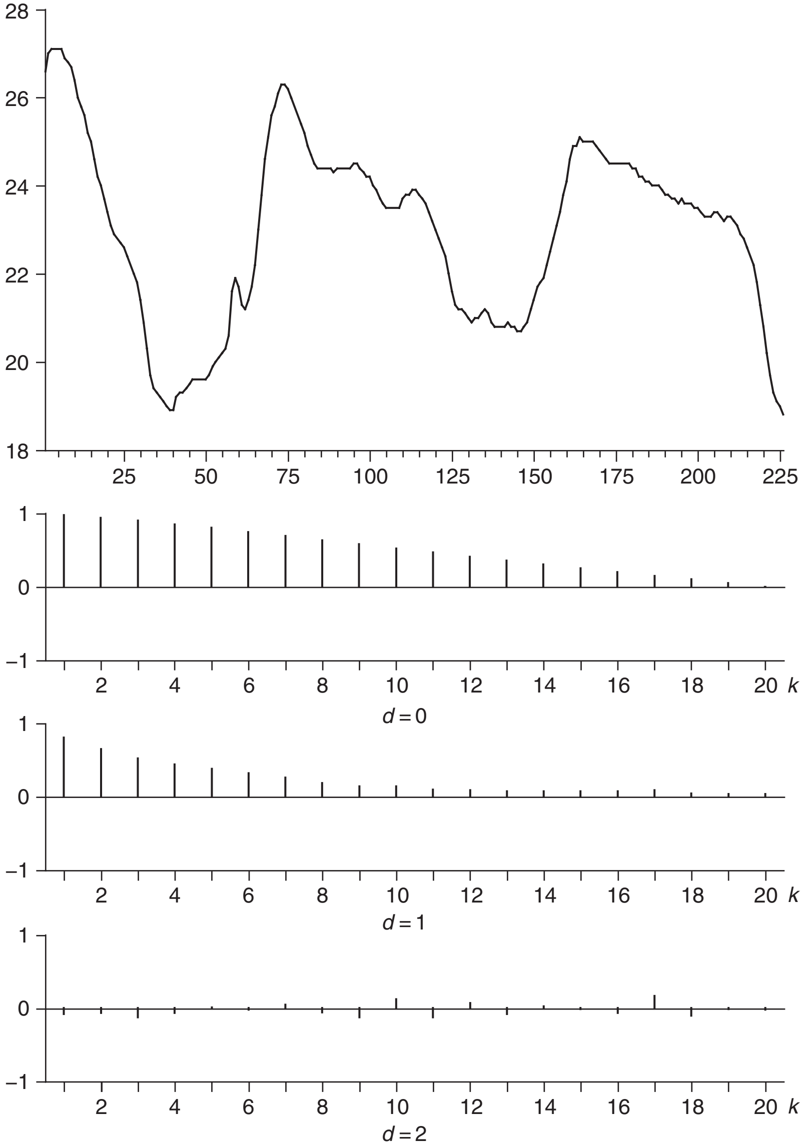


Figure 10.8 Series C from Box and Jenkins (1970): chemical process temperature readings: every minute

For Series C (the 226 minute-by-minute temperature readings of a chemical process), there is some indication that the sample autocorrelations for $d = 1$ are decaying only slowly, which might suggest that second differencing is required. Such a conclusion would be consistent with the changes in level and slope that are observed in the series. If $d = 2$ is chosen, then it would appear from the associated sample autocorrelations that $\Delta^2 x_t$ is white noise. Box and Jenkins were not convinced that second differencing was required, however, for the autocorrelations for $d = 1$ could equally be argued to be declining exponentially from ${}_1r_1 \approx 0.8$, which would identify the ARIMA(1, 1, 0) model $(1 - 0.8B)(1 - B)x_t = a_t$ rather than the ARIMA(0, 2, 0) model $\Delta^2 x_t = a_t$.

This difficulty of deciding the appropriate order of differencing from the behaviour of sample autocorrelations alone was to become a major drawback of the Box and Jenkins identification procedure and, subsequently, led to a massive research project on the subject of testing for unit roots (see §**16.2**). Nevertheless, determining the order of differencing in this way was the final piece in establishing a workable method of identifying ARIMA processes and subsequently had a major impact on getting these models accepted and used across a wide range of time series applications.

11 Forecasting Nonstationary Time Series

Early attempts at economic and financial forecasting

11.1 Forecasting time series, particularly economic ones, has had a long, and often chequered, history. Attempts to find temporal patterns in economic data that might enable predictions to be made about future events stretch all the way back to a London cloth merchant, John Graunt, who in 1662 published several ingenious comparisons using bills of mortality.[1] For example, in an attempt to make trade and government 'more certain and regular', Graunt searched for seasonal and other periodic patterns in mortality, conditioned the data on the plague, and determined the temporal pattern of 'sickliness' that would enable him to predict 'by what spaces, and intervals we may hereafter expect such times again', as quoted in Klein (1997, page 55), who provides an authoritative account of these early attempts at statistical analysis using economic data.

As Klein recounts, attempts to detect periodic fluctuations in economic data continued throughout the eighteenth and nineteenth centuries, and links with meteorology became particularly fashionable, culminating in William Stanley Jevons' advocacy of the sunspot theory of the business cycle, in which he convinced himself that sunspot cycles and business cycles were of the same length, around ten and a half years (cf. the sunspot cycle estimate provided in §**6.4**), and that the causal relationship ran from sunspots to the economy, so that he could predict commercial cycle turning points from sunspot cycle peaks (see Jevons, 1884). Indeed, Jevons even used the evidence of a cyclical peak in corn prices to infer the presence of a preceding peak in the sunspot data. Jevons' sunspot theory was reviewed in Morgan (1990, chapter 1), which emphasized the ridicule, rather than simply criticism, that the theory received from his peers, possibly the first in a long line of critiques of economic forecasting!

The presumed link between meteorology and the economy reached its apogee with Henry Moore's (1923) 'Venus theory', in which he outlined how recent discoveries in physics had shed light on exactly how Venus, in conjunction with the Earth and the Sun, might cause rainfall cycles on Earth, which then led on to produce economic cycles. As Morgan relates, Moore's theories had about as much support from other economists as had those of Jevons, although the reaction was more of polite scepticism than the previous outrage and ridicule.

At the same time as Moore was publishing his ideas about the causes of economic cycles, a different, more mainstream, view about measuring and using business cycles was coming to prominence through the statistical work of Wesley Mitchell (1913). Mitchell's analysis was primarily descriptive, and his belief that each cycle was different effectively excluded formal statistical analysis and, by extension, any role for forecasting. A parallel statistical approach to cycles did, however, purport to provide a methodology for forecasting. This was the business barometer approach of Warren Persons and the Harvard Economic Society, a commercial venture set up by several of the economics staff at Harvard University and which specialized in providing a business forecasting service. This approach rapidly gained popularity during the early 1920s, with institutes devoted to business cycle research being established in a number of countries (see Persons, 1924a, 1924b, and Bullock, Persons and Crum, 1927).

Persons' view of statistics, that probability theory was unsuitable for business cycle analysis and forecasting, was shared by Oskar Morgenstern, who in 1928 published a thesis arguing that economic forecasting based on probability reasoning was impossible because economic data did not satisfy conditions such as homogeneity and independence. Morgenstern's thesis has never been published in English, but a reply to his critique was written by Marget (1929). Although Marget accepted Morgenstern's claim that probability was not applicable to economic forecasting, he nevertheless disputed the conclusion that forecasting was therefore impossible in principle, arguing that most of the forecasting of the period used extrapolative methods rather than formal statistical techniques. Of course, a decade later Wold's treatise showed conclusively that probabilistic concepts could indeed be employed in the forecasting of (stationary) time series (cf. §**7.17**).

11.2 Unsurprisingly, the predictability of financial markets also became of great interest in the run up to, and aftermath of, the 1929 Crash. The early years of the twentieth century saw stock prices in the United States soar: prices rose 60 per cent between 1900 and 1916 and a further *sixfold* between 1921 and 1929. Unsurprisingly, too, this period also saw the appearance of many investment forecasting services, not the least of which was the Harvard A-B-C barometers, originally devised by Persons. These were so termed because groups of variables were combined into three indices: A was the index, or barometer, of the group of leading series, B the index of current indicators, and C the index of lagging series. This approach attempted to identify series, contained in A, that could be used to predict future movements in the stock market. For several years the barometers seemed to be a successful forecasting device, but, after they failed to predict the 1929 Crash, in which stock prices fell 90 per cent from their peak before bottoming out in 1932, they swiftly fell out of favour and eventually into disuse. This episode is entertainingly recounted by Samuelson (1987), who also provides trenchant comments on the supposed forecasting ability of academic economists![2]

In 1928, just before the Crash, Alfred Cowles III, an independent investor, began to question the reliability of the vast array of investment information that was being published and started keeping track records of the most widely circulated services. As the crash and the subsequent bear market unfolded, it seemed to come as a total surprise to the services that Cowles was subscribing to, and this prompted him to set out to find whether stock prices actually were predictable. Cowles got in touch with the Econometric Society, which had been established in 1929 to encourage scholars interested in combining mathematics, statistics and economics. Irving Fisher was the President of the Society, an academic who (recall §**10.16**) had a worldwide reputation for his work in monetary economics, business cycles, and index numbers. By coincidence, he had also achieved a different, and less enviable, reputation for his attempts to forecast the stock market, losing a substantial fortune in the wake of the crash (see the comments in Samuelson, 1987)!

Cowles offered to finance both the publication of the Society's journal and the establishment of an organization to promote and publish econometric research. In 1932, the Cowles Commission for Research in Economics was established in Colorado Springs, from where it moved to Chicago in 1939 and thence to its present home at Yale in 1950, where it is now known as the Cowles Foundation. The first issue of the journal, *Econometrica*, appeared in January 1933 and the July issue of that year contained the first fruit of Cowles' research, an article entitled 'Can Stock Market Forecasters Forecast?', for which there was the three-word abstract – 'It is doubtful'. This article (Cowles, 1933) investigated the track records of 45 professional forecasting agencies, concluding that the results were basically no better than what could have been achieved through a random selection of stocks. Two further articles followed in due course, Cowles and Jones (1937) and Cowles (1944). The latter covered almost 7,000 forecasts over a period of more than fifteen years and again failed to find evidence of any ability to forecast successfully future stock market movements.

Apart from these papers, only two analyses of any note on the subject of stock market forecasting appeared in the quarter of a century following Cowles' original contribution. Holbrook Working's (1934) paper, the statistical aspects of which were discussed in §§**5.19–5.20**, sought to explain why graphs of price levels displayed trends and fluctuations that appeared to show identifiable and repetitive patterns, but that when the levels were differenced to obtain price changes, any such patterns disappeared. This, of course, is arriving at the random walk from an empirical, rather than a theoretical, perspective. The second paper was published some twenty years later by Kendall (1953), who analysed many different weekly financial price series and came to the same conclusion as Working, that there was no structure of any sort in the history of price patterns.

> Broadly speaking the results are these:
>
> (*a*) In series of prices which are observed at fairly close intervals the random changes from one term to the next are so large as to swamp any systematic effect which may be present. The data behave almost like wandering series.
>
> (*b*) It is therefore difficult to distinguish by statistical methods between a genuine wandering series and one wherein the systematic element is weak.
>
> ⋮
>
> (*e*) An analysis of stock-exchange movements revealed little serial correlation within series and little correlation between series. Unless individual stocks behave differently from the average of similar stocks, there is no hope of being able to predict movements on the exchange for a week ahead without extraneous information. (*ibid.*, page 11)

Kendall was clearly surprised by these empirical findings.

> At first sight the implications of these results are disturbing. If the series is homogeneous, it seems that the change in price from one week to the next is practically independent of the change from that week to the week after. This alone is enough to show that it is impossible to predict the price from week to week from the series itself. And if the series really is wandering, any systematic movements such as trends and cycles which may be 'observed' in such series are illusory. The series looks like a 'wandering' one, *almost as if once a week the Demon of Chance drew a random number from a symmetrical population of fixed dispersion and added it to the current price to determine the next week's price.* (*ibid.*, page 13: italics added for emphasis)

Interestingly, Kendall, for all his great knowledge of time series, which has been amply demonstrated in previous chapters, did not appear to be familiar with the term 'random walk'. Even though such a model is clearly implied from the quotes above, he preferred to state that '(i)t may be that the motion is genuinely random and that what looks like a purposive movement over a long period is merely a kind of economic Brownian motion' (*ibid.*, page 18).

11.3 Stimulated by the research of Working and Kendall, Roberts (1959) demonstrated, by way of a simulation using random number tables, why successive price changes should be independent and why analysts could get 'fooled' into believing that the evolution of price levels contained patterns that could be exploited for forecasting. Researching independently and in complete ignorance of these papers, Osborne (1959), an astrophysicist at the US Naval Research Laboratory, hypothesized that the *percentage* change in stock prices would fluctuate as Brownian motion, finding empirical support for this from both Cowles' stock price index and the Dow Jones Industrial Average. A follow-up paper, Osborne (1962), acknowledged the roles played by earlier researchers, including Bachelier.

Alexander (1961) also asked whether stock prices were predictable, but decided to answer this question by comparing a buy-and-hold investment strategy with a 'filter' strategy, where the investor buys after prices have moved up by some predetermined amount, say x per cent, and then sells after they have fallen by x per cent. Examining daily data for the Standard & Poor's Industrial Average from 1857 to 1959 and using various values for x, Alexander found that such a filter strategy did indeed produce greater profits than the buy-and-hold strategy. Unfortunately, however, his procedure drew a great deal of criticism, particularly for ignoring dividends and transaction costs associated with trading stocks, and this led Alexander (1964) to repeat the analysis taking such factors into account. The conclusions from this second attempt were much weaker: 'the big bold profits of Paper 1 must be replaced with rather puny ones. The question still remains whether even these profits could plausibly be the result of a random walk. But I admit that the fun has gone out of it somehow' (Alexander, 1964, page 27).

Tests of the predictability or otherwise of stock market prices or returns now began in earnest and during the 1960s several important studies were published examining some feature or other of the random walk's implications for the predictability of stock market prices: see, inter alia, Cootner (1962), Fama (1965), Fama and Blume (1966), Godfrey, Granger and Morgenstern (1964) and Niederhoffer and Osborne (1966). A variety of methods and approaches to examining the predictability of stock prices were used in these papers, suggesting that a systematic approach to the forecasting of time series might prove to be particularly useful.

Forecasting using ARIMA models

11.4 In fact, concurrently with these attempts to ascertain whether the random walk model could be outperformed in terms of forecasting stock prices, Box and Jenkins (1968, 1970, chapter 5) were indeed providing a synthesis of the theory of forecasting from an ARIMA(p, d, q) model, of which the random walk, of course, is but a special case. Box and Jenkins focused on the general model

$$\varphi(B)x_t = \theta(B)a_t \tag{11.1}$$

where $\varphi(B)=\phi(B)\Delta^d$ is the 'generalized autoregressive operator', to answer the question of how a *future* value, x_{t+l}, $l \geq 1$, could be forecast at the *current* time t. Such a forecast is said to be made at *origin t* for *lead time l*.

An observation x_{t+l} generated by the process (11.1) can be expressed in three equivalent forms. First, it can be written directly as the difference equation

$$x_{t+l} = \varphi_1 x_{t+l-1} + \cdots + \varphi_{p+d} x_{t+l-p-d} - \theta_1 a_{t+l-1} - \cdots - \theta_q a_{t+l-q} + a_{t+l} \qquad (11.2)$$

Second, it can be written as an infinite weighted sum of current and past shocks $a_{t+l}, a_{t+l-1}, \ldots,$

$$x_{t+l} = \sum_{j=-\infty}^{t+l} \psi_{t+l-j} a_j = \sum_{j=0}^{\infty} \psi_j a_{t+l-j} \qquad (11.3)$$

where $\psi_0 = 1$ and the 'ψ-weights' are obtained by equating the coefficients of powers of B in

$$\varphi(B)(1 + \psi_1 B + \psi_2 B^2 + \cdots) = \theta(B)$$

Equivalently, for positive $l > q$, the model may be written in the truncated form

$$x_{t+l} = C_t(l) + a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{t-1} a_{t+1} \qquad (11.4)$$

where

$$C_t(l) = \sum_{j=-\infty}^{t} \psi_{t+l-j} a_j = \sum_{j=0}^{\infty} \psi_{l+j} a_{t-j}$$

has the interpretation of being the 'complementary function'. Finally, x_{t+l} can be written as an infinite weighted sum of previous observations plus a random shock

$$x_{t+l} = \sum_{j=1}^{\infty} \pi_j x_{t+l-j} + a_{t+l} \qquad (11.5)$$

The 'π-weights' may be obtained by equating the coefficients in

$$\varphi(B) = (1 - \pi_1 B - \pi_2 B^2 - \cdots)\theta(B)$$

and, if $d \geq 1$,

$$\overline{x}_{t+l-1}(\pi) = \sum_{j=1}^{\infty} \pi_j x_{t+l-j}$$

will be a weighted moving average, since $\sum_{j=1}^{\infty} \pi_j = 1$.

11.5 Suppose that, at origin t, a forecast $\hat{x}_t(l)$ is to be made of x_{t+l} which is required to be a linear function of current and previous observations

$x_t, x_{t-1}, x_{t-2}, \ldots$. It will then also be a function of the current and previous shocks $a_t, a_{t-1}, a_{t-2}, \ldots$. The best forecast, in the minimum mean square error (MMSE) sense, will be

$$\hat{x}_t(l) = \psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \cdots$$

where the weights $\psi_l^*, \psi_{l+1}^*, \psi_{l+2}^*, \ldots$ minimize the mean square error of the forecast,

$$E[x_{t+l} - \hat{x}_t(l)]^2 = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 + \sum_{j=0}^{\infty} (\psi_{t+j} - \psi_{t+j}^*)^2 \sigma_a^2$$

This expectation will be minimized by setting $\psi_{t+j}^* = \psi_{t+j}$, in which case

$$x_{t+l} = (a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}) + (\psi_l a_t + \psi_{l+1} a_{t-1} + \cdots) = e_t(l) + \hat{x}_t(l) \tag{11.6}$$

where $e_t(l)$ is the error of the forecast $\hat{x}_t(l)$ at lead time l.

On denoting the conditional expectation of x_{t+l}, given knowledge of all the x's up to time t, as (cf. §**9.35**)

$$[x_{t+l}] = E[x_{t+l} | x_t, x_{t-1}, \ldots]$$

then

$$\hat{x}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots = [x_{t+l}] \tag{11.7}$$

The MMSE forecast at origin t, for lead time l, is thus the conditional expectation of x_{t+l} at time t. When $\hat{x}_t(l)$ is regarded as a function of l for fixed t, Box and Jenkins referred to it as the *forecast function* for origin t. Indeed, not only is $\hat{x}_t(l)$ the MMSE forecast of x_{t+l}, but any linear function $\sum_{l=1}^{L} w_l \hat{x}_t(l)$ of the forecasts will be a MMSE forecast of the corresponding linear function $\sum_{l=1}^{L} w_l x_{t+l}$ of the future observations, which is a useful property when, for example, constructing annual forecasts from monthly data.

11.6 The forecast error for lead time l is

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}$$

Since $[e_{t+l}] = 0$ the forecast is unbiased and the variance of the forecast error is

$$V(l) = Var[e_t(l)] = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \tag{11.8}$$

From (11.6), the one-step ahead forecast error is

$$e_t(1) = x_{t+1} - \hat{x}_t(1) = a_{t+1}$$

Hence, the residuals a_t are the *one-step ahead forecast errors*, so that the sequence of such errors must be uncorrelated: 'this is eminently sensible, for if one step ahead errors were correlated, then the forecast error a_{t+1} could, to some extent, be predicted from available forecast errors $a_t, a_{t-1}, a_{t-2}, \ldots$. If the prediction so obtained was $\hat{a}_{t+1}$, then $\hat{x}_t(1) + \hat{a}_{t+1}$ would be a better forecast of x_{t+1} than was $\hat{x}_t(1)$' (Box and Jenkins, 1970, page 129).

However, this result does not extend to higher lead times. Box and Jenkins (*ibid.* Appendix 5.1.1) showed that the correlation between the forecast errors $e_t(l)$ and $e_{t-j}(l)$ made for the *same* lead time l, but at *different* origins t and $t-j$, was given by

$$\rho[e_t(l), e_{t-j}(l)] = \frac{\sum_{i=j}^{l-1} \psi_i \psi_{i-j}}{\sum_{i=0}^{l-1} \psi_i^2}$$

for $0 \leq j < l$ and would be zero for $j \geq l$. Furthermore, the forecast errors $e_t(l)$ and $e_t(l+j)$, i.e., those made for *different* lead times from the *same* origin, will also be correlated: from Box and Jenkins (*ibid.*, Appendix 5.1.2)

$$\rho[e_t(l), e_t(l+j)] = \frac{\sum_{i-0}^{l-1} \psi_i \psi_{j+i}}{\left\{\sum_{h=0}^{l-1} \psi_h^2 \sum_{g=0}^{l+j-1} \psi_g^2\right\}^{\frac{1}{2}}}$$

For example, setting $l=2$ and $j=1$ in these formulae yield

$$\rho[e_t(2), e_{t-1}(2)] = \frac{\psi_1}{(1+\psi_1^2)}$$

and

$$\rho[e_t(2), e_t(3)] = \frac{\psi_2 + \psi_1 \psi_3}{\{(1+\psi_1^2)(1+\psi_1^2+\psi_2^2)\}^{\frac{1}{2}}}$$

'One consequence of this is that there will often be a tendency for the forecast function to be wholly above or below the values of the series when they eventually come to hand' (*ibid.*, page 129).

Alternative forms of the ARIMA forecast

11.7 The forecasts from the ARIMA model (11.1) can be written down in three different ways, corresponding to the three equivalent expressions in **§11.3**.

Taking conditional expectations of the difference equation (11.2) yields

$$[x_{t+l}] = \hat{x}_t(l) = \varphi_1[x_{t+l-1}] + \cdots + \varphi_{p+d}[x_{t+l-p-d}] - \theta_1[a_{t+l-1}] - \cdots - \theta_q[a_{t+l-q}] + [a_{t+l}]$$

while using (11.3) and (11.4), respectively, give

$$\begin{aligned}[x_{t+l}] = \hat{x}_t(l) &= [a_{t+l}] + \psi_1[a_{t+l-1}] + \cdots + \psi_{t-1}[a_{t+1}] + \psi_t[a_t] \\ &\quad + \psi_{t+1}[a_{t-1}] + \cdots + [a_{t+l}]\end{aligned}$$

and

$$[x_{t+l}] = \hat{x}_t(l) = C_t(l) + [a_{t+l}] + \psi_1[a_{t+l-1}] + \cdots + \psi_{t-1}[a_{t+1}]$$

Finally, taking conditional expectations of (11.5) yields

$$[x_{t+l}] = \hat{x}_{t+l} = \sum_{j=1}^{\infty} \pi_j[x_{t+l-j}] + [a_{t+l}] \tag{11.9}$$

Box and Jenkins (*ibid.*, page 130) noted that, although the MMSE forecast was defined in terms of the conditional expectation $[x_{t+l}] = E[x_{t+l}|x_t, x_{t-1}, \ldots]$, which theoretically requires knowledge of the x's stretching back into the infinite past,

> the requirement of invertibility, which we have imposed on the general ARIMA model, ensures that the π weights in [11.9] form a convergent series. Hence, for the computation of a forecast to a given degree of accuracy, for some k, the dependence on x_{t-j} for $j > k$ can be ignored. In practice, the π weights usually decay rather quickly, so that whatever form of the model is employed in the computation, only a moderate length of series $x_t, x_{t-1}, \ldots, x_{t-k}$ is needed to calculate the forecasts to sufficient accuracy.

The conditional expectations can be calculated using the results

$$\begin{aligned}
&[x_{t-j}] = x_{t-j} && j = 0, 1, 2, \ldots \\
&[x_{t+j}] = \hat{x}_t(j) && j = 1, 2, \ldots \\
&[a_{t-j}] = a_{t-j} = x_{t-j} - \hat{x}_{t-j-1}(1) && j = 0, 1, 2, \ldots \\
&[a_{t+j}] = 0 && j = 1, 2, \ldots
\end{aligned}$$

Thus, to obtain the forecast $\hat{x}_t(l)$, the model for x_{t+l} can be written in any one of the above forms, with the terms on the right-hand side of these forms being treated according to the following rules:

The $x_{t-j} (j=0,1,2,\ldots)$, which have already occurred at origin t, are left unchanged.

The $x_{t+j} (j=1,2,\ldots)$, which have yet to occur, are replaced by their forecasts $\hat{x}_t(j)$ at origin t.

The $a_{t-j} (j=0,1,2,\ldots)$, which have occurred, are calculated as $x_{t-j}-\hat{x}_{t-j-1}(1)$.

The $a_{t+j} (j=1,2,\ldots)$, which have yet to occur, are replaced by their expectation of zero.

11.8 As an example of constructing ARIMA forecasts, consider Box and Jenkins' Series C, which in §**10.22** was identified as likely to have been generated by the ARIMA(1, 1, 0) model

$$(1-0.8B)(1-B)x_{t+l} = (1-1.8B+0.8B^2)x_{t+l} = a_{t+l}$$

The difference equation form, which is usually the simplest to work with for computing forecasts, is thus

$$x_{t+l} = 1.8x_{t+l-1} - 0.8x_{t+l-2} + a_{t+l}$$

The forecasts at origin t are then given by

$$\begin{aligned}
\hat{x}_t(1) &= 1.8x_t - 0.8x_{t-1} \\
\hat{x}_t(2) &= 1.8\hat{x}_t(1) - 0.8x_t \\
\hat{x}_t(l) &= 1.8\hat{x}_t(l-1) - 0.8\hat{x}_t(l-2) \quad l=3,4,5,\ldots
\end{aligned}$$

and are readily generated recursively in the order $\hat{x}_t(1), \hat{x}_t(2), \ldots$.

Thus suppose that we wish to forecast Series C from origin $t=20$. The observed values that are required are $x_{19}=23.7$ and $x_{20}=23.4$, using which

$$\begin{aligned}
\hat{x}_{20}(1) &= (1.8\times 23.4) - (0.8\times 23.7) = 23.16 \\
\hat{x}_{20}(2) &= (1.8\times 23.16) - (0.8\times 23.4) = 22.97
\end{aligned}$$

and so on. As soon as x_{21} becomes available, a new set of forecasts $\hat{x}_{21}(1)$, $\hat{x}_{21}(2), \ldots$ can be generated. Since $x_{21}=23.1$, $\hat{x}_{21}(1)=(1.8\times 23.1)-(0.8\times 23.7)=22.86$, etc. Using $a_t = x_t - \hat{x}_t(1)$, the residual $a_{21}=23.1-23.16=-0.06$ may be calculated as soon as x_{21} becomes known.

Calculation of the ψ-weights and the construction of probability limits

11.9 The ψ-weights are obtained by equating the coefficients of powers of B in

$$(1-\varphi_1 B-\cdots-\varphi_{p+d}B^{p+d})(1+\psi_1 B+\psi_2 B^2+\cdots)=(1-\theta_1 B-\cdots-\theta_q B^q)$$

i.e., as

$$\begin{aligned}
&\psi_1=\varphi_1-\theta_1\\
&\psi_2=\varphi_1\psi_1+\varphi_2-\theta_2\\
&\vdots\\
&\psi_j=\varphi_1\psi_{j-1}+\cdots+\varphi_{p+d}\psi_{j-p-d}-\theta_j
\end{aligned}$$

where $\psi_0=1$, $\psi_j=0$ for $j<0$ and $\theta_j=0$ for $j>q$. If K is the greater of the integers $p+d-1$ and q, then for $j>K$ the ψ-weights satisfy the difference equation

$$\psi_j=\varphi_1\psi_{j-1}+\cdots+\varphi_{p+d}\psi_{j-p-d}$$

which enables them to be calculated recursively. Thus, for the model $(1-1.8B+0.8B^2)x_t=a_t$, which is appropriate for Series C,

$$(1-1.8B+0.8B^2)(1+\psi_1 B+\psi_2 B^2+\cdots)=1$$

from which $\psi_0=1$, $\psi_1=1.8$ and $\psi_j=1.8\psi_{j-1}-0.8\psi_{j-2}$, $j=2,3,\ldots$. Hence

$$\begin{aligned}
\psi_2&=(1.8\times 1.8)-(0.8\times 1.0)=2.44\\
\psi_3&=(1.8\times 2.44)-(0.8\times 1.8)=2.95
\end{aligned}$$

and so on.

From (11.7) the forecasts $\hat{x}_{t+1}(l)$ and $\hat{x}_t(l+1)$ of the future observation x_{t+l+1} made at origins $t+1$ and t can be written as

$$\begin{aligned}
\hat{x}_{t+1}(l)&=\psi_l a_{t+1}+\psi_{l+1}a_t+\psi_{l+2}a_{t-1}+\cdots\\
\hat{x}_t(l+1)&=\psi_{l+1}a_t+\psi_{l+2}a_{t-1}+\cdots
\end{aligned}$$

from which it follows that

$$\hat{x}_{t+1}(l)=\hat{x}_t(l+1)+\psi_l a_{t+1}$$

Thus the t-origin forecast of x_{t+l+1} can be updated to become the $(t+1)$-origin forecast of the same x_{t+l+1} by adding a multiple, given by ψ_l, of the one-step ahead forecast error a_{t+1}. For example, when forecasting Series C, once $x_{21}=23.1$ is known, from which $a_{21}=23.1-23.16=-0.06$ has been computed, new forecasts for all lead times may then be calculated as

$$\begin{aligned}\hat{x}_{21}(1) &= 22.97+(1.8\times -0.06)=22.86\\ \hat{x}_{21}(2) &= 22.81+(2.44\times -0.06)=22.67\\ \hat{x}_{21}(3) &= 22.69+(2.95\times -0.06)=22.51\end{aligned}$$

and so on.

11.10 The expression (11.8) shows that the variance of the l-step ahead forecast error for any origin t is given by

$$V(l)=\left(1+\sum_{j=1}^{l-1}\psi_j^2\right)\sigma_a^2$$

Assuming the a's are normally distributed, it then follows that, given information up to time t, the conditional probability distribution of a future value x_{t+l} will be normal with mean $\hat{x}_t(l)$ and standard deviation

$$SE(l)=\left(1+\sum_{j=1}^{l-1}\psi_j^2\right)^{\frac{1}{2}}\sigma_a$$

$(1-\varepsilon)$% probability limits, $x_{t+l}(-)$ and $x_{t+l}(+)$, for x_{t+l} will then be given by $x_{t+l}(\pm)=\hat{x}_t(l)\pm z_{\varepsilon/2}SE(l)$, where $z_{\varepsilon/2}$ is the $\varepsilon/2$ percentage point of the standard normal distribution.

Of course, σ_a is typically unknown and must be estimated along with the θ's and ϕ's using the methods of §§**9.31–9.44**. Such an estimate for Series C is $\hat{\sigma}_a=0.134$ and, since the length of the series, $T=226$, is reasonably large, this value can be substituted into $SE(l)$ to obtain, for example, 50% and 95% limits for $\hat{x}_t(2)$:[3]

$$\begin{aligned}&\text{50\% limits:}\quad \hat{x}_t(2)\pm 0.674\times(1+1.8^2)^{1/2}\times 0.134=\hat{x}_t(2)\pm 0.19\\ &\text{95\% limits:}\quad \hat{x}_t(2)\pm 1.960\times(1+1.8^2)^{1/2}\times 0.134=\hat{x}_t(2)\pm 0.55\end{aligned}$$

The interpretation of the limits $x_{t+l}(-)$ and $x_{t+l}(+)$ should be carefully noted. These limits are such that, *given the information available at origin t*, there is a probability of $1-\varepsilon$ that the actual value x_{t+l}, when it occurs, will be within them.

> It should also be explained that the probabilities quoted apply to *individual* forecasts and not jointly to the forecasts at all the different lead times. For example, it is true with 95% probability, the limits for lead time 10 will include the value x_{t+10} when it occurs. It is not true that the series can be expected to remain within *all* the limits simultaneously at this level of probability. (*ibid.*, page 138: italics in original)

The eventual forecast function and forecast weights

11.11 At time $t+l$ the ARIMA model may be written

$$x_{t+l} - \varphi_1 x_{t+l-1} - \cdots - \varphi_{p+d} x_{t+l-p-d} = a_{t+l} - \theta_1 a_{t+l-1} - \cdots - \theta_q a_{t+l-q} \qquad (11.10)$$

Taking conditional expectations at time t yields, for $l > q$,

$$\hat{x}_t(l) - \varphi_1 \hat{x}_t(l-1) - \cdots - \varphi_{p+d} \hat{x}_t(l-p-d) = 0 \quad l > q$$

where it is understood that $\hat{x}_t(-j) = x_{t-j}$ for $j \geq 0$. This difference equation has the solution

$$\hat{x}_t(l) = b_1^{(t)} f_1(l) + b_2^{(t)} f_2(l) + \cdots + b_{p+d}^{(t)} f_{p+d}(l) \qquad (11.11)$$

for $l > q-p-d$. Box and Jenkins referred to (11.11) as the *eventual forecast function*, whose mathematical form is decided by the general autoregressive operator $\varphi(B)$, which determines whether the functions $f_j(l)$, $j=1,\ldots,p+d$, are polynomials, exponentials, a mixture of sines and cosines, or some combination of these functions. For example, suppose $d=0$ so that $\varphi(B)=\phi(B)$. Using the factorization of §**10.21** and assuming that all the roots G_i, $i=1,\ldots,p$, are distinct, then if G_1, say, is real, $f_1(l)=G_1^l$. If, on the other hand, G_1 and G_2 are a pair of complex roots, then they will contribute a damped sine wave to (11.11). If $\varphi(B)$ has d equal roots of G_0 then this imposes the forms $f_{p+j}(l)=l^{j-1}G_0$, $j=1,\ldots,d$, onto (11.11). If these roots are equal to unity then, since now $f_{p+j}(l)=l^{j-1}$, a polynomial in l of order $d-1$ is introduced into the eventual forecast function.

For a *given origin t*, the coefficients $b_j^{(t)}$ are constants applying for all lead times l, but they change from one origin to the next. It can be shown that the updating equations of these coefficients can be written as (Box and Jenkins, 1970, Appendix A5.3.3)

$$\mathbf{b}^{(t)} = (\mathbf{F}_l^{-1}\mathbf{F}_{l+1})\,\mathbf{b}^{(t-1)} + (\mathbf{F}_l^{-1}\psi_l)a_t \qquad (11.12)$$

where

$$\mathbf{F}_l = \begin{bmatrix} f_1(l) & f_2(l) & \cdots & f_{p+d}(l) \\ f_1(l+1) & f_2(l+1) & \cdots & f_{p+d}(l+1) \\ \vdots & \vdots & & \vdots \\ f_1(l+p+d) & f_2(l+p+d) & \cdots & f_{p+d}(l+p+d) \end{bmatrix}$$

$$\mathbf{b}^{(t)} = \begin{bmatrix} b_1^{(t)} \\ b_2^{(t)} \\ \vdots \\ b_{p+d}^{(t)} \end{bmatrix} \quad \psi_l = \begin{bmatrix} \psi_l \\ \psi_{l+1} \\ \vdots \\ \psi_{l+p+d} \end{bmatrix}$$

While $\varphi(B)$ decides the nature of the eventual forecast function, the moving average operator $\theta(B)$, through the ψ-weights, determines how the function is to be 'fitted' to the data, i.e., how the $b_j^{(t)}$ are to be calculated and updated.

> In general, since only one function of the form [11.11] can pass through $p+d$ points, the eventual forecast function is that unique curve of the form required by $\varphi(B)$, which passes through the $p+d$ 'pivotal' values $\hat{x}_t(q), \hat{x}_t(q-1), \ldots, \hat{x}_t(q-p-d+1)$, where $\hat{x}_t(-j) = x_{t-j}$ $(j = 0, 1, 2, \ldots)$. In the extreme case where $q = 0$, so that the model is of the purely autoregressive form $\varphi(B)x_t = a_t$, the curve passes through the points $x_t, x_{t-1}, \ldots, x_{t-p-d+1}$. Thus, the pivotal values can consist of forecasts or of actual values of the series....
>
> The moving average terms ... help to decide the way in which we 'reach back' into the series to fit the forecast function determined by the autoregressive operator $\varphi(B)$. (*ibid.*, page 140)

11.12 Substituting for the conditional expectations in (11.9) obtains

$$\hat{x}_t(l) = \sum_{j=1}^{\infty} \pi_j \hat{x}_t(l-j) = \pi_1 \hat{x}_t(l-1) + \cdots + \pi_{l-1}\hat{x}_t(1) + \pi_l x_t + \pi_{l+1} x_{t-1} + \cdots$$

on using $\hat{x}_t(l) = x_{t-l}$ for $l \geq 0$. In particular,

$$\hat{x}_t(1) = \pi_1 x_t + \pi_2 x_{t-1} + \cdots$$

and the forecasts for higher lead times may also be expressed directly as linear functions of the observations $x_t, x_{t-1}, \ldots$. For example, the lead-two forecast at origin t is

$$\begin{aligned}\hat{x}_t(2) &= \pi_1\hat{x}_t(1) + \pi_2 x_t + \cdots \\ &= \pi_1\sum_{j=1}^{\infty}\pi_j x_{t-j+1} + \sum_{j=1}^{\infty}\pi_{j+1}x_{t-j+1} \\ &= \sum_{j=1}^{\infty}\pi_j^{(2)}x_{t-j+1}\end{aligned}$$

where

$$\pi_j^{(2)} = \pi_1\pi_j + \pi_{j+1} \quad j = 1, 2, \ldots$$

More general results and alternative methods of computing these weights are given in Box and Jenkins (*ibid.*, page 142 and Appendix 5.2).

Forecasting with some special cases of ARIMA models

11.13 Consider the ARIMA(0, 1, 1) process $\Delta x_t = (1 - \theta B)a_t$, which at time $t + l$ may be written

$$x_{t+l} = x_{t+l-1} + a_{t+l} - \theta a_{t+l-1}$$

Taking conditional expectations at origin t gives

$$\begin{aligned}\hat{x}_t(1) &= x_t - \theta a_t \\ \hat{x}_t(l) &= \hat{x}_t(l-1) \quad l \geq 2\end{aligned}$$

so that, for all lead times, the forecasts at origin t will follow a straight line parallel to the time axis. Using $x_t = \hat{x}_{t-1}(1) + a_t$, it is clear that

$$\hat{x}_t(l) = \hat{x}_{t-1}(l) + \lambda a_t \tag{11.13}$$

where $\lambda = 1 - \theta$.

> This implies that, having seen that our previous forecast $\hat{x}_{t-1}(l)$ falls short of the realized value by a_t, we adjust it by an amount λa_t. ...λ measures the proportion of any given shock a_t, which is permanently absorbed by the 'level' of the process. Therefore it is reasonable to increase the forecast by that part λa_t of a_t, which we expect to be absorbed. (*ibid.*, page 144)

Alternatively,

$$\hat{x}_t(l) = \lambda x_t + (1 - \lambda)\hat{x}_{t-1}(l) \tag{11.14}$$

> This implies that the new forecast is a linear interpolation at argument λ between old forecast and new observation. The form [11.14] makes it clear that if λ is very small, we shall be relying principally on a weighted average of past data and heavily discounting the new observation x_t. By contrast, if $\lambda = 1$, the evidence of past data is completely ignored, $\hat{x}_t(1) = x_t$, and the forecast for all future time is the current value. With $\lambda > 1$, we induce an extrapolation rather than an interpolation between $\hat{x}_{t-1}(l)$ and x_t. The forecast error must now be *magnified* in [11.13] to indicate the change in the forecast. (*ibid.*, pages 144–5: italics in original)

The ψ-weights are obtained from

$$\psi(B) = \frac{1 - \theta B}{1 - B} = 1 + (1 - \theta)B + (1 - \theta)B^2 + \cdots = 1 + \lambda B + \lambda B^2 + \cdots$$

The eventual forecast function is the solution of $(1 - B)\hat{x}_t(l) = 0$. From §**11.11**, $f_1(l) = 1$ and $\hat{x}_t(l) = b_1^{(t)}$ for $l > q - p - d = 0$. For *any fixed origin*, $b_1^{(t)}$ will be a constant and, as has been shown above, the forecasts for all lead times will follow a straight line parallel to the time axis. However, $b_1^{(t)}$ will get updated when a new observation becomes available and the origin advances. From (11.12), the updating equation is

$$b_1^{(t+1)} = b_1^{(t)} + \lambda a_{t+1}$$

The forecast function can therefore be thought of as a polynomial of degree zero in the lead time l, with a coefficient which is adaptive with respect to the origin t.

The π-weights are obtained from

$$(1 - \theta B)\pi(B) = 1 - B$$

as

$$\begin{aligned}\pi(B) = \frac{1 - B}{1 - \theta B} &= \frac{1 - \theta B - (1 - \theta)B}{1 - \theta B} \\ &= 1 - (1 - \theta)(B + \theta B^2 + \theta^2 B^3 + \cdots)\end{aligned}$$

i.e.,

$$\pi_j = (1 - \theta)\theta^{j-1} = \lambda(1 - \lambda)^{j-1}$$

Hence

$$\hat{x}_t(l) = \lambda x_t + \lambda(1 - \lambda)x_{t-1} + \lambda(1 - \lambda)^2 x_{t-2} + \cdots$$

and the forecast for all future values of an ARIMA(0, 1, 1) process is an *exponentially weighted moving average* (EWMA) of all current and past x's.

The variance of the lead l forecast is

$$V(l) = (1 + (l-1)\lambda^2)\sigma_a^2$$

so that the variance increases linearly with l.

11.14 Now consider the ARIMA(0, 2, 2) process $\Delta^2 x_t = (1 - \theta_1 B - \theta_2 B^2)a_t$, which at time $t+l$ may be written

$$x_{t+l} = 2x_{t+l-1} - x_{t+l-2} + a_{t+l} - \theta_1 a_{t+l-1} - \theta_2 a_{t+l-2}$$

On taking conditional expectations at time t

$$\begin{aligned}\hat{x}_t(1) &= 2x_t - x_{t-1} - \theta_1 a_t - \theta_2 a_{t-1}\\ \hat{x}_t(2) &= 2\hat{x}_t(1) - x_t - \theta_2 a_t\\ \hat{x}_t(l) &= 2\hat{x}_t(l-1) - \hat{x}_t(l-2) \quad l \geq 3\end{aligned}$$

from which forecasts are most naturally calculated. These forecasts are seen to follow a straight line passing through the forecasts $\hat{x}_t(1)$ and $\hat{x}_t(2)$. The ψ-weights are calculated from

$$\begin{aligned}\psi(B) &= \frac{1 - \theta_1 B - \theta_2 B^2}{(1-B)^2}\\ &= 1 + (2 - \theta_1)B + (3 - 2\theta_1 - \theta_2)B^2 + \cdots + (1 + \theta_2 + j(1 - \theta_1 - \theta_2))B^j + \cdots\end{aligned}$$

The eventual forecast function is the solution of $(1-B)^2\hat{x}_t(l) = 0$, which from §**11.11** is

$$\hat{x}_t(l) = b_1^{(t)} + b_2^{(t)} l \quad l > 0$$

since $q - p - d = 0$. The forecast function is thus a linear function of the lead time l with coefficients that are adaptive with respect to the origin t. Here

$$\mathbf{F}_l = \mathbf{F}_{l+1} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \psi_l = \begin{bmatrix} 2 - \theta_1 \\ 3 - 2\theta_1 - \theta_2 \end{bmatrix}$$

so that (11.12) yields the following updating equations

$$b_1^{(t+1)} = b_1^{(t)} + b_2^{(t)} + (1 + \theta_2)a_{t+1}$$

$$b_2^{(t+1)} = b_2^{(t)} + (1 - \theta_1 - \theta_2)a_{t+1}$$

The variance of the lead-l forecast is (*ibid.*, page 149)

$$V(l) = \sigma_a^2 \begin{pmatrix} 1 + (l-1)(1+\theta_2)^2 + \frac{1}{6}l(l-1)(2l-1)(1-\theta_1-\theta_2)^2 \\ +l(l-1)(1+\theta_2)(1-\theta_1-\theta_2) \end{pmatrix}$$

which again increases with l, although now in a rather complicated manner.

11.15 A model that has been found to be useful in a variety of applications is the ARIMA(0, 1, 1) process 'with deterministic drift', $\Delta x_t = \theta_0 + (1-\theta_1 B)a_t$. This has the eventual forecast function

$$\hat{x}_t(l) = b_0 + b_1^{(t)} l = (l-1)\theta_0 + \frac{\theta_0}{1-\theta_1} + b_1^{(t)} l \quad l > 0$$

where, as in §**11.13**,

$$b_1^{(t)} = b_1^{(t-1)} + (1-\theta_1)a_t$$

The forecast function thus contains a deterministic slope, or 'drift', due to the term $(l-1)\theta_0$. This forecast function should be compared with that obtained from the ARIMA(0, 2, 2) model, which is also a linear function but with an adaptive intercept. A special case, of course, is the random walk with drift, obtained when $\theta_1 = 0$. In this case the eventual forecast function becomes

$$\hat{x}_t(l) = l\theta_0 + b_1^{(t)} l$$

with

$$b_1^{(t)} = b_1^{(t-1)} + a_t$$

i.e.,

$$\hat{x}_t(l) = l\theta_0 + x_t \quad l > 0$$

In general, if an intercept is included in the ARIMA model then an additional term, $b_0 = \xi \sum_{j=t+1}^{t+l} \psi_{t+l-j}$, where $\xi = \theta_0/(1-\theta_1 - \cdots - \theta_q)$, appears in the eventual forecast function (11.11).

11.16 These examples lead to the following summarization. For an ARIMA$(0, d, q)$ process with drift, the eventual forecast function satisfies $(1-B)^d \hat{x}_t(l) = 0$ and has for its solution a polynomial in l of degree $d-1$:

$$\hat{x}_t(l) = b_0 + b_1^{(t)} + b_2^{(t)} l + \cdots + b_d^{(l)} l^{d-1}$$

which provides forecasts for $l > q-d$. The coefficients $b_1^{(t)}, \ldots, b_d^{(t)}$ are progressively updated as the origin advances. The forecast for origin t makes $q-d$ initial jumps, which depend upon $a_t, a_{t-1}, \ldots, a_{t-q+1}$, before following this polynomial, whose position is uniquely determined by the 'pivotal' values $\hat{x}_t(q), \hat{x}_t(q-1), \ldots, \hat{x}_t(q-d+1)$, where $\hat{x}_t(j) = x_{t-j}$ for $j \le 0$.

Analogous results can be obtained for an ARIMA$(p, d, 0)$ process. Here the eventual forecast function satisfies $\phi(B)(1-B)^d \hat{x}_t(l) = 0$ and has for its solution

$$\hat{x}_t(l) = b_0 + \sum_{j=1}^{p} b_j^{(t)} f_j(l) + \sum_{j=p+1}^{p+d} b_j^{(t)} l^{j-p-1} \tag{11.15}$$

This provides forecasts for all $l > 0$ and passes through the last $p+d$ available values, $x_t, x_{t-1}, \ldots, x_{t-p-d+1}$, these being the pivotal values.

For the mixed ARIMA(p, d, q) process, equation (11.15) holds for $l > q-p-d$ if $q > p+d$ and for $l > 0$ if $q < p+d$. In both cases the forecast function is uniquely determined by the pivotal values $\hat{x}_t(q), \hat{x}_t(q-1), \ldots, \hat{x}_t(q-d+1)$. Thus, for the ARIMA$(1, 1, 1)$ process $(1-\phi B)\Delta x_t = (1-\theta B)a_t$, forecasts are readily obtained from

$$\begin{aligned}\hat{x}_t(1) &= (1+\phi)x_t - \phi x_{t-1} - \theta a_t \\ \hat{x}_t(l) &= (1+\phi)\hat{x}_t(l-1) - \phi\hat{x}_t(l-2) \quad l > 1\end{aligned}$$

Since $q < p+d$, the eventual forecast function for all l is the solution of $(1-\phi B)(1-B)\hat{x}_t(l) = 0$, which is

$$\hat{x}_t(l) = b_1^{(t)} + b_2^{(t)}\phi^l$$

Here

$$\mathbf{F}_l = \begin{bmatrix} 1 & \phi \\ 1 & \phi^2 \end{bmatrix} \quad \mathbf{F}_{l+1} = \begin{bmatrix} 1 & \phi^2 \\ 1 & \phi^2 \end{bmatrix} \quad \psi_l = \begin{bmatrix} \dfrac{1-\theta}{1-\phi} + \dfrac{\theta-\phi}{1-\phi}\phi \\ \dfrac{1-\theta}{1-\phi} + \dfrac{\theta-\phi}{1-\phi}\phi^2 \end{bmatrix}$$

so that the updating equations are

$$\begin{aligned}b_1^{(t)} &= b_1^{(t-1)} + \frac{(1-\theta)}{(1-\phi)}a_t \\ b_2^{(t)} &= b_2^{(t-1)} + \frac{(\theta-\phi)}{(1-\phi)}a_t\end{aligned}$$

Substituting for $\hat{x}_t(1)$ and $\hat{x}_t(2)$ in terms of $b_1^{(t)}$ and $b_2^{(t)}$ obtains

$$\begin{aligned}b_1^{(t)} &= x_t + \frac{\phi}{1-\phi}(x_t - x_{t-1}) - \frac{\theta}{1-\phi}a_t \\ b_2^{(t)} &= \frac{\theta a_t - \phi(x_t - x_{t-1})}{1-\phi}\end{aligned}$$

so that

$$\hat{x}_t(l) = x_t + \phi\frac{(1-\phi^l)}{1-\phi}(x_t - x_{t-1}) - \theta\frac{(1-\phi^l)}{1-\phi}a_t \to b_1^{(t)} \quad \text{as } l \to \infty$$

The ARIMA(1, 1, 0) model used to forecast Series C in §**11.8** has $\phi = 0.8$ and $\theta = 0$, which leads to the eventual forecast function

$$\hat{x}_t(l) = b_1^{(t)} + b_2^{(t)}0.8^l$$

with

$$\begin{aligned} b_1^{(t)} &= b_1^{(t-1)} + 5a_t = x_t + 4(x_t - x_{t-1}) \\ b_2^{(t)} &= b_2^{(t-1)} - 4a_t = -4(x_t - x_{t-1}) \end{aligned}$$

Hence

$$\hat{x}_t(l) = x_t + 4(1 - 0.8^l)(x_t - x_{t-1}) \to b_1^{(t)}$$

Thus l-step ahead forecasts tend to the constant $x_t + 4(x_t - x_{t-1})$. If a constant is included then these forecasts will tend to a straight line with slope given by the constant (see Box and Jenkins, *ibid.*, page 152 and their Figure 5.10).

Exponential and adaptive smoothing for inventory control and sales forecasting

11.17 The forecasting technique known as exponential smoothing originated in the operations research activities of the US Navy during the Second World War. In 1944, Robert G. Brown was given the job of developing a model for tracking enemy submarines; this model was essentially an exponentially weighted moving average applied to continuous data. During the early 1950s, Brown extended the approach to discrete data, developing models that could deal with trends and seasonal patterns. A particular application was in forecasting the demand for spare parts in Navy inventory systems, which was so successful in terms of forecast accuracy and data storage savings that exponential smoothing, as the technique quickly became known, was adopted throughout the Navy's inventory systems (Gardner, 2006). The methodology was formalized, generalized and extended in Brown (1959, 1963), Brown and Meyer (1961) and D'Esopo (1961).

Working independently of Brown, Charles C. Holt, with support from the Logistics Branch of the Office of Naval Research (ONR), developed a similar method for exponential smoothing of trending time series. Originally documented as an ONR memorandum (Holt, 1957), it has recently been republished

(Holt, 2004a) along with a short reflection by the author on the genesis of the method (Holt, 2004b). Holt's ideas gained much wider acceptance with the publication of Winters (1960), which tested the methods on sales data with such success that they became known as the Holt–Winters forecasting system. Further developments were soon made by Muth (1960), Theil and Wage (1964) and Nerlove and Wage (1964), the latter articles coining the term 'adaptive forecasting' for the technique, in view of it being a formulation of the adaptive expectations mechanism used in various economic models, notably for investment and consumption (see Koyck, 1954, and Friedman, 1957, respectively, and Muth, 1960, for a general economic analysis).

11.18 Harrison (1965, 1967) provided the first synthesis of the exponential smoothing methodology, showing that the various types of exponential smoothing were all particular forms of the Box and Jenkins (1962) *polynomial predictor* (see §**11.22** below for further discussion of this predictor):[4]

$$\hat{x}_t(1) = \hat{x}_{t-1}(1) + \sum_{i=0}^{n-1} \eta_i S^i e_t \tag{11.16}$$

where $e_t = e_{t-1}(1) = x_t - \hat{x}_{t-1}(1)$ is the current one-step ahead forecast error and $S^i e_t$ is the ith multiple sum of the past errors:

$$S^0 e_t = e_t; \quad S^1 e_t = \sum_{j=0}^{\infty} e_{t-j}; \quad S^2 e_t = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} e_{t-j-k}; \ \ldots$$

Harrison (1967) showed that (11.16) could equivalently be written in the more convenient form

$$\hat{x}_t(1) = \sum_{i=1}^{n} m_t^{(i)}$$

where

$$m_t^{(i)} = \sum_{j=1}^{n} m_{t-1}^{(j)} + \alpha_i e_t$$

The α_i, $i = 1, \ldots, n$, are the *forecasting*, or *smoothing, parameters*. The first-order predictor ($n = 1$) is thus

$$\hat{x}_t(1) = m_t$$

$$\begin{aligned} m_t &= m_{t-1} + \alpha e_t = \alpha x_t + (1-\alpha) m_{t-1} = \alpha x_t + (1-\alpha)\hat{x}_{t-1}(1) \\ &= \alpha \sum_{i=0}^{\infty} (1-\alpha)^i x_{t-i} \end{aligned}$$

This is therefore the simple EWMA of Holt (1957) and Brown (1959). Since $\hat{x}_t(l)$ will also be equal to m_t, x_t is said to be *locally constant* and the forecasts are *steady*.

The second-order predictor ($n = 2$) is Holt's two-parameter growth model, in which x_t is represented as a *local linear trend*, so that

$$\hat{x}_t(1) = m_t + b_t$$

and, in general, l-step ahead forecasts follow the straight line

$$\hat{x}_t(l) = m_t + lb_t \tag{11.17}$$

Here the *updating* (or *error correction*) equations are

$$\begin{aligned} m_t &= m_{t-1} + b_{t-1} + \alpha_1 e_t \\ b_t &= b_{t-1} + \alpha_2 e_t \end{aligned} \tag{11.18}$$

where $m_t^{(1)} = m_t$ and $m_t^{(2)} = b_t$. The equivalent EWMA (or *recurrence*) equations are

$$\begin{aligned} m_t &= \alpha_1 x_t + (1 - \alpha_1)(m_{t-1} + b_{t-1}) \\ b_t &= \beta_1 (m_t - m_{t-1}) + (1 - \beta_1) b_{t-1} \end{aligned} \tag{11.19}$$

where $\beta_1 = \alpha_2/\alpha_1$. Brown's second-order predictor, otherwise known as *double exponential smoothing*, is a particular form of Holt's, restricting the two parameters α_1 and α_2 so that $\alpha_1 = 1 - \gamma^2$ and $\alpha_2 = (1 - \gamma)^2$: γ has the interpretation of being the rate at which an observation loses its importance each period, i.e., the effect of x_t on the determination of m_t and b_t, and hence on $\hat{x}_t(l)$, is discounted by a factor γ each period.

The predictor of Theil, Nerlove and Wage also restricts the parameters of the local linear trend model such that $\alpha_1 = 2\delta/(1 - \delta)$ and $\alpha_2 = \delta\alpha_1$, implying that $\alpha_2 = \alpha_1^2/(2 + \alpha_1)$. For the general nth-order predictor, Brown's method again restricts the forecasting parameters so that they are all functions of the discount parameter γ. An exponential trend formulation was proposed by Pegels (1969) in which equations (11.17)–(11.19) are replaced by

$$\begin{aligned} \hat{x}_t(l) &= m_t b_t^l \\ m_t &= m_{t-1} b_{t-1} + \alpha_1 e_t = \alpha_1 x_t + (1 - \alpha_1) m_{t-1} b_{t-1} \\ b_t &= b_{t-1} + \alpha_2 e_t / m_{t-1} = \beta_1 m_t / m_{t-1} + (1 - \beta_1) b_{t-1} \end{aligned}$$

although no empirical applications appear to have been made using such an extension.

Optimality of exponential smoothing

11.19 Muth (1960) and Harrison (1967) showed that the simple exponential smoothing predictor was optimal if the observations were generated by the unobserved component model

$$x_t = m_t + u_t$$

$$m_t = m_{t-1} + v_t$$

where m_t may now be thought of as the 'underlying' or 'permanent' level of x_t, and which evolves as a driftless random walk with a mean zero, variance σ_v^2, innovation v_t. u_t is a superimposed random noise with mean zero and variance σ_u^2 which is uncorrelated with v_t. Harrison (1967) referred to this as the *steady* model, as the underlying level m_t only randomly changes from its previous level. These two equations imply that

$$\Delta x_t = v_t + u_t - u_{t-1}$$

so that the autocorrelation function for Δx_t cuts-off at lag one with coefficient

$$\rho_1 = -\frac{\sigma_u^2}{\sigma_v^2 + 2\sigma_u^2}$$

Assuming that the two variances are positive, it is clear that $-0.5 < \rho_1 < 0$, the exact value depending on the relative sizes of the variances, and that Δx_t must therefore be generated by an MA(1) process

$$\Delta x_t = e_t - \theta e_{t-1}$$

for which optimal forecasts are known to be generated by a EWMA (see §**11.13**). It can then be shown that the optimal value of the smoothing parameter is (see Harrison, 1967, section 6.3)

$$\alpha = \frac{(1 + 4\kappa)^{1/2} - 1}{2\kappa} = 1 - \theta \quad \kappa = \sigma_u^2/\sigma_v^2$$

and the minimum variance of e_t is $\sigma_e^2 = \sigma_u^2/(1 - \alpha) = \sigma_u^2/\theta$. Using the result from equation (7.34) that $\rho_1 = -\theta/(1 + \theta^2)$, the condition on ρ_1 restricts the moving average parameter to the range $0 < \theta < 1$ and hence the smoothing parameter to the range $0 < \alpha < 1$, which ensures that $\sigma_e^2 > 0$.

However, a more general result is possible. The unobserved component model formulation can be dispensed with and replaced simply with x_t following an ARIMA(0, 1, 1) process. Such a generalization allows Δx_t to have positive first-order autocorrelation, which is ruled out by the unobserved component formulation. Since $|\theta| < 1$ is the condition for the process to be invertible, this implies that $0 < \alpha < 2$, so simple exponential smoothing will be optimal in this case with the smoothing parameter set at $\alpha = 1 - \theta$.

11.20 The local linear trend model, termed by Harrison (1967) the *linear growth* model, is optimal if the observations are generated by the equations

$$\begin{aligned} x_t &= m_t + u_t \\ m_t &= m_{t-1} + b_t + v_t \\ b_t &= b_{t-1} + w_t \end{aligned}$$

This is seen to extend the steady model by including a second random walk component which imparts a slope into the underlying level m_t, with the slope itself subject to a random change, the innovation w_t being uncorrelated with u_t and v_t. Harrison (1967) showed that the relationship between the smoothing parameters and the variances of the innovations are given by the equations

$$\begin{aligned} \alpha_2^2 \kappa_1 &= 1 - \alpha_1 \\ \alpha_2^2 \kappa_2 &= \alpha_1^2 + \alpha_1 \alpha_2 - 2\alpha_2 \end{aligned}$$

where

$$\kappa_1 = \sigma_u^2 / \sigma_w^2 \quad \kappa_2 = \sigma_v^2 / \sigma_w^2$$

The smoothing parameters must then lie within the region

$$\begin{aligned} &0 < \alpha_2 \le (\alpha_1^2 + \alpha_2^2)^2 / (2 + \alpha_1 + \alpha_2) < 1 \\ &0 < \alpha_1 < 1 \end{aligned}$$

The Theil–Nerlove–Wage version of the linear growth model does not include an innovation to the level, so that $v_t = \sigma_v^2 = \kappa_2 = 0$, while if the innovation to the slope is set to zero the slope is constant and m_t becomes a random walk with drift. If both innovations are excluded m_t becomes a linear trend.

It also follows that the linear growth model is optimal if x_t follows an ARIMA(0, 2, 2) process $\Delta^2 x_t = (1 - \theta_1 B - \theta_2 B^2)\, a_t$ in which $\theta_1 = 2 - \alpha_1 - \alpha_2$ and $\theta_2 = \alpha_1 - 1$. Brown's double exponential smoothing restriction that $\alpha_1 = 1 - \gamma^2$ and $\alpha_2 = (1 - \gamma)^2$ then implies that the ARIMA process has equal moving average roots: $\Delta^2 x_t = (1 - 2\gamma B + \gamma^2 B^2)\, a_t = (1 - \gamma B)^2 a_t$.[5]

Using the expression for the forecast error variance of the ARIMA(0, 2, 2) process given in §**11.14**, Harrison (1967) showed that using Brown's double exponential smoothing rather than Holt's two smoothing parameters procedure would increase the one-step ahead forecast standard error by at most 1.6 per cent as long as $\alpha_1 < 0.25$. Since values of α_1 in this range were typically used in applications of exponential smoothing to sales forecasting, this led him to recommend the former procedure for short-term forecasting of sales.

11.21 The potential robustness of exponential smoothing was demonstrated by Cox (1961), who showed that, for one-step ahead forecasts, EWMAs provided forecasts that were comparable to the optimal predictors from a range of more complicated models. He also showed that, for l-step ahead forecasts, a modified EWMA of the form

$$\hat{x}_t(l) = (1 - \theta + \theta\phi^l)x_t + \phi(1 - \theta^l)\hat{x}_{t-1}(l)$$

closely approximated the optimal predictor for the stationary AR(1) process $x_t = \phi x_{t-1} + a_t$.

Box and Jenkins' polynomial predictor

11.22 Box and Jenkins' (1962) polynomial predictor of equation (11.16) is an interesting and general formulation that can be used for forecasting nonstationary time series. A general form of the predictor is

$$\hat{x}_t(1) = \hat{x}_{t-1}(1) + \sum_{i=-m}^{n-1} \eta_i S^i e_t \tag{11.20}$$

and Box and Jenkins showed that this will be an optimal predictor if the observed series x_t follows an ARIMA$(0, n, m+n)$ process:

> (t)hus we have a result which is of considerable practical value. If, after differencing our series x, which in general will be non-stationary, n times, we could render it stationary and if the population serial covariances of lag greater than some value $m+n$ were then zero, a predictor of the type [11.20] would then be optimal. (Box and Jenkins, *ibid.*, page 313)

Box and Jenkins then pointed out that the EWMA corresponded to setting $m = 0$ and $n = 1$, as in §**11.18**, and that, given the success of this predictor, the simple generalization

$$\hat{x}_t(1) = \hat{x}_{t-1}(1) + \eta_{-1}\Delta e_t + \eta_0 e_t + \eta_1 S e_t \tag{11.21}$$

might be an adequate model for many practical forecasting purposes. They also noted that the form of (11.21) corresponded to the type of model used in automatic control, with the terms in Δe_t, e_t and Se_t corresponding to discrete analogues of derivative, proportional and integral control mechanisms, respectively.

The stochastic process for which (11.21) is optimal is

$$x_t = \mu + a_t + \eta_{-1} a_t + \eta_0 S a_t + \eta_1 S^2 a_t$$

which can be written as the ARIMA(0, 2, 3) process

$$\Delta^2 x_t = a_t + (\eta_1 + \eta_0 + \eta_{-1} - 2)\, a_{t-1} + (1 - 2\eta_{-1} - \eta_0)\, a_{t-2} + \eta_{-1} a_{t-3} \qquad (11.22)$$

If $\eta_{-1} = 0$, so that no difference term is needed, (11.22) reduces to Holt's linear growth model of §**11.19**, while if, as well, $\eta_0 = \eta_1 = 1$, the model reduces to $\Delta^2 x_t = a_t$, in which the second differences of x_t are uncorrelated.

Other issues involved with exponential smoothing

11.23 The whole area of exponential smoothing developed rapidly during the 1960s, with much attention being focused on issues such as parameter selection, choice of starting values for the recurrence relationships used to compute forecasts, the monitoring of such forecasts, and the adaptive control of the smoothing parameters. As none of these issues are germane to the development of time series analysis being undertaken here, the interested reader is referred to the survey by Gardner (1985) for discussion and key references.

Examples of exponential smoothing

11.24 Figure 11.1 shows Series A from Box and Jenkins (1970), who identify it as potentially an ARIMA(0, 1, 1) process. Fitting such a model yields an estimate of the moving average parameter of $\hat{\theta} = 0.7$. Also shown in Figure 11.1 are the one-step ahead EWMA forecasts using $\alpha = 1 - \hat{\theta} = 0.3$

$$\hat{x}_t(1) = 0.3x_t + 0.7\hat{x}_{t-1}(1) = 0.3\sum_{i=0}^{\infty} 0.7^i x_{t-i}$$

Since σ_e^2 is estimated to be 0.101, the implied steady model of §**11.19** has $\sigma_u^2 = 0.071$, $\sigma_v^2 = 0.009$ and $\kappa = 7.8$. Thus the random shocks to the series have almost eight times the variance of the random shocks to the underlying level, and this is seen clearly in the relative smoothness of the forecasts when compared to the observations themselves. Estimating the smoothing parameter directly by minimizing the MSE of the forecasts for alternative values of α also leads to a value of 0.3 being selected.

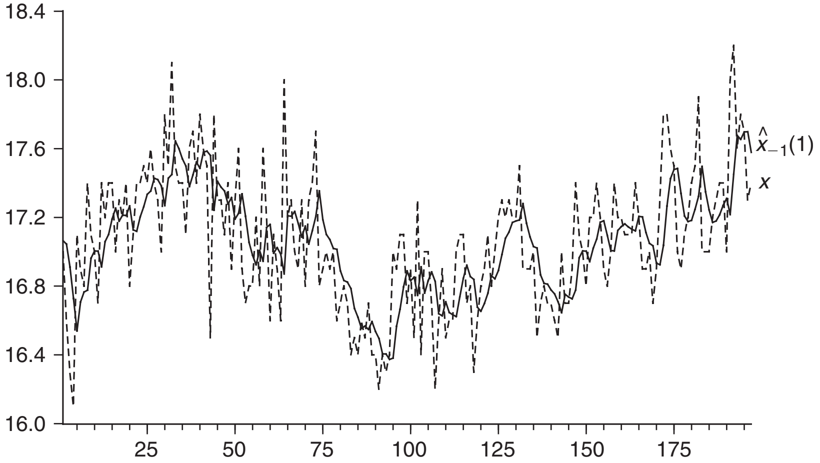


Figure 11.1 Series A from Box and Jenkins (1970) (chemical process concentration readings every two hours) with one-step ahead EWMA forecasts using $\alpha = 0.3$

Fitting an ARIMA(0, 1, 1) process to Series B, analysed in §**10.22**, obtains $\hat{\theta} = 0.08$, which implies that $\alpha = 0.92$, while estimating the smoothing parameter directly obtains $\hat{\alpha} = 0.999$. It is thus clear that this series, daily IBM stock prices, is indeed effectively a random walk, so that the optimal forecast of a future price of the stock is essentially the current price.

There was some indication in §**10.22** that Series C could need second differencing to render it stationary, thus making it a candidate for fitting by either Holt's linear growth model or by Brown's double exponential smoothing. Holt's smoothing parameters are estimated as $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0.87$, the forecasts from which yield a root mean squared error (RMSE) of 0.142. The single parameter of double exponential smoothing is estimated as $\hat{\gamma} = 0.69$ with an accompanying RMSE of 0.315, showing that Holt's linear growth model provides much more accurate forecasts than double exponential smoothing, which should come as no surprise as α_1 is so much greater than 0.25, the value below which Harrison (1967) found little difference between the RMSEs of the two approaches.

Note that these estimates of the smoothing parameters imply that the moving average coefficients of the accompanying ARIMA(0, 2, 2) process are $\theta_1 = 0.13$ and $\theta_2 = 0$. Fitting the ARIMA(0, 2, 3) model implied by the Box and Jenkins polynomial predictor (11.21) yields the coefficient estimates $\hat{\theta}_1 = 0.15$, $\hat{\theta}_2 = 0.13$ and $\hat{\theta}_3 = 0.20$, all accompanied by standard errors of 0.07. The implied polynomial predictor (11.21) is

$$\hat{x}_t(1) = \hat{x}_{t-1}(1) - 0.20\Delta e_t + 0.73e_t + 1.32Se_t \tag{11.23}$$

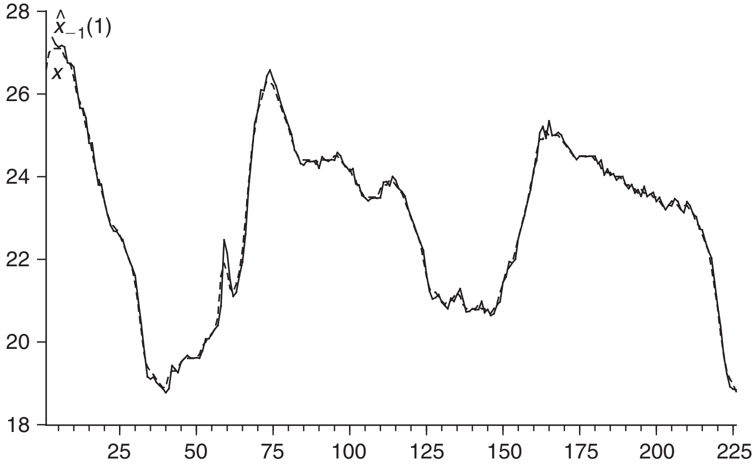


Figure 11.2 Series C and its one-step ahead forecasts from the polynomial predictor (11.20)

which is optimal for the model

$$x_t = 478.5 + a_t - 0.20a_t + 0.73Sa_t + 1.32S^2a_t$$

The RMSE for this model is 0.137, which is smaller than that from Holt's linear growth model and demonstrates the usefulness of the polynomial predictor. Figure 11.2 shows just how accurate the forecasts are: since the sample standard deviation of Series C is 2.059, the polynomial predictor explains in excess of 99.5 per cent of the variation in the series. However, note that the RMSE of the ARIMA(1, 1, 0) model fitted to this series is, from §**11.10**, 0.134, which represents a further marginal improvement. Of course, the eventual forecast functions of the two ARIMA models are very different, as the analysis of §§**11.14–11.16** indicates. While this may not make a great deal of difference when making just one-step ahead forecasts, for longer lead times very different forecasts will be obtained and in these circumstances the selection of the appropriate model becomes paramount.

12
Modelling Dynamic Relationships Between Time Series

Testing the correlation between two time series

12.1 As the theory of testing the significance of autocorrelation coefficients was being developed (see §§**9.1–9.7**), so the related theory of testing the significance of the correlation between two time series was being investigated in tandem. Again, this began with Bartlett's (1935) seminal paper, which derived the result that the variance of the sample correlation, r_{xy}, between two autocorrelated series x_t and y_t, when the true correlation, ρ_{xy}, was zero, could be approximately written as

$$\begin{aligned} V(r_{xy}) &\sim \frac{1}{T^2}(T + 2[(T-1)\rho_x(1)\rho_y(1) + (T-2)\rho_x^2(1)\rho_y^2(1) + \cdots \\ &\quad + \rho_x^{T-1}(1)\rho_y^{T-1}(1)]) \\ &\sim \frac{1}{T}\frac{1+\rho_x(1)\rho_y(1)}{1-\rho_x(1)\rho_y(1)} \qquad (12.1) \end{aligned}$$

where the notation $\rho_x(1)$ and $\rho_y(1)$ is now used for the first-order autocorrelations of x_t and y_t, it being assumed in (12.1) that these series are generated as the AR(1) processes $x_t = \rho_x(1)x_{t-1} + a_{x,t}$ and $y_t = \rho_y(1)y_{t-1} + a_{y,t}$. This result was generalized by Bartlett (1946) and Quenouille (1947b) to the case where x_t and y_t have autocorrelations $\rho_x(i)$ and $\rho_y(i)$, $i = 1, 2, \ldots$:

$$\begin{aligned} V(r_{xy}) &\sim \frac{1}{T^2}(T + 2[(T-1)\rho_x(1)\rho_y(1) + (T-2)\rho_x(2)\rho_y(2) + \cdots \\ &\quad + \rho_x(T-1)\rho_y(T-1)]) \\ &\sim \frac{1}{T}\sum_{i=-\infty}^{\infty}\rho_x(i)\rho_y(i) \qquad (12.2) \end{aligned}$$

Moran (1947) was able to derive the variance of the covariance between x_t and y_t in this set-up and to show that the covariance was asymptotically normally distributed, although his results were too complicated to be used in practice.

12.2 The formulae (12.1) and (12.2) have obvious practical limitations, being based on large sample assumptions and requiring the true but unknown autocorrelations of the two series. An interesting simulation study was thus carried out by Orcutt and James (1948) to investigate the small sample properties of tests of significance based on the above formulae (a similar, but smaller-scale simulation was also reported by G.T. Walker, 1950). Orcutt and James (1948, page 398) were 'anxious [to ensure] that the sampling model used should generate unrelated series which were as analogous as possible to economic time series'. To this end they drew on the findings of Orcutt (1948), who had shown that the 52 series used by Tinbergen (1939) in his business cycle model of the United States might all have been generated by the model

$$x_t = x_{t-1} + 0.3(x_{t-1} - x_{t-2}) + a_t \tag{12.3}$$

which was thus used to generate all the series used in Orcutt and James' simulations. They were clear that (12.3) could not generate stationary series (the model would, of course, later be described as an ARIMA(1,1,0) process), but was

> rather a Brownian type of movement having no true mean. ... On the other hand, the series...are not explosive in the sense that they tend to deviate from any given point or set up oscillations of ever increasing amplitude. ... Since the formulae given earlier for $V(r_{xy})$ were derived on the assumption of stationary autoregressive processes, it is clear that on this account alone it would not be safe without additional evidence to apply them to correlations between non-stationary series such as generated by equation [12.3]. (*ibid.*, page 399)

Orcutt and James focused attention on an alternative approximation to $V(r_{xy})$:

$$V(r_{xy}) \sim \frac{(1 + r_x(1)r_y(1))}{T(1 - r_x(1)r_y(1))} - \frac{2r_x(1)r_y(1)(1 - r_x^T(1)r_y^T(1))}{T^2(1 - r_x(1)r_y(1))^2} \tag{12.4}$$

The sample autocorrelations were estimated, taking $r_x(1)$ for example, by

$$r_x(1) = 1 - \tfrac{1}{2}\delta^2/s^2$$

where

$$\delta^2 = (T-1)^{-1}\sum_{t=1}^{T-1}(x_{t+1}-x_t)^2$$

and

$$s^2 = T^{-1}\sum_{t=1}^{T}(x_t-\bar{x})^2$$

were calculated from the simulated sample of T observations $x_1, x_2, \ldots, x_T$.[1]

After a simulation experiment that was, in its extent, then unprecedented, Orcutt and James concluded that a 'reasonable way of testing the significance on the null hypothesis of a correlation between economic time series is first to estimate $V(r_{xy})$ by means of equation [12.4] and, in so doing, to use the sample values of the first lag autocorrelations of the two series' (*ibid.*, page 409). However, if the estimated value of $V(r_{xy})$ was less than about 0.25, they suggested a simpler alternative. Noting that the variance of the correlation coefficient between two independent and random series of length T' is $1/(T'-1)$, Orcutt and James suggested equating this variance with $V(r_{xy})$ and solving for T'. Rounding off T' to the nearest integer then enables the standard probability distribution of the correlation coefficient to be used. For example, if $r_x(1)=r_y(1)=0.6$ and $T=30$, $V(r_{xy})=0.07$ and $T'=16$. The use of sample autocorrelations in (12.4) was justified from the simulation experiments of Orcutt and James, for they found that the distribution of the correlation coefficient between non-related series depended primarily on the sample autocorrelations of the series and very little, if at all, on the true autocorrelations once given the sample values.

Generally, Orcutt and James found that high correlations between economic time series, at least those generated by processes similar to (12.3), often occurred by chance so that detecting real relationships between such series could be quite difficult. To mitigate this, they suggested making autoregressive transformations of the series involved in such a way that at least one of the series became approximately random. This is best seen by working within a regression framework. Suppose that

$$y_t = \beta x_t + u_t$$

where the error term u_t is generated by

$$u_t = \alpha u_{t-1} + \varepsilon_t$$

ε_t being a random variable. If $\beta=0$ then $y_t=u_t$ and an appropriate autoregressive transformation is

$$y'_t = y_t - \alpha y_{t-1} \quad x'_t = x_t - \alpha x_{t-1}$$

leading to

$$y'_t = \beta x'_t + \varepsilon_t$$

Under $\beta = 0$, $y'_t = \varepsilon_t$ is random so that $\rho_{y'}(1) = 0$, $V(r_{x'y'}) = 1/T$ and the usual test of significance can be applied to $r_{x'y'}$ but now with T rather than $T' << T$ degrees of freedom. When $\beta \neq 0$ this autoregressive transformation offers an estimation technique, known as *Cochrane–Orcutt*, which allows consistent and efficient estimation of β (Cochrane and Orcutt, 1949).

12.3 Quenouille (1949c) proposed using partial correlation coefficients to test for correlation between two autocorrelated series. On the assumption that, as in §**12.1**, x_t and y_t are AR(1) processes, Quenouille recommended using the partial correlation coefficient $r_{xy \cdot x(-1)y(-1)}$, the partial correlation between x_t and y_t with the effects of x_{t-1} and y_{t-1} removed, rather than the simple correlation r_{xy}. Asymptotically such a statistic will have variance T^{-1}, since at least one of the 'residual' series that are, in effect, being correlated will approach independence. Quenouille showed that this will also be the case when the two series are uncorrelated even in small samples and that any bias will be small. Hannan (1955) analysed the properties of this partial correlation as a test of significance, along with those of an alternative partial correlation, that between x_{2t} and y_{2t} when the effects of $y_{2t-1} + y_{2t+1}$, x_{2t-1} and x_{2t+1} have been removed, where here $t = 1, 2, \ldots, \left[\frac{1}{2}(T-1)\right]$. This latter statistic is always asymptotically more efficient than r_{xy} but, although it is an exact test, it is less efficient in most cases than $r_{xy \cdot x(-1)y(-1)}$, the exception being when the first partial correlation of the x_t process is high and positive.[2]

The differing implications of correlating first differences and deviations-from-trends

12.4 The analysis so far has been couched within the framework of correlating stationary time series. As the discussion in Chapter 11 has emphasized, many naturally occurring time series are nonstationary and thus need to be transformed to stationarity before correlation techniques or, equivalently, regression analysis can be used. The two basic methods of detrending time series are to take differences or to use the deviations from a fitted trend. In an extraordinarily prescient article from a modern time series perspective, Bradford Smith (1926) considered the implications for regression analysis of the simplest forms of these alternative detrending methods, that of taking first differences and deviations from a linear trend.[3] The first paragraph of his article stated the problem with great clarity:

> The statistical investigator is not infrequently required to decide between two methods of correlating time series – the deviations-from-trend and the first difference. This choice is important, for the two methods often yield correlation coefficients markedly varying in magnitude. This choice of method is properly made on the basis of the applicability of the implicit assumptions in the two. Certain of these assumptions are in direct contrast. (Smith, 1926, page 55)

Although originally couching the problem in terms of correlations, Smith immediately proceeded to set out the two competing regression models. First, the 'deviation-from-trend' model was developed from the assumption that

> there is a linear relation between deviations from trend in one variable and deviations from trend in the other variable. Or, expressing it for convenience in algebraic form,
>
> $$Y - T_y = b_1(X - T_x) \tag{12.5}$$
>
> wherein $Y$and T_y represent the dependent variable and associated values of its trend and X and T_x similar values for the independent ...
>
> The (linear) trends of the two series can obviously be stated in terms of the values from which they are computed. Thus
>
> $$\begin{aligned} T_y &= b_2 t + K_2 \\ T_x &= b_3 t + K_3 \end{aligned} \tag{12.6}$$
>
> wherein t is that numerical designate of time which has been assigned ..., b_2 and b_3 the annual increments ... and K are the appropriate constants ...
>
> If these expressions for trend are substituted in [12.5] an equation may be formed which shows the type of relationship between the three original measurements, Y, X and t, which are assumed in correlating deviations from trend. Thus
>
> $$\begin{aligned} Y - b_2 t - K_2 &= b_1(X - b_2 t - K_3) \\ Y &= (K_2 - b_1 K_3) + (b_2 - b_1 b_3)\, t + b_1 X \end{aligned} \tag{12.7}$$
>
> Since the terms enclosed in the parentheses are constant, [12.7] may be written,
>
> $$Y = K + b_4 t + b_1 X \tag{12.8}$$
>
> which is recognisable as a multiple regression equation in which t and X are the independents. The term, $b_4 t$, may be interpreted as evidencing the trend in the *relationship between* y *and* X. (*ibid.*, pages 55–6, italics in original)

Next, the first difference model is introduced.

> The basic assumption implicit in correlating first differences is that *changes* in Y over the preceding value are a linear function of similar changes in X. Thus
>
> $$Y - Y_{-1} = K + b_1(X - X_{-1}) \tag{12.9}$$
>
> $$Y = K + Y_{-1} + b_1(X - X_{-1}) \tag{12.10}$$
>
> (*ibid.*, page 56)

Smith then considered the 'differences' between the two approaches, bearing in mind that he was writing before the publication of the seminal articles on the influence of shocks on time series by Yule (1927), Slutzky (1937) and Frisch (1933), as discussed in Chapter 5.

> A consideration of the 'estimating' properties of the two regression equations, [12.8] and [12.10], permits throwing into direct contrast the assumptions of the two methods.
>
> If the constants of formulae [12.8] and [12.10] be determined by usual methods, the right-hand sides of the equations may be evaluated for associated values of Y, the evaluations being designated Y' or estimates of Y. The difference, $Y - Y'$, is, of course, the residual, Z, or error of estimate.
>
> The implicit assumption of the first-difference method, then, is that whatever influences there were causing an error in estimating for any given observation, these influences tend to persist into the ensuing observation. For, by taking the original value of Y_{-1} as a base in computing an estimate of Y, the error of estimating the original Y_{-1}, or $Y_{-1} - Y'_{-1} = Z_{-1}$, is added to the estimate of that Y_{-1}, or Y'_{-1}, the base thus being changed from one of average or normal relationship, Y'_{-1}, to one which has been corrected for the error, $Y_{-1} = Y'_{-1} + Z_{-1}$, and hence supposing the continuation of influences producing that error.
>
> On the other hand, since a normal or trend value is taken as a base for estimating values in the deviation-from-trend method, the assumption is here implicit that whatever influences there were causing errors of estimate in any given observation, those influences are peculiar to that observation alone and do not persist to the next. The errors are automatically dropped out of the computations by making the estimates for succeeding observations from a trend or normal base. (*ibid.*, page 56)

Although errors were not explicitly included in Smith's models, it is nevertheless clear that not only did he understand that they were present, but that he also understood their differing interpretations: hence his statement that, in the first-difference method, 'whatever influences there were causing an error in estimating for any given observation, these influences tend to persist into the ensuing observation' (*ibid.*, page 56) clearly showed that he realized that the shocks in this model were permanent. By contrast, for the deviations-from-trend model, 'those influences are peculiar to that observation alone and do not persist to the next' (*ibid.*, page 56). To emphasize this distinction between the two models, Smith then embarked on a 'counterfactual' analysis.

> This contrast between the assumptions of the two methods can be easily proved by arbitrarily applying to the first-difference method the assumption of the deviation-from-trend method and showing the identity of the two under these circumstances.
>
> Thus, if in the first-difference method the estimates uncorrected for errors were used as a base for computing succeeding estimates, instead of the estimates corrected by the amount of error, then this assumption could be given algebraic expression by substituting Y'_{-1} for Y_{-1} in [12.10]. Thus
>
> $$Y = K + Y'_{-1} + b_1(X - X_{-1}) \tag{12.11}$$
>
> But this calls for the value of Y'_{-1} which can only be determined in terms of Y'_{-2} and so on. If carried back to the first case and terms collected, the formula then becomes
>
> $$Y_n = Y'_0 + nK + b_1(X_n - X_0) \tag{12.12}$$
>
> wherein n is a number designating the place in the series of the given observation, and subscript, 0, designates the initial value from which the first of the differences in the series was secured. The value of Y'_0 is, of course, unobtainable; but it is a constant for all evaluations of the equations, as also is X_0. Hence [12.12] may be written
>
> $$Y_n = K_2 + nK + b_1X_n \tag{12.13}$$
>
> [12.13] is identical in form to [12.8], since n varies with t, thus showing the identity of the first difference and deviation-from-trend methods when the assumptions of the latter are applied to the former. The difference between the two, then, is the difference in using Y' and Y as bases for computing estimates in succeeding years; and this, in turn, is the difference between assuming that the effect of influences causing deviation is peculiar to a

> given observation in one case, and persists unchanged to the succeeding observation in the other case. (*ibid.*, 1926, page 57)

Smith then discussed some practical implications of the first-difference model and emphasized some situations when these might need to be mitigated:

> The implicit assumption in the first-difference method makes that method peculiarly suitable for certain types of analyses – notably those concerned with prices which have a large element of mass psychology in them, it being very common and human, for example, to judge a price by comparing it with the preceding year, forgetting whether or not the preceding year was itself out of line. On the other hand, the assumption is of unreasonable rigidity for various other types of analyses. It may be felt that there is some persistence in influences producing errors from year to year; but, on the other hand, it could seem highly improbable that such influences tend to be exerted in the given years to the full and exact force they were exerted in the preceding. Such, however, is the implicit assumption of the first-difference method. Again, it may seem perfectly feasible that an influence effective in one year may be of a pendulum nature and hence be a reverse influence in the succeeding year. The use of first differences here would aggravate, rather than reduce, the error of estimate. (*ibid.*, 1926, page 57–58)

Smith then suggested that a composite model combining the benefits of the first-difference and deviation-from-trend models could be constructed:

> Evidently, what is needed is a method which will permit the inclusion of such persisting influences but allow them importance only in proportion to the degree to which they persist.
>
> Such a method may be devised as follows:
>
> Equation [12.10] may be written in the form
>
> $$Y = b_1 Y_{-1} + K + b_2 X + b_3 X_{-1} \tag{12.14}$$
>
> wherein b_1 must equal 1.0 and b_2 must equal $-b_3$ if the equation is to be identical with equation [12.10] and, thus, if the assumptions of the first difference method are strictly appropriate.
>
> But if a solution for the best values of b and K in [12.14] is made by methods of least squares, the equivalents of b just cited are no longer predetermined by mathematical necessity as in the case of the more usual solution for b and K by methods illustrated in formula [12.9]. Hence these equivalents will only result in case the assumption as to persistence of errors is appropriate. Formula [12.14] thus typifies a method of correlating time series which has

> all the advantages of the first-difference method without that method's rigid implicit assumption. At the same time it obviously has the advantages of the deviation-from-trend method, since X is one of the independents.
>
> Probably in no single instance of correlating time series can it with strict truth be said either (a) that there is no association between influences producing errors in relationship from one observation to the next or (b) that there is a direct, 'one-to-one' association. But it is equally true that in every instance of correlating time series some element of *both* these hypotheses is present. A correlation of the variables by methods represented in formula [12.14] permits the data, in themselves, to determine the best values of the constants by criteria of least squares.
>
> If by this method b_1 equals 1.0 and b_2 equals $-b_3$ it is evidence that the assumptions of first-differences are appropriate. To the degree that these fail to result the assumptions are inappropriate. If b_1 and b_3 tend to be very small in comparison with b_2 then the assumptions of deviation-from-trend methods are appropriate. Although experience in using the method shows that it is practically never necessary to introduce the term in t shown in formula [12.8], such a term may for completeness be introduced in instances where the value and relationship of the b terms indicate appropriateness of deviation-from-trend assumptions. The formula then becomes
>
> $$Y = K + b_1 Y_{-1} + b_2 X + b_3 X_{-1} + b_4 t$$
>
> The constant values may be found by methods of least squares. It is possible to add terms in X_{-2} and Y_{-2} and so on, in case it is felt that errors are of a cyclical nature, or persist over more than one observation. The measure of the relationship of the independents to Y is, of course, given by the coefficient of multiple correlation. (*ibid.*, 1926, pages 58–9)

The restrictions placed upon [12.14] to obtain the first-difference model are of the 'common factor (COMFAC)' type considered a half-century later by Hendry and Mizon (1978) and whose testing was developed by Sargan (1980) and Mizon and Hendry (1980). It is also clear that Smith was explicitly advocating a 'general-to-specific' modelling strategy (see, for example, Hoover and Perez, 1999), even to the extent of including higher-order lags to model cyclicality and persistence. It is thus extraordinary that this analysis disappeared completely from view, only being rediscovered, or more accurately, reinvented, by econometricians working fifty or so years later.

Fisher's concept of a 'distributed lag'

12.5 As well as examining the differencing and percentage change transformations as means of inducing stationarity (see §**10.16**), Irving Fisher's 1925 paper

'Our unstable dollar and the so-called business cycle' was also important for two other reasons. Concentration is focused here on his introduction of the concept of a *distributed lag*.[4] Recalling that Fisher was looking at the relationship between the percentage change in prices, P', and a trade index, I, he reported that, using monthly data from 1915 to 1923, the

> correlation between P' and I, with a lag of seven months, is 72.7 per cent. This is the maximum correlation; that is, it is greater than that (71.9 per cent) for six months or that (71.5 per cent) for eight months or at any other length of lag. This is a high degree of correlation. But it can be greatly bettered – increased, in fact, to 94.1 – simply by substituting for this *fixed* lag of seven months a *distributed* lag spread over one, two, three, etc., months according to the principles of probability.
>
> So far as I know, this is the first attempt to distribute a statistical lag. (I. Fisher, 1925, page 183; italics in original)

Fisher explained his idea thus.

> The reason for distributing the lag is that the full effect of each P' item is extremely unlikely to be felt at only one instant, such as seven months later, and not felt at any other time either earlier or later than this seven months. ... It is far more probable that the influence began at once, ... and it then gradually increased to a maximum a few months later and therefore tapered off indefinitely according to a probability distribution. (*ibid.*, page 184)

The obvious question is what this probability distribution might be and how it might be chosen: '(i)t would seem that we should select ... that type which most nearly accounts for the behavior of I' (*ibid.*, page 184). Fisher used a rather convoluted gunnery argument to explain the construction of this distribution (in fact a log-normal), which eventually was selected to be that shown in Figure 12.1. The immediate influence of a rise in prices on the trade index is quite small, approximately 3 per cent of the total influence in the first month. This rises to 6 per cent in the second month and reaches a maximum of 7 per cent in the third and fourth months, after which point the 'intensity' gradually diminishes, so that the distribution shown in Figure 12.1 is asymmetric about its mode of $3\frac{1}{2}$ months. The 'shape' of this distribution was selected to provide the closest fit of the 'predicted' trade index to I itself.[5]

Transfer function analysis

12.6 Fisher's distributed lag concept resurfaced in the time series literature during the 1960s in the guise of the *linear transfer function model*, whose

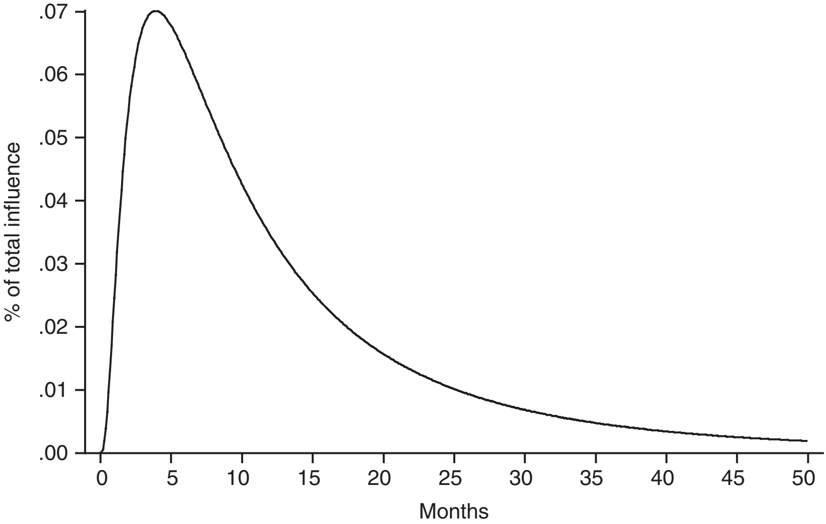


Figure 12.1 Fisher's distributed lag distribution showing the percentage of the total influence of P' on I contributed in each month

development formed chapters 10 and 11 of Box and Jenkins (1970).[6] While Fisher did not formally set out his concept of a distributed lag, the distributed lag/transfer function model was formalized by Box and Jenkins in the following way. Given observations on an 'output' variable Y_t and an 'input' variable X_t, attention is often focused on the value at which the output *eventually* comes to equilibrium when the input is held at a fixed level X. This *steady state* relationship can be denoted $Y_\infty = gX$, where g is the *steady state* gain.

If the level of the input is varied and X_t and Y_t represent *deviations* at time t from equilibrium then the inertia in the system can often be adequately approximated by the *linear filter*

$$\begin{aligned} Y_t &= \upsilon_0 X_t + \upsilon_1 X_{t-1} + \upsilon_2 X_{t-2} + \cdots \\ &= (\upsilon_0 + \upsilon_1 B + \upsilon_2 B^2 + \cdots) X_t \qquad (12.15) \\ &= \upsilon(B) X_t \end{aligned}$$

in which the output deviation at some time t is represented as a linear aggregate of input deviations at times $t, t-1, \ldots$: the operator $\upsilon(B)$ is the *transfer function* of the filter, with the weights $\upsilon_0, \upsilon_1, \upsilon_2, \ldots$ being known as the *impulse response function*.

The *incremental changes* in Y and X are $y_t = \Delta Y_t$ and $x_t = \Delta X_t$, which, on differencing (12.15), are related by $y_t = \upsilon(B)x_t$ and so satisfy the same transfer function model as do Y and X.

It is assumed that the infinite series $\upsilon_0 + \upsilon_1 B + \upsilon_2 B^2 + \cdots$ converges for $|B| \leq 1$ so that the system is *stable*, which implies that a finite incremental change in the input results in a finite incremental change in the output. If X is then held indefinitely at the value $+1$, Y will adjust and maintain itself at the value g. Substituting the values $Y_t = g$ and $1 = X_t = X_{t-1} = X_{t-2} = \cdots$ into (12.15) then obtains

$$g = \sum_{j=0}^{\infty} \upsilon_j$$

so that, for a stable system, the sum of the impulse response weights converges and is equal to the steady state gain of the system.

12.7 The transfer function $\upsilon(B)$ is of infinite extent and thus has limited use for empirically representing such dynamic systems. A parsimonious representation is given by the general linear *difference* equation

$$(1 + \xi_1 \Delta + \cdots + \xi_r \Delta^r) Y_t = g(1 + \eta_1 \Delta + \cdots + \eta_s \Delta^s) X_{t-b} \tag{12.16}$$

known as a transfer function model of order (r, s). This may also be written in terms of $B = 1 - \Delta$ as

$$(1 - \delta_1 B - \cdots - \delta_r B^r) Y_t = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s) X_{t-b} \tag{12.17}$$

or

$$\delta(B) Y_t = \omega(B) X_{t-b} = \omega(B) B^b X_t = \Omega(B) X_t$$

so that the transfer function is $\upsilon(B) = \delta^{-1}(B)\Omega(B)$, a ratio of two polynomials in B. With this representation, an ARIMA model can thus be regarded as a dynamic system having a white noise input for which the transfer function can be expressed as the ratio of two polynomials. The stability of the system requires that the roots of the characteristic equation $\delta(B) = 0$ all lie outside the unit circle. From (12.17), if X_t is held indefinitely at $+1$, Y_t will eventually reach the steady state gain

$$g = \frac{\omega_0 - \omega_1 - \cdots - \omega_s}{1 - \delta_1 - \cdots - \delta_r}$$

Substituting $y_t = \upsilon(B) x_t$ into (12.17) yields the identity

$$(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)(\upsilon_0 + \upsilon_1 B + \upsilon_2 B^2 + \cdots) = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s) B^b$$

On equating coefficients of B, the following relationships are obtained

$$\begin{aligned}
\upsilon_j &= 0 && j < b \\
\upsilon_j &= \delta_1\upsilon_{j-1} + \delta_2\upsilon_{j-2} + \cdots + \delta_r\upsilon_{j-r} + \omega_0 && j = b \\
\upsilon_j &= \delta_1\upsilon_{j-1} + \delta_2\upsilon_{j-2} + \cdots + \delta_r\upsilon_{j-r} - \omega_{j-b} && j = b+1, b+2, \ldots\ldots, b+s \\
\upsilon_j &= \delta_1\upsilon_{j-1} + \delta_2\upsilon_{j-2} + \cdots + \delta_r\upsilon_{j-r} && j > b+s \qquad (12.18)
\end{aligned}$$

The weights $\upsilon_{b+s}, \upsilon_{b+s-1}, \ldots, \upsilon_{b+s-r+1}$ supply r starting values for the difference equation

$$\delta(B)\upsilon_j = 0 \quad j > b+s$$

the solution of which applies to all values υ_j for which $j \geq b+s-r+1$. In general, the impulse response weights consist of

(i) b zero values $\upsilon_0, \upsilon_1, \ldots, \upsilon_{b-1}$;
(ii) a further $s-r+1$ values $\upsilon_b, \upsilon_{b+1}, \ldots, \upsilon_{b+s-r}$ following no fixed pattern (although no such values occur if $s<r$);
(iii) for $j \geq b+s-r+1$, values υ_j that follow the pattern dictated by the rth-order difference equation $\delta(B)\,\upsilon_j = 0$, which has r starting values $\upsilon_{b+s}, \upsilon_{b+s-1}, \ldots, \upsilon_{b+s-r+1}$. Starting values for $j<b$ will be zero.

12.8 The *step response* weights V_j are defined through the identity $\upsilon(B) = (1-B)V(B)$, so that

$$V(B) = V_0 + V_1B + V_2B^2 + \cdots = \upsilon_0 + (\upsilon_0 + \upsilon_1)B + (\upsilon_0 + \upsilon_1 + \upsilon_2)B^2 + \cdots$$

from which it follows that

$$\begin{aligned}
(1 - \delta_1^*B - \delta_2^*B^2 - \cdots - \delta_{r+1}^*B^{r+1})\,(V_0 + V_1B + V_2B^2 + \cdots) \\
= (\omega_0 - \omega_1B - \cdots - \omega_sB^s)B^b
\end{aligned}$$

with

$$(1 - \delta_1^*B - \delta_2^*B^2 - \cdots - \delta_{r+1}^*B^{r+1}) = (1-B)(1 - \delta_1B - \delta_2B^2 - \cdots - \delta_{r+1}B^{r+1})$$

Using (12.18), it follows that the step response function is defined by

(i) b zero values $V_0, V_1, \ldots, V_{b-1}$;
(ii) a further $s-r$ values $V_b, V_{b+1}, \ldots, V_{b+s-r-1}$ following no fixed pattern (no such values occur if $s<r+1$);

(iii) for $j \geq b+s-r$, V_j values that follow the pattern dictated by the $(r+1)$th-order difference equation $\delta^*(B)V_j = 0$ which has $r+1$ starting values $V_{b+s}, V_{b+s-1}, \ldots, V_{b+s-r}$. Starting values for $j < b$ will be zero.

12.9 An example of the representations (12.17) and (12.18) is the transfer function of order $(2,2)$:

$$(1 + \xi_1\Delta + \xi_2\Delta^2)\, Y_t = g(1 + \eta_1\Delta + \eta_2\Delta^2)\, X_{t-b}$$

$$(1 - \delta_1 B - \delta_2 B^2)\, Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)\, X_{t-b}$$

The links between the parameters in these 'Δ' and 'B' forms are

$$\xi_1 = \frac{\delta_1 + 2\delta_2}{1 - \delta_1 - \delta_2} \qquad \xi_2 = \frac{-\delta_2}{1 - \delta_1 - \delta_2}$$

$$\eta_1 = \frac{\omega_1 + 2\omega_2}{1 - \omega_1 - \omega_2} \qquad \eta_2 = \frac{-\omega_2}{\omega_0 - \omega_1 - \omega_2}$$

and

$$\delta_1 = \frac{\xi_1 + 2\xi_2}{1 + \xi_1 + \xi_2} \qquad \delta_2 = \frac{-\xi_2}{1 + \xi_1 + \xi_2}$$

$$\omega_0 = \frac{g(1 + \eta_1 + \eta_2)}{1 + \xi_1 + \xi_2} \qquad \omega_1 = \frac{g(\eta_1 + 2\eta_2)}{1 + \xi_1 + \xi_2} \qquad \omega_2 = \frac{-g\eta_2}{1 + \xi_1 + \xi_2}$$

where

$$g = \frac{\omega_0 - \omega_1 - \omega_2}{1 - \delta_1 - \delta_2}$$

The general behaviour of the transfer function $y_t = v(B)x_t$ may be characterized thus:

Models with $r = 0$. With r and s both equal to zero, the impulse response consists of a single value $v_b = \omega_0 = g$, so that the output is proportional to the input but is displaced by b time periods. More generally, if s is positive, after the displacement the input will be spread over $s+1$ periods in proportion to $v_b = \omega_0, v_{b+1} = -\omega_1, \ldots, v_{b+s} = -\omega_s$. The step response is obtained by summing the impulse response and will eventually satisfy the difference equation $(1-B)\, V_j = 0$ with starting value $V_{b+s} = g = \omega_0 - \omega_1 - \cdots - \omega_s$.

Models with $r = 1$. For $s = 0$, the impulse response tails off geometrically from the initial starting value $v_b = \omega_0 = g/(1+\xi_1) = g(1-\delta_1)$. The step response, on the other hand, increases geometrically to g, being the solution of $(1-\delta_1 B)(1-B)\, V_j = 0$ with starting values $V_b = v_b$ and $V_{b-1} = 0$. For $s = 1$ the initial impulse response $v_b = \omega_0 = g(1+\eta_1)/(1+\xi_1)$ follows no pattern, with the geometric decline induced by the difference equation $\upsilon_j = \delta_1 \upsilon_{j-1}$ beginning with the starting value $\upsilon_{b+1} = \delta_1 \omega_0 - \omega_1 = g(\xi_1 - \eta_1)/(1+\xi_1)^2$. The step response is again determined by the difference equation $(1-\delta_1 B)(1-B)\, V_j = 0$ and again approaches g asymptotically from the starting values $V_b = \upsilon_b$ and $V_{b+1} = \upsilon_b + \upsilon_{b+1}$. With $s=2$ neither υ_b or υ_{b+1} follow a pattern, the geometric decline beginning at υ_{b+2}. Correspondingly, the step response has a single preliminary value $V_b = \upsilon_b$, after which it is again determined by $(1-\delta_1 B)(1-B)\, V_j = 0$ but with starting values V_{b+1} and V_{b+2}.

Models with r = 2. Here the impulse responses eventually satisfy the difference equation

$$\upsilon_j - \delta_1 \upsilon_{j-1} - \delta_2 \upsilon_{j-2} = 0 \quad j > b + s \tag{12.19}$$

the nature of which depends on the roots S_1^{-1} and S_2^{-1} of the associated characteristic equation

$$1 - \delta_1 B - \delta_2 B^2 = (1 - S_1 B)\,(1 - S_2 B) = 0$$

If the roots are real ($\delta_1^2 + 4\delta_2 \geq 0$) the solution to (12.19) is the sum of two exponentials and the system can be thought of as being equivalent to two first-order systems arranged in tandem and having parameters S_1 and S_2. If the roots are complex ($\delta_1^2 + 4\delta_2 < 0$) the solution will follow a damped sine wave.

The weights in the step response function eventually satisfy the difference equation

$$(V_j - g) - \delta_1 (V_{j-1} - g) - \delta_2 (V_{j-2} - g) = 0$$

As this is of the same form as (12.19), the asymptotic behaviour of the step response V_j about its asymptotic value g will parallel the behaviour of the impulse response about zero. If there are complex roots the step response 'overshoots' g and then oscillates about this value until it reaches equilibrium. When the roots are real and positive the step response approaches its asymptote without crossing it. If there are negative real roots, the step response may once again overshoot and then oscillate.

12.10 Box and Jenkins discussed in detail how the discrete dynamic systems developed above may be linked to continuous systems, either directly or as approximations. This analysis will not be discussed here but the interested reader may consult Box and Jenkins (1970, chapter 10.1.2, 10.3, A10.1) for details.

Empirical identification of transfer function models

12.11 In practice, the output Y would not be expected to follow *exactly* the pattern determined by the transfer function model since disturbances of various kinds other than X will normally 'corrupt' the system. Box and Jenkins therefore assumed that all such disturbances are captured by a *noise*, N_t, which is independent of the level of X and additive with respect to the influence of X. Hence the transfer function with added noise model may be specified as

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t \tag{12.20}$$

Representing the noise as the ARMA (p, q) process

$$N_t = \varphi^{-1}(B)\theta(B)a_t$$

leads to the representation

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \varphi^{-1}(B)\theta(B)a_t$$

the actual form of which may then be identified, fitted and checked using an extension of the three stage procedure for individual series discussed in §§**9.41–9.48**.

12.12 The procedure begins by assuming that there are T simultaneous pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_T, Y_T)$ available and uses the cross-covariance and cross-correlation functions (cf. §**4.12**, **8.6**). It is assumed that, if X_t and Y_t are individually nonstationary, then they may be transformed to stationarity by differencing. If the order of differencing is assumed, for simplicity, to be the same in both cases, then the cross-covariance between $x_t = \Delta^d X_t$ and $y_t = \Delta^d Y_t$ at lag k is defined as

$$\gamma_{xy}(k) = E[(x_t - \mu_x)(y_{t+k} - \mu_y)] \quad k = 0, \pm 1, \pm 2, \pm \cdots$$

from which the *cross-correlation function* may be defined as

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \quad k = 0, \pm 1, \pm 2, \pm \cdots$$

As usual, μ_x, μ_y, σ_x and σ_y are the means and standard deviations, respectively, of x and y and it should be noted that $\gamma_{xy}(k) = \gamma_{yx}(-k) \neq \gamma_{xy}(-k)$ and $\rho_{xy}(k) = \rho_{yx}(-k) \neq \rho_{xy}(-k)$, so that the cross-covariance and cross-correlation functions are not symmetric about $k = 0$.

These functions may be estimated from the $\tau = T - d$ pairs of values $(x_1, y_1), (x_2, y_2), \ldots, (x_\tau, y_\tau)$ available for analysis. Thus the sample cross-covariance at lag k is

$$c_{xy}(k) = \begin{cases} \tau^{-1} \sum_{t=1}^{\tau-k} (x_t - \bar{x})(y_{t+k} - \bar{y}) & k = 0, 1, 2, \ldots \\ \tau^{-1} \sum_{t=1}^{\tau-k} (x_{t-k} - \bar{x})(y_t - \bar{y}) & k = 0, -1, -2, \ldots \end{cases}$$

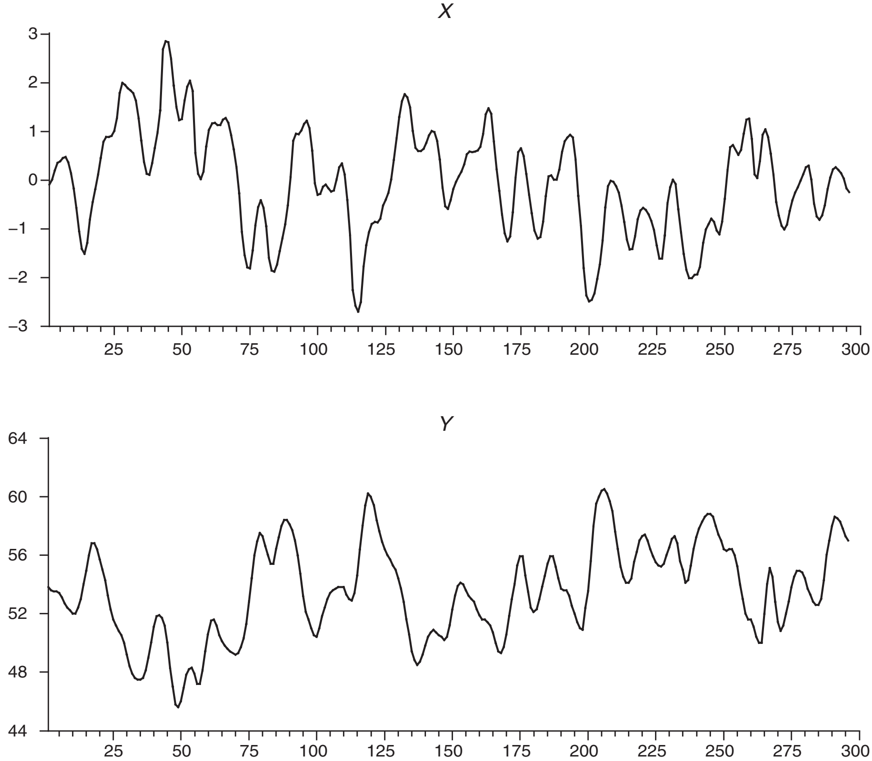


Figure 12.2 Series J from Box and Jenkins (1970): X is the input gas feed rate into a furnace; Y is the percentage output CO_2 concentration

from which the sample cross-correlation at lag k is defined as

$$r_{xy}(k) = \frac{c_{xy}(k)}{s_x s_y} \quad k = 0, \pm 1, \pm 2, \pm \cdots$$

Here $\bar{x}$ and $\bar{y}$ are the sample means and $s_x = \sqrt{c_{xx}(0)}$ and $s_y = \sqrt{c_{yy}(0)}$ are the sample standard deviations of x and y.

12.13 Figure 12.2 shows the pair of observations denoted Series J in Box and Jenkins (1970). X_t is the input gas feed rate into a gas furnace and Y_t is the output CO_2 concentration rate, observed at a nine-second sampling interval, with $T = 296$. Both series are clearly stationary and hence no differencing is required prior to cross-correlation analysis, i.e., d is set equal to zero. Figure 12.3 shows the cross-correlation function $r_{XY}(k)$, which is not symmetrical about $k = 0$ and has a well-defined peak at $k = +5$, indicating that the output lags behind the input, as one might expect. The cross-correlations are negative, which is also to

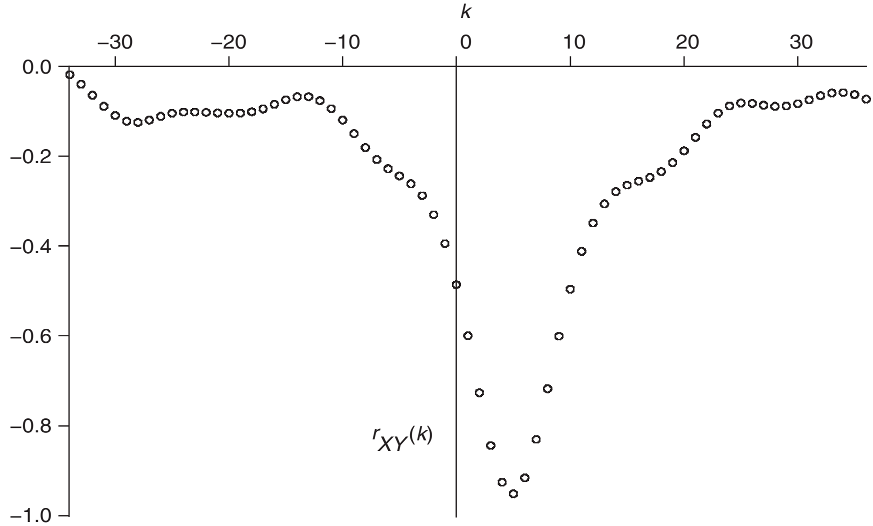


Figure 12.3 Cross-correlation function between *X* and *Y* of Figure 12.2

be expected since an increase in the input *X* produces a decrease in the output *Y*, as can be seen from Figure 12.2.

12.14 Box and Jenkins used the following formula, originally obtained by Bartlett (1955) as an extension of the univariate formulae of §§**9.2–9.5**, to obtain standard errors to attach to cross-correlations:

$$\begin{aligned} V[r_{xy}(k)] \approx\ & (\tau - k)^{-1} \sum_{v=-\infty}^{+\infty} \rho_{xx}(v)\rho_{yy}(v) + \rho_{xy}(k+v)\rho_{xy}(k-v) \\ & + \rho_{xy}^2(k)\left\{\rho_{xy}^2(v) + \frac{1}{2}\rho_{xx}^2(v) + \frac{1}{2}\rho_{yy}^2(v)\right\} \\ & - 2\rho_{xy}(k)\{\rho_{xx}(v)\rho_{xy}(v+k) + \rho_{xy}(-v)\rho_{yy}(v+k)\} \end{aligned} \tag{12.21}$$

Here $\rho_{xx}(v)$ and $\rho_{yy}(v)$ are the individual autocorrelation functions of x_t and y_t themselves and replacing each correlation with their sample counterpart will provide, on taking the square root of (12.21), an approximate standard error for a sample cross-correlation.

There are some interesting special cases of (12.21) that can be very useful in practical applications. For example, on the null hypothesis that x_t and y_t have *no cross-correlation*, (12.21) simplifies to

$$V[r_{xy}(k)] \approx (\tau - k)^{-1} \sum_{v=-\infty}^{\infty} \rho_{xx}(v)\rho_{yy}(v)$$

If, in addition, one of the series is white noise, say $y_t = a_t$, this simplifies further to

$$V[r_{xy}(k)] \approx (\tau - k)^{-1}$$

In such circumstances, it can then be shown that the cross-correlation function will vary about zero with standard deviation $(\tau - k)^{-1/2}$ in a systematic pattern given by the autocorrelation function of x_t.

12.15 The procedure for identifying a transfer function model of the form (12.20) consists of the following steps:

(i) deriving rough estimates $\hat{\upsilon}_j$ of the impulse response weights;
(ii) using these estimates to make guesses of the orders r and s of the polynomials $\delta(B)$ and $\omega(B)$ and the delay parameter b;
(iii) substituting the estimates $\hat{\upsilon}_j$ into equations (12.18) to obtain initial estimates of the parameters in $\delta(B)$ and $\omega(B)$.

The properties of the υ_j implied by (12.18) and outlined in §**12.9** can be used to guess the values of b, r and s, while the appropriate order of differencing for the individual series may be identified by the standard methods of §§**10.21–10.22**. Given this value of d, the model can be written as

$$y_t = \upsilon(B)x_t + n_t \tag{12.22}$$

where $n_t = \Delta^d N_t$.

Box and Jenkins argued that the identification procedure is considerably simplified if the input series is white noise or, if x_t follows an ARMA process, if it is 'prewhitened', i.e., if it is transformed using the ARMA process to the white noise

$$\alpha_t = \phi_x(B)\theta_x^{-1}(B)x_t$$

which will also supply an estimate s_x^2 of σ_x^2 (recall from §**12.2** that transforming to white noise was also advocated some two decades earlier by Orcutt and James, 1948, in a static regression setting). If the same transformation is applied to y_t to obtain

$$\beta_t = \phi_x(B)\theta_x^{-1}(B)y_t$$

then (12.22) may be written

$$\beta_t = \upsilon(B)\alpha_t + \varepsilon_t \tag{12.23}$$

where $\varepsilon_t = \phi_x(B)\theta_x^{-1}(B)n_t$ is the transformed noise. Multiplying (12.23) through by α_{t-k} and taking expectations yields

$$\upsilon_k = \frac{\gamma_{\alpha\beta}(k)}{\sigma_\alpha^2} = \frac{\rho_{\alpha\beta}(k)\sigma_\beta}{\sigma_\alpha}$$

so that, after prewhitening the input, the cross-correlation function between the prewhitened input and the correspondingly transformed output is directly proportional to the impulse response function. The preliminary estimates $\hat{\upsilon}_k$ will then be given by

$$\hat{\upsilon}_k = \frac{r_{\alpha\beta} s_\beta}{s_\alpha}$$

12.16 To identify a transfer function model for the gas furnace data of Figure 12.2, an ARMA model for the input X_t was first obtained, this being the AR(3) process

$$(1 - 1.98B + 1.37B^2 - 0.34B^3) X_t = \alpha_t \quad s_\alpha^2 = 0.0360$$

On defining $\beta_t = (1 - 1.98B + 1.37B^2 - 0.34B^3) Y_t$, and with $s_\alpha = 0.190$ and $s_\beta = 0.360$, the estimated cross-correlation function between α_t and β_t is shown in Figure 12.4, along with two standard error bounds of $2T^{-1/2} = 0.12$, which are appropriate if the series are uncorrelated. The impulse responses are then preliminarily estimated as

k	0	1	2	3	4	5	6	7	8	9	10
$\hat{\upsilon}_k$	0.00	0.11	0.04	0.54	0.63	0.86	0.49	0.29	0.01	0.10	0.07

The values $\hat{\upsilon}_0$, $\hat{\upsilon}_1$ and $\hat{\upsilon}_2$ are all small compared with their standard errors (approximately 0.11), suggesting that $b = 3$. Using the results of §**12.7**, the subsequent pattern of the $\hat{\upsilon}$'s might be accounted for by a model with (r, s, b) either equal to $(1, 2, 3)$ or $(2, 2, 3)$. The former model implies that υ_3 and υ_4 are preliminary values following no fixed pattern and that υ_5 provides the starting value

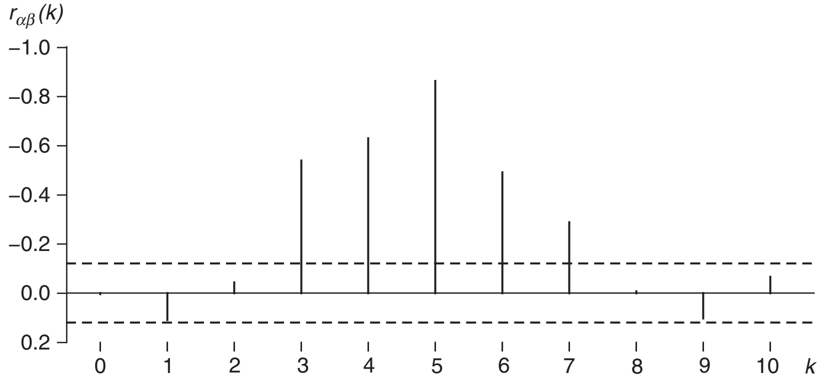


Figure 12.4 Estimated cross-correlation function for the gas furnace data

for a geometric decay determined by the difference equation $\upsilon_j - \delta\upsilon_{j-1} = 0, j > 5$. The latter model implies that υ_3 is a single preliminary value and that υ_4 and υ_5 provide the starting values for a pattern of double exponential decay determined by the difference equation $\upsilon_j - \delta_1\upsilon_{j-1} - \delta_2\upsilon_{j-2} = 0, j > 5$. This preliminary identification suggests the transfer function model

$$(1 - \delta_1 B - \delta_2 B^2)Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)X_{t-3} \tag{12.24}$$

or some simplification of it. Assuming this model, the equations (12.18) become

$$\begin{aligned}
\upsilon_j &= 0 \quad j < 3 \\
\upsilon_3 &= \omega_0 \\
\upsilon_4 &= \delta_1\upsilon_3 - \omega_1 \\
\upsilon_5 &= \delta_1\upsilon_4 + \delta_2\upsilon_3 - \omega_2 \\
\upsilon_6 &= \delta_1\upsilon_5 + \delta_2\upsilon_4 \\
\upsilon_7 &= \delta_1\upsilon_6 + \delta_2\upsilon_5
\end{aligned}$$

Substituting the estimates $\hat{\upsilon}_k$ into the last two of these equations yield

$$\begin{aligned}
-0.86\hat{\delta}_1 - 0.63\hat{\delta}_2 &= -0.49 \\
-0.49\hat{\delta}_1 - 0.86\hat{\delta}_2 &= -0.29
\end{aligned}$$

which give $\hat{\delta}_1 = 0.55$ and $\hat{\delta}_2 = 0.02$. Substituting these values into the second, third and fourth equations yields

$$\begin{aligned}
\hat{\omega}_0 &= \hat{\upsilon}_3 = -0.54 \\
\hat{\omega}_1 &= \hat{\delta}_1\hat{\upsilon}_3 - \hat{\upsilon}_4 = (0.55)(-0.54) + 0.63 = 0.33 \\
\hat{\omega}_2 &= \hat{\delta}_1\hat{\upsilon}_4 + \hat{\delta}_2\hat{\upsilon}_3 - \hat{\upsilon}_5 = (0.55)(-0.63) + (0.02)(-0.54) + 0.86 = 0.50
\end{aligned}$$

Hence preliminary identification leads to the transfer function model

$$(1 - 0.55B - 0.02B^2)\,Y_t = -(0.54 + 0.33B + 0.50B^2)\,X_{t-3}$$

Note that $\hat{\delta}_2$ is very small, suggesting that this parameter may be omitted from the model.

12.17 In general, given an estimate of the transfer function $\hat{v}(B)$, an estimate of the noise series is provided, from (12.22), by

$$\begin{aligned}\hat{n}_t &= y_t - \hat{v}(B)x_t = y_t - \hat{\delta}^{-1}(B)\hat{\omega}(B)x_{t-b}\\ &= y_t + \hat{\delta}_1(\hat{n}_{t-1} - y_{t-1}) + \hat{\delta}_2(\hat{n}_{t-2} - y_{t-2}) + \cdots + \hat{\delta}_r(\hat{n}_{t-r} - y_{t-r})\\ &\quad - \hat{\omega}_0 x_{t-b} + \hat{\omega}_1 x_{t-b-1} + \cdots + \hat{\omega}_s x_{t-b-s}\end{aligned}$$

A straightforward approach to identifying the ARMA structure of the noise is to use the conventional identification procedure of §§**9.41–9.43** on $\hat{n}_t$. This suggested an AR(2) structure and the first two sample autocorrelations, $r_{\hat{n}}(1) = 0.886$ and $r_{\hat{n}}(2) = 0.743$, yielded the initial autoregressive parameter estimates of $\hat{\varphi}_1 = 1.06$ and $\hat{\varphi}_2 = -0.20$.[7] Thus the identified model for the gas furnace data is

$$Y_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1 - \delta_1 B - \delta_2 B^2} X_{t-3} + \frac{1}{1 - \varphi_1 B - \varphi_2 B^2} a_t$$

12.18 Box and Jenkins stressed the importance of using the rational form $v(B) = \delta^{-1}(B)\omega(B)$ for the transfer function to reduce the number of parameters that need to be estimated, particularly as the impulse response weights will typically have large variances and be highly correlated. Related to this, the identification procedure requires that the variation in the input X be reasonably large compared with the variation in the noise and/or a large amount of data is available, otherwise identification may fail, although even then a process of beginning with a simple and rudimentary model and extending it if necessary after estimation and checking (see §**12.19** below) may prove successful.

Box and Jenkins also emphasized the problems that may arise through *lack of uniqueness*. Since the model (12.20) could equally well be represented by

$$L(B)Y_t = L(B)\delta^{-1}(B)\omega(B)X_{t-b} + L(B)\varphi^{-1}(B)\theta(B)a_t$$

it is possible that the identification strategy could lead to a model of unnecessarily complicated form. This possibility is reduced if simple rational forms of the transfer function are employed initially – these are often found to be adequate so that more complicated models should only be considered if the need is demonstrated. Potential common factors in the operators on Y_t, X_t and a_t should be investigated and, if possible, removed, as their presence can lead to instability in estimation. Considerable ingenuity may be needed in order to do this, as estimated coefficients will often be accompanied by large standard errors, but parameter *redundancy* should be avoided at all costs, with a *parsimonious* parameterization always being the goal of model building.

Estimation and checking of transfer function models

12.19 Box and Jenkins estimated the transfer function with noise model (12.20) using an extension of the nonlinear least squares method outlined in §§**9.31–9.40**: Box and Jenkins (1970, chapter 11.3) may be consulted for details.

After estimation, serious model inadequacy can usually be detected by examining

(a) the autocorrelation function $r_{\hat{a}\hat{a}}(k)$ of the residuals $\hat{a}_t$ from the fitted model, and

(b) certain cross-correlation functions involving the input and the residuals, in particular the cross-correlation function $r_{\alpha\hat{a}}(k)$ between the prewhitened input α_t and the residuals $\hat{a}_t$.

The model (12.20) can be written as

$$\begin{aligned} y_t &= \delta^{-1}(B)\omega(B)x_{t-b} + \varphi^{-1}(B)\theta(B)a_t \\ &= \upsilon(B)x_t + \psi(B)a_t \end{aligned}$$

Suppose that an incorrect model has been identified, producing the residuals a_{0t}, where

$$y_t = \upsilon_0(B)x_t + \psi_0(B)a_{0t} \tag{12.25}$$

These residuals can be written as

$$a_{0t} = \psi_0^{-1}(B)\{\upsilon(B) - \upsilon_0(B)\}x_t + \psi_0^{-1}(B)\psi(B)a_t \tag{12.26}$$

so that it is apparent that, if a wrong model is selected, the a_{0t}s will be autocorrelated and also cross-correlated with the x_ts and hence the α_ts which generate the x_ts. Two important cases need considering.

Transfer function model correct: noise model incorrect. If $\upsilon_0(B) = \upsilon(B)$ then (12.26) becomes

$$a_{0t} = \psi_0^{-1}(B)\psi(B)a_t$$

The a_{0t}s would *not* be cross-correlated with the input but they would be autocorrelated and the form of the autocorrelation function may indicate how the noise structure could be modified.

Transfer function model incorrect. From (12.26), if the transfer function is incorrect then, as stated above, the a_{0t}s will be autocorrelated and cross-correlated with both the x_ts and the α_ts, *even if the noise model were correct*, so that a cross-correlation analysis could indicate the modifications needed in the transfer function model.

12.20 Of course, in practice the parameters of the transfer function model are unknown and must be estimated, so that the checks suggested in the previous section must be applied to the residuals $\hat{a}_t$ computed after least squares fitting. This will introduce discrepancies into autocorrelation and cross-correlation functions so that some caution is warranted when using these results. Nevertheless, if the estimated autocorrelation function of the residuals, $r_{\hat{a}\hat{a}}(k)$, shows marked correlation patterns then model inadequacy is suggested, while if the cross-correlation checks do not indicate that the transfer function model is inadequate, then the problems will tend to lie in the fitted noise model $n_t = \psi_0(B)a_t$. In this latter case, identification of the 'subsidiary' model $\hat{a}_{0t} = \mathrm{T}(B)a_t$ to represent the autocorrelation of the residuals from the 'primary' model (12.25) will indicate that the modification of the noise model should take the form $n_t = \psi_0(B)\mathrm{T}(B)a_t$.

Determining the significance of a residual autocorrelation departing from zero needs to take account of the issues previously discussed in §**9.46** when dealing with residuals from univariate models, although individual tests of significance and a joint test using the $Q(K)$ statistic can continue to be used. Similar statistics may be employed for assessing the significance of the cross correlations of the residuals with the prewhitened input, $r_{\alpha\hat{a}}(k)$: for example, a test of the joint significance of the first K of these cross-correlations is given by the statistic

$$S(K) = T'\sum_{k=0}^{K} r^2_{\alpha\hat{a}}(k)$$

which will be approximately distributed as $\chi^2(K-r-s)$, T' being the effective sample size used for estimation, while an individual cross-correlation will have a variance of $1/T'$, although in practice low order correlations may have a considerably smaller variance than this. (Note that the degrees of freedom in $S(K)$ are independent of the number of parameters fitted in the noise model.)

12.21 The transfer function model fitted to the gas furnace data was

$$\begin{pmatrix} 1 - \underset{(\pm 0.21)}{0.57}\,B - \underset{(\pm 0.14)}{0.01}\,B^2 \end{pmatrix} Y_t = -\begin{pmatrix} \underset{(\pm 0.08)}{0.53} + \underset{(\pm 0.15)}{0.37}\,B + \underset{(\pm 0.16)}{0.51}\,B^2 \end{pmatrix} X_{t-3} + \frac{1}{\begin{pmatrix} 1 - \underset{(\pm 0.05)}{1.53}\,B + \underset{(\pm 0.05)}{0.63}\,B^2 \end{pmatrix}} a_t$$

where ± one standard error limits are shown in parentheses and $\hat{\sigma}_a^2 = 0.0561$. Diagnostic checks showed no evidence of model inadequacy with $Q(36) = 41.7 \sim \chi^2(34)$ and $S(35) = 29.4 \sim \chi^2(31)$ both being insignificant. The estimate $\hat{\delta}_2 = 0.01$ is very small when compared to its standard error of ± 0.14 and omitting it from the model has no effect on the estimates of the other parameters.

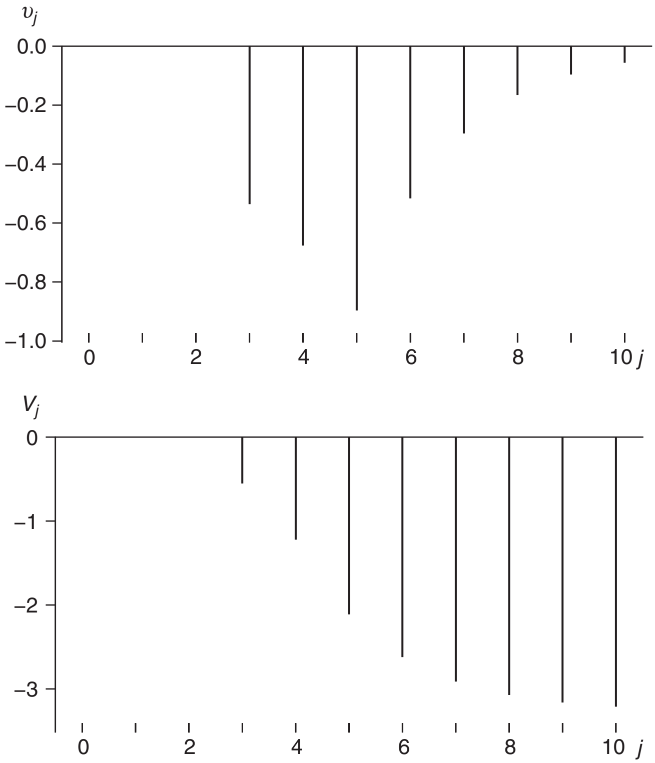


Figure 12.5 Impulse and step responses for the transfer function model $(1-0.57B)Y_t = -(0.53+0.57B+0.51B^2)X_{t-3}$ fitted to the gas furnace data

The impulse response and step functions are shown in Figure 12.5: the former has two initial values of $\hat{v}_3 = -0.53$ and $\hat{v}_4 = 0.67$, after which the weights decay geometrically from $\hat{v}_5 = 0.89$ as $\hat{v}_j = 0.57\hat{v}_{j-1}$; the latter tends, without overshooting, to the steady state gain

$$g = \frac{-(0.53+0.37+0.51)}{1-0.57} = 3.28$$

Forecasting using leading indicators

12.22 Box and Jenkins utilized the transfer function model to develop the technique of forecasting Y_t from the 'leading indicator' X_t. This is essentially a generalization of their approach to forecasting individual time series, as discussed in §§**11.4–11.10**, and will not be developed here: see Box and Jenkins (1970, chapter 11.5) for details.

Multiple time series models

12.23 The transfer function model can, at least conceptually, be extended to multiple inputs, although the full identification of such models was not considered by Box and Jenkins. A more natural generalization of the single input–single output transfer function model might be to consider a vector of time series, thus leading to the analysis of *multiple* time series. Early contributions to the field were Bartlett and Rajalakshman (1953) and Whittle (1953b), but the seminal study was Quenouille (1957), which represented the state of the art of the subject until well into the 1970s. The following sections outline the approach taken by Quenouille, but it is emphasized that these provide very much a 'bare bones' treatment of the much more detailed analysis provided in Quenouille's monograph.

Thus, consider n time series of length T, with the tth observation on the ith series denoted as $x_{i,t}$ and the corresponding column vector of the n variables denoted as $\mathbf{x}_t$. The covariance between $x_{i,t}$ and $x_{j,t-k}$ is denoted as $\gamma_{ij}(k)$ and the correlation by $\rho_{ij}(k)$. The sample estimates of these quantities are $c_{ij}(k)$ and $r_{ij}(k)$. The $n \times n$ matrices $\mathbf{\Gamma}_k$, $\mathbf{P}_k$, $\mathbf{C}_k$ and $\mathbf{R}_k$ contain the collections of each of these quantities at lag k, from which it follows that $\mathbf{\Gamma}_{-k} = \mathbf{\Gamma}'_k$, etc. Using the lag operator B, define

$$\mathbf{\Gamma} = \sum_{k=-\infty}^{\infty} \mathbf{\Gamma}_k B^k$$

with corresponding definitions for $\mathbf{P}$, $\mathbf{C}$ and $\mathbf{R}$.

The observations $x_{i,t}$ are assumed to be generated from n infinite random series $\varepsilon_{i,t}$ by the infinite linear moving average process

$$x_{i,t} = \sum_{j=1}^{n} \sum_{k=0}^{\infty} f_{ij,k} \varepsilon_{j,t-k} \tag{12.27}$$

In vector form,

$$\mathbf{x}_t = \sum_{k=0}^{\infty} \mathbf{F}_k \boldsymbol{\varepsilon}_{t-k} = \sum_{k=0}^{\infty} \mathbf{F}_k B^k \boldsymbol{\varepsilon}_t = \mathbf{F}(B)\boldsymbol{\varepsilon}_t, \quad \text{where } \mathbf{F}(B) = \sum_{k=0}^{\infty} \mathbf{F}_k B^k$$

in which $\boldsymbol{\varepsilon}_t$ and $\mathbf{F}_k$ are defined analogously to $\mathbf{x}_t$ and $\mathbf{\Gamma}_k$ respectively. For an autoregressive scheme, $\mathbf{F}^{-1}(B)$ will be a finite-order polynomial in B such that

$$\begin{aligned}\mathbf{F}^{-1}(B) &= \mathbf{A}_0 + \mathbf{A}_1 B + \cdots + \mathbf{A}_p B^p \\ &= \mathbf{A}_0(\mathbf{I} - \mathbf{U}_1 B - \cdots - \mathbf{U}_p B^p)\end{aligned}$$

with determinant

$$\begin{aligned}|\mathbf{F}^{-1}| &= a_0 + a_1B + \cdots + a_mB^m \\ &= a_0(1 - u_1B - \cdots - u_mB^m)\end{aligned}$$

The quantity p is the order of the autoregressive scheme and, unless $\mathbf{A}_p$ is singular, $m = np$. If $\mathbf{A}_p$ is singular then there is some r, called the minimum order, for which $rn \geq m > (r-1)n$.

As in the univariate case, three schemes are available to model $\mathbf{x}_t$: the finite moving average, where $\mathbf{F}(B)$ is a finite-order polynomial in B; the autoregression $\mathbf{F}^{-1}(B)\mathbf{x}_t = \boldsymbol{\varepsilon}_t$; and the autoregressive scheme with moving average residuals, where $\mathbf{F}(B) = \mathbf{G}(B)/\mathbf{H}(B)$, so that $\mathbf{H}(B)\mathbf{x}_t = \mathbf{G}(B)\boldsymbol{\varepsilon}_t$.[8]

12.24 Assuming, for simplicity, that $x_{i,t}$ and $\varepsilon_{i,t}$ have zero means and that $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \sigma^2\mathbf{I}$, it then follows that

$$\boldsymbol{\Gamma}_k = \sigma^2\sum_{i=0}^{\infty}\mathbf{F}_{k+i}\mathbf{F}'_i \quad \text{and} \quad \boldsymbol{\Gamma} = \sigma^2\mathbf{F}(B)\mathbf{F}'(B^{-1})$$

If fewer than n random variables are required to generate $\mathbf{x}_t$ then $\mathbf{F}$ ($\boldsymbol{\Gamma}$) is singular and vice versa. The rank of $\mathbf{F}$ ($\boldsymbol{\Gamma}$) gives the number of random variables needed and the difference between this and n gives the number of identities between the variables of $\mathbf{x}_t$. The latent vectors of $\mathbf{F}$ and $\boldsymbol{\Gamma}$ corresponding to the zero latent roots of these two matrices give the identities existing between the variables. For example, if

$$\begin{aligned}x_{1,t} &= \varepsilon_{1,t} + \varepsilon_{2,t-1} \\ x_{2,t} &= \varepsilon_{1,t} + \varepsilon_{2,t} \\ x_{3,t} &= \varepsilon_{2,t} + \varepsilon_{2,t-1}\end{aligned}$$

then

$$\mathbf{F} = \begin{bmatrix} 1 & B \\ 1 & 1 \\ 0 & 1+B \end{bmatrix}$$

and

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & B \\ 1 & 1 \\ 0 & 1+B \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ B^{-1} & 1 & 1+B^{-1} \end{bmatrix} = \begin{bmatrix} 2 & 1+B & 1+B \\ 1+B^{-1} & 2 & 1+B^{-1} \\ 1+B^{-1} & 1+B & 2+B+B^{-1} \end{bmatrix}$$

$|\boldsymbol{\Gamma}| = 0$ and the latent vector corresponding to the zero latent root of $\boldsymbol{\Gamma}$ is proportional to $[1+B, -(1+B), 1-B]$, showing that the single identity existing between the variables is

$$(1+B)x_{1,t} - (1+B)x_{2,t} + (1-B)x_{3,t} = 0$$

i.e.,

$$x_{1,t} + x_{1,t-1} + x_{3,t} = x_{3,t} + x_{2,t-1} + x_{3,t-1}$$

12.25 Any restriction on **F** will give rise to corresponding restrictions on $\mathbf{\Gamma}$ and **P** and hence upon the covariances and correlations of the variables in $\mathbf{x}_t$. For example, if $n=2$ and $x_{1,t}$ and $x_{2,t}$ are generated just by $\varepsilon_{1,t}$, then $|\mathbf{F}|=|\mathbf{\Gamma}|=|\mathbf{P}|=0$ and hence

$$\sum_{i=-\infty}^{\infty} \rho_{11}(i)\rho_{22}(k-i) = \sum_{i=-\infty}^{\infty} \rho_{12}(i)\rho_{21}(i) \quad \text{for all } k,$$

which is a necessary and sufficient condition for such dependence. Similarly, for a finite moving average scheme, $\rho_{ij}(s)=0$ for all s greater than q, the order of the moving average. For an autoregressive scheme, the fundamental equations linking the covariance matrices are (cf. §**7.18** for their univariate counterparts)

$$\mathbf{A}_0\mathbf{\Gamma}_s + \mathbf{A}_1\mathbf{\Gamma}_{s-1} + \cdots + \mathbf{A}_p\mathbf{\Gamma}_{s-p} = 0 \quad s>0$$

i.e.,

$$\mathbf{\Gamma}_s = \mathbf{U}_1\mathbf{\Gamma}_{s-1} + \mathbf{U}_2\mathbf{\Gamma}_{s-2} + \cdots + \mathbf{U}_p\mathbf{\Gamma}_{s-p} \quad s>0$$

or equivalently

$$a_0\mathbf{\Gamma}_s + a_1\mathbf{\Gamma}_{s-1} + \cdots + a_m\mathbf{\Gamma}_{s-m} = 0 \quad s>n$$

and

$$\mathbf{\Gamma}_s = u_1\mathbf{\Gamma}_{s-1} + u_2\mathbf{\Gamma}_{s-2} + \cdots + u_m\mathbf{\Gamma}_{s-m} \quad s>n$$

The same formulae hold for autoregressive schemes whose errors follow moving averages except that the limits are changed to $s>q$.

12.26 The characteristics of any autoregressive scheme are determined by the number of non-zero roots of $|\mathbf{F}^{-1}|=0$, i.e., m, and its order n. If, for example, $m<n$, then the scheme either degenerates into separate schemes, some of which will be of lower order, e.g.,

$$\mathbf{F}^{-1}(B) = \begin{bmatrix} 1 & 0 \\ 0 & a_0 + a_1B + a_2B^2 \end{bmatrix}$$

or it is possible to construct a matrix $\mathbf{U}_1$ such that $\mathbf{\Gamma}_s=\mathbf{U}_1\mathbf{\Gamma}_{s-1}$ for $s>1$, so that the whole scheme behaves as if it was of lower order. Similarly, if $n<m\leq 2n$ and the scheme does not break into separate schemes, it is possible to satisfy

$\mathbf{\Gamma}_s = \mathbf{U}_1\mathbf{\Gamma}_{s-1} + \mathbf{U}_2\mathbf{\Gamma}_{s-2}$ for $s > 2$, so that the scheme behaves like a lower-order scheme except that $\mathbf{\Gamma}_0$ fails to satisfy the recurrence relationships. This would be characteristic of schemes with correlated or even superposed errors, so that such degenerate schemes may not be able to be distinguished from these other types of scheme.

Canonical variables

12.27 Quenouille (1957, section 2.5) used an important result to provide a means of calculating the covariance matrices. This states that if $z_1, z_2, \ldots, z_m$ are the roots of $|\mathbf{F}^{-1}| = 0$, then

$$\mathbf{F}(B) = \sum_{i=1}^{m} \mathbf{u}_i\mathbf{v}'{}_i \frac{1}{B - z_i}$$

where

$$\mathbf{u}_i\mathbf{v}'{}_i = \frac{\mathrm{adj}\,(\mathbf{F}^{-1}(z_i))}{a_m \prod\limits_{j \neq i} (z_i - z_j)}$$

with adj($\mathbf{X}$) denoting the adjoint matrix of $\mathbf{X}$. It is then straightforward to show that

$$\mathbf{F}^{-1}(z_i)\mathbf{u}_i = \mathbf{F}'^{-1}(z_i)\mathbf{v}_i = 0$$

From this result it is then possible to calculate a set of vectors, $\mathbf{t}_i$, defined as

$$\mathbf{v}_i'\mathbf{F}^{-1}(B) = (B - z_i)\mathbf{t}'{}_i$$

which give rise to a set of *canonical* variables $\mathbf{t}'{}_i\mathbf{x}_t$ with the property that

$$(B - z_i)\mathbf{t}_i\mathbf{x}_t = \mathbf{v}_i'\mathbf{F}^{-1}(B)\mathbf{x}_t$$

i.e.,

$$(1 - z_i^{-1}B)\mathbf{t}_i\mathbf{x}_t = -z_i^{-1}\mathbf{v}_i'\boldsymbol{\varepsilon}_t$$

so that each of the canonical variables follows a Markov scheme with parameter z_i^{-1}. As an illustration of this property, consider the scheme

$$2x_{1,t} + x_{1,t-1} + x_{2,t} = \varepsilon_{1,t}$$

$$x_{1,t-1} + 6x_{2,t} + x_{2,t-1} = \varepsilon_{2,t}$$

Here

$$\mathbf{F}^{-1}(B) = \begin{bmatrix} 2+B & 1 \\ B & 6+B \end{bmatrix} \quad \mathrm{adj}(\mathbf{F}^{-1}(B)) = \begin{bmatrix} 6+B & -1 \\ -B & 2+B \end{bmatrix}$$

and

$$|\mathbf{F}^{-1}| = 12 + 7B + B^2$$

so that $z_1 = -3$, $z_2 = -4$ and $a_2 = 1$. Thus

$$\mathbf{u}_1\mathbf{v}'_1 = \begin{bmatrix} 3 & -1 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} [3 \quad -1]$$

$$\mathbf{u}_2\mathbf{v}'_2 = \begin{bmatrix} -2 & 1 \\ -4 & 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} [-2 \quad 1]$$

It then follows that

$$[3 \quad -1] \begin{bmatrix} 2+B & 1 \\ B & 6+B \end{bmatrix} = (B+3)\mathbf{t}'_1$$

to give $\mathbf{t}'_1 = [2 \quad -1]$ and, by a similar calculation, $\mathbf{t}'_2 = [-1 \quad 1]$. The canonical variables are therefore $2x_{1,t} - x_{2,t}$ and $-x_{1,t} + x_{2,t}$ and an alternative representation of the scheme is provided by

$$\left(1 + \frac{1}{3}B\right)(2x_{1,t} - x_{2,t}) = \varepsilon_{1,t} - \frac{1}{3}\varepsilon_{2,t}$$

$$\left(1 + \frac{1}{4}B\right)(-x_{1,t} + x_{2,t}) = -\frac{1}{2}\varepsilon_{1,t} + \frac{1}{4}\varepsilon_{2,t}$$

This result is interesting in that the individual variables will normally each follow an autoregressive scheme of order *m* with moving average errors of order *m*, as is seen by noting that the autoregression $\mathbf{F}^{-1}(B)\mathbf{x}_t = \boldsymbol{\varepsilon}_t$ can be written as $\mathrm{adj}(\mathbf{F}(B))\mathbf{x}_t = |\mathbf{F}(B)|\boldsymbol{\varepsilon}_t$. The canonical variables, however, comprise those combinations of the individual variables having the simplest possible serial correlation properties. The canonical variables associated with real roots will then, as in the example above, follow Markov schemes. If there are complex roots then the associated canonical variables will have real and imaginary parts that jointly follow a Markov scheme.

There will also be a relationship between the $\mathbf{t}_i$ and $\mathbf{u}_j$ vectors. Since $(B - z_i)\mathbf{t}_i\mathbf{u}_j = \mathbf{v}'_i\mathbf{F}^{-1}(B)\mathbf{u}_j$ disappears for $B = z_j$, it therefore follows that $\mathbf{t}_i\mathbf{u}_j = 0$.

The identification problem

12.28 Is it possible to determine the moving average structure from a knowledge of the covariance structure, i.e., to determine **F** from **Γ**? If $\mathbf{F}(B)$ satisfies $\boldsymbol{\Gamma} = \sigma^2\mathbf{F}(B)\mathbf{F}'(B^{-1})$ then so will $\mathbf{F}(B)\boldsymbol{\Phi}(B)\mathbf{J}\boldsymbol{\Psi}(B)$, where $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Psi}(B)$ are diagonal matrices with elements $\phi_i(B^{-1})/\phi_i(B)$ and $\psi_i(B^{-1})/\psi_i(B)$, respectively, and **J** is any matrix satisfying $\mathbf{JJ}' = \mathbf{I}$. Thus, if any solution exists, there must be an infinite number of solutions.

If the structure of the scheme, i.e., **F**, is known, is it possible to uniquely determine **Γ**? Again, this would not seem to be generally possible, as Quenouille (1957, section 3.1) showed, by way of examples, that different schemes may give rise to the same **Γ**.

Effects of model misspecification

12.29 It is quite possible that a scheme might be specified that it incorrect in one or more ways. Of particular interest is the case where some variables have been incorrectly included or excluded. Suppose first that the true specification is

$$\begin{bmatrix} \mathbf{K} & \mathbf{L} \\ \mathbf{M} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_t \end{bmatrix}$$

where **K**, **L**, **M** and **N** are polynomials in *B* whose arguments have been suppressed for clarity of notation. If the set of variables $\mathbf{y}_t$ are omitted the scheme becomes

$$(\mathbf{K} - \mathbf{LN}^{-1}\mathbf{M})\mathbf{x}_t = \boldsymbol{\varepsilon}_t - \mathbf{LN}^{-1}\boldsymbol{\eta}_t \tag{12.28}$$

If some of the $\mathbf{x}_t$ variables, say $\mathbf{x}_{1,t}$, do not directly affect $\mathbf{y}_t$ and are themselves not directly affected by $\mathbf{y}_t$, the scheme may be written as

$$\begin{bmatrix} \mathbf{G} & \mathbf{H} & \mathbf{0} \\ \mathbf{J} & \mathbf{K} & \mathbf{L} \\ \mathbf{0} & \mathbf{M} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{1,t} \\ \mathbf{x}_{2,t} \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_{1,t} \\ \boldsymbol{\varepsilon}_{2,t} \\ \boldsymbol{\eta}_t \end{bmatrix} \tag{12.29}$$

Now the omission of $\mathbf{y}_t$ gives

$$\begin{bmatrix} \mathbf{G} & \mathbf{H} \\ \mathbf{J} & \mathbf{K} - \mathbf{LN}^{-1}\mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{1,t} \\ \mathbf{x}_{2,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_{1,t} \\ \boldsymbol{\varepsilon}_{2,t} - \mathbf{LN}^{-1}\boldsymbol{\eta}_t \end{bmatrix}$$

Alternatively, if the variables $\mathbf{y}_{1,t}$ from an unrelated scheme are included, so that the true scheme is

$$\begin{bmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{K} & \mathbf{L} \\ \mathbf{0} & \mathbf{M} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_{1,t} \\ \mathbf{y}_{2,t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_{1,t} \\ \boldsymbol{\eta}_{2,t} \end{bmatrix} \tag{12.30}$$

the observed variables will follow the scheme

$$\begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{K} - \mathbf{L}\mathbf{N}^{-1}\mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_{1,t} - \mathbf{L}\mathbf{N}^{-1}\boldsymbol{\eta}_{2,t} \end{bmatrix}$$

It is clear that these misspecifications are closely related. Noting that $|\mathbf{N}|\mathbf{N}^{-1} = \text{adj}(\mathbf{N})$ is a polynomial in B, premultipying (12.28) by $|\mathbf{N}|$ reduces both sides of the equation to polynomials in B, with the order of the polynomial on the left-hand side potentially being increased substantially, while the appearance of a polynomial on the right-hand side makes the errors now serially correlated. In other words, omitting relevant variables can quickly result in the scheme for the remaining variables being too complicated to be analysed easily, and making the simplicity of the original scheme completely undetectable. Whether this is so will depend upon the importance of the omitted variables, as shown by the matrices **K**, **L**, **M** and **N**. For example, if $\mathbf{L} = 0$, the $\mathbf{y}_t$ variables do not affect the $\mathbf{x}_t$ variables and the scheme is unaltered. Similarly, if **L** and **N** are independent of B, i.e., if previous values of $\mathbf{y}_t$ do not directly influence either $\mathbf{y}_t$ itself or $\mathbf{x}_t$, the scheme reduces to a simple autoregression of order equal to the larger of the orders of **K** and **M**. If the order of **N** is large relative to those of **L** and **M**, the disturbance to the scheme will be small: in particular, if, as in (12.29), part of the scheme is not directly related to the omitted variables, that part will remain unaffected. Conversely, in (12.30), which is a special case of (12.29), the order of **N** may be expected to be larger and that of **K** smaller, which would mean that $\mathbf{y}_{1,t}$ will follow a very different scheme to that of $\mathbf{x}_t$.

Dealing with nonstationarities

12.30 Quenouille was well aware of the importance of dealing with nonstationarities and the alternative forms that trending time series could take, stating that

> (t)wo types of trend suggest themselves: trend of a polynomial or functional nature and trend of a stochastic nature.
>
> If the trend is assumed to be a polynomial in t, the easiest and best method of analysis is to include extra terms in the analysis corresponding to linear, quadratic, ... components of the trend. Although these terms are completely predictable from previous terms, they may be treated as variables for which the error is zero. (Quenouille, 1957, page 51)

However, he continued that

> (i)t is my opinion that many economic time series trends will have a stochastic rather than a polynomial nature. That is to say that certain terms in equation [12.27] will combine to give rise to what is known as 'trend'. ...
>
> The position here, then, is that, if the situation discussed in [§**12.13**] is not to obtain, all the variables relevant to the trend should be included in the scheme. This is likely to be difficult to achieve, and methods adopted to overcome the resulting complications differ.
>
> Most commonly, a moving-difference formula, perhaps of the simple type $x_{i,t} - x_{i,t-1}$ is applied to the original series. *Uncritical application of this procedure is, however, more dangerous than the application of no trend-reducing procedure at all. It is essential that the decision as to the procedure should not be made without regard to the likely character of the series.*
>
> ... *Thus, whether the trend is a polynomial or not, the procedure of including polynomial terms as extra variables is not a bad one, since it does not commit us to their acceptance without further consideration, and, if they are rejected, they may easily be dropped from the analysis. For similar reasons, it is preferable not to carry out any differencing procedure for any initial analysis.* (*ibid.*, pages 54–5: italics added for emphasis)

To illustrate the potential impact of differencing in the presence of stochastic trends, Quenouille considered decomposing each individual $x_{i,t}$ into independent trend, $\mu_{i,t}$, and stationary, $z_{i,t}$, components such that $G(B)\mu_{i,t} = \xi_{i,t}$ and $F(B)z_{i,t} = \zeta_{i,t}$. This implies that (cf. §**11.19**)

$$F(B)G(B)x_{i,t} = G(B)\zeta_{i,t} + F(B)\xi_{i,t}$$

and, since the autocorrelations of $\mu_{i,t}$ are likely to be slowly decaying from unity, $G(B) \approx 1 - B$ and it will appear that the first differences of $x_{i,t}$ will follow an ARMA process. In terms of the vector time series $\mathbf{x}_t$, Quenouille suggested that the following scheme was likely to provide a good representation of many nonstationary time series:

$$\mathbf{x}_t = \boldsymbol{\mu}_t + \mathbf{z}_t$$

where

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \mathbf{B}_0^{-1}\boldsymbol{\eta}_t$$

and

$$\mathbf{z}_t = \mathbf{U}_1\mathbf{z}_{t-1} + \mathbf{A}_0^{-1}\boldsymbol{\varepsilon}_t$$

which is a multivariate extension of the exponential smoothing type model of §**11.18**.

Empirical modelling of multiple time series

12.31 The basic statistics required for building a multiple time series model are the sample covariance and autocorrelation matrices $\mathbf{C}_k$ and $\mathbf{R}_k$, the latter comprising the elements

$$r_{ij}(k) = \frac{\sum_{t=s+1}^{T}(x_{i,t} - \bar{x}_i)(x_{j,t-s} - \bar{x}_j)}{\left(\sum_{t=s+1}^{T}(x_{i,t} - \bar{x}_i)^2 \sum_{t=s+1}^{T}(x_{j,t-s} - \bar{x}_j)^2\right)^{\frac{1}{2}}}$$

After some experimentation using simulated series, Quenouille suggested that the correlation *matrix quotients* $\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$ should be examined, along with their successive differences, $\mathbf{R}_k\mathbf{R}_{k-1}^{-1} - \mathbf{R}_{k-1}\mathbf{R}_{k-2}^{-1}$. For large samples, these quotients should take approximately the same value, namely $\mathbf{U}_1$, for Markov schemes, so that their successive differences should be close to zero. The latent roots of the quotients (or their moduli if the roots are complex) will reflect the number of independent factors significantly affecting the correlation structure of the scheme generating the data.

12.32 A more sensitive approach to determining the order of an autoregressive scheme is to use the partial autocorrelations based test proposed by Quenouille (1949c) and extended to the multivariate context. Suppose $\mathbf{x}_t$ is generated by the scheme

$$\sum_{i=0}^{p} \mathbf{A}_i \mathbf{x}_{t-i} = \boldsymbol{\varepsilon}_t \quad E(\boldsymbol{\varepsilon}_{t+s}\boldsymbol{\varepsilon}_t') = \begin{cases} 0, & s \neq 0 \\ \mathbf{I}, & s = 0 \end{cases}$$

and

$$\mathbf{x}_t = \sum_{i=1}^{p} \mathbf{U}_i \mathbf{x}_{t-i} + \mathbf{A}_0^{-1}\boldsymbol{\varepsilon}_t$$

There will then be a second set of related random variables $\boldsymbol{\eta}_t$ such that

$$\sum_{i=0}^{p} \mathbf{B}_i \mathbf{x}_{t+i} = \boldsymbol{\eta}_t \quad E(\boldsymbol{\eta}_{t+s}\boldsymbol{\eta}_t') = \begin{cases} 0, & s \neq 0 \\ \mathbf{I}, & s = 0 \end{cases}$$

and

$$\mathbf{x}_t = \sum_{i=1}^{p} \mathbf{W}_i \mathbf{x}_{t-i} + \mathbf{B}_0^{-1}\boldsymbol{\eta}_t$$

with $E(\boldsymbol{\varepsilon}_{t+s}\boldsymbol{\eta}'_t) = 0,\ s > p$. The connection between the two representations is given by the equations

$$\mathbf{A}'_0\mathbf{A}_0 = (\boldsymbol{\Gamma}_0 - \mathbf{U}_1\boldsymbol{\Gamma}'_1 - \mathbf{U}_2\boldsymbol{\Gamma}'_2 - \cdots)^{-1}$$

$$\mathbf{B}'_0\mathbf{B}_0 = (\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}'_1\mathbf{W}'_1 - \boldsymbol{\Gamma}'_2\mathbf{W}'_2 - \cdots)^{-1}$$

$$\mathbf{A}_i = -\mathbf{A}_0\mathbf{U}_i \quad \mathbf{B}_i = -\mathbf{B}_0\mathbf{W}_i$$

Noting that

$$\boldsymbol{\varepsilon}_{t+p+1}\boldsymbol{\eta}'_t = \sum_{i=0}^{p}\sum_{j=0}^{p}\mathbf{A}_i\mathbf{x}_{t+p+1-i}\mathbf{x}'_{t+j}\mathbf{B}'_j$$

may be interpreted as the partial correlation between $\mathbf{x}_t$ and $\mathbf{x}_{t+p+1}$ when the effects of $\mathbf{x}_{t+i},\ i = 1, \ldots, p$, have been removed, Quenouille proposed using a statistic based on the $n \times n$ matrix

$$T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_{t+p+1}\hat{\boldsymbol{\eta}}'_t = T^{-1}\sum_{t=1}^{T}\sum_{i=0}^{p}\sum_{j=0}^{p}\hat{\mathbf{A}}_i\mathbf{x}_{t+p+1-i}\mathbf{x}'_{t+j}\hat{\mathbf{B}}'_j = \left(\sum_{i=0}^{p}\hat{\mathbf{A}}_i\mathbf{C}_{n+1-i}\right)\hat{\mathbf{B}}'_0 \quad (12.31)$$

each element of which is independent standard normal in large samples. T times the sum of squares of these elements will therefore be approximately distributed as $\chi^2(n^2)$. For example, a Markov scheme has $p = 1$ and (12.31) becomes

$$(\hat{\mathbf{A}}_0\mathbf{C}_2 + \hat{\mathbf{A}}_1\mathbf{C}_1)\hat{\mathbf{B}}_0 = \hat{\mathbf{A}}_0(\mathbf{C}_2 + \hat{\mathbf{A}}_0^{-1}\hat{\mathbf{A}}_1\mathbf{C}_1)\hat{\mathbf{B}}_0 = \hat{\mathbf{A}}_0(\mathbf{C}_2 - \mathbf{C}_1\mathbf{C}_0^{-1}\mathbf{C}_1)\hat{\mathbf{B}}_0$$

with a significant test statistic indicating that the Markov scheme is misspecified and that a higher autoregressive order is warranted.

Quenouille's hog series example

12.33 Quenouille illustrated this modelling procedure by way of a detailed example comprising five annual US time series from 1867 to 1948 on the number and price of hogs ($x_{1,t}$ and $x_{2,t}$), the price and supply of corn ($x_{3,t}$ and $x_{4,t}$) and the farm wage rate ($x_{5,t}$). Exact definitions of each of these variables are given in Quenouille (1957, section 8.1), with the actual data recorded in Table 8.1a of the monograph. Figure 12.6 shows the five series while the sample autocorrelation matrices $\mathbf{R}_k$ for $k = 0, 1, \ldots, 5$ are reported in Table 12.1. The correlation quotients $\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$ and their successive differences $\mathbf{R}_k\mathbf{R}_{k-1}^{-1} - \mathbf{R}_{k-1}\mathbf{R}_{k-2}^{-1}$ are given in Table 12.2, with the latent roots of $\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$ being shown in Table 12.3. The autocorrelation matrices show that there are close correlations between the variables but the successive quotients $\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$ vary considerably, suggesting that a Markov scheme is unlikely to be operating. Each of the differences $\mathbf{R}_k\mathbf{R}_{k-1}^{-1} - \mathbf{R}_{k-1}\mathbf{R}_{k-2}^{-1}$ are nearly

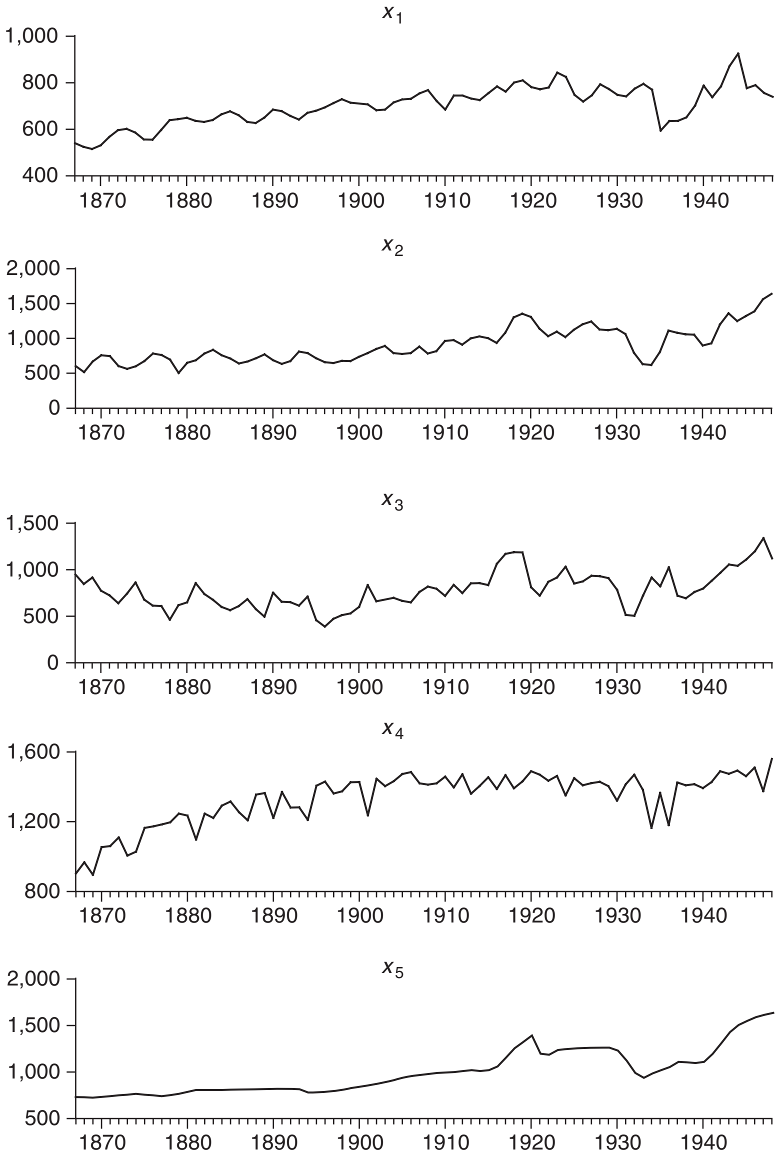


Figure 12.6 Quenouille's US hog series, 1867–1948

singular (the determinants for $k=2,\ldots,5$ being 0.0005, 0.0004, 0.0110 and 0.0004 respectively). This is because the row elements are roughly proportional to [2, 1, 0, −2, −1].

The latent roots of $\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$ show two large real roots, a generally smaller, but rather volatile, third real root and a pair of imaginary roots giving rise to

Table 12.1 Correlation matrices of the hog series

k	$\mathbf{R}_k$
0	$\begin{bmatrix} 1.000 & 0.625 & 0.413 & 0.784 & 0.743 \\ 0.625 & 1.000 & 0.695 & 0.604 & 0.937 \\ 0.413 & 0.695 & 1.000 & 0.124 & 0.726 \\ 0.784 & 0.604 & 0.124 & 1.000 & 0.644 \\ 0.743 & 0.937 & 0.726 & 0.644 & 1.000 \end{bmatrix}$
1	$\begin{bmatrix} 0.882 & 0.705 & 0.315 & 0.857 & 0.730 \\ 0.609 & 0.911 & 0.792 & 0.555 & 0.923 \\ 0.481 & 0.584 & 0.790 & 0.299 & 0.686 \\ 0.757 & 0.648 & 0.267 & 0.797 & 0.645 \\ 0.770 & 0.907 & 0.757 & 0.627 & 0.983 \end{bmatrix}$
2	$\begin{bmatrix} 0.753 & 0.733 & 0.272 & 0.831 & 0.722 \\ 0.677 & 0.765 & 0.806 & 0.529 & 0.874 \\ 0.535 & 0.507 & 0.608 & 0.380 & 0.645 \\ 0.705 & 0.650 & 0.252 & 0.807 & 0.633 \\ 0.775 & 0.855 & 0.712 & 0.649 & 0.949 \end{bmatrix}$
3	$\begin{bmatrix} 0.684 & 0.724 & 0.334 & 0.755 & 0.720 \\ 0.730 & 0.664 & 0.693 & 0.550 & 0.810 \\ 0.567 & 0.550 & 0.455 & 0.426 & 0.628 \\ 0.674 & 0.569 & 0.295 & 0.295 & 0.615 \\ 0.758 & 0.803 & 0.641 & 0.641 & 0.905 \end{bmatrix}$
4	$\begin{bmatrix} 0.653 & 0.696 & 0.441 & 0.676 & 0.699 \\ 0.724 & 0.637 & 0.502 & 0.604 & 0.751 \\ 0.511 & 0.598 & 0.379 & 0.439 & 0.621 \\ 0.701 & 0.487 & 0.231 & 0.775 & 0.561 \\ 0.739 & 0.763 & 0.563 & 0.635 & 0.860 \end{bmatrix}$
5	$\begin{bmatrix} 0.605 & 0.634 & 0.409 & 0.683 & 0.652 \\ 0.653 & 0.642 & 0.347 & 0.624 & 0.710 \\ 0.444 & 0.580 & 0.376 & 0.436 & 0.595 \\ 0.683 & 0.455 & 0.181 & 0.728 & 0.514 \\ 0.677 & 0.736 & 0.500 & 0.619 & 0.815 \end{bmatrix}$

oscillatory behaviour. The first real root corresponds to a canonical variable largely dominated by $x_{5,t}$ and thus captures the major trend component in the data; the second real root, although less stable than the first, captures a further trend component.

Quenouille argued that the stationary components could be captured by three canonical variables defined from the real and imaginary parts of the 3rd/4th

Table 12.2 Correlation quotients and their successive differences

k	$\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$	$\mathbf{R}_k\mathbf{R}_{k-1}^{-1} - \mathbf{R}_{k-1}\mathbf{R}_{k-2}^{-1}$
1	$\begin{bmatrix} 0.64 & 0.56 & -0.11 & 0.26 & -0.37 \\ -0.22 & 0.31 & 0.40 & 0.28 & 0.42 \\ -0.17 & -0.68 & 0.84 & 0.34 & 0.62 \\ 0.43 & 0.62 & -0.03 & 0.41 & -0.58 \\ 0.11 & -0.07 & 0.12 & 0.00 & 0.88 \end{bmatrix}$	
2	$\begin{bmatrix} -1.45 & -0.39 & -0.33 & 2.28 & 0.91 \\ -1.54 & -0.74 & 0.79 & 1.71 & 1.06 \\ -2.57 & -1.74 & 0.93 & 2.71 & 1.77 \\ -1.67 & -0.39 & 0.03 & 2.71 & 0.45 \\ -1.35 & -0.77 & 0.21 & 1.50 & 1.56 \end{bmatrix}$	$\begin{bmatrix} -2.09 & -0.96 & -0.22 & 2.02 & 1.27 \\ -1.32 & -0.95 & 0.39 & 1.43 & 0.63 \\ -2.40 & -1.06 & 0.09 & 2.38 & 1.15 \\ -2.10 & -1.01 & 0.06 & 2.31 & 0.95 \\ -1.46 & -0.70 & 0.09 & 1.50 & 0.68 \end{bmatrix}$
3	$\begin{bmatrix} 7.42 & 3.78 & 0.49 & -5.23 & -5.22 \\ -8.19 & -4.55 & 0.22 & 6.59 & 6.72 \\ -12.05 & -6.96 & -0.94 & 9.45 & 10.57 \\ 12.11 & 6.11 & 2.27 & -8.53 & -10.04 \\ 2.56 & 0.86 & 0.62 & -1.99 & -0.88 \end{bmatrix}$	$\begin{bmatrix} 8.87 & 4.18 & 0.82 & -7.51 & -6.12 \\ -6.64 & -3.80 & -0.57 & 4.87 & 5.66 \\ -9.48 & -5.23 & -1.86 & 6.73 & 8.81 \\ 13.78 & 6.50 & 2.24 & -11.24 & -10.50 \\ 3.91 & 1.63 & 0.41 & -3.49 & -2.44 \end{bmatrix}$
4	$\begin{bmatrix} 2.13 & 1.63 & -0.90 & -0.96 & -1.11 \\ -0.81 & -0.55 & 1.32 & 0.85 & 0.47 \\ 0.62 & -0.16 & 0.49 & -0.44 & 0.30 \\ -1.04 & -0.76 & 1.20 & 1.92 & -0.01 \\ -0.44 & -0.61 & 0.59 & 0.39 & 1.17 \end{bmatrix}$	$\begin{bmatrix} -5.29 & -2.16 & -1.39 & 4.27 & 4.11 \\ 7.38 & 4.00 & 1.10 & -5.73 & -6.25 \\ 12.67 & 6.80 & 1.43 & -9.89 & -10.27 \\ -13.15 & -6.87 & -1.07 & 10.45 & 10.03 \\ -3.00 & -1.47 & -0.03 & 2.38 & 2.05 \end{bmatrix}$
5	$\begin{bmatrix} 1.05 & -0.85 & -1.03 & 0.15 & 1.29 \\ -0.60 & -0.72 & 0.95 & 0.69 & 0.81 \\ 0.32 & -1.05 & -0.05 & -0.00 & 1.39 \\ -0.19 & 0.60 & 0.90 & 1.02 & -1.08 \\ -0.12 & -0.79 & 0.00 & 0.21 & 1.60 \end{bmatrix}$	$\begin{bmatrix} -1.08 & -2.48 & -0.13 & 1.11 & 2.40 \\ 0.21 & -0.17 & -0.37 & -0.16 & 0.34 \\ -0.31 & -0.89 & -0.54 & 0.44 & 1.09 \\ 0.85 & 1.36 & -0.31 & -0.91 & -1.07 \\ 0.32 & -0.18 & -0.59 & -0.18 & 0.43 \end{bmatrix}$

Table 12.3 Latent roots of $\mathbf{R}_k\mathbf{R}_{k-1}^{-1}$

		Root			
k	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
1	0.98	0.81	$0.56 \pm 0.40i$		0.08
2	0.97	1.27	$0.44 \pm 0.40i$		−0.14
3	$0.93 \pm 0.03i$				−9.31
4	0.98	1.07	$-0.54 \pm 0.56i$		2.02
5	0.96	1.02	$0.70 \pm 0.56i$		−0.40

Table 12.4 Correlation matrices of the canonical variables $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$

k	$\mathbf{R}_k$
0	$\begin{bmatrix} 1.00 & -0.10 & 0.32 \\ -0.10 & 1.00 & 0.42 \\ 0.32 & 0.42 & 1.00 \end{bmatrix}$
1	$\begin{bmatrix} 0.62 & -0.59 & -0.04 \\ 0.26 & 0.52 & 0.35 \\ 0.03 & 0.02 & 0.07 \end{bmatrix}$
2	$\begin{bmatrix} 0.18 & -0.68 & -0.22 \\ 0.23 & -0.02 & 0.06 \\ -0.00 & 0.03 & 0.06 \end{bmatrix}$
3	$\begin{bmatrix} -0.12 & -0.44 & -0.25 \\ 0.05 & -0.34 & -0.04 \\ -0.05 & -0.09 & 0.11 \end{bmatrix}$
4	$\begin{bmatrix} -0.26 & -0.01 & -0.04 \\ -0.14 & -0.19 & -0.01 \\ -0.26 & 0.01 & 0.16 \end{bmatrix}$
5	$\begin{bmatrix} -0.25 & 0.15 & -0.06 \\ -0.26 & 0.15 & 0.09 \\ -0.35 & 0.05 & -0.07 \end{bmatrix}$

roots and the trend-free fifth root. These were defined as

$$\begin{aligned} y_{1,t} &= 380x_{1,t} - 68x_{2,t} - 111x_{3,t} - 133x_{4,t} + 80x_{5,t} \\ y_{2,t} &= -109x_{1,t} - 251x_{2,t} + 177x_{3,t} + 18x_{4,t} + 165x_{5,t} \\ y_{3,t} &= 1216x_{1,t} + 139x_{2,t} - 7x_{3,t} - 711x_{4,t} - 154x_{5,t} \end{aligned}$$

The autocorrelation matrices of these variables, shown in Table 12.4, clearly indicate that all three of the variables are stationary and trend free, as is confirmed by the plots of the series in Figure 12.7.

12.34 Quenouille subjected the canonical variables to various partial correlation tests. Tests of all three as a group, of just the pair $y_{1,t}$ and $y_{2,t}$, and of various linear combinations all refuted the hypothesis that they were generated by Markov schemes. He concluded that there were two likely possibilities: either

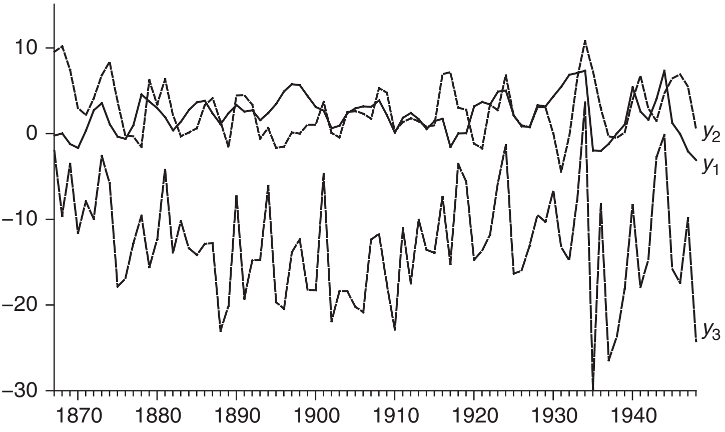


Figure 12.7 Canonical variables $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$

there was serial correlation in the residuals (or superposed error), or the parameters of the scheme were changing over time, or indeed both of these possibilities were occurring. While he gave some evidence in favour of these possibilities he had to conclude that this was not conclusive and that a much fuller analysis was probably necessary.

Nevertheless, Quenouille's analysis, the first of its kind and one which would not be repeated for many years, was able to make several conjectures. First, there was clear evidence of an oscillatory component. This he explained by high hog prices or low corn prices tending to be followed by an increase in hog numbers, which were in turn followed by lower hog prices, while high corn prices or farm wage rates would tend to be followed by higher hog prices. Second, there existed non-oscillatory movements with very high serial correlations, which Quenouille characterized as 'trend' components. Third, part of the trend was accounted for by the farm wage rate, $x_{5,t}$, which appeared to act as an exogenous canonical variable (this may be seen from the very small non-diagonal elements in the last row of $\mathbf{R}_1\mathbf{R}_0^{-1}$).

12.35 The dynamic modelling of the relationships between two or more time series developed in this chapter, beginning with the ideas of Fisher and Smith in the mid-1920s, and culminating with the transfer function approach of Box and Jenkins and the full multiple time series framework of Quenouille, provided a formal basis for the econometric modelling of time series, a discipline that expanded dramatically in the last forty years, with some of the links being drawn out in the concluding Chapter 16.

13

Spectral Analysis of Time Series: The Periodogram Revisited and Reclaimed

Revisiting and revealing the periodogram

13.1 The periodograms of the sunspot numbers and Beveridge's wheat price index calculated in Chapter 3 are rather 'jumpy' and show numerous peaks, particularly in the latter, which led to much disquiet from commentators and discussants when they were first published and a rather unconvincing explanation by Beveridge (§**3.8**). Subsequently, periodogram analysis lost much of its appeal, but it was only in the late 1940s that a convincing explanation was offered for this erratic behaviour of the periodogram.

The key to this explanation was Wold's (1938) Theorem 5 (§**7.12**), which linked the autocorrelations ρ_k, $k=0,1,2,\ldots$, of a stationary process X_t to the Fourier coefficients of a non-decreasing function $F(\omega)$, which has the form of a cumulative distribution function and is known as the *integrated spectrum* of the process. A convenient way of writing this relationship is, from Moran (1949) and Bartlett (1950),

$$\rho_k = \int_{-\pi}^{\pi} e^{ik\omega} dF(\omega)$$

or, by using its Fourier transform (cf. §**8.11**), as

$$dF(\omega) = f(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{k=1}^{\infty} \rho_k \cos \omega k\right) \quad 0 \leq \omega \leq \pi$$

The relevance of this result becomes clear when it is realized that the *spectrum*, or *spectral density*, $f(\omega)=dF(\omega)$, may be either discrete or continuous or, possibly, a combination of the two. Only for a discrete distribution will the concept of a classical harmonic series of the type analysed in Chapter 3 be valid, as a harmonic series will have a discrete spectrum containing 'spikes' at the harmonic frequencies. In general, though, the possibility of any kind of continuous component to the spectrum should be allowed for.[1]

As an example, consider the AR(2) process

$$X_t + aX_{t-1} + bX_{t-2} = \varepsilon_t$$

Bartlett (1948, 1950) showed that the spectrum of this process is given by

$$2\pi f(\omega) = \frac{(1-b)(1-a^2+b^2+2b)}{(1+b)\{1+a^2+b^2-2b+2a(1+b)\cos\omega+4b\cos^2\omega\}} \quad (-\pi \le \omega \le \pi) \tag{13.1}$$

Hence the spectra for white noise ($a=b=0$) and for an AR(1) process ($b=0$) are, respectively, unity and $(1-a^2)/(1+a^2+2a\cos\omega)$. If a is negative this latter spectrum will be large at low frequencies ω and small at high frequencies, the reverse being true for a positive: these are termed low- and high-frequency spectra respectively. The AR(2) spectrum (13.1) will also produce low- and high-frequency spectra for certain values of a and b. It is also possible to obtain spectra with a peak or a trough at an intermediate frequency ω_0, given by

$$\omega_0 = \cos^{-1}\left(-\frac{a(1+b)}{4b}\right)$$

which will occur when $\left|a(1+b)\right| < |4b|$. This has an interesting implication, which was later pointed out in Jenkins and Watts (1968, section 6.2.5). The case where $a^2-4b<0$, for which the autocorrelation function of X is a damped sine wave, lies partially in the region where the spectrum has no intermediate peak, so that a periodicity in the autocorrelation function need not appear as a peak in the spectrum unless the amplitude of the damped sine wave is large enough.

13.2 For a time series of length T, the classic estimator of the periodogram was given in §**3.6**, which Bartlett (1950) showed could be written as the *sample spectrum* or 'intensity'

$$I_p = 2\sum_{k=-T+1}^{T-1}\left(1-\frac{|k|}{T}\right)c_k\cos\omega_p k = \frac{1}{\pi}\left(c_0+2\sum_{k=1}^{T-1}c_k\cos\omega_p k\right) \tag{13.2}$$

where $\omega_p = 2\pi p/T$ and the c_k are the sample autocovariances

$$c_k = \frac{1}{T-k}\sum_{t=1}^{T-k}X_tX_{t+k} \quad (k>0, c_{-k}=c_k)$$

Assuming $E(X_t)=0$, $E(X_t^2)=\sigma^2$ and $E(X_tX_{t+k})=\sigma^2\rho_k$, then the expectation of (13.2) is

$$E(I_p) = 2\sigma^2\sum_{k=-T+1}^{T-1}\left(1-\frac{|k|}{T}\right)\rho_k\cos\omega_p k = \frac{\sigma^2}{\pi}\left(1+2\sum_{k=1}^{T-1}\left(1-\frac{k}{T}\right)\rho_k\cos\omega_p k\right)$$

It can be shown that the limiting value of this expectation as $T \to \infty$ is $\sigma^2 f(\omega_p)$, which is known as the *power spectrum*.

If X_t contains a harmonic component of frequency ω then the autocorrelation function will have a component $\lambda \cos \omega j$ and the spectrum will be discrete, having a spike at ω, at which frequency $E(I_p)$ will tend to infinity as T increases, while tending to zero at all other frequencies. Since $E(I_p) = 2\sigma^2$ for a completely random series, it follows that, when X_t is normal, $P(I_p \geq z) = \exp(-z/E(I_p))$, i.e., I_p is exponentially distributed with mean $E(I_p)$, a result shown originally by Fisher (1929) and which can be used to form the basis for hypothesis tests concerning frequency components (Hartley, 1949).

When the spectrum is continuous, as it will be for the AR(2) process (13.1) and, indeed, for any *linear* process (see Bartlett, 1946), the position changes completely. Bartlett (1950) and Grenander (1951) showed that, in this case, although I_p remains an unbiased estimator, it is inconsistent: although it will fluctuate about $f(\omega_p)$, I_p will not tend to this or to any other value as the sample size T increases, a consequence of its variance being $\sigma^4 f(\omega_p)$, which obviously does not decrease to zero as $T \to \infty$. In fact, for large T, the distribution of I_p becomes a multiple of a χ^2 distribution with two degrees of freedom, independently of T, so that there is no statistical sense in which I_p converges to $f(\omega_p)$ as T becomes large. As Daniell (1946) had already derived the property that I_p and I_q for $p \neq q$ were asymptotically uncorrelated, these results therefore imply that, for a time series with a continuous spectrum, the traditional periodogram estimator will have a very jumpy and irregular appearance, thus explaining the behaviour of the periodograms calculated in Chapter 3 and also that of the periodogram of Kendall's Series I shown in Figure 8.2, which is a simulated AR(2) process with $a = -1.1$ and $b = 0.5$, so that it has a continuous spectrum with a peak at $\omega_0 = 0.191$ radians. Bartlett (1950, page 1) remarked that problems of this type meant that 'the classical method of searching for periodicities in time series, the so-called periodogram analysis of the series, is useless in many cases', while Jenkins (1961, page 149) was later of the opinion that the misuse of periodogram analysis 'has been responsible for the acceptance of probably more spurious hypotheses than any other statistical or mathematical tool'.

Smoothing the periodogram

13.3 To circumvent these problems, Bartlett (1950) extended a suggestion made by Daniell (1946), that of averaging alternative estimates of the periodogram, by invoking the idea that an average taken from m independent samples would possess the usual sampling property of its fluctuations being proportional to $1/\sqrt{m}$. Specifically, he suggested calculating the sample spectrum over m contiguous portions of length n of the observed series (so that $mn = T$)

and then taking the average across the m subseries. If the subseries periodogram estimates are denoted $I_{p,r}$, $r = 1, 2, \ldots, m$, this leads to the estimator

$$\bar{I}_p = m^{-1}\sum_{r=1}^{m} I_{p,r}$$

for which the variance will be given by $\sigma^4 f^2/m$, which can be made as small as required by taking m large or, equivalently, n small. It can also be shown that, apart from a few end corrections in going from one subseries to the next, $\bar{I}_p$ may be closely approximated by

$$\hat{f}(\omega) = \frac{1}{\pi}\left(c_0 + \sum_{k=1}^{T-1} \lambda_k c_k \cos \omega k\right) \tag{13.3}$$

where $\lambda_k = 1 - k/n$ for $k \leq n$ and $\lambda_k = 0$ for $k > n$. The covariances are thus weighted in the computation of the periodogram, thus leading to the term '*smoothed* periodogram' for this type of estimator.

Figure 13.1 shows the true spectrum of Kendall's Series I along with two smoothed periodograms calculated using (13.3) with n set to 15 and 30, respectively: following Bartlett (1950, Figure 1), the values were calculated for $q = 30\omega/\pi$, corresponding to $m = n\omega/2\pi$. Compared to the unsmoothed periodogram shown in Figure 8.2, the two smoothed periodograms are certainly closer to the true spectrum and a test of goodness of fit calculated by Bartlett

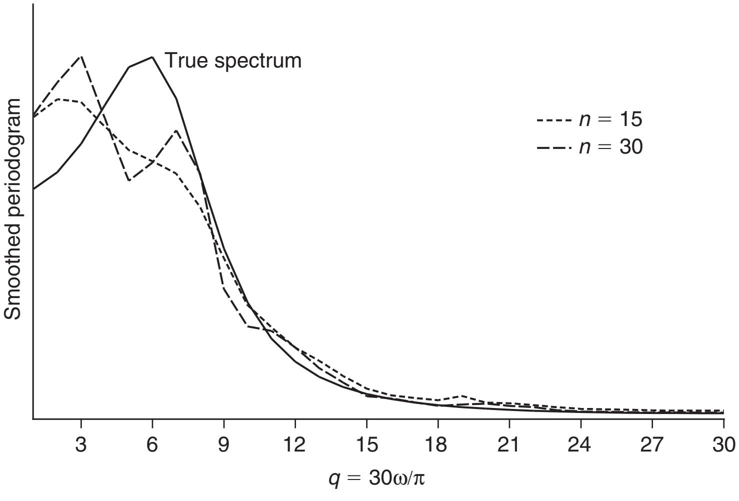


Figure 13.1 Periodogram analysis of Kendall's series I: smoothed periodograms ($n = 15$ and 30) compared with the true spectrum

reveals little evidence of systematic differences. Nevertheless, such calculations still leave open the question of how m and n should be chosen and, indeed, of whether the proposed smoothing function λ_k is the best that is available.

Spectral analysis comes of age

13.4 Unsurprisingly, these questions provoked a major research effort throughout the 1950s, with the theory of estimating continuous spectra being taken forward by, most notably, Grenander and Rosenblatt (1953), Bartlett and Mehdi (1955), Lomnicki and Zaremba (1957, 1959), Parzen (1957, 1958), Whittle (1957), Jenkins and Priestley (1957) and Grenander (1958).

The position at the beginning of the 1960s was summarized by two papers in volume 3 of *Technometrics*, Jenkins (1961) and Parzen (1961), and the subsequent discussions by Tukey (1961) and Goodman (1961). Jenkins' focus was on the physical and statistical aspects of spectral analysis, while Parzen was more concerned with technical and mathematical considerations.

Frequency response functions and filters

13.5 Suppose that $X_t = A\cos\omega t = Ae^{i\omega t}$ is an input into a simple linear system and that the output is characterized as

$$Y_t = G(\omega)A\cos(\omega t + \phi(\omega)) = G(\omega)Ae^{i(\omega t+\phi(\omega))}$$

$G(\omega)$ is referred to as the *gain* and $\phi(\omega)$ as the *phase shift*, both being functions of the frequency ω. The *frequency response function* is then defined as $\psi(\omega) = G(\omega)e^{i\phi(\omega)}$, so that $G(\omega) = |\psi(\omega)|$ and $\phi(\omega) = \arg\psi(\omega)$.

In general, if X_t and Y_t have spectra $f_X(\omega)$ and $f_Y(\omega)$ and variances σ_X^2 and σ_Y^2, then these are related by the gain through the formula

$$\sigma_Y^2 f_Y(\omega) = G^2(\omega)\,\sigma_X^2 f_X(\omega) \tag{13.4}$$

A *linear filter* may be defined as

$$Y_t = \sum_{j=-h}^{h} \lambda_j X_{t+j} \quad \lambda_j = \lambda_{-j}$$

or as $Y_t = G(\omega)Ae^{i\omega t}$, where

$$G(\omega) = \sum_{j=-h}^{h} \lambda_j e^{i\omega t}$$

is the frequency response function of the filter. If $G(\omega)$ is defined to accept only low (high) frequencies then it is referred to as a low (high)-pass filter. If it is allowed to accept a band of intermediate frequencies (so that it is defined as the difference between a low- and a high-pass filter) then it is referred to as a band-pass filter. It also then follows that

$$\sigma_Y^2 = \int_0^\infty f_X(\omega)|G(\omega)|^2 d\omega$$

which shows that the variance of the output is a weighted average, with weights given by the squared gain, over the spectrum of the input.

Nyquist frequency and aliasing

13.6 If the data consist of a continuous *trace* $X(t)$ then it is often read only at discrete intervals ∇t, which will obviously lead to a loss of information. In terms of the spectrum, all information will be lost for frequencies above what is called the *Nyquist frequency* $\omega_N = \pi/\nabla t$, as what is measured at ω_N is not $f(\omega_N)$ but the latter confounded with all frequencies which are indistinguishable from ω_N. In general, if $f^*(\omega)$ is the spectral density corresponding to $X(t)$, then the spectral density of the sampled trace is given by

$$f(\omega) = \sum_{k=0}^{\infty} \left\{ f^*\left(\frac{2\pi k}{\nabla t} + \omega\right) + f^*\left(\frac{2\pi k}{\nabla t} - \omega\right) \right\}$$

This may be interpreted as being obtained by 'folding' the unsampled spectrum about even multiples $2\pi k/\nabla t$ of the Nyquist frequency and then adding these contributions in the range $(0, \omega_N)$, a practice known as *aliasing*. It is clear that, for this to work, $f^*(\omega)$ should be (approximately) zero for $\omega > \omega_N$ and Jenkins (1961, pages 144–5) offered some guidance on how this could be achieved.

Kernels and windows

13.7 Jenkins (1961) showed that (13.3) could, in general, be expressed equivalently as

$$\hat{f}(\omega) = \int_0^\pi I(y)K(\omega, y)dy$$

where

$$K(\omega, y) = \frac{1}{2}(\mu(\omega + y) + \mu(\omega - y)) \quad \mu(y) = \frac{1}{\pi}\sum_{k=-T}^{T} \lambda_k e^{iyk}$$

from which it follows that the *kernel* or *window* $K(\omega, y)$ is such that

$$\int_0^{\pi} K(\omega, y) dy = 1$$

For example, the kernel corresponding to the 'Bartlett weights' $\lambda_k = 1 - k/n$ for $k \leq n$ and $\lambda_k = 0$ for $k > n$ is

$$K(\omega, y) = \frac{1}{\pi n}\left(\frac{\sin^2(n/2)(\omega + y)}{\sin^2 \frac{1}{2}(\omega + y)} + \frac{\sin^2(n/y)(\omega - y)}{\sin^2 \frac{1}{2}(\omega - y)}\right)$$

This kernel has a shape that falls off rapidly from its maximum at the peak frequency $y = \omega$ and reaches zero at $y = \pm\pi/n$, beyond which it oscillates with decreasing amplitude.

Associated with a kernel is its *bandwidth*. Parzen (1961) defined this to be half the base width, $2\pi/n$, of a rectangular kernel which has the same height and same area as $K(\omega, y)$, although other definitions have been suggested. Increasing n thus has the effect of reducing the bandwidth, which increases the 'focusing power' of the kernel and hence decreases the sampling distortion: unfortunately, it will also increase the variance of the estimated spectrum. The trade-off between these considerations led to a variety of kernels being suggested. Writing the weight function as $\lambda(u)$, $u = k/n$, then the Bartlett weights correspond to setting

$$\begin{aligned} \lambda(u) &= 1 - |u|, \quad && |u| \leq 1 \\ &= 0, && |u| > 1 \end{aligned}$$

The 'hanning' estimate of Blackman and Tukey (1958) is

$$\begin{aligned} \lambda(u) &= \tfrac{1}{2}(1 + \cos \pi u), \quad && |u| \leq 1 \\ &= 0, && |u| > 1 \end{aligned}$$

while their 'hamming' estimate is

$$\begin{aligned} \lambda(u) &= 0.54 + 0.46 \cos \pi u \quad && |u| \leq 1 \\ &= 0, && |u| > 1 \end{aligned}$$

A generalization of these two weight functions is

$$\begin{aligned} \lambda(u) &= 1 - 2a + 2a \cos \pi u, \quad && |u| \leq 1 \\ &= 0, && |u| > 1 \end{aligned}$$

in which hanning is obtained by setting $a = 0.25$ and hamming by setting $a = 0.23$. Parzen (1957, 1961) suggested the weight functions

$$\begin{aligned} \lambda(u) &= 1 - u^2, \quad && |u| \leq 1 \\ &= 0, && |u| > 1 \end{aligned}$$

and

$$\begin{aligned}\lambda(u) &= 1 - 6u^2(1 - |u|), & |u| &\leq \tfrac{1}{2}\\ &= 2(1 - |u|)^2, & \tfrac{1}{2} &< |u| \leq 1\\ &= 0, & |u| &> 1\end{aligned}$$

Finally, the Daniell weight function sets $\lambda(u) = \sin u/u$ and so involves no weight truncation. These weight functions and their associated kernels are conveniently tabulated in Jenkins (1961, Table 1) and Parzen (1961, Table I and II).

13.8 The 'design' considerations involved in choosing a weight function/kernel and a bandwidth were discussed in detail in Jenkins (1961), where a convenient summary is to be found, and in Blackman and Tukey (1958), to which readers interested in an engineering perspective are referred, while further extensions were provided by Daniels (1962) and Priestley (1962) and surveyed in Jenkins (1965).

The Fast Fourier Transform

13.9 Because of the ubiquitous nature of the Fourier Transform to spectral analysis, early computation of the spectrum was hampered by the limited computing power then available, particularly when analysing long time series. A major innovation was the development of the Fast Fourier Transform (FFT) by Cooley and Tukey (1965), to which an issue of the IEEE *Transactions on Audio and Electroacoustics* (IEEE, 1967) was devoted, covering the FFT's history of discovery and rediscovery and various of its applications. The algorithm's importance was that, while direct computation of a Fourier transform of a series containing T observations required approximately T^2 operations, the FFT required only $T \log_2 T$, thus constituting a huge saving in computer time. A convenient summary of the FFT algorithm was provided in Jenkins and Watts (1968, Appendix A7.3) and a useful contemporary monograph on the topic is Brigham (1974).

Dealing with nonstationarity

13.10 It must be emphasized that spectral analysis assumes that the underlying series is stationary about a zero mean, so that the usual considerations involved with detrending need to be considered before spectral analysis can be attempted. If the series being analysed has a non-periodic trend component, say $Y_t = \mu(t) + X_t$, where X_t is stationary, then the presence of $\mu(t)$ will introduce a jump in the spectrum at $\omega = 0$. When estimating the spectrum from a sample, it may then be difficult to differentiate between a true trend and a component having a very low frequency, although whether this has any practical implications is perhaps arguable in many applications. Granger and Hatanaka (1964, chapter 8) discussed how trends should be dealt with prior to spectral analysis, although they were unable to come to any firm recommendations, as these would depend upon the type of analysis that was to be undertaken: for example, on whether low-frequency components were of interest or not. Nerlove (1964), however, felt able to propose the use of *prewhitening*, a technique suggested by Blackman and Tukey (1958) for filtering out power at low frequencies prior to computing the spectrum. Nerlove proposed using a filter of the form of repeated 'quasi'-differences $\Delta(k) = x_t - kx_{t-1}$, for which the pth repeat has the gain function $G(\omega) = (1 - 2k\cos 2\pi\omega + k^2)^p$. For $\omega \approx 0$, this function is approximately $(1 - k)^{2p}$ and at frequencies near 0.5 it is approximately $(1 + k)^{2p}$. Setting k to be a positive fraction will thus have the effect of raising power at higher frequencies and lowering the power at low frequencies, rather than annihilating it completely by setting k to unity. Nerlove suggested setting $k = 0.75$ and $p \leq 3$, with the object being to obtain an appropriately flat spectrum. An estimate of the original spectrum may then be obtained by dividing the prewhitened spectrum by the gain function, a process termed by Nerlove as *recoloring*.

13.11 The above considerations suggest that the spectrum fitted to trending series would look similar to that depicted in Figure 13.2. Granger (1966) surveyed the spectra that had been fitted to a wide range of economic time series and suggested that this was the 'typical spectral shape' of an economic time series, *even after a trend had been removed*: 'if one estimates the power spectrum of an economic series containing an important trend, a "typical shape" spectral estimate is likely to result. The important point about the typical spectral shape, however, is that it still appears even if trend in mean is removed' (Granger, 1966, page 154). To Granger, this suggested that 'the long-term fluctuations in economic variables, if decomposed into frequency components, are such that

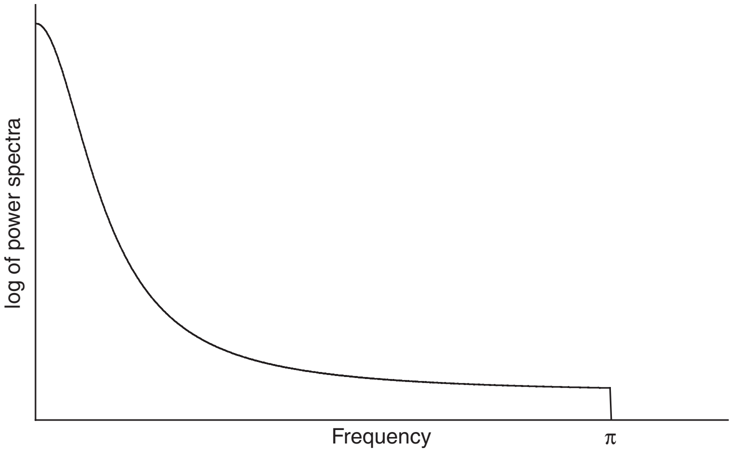


Figure 13.2 Granger's (1966) 'typical spectral shape'

the amplitudes of the components decrease smoothly with decreasing period' or, in other words, 'events which affect the economy for a long period are more important than those which affect it only for a short time' (*ibid.*, page 155).

Cross-spectral analysis

13.12 Suppose that §**13.1** is generalized to the case where there is a bivariate stationary generating process (X_t, Y_t), with spectral representations

$$f_x(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{k=1}^{\infty} \rho_{xx,k} \cos \omega k\right)$$

and

$$f_y(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{k=1}^{\infty} \rho_{yy,k} \cos \omega k\right)$$

There will also be a *cross-spectrum*, defined as

$$f_{xy}(\omega) = c(\omega) + iq(\omega)$$

where $c(\omega)$ and $q(\omega)$ are the *co-spectrum* and *quadrature spectrum*, which obey the 'coherence-inequality'

$$c^2(\omega) + q^2(\omega) \leq f_x(\omega) f_y(\omega)$$

and are defined, using the cross-correlations $\rho_{xy,k}$, as

$$c(\omega) = \frac{1}{\pi}\left(1 + \sum_{k=1}^{\infty} (\rho_{xy,k} + \rho_{yx,k}) \cos k\omega\right) \tag{13.5}$$

and

$$q(\omega) = \frac{1}{\pi}\sum_{k=1}^{\infty} (\rho_{xy,k} - \rho_{yx,k}) \sin k\omega \tag{13.6}$$

Granger and Hatanaka (1964, chapter 5) provided an interpretation of these concepts. If X_t and Y_t are real, then they have *Cramér representations* (Cramér, 1940)

$$\begin{aligned} X_t &= \int_{-\pi}^{\pi} e^{it\omega} dz_x(\omega) = \int_0^{\pi} \cos t\omega\, du_x(\omega) + \int_0^{\pi} \sin t\omega\, dv_x(\omega) \\ Y_t &= \int_{-\pi}^{\pi} e^{it\omega} dz_y(\omega) = \int_0^{\pi} \cos t\omega\, du_y(\omega) + \int_0^{\pi} \sin t\omega\, dv_y(\omega) \end{aligned} \tag{13.7}$$

where the $dz_x(\omega)$, etc., are random and uncorrelated processes. Each process can thus be represented by the integral over all frequencies in $0 \leq \omega \leq \pi$, with each frequency being decomposed into two components $\pi/2$ out of phase with each other. Each of these components has a random amplitude, $du_x(\omega)$, etc., and Granger and Hatanaka showed that for each process the amplitudes are uncorrelated not only between the components for any particular frequency but also with the random amplitudes of the components for all other frequencies. The random amplitudes for frequency ω_1 for one process are also uncorrelated with the frequencies, other than ω_1, of the other process. Consequently, only the relationships between a particular frequency in one process and the *same* frequency in the other process need to be considered.

Granger and Hatanaka also showed that

$$E(du_x(\omega)du_y(\omega)) = E(dv_x(\omega)dv_y(\omega)) = 2c(\omega)d\omega$$

and

$$\begin{aligned} E(du_x(\omega)dv_y(\omega)) &= 2q(\omega)d\omega \\ E(du_y(\omega)dv_x(\omega)) &= -2q(\omega)d\omega \end{aligned}$$

Thus (twice) the co-spectral density gives the covariance between the components that are ‘in phase’, while (twice) the quadrature spectral density gives the covariance between the components that are ‘in quadrature’ (i.e., $\pi/2$ out of phase). If $q(\omega)=0$ ($\neq 0$) the components of the two processes at frequency ω are exactly in (out of) phase with each other, while if $c(\omega)=0$ ($\neq 0$) the two processes at frequency ω are uncorrelated (correlated).

Coherence, phase and gain

13.13 To measure the correlation between the frequency components of the two processes, the *coherence at ω* is used:

$$0 \leq C(\omega) = \frac{c^2(\omega) + q^2(\omega)}{f_x(\omega)f_y(\omega)} \leq 1$$

$C(\omega)$ is analogous to the square of the correlation coefficient between two samples and may be interpreted in a similar way: the larger is $C(\omega)$, the more closely related are the two components at frequency ω. The *gain* is then defined as (cf. equation (13.4))

$$G^2(\omega) = C(\omega)\frac{f_x(\omega)}{f_y(\omega)}$$

A plot of $C(\omega)$ against ω over $0 \le \omega \le \pi$ is called the coherence diagram.

If the two series are given by $Y_t = a_t \cos(\omega t + \theta)$ and $X_t = a_t \cos(\omega t + \varphi)$, with $\theta > \varphi$, $\psi = \theta - \varphi$ is termed the *phase-difference* (or *phase-lag*) at frequency ω. In general, a measure of the phase difference between two frequency components is given by

$$\psi(\omega) = \tan^{-1}\left(\frac{q(\omega)}{c(\omega)}\right)$$

and the plot of $\psi(\omega)$ against ω is called the phase diagram, with $\psi(\omega)/\omega$ measuring the extent of the time lag. In general, two series that differ only in phase will have a coherency of unity.

Estimating the cross-spectrum

13.14 Estimates of the cross-spectrum can be obtained by analogy to and extension of the smoothed periodogram (13.3). Thus, from (13.5) and (13.6), estimates of the co-spectrum and the quadrature spectrum at frequency ω_j, $j = 0, 1, \ldots, m$, are given by

$$\begin{aligned} \hat{c}(\omega_j) &= \frac{1}{\pi}\left(c_{xy,0} + \sum_{k=1}^{m-1} \lambda_k (c_{xy,k} + c_{yx,k}) \cos \omega_j k\right) \\ \hat{q}(\omega_j) &= \frac{1}{\pi}\sum_{k=1}^{m-1} \lambda_k (c_{xy,k} - c_{yx,k}) \cos \omega_j k \quad \omega_j = \frac{\pi j}{m} \end{aligned}$$

$c_{xy,k}(c_{yx,k})$ being the sample covariance between X_t and Y_{t-k} (Y_t and X_{t-k}). The sample coherence is then given by

$$\hat{C}(\omega_j) = \frac{\hat{c}^2(\omega_j) + \hat{q}^2(\omega_j)}{\hat{f}_x(\omega_j)\hat{f}_y(\omega_j)}$$

where $\hat{f}_x(\omega_j)$ and $\hat{f}_y(\omega_j)$ are the sample spectra of X and Y at frequency ω_j (cf. equation (13.3)). Granger and Hatanaka (1964, chapter 5.2) gave the distribution of $u = \sqrt{\hat{C}(\omega)}$ under the null $C(\omega) = 0$:

$$F(u) = 1 - (1 - u^2)^{T/m-1}$$

and provided critical values. They also provided confidence intervals for the estimated phase angle $\hat{\phi}(\omega) = \tan^{-1}(\hat{q}(\omega)/\hat{c}(\omega))$.

13.15 Granger and Hatanaka (1964) provided the first applications of cross-spectral analysis and the underlying theoretical framework was further advanced in the treatment by Jenkins and Watts (1968, chapters 8–11).

The partial cross-spectrum

13.16 Paralleling the use of partial correlation coefficients, partial cross-spectra may be defined to help in assessing the spectral relationships between sets of time series. Granger and Hatanaka (1964, chapter 5.8) thus considered the set of M stationary series $(X_{1i}, X_{2t}, \ldots, X_{Mt})$, each series having its own (auto) spectra, $f_{ii}(\omega)$, as well as there being a set of cross-spectra, $f_{ij}(\omega)$, $i, j = 1, \ldots, M$, noting that these will typically be complex quantities. The matrix of these spectra, $\Sigma(\omega)$, was regarded by Granger and Hatanaka (*ibid.*, page 91: italics in original) as '*estimating the covariance matrix of the time series around frequency* ω, the term "around" being deliberately chosen as a reminder that spectral estimates are estimates of an average over a frequency band'.

Concentrating on the partial cross-spectrum between $X_1(\omega)$ and $X_2(\omega)$, these being the components of X_{1t} and X_{2t} around frequency ω, consider the following partition of the cross-spectral matrix

$$\Sigma(\omega) = \left[\begin{array}{cc|cccc} f_{11}(\omega) & f_{12}(\omega) & f_{13}(\omega) & \ldots & f_{1M}(\omega) \\ f_{21}(\omega) & f_{22}(\omega) & f_{23}(\omega) & \ldots & f_{2M}(\omega) \\ \hline f_{31}(\omega) & f_{32}(\omega) & f_{33}(\omega) & \ldots & f_{3M}(\omega) \\ \vdots & \vdots & \vdots & & \vdots \\ f_{M1}(\omega) & f_{M2}(\omega) & f_{M3}(\omega) & \ldots & f_{MM}(\omega) \end{array}\right] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

and define the matrix

$$\Sigma_{12\cdot k}(\omega) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{bmatrix} f_{11\cdot k}(\omega) & f_{12\cdot k}(\omega) \\ f_{21\cdot k}(\omega) & f_{22\cdot k}(\omega) \end{bmatrix}$$

where k denotes the set $3,4,\ldots,M$. This is the partial cross-spectral matrix for X_{1t} and X_{2t} and from it the definitions of the partial coherence and partial phase angle follow naturally:

$$C_{12\cdot k}(\omega) = \frac{\left|f_{12\cdot k}(\omega)\right|^2}{f_{11\cdot k}(\omega) f_{22\cdot k}(\omega)} \qquad \psi_{12\cdot k}(\omega) = \frac{\text{Imaginary part of } f_{12\cdot k}(\omega)}{\text{Real part of } f_{12\cdot k}(\omega)}$$

These concepts have the following interpretation. Suppose that an optimum linear combination of the series $X_{3t}, X_{4t}, \ldots, X_{Mt}$ has been subtracted from X_{1t} and X_{2t} to form $\hat{X}_{1t}$ and $\hat{X}_{2t}$ (how this optimum combination might be arrived at is discussed in §**13.18**). $f_{11\cdot k}(\omega)$ will thus be the spectrum of $\hat{X}_{1t}$, and $C_{12\cdot k}(\omega)$ and $\psi_{12\cdot k}(\omega)$ will be the coherence and phase angle, respectively, between $\hat{X}_{1t}$ and $\hat{X}_{2t}$.

As a simple example, consider a three variable set of series X_{1t}, X_{2t} and X_{3t}. If X_{1t} and X_{3t} are related for all frequencies and X_{2t} and X_{3t} are also related,

> there is no reason why X_{1t} and X_{2t} should be. If, for example, X_{1t} were sale of ice cream, X_{2t} sale of air conditioners, and X_{3t} was a temperature series, then the coherence between X_{1t} and X_{3t} and between X_{2t} and X_{3t} would probably be large for many frequencies. The coherence between X_{1t} and X_{2t} might also be large but this would be a spurious relationship, as X_{1t}, X_{2t} are connected only via X_{3t}. In such a case the *partial* coherence between X_{1t}, X_{2t} ought to be zero (in theory) or small (in practice) for all frequencies. (Granger and Hatanaka, 1964, pages 92–3)

Cross-spectral analysis, feedback and causality

13.17 The concepts introduced in §§**13.14–13.16** were utilized by Granger and Hatanaka (1964, chapter 7) and Granger (1963, 1969) to develop one of the most enduring concepts in time series analysis. Through the phase-lag and coherence, cross-spectral methods provide a useful way of describing the relationship between two (or more) variables when one is leading in time, so 'causing' (in a very precise way to be defined below) the other(s). Suppose X_t and Y_t have the Cramér representations

$$X_t = \int_{-\pi}^{\pi} e^{it\omega} dz(\omega) = \int_0^{\pi} \cos t\omega \, du_x(\omega) + \int_0^{\pi} \sin t\omega \, dv_x(\omega)$$

and

$$\begin{aligned} Y_t &= \int_{-\pi}^{\pi} e^{it\omega} a(\omega) e^{-i\Phi\omega} dz(\omega) = a(\omega) \int_0^{\pi} \cos t\omega\Phi(\omega) du_x(\omega) \\ &\quad + a(\omega) \int_0^{\pi} \sin t\omega \, \Phi(\omega) dv_x(\omega) \end{aligned}$$

where $\Phi(\omega) = \phi(\omega)$, $\omega > 0$ and $\Phi(0) = 0$. The spectrum of Y_t is then given by $f_y(\omega) = a^2(\omega) f_x(\omega)$ and the relationship between the two series can be expressed as

$$Y_t = X_t(a(\omega), \phi(\omega)) + U_t \tag{13.8}$$

where U_t is some stationary series such that $C_{xu} = 0$, so that

$$0 < C_{yx} = \frac{a^2 f_x(\omega)}{f_y(\omega)} < 1$$

If, as well as (13.8),

$$X_t = Y_t(b(\omega), \theta(\omega)) + V_t \tag{13.9}$$

where V_t has similar properties to U_t, then there is said to be *feedback* between X_t and Y_t. In the presence of feedback the phase diagram is unlikely to provide much useful information as no one process continually lags the other.

13.18 To provide a formal definition of feedback from which tests may be developed, Granger (1969) set up the following framework. Suppose, in general, that A_t is a stationary stochastic process and that $A(k) = \{A_{t-k}, A_{t-k-1}, \ldots\}$. Then $\overline{A} = A(1)$ and $\overline{\overline{A}} = A(0)$ represent the sets of *past* and *past and present* values of A_t. The optimum, unbiased, least squares predictor of A_t using the set of values B is denoted $P_t(A|B)$, so that $P_t(X|\overline{X})$ is the optimum predictor of X_t using only past values of X_t. The predictive error series is then denoted $\varepsilon_t(A|B) = A_t - P_t(A|B)$, with variance $\sigma^2(A|B)$. Let I_t be all the information in the universe accumulated since time $t-1$ and let $I_t - Y_t$ denote all this information *apart* from the specified series Y_t. Granger then introduced the following definitions.

Causality

If $\sigma^2(X|\overline{I}) < \sigma^2(X|\overline{I} - \overline{Y})$ then Y is said to cause X, denoted $Y \Rightarrow X$: X_t is better able to be predicted using all available past information than if the information apart from past Y had been used.

Feedback

If $\sigma^2(X|\overline{I}) < \sigma^2(X|\overline{I} - \overline{Y})$ and $\sigma^2(Y|\overline{I}) < \sigma^2(Y|\overline{I} - \overline{X})$ then feedback is said to occur, denoted $Y \Leftrightarrow X$: feedback thus occurs when Y causes X and, at the same time, X causes Y.

Instantaneous causality

If $\sigma^2(X|\overline{I}, \overline{\overline{Y}}) < \sigma^2(X|\overline{I})$ instantaneous causality is occurring, denoted $Y_t \Rightarrow X_t$: X_t is better predicted if the current value of Y is included in the prediction than if it is not.

Causality lag

If $Y \Rightarrow X$, the causality lag m is defined to be the least value of k such that $\sigma^2(X|\overline{I} - Y(k)) < \sigma^2(X|\overline{I} - Y(k+1))$: knowing the values $Y_t, Y_{t-1}, \ldots, Y_{t-m+1}$ is of no help in improving the prediction of X_t.

The assumption that only stationary series are involved ensures that prediction variances remain constant. If nonstationarity is allowed such variances would depend upon time, implying that the existence of causality could alter over time.

Granger argued that the unrealistic use of the universal information set I could easily be modified so that it was defined to contain only those series that are relevant. For example, if it is restricted to just the two series X_t and Y_t then

$Y \Rightarrow X$ if $\sigma^2(X|\overline{X}) > \sigma^2(X|\overline{X}, \overline{Y})$. Use of restricted data sets opens up the possibility of *spurious causality* in a way analogous to that of spurious correlation: if a third series, Z_t, is actually causing both X_t and Y_t, but is omitted from the analysis, spurious causality patterns may result. Spurious instantaneous causality is another possibility when the sampling interval is greater than the causality lag.

13.19 In practice linear predictors will tend to replace optimum predictors in these definitions and it might be argued that the prediction error variance is not always the appropriate criterion to employ, although it is natural to use it in connection with linear predictors. Granger suggested that 'causality in mean' might be a more accurate term in these circumstances.

13.20 These definitions of feedback and causality have implications for the cross-spectrum between X_t and Y_t and the related measures of coherence and phase. Using the notation of Chapter 12, suppose these series are generated by the bivariate process

$$\begin{aligned} X_t &= \sum_{j=1}^{p} a_j X_{t-j} + \sum_{j=1}^{p} b_j Y_{t-j} + \varepsilon_t = a(B)X_t + b(B)Y_t + \varepsilon_t \\ Y_t &= \sum_{j=1}^{p} c_j X_{t-j} + \sum_{j=1}^{p} d_j Y_{t-j} + \eta_t = c(B)X_t + d(B)Y_t + \eta_t \end{aligned} \qquad (13.10)$$

where ε_t and η_t are two uncorrelated white noises with variances σ_ε^2 and σ_η^2 respectively. From the definitions above, $Y \Rightarrow X$ if some b_j is not zero, while $X \Rightarrow Y$ if some c_j is not zero. Using Cramér representations, the lag polynomial $a(B)X_t$ in (13.10), for example, can be written as

$$a(B)X_t = \int_{-\pi}^{\pi} e^{it\omega} a(e^{-i\omega})\, dz_x(\omega)$$

so that (13.10) can be written

$$\int_{-\pi}^{\pi} e^{it\omega}((1 - a(e^{-i\omega}))dz_x(\omega) - b(e^{-i\omega})dz_y(\omega) - dz_\varepsilon(\omega)) = 0$$

$$\int_{-\pi}^{\pi} e^{it\omega}(-c(e^{-i\omega})dz_x(\omega) + (1 - d(e^{-i\omega}))dz_y(\omega) - dz_\eta(\omega)) = 0$$

From this representation, Granger (1969) showed that the spectra of X_t and Y_t are given by

$$\begin{aligned} f_x(\omega) &= \frac{1}{2\pi\Delta}(|1 - d|^2\sigma_\varepsilon^2 + |b|^2\sigma_\eta^2) \\ f_y(\omega) &= \frac{1}{2\pi\Delta}(|c|^2\sigma_\varepsilon^2 + |1 - a|^2\sigma_\eta^2) \end{aligned}$$

in which a is written for $a(e^{-i\omega})$, etc., and where $\Delta = |(1-a)(1-d)-bc|^2$. The cross-spectrum takes the form

$$C(\omega) = \frac{1}{2\pi\Delta}((1-d)c\sigma_\varepsilon^2 + (1-a)b\sigma_\eta^2) = C_1(\omega) + C_2(\omega)$$

where

$$C_1(\omega) = \frac{\sigma_\varepsilon^2}{2\pi\Delta}(1-d)c \quad C_2(\omega) = \frac{\sigma_\eta^2}{2\pi\Delta}(1-a)b$$

Thus, if Y_t is *not* causing X_t then $b=0$ and $C_2(\omega)$ vanishes and, similarly, if X_t is *not* causing Y_t then $c=0$ and $C_1(\omega)$ vanishes. Hence the cross-spectrum may be decomposed into the sum of two components: $C_1(\omega)$, depending upon the causality of Y by X, and $C_2(\omega)$, depending on the causality of X by Y. In general, these may be treated separately and coherences can be defined for $X \Rightarrow Y$ and $Y \Rightarrow X$: for example, the *causality coherence*,

$$C_{\vec{xy}}(\omega) = \frac{|C_1(\omega)|^2}{f_x(\omega)f_y(\omega)} = \frac{\sigma_\varepsilon^4|(1-d)c|^2}{(\sigma_\varepsilon^2|1-d|^2 + \sigma_\eta^2|b|^2)(\sigma_\varepsilon^2|c|^2 + \sigma_\eta^2|1-a|^2)}$$

may be considered to be the strength of the causality $X \Rightarrow Y$ at frequency ω. Similarly,

$$\phi_{\vec{xy}}(\omega) = \tan^{-1}\frac{\text{imaginary part of}\, C_1(\omega)}{\text{real part of}\, C_1(\omega)}$$

will measure the phase lag at frequency ω of $X \Rightarrow Y$. Similar functions can be defined for $Y \Rightarrow X$ using $C_2(\omega)$.

Instantaneous causality may be allowed for by including the terms b_0Y_t and c_0X_t in the respective equations in the representation (13.10). The cross-spectrum is then given by

$$C(\omega) = \frac{1}{2\pi\Delta'}((1-d)(c+c_0)\sigma_\varepsilon^2 + (1-a)(b+b_0)\sigma_\eta^2) = C_1'(\omega) + C_2'(\omega) + C_3'(\omega)$$

where $\Delta' = |(1-a)(1-d)-(b+b_0)(c+c_0)|^2$, $C_1'(\omega)$ and $C_2'(\omega)$ are defined as $C_1(\omega)$ and $C_2(\omega)$ but using Δ' rather than Δ, and

$$C_3(\omega) = \frac{1}{2\pi\Delta'}(c_0(1-d)\sigma_\varepsilon^2 + b_0(1-a)\sigma_\eta^2)$$

The presence of instantaneous causality clearly means that the measures of causal strength and phase lag lose their distinct interpretations.

Granger (1969) provided an illustrative example to show the potential usefulness of these definitions and also considered extensions to more than two variables. However, an estimation and testing methodology for causal cross spectra was not presented and the importance of what was later to be termed 'Granger causality' had to wait until a time domain approach to estimation and testing was developed, as is discussed in §**16.17**.

W(h)ither spectral analysis?

13.21 By the end of the 1960s major developments had thus been made in both the theoretical foundations and the computational aspects of spectral analysis. Yet, as Jenkins (1965, page 2) was able to remark, '(i)n no sense ... can it be said that spectral analysis is widely used or understood by statisticians'. Most of the applications of spectral analysis had been made by physicists and engineers, possibly because of the 'genuine difficulties which statisticians (as opposed to physicists and engineers) face in thinking in terms of frequency concepts' (but see the discussion of seasonal adjustment procedures in Chapter 14). Jenkins argued that the advantages of spectral analysis were: (i) that it was able to convey a great deal of visual information about the underlying process generating the data and hence could suggest potentially useful models; and (ii) that the basic gain relationship (13.4), which shows that the output of a linear system has a spectrum which is the spectrum of the input multiplied by a factor that is proportional to the squared gain of the system, may be used as an aid in system and experimental design.

The major disadvantages of spectral analysis were its non-parametric nature, which necessitated fitting either a whole function or a very large set of parameters, with a corresponding loss of efficiency, and its reliance on the assumption of stationarity. By the end of the 1960s it was far from clear whether spectral techniques would find a place in the mainstream of time series analysis or whether they would remain in a backwater helping engineers and physicists to design physical systems by experimentation.

14
Tackling Seasonal Patterns in Time Series

Early interest in seasonal fluctuations

14.1 Seasonal patterns in economic and meteorological time series were first investigated in the middle of the nineteenth century, with Gilbart (1854), Babbage (1856) and Jevons (1866) all uncovering seasonal fluctuations in currency data, but an, albeit informal, definition of seasonality had to wait a further half-century until Persons (1919, page 18): '(b)y seasonal movement is meant a consistent variation from one month to the next. Are the items for certain months of the year systematically or regularly different from the items for other months? If so, there is a seasonal variation.'

Persons' main intention as editor of the newly published *Review of Economic Statistics*, from which the above quote is taken, was to construct indices of business conditions. Clearly, many of the extant business statistics contained marked seasonal patterns, so that these indices needed to take such patterns into account:

> (t)he object of the present study of seasonal fluctuations is, first, to determine the existence of such fluctuations in various series of monthly data; second, if fluctuations exist, to measure them; and third, to correct the items for seasonal movement. … That many series of business statistics consisting of monthly or quarterly items present marked seasonal variations is recognized both by producers and by consumers of such statistics. That such seasonal variations must be taken into account if we are to use the data as indices of business conditions is also recognized. (*ibid.*, pages 18–19)

Indices of seasonal variation

14.2 A prior attempt at constructing a business conditions index had actually been made by Copeland, who dealt with seasonal fluctuations by 'dividing the actual figure for the month by the average for that month during the ten preceding years', since 'by using the ten-year monthly averages, seasonal fluctuations are automatically allowed for' (Copeland, 1915, pages 554, 556). Even earlier, Kemmerer (1910) had constructed indices by using both simple monthly averages and monthly means of what were, effectively, annual ranks.

Persons was not impressed by any of these approaches, preferring to use what he termed the *link-relative* method, which he described in detail in Persons (1923). Given a monthly time series Y_t, the link relative is simply the ratio of each observation to the previous observation, Y_t/Y_{t-1}. The medians of the monthly link relatives are then denoted $r_1, r_2, \ldots, r_{12}$, where r_1 is the median January link relative, etc. Next, the *initial chain relatives* are obtained as the sequence $c_2 = 100r_2$, $c_3 = c_2r_3, \ldots, c_{12} = c_{11}r_{12}$. By circularity, the January chain relative will then be $c_1 = c_{12}r_1$, but this will not, in general, be equal to 100, as is implied by the sequence above. This 'discrepancy' is assumed to be distributed across the chain relatives using the factor $1+d$, obtained by solving $100(1+d)^{12} = c_{12}r_1$, thus leading to the *adjusted chain relatives*

$$\frac{c_{12}r_1}{(1+d)^{12}} = 100, \frac{c_2}{1+d}, \frac{c_3}{(1+d)^2}, \ldots, \frac{c_{12}}{(1+d)^{11}}$$

These relatives are then taken as the index of seasonal variation (after any rescaling to ensure that the factors average to 100), which is constant across years and, if denoted S_t, has the property that $S_t = S_{t-12}$. 'Seasonal adjustment' can then be carried out: '(i)f it be desired to correct the original items for seasonal variation ..., the form Y_t/S_t may be used, the result being expressed in terms of the units of the original series' (Persons, 1923, page 722).

14.3 Contemporaneously with Persons, statisticians at the Federal Reserve Board and the National Bureau of Economic Research (NBER), led by Frederick Macaulay, were developing an extension of the Copeland approach. Rather than dividing Y_t by what is, in effect, $\sum_{j=1}^{10} Y_{t-12j}/10$, they preferred to use the divisor $\sum_{j=-6}^{5} Y_{t-j}/12$, i.e., a 12-month moving average 'centred' on the seventh month. For each month a 'central value' of this ratio was then computed, the exact manner of doing this depending on the length of the series, often being a trimmed mean using the four middle values when ranked by size. These twelve 'average ratios' were then adjusted to total 1,200 by 'prorating the difference between their sum and 1,200 in proportion to the size of the ratio' (Joy and Thomas, 1928, page 245). According to Joy and Thomas, the resulting 'relatives' were then treated as preliminary seasonal factors which 'were then inspected to determine whether they gave evidence in any way of accidental variations of a non-seasonal character' (*ibid.*, page 246). This approach was termed the *ratio-to-moving average* index of seasonal variation by Joy and Thomas.

14.4 These two methods of seasonal adjustment are applied to the series shown in Figure 14.1, which is Series G from Box and Jenkins (1970), originally

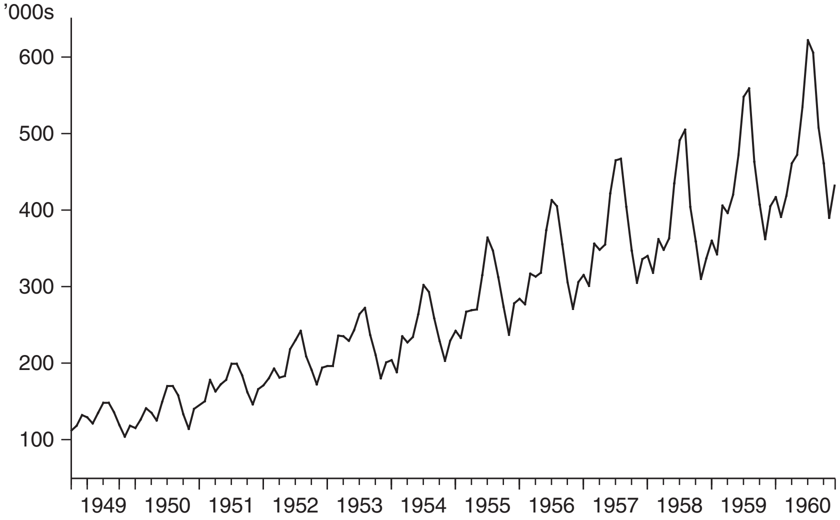


Figure 14.1 Series G from Box and Jenkins (1970): international airline passengers (in thousands), monthly, 1949–1960

provided by Brown (1963). These are monthly observations from 1949 to 1960 on international airline passengers and have become a stock series for analysing seasonality, often being referred to as the 'airline data'. The seasonal factors obtained from the two methods are as follows:

	Jan.	**Feb.**	**Mar.**	**Apr.**	**May**	**June**	**July**	**Aug.**	**Sep.**	**Oct.**	**Nov.**	**Dec.**
Link-relative	91.7	88.0	99.9	96.6	97.5	111.1	123.7	122.5	106.2	92.6	80.4	89.8
Ratio-to-ma	91.7	89.1	101.5	98.4	98.9	112.2	123.6	123.0	106.9	92.9	80.8	90.6

Although there are some minor differences in the factors, overall the seasonal patterns are very similar, as one might expect for a series with a very pronounced seasonal fluctuation and little volatility. Note that the amplitude of the seasonal pattern increases with the trend, so that constructing seasonal factors based on ratios is more appropriate than constructing them using differences. The two seasonally adjusted series are shown in Figure 14.2 and, unsurprisingly given the similarity of the seasonals, they exhibit only minor differences, although that obtained from the link-relative method has a tendency to be slightly larger than that obtained from the ratio-to-moving average method.

14.5 The introduction of these two methods provoked a great deal of interest, with numerous refinements, extensions and critiques being proposed, all of which were surveyed in very systematic fashion by Mendershausen (1937).[1]

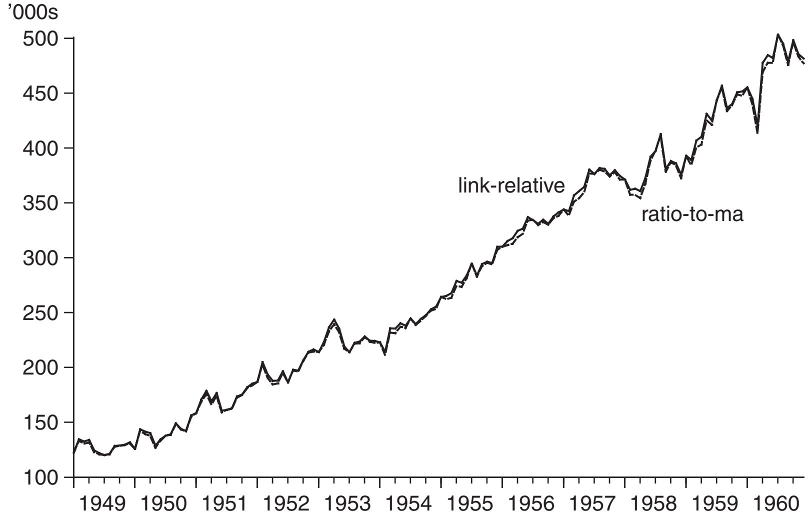


Figure 14.2 Seasonally adjusted airline passenger miles using the link-relative (—) and ratio-to-moving average (- - -) methods

For all these efforts, however, these two basic approaches, particularly the ratio-to-moving average, quickly became the 'industry standard' for seasonally adjusting economic time series: see, for example, Barton (1941) and Burns and Mitchell (1946, pages 43–55), which was the recognized reference on business cycles at the time.

Evolving seasonality and causal explanations

14.6 A major aim of many of the extensions surveyed by Mendershausen was to propose ways of allowing the seasonal factors to evolve over time, rather than remaining fixed as in the basic methods. Unfortunately, Mendershausen (1937, pages 252–3) was forced to conclude that 'the results of these methods in measuring variable seasonal movements are no more significant than those obtainable by a much simpler mechanical method; and that the proposed measures of changes in amplitude or pattern would be useful only *after* the variable seasonal had been exactly determined'.

Once evolving seasonality is entertained, causal explanations of seasonal fluctuations become of interest. Mendershausen surveyed the embryonic attempts to provide such causal explanations, typically by relating seasonal amplitudes to business cycle conditions and/or weather variables, such as temperature, concluding that such studies 'are tending in the right direction, i.e., to measure the changes of the seasonal movement as determined by changes in their causes. It would seem that a *general method* of dealing with this problem must proceed along these lines' (*ibid.*, page 262). Suffice to say that it was many years before much progress was to be made in this area: see Hylleberg (1992) for several notable attempts and further references.

Electronic computations and the Census seasonal adjustment programmes

14.7 The mechanical nature of the calculations involved in the ratio-to-moving average method had one important advantage: with the advent of electronic computers in the early 1950s, these calculations could be straightforwardly programmed and computed on the new machines. Such programmes for seasonally adjusting time series were developed by the US Bureau of the Census, improved and extended by the NBER, and made available to other organizations. These efforts were described in Eisenpress, McPherson and Shiskin (1955), Shiskin (1955) and Shiskin and Eisenpress (1957), and led directly to the Census Methods I and II seasonal adjustment programs.

The basic approach of Method I was to consider a *multiplicative decomposition* of the original monthly time series Y, this being $Y = TC \times S \times I$, where TC, S and I are the trend-cycle, seasonal and irregular components, respectively. The trend-cycle component is first estimated by a 12-month moving average of Y and this estimate is then divided into Y to obtain an estimate of $S \times I = Y/TC$. For each month, a moving average is then fitted to the $S \times I$ component for that month in successive years to obtain estimates of the seasonal factor, S, alone. The 12 seasonal factors for each year are then 'centred' so that their sum equals 1,200: note that the method automatically allows for evolving seasonal factors throughout the series. An iterative procedure is then used for seasonal adjustment. First, the seasonal factors are divided into the original observations to obtain a preliminary seasonally adjusted series, $Y^a = TC \times I/S$. This series is then smoothed by a five-month moving average to provide a more flexible trend-cycle estimate and the above procedure repeated to obtain a new set of seasonal factors and a final seasonally adjusted series.

Method II followed the general procedure of Method I but made further refinements, based on the experiences of using Method I, which could be afforded by the increasing computational capacity of electronic computers. The main change was that the five-month moving average used in the iterative stage was replaced by a Spencer 15-term moving average (see §**10.3**). Because of the loss of seven observations at the start and end of the sample period brought about by using the Spencer moving average, these missing values were estimated using an average of the first (last) four months of the preliminary seasonally adjusted series. Further improvements were also made to the methods used to obtain the seasonal factors and to isolate extreme ratios that might unduly influence these factors. Method II also contained methods for dealing with trading day variations within a month and also provided several tests of the efficacy of the seasonal adjustments (for a full description of the two methods, see Shiskin and Eisenpress, 1957).

14.8 Over the next decade several variants of Census Method II were developed, these being identified with the letter 'X' and a sequence number, beginning with X-3. In October 1965 the X-11 variant became the standard seasonal adjustment programme. This was described in detail in Shiskin, Young and Musgrave (1967), where it was summarized thus.

> (X-11) includes several improvements over earlier versions. Several of the new features in X-11 provide additional tools for the time series analyst. While the computations in the standard program are sufficient for most applications, the analyst can select optional features peculiar to his own needs. For example, he may choose between the additive and multiplicative versions and between the full seasonal-adjustment routine and one limited to the calculation of summary measures computed from seasonally adjusted data obtained from other sources; the σ limits for identifying extreme values may be varied, providing for contingencies such as strikes; and he may specify the moving averages to be used in estimating the trend-cycle and seasonal components. ... As a result of the availability of these options, X-11 is an instrument, not only for the massive seasonal adjustment of time series, but also for seasonally adjusting unusual series, for research into new techniques of time series analysis, and for studies of the relations among different types of fluctuations. (*ibid.*, page 1)

A particularly interesting extension was the replacement of the preliminary 12-month moving average for estimating the trend-cycle component with a (Whittaker–)Henderson 13-term moving average (§**10.6**) and the replacement of the Spencer 15-term moving average in the iterative stage with Henderson moving averages whose orders were based on the ratio of the average absolute month-to-month change in the irregular to that in the trend-cycle (known as the $\overline{I}/\overline{C}$ ratio). X-11 was subsequently made available to other users and was adopted by many statistical agencies throughout North America and later across the world.[2]

14.9 An interesting analysis of X-11 was provided by Young (1968). The multiplicative decomposition $Y = TC \times S \times I$ has, of course, an equivalent additive decomposition in the logarithms of Y and its components, i.e., $y = tc + s + i$. Young showed that these log-components could each be represented as a combination of linear and nonlinear operations on y. The linear operations could be expressed as weighted moving averages, with the weights being obtained from the various moving averages used in Method II: to be precise, s could be estimated by a 145-term moving average and tc and i by 157-term moving averages.

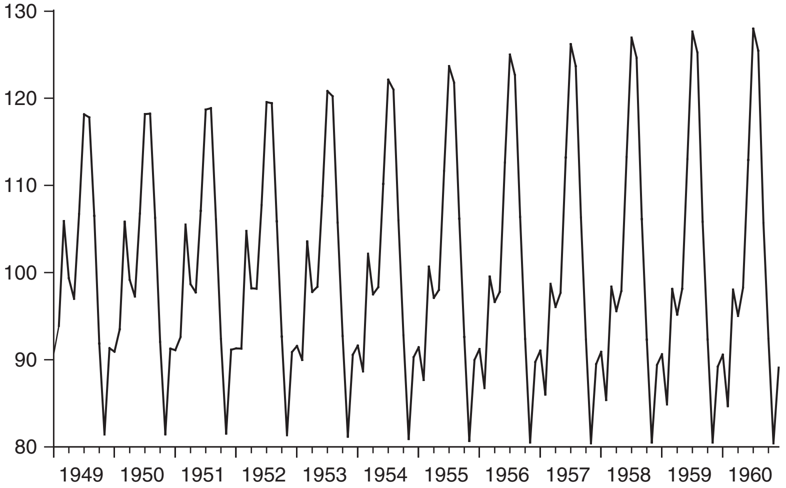


Figure 14.3 X-11 seasonal factors for the airline data

The nonlinearities arose chiefly through X-11's requirement that 12-month sums of the seasonally adjusted and adjusted data should be equal, rather than their products being equal, as would seem logical for a multiplicative decomposition. However, for many purposes the effect of these nonlinear operations on the seasonal factors turned out to be negligible, so that ratio-to-moving average based methods, such as X-11 and the BLS approach, could be closely approximated by weighted linear moving averages to yield the estimates of the underlying components. The weight functions for various choices of the underlying Henderson moving averages were given in Young (1968).

14.10 The seasonal weights for the airline data are shown in Figure 14.3 and they evolve gradually over the sample period with increasing amplitude, reflecting the trend movement of the series shown in Figure 14.1. The X-11 seasonally adjusted series is compared with that obtained by using the classic ratio-to-moving average method in Figure 14.4. Although both series are almost identical in the early years of the sample, the divergence between them increases over time.

Unobserved component and spectral approaches to seasonal adjustment

14.11 In a sequence of papers in the early 1960s, Hannan (1960, 1963, 1964) examined seasonality from an unobserved component (UC) perspective. Recalling Muth's (1960) UC decomposition of §**11.19**, Hannan formally considered

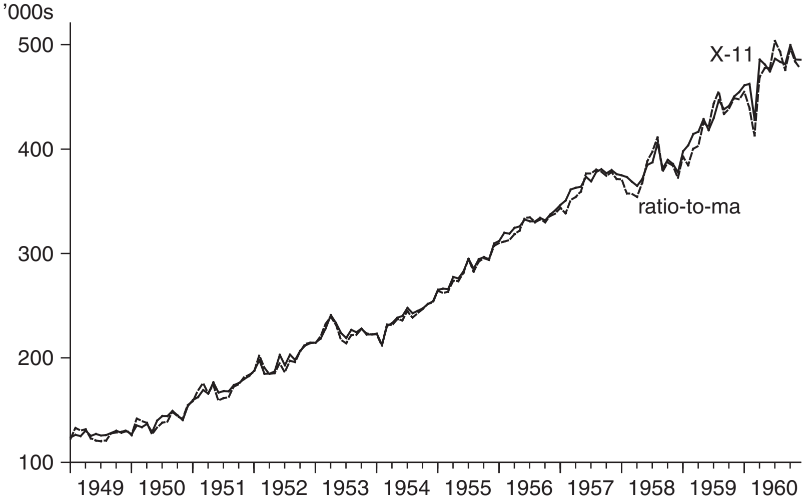


Figure 14.4 Seasonally adjusted airline passenger miles using the X-11 (—) and ratio-to-moving average (- - -) methods

the decomposition used in §**14.9**:

$$y_t = tc_t + s_t + i_t = s_t + z_t \tag{14.1}$$

where the component $z_t = tc_t + i_t$ is the 'remainder of the series over and above the seasonal', i.e., the seasonally adjusted series, tc_t is a smooth function of time capturing the low-frequency movements of y_t, and i_t, representing the month-to-month non-seasonal fluctuations, is assumed to be a zero mean stationary process.

All pairs of components in (14.1) are assumed to be uncorrelated. Hannan defined a *stable* seasonal process to be one for which $s_t = s_{t-12}$, and this leads to two equivalent representations of such a process:

$$s_t = \begin{cases} a_j & \text{for } t = j \text{ or } t - j \text{ divisible by 12} \\ 0 & \text{otherwise} \end{cases} \qquad \sum_{j=1}^{12} a_j = 0 \tag{14.2}$$

and

$$s_t = \sum_{k=1}^{6} (\alpha_k \cos \omega_k t + \beta_k \sin \omega_k t) \quad \omega_k = 2\pi k/12 \tag{14.3}$$

Note that $\sin \omega_6 t = \sin \pi = 0$, so that the corresponding term in (14.3) has been included only for notational convenience. A demonstration of this equivalence, in which the α_k and β_k coefficients in (14.3) uniquely determine, and are uniquely determined by, the seasonal shift factors a_j in (14.2), was given by Nerlove (1965, footnote 14).

A changing seasonal component can be introduced by extending (14.3) to

$$s_t = \sum_{k=1}^{6} (\alpha_{k,t} \cos \omega_k t + \beta_{k,t} \sin \omega_k t) \tag{14.4}$$

Hannan (1964) regarded the time-varying $\alpha_{k,t}$ and $\beta_{k,t}$ as changing due to chance causes, but thought that these changes should only be gradual, achieving this by assuming that

$$E(\alpha_{k,s}\alpha_{k,s-t}) = E(\beta_{k,s}\beta_{k,s-t}) = \sigma_k^2 \rho_{k,t}$$

where $\rho_{k,0} \equiv 1$ and $\rho_{k,t} \approx 1$ for small t before gradually dying away to zero. All means and other cross-moments are assumed to be zero. Hannan then showed that, when $\rho_{k,t} = \rho_k^t$, the kth term in (14.4) has spectral density

$$f_k(\omega) = \frac{\sigma_k^2}{2\pi} \left(\frac{1-\rho_k^2}{1+\rho_k^2 - 2\rho_k \cos(\omega - \omega_k)} + \frac{1-\rho_k^2}{1+\rho_k^2 - 2\rho_k \cos(\omega + \omega_k)} \right)$$

which will be highly concentrated around $\omega = \omega_k$ when ρ_k is close to unity, a condition which is certainly required for a changing seasonal pattern. As Hannan pointed out, setting $\rho_k = 0.95$, so that $0.95^{48} = 0.085$ and $0.95^{24} = 0.29$, implies that the seasonal component can change almost completely after four years and that even after two years it could differ quite radically, so that setting ρ_k even closer to unity would often be appropriate.

14.12 Before estimation of s_t can be attempted, it will usually be necessary to filter y_t to reduce the effects of the very low-frequency components in z_t (i.e., tc_t) on the estimate. This can be accomplished by subtracting a symmetric two-sided moving average trend estimate from the data to obtain

$$y_t' = y_t - \sum_{j=-q}^{q} \delta_j y_{t-j}, \quad \delta_q = \delta_{-q}$$

Any of the moving averages discussed previously could be used for this purpose (see the discussion in Hannan, 1963, section 2). On defining

$$h(\omega) = \sum_{j=-q}^{q} \delta_j \cos j\omega = \delta_0 + 2\sum_{j=1}^{q} \delta_j \cos j\omega$$

the spectral density of y_t, $f_y(\omega)$, will then be 'modified' to $f_y'(\omega) = (1 - h(\omega))^2 f_y(\omega)$ and, using $y_t' = s_t' + z_t'$, the spectra of s_t and z_t will be modified to

$$f_z'(\omega) = (1 - h(\omega))^2 f_z(\omega)$$

and

$$f_s'(\omega) = \sum_k (1 - h(\omega))^2 f_k(\omega)$$

Since $f_k(\omega)$ is concentrated at ω_k, it follows that $\omega \approx \omega_k$, $1 - h(\omega) \approx 1 - h(\omega_k) \approx 1$ and the spectrum of the seasonal will not be altered by too much. Hannan (1964) utilized this idea to propose a method for computing the seasonal component that used signal extraction techniques based on the formulae given by Whittle (1963). Hannan (1967) later obtained a formula for the frequency response function of a filter which extracts a signal generated by a nonstationary process buried in noise and applied this to extracting a seasonal component of the type discussed here.

14.13 Nerlove (1964, 1965) utilized spectral techniques to analyse further Hannan's procedure and also to assess the BLS seasonal adjustment method. Burman (1965) also used the UC framework to extend the methodology underlying the Census II and BLS methods. Leser (1963, 1966) extended his trend removal method, discussed in §**10.7**, to jointly estimate the trend and seasonal variation. The use of moving averages for trend elimination and for measuring seasonal variation was extensively discussed by Durbin (1962, 1963) and Leong (1962).

Spectral evaluation of seasonal adjustment procedures

14.14 Rosenblatt (1968) listed the following spectral criteria needed for a satisfactory decomposition of a time series.

1. Spectral criteria for a good seasonal adjustment
   - 1.1 Seasonal peaks in the unadjusted series should be removed and should not appear in the spectrum of the seasonally adjusted series. The spectral power at seasonal frequencies should not be reduced to zero but to a uniform level consistent with the expected power of the irregular component. The spectrum of a good seasonally adjusted series should be relatively flat with no dominant peaks or troughs at the seasonal periods of $12/k$ months (seasonal frequencies $2\pi k/12$), $k = 1, 2, \dots, 6$.
   - 1.2 The coherence between the seasonally adjusted and unadjusted series should be very low at seasonal frequencies but should be high at nonseasonal frequencies, although the strength of the coherence may be reduced if significant moving seasonality is present.
   - 1.3 Since the adjustment process should not alter the timing between the unadjusted and seasonally adjusted series, the phase should be zero, although large deviations from zero are likely to occur at seasonal frequencies where low coherence is expected, given the sampling properties of phase estimates.

   1.4 The co-spectrum between the seasonally adjusted series and the seasonal component should be zero at all frequencies.
2. Spectral criteria for the separation of the irregular component from the seasonally adjusted series
   2.1 The spectrum of the irregular component should be relatively flat over the entire range of frequencies.
   2.2 Over the frequency range $2\pi/12$ to π (periods from 12 to 2 months) the spectrum of the seasonally adjusted and irregular series should be similar in appearance, in that they should have the same spectral power, both being relatively flat.
   2.3 The coherence between the irregular and seasonally adjusted series for periods from 12 to 2 months should be very high and the phase should be zero. For periods greater than 12 months the coherence should be low and the phase arbitrary.
   2.4 The trend-cycle and irregular components should have zero co-spectrum at all frequencies.

In analysing the X-11 and BLS seasonal adjustment procedures, Rosenblatt found that all the criteria were reasonably well satisfied except for 1.4: the co-spectra between the seasonally adjusted series and the estimated seasonal component were not close to zero and, in fact, were often negative for many frequency bands.

Grether and Nerlove (1970) devised several methods of seasonal adjustment based on a MMSE criterion of optimality and showed that these methods, in a simulation of a simple three component model, produced seasonally adjusted series bearing the same relationship to the unadjusted series as found with the BLS and X-11 methods, thus further confirming the usefulness of these adjustment procedures.

Regression methods of seasonal adjustment

14.15 Seasonal adjustment by regression methods was first proposed by Mendershausen (1937, 1939), Cowden (1942) and Jones (1943), after which it was ignored for twenty years as the Census and BLS techniques, based on moving averages, came to prominence. The regression approach was then given renewed impetus by the publication of Lovell (1963) and Jorgenson (1964, 1967). The primary aims of these articles were to propose alternative criteria for optimal seasonal adjustment and to show that seasonal adjustment by regression methods satisfied these criteria.

Lovell stated that an adjustment procedure should have the properties of orthogonality, idempotency and symmetry. In brief, orthogonality means that the seasonal component of a series is uncorrelated with the non-seasonal component: if y_t^a is the seasonal adjustment of y_t, then this implies that $\sum(y_t - y_t^a)y_t^a = 0$. Idempotency is the property that if an adjusted series is adjusted again, it remains unchanged, i.e., $(y_t^a)^a = y_t^a$. Symmetry means that if a linear model of adjustment is used such that $\mathbf{y}^a = \mathbf{Ay}$, where $\mathbf{y}$ and $\mathbf{y}^a$ are column vectors made up of y_t and y_t^a, then the matrix $\mathbf{A}$ is symmetric. Lovell proved that any sum-preserving procedure, i.e., one for which $(x_t + y_t)^a = x_t^a + y_t^a$, that satisfies two of these properties necessarily satisfies the third. A corollary to this is that any sum-preserving procedure that is orthogonal and idempotent (and hence symmetric) can be executed by regressing the unadjusted time series on an appropriate set of explanatory variables. Lovell also proved an extension of the Frisch–Waugh theorem (see §**12.4**): in terms of regression coefficients it is immaterial whether variables in a regression are deseasonalized by linear regression prior to the regression or whether unadjusted variables are used with the variables used for deseasonalization included as additional regressors.

Jorgenson (1964) developed a theory of seasonal adjustment based on the general linear statistical model:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\delta} + \mathbf{S}\boldsymbol{\sigma} + \boldsymbol{\varepsilon} \tag{14.5}$$

where $\mathbf{D}$ is a matrix whose columns represent the non-seasonal, deterministic (i.e., trend-cycle) variables, $\mathbf{S}$ is a matrix whose columns represent the seasonal deterministic variables, $\boldsymbol{\delta}$ and $\boldsymbol{\sigma}$ are vectors of constant coefficients and $\boldsymbol{\varepsilon}$ is a random component satisfying the assumptions $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. The matrix $\mathbf{D}\vdots\mathbf{S}$ is assumed to be fixed and of full rank.

Within this framework, Jorgenson derived linear estimators of the trend-cycle and seasonal components that are unique, minimum variance and unbiased. The seasonally adjusted series is obtained by subtracting the estimated seasonal component

$$\mathbf{y}^a = \mathbf{y} - \mathbf{S}\hat{\boldsymbol{\sigma}} = \mathbf{D}\hat{\boldsymbol{\delta}} + \hat{\boldsymbol{\varepsilon}} \tag{14.6}$$

and the adjustment procedure is termed minimum variance, linear, unbiased seasonal adjustment.

Although the Lovell and Jorgenson procedures both lead to regression methods of seasonal adjustment, they do not satisfy the same optimality criteria nor lead to identical adjustment procedures. Lovell (1966) pointed out that the Jorgenson method, while sum-preserving and idempotent, was not orthogonal or symmetric, so that it would be expected to yield an adjusted series that was correlated with the seasonal component.

If it is felt that linearity, unbiasedness and minimum variance are desirable criteria, then a definition of seasonal adjustment that would also satisfy Lovell's criteria is

$$\mathbf{y}^a = \mathbf{y} - \mathbf{S}\hat{\boldsymbol{\sigma}} - \mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{D}\hat{\boldsymbol{\delta}} \tag{14.7}$$

In contrast to (14.6), this seasonal adjustment also eliminates that part of the trend-cycle component which is correlated with the seasonal component.

14.16 Stephenson and Farr (1972) pointed out that there was no way of disentangling the trend-cycle component from the changing seasonal (the trend-related) component in (14.7), which could be a major drawback when the causes of changing seasonality were of interest. They therefore proposed employing the seasonal adjustment (14.6) but also specifying explicitly the changing seasonal variables along with the constant seasonals in the **S** matrix. The set-up chosen by Stephenson and Farr was to include powers of time in the **D** matrix, although they suggested that a set of low-order *grafted* (or segmented) polynomials might be preferable if a high-order polynomial was needed to adequately model the trend-cycle. For the seasonal variables making up **S** the set-up of (14.3) was adopted for the deterministic seasonal variables. To model changing seasonality, interactions of the components of (14.3) and powers of time could be defined, such as $t\cos(2\pi k/12)$, $t^2\cos(2\pi k/12)$, etc.

14.17 This approach was used to make a seasonal adjustment of the *logarithms* of the airline data, as such a transformation was more likely to support an additive component decomposition, and initial investigation suggested that a quadratic trend and linear seasonal-trend interactions were all that were required. The estimated seasonal factors are shown in Figure 14.5 with the X-11 factors obtained from an additive decomposition shown for comparison. The linear seasonal-trend interactions induce an evolving seasonal pattern, with the seasonal 'shape' being very different at the end of the sample period to what it was at the beginning, whereas the X-11 factors, although changing slowly, nevertheless keep much the same shape throughout the sample.

Plots of the two adjusted series are shown, along with the unadjusted logarithms, in Figure 14.6. Equation (14.6) is seen to effectively remove the seasonality in the series with the quadratic trend-cycle having a 'flattening' effect on the seasonally adjusted series, although both methods of seasonal adjustment produce very similar adjusted series.

Seasonal exponential smoothing

14.18 The exponential smoothing techniques introduced in §§**11.17–11.21** can straightforwardly be extended to deal with seasonal series. The most general version, Holt's two-parameter growth model with additive seasonals, now has l-step ahead forecasts of the monthly series x_t given by

$$\hat{x}_t(l) = m_t + lb_t + s_{t-12+l}$$

where s_t is the seasonal component, having the updating equation

$$s_t = s_{t-12} + \delta(1-\alpha_1)e_t$$

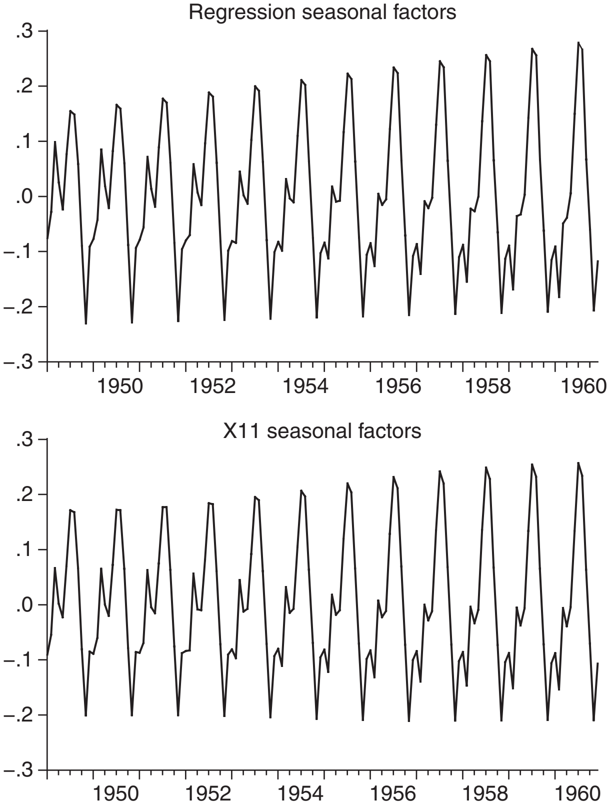


Figure 14.5 Seasonal factors for the logarithms of the airline data from the regression and X-11 approaches to seasonal adjustment

The corresponding recurrence equation is

$$s_t = \gamma_1(x_t - m_t) + (1 - \gamma_1)s_{t-12}$$

with that for m_t becoming

$$m_t = \alpha_1(x_t - s_{t-12}) + (1 - \alpha_1)(m_{t-1} + b_{t-1})$$

and that for b_t remaining unchanged. Multiplicative seasonality can be modelled by making various straightforward adjustments (see, for example, exhibit 4 of Gardner, 1985, where this formulation of the additive seasonal model is taken from).

14.19 Fitting this model to the logarithms of the airline data produced estimates of the smoothing parameters of $\hat{\alpha}_1 = 0.75$ and $\hat{\beta}_1 = \hat{\gamma}_1 = 0$, so that the

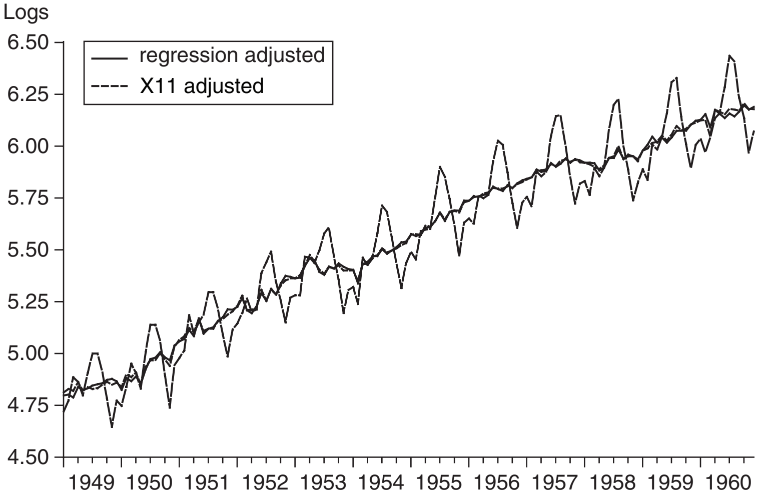


Figure 14.6 Seasonally adjusted logarithms of the airline data

updating equations are

$$\begin{aligned} m_t &= 0.75(x_t - s_{t-12}) + 0.25(m_{t-1} + b_{t-1}) \\ b_t &= b_{t-1} = b \\ s_t &= s_{t-12} \end{aligned}$$

and the l-step ahead forecast is $\hat{x}_t(l) = m_t + lb + s_{t-12+l}$, i.e., it is a constant slope forecast from the current seasonally adjusted level, the seasonal factors being fixed.

The Box–Jenkins approach to modelling seasonality

14.20 The Box–Jenkins approach to modelling time series revolved around the ARMA process (recall, for example, §**10.19**)

$$\varphi(B)x_t = \theta(B)a_t$$

which has an eventual forecast function that is the solution to the difference equation $\varphi(B)\hat{x}_t(l) = 0$, where B is understood to operate on l (cf. §**11.11**). Box and Jenkins (1970, chapter 9) argued that, to be able to represent seasonal behaviour, the forecast function would need to trace out a periodic pattern. This could be achieved by allowing the autoregressive operator $\varphi(B)$ to consist of a mixture of sines and cosines, possibly mixed with polynomial terms to allow for changes in the level of x_t and changes in the seasonal pattern. For example, a forecast function containing a sine wave with a 12-month period which is adaptive in both phase and amplitude will satisfy the difference equation

$$(1-\sqrt{3}B+B^2)\,\hat{x}_t(l)=0$$

The operator $1-\sqrt{3}B+B^2$ has roots of $\exp(\pm i2\pi/12)$ on the unit circle and is thus homogeneously nonstationary. Box and Jenkins pointed out, however, that periodic behaviour would not necessarily be represented parsimoniously by mixtures of sines and cosines. Taking their cue from their earlier use of the differencing operator $\Delta^d=(1-B)^d$ to effectively model homogenously nonstationary series, so that setting $\varphi(B)=\Delta^d\phi(B)$ allowed for d roots of the equation $\varphi(B)=0$ to be equal to unity, Box and Jenkins considered the seasonal difference operator $\Delta_s=1-B^s$, where s is the period of seasonality (e.g., $s=12$ for monthly data). Δ_s is a stable nonstationary operator having s roots of $\exp(i2\pi k/s)$, $k=0,\ 1,\ldots,s-1$, evenly spaced on the unit circle. The eventual forecast function will then satisfy $(1-B^s)\hat{x}_t(l)=0$ and so may (but need not) be represented by a full complement of sines and cosines:

$$\hat{x}_t(l)=b_0^{(t)}+\sum_{j=1}^{[s/2]}\left\{b_{1j}^{(t)}\cos\frac{2\pi jl}{s}+b_{2j}^{(t)}\sin\frac{2\pi jl}{s}\right\}$$

The b's are adaptive coefficients and $[s/2]=s/2$ if s is even and $[s/2]=(s-1)/2$ if s is odd.

14.21 When analysing seasonal data, say monthly, Box and Jenkins pointed out that relationships would be expected to occur (a) between observations for successive months in a particular year, and (b) between observations for the same month in successive years. They suggested that observations one year apart might be linked by a model of the form

$$\Phi(B^s)\Delta_s^D x_t=\Theta(B^s)\alpha_t \qquad (14.8)$$

Here $\Phi(B^s)$ and $\Theta(B^s)$ are polynomials in B^s of degrees P and Q, respectively, which satisfy the appropriate stationarity and invertibility conditions.

In general, the error component α_t would be expected to be correlated and, to take care of such relationships, a second model was introduced, this being an ARIMA (p,d,q) process for α_t

$$\phi(B)\Delta^d\alpha_t=\theta(B)a_t \qquad (14.9)$$

Substituting (14.9) into (14.8) obtains the general *multiplicative* model

$$\phi(B)\Theta(B^s)\Delta^d\Delta_s^D x_t=\theta(B)\Theta(B^s)\,a_t \qquad (14.10)$$

This process is said to be of *order* $(p, d, q) \times (P, D, Q)_s$. A similar argument can be used to obtain models with more periodic components to take care of multiple seasonalities.

The 'airline model'

14.22 The model (14.10) contains a high level of generality and, in accordance with their principle of parsimony, Box and Jenkins focused attention on generalizing a simple and widely applicable stochastic process for modeling nonstationary time series, the ARIMA(0, 1, 1) process, to the seasonal case. This leads to the component models (setting $s=12$ for convenience)

$$\begin{aligned} \Delta_{12} x_t &= (1 - \Theta B^{12}) \alpha_t \\ \Delta \alpha_t &= (1 - \theta B) a_t \end{aligned}$$

and the multiplicative $(0,1,1) \times (0,1,1)_{12}$ model

$$\Delta \Delta_{12} x_t = (1 - \theta B)(1 - \Theta B^{12}) a_t \tag{14.11}$$

which can be written explicitly as

$$x_t - x_{t-1} - x_{t-12} + x_{t-13} = a_t - \theta a_{t-1} - \Theta a_{t-12} - \theta\Theta a_{t-13}$$

Since the roots of $(1 - \theta B)(1 - \Theta B^{12}) = 0$ must lie outside the unit circle for invertibility, this imposes the conditions $|\theta| < 1$, $|\Theta| < 1$ on the parameters of the model.

Box and Jenkins found that (14.11) provided an adequate fit to the logarithms of the airline data with $\hat{\theta} = 0.4$, $\hat{\Theta} = 0.6$ and $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$ and hence the $(0, 1, 1) \times (0, 1, 1)_{12}$ model often became referred to as the 'airline model'.[3]

14.23 Forecasts from (14.11) can be made directly by using the difference equation approach of §**11.7**. Thus, using the airline parameter estimates, the first three months ahead forecasts are given by

$$\begin{aligned} \hat{x}_t(1) &= x_t + x_{t-11} - x_{t-12} - 0.4\hat{a}_t - 0.6\hat{a}_{t-11} + 0.24\hat{a}_{t-12} \\ \hat{x}_t(2) &= \hat{x}_t(1) + x_{t-10} - x_{t-11} - 0.6\hat{a}_{t-10} + 0.24\hat{a}_{t-11} \\ \hat{x}_t(3) &= \hat{x}_t(2) + x_{t-9} - x_{t-10} - 0.6\hat{a}_{t-9} + 0.24\hat{a}_{t-10} \end{aligned}$$

Figure 14.7 shows the forecasts of the logarithms of the airline data made at July 1957 for lead times up to 36 months: 'we see that the simple model, containing only two parameters, faithfully reproduces the seasonal pattern and supplies excellent forecasts' (Box and Jenkins, 1970, page 307).

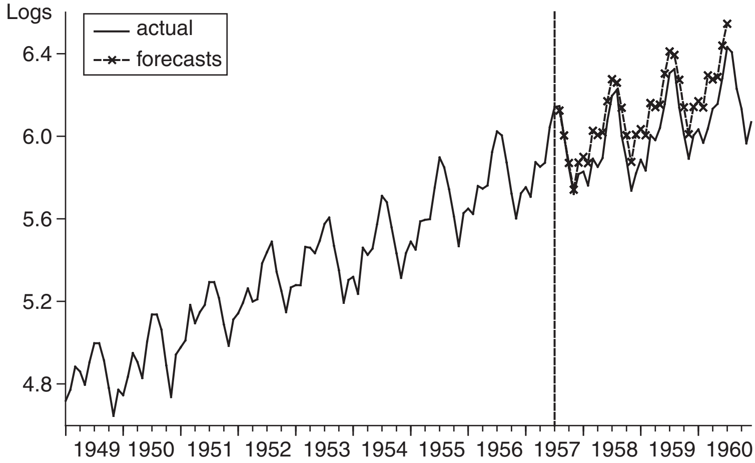


Figure 14.7 Logarithms of the airline data with forecasts for 1,2,3,...,36 months ahead made from the origin July 1957

On defining $\lambda = 1 - \theta$ and $\Lambda = 1 - \Theta$, the ψ-weights of (14.11) (cf. §**11.3**) are given by

$$\psi_{12r+m} = \lambda(1 + r\Lambda) + \delta\Lambda \quad r = 0, 1, 2, \ldots \;\; m = 1, 2, 3, \ldots, 12$$

where

$$\delta = \begin{cases} 1 & \text{when } m = 12 \\ 0 & \text{when } m \neq 12 \end{cases}$$

Given these ψ-weights, the forecast error variance at lead l is then given by (11.8) and, for the airline data and parameter estimates, the forecast error standard deviations increase from 3.7×10^{-2} at lead $l = 1$ to 19.6×10^{-2} at lead $l = 36$.

14.24 The π-weights of the airline model are obtained by equating coefficients in

$$(1 - B)(1 - B^{12}) = (1 - \theta B)(1 - \Theta B^{12})(1 - \pi_1 B - \pi_2 B^3 - \cdots)$$

to give

$$\begin{aligned} \pi_j &= \theta^{j-1}(1 - \theta) \quad j = 1,\ 2,\ \ldots,\ 11 \\ \pi_{12} &= \theta^{11}(1 - \theta) + (1 - \Theta) \\ \pi_{13} &= \theta^{12}(1 - \theta) - (1 - \theta)(1 - \Theta) \\ (1 - \theta B - \Theta B^{12} + \theta\Theta B^{13})\pi_j &= 0 \quad j > 14 \end{aligned}$$

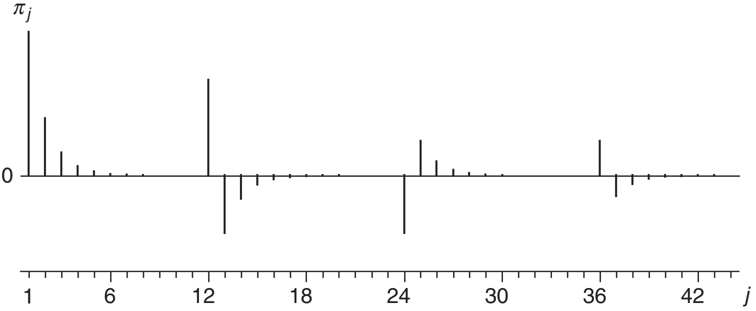


Figure 14.8 π-weights of the airline model for $\theta = 0.4$ and $\Theta = 0.6$

These are plotted in Figure 14.8 for the parameter values $\theta = 0.4$ and $\Theta = 0.6$. The reason why the weight function takes this particular form stems from the fact that (14.11) can be written as

$$a_{t+1} = \left\{1 - \frac{\lambda B}{1 - \theta B}\right\}\left\{1 - \frac{\Lambda B^{12}}{1 - \Theta B^{12}}\right\} x_{t+1} \tag{14.12}$$

From the definition of a EWMA in §**11.12**, (14.12) can be written as

$$a_{t+1} = (1 - \text{EWMA}_\lambda(x_t))(1 - \text{EWMA}_\Lambda(x_t)B^{12})x_{t+1}$$

where

$$\begin{aligned} \text{EWMA}_\lambda(x_t) &= \frac{\lambda}{1 - \theta B} x_t \\ \text{EWMA}_\Lambda(x_t) &= \frac{\Lambda}{1 - \Theta B^{12}} x_t \end{aligned}$$

On substituting $\hat{x}_t(1) = x_{t+1} - a_{t+1}$, (14.12) then becomes

$$\hat{x}_t(1) = \text{EWMA}_\lambda(x_t) + \text{EWMA}_\Lambda(x_{t-11} - \text{EWMA}_\lambda(x_{t-12})) \tag{14.13}$$

The one-step ahead forecast is thus a EWMA taken over previous months, modified by a second EWMA of discrepancies found between similar monthly EWMAs and actual observations in previous years. As Box and Jenkins (1970, page 313) put it

> For example, suppose we are attempting to predict December sales for a department store. These sales would include a heavy component from Christmas buying. The first term on the right of [14.13] would be an EWMA taken

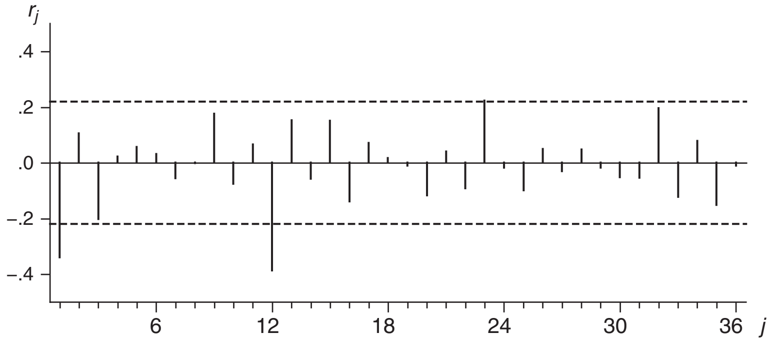


Figure 14.9 Sample autocorrelations of $\Delta\Delta_{12}x_t$ for the airline data with ±2 standard error bounds

over previous months up to November. However, we know this will be an underestimate, so we correct it by taking a second EWMA over previous years of the *discrepancies* between actual December sales and the corresponding monthly EWMA's taken over previous months in those years.

14.25 Recall from Table 9.7 that, for a nonseasonal IMA(0,1,1) process, the autocorrelations of the first differences beyond the first lag are all zero. For the multiplicative (0,1,1) × $(0,1,1)_{12}$ process (14.11) the only non-zero autocorrelations of $\Delta\Delta_{12}x_t$ are those at lags 1, 11, 12 and 13, which take the values

$$\rho_1 = -\frac{\theta}{1+\theta^2} \qquad \rho_{11} = \frac{\theta\,\Theta}{(1+\theta^2)(1+\Theta^2)} = \rho_{13}$$

$$\rho_{12} = -\frac{\Theta}{1+\Theta^2}$$

The sample autocorrelations of $\Delta\Delta_{12}x_t$ for the airline data are shown in Figure 14.9. On the assumption that the model is of the form (14.11), the variances for the higher-order sample autocorrelations are given by

$$V(r_j) \approx (T-13)^{-1}(1+2(\rho_1^2+\rho_{11}^2+\rho_{12}^2+\rho_{13}^2)) \quad j > 13$$

The standard errors to be attached to the higher-order sample autocorrelations for the airline data are approximately 0.11 and two standard error bounds are also shown in Figure 14.9. The sample autocorrelations at lags 1 and 12 are clearly significant and of the correct sign, those at 11 and 13 are correctly signed and approximately equal, and no others are significant, thus suggesting that the (0,1,1) × $(0,1,1)_{12}$ process might provide an adequate fit to the airline data.

Seasonal ARMA models

14.26 More general seasonal ARMA models of the form (14.10) were discussed in Box and Jenkins (1970, chapter 9.3 and Appendix A9.1), where the autocovariance structures of numerous seasonal models are provided, including models for which a non-multiplicative seasonal structure is allowed for. The identification, estimation and diagnostic checking of seasonal ARMA models essentially follow obvious generalizations of the principles outlined in Chapter 9 (Box and Jenkins find no major inadequacies in the model fitted to the airline data).

Implications of seasonal adjustment on regression models

14.27 Thus, by the beginning of the 1970s there was a well-used and widely available general method of seasonal adjustment, X-11, well-established seasonal exponential smoothing models for short-term forecasting, and an explicit extension of the ARIMA class of models for modelling and predicting seasonal time series. What was lacking was a framework for assessing the implications of seasonal adjustment for regressions containing time series. A succession of papers in the early 1970s, particularly those by Thomas and Wallis (1971), Sims (1974) and Wallis (1974), quickly produced such an assessment, although discussion of this lies outside our remit here.

15
Emerging Themes

The swinging sixties

15.1 This chapter discusses four research themes that began to emerge during the late 1950s and 1960s but whose real importance, like many aspects of this latter decade, only became apparent from the late 1970s onwards. These themes are: (i) inference in nonstationary autoregressive models; (ii) the use of model selection criteria; (iii) the Kalman filter, state space formulations and recursive estimation of time series models; and (iv) the specification and modelling of nonlinear time series processes.

Statistical inference in nonstationary autoregressions

15.2 The theory of statistical inference and estimation for stationary time series, which was discussed extensively in Chapter 9, began to be extended to nonstationary situations during the 1950s. Rubin (1950) had considered the first-order autoregressive model of §**9.22**, showing that the least squares estimator of the autoregressive parameter α was consistent (i.e., $\operatorname{p\,lim}\hat{\alpha}=\alpha$) for *all* values of α, including the *explosive* case $\alpha>1$, and not just for $|\alpha|<1$, as was implied by the results of Mann and Wald (1943) (see §**9.21**). White (1958) later demonstrated that, for $\alpha>1$, a suitably standardized function of $\hat{\alpha}$ had a well-defined limiting distribution. To be precise, under the assumption that $x_0=0$, $|\alpha|^T(\alpha^2-1)^{-1}(\hat{\alpha}-\alpha)$ has a limiting Cauchy distribution.[1] White (1959) then showed that, if the innovations could be assumed to be normal and independent, $\left(\sum x_{t-1}^2\right)^{1/2}(\hat{\alpha}-\alpha)$ has a limiting normal distribution if $|\alpha|\neq 1$, so that, for example, a symmetric 95% confidence interval for α would be given by

$$\alpha=\hat{\alpha}\pm 1.96\frac{\hat{\sigma}}{\left(\sum x_{t-1}^2\right)^{1/2}}\qquad \hat{\sigma}^2=T^{-1}\sum\left(x_t-\hat{\alpha}x_{t-1}\right)^2$$

and a large sample test of the hypothesis $\alpha = \alpha_0$ would be given by the likelihood ratio statistic $-2\log\lambda$, where

$$\lambda = \left(1 - \frac{(\hat{\alpha} - \alpha_0)^2 \sum x_{t-1}^2}{\sum (x_t - \alpha_0 x_{t-1})^2}\right)^{T/2}$$

which is asymptotically distributed as $\chi^2(1)$.

Rao (1961) extended White's results to higher-order autoregressive models of the type (9.14) whose characteristic equations have a single root exceeding unity with all remaining roots being less than one in absolute value. Anderson (1959) obtained the limiting distribution of estimators for higher-order models when more than one of the roots exceed unity in absolute value.

15.3 None of the above results relate to the borderline but extremely important case of $\alpha = 1$, when the series follows a random walk. For this case, White (1958) and Anderson (1959) showed that the limiting distribution of $T(\hat{\alpha} - 1)$ is the 'functional'

$$\frac{1}{2}\frac{x^2(1) - 1}{\int_0^1 x^2(t)dt}$$

where $x(t)$ is a 'Weiner process' with $Ex(t) = 0$ and $Ex^2(t) = t$, and White (1959) conjectured that the inferential results of the previous section would not hold.

Because the random walk is such an important model in time series analysis, the absence of any standard result for the $\alpha = 1$ case was a serious lacuna in the theory of statistical inference for autoregressive models. As is discussed in §**16.2**, when breakthroughs in this area eventually came in the 1970s and 1980s, these led to major research developments in the theory of nonstationary time series.

Determining the order of an autoregression by selection criteria

15.4 Akaike (1969) considered the problem of determining the order p of the autoregression (9.14). Denoting estimates of the innovation variance σ^2 obtained by successively fitting the autoregression by least squares for $p = 0, 1, 2, \ldots, P$ as $\hat{\sigma}^2(p)$, Akaike defined the *final prediction error* (*FPE*) to be

$$FPE(p) = \left(1 + \frac{p+1}{T}\right)\hat{\sigma}^2(p) \qquad (15.1)$$

The model with the *minimum FPE* is then selected as the most appropriate. The idea here is that $\hat{\sigma}^2(p)$ will be equal to the innovation variance when the correct autoregression is of order equal to or less than p, in which case *FPE* gives the asymptotic mean square prediction error. *FPE* will tend to be large when an unnecessarily large value of p is adopted and when p is less than the true order, *FPE* will also be large because $\hat{\sigma}^2(p)$ will contain model bias from the too small a value being used. Minimizing the *FPE* thus adopts a compromise between bias and too large a mean square prediction error.

Akaike (1971) generalized the definition (15.1) to the multivariate case (cf. §**12.33**) of an n-dimensional time series $\mathbf{x}_t$:

$$\mathbf{x}_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}_{t-i} + \boldsymbol{\varepsilon}_t \quad E(\boldsymbol{\varepsilon}_{t+s}\boldsymbol{\varepsilon}_t') = \begin{cases} \mathbf{0} & s \neq 0 \\ \Sigma & s = 0 \end{cases}$$

The *multiple FPE* (*MFPE*) is then defined as

$$MFPE(p) = \left(1 + \frac{pn}{T}\right)^{n} \left(1 - \frac{pn}{T}\right)^{-n} |\hat{\Sigma}(p)|$$

The *FPE* was the first *information criterion* to be proposed and quickly became a popular method of choosing the order of an autoregressive model. Variations and extensions to other types of model were soon to follow (see §**16.12**).

Recursive estimation, state space models and the Kalman filter

15.5 The idea of sequentially updating or recursively estimating the parameters of a model has a history stretching back to Gauss in the 1820s, but was only rediscovered by Plackett (1950).[2] A decade later, Rudolf Kalman published a recursive state estimation algorithm for stochastic dynamic systems described by discrete-time state space equations (Kalman, 1960), at the core of which was a modified Gauss–Plackett RLS algorithm (although it was unlikely that Kalman was aware of this at the time). After something of a delay, Kalman's idea led to a huge body of research on recursive estimation across a range of different disciplines, with the algorithm being referred to universally as the *Kalman filter*.[3]

It is generally accepted that the reasons for this delay in the take-up of Kalman's procedure in the time series community were twofold. First, the original paper and its continuous time counterpart (Kalman and Bucy, 1961) were written for an engineering audience and so used a language, notation and style that was alien to statisticians. Second, the original set-up of the model assumed that the parameters of the underlying state space model were known exactly, so that it could only provide estimates and forecasts of the state variables of the system. This latter restriction was lifted with the development of methods for computing the likelihood function for state space models (see Schweppe, 1965), while several papers in the early 1970s introduced the Kalman filter to a wider audience by casting it in more familiar terminology (see, especially, Harrison and Stevens, 1971, and Duncan and Horn, 1972).

15.6 The Kalman filter has been expressed in several ways but essentially contains two recursive equations. Defining $\mathbf{y}_t$ to be an n-dimensional time series, and following, with some modifications, Duncan and Horn (1972), these two equations are an *observation equation*

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \quad t = 1, 2, \ldots, T \tag{15.2}$$

and a *dynamic regression coefficients transition equation,*

$$\boldsymbol{\beta}_t = \mathbf{T}_t\boldsymbol{\beta}_{t-1} + \mathbf{W}_t\mathbf{u}_t \quad t = 2, \ldots, T \tag{15.3}$$

with initial value $\boldsymbol{\beta}_1 = \gamma_1 + \mathbf{u}_1$, where γ_1 is a known prior mean for $\boldsymbol{\beta}_1$. Each of the observation equations is of the familiar regression form in which the $r \times 1$ regression coefficient vector $\boldsymbol{\beta}_t$ is called the *state* at time t, $\mathbf{X}_t$ is a known $n \times r$ matrix of regressors and $\boldsymbol{\varepsilon}_t$ is an $r \times 1$ vector of errors. The dynamic state transition equation (15.3) expresses the state $\boldsymbol{\beta}_t$ as a known linear transformation, given by the $r \times r$ matrix $\mathbf{T}_t$, of the previous state $\boldsymbol{\beta}_{t-1}$ plus a linear combination, given by the $r \times g$ matrix $\mathbf{W}_t$, of the $g \times 1$ vector of errors, $\mathbf{u}_t$.

The dynamic state transition equations, the starting equation $\boldsymbol{\beta}_1 = \gamma_1 + \mathbf{u}_1$, and assumptions about the error vectors $\mathbf{u}_t$, $t = 1, 2, \ldots, T$, provide a model for the complete vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_T)$ of state elements (regression coefficients). The stronger set of assumptions concerning the error terms are $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{Q}_t)$ and $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{R}_t)$ with $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$ being independent of each other, i.e., that the errors follow independent multivariate normal distributions with known covariance matrices. These are known as the *Gaussian assumptions*, while the weaker set, the *wide sense* (*WS*) *assumptions*, make no distributional assumptions, so that

$$\begin{bmatrix} \mathbf{u}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} \sim WS\begin{bmatrix} \mathbf{Q}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_t \end{bmatrix}$$

Kalman (1960) showed that the minimum mean square linear estimator (MMSLE) $\mathbf{b}_t$ of the state $\boldsymbol{\beta}_t$, based on all the data γ_1, $\mathbf{y}_1$, $\mathbf{y}_2, \ldots, \mathbf{y}_t$ through to time t, was given by the recursive *updating* equations:

$$\mathbf{b}_t = \mathbf{b}_{t|t-1} + \mathbf{S}_{t|t-1}\mathbf{X}_t'\mathbf{D}_t^{-1}(\mathbf{y}_t - \mathbf{X}_t\mathbf{b}_{t|t-1}) \tag{15.4}$$

and

$$\mathbf{S}_t = \mathbf{S}_{t|t-1} - \mathbf{S}_{t|t-1}\mathbf{X}_{t-1}'\mathbf{D}_t^{-1}\mathbf{X}_t\mathbf{S}_{t|t-1}$$

where

$$\begin{aligned} \mathbf{b}_{1|0} &= \gamma_1 & \mathbf{S}_{1|0} &= \mathbf{Q}_1 & & \\ \mathbf{b}_{t|t-1} &= \mathbf{T}_t\mathbf{b}_{t-1} & \mathbf{S}_{t|t-1} &= \mathbf{T}_t\mathbf{S}_{t-1}\mathbf{T'}_t + \mathbf{W}_t\mathbf{Q}_t\mathbf{W'}_t & & t = 2, \ldots, T \\ \mathbf{D}_t &= \mathbf{R}_t + \mathbf{X}_t\mathbf{S}_{t|t-1}\mathbf{X}_t' & & & & t = 1, 2, \ldots, T \end{aligned}$$

It can also be shown that

$$(\mathbf{b}-\boldsymbol{\beta}_t)\sim WS(\mathbf{0},\mathbf{S}_t)$$

and that

$$E(\mathbf{b}_t-\boldsymbol{\beta}_t)\tilde{\mathbf{y}}_t=\mathbf{0}$$

where $\tilde{\mathbf{y}}_t=(y_1',\mathbf{y}_1',\ldots,\mathbf{y}_t')'$.

Kalman derived these results using wide-sense conditional distributions and expectations and orthogonal projection theory and it is important to emphasize that they do not rely on Gaussian assumptions. This lent a degree of robustness to the algorithm which made it ideal for many practical applications. If Gaussianity is assumed then (15.4) may be given a Bayesian interpretation in that $\mathbf{b}_t$ can be shown to be the posterior mean for $\boldsymbol{\beta}_t$ given $\tilde{\mathbf{y}}_t$ and is thus the Bayesian estimator under squared error loss or, in other words, the MMSE for $\boldsymbol{\beta}_t$. A 'full' Bayesian approach for a state space formulation of a Holt-Winters type UC model was provided by Harrison and Stevens (1971) in the context of short-term sales forecasting.

15.7 At first sight, the *state space representation* of (15.2) and (15.3), along with the associated Kalman filter, look to have little in common with the typical time series models developed over the years. This lack of connection is illusory, however. Consider, for example, the AR(2) model $y_t=\phi_1 y_{t-1}+\phi_2 y_{t-2}+u_t$. This can be written as the state space model

$$\begin{aligned} y_t &= [1\ \ 0]\boldsymbol{\beta}_t \\ \boldsymbol{\beta}_t &= \begin{bmatrix} y_t \\ \phi_2 y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix}\boldsymbol{\beta}_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_t \end{aligned}$$

Similarly, the MA(1) model $y_t=u_t+\theta u_{t-1}$ has the state space form

$$\begin{aligned} y_t &= [1\ \ 0]\boldsymbol{\beta}_t \\ \boldsymbol{\beta}_t &= \begin{bmatrix} y_t \\ \theta u_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\boldsymbol{\beta}_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} u_t \end{aligned}$$

On the other hand, the random walk plus noise model $y_t=\beta_t+\varepsilon_t$, with $\beta_t=\beta_{t-1}+u_t$, is already in state space form.

15.8 There were various extensions to the Kalman filter suggested during the 1960s, most notably the *extended Kalman filter*, which was a simple, approximate solution to the underlying optimal, nonlinear, estimation and filtering problem. This and other alternatives were discussed in the early but very influential book by Jazwinski (1970). Further applications and extensions of the Kalman filter and recursive least squares in general are discussed in §§**16.8–16.9**.

Nonlinearities in time series

15.9 A general nonlinear time series model would take the form

$$h(y_t, y_{t-1}, y_{t-2}, \ldots) = \varepsilon_t \tag{15.5}$$

Suppose the function $h()$ was such that (15.5) was 'invertible', so that

$$y_t = h'(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$$

The basis for the nonlinear modelling of time series is to write this as the *Volterra series expansion*

$$y_t = \mu + \sum_{u=0}^{\infty} g_u \varepsilon_{t-u} + \sum_{u=0}^{\infty}\sum_{v=0}^{\infty} g_{uv}\varepsilon_{t-u}\varepsilon_{t-v} + \sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\sum_{w=0}^{\infty} g_{uvw}\varepsilon_{t-u}\varepsilon_{t-v}\varepsilon_{t-w} + \cdots \tag{15.6}$$

where $\mu = h'(0, 0, 0, \ldots)$ and

$$g_u = \left(\frac{\partial h'}{\partial e_{t-u}}\right)_0 \quad g_{uv} = \left(\frac{\partial^2 h'}{\partial e_{t-u}\partial e_{t-v}}\right)_0 \quad g_{uvw} = \left(\frac{\partial^2 h'}{\partial e_{t-u}\partial e_{t-v}\partial e_{t-w}}\right)_0 \quad \text{etc.}$$

The notation for the derivatives denotes that they have been taken at the point $\mathbf{0} = (0, 0, 0, \ldots)$.

Models of this type formed the basis for Weiner's (1958) treatment of nonlinear processes but, although there was much theoretical research done on Volterra series, until the 1970s the formidable complexities of both the theory and the computations involved meant that there was almost no progress made on modelling time series in this way.

15.10 Some progress, however, was made by focusing attention on spectral representations of time series and, in particular, on higher-order counterparts to the spectrum and cross-spectrum discussed in Chapter 13. If, rather than the full Volterra expansion (15.6), y_t is generated by the stationary, linear process

$$y_t = \sum_{s=0}^{\infty} g_u \varepsilon_{t-u}$$

it will have the spectrum (cf. (13.4))

$$f(\omega) = |g(e^{-i\omega})|^2 \sigma_\varepsilon^2$$

It will also have the *bi-spectrum*

$$f(\omega_1, \omega_2) = f(\omega_1) f(\omega_2) f(\omega_1 + \omega_2)\lambda_3$$

where $\lambda_3 = E\varepsilon_t^3$. Thus the bi-spectrum will be zero if $\lambda_3 = 0$, as would be the case for a Gaussian process. Higher-order functions of spectra and cross-spectra are called *polyspectra*, for which Shiryaev (1960) and Brillinger (1965) provided detailed theoretical treatments, with Godfrey (1965) being the first to compute bi-spectral estimates for economic time series.

Although progress was thus rather slow during the 1960s, the topic of non-linear time series modelling began to advance rapidly over the following two decades, as is discussed in §§**16.10–16.11**.

15.11 It is thus apparent that several major research areas were in gestation during the 1960s: how they began to flourish in the decades after 1970, along with other topics that developed in tandem, is the theme of our final chapter.

16
The Scene is Set

Box and Jenkins as a watershed

16.1 The publication of Box and Jenkins' book in 1970 represented a watershed in the development of time series analysis, for it provided a systematic framework for identifying, estimating and checking a range of models that have had a great impact on the practical modelling of time series, particularly for forecasting. This synthesis also provided the impetus for major theoretical developments which, when allied with rapidly increasing computing power and enhanced computational algorithms, opened up many new areas for empirical analysis.

This final chapter looks at the major advances in time series modelling that have occurred since 1970, linking them with the historical developments of the previous chapters, particularly Chapter 15. The discussion, however, aims to be as concise as possible, as these advances lead naturally into material that constitutes the core of all modern textbooks on time series analysis.

Unit roots and trend and difference stationarity

16.2 Box and Jenkins' championing of differencing to induce stationarity was an important practical step that attracted great interest. Kendall's (1971) review of their book professed some disquiet about the use of differencing, offering an argument for why d was rarely found to exceed two in practice that was firmly demolished by Box and Jenkins (1973: see also Box and Newbold, 1971). Attention then became focused on whether, in the first-order autoregressive model, it was possible to test the hypothesis that $\alpha = 1$, which became known as the *unit root hypothesis* (recall §**15.3**). The limit distributions of $\hat{\alpha}$ and $\hat{\alpha} - 1$ under this null hypothesis were obtained by Dickey and Fuller (1979), with percentiles of test statistics based on these distributions being provided in Fuller (1976, chapter 8). This was just the start of a major research agenda which, over the last thirty years, has come to dominate much of the time series literature, particularly in econometrics, with major contributions being provided by Evans and Savin (1981, 1984), Phillips (1987a, 1987b), Elliott, Rothenberg and Stock (1996) and Müller and Elliott (2003). A recent and detailed textbook exposition of this vast literature is Patterson (2010).

There have been two particularly interesting themes of this research agenda. The first is the distinction introduced by Nelson and Plosser (1982) between *trend stationary* and *difference stationary* processes: the former represent time series that can be expressed as stationary deviations about a deterministic trend, typically linear, so that they do not contain a unit root in their autoregressive component; the latter represent time series that require differencing to induce stationarity, and hence do have a unit root and embody a stochastic trend (recall from §**12.5** the prescient distinction made half a century earlier between these formulations by Smith, 1926). Chan, Hayya and Ord (1977) and Nelson and Kang (1981, 1984) illustrated the difficulties caused by incorrectly assuming that a series was trend stationary when it was, in fact, difference stationary, but simple extensions of the Dickey–Fuller testing procedure (see Dickey and Fuller, 1981) allow discrimination to be made between the two processes.

The second theme is the impact that shifts and breaks in the trend function have on inferences about the unit root hypothesis. This was first demonstrated by Perron (1989) and the theme has been developed substantially over the last twenty years, an important recent contribution being Harris et al. (2009), with Perron (2006) surveying the literature.

Signal extraction using unobserved component models

16.3 The *steady* unobserved component model used to assess the optimality of exponential smoothing in §**11.19** assumed that an observed series x_t was generated by a random walk permanent or trend component plus an independent white noise error (the Muth, 1960, model). Such a representation gives rise to an ARIMA(0,1,1) process for which the first-order autocorrelation for Δx_t is restricted to the interval $-0.5 < \rho_1 < 0$, so that the process cannot account for positive autocorrelation in Δx_t. To allow for positive autocorrelation requires either relaxing the assumption that the trend component is a random walk, so that it contains both permanent and transitory components, or allowing the error and the innovation to the random walk trend to be correlated.

This idea led to an interesting decomposition of an $I(1)$ series proposed by Beveridge and Nelson (1981), who thought that a random walk trend was not as restrictive as it might at first seem, in essence asking the question of why a trend should contain a transitory component. They thus relaxed the assumption that the component innovations had to be independent and considered the Wold decomposition

$$\Delta x_t = \mu + \psi\,(B)\,a_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j} \tag{16.1}$$

Since $\psi(1) = \sum \psi_j$ is a constant, the polynomial $\psi(B)$ can be written as $\psi(B) = \psi(1) + C(B)$, so that

$$\begin{aligned} C(B) &= \psi(B) - \psi(1) \\ &= 1 + \psi_1 B + \psi_2 B^2 + \cdots - (1 + \psi_1 + \psi_2 + \psi_3 + \cdots) \\ &= -\psi_1(1-B) - \psi_2(1-B^2) - \psi_3(1-B^3) - \cdots \\ &= (1-B)(-\psi_1 - \psi_2(1+B) - \psi_3(1+B+B^2) - \cdots) \end{aligned}$$

i.e.,

$$\begin{aligned} C(B) &= (1-B)\left(-\sum_{j=1}^{\infty}\psi_j - \left(\sum_{j=2}^{\infty}\psi_j\right)B - \left(\sum_{j=3}^{\infty}\psi_j\right)B^2 - \cdots\right) \\ &= \Delta\tilde{\psi}(B) \end{aligned}$$

Thus $\psi(B) = \psi(1) + \Delta\tilde{\psi}(B)$, implying that

$$\Delta x_t = \mu + \psi(1)a_t + \Delta\tilde{\psi}(B) + a_t$$

Hence, if $x_t = m_t + u_t$, as in §**11.19**, then the components are defined as

$$\Delta m_t = \mu + \left(\sum_{j=0}^{\infty}\psi_j\right) a_t = \mu + \psi(1)\,a_t$$

and

$$u_t = -\left(\sum_{j=1}^{\infty}\psi_j\right) a_t - \left(\sum_{j=2}^{\infty}\psi_j\right) a_{t-1} - \left(\sum_{j=3}^{\infty}\psi_j\right) a_{t-2} - \cdots = \tilde{\psi}(B)\,a_t$$

Since a_t is white noise, the trend component is therefore a random walk with a rate of drift equal to μ and an innovation equal to $\psi(1)\,a_t$, which is thus proportional to the innovation of the original series. The variance of this innovation is $(\psi(1))^2\sigma_a^2$, which may be larger or smaller than σ_a^2 depending on the signs and patterns of the ψ–weights. In particular, the innovations to the trend component will be 'noisier' than those to x if the ψ–weights are positive, which would typically be the case if the changes in x are positively correlated. The error component is clearly stationary but, since it is driven by the same innovation as the trend, Δm_t and u_t must be *perfectly correlated*, in direct contrast to the Muth decomposition.

As an example, the Beveridge–Nelson decomposition of an ARIMA(0,1,1) process is $\Delta m_t = \mu + (1-\theta)\, a_t$ and $u_t = \theta\, a_t$. In terms of the original observations, these can be written as

$$m_t = (1-\theta)\sum_{j=0}^{\infty} \theta^j x_{t-j}$$

and

$$u_t = \theta \sum_{j=0}^{\infty} \theta^j x_{t-j}$$

16.4 In more general set-ups the unobserved components m_t and u_t can be estimated by the technique of *signal extraction*, which is based on the theory developed in Whittle (1963), itself an extension of Weiner–Kolmogorov prediction theory (see Weiner, 1949). This approach was extended by Pierce (1979) and Bell (1984) to cover nonstationarities. Suppose that the component models are $\Delta m_t = \mu + \gamma(B)\, v_t$ and $u_t = \lambda(B)\, \varepsilon_t$, where $\gamma(B)$ and $\lambda(B)$ have no common roots and v_t and ε_t are independent white noises with variances σ_v^2 and σ_ε^2. The observed series x_t can then be written as $\Delta x_t = \mu + \theta(B)\, a_t$, where $\theta(B)$ and σ_a^2 can be obtained from

$$\sigma_a^2 \frac{\theta(B)\theta(B^{-1})}{(1-B)(1-B^{-1})} = \sigma_v^2 \frac{\gamma(B)\gamma(B^{-1})}{(1-B)(1-B^{-1})} + \sigma_\varepsilon^2 \lambda(B)\lambda(B^{-1}) \qquad (16.2)$$

Given an infinite sample $\ldots, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, \ldots$ of observations, the MMSE estimate of m_t is (the following analysis can be straightforwardly amended when only a finite sample is available: see Pierce, 1979)

$$\hat{m}_t = v_m(B) x_t = \sum_{j=-\infty}^{\infty} v_{mj} x_{t-j}$$

where the filter $v_m(B)$ is defined as

$$v_m(B) = \frac{\sigma_v^2 \gamma(B)\gamma(B^{-1})}{\sigma_a^2 \theta(B)\theta(B^{-1})}$$

An estimate of u_t will then be given by

$$\hat{u}_t = x_t - \hat{m}_t = (1 - v_m(B)) x_t = v_u(B) x_t$$

For example, for the Muth model of a random walk overlaid with stationary noise

$$v_m(B) = \frac{\sigma_v^2}{\sigma_a^2}(1-\theta B)^{-1}(1-\theta B^{-1})^{-1} = \frac{\sigma_v^2}{\sigma_a^2}\frac{1}{(1-\theta^2)}\sum_{j=-\infty}^{\infty}\theta^{|j|}B^j$$

and

$$\hat{m}_t = \frac{(1-\theta)^2}{1-\theta^2}\sum_{j=-\infty}^{\infty}\theta^{|j|}x_{t-j}$$

Thus, for values of θ close to unity, $\hat{m}_t$ will be given by a very long moving average of future and past values of x. If θ is close to zero, however, $\hat{m}_t$ will be almost equal to the most recently observed value of x.

Note that it will not necessarily be the case that the parameters of the component models can be identified from $\theta(B)$ and σ_a^2. For example, if $\Delta x_t = (1-\theta B)\,a_t$ then, if u_t is to be white noise, the most general model for m_t is $\Delta m_t = (1-\Theta B)\,v_t$, where $-1 \leq \Theta \leq \theta$. Setting $\Theta = -1$ can be shown to minimize σ_v^2 and maximize σ_ε^2, subject to the condition (16.2), so making the trend as smooth as possible. This is known as the *canonical decomposition* of x_t, for which

$$\hat{m}_t = \frac{\sigma_v^2}{\sigma_a^2}\frac{(1+B)(1+B^{-1})}{(1-\theta B)(1-\theta B^{-1})}$$

A more general framework of signal extraction for ARIMA models was developed in Tiao and Hillmer (1978), with further extensions being provided by Bell and Martin (2004).

16.5 The UC approach provides an alternative formulation of the H–P filter discussed briefly in §**10.7**. The first-order conditions given there can be expressed, using current notation, as

$$x_t = (1+\delta(1-B)^2(1-B^{-1})^2)\,m_t$$

where δ now denotes the smoothing parameter. The H–P trend estimator is thus

$$\hat{m}_t(\delta) = (1+\delta(1-B)^2(1-B^{-1})^2)^{-1}x_t$$

The MMSE trend estimator can be written, using (16.2), as

$$\hat{m}_t = \frac{\sigma_v^2}{\sigma_a^2}\frac{\gamma(B)\gamma(B^{-1})}{\theta(B)\theta(B^{-1})}x_t = \frac{\gamma(B)\gamma(B^{-1})}{\gamma(B)\gamma(B^{-1}) + \left(\sigma_\varepsilon^2/\sigma_v^2\right)\lambda(B)\lambda(B^{-1})}x_t$$

Comparing this expression with $\hat{m}_t(\delta)$ shows that, for this to be optimal in the MMSE sense,

$$\gamma(B) = (1-B)^{-1}, \quad \lambda(B) = 1, \quad \delta = \sigma_\varepsilon^2/\sigma_v^2$$

In other words, the underlying UC model must have the trend component $\Delta^2 m_t = v_t$ with the irregular component u_t being white noise.

Several competitors to the H–P filter have since been proposed: see Baxter and King (1999), Pollock (2000) and Christiano and Fitzgerald (2003) for three such filters and Mills (2003) for textbook exposition.

Advances in seasonal adjustment techniques

16.6 These signal extraction concepts were subsequently employed to develop model-based methods of seasonal adjustment, most notably by Box, Hillmer and Tiao (1978), Burman (1980), Hillmer and Tiao (1982), Hillmer, Bell and Tiao (1983), Bell and Hillmer (1984) and Maravall and Pierce (1987). Cleveland and Tiao (1976), Wallis (1982) and Burridge and Wallis (1984) were able to provide a signal extraction interpretation of the X-11 filters. If the UC decomposition of a seasonal series is defined as $x_t = n_t + s_t$, where n_t and s_t are non-seasonal and seasonal components following independent ARMA processes, then Cleveland and Tiao (1976) found that the following models provide a close approximation to the monthly X-11 filter:

$$\begin{aligned}\Delta^2 n_t &= (1 - 1.252B + 0.4385B^2)\, u_t \\ \Delta_{12} s_t &= (1 + 0.64B^{12} + 0.83B^{24})\, v_t \quad \sigma_u^2/\sigma_v^2 = 24.5\end{aligned}$$

These lead to an ARIMA model for x_t of the form $\Delta\Delta^{12} x_t = \theta(B)\, a_t$, where the moving average polynomial $\theta(B)$ is of order 25.

Burridge and Wallis (1984) used the polynomial $U(B) = 1 + B + \ldots + B^{11}$ as the seasonal operator rather than Δ_{12} and obtained the component models

$$\begin{aligned}\Delta^2 n_t &= (1 - 1.59B + 0.86B^2)\, u_t \\ U(B) s_t &= (1 + 0.71B^{12} + B^{24})\, v_t \quad \sigma_u^2/\sigma_v^2 = 90.9\end{aligned}$$

These lead to the same form of model for x_t except that $\theta(B)$ is of order 26. In both cases the moving average coefficients are very small after lag 13, but those for lags 2–11 are quite substantial, implying that a multiplicative moving average specification of the type $(1 - \theta B)(1 - \Theta B^{12})$ would be a rather poor approximation to $\theta(B)$: as Burridge and Wallis emphasized, simple specifications for the component models typically do not yield simple composite models.

Of course, within a UC framework it is also possible, and perhaps even desirable, to estimate the trend component itself rather than to simply seasonally adjust the actual observations, which will contaminate the trend with the noise component. An interesting framework for doing this was proposed by Box, Pierce and Newbold (1987).

16.7 The X-11 seasonal adjustment procedure was upgraded in the late 1970s to X-11-ARIMA by Dagum (1978, 1982). The major innovation is revealed by the name: the need for asymmetric filters to deal with 'end-point' problems was reduced by extending the series being adjusted with ARIMA forecasts. This was itself upgraded in the mid-1990s to X-12-ARIMA (Findley et al., 1998). Scott (1992, 1997) provided extended reviews of the two packages and Ladiray and Quenneville (2001) gave detailed descriptions of them. Further analysis of the underlying trend-cycle filters was provided by Doherty (2001), Gray and Thompson (2002) and Quenneville, Ladiray and Lefrançois (2003).

Several other seasonal adjustment programmes have been developed over the years, such as SABL (Cleveland, Dunn and Terpenning, 1978) and STL (Cleveland et al., 1990). Perhaps the most important of these is the component model-based TRAMO/SEATS package, which is intensively used at Eurostat and the European Central Bank and, indeed, by many other central banks, with parts of the package having since been incorporated into X-12-ARIMA: see, for example, Gómez and Maravall (1996) and Maravall (2000) for details.

Structural models, state space representations and the Kalman filter

16.8 As noted in §**15.6**, UC, or structural models, as they came to be more popularly known, fit naturally into a state space representation. For example, the model $x_t = m_t + u_t$ with $m_t = \mu + m_{t-1} + v_t$ and u_t white noise has the observation equation

$$x_t = [0 \ \ 1]\boldsymbol{\beta}_t + u_t$$

and transition equation

$$\boldsymbol{\beta}_t = \begin{bmatrix} \mu_t \\ m_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ v_t \end{bmatrix}$$

on defining $\mu_t = \mu_{t-1} = \mu$. Once given a state space representation, the parameters of the structural model can be estimated via the Kalman filter and forecasts straightforwardly obtained.

16.9 Such formulations allowed state space modelling to develop rapidly across many areas of time series analysis, leading to far too many major contributions to reference individually. Notable texts, however, are Anderson and Moore (1979), Harvey (1989) and Durbin and Koopman (2001), and reference should also be made to the dedicated software package STAMP (Structural Time Series Analyser, Modeller and Predictor: Koopman et al., 2009). A related theme of some interest is that of Bayesian forecasting using dynamic linear models

(Harrison and Stevens, 1971, 1976; West, Harrison and Migon, 1985; West and Harrison, 1997).

Nonlinearities

16.10 Perhaps the greatest advances in time series modelling over the last forty years have been made in the area of nonlinear processes. At least three themes have emerged from this research: the use of nonlinear parametric models, for example regime switching models such as SETAR (self-exciting threshold autoregressive), STAR (smooth transition autoregressive) and Markov-switching processes; neural networks; and chaotic processes. Tong (1990), Granger and Teräsvirta (1993), Kantz and Schreiber (1997) and Franses and van Dijk (2000) are texts that may be consulted on these themes, while Teräsvirta (2006) provides a more recent survey.

16.11 A class of models, first introduced in the early 1980s to model inflation volatility by Engle (1982), have since become extremely influential for modelling variances that change through time. This is, of course, the GARCH class of processes (see, for example, Gouriéroux, 1997), which often show that the impact of shocks to the variance can be long-lasting and persistent, especially for financial time series. There is an enormous literature on this topic and recent surveys of univariate and multivariate volatility models based on GARCH processes are Baillie (2006) and Brooks (2006).

Persistency, often referred to as *long memory*, has also been found in the levels of time series as well as in their variances, being introduced by Mandelbrot and Wallis (1969), where it was named the Hurst effect, after the hydrologist Harold E. Hurst, who encountered this phenomenon when analysing records of river flows for the Nile (Hurst, 1951: for further evidence of long memory in hydrology, see Hosking, 1984). Long memory is associated with an autocorrelation function that decays hyperbolically, i.e., slower than the exponential decline of a stationary process but faster than the linear decline associated with an $I(1)$ process. It can be characterized by the use of *fractional differencing* through the operator

$$\Delta^d = (1-B)^d = 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots$$

where d is now allowed to take *any* value greater than -1, not just integers as in the typical ARIMA framework. $\Delta^d x_t = a_t$ then defines *fractional white noise* and, if $|d| < 0.5$, x_t is stationary and invertible and will exhibit long memory. If a_t is autocorrelated it may be modelled as an ARMA process, thus leading to the AR-*fractionally integrated*-MA, or ARFIMA, process.

The notion of fractional differencing seems to have been proposed contemporaneously and independently by Hosking (1981) and Granger and Joyeux

(1980). Detailed surveys of what is another large literature have been provided over the years by, for example, Beran (1992), Baillie (1996), Robinson (2003) and Velasco (2006).

Information criteria and diagnostic checking

16.12 Akaike quickly replaced the *FPE* autoregressive model selection criterion with the *AIC*, which evidently stood originally for 'An Information Criterion', but was quickly tagged Akaike's Information Criterion (Akaike, 1974). During the late 1970s several other criteria were also proposed, all taking the general form (cf. (15.1))

$$IC(p) = \log \hat{\sigma}^2(p) + c_T p \tag{16.3}$$

Here c_T is a weighting factor that typically depends upon the sample size T and its form effectively distinguishes the alternative criteria. The first term of (16.3) measures the fit of the AR(p) model and will decrease for increasing p, as long as there are no degrees of freedom correction in the variance estimator and the sample size remains fixed at T for all p. The order that minimizes this criterion is chosen as the estimator $\hat{p}$ of the true order p.

Akaike's *AIC* takes the form

$$AIC(p) = \log \hat{\sigma}^2(p) + (2/T)\, p$$

thus setting $c_T = 2/T$, while the Hannan–Quinn (1979) *HQ* and Schwarz (1978) and Rissanen (1978) *SC* criteria (the latter also being known as the *BIC*) set c_T to $2 \log \log T/T$ and $\log T/T$ respectively. The criteria have the following properties. *AIC* asymptotically overestimates the true order p with positive probability, while *HQ* and *SC* estimate the order consistently as long as the maximum order considered exceeds the true order, a result that holds for both stationary and integrated processes (Paulsen, 1984). Denoting $\hat{p}(AIC)$ as the order selected by the *AIC*, etc., then for $T \geq 16$ the following inequality holds (Lütkepohl, 1991)

$$\hat{p}(SC) \leq \hat{p}(HQ) \leq \hat{p}(AIC)$$

Thus the *SC* will select the most parsimonious specification when different orders are chosen by the criteria.

These criteria were quickly adapted for ARMA processes and, eventually, for any type of model, univariate and multivariate, with obvious extensions to the notation. For example, the corresponding *AIC* for an ARMA(p,q) process is

$$AIC(p, q) = \log \hat{\sigma}^2(p, q) + (2/T)\,(p + q)$$

16.13 The Box–Pierce (1970) statistic introduced in §**9.46** to check the adequacy of a fitted ARMA process, $Q(K)=T\sum_{k=1}^{K} r_k^2(\hat{a})$, was shown by Davies, Trigg and Newbold (1977) to have actual significance levels that could be much smaller than those provided by the asymptotically justified $\chi^2(K-p-q)$ distribution, implying that the chance of incorrectly rejecting the null hypothesis of model adequacy would be smaller than the chosen significance level. Ljung and Box (1978) proposed the modified statistic

$$Q^*(K) = T(T+2)\sum_{k=1}^{K} \frac{r_k^2(\hat{a})}{T-k}$$

which they claimed would follow the asymptotic distribution more closely. However, the power of this modified statistic can still be quite low, even in the presence of severe misspecification (Davies and Newbold, 1979; Godfrey, 1979).

16.14 'Portmanteau' statistics of this type are derived without explicitly formulating an alternative hypothesis, which is a sensible strategy when there is little information available about the likely nature of any model misspecification. This may lead to problems, however, as a strong rejection of model adequacy might not provide much of a clue as to the type and magnitude of the inadequacy.

In response to this, tests were proposed that were based on the Lagrange Multiplier (LM) principle: see, for example, Godfrey (1979) and Poskitt and Tremayne (1980). Unlike the portmanteau statistics, an LM-based test requires an explicit alternative hypothesis, although the model given by this alternative does not need to be estimated, which is an attractive advantage of this type of test. For example, suppose the null hypothesis is that of an AR(p) process while the alternative is that of an AR($p+r$) process. Godfrey (1979) shows that the LM test can be calculated as the sample size, T, multiplied by the coefficient of determination (R^2) of an auxiliary regression using the fitted residuals from the AR(p) process, this being

$$\hat{a}_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + \beta_1 \hat{u}_{t-1} + \cdots + \beta_r \hat{u}_{t-r}$$

On the null hypothesis, $T\cdot R^2$ will be asymptotically distributed as $\chi^2(r)$, although for small samples an F-test of the hypothesis $\beta_1=\cdots=\beta_r=0$ has better properties (Kiviet, 1986). LM tests were developed for a wide range of alternatives and have since become a staple test of misspecification in a range of time series models: see Godfrey (1988).

McLeod (1978) derived the large sample distribution of the residual autocorrelations from an ARMA(p,q) model and suggested a further portmanteau test. Denoting the estimated covariance matrix of $\mathbf{r}' = (r_1(\hat{a}), \ldots, r_k(\hat{a}))$ as $\hat{\mathbf{C}}$, this statistic is defined as $T\mathbf{r}'\hat{\mathbf{C}}^{-1}\mathbf{r} \sim \chi^2(k)$. Newbold (1980) showed that this was equivalent to an LM test of the alternative ARMA$(p+k, q)$ specification.

Intervention analysis and the detection of outliers

16.15 In the iterative ARIMA model-building strategy developed by Box and Jenkins (1970) and discussed in §§**9.41–9.47**, little attention was paid to the behaviour of specific observations from the series under analysis, or to the residuals associated with them, with attention being devoted to the overall patterns of serial correlation. However, many time series are influenced by specific events and policy changes that occur at known points of time and ignoring these could lead to an inadequate model being fitted and poor forecasts being made.

Events of this type, whose timing are known, were termed *interventions* by Box and Tiao (1975), who incorporated them into, say, an ARIMA model by extending it to include deterministic, or 'dummy', input variables. For example, suppose that there is a single intervention, $I_t(\tau)$, known to have occurred at time τ. If x_t is generated by an ARMA(p, q) process then an intervention model may be defined as

$$x_t = \frac{\omega(B)}{\delta(B)} B^b I_t(\tau) + \frac{\theta(B)}{\phi(B)} a_t \tag{16.4}$$

The parameter b measures the delay in effect, or 'dead time', of the intervention, which itself is an indicator sequence taking the values of 1 and 0 to denote the occurrence or nonoccurrence of the exogenous intervention. Forms of intervention that have been found to be useful are

(i) A *pulse* or *spike*, which models an intervention lasting only for the observation τ: $I_t(\tau) = 1$ for $t = \tau$, and $I_t(\tau) = 0$ for $t \neq \tau$.
(ii) A *step*, which models a step change in x_t beginning at τ: $I_t(\tau) = 0$ for $t < \tau$ and $I_t(\tau) = 1$ for $t \geq \tau$.
(iii) An *extended pulse*, useful for modelling 'policy on–policy off' interventions, defined as $I_t(\tau_1, \tau_2) = 1$ for $\tau_1 \leq t \leq \tau_2$ and 0 otherwise.

Box and Tiao (1975) developed a procedure for identifying models of the form (16.4) and estimating them via nonlinear least squares and the models may naturally be extended to include several interventions. Box and Tiao's illustrative applications were in the economic and environmental areas, and Jenkins and McLeod (1982) later presented a number of case studies that employed intervention effects, such as changing price structures in the US telephone industry, competition between rail and air on London to Scotland passenger routes, and the influence of advertising promotions on product sales. An area that for a time became a favourite application was the effect of seatbelt legislation on road traffic accidents: see Bhattacharyaa and Layton (1979), Harvey and Durbin (1986) and Abraham (1987) for the analysis of Australian, British and Canadian data, respectively.

16.16 In these models the exact timing of the interventions were known, but often this is not the case and a variant of this methodology was developed for handling situations in which the exact timings of the events were unknown and whose effects lead to what are called aberrant observations or *outliers*. The analysis developed by Hillmer, Bell and Tiao (1983), Tiao (1985), Tsay (1986) and Chang, Tiao and Chen (1988) concentrated on identifying two types of outliers, additive and innovational. An *additive outlier* (AO) model is defined as

$$x_t = \omega I_t(\tau) + \frac{\theta(B)}{\phi(B)} a_t$$

while an *innovational outlier* (IO) model is

$$x_t = \frac{\theta(B)}{\phi(B)} (\omega I_t(\tau) + a_t)$$

where in both cases $I_t(\tau)$ is a pulse intervention. Thus the AO case may be called a 'gross error' model, since only the level of the τ^{th} observation is affected, whereas an IO represents an extraordinary shock at τ influencing $x_\tau, x_{\tau+1}, \ldots$ through the memory of the model, given by $\theta(B)/\phi(B)$.

Iterative methods of identifying outliers and building intervention models were developed in Tiao (1985) and Tsay (1986) and the approach was extended by Tsay (1988) to deal with both level shifts and variance changes.

Granger causality in the time domain

16.17 The concept of Granger causality was introduced in §§**13.17–13.20** within a frequency domain framework. It began to have a major impact on time series modelling, particularly in economics, after Sims (1972) 'translated' the idea into the time domain and provided a dynamic regression framework within which to discuss tests of causality, which have since become part of the standard practice of time series econometrics and discussions of which may be found in many textbooks: see, for example, Mills (1990, chapter 14.4) or Mills and Markellos (2008, chapter 8.6). The Granger causality literature subsequently expanded rapidly, with notable extensions and discussions being Chamberlain (1982), Granger (1980), Geweke (1982, 1984) and Dufour and Renault (1998).

16.18 The standard framework for investigating Granger causality became, during the 1980s, the vector autoregression (VAR), after the publication of Sims

(1980) (cf. §**12.23**), although Lütkepohl (1991, chapter 6.7) discussed Granger causality within a VARMA setting. Indeed, VARs, with their associated impulse response functions and variance decompositions, have since become the standard multivariate time series model: for a recent survey of the area, see Lütkepohl (2006).

Spurious regression and cointegration

16.19 The possibility of nonsense regressions, recognized by Yule, Slutzky and Working and discussed in Chapter 5, re-emerged in the 1970s with the publication of Granger and Newbold (1974) on the possibility of 'spurious regression', in which they considered the empirical consequences of regressing two *independent* random walks. Although the true regression coefficient must be zero, they found, in an influential simulation experiment, that a conventional t-test would reject this correct null approximately three quarters of the time using a standard 5% significance level. When five independent random walks were included as regressors in a multiple regression, the rejection rate of the conventional F-statistic for testing that the coefficient vector was zero was found to be over 95%. Further extensions involving correlated integrated processes produced similar results, leading Granger and Newbold to conclude that conventional significance tests were seriously biased towards rejection of the null hypothesis of no relationship and hence towards acceptance of a *spurious* relationship when time series were generated as statistically independent integrated processes.

Such regression results were frequently accompanied by large R^2 values and highly autocorrelated residuals, as indicated by very low Durbin-Watson (d) statistics: recall §**9.18**. These findings led Granger and Newbold to suggest that, in the joint circumstances of a high R^2 and a low Durbin–Watson statistic (a useful rule of thumb being $R^2 > d$), regressions should be run on the first differences of the variables, and further support for this suggestion was provided by Plosser and Schwert (1978).

These essentially empirical conclusions were given an analytical foundation by Phillips (1986), who showed that the regression coefficients do not converge in probability to constants as the sample size increases but have non-degenerate limiting distributions, so that different arbitrary large samples will yield randomly differing coefficient estimates. The conventional t-ratio does not have a t-distribution and, indeed, does not have any limiting distribution, so that there are *no* asymptotically correct critical values and it should thus be expected that the rejection rate when using conventional critical values will continue to increase with sample size. R^2 has a non-degenerate limiting distribution and the Durbin–Watson statistic converges in probability to zero, so that low values for d and moderate values for R^2 are to be expected in spurious regressions.

16.20 The spurious nature of regressions of this type is a consequence of the fact that the regression error, being a linear combination of $I(1)$ processes, must itself be $I(1)$. In general, if y_t and x_t are both $I(d)$ then the linear combination $u_t = y_t - ax_t$ will usually also be $I(d)$. It is possible, however, that u_t may be integrated of a lower order, say $I(d-b)$, in which case a special constraint operates on the long-run components of the two series. If $d=b=1$, so that y_t and x_t are both $I(1)$ and hence dominated by 'long wave' components, u_t will be $I(0)$ and so will not have such a component: y_t and ax_t must therefore have long-run components that cancel out to produce u_t. In such circumstances, y_t and x_t are said to be *cointegrated*, a concept first introduced in Granger (1981), although it should be emphasized that it will not generally be true that there will exist such an a that makes u_t stationary.

The concept of cointegration had an enormous impact on time series analysis with several tests for cointegration quickly being proposed (see, for example, Sargan and Bhargava, 1983; Engle and Granger, 1987; and Phillips and Ouliaris, 1990). A modelling strategy, based on *Granger's representation theorem*, was also developed, being known as *error correction modelling* (ECM: Engle and Granger, 1987; Granger, 1986). Mills and Markellos (2008, chapter 9), for example, provide further textbook discussion of testing and estimating cointegrating regressions. Long memory and cointegration may be combined to produce a fractional cointegration framework (see Gil-Alana and Hualde, 2009).

Vector error correction models

16.21 Setting cointegration within a VAR framework led to a further class of models and sets of testing and estimation techniques. This was first done in Johansen (1988a, 1988b) and subsequently led to the vector error correction model (VECM) representation of a set of multivariate cointegrated time series. Within this multivariate setting the presence of cointegration leads to many complications and refinements and is currently a dominant research area. Johansen (1995) and Juselius (2006) are prominent texts and Johansen (2006) provides a current overview of the *cointegrated VAR* model.

Forecast comparisons and evaluation

16.22 Forecasts of economic time series using the framework developed by Box and Jenkins, as set out in §§**11.3–11.15**, began to be compared with forecasts from much larger-scale econometric models in the early 1970s. Cooper (1972), Cooper and Nelson (1975), Nelson (1972) and Naylor, Seaks and Wichern (1972) all provided results that favoured the forecasts of the 'naïve' time series models over those from the larger models, findings that provoked a lengthy debate in the economics literature: see Granger and Newbold (1986, chapter 9.4) for discussion and also references to many other issues in forecasting, such as combining forecasts from different methods. At the same time Granger and Newbold (1973) questioned the evaluation criteria used by many econometric forecasters, arguing that these were insufficiently demanding. Out of this debate eventually arose a theory of economic forecasting which may be used to place the forecasting exercise on much firmer theoretical foundations (see, in particular, Clements and Hendry, 1998).

Spectral extensions

16.23 A comprehensive review of the theory of spectral analysis was provided by Priestley (1981). Developments since 1970 include *band spectrum regression*, in which components of variables in different frequency bands are related (Engle, 1974, 1980), and *frequency domain factor analysis*, which considers the question of whether the dynamic relationships between a number of variables can be explained by the presence of a small number of unobserved common factors (Geweke, 1977; Sargent and Sims, 1977). Spectral regression techniques were extended to cointegrated systems by Phillips (1989).

New developments

16.24 Of the numerous new developments in time series analysis in recent years, two areas of research particularly stand out. The first is the use of panel data sets, especially nonstationary panels, for which techniques began to be developed in the early 1990s. Choi (2006) and Banerjee and Wagner (2009) provide recent surveys of this burgeoning literature.

The second is the Bayesian treatment of time series models. Bayesian VARs began to be developed in the 1980s (see Litterman, 1986) and Bayesian analysis of time series models really took off in the 1990s as powerful computational tools, such as the Gibbs sampler and the Metropolis–Hastings algorithm, became available to a wider set of researchers to use in Markov chain Monte Carlo (MCMC) simulation (see Gelfand et al, 1990; Chib and Greenberg, 1995, 1996; and Poirier and Tobias, 2006). The Bayesian literature on cointegration is surveyed by Koop et al. (2006).

Some final thoughts

16.25 Studying the historical development of a subject should be a fascinating experience and it certainly has been for this author. While the first formal foundations of the subject are only just over a century old, and much of the development is contained in reasonably accessible journal articles, one or two unnoticed nuggets have been unearthed, notably the article by Bradford Smith on the implications of using detrended or first differenced variables in regression analysis (see §**12.5** and Mills, 2011).

Where does the subject go from here? There is no doubt that the underlying theory has advanced tremendously over recent years, but has this enhanced the quality of applied research? Some may doubt this, even as theory becomes ever more refined, data become more extensive, computing power increases rapidly and software becomes more powerful and accessible. On the other hand, advanced time series techniques are becoming more extensively used in a wider range of disciplines, a particular case in point being meteorology and climatology, where, for example, the use of time series techniques and modelling has helped to enhance understanding of some key issues (see, for example, Mills, 2010a, 2010b). It is to be hoped that the interaction between theory and practice in an expanding number of fields continues to advance the fascinating subject of time series analysis whose historical development has been the focus of this book.

Notes

2 Yule and Hooker and the Concepts of Correlation and Trend

1. Galton is often regarded as the father of correlation, but the concept – or at least the word – had been in common use, particularly in physics, since the middle of the century. The historical development of correlation, although fascinating, lies outside the scope of this study. A contemporary account of the development is provided by Pearson (1920): for a later, rather more detached, discussion, see Stigler (1986, chapter 8).
2. George Udny Yule plays a major and recurring role in our story. Born on 18 February 1871 in Beech Hill near Haddington, Scotland, Yule was a member of an established Scottish family composed of army officers, civil servants, scholars and administrators and both his father, also named George Udny, and a nephew were knighted. Although he originally studied engineering and physics at University College, London (UCL), and Bonn, Germany, publishing four papers on electric waves, Yule returned to UCL in 1893, becoming first a demonstrator for Karl Pearson, then an Assistant Professor. Yule left UCL in 1899 to work for the City and Guilds of London Institute but also later held the Newmarch Lectureship in Statistics at UCL. In 1912 he became lecturer in statistics at the University of Cambridge (later being promoted to Reader) and in 1913 began his long association with St. John's College, becoming a Fellow in 1922. Yule was also very active in the Royal Statistical Society: elected a Fellow in 1895, he served as Honorary Secretary, was President and was awarded the prestigious Guy Medal in Gold in 1911. His textbook, *Introduction to the Theory of Statistics*, ran to 14 editions during his lifetime and, as well as contributing massively to the foundations of time series analysis, he also researched on Mendelian inheritance and on the statistics of literary style, as well as other aspects of statistics. Retiring from his readership at the age of 60, and having always been a very fast driver, he decided to learn to fly, eventually buying his own plane and acquiring a pilot's licence. Unfortunately, from 1932 heart problems curtailed his flying experiences and he became a quasi-invalid for the rest of his life, dying on 26 June 1951 in Cambridge. For further biographical details and a full list of publications, see Kendall (1952) and also Williams (2004).
3. Historical perspectives on Yule's development of correlation and regression, which are not our major concern or focus here, but are arguably extremely important for the development of applied statistical techniques, are provided by Aldrich (1995, 1998) and Hepple (2001).
4. Expressing the k arrays in a tabular form gives what Yule refers to as a *correlation table*.
5. In terms of the original variables Y and X, the regression line is $Y = a + bX$ with $a = \overline{Y} - b\overline{X}$.
6. Or, as Yule (1897b, page 818) put it, 'errors of mean square, measuring the degree of scatter of the X's and Y's around their mean values'.
7. R_1^2 is, of course, the coefficient of multiple correlation in modern econometric parlance.
8. We follow Yule (1907) is using the notation $r_{12.3}$ here, rather than ρ_{12} as used in Yule (1897b).
9. Meteorological applications of correlation were also beginning to appear around this time: see Pearson and Lee (1897) and Cave-Browne-Cave and Pearson (1902).

10. Reginald Hawthorn Hooker (1867–1944) was a contemporary and close friend of Yule. Educated at the Collège Rollin, Paris, and Trinity College, Cambridge, Hooker was a career civil servant at the Board (later Ministry) of Agriculture, retiring in 1927, although before he joined the board in 1895 he was for four years Assistant Secretary of the Royal Statistical Society. A Fellow of both this and the Royal Meteorological Society, of which he was twice President, Hooker made contributions in the use of correlation to analyse economic, agricultural and meteorological data: see the obituary by Yule (1944a) for biographical detail.
11. Some years earlier, Poynting (1884) had used the ratio of a four-year moving average to a ten-year moving average of wheat prices and imports of cotton and silk to remove both secular movements and short-run fluctuations from the series, but his analysis was purely descriptive.
12. Hooker's results from using this technique were reported in further detail in Yule (1906), which also contained an extended discussion about the economic factors influencing marriage and birth-rates.
13. Hooker (1901b) used the notation $r_{m(t-k)}$, with m and t denoting the marriage rate and trade respectively. Hooker also constructed 'coefficient curves' in which a curve was interpolated through the scatterplot of $r_{m(t-k)}$ against k, from which he was able to estimate the maximum correlation and the (non-integer) value of k producing that maximum; we have not felt the desire to repeat this procedure here!
14. As time series data are now explicitly being considered, generic observations are denoted by t subscripts.

3 Schuster, Beveridge and Periodogram Analysis

1. Franz Arthur Friedrich Schuster (1851–1934) was born and educated in Frankfurt am Main, Germany before joining his parents in Manchester, where the family textile business was based, in 1870, becoming a British citizen in 1875. Independently wealthy, he studied physics at Heidelberg, Gottingen and Berlin before spending five years at the Cavendish Laboratory in Cambridge. In 1881 Schuster became Professor of Applied Mathematics and then, in 1887, Langworthy Professor of Physics, at Owens College, part of Manchester University. After retiring in 1907 (and being succeeded by Ernest Rutherford) he devoted the remainder of his professional life to promoting the cause of international science, being knighted in 1920. Schuster's interest in meteorology, earthquakes and terrestrial magnetism was longstanding and he took part in four eclipse expeditions.
2. The earliest reference to periodogram analysis appears to be in a note by G.G. Stokes to the paper by Stewart and Dodgson (1879) reporting on attempts to detect links between sunspots and magnetic and meteorological changes on earth.
3. A modern and comprehensive treatment of Fourier analysis may be found in Pollock (1999, chapters 13–15), where the derivation of these equations may be found.
4. The series is obtained from the National Geophysical Data Center (NGDC) website: www.ngdc.noaa.gov.
5. The periodogram is calculated with $s = 10$ and $p \leq 240$, so that all one-tenth of a year cycles up to 24 years are computed.
6. Schuster (1906) made a concerted attempt to investigate many of the minor peaks in the periodogram but, given the difficulties that later became apparent in trying to interpret calculated periodograms (see Chapters 5 and 13), we content ourselves with just this analysis.

7. William Beveridge, 1st Baron Beveridge (1879–1963), was a British economist and social reformer. After being educated at Charterhouse School and Balliol College, Oxford, Beveridge's varied career encompassed the law, journalism, the civil service and politics. Knighted in 1919, he was made a baron in 1946 and eventually became leader of the Liberals in the House of Lords. He was also a leading academic, being Director of the London School of Economics and Master of University College, Oxford. Probably best remembered for his 1942 report *Social Insurance and Allied Services* (known as the *Beveridge Report*), which served as the basis for the welfare state, especially the National Health Service, Beveridge was an expert on unemployment and labour markets. Although he had a long-standing interest in the history of prices, his work on periodogram analysis stands apart from his usual areas of research.
8. The periodogram was constructed for $s = 2$ and $p \leq 168$, so that all half-year cycles up to 84 years were calculated. This was felt to be as fine a 'mesh', to use Beveridge's terminology, as was needed for the purposes of the example. Beveridge used an even finer mesh and, given the rudimentary calculating technology available at the time, this obviously took considerable time and effort. The periodogram was recalculated by Gower (1955) using a computer for the first time: the computations took approximately two hours on the Manchester University Mark II Electronic Computer. The calculations reported here were produced using a Gauss program written by the author and took less than 0.008 seconds using a standard desktop PC.

4 Detrending and the Variate Differencing Method: Student, Pearson and Their Critics

1. William S. Gosset (1876–1937) was one of the most influential statisticians of the twentieth century. As an employee of Guinness, the famous Irish brewer of stout, he was precluded from publishing under his own name and thus took the pseudonym 'Student'. Gosset worked primarily on experimental design and small sample problems and was the inventor of the eponymous t-distribution. His research focus is tangential to our theme and this was his only paper on the analysis of time series. For further biographical details see Pearson (1950, 1990).
2. Karl Pearson (1857–1936) was another of the greats of statistics in the early twentieth century whose interest in time series was only tangential. His son, Egon, another extremely influential statistician, produced a biography of his father soon after Karl's death (Pearson, 1936/38, 1938). More recently, Porter (2004) concentrates on Pearson's life before 1900, focusing on his wide range of interests and the factors that influenced him into becoming a statistician. John Aldrich's Pearson website provides easily accessible detail and numerous links: www.economics.soton.ac.uk/staff/aldrich/kpreader.html.
3. Ritchie-Scott's (1915) contribution was restricted to providing an adjustment to this formula when additional observations were included to compensate for those lost due to differencing.
4. The data are tabulated in Cave and Pearson (1914, Table 1). An average, or 'synthetic', index was also provided. Because of the complications arising from the correlation between this average and its constituent indices, which are discussed in detail by Cave and Pearson, we exclude this average from our discussion of their results.
5. A further criticism was that the data must be very precisely measured to enable the accurate calculation of higher differences. Yule was certainly aware of this point,

and it particularly exercised Bowley (1920, page 376), who was of the view that the method was 'too refined and too sensitive for ordinary statistical analysis'.

6. The rather odd reference in the last sentence of the first paragraph of this quotation disparaging knighthoods may perhaps explain the fact that, according to Aldrich (1995, page 365), Pearson later refused a knighthood in 1935!
7. Sir Ronald Fisher also picked up this implication of the variate difference method: see his comments (pages 534–6) on Yule (1921) and also R.A. Fisher (1925).
8. For one series the trend was assumed to follow 'the compound interest law' and a nonlinear trend was fitted (in effect, a linear trend was fitted to the logarithms of the series).
9. The method of construction is described in Persons (1916).
10. The method actually employed was set out in Persons (1917, pages 612–13) and involved randomly drawing from tables of logarithms. In effect, Persons was creating random samples drawn from a uniform distribution ranging from 0 to 9. This is the distribution used to construct the sample series in our own simulations.
11. Yule admitted that he could not produce a proof of this result. It is, in fact, a special case of a more general result proved by Egon Pearson for a non-random series: see Pearson (1922, pages 37–40, and in particular his equation (xviii)).
12. 'Intercorrelated' is the term used by Pearson and Elderton for series that exhibit serial correlation, a term not coined until Yule (1926): see §**5.6.**
13. The equations were estimated using ordinary least squares in EViews 6, with the coefficients showing slight differences to those reported by Persons. Likewise, the correlations reported in Tables 4.4 and 4.5 below were also computed using EViews 6.
14. A more accessible derivation of these results than is given by O. Anderson, which is in German, may be found in T.W. Anderson (1971, chapter 3): see also Kendall, Stuart and Ord (1983, chapter 46.28).

5 Nonsense Correlations, Random Shocks and Induced Cycles: Yule, Slutzky and Working

1. The method used by Yule to produce his Figs. 5–9 is discussed in Yule (1926, page 55). We approximate it here to recreate these distributions in our composite Figure 5.5.
2. The calculations required to construct Figure 5.6 are outlined in Yule (1926, page 56).
3. For a derivation of this result in the more general context of calculating 'intraclass' correlation, see Yule and Kendall (1950, §§**11.40–11.41**).
4. A small Gauss program was written to simulate Yule's sampling procedure and to compute the results shown in Table 5.1 and later in Tables 5.4 and 5.7. Necessarily, the results differ numerically from those of Yule because of the sampling process.
5. Yule states that the maximum negative correlation is that between 'terms 2 and 8 or 3 and 9, and is −0.988', which is clearly a misreading of his Table VI, which gives correlations to three decimal places, unlike our Table 5.6, which retains just two to maintain consistency with earlier tables.
6. The serial correlations were obtained using the correlogram view in EViews 6. Compared to Yule's own heroic calculations, which he described in detail (Yule, 1926, page 42) and reported as Table XIII and Fig. 19, they are uniformally smaller but display an identical pattern.
7. Eugen Slutzky (1880–1948), as his name is given on his 1937 *Econometrica* paper (Slutzky, 1937), a translation and update of Slutzky (1927), which was written in Russian, was a Ukrainian statistician and economist. Slutzky's contributions to time

series analysis have recently been revisited by Barnett (2006). Also known as Eugene Slutsky, his other major contribution was to introduce in 1915 the 'Slutsky decomposition' of demand functions into substitution and income effects. Although ignored at the time, it was famously resurrected some two decades later by Hicks and Allen (1934) and Allen (1936), who coined the term which has since become a cornerstone of demand theory in microeconomics: see also Allen (1950). Weber (1999) and Barnett (2004) provide recent appreciations of Slutzky's work in economics and statistics.

8. We shall continue to use Yule's terminology of serial correlation as this has since entered into the lexicon of time series analysis.
9. See Slutzky (*ibid.*, page 108) for more details.
10. The method used by Working to randomly draw such normal variates was discussed in detail in (*ibid,*, pages 16–20).
11. Chartism and technical analysis as a means of forecasting financial markets have gathered many proponents and detractors over the years: for an accessible discussion by a member of the proponent camp, see Plummer (1989); for a well-known and entertaining rebuttal, see Malkiel (2007).

6 Periodicities in Sunspots and Air Pressure: Yule, Walker and the Modelling of Superposed Fluctuations and Disturbances

1. Yule (1927, pages 284–6) discussed further features of the sunspot numbers and their related disturbances estimated from equation (6.10). These features do not seem to appear in the extended series and are therefore not discussed here.
2. Indeed, the unusual behaviour of sunspots over the last decades of the twentieth century has been a subject of great interest: see, for example, Solanki *et al.* (2004).
3. Sir Gilbert Thomas Walker (1868–1958) was, successively, a fellow of Trinity College and lecturer in mathematics, Cambridge University, Director of the Indian Meteorological Department, and Professor of Meteorology at Imperial College, London. Knighted in 1924 on his return to England, Walker was perhaps the first to rigorously employ statistical techniques in attempting to forecast meteorological variables, notably monsoon rains. He also published in many other areas, including the flight of birds and on mathematical aspects of sports and games, where he was a particular expert with boomerangs, earning the sobriquet 'Boomerang Walker' at Cambridge. For biographical details and more on his statistical research in meteorology, see Walker (1997) and Katz (2002).

7 The Formal Modelling of Stationary Time Series: Wold and the Russians

1. Herman Ole Andreas Wold (1908–1992) was a Swedish statistician who held chairs at the universities of Uppsala and Gothenburg. As well as his early research in time series analysis (Wold, 1938, was his doctoral dissertation), he was also known for his work in developing partial least squares and causal chain modelling. For biographical details, see Wold (1982) and Hendry and Morgan (1994).
2. Aleksandr Khinchin (1894–1959) was a significant figure in the Soviet school of probability theory. For a bibliography of his research in probability, see Cramér (1962). Andrey Kolmogorov (1903–1987) was a major figure in twentieth-century mathematics, making seminal contributions in topology, turbulence and mechanics as well

as in probability. In further related research, he established the basic theorems for smoothing and predicting stationary stochastic processes, thus forming the basis for the famous Weiner–Kolmogorov prediction formulae (Kolmogorov, 1941): see §**16.4**.

3. Wold refers to the standard deviation σ as the dispersion, which he denotes as D. We choose to use the more familiar term variance for the 'squared dispersion': in a similar vein, we denote the mean as μ rather than use Wold's m.
4. Wold (1938) does not refer to this theorem as a decomposition (indeed, the word does not appear in the index). Whittle (1954), in Appendix 2 of the second edition (Wold, 1954, page 106), appears to be the first to refer to it as such.
5. The Port Darwin air pressure data is not directly available and Figure 6.8 was drawn by reading off the values of the series from the original plot in Walker (1931). Interestingly, although the data are thus subject to a degree of measurement error, a linear regression of air pressure on lagged air pressure gives a coefficient estimate of 0.74.
6. There are some minor differences between the serial correlations reported by Wold in his Table 6 (page 151) and those computed here using EViews 6. In particular, Wold's estimate for r_1 is 0.614.
7. As these estimates are extremely close to those obtained using EViews 6, which were 0.523 and −0.223, Wold's values will continue to be used.

8 Generalizations and Extensions of Stationary Autoregressive Models: From Kendall to Box and Jenkins

1. Sir Maurice George Kendall (1907–1983) was a major UK figure in twentieth-century statistics. A close friend of the ageing Yule after becoming co-author of his textbook, Kendall is perhaps best known for his co-authorship of the *Advanced Theory of Statistics*, Kendall, Stuart and Ord (1983) being the fourth edition. He made seminal contributions in many fields of statistics apart from time series analysis: for example, in random number generation, multivariate analysis, rank correlation and *k*-statistics. For biographical details see Ord (1984) and Stuart (1984).
2. We revert to the notation that has since become standard in time series analysis.
3. Kendall eliminated trends in the series by taking nine-year moving averages, a method suggested by his earlier analysis of detrending oscillatory time series (Kendall, 1941). The issue of the appropriate method of detrending will be returned to in Chapter 10. Serial and partial correlations are calculated using EViews 6, rather than using those reported by Kendall.
4. The 'significance test' employed by Kendall would appear to be that proposed by Bartlett (1935). Inferential issues in time series analysis are developed further in Chapter 9.
5. Most of Yule's publications from 1939 onwards were philologically based, principally related to the occurrence of words as a means of identifying the statistical characteristics of an author's style: see, in particular, Yule (1944b).
6. George E.P. Box (born 1919) is an English statistician who has made fundamental contributions in the areas of quality control, the design of experiments and Bayesian inference as well as time series analysis. Gwilym M. Jenkins (1933–1982) was a Welsh statistician and systems engineer. For autobiographical details of Box and his relationship with Jenkins see DeGroot (1987) and Peña (2001): also see Box's (1983) obituary of Jenkins. The Box–Jenkins approach to time series modelling building is extensively discussed in later chapters.

9 Statistical Inference, Estimation and Model Building for Stationary Time Series

1. A Markov (sometimes spelt Markoff) process is a (possibly multivariate) stochastic process such that the conditional probability distribution for the state at any future time, given the present state, is unaffected by any additional knowledge of the past history of the system. The AR(1) scheme is thus seen to be an example of a univariate Markov process. The process was named after the Russian probabalist Andrei Andreevich Markov (1856–1922).
2. Later studies by Craddock (1965) and Morris (1977) suggested that an autoregression of order nine was required to model this series and evidence in favour of such a scheme is presented later in §§**9.43–9.46**. There have also been many attempts to fit nonlinear schemes to the sunspot index: see the references contained in Morris (1977), for example.
3. A number of papers, beginning with Anderson (1942), refer to the cyclic definition as being suggested by Harold Hotelling although, intriguingly, no reference to a paper by Hotelling is ever given! The genesis of this 'suggestion' may have been the theorem on circularly distributed variates on pages 371–2 of Hotelling (1936) and, more generally, the discussion contained in sections 12–15 of that paper. The cyclic definition was used regularly in the theoretical developments to be discussed later in this chapter as it enabled tractable sampling distributions to be obtained.
4. A formal definition of this order notation is that $f(T)$ will be $O(g(T))$ if and only if there exist positive real numbers M and T_0 such that $|f(T)| \leq M|g(T)|$ for all $T > T_0$.
5. Anderson (1942) constructed this figure and associated critical values using the asymptotic variance. Dixon (1944) also provided critical values based on a Pearson Type I approximation.
6. Quenouille (1948) observed that this is, in fact, the distribution of the ordinary correlation coefficient based on $T+3$ observations.
7. These moment assumptions have since been progressively weakened. Anderson and Walker (1964) show that only a finite second moment is, in fact, required.
8. Walker (1961, Appendix) provided a more complicated adjustment for the bias in this estimator. For this simulation it leads to a bias adjustment of the order 0.02, which would almost eradicate the bias. He also showed that Durbin's adjustment would be approximately 0.004, which is what we find.
9. The likelihood principle states that everything the data has to tell about the parameters of an assumed model is contained in the likelihood function, with all other aspects of the data being irrelevant. From a Bayesian perspective, the likelihood function is that part of the posterior distribution of the parameters which comes from the data. Although the principle has by no means uniform support amongst statisticians, it does underpin a large body of modern statistical analysis: see Barnard, Jenkins and Winsten (1962) for a contemporary discussion of its importance to time series analysis.

10 Dealing with Nonstationarity: Detrending, Smoothing and Differencing

1. Working concurrently with Hooker, John Norton (1902) superimposed linear time trends on graphs of banking data from which he calculated percentage deviations (see Klein, 1997, pages 236–40, for discussion).

2. That the penalized least squares approach, in various forms, anteceded the H–P filter by several decades was well known to both Leser and Hodrick and Prescott, although the latter appear to be unaware of Leser's paper. Pedregal and Young (2001) provide both a historical and a multidisciplinary perspective. The Hodrick and Prescott paper was first circulated as a working paper in 1980, only being published some 17 years later as a journal article and many years after it had become a staple approach to detrending macroeconomic time series. Further discussion of the H–P filter is provided in §**16.5**.
3. This is the approach set out in Kendall (1946). Hall (1925) suggested the same approach but restricted attention to local linear trends and failed to make the link with moving averages as set out below. Interestingly, however, he introduced moving sums to remove seasonal and cyclical components, referring to these as moving integrals and the process of computing them as moving integration, thus predating the use of the term integrated process to refer to a cumulated variable, introduced by Box and Jenkins (see §**10.18**), by some four decades.
4. A fuller discussion of this approach to trend fitting may be found in Kendall, Stuart and Ord (1983, §§46.4–46.7), whose derivation is followed here. Further, although rather arcane, properties of the method are provided by Kendall (1961). A recent application of the approach is to be found in Mills (2007), where it is used to obtain recent trends in temperatures.
5. An extension of this approach was suggested by Quenouille (1949b), who dispensed with the requirement that the local polynomial trends should smoothly 'fit together', thus allowing discontinuities in their first derivatives. As with many of Quenouille's contributions, this approach was both technically and computationally demanding and does not seem to have captured the attention of practitioners of trend fitting!
6. Spencer-Smith (1947, page 105) returned to this point about moving averaging inducing spurious oscillations, but defined trend in a manner rather different to that considered here, being more akin to a long cycle: 'such series do not contain very prolonged steady increases or decreases in the general values of these terms, as may happen in economic series, and where such movements occur the use of the moving average method may be valid'.
7. Figure 10.4 is a reconstruction of the (unnumbered) figure on page 74 of Working and Hotelling (1929). As the actual potato yield values are not given, for the purposes of constructing the observed series and computing the trend and its probable error, they have been 'estimated' from the original figure. Parameter estimates and standard errors thus differ from those reported by Working and Hotelling.
8. The main thrust of Fisher's paper was to investigate the dynamic relationship between prices and trade and this led him to develop the concept of *distributed lags*, which has since become a basic technique in time series econometrics. This aspect of the paper is returned to in §**12.6**. Hendry and Morgan (1995, pages 45–8) reanalyse and reinterpret Fisher's econometric modelling.
9. As well as Tintner's (1940) advocacy of the variate differencing method for inducing stationarity, as discussed in Chapter 4, Yaglom (1955) also advocated the use of differences, while a number of econometricians also proposed the differencing of variables in regression analysis. Tintner and Kadekodi (1973) provide numerous references to research in these areas during the period 1940 to 1970.
10. Pearson's metaphor was, of course, in terms of *spatial* displacement, but the time series analogy should be clear. Random walks were, in fact, first formally introduced in 1900 by Louis Bachelier in his doctoral dissertation *Théorie de Speculation*,

although he never used the term. Under the supervision of Henri Poincaré, Bachelier developed the mathematical framework of random walks in continuous time (where it is termed Brownian motion) in order to describe the unpredictable evolution of stock prices (biographical details of Bachelier may be found in Mandelbrot, 1989: see also Dimand, 1993). The dissertation remained unknown until it was rediscovered in the mid-1950s after the mathematical statistician Jimmie Savage had come across a later book by Bachelier on speculation and investment (Bachelier, 1914). A translation of the dissertation by James Boness was eventually published as Bachelier (1964). Random walks were independently discovered by Albert Einstein in 1905 and, of course, have since played a fundamental role in mathematics and physics as models of, for example, waiting times, limiting diffusion processes, and first-passage-time problems: see Weiss (1986).

11 Forecasting Nonstationary Time Series

1. The words 'predictable' and 'forecastable' are used interchangeably throughout this chapter and, indeed, throughout the book, even though in the modern literature on the econometrics of forecasting they have different, albeit subtle and rather deep, definitions. For a detailed discussion of these different definitions, see Clements and Hendry (1998, chapter 2). Briefly, predictability is defined as a property of a random variable in relation to an information set (the conditional and unconditional distributions of the variable do not coincide). It is a necessary, but not sufficient, condition for forecastability, as the latter requires knowledge of what information is relevant and how it enters the causal mechanism. Along with the great majority of researchers and practitioners, such technical niceties will be ignored here and the two words will continue to be regarded as synonyms.
2. This paper also includes Samuelson's memorable and often repeated quote that the 'Dow Jones index has predicted *nine* of the last five recessions' (*ibid.*, page 5).
3. Estimation of the autoregressive parameter obtains a value of 0.813 for ϕ, but the value of 0.8 continues to be used for simplicity.
4. Attention is focused here on the non-seasonal variants of exponential smoothing. Discussion of the seasonal models is to be found in §**14.18**.
5. Cogger (1974) was later to show that *n*th-order exponential smoothing was optimal for an ARIMA(0,*n*,*n*) process with equal moving average roots.

12 Modelling Dynamic Relationships Between Time Series

1. δ^2/s^2 is the von Neumann (1941, 1942) ratio which was proposed as a test of (first-order) autocorrelation and was a forerunner of the Durbin–Watson test (see §**9.18**).
2. Hannan's efficiency analysis required some knowledge of the distribution of these statistics when the null of no correlation was false, in particular the expected values under the alternative. With regard to actual distributions, McGregor (1962) and McGregor and Bielenstein (1965) derived the distribution of the correlation coefficient between two normal AR(1) processes, while Ames and Reiter (1961) obtained some sampling distributions via a further simulation experiment.
3. Bradford Bixby Smith was an economist employed by the United States Bureau of Agricultural Economics. As will be seen from these comprehensive excerpts from Smith (1926), it is clear that his ideas were half a century ahead of their time, covering as

they do almost all of the concerns about detrending and differencing that were later raised by the numerous econometricians researching this area during the decade and a half from the mid-1970s to the end of the 1980s (see Mills, 2011, and the discussion in §**16.2**). Yet a JSTOR search reveals not a single reference to the paper!

Interestingly, this search (for 'Bradford Smith') produced a single reference to an earlier paper (Smith, 1925) in which he advocated that trends and seasonal patterns should not be eliminated before time series were correlated: 'the unconsidered practice of eliminating trend and seasonal from series prior to their correlation is to be looked upon askance, therefore. It is often a serious error' (Smith, 1925, page 543) and '… correlation coefficients secured by *simultaneous*, or multiple, correlation methods will be as high or higher, and never less, than those resulting from any possible sequence of *consecutive* elimination of the influence of independent factors from the dependent, of which current methods of eliminating seasonal variations *before* correlating are an example' (Smith, 1925, page 545). For holding these views, Smith was taken to task by Frisch and Waugh (1933), who showed that it was irrelevant whether such trends and seasonal movements were eliminated before regressing two series or whether they were included as regressors in a multiple regression in the sense that the resulting coefficient estimates would be identical. The resulting 'Frisch–Waugh Theorem' was the basis for Lovell's (1963) theory of seasonal adjustment by regression methods (see §**14.16**). Notwithstanding this error of interpretation (but see Mills, 2011, for mitigatory arguments from a modern perspective), it is clear that Bradford Smith was a prescient analyser of economic time series and that his work should be more widely known amongst both practitioners of econometric time series modelling and historians of the subject.

4. The second reason that Fisher's paper was of particular importance was for his interpretation of the economic 'business cycle' which, although certainly of historical interest, is somewhat tangential to the present focus. Fisher concluded from his statistical analysis of the relationship between trade and price changes that the business cycle was

> simply the fluctuation about its own mean [rather than] a regular succession of *similar* fluctuations, constituting some sort of *recurrence*, so that, as in the case of the phases of the moon, the tides of the sea, wave motion, or pendulum swing, we can forecast the future on the basis of a pattern worked out from past experience and which we have reason to think will be copied in the future. We certainly cannot do that in predicting the weather or Monte Carlo luck. Can we do so as to business? Not so long as business is dominated by changes in the price level! For changes in the price level show no regular occurrence … .
>
> In short, if the one *non-cyclical* or irregular factor, price-change, can so nearly explain the behaviour of business, there is little room left for any *cyclical*, or regular, factors, especially as there must be numerous other non-cyclical ones always at work. (I. Fisher, 1925, page 192: italics in original)

Fisher developed an extended physical analogue for business cycle movements which may be contrasted with the randomly struck pendulum offered almost contemporaneously by Yule (1927) (§**6.1**):

> it would be not the swing of a clock pendulum but the swaying of the trees or their branches. If, in the woods, we pull a twig and let it snap back, we set up a swaying movement back and forth. That is a real cycle, but if there is no further disturbance, the swaying soon ceases and the twig becomes motionless again.

> In actual experience, however, twigs or tree-tops seldom oscillate so regularly, even temporarily. They register, instead, chiefly the variations in wind velocity. A steady wind may keep the tree for weeks at a time, leaning almost continuously in one direction and its natural tendency to swing back is thereby defeated or blurred. Its degree of bending simply varies with the wind. That is, the inherent pendulum tendency is ever being smothered. In the net resultant motion there is no perceptible trace left of real recurrences and no would-be forecaster would attempt to base his estimate of the future behaviour of a tree-top on a statistical average period of recorded swayings taken in gusts of wind.' (*ibid.*, page 192)

He then extended the analogy to take into account the outside forces that are necessary to counteract 'frictions' that would otherwise bring the system to rest: 'these … will not perpetuate, but obfuscate, the cycle, like the wind blowing on the trees' (*ibid.*, page 192). Consequently, business fluctuations were characterized not by a clock pendulum but by 'a bough of a tree, in the grip of outside forces of which the chief … was the rise and fall of the price level, *the dance of the dollar*' (*ibid.*, page 194: italics added for emphasis). Thus, rather than an oscillatory autoregressive process of the type suggested by Yule, business fluctuations were more akin to random variations, reacting to various extraneous and unpredictable forces.

5. Hendry and Morgan (1995, pages 45–8) attempted to reproduce Fisher's statistical analysis, uncovering a number of problems and inconsistencies that, in hindsight, cast considerable doubt on the veracity of the conclusions drawn by Fisher.
6. Distributed lags had also made a considerable impact in econometrics through the work of, most notably, Koyck (1954), Jorgensen (1963) and Almon (1965).
7. Box and Jenkins (1970, pages 384–6) suggested an alternative procedure for identifying the noise through the prewhitened input and output.
8. In modern terminology, these schemes are known as the vector AR, vector MA and vector ARMA models, with common acronyms VAR, VMA and VARMA, but such terms only came into common parlance some two decades after the publication of Quenouille's monograph: see, for example, Anderson (1980), Sims (1980), Tiao and Box (1981) and the book by Lütkepohl (1991).

13 Spectral Analysis of Time Series: The Periodogram Revisited and Reclaimed

1. One of the difficulties with spectral analysis is the multitude of different notations and definitions that are available. A consistent notation has been adhered to throughout this chapter which may differ from that used in the cited references.

14 Tackling Seasonal Patterns in Time Series

1. Hylleberg (1992) provides a convenient and useful schematic presentation, complete with references, of Mendershausen's survey.
2. Concurrently with the Census Bureau, the US Bureau of Labor Statistics developed their own technique to seasonally adjust monthly employment, unemployment and labour force data: see BLS (1966) for details. Although this method certainly provoked interest amongst statisticians (see, for example, Nerlove, 1965, and Young, 1968), it never found its way into as widespread use as X-11.

3. Estimation of the model in EViews 6 obtained the estimates $\hat{\theta} = 0.403$, $\hat{\Theta} = 0.636$ and $\hat{\sigma}_a^2 = 1.332 \times 10^{-3}$.

15 Emerging Themes

1. The Cauchy distribution is the ratio of two independent standard normal distributions and has no defined moments, although its location parameter defines both the median and the mode. Allowing the initial value of x to be non-zero complicates this limiting distribution, although White (1958, equation (4.25)) was still able to provide an explicit form for it.
2. Gauss' original derivation of the *recursive least squares* (RLS) algorithm is given in Young (1984, Appendix 2), which provides the authorized French translation of 1855 by Bertrand along with comments by the author designed to 'translate' the derivation into modern statistical notation and terminology. Young (2010, Appendix A) provides a simple vector-matrix derivation of the RLS algorithm.
3. Hald (1981) and Lauritzen (1981, 2002) claim that T.N. Thiele, a Danish astronomer and actuary, proposed in an 1880 paper a recursive procedure for estimating the parameters of a model that contained, as we know them today, a regression component, a Brownian motion and a white noise, that was a direct antecedent of the Kalman filter. Closer antecedents were Peter Swerling, an eminent radar theoretician then working for the RAND Corporation (see Swerling, 1959) and the Soviet mathematician Ruslan Stratonovich, who had developed a more general nonlinear filter of which Kalman's filter was a linear special case (see Stratonovich, 1960).

References

Abraham, B. (1987). 'Application of intervention analysis to a road fatality series in Ontario'. *Journal of Forecasting* 6, 211–20.

Akaike, H. (1969). 'Fitting autoregressive models for prediction'. *Annals of the Institute of Statistical Mathematics* 21, 243–7.

Akaike, H. (1971). 'Autoregressive model fitting for control'. *Annals of the Institute of Statistical Mathematics* 23, 163–80.

Akaike, H. (1974). 'A new look at the statistical model identification'. *IEEE Transactions on Automatic Control* AC-19, 716–23.

Aldrich, J. (1995). 'Correlations genuine and spurious in Pearson and Yule'. *Statistical Science* 10, 364–76.

Aldrich, J. (1998). 'Doing least squares: Perspectives from Gauss and Yule'. *International Statistical Review* 68, 155–72.

Alexander, S.S. (1961). 'Price movements in speculative markets: trends or random walks'. *Industrial Management Review* 2, 7–26.

Alexander, S.S. (1964). 'Price movements in speculative markets: trends or random walks, Number 2'. *Industrial Management Review* 5, 25–46.

Allen, R.G.D. (1936). 'Professor Slutsky's theory of consumer choice'. *Review of Economic Studies* 3, 120–9.

Allen, R.G.D. (1950). 'The work of Eugen Slutsky'. *Econometrica* 18, 209–16.

Almon, S. (1965). 'The distributed lag between capital appropriations and expenditures'. *Econometrica* 33, 178–96.

Ames, E. and Reiter, S. (1961). 'Distributions of correlation coefficients in economic time series'. *Journal of the American Statistical Association* 56, 637–56.

Anderson, B.D.O. and Moore, J.B. (1979). *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall.

Anderson, O. (1914). 'Nochmals Über "The elimination of spurious correlations due to position in time and space"'. *Biometrika* 10, 269–79.

Anderson, O. (1923). 'Uber ein neues verfahren bei anwendung der "variate-difference" methode'. *Biometrika*, 15, 134–49.

Anderson, O. (1926). 'Uber die anwendung der differenzenmethode ("variate difference method") bei reihenausgleichungen, stabilitätsuntersuchungen und korrelationsmessungen. Part 1'. *Biometrika* 18, 293–320.

Anderson, O. (1927a). 'Uber die anwendung der differenzenmethode ("variate difference method") bei reihenausgleichungen, stabilitätsuntersuchungen und korrelationsmessungen. Part 2'. *Biometrika* 19, 53–86.

Anderson, O. (1927b). 'On the logic of the decomposition of statistical series into separate components'. *Journal of the Royal Statistical Society* 90, 548–69.

Anderson, O. (1929). *Die Korrelationsrechnung in der Konjunkturforschung*. Bonn: Scroeder.

Anderson, R.L. (1942). 'Distribution of the serial correlation coefficient'. *Annals of Mathematical Statistics* 13, 1–13.

Anderson, T.W. (1959). 'On asymptotic distributions of estimates of parameters of stochastic difference equations'. *Annals of Mathematical Statistics* 30, 676–87.

Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.

Anderson, T.W. (1980). 'Maximum likelihood estimation for vector autoregressive moving average models', in D.R. Brillinger and G.C. Tiao (eds), *Directions in Time Series*. Institute of Mathematical Statistics, pp. 49–59.

Anderson, T.W. and Walker, A.M. (1964). 'On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process'. *Annals of Mathematical Statistics* 35, 1296–303.

Babbage, C. (1856). 'Analysis of the statistics of the Clearing House during the year 1839'. *Journal of the Statistical Society of London* 19, 28–48.

Bachelier, L.J.B. (1914). *Le Jeu, la Chance, et le Hasard*. Paris: E. Flammarion.

Bachelier, L.J.B. (1964, [1900]). 'Theory of speculation', in P. Cootner (ed.), *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press, pp. 17–78.

Baillie, R.T. (1996). 'Long memory processes and fractional integration in econometrics'. *Journal of Econometrics* 73, 5–59.

Baillie, R.T. (2006). 'Modelling volatility', in T.C. Mills and K.D. Patterson (eds) *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 737–64.

Banerjee, A. and Wagner, M. (2009). 'Panel methods to test for unit roots and cointegration', in T.C. Mills and K.D. Patterson (eds) *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*. Basingstoke: Palgrave Macmillan, pp. 632–728.

Barnard, G.A., Jenkins, G.M. and Winsten, C.B. (1962). 'Likelihood inference in time series'. *Journal of the Royal Statistical Society, Series A* 125, 321–72.

Barnett, V. (2004). 'E.E. Slutsky: Mathematical statistician, economist, and political economist?' *Journal of the History of Economic Thought* 26, 5–18.

Barnett, V. (2006). 'Chancing an interpretation: Slutsky's random cycles revisited'. *European Journal of the History of Economic Thought* 13, 411–32.

Bartlett, M.S. (1935). 'Some aspects of the time-correlation problem in regards to tests of significance'. *Journal of the Royal Statistical Society* 98, 536–43.

Bartlett, M.S. (1946). 'On the theoretical specification and sampling properties of autocorrelated time series'. *Journal of the Royal Statistical Society, Series B, Supplement* 8, 27–41.

Bartlett, M.S. (1948). 'Smoothing periodograms from time series with continuous spectra'. *Nature* 161, 686–87.

Bartlett, M.S. (1950). 'Periodogram analysis and continuous spectra'. *Biometrika* 37, 1–16.

Bartlett, M.S. (1955). *Stochastic Processes*. Cambridge: Cambridge University Press.

Bartlett, M.S. and Diananda, P.H. (1950). 'Extensions of Quenouille's test for autoregressive schemes'. *Journal of the Royal Statistical Society, Series B* 12, 108–15.

Bartlett, M.S. and Medhi, J. (1955). 'On the efficiency of procedures for smoothing periodograms from time-series with continuous spectra'. *Biometrika* 42, 143–50.

Bartlett, M.S. and Rajalakshman, D.V. (1953). 'Goodness-of-fit-tests for simultaneous autoregressive series'. *Journal of the Royal Statistical Society, Series B* 15, 107–24.

Barton, H.C. (1941). 'Adjustment for seasonal variation'. *Federal Reserve Bulletin* 27, 518–28.

Baxter, M. and King, R.G. (1999). 'Measuring business cycles: approximate band-pass filters for economic time series'. *Review of Economics and Statistics* 81, 575–93.

Bell, W.R. (1984). 'Signal extraction for nonstationary time series'. *Annals of Statistics* 12, 646–64.

Bell, W.R. and Hillmer, S.C. (1984). 'Issues involved with the seasonal adjustment of economic time series'. *Journal of Business and Economic Statistics* 2, 291–320.

Bell, W.R. and Martin, D.E.K. (2004). 'Computation of asymmetric signal extraction filters and mean squared error for ARIMA component models'. *Journal of Time Series Analysis* 25, 603–25.

Beran, J.A. (1992). 'Statistical methods for data with long-range dependence'. *Statistical Science* 7, 404–27.

Beveridge, S. and Nelson, C.R. (1981). 'A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the "business cycle"'. *Journal of Monetary Economics* 7, 151–74.

Beveridge, W. (1921). 'Weather and harvest cycles'. *Economic Journal* 31, 429–52.

Beveridge, W. (1922). 'Wheat prices and rainfall in Western Europe'. *Journal of the Royal Statistical Society* 85, 412–78.

Bhattacharyaa, M.N. and Layton, A.P. (1979). 'Effectiveness of seat belt legislation on the Queensland road toll – an Australian case study in intervention analysis'. *Journal of the American Statistical Association* 74, 596–603.

Blackman, R.B. and Tukey, J.W. (1958). *The Measurement of Power Spectra from the Point of View of Communication Engineering.* New York: Dover Publications, Inc.

BLS (1966). *The BLS Seasonal Factor Method.* Washington, DC: Department of Labor, Bureau of Labor Statistics.

Bowley, A.L. (1920). *Elements of Statistics*, 4th edition. London: P.S. King and Sons, Ltd.

Box, G.E.P. (1983). 'G.M. Jenkins, 1933–1982'. *Journal of the Royal Statistical Society, Series A* 146, 205–6.

Box, G.E.P., Hillmer, S.C. and Tiao, G.C. (1978). 'Analysis and modeling of seasonal time series', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series.* Washington, DC: US Dept of the Commerce, Bureau of the Census, pp. 309–34.

Box, G.E.P. and Jenkins, G.M. (1962). 'Some statistical aspects of adaptive optimization and control'. *Journal of the Royal Statistical Society, Series B* 24, 297–343.

Box, G.E.P. and Jenkins, G.M. (1968). 'Some recent advances in forecasting and control'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 17, 91–109.

Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Box, G.E.P. and Jenkins, G.M. (1973). 'Some comments on a paper by Chatfield and Prothero and on a review by Kendall'. *Journal of the Royal Statistical Society, Series A* 136, 337–52.

Box, G.E.P. and Newbold, P. (1971). 'Some comments on a paper of Coen, Gomme and Kendall'. *Journal of the Royal Statistical Society, Series A* 134, 229–40.

Box, G.E.P, and Pierce, D.A. (1970). 'Distribution of the residual autocorrelations in autoregressive-moving average time series models'. *Journal of the American Statistical Association* 64, 1509–26.

Box, G.E.P., Pierce, D.A. and Newbold, P. (1987). 'Estimating current trend and growth rates in seasonal time series'. *Journal of the American Statistical Association* 82, 276–82.

Box, G.E.P. and Tiao, G.C. (1975). 'Intervention analysis with application to economic and environmental problems'. *Journal of the American Statistical Association* 70, 70–9.

Brigham, E.O. (1974). *The Fast Fourier Transform.* Englewood Cliffs, NJ: Prentice-Hall.

Brillinger, D.R. (1965). 'An introduction to polyspectra'. *Annals of Mathematical Statistics* 36, 1351–74.

Brooks, C. (2006). 'Multivariate volatility models', in T.C. Mills and K.D. Patterson (eds) *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 756–83.

Brown, R.G. (1959). *Statistical Forecasting for Inventory Control.* New York: McGraw-Hill.

Brown, R.G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time Series.* Englewood Cliffs, NJ: Prentice-Hall.

Brown, R.G. and Meyer, R.F. (1961). 'The fundamental theorem of exponential smoothing'. *Operations Research* 9, 673–87.

Bullock, C.J., Persons, W.M. and Crum, W.L. (1927). 'The construction and interpretation of the Harvard Index of Business Conditions'. *Review of Economics and Statistics* 9, 74–92.

Burman, J.P. (1965). 'Moving seasonal adjustment of economic time series'. *Journal of the Royal Statistical Society, Series A* 128, 534–58.

Burman, J.P. (1980). 'Seasonal adjustment by signal extraction'. *Journal of the Royal Statistical Society, Series A* 143, 321–37.

Burridge, P. and Wallis, K.F. (1984). 'Unobserved-components models for seasonal adjustment filters'. *Journal of Business and Economic Statistics* 2, 350–9.

Burns, A.F. and Mitchell, W.C. (1946). *Measuring Business Cycles*. New York: NBER.

Cave, B.M. and Pearson, K. (1914). 'Numerical illustrations of the variate-difference correlation method'. *Biometrika* 10, 340–55.

Cave-Browne-Cave, F.E. (1905). 'On the influence of the time factor on the correlation between the barometric heights at stations more than 1000 miles apart'. *Proceedings of the Royal Society of London* 74, 403–13.

Cave-Browne-Cave, F.E. and Pearson, K. (1902). 'On the correlation between the barometric heights at stations on the eastern side of the Atlantic'. *Proceedings of the Royal Society of London* 70, 465–70.

Chamberlain, G. (1982). 'The general equivalence of Granger and Sims causality'. *Econometrica* 50, 569–81.

Chan, K.H., Hayya, J.C. and Ord, J.K. (1977). 'A note on trend removal methods: the case of polynomial versus variate differencing'. *Econometrica* 45, 737–44.

Chang, I, Tiao, G.C. and Chen, C. (1988). 'Estimation of time series parameters in the presence of outliers'. *Technometrics* 30, 193–204.

Chib, S. and Greenberg, E. (1995). 'Understanding the Metropolis–Hastings algorithm'. *American Statistician* 49, 327–35.

Chib, S. and Greenberg, E. (1996). 'Markov chain Monte Carlo simulation methods in econometrics'. *Econometric Theory* 12, 409–31.

Choi, I. (2006). 'Nonstationary panels', in T.C. Mills and K.D. Patterson (eds) *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 511–39.

Christiano, L.J. and Fitzgerald, T.J. (2003). 'The band pass filter'. *International Economic Review* 44, 435–65.

Clements, M.P. and Hendry, D.F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Cleveland, R.B., Cleveland, W.S., McRae, J.E. and Terpenning, I. (1990). 'STL: a seasonal-trend decomposition procedure based on loess'. *Journal of Official Statistics* 6, 3–73.

Cleveland, W.S., Dunn, D.M. and Terpenning, I. (1978). 'SABL – a resistant seasonal adjustment procedure with graphical methods for interpretation and diagnosis', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*. Washington, DC: US Dept of Commerce, Bureau of the Census, pp. 201–31.

Cleveland, W.P. and Tiao, G.C. (1976). 'Decomposition of seasonal time series: a model for the X-11 program'. *Journal of the American Statistical Association* 71, 581–7.

Cochran, W.G. (1934). 'Distribution of quadratic forms in a normal system with applications to the analysis of covariances'. *Proceedings of the Cambridge Philosophical Society* 30, 178–91.

Cochrane, D. and Orcutt, G.H. (1949). 'Application of least squares regression to relationships containing autocorrelated error terms'. *Journal of the American Statistical Association* 44, 32–61.

Cogger, K.O. (1974). 'The optimality of general-order exponential smoothing'. *Operations Research* 22, 858–67.

Cooley, J.W. and Tukey, J.W. (1965). 'An algorithm for the machine calculation of complex Fourier series'. *Mathematics of Computation* 19, 297–301.

Cooper, J.P. and Nelson, C.R. (1975). 'The *ex ante* prediction performance of the St Louis and F.R.B.–M.I.T.–Penn econometric models and some results on composite predictors'. *Journal of Money, Credit and Banking* 7, 1–32.

Cooper, R.L. (1972). 'The predictive performance of quarterly econometric models of the United States', in B.G. Hickman (ed.) *Econometric Models of Cyclical Behaviour.* New York: Columbia University Press, pp. 813–947.

Cootner, P.H. (1962). 'Stock prices: random vs. systematic changes'. *Industrial Management Review* 3, 24–45.

Copeland, M.T. (1915). 'Statistical indices of business conditions'. *Quarterly Journal of Economics* 29, 522–62.

Cowden, D.J. (1942). 'Moving seasonal indexes'. *Journal of the American Statistical Association* 37, 523–4.

Cowles, A. (1933). 'Can stock market forecasters forecast?' *Econometrica* 1, 309–24.

Cowles, A. (1944). 'Stock market forecasting'. *Econometrica* 12, 206–14.

Cowles, A. and Jones, H.E. (1937). 'Some a posteriori probabilities in stock market action'. *Econometrica* 5, 280–94.

Cox, D.R. (1961). 'Prediction by exponentially weighted moving averages and related methods'. *Journal of the Royal Statistical Society, Series B* 23, 414–22.

Cox, D.R. (1966). 'The null distribution of the first serial correlation coefficient'. *Biometrika* 53, 623–6.

Craddock, J.M. (1967). 'An experiment in the analysis and prediction of time series'. *The Statistician* 17, 257–68.

Cramér, H. (1940). 'On the theory of stationary random processes'. *Annals of Mathematics* 41, 215–30.

Cramér, H. (1962). 'A.I. Khinchin's work in mathematical probability'. *Annals of Mathematical Statistics* 33, 1227–37.

Dagum, E.B. (1978). 'Modelling, forecasting and seasonally-adjusting economic time series with the X-11-ARIMA method'. *The Statistician* 27, 203–16.

Dagum, E.B. (1982). 'The effects of asymmetric filters on seasonal factor revisions'. *Journal of the American Statistical Association* 77, 732–8.

Daniell, P.J. (1946). 'Discussion on "Symposium on autocorrelation in time series"'. *Journal of the Royal Statistical Society, Series B* 8, 88–90.

Daniels, H.E. (1956). 'The approximate distribution of serial correlation coefficients'. *Biometrika* 43, 169–85.

Daniels, H.E. (1962). 'The estimation of spectral densities'. *Journal of the Royal Statistical Society, Series B*, 24, 185–98.

Davidson, J.E.H., Hendry, D.F., Srba, F. and Yeo, S. (1978). 'Econometric modeling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom'. *Economic Journal* 88, 861–92.

Davies, N. and Newbold, P. (1979). 'Some power studies of a portmanteau test of time series model specification'. *Biometrika* 66, 153–5.

Davies, N., Trigg, C.M. and Newbold, P. (1977). 'Significance levels of the Box–Pierce portmanteau statistic in finite samples'. *Biometrika* 64, 517–22.

Davis, H.T. (1941). *The Analysis of Economic Time Series*. Bloomington, IL: The Principia Press.

De Groot, M.H. (1987). 'A conversation with George Box'. *Statistical Science* 2, 239–58.

D'Esopo, D.A. (1961). 'A note on forecasting by the exponential smoothing operator'. *Operations Research* 9, 686–7.

Dickey, D.A. and Fuller, W.A. (1979). 'Distribution of the estimators for autoregressive time series with a unit root'. *Journal of the American Statistical Association* 74, 427–31.

Dickey, D.A. and Fuller, W.A. (1981). 'Likelihood ratio statistics for autoregressive time series with a unit root'. *Econometrica* 49, 1057–72.

Dimand, R.W. (1993). 'The case of Brownian motion: a note on Bachelier's contribution'. *British Journal for the History of Science* 26, 233–4.

Dixon, W.J. (1944). 'Further contributions to the problem of serial correlation'. *Annals of Mathematical Statistics* 15, 119–44.

Dodd, E.L. (1939). 'The length of the cycles which result from the graduation of chance elements'. *Annals of Mathematical Statistics* 10, 254–64.

Dodd, E.L. (1941a). 'The problem of assigning a length to the cycle to be found in a simple moving average and in a double moving average of chance data'. *Econometrica* 9, 25–37.

Dodd, E.L. (1941b). 'The cyclic effects of linear graduations persisting in the differences of the graduated values'. *Annals of Mathematical Statistics* 12, 127–36.

Doherty, M. (2001). 'Surrogate Henderson filters in X-11'. *Australian and New Zealand Journal of Statistics* 43, 385–92.

Dufour, J.-M. and Renault, E. (1998). 'Short run and long run causality in time series: Theory'. *Econometrica* 66, 1099–125.

Duncan, D.B. and Horn, S.D. (1972). 'Linear dynamic recursive estimation from the viewpoint of regression analysis'. *Journal of the American Statistical Association* 67, 815–21.

Durbin, J. (1959). 'Efficient estimation of parameters in moving-average models'. *Biometrika* 46, 306–16.

Durbin, J. (1960). 'The fitting of time-series models'. *Review of the International Statistical Institute* 28, 233–44.

Durbin, J. (1962). 'Trend elimination by moving-average and variate-differencing filters'. *Bulletin of the International Statistical Institute* 34, 131–41.

Durbin, J. (1963). 'Trend elimination for the purposes of estimating seasonal and periodic components of time series', in M. Rosenblatt (ed.), *Proceedings of the Symposium on Time Series Analysis*. New York: John Wiley, pp. 3–16.

Durbin, J. (1970). 'Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables'. *Econometrica* 38, 410–21.

Durbin, J. (1982). 'More than twenty-five years of testing serial correlation in least squares regression', in M. Hazewinkel and A.H.G. Rinnooy Kan (eds), *Current Developments in the Interface: Economics, Econometrics, Mathematics*. Rotterdam: Econometric Institute, pp. 59–71.

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durbin, J. and Watson, G.S. (1950). 'Testing for serial correlation in least squares regression: I'. *Biometrika* 37, 409–28.

Durbin, J. and Watson, G.S. (1951). 'Testing for serial correlation in least squares regression: II'. *Biometrika* 38, 159–77.

Durbin, J. and Watson, G.S. (1971). 'Testing for serial correlation in least squares regression: III'. *Biometrika* 58, 1–19.

Edgeworth, F.Y. (1892). 'On correlated averages'. *Philosophical Magazine* 84, 194–204.

Edgeworth, F.Y. (1893). 'Statistical correlation between social phenomena'. *Journal of the Royal Statistical Society* 56, 670–5.

Edgeworth, F.Y. (1894). 'Asymmetrical correlation between social phenomena'. *Journal of the Royal Statistical Society* 57, 563–8.

Eisenpress, H., McPherson, J.L. and Shiskin, J. (1955). 'Charting on automatic data processing systems'. *Computers and Automation*, August.

Elderton, E.M. and Pearson, K. (1915). 'Further evidence of natural selection in men'. *Biometrika* 10, 488–506.

Elliott, G., Rothenberg, T. and Stock, J.H. (1996). 'Efficient tests for an autoregressive unit root'. *Econometrica* 64, 813–36.

Engle, R.F. (1974). 'Band spectrum regression'. *International Economic Review* 15, 1–11.

Engle, R.F. (1980). 'Exact maximum likelihood methods for dynamic regressions and band spectrum regression'. *International Economic Review* 21, 391–407.

Engle, R.F. (1982). 'Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation'. *Econometrica* 50, 987–1008.

Engle, R.F. and Granger, C.W.J. (1987). 'Cointegration and error correction: representation, estimation and testing'. *Econometrica* 55, 251–76.

Evans, G.B.A. and Savin, N.E. (1981). 'Testing for unit roots: 1'. *Econometrica* 49, 753–79.

Evans, G.B.A. and Savin, N.E. (1984). 'Testing for unit roots: 2'. *Econometrica* 52, 1249–69.

Fama, E.F. (1965). 'The behavior of stock market prices'. *Journal of Business* 38, 34–105.

Fama, E.F. and Blume, M.E. (1966). 'Filter rules and stock-market trading'. *Journal of Business* 39, 226–41.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. and Chen, B.C. (1998). 'New capabilities and methods in the X-12-ARIMA seasonal-adjustment package'. *Journal of Business and Economic Statistics* 16, 127–77.

Fisher, I. (1925). 'Our unstable dollar and the so-called business cycle'. *Journal of the American Statistical Association* 20, 179–202.

Fisher, R.A. (1925). 'The influence of rainfall on the yield of wheat at Rothampstead'. *Philosophical Transactions of the Royal Society of London, Series B* 213, 89–142.

Fisher, R.A. (1929). 'Tests of significance in harmonic analysis'. *Journal of the Royal Statistical Society, Series A* 125, 54–9.

Franses, P.H. and van Dijk, D. (2000). *Non-linear Time Series Models in Empirical Finance*. Cambridge: Cambridge University Press.

Frickey, E. (1934). 'The problem of secular trend'. *Review of Economic Statistics* 16, 199–206.

Friedman, M. (1957). *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.

Frisch, R. (1933). 'Propagation problems and impulse problems in dynamic economics'. In *Economic Essays in Honour of Gustav Cassel*. London: Allen & Unwin, pp. 171–205.

Frisch, R. and Waugh, F.V. (1933). 'Partial time regressions as compared with individual trends'. *Econometrica* 1, 387–401.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

Galton, F. (1888). 'Co-relations and their measurement: chiefly from anthropological data'. *Proceedings of the Royal Society of London* 45, 135–45.

Galton, F. (1890). 'Kinship and correlation'. *North American Review* 150, 419–31.

Gardner Jr., E.S. (1985). 'Exponential smoothing: the state of the art'. *Journal of Forecasting* 4, 1–28.

Gardner Jr., E.S. (2006). 'Exponential smoothing: the state of the art – Part II'. *International Journal of Forecasting* 22, 637–66.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1990). 'Illustration of Bayesian inference in normal data models using Gibbs sampling'. *Journal of the American Statistical Association* 85, 972–85.

Geweke, J. (1977). 'The dynamic factor analysis of economic time series models', in D.J. Aigner and A.S. Goldberger (eds), *Latent Variables in Socio-Economic Models*. Amsterdam: North-Holland.

Geweke, J. (1982). 'Measurement of linear dependence and feedback between time series'. *Journal of the American Statistical Association* 79, 304–24.

Geweke, J. (1984). 'Inference and causality in economic time series models', in Z. Griliches and M.D. Intriligator (eds), *Handbook of Econometrics, Volume II*. Amsterdam: North-Holland, pp. 1101–44.

Gil-Alana, L.A. and Hualde, J. (2009). 'Fractional integration and cointegration: an overview and an empirical application', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*. Basingstoke: Palgrave Macmillan, pp. 434–69.

Gilbart, J.W. (1854). 'The laws of the currency, as exemplified in the circulation of country bank notes in England, since the passing of the Act of 1844'. *Journal of the Statistical Society of London* 17, 289–321.

Godfrey, L.G. (1979). 'Testing the adequacy of a time series model'. *Biometrika* 66, 67–72.

Godfrey, L.G. (1988). *Misspecification Tests in Econometrics*. Cambridge: Cambridge University Press.

Godfrey, M.D. (1965). 'An exploratory study of the bi-spectrum of economic time series'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 14, 48–69.

Godfrey, M.D., Granger, C.W.J. and Morgenstern, O. (1964). 'The random-walk hypothesis of stock market behavior'. *Kyklos* 17, 1–29.

Gómez, V. and Maravall, A. (1996). 'Programs TRAMO (Time series Regression with Arima noise, Missing observations and Outliers) and SEATS (Signal Extraction in Arima Time Series). Instructions for the user'. Working Paper 9628, Research Department, Banco de Espaňa.

Goodman, N.R. (1961). 'Some comments on spectral analysis of time series'. *Technometrics* 3, 221–8.

Gouriéroux, C. (1997). *ARCH Models and Financial Applications*. Berlin: Springer-Verlag.

Gower, J.C. (1955). 'A note on the periodogram of the Beveridge wheat price index'. *Journal of the Royal Statistical Society, Series B* 17, 228–34.

Granger, C.W.J. (1963). 'Economic processes involving feedback'. *Information and Control* 6, 28–48.

Granger, C.W.J. (1966). 'The typical spectral shape of an economic variable'. *Econometrica* 34, 150–61.

Granger, C.W.J. (1969). 'Investigating causal relations by econometric methods and cross-spectral methods'. *Econometrica* 37, 424–38.

Granger, C.W.J. (1980). 'Testing for causality: a personal viewpoint'. *Journal of Economic Dynamics and Control* 2, 329–52.

Granger, C.W.J. (1981). 'Some properties of time series data and their use in econometric model specification'. *Journal of Econometrics* 16, 121–30.

Granger, C.W.J. (1986). 'Developments in the study of cointegrated economic variables'. *Oxford Bulletin of Economics and Statistics* 48, 213–28.

Granger, C.W.J. and Hatanaka, M. (1964). *Spectral Analysis of Economic Time Series*. Princeton, NJ: Princeton University Press.

Granger, C.W.J. and Hughes, A.O. (1971). 'A new look at some old data: the Beveridge wheat price series'. *Journal of the Royal Statistical Society, Series A* 134, 413–28.

Granger, C.W.J. and Joyeux, R. (1980). 'An introduction to long memory time series models and fractional differencing'. *Journal of Time Series Analysis* 1, 15–29.

Granger, C.W.J. and Newbold, P. (1973). 'Some comments on the evaluation of economic forecasts'. *Applied Economics* 5, 35–47.

Granger, C.W.J. and Newbold, P. (1974). 'Spurious regressions in econometrics'. *Journal of Econometrics* 2, 111–20.

Granger, C.W.J. and Newbold, P. (1986). *Forecasting Economic Time Series* 2nd edition, San Diego: Academic Press.

Granger, C.W.J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Gray, A. and Thompson, P. (2002). 'On a family of moving-average filters for the ends of series'. *Journal of Forecasting* 21, 125–49.

Grenander, U. (1951). 'On empirical analysis of stochastic processes'. *Arkiv fő r Matematik* 1, 503–31.

Grenander, U. (1958). 'Bandwidth and variance in the estimation of the spectrum'. *Journal of the Royal Statistical Society, Series B* 20, 152–7.

Grenander, U. and Rosenblatt, M. (1953). 'Statistical spectral analysis of time series arising from stationary stochastic processes'. *Annals of Mathematical Statistics* 24, 537–58.

Grether, D.M. and Nerlove, M. (1970). 'Some properties of "optimal" seasonal adjustment'. *Econometrica* 38, 682–703.

Hald, A. (1981). 'T.N. Thiele's contributions to statistics'. *International Statistical Review* 49, 1–20.

Hall, L.W. (1925). 'A moving secular trend and moving integration'. *Journal of the American Statistical Association* 20, 13–24.

Hannan, E.J. (1955). 'An exact test for correlation between time series'. *Biometrika* 42, 316–26.

Hannan, E.J. (1960). 'The estimation of seasonal variation. *Australian Journal of Statistics* 2, 1–15.

Hannan, E.J. (1963). 'The estimation of seasonal variation in economic time series'. *Journal of the American Statistical Association* 58, 31–44.

Hannan, E.J. (1964). 'The estimation of a changing seasonal pattern'. *Journal of the American Statistical Association* 59, 1063–77.

Hannan, E.J. (1967). 'Measurement of a wandering signal amid noise'. *Journal of Applied Probability* 4, 90–102.

Hannan, E.J. and Quinn, B.G. (1979). 'The determination of the order of an autoregression'. *Journal of the Royal Statistical Society, Series B* 41, 190–5.

Harris, D., Harvey, D.I., Leybourne, S.J. and Taylor, A.M.R. (2009). 'Testing for a unit root in the presence of a possible break in trend'. *Econometric Theory* 25, 1545–88.

Harrison, P.J. (1965). 'Short-term sales forecasting'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 14, 102–39.

Harrison, P.J. (1967). 'Exponential smoothing and short-term sales forecasting'. *Management Science* 13, 821–42.

Harrison, P.J. and Stevens, C.F. (1971). 'A Bayesian approach to short-term forecasting'. *Operations Research Quarterly* 22, 341–62.

Harrison, P.J. and Stevens, C.F. (1976). 'Bayesian forecasting'. *Journal of the Royal Statistical Society, Series B* 38, 205–47.

Hartley, H.O. (1949). 'Tests of significance in harmonic analysis'. *Biometrika* 36, 194–201.

Hartley, H.O. (1961). 'The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares'. *Technometrics* 3, 269–80.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Harvey, A.C. and Durbin, J. (1986). 'The effects of seat belt legislation on British road casualties: a case study in structural time series modeling (with discussion)'. *Journal of the Royal Statistical Society, Series A* 149, 187–227.

Henderson, R. (1916). 'Note on graduation by adjusted average'. *Transactions of the American Society of Actuaries* 17, 43–8.

Henderson, R. (1924). 'A new method of graduation'. *Transactions of the American Society of Actuaries* 25, 29–40.

Hendry, D.F. (1977). 'Comments on Granger–Newbold's "Time series approach to econometric model building" and Sargent-Sims "Business cycle modeling without pretending to have too much *a priori* theory"', in C.A. Sims (ed.), *New Methods in Business Cycle Research.* Minneapolis, MN: Federal Reserve Bank of Minneapolis, pp. 183–202.

Hendry, D.F. and Mizon, G.E. (1978). 'Serial correlation as a convenient simplification, not a nuisance: a comment on a study of the demand for money by the Bank of England'. *Economic Journal* 88, 549–63.

Hendry, D.F. and Morgan, M.S. (1994). 'The ET interview: Professor H.O.A. Wold: 1908–1992'. *Econometric Theory* 10, 419–33.

Hendry, D.F. and Morgan, M.S. (1995). 'Introduction'. In D.F. Hendry and M.S. Morgan (eds), *The Foundations of Econometric Analysis.* Cambridge: Cambridge University Press, pp. 1–82.

Hepple, L.W. (2001). 'Multiple regression and spatial policy analysis: George Udny Yule and the origins of statistical social science'. *Environment and Planning D: Society and Space* 19, 385–407.

Hicks, J.R. and Allen, R.G.D. (1934). 'A reconsideration of the theory of value'. *Economica* 1, 52–76.

Hillmer, S.C., Bell, W.R. and Tiao, G.C. (1983). 'Modelling considerations in the seasonal adjustment of economic time series', in A. Zellner (ed.) *Applied Time Series Analysis of Economic Data*. Washington, DC: US Dept. of Commerce, Bureau of the Census, pp. 74–100.

Hillmer, S.C. and Tiao, G.C. (1982). 'An ARIMA-model-based approach to seasonal adjustment'. *Journal of the American Statistical Association* 77, 63–70.

Hodrick, R.J. and Prescott, E.C. (1997). 'Postwar U.S. business cycles: an empirical investigation'. *Journal of Money, Credit and Banking* 29, 1–16.

Holt, C.C. (1957). 'Forecasting seasonals and trends by exponentially weighted moving averages'. *ONR Memorandum* 52. Pittsburgh: Carnegie Institute of Technology.

Holt, C.C. (2004a). 'Forecasting seasonals and trends by exponentially weighted moving averages'. *International Journal of Forecasting* 20, 5–10.

Holt, C.C. (2004b). 'Author's retrospective on "Forecasting seasonals and trends by exponentially weighted moving averages"'. *International Journal of Forecasting* 20, 11–13.

Hooker, R.H. (1901a). 'The suspension of the Berlin Produce Exchange and its effect upon corn prices'. *Journal of the Royal Statistical Society* 64, 574–613.

Hooker, R.H. (1901b). 'Correlation of the marriage-rate with trade'. *Journal of the Royal Statistical Society* 64, 485–92.

Hooker, R.H. (1905). 'On the correlation of successive observations'. *Journal of the Royal Statistical Society* 68, 696–703.

Hoover, K.D. and Perez, S.J. (1999). 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search'. *Econometrics Journal* 2, 167–91.

Hosking, J.R.M. (1981). 'Fractional differencing'. *Biometrika* 68, 165–76.

Hosking, J.R.M. (1984). 'Modelling persistence in hydrological time series using fractional differencing'. *Water Resources Research* 20, 1898–908.

Hotelling, H. (1936). 'Relations between two sets of variates'. *Biometrika* 28, 321–77.

Hurst, H.E. (1951). 'Long term storage capacity of reservoirs'. *Transactions of the American Society of Civil Engineers* 116, 770–99.

Hurwicz, L. (1945), 'Least-squares bias in time series', in T.C. Koopmans (ed.), *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph, University of Chicago Press, pp. 365–83.

Hylleberg, S. (1992). 'The historical perspective', in S. Hylleberg (ed.), *Modelling Seasonality.* Oxford: Oxford University Press, pp. 15–25.

IEEE (1967). *Transactions on Audio and Electroacoustics.* Volume AU-15, No. 2.

Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory*. San Diego: Academic Press.

Jenkins, G.M. (1954a). 'An angular transformation for the serial correlation coefficient'. *Biometrika* 41, 261–5.

Jenkins, G.M. (1954b). 'Tests of hypotheses in the linear autoregressive model. I Null hypothesis distributions in the Yule scheme'. *Biometrika* 41, 405–19.

Jenkins, G.M. (1956). 'Tests of hypotheses in the linear autoregressive model. II Null distributions for higher order schemes: non-null distributions'. *Biometrika* 43, 186–99.

Jenkins, G.M. (1961). 'General considerations in the analysis of spectra'. *Technometrics* 3, 133–66.

Jenkins, G.M. (1965). 'A survey of spectral analysis'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 14, 2–32.

Jenkins, G.M. and McLeod, G. (1982). *Case Studies in Time Series Analysis*. Lancaster: GJP.

Jenkins, G.M. and Priestley, M. (1957). 'The spectral analysis of time series'. *Journal of the Royal Statistical Society, Series B* 19, 1–12.

Jenkins, G.M. and Watts, D.G. (1968). *Spectral Analysis and its Applications*. San Francisco: Holden-Day.

Jevons, W.S. (1866). 'On the frequent autumnal pressure in the money market, and the action of the Bank of England', in Jevons (1884), 160–93.

Jevons, W.S. (1884). *Investigations in Currency and Finance*. London: Macmillan.

Johansen, S. (1988a). 'The mathematical structure of error correction models'. *Contemporary Mathematics* 80, 359–86.

Johansen, S. (1988b). 'Statistical analysis of cointegration vectors'. *Journal of Economic Dynamics and Control* 12, 231–54.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Johansen, S. (2006). 'Cointegration: an overview', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 540–77.

Johnson, N.L. (1948). 'Tests of significance in the variate difference method'. *Biometrika* 35, 206–9.

Jones, H.L. (1943). 'Fitting polynomial trends to seasonal data by the method of least squares'. *Journal of the American Statistical Association* 38, 453–65.

Jorgenson, D.W. (1963). 'Capital theory and investment behavior'. *American Economic Review* 53, 247–59.

Jorgenson, D.W. (1964). 'Minimum variance, linear, unbiased seasonal adjustment of economic time series'. *Journal of the American Statistical Association* 59, 681–724.

Jorgenson, D.W. (1967). 'Seasonal adjustment of data for econometric analysis'. *Journal of the American Statistical Association* 62, 147–50.

Joy, A. and Thomas, W. (1928). 'The use of moving averages in the measurement of seasonal variations'. *Journal of the American Statistical Association* 23, 241–52.

Juselius, K. (2006). *The Cointegrated VAR Model: Methodology and Applications*. Oxford: Oxford University Press.

Kalman, R.E. (1960). 'A new approach to linear filtering and prediction problems'. *Journal of Basic Engineering* 82, 34–45.

Kalman, R.E. and Bucy, R.S. (1961). 'New results in linear prediction and filtering theory'. *Journal of Basic Engineering* 83, 95–108.

Kantz, H. and Schreiber, T. (1997). *Nonlinear Time Series Analysis*. Cambridge: Cambridge University Press.

Katz, R.W. (2002). 'Sir Gilbert Walker and a connection between El Nino and statistics'. *Statistical Science* 17, 97–112.

Kemmerer, E.W. (1910). *Seasonal Variations in the Relative Demand for Money and Capital in the United States*. Washington, DC, Report of the National Monetary Commission.

Kemp, A.W. (1970). 'General formulae for the central moments of certain serial correlation coefficient approximations'. *Annals of Mathematical Statistics* 41, 1363–8.

Kendall, M.G. (1941). 'The effect of the elimination of trend on oscillation in time-series'. *Journal of the Royal Statistical Society* 104, 43–52.

Kendall, M.G. (1943). 'Oscillatory movements in English agriculture'. *Journal of the Royal Statistical Society* 106, 43–52.

Kendall, M.G. (1944). 'On autoregressive time series'. *Biometrika* 33, 105–22.

Kendall, M.G. (1945a). 'On the analysis of oscillatory time-series'. *Journal of the Royal Statistical Society* 108, 93–141.

Kendall, M.G. (1945b). 'Note on Mr. Yule's paper', *Journal of the Royal Statistical Society* 108, 226–30.

Kendall, M.G. (1946). *The Advanced Theory of Statistics, Volume II*. London: Griffin.

Kendall, M.G. (1949). 'The estimation of parameters in linear autoregressive time series'. *Econometrica* 17 Supplement, 44–57.

Kendall, M.G. (1952). 'Obituary: George Udny Yule'. *Journal of the Royal Statistical Society, Series A* 115, 156–61.

Kendall, M.G. (1953). 'The analysis of economic time series, Part I: Prices'. *Journal of the Royal Statistical Society, Series A* 96, 11–25.

Kendall, M.G. (1954). 'Note on the bias in the estimation of autocorrelations'. *Biometrika* 41, 403–4.

Kendall, M.G. (1957). 'The moments of the Leipnik distribution'. *Biometrika* 44, 270–2.

Kendall, M.G. (1961). 'A theorem in trend analysis'. *Biometrika* 48, 224–7.

Kendall, M.G. (1971). 'Review of "Time Series Analysis, Forecasting and Control" by G.E.P. Box and G.M. Jenkins'. *Journal of the Royal Statistical Society, Series A* 134, 450–3.

Kendall, M.G., Stuart, A. and Ord, J.K. (1983). *The Advanced Theory of Statistics, Volume 3*, 4th edition. London: Griffin.

Kenny, P.B. and Durbin, J. (1982). 'Local trend estimation and seasonal adjustment of economic and social time series'. *Journal of the Royal Statistical Society, Series A* 145, 1–41.

Khinchin, A.Y. (1932). 'Sulle successioni stazionarie di eventi'. *Gionale Institute Italian Atturi* 3, 267–323.

Khinchin, A.Y. (1933). 'Ǘber stationäre reihen zufalliger variablen'. *Matematicheskii Sbornik* 40, 124–8.

Khinchin, A.Y. (1934). 'Korrleationstheorie der stationären stochastischen prozesse'. *Mathematische Annalen* 109, 604–15.

Kiviet, J.F. (1986). 'On the rigor of some mis-specification tests for modeling dynamic relationships'. *Review of Economic Studies* 53, 241–61.

Klein, J.L. (1997). *Statistical Visions in Time: A History of Time Series Analysis 1662–1938*. Cambridge: Cambridge University Press.

Kolmogorov, A.N. (1931). 'Ǘber die analytischen methoden in der wahrscheinlichkeitsrechnung'. *Mathematische Annalen* 104, 415–58.

Kolmogorov, A.N. (1933 [1950]). *Foundations of the Theory of Probability*. Translated by Nathan Morrison. New York: Chelsea.

Kolmogorov, A.N. (1941). 'Interpolation and extrapolation'. *Bulletin de l'academie des sciences de U.S.S.R., Ser. Math.* 5, 3–14.

Koop, G., Strachan, R., van Dijk, H. and Villani, M. (2006). 'Bayesian approaches to cointegration', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 871–98.

Koopman, S.J., Harvey, A.C., Doornik, J.A. and Shephard, N. (2009). *STAMP™ 8*. London: Timberlake Consultants Ltd.

Koopmans, T. (1942). 'Serial correlation and quadratic forms in normal variables'. *Annals of Mathematical Statistics* 13, 14–33.

Koyck, L.M. (1954). *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.

Kuznets, S. (1929). 'Random events and cyclical oscillations'. *Journal of the American Statistical Association* 24, 258–75.

Ladiray, D. and Quenneville, B. (2001). *Seasonal Adjustment with the X-11 Method*. Lecture Notes in Statistics 158. New York: Springer-Verlag.

Larmor, J. and Yamaga, N. (1917). 'On permanent periodicity in sunspots'. *Proceedings of the Royal Society of London, Series A* 93, 493–506.

Lauritzen, S.L. (1981). 'Time series analysis in 1880: a discussion of contributions made by T.N. Thiele'. *International Statistical Review* 49, 319–31.

Lauritzen, S.L. (2002). *Thiele: Pioneer in Statistics*. Oxford: Oxford University Press.

Leong, Y.S. (1962). 'The use of an iterated moving average in measuring seasonal variation'. *Journal of the American Statistical Association* 57, 149–71.

Leipnik, R.B. (1947). 'Distribution of the serial correlation coefficient in a circularly correlated universe'. *Annals of Mathematical Statistics* 18, 80–7.

Leser, C.E.V. (1961). 'A simple method of trend construction'. *Journal of the Royal Statistical Society, Series B* 23, 91–107.

Leser, C.E.V. (1963). 'Estimation of quasi-linear trend and seasonal variation'. *Journal of the American Statistical Association* 58, 1033–43.

Leser, C.E.V. (1966). 'Direct estimation of seasonal variation'. *International Statistical Review* 34, 369–75.

Litterman, R.B. (1986). 'Forecasting with Bayesian vector autoregressions – Five years of experience'. *Journal of Business and Economic Statistics* 4, 25–38.

Ljung, G.M. and Box, G.E.P. (1978). 'On a measure of lack of fit in time series'. *Biometrika* 65, 297–303.

Lomnicki, Z.A. and Zaremba, S.K. (1957). 'On estimating the spectral density function of a stochastic process'. *Journal of the Royal Statistical Society, Series B* 19, 13–37.

Lomnicki, Z.A. and Zaremba, S.K. (1959). 'Bandwidth and resolvability in statistical spectral analysis'. *Journal of the Royal Statistical Society, Series B* 21, 169–71.

Lovell, M.C. (1963). 'Seasonal adjustment of economic time series and multiple regression analysis'. *Journal of the American Statistical Association* 58, 993–1010.

Lovell, M.C. (1966). 'Alternative axiomatizations of seasonal adjustment'. *Journal of the American Statistical Association* 61, 800–2.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Lütkepohl, H. (2006). 'Vector autoregressive models', in T.C. Mills and K.D. Patterson (eds) *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 477–510.

Macaulay, F.R. (1931). *The Smoothing of Time Series*. New York: NBER.

Madow, W.G. (1945). 'Note on the distribution of the serial correlation coefficient'. *Annals of Mathematical Statistics* 16, 308–10.

Malkiel, B.G. (2007). *A Random Walk Down Main Street*, 9th edition. New York: Norton.

Mandelbrot, B.B. (1989). 'Louis Bachelier', in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave: Finance*. London: Macmillan, pp. 86–8.

Mandelbrot, B.B. and Wallis, J.R. (1969). 'Some long-run properties of geophysical records'. *Water Resources Research* 5, 321–40.

Mann, H.B. and Wald, A. (1943). 'On the statistical treatment of linear stochastic difference equations'. *Econometrica* 11, 173–220.

Maravall, A. (2000). 'An application of TRAMO and SEATS'. *Annali di Statistica* 20, 271–344.

Maravall, A. and Pierce, D.A. (1987). 'A prototypical seasonal adjustment model'. *Journal of Time Series Analysis* 8, 177–93.

Marget, A.W. (1929). 'Morgenstern on the methodology of economic forecasting'. *Journal of Political Economy* 37, 312–39.

Marquardt, D.W. (1963). 'An algorithm for least squares estimation of nonlinear parameters'. *Journal of the Society for Industrial and Applied Mathematics* 11, 431–41.

Marriott, F.H.C. and Pope, J.A. (1954). 'Bias in the estimation of autocorrelations'. *Biometrika* 41, 390–402.

McGregor, J.R. (1962). 'The approximate distribution of the correlation between two stationary linear Markov series'. *Biometrika* 49, 379–88.

McGregor, J.R. and Bielenstein, U.M. (1965). 'The approximate distribution of the correlation between two stationary linear Markov series. II'. *Biometrika* 52, 301–2.

McLeod, A.I. (1978). 'On the distribution of residual autocorrelations in Box–Jenkins models'. *Journal of the Royal Statistical Society, Series B* 40, 296–302.

Mendershausen, H. (1937). 'Annual survey of statistical technique: methods of computing and eliminating changing seasonal fluctuations'. *Econometrica* 5, 234–62.

Mendershausen, H. (1939). 'Eliminating changing seasonal by multiple regression analysis'. *Review of Economic Statistics* 21, 171–7.

Mills, T.C. (1982a). 'Signal extraction and two illustrations of the quantity theory'. *American Economic Review* 72, 1162–8.

Mills, T.C. (1982b). ‘The use of unobserved component and signal extraction techniques in modeling economic time series’. *Bulletin of Economic Research* 34, 92–108.

Mills, T.C. (1990). *Time Series Techniques for Economists*. Cambridge: Cambridge University Press.

Mills, T.C. (1999). *Economic Forecasting*. Cheltenham: Edward Elgar.

Mills, T.C. (2002a). *Forecasting Financial Markets*. Cheltenham: Edward Elgar.

Mills, T.C. (2002b). *Long Term Trends and Business Cycles*. Cheltenham: Edward Elgar.

Mills, T.C. (2003). *Modelling Trends and Cycles in Economic Time Series*. Basingstoke: Palgrave Macmillan.

Mills, T.C. (2007). ‘A note on trend decompositions: the “classical” approach revisited with an application to surface temperature trends’. *Journal of Applied Statistics* 34, 963–72.

Mills, T.C. (2009). ‘Modelling trends and cycles in economic time series: Historical perspective and future development’. *Cliometrica* 3, 221–44.

Mills, T.C. (2010a). ‘Skinning a cat: stochastic models for assessing temperature trends’. *Climatic Change* 101, 415–26.

Mills, T.C. (2010b). ‘Is global warming real? Analysis of structural time series models of global and hemispheric temperatures?’ *Journal of Cosmology* 8, 1947–54.

Mills, T.C. (2011). ‘Bradford Smith: an econometrician decades ahead of his time’. *Oxford Bulletin of Economics and Statistics* 73, 276–85.

Mills, T.C. and Markellos, R.N. (2008). *The Econometric Modelling of Financial Time Series*, 3rd edition. Cambridge: Cambridge University Press.

Mitchell, B.R. (1998). *International Historical Statistics: Europe 1750–1993*, 4th edition. London: Macmillan Reference.

Mitchell, W.C. (1913). *Business Cycles and their Causes*. Berkeley: California University Memoirs, Vol. III.

Mizon, G.E. and Hendry, D.F. (1980). ‘An empirical application and Monte Carlo analysis of tests of dynamic specification’. *Review of Economics and Statistics* 47, 1221–42.

Moore, H.L. (1923). *Generating Economic Cycles*. New York: Macmillan.

Moran, P.A.P. (1947). ‘Some theorems on time series, I’. *Biometrika* 34, 281–91.

Moran, P.A.P. (1948). ‘Some theorems on time series, II. The significance of the serial correlation coefficient’. *Biometrika* 35, 255–60.

Moran, P.A.P. (1949). ‘The spectral theory of discrete stochastic processes’. *Biometrika* 36, 63–70.

Moran, P.A.P. (1950). ‘The oscillatory behavior of moving averages’. *Proceedings of the Cambridge Philosophical Society* 46, 272–80.

Moran, P.A.P. (1967a). ‘Testing for serial correlation with exponentially distributed variates’. *Biometrika* 54, 395–401.

Moran, P.A.P. (1967b). ‘Testing for correlation between non-negative variates’. *Biometrika* 54, 385–94.

Moran, P.A.P. (1970). ‘A note on serial correlation coefficients’. *Biometrika* 57, 670–3.

Morgan, M.S. (1990). *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Morris, J.M. (1977). ‘Forecasting the sunspot cycle’. *Journal of the Royal Statistical Society, Series A* 140, 437–48.

Müller, U.K. and Elliott, G. (2003). ‘Tests for unit roots and the initial condition’. *Econometrica* 71, 1269–86.

Muth, J.F. (1960). ‘Optimal properties of exponentially weighted forecasts’. *Journal of the American Statistical Association* 55, 299–306.

Naylor, T.H., Seaks, T.G. and Wichern, D.W. (1972). ‘Box-Jenkins models: an alternative to econometric models’. *International Statistical Review* 40, 123–37.

Nelson C.R. (1972). ‘The prediction performance of the F.R.B.-M.I.T.-PENN models of the U.S. economy’. *American Economic Review* 62, 902–17.

Nelson, C.R. and Kang, H. (1981). 'Spurious periodicity in inappropriately detrended time series'. *Econometrica* 49, 741–51.

Nelson, C.R. and Kang, H. (1984). 'Pitfalls in the use of time as an explanatory variable in regression'. *Journal of Business and Economic Statistics* 2, 73–82.

Nelson, C.R. and Plosser, C.I. (1982). 'Trends and random walks in macroeconomic time series: some evidence and implications'. *Journal of Monetary Economics* 10, 139–62.

Nerlove, M. (1964). 'Spectral analysis of seasonal adjustment procedures'. *Econometrica* 32, 241–86.

Nerlove, M. (1965). 'A comparison of a modified "Hannan" and the BLS seasonal adjustment filters'. *Journal of the American Statistical Association* 60, 442–91.

Nerlove, M., Grether, D.M and Carvalho, J.L. (1979). *Analysis of Economic Time Series: A Synthesis*, New York: Academic Press.

Nerlove, M. and Wage, S. (1964). 'On the optimality of adaptive forecasting'. *Management Science* 10, 207–24.

Newbold, P. (1980). 'The equivalence of two tests of model adequacy'. *Biometrika* 67, 463–5.

Niederhoffer, V. and Osborne, M.F.M. (1966). 'Market making and reversal on the stock exchange'. *Journal of the American Statistical Association* 61, 897–916.

Norton, J. (1902). *Statistical Studies in the New York Money Market.* New York: Macmillan.

Orcutt, G.H. (1948). 'A study in the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States, 1919–1932'. *Journal of the Royal Statistical Society, Series B* 10, 1–53.

Orcutt, G.H. and James, S.F. (1948). 'Testing the significance of correlation between time series'. *Biometrika* 35, 397–413.

Ord, J.K. (1984). 'In memoriam: Maurice George Kendall, 1907–1983'. *The American Statistician* 38, 36–7.

Osborne, M.F.M. (1959). 'Brownian motion in the stock market'. *Operations Research* 7, 145–73.

Osborne, M.F.M. (1962). 'Periodic structure in the Brownian motion of stock market prices'. *Operations Research* 10, 345–79.

Parzen, E. (1957a). 'On consistent estimates of the spectrum of a stationary series'. *Annals of Mathematical Statistics* 28, 329–48.

Parzen, E. (1957b). 'On choosing an estimate of the spectral density function of a stationary time series'. *Annals of Mathematical Statistics* 28, 921–32.

Parzen, E. (1958). 'On asymptotically efficient consistent estimates of the spectral density function of a stationary time series'. *Journal of the Royal Statistical Society, Series B* 20, 303–22.

Parzen, E. (1961). 'Mathematical considerations in the estimation of spectra'. *Technometrics* 3, 167–90.

Patterson, K.D. (2010) *A Primer for Unit Root Testing.* Basingstoke: Palgrave Macmillan.

Paulsen, J. (1984). 'Order determination of multivariate autoregressive time series with unit roots'. *Journal of Time Series Analysis* 5, 115–27.

Pearson, E.S. (1922). 'On the variations in personal equation and the correlation of successive judgments'. *Biometrika* 14, 23–102.

Pearson, E.S. (1936/38). 'Karl Pearson: An appreciation of some aspects of his life and work, in two parts'. *Biometrika* 28, 193–257; 29, 161–247.

Pearson, E.S. (1938). *Karl Pearson: An Appreciation of Some Aspects of his Life and Work*. Cambridge: Cambridge University Press.

Pearson, E.S. (1950). "Student' as statistician'. *Biometrika* 30, 205–50.

Pearson, E.S. (1990). *'Student': A Statistical Biography of William Sealy Gosset*. Edited and augmented by R.L. Plackett with the assistance of G.A. Barnard. Oxford: Oxford University Press.

Pearson, K. (1896). 'Mathematical contributions to the theory of evolution, III: regression, heredity and panmixia'. *Philosophical Transactions of the Royal Society of London, Series A* 187, 253–318.

Pearson, K. (1912). 'The intensity of natural selection in man'. *Proceedings of the Royal Society of London, Series B* 85, 469–76.

Pearson, K. (1920). 'Notes on the history of correlation'. *Biometrika* 13, 25–45.

Pearson, K. and Elderton, E.M. (1923). 'On the variate difference method'. *Biometrika* 14, 281–310.

Pearson, K. and Filon, L.N.G. (1898). 'Mathematical contributions to the theory of evolution, IV: on the probable errors of frequency constants and on the influence of random selection on variation and correlation'. *Philosophical Transactions of the Royal Society of London, Series A* 191, 229–311.

Pearson, K. and Lee, A. (1897). 'On the distribution of frequency (variation and correlation) of the barometric height of divers stations'. *Philosophical Transactions of the Royal Society of London, Series A* 190, 423–69.

Pearson, K. and Rayleigh, Lord (1905). 'The problem of the random walk', *Nature* 72, 294, 318, 342.

Pedregal, D.J. and Young, P.C. (2001). 'Some comments on the use and abuse of the Hodrick–Prescott filter'. *Review on Economic Cycles* 2 (www.uned.es/imaec2000/issue2.htm).

Pegels, C. (1969). 'Exponential forecasting: some new variations'. *Management Science* 15, 311–15.

Peña, D. (2001). 'George Box: An interview with the *International Journal of Forecasting*'. *International Journal of Forecasting* 17, 1–9.

Perron, P. (1989). 'The Great Crash, the oil price shock, and the unit root hypothesis'. *Econometrica* 57, 1361–401.

Perron, P. (2006). 'Dealing with structural breaks', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 278–352.

Persons, W.M. (1916). 'Construction of a business barometer based upon annual data'. *American Economic Review* 6, 739–69.

Persons, W.M (1917). 'On the variate difference correlation method and curve-fitting'. *Publications of the American Statistical Association* 15, 602–42.

Persons, W.M. (1919). 'Indices of business conditions'. *Review of Economic Statistics* 1, 5–107.

Persons, W.M. (1923). 'Correlation of time series'. *Journal of the American Statistical Association* 18, 713–26.

Persons, W.M. (1924a). 'Some fundamental concepts of statistics'. *Journal of the American Statistical Association* 19, 1–8.

Persons, W.M. (1924b). *The Problem of Business Forecasting.* Pollak Foundation for Economic Research Publications, No. 6, London: Pitman.

Phillips, P.C.B. (1986). 'Understanding spurious regressions in econometrics'. *Journal of Econometrics* 33, 311–40.

Phillips, P.C.B. (1987a). 'Towards a unified asymptotic theory for autoregression'. *Biometrika* 74, 535–47.

Phillips, P.C.B. (1987b). 'Time series regression with a unit root'. *Econometrica* 55, 277–301.

Phillips, P.C.B. (1989). 'Spectral regression for cointegrated time series', in W.A. Barnett, J. Powell and G. Tauchen (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press, pp. 413–35.

Phillips, P.C.B. and Ouliaris, S. (1990). 'Asymptotic properties of residual based tests for cointegration'. *Econometrica* 58, 165–94.

Pierce, D.A. (1977). 'Relationships – and lack thereof – between economic time series, with special reference to money and interest rates'. *Journal of the American Statistical Association* 72, 11–22.

Pierce, D.A. (1979). 'Signal extraction error in nonstationary time series'. *Annals of Statistics* 7, 1303–20.

Plackett, R.L. (1950). 'Some theorems in least squares'. *Biometrika* 37, 149–57.

Plosser, C.I. and Schwert, G.W. (1978). 'Money, income and sunspots: measuring economic relationships and the effects of differencing'. *Journal of Monetary Economics* 4, 637–60.

Plummer, T. (1989). *Forecasting Financial Markets: The Truth Behind Technical Analysis*. London: Kogan Page.

Pollock, D.S.G. (1999). *A Handbook of Time-Series Analysis, Signal Processing and Dynamics*. San Diego: Academic Press.

Pollock, D.S.G. (2000). 'Trend estimation and de-trending via rational square wave filters'. *Journal of Econometrics* 99, 317–34.

Poirier, D.J. and Tobias, J.L. (2006). 'Bayesian econometrics', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 841–70.

Porter, T.M. (2004). *Karl Pearson: the Scientific Life in a Statistical Age*. Princeton, NJ: Princeton University Press.

Poskitt, D.S. and Tremayne, A.R. (1980). 'Testing the specification of a fitted autoregressive-moving average model'. *Biometrika* 67, 359–63.

Poynting, J.H. (1884). 'A comparison of the fluctuations in the price of wheat and in cotton and silk imports into Great Britain'. *Journal of the Royal Statistical Society* 47, 34–74.

Priestley, M.B. (1962). 'Basic considerations in the estimation of spectra'. *Technometrics* 4, 551–64.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. San Diego: Academic Press.

Quenneville, B., Ladiray, D. and Lefrançois, B. (2003). 'A note on Musgrave asymmetrical trend-cycle filters'. *International Journal of Forecasting* 19, 727–34.

Quenouille, M.H. (1947a). 'Notes on the calculation of autocorrelations of linear autoregressive schemes'. *Biometrika* 34, 365–7.

Quenouille, M.H. (1947b). 'A large sample test for the goodness of fit of autoregressive schemes'. *Journal of the Royal Statistical Society, Series A* 110, 123–9.

Quenouille, M.H. (1948). 'Some results in the testing of serial correlation coefficients'. *Biometrika* 35, 261–7.

Quenouille, M.H. (1949a). 'Approximate tests of correlation in time-series'. *Journal of the Royal Statistical Society, Series B* 11, 68–84.

Quenouille, M.H. (1949b). 'On a method of trend elimination'. *Biometrika* 36, 75–91.

Quenouille, M.H. (1949c). 'The joint distribution of serial correlation coefficients'. *Annals of Mathematical Statistics* 20, 561–71.

Quenouille, M.H. (1951). 'The variate-difference method in theory and practice'. *Review of the International Statistical Institute* 19, 121–9.

Quenouille, M.H. (1953). 'Modifications to the variate-difference method'. *Biometrika* 40, 383–408.

Quenouille, M.H. (1957). *The Analysis of Multiple Time-Series*. London: Griffin.

Rao, M.M. (1961). 'Consistency and limit distributions of estimators of parameters in explosive stochastic difference equations'. *Annals of Mathematical Statistics* 32, 195–218.

Ritchie-Scott, A. (1915). 'Note on the probable error of the coefficient of correlation in the variate difference method'. *Biometrika* 11, 136–8.

Rissanen, J. (1978). 'Modeling by shortest data description'. *Automatica* 14, 465–71.

Roberts, H.V. (1959). 'Stock-market "patterns" and financial analysis: Methodological suggestions'. *Journal of Finance* 14, 1–10.

Robinson, P.M. (2003). 'Long memory time series', in P.M. Robinson (ed.), *Time Series with Long Memory*. Oxford: Oxford University Press, pp. 4–32.

Romanovsky, V.I. (1932). 'Sur la loi sinusoidale limite'. *Rendiconti del Circolo Matematico di Palermo* 56, 82–111.

Romanovsky, V.I. (1933). 'Sur une généralization de la loi sinusoidale limite'. *Rendiconti del Circolo Matematico di Palermo* 57, 130–6.

Rosenblatt, H.M. (1968). 'Spectral evaluation of BLS and Census revised seasonal adjustment procedures'. *Journal of the American Statistical Association* 63, 472–501.

Rubin, H. (1945). 'On the distribution of the serial correlation coefficient'. *Annals of Mathematical Statistics* 16, 211–15.

Rubin, H. (1950). 'Consistency of maximum likelihood estimates in the explosive case', in T.C. Koopmans (ed.), *Statistical Inference in Dynamic Linear Models*. New York: Wiley, pp. 356–64.

Samuelson, P.A. (1987). 'Paradise lost and refound: the Harvard ABC barometers'. *Journal of Portfolio Management* 4, Spring, 4–9.

Sargan, J.D. (1980). 'Some tests of dynamic specification for a single equation'. *Econometrica* 48, 879–97.

Sargan, J.D. and Bhargava, A.S. (1983). 'Testing residuals from least squares regression for being generated by the Gaussian random walk'. *Econometrica* 51, 153–74.

Sargent, T.J. and Sims, C.A. (1977). 'Business cycle modeling without pretending to have too much a priori economic theory', in C.A. Sims (ed.) *New Methods of Business Cycle Research*. Minneapolis, MN: Federal Reserve Bank of Minneapolis, pp. 45–110.

Schuster, A. (1897). 'On lunar and solar periodicities of earthquakes'. *Proceedings of the Royal Society of London* 61, 455–65.

Schuster, A. (1898). 'On the investigation of hidden periodicities with application to a supposed 26 day period in meteorological phenomena'. *Terrestrial Magnetism* 3, 13–41.

Schuster, A. (1906). 'On the periodicities of sunspots'. *Philosophical Transactions of the Royal Society of London, Series A* 206, 69–100.

Schwarz, G. (1978). 'Estimating the dimension of a model'. *Annals of Statistics* 6, 461–4.

Schweppe, F.C. (1965). 'Evaluation of likelihood functions for Gaussian signals'. *IEEE Transactions on Information Theory* 11, 61–70.

Scott, S. (1992). 'An extended review of the X11ARIMA seasonal adjustment package'. *International Journal of Forecasting* 8, 627–33.

Scott, S. (1997). 'Adjusting from X-11 to X-12'. *International Journal of Forecasting* 13, 567–82.

Shenton, L.R. and Johnson, W.L. (1965). 'Moments of a serial correlation coefficient'. *Journal of the Royal Statistical Society, Series B* 27, 308–20.

Shiryaev, A.N. (1960). 'Some problems in the spectral theory of higher-order moments, I'. *Theory of Probability and its Applications* 5, 265–84.

Shiskin, J. (1955). 'Seasonal computations on Univac'. *American Statistician* 9, 19–23.

Shiskin, J. and Eisenpress, H. (1957). 'Seasonal adjustments by electronic computer methods'. *Journal of the American Statistical Association* 52, 415–49.

Shiskin, J., Young, A.H. and Musgrave, J.C. (1967). 'The X-11 variant of the Census Method II seasonal adjustment program'. Bureau of the Census Technical Paper No. 15, US Department of Commerce. Available at www.census.gov/srd/www/sapaper/historicpapers.html.

Sims, C.A. (1972). 'Money, income and causality'. *American Economic Review* 62, 540–52.

Sims, C.A. (1974). 'Seasonality in regression'. *Journal of the American Statistical Association* 69, 618–27.

Sims, C.A. (1980). 'Macroeconomics and reality'. *Econometrica* 48, 1–48.

Slutzky, E. (1927). 'The summation of random causes as the source of cyclic processes'. *The Problems of Economic Conditions*, edited by the Conjucture Institute, Moscow, 3:1, 34–64 (English summary, 156–61).

Slutzky, E. (1937). 'The summation of random causes as the source of cyclic processes'. *Econometrica* 5, 105–46.

Smith, B.B. (1925). 'The error in eliminating secular trend and seasonal variation before correlating time series'. *Journal of the American Statistical Association* 20, 543–5.

Smith, B.B. (1926). 'Combining the advantages of first-difference and deviation-from-trend methods of correlating time series'. *Journal of the American Statistical Association* 21, 55–9.

Solanki, S.K., Usoskin, I.G., Kromer, B., Schűssler, M. and Beer, J. (2004). 'Unusual behaviour of the Sun during recent decades compared to the previous 11,000 years'. *Nature* 431, 1084–7.

Spencer, J. (1904). 'On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893–97'. *Journal of the Institute of Actuaries* 38, 334–43.

Spencer, J. (1907). 'Some illustrations of the employment of summation formulas in the graduation of mortality tables'. *Journal of the Institute of Actuaries* 41, 361–408.

Spencer-Smith, J.L. (1947). 'The oscillatory properties of the moving average'. *Journal of the Royal Statistical Society, Series B* 9, 104–13.

Stephenson, J.A. and Farr, H.T. (1972). 'Seasonal adjustment of economic data by application of the general linear statistical model'. *Journal of the American Statistical Association* 67, 37–45.

Stewart, B. and Dodgson, W. (1879). 'Preliminary report to the Committee on Solar Physics on a method of detecting the unknown inequalities of a series of observations'. *Proceedings of the Royal Society of London* 29, 106–23.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stratonovich, R.L. (1960). 'Application of the Markov processes theory to optimal filtering'. *Radio Engineering and Electronic Physics* 5:11, 1–19.

Stuart, A. (1984). 'Sir Maurice Kendall, 1907–1983'. *Journal of the Royal Statistical Society, Series A* 147, 120–2.

Student (W.S. Gosset) (1914). 'The elimination of spurious correlation due to positions in time and space'. *Biometrika* 10, 179–80.

Swerling, P. (1959). 'First-order error propagation in a stagewise smoothing procedure for satellite observations'. *US Air Force Project Rand,* Research Memorandum-2329.

Teräsvirta, T. (2006). 'Univariate nonlinear time series models', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 396–424.

Theil, H. and Wage, S. (1964). 'Some observations on adaptive forecasting'. *Management Science* 10, 198–206.

Thomas, J.J. and Wallis, K.F. (1971). 'Seasonal variation in regression analysis'. *Journal of the Royal Statistical Society, Series A* 134, 57–72.

Tiao, G.C. (1985). 'Autoregressive moving average models, intervention problems and outlier detection in time series', in E.J. Hannan, P.R. Krishnaiah and M.M. Rao (eds), *Handbook of Statistics, Volume 5: Time Series in the Time Domain.* Amsterdam: North-Holland, pp. 85–118.

Tiao, G.C. and Box, G.E.P. (1981). 'Modelling multiple time series with applications'. *Journal of the American Statistical Association* 76, 802–16.

Tiao, G.C. and Hillmer, S.C. (1978). 'Some consideration of decomposition of a time series'. *Biometrika* 65, 497–502.

Tinbergen, J. (1937). *An Econometric Approach to Business Cycle Problems*. Paris: Hermann & Cie.

Tinbergen, J. (1939). *Statistical Testing of Business Cycle Theories. Volume II: Business Cycles in the United States of America, 1919–1932*. Geneva: League of Nations.

Tintner, G. (1940). *The Variate Difference Method*. Bloomington, IN: Principia Press.

Tintner, G. and Kadekodi, G. (1973). 'A note on the use of differences and transformations in the estimation of econometric relations'. *Sankhyä: The Indian Journal of Statistics, Series B* 35, 268–77.

Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*. Oxford: Oxford University Press.

Tsay, R.S. (1986). 'Time series model specification in the presence of outliers'. *Journal of the American Statistical Association* 81, 132–41.

Tsay, R.S. (1988). 'Outliers, level shifts, and variance changes in time series'. *Journal of Forecasting* 7, 1–20.

Tukey, J.W. (1961). 'Discussion, emphasizing the connection between analysis of variance and spectrum analysis'. *Technometrics* 3, 191–219.

Velasco, C. (2006). 'Semiparametric estimation of long-memory models', in T.C. Mills and K.D. Patterson (eds) *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*. Basingstoke: Palgrave Macmillan, pp. 353–95.

Von Neumann, J. (1941). 'Distribution of the ratio of the mean square successive differences to the variance'. *Annals of Mathematical Statistics* 12, 367–95.

Von Neumann, J. (1942). 'A further remark concerning the distribution of the ratio of the mean square successive differences to the variance'. *Annals of Mathematical Statistics* 13, 86–8.

Walker, A.M. (1950). 'Note on a generalization of the large sample goodness of fit test for linear autoregressive schemes'. *Journal of the Royal Statistical Society, Series B* 12, 102–7.

Walker, A.M. (1954). 'The asymptotic distribution of serial correlation coefficients for autoregressive processes with dependent residuals'. *Proceedings of the Cambridge Philosophical Society* 50, 60–4.

Walker, A.M. (1961). 'Large-sample estimation of parameters for moving average models'. *Biometrika* 48, 343–57.

Walker, A.M. (1962). 'Large-sample estimation of parameters for autoregressive processes with moving-average residuals'. *Biometrika* 49, 117–31.

Walker, G.T. (1925). 'On periodicity'. *Quarterly Journal of the Royal Meteorological Society* 51, 337–45.

Walker, G.T. (1931). 'On periodicity in series of related terms'. *Proceedings of the Royal Society of London, Series A* 131, 518–32.

Walker, G.T. (1950). 'Apparent correlation between independent series of autocorrelated observations'. *Biometrika* 37, 184–5.

Walker, J.M. (1997). 'Pen portrait of Sir Gilbert Walker, CSI, MA, ScD, FRS'. *Weather* 52, 217–20.

Wallis, K.F. (1974). 'Seasonal adjustment and relations between variables'. *Journal of the American Statistical Association* 69, 18–31.

Wallis, K.F. (1982). 'Seasonal adjustment and revision of current data: linear filters for the X-11 method'. *Journal of the Royal Statistical Society, Series A* 145, 74–85.

Watson, G.S. and Durbin, J. (1951). 'Exact tests of serial correlation using noncircular statistics'. *Annals of Mathematical Statistics* 22, 446–51.

Watson, M.W. (1986). 'Univariate detrending methods with stochastic trends'. *Journal of Monetary Economics* 18, 49–75.

Weber, C.E. (1999). 'Slutsky (1915) and additive utility functions, 1947–1972'. *History of Political Economy* 31, 393–416.

Weiner, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York: Wiley.

Weiner, N. (1958). *Nonlinear Problems in Random Theory*. Cambridge, MA: MIT Press.

Weiss, G. (1986). 'Random walks', in *Encyclopedia of Statistical Sciences*, vol. 7. New York: Wiley, pp. 574–80.

West, M. and Harrison, P.J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd edition. New York: Springer-Verlag.

West, M., Harrison, P.J. and Migon, H.S. (1985). 'Dynamic generalized linear models and Bayesian forecasting'. *Journal of the American Statistical Association* 80, 73–97.

White, J.S. (1958). 'The limiting distribution of the serial correlation coefficient in the explosive case'. *Annals of Mathematical Statistics* 29, 1188–97.

White, J.S. (1959). 'The limiting distribution of the serial correlation coefficient in the explosive case II'. *Annals of Mathematical Statistics* 30, 831–4.

White, J.S. (1961). 'Asymptotic expansions for the mean and variance of the serial correlation coefficient'. *Biometrika* 48, 85–94.

Whittaker, E.T. (1923). 'On a new method of graduation'. *Proceedings of the Edinburgh Mathematical Society* 41, 63–75.

Whittaker, E.T. (1924). 'On a theory of graduation'. *Proceedings of the Royal Society of Edinburgh* 44, 77–83.

Whittaker, E.T. and Robinson, G. (1924). *The Calculus of Observations: A Treatise on Numerical Mathematics*. Glasgow: Blackie & Son.

Whittle, P. (1952). 'Tests of fit in time series'. *Biometrika* 39, 309–18.

Whittle, P. (1953a). 'Estimation and information in stationary time series analysis'. *Arkiv fő r Matematik* 2, 423–34.

Whittle, P. (1953b). 'The analysis of multiple stationary time series'. *Journal of the Royal Statistical Society, Series B* 15, 125–39.

Whittle, P. (1954). 'Some recent contributions to the theory of stationary processes'. Appendix 2 of Wold (1954).

Whittle, P. (1957). 'Curve and periodogram smoothing'. *Journal of the Royal Statistical Society, Series B* 19, 38–47.

Whittle, P. (1963). *Prediction and Regulation by Linear Least-Square Methods*. London: The English Universities Press.

Williams, R.H. (2004). 'George Udny Yule: Statistical Scientist'. *Human Nature Review* 4, 31–7.

Winters, P.R. (1960). 'Forecasting sales by exponentially weighted moving averages'. *Management Science* 6, 324–42.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist & Wiksell.

Wold, H. (1949). 'A large-sample test for moving averages'. *Journal of the Royal Statistical Society, Series B* 11, 297–305.

Wold, H. (1954). *A Study in the Analysis of Stationary Time Series, 2nd edition with an appendix by Peter Whittle*. Stockholm: Almqvist & Wiksell.

Wold, H. (1982). 'Models for knowledge', in J. Gani (ed.), *The Making of Statisticians*. Berlin: Springer.

Working, H. (1934). 'A random-difference series for use in the analysis of time series'. *Journal of the American Statistical Association* 29, 11–24.

Working, H. and Hotelling, H. (1929). 'Applications of the theory of error to the interpretation of trends'. *Journal of the American Statistical Association* 24, 73–85.

Yaglom, A.M. (1955). 'The correlation theory of processes whose *n*th differences constitute a stationary process'. *Matem. Sbornik* 37, 141–96.

Young, A.H. (1968). 'Linear approximations to the Census and BLS seasonal adjustment methods'. *Journal of the American Statistical Association* 63, 445–71.

Young, P.C. (1984). *Recursive Estimation and Time Series Analysis*. Berlin: Springer-Verlag.

Young, P.C. (2011). 'Gauss, Kalman and advances in recursive parameter estimation'. *Journal of Forecasting* 30, 104–46.

Yule, G.U. (1895). 'On the correlation of total pauperism with proportion of out-relief, I: all ages'. *Economic Journal* 5, 603–11.

Yule, G.U. (1896). 'On the correlation of total pauperism with proportion of out-relief, II: males over sixty-five'. *Economic Journal* 6, 613–23.

Yule, G.U. (1897a). 'On the significance of Bravais' formula for regression, &c, in the case of skew variables'. *Proceedings of the Royal Society of London* 60, 477–89.

Yule, G.U. (1897b). 'On the theory of correlation'. *Journal of the Royal Statistical Society* 60, 812–54.

Yule, G.U. (1899). 'An investigation into the causes of changes in pauperism in England during the last two intercensal decades, I'. *Journal of the Royal Statistical Society* 62, 249–95.

Yule, G.U. (1906). 'On changes in the marriage- and birth-rates in England and Wales during the past half-century, with an inquiry as to their probable causes'. *Journal of the Royal Statistical Society* 69, 88–147.

Yule, G.U. (1907). 'On the theory of correlation for any number of variables, treated by a new system of notation'. *Proceedings of the Royal Society of London, Series A* 79, 182–93.

Yule, G.U. (1921). 'On the time-correlation problem, with especial reference to the variate-difference correlation method'. *Journal of the Royal Statistical Society* 84, 497–537.

Yule, G.U. (1926). 'Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time series'. *Journal of the Royal Statistical Society* 89, 1–63.

Yule, G.U. (1927). 'On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers'. *Philosophical Transactions of the Royal Society of London, Series A* 226, 267–98.

Yule, G.U. (1944a). 'Reginald Hawthorn Hooker, M.A.'. *Journal of the Royal Statistical Society* 107, 74–7.

Yule, G.U. (1944b). *The Statistical Study of Literary Vocabulary.* Cambridge, Cambridge University Press.

Yule, G.U. (1945). On a method of studying time-series based on their internal correlations'. *Journal of the Royal Statistical Society* 108, 208–25.

Yule, G.U. and Kendall, M.G. (1950). *An Introduction to the Theory of Statistics*, 14th edition, London: Griffin.

Zaycoff, R. (1937). 'Ueber die ausschalung der zufälligen komponente nach der "variate-difference" methode'. *Publications of the Statistical Institute for Economic Research*, No. 1, State University of Sofia.

Index