PALGRAVE ADVANCED TEXTS IN ECONOMETRICS

A VERY BRITISH AFFAIR

SIX BRITONS AND THE DEVELOPMENT OF TIME SERIES ANALYSIS DURING THE 20TH CENTURY

TERENCE C. MILLS

A Very British Affair

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.

Titles include:

Terence C. Mills
THE FOUNDATIONS OF MODERN TIME SERIES ANALYSIS

Terence C. Mills
A VERY BRITISH AFFAIR: SIX BRITONS AND THE DEVELOPMENT OF TIME SERIES ANALYSIS DURING THE 20TH CENTURY

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.

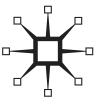
Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke, Hampshire RG21 6XS, England.

A Very British Affair

Six Britons and the Development of Time Series Analysis During the Twentieth Century

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

palgrave
macmillan



Softcover reprint of the hardcover 1st edition 2013 978-0-230-36911-5

First published 2013 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 978-1-349-35027-8 ISBN 978-1-137-29126-4 (eBook)
DOI 10.1057/9781137291264

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
22 21 20 19 18 17 16 15 14 13

Contents

List of Tables

List of Figures

1
Time Series Analysis and the British

1.1 During the writing of *The Foundations of Modern Time Series Analysis* (Mills, 2011a), it became apparent to me just how important was the role played by British statisticians in the development of the subject. Although that book focused on the period up to 1970, British statisticians have since continued to be at the forefront of time series research and this led naturally to the conceit of the present book: that the story of time series analysis can be told largely through the published research of six Britons, beginning with George Udny Yule and seguing through Maurice Kendall, James Durbin, George Box and Gwilym Jenkins, before reaching finally Clive Granger. This distinguished group of statisticians includes three Englishmen, two Welshmen and a Scot, two knights of the realm, one Nobel prize winner, a holder of the UN Peace Medal, three Presidents of the Royal Statistical Society (RSS) and four recipients of the Guy Medal in Gold, the highest honour awarded by the RSS, as well as the holding of numerous international awards and honours. In terms of university affiliations, three have been associated with St John's College, Cambridge and four with University College, London (UCL), with Imperial College, the London School of Economics (LSE), Lancaster and Nottingham also featuring, along with the American universities of Stanford, Princeton, North Carolina, Wisconsin at Madison and San Diego.

1.2 The links between time series analysis and economic statistics and econometrics have always been particularly strong and have become increasingly so over recent years. I must therefore emphasize that my focus in this book is on time series analysis and time series statisticians, not on the equally eminent and influential group of British time series econometricians, such as Denis Sargan and David Hendry. Econometrics

already has its own histories – Epstein (1987), Morgan (1990), Qin (1993), Hendry and Morgan (1995), Spanos (2006), Farebrother (2006) and Gilbert and Qin (2006) are notable examples – with the influential work of the 'LSE group' on time series econometrics being the topic of Gilbert (1989). Time series analysis, on the other hand, appears to have just Klein (1997) and Mills (2011a).

1.3 My choice of these six British time series analysts seemed to me uncontroversial but when I mentioned this project to Robert Taylor, he immediately mentioned our mutual friend and colleague Paul Newbold, arguing that if he was included in the group I could then refer to the 'Magnificent Seven'! Much as I was tempted, Paul's most influential work has been written jointly with George Box and, in particular and most memorably, with Clive Granger, so he appears regularly as a co-author in Chapters 8 to 10. Nevertheless, in recognition of Paul's contribution to the subject, it gives me great pleasure to dedicate this book to him. Those colleagues of Paul who know of his struggles with ill-health over the last few years will, I hope, join with me in honouring him here as a fine time series analyst and colleague whose presence always enlivened academic discussion. Granger and Leybourne (2009) provide an appreciation of Paul Newbold's contributions to time series and econometrics as part of the *Econometric Theory* special issue (Volume 25, No. 6) dedicated to him.

1.4 Some of the material in Mills (2011a) unavoidably finds its way into this book and, although the theme here is narrower in scope, it takes the story on a further 40 years. I have again employed the format of subheading and section number used in that book, so that a cross-reference to section **y** of Chapter **x** is denoted **§x.y** in subsequent chapters. The structure of the book is straightforward and, as it turns out, almost chronologically ordered, beginning with Yule's initial forays into time series analysis in the early 1920s and ending with Granger's contemporary research throughout the first decade of the twenty-first century, with a final chapter praising the pragmatism of these giants of the subject and presenting my views on the current state of the subject and the likelihood of the further involvement by British statisticians in the future. Nothing more needs to be said by way of preamble, so let us begin our story of a fascinating and increasingly important area of statistics.

2

Yule: The Time–Correlation Problem, Nonsense Correlations, Periodicity and Autoregressions

George Udny Yule

2.1 George Udny Yule was born on 18 February 1871 in Beech Hill near Haddington, Scotland, into 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 UCL and Bonn, Germany, publishing four papers on electric waves, Yule returned to UCL in 1893 as a demonstrator for the distinguished statistician (and much else besides!) Karl Pearson, later becoming an Assistant Professor. Yule left UCL in 1899 to work for the City and Guilds of London Institute, although he was also to hold 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 active in the RSS: elected a Fellow in 1906, he served as Honorary Secretary, was President from 1924 to 1926, and was awarded the prestigious Guy Medal in Gold in 1911. His textbook, *Introduction to the Theory of Statistics*, ran to 14 editions (the last four being co-authored with Maurice Kendall: see Chapter 3) and, as well as contributing massively to the foundations of time series analysis, he also researched on Mendelian inheritance (see, in particular, Yule, 1902, 1914, and for discussion, Tabery, 2004) and on the statistics of literary style (Yule, 1944, 1946), as well as on many 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, heart problems from 1932 curtailed his

flying experiences and he became a semi-invalid for the rest of his life, dying on 26 June 1951 in Cambridge.

Much of Yule's early research in statistics was on developing the theory of correlation and regression, with applications to both economic and sociological topics (see, in particular, Yule 1897a, 1897b, 1907, 1910). Historical perspectives on these aspects of Yule's work, which are not our major concern or focus here, but are arguably extremely important for the development of applied statistical techniques, have been provided by Aldrich (1995, 1998) and Hepple (2001). For further biographical details and a full list of publications, see Kendall (1952) and Williams (2004), while Stuart and Kendall (1971) provided a collection of Yule's major papers for the centenary of his birth.

The time–correlation problem

2.2 Yule's first major foray into time series analysis was a paper in the *Journal of the Royal Statistical Society* in 1921, which he began by surveying the literature on what he termed the 'time correlation problem' as at 1914, summarizing it with his customary clarity thus:

> the 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. It is unfortunate that the word 'periodic' implies rather too much as to the character of such more rapid movements; few of us, I suppose, now believe that they are strictly periodic in the proper sense of the term, and hence the occurrence in writings on the subject of such terms as 'quasi-periodic' and 'pseudo-periodic'. They are wave-like movements, movements which can be readily represented with a fair degree of accuracy over a moderate number of years by a series of harmonic terms but which cannot be represented in the same way, for example, by a polynomial; movements in which the length of time from crest to crest of successive waves is not constant, and in which, it may be added, the amplitude is not constant either, but would probably, if we could continue our observations over a sufficient number of waves, exhibit a frequency distribution with a fairly definite mode;

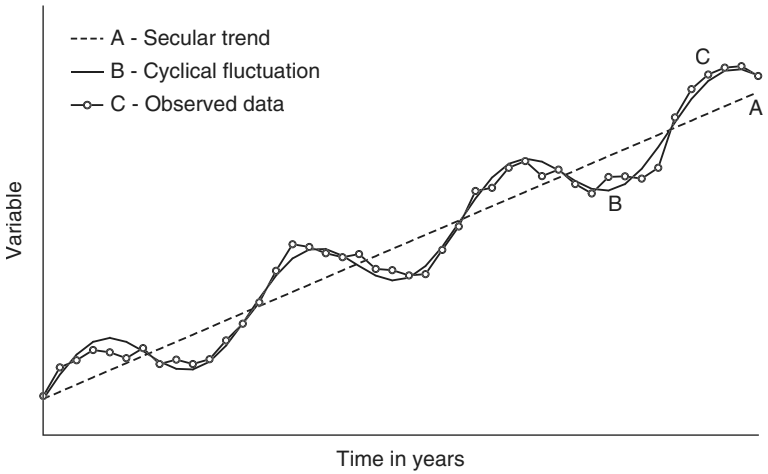


Figure 2.1 Idealistic representation of a time series as the sum of trend, cyclical and irregular components

> to avoid the suggestion of strict periodicity and the use of the term *period* I propose to speak of them as *oscillations* of a given *duration* to imply, not a fixed and constant duration, but an average only. In these terms, the problem of time-correlation may be said to be the isolation, for separate study, of oscillations of differing durations. Most writers up to 1914 – indeed all writers so far as I am aware – seem to be agreed on this. (Yule, 1921, page 501; italics in original)

An idealistic graphical representation of such a component structure of a time series, much in the spirit of Warren Persons (1917, Figure 1), whose views were similar to Yule's, is shown in Figure 2.1. 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.

2.3 Why Yule emphasized 1914 was because in that year the famous statistician Student, in his only published paper on time series analysis, introduced what subsequently became known as the *variate-difference* method, in which he advocated correlating the successive differences of pairs of time series:[1]

> 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 … dth differences between successive values of our variable with the 1st, 2nd, 3rd … dth differences between successive values of the other variable. When the correlation between the two dth differences is equal to that between the two $(d + 1)$th differences, this value gives the correlation required. (Student, 1914, page 180)

The variate-difference method had quickly been taken up with enthusiasm by Karl Pearson and his co-workers.

> 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. (Cave and Pearson, 1914, page 354)

> (T)here 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. (Elderton and Pearson, 1915, page 489)

Notwithstanding such enthusiasm, and the empirical results that the method generated in both economic and medical applications, Persons and Yule were both perplexed by the approach, the latter stating

> (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. (Yule, 1921, 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)

2.4 Yule then focused attention on the correlation that was induced into a series by differencing. Student (1914) had begun by assuming that two time series 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 then 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

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

Consider first $d = 1$. Then

$$\begin{aligned} \sigma_{\Delta y}^2 &= E(\Delta y_t^2) = E(y_t^2 - 2y_t y_{t-1} + y_{t-1}^2) = 2\sigma_y^2 \\ \sigma_{\Delta x}^2 &= 2\sigma_x^2 \\ E(\Delta y_t \Delta x_t) &= E(y_t x_t + y_{t-1} x_{t-1} - y_t x_{t-1} - y_{t-1} x_t) = 2r_{yx}\sigma_y\sigma_x \end{aligned} \tag{2.1}$$

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 generalized the argument by assuming that y_t and x_t were given by polynomials in time:

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

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,

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

we have

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

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 his conclusions quoted in **§2.3** above.

While Student was concerned with the correlation between pairs of time series, Yule investigated the correlation between *adjacent differences* of an individual random series, pointing out that, in our notation,

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

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

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

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 ${}_1 r_y(2) = 0$ and, by extension, ${}_1 r_y(k) = 0$ for $k > 1$. Having shown this, Yule then generalized these results to dth differences:

$$\begin{aligned} {}_d r_y(1) &= -\frac{d}{d+1}, \qquad {}_d r_y(2) = \frac{d(d-1)}{(d+1)(d+2)}, \\ {}_d r_y(k) &= (-1)^k \frac{d(d-1)\cdots(d-k+1)}{(d+1)(d+2)\cdots(d+k)} = (-1)^k \frac{d!d!}{(d-k)!(d+k)!} \qquad k \le d \end{aligned}$$

with ${}_d r_y(k) = 0$ for $k > d$.[2] Hence dth differences of a random series have non-zero 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)

2.5 Yule then extended the analysis by considering differences of the periodic function

$$y_t = \rho \sin\left(2\pi \frac{t+\alpha}{n}\right) = \rho \sin\left(\frac{2\pi t}{n} + \frac{2\pi\alpha}{n}\right) \tag{2.2}$$

where 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 (2.2) is

$$\begin{aligned}\Delta y_{t+h} &= \rho\left(\sin\left(2\pi\frac{t+\alpha+h}{n}\right)-\sin\left(2\pi\frac{t+\alpha}{n}\right)\right)\\ &= 2\rho\sin\left(\pi\frac{h}{n}\right)\cos\left(2\pi\frac{t+\alpha+0.5h}{n}\right)\\ &= 2\rho\sin\left(\pi\frac{h}{n}\right)\sin\left(2\pi\frac{t+\alpha+0.5h+0.25n}{n}\right)\end{aligned} \tag{2.3}$$

The second equality in (2.3) 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 eliminating 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. (ibid., page 509; italics in original)

2.6 Yule's overall conclusions concerning the time-correlation problem were that it was a

> problem of isolating, for the purpose of discussing the relations between them, oscillations of different durations – such oscillations being, in all probability, not strictly periodic but up-and-down movements of greater or less rapidity. …
>
> The problem is not that of isolating uncorrelated residuals.
>
> The variate-differencing method does not tend to isolate nor lay most stress on such uncorrelated residuals. It tends to stress preponderantly oscillations with a duration of two years, the actual weight of oscillations of two, three, four, five … year durations of *k*th differences naturally depending, however, on their relative amplitudes in the data.
>
> In so far as the problem consists in finding the relations between such shorter oscillations, eliminating or reducing the effects of others so far as may be possible, the variate-difference method may possibly be of service on appropriate data in which short oscillations are significant. The work already done by the method requires reinterpretation, however, in the light of the present discussion. (ibid., page 524)

Yule expressed an, albeit tentative, preference for isolating the components of a time series in the manner of Figure 2.1, as the method

> isolates or may isolate (subject to the use of suitable processes for determining the trend) the oscillations with relatively little distortion and – no mean advantage – they can be exhibited graphically, so that investigator and reader can see what are actually the movements considered. The variate-difference method distorts the actual oscillations, altering the various harmonic components in amplitude and phase. (ibid., page 524)

Not surprisingly, this critique produced a vigorous response by Pearson and Elderton (1923), in which it became clear that proponents of the

variate-differencing method conceived of a time series as being decomposed into just two components, a 'catch-all' component comprising both the secular trend and periodic fluctuations, modeled 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). 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 (UC) 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 non-stationarity (see **§§6.10–6.14**), 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 UC 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 'pre-whitening' approach (see **§§8.14–8.18** and Pierce, 1977), while the Persons–Yule idea of correlating the components has its descendants in Mills (1982) and Watson (1986).

Nonsense correlations between time series

2.7 In his Presidential Address to the RSS in November 1925, Yule considered a problem that had puzzled him for many years. Since it lies at the centre of all attempts to analyze the relationships between time series, Yule's statement of the problem is worth setting out in his own words:

> 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 between time-series ... it is important to clear up the problem of how they arise and in what special cases. [Figure 2.2] gives a very good illustration.

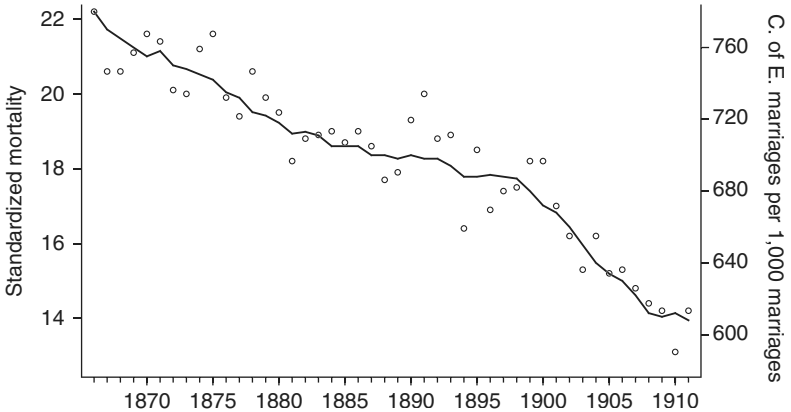


Figure 2.2 Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0.9512$. (Recreated from Yule, 1926, Fig. 1, page 3)

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 $(1 - r^2)/\sqrt{T}$, where T 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 Tcards (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 $(1 - r^2)/\sqrt{T}$. 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; notation altered for consistency)

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 2.2 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 analyze in his Presidential address.

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

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

where, as in (2.2), n is the period and α is now 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 $u \pm h$:

$$\bar{x}(u \pm h) = \frac{n}{2\pi h}\sin\left(2\pi\frac{u+\alpha}{n}\right)\sin\left(2\pi\frac{h}{n}\right) \tag{2.4}$$

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

$$s_x^2(u \pm h) = \frac{1}{2} - \frac{n}{8\pi h} \cos\left(4\pi \frac{u+\alpha}{n}\right) \sin\left(4\pi \frac{h}{n}\right) - \bar{x}^2(u \pm h) \qquad (2.5)$$

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

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

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

$$r_{yx}(u \pm h) = \frac{\overline{yx}(u \pm h) - \bar{y}(u \pm h)\bar{x}(u \pm h)}{s_y(u \pm h) s_x(u \pm h)} \qquad (2.6)$$

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 (2.4) and (2.5).

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 shown in Figure 2.3. 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, ... 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 2.4).

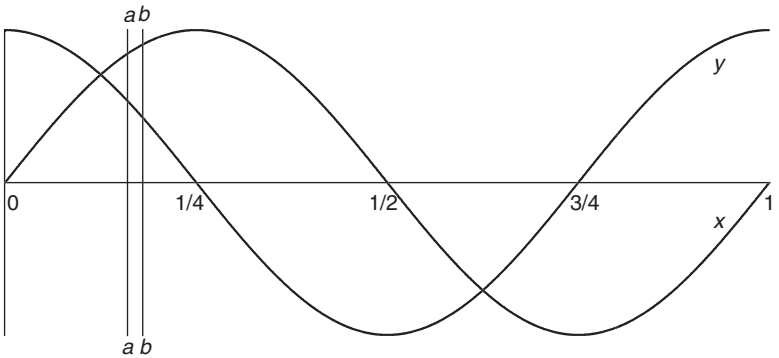


Figure 2.3 Two sine curves differing by a quarter-period in phase, and consequently uncorrelated when the correlation is taken over a whole period

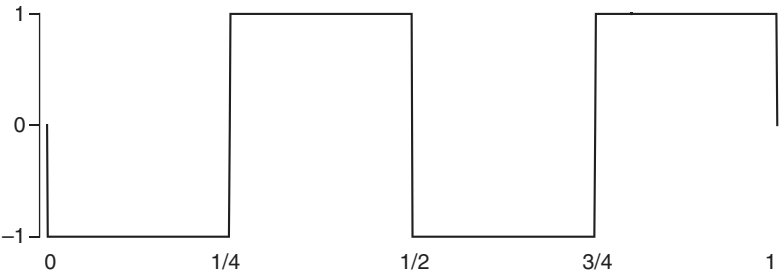


Figure 2.4 Variation of the correlation between two simultaneous intervals of the sine curves of Figure 2.3, as the centre of the interval is moved across from left to right

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

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

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

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

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

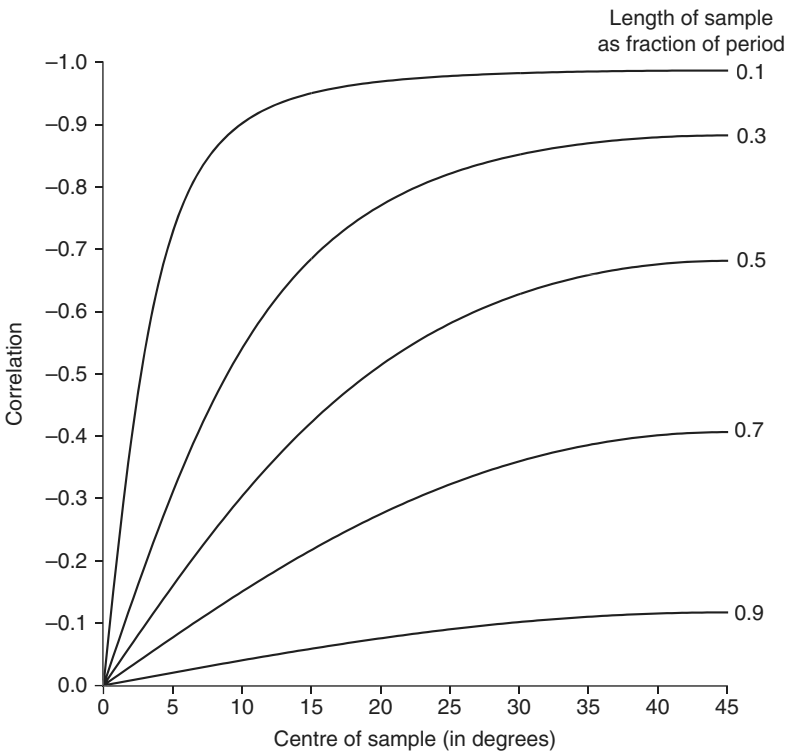


Figure 2.5 Variation of the correlation coefficient between two simultaneous finite intervals of the harmonic curves of Figure 2.3, when the length of the interval is 0.1, 0.3, ..., 0.9 of the period, as the centre of the interval is moved across from left to right; only one-eighth of the whole period shown

and

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

from which the correlation as u varies for a given value of h can be calculated from (2.6). Figure 2.5 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

> (t)he 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 2.4], 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 2.6 and led Yule to conclude that the[3]

> 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 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 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

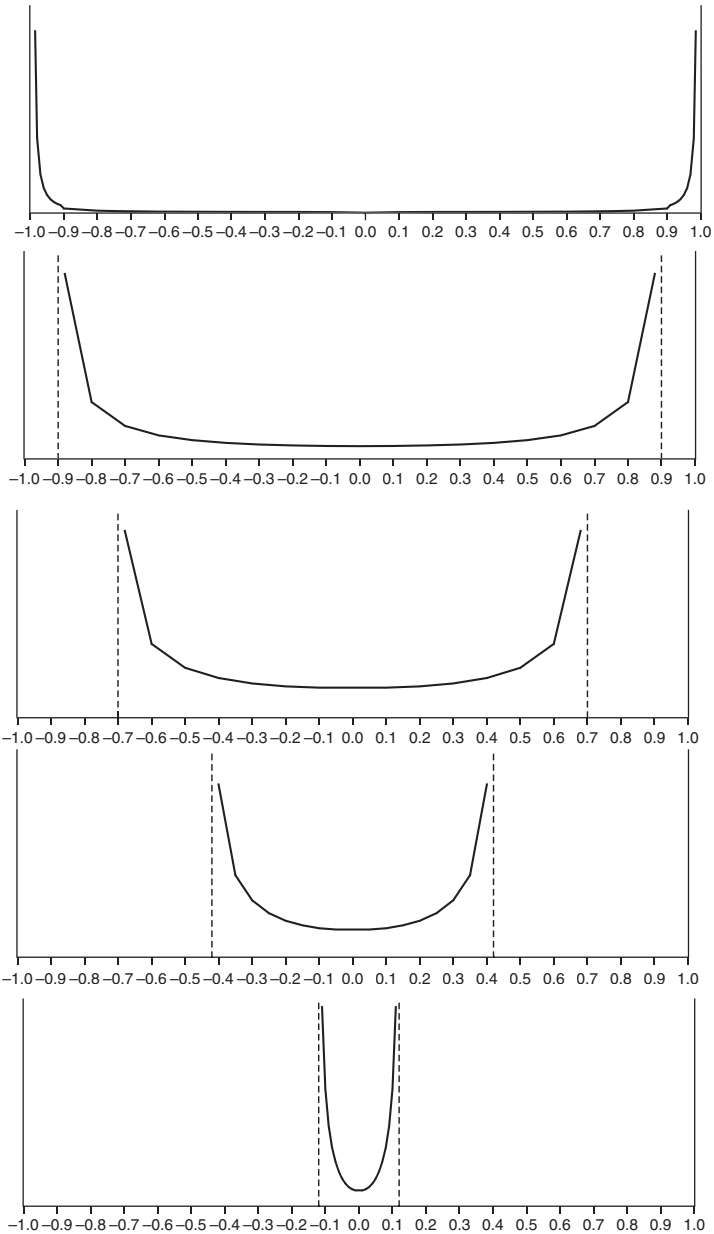


Figure 2.6 Frequency distribution of correlations between simultaneous intervals of the sine curves of Figure 2.3 when the interval is, from the top, 0.1, 0.3, 0.5, 0.7 and 0.9, respectively, of the period

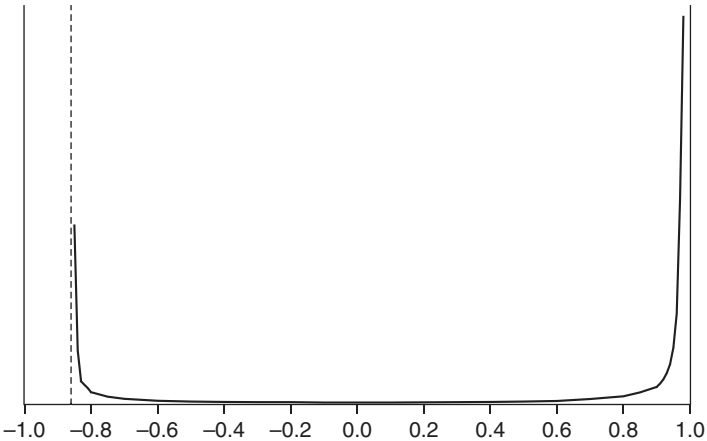


Figure 2.7 Frequency distribution of correlations between two simultaneous intervals of sine curves differing by 60° in phase (correlation over a whole period +0.5) when the length of interval is 0.2 of the period

frequency distribution is shown in Figure 2.7, and was described by Yule thus:[4]

> (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)

2.9 The implication of the analysis in **§2.8** 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

$$\begin{aligned}
\bar{x} &= \bar{x}(u \pm n/2) = \bar{y} = \bar{y}(u \pm n/2) = 0 \\
s_x^2 &= s_x^2(u \pm n/2) = s_y^2 = s_y^2(u \pm n/2) = 0.5 \\
\overline{yx} &= \overline{yx}(u \pm n/2) = \frac{1}{2} \cos\left(2\pi \frac{\alpha}{n}\right)
\end{aligned}$$

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, adapting the notation used previously in **§2.4**, 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 downwards movement from left to right. Conversely, if the first term is below the average such a plot will

show an upwards 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 2.3.

> 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)

2.10 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 $-1/(T-1)$.[5] 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 1000, the most probable deviation of each of the other terms is $-1000/9$ 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 2.1 but, to avoid having to physically repeat Yule's rather heroic sampling procedure, we utilize modern computing power and software![6] Column (3) gives the average deviations for the first deviation 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 order. (ibid., page 16)

He then constructed a correlated series by cumulating a random series (see Table 2.1).

> 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, $a+b$, $a+b+c\ldots$. In this new series

Table 2.1 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 +1000

Term	Expectation	Experimental results		
		First term +	First term −	Together
(1)	(2)	(3)	(4)	(5)
1	+1000	+1000	+1000	+1000
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

> 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 2.2 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 2.2 (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

Table 2.2 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 $a, b, c, \ldots, l$

Term	(1) b	(2) c	(3) d	(4) e	(5) f	(6) g	(7) h	(8) k	(9) l	Coefficient 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 2.3 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

product of their respective standard deviations. These coefficients are shown in Table 2.3. 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

Table 2.4 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 +, (*b*) first deviation is −, (*c*) last deviation is +, (*d*) last deviation is −. The average of first or last deviations, respectively, called +1000

Term	Expectation	Experimental results *a* and *b*	Term	Experimental results	
				c and *d*	Together
(1)	(2)	(3)	(4)	(5)	(6)
1	+1000	+1000	10	+1000	+1000
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

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) are shown in column (2) of Table 2.4: they show a continuous decline from +1000 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 2.1 is shown in column (3) of Table 2.4.

As Yule noted, since the correlations and standard deviations in Table 2.2 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 2.4 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 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 2.3]. (ibid., page 19: italics in original)

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

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

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 2.2 and 2.3 are shown as Tables 2.5 and 2.6:

> 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 2.2], the correlations are much higher, and again they are not symmetrical with respect to the two ends of the sample. But the magnitude of 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)[7]

Table 2.5 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) *b*	(2) *c*	(3) *d*	(4) *e*	(5) *f*	(6) *g*	(7) *h*	(8) *k*	(9) *l*	Coefficient 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 2.6 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

Table 2.7 reports analogous simulations to those given in Table 2.4 and, as should be expected, these show an appropriate degree of conformity, with the experimental results being very close to their expectations.

2.11 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 **§6.10**):

> Now this argument has led us to a remarkable result, which at first sight may seem paradoxical: namely, that for the present purpose we

Table 2.7 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 +, (*b*) first deviation is −, (*c*) last deviation is +, (*d*) last deviation is −. The average of first or last deviations, respectively, called +1000

Term	Expectation	Experimental results *a* and *b*	Term	Expectation	Experimental results *c* and *d*
(1)	(2)	(3)	(4)	(5)	(6)
1	+1000	+1000	10	+1000	+1000
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

> 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 important insight, suppose that $x_1, x_2, \ldots, x_T$ is a zero mean series with standard deviation σ_x for which the serial correlations are $r_x(1), r_x(2), \ldots, r_x(k)$, using the notation of **§2.9**. 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 (cf. equation (2.1)),

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

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 **§2.4**), we thus have

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

that is,

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

Generalizing this result gives

$${}_1r_x(k) = \frac{2r_x(k)-r_x(k+1)-r_x(k-1)}{2(1-r_x(1))} = -\frac{1}{2(1-r_x(1))}\Delta^2 r_x(k+1) \quad (2.7)$$

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

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

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 **§2.10** above, these being shown in Figure 2.9 with the underlying random series, A_0, B_0 and C_0, being shown in Figure 2.8. 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 $k = 10$, $\hat{r}_x(1)$ 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 2.8 and plotted in Figure 2.10. 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 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 2.8] 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

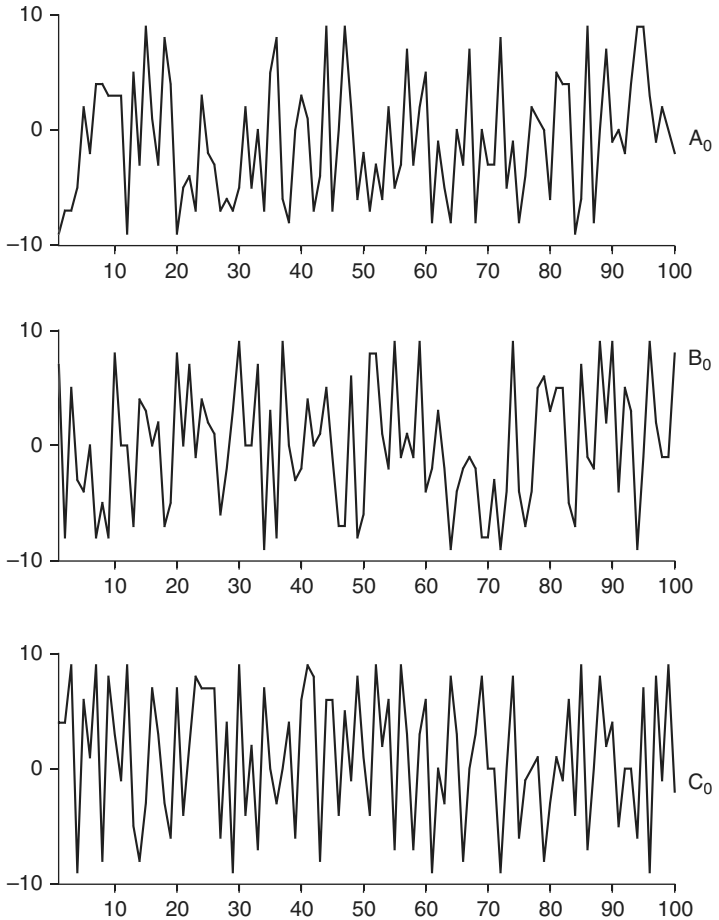


Figure 2.8 Three random series

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 ${}_1r_x(k) = 1 - \alpha k$ since ${}_1r_x(0) = 1$. From (2.7) we then have

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

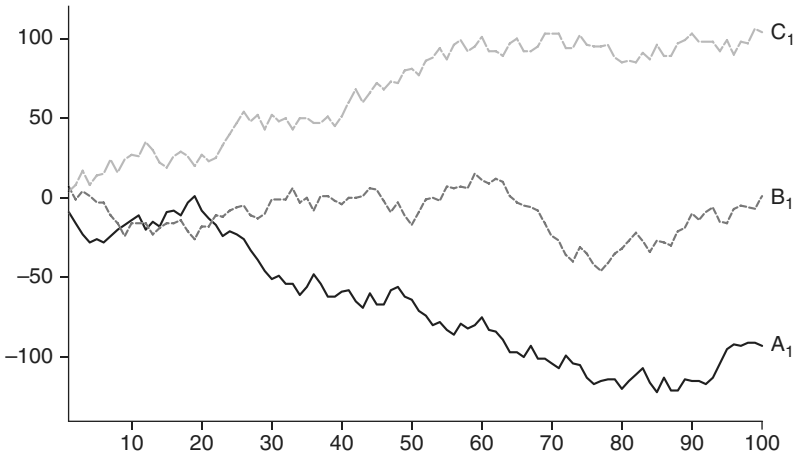


Figure 2.9 Three series with random differences (conjunct series with random differences)

Table 2.8 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

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:

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

This implies that

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

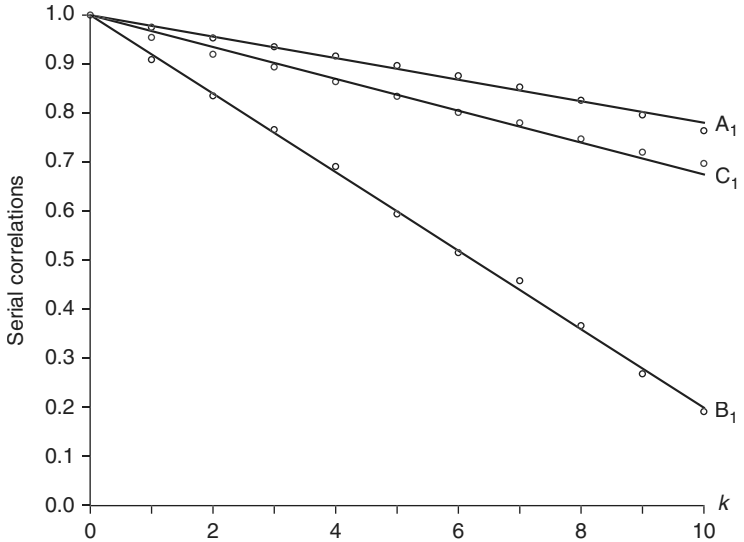


Figure 2.10 Serial correlations up to $r(10)$ for three experimental series (of 100 terms) with random differences

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

$$r_x(k) = 1 - mk^2 + \tfrac{1}{3}\alpha mk(k^2 - 1) \tag{2.8}$$

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

$$\sum_{j=1}^{k} r_x(j) = k - \hat{m}\left\{\tfrac{1}{6}k(k+1)(2k+1) + \tfrac{1}{6}\alpha k(k+1) - \tfrac{1}{12}\alpha k^2(k+1)^2\right\} \tag{2.9}$$

To utilize this result, Yule constructed a series with correlated differences by taking the random series u_t and cumulating 11-period

moving sums, that is, by calculating

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

It is then 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 (2.9) 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 2.11, with their observed serial correlations and the serial correlations calculated from the cubic in k reported in Table 2.9 and plotted in Figure 2.12. 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.

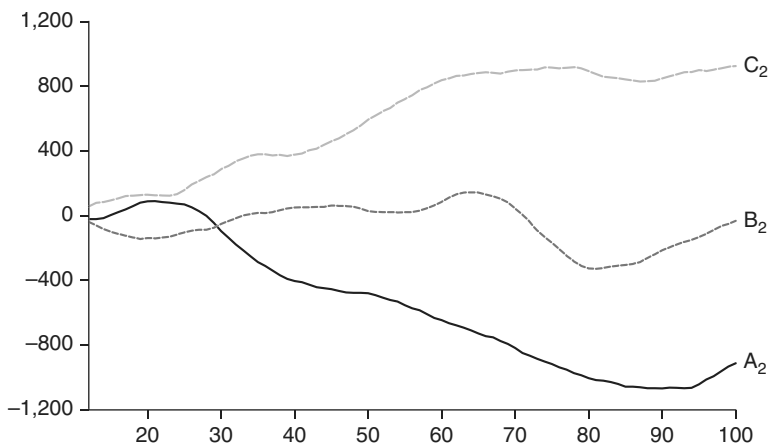


Figure 2.11 Three series with positively correlated differences (conjunct series with conjunct differences)

Table 2.9 Comparison of serial correlations for three series with correlated differences, with fitted cubic series

	Series A_2		Series B_2		Series C_2	
	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

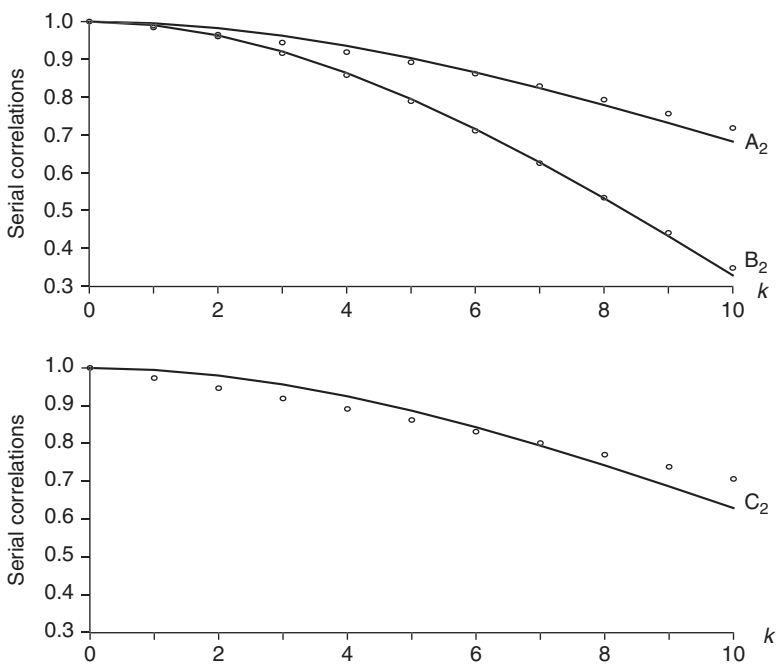


Figure 2.12 Serial correlations up to $r(10)$ for three experimental series (of 100 terms) with positively correlated (conjunct) differences

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, that is,

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

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 $\Delta^2{}_1 r_x(k+1) = 0$, or

$${}_1 r_x(k) = k_1 r_x(1) - (k-1) = 1 - k(1 - {}_1 r_x(1)) = 1 - \alpha k$$

with $\alpha = 1 - {}_1 r_x(1)$. Thus, the $r_x(k)$ are given by (2.8) and the analysis is identical to that above.

2.12 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 **§2.4**).

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

Yule regarded these classifications as simply building blocks: 'clearly in the endless variety presented by facts we may expect to meet with compound series of any type, for example, conjunct series with an oscillatory series superposed' (ibid., page 26). Nevertheless, his focus continued to be on the three types of series analyzed in **§2.11**: (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 2.8 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 2.9 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 2.11 shows A_2, B_2 and C_2, conjunct series with conjunct differences: 'the curves are smoothed out, the secular movements or long waves are conspicuous, but there are no evident oscillations of short duration' (ibid., page 26).

2.13 Having considered the various 'internal' properties of the different types of time series, Yule then turned his attention to his primary aim, that of analyzing 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 **§2.10**. We again recreate Yule's calculations and show in Figures 2.13–2.15 the frequency distributions of the correlations between pairs drawn from the three

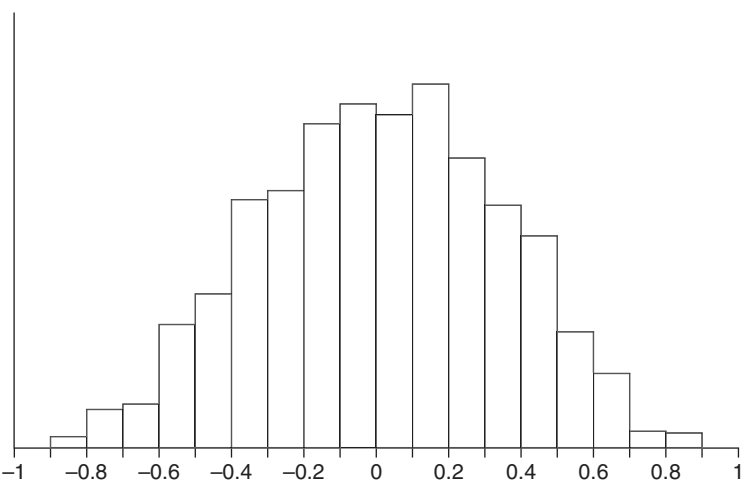


Figure 2.13 Frequency distribution of 600 correlations between samples of 10 observations from random series

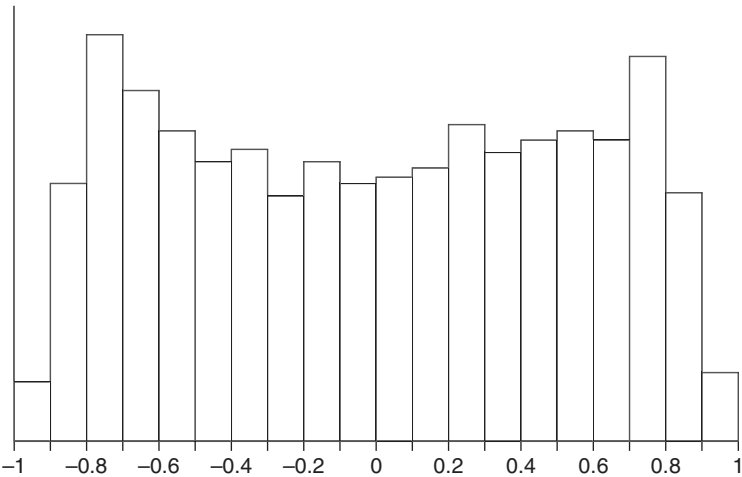


Figure 2.14 Frequency distribution of 600 correlations between samples of 10 observations from conjunct series with random differences

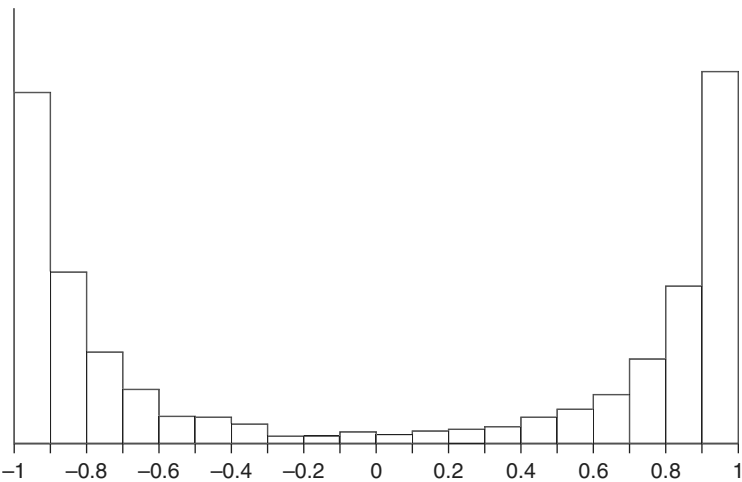


Figure 2.15 Frequency distribution of 600 correlations between samples of 10 observations from conjunct series with conjunct differences

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 2.13) 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 2.14]; correlations between samples of the second type tend definitely to be 'nonsense-correlations' – correlations approaching plus or minus unity in value [Figure 2.15]. 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 2.9). 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 2.11). With this analysis, Yule was making the first tentative steps towards identifying what are now referred to as *stochastic trends* (see **§10.40**).

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 T 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 serial 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 6, 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 **§10.19**.

2.14 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 and concentrate, as did Yule, on the 300-year period from 1545 to 1844, the series being shown in Figure 2.16 with the serial correlations up to $k = 40$ displayed in Figure 2.17.[8]

> 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. 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 42–3; 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 **§2.11**, the serial correlations of the h-step differences $x_{t+h} - x_t$, which we denote as $r^h(k)$ (noting that $r^1(k) \equiv {}_1r(k)$), are given by a generalization of (2.7)

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

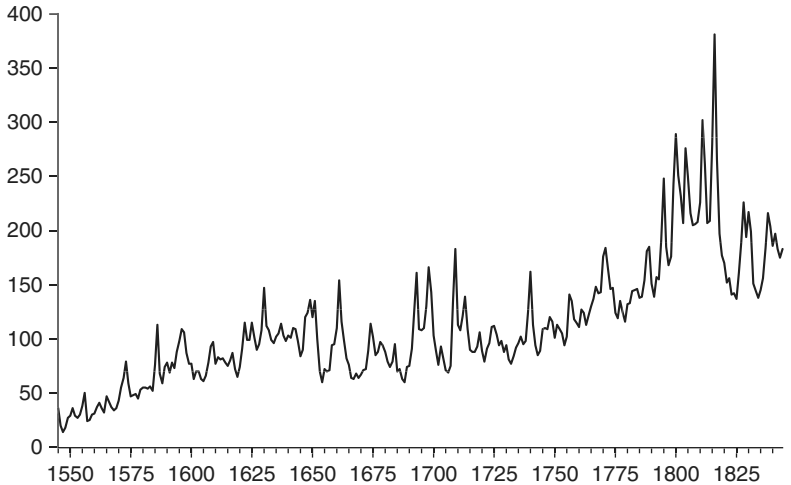


Figure 2.16 Beveridge's index numbers of wheat prices in Western Europe, 1545–1844

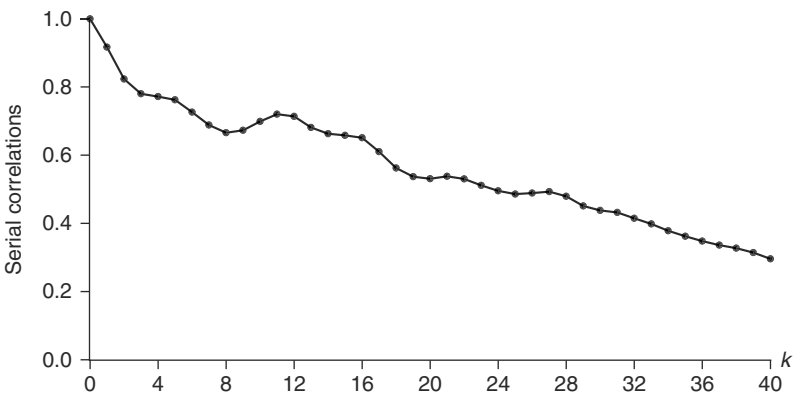


Figure 2.17 Serial correlations up to $r(40)$ for Beveridge's index numbers of wheat prices in Western Europe, 1545–1844

on noting that if $k < h$, $r(k - h) = r(h - k)$. The 'serial difference correlations' for various values of h are plotted in Figure 2.18. 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

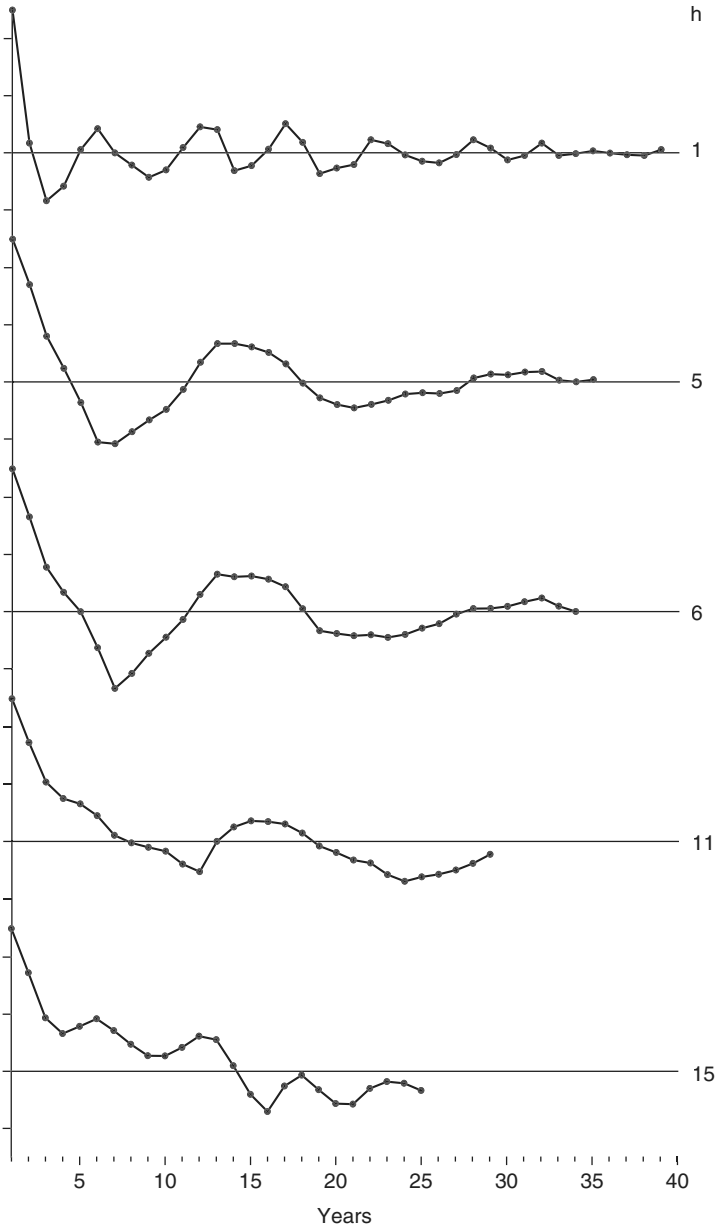


Figure 2.18 Serial difference correlations $r^h(k)$ for the index numbers of wheat prices in Western Europe; intervals for differencing $h = 1, 5, 6, 11$ and 15 years respectively

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 2.16). 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 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 a trough-to-trough period of 12 years.

Yule's final choice was $h = 15$, which produces many minor oscillations in the serial correlations. By 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 2.19. 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).

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

> 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

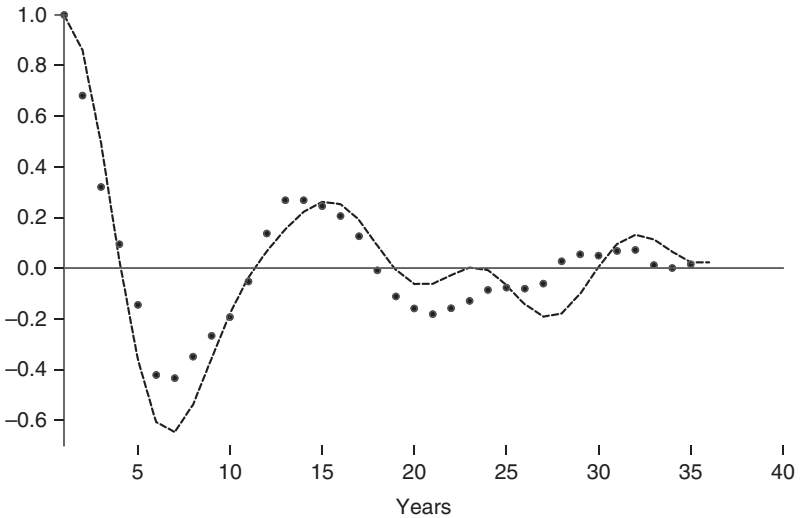


Figure 2.19 Serial difference correlations for $h = 5\,(r^5(k))$ (dots) and a curve constructed from certain of the periodicities given by Beveridge (dashed line)

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)

Superposed fluctuations and disturbances

2.16 At the same time as he was analyzing 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 (Yule, 1927), in which 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' (Yule, 1927, page 267), and an example of this situation is shown in Figure 2.20(a). If the errors were increased in size, as in Figure 2.20(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). Yule referred to this set-up as one of *superposed fluctuations* – 'fluctuations which do not in

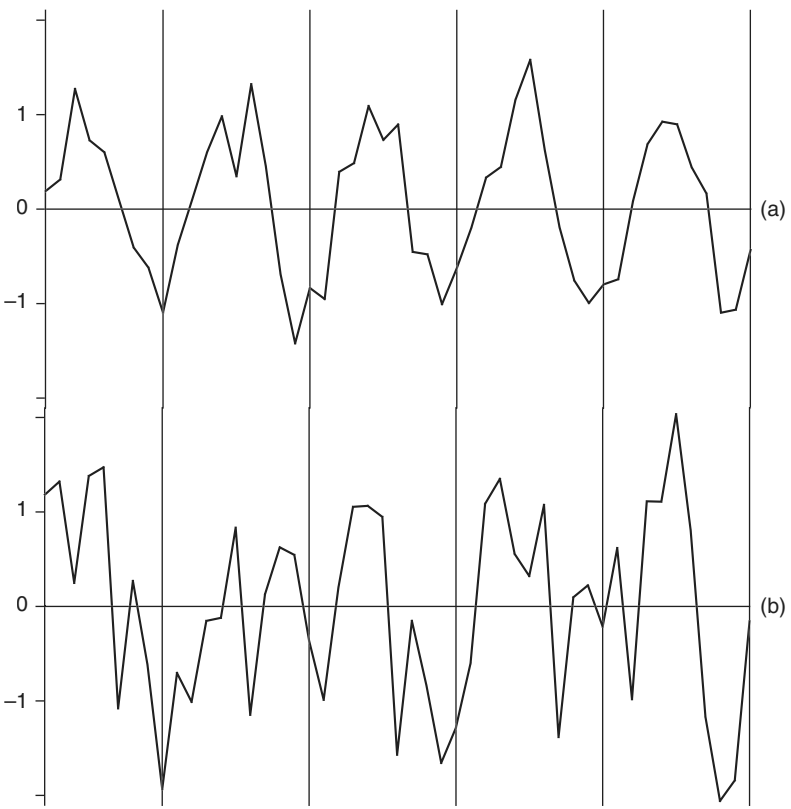


Figure 2.20 Graphs of simple harmonic functions of unit amplitude with superposed random fluctuations: (a) smaller fluctuations; (b) larger fluctuations

any way disturb the steady course of the underlying periodic function' (*ibid.*, page 268).

But Yule did not see this setup 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 2.20]. But by improvement of apparatus and automatic methods of recording, let us say, errors of observation are practically eliminated. (ibid., page 268)

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)

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

$$x_t = \rho \sin 2\pi \frac{t}{n} \tag{2.10}$$

where, once again, ρ is the amplitude of the sine wave and n is the period. The function (2.10) can be written as

$$\Delta^2 x_t = -4 \sin^2 \pi \frac{1}{n} = -\theta x_{t+1} \tag{2.11}$$

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 $\vartheta = 2\pi/n$. The proof of this fundamental result uses standard trigonometric identities. If we define

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

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 $\theta = 4\sin^2(2\pi/n)$.

Equation (2.11) may be written equivalently as

$$x_{t+2} = (2-\theta)x_{t+1} - x_t \tag{2.12}$$

The 'errors' produced by the boys pelting the pendulum with peas leads to the inclusion of an error, ε_{t+2}, in (2.12), which we may then rewrite in the more convenient form

$$x_t = (2-\theta)x_{t-1} - x_{t-2} + \varepsilon_t \tag{2.13}$$

Figure 2.21 shows a graph of x_t constructed from (2.13) by setting $n = 10$, so that $\theta = 4\sin^2 18^\circ = 4 \times 0.3090^2 = 0.382$ and thus

$$x_t = 1.618x_{t-1} - x_{t-2} + \varepsilon_t \tag{2.14}$$

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 2.21 shows the simulation of (2.14) 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'.

2.17 Why does the simulated series in Figure 2.21 present such a smooth appearance? An undisturbed harmonic function may be regarded as the solution of the difference equation

$$\Delta^2 x_t + \theta x_{t+1} = 0 \tag{2.15}$$

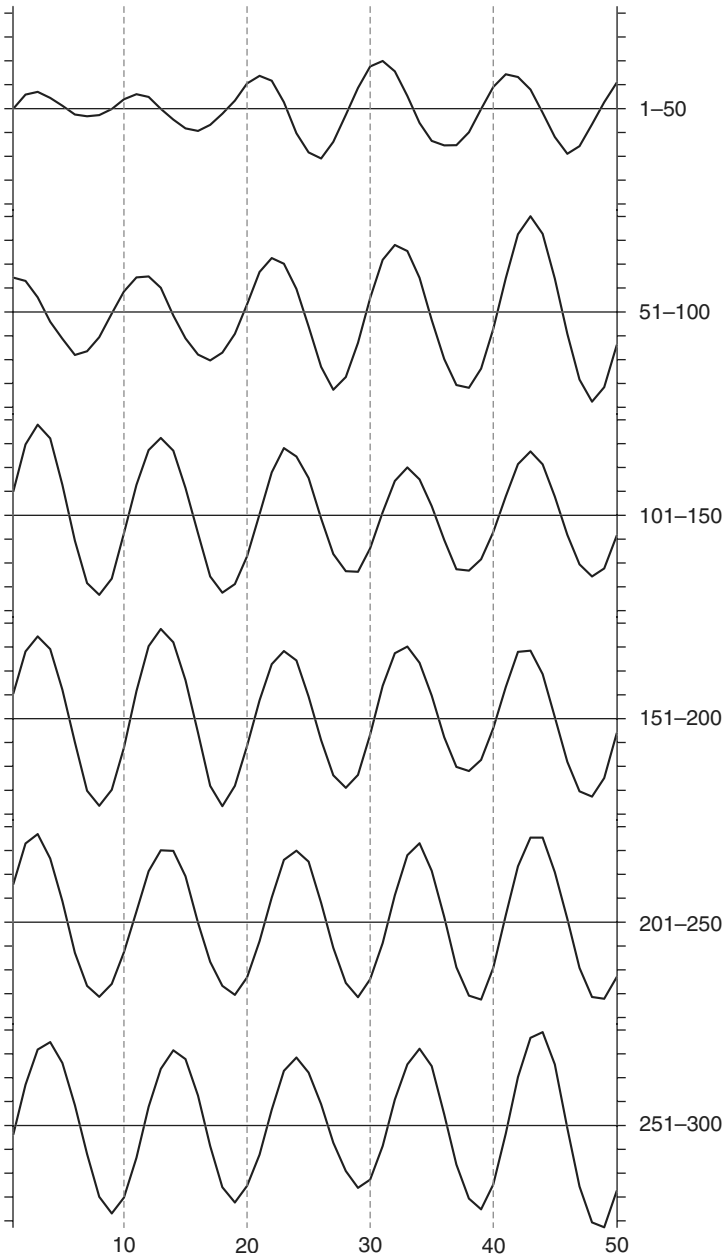


Figure 2.21 Graph of a disturbed harmonic function, equation (2.14)

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

$$\Delta^2 x_t + \theta x_{t+1} = \phi(t) \tag{2.16}$$

where $\phi(t)$ is some 'disturbance function'. Hence we see that (2.15) is the complementary function of the solution to (2.16) and $\phi(t)$ is the particular integral.

The solution to (2.13), 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 $1, k, k^2 - 1, k(k^2 - 1) - k$, ... 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 (2.12), so that the coefficients on the ε's are therefore the terms of a sine function having the same period as the complementary function (2.15) 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, along with its complementary function, particular integral and disturbances, are shown in Table 2.10.

> 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)

Table 2.10 Decomposition of the first 30 terms of the simulated series used in Figure 2.21 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

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 2.21 – '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 [Figure 2.21] 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. (ibid., page 278)

2.18 Yule proposed that models of the type (2.13) should be analyzed by least squares regression. If (2.13) is written

$$x_t = kx_{t-1} - x_{t-2} + \varepsilon_t \tag{2.17}$$

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 2.21, 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

$$\begin{aligned} x_t &= 1.62338x_{t-1} - x_{t-2} \\ \cos\vartheta &= 0.81169; \quad \vartheta = 35^\circ.74; \quad \text{period} = 10.07 \end{aligned}$$

First 150 terms

$$\begin{aligned} x_t &= 1.62897x_{t-1} - x_{t-2} \\ \cos\vartheta &= 0.81448; \quad \vartheta = 35^\circ.46; \quad \text{period} = 10.15 \end{aligned}$$

Second 150 terms

$$\begin{aligned} x_t &= 1.62026x_{t-1} - x_{t-2} \\ \cos\vartheta &= 0.81013; \quad \vartheta = 35^\circ.89; \quad \text{period} = 10.03 \end{aligned}$$

The periods thus found are not far from those obtained in **§2.17** 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).

2.19 Yule then turned his attention to annual sunspot numbers between 1749 and 1924. 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 extended sample period 1700 to 2011 are shown in the top two panels of Figure 2.22. To Yule

> the upper curve in [Figure 2.22] ... suggests quite definitely to my eye that we have to deal with a graph of the type of [Figure 2.21], not of the type of [Figure 2.20], 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 2.21] – 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 (2.17). Estimation over both the extended sample period 1700 to 2011 and the period available to Yule gave the following results (the first

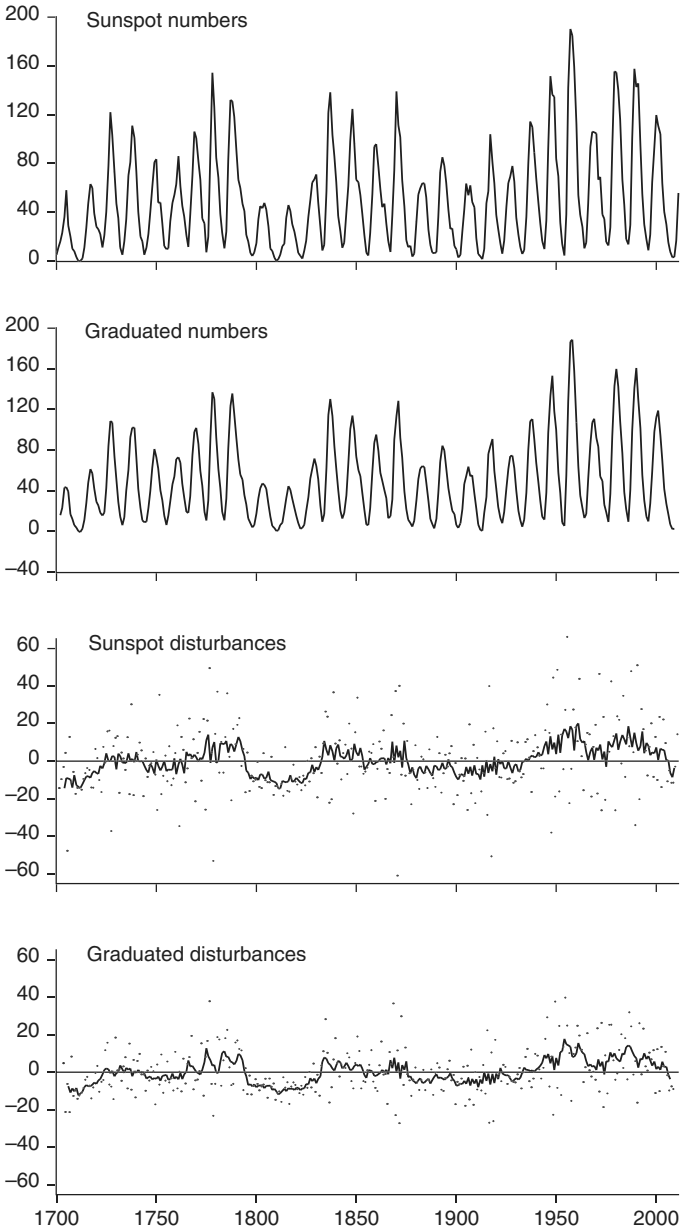


Figure 2.22 Graphs of the sunspots and graduated numbers, and of the disturbances given by equation (2.16): the lines on the disturbance graphs show quinquennial averages

two observations in each period are lost due to the construction of x_{t-1} and x_{t-2}):

Sunspot Numbers, 1749–1924

$$\begin{aligned} \text{s.d. of whole series} &= 34.75 \\ x_t &= 1.61979x_{t-1} - x_{t-2} + 17.06 \\ \cos\vartheta &= 0.80989; \quad \vartheta = 35^\circ.91; \quad \text{period} = 10.02 \\ \text{s.d. of disturbances} &= 17.08 \end{aligned}$$

Sunspot Numbers, 1700–2011

$$\begin{aligned} \text{s.d. of whole series} &= 40.39 \\ x_t &= 1.64723x_{t-1} - x_{t-2} + 17.62 \\ \cos\vartheta &= 0.82361; \quad \vartheta = 34^\circ.55; \quad \text{period} = 10.42 \\ \text{s.d. of disturbances} &= 18.05 \end{aligned}$$

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 0.4 of a year and the variability of the series has also increased somewhat. The disturbances estimated from the extended sample regression are plotted in the third panel of Figure 2.22 with a quinquennial moving average superimposed. Focusing on the sample from 1749 to 1924, Yule described their behavior 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

$$\begin{aligned} \text{s.d. of whole series} &= 34.10 \\ x'_t &= 1.68431x'_{t-1} - x'_{t-2} + 14.23 \\ \cos\vartheta &= 0.84216; \quad \vartheta = 32^\circ.63; \quad \text{period} = 11.03 \\ \text{s.d. of disturbances} &= 11.50 \end{aligned}$$

Graduated Sunspot Numbers, 1704–2009

$$\begin{aligned} \text{s.d. of whole series} &= 39.52 \\ x'_t &= 1.69664x'_{t-1} - x'_{t-2} + 15.19 \\ \cos\vartheta &= 0.84832; \quad \vartheta = 31^\circ.97; \quad \text{period} = 11.26 \\ \text{s.d. of disturbances} &= 12.15 \end{aligned}$$

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, 11.26, turns out

to be very similar to the period obtained using Fourier analysis for the period 1750–1914 by Larmor and Yamaga (1917), 11.21. The calculated disturbances are shown as the bottom panel of Figure 2.22: 'the scatter is greatly reduced (s.d. of disturbances [12.15] against [18.05]), but the general course of affairs is very similar to that shown from the graph for the ungraduated numbers' (ibid., page 276).

Figure 2.23 shows a scatterplot of $x_t + x_{t-2}$ on x_{t-1} and its graduated counterpart and these provide scant indication of any non-linearity in the relationships. The proportion of the variance of x_t that has been

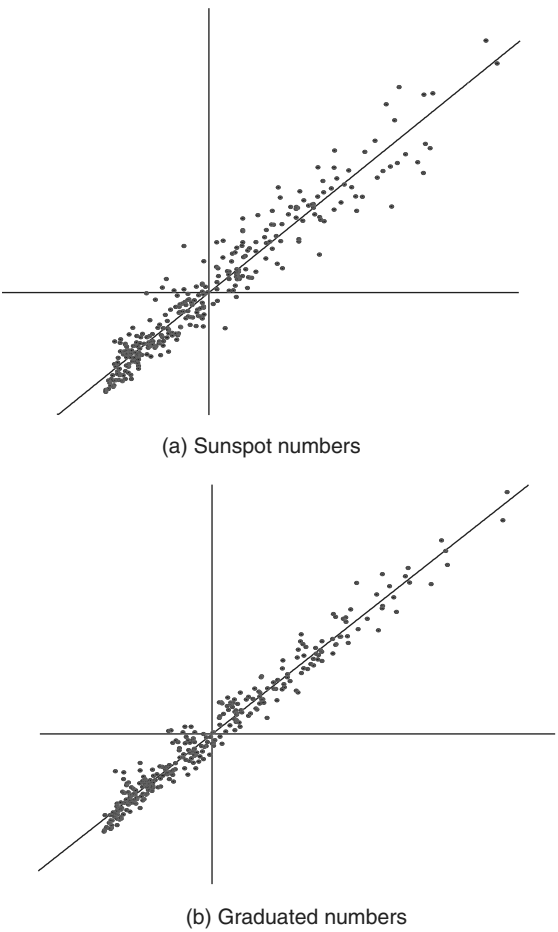


Figure 2.23 Scatterplot of $x_t + x_{t-2}$ (horizontal) on x_{t-1} (vertical)

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.

2.20 Yule then extended the model (2.10) 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, (2.11) extends naturally to

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

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. (2.12))

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

and, if a disturbance is again appended, we can write (cf. (2.13))

$$x_t = k_1(x_{t-1} + x_{t-3}) - k_2 x_{t-2} - x_{t-4} + \varepsilon_t \tag{2.18}$$

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 [2.18] 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 [2.18] 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 (2.18) for the various series and samples were obtained as follows:

Sunspot Numbers 1749–1924

$x_t = 1.15975(x_{t-1} + x_{t-3}) - 1.016367x_{t-2} - x_{t-4} + 31.21$
$\theta_1 = 2.56899 \quad \cos\vartheta_1 = -0.284495$
$\vartheta_1 = 106°53 \text{ or } 253°47 \quad \text{period} = 1.42 \text{ or } 3.38 \text{ years}$
$\theta_2 = 0.27126 \quad \cos\vartheta_1 = -0.86437$
$\vartheta_1 = 30°19 \quad \text{period} = 11.91 \text{ years}$
s.d. of disturbances = 21.97 years

Graduated Sunspot Numbers 1753–1920

$x'_t = 1.67128(x'_{t-1} + x'_{t-3}) - 1.86233x'_{t-2} - x'_{t-4} + 23.48$
$\theta_1 = 2.07867 \quad \cos\vartheta_1 = -0.03933$
$\vartheta_1 = 92°25 \text{ or } 267°75 \quad \text{period} = 1.34 \text{ or } 3.90 \text{ years}$
$\theta_2 = 0.25005 \quad \cos\vartheta_1 = 0.87498$
$\vartheta_1 = 28°96 \quad \text{period} = 12.43 \text{ years}$
s.d. of disturbances = 17.47 years

Sunspot Numbers 1700–2011

$x_t = 1.17239(x_{t-1} + x_{t-3}) - 1.02072x_{t-2} - x_{t-4} + 33.82$
$\theta_1 = 2.56398 \quad \cos\vartheta_1 = -0.28199$
$\vartheta_1 = 106°38 \text{ or } 253°62 \quad \text{period} = 1.42 \text{ or } 3.38 \text{ years}$
$\theta_2 = 0.26363 \quad \cos\vartheta_1 = 0.868185$
$\vartheta_1 = 29°75 \quad \text{period} = 12.10 \text{ years}$
s.d. of disturbances = 23.99 years

Graduated Sunspot Numbers 1704–2009

$x'_t = 1.70403(x'_{t-1} + x'_{t-3}) - 1.92042x'_{t-2} - x'_{t-4} + 25.73$
$\theta_1 = 2.09097 \quad \cos\vartheta_1 = -0.04549$
$\vartheta_1 = 92°61 \text{ or } 267°39 \quad \text{period} = 1.35 \text{ or } 3.89 \text{ years}$
$\theta_2 = 0.20500 \quad \cos\vartheta_1 = 0.89750$
$\vartheta_1 = 26°17 \quad \text{period} = 13.75 \text{ years}$

s.d. of disturbances = 19.02 years

Since the values of the θs 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 [2.17] with only two terms on the right. (ibid., page 280)

2.21 As a second approach, Yule considered the 'ordinary regression equation'

$$x_t = b_1 x_{t-1} - b_2 x_{t-2} \tag{2.19}$$

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 (2.19) is of the form

$$x_t = e^{\lambda t}(A\cos\vartheta t + B\sin\vartheta t) \tag{2.20}$$

For a real physical phenomenon, λ would be expected to be either negative ($b_2 < 1$), so that the solution (2.20) would be a damped harmonic vibration, or zero ($b_2 = 1$), in which case the solution would be simple harmonic vibration.

The regression (2.19), with a disturbance term ε_t implicitly appended, was first fitted to the simulated series of Figure 2.21, producing the following results.

Complete sample of 300 terms

$x_t = 1.6220x_{t-1} - 0.9983x_{t-2}$
Roots: $0.8110 \pm 0.5836i$
$\tan\vartheta = 0.71959 \quad \vartheta = 35°74 \quad \text{Period} = 10.07 \quad \lambda = -0.0009$

First 150 terms

$x_t = 1.6253x_{t-1} - 0.9955x_{t-2}$
Roots: $0.8126 \pm 0.5789i$
$\tan\vartheta = 0.712334 \quad \vartheta = 35°46 \quad \text{Period} = 10.15 \quad \lambda = -0.0023$

Second 150 terms

$x_t = 1.601x_{t-1} - 0.9999x_{t-2}$
Roots: $0.8101 \pm 0.5863i$
$\tan\vartheta = 0.723753 \quad \vartheta = 35°89 \quad \text{Period} = 10.03 \quad \lambda = -0.0001$

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: $0.66798 \pm 0.45128i$
$\tan\vartheta = 0.67559 \quad \vartheta = 34°04 \quad \text{Period} = 10.58 \quad \lambda = -0.21550$
s.d. of disturbances = 15.56

Graduated Sunspot Numbers 1753–1920

$x'_t = 1.51975x'_{t-1} - 0.80457x'_{t-2} + 12.84$
Roots: $0.75987 \pm 0.47661i$
$\tan\vartheta = 0.62723 \quad \vartheta = 32°10 \quad \text{Period} = 11.22 \quad \lambda = -0.10872$
s.d. of disturbances = 10.96

Sunspot Numbers 1700–2011

$x_t = 1.39328x_{t-1} - 0.69239x_{t-2} + 14.97$
Roots: $0.69664 \pm 0.45600i$
$\tan\vartheta = 0.65457 \quad \vartheta = 33°21 \quad \text{Period} = 10.84 \quad \lambda = -0.18380$
s.d. of disturbances = 16.65

Graduated Sunspot Numbers 1704–2009

$x'_t = 1.55555x'_{t-1} - 0.83341x'_{t-2} + 13.90$
Roots: $0.77777 \pm 0.47799i$
$\tan\vartheta = 0.61456 \quad \vartheta = 31°57 \quad \text{Period} = 11.40 \quad \lambda = -0.09111$
s.d. of disturbances = 11.65

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

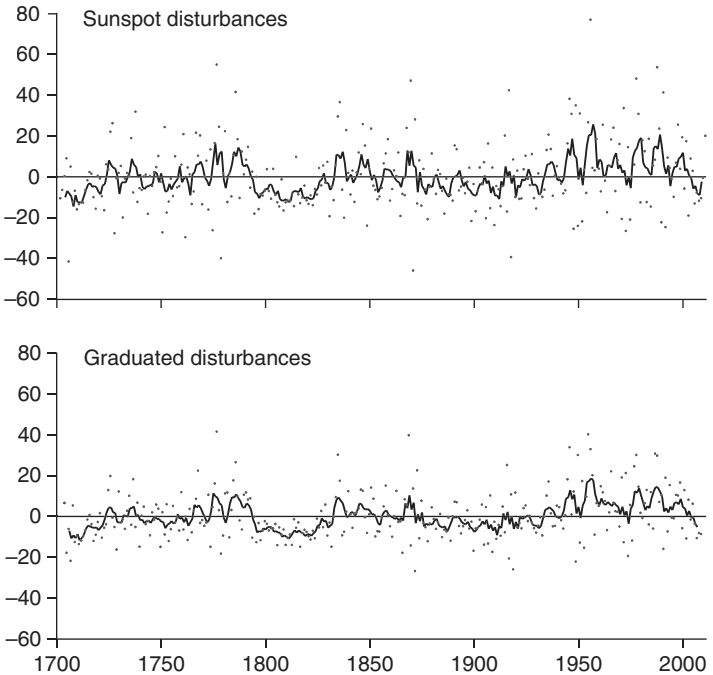


Figure 2.24 Graphs of the disturbances given by equation (2.19): the lines on the graphs show quinquennial averages

and Yamaga (1917). For the extended samples, both periods are increased to 10.84 and 11.40, respectively (as against 10.42 and 11.26 from the harmonic formula).

2.22 The two disturbance series for the extended sample period are shown in Figure 2.24. Yule analysed these in the context of the alternating quiet and disturbed periods of approximately 42 years alluded to in **§2.19**. Table 2.11 extends Yule's periods both backwards and forwards in time to cover the extended sample now available. As discussed in **§2.19**, 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.

In the extended period covered in Table 2.11, this alternating pattern continues to be found from the early 1700s up until 1960, but the final

Table 2.11 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	−2.31	12.06	−2.64	8.69
1751–1792 (Y)	2.49	18.92	2.42	11.84
1793–1834 (Y)	−7.21	7.41	−6.29	5.86
1835–1876 (Y)	2.56	18.00	2.20	12.65
1877–1918 (Y)	−4.33	13.95	−3.63	8.66
1919–1960	4.20	19.21	3.01	14.25
1961–2002	6.78	19.66	6.25	12.71

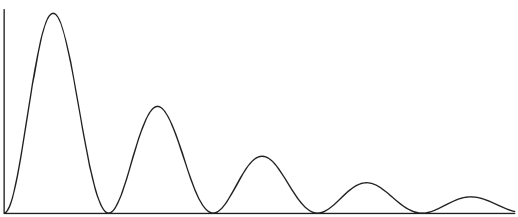


Figure 2.25 Graph of the square of a damped harmonic vibration, (2.21)

period 'bucks the trend', having a positive mean accompanied by a high standard deviation, rather than a negative mean and a low standard deviation.[9,10]

2.23 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,

$$x_t = Ae^{-at}(1 - \cos \vartheta t) \tag{2.21}$$

is shown in Figure 2.25. Figure 2.26 shows a train of such functions, each with different amplitude A and each starting when the one before it reaches its first minimum. This looks much more like the graph of the sunspot numbers. However, if (2.21) is regarded as the solution of a difference equation, then it is seen to imply that there must be a real

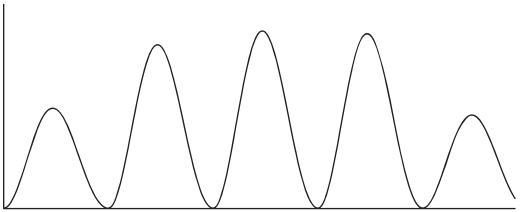


Figure 2.26 Graph of a series of superposed functions of the form of Figure 2.25, each one starting when the one before reaches its first minimum

root, thus giving rise to at least a third-order difference equation. The difference equation (2.19) 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.

Yule thus considered the serial correlations of the sunspot numbers up to lag five, i.e., he computed $r(1), \ldots, r(5)$ using the notation of **§2.9**. He then computed corresponding partial correlations, which we may denote as $r(1)$, $r(2 \cdot 1)$, $r(3 \cdot 12)$, $r(4 \cdot 123)$ and $r(5 \cdot 1234)$, where $r(k \cdot 12 \ldots (k-1))$ 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 2.12. The third column, labelled $1 - r^2$, 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. $1 - R^2_{1\cdots k}$ 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.

Table 2.12 Serial correlations of the sunspot numbers and the deduced partial correlations for the extended sample period 1700–2011. 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, that is, $r(2 \cdot 1)$, and so on

Serial correlations		Partial correlations		$1 - r^2$	Continued product of $1 - r^2$
Sunspot Numbers					
1	0.822	1	0.822	0.324	0.324
2	0.455	2.1	−0.680	0.540	0.175
3	0.045	3.12	−0.140	0.980	0.171
4	−0.270	4.123	0.049	0.998	0.171
5	−0.421	5.1234	0.018	0.999	0.171
Graduated Sunspot Numbers					
1	0.846	1	0.846	0.284	0.284
2	0.483	2.1	−0.816	0.334	0.095
3	0.054	3.12	0.034	0.999	0.095
4	−0.277	4.123	0.297	0.912	0.086
5	−0.435	5.1234	0.174	0.970	0.084

> … 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 [2.21], it may be noted that $r(3 \cdot 12)$ 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)

2.24 After conducting some experiments that used periodogram analysis on 'disturbed' harmonic functions, which need not concern us here, Yule concluded his paper with the following observation:

> many series … may be subject to 'disturbance' in the sense in which the term is here used, and … 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)

Yule's legacy to time series analysis

2.25 The three papers discussed in this chapter produced major advances in time series analysis so that, by the mid-1930s, many of the foundations of the subject had been lain. Along with Persons, Yule introduced the unobserved component formulation of a time series and emphasized the difficulties that may be encountered when correlating differenced series having cyclical components. Yule also considered the implications for correlating time series when they individually had internal correlation structures, which he termed *serial correlations* (he also introduced the concept of *partial serial correlations*). Two of these structures – which Yule regarded as building blocks – correspond to integrated processes of orders one and two, that is, series obtained by accumulation, either once or twice, of basic series, typically random but not necessarily so. The implications of taking sums and differences of a time series was considered in great detail by Eugene Slutzky (1927, 1937), and later Holbrook Working (1934) analyzed further the implications of summing a time series.

Yule also provided an alternative to the then conventional model of harmonic motion disturbed by superposed fluctuations, typically measurement errors. This was a model in which the evolution of a time series was dependent upon both previous values and on disturbances. This 'ordinary regression' was analyzed further by Gilbert Walker (1931) using a difference equation framework and led to Herman Wold (1938), in his treatise introducing the concept and theoretical structure of stationary time series, terming such a model an *autoregression*. Wold's contribution was essential because it was able to provide a probabilistic basis for the models developed essentially intuitively by Yule, Walker and Working and consequently paved the way for the explosion of research in theoretical time series analysis that was to come over the subsequent two decades.

3
Kendall: Generalizations and Extensions of Stationary Autoregressive Models

Sir Maurice Kendall

3.1 After being introduced by Yule and Walker and having its theoretical foundations established by Wold (recall **§2.25**), the autoregressive model was further developed in a trio of papers written during the Second World War by Kendall (1943a, 1944, 1945a). Sir Maurice Kendall (he was knighted in 1974 for his services to the theory of statistics) was born in Kettering, in Northamptonshire, on September 6, 1907 and grew up in Derby. After graduating as a mathematics wrangler (i.e., he received first-class honours) from St John's, Cambridge, he joined the Ministry of Agriculture in 1930 before moving to the Chamber of Shipping as Assistant General Manager in 1940, combining this with nightly war-time duties as an air-raid warden.

During the early 1930s Kendall became interested in using statistics to analyze agricultural problems and returned to St John's in the summer of 1935 to consult the statistics collection in the college's library. This produced a chance encounter with Yule, who kept the key to the library, and subsequently they became close friends until Yule's death in 1951, with Yule becoming godfather to Kendall's second son. They also became professionally close with Kendall becoming co-author of Yule's *Introduction to the Theory of Statistics* for the 11th edition and for three subsequent editions (the last being Yule and Kendall, 1950), the textbook becoming commonly known as 'Yule and Kendall'.

By the end of the 1930s Kendall had published on random number generation, on non-parametric statistics (Kendall's tau), and had become part of a group of statisticians aiming to produce a reference work summarizing recent developments in statistical research. The project was cancelled at the onset of war, but Kendall somehow managed to continue

the project on his own, producing Volume 1 of *The Advanced Theory of Statistics* in 1943 and Volume 2 in 1946 (Kendall, 1943b, 1946). He was later joined by Alan Stuart for a three-volume revision published between 1958 and 1966 and it is now in its sixth edition (Stuart and Ord, 1994), being titled eponymously as *Kendall's Advanced Theory of Statistics*.

In 1949 Kendall accepted a chair in statistics at the LSE, which he left in 1961 to become managing director and then chairman of a computer consulting company, later known as Scientific Control Systems (SciCon). On retiring from this position in 1972 he became director of the World Fertility Survey, a project sponsored jointly by the United Nations and the International Statistical Institute (of which he later became honorary president), which aimed to study fertility in both developed and developing nations. He continued with this work until 1980, when illness forced him to retire, and he died on 29 March 1983.

As well as his knighthood, Kendall was also awarded the Peace Medal of the United Nations in 1980 in recognition for his work on the World Fertility Survey. He also held many other honorary positions and received several awards, including the Presidency of the RSS, 1960–2, and the Guy Medal in Gold in 1968. For further biographical details and discussion of Kendall's many and varied contributions to statistics see the obituaries by Ord (1984) and Stuart (1984), the biographical sketch of Barnard (1997) and the centenary tribute from David and Fuller (2007).

Oscillations induced by taking moving averages

3.2 In his first paper on time series, Kendall (1941) considered a problem initially discussed by Yule but later studied in depth by Slutzky. Suppose an observed series y_t has a decomposition into trend, τ_t, oscillatory, γ_t, and random, ε_t, components of the form

$$y_t = \tau_t + \gamma_t + \varepsilon_t$$

and a moving average, denoted

$$Wy_t = [w_{-m}, \ldots, w_0, \ldots, w_m]y_t = \sum_{j=-m}^{m} w_j y_{t+j} \qquad \sum_{j=-m}^{m} w_j = 1$$

is applied, so that

$$Wy_t = W\tau_t + W\gamma_t + W\varepsilon_t$$

As Kendall (1941) pointed out, the ideal moving average would be one that reproduces the trend exactly, that is, one for which $W\tau_t = \tau_t$, in which case the 'detrended' series is

$$y_t - Wy_t = \gamma_t + \varepsilon_t - W\gamma_t - W\varepsilon_t \tag{3.1}$$

> (T)he point to be emphasized is that the existence of the terms $W\gamma_t$ and $W\varepsilon_t$ in [3.1] 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)

3.3 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) and those later provided by Dodd (1939, 1941a, 1941b) on the effect of summing random series may be used. Thus suppose that, on setting $w_j = 1/(2m+1)$, $-m \le j \le m$,

$$\varepsilon_t^{[2]} = W\varepsilon_t = \frac{1}{2m+1}\sum_{j=-m}^{m}\varepsilon_{t+j} = \frac{1}{n}\sum_{j=-m}^{m}\varepsilon_{t+j}$$

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 with 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 signs. 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 'up-cross', so that the mean distance between up-crosses is $2\pi/\theta$: this, of course, is also the mean distance between 'down-crosses' – changes of sign from positive to negative. For

$$\Delta x_t = x_{t+1} - x_t = c_1\varepsilon_t + c_2\varepsilon_{t+1} + \cdots + c_n\varepsilon_{t+n-1} + c_{n+1}\varepsilon_{t+n}$$

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

$$\cos\theta' = \frac{\sum c_j c_{j+1}}{\sum c_j^2}$$

$$c_1 = -a_1, \quad c_{n+1} = a_n, \quad c_j = a_{j-1} - a_j, \quad j = 2, 3, \ldots, n-1$$

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 can be relaxed with the results seeming to be applicable to various other distributional assumptions.

3.4 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, that is, the individual weights should be as small and as equal as possible.

3.5 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

$$\sum_{j=1}^{n} \sin(\alpha + j\lambda) = \frac{\sin \frac{1}{2}n\lambda}{\sin \frac{1}{2}\lambda} \sin\left(\alpha + \tfrac{1}{2}(n-1)\lambda\right)$$

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 $\frac{1}{2}n\lambda = 0 \pmod{\pi}$, that is, 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)

Oscillatory autoregressions

3.6 Kendall's interest in oscillatory autoregressions was kindled by his work at the Ministry of Agriculture, where he was examining a wide variety of agricultural time series which he had detrended by using a nine-term moving average. Given the above analysis, Kendall was able to argue that, although the mean period of the oscillations induced by taking such a moving average to eliminate the trend 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.[1]

This enabled him to concentrate on analyzing the detrended observations as an 'oscillatory' time series generated by the second-order autoregressive process studied by Yule (1927)

$$x_t + ax_{t-1} + bx_{t-2} = \varepsilon_t \tag{3.2}$$

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 (3.2) is then

$$p^t(A\cos\theta t + B\sin\theta t) \tag{3.3}$$

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. Assuming $b > 0$, $0 < p < 1$, and $4b > a^2$, the complementary function (3.3) represents a damped harmonic with a fundamental period of $2\pi/\theta$. If ξ_t is a particular value of (3.3) 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 (3.3) 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[2]

$$\rho_k = \frac{p^k \sin(k\theta + \psi)}{\sin \psi} \qquad \tan \psi = \frac{1+p^2}{1-p^2} \tan \theta$$

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 (3.2). Using results in Wold (1938), the relationship between the autocorrelations and the coefficients of the autoregression may be shown to be

$$\rho_1 = -\frac{a}{(1+b)} \qquad \rho_2 = \frac{a^2 - b(1+b)}{1+b}$$

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 (as the plot of the autocorrelations against k had become known), 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 up-cross to up-cross (that is, 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 up-crosses was

$$\text{m.d. (up-crosses)} = \frac{2\pi}{\cos^{-1}\rho_1} = \frac{2\pi}{\cos^{-1}(-a/(1+b))}$$

while the mean distance between peaks was

$$\text{m.d. (peaks)} = \frac{2\pi}{\cos^{-1}\tau_1}, \qquad \tau_1 = \frac{-1 + 2\rho_1 - \rho_2}{2(1-\rho_1)} = \frac{b^2 - (1+a)^2}{2(1+a+b)}$$

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

$$\frac{\sigma_\varepsilon^2}{\sigma_x^2} = \frac{1-b}{1+b}((1+b)^2 - a^2) \tag{3.4}$$

a result that will be found to be useful in **§3.13**.

3.7 Kendall illustrated these properties of an oscillatory autoregressive process by generating 480 observations from the model (3.2) with $a = -1.1$ and $b = 0.5$, that is,

$$x_t = 1.1x_{t-1} - 0.5x_{t-2} + \varepsilon_t \tag{3.5}$$

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 3.1. '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 up-cross to up-cross intervals are shown in Table 3.1. 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 up-crosses 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 (see Mills, 2011a, chapter 2, for discussion of the contemporary development of the periodogram: the rudiments of periodogram analysis are sketched out in **§§5.4–5.5**). The periodogram calculated by Kendall is shown in Figure 3.2, 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)

Kendall compared 'the ambiguous and confusing picture presented by the periodogram' with the correlogram of Series I, shown in Figure 3.3.

> The damped oscillatory effect is now clearly evident, and the only doubt that would occur is that after a point the oscillations do not

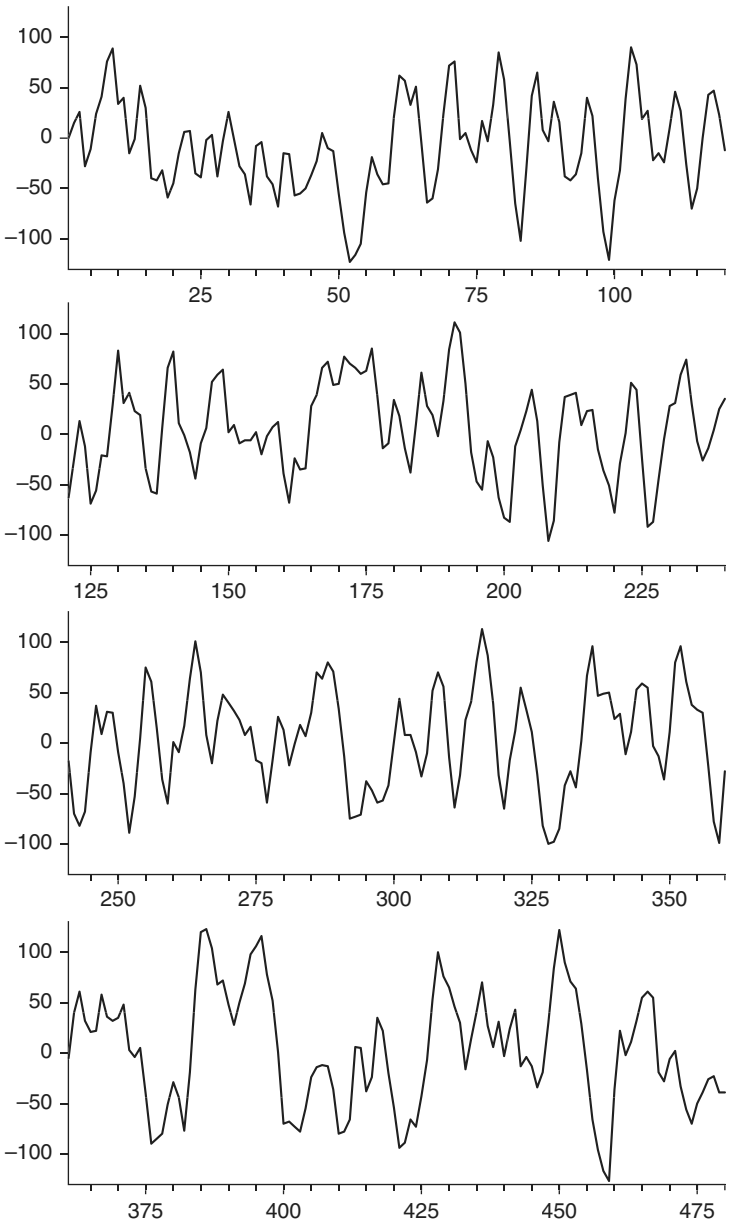


Figure 3.1 480 observations of Kendall's Series I

Table 3.1 Distribution of intervals from peak-to-peak and up-cross to up-cross for Series I

Interval (units)	Peak-to-peak frequency	Up-cross to up-cross 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

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 up-crosses (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 (3.5).

3.8 Kendall (1943a, 1944) applied these ideas to several agricultural series for England and Wales. Figure 3.4 shows the annual observations from 1871 to 1934/1935 for wheat prices and sheep population taken

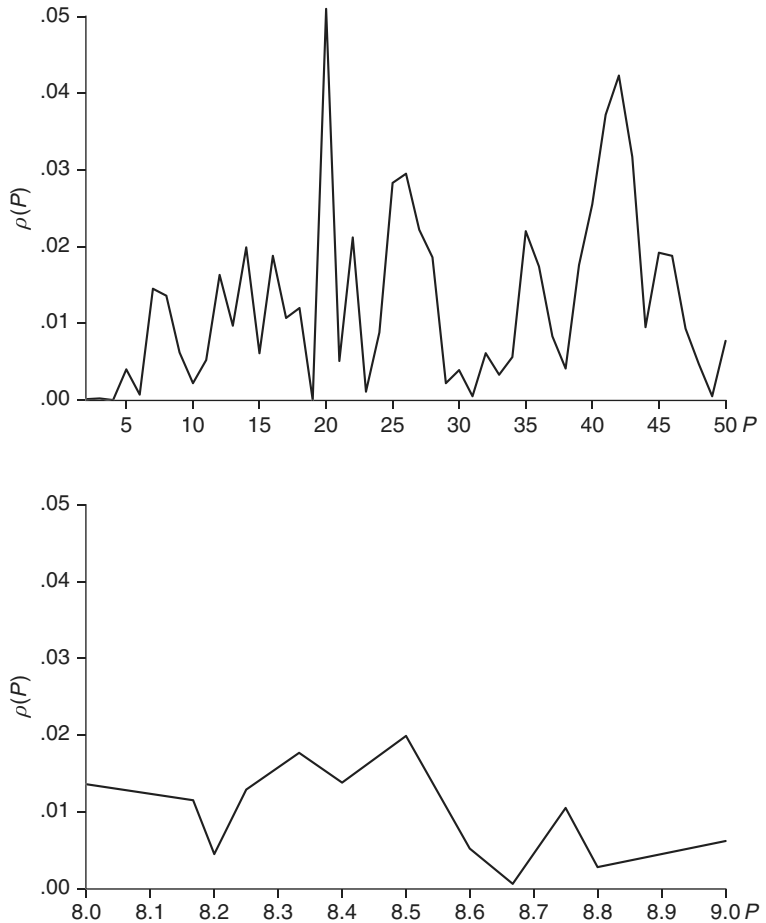


Figure 3.2 Periodogram of Series I: $\rho(P)$ is the value of the periodogram for period P

from Kendall (1943a, Table 1), while Figure 3.5 shows their correlograms. Kendall concluded that both show 'real systematic fluctuations' and he used, for the first time, concepts of statistical significance to support 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

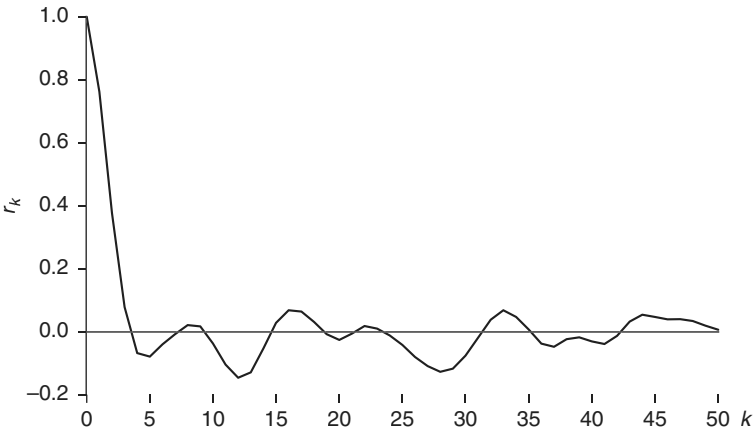


Figure 3.3 Correlogram of Series I

significance of a correlogram. For any *given* serial correlation the theory of large samples may be used to 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 $r_2 \ldots 35$ 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, 1943a, pages 102–3; italics in original)

Focusing first on the sheep population data, Kendall considered the partial correlations of the series, the first six being shown in Table 3.2, along with the continued product of $1 - r^2$ (as in Table 2.12), 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'

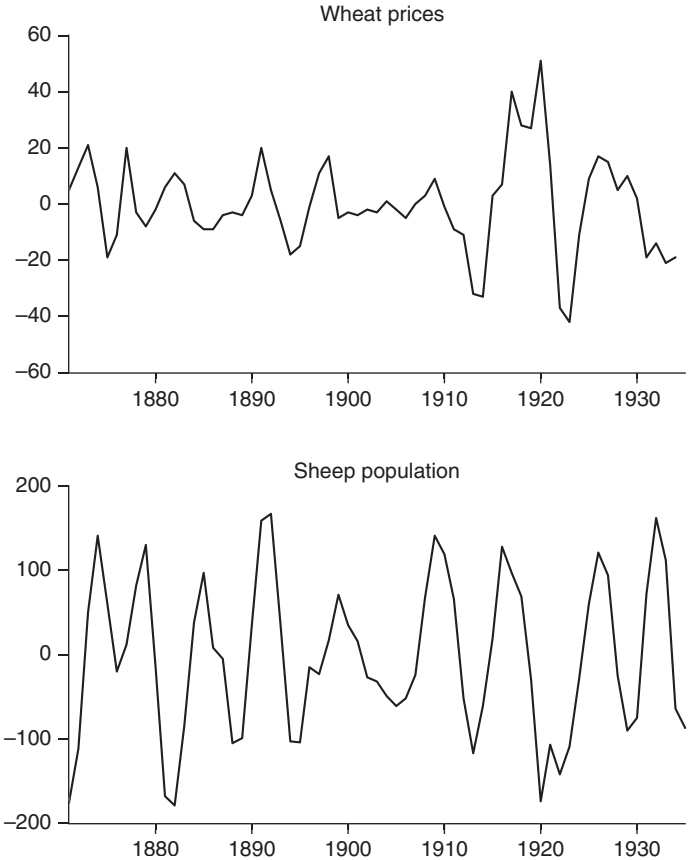


Figure 3.4 Detrended wheat prices and sheep population for England and Wales: 1871–1934/5

(*ibid.*, page 104). A similar pattern of partial correlations is found for the wheat price series, also shown in Table 3.2.

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

$$\tan\theta = \sqrt{\left(\frac{4b}{a^2} - 1\right)} = 1.341, \qquad \theta = 53.3^\circ$$

Figure 3.5 Correlograms of wheat prices and sheep population

Table 3.2 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

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 **§3.6**), 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 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.

3.9 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 **§2.16**). 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

$$c = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2} \tag{3.6}$$

To illustrate this effect, Kendall constructed an autoregressive series of 65 terms as

$$u_t = 1.5u_{t-1} - 0.9u_{t-2} + \varepsilon_t \tag{3.7}$$

where the ε_t are rectangular random variables in the range $-49.5(1)\ldots +49.5$. 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 $-199.5(1)\ldots +199.5$, 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 3.6. Kendall showed that, for a series of infinite length, 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

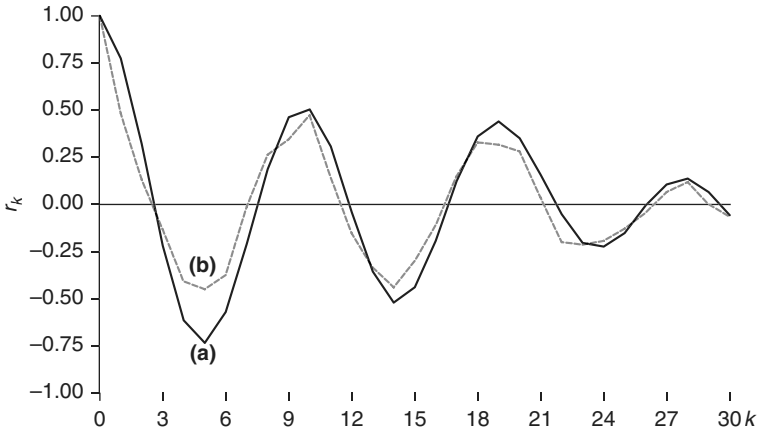


Figure 3.6 Correlograms of two artificial series with (a) a slight superposed variation, and (b) a large superposed variation

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 **§3.6**, the coefficients a and b can be written in terms of the serial correlations r_1 and r_2 as

$$-a = \frac{r_1(1 - r_2)}{1 - r_1^2} \qquad -b = \frac{r_2 - r_1^2}{1 - r_1^2} \tag{3.8}$$

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 a and b, the presence of superposed variation will reduce the autocorrelations by a factor c, leading to the estimates

$$-a' = \frac{cr_1(1 - cr_2)}{1 - c^2r_1^2} \qquad -b' = \frac{cr_2 - c^2r_1^2}{1 - c^2r_1^2}$$

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 (that is., 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

$$b > \tfrac{1}{6}(-1 + \sqrt{(12a^2 + 1)})$$

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 [3.8] 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 [**§2.19**]. 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 [3.8] 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 3.6 has $r'_1 = 0.486$ and $r'_2 = 0.133$, thus giving, according to (3.8),

$$-a' = 0.552 \quad b' = 0.135 \qquad \cos\theta' = \frac{-a'}{2\sqrt{b'}} = 0.751 \qquad \theta' = 41.3°$$

which corresponds to a period of about 8.7 years. In contrast, since it is known that $a = -1.5$, $b = 0.9$ and $\theta = 37.7°$, 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 are 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 difficult 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 (3.7) in which a superposed variation of about 10% of the total ($c = 0.9$) shortened the period by around one year.

3.10 Kendall employed these results to investigate the oscillatory properties of the wheat price series of Figure 3.4. The correlogram shown in Figure 3.5 has up-crosses 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 (3.8) with $r'_1 = 0.5773$ and $r'_2 = 0.0246$ gives

$$a' = -0.8446 \qquad b' = 0.4630$$

so that

$$\cos\theta' = 0.6206 \qquad \theta' = 51.63^\circ$$

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 (cf. **§2.4**). By taking up to tenth differences of the original series (that is, 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%, rather than 10%, 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

3.11 After mentioning extensions to higher order and non-linear 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 (1943a) 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 analyze such a situation, Kendall used cross-correlations, which were first introduced over forty years earlier by Hooker (1901) and which may be denoted $r_{xy}(k)$. Suppose there are two series of the form (3.2) 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

$$\xi_j = A_1 p_1^j \sin \theta_1 j \qquad \chi_j = A_2 p_2^j \sin \theta_2 j$$

then

$$r_{xy}(k) \propto p_2^k \sum_{j=0}^{\infty} p_1^j p_2^j \sin \theta_1 j \sin \theta_2 (j+k) \tag{3.9}$$

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 $r_{xy}(k) \neq r_{xy}(-k)$.

3.12 Figure 3.7 shows the sheep and cow population series, while the cross-correlation function $r_{cs}(k)$, using an obvious nomenclature, is shown in Figure 3.8: 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 to reach (3.9) 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)

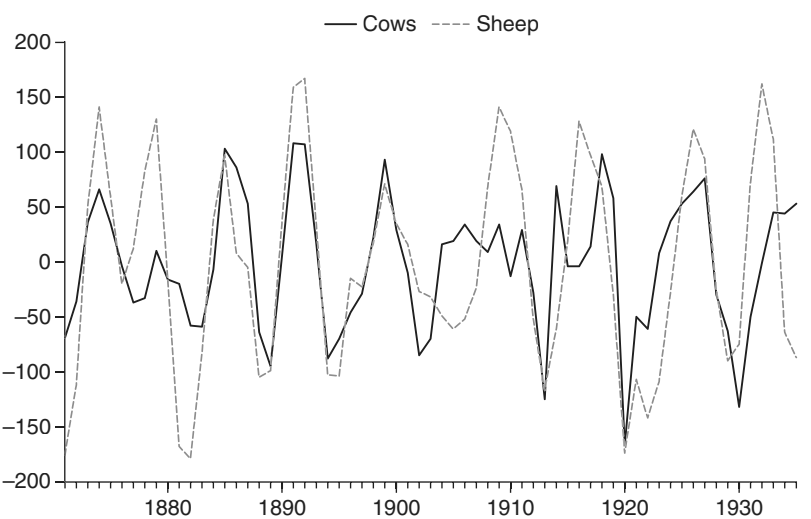


Figure 3.7 Cow and sheep populations for England and Wales, 1871–1935

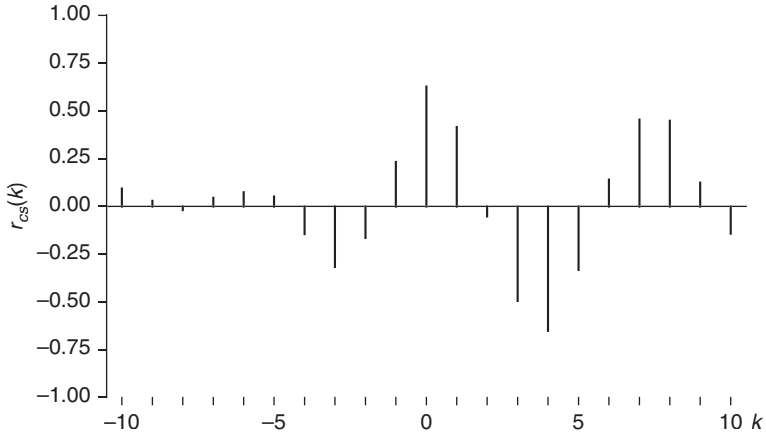


Figure 3.8 Cross-correlations between cow and sheep populations, $-10 \leq k \leq 10$

'Internal' correlations and the lambdagram

3.13 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. 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 behavior of the quantity

$$\lambda_n = \frac{2}{n}((n-1)\rho_1 + (n-2)\rho_2 + \cdots + \rho_{n-1}) \tag{3.10}$$

as n increases. This can be written as

$$\lambda_n = \frac{2}{n} T_n$$

where

$$T_n = \sum_{i=1}^{n-1} S_i \qquad S_i = \sum_{j=1}^{i} \rho_j$$

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 $\rho_i = (k - i)/k$, $i = 1, \ldots, k - 1$, $\rho_i = 0$, $i \geq k$, $S_n = \frac{1}{2}(k - 1)$ 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 3.9, which displays the lambdagram for $k = 5$.

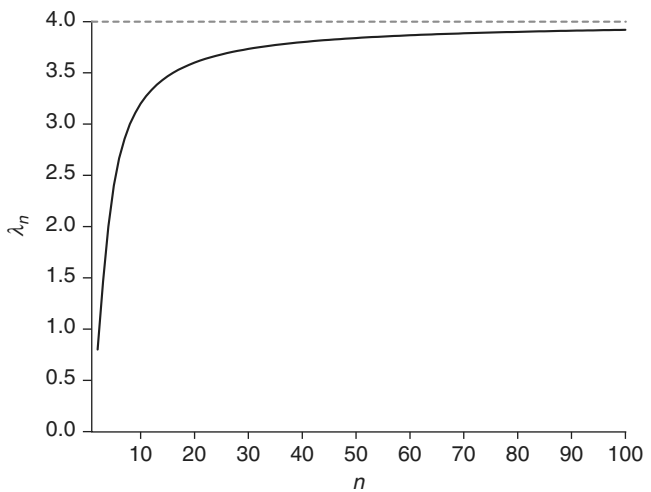


Figure 3.9 Lambdagram for a correlated series formed by summing the terms of a random series in overlapping groups of five

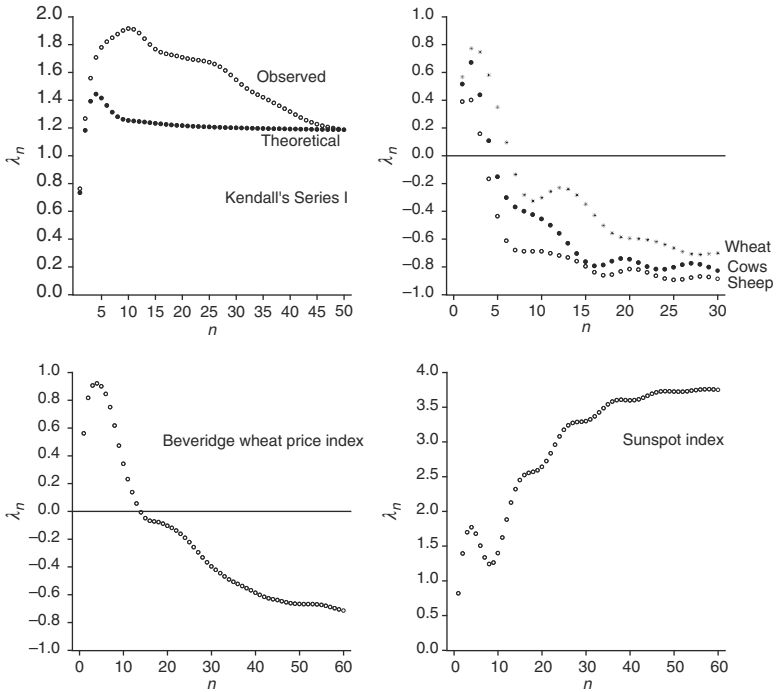


Figure 3.10 Calculated lambdagrams for a variety of time series

Figure 3.10 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 the Beveridge 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 (3.5), Kendall (1945b) analysed the implications for the lambdagram of this underlying generating process. For the process of **§3.6**, Kendall showed that the limiting value of the lambdagram for large n is

$$\lambda = \frac{-2(a+b+b^2)}{(1+b)(1+a+b)} \tag{3.11}$$

If $b = 1$ then it is easy to see that $\lambda = -1$, while using (3.4) and (3.8) 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.

3.14 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 3.10. The limiting value from (3.11) 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 **§3.7**]. If this is the correct explanation I should expect it to be equally 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 of a and b for each of the other series shown in Figure 3.10 can be computed using (3.10) and the limiting values of the lambdagram calculated using (3.11). 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 price index and 0.935 for the sunspot index. From Figure 3.10 it is clear that none of these limiting values look to be very close to the values that the empirical lambdagrams appear 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.

Estimation of autoregressive models

3.15 Yule had estimated the coefficients of his autoregressions by ordinary least squares. Using the second-order autoregression (3.2) as an example, then (3.8), which expresses the coefficients a and b in terms of the first two serial correlations r_1 and r_2 as

$$a = -\frac{r_1(1 - r_2)}{1 - r_1^2} \qquad b = \frac{r_1^2 - r_2}{1 - r_1^2}$$

provides the least squares estimates of the coefficients (strictly, these expressions are asymptotically equivalent to the least squares estimates, that is, they are identical if the serial correlations are estimated as $r_k = \sum_{t=1}^{T-k} x_t x_{t+k} \big/ \sum_{t=1}^{T} x_t^2$). The properties of the least squares estimator were analyzed by Mann and Wald (1943), who showed that they were equivalent to maximum likelihood estimators on the assumption that ε_t was identically and independently distributed.

The equations for a and b can be expressed as the pair

$$\begin{aligned} r_1 + a + r_1 b &= 0 \\ r_2 + a r_1 + b &= 0 \end{aligned}$$

Wold (1938, chapter III. 24) showed that, in fact, these were only the first two equations in the extended system

$$\begin{aligned} &r_1 + a + r_1 b = 0 \\ &r_2 + a r_1 + b = 0 \\ &r_3 + a r_2 + b r_1 = 0 \\ &\vdots \end{aligned}$$

Kendall (1949) termed this system the *Yule–Walker equations*. The solution of the first two equations yields the standard least squares estimates. The least-squares solution to the first m of these equations is obtained by minimizing

$$\sum_{i=1}^{m} (r_i + a r_{i-1} + b r_{i-2})^2$$

For example, using the first three Yule–Walker equations leads to the pair of equations

$$\begin{aligned}(r_1 + r_1r_2 + r_2r_3) + a(1 + r_1^2 + r_2^2) + b(2r_1 + r_1r_2) &= 0\\ (r_1^2 + r_2 + r_1r_3) + a(2r_1 + r_1r_2) + b(1 + 2r_1^2) &= 0\end{aligned}$$

and the solutions

$$a = \frac{(2r_1 + r_1r_2)(r_1^2 + r_2 + r_1r_3) - (1 + 2r_1^2)(r_1 + r_1r_2 + r_2r_3)}{(1 + 2r_1^2)(1 + r_1^2 + r_2^2) - (2r_1 + r_1r_2)^2}$$

$$b = \frac{(2r_1 + r_1r_2)(r_1 + r_1r_2 + r_2r_3) - (1 + r_1^2 + r_2^2)(r_1^2 + r_2 + r_1r_3)}{(1 + 2r_1^2)(1 + r_1^2 + r_2^2) - (2r_1 + r_1r_2)^2}$$

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 the 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 (1947). Again, neither method proved superior to least squares, which has since become the standard method of estimating the coefficients of autoregressions.

Fitting polynomial trends

3.16 As we discussed in **§3.2–3.5**, a popular method of detrending during the first half of the twentieth century was to use a moving average and many of the variants are discussed in detail in Mills (2011a, chapter 10). An important reason for their popularity was that they could be computed essentially as a sequence of summations, which substantially minimized the arithmetic burden. As computational requirements became less of a concern, attention focused on the direct fitting of local polynomials. The general approach was set out by Kendall in Volume 2 of his *Advanced Theory of Statistics* (Kendall, 1946). This 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 date). The operation is then repeated using the terms $u_2, \ldots, u_{n+1}$ to obtain the next trend value $v_{(n+3)/2}$, and then repeated throughout the time series, finally obtaining, for the terms $u_{T-n+1}, \ldots, u_T$, 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 enables the trend values to be calculated directly as a weighted moving average. To see this, again put $n = 2m+1$ and, without loss of generality, consider the sequence of terms u_{-m}, $u_{-m+1}, \ldots, u_0, \ldots, u_{m-1}, u_m$. To fit a polynomial of degree p by least squares to this sequence requires solving the $p+1$ equations

$$\frac{\partial}{\partial a_j} \sum_{t=-m}^{m} (u_t - a_0 - a_1 t - \cdots - a_p t^p)^2 = 0 \qquad j = 0, 1, \ldots, p$$

which gives equations of the form

$$\sum t^j u_t - a_0 \sum t^j - a_1 \sum t^{j+1} - \cdots - a_p \sum t^{j+p} = 0 \quad j = 0, 1, \ldots, p \tag{3.12}$$

Since the summations in (3.12) are functions of m only, solving for a_0 yields an equation of the form

$$a_0 = c_0 + c_1 u_{-m} + c_2 u_{-m+1} + \cdots + c_{2m+1} u_m \tag{3.13}$$

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 (3.13), 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 (3.13): if the sequence is $u_k, \ldots, u_{2m+k}$, 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 (3.12) 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

$$\begin{aligned} \sum u &= 7a_0 + 28a_2 \\ \sum tu &= 28a_1 + 196a_3 \\ \sum t^2u &= 28a_0 + 196a_2 \\ \sum t^3u &= 196a_1 + 1588a_3 \end{aligned} \tag{3.14}$$

These may be solved to give, for a_0,

$$\begin{aligned} a_0 &= \tfrac{1}{21}\left(7\sum u - \sum t^2u\right) \\ &= \tfrac{1}{21}(-2u_{-3} + 3u_{-2} + 6u_{-1} + 7u_0 + 6u_1 + 3u_2 - 2u_3) \\ &= \tfrac{1}{21}[-2, 3, 6, 7, 6, 3, -2] \end{aligned}$$

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

$$\begin{aligned} a_0 &= \tfrac{1}{21}((-2 \times 0) + (3 \times 1) + (6 \times 8) + \cdots - (2 \times 216)) \\ &= \tfrac{1}{21}567 = 27 \end{aligned}$$

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 (3.14) 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] Further, although rather arcane, properties of the method were later derived in Kendall (1961).

The sampling theory of serial correlations

3.17 As we saw in **§3.8**, Kendall (1945a) had expressed frustration at the lack of a sampling theory related to serial correlations when attempting to interpret the correlograms obtained from his experimental series, going on to say that

> '(t)he 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, Maurice Bartlett 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, and whose aim was 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 main result was to show 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. The actual formula is

$$V(r_k) = \frac{1}{T} \sum_{i=-\infty}^{\infty} (\rho_i^2 + \rho_{i-k}\rho_{i+k} - 4\rho_k\rho_i\rho_{i+k} + 2\rho_i^2\rho_k^2) \tag{3.15}$$

but useful approximations may be obtained in certain cases. If x_t is random, so that $\rho_k = 0$, $k \neq 0$, then, from (3.15), $V(r_k) = 1/T$, which is the variance of a correlation coefficient from a bivariate normal sample and was the formula employed by Kendall (1943): cf. **§3.8**. 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 (3.15) we have

$$V(r_k) = \frac{1}{T} \sum_{i=-(k-1)}^{k-1} \rho_i^2 = \frac{1}{T}(1 + 2\rho_1^2 + \cdots + 2\rho_{k-1}^2) \tag{3.16}$$

a formula whose square root has since become known as the 'Bartlett standard error'. Suppose that x_t is generated by a first-order autoregression, now denoted as AR(1), and which is also known as a Markov process. We then have $\rho_k = \rho^k$, and

$$V(r_k) = \frac{1}{T}\left(\frac{(1+\rho^2)(1-\rho^{2k})}{1-\rho^2} - 2k\rho^{2k}\right)$$

which, for large k, becomes

$$V(r_k) = \frac{1}{T} \sum_{i=-\infty}^{\infty} \rho^{|2i|} = \frac{1}{T}\frac{1+\rho^2}{1-\rho^2}$$

3.18 Bartlett (1946) used these results to analyze the correlogram of Kendall's (1944) artificial series of length $T = 65$ generated as (3.5) but with the error process now an integer rectangular random variable ranging from -9.5 to $+9.5$ (cf. the process in **§3.7**). 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 3.11 (two-standard error bounds under the null

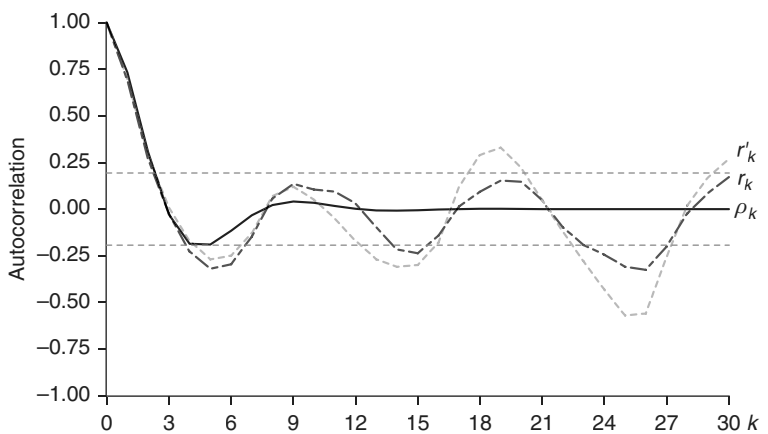


Figure 3.11 Correlogram and autocorrelations of Kendall's (1944) artificial series $x_t - 1.5x_{t-1} + 0.5x_{t-2} = u_t$

hypothesis that the series is random are $2/\sqrt{65} \approx 0.25$). The first estimate of the correlogram uses the large sample formula

$$r_k = \frac{\sum_{t=1}^{T-k} x_t x_{t+k}}{\sum_{t=1}^{T} x_t^2},$$

while 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 (3.16), $V(\rho_k) \approx 2.44/T$ 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 3.11).

Bias in the estimation of serial correlations

3.19 As well as developing the sampling theory of serial correlations, there was also great attention paid to examining possible biases in the estimates themselves. Kendall (1954) (along with Marriott and Pope, 1954) showed that, for the Markov AR(1) 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}, so that, for example,

$$E(r_1) = \rho - \frac{1}{T-1}(1+3\rho)$$

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 first-order moving average scheme

(denoted MA(1)) $x_t = \eta_t + \theta\eta_{t-1}$, so that $\rho_1 = \theta/(1+\theta^2) = \rho$ and $\rho_k = 0$, $k \geq 2$, we obtain, using the method of Kendall (1954),

$$E(r_1) = \rho + \frac{1}{T-1}(1+\rho)(4\rho^2 - 2\rho - 1)$$
$$E(r_2) = -\frac{1}{T-2}(1 + 2\rho + 2\rho^2)$$
$$E(r_k) = -\frac{1}{T-k}(1+2\rho) \qquad k > 2$$

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.).

Kendall (1954) 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. Moreover, Kendall argued that expansions of the type being used above are asymptotic and may not be accurate unless the serial correlations decline rapidly.

Kendall's later time series research

3.20 Kendall's later work on time series was restricted (apart from a few peripheral book reviews) to analyzing share prices (Kendall, 1953), to investigating the higher moments of the 'Leipnik distribution', that is, the distribution of the serial correlation coefficient of a Markov process (Kendall, 1957), to a study of economic forecasting (Coen, Gomme and Kendall, 1969), to a review of Box and Jenkins' book (Kendall, 1971), to a very short discussion of spectral analysis aimed at geophysicists (Kendall, 1973a) and to a textbook (Kendall, 1973b). The review of Box and Jenkins will be commented upon in **§8.2**, while the papers on the Leipnik distribution and on spectral analysis are tangential to our theme and will be ignored.

3.21 Kendall (1953) analyzed many different weekly financial price series and came to the same conclusion as Holbrook Working (1934) some two decades earlier, 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. (Kendall, 1953, 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 the history of statistics and of time series, did not appear to be familiar with the term 'random walk', even though the term had first been used almost half a century earlier. Although 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).

3.22 Coen, Gomme and Kendall (1969) carried out an exercise in economic forecasting in which the focus was on uncovering and using dynamic relationships existing between, in standard regression terminology, the dependent variable and lagged values of the regressors. Notwithstanding Kendall's findings a decade and a half earlier (see **§3.21**

above), the focus was on forecasting quarterly values of the FT ordinary share index and, although a variety of regression models were investigated using a selection of regressors (or indicator variables), concentration first fell on the following model:

$$Y_t = \beta_0 + \beta_1 X_{1,t-6} + \beta_2 X_{2,t-7} + n_t \tag{3.17}$$

Here Y_t is the share price index in quarter t, $X_{1,t-6}$ is UK car production in quarter $t-6$ and $X_{2,t-7}$ is the FT commodity index in quarter $t-7$, so that share prices react to car production six quarters earlier and to commodity prices seven quarters earlier. The residual n_t was assumed to be independently and identically distributed with a zero mean and constant variance – the typical regression assumptions.

The lags were selected on the basis of initial and exploratory graphical examination ('by graphing the series on transparencies to a roughly comparable scale and then superposing them, sliding them along the time-axis to see whether there was any fairly obvious coincident variation', Coen et al., 1969, page 136) and calculating cross-correlation coefficients between the share price index and the indicator variables (either linearly detrended or annually differenced). Once a lag length was tentatively identified a regression analysis was undertaken, which was 'conducted by including among the regressors a variable at several lags around the value where the cross-correlation was a maximum in absolute value, for example, if Y_t had a high correlation with X_{t-u} we might include among the regressors $X_{t-u-4}, X_{t-u-3}, \ldots, X_{t-u+4}$' (ibid., page 140, notation altered for consistency). The step-wise algorithm of Beale, Kendall and Mann (1967) was then employed to reject redundant regressors, so that 'the regression analysis effectively determines the lags to be included in the final equations' (Coen et al., page 140). Figure 3.12 shows plots of the share index against car production six quarters earlier and the commodity index seven quarters earlier and appears to show a positive relationship between the former pair and a negative relationship between the latter pair, this being confirmed by the fitted regression (t-ratios shown in parentheses)[5]

$$\underset{}{Y_t} = \underset{}{653.2} + \underset{(14.1)}{0.00047} X_{1,t-6} \underset{(-9.9)}{-6.128}\ X_{2,t-7} \quad R^2 = 0.90 \quad 1954.2\text{–}1966.4$$

Forecasts for 1967 from this regression are shown in Figure 3.13: 'the resulting prediction was quite unambiguous: before the end of the year a downward swing would start on the stock market and would be equivalent to a serious recession' (ibid., 1969, page 140). Unfortunately,

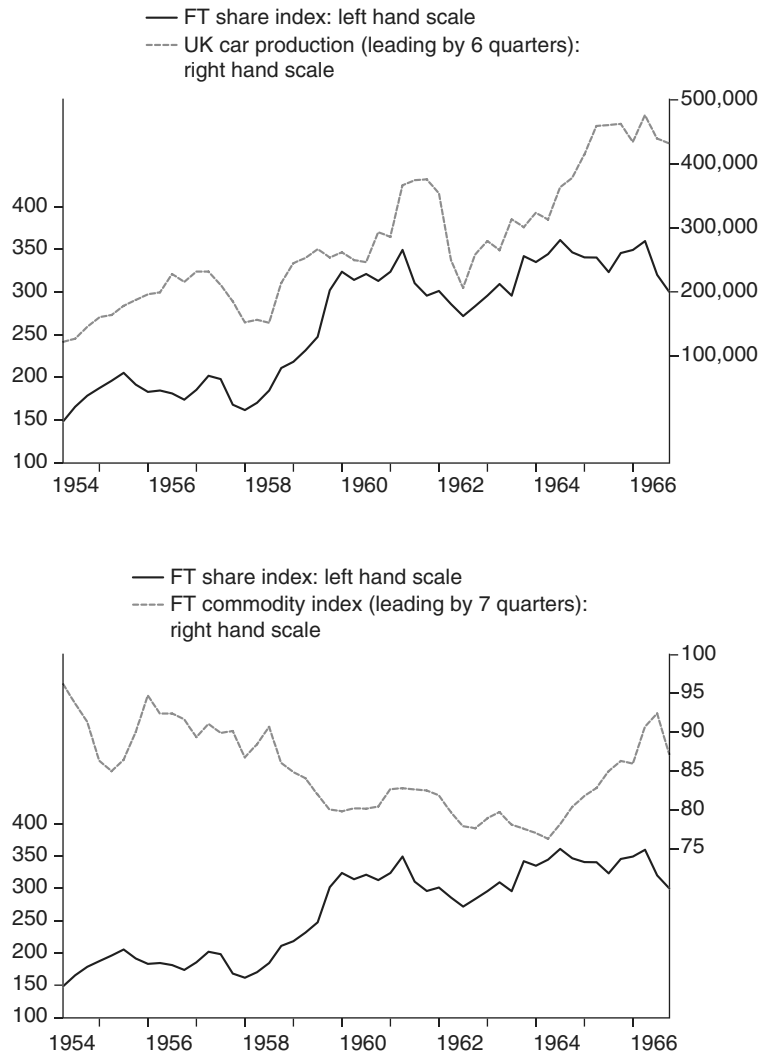


Figure 3.12 FT ordinary share index plotted against UK car production six quarters earlier and the FT commodity index seven quarters earlier: 1954.1–1967.4

'(a)s things turned out the market went on rising until the most optimistic bulls had doubts about its stability. It looked as if our first attempts at forecasting were a spectacular failure' (ibid., page 140), a fact that Coen et al. attributed to the subsequent sterling crisis and

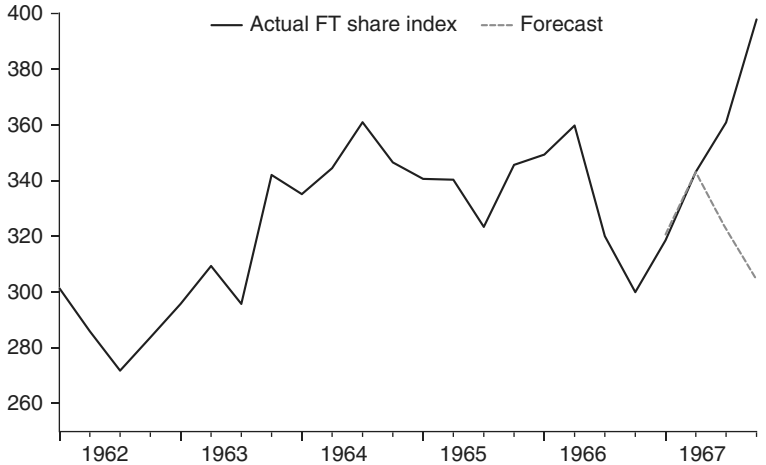


Figure 3.13 Forecasts from equation (3.17) for 1967

devaluation of 1967, after which 'there was a consumer-spending spree, and a stock market boom and imports rose to a level which caused acute concern' (ibid., page 143).

In response to this forecasting failure, Coen et al. included a constructed variable, termed the 'euphoria index', in (3.17) and also considered further regressors, reporting the following regression[6]

$$\underset{}{Y_t} = \underset{}{539.0} + \underset{(8.9)}{2.127}X_{3,t-17} - \underset{(3.4)}{0.211}X_{4,t-8} - \underset{(6.0)}{3.042}X_{5,t-13} \tag{3.18}$$

$$R^2 = 0.96 \quad 1952.2\text{–}1966.4$$

Here X_3 is the Standard and Poor stock index, X_4 is Reuter's commodity index and X_5 is a UK government securities index. The forecasts out to 1968.1 from this model are shown in Figure 3.14 and they are much more accurate than those from (3.17) during 1967, but are again 'out of control', to use Coen et al.'s phrase, by early 1968.

3.23 Coen et al.'s paper elicited a good deal of interest and was accompanied by detailed comments from a number of discussants from both the statistics and econometrics communities. Chief amongst the concerns of the discussants was the lack of a theoretical economic model underlying the forecasting equations that would explain the length of the lags attached to the regressors, the question of endogeneity and,

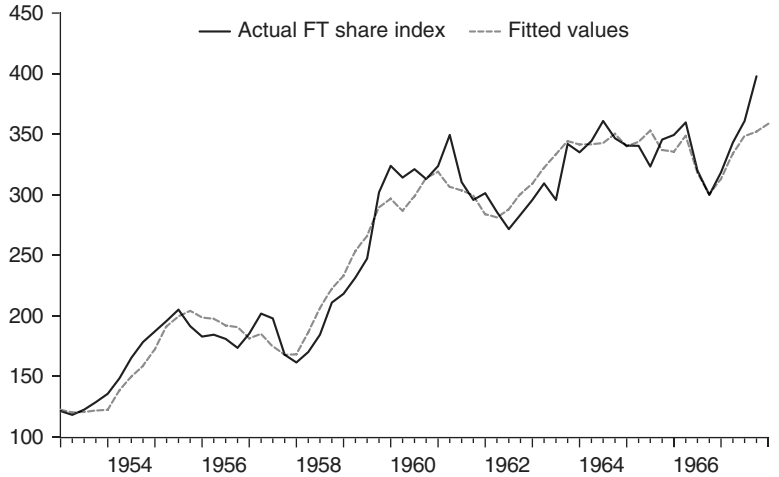


Figure 3.14 Forecasts from equation (3.18) for 1967 and early 1968

most importantly, the complete absence of lagged dependent variables as regressors and the consequent possibility that the regressions suffered from serial correlation, which might be thought to be an odd omission given Kendall's seminal research on autoregressive processes discussed earlier. Although the authors made a spirited defense of their paper, many of the above criticisms were later formalized in quite devastating fashion by Box and Newbold (1971), which will be discussed in **§8.3**.

3.24 Kendall's textbook, published in 1973, was one of the first to provide an explicitly introductory exposition of time series analysis at the expense of giving a comprehensive and rigorous treatment of all aspects of the subject. Much of the material was taken from the chapters on time series contained in the third edition of *The Advanced Theory of Statistics*, while another chapter was devoted to the analysis of forecasting discussed in **§3.22–3.23** above. While reviewers were generally impressed by the informal style and lucidity of the exposition, the overall impression given by the reviews was that the book represented the swansong of an ageing master of the subject who had not really kept up with, or indeed was even very sympathetic to, developments that had taken place over the last 15 years or so. In fact, there had been several major developments in this period that had transformed the subject completely, both theoretically and empirically, and it to those developments that we must now necessarily turn to.

4

Durbin: Inference, Estimation, Seasonal Adjustment and Structural Modelling

James Durbin

4.1 James Durbin was born in 1923 and, like Yule and Kendall before him, has a St John's, Cambridge connection, for he obtained a 'wartime' BA in Mathematics there before spending the rest of the war attached to the Army Operational Research Group and then the British Boot, Shoe and Allied Trades Research Association as a statistician. After demobilization, he returned to Cambridge to become part of the first intake into the Postgraduate Diploma in Mathematical Statistics before joining the research staff at the Department of Applied Economics in 1948. In 1950 he was appointed, as part of Maurice Kendall's professorial deal with the LSE, to an assistant lectureship in statistics and thus began his long association with that institution. After becoming a Reader in Statistics in 1953 he was promoted to Professor in 1961 on Kendall's departure to SciCon, remaining in this post until official retirement in 1988. Since then James Durbin has remained professionally active, continuing to publish well into the first decade of the twenty-first century: see, for example, Durbin and Koopman (2001) and Durbin (2004). In 2007 he became an honorary fellow of the Centre for Microdata Methods and Practice (CeMMAP) at UCL.

Having been the recipient of the RSS's Guy Medals in Bronze and Silver, in 1966 and 1976 respectively, and been president of the Society in 1986–7, Durbin was awarded the Guy Medal in Gold in 2008 for

> a life-time of highly influential contributions which have given him outstanding international recognition as a leader in [the] field, taking particular account of his pioneering work on testing for serial correlation in regression, on estimating equations, on Brownian

motion and other processes crossing curved boundaries, on goodness of fit tests with estimated parameters, and on many aspects of time series analysis especially in areas relevant to econometrics, and also his remarkable service to the wider statistical profession on the international stage.

It is, of course, these 'many aspects of time series analysis' that we shall focus upon in this chapter.

Inference on the first-order serial correlation coefficient

4.2 Along with the large-sample theory of serial correlation that developed from the seminal research of Bartlett (1946), the 1940s also saw great progress made on developing exact, small-sample, inferential methods. Beginning with Anderson (1942), the *cyclic* definition of the first-order serial correlation of the observed series $X_1, X_2, \ldots, X_T$, with sample mean $\bar{X} = \sum_{t=1}^{T} X_t/T$, was the prime focus of attention, mainly for reasons of analytical tractability. This correlation coefficient may be expressed as

$$r_1^c = \frac{\sum_{t=1}^{T} x_t x_{t+1}}{\sum_{t=1}^{T} x_t^2} \tag{4.1}$$

where $x_t = X_t - \bar{X}$, and assumes 'circularity', so that $x_{T+t} = x_t$. The distribution of r_1^c was known to have an exact form under the assumption of independence, being a piecewise density function for very small samples but quickly approaching normality centered on a mean of $-1/(T-1)$. To illustrate this, Figures 4.1 and 4.2 show, respectively, the exact distribution of r_1^c for $T = 6$ and 7 (labelled $r(6)$ and $r(7)$), and for $T = 15$ ($r(15)$) with its normal approximation (see Dixon, 1944).

Watson and Durbin (1951) argued that this circular conception of the stochastic process generating X_t, as embodied in the cyclic definition (4.1) of r_1^c, while being analytically convenient, was rarely plausible in practice. They thus relaxed the assumption of circularity and considered the following statistic for testing independence:

$$d = \frac{\sum_{i=2}^{T} (X_{i-1} - X_i)^2 - (X_n - X_{n+1})^2}{\sum_{i=1}^{T} (X_i - \bar{X})^2}$$

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

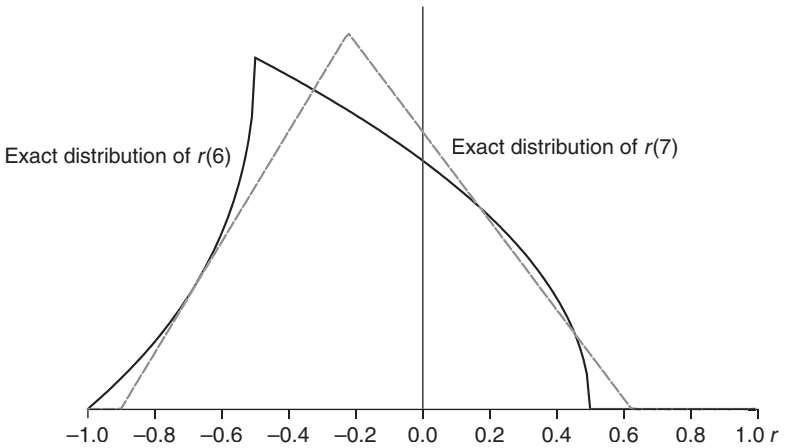


Figure 4.1 Exact distributions of the first-order serial correlation coefficient for $T = 6$ and $T = 7$

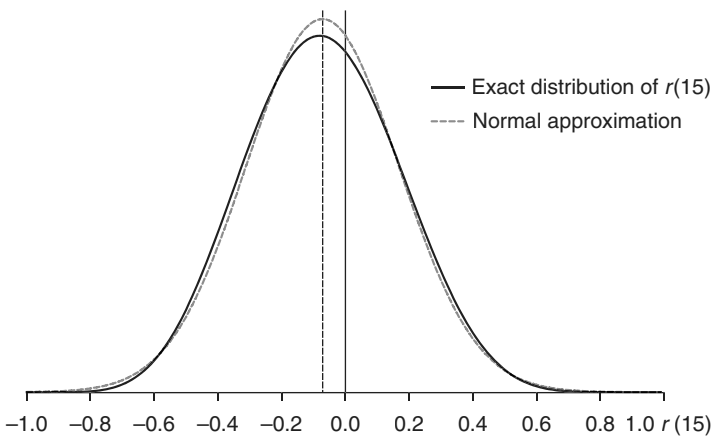


Figure 4.2 Exact distribution of the first-order serial correlation coefficient for $T = 15$ with its normal approximation

of Anderson (1942), Watson and Durbin (1951) showed that, for $\zeta_i = 4\sin^2(n-i)\pi/2n$, the distribution of d is

$$P(d > d') = \sum_{i=1}^{s} \frac{(\zeta_i - d')^{n-\frac{3}{2}}}{\zeta_i^{\frac{1}{2}} \prod_{j=1, j\neq i}^{n-1} (\zeta_i - \zeta_j)} \qquad \zeta_{s+1} \leq d' \leq \zeta_s \quad s = 1, 2, \ldots, n-1$$

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: for example, for $T = 12$ the 5% critical value is 0.967 while for $T = 30$ it is 1.35. This statistic was extended by Durbin and Watson (1950, 1951, 1971: see also Durbin, 1982) to test for first-order serial correlation in regression models using the regression residuals in place of the observed values of the dependent variable, becoming probably the most recognized test statistic in econometrics. Known eponymously as the *Durbin–Watson test*, it is a staple output of most econometric packages, even though its performance can be beaten by several tests that have since been developed. However, since the test is not applicable to models containing lagged dependent variables, thus ruling out autoregressions, for example, this extension will not be considered further here, although a later extension, *Durbin's h-test* (Durbin, 1970), was explicitly designed to deal with regression models containing a single lagged dependent variable.

Estimation and inference in moving average models

4.3 While the estimation of autoregressive models was the main focus of attention during the 1940s and 1950s, much less progress was made on the estimation of the moving average schemes introduced by Wold (1938) and hence, not surprisingly, on the combined class of mixed autoregressive-moving average models introduced by Walker (1950). Whittle (1953, 1954a) developed a large sample approach to the estimation of moving average models that, while providing 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 subsequently extended by Walker (1961).

4.4 Durbin began by focusing on estimating the parameter β in the first-order moving average model

$$x_t = \varepsilon_t + \beta\varepsilon_{t-1} \qquad t = 1, 2, \ldots, T \tag{4.2}$$

where ε_t is identically and independently distributed and it is assumed that $|\beta| < 1$, so that the moving average is 'regular' in Wold's (1938, III.26) terminology, although it is now more commonly referred to as being 'invertible'. A perhaps obvious estimator is to use the result that

$\rho_1 = \beta/(1+\beta^2)$ (obtained by setting $k = h = 1$ in equation (250) of Wold, *ibid.*, page 122), solve the quadratic $r_1\tilde{\beta}^2 - \tilde{\beta} + r_1 = 0$, and use the regular solution $|\tilde{\beta}| < 1$. Whittle (1953), however, showed that this estimator was very inefficient but his proposed adjustment was extremely complicated. Durbin thus considered the infinite autoregressive representation of (4.2) truncated at lag p:

$$\begin{aligned} &x_t - \beta x_{t-1} + \beta^2 x_{t-2} - \cdots + (-\beta)^p x_{t-p} \\ &\quad = x_t + \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} = \varepsilon_t \end{aligned}$$

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 (ML) estimator of β is given by

$$\hat{\beta} = -\frac{\sum_{k=0}^{p-1} \hat{\alpha}_k \hat{\alpha}_{k+1}}{\sum_{k=0}^{p} \hat{\alpha}_k^2} \tag{4.3}$$

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 (1953) 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 (4.2) Durbin showed that the statistic

$$T\left((1-\beta^2)\sum_{k=0}^{p}\alpha_k^2 - 1\right) \sim \chi^2(p-1)$$

can be employed.

Durbin then considered the extension to higher-order moving averages, which is straightforward, at least in theory. For the model

$$x_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q} \tag{4.4}$$

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

$$\begin{bmatrix} \sum_{k=0}^{p}\hat{\alpha}_k^2 & \sum_{k=0}^{p-1}\hat{\alpha}_k\hat{\alpha}_{k+1} & \cdots & \sum_{k=0}^{p-q+1}\hat{\alpha}_k\hat{\alpha}_{k+q-1} \\ \sum_{k=0}^{p-1}\hat{\alpha}_k\hat{\alpha}_{k+1} & \sum_{k=0}^{p}\hat{\alpha}_k^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{k=0}^{p-q+1}\hat{\alpha}_k\hat{\alpha}_{k+q-1} & \cdots & \cdots & \sum_{k=0}^{p}\hat{\alpha}_k^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_q \end{bmatrix} = - \begin{bmatrix} \sum_{k=0}^{p-1}\hat{\alpha}_k\hat{\alpha}_{k+1} \\ \sum_{k=0}^{p-2}\hat{\alpha}_k\hat{\alpha}_{k+2} \\ \vdots \\ \sum_{k=0}^{p-q}\hat{\alpha}_k\hat{\alpha}_{k+q} \end{bmatrix}$$

The asymptotic variance matrix of $\hat{\beta}_1, \ldots, \hat{\beta}_q$ is $T^{-1}\mathbf{V}_q$, where

$$\mathbf{V}_q = \begin{bmatrix} 1-\beta_q^2 & \beta_1-\beta_{q-1}\beta_q & \beta_2-\beta_{q-2}\beta_q & \cdots & \beta_{q-1}-\beta_1\beta_q \\ \beta_1-\beta_{q-1}\beta_q & 1+\beta_1^2-\beta_{q-1}^2-\beta_q^2 & & & \vdots \\ \beta_2-\beta_{q-2}\beta_q & \beta_1+\beta_1\beta_2-\beta_{q-2}\beta_{q-1}-\beta_{q-1}\beta_q & \ddots & & \vdots \\ \vdots & & & 1+\beta_1^2-\beta_{q-1}^2-\beta_q^2 & \beta_1-\beta_{q-1}\beta_q \\ \beta_{q-1}-\beta_1\beta_q & \cdots & \cdots & \beta_1-\beta_{q-1}\beta_q & 1-\beta_q^2 \end{bmatrix}$$

Thus, for $q = 1, 2, 3$,

$$\mathbf{V}_1 = 1-\beta_1^2 \qquad \mathbf{V}_2 = \begin{bmatrix} 1-\beta_2^2 & \beta_1-\beta_1\beta_2 \\ \beta_1-\beta_1\beta_2 & 1-\beta_2^2 \end{bmatrix}$$

$$\mathbf{V}_3 = \begin{bmatrix} 1-\beta_3^2 & \beta_1-\beta_2\beta_3 & \beta_2-\beta_1\beta_3 \\ \beta_1-\beta_2\beta_3 & 1+\beta_1^2-\beta_2^2-\beta_3^2 & \beta_1-\beta_2\beta_3 \\ \beta_2-\beta_1\beta_3 & \beta_1-\beta_2\beta_3 & 1-\beta_3^2 \end{bmatrix}$$

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 ijth element of $\mathbf{V}_q^{-1}$ evaluated at $\beta_k = \beta_{0k}$, $k = 1, \ldots, q$. The goodness-of-fit of (4.4) 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.

4.5 Durbin (1959) examined this method by simulating twenty series of length $T = 100$ from the model (4.2) with $\beta = 0.5$ and $\varepsilon_t \sim N(0, 1)$, and computing $\hat{\beta}$ from (4.3) 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 $r_1\tilde{\beta}^2 - \tilde{\beta} + r_1 = 0$ are imaginary $\tilde{\beta}$ was taken to be one: this will occur when $r_1 > 0.5$). Table 4.1 shows the results, along with summary statistics, obtained by recreating Durbin's simulation. $\hat{\beta}_C$ 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 (4.3). 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 $(1 - 0.5^2)/100 = 0.0075$, 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.[1]

4.6 Walker (1961) was concerned that the truncation of the infinite autoregression to a finite-order p might lead to problems in some

Table 4.1 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

	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

circumstances and thus proposed an extension of Durbin's method which had the added advantage of allowing bias adjustments to be made fairly straightforwardly. Walker showed that his estimator had excellent asymptotic efficiency for p as small as 4 unless ρ was close to its maximum value of 0.5 for a first-order moving average. He also showed that his estimator suffered less bias than Durbin's, a property that continued to hold when higher-order moving averages were considered. Interestingly, however, Walker argued that it was by no means clear why his method appeared to suffer from less bias than Durbin's, suggesting that the improvements that he found in his simulations were 'probably fortuitous'.

Estimation and inference in autoregressive-moving average models

4.7 Durbin (1960a) and Walker (1962) extended their methods for estimating moving averages to mixed autoregressive-moving average models of the general form

$$x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \tag{4.5}$$

which may be termed an ARMA(p, q) process and where it is assumed that all the roots of the equations $z^p - \phi_1 z^{p-1} - \cdots - \phi_p = 0$ and $z^q + \theta_1 z^{q-1} + \cdots + \theta_q = 0$ have modulus less than unity (models of this type seem to have been first considered by Walker, 1950). Durbin pointed out that, apart from obviously containing the autoregressive and moving average models as special cases, when $q = p - 1$ (4.5) is invariant under changes in the time period between successive observations, which is not the case for the simpler models (a point made earlier by Quenouille, 1958). Furthermore, equi-spaced observations from a continuous stochastic process generated by a linear stochastic differential equation will also conform to (4.5) with $q = p - 1$. Nevertheless, notwithstanding the theoretical importance of the ARMA process, until Durbin only Walker (1950) and Quenouille (1958) had considered fitting the model, with neither attempting an efficient method of estimation.

4.8 Durbin began by focusing attention on the ARMA(1,1) process

$$x_t - \phi x_{t-1} = \varepsilon_t + \theta \varepsilon_{t-1} \tag{4.6}$$

and suggested fitting an autoregression of order p, as in **§4.4**, and estimating the parameters by

$$\hat{\phi} = \frac{\hat{\alpha}_1 r_2 + \hat{\alpha}_2 r_3 + \cdots + \hat{\alpha}_p r_{p+1}}{\hat{\alpha}_1 r_1 + \hat{\alpha}_2 r_2 + \cdots + \hat{\alpha}_p r_p} \tag{4.7}$$

and

$$\hat{\theta} = -\hat{\phi} + \frac{r_1 + \hat{\alpha}_1 r_2 + \cdots + \hat{\alpha}_p r_{p+1}}{1 + \hat{\alpha}_1 r_1 + \cdots + \hat{\alpha}_p r_p} \tag{4.8}$$

showing that this was the solution obtained by minimizing the sum of squared residuals from (4.6) 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

$$\ell_i = \hat{\alpha}_i + \hat{\phi}\ell_{i-1} \qquad \ell_0 = 1 \quad i = 1, \ldots, p$$

Durbin showed that an efficient estimator of θ was

$$\hat{\theta} = \frac{\sum_{i=0}^{p} \ell_i \ell_{i+1}}{\sum_{i=0}^{p} \ell_i^2} \tag{4.9}$$

Given $\hat{\theta}$, and now defining

$$w_t = x_t - \phi w_{t-1} = \phi w_{t-1} + \varepsilon_t, \qquad w_0 = 0, \quad t = 1, \ldots, T$$

an efficient estimator of ϕ is then

$$\hat{\phi} = \frac{\sum_{t=1}^{T-1} w_t w_{t+1}}{\sum_{t=1}^{T-1} w_t^2} \tag{4.10}$$

Thus, given either (4.7) or (4.8) as an initial condition, (4.9) and (4.10) can be used iteratively to obtain estimates of ϕ and θ. Durbin then showed how this approach can readily be extended to the general ARMA(p, q) model.

4.9 Durbin provided no simulation evidence on the properties of his procedure, but this is not difficult to do and hence the simulation of **§4.5** was repeated for the ARMA(1,1) process (4.6) with $\phi = 0.8$ and $\theta = 0.5$, Table 4.2 presenting the results. 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. However, the results are by no means satisfactory, with ϕ being consistently underestimated and θ being consistently overestimated.

Table 4.2 Twenty simulations of length $T = 100$ from a first-order autoregressive-moving average model with $\phi = 0.8$ and $\theta = 0.5$

Series	$\hat{\phi}$	$\hat{\theta}$	Series	$\hat{\phi}$	$\hat{\theta}$
1	0.649	0.718	11	0.751	0.670
2	0.603	0.650	12	0.673	0.696
3	0.578	0.643	13	0.749	0.733
4	0.578	0.582	14	0.714	0.688
5	0.738	0.752	15	0.787	0.744
6	0.575	0.610	16	0.691	0.758
7	0.712	0.728	17	0.617	0.678
8	0.762	0.699	18	0.717	0.708
9	0.653	0.686	19	0.734	0.735
10	0.774	0.760	20	0.666	0.684
			$\hat{\phi}$		$\hat{\theta}$
Mean			0.686		0.696
Std. Dev.			0.070		0.045
SE of mean			0.065		0.097

Walker (1962) extended his method (**§4.6**) to the ARMA case and provided simulation evidence, which suggested that his estimates were superior to Durbin's in terms of both smaller bias and in the ratio of the bias to the standard error. Walker also proposed a bias adjustment that reduced the bias in ϕ but made the bias in θ worse.

4.10 The fitting of continuous time models from discrete data was considered in Durbin (1961), while Durbin (1960a, 1960b) investigated the estimation of regression models of the form

$$y_t = \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \beta_1 x_{1,t} + \cdots + \beta_q x_{q,t} + \varepsilon_t$$

that is, a regression model in which p lags of the dependent variable were included as regressors, and

$$y_t = \beta_1 x_{1,t} + \cdots + \beta_q x_{q,t} + u_t \qquad u_t - \theta_1 u_{t-1} - \cdots - \theta_p u_{t-p} = \varepsilon_t$$

a model in which the regression errors followed an AR(p) process. The analysis of the estimation of such models subsequently paved the way for many of the major developments in the econometric modelling of time series that were to be made over the following decades (see **§§10.21–10.22**).

Trend elimination and seasonal adjustment

4.11 While visiting Stanford during the summer of 1960, Durbin began working on the nature of trend and the consequences of attempting to eliminate it by various methods. This quickly led him on to issues related to seasonal adjustment, a topic that was becoming of major practical concern in the early 1960s (see Mills, 2011a, chapter 14). Durbin (1962) considered, from the standpoint of estimating the 'residuals', the deviations of the observations from the fitted trend, the consequences of detrending by either taking variate-differences, calculating moving averages or subtracting a low-order polynomial from the original observations. He showed that the three methods could be regarded as essentially equivalent, at least in providing estimates of the periodogram of the series, although differencing could confer some advantages in that a greater concentration of residual trend would be left in the low frequencies of the periodogram.

Durbin (1963) used these ideas to investigate some issues in seasonal adjustment. He began with the customary additive model for an observed series y_t

$$y_t = m_t + s_t + u_t \qquad t = 1, \ldots, T \tag{4.11}$$

in which m_t is the trend, assumed to be a smooth deterministic function of time, s_t is the seasonal component, regarded as strictly periodic with a period of one year, and u_t is assumed to be a stationary random disturbance. It is assumed that there are $p + 1$ years of monthly data available for which (4.11) takes the form

$$y_t = m_t + \alpha_i + u_t \qquad t = 12j + i; \quad i = 1, \ldots, 12; \quad j = 0, \ldots, p \tag{4.12}$$

The monthly constants $\alpha_1, \ldots, \alpha_{12}$ measure deviations of the monthly means from the overall mean of y_t and so may be constrained to sum to zero: $\sum_{i=1}^{12} \alpha_i = 0$. A simple moving average estimator of trend is obtained by taking the mean of two successive arithmetic means of 12 successive observations, so that the 'centred' moving average

$$\begin{aligned} \hat{m}_t &= \tfrac{1}{24}(y_{t-6} + 2y_{t-5} + \cdots + 2y_{t+5} + y_{t+6}) \\ &= \tfrac{1}{24}[1, 2, \ldots, 2, 1]y_t \qquad t = 7, 8, \ldots, 12p + 6 \end{aligned}$$

is used as an estimator of m_t. This will be free of seasonal variation by virtue of the sum constraint placed on the α_is. Let $x_t = y_t - \hat{m}_t$ be the

deviation from trend and let $\bar{x}_1, \ldots, \bar{x}_{12}$ denote the monthly means of the x's:

$$\begin{aligned} \bar{x}_i &= \frac{1}{p}\sum_{j=1}^{p} x_{12j+i} \qquad i = 1, \ldots, 6 \\ &= \frac{1}{p}\sum_{j=0}^{p-1} x_{12j+i} \qquad i = 7, \ldots, 12 \end{aligned} \tag{4.13}$$

The difference in summation limits arises from the fact that six values are 'lost' from each end of the series when the trend is estimated so that x_t can only be obtained for $t = 7$ to $12p + 6$ inclusive. The monthly seasonal constants can then be estimated as

$$a_i = \bar{x}_i - \bar{x} \qquad \bar{x} = \sum_{i=1}^{12} \bar{x}_i / 12$$

After some algebra, Durbin showed that (4.13) can be written as

$$\begin{aligned} \bar{x}_1 &= \bar{y}_1 - \bar{y} + \frac{1}{24p}(y_7 - y_{12p+7}) \\ \bar{x}_i &= \bar{y}_i - \bar{y} + \frac{1}{24p}\left[y_{i+6} - y_{12p+i+6} + 2\sum_{r=7}^{i+5}(y_r - y_{12p+r})\right] \qquad i = 2, \ldots, 6 \\ \bar{x}_i &= \bar{y}_i - \bar{y} - \frac{1}{24p}\left[y_{i-6} - y_{12p+i-6} + 2\sum_{r=i-5}^{6}(y_r - y_{12p+r})\right] \qquad i = 7, \ldots, 11 \\ \bar{x}_{12} &= \bar{y}_{12} - \bar{y} - \frac{1}{24p}(y_6 - y_{12p+6}) \end{aligned}$$

where $\bar{y}$ and $\bar{y}_i$ are defined analogously to $\bar{x}$ and $\bar{x}_i$. Each $\bar{x}_i$ can thus be obtained by adding to the deviation $\bar{y}_i - \bar{y}$ an adjustment term depending only on 12 observations at the beginning and end of the series. There is thus no need to calculate the trend estimate $\hat{m}_t$ and hence the individual x_t values, since the same estimates of the seasonal constants may be obtained directly by applying end-correction terms to the simple monthly means. These results led Durbin to the view that their importance lay in

> the light they throw on the nature of the trend elimination implicit in the method. I feel that the loyalty many economic statisticians have toward the moving-average method arises from their skepticism

> of the adequacy of any simple mathematical model for the fitting of the trends found in many economic time series. The flexibility of the moving average in following a trend of arbitrary shape has a considerable intuitive appeal, and it will doubtless come as a shock to many to realize that what the method reduces to in the end is the adjustment of the raw monthly means by crude estimates of trend derived from a few observations taken from each end of the series. The fact that the behavior of the series at intermediate points of time has no effect whatever on the adjustments for trend demonstrates conclusively that the apparent fidelity with which the moving average reproduces the true trend is illusory as far as the estimation of seasonal variation is concerned. (Durbin, 1963, pages 6–7)

Note that the deviations from trend can be written in the form

$$x_t = -\frac{1}{24}\Delta^2 \sum_{s=-6}^{4} (6 - |s+1|)^2 y_{t+s}$$

Since taking second differences of a quadratic in t gives a constant which will disappear on taking deviations from means to get $\bar{x}_t$, using this moving average to remove trend will eliminate a quadratic trend in the data exactly.

4.12 If the monthly seasonal constants had been estimated from a series of length $12p$ without any trend-elimination procedure having been applied, each estimate would have standard error $\sigma\sqrt{11/12p}$ on the assumption that the disturbances u_t were uncorrelated with constant variance σ^2. Durbin showed that using the centred moving average of the previous section induces only small effects on these standard errors. For example, if $p = 6$, the standard error will be increased from 0.391σ to a maximum of 0.392σ (for a_1 and a_{12}), while several of the standard errors will actually be reduced, the minimum being 0.389σ (for a_6 and a_7).

A more important factor affecting the precision with which the seasonal constants are estimated is the loss of observations at the beginning and end of the series brought about by taking a moving average. The 12-month centred moving average will lose six observations from each end of the series and if these were available then the standard error of the estimates would be $\sigma\sqrt{11/12(p+1)}$: with $p = 6$ this is 0.362σ. Durbin thus considered applying the adjustments to the entire set of $12(p+1)$ observations and showed that these new estimates have a maximum standard error of 0.364σ with $p = 6$, with no standard error being reduced in size.

4.13 Durbin extended his analysis to the case when the seasonal pattern is changing slowly over time and a linear regression on time is fitted to the deviations from trend for each month, showing that a similar set of adjustments to those outlined above are obtained. He then analyzed more general moving averages of length $2m + 1$, in particular those considered by Kendall (1946) (see **§3.16**), and again obtained analogous results, enabling him to argue that

> the behavior of the series between the first $2m + 1$ and the last $2m + 1$ observations has no effect whatever on the allowance made for trend in the final estimates, and we are forced to conclude, possibly with reluctance, that the faithful manner in which a moving-average estimator appears to follow a trend of arbitrary shape is deceptive and misleading as far as the estimation of seasonal variation is concerned. (Durbin, 1963, page 14)

4.14 This purely theoretical foray into the estimation of seasonal factors was the prelude to a more sustained interest in the practicalities of the seasonal adjustment process as undertaken by official statistical agencies. In 1968 the Central Statistical Office (CSO) of the UK set up a Research and Special Studies Section to investigate methodological problems, to which Durbin was asked to act as an academic consultant on time series problems. One of Durbin's first investigations was the seasonal adjustment of unemployment and this led to the development of a mixed additive-multiplicative model for seasonal adjustment, which resulted some years later in the publication of Durbin and Murphy (1975) and Durbin and Kenny (1979) and the related paper by Kenny and Durbin (1982) on local trend estimation and seasonal adjustment.[2] These papers are both interesting and important because they demonstrate how technical statistical virtuosity and applied statistical experience can be combined to develop robust statistical procedures that can be programmed and thus used by non-specialists to carry out the routine tasks of statistical analysis that are the essential *raison d'être* of government statistical agencies.

4.15 The papers on seasonal adjustment began by contrasting the additive model, now written as

$$y_t = \xi_t + \alpha_t + \varepsilon_t \tag{4.14}$$

where ξ_t is the trend, ε_t is the irregular and α_t is the additive seasonal factor, with both the multiplicative model

$$y_t = \xi_t(1 + \beta_t) + \varepsilon_t \tag{4.15}$$

where β_t is a multiplicative seasonal factor, and the 'mixed' additive-multiplicative model

$$y_t = \xi_t(1 + \beta_t) + \alpha_t + \varepsilon_t = \xi_t + \alpha_t + \beta_t\xi_t + \varepsilon_t \tag{4.16}$$

The additive and multiplicative models are the basis for, respectively, 'difference-from-trend' seasonal adjustment, which uses $y_t - \xi_t$ as the seasonally adjusted series, and 'ratio-to-trend' adjustment, which uses y_t/ξ_t.

Let z_t, a_t and b_t denote estimates of ξ_t, α_t and β_t obtained from a sample of monthly data. Assuming that the seasonal variation is constant from year to year, the fitted mixed model can then be written, with $t = 12(i-1) + j$, $i = 1, 2, \ldots,$ $j = 1, 2, \ldots, 12$, as

$$y_{i,j} = z_{i,j} + a_j + b_j z_{i,j} + e_{i,j} \tag{4.17}$$

where $e_{i,j}$ is the estimated irregular component. The additive and multiplicative factors are constrained by the relations

$$\sum_{j=1}^{12} a_j = \sum_{j=1}^{12} b_j = 0$$

to ensure they do indeed measure departures from the general level of the series. The estimated model (4.17) may be written as

$$\frac{y_{i,j} - a_j}{1 + b_j} = z_{i,j} + \frac{e_{i,j}}{1 + b_j}$$

so that the seasonally adjusted series can be defined as

$$y_{i,j}^{SA} = \frac{y_{i,j} - a_j}{1 + b_j} \tag{4.18}$$

By suppressing either the multiplicative terms b_j or the additive terms a_j the standard forms $y_{i,j} - a_j$ or $y_{i,j}/(1 + b_j)$ for additively and multiplicatively adjusted series are obtained.

4.16 Although (4.18) does not depend explicitly on the trend, this will typically be estimated by a moving average filter of the form

$\hat{z}_t = \sum w_p y_{t+p}$. The first step of the seasonal adjustment method proposed by Durbin and Murphy is to obtain a preliminary estimate of the trend using a specially constructed 21-term filter. This filter was designed to pass a cubic polynomial unchanged while eliminating all additive seasonal waves, and to minimize the amount of the irregular component passed through (which is equivalent to minimizing the variance of the estimated trend). The filter weights are found by solving the following model for the trend value $\sum \alpha_j t^j$ over the 21 terms from $t = -10$ to $+10$

$$x_t = \sum_{j=0}^{3} \alpha_j t^j + \sum_{j=1}^{6} \beta_j \cos(2\pi j/12) + \sum_{j=1}^{6} \gamma_j \sin(2\pi j/12)$$

For $t = 0$ the filter weights are

$$[-0.04769, -0.02535, -0.00301, 0.01933, 0.04167, 0.06401$$
$$0.08634, 0.10868, 0.13102, 0.08333, \underline{0.08333}, \ldots]$$

Given the trend estimate $\hat{z}_{i,j}$, the second step in the Durbin–Murphy method is to estimate (4.17), now expressed as

$$y_{i,j} - \hat{z}_{i,j} = a_j + b_j \hat{z}_{i,j} + e_{i,j}$$

There will be 24 constants in this model so that estimation could be problematic when only short stretches of data are available (the typical length of a model fit is seven years, with the model being refitted over each window of seven calendar years). Durbin and Murphy's solution is to express the constants a_j and b_j in terms of (orthogonal) Fourier components and then to employ a stepwise regression procedure, which typically reduces the number of parameters to between 9 and 12.

In some series the pattern of seasonal variation remains relatively stable over time but the amplitude of the seasonal component changes quite rapidly. Durbin and Murphy thus introduced the concept of a *local amplitude scaling factor*. If $s_{i,j} = a_j + b_j \hat{z}_{i,j}$ is the seasonal component estimated by the stepwise regression, the further regression

$$y_{i,j} - x_{i,j} = d_{i,j} s_{i,j} + e^*_{i,j}$$

is then fitted over a relatively short sample of observations, typically 15 or 25. The estimate of the local amplitude scaling factor $d_{i,j}$ is then used to amend the seasonally adjusted values to

$$y_{i,j}^{SA} = \frac{y_{i,j} - d_{i,j}a_j}{1 + d_{i,j}b_j}$$

The final major step of the procedure is to filter the adjusted series, this time using the 13-term filter suggested by Burman (1965), which has weights

$$[-0.0331, -0.0208, 0.0152, 0.0755, 0.1462, 0.2039, \underline{0.2262}, \ldots]$$

and to run through the stepwise regression and local amplitude scaling factor steps again to obtain a final seasonally adjusted series. In practice, however, extreme values, arising from causes such as strikes and exceptional weather conditions, need to be identified and modified before refitting as necessary.

Further features of the methodology include an extensive testing program to determine the actual form of the model, be it additive, multiplicative or mixed, and further extensions deal with changing seasonal patterns. Extensive details of all these procedures may be found in the papers by Durbin and Murphy and Durbin and Kenny, where several examples using various unemployment series are presented in detail.

4.17 Kenny and Durbin focused on local and current trend estimation as well as on current seasonal adjustment: 'what estimates of trend should be employed for the current and recent months as each new monthly value of the underlying series becomes available(?)' (Kenny and Durbin, 1982, page 1). They were at pains to point out that

> our approach … has been entirely pragmatic and empirical … we have not attempted to set up mathematical models to represent the trend, seasonal and irregular components for the purpose of obtaining formulae by applying some kind of optimality criterion. Instead we have applied the standard *X*-11 trend estimation procedure to observations in the central part of the series to define the trend we wish to estimate; we have considered as possible methods of estimating local trend a wide variety of techniques including those used in current practice, our choice being determined essentially by intuitive and pragmatic considerations, and we have based our recommendations for practical implementation on empirical investigation of the performance of the methods on a variety of real time series. (ibid., page 2)

Kenny and Durbin took the trend to be that given by the Henderson moving average used in the X-11 seasonal adjustment program for

observations in the central part of the time series record, these being assumed to be far enough from the beginning and end of the series such that, even if more observations were added at the beginning or end of the record, the trend value produced by the program would be essentially unaltered.[3] The idea was then to investigate how accurate the first estimate of local trend or local seasonal adjustment, made when a monthly observation first becomes available, could be made to be. Accuracy was measured as the discrepancy between the estimate and the value subsequently given by the Henderson trend or the X-11 program, respectively, after three further years of data has been added. A range of estimation and seasonal adjustment methods were investigated using a wide set of economic and social time series embodying various trend and seasonal characteristics. Two features stood out from the results, extensive details of which may be found in the Kenny and Durbin paper. For local trend estimation, forecasting future values of the series by, in particular, stepwise autoregression, and then using the Henderson trend from X-11 produced by far the most accurate local trend estimates, while for seasonal adjustment the greatest improvements were obtained by using 'current updating', i.e., using the current X-11 seasonal adjustment based on forecasting future values. They summarized their conclusions and recommendations as follows.

> As each new observation is added to the series, forecast the next 12 values by stepwise autoregression … or any other suitable method. Then seasonally adjust the augmented series by *X*-11 in the current updating mode discarding output values corresponding to future time points. Take as an estimate of current and recent trends the value given by the Henderson moving average in the *X*-11 program.
>
> …
>
> We recommend, therefore, that each observation is seasonally adjusted when it first enters the series, is revised 1 month later, and thereafter is revised as at present at the end of each calendar year. (*ibid.*, page 24)

Kenny and Durbin expressed a preference for using stepwise autoregression for forecasting because 'our view is that many applied workers would prefer to use a forecasting method based on a simple regression formula to one derived from a more sophisticated approach provided there is relatively little difference in overall performance. The completely mechanical nature of the stepwise method is also an advantage' (ibid., page 25). They thus termed their procedure '*X*-11

Stepwise' and recommended its adoption by the Government Statistical Service.

Testing the constancy of regressions over time

4.18 Durbin's involvement with the CSO also bore fruit in another area, that of testing the constancy of regression relationships over time using recursive residuals and rolling regressions. A preliminary account of the underlying theory was given in Brown and Durbin (1968) but the main results, with applications, were presented in Brown, Durbin and Evans (1975), henceforth denoted as BDE. Such was the practical usefulness of the techniques proposed by BDE that they have since been programmed into many time series and econometric packages.

4.19 The basic regression model considered by BDE is

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_t + u_t \qquad t = 1, \ldots, T \tag{4.19}$$

where at time t, y_t is the observation on the dependent variable and $\mathbf{x}_t$ is a column vector of observations on k regressors. The first regressor, x_{1t}, is taken to be equal to unity for all t if the model contains a constant while the other regressors are assumed to be non-stochastic, thus ruling out autoregressive models, for example.[4] The column vector of regression coefficients, $\boldsymbol{\beta}_t$, is allowed to vary over time and it is assumed that the error terms, u_t, are independent and normally distributed with zero means but possibly changing variances σ_t^2. The hypothesis of constancy over time, denoted H_0, is

$$\begin{aligned} \boldsymbol{\beta}_1 &= \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T = \boldsymbol{\beta} \\ \sigma_1^2 &= \sigma_2^2 = \cdots = \sigma_T^2 = \sigma^2 \end{aligned}$$

BDE noted that, in order to assess departures from H_0, a natural thing to do would be to look at the regression residuals from (4.19). Unfortunately, plots against time of the ordinary least-squares residuals $e_t = y_t - \mathbf{x}_t'\mathbf{b}$, where $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $\mathbf{X}' = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ and $\mathbf{y}' = [y_1, \ldots, y_T]$, tend not to be very sensitive indicators of small or gradual changes in the regression coefficients.

By analogy to industrial quality control problems, BDE therefore considered plotting the cumulative sum (cusum) of the residuals, $Z_r = \hat{\sigma}^{-1}\sum_{t=1}^{r} e_t$, where the cumulative sum has been divided by the estimated standard deviation $\hat{\sigma} = \left(\sum_{t=1}^{T} e_t^2/(T-k)\right)^{\frac{1}{2}}$ to eliminate the irrelevant

scale factor. Unfortunately, there does not seem to be any easy way of assessing the significance of the departure of Z_r from $E(Z_r) = 0$ and this also goes for the cumulative sum of squared residuals (cusum of squares), $\hat{\sigma}^{-2} \sum_{t=1}^{r} e_r^2$.

4.20 BDE thus preferred to make the transformation to *recursive residuals*, which may be specified in the following way. Define $\mathbf{X}_r'$ and $\mathbf{y}_r'$ to be the first r columns of $\mathbf{X}'$ and $\mathbf{y}'$ and let $\mathbf{b}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{y}_r$ be the least-squares estimate of $\boldsymbol{\beta}_r$ based on the first r observations. The recursive residuals are then defined as

$$w_r = \frac{y_r - \mathbf{x}_r'\mathbf{b}_{r-1}}{\sqrt{(1 + \mathbf{x}_r'(\mathbf{X}_{r-1}'\mathbf{X}_{r-1})^{-1}\mathbf{x}_r)}} \qquad r = k+1, \ldots, T \tag{4.20}$$

BDE proved that, under H_0, the $w_{k+1}, \ldots, w_T$ are independent and distributed as $N(0, \sigma^2)$.

Implicit in formula (4.20) is the assumption that the matrix $\mathbf{X}_r'\mathbf{X}_r$ is non-singular. This will not be the case if the regression model contains a constant and one of the regressor variables is itself constant for the first $r_1 \geq k$ observations, a situation that arises quite frequently in practice. BDE provided a method of computing the w_r in these circumstances which involves deleting the initially constant regressor from the recursions up to r_1 and then bringing it in when it has changed.

4.21 If $\boldsymbol{\beta}_t$ is constant up to time $t = t_0$ and then changes, the w_rs will have zero means up to t_0 but will have non-zero means subsequently, which suggests that plots aimed at revealing departures of the means of the recursive residuals from zero through time could be worth examining. BDE therefore proposed plotting the cusum quantity

$$W_r = \frac{1}{\hat{\sigma}} \sum_{j=k+1}^{r} w_j$$

against r for $r = k+1, \ldots, T$. Here $\hat{\sigma}$ is the residual standard deviation determined by $\hat{\sigma}^2 = S_T/(T-k)$, where $S_r = S_{r-1} + w_r^2$, $r = k+1, \ldots, T$. Under H_0, $W_{k+1}, \ldots, W_r$ is a sequence of approximately normal variables such that W_r has mean zero and variance $r - k$, with the covariance between W_r and W_s being $\min(r, s) - k$. BDE then showed that the significance of any departure of W_r from zero may be assessed from the cusum plot by reference to a pair of significance lines found by connecting the points $\{k, \pm a\sqrt{T-k}\}$ and $\{T, \pm 3a\sqrt{T-k}\}$. For a 5% test, a should

be set at 0.948, while for a 1% test $a = 1.143$. A movement of W_r outside of these lines represents a rejection of H_0 and evidence of some form of parameter instability.

4.22 BDE suggested a second test that uses the squared recursive residuals, w_r^2, to define the statistic

$$W_r' = \left(\sum_{j=k+1}^{r} w_j^2 \right) \Big/ \left(\sum_{j=k+1}^{T} w_j^2 \right) = \frac{S_r}{S_T}$$

This cusum of squares test is a useful complement to the cusum test, particularly when the departure from constancy of the $\boldsymbol{\beta}_t$s is haphazard rather than systematic. On H_0 W_r' has a beta distribution with mean $(r-k)/(T-k)$, so BDE recommended drawing a pair of parallel straight lines, given by $\pm c_0 + (r-k)/(T-k)$, on the cusum of squares plot. For a given significance level, c_0 may be obtained from Table 1 of Durbin (1969) and BDE provided details of how to do this. Again, H_0 is rejected if W_r' cuts these lines, although BDE preferred to regard the lines constructed in this way as 'yardsticks' against which to assess the observed sample path rather than providing formal tests of significance.

4.23 BDE also suggested plotting the components of $\mathbf{b}_r$ against time to try to identify the source of any departure from constancy indicated by the cusum tests, perhaps running the analysis backwards through time as well as forwards: this has since become known as recursive regression and the $\mathbf{b}_r$ recursive coefficients.

They also considered *moving regressions* (also known as rolling regressions), where the regression is fitted to short segments of n observations which are then 'rolled' through the sample by dropping the first observation of the segment and adding a new observation at its end. Analogues of the recursive formulae are provided by BDE to facilitate the estimation of moving regression coefficients.

4.24 As an example of the BDE procedures, Figure 4.3 shows the observations on a variable defined as $y_t = 1 + 0.5t + u_{1t}$, $u_{1t} \sim N(0, 9)$, for $t = 1, \ldots, 50$, and $y_t = 1.2 + 0.8t + u_{2t}$, $u_{2t} \sim N(0, 25)$, for $t = 51, \ldots, 100$, along with the ordinary least squares fitted line $\hat{y}_t = -5.98 + 0.87t$ on the assumption of parameter constancy. As there are intercept, slope and error variance shifts midway through the sample, the fit, although accompanied by an R^2 of 0.95, is clearly unsatisfactory, having a large

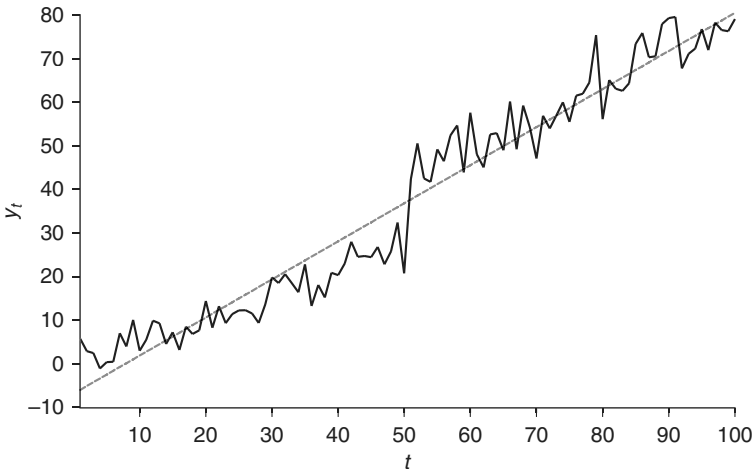


Figure 4.3 Linear trend fitted to y_t

negative intercept, a residual standard deviation of $\hat{\sigma} = 5.86$ and displaying evidence of residual autocorrelation (the Durbin-Watson statistic is 1.09).

Figure 4.4 shows cusum and cusum of squares plots accompanied by 5% significance lines and both indicate a clear shift in the regression at $t = 50$ (here $c_0 = 0.1795$). Figure 4.5 shows the plots of the recursive intercept and slope estimates, which again indicate shifts in the coefficients at the mid-sample point. Since the magnitudes of the shifts are reasonably small, the recursive estimates do not alter as markedly or abruptly as the cusum plots, although all plots show that a shift has taken place by $t = 60$.

Structural models and the Kalman filter

4.25 Although the idea of sequentially updating or recursively estimating the parameters of a model has a history stretching back to Gauss in the 1820s, it was only rediscovered in the middle of the twentieth century by Plackett (1950).[5] 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

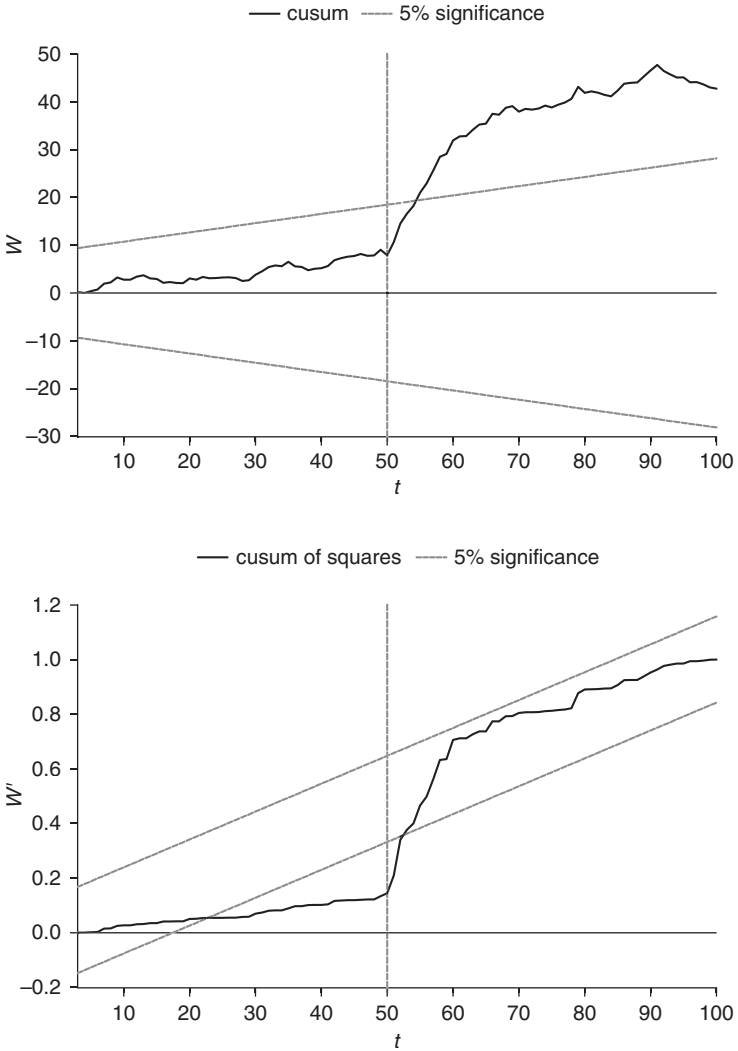


Figure 4.4 Cusum and cusum of squares plots

across a range of different disciplines, with the algorithm being referred to universally as the *Kalman filter*.[6]

It is generally accepted that the reasons for this delay in the takeup of Kalman's procedure by the time series community were twofold. First, the original paper and its continuous time counterpart (Kalman and

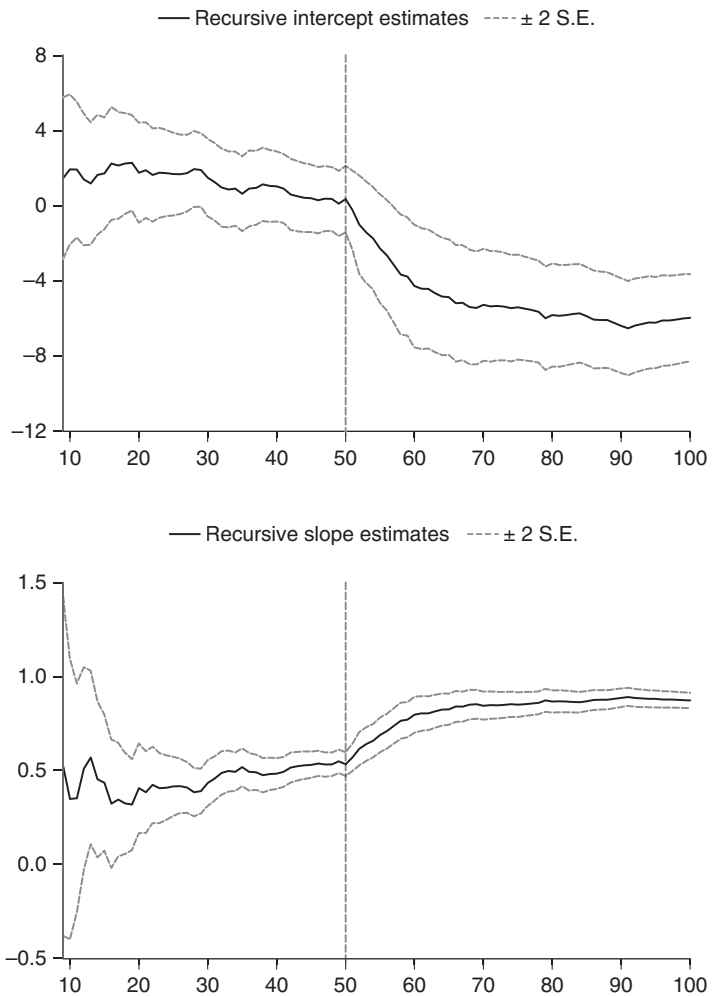


Figure 4.5 Recursive intercept and slope estimates

Bucy, 1961) were written for an engineering audience and so used a language, notation and style that was alien to statisticians. Second, the original setup 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, 1976, and Duncan and Horn, 1972).

4.26 Durbin's first discussion of the Kalman filter was in the context of forecasting models in his contribution to the *Journal of the Royal Statistical Society, Series A* issue commemorating the 150th Anniversary of the RSS (Durbin, 1984). Here he analyzed the non-stationary structural model proposed by Theil and Wage (1964),

$$y_t = \mu_t + \varepsilon_t \qquad \mu_t = \mu_{t-1} + \delta_{t-1} + \eta_t \qquad \delta_t = \delta_{t-1} + \zeta_t \tag{4.21}$$

in which the observed value y_t is made up of a trend component, μ_t, whose level and slope are both determined by uncorrelated random walks, together with a disturbance ε_t. Theil and Wage showed that optimal one step-ahead forecasts, in the minimum mean square error sense, were equivalent to those obtained by the first two recursions of the Holt–Winters extension of exponential smoothing (see, for example, Mills, 2011a, §§11.16–11.20, for details and historical perspective). Durbin pointed out that (4.21) was a simple special case of the Kalman filter-state space model

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \qquad t = 1, 2, \ldots, T \tag{4.22}$$

$$\boldsymbol{\beta}_t = \mathbf{T}_t\boldsymbol{\beta}_{t-1} + \mathbf{R}_t\boldsymbol{\eta}_t \qquad t = 1, 2, \ldots, T \tag{4.23}$$

where $\mathbf{y}_t$ is an n-dimensional vector of observed time series and $\boldsymbol{\beta}_t$ is an m-dimensional vector of state variables. $\mathbf{X}_t$, $\mathbf{T}_t$ and $\mathbf{R}_t$ are non-stochastic matrices, and $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are n-dimensional and g-dimensional vectors of uncorrelated and non-autocorrelated unobserved disturbances with variance matrices $\mathbf{H}_t$ and $\mathbf{Q}_t$ respectively, typically assumed to be normally distributed. Equation (4.22) is often referred to as the *measurement* or *observation* equation and (4.23) as the *transition* or *state* equation. The model (4.21) is obtained by setting

$$\mathbf{y}_t = y_t \qquad \mathbf{X}_t = [1 \quad 0] \qquad \boldsymbol{\beta}_t = [\mu_t \quad \delta_t]'$$

$$\boldsymbol{\varepsilon}_t = \varepsilon_t \qquad \mathbf{T}_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{R}_t = \mathbf{I}_2 \qquad \boldsymbol{\eta}_t = [\eta_t \quad \zeta_t]'$$

$$\mathbf{H}_t = \sigma_\varepsilon^2 \qquad \mathbf{Q}_t = \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}$$

Durbin was rather enthusiastic about this approach, stating that the

> beauty of Kalman's formulation is that there is a routine mechanical way of updating estimates, whether of parameters or of past, present or future values of some unknown vector, each time a new observational vector arrives; moreover, the updating is essentially straightforward on a modern computer. (Durbin, 1984, page 169)

This updating could be carried out in the following way. The minimum mean square linear estimator $\mathbf{b}_t$ of the state $\boldsymbol{\beta}_t$, based on all the data, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t$ through to time t, is given by the set of recursive *updating* equations

$$\mathbf{b}_t = \mathbf{b}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{X}_t'\mathbf{D}_t^{-1}(\mathbf{y}_t - \mathbf{X}_t\mathbf{b}_{t|t-1})$$

The variance matrix associated with $\mathbf{b}_t$ is given by the set of recursive equations

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{X}_{t-1}'\mathbf{D}_t^{-1}\mathbf{X}_t\mathbf{P}_{t|t-1}$$

In these recursions

$$\begin{array}{lll} \mathbf{b}_{t|t-1} = \mathbf{T}_t\mathbf{b}_{t-1} & \mathbf{P}_{t|t-1} = \mathbf{T}_t\mathbf{P}_{t-1}\mathbf{T}_t' + \mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t' & t = 2, \ldots, T \\ \mathbf{D}_t = \mathbf{H}_t + \mathbf{X}_t\mathbf{P}_{t|t-1}\mathbf{X}_t' & & t = 1, 2, \ldots, T \end{array}$$

The one step-ahead forecast of $\mathbf{y}_t$ is then $\mathbf{y}_t(1) = \mathbf{X}_{t+1}\mathbf{b}_{t+1|t}$. The quantities $\mathbf{y}_t - \mathbf{X}_t\mathbf{b}_{t|t-1}$ are called innovations and are uncorrelated, and independent if normality of the disturbances can be assumed. This property can be exploited to compute the likelihood function efficiently and can thus be used to obtain exact ML estimators of parameters in very general dynamic models, as shown by Harvey (1981).

4.27 Durbin (1984) claimed that the first explicit use of the Kalman filter for statistical forecasting was Jones (1966), who employed it in the context of multivariate exponential smoothing, with Enns et al. (1982) later extending the procedure to obtain exact ML estimates of the model parameters. A general treatment of Kalman forecasting was provided by Harrison and Stevens (1976), ostensibly from a Bayesian perspective although, as Durbin pointed out, their model can easily be treated in a non-Bayesian way.

Durbin's LSE colleague, Andrew Harvey, had contemporaneously given a comprehensive review of statistical forecasting from a Kalman filter standpoint, emphasizing its flexibility and suitability for structural

modelling (Harvey, 1984). Durbin agreed: 'Harvey's overall conclusion, with which I concur, is that the Kalman formulation offers considerable advantages to the forecaster because of its flexibility and comprehensive scope' (Durbin, 1984, page 170). He also pointed out the potential of the approach for analyzing a variety of issues in the decomposition of time series. 'My own belief is that because of its capacity to cope easily with such things as data revisions, calendar variations, extreme values and current updating as well as appropriateness of its structural form, the Kalman model should provide a preferable basis for a modern treatment of the decomposition problem' (ibid., page 170).

4.28 This enthusiasm for structural models analyzed within the state space form and estimated using the Kalman filter was quickly put to practical use by Durbin and Harvey in a project commissioned by the UK Department of Transport on the effect of seat belt legislation on British road casualties.[7] While a full description of the findings of the investigation was given in Durbin and Harvey (1985), Harvey and Durbin (1986) focused on providing an opportunity for public discussion of their results on the effects of the seat belt law on road casualties and on inviting a technical debate on the methodology they had used.[8]

The model fitted by Harvey and Durbin was of the form (4.21) but extended to include a seasonal component, exogenous explanatory variables and an intervention variable in the measurement equation

$$y_t = \mu_t + \gamma_t + \sum_{j=1}^{k} \delta_j x_{jt} + \lambda w_t + \varepsilon_t \tag{4.24}$$

The seasonal component is modelled by the trigonometric function

$$\gamma_t = \sum_{j=1}^{s/2} \gamma_{jt}$$

With s even (typically 4 or 12) and $\vartheta_j = 2\pi j/s$, a fixed seasonal pattern could be modelled by setting (on noting that $\sin \vartheta_{s/2} t = \sin \pi t = 0$)

$$\begin{aligned} \gamma_{jt} &= \gamma_j \cos \vartheta_j t + \gamma_j^* \sin \vartheta_j t, \qquad j = 1, \ldots, \tfrac{1}{2}s - 1 \\ \gamma_{s/2,t} &= \gamma_{s/2} \cos \vartheta_{s/2} t \end{aligned} \tag{4.25}$$

It is easily shown that this is equivalent to including a set of seasonal dummies as regressors with the constraint that the seasonal effects sum

to zero. The seasonal pattern can be allowed to evolve over time by replacing (4.25) with

$$\begin{aligned}\gamma_{jt} &= \gamma_{j,t-1}\cos\vartheta_j t + \gamma^*_{j,t-1}\sin\vartheta_j t + \omega_{jt}\\ \gamma^*_{jt} &= \gamma^*_{j,t-1}\cos\vartheta_j t - \gamma_{j,t-1}\sin\vartheta_j t + \omega^*_{jt}\\ \gamma_{s/2,t} &= \gamma_{s/2,t-1}\cos\vartheta_{s/2}t + \omega_{s/2,t}\end{aligned} \qquad j = 1,\ldots,\tfrac{1}{2}s-1$$

where ω_{jt} and ω^*_{jt} are zero mean non-autocorrelated errors uncorrelated with each other and having common variance σ^2_ω. The larger is σ^2_ω the greater the evolution of the seasonal component through time.

The x_{jt} are explanatory variables while w_t is an intervention variable (see **§8.10–8.13**) defined by Harvey and Durbin to be the simple dummy variable

$$w_t = \begin{cases} 0 & t < \tau \\ 1 & t \geq \tau \end{cases}$$

4.29 The model (4.24) was fitted to monthly series on numbers killed or killed and seriously injured in various road user categories for the period January 1969 to December 1984. Typically two explanatory variables were included, an appropriate traffic index and the real price of petrol, and these, along with the dependent variable, were logarithmically transformed. The intervention variable was defined by setting $\tau =$ January 1983 and estimates of the intervention effect of the seat belt law, $100(1-\exp\lambda)$, measuring the percentage change in the monthly level of casualty rates after January 1983 over and above that predicted by the model without the intervention, were obtained for the various road user categories. These ranged from −18 per cent for drivers killed to −30.3 per cent for front seat passengers killed and seriously injured, but other categories of road user, such as rear seat passengers, pedestrians and cyclists, saw percentage increases (or insignificant declines) in casualty rates. Drawing firm conclusions from these results proved rather tricky, however, although Harvey and Durbin felt able to 'conclude that, whether we concentrate on those directly affected or also include those indirectly affected, there have been substantial net reductions in numbers [killed and seriously injured] and numbers killed due to the introduction of the seat belt law' (ibid., page 208).

4.30 The above analysis was restricted to casualties to car occupants, pedestrians and cyclists because, although the series are obviously 'counts', and hence can only be positive integers, the range and extent

of the data was sufficient for Harvey and Durbin to feel comfortable using an approximating linear, Gaussian state space model. This was not the case, however, for the light goods vehicles (van) drivers fatality series, whose numbers, according to Durbin and Koopman (2000), were regarded as being too small to justify the use of the linear Gaussian model. State space models with non-Gaussian observations were thus analyzed in Durbin and Koopman (1997, 2000). The linear Gaussian model of (4.22) and (4.23),

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \qquad & \boldsymbol{\varepsilon}_t &\sim N(\mathbf{0}, \mathbf{H}_t) \\ \boldsymbol{\beta}_t &= \mathbf{T}_t\boldsymbol{\beta}_{t-1} + \mathbf{R}_t\boldsymbol{\eta}_t \qquad & \boldsymbol{\eta}_t &\sim N(\mathbf{0}, \mathbf{Q}_t) \end{aligned}$$

where it is now assumed that the elements of the matrices $\mathbf{H}_t$, $\mathbf{Q}_t$, $\mathbf{X}_t$ and $\mathbf{T}_t$ may be functions of an unknown parameter vector $\boldsymbol{\psi}$, can be extended to the non-Gaussian case by writing the observation equation as

$$p(\mathbf{y}_t|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_t, \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}) = p(\mathbf{y}_t|\mathbf{X}_t\boldsymbol{\beta}_t)$$

and the transition equation as

$$\boldsymbol{\beta}_t = \mathbf{T}_t\boldsymbol{\beta}_{t-1} + \mathbf{R}_t\boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim p(\boldsymbol{\eta}_t)$$

where the notation $p(\,\cdot\,)$, $p(\cdot,\cdot)$ and $p(\,\cdot\,|\,\cdot\,)$ is used to denote generic marginal, joint and conditional densities respectively. Durbin and Koopman define $\boldsymbol{\theta}_t = \mathbf{X}_t\boldsymbol{\beta}_t$, calling it the signal, and pay particular attention to two special cases:

(a) observations which come from exponential family distributions with densities of the form

$$p(\mathbf{y}_t|\boldsymbol{\theta}_t) = \exp\{\mathbf{y}_t'\boldsymbol{\theta}_t - b_t(\boldsymbol{\theta}_t) + c_t(\mathbf{y}_t)\}$$

where $b_t(\boldsymbol{\theta}_t)$ is twice differentiable and $c_t(\mathbf{y}_t)$ is a function of $\mathbf{y}_t$ only;

(b) observations generated by the relationship

$$\mathbf{y}_t = \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t \sim p(\boldsymbol{\varepsilon}_t)$$

where the $\boldsymbol{\varepsilon}_t$ are non-Gaussian and serially independent.

While such models had previously been analyzed by several methods, Durbin and Koopman argued that all of these involved approximation

errors of unknown magnitude, in contrast to their own approach, where the only errors were due to simulation and thus their size could be measured and made as small as desired. Essentially, their approach is to consider the stacked vectors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T')'$ and $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$ and to attempt to estimate by simulation the conditional mean

$$\bar{x} = E[x(\boldsymbol{\beta})|\mathbf{y}]$$

of an arbitrary function $x(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ given the observation vector $\mathbf{y}$.[9] Denoting Gaussian marginal, joint and conditional densities generically as $g(\cdot)$, $g(\cdot,\cdot)$ and $g(\cdot\,|\,\cdot)$ respectively, Durbin and Koopman showed that this conditional mean can be written as

$$\bar{x} = \int x(\boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{y}) d\boldsymbol{\beta} = \int x(\boldsymbol{\beta}) \frac{p(\boldsymbol{\beta}|\mathbf{y})}{g(\boldsymbol{\beta}|\mathbf{y})} g(\boldsymbol{\beta}|\mathbf{y}) d\boldsymbol{\beta} = E_g \left[x(\boldsymbol{\beta}) \frac{p(\boldsymbol{\beta}|\mathbf{y})}{g(\boldsymbol{\beta}|\mathbf{y})} \right] \tag{4.26}$$

where E_g denotes expectation with respect to the Gaussian importance density $g(\boldsymbol{\beta}|\mathbf{y})$, which is chosen to resemble $p(\boldsymbol{\beta}|\mathbf{y})$ as closely as possible. Since $g(\boldsymbol{\beta}|\mathbf{y})$ and $p(\boldsymbol{\beta}|\mathbf{y})$ are typically algebraically complicated, whereas the corresponding joint densities $g(\boldsymbol{\beta}, \mathbf{y})$ and $p(\boldsymbol{\beta}, \mathbf{y})$ tend to be relatively straightforward, Durbin and Koopman rewrite (4.26) as

$$\bar{x} = \frac{E_g[x(\boldsymbol{\beta}) w(\boldsymbol{\beta}, \mathbf{y})]}{E_g[w(\boldsymbol{\beta}, \mathbf{y})]} \qquad w(\boldsymbol{\beta}, \mathbf{y}) = \frac{p(\boldsymbol{\beta}, \mathbf{y})}{g(\boldsymbol{\beta}, \mathbf{y})} \tag{4.27}$$

In principle, a Monte Carlo estimate $\hat{x}$ of $\bar{x}$ could be obtained by taking a series of independent draws $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}$ from $g(\boldsymbol{\beta}, \mathbf{y})$ and computing

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i} \qquad x_i = x(\boldsymbol{\beta}^{(i)}), \qquad w_i = w(\boldsymbol{\beta}^{(i)}, \mathbf{y})$$

Since the draws are independent $\hat{x}$ will converge to $\bar{x}$ as $N \to \infty$.

4.31 This simple estimator is, however, numerically inefficient, so Durbin and Koopman refined it in several ways. Denoting $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_T')'$, (4.27) is rewritten as

$$\bar{x} = \frac{E_g[x(\boldsymbol{\eta}) w(\boldsymbol{\eta}, \mathbf{y})]}{E_g[w(\boldsymbol{\eta}, \mathbf{y})]} \qquad w(\boldsymbol{\eta}, \mathbf{y}) = \frac{p(\boldsymbol{\eta}, \mathbf{y})}{g(\boldsymbol{\eta}, \mathbf{y})}$$

where the relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ is determined from $\boldsymbol{\beta}_t = \mathbf{T}_t \boldsymbol{\beta}_{t-1} + \mathbf{R}_t \boldsymbol{\eta}_t$ and E_g now denotes expectations with respect to the importance

density $g(\boldsymbol{\eta}, \mathbf{y})$. The simulations are then based on random draws of $\boldsymbol{\eta}$ from this importance density using the very efficient simulation smoother of de Jong and Shephard (1995). The efficiency is increased by the use of antithetic variables, which are variables constructed from $\boldsymbol{\eta}$ that have negative covariance and so both increase the size of the simulation sample and decrease the sampling variance at no extra computing cost.

Durbin and Koopman also provided an importance sampling ML estimator of $\boldsymbol{\psi}$ and several examples of the overall method. Assuming that van drivers killed can be modeled by the Poisson density with mean $\exp(\theta_t)$ (the Poisson, of course, being a special case of the exponential)

$$p(y_t|\theta_t) = \exp(\theta_t' y_t - \exp(\theta_t) - \log(y_t!))$$

with the signal θ_t generated by

$$\theta_t = \mu_t + \gamma_t + \lambda w_t$$

where $\mu_t = \mu_{t-1} + \eta_t$, γ_t is the evolving seasonal component and w_t is the seatbelt legislation dummy, Durbin and Koopman found that both classical and Bayesian analysis produced an estimate of around 24 per cent for the reduction in deaths after the introduction of the legislation.

4.32 Durbin's long-standing interest in official statistics continued in Durbin and Quenneville (1997), in which a state space approach to the benchmarking of official statistics was proposed, and Durbin (2000), which was a more general survey of the usefulness of state space models to official statistics. These papers resulted from visits to Statistics Canada and Statistics New Zealand, respectively, as Durbin continued to hold visiting posts overseas well into his 70s.

Durbin's research on filtering and smoothing algorithms for state space models was extended with Siem Jan Koopman in Durbin and Koopman (2002) and Koopman and Durbin (2000, 2003). They also published a book on state space modelling (Durbin and Koopman, 2001) while 2004 saw the publication of the proceedings of a conference held in Durbin's honor to which he contributed the introduction (Durbin, 2004), which appears to be his final publication.

5

Jenkins: Inference in Autoregressive Models and the Development of Spectral Analysis

Gwilym Jenkins

5.1 Gwilym Meirion Jenkins was born in 1933 in Gowerton, Swansea, in the principality of Wales and was very much a Welshman, speaking only Welsh until the age of seven and even in later life often thinking in the language. He obtained a first class honours degree in Mathematics from UCL in 1953 and remained there to complete a doctorate, in the area of time series analysis, in 1956. He then spent two years as a junior fellow at the Royal Aircraft Establishment (RAE) at Farnborough helping to design aircraft before being appointed to a lectureship in statistics at Imperial College. During 1959–60 Jenkins was invited to Stanford and Princeton as a visiting professor, where he began his long collaboration with George Box (see Chapter 6), and he returned to the United States in 1964–5 as a visiting professor at the University of Wisconsin, where the collaboration continued. It was during this visit that the seriousness of his medical condition was first realized: for the next 17 years he would fight a slowly losing battle against Hodgkin's disease.

Although promoted to Reader at Imperial in 1964, Jenkins, through his extensive consultancy work, wanted to place his statistical expertise within a wider, systems, context, and consequently took up the offer of a chair in Systems Engineering at Lancaster in 1965. In 1969 he founded, and became co-ordinating editor of, the *Journal of Systems Engineering*, contemporaneously writing an introductory book on the subject (Jenkins and Youle, 1971). His consultancy work expanded with the help of industrial and governmental projects undertaken by students on the MSc degree that he had devised and this led Jenkins to found and become Managing Director of the consultancy enterprise ISCOL (International Systems Corporation of Lancaster). Although initially wholly owned by

Lancaster University, this proved to be an unworkable arrangement and Jenkins left in 1974 to set up his own consultancy company, at the same time taking up a visiting professorial appointment at the London Business School. The success and variety of these consultancy projects eventually led directly to two volumes of case studies, Jenkins (1979) and Jenkins and McLeod (1982), and indirectly to the related article by Jenkins and Alavi (1981).

All the while, however, his health continued to deteriorate and Jenkins eventually succumbed to Hodgkin's lymphoma on 10 July 1982 at the tragically young age of 49. Further biographical details, particularly of his collaboration with George Box, may be found in Box (1983a, 1983b) and also see De Groot (1987) and Pêna (2001) for additional reflections by Box.

Inference in autoregressive models

5.2 Jenkins' first published papers appeared in *Biometrika* in 1954 which, given that he had only graduated the previous year, certainly demonstrated Jenkins' tremendous statistical precocity and, as we shall see, his extraordinary technical virtuosity. Jenkins (1954a) considered testing the significance of ϕ in the Markov AR(1) model

$$x_t = \phi x_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim iidN(0, \sigma^2) \tag{5.1}$$

by using an inverse sine transformation of the circular serial correlation coefficient (4.1) calculated from a sample of size T, which we now denote r for convenience (note that it is thus assumed that the theoretical mean of the series is zero):

$$r = \frac{x_1x_2 + x_2x_3 + \cdots + x_Tx_1}{x_1^2 + x_2^2 + \cdots + x_T^2} \tag{5.2}$$

Leipnik (1947) had derived the distribution of r as

$$p(r) = \kappa_1(1 - r^2)^{\frac{1}{2}(T-1)}(1 + \phi^2 - 2\phi r)^{-\frac{1}{2}T}$$

where κ_1 is a ratio of gamma functions. The mean and variance of r are then given by

$$E(r) = \alpha = \frac{\phi T}{T+2}$$

and

$$\sigma_r^2 = \frac{1}{T+2}\left(1 - \frac{\phi^2 T(T+1)}{(T+2)(T+4)}\right)$$
$$= \frac{1}{T+2}(1-\lambda\alpha^2), \qquad \lambda = \frac{(T+2)(T+1)}{T(T+4)}$$

Since $\lambda \approx 1$,

$$\sigma_r^2 \approx \frac{1}{T+2}(1-\alpha^2)$$

Jenkins' approach was to seek a transformation $y = f(r)$ such that the variance of y is approximately independent of ϕ and he showed that the inverse sine transformation $y = \sin^{-1} r$ would accomplish this. On defining $\eta = \sin^{-1}\phi$, Jenkins then obtained the distribution of $z = y - \eta$, which has mean and variance given by, respectively,

$$E(z) = \frac{\beta(\phi)}{T} + \frac{\gamma(\phi)}{T^2} + O(T^{-2})$$

and

$$\sigma_z^2 = \frac{1}{T} + \frac{\delta(\phi)}{T^2} + O(T^{-2})$$

where

$$\beta(\phi) = -\frac{3}{2}\frac{\phi}{(1-\phi^2)^{\frac{1}{2}}} \qquad \gamma(\phi) = \frac{1}{8}\frac{\phi}{(1-\phi^2)^{\frac{1}{2}}}(17-2\phi^2)$$
$$\delta(\phi) = -\frac{1}{2T^2}\frac{(2-5\phi^2)}{(1-\phi^2)}$$

The variance is very stable, altering from $1/T - 1/T^2 \approx 1/(T+1)$ to $1/T - 1/(2T^2) \approx 1/(T+2)$ in the range $0 < \phi < 0.5$. Jenkins thus suggested that $z = y - \eta$ may be taken to be approximately normally distributed with zero mean and variance $1/(T+1)$ provided that ϕ was small and T reasonably large. This led him to define approximate confidence intervals for ϕ of the form

$$\phi = r\cos\frac{Z_\alpha}{\sqrt{T+1}} \pm (1-r^2)^{\frac{1}{2}}\sin\frac{Z_\alpha}{\sqrt{T+1}}$$

or, since $\cos Z_\alpha/\sqrt{T+1} \approx 1$ and $\sin Z_\alpha/\sqrt{T+1} \approx Z_\alpha/\sqrt{T+1}$ for T large enough,

$$\phi = r \pm \frac{Z_\alpha(1-r^2)^{\frac{1}{2}}}{\sqrt{T+1}}$$

where Z_α is the α percentage point of the standard normal distribution. Thus, if $r = 0.4$ is calculated from a sample of $T = 35$, a 95% confidence interval for the autoregressive coefficient is given approximately by $\phi = 0.4 \pm 0.3$.

5.3 Jenkins (1954b) considered the 'Yule scheme', or AR(2) model, and showed that the partial correlation $v_2 = (r_2 - r_1^2)/(1 - r_1^2)$ calculated from the lag 1 and lag 2 circular serial correlations $r_1 = r$ and r_2, defined analogously to (5.2), has a 'smoothed' distribution of the form

$$p(v_2) = \kappa_2(1 - v_2^2)^{\frac{1}{2}(T-2)}(1 - v_2) \tag{5.3}$$

The term 'smoothed' reflects the assumption made by Jenkins that the exact characteristic function generating the moments of v_2 may be replaced by a smoothed function that produces moments that are correct up to $O(T)$. Jenkins (1956) then showed that the distribution (5.3) also holds for v_4, while for v_1 and v_3 the distribution is

$$p(v_k) = \kappa_3(1 - v_k^2)^{\frac{1}{2}(T-1)} \qquad k = 1, 3 \tag{5.4}$$

with κ_2 and κ_3 again being ratios of gamma functions. Daniels (1956), using the more elegant method of saddlepoint approximation, subsequently generalized this result to all k, with (5.3) holding for even k and (5.4) for odd k.

Jenkins also considered the modifications required when the circular definition (5.2) was corrected for the sample mean. The distribution of $\bar{v}_2$, to use an obvious notation, is

$$p(\bar{v}_2) = \kappa_4(1 - \bar{v}_2^2)^{\frac{1}{2}(T-3)}\left((1 - \bar{v}_2)^2 - \frac{2}{T}(1 - \bar{v}_2^2)\right)$$

with κ_4^{-1} being a linear combination of beta functions. The first two moments of this distribution are

$$E(\bar{v}_2) = -\frac{2}{T-1} \qquad E(\bar{v}_2^2) = \frac{1}{T+2}\left(\frac{T+3}{T-1}\right)$$

and Jenkins (1956) provided general expressions for higher moments. This paper also presented results for the moments of $\bar{v}_3$ and also results for non-null distributions for general autoregressive schemes. Unfortunately, and in stark contrast to his later work, no empirical examples of testing procedures were provided by Jenkins, presumably because of the highly complicated and abstract nature of the results that he had obtained, and he never returned to the topic again.

Spectral analysis in the mid-1950s

5.4 During his time at the RAE, Jenkins produced one of the first reviews of spectral analysis, in collaboration with Maurice Priestley, who subsequently became a world leader in the subject (see, for example, Priestley, 1981). Jenkins and Priestley (1957) began by considering the trend-free, and hence stationary, discrete time series x_t, $t = 1, 2, \ldots, T$, taken to have zero mean for simplicity, which may be decomposed, following Wold (1938) and Rudra (1955), as

$$\begin{aligned} x_t &= \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{j=0}^{q} \beta_j \varepsilon_{t-j} + \sum_{r=1}^{u} \sum_{s=1}^{n} \gamma_{rs} \cos(2\pi ts/n + \delta_{rs}) \\ &= u_t + v_t + w_t \end{aligned} \tag{5.5}$$

where the ε_t's are uncorrelated random variables and β_0 is normalized to unity. The components u_t and v_t are the familiar autoregressive and moving average processes while w_t is termed a linear cyclical process. The *periodogram* of x_t is defined as

$$I_T(\omega_i) = A_i^2 + B_i^2 = 2 \sum_{s=-(T-1)}^{T-1} (1 - |s|/T) C_s \cos \omega_i s$$

where

$$A_i = \left(\frac{2}{T}\right)^{\frac{1}{2}} \sum_{t=1}^{T} x_t \cos \omega_i t \qquad B_i = \left(\frac{2}{T}\right)^{\frac{1}{2}} \sum_{t=1}^{T} x_t \sin \omega_i t$$

and the

$$C_s = \frac{1}{T - |s|} \sum_{t=1}^{T-|s|} x_t x_{t+|s|}$$

are the serial covariances. The *autocorrelation function* is defined as

$$\rho_s = \frac{Ex_t x_{t+|s|}}{Ex_t^2} = \int_0^T \cos \omega s \, dF(\omega) \tag{5.6}$$

$F(\omega)$ is the *integrated spectrum* and is a distribution function in $(0, \pi)$, so that it is non-decreasing with $F(0) = 0$ and $F(\pi) = 1$. Inverting (5.6) obtains

$$F(\omega) = \pi^{-1} \left(\omega + 2 \sum_{s=1}^{\infty} \rho_s \frac{\sin \omega s}{s} \right)$$

When $F(\omega)$ is differentiable, its derivative

$$f(\omega) = \pi^{-1} \sum_{s=-\infty}^{\infty} \rho_s \cos \omega s$$

is called the *spectral density*, *power spectrum* or just *spectrum*, reserving the term power for the value of the spectral density at a particular frequency.

5.5 The traditional method of *harmonic analysis* is concerned with the estimation of the principal harmonic components of a time series. The amplitudes of these harmonics are given by $\sqrt{(2/T)(A_i^2 + B_i^2)}$ with frequencies $\omega_i = 2\pi i/T$. For each (integral) period $P_i = 2\pi/\omega_i$, the coefficients A_i and B_i may be estimated after arranging the observations in a Buys-Ballot table with each row corresponding to a period.

In *periodogram analysis* a period P_i (either integral or fractional) is tested for significance with the sample size being adjusted to $T' < T$ if necessary so that T' is a multiple of P_i. Although such tests are strictly applicable to harmonic analysis only as the periodogram coordinates are no longer independent for non-integral periods, Jenkins and Priestley argued that the tests were unlikely to be seriously affected in such circumstances as the correlation between two ordinates was only of $O(T^{-2})$. The first test of significance in harmonic analysis was given by Schuster (1898, 1906), who showed that, asymptotically, under the null hypothesis that the series is random,

$$P[I_T(\omega_i) \geq z] \sim e^{-z/c}$$

where

$$c = \sum_{i=1}^{[T/2]} I_T(\omega_i)/[T/2]$$

This was later modified to take account of the fact that, in practice, the *largest* peak of the periodogram is tested. If $I_T(\omega_g)$ is the largest peak, then asymptotically

$$P[I_T(\omega_g) \geq z] \sim 1 - (1 - e^{-z/c})^{[T/2]}$$

An exact test for the largest peak when T is odd was developed by Fisher (1929) using the statistic

$$g = \frac{I_T(\omega_g)}{\sum_{i=1}^{\frac{1}{2}(T+1)} I_T(\omega_i)}$$

The distribution of g was also derived by Fisher, who showed that the 100α per cent critical values could be closely approximated by

$$g_\alpha = 1 - (\alpha/T)^{\frac{1}{T-1}}$$

Jenkins and Priestley discussed further modifications to Fisher's statistic which involved testing the significance of the second largest peak and dealing with a continuous spectrum with spectral density $f(\omega)$. However, their overall conclusion was that the periodogram displayed very erratic behavior and therefore should not be used for estimating the spectrum of a time series.

5.6 If the structure of (5.5) is fairly simple then there is, in fact, no need to estimate the spectrum as it may be written down directly in terms of the estimated parameters. For example, if the parameters α_i and β_j are estimated by a_i and b_j and if $w_t = 0$ then the spectrum of x_t may be written as

$$f_x(\omega) = \frac{\sigma_\varepsilon^2}{\sigma_x^2} \left| \frac{1 + b_1 e^{-i\omega} + b_2 e^{-2i\omega} + \cdots + b_q e^{-qi\omega}}{1 - a_1 e^{-i\omega} - a_2 e^{-2i\omega} - \cdots - a_p e^{-pi\omega}} \right| f_\varepsilon(\omega) \qquad (5.7)$$

Here $f_\varepsilon(\omega)$ is the spectrum of ε_t, typically assumed to be uniform, and σ_x^2 and σ_ε^2 are the variances of x_t and ε_t respectively. An equivalent way of writing (5.7) is as

$$\sigma_x^2 f_x(\omega) = \psi(\omega) f_\varepsilon(\omega) \sigma_\varepsilon^2$$

where $\psi(\omega)$ is known as the *transfer*, or *frequency response*, *function*. For example, if x_t is an AR(1) process with autocorrelation function $\rho_s = \rho^s$

and $f_\varepsilon(\omega) = \sigma_\varepsilon^2/\pi$, (5.7) becomes

$$f_x(\omega) = \frac{1}{\pi}\left(\frac{1-\rho^2}{(1-\rho e^{-i\omega})(1+\rho e^{-i\omega})}\right) = \frac{1}{\pi}\left(\frac{1-\rho^2}{1+\rho^2-2\rho\cos\omega}\right)$$

on noting that $\sigma_\varepsilon^2 = (1-\rho^2)\sigma_x^2$. The parameter ρ may be estimated from the correlogram and the spectrum then computed.

5.7 In most practical applications, however, it will be necessary to estimate the spectrum. Although the periodogram does provide an estimate of the spectral density, unfortunately it is not consistent, since its variance is approximately $(2\sigma_x^2 f_x(\omega))^2$, which remains positive as $T \to \infty$. Bartlett (1948, 1950) therefore suggested estimates of the form

$$f_T(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{m-1}\left(1-\frac{s}{T}\right) r_s \cos\omega s\right) \tag{5.8}$$

$$f_S(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{m-1}\left(1-\frac{s}{m}\right) r_s \cos\omega s\right) \tag{5.9}$$

where $r_s = C_s/C_0$ are the serial correlations. Both estimates use the first $m < T$ serial correlations and are known as the *truncated* and *smoothed* periodograms respectively and may be regarded as the average of m/T periodogram estimates computed from sub-series of length m.

A more general class of estimates, suggested by Grenander and Rosenblatt (1953), is

$$f_G(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{T-1}\lambda_s(\omega) r_s \cos\omega s\right) \tag{5.10}$$

where the weighting function is chosen to make the estimate consistent. Grenander and Rosenblatt also introduced the concepts of resolvability and statistical reliability. The former is defined to be the ability to distinguish between the power of the spectrum at a given frequency and that at a neighbouring frequency; the latter relates to the fact that power estimates are subject to errors induced by recording at discrete intervals and using a series of finite length. They proved a theorem stating that the product of the errors made in estimating the amplitude and frequency have a certain lower bound.

Lomnicki and Zaremba (1957) argued that resolvability and reliability were not 'antagonistic', particularly when the estimation problem was viewed as that of estimating a function, so that standard errors

attached to individual frequency estimates were of lesser importance. They defined 'best' estimators in the sense that the expected mean square error over the whole range of ω was to be a minimum, leading to the set of optimum weights

$$\lambda_s = \frac{\rho_s^2}{V(r_s) + \rho_s^2}$$

although how this was to be put into practical use remained unclear.

5.8 Jenkins and Priestley emphasized that, while a satisfactory theoretical solution to estimating the spectral density was clearly emerging, estimates of the density at individual frequencies were often subject to large sampling fluctuations. They suggested that it might be safer in practice to estimate the average spectrum over a band width, which was often of direct interest anyway. For a bandwidth $(\omega - \xi, \omega + \xi)$ the equivalent smoothed band spectrum estimate (5.9) is

$$f_{S,B}(\omega) = \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{m-1}\left(1 - \frac{s}{m}\right) C_s \cos \omega s \frac{\sin s\xi}{s\xi}\right) \tag{5.11}$$

5.9 Using (5.10), the (non-normalized) integrated spectrum can be estimated as

$$F_G(\omega) = \frac{1}{\pi}\left(\omega + 2\sum_{s=1}^{m-1} \lambda_s C_s \frac{\sin \omega s}{s}\right)$$

where $\lambda_s = 1 - s/h_T$, $h_T = [kT^\theta]$ and $\frac{1}{2} < \theta < 1$.

5.10 Jenkins and Priestley proposed that a confidence interval for the spectral density $f(\omega)$ be derived from taking $\log(c + f_{S,T}(\omega))$ to be asymptotically distributed with mean $\log(c + f(\omega))$ and variance m/T, where c is a suitably chosen positive constant, although they pointed out that this method only provides a confidence interval at a particular frequency ω and not for the entire spectral density function.

A confidence interval for $f(\omega)$ may also be derived from the band spectrum estimate (5.11): a $100(1-\alpha)\%$ interval is given by

$$\left(\frac{\upsilon}{\chi^2_{1-\alpha,\upsilon}} f_{S,B}(\omega), \frac{\upsilon}{\chi^2_{\alpha,\upsilon}} f_{S,B}(\omega)\right)$$

where $\upsilon = 2\xi T/\pi$.

5.11 Jenkins and Priestley concluded their survey by summarizing the two main difficulties facing spectral analysis at this point, these being: (i) the choice of the optimum truncation point m; and (ii) the practical problems that arise when there are both continuous and discrete components present in a time series. The correlogram will then not be damped but will oscillate widely, in which case the surveyed methods are no longer applicable without further modification, although Jenkins and Priestley did offer some tentative suggestions for dealing with this situation of a mixed spectra. Finally, they offered a brief description of some computer programs that had been developed at the RAE for computing serial and cross-correlations, spectra and periodograms.

Spectral analysis at the start of the 1960s

5.12 Jenkins (1961) represented an updated survey 'written with the intention of giving some motivation for spectral analysis to statisticians who might be puzzled by the fact that one would want to work in the frequency domain at all' (pages 140–1). While it necessarily repeated much of the preliminary ground covered in Jenkins and Priestley (1957), it does contain much new material that is worth discussing.

After giving some physical/engineering examples to motivate spectral analysis, Jenkins introduced the concept of *white noise*. This is a series, such as ε_t in **§5.4**, which has a constant spectral density $f(\omega) = 1/\pi$, which corresponds to the series having zero autocovariances and thus being uncorrelated through time. Jenkins regarded white noise and its associated spectrum as having special theoretical importance, making it an essential building block for time series processes.[1]

5.13 Jenkins then returned to the analysis of **§5.6** related to frequency response functions, introducing a finer distinction between the concepts of *gain* and *phase* and the frequency response. He began by considering the, possibly continuous time, 'cosinusoidal disturbance' $x(t) = A\cos\omega t$ and the output, $y(t)$, from the linear system

$$y(t) = G(\omega)A\cos(\omega t + \Phi(\omega)) = \int_0^\infty W(\tau)x(t-\tau)d\tau$$

where $G(\omega)$ is the gain or attenuation factor and $\Phi(\omega)$ is the phase shift. The frequency response is then defined as

$$\psi(\omega) = G(\omega)e^{i\Phi(\omega)} = \int_0^\infty e^{i\omega\tau}W(\tau)d\tau$$

and it therefore follows that $G(\omega) = |\Phi(\omega)|$ and $\Phi(\omega) = \arg \psi(\omega)$. It is also the case that, cf. **§5.6** and the notation used there,

$$\sigma_y^2 f_y(\omega) = \sigma_x^2 f_x(\omega) G^2(\omega) \tag{5.12}$$

so that $G(\omega)$ may be calculated if the spectral densities $f_x(\omega)$ and $f_y(\omega)$ have been estimated.

Closely related to the notion of a frequency response is the idea of a (*digital*) *filter*, which essentially is a tailored $G(\omega)$ which may be used to produce an output falling in a required frequency range. In effect, such a filter is a symmetrical moving average whose weights have been selected to produce particular frequency domain properties. Thus the (discrete) periodic input $x_t = Ae^{i\omega t}$ (now being represented in complex form) may be converted into an output $y_t = Ae^{i\omega t} T(\omega)$ through the use of the filter

$$T(\omega) = \sum_{i=-k}^{k} \delta_i e^{i\omega t}$$

A *low* (*high*)*-pass* filter is such that $T(\omega)$ is designed so that the output y_t only contains frequencies up to (above) a chosen cut-off frequency. A *band-pass* filter produces an output that contains only frequencies in a certain interval and thus can be considered as the difference between two low-pass or high-pass filters.

Replacing $G^2(\omega)$ by $|G(\omega)|^2$ and integrating (5.12) over all frequencies obtains

$$\sigma_y^2 = \sigma_x^2 \int_0^\infty f_x(\omega) |G(\omega)|^2 d\omega$$

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.

5.14 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.

5.15 Jenkins showed that (5.9), say, could, in general, be expressed equivalently as

$$\hat{f}(\omega) = \int_0^{\pi} I(y)K(\omega, y)dy$$

where

$$K(\omega, y) = \tfrac{1}{2}(\lambda(\omega + y) + \lambda(\omega - y)) \qquad \lambda(y) = \frac{1}{\pi} \sum_{s=-T}^{T} \lambda_s e^{iys}$$

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' of (5.9), $\lambda_s = 1 - s/m$ for $s \leq m$ and $\lambda_s = 0$ for $s > m$, is

$$K(\omega, y) = \frac{1}{\pi m} \left(\frac{\sin^2 (m/2)(\omega + y)}{\sin^2 \tfrac{1}{2}(\omega + y)} + \frac{\sin^2 (m/y)(\omega - y)}{\sin^2 \tfrac{1}{2}(\omega - y)} \right)$$

This kernel has a shape that falls off rapidly from its maximum at the peak frequency $y = \omega$ and reaches zero at $y = \pm\pi/m$, beyond which it oscillates with decreasing amplitude.

Associated with a kernel is its *bandwidth*. This is defined to be half the base width, $2\pi/m$, of a rectangular kernel which has the same height and same area as $K(\omega, y)$, although other definitions have been suggested. Increasing m 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 weight functions being suggested, which are summarized, along with their associated kernels, as Jenkins (1961, Table 1).

Writing the weight function as $\lambda(u)$, $u = k/m$, then the Bartlett weights correspond to setting

$$\begin{aligned} \lambda(u) &= 1 - |u|, & |u| \leq 1 \\ &= 0, & |u| > 1 \end{aligned}$$

The 'hanning' estimate of Blackman and Tukey (1958) is

$$\begin{aligned} \lambda(u) &= \tfrac{1}{2}(1 + \cos \pi u), & |u| \leq 1 \\ &= 0, & |u| > 1 \end{aligned}$$

while their 'hamming' estimate is

$$\begin{aligned} \lambda(u) &= 0.54 + 0.46 \cos \pi u, & |u| \leq 1 \\ &= 0, & |u| > 1 \end{aligned}$$

A generalization of these two weight functions is

$$\begin{aligned} \lambda(u) &= 1 - 2a + 2a \cos \pi u, & |u| \leq 1 \\ &= 0, & |u| > 1 \end{aligned}$$

in which hanning is obtained by setting $a = 0.25$ and hamming by setting $a = 0.23$. Parzen (1957, 1961) suggested the weight functions

$$\begin{aligned} \lambda(u) &= 1 - u^2, & |u| \leq 1 \\ &= 0, & |u| > 1 \end{aligned}$$

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. The 'design' considerations involved in choosing a weight function/kernel and a bandwidth were then discussed in detail by Jenkins, who concluded the paper with the following convenient summary of the various stages of a spectral analysis.

(1) *Preliminary considerations*: It will be very important to know what relation the estimated spectrum bears to that of the spectrum of interest. Most of the important errors have been introduced before the statistical considerations have begun. The interpretation of the spectrum is then dictated almost entirely by non-statistical considerations.

(2) *Choice of Spectral Window*: The most important feature of spectral estimation is that *some* sort of window should be used with a bandwidth considerably greater than $1/T$ where T is the total number of observations.

There is still some doubt as to what constitutes a good spectral shape but [Tukey's hamming and Parzen's second window] seem to have attractive properties.

(3) *Analysis Considerations*

(a) Calculate the autocovariances having removed a mean and possibly a linear trend. For fixed T, all that is needed is to select the number of lags.

(b) There seem to be three ways of doing this:

(i) Plot the autocovariance function up to 25%–30% of the total sample size and determine a reasonable truncation point empirically.

(ii) Specify a bandwidth for the spectral window chosen taking care that there are enough degrees of freedom per estimate.

(iii) Choose m from a mean square error criterion using some knowledge of $f(\omega)$ obtained from (i) or from pilot spectral analysis.

…

Ultimately there is not going to be a great deal to choose between the above three methods of choosing m from a practical point of view since one will be wise to base spectra *on a few selections of m* anyway. The author would prefer to work with (i) backed up by (iii) on the grounds that he is unable to specify a bandwidth in a vacuum unless possibly there are special objectives which restrict the choice of bandwidth. In addition to basing spectra on a few choices of m, it is suggested that one should feel free to calculate spectral ordinates at any frequencies.

In the last resort, if it is difficult to make sense of the spectrum from a physical point of view, *then the more refined statistical considerations are irrelevant*. In particular, if taking the two halves of the same series gives widely differing answers or if the next experiment produces a different spectral shape, then one has far greater problems than statistical ones. (Jenkins, 1961, pages 164–5: italics in original)

This summary represents a good reflection of Jenkins' data modelling philosophy, in that progress can only be made by experimentation using robust techniques within a wider context of the physical 'system' producing the data.

Open loop transfer functions and cross-spectral analysis

5.16 Jenkins (1963a, 1963b) considered the more general frequency response/transfer function model

$$y(t) = \phi \boldsymbol{x}(t) + n(t) = \int_0^\infty W(\tau)\boldsymbol{x}(t-\tau)d\tau + n(t) \tag{5.13}$$

where $\boldsymbol{x}(t)$ is a continuous vector of past values of the input and $n(t)$ is a noise term containing the influences of other inputs on the output and also any non-linear effects that have been omitted from the linear approximation implicit in (5.13): Jenkins referred to the presence of $n(t)$ as 'smudging' the relationship between $y(t)$ and $\boldsymbol{x}(t)$, which may be termed a *linear dynamic equation* with the regression function $W(\tau)$ often being known as the *impulse response*. The frequency response function is given by (cf. **§5.13**)

$$T(\omega) = G(\omega)e^{-\Phi(\omega)} = \int_0^\infty W(\tau)e^{-i\omega\tau}d\tau$$

The modulus $G(\omega) = |T(\omega)|$ then represents the gain and $\Phi(\omega)$ the phase shift. The system is said to be *open loop* if there is *no feedback* from $y(t)$ to $\boldsymbol{x}(t)$.

Assuming that $y(t)$ and $\boldsymbol{x}(t)$ are stationary and $\boldsymbol{x}(t)$ and $n(t)$ are uncorrelated then multiplying (5.13) throughout by $\boldsymbol{x}(t-s)$ and averaging obtains

$$\gamma_{xy}(s) = \int_0^\infty W(\tau)\gamma_{xx}(s-\tau)d\tau$$

where $\gamma_{xy}(s) = E(\boldsymbol{x}(t-s)y(t)) = \sigma_x\sigma_y\rho_{xy}(s)$ is the cross-covariance function at lag s between $\boldsymbol{x}(t)$ and $y(t)$ and $\gamma_{xx}(v)$ is the autocovariance of $\boldsymbol{x}(t)$ at lag v. The frequency response function can then be written as

$$T(\omega) = \frac{\sigma_y}{\sigma_x}\frac{g_{xy}(\omega)}{g_{xx}(\omega)} \tag{5.14}$$

where $g_{xy}(\omega)$ is the cross-spectral density function given by

$$\sigma_x \sigma_y g_{xy}(\omega) = \frac{1}{\pi} \int_{-\pi}^{\pi} \gamma_{xy}(s) e^{-i\omega s} ds$$

This cross-spectrum can be written as

$$g_{xy}(\omega) = c_{xy}(\omega) - i q_{xy}(\omega)$$

where $c_{xy}(\omega)$ and $q_{xy}(\omega)$ are the *cospectrum* and *quadrature spectrum*, respectively, given by

$$\begin{aligned} c_{xy}(\omega) &= \frac{2}{\pi} \int_0^{\infty} \alpha_{xy}(s) \cos \omega s \, ds \\ q_{xy}(\omega) &= \frac{2}{\pi} \int_0^{\infty} \beta_{xy}(s) \sin \omega s \, ds \end{aligned}$$

where

$$\begin{aligned} \alpha_{xy}(s) &= \tfrac{1}{2}(\rho_{xy}(s) + \rho_{xy}(-s)) \\ \beta_{xy}(s) &= \tfrac{1}{2}(\rho_{xy}(s) - \rho_{xy}(-s)) \end{aligned}$$

so that

$$\rho_{xy}(s) = \alpha_{xy}(s) + \beta_{xy}(s)$$

$\alpha_{xy}(s) = \alpha_{xy}(-s)$ is known as the *even part* of the spectrum and $\beta_{xy}(s) = -\beta_{xy}(-s)$ as the *odd part*. It then follows that

$$\begin{aligned} G(\omega) &= \frac{\sigma_y}{\sigma_x} \frac{(c_{xy}^2(\omega) + q_{xy}^2(\omega))^{1/2}}{g_{xx}(\omega)} \\ \Phi(\omega) &= \tan^{-1} \frac{q_{xy}(\omega)}{c_{xy}(\omega)} \end{aligned}$$

The *cross-amplitude spectrum* is defined as

$$R_{xy}(\omega) = \sqrt{c_{xy}^2(\omega) + q_{xy}^2(\omega)}$$

and the squared *coherency*, a measure of the correlation between $y(t)$ and $\boldsymbol{x}(t)$ at frequency ω, is then defined as

$$0 \leq C_{xy}^2(\omega) = \frac{R_{xy}^2(\omega)}{g_{xx}(\omega)g_{yy}(\omega)} = \frac{|g_{xy}(\omega)|^2}{g_{xx}(\omega)g_{yy}(\omega)} \leq 1$$

Since from (5.13) it follows that, using an obvious notation,

$$\sigma_y^2 g_{yy}(\omega) = \sigma_x^2 G^2(\omega) g_{xx}(\omega) + \sigma_n^2 g_{nn}(\omega)$$

the coherency can then be written as

$$C_{xy}^2(\omega) = \frac{\sigma_x^2 G^2(\omega) g_{xx}(\omega)}{\sigma_n^2 g_{nn}(\omega) + \sigma_x^2 G^2(\omega) g_{xx}(\omega)} \tag{5.15}$$

The coherency will therefore be 1 if $g_{nn}(\omega) = 0$ and will tend to zero as $\sigma_{nn}^2 g_{nn}(\omega)$ becomes large compared to $\sigma_x^2 G^2(\omega) g_{xx}(\omega)$. The relationship (5.15) can also be written as

$$\sigma_n^2 g_{nn}(\omega) = \sigma_y^2 g_{yy}(\omega)(1 - C_{xy}^2(\omega)) \tag{5.16}$$

Recall that the slope of the linear regression of y_t on $\boldsymbol{x}_t$ is given by $\gamma_{xy}(0)/\gamma_{xx}(0)$ and that the residual sum of squares from the fitted regression is $\sigma_n^2 = \sigma_y^2(1 - \rho_{xy}^2(0))$. These formulas are seen to be mimicked by (5.14) and (5.16) so that transfer function estimation is effectively linear regression analysis at each frequency ω.

5.17 The quantities required for estimating the gain, phase and coherence can be obtained using obvious extensions and modifications of (5.10), viz.,

$$\begin{aligned}
\hat{g}_{xx}(\omega) &= \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{T-1} \lambda_s r_x(s) \cos \omega s\right) \\
\hat{g}_{yy}(\omega) &= \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{T-1} \lambda_s r_y(s) \cos \omega s\right) \\
\hat{c}_{xy}(\omega) &= \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{T-1} \lambda_s a_{xy}(s) \cos \omega s\right) \\
\hat{q}_{xy}(\omega) &= \frac{1}{\pi}\left(1 + 2\sum_{s=1}^{T-1} \lambda_s b_{xy}(s) \cos \omega s\right)
\end{aligned}$$

where

$$a_{xy}(s) = \tfrac{1}{2}(r_{xy}(s) + r_{xy}(-s))$$
$$b_{xy}(s) = \tfrac{1}{2}(r_{xy}(-s) - r_{xy}(s))$$

For large T, the covariance matrix of these estimates is given by, after dropping the dependence on ω for notational convenience, $(\ell/T)\mathbf{V}$, where

$$\mathbf{V} = \begin{bmatrix} g_{xx}^2 & c_{xy}^2 + q_{xy}^2 & g_{xx}c_{xy} & g_{xx}q_{xy} \\ & g_{yy}^2 & g_{yy}c_{xy} & g_{yy}q_{xy} \\ & & \tfrac{1}{2}(g_{xx}g_{yy} + c_{xy}^2 - q_{xy}^2) & c_{xy}q_{xy} \\ & & & \tfrac{1}{2}(g_{xx}g_{yy} + q_{xy}^2 - c_{xy}^2) \end{bmatrix} \tag{5.17}$$

and

$$\ell = \tfrac{1}{2}\sum_{s=-T}^{T} \lambda_s^2$$

depends upon the weight function being used. Using (5.17), the variances of the estimated cross-amplitude spectrum, $\hat{R}_{xy}(\omega)$, gain, $\hat{G}(\omega)$, coherence, $\hat{C}_{xy}(\omega)$, and phase, $\hat{\Phi}(\omega)$, are given by

$$V(\hat{R}_{xy}(\omega)) = \frac{m}{T}R_{xy}^2(\omega)\frac{1}{2}\left(1 + \frac{1}{C_{xy}^2(\omega)}\right)$$
$$V(\hat{G}(\omega)) = \frac{m}{T}G^2(\omega)\frac{1}{2}\left(\frac{1}{C_{xy}^2(\omega)} - 1\right)$$
$$V(\hat{C}_{xy}(\omega)) = \frac{m}{T}\frac{1}{2}(1 + C_{xy}^2(\omega))$$
$$V(\hat{\Phi}(\omega)) = \frac{m}{T}\frac{1}{2}\left(\frac{1}{C_{xy}^2(\omega)} - 1\right)$$

The formulae for the variance of the estimated gain and phase are seen to depend on the factor

$$\vartheta^2(\omega) = \frac{1}{C_{xy}^2(\omega)} - 1$$

which tends to zero as $C_{xy}^2(\omega)$ tends to one and tends to infinity as $C_{xy}^2(\omega)$ tends to zero. Since the coherency is determined by the size

of the noise, the sampling properties of the estimated gain and phase may become dominated by the uncontrollable influence of $\vartheta^2(\omega)$ rather than the controllable influence of m/T. It may also be shown that $\text{cov}(\hat{G}(\omega), \hat{\Phi}(\omega)) = 0$ so that the phase and gain can be treated separately.

A confidence interval for the gain may be constructed using the result that $f\hat{G}(\omega)/G(\omega)$ is approximately distributed as χ^2 with $4T/(m\vartheta^2(\omega))$ degrees of freedom, so that a $100(1-\alpha)\%$ interval is given by

$$\left(\frac{f\hat{G}(\omega)}{\chi^2_{1-\alpha/2}(f)}, \frac{f\hat{G}(\omega)}{\chi^2_{\alpha/2}(f)} \right)$$

A confidence interval for the phase may be obtained by taking $\tan \Phi(\omega)$ to be approximately normally distributed with variance

$$V(\tan \Phi(\omega)) = \frac{m}{T} \sec^4 \Phi(\omega) \frac{\vartheta^2(\omega)}{2}$$

from which it follows that

$$V(\Phi(\omega)) = \frac{m}{T} \frac{\vartheta^2(\omega)}{2}$$

Spectral analysis at the end of the 1960s

5.18 Jenkins (1965) returned to surveying the spectral scene, including a discussion of cross-spectral analysis and a detailed treatment of transfer function modelling and estimation. His experiences of spectral analysis were then brought together in a major textbook written with Donald Watts. Jenkins and Watts (1968) remains a key reference to the subject, but it was Jenkins' last published foray in spectral analysis although, as will be seen in Chapter 8, several of the topics discussed in this chapter subsequently became important features and building blocks of later developments in time series analysis.

The final section of Jenkins (1965) discussed the uses and limitations of spectral analysis. The uses are summarized in the following table, and reflect Jenkins' burgeoning interest in systems engineering: for an expanded discussion of these areas see Jenkins (1965, pages 26–30).

1. *Suggesting Models*	(*a*) Presence of peaks
	(*b*) Variation of spectra with controlled or uncontrolled variables
	(*c*) Shape of cross amplitude, gain and phase plots

(*d*) Importance of digital filters
(*e*) Non-linearities
(*f*) Non-Poisson nature of point processes

2. *Systems Design*
$\sigma_y^2 f_y(\omega) = \sigma_x^2 f_x(\omega) G^2(\omega)$

(a) Choose $G(\omega)$ so as to modify $f_x(\omega)$
(*b*) Design of experiments to which block size is chosen from the minimum of the spectral density

Perhaps of more interest are Jenkins' views concerning the limitations of spectral analysis. Jenkins had, in fact, begun the paper by stating that

> (i)n no sense … can it be said that spectral analysis is widely used or even understood by statisticians and many of the applications of the technique have in fact been made by physicists and engineers. It is suggested that there are two reasons for this:
>
> (1) The genuine difficulties which statisticians (as opposed to physicists and engineers) face in thinking in terms of frequency concepts.
> (2) The highly mathematical nature of papers written on spectral analysis.
>
> This undue emphasis on mathematical work has led many statisticians to believe that spectral analysis is very difficult to apply. (*ibid.*, page 2)

The paper thus tried 'to present, using the minimum of mathematics, all those ideas in spectral analysis which are necessary *in order to be able to apply* the technique' (ibid., page 2: italics in original) and Jenkins and Watts (1968) represented a further attempt to do this with greater detail and mathematical rigour. Nonetheless, Jenkins felt compelled to end the paper by discussing what he felt to be the major limitations of spectral analysis. He thought that the fact that it was a non-parametric technique was a big disadvantage because it was then necessary to estimate a whole function, or at least a very large set of parameters, something that statisticians were not usually accustomed to do, with a consequent impact upon the efficiency with which the parameters could be estimated. The availability of a parametric model, say of the ARMA type, could place certain assumptions upon the smoothness of the spectrum and thus lead to better behaved estimates.

A further disadvantage of spectral analysis was its dependence on stationarity, for during the 1960s the presence of trends in time series

and the implications these had for statistical modelling and inference were becoming increasingly of interest. Parametric estimation of non-stationary models involving only a few parameters was becoming possible and, as will be emphasized in subsequent chapters, this was an area in which Jenkins was to become increasingly involved in.

6
Box and Jenkins: Time Series Analysis, Forecasting and Control

George Box

6.1 George Box was born in Gravesend, Kent on 18 October 1919 and, after being educated at grammar school, went to the local polytechnic to study chemistry. When the war intervened he was posted to the British Army Engineers to work as a laboratory assistant in a chemical defence experiment station investigating the effects of poison gas. His job was to carry out tests on small animals and determine the effects of gassing and subsequent treatment but, as the test results varied considerably, Box realized that statistical analysis was required and that any such analysis would have to be done by himself! Being 1942, all that he could do was to purchase some books and teach himself enough statistics to analyze the data. This he certainly did 'beyond the call of duty' and his work on experimental design in this area of pathology was recognized with the award of a British Empire Medal at the end of the war.

On returning to higher education after the war, Box saw that his interests lay in statistics rather than chemistry and he obtained a BSc in mathematical statistics from UCL in 1947. He then embarked on a Master's degree, later upgraded to a doctorate, at UCL but, after a summer placement, accepted a job with Imperial Chemical Industries (ICI) which consequently delayed the completion of his PhD until 1953. During his time at ICI Box worked on experimental design and began to publish papers that got the attention of statisticians in the United States, leading to a visiting professorship at the University of North Carolina in 1953–4 and eventually a permanent move, first as Director of the Statistical Techniques Research Group at Princeton in 1957 and then to the University of Wisconsin, Madison in 1960, where he became professor and chairman of the new Statistics department that he was asked to set up. In 1971

he was appointed to the Ronald Aylmer Fisher chair in Statistics and in 1980 was named the Vilas Research Professor of Statistics before finally retiring from Wisconsin in 1992 with the title Professor Emeritus.

Box continued to work in experimental design as well as many other areas of statistics and, in later years, became especially interested in quality control (see, for example, Box, 1989), establishing the Centre for Quality and Productivity Improvement at Madison. By his own admission, until he met Gwilym Jenkins, Box 'regarded time series as a very boring subject. ... Indeed, I think that Gwilym was the first person I'd ever met who talked coherently about time series in terms of actually doing something with it' (Box, 1983b, page 516). Their initial collaboration was essentially about a problem in experimental design in which it was required to make an industrial process track a moving optimum. They gradually realized that this problem in automatic optimization was actually one of control, and involved forecasting because simple control algorithms can be regarded as mechanisms for forecasting how big the process deviation will be at the end of the next process interval and then taking action which will cancel out the forecast deviation. This led them to forecasting techniques such as exponential smoothing and then naturally on to methods of modeling and forecasting non-stationary time series:

> The book that finally came out [Box and Jenkins, 1970] is sort of backward to the way we got in. The control part is at the end and I don't think the actual problem we started with, pursuing the maximum, even gets mentioned in the book. But that's the place we actually started. We worked back and then realized we had to do something about non-stationary time series in order to do that. (de Groot, 1987, page 251)

In 1964 Box was awarded the RSS's Guy Medal in Silver and in 1993 the Guy Medal in Gold. He has also been awarded the Wilks Memorial Medal from the American Statistical Association in 1972 and the Shewhart Medal from the American Society for Quality Control in 1968. Further biographical details on Box, plus a host of reminiscences, may be found in Box (1983b), de Groot (1987) and Peňa (2001).

Statistical aspects of adaptive optimization and control

6.2 The first fruit of their collaboration was Box and Jenkins (1962), in which they addressed the issues of adaptive optimization and control

within a statistical context. The simple discrete adaptive optimization model they began with has data that is available only at equal intervals of time, which they refer to as a *phase*, during which the underlying process remains constant but which may change from phase to phase. The model thus provides a discrete approximation to the operation of a continuous process from which discrete data are taken at equal intervals of time.

In *discrete adaptive optimization* the uncontrollable and immeasurable variables have levels ξ_t during the tth phase which change from one phase to the next. The controllable variable X then determines the conditional response function $\eta(X|\xi_t)$, which is assumed to be approximated by the quadratic function

$$\eta_t = \eta(X|\xi_t) = \eta(\theta_t) - \tfrac{1}{2}\beta(X - \theta_t)^2 \tag{6.1}$$

where $\theta_t = X_{\max}|\xi_t$ is the conditional maximal setting during the tth phase and β is known from prior calibration and does not change appreciably with ξ_t. Because of fluctuations in ξ_t, the conditional maximal value θ_t follows, typically, a non-stationary process. If X_t is the *set point* at which X is held during the tth phase and $\varepsilon_t = \theta_t - X_t$ measures the extent to which X_t deviates from θ_t, then an estimate of ε_t is given by

$$e_t = \varepsilon_t + u_t = \theta_t - X_t + u_t = z_t - X_t \tag{6.2}$$

where u_t is the measurement error and $z_t = \theta_t + u_t$ is an estimate of the optimal setting θ_t during the tth phase.

If a series of adjustments have actually been made, then there will be available a record of the set points $X_t, X_{t-1}, X_{t-2}, \ldots$ and deviations $e_t, e_{t-1}, e_{t-2}, \ldots$, from which the sequence $z_t, z_{t-1}, z_{t-2}, \ldots$ of estimated positions of the maxima in phases $t, t-1, t-2, \ldots$ may be calculated. From these data an adjustment x_{t+1} to the set point X_t is required to be calculated so that the adjusted set point $X_{t+1} = X_t + x_{t+1}$, which will be maintained during the *coming* $(t+1)$-th phase, will, in some sense, be 'best' in relation to the coming and unknown value of θ_{t+1}.

The loss sustained during the $(t+1)$-th phase is assumed, using (6.1), to be measured by

$$\eta(\theta_{t+1}) - \eta(X_{t+1}|\xi_{t+1}) = \tfrac{1}{2}\beta(X_{t+1} - \theta_{t+1})^2$$

The adjustment x_{t+1} will then minimize the expected loss if it is chosen so that $E(X_{t+1} - \theta_{t+1})^2$ is minimized. This will be achieved if X_{t+1} is set

equal to

$$\hat{\theta}_{t+1} = \sum_{j=0}^{\infty} \mu_j z_{t-j}$$

where $\hat{\theta}_{t+1}$ is the minimum mean square linear estimate (MMSLE) or *predictor* of θ_{t+1} based on $z_t, z_{t-1}, z_{t-2}, \ldots$ and the μ_j's are referred to as the *predictor weights*.

At the beginning of the $(t+1)$-th phase an adjustment of $x_{t+1} = \hat{\theta}_{t+1} - \hat{\theta}_t$ should be applied to the previous setting X_t. In practice the z's will not be observed, only the e's, so that the adjustment is calculated from

$$x_{t+1} = X_{t+1} - X_t = \hat{\theta}_{t+1} - \hat{\theta}_t = \sum_{j=0}^{\infty} w_j e_{t-j} \tag{6.3}$$

with the w_j being the *controller weights*. The optimal adjustment is then obtained by choosing the w's to minimize $E(\varepsilon_{t+1}^2) = E(\theta_{t+1} - \hat{\theta}_{t+1})^2$.

Assuming that the measurement error $u_{t+1} = z_{t+1} - \theta_{t+1}$ is distributed about zero with variance σ_u^2 independently of $u_t, u_{t-1}, u_{t-2}, \ldots$ and $\theta_t, \theta_{t-1}, \theta_{t-2}, \ldots$, then

$$E(e_{t+1}^2) = E(z_{t+1} - \hat{\theta}_{t+1})^2 = E(\varepsilon_{t+1}^2) + \sigma_u^2$$

The loss $E(\varepsilon_{t+1}^2)$ is then minimized when $E(z_{t+1} - \hat{\theta}_{t+1})^2 = E(e_{t+1})^2$ is minimized and $\hat{x}_{t+1}$, the best predictor of x_{t+1}, is then the best predictor $\hat{\theta}_{t+1}$ of θ_{t+1}.

6.3 Equation (6.2) can be placed in an alternative context in which the object is to hold θ as closely as possible to some *target value*, taken without loss of generality to be zero, and to achieve this by adjusting the process z up or down at will. Suppose that by the tth phase a *total correction* $-X_t$ has been applied to z so that the apparent deviation from target is given by (6.2). A further adjustment x_{t+1} is then to be made so that, when the total correction $-X_{t+1} = -(X_t + x_{t+1})$ is applied, the actual deviation from target $\varepsilon_{t+1} = \theta_{t+1} - X_{t+1}$ will hopefully be small. If a quadratic loss function is assumed then x_{t+1} should be chosen as before so that $E(\varepsilon_{t+1}^2) = E(\theta_{t+1} - X_{t+1})^2$ is minimized. Once more, this requires that (6.3) holds with the w_j's chosen so that $\hat{\theta}_{t+1}$ is the MMSLE of θ_{t+1} and, if the measurement errors are uncorrelated with each other and with the θ's, then $\hat{z}_{t+1} = \hat{\theta}_{t+1}$ is the MMSLE of z_{t+1}. Box and Jenkins referred to this problem as *discrete adaptive quality control*.

6.4 The adaptive optimization and quality control problems of **§6.2** and **§6.3** are identical, as in both cases optimal action is taken when X_t is adjusted either directly or indirectly to a value

$$X_{t+1} = X_t + x_{t+1} = \hat{\theta}_{t+1}$$

where $\hat{\theta}_{t+1}$ is the mean square error predictor of θ_{t+1}.

Unlike in the prediction problem, where the z's are directly observed and the predictor $\hat{\theta}_{t+1}$ can be conveniently calculated from

$$\hat{\theta}_{t+1} = \sum_{j=0}^{\infty} \mu_j z_{t-j} \tag{6.4}$$

where the μ's are chosen to minimize $E(\theta_{t+1} - \hat{\theta}_{t+1})^2$, in the optimization and control problems optimal action is taken by setting X_{t+1} equal to $\hat{\theta}_{t+1}$, so that only the e's are observed and not the z's. The optimal adjustments x_{t+1} can then be calculated from the e's using (6.3) according to

$$x_{t+1} = \hat{\theta}_{t+1} - \hat{\theta}_t = \sum_{j=0}^{\infty} w_j e_{t-j} \tag{6.5}$$

On defining the lag operator B such that $B^j z_t \equiv z_{t-j}$, equation (6.4) can be written as

$$\hat{\theta}_{t+1} = \sum_{j=0}^{\infty} \mu_j B^j z_t \tag{6.6}$$

and (6.5) as

$$(1 - B)\hat{\theta}_{t+1} = \sum_{j=0}^{\infty} w_j B^j (z_t - \hat{\theta}_t)$$

so that

$$\left(1 - B + B \sum_{j=0}^{\infty} w_j B^j\right) \hat{\theta}_{t+1} = \sum_{j=0}^{\infty} w_j B^j z_t \tag{6.7}$$

Since both (6.6) and (6.7) express $\hat{\theta}_{t+1}$ as an infinite series in powers of B, equating coefficients on B^j, $j = 0, 1, 2, \ldots$, leads to the following recurrence relationship between the controller and predictor weights

$$w_j = \mu_j - \mu_{j-1} + \sum_{k=0}^{j-1} \mu_k w_{j-1-k} \tag{6.8}$$

6.5 Box and Jenkins then considered various specifications of the series z_t and how to choose corresponding predictor weights having desirable properties. The associated controller weights can then be obtained from the recurrence relation (6.8) and will, of course, have the same desirable properties.

The approach they suggested begins by assuming once again that the measurement errors u are uncorrelated with each other and with the θ's, so that $\hat{z}_{t+1} = \hat{\theta}_{t+1}$. Now the change in the predictor $\hat{z}$ at the $(t+1)$-th phase may be written, using the difference operator $\Delta = 1 - B$, as (cf. (6.5))

$$\Delta\hat{z}_{t+1} = \Delta\hat{\theta}_{t+1} = \sum_{j=0}^{\infty} w_j e_{t-j}$$

If the predictor model is such that it fits a polynomial of degree d to the last $d+1$ observations then Box and Jenkins showed that

$$\Delta\hat{z}_{t+1} = e_t + S^1 e_t + S^2 e_t + \cdots + S^d e_t$$

where

$$S^1 e_t = \sum_{j=0}^{\infty} e_{t-j}, \quad S^2 e_t = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} e_{t-j-k} \quad \cdots$$

$$S^d e_t = \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}\cdots\sum_{j_d=0}^{\infty} e_{t-j_1-j_2\cdots-j_d}$$

so that $S^j e_t$ denotes the jth multiple sum over the past history of the e's.

6.6 Box and Jenkins modified this model to

$$\Delta\hat{z}_{t+1} = (\gamma_{-l}S^{-l} + \cdots + \gamma_{-1}S^{-1} + \gamma_0 + \gamma_1 S^1 + \cdots + \gamma_m S^m)e_t \qquad (6.9)$$

where $S^{-j} = \Delta^j e_t$, and asked what stochastic process would z need to follow for such a predictor to be optimal. To answer this, they considered the, generally non-stationary, process

$$z_{t+1} = \sum_{j=0}^{\infty} \eta_j z_{t-j} + \alpha_{t+1}$$

where $\alpha_{t+1}, \alpha_t, \alpha_{t-1}, \ldots$ are zero mean, uncorrelated and identically distributed random variables, and proposed predicting the series by

$$\hat{z}_{t+1} = \sum_{j=0}^{\infty} \mu_j z_{t-j}$$

Since

$$E(e_{t+1}^2) = E(z_{t+1} - \hat{z}_{t+1})^2 = E\left\{\sum_{j=0}^{\infty} (\eta_j - \mu_j) z_{t-j}\right\}^2 + E(\alpha_{t+1}^2)$$

such a prediction will be optimal if $\mu_j = \eta_j$, $j = 0, 1, 2, \ldots$. It then follows that the equivalent predictor

$$\Delta\hat{z}_{t+1} = \sum_{j=0}^{\infty} w_j e_{t-j}$$

will be optimal for the equivalent stochastic process

$$\Delta z_{t+1} = \Delta\alpha_{t+1} + \sum_{j=0}^{\infty} w_j \alpha_{t-j}$$

and that when the optimal predictor is used the e's become α's and are uncorrelated. It then follows that the predictor (6.9) will be optimal for a series generated by

$$\Delta z_{t+1} = \Delta\alpha_{t+1} + \sum_{j=-l}^{m} \gamma_j S^j \alpha_t$$

Differencing m times gives

$$\Delta^{m+1} z_{t+1} = \Delta^{m+1}\alpha_{t+1} + \sum_{j=0}^{l+m} \gamma_j \Delta^{l+m-j}\alpha_t$$

which can be rearranged to yield

$$\Delta^{m+1} z_{t+1} = \alpha_{t+1} + \sum_{j=0}^{l+m} \delta_j \alpha_{t-j}$$

Thus the predictor (6.9) would be optimal for a stochastic variable z whose $(m+1)$-th difference can be represented by a moving average of order $l+m+1$:

> Thus we have a result which is of considerable practical value. If, after differencing our series z, which in general will be non-stationary, $m+1$ times, we could render it stationary and if the population serial covariances of lag greater than some value $l+m+1$ were then zero, a predictor of the type [6.9] would be optimal. (ibid., page 313)

Box and Jenkins pointed out that the widely used predictor of taking the exponentially weighted mean

$$\hat{z}_{t+1} = \gamma_0 \sum_{j=0}^{\infty} (1-\gamma_0)^j z_{t-j}$$

corresponded to taking the central term of (6.9), $\Delta\hat{z}_{t+1} = \gamma_0 e_t$, and would be optimal for the stochastic process

$$\Delta z_{t+1} = \Delta\alpha_{t+1} + \gamma_0\alpha_t = \alpha_{t+1} - (1-\gamma_0)\alpha_t$$

that is,

$$z_{t+1} = m + \alpha_{t+1} + \gamma_0 S^1\alpha_t$$

for which the first difference is a moving average of order one. The addition of further terms can thus be considered to be a generalization of this exponential predictor: for example, for series that are highly non-stationary and exhibit marked trends, the additional term in $S^1 e_t$ should be of particular value since it will allow the predictor to adjust to changes in linear trend as well as to changes in mean.

6.7 Bearing in mind the great success of the exponential predictor (see Mills, 2011a, chapter 11, for historical development), Box and Jenkins thought that a simple generalization to the *three-term model*

$$\Delta\hat{z}_{t+1} = \gamma_{-1}\Delta e_t + \gamma_0 e_t + \gamma_1 S e_t \tag{6.10}$$

could be useful in practice. One reason for this is that (6.10) can be considered to be the discrete time analogue to a form of continuous automatic control in which corrections are made proportional to a linear combination of: (i) the first derivative of the current deviation; (ii) the

deviation itself; and (iii) the integral of the deviations over all past history. These types of continuous control are called *derivative*, *proportional* and *integral* respectively and the corresponding terms in (6.10) are referred to as *first difference*, *proportional* and *cumulative* controls.

The stochastic process for which (6.10) is optimal is

$$z_{t+1} = m + \alpha_{t+1} + \gamma_{-1}\alpha_t + \gamma_0 S^1\alpha_t + \gamma_1 S^2\alpha_t$$

for which the second difference is the third-order moving average

$$\Delta^2 z_{t+1} = \alpha_{t+1} + (\gamma_1 + \gamma_0 + \gamma_{-1} - 2)\alpha_t + (1 - 2\gamma_{-1} - \gamma_0)\alpha_{t-1} + \gamma_{-1}\alpha_{t-2}$$

Box and Jenkins noted that, when no difference term is needed in (6.10), $\gamma_{-1} = 0$ and only the first two serial correlations of $\Delta^2 z_{t+1}$ would be nonzero. If, as well, $\gamma_1 = \gamma_0 = 1$ the model reduces to $\Delta^2 z_{t+1} = \alpha_{t+1}$ with predictor $\Delta^2 \hat{z}_{t+1} = e_t + S^1 e_t$.

Box and Jenkins found that, in their experience, second differencing ($m = 1$) was always adequate but a higher order would be suggested if positive serial correlations for the second differences persisted for higher lags. If serial correlations at the fourth or higher lags were appreciable, but much higher lag correlations were small, further parameters corresponding to higher derivative control, that is, $S^j e_t$, could be introduced.

6.8 Although they described a spectral method of estimating the parameters of (6.10), Box and Jenkins preferred to evaluate the sum of squares function $S(\gamma_1, \gamma_0, \gamma_{-1}) = e_1^2 + e_2^2 + \cdots + e_T^2$ for a grid of γ values and to pick out the best values of the parameters. Given a set of data $z_1, z_2, \ldots, z_T$, the sequence $e_1, e_2, \ldots, e_T$ needed to compute $S(\gamma_1, \gamma_0, \gamma_{-1})$ may be obtained by first setting $\hat{z}_1 = z_1$, so that $e_1 = \Delta e_1 = S^1 e_1 = 0$, and then using

$$z_2 = z_1 + \gamma_{-1}\Delta e_2 + \gamma_0 e_2 + \gamma_1 S^1 e_2$$

to obtain

$$e_2 = \frac{z_2 - z_1}{\gamma_{-1} + \gamma_0 + \gamma_1}$$

Repeating the calculation for t incremented one period yields the general expression

$$e_t = \frac{\Delta z_t + \gamma_{-1} e_{t-1} + \gamma_1 S^1 e_{t-1}}{\gamma_{-1} + \gamma_0 + \gamma_1}, \qquad t = 1, 2, \ldots, T$$

which thus enables $S(\gamma_1, \gamma_0, \gamma_{-1})$ to be calculated.

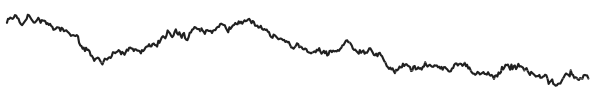
Box and Jenkins showed that it was necessary to restrict the grid of values $(\gamma_1, \gamma_0, \gamma_{-1})$ to a certain region or else the control procedure would become unstable because the prediction variances become infinite. For the general model the stability condition is that the roots of the characteristic equation

$$1 + \sum_{j=0}^{l+m} \delta_j x^{j+1} = 0$$

should not lie on or outside the unit circle with the possible exception of $m + 1$ roots which could all be equal to unity.

6.9 Box and Jenkins' predictor (6.9) clearly reduces a non-stationary time series to stationarity by successive differencing and then fits a moving average model to the differenced data. While differencing had earlier been proposed by Irving Fisher (1925) as a means of inducing stationarity, the success of their *polynomial predictor*, as it became known, led Box and Jenkins to a more systematic study of the role of differencing in time series model building. This eventually became 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 non-stationary*.[1]

Figure 6.1(a) shows a non-stationary series that is homogeneous in its *level*: except for a vertical translation, one part of the series looks much



(a) A series showing non-stationarity in level

(b) A series showing non-stationarity in level and in slope

Figure 6.1 Two kinds of homogeneous non-stationary behavior

the same as any other. Such a series can be rendered stationary by differencing once, that is, by analyzing $z_t = \Delta x_t$ rather than x_t. Figure 6.1(b) shows a second type of non-stationarity of fairly common occurrence, where the series has neither a fixed level nor a fixed slope but exhibits homogeneous behavior if differences in these characteristics are allowed for, for example, if second differences $z_t = \Delta^2 x_t$ are considered.

Integrated processes

6.10 In general, if dth differences are required to render x_t stationary then the series to be analyzed 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 introduced in **§6.5** and defined in terms of the lag operator B 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 $\Delta^{-2} z_t$ and so on. Thus x_t can be obtained by summing (or 'integrating') z_t d times and is therefore said to be an *integrated process* of order d, a terminology first introduced by Hall (1925) (recall **§3.16**).

Autoregressive-integrated-moving average processes

6.11 If the stationary dth differences $z_t = \Delta^d x_t$ can be represented by an ARMA(p, q) process (cf. **§§4.7–4.9**), then this process can be written as

$$\phi(B) z_t = \theta(B) a_t \tag{6.11}$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ are polynomials in B of order p and q respectively and a_t is a zero mean white noise process. Box and Jenkins (1970, chapter 4) called the equivalent model for x_t itself,

$$\phi(B) \Delta^d x_t = \theta(B) a_t \tag{6.12}$$

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 non-stationary processes.

Recall the simple AR(1) process, now written as $(1 - \phi B)x_t = a_t$. 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 behavior of the series shown in Figure 6.1(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 (6.12) 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 6.1(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 non-stationary, but nevertheless exhibit homogeneity, the autoregressive operator must be of the form shown in (6.12).

6.12 For the AR(1) process, the requirement that $\phi(1) = 0$ implies $\phi = 1$, 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).[2] 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 (**§§2.12–2.13**), to Working's (1934) 'random-difference series', and to Macaulay's (1931) 'cumulated chance series'. 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 6.1 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

$$x_t = x_{t-1} + \theta_0 + a_t \tag{6.13}$$

is known as a *random walk with drift.* Figure 6.2 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.

Figure 6.2 A random walk with drift

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

$$\mu_t = E(x_t) = x_0 + t\theta_0$$

as will be the variance and all the auto-covariances

$$\gamma_{k,t} = Cov(x_t, x_{t-k}) = (t-k)\sigma_a^2 \qquad k > 0$$

where $\sigma_a^2 = E(a_t^2)$. 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 non-stationary 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 non-stationary 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, for example

$$x_t = \sum_{j=0}^{d} \beta_j t^j + b_t \qquad \phi(B)b_t = \theta(B)a_t$$

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

6.13 The autocorrelations of an ARIMA$(p, 0, q)$ process will satisfy the difference equation $\phi(B)\rho_k = 0$ for $k > q$ (see Box and Jenkins, 1970, section 3.4.2). 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,

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \qquad k > q - p \tag{6.14}$$

The stationarity requirement is that the roots of $\phi(B)$ must lie outside the unit circle, thus implying that $|G_i| < 1$, $i = 1, \ldots, p$. From (6.14) 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 non-stationary 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 non-stationarity 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, as with their experience of the polynomial predictor, 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.

6.14 Figures 6.3 and 6.4 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 non-stationary 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 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 and that the series itself follows a random walk, the traditional model used for stock prices.

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 an initial value of 0.8, which would suggest 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 behavior 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

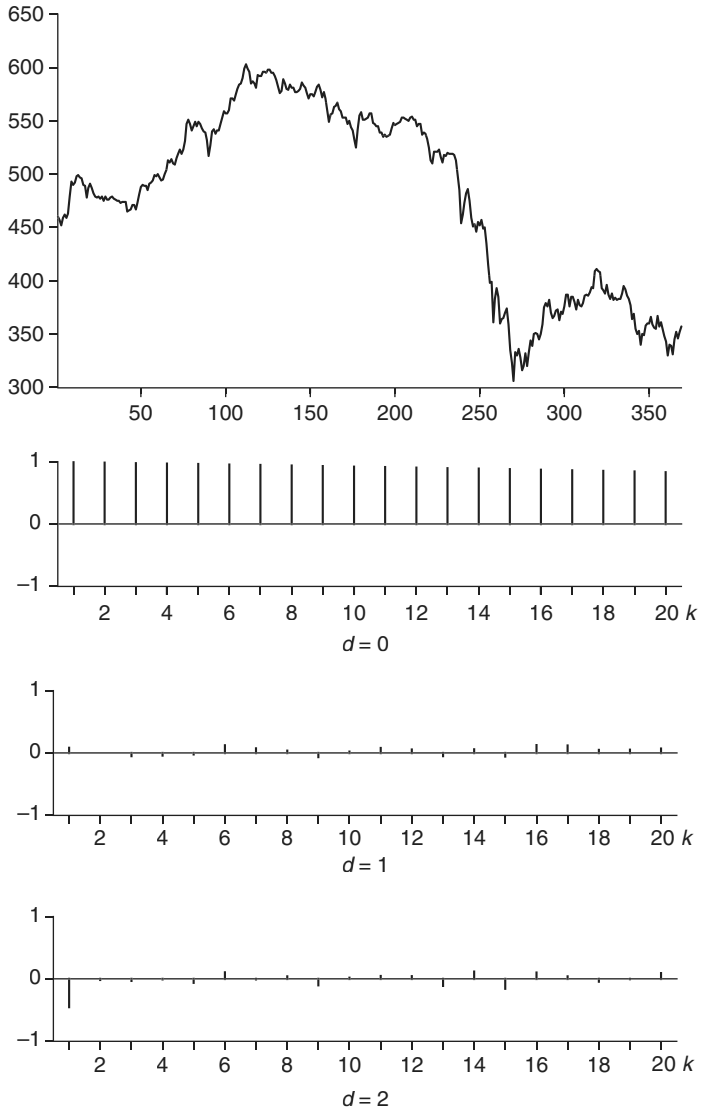


Figure 6.3 Series B from Box and Jenkins (1970); IBM common stock closing prices: daily May 17, 1961–November 2, 1962

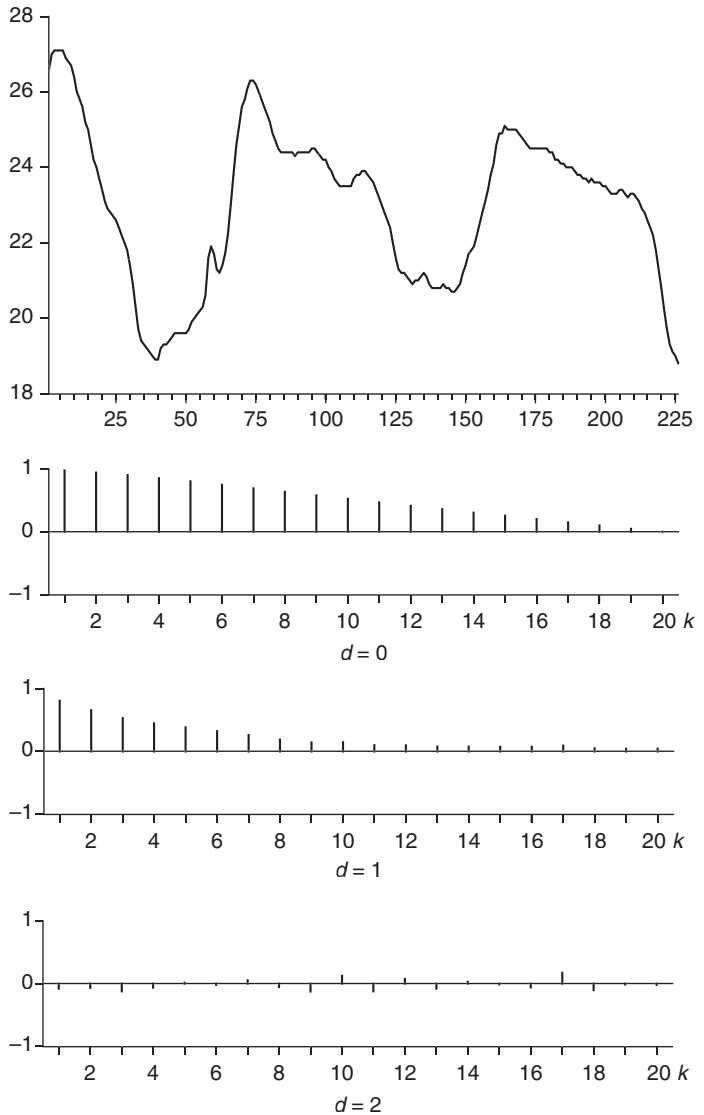


Figure 6.4 Series C from Box and Jenkins (1970); chemical process temperature readings: every minute

for unit roots (see **§10.32**). Nevertheless, determining the order of differencing in this way was an important part of establishing a workable method of identifying ARIMA processes and certainly had a major impact on getting those models accepted and used across a wide range of time series applications.

Identification of ARMA models

6.15 Having chosen the order of differencing d, so that $z_t = \Delta^d x_t$, the orders p and q of the ARMA model generating z_t need to be selected. 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 sample and theoretical 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).

Identification involves studying the general appearance of the sample autocorrelation 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

behavior of the theoretical autocorrelation and partial autocorrelation functions for moving average, autoregressive, and mixed processes. Such characteristic behavior is developed in detail in Box and Jenkins (1970, chapter 3) and subsequently in many time series texts, being succinctly summarized by them in the following way.

> 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 6.1, which has been adapted from Box and Jenkins' own Table 6.1.

6.16 Comparing the behavior of the sample and theoretical autocorrelation functions is by no means straightforward, particularly with small sample sizes. As was discussed in Chapter 3 (particularly **§§3.8–3.9**), 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 appear

Table 6.1 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} nonzero	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	$\phi_1 = \dfrac{\rho_2(1-\rho_2)}{1-\rho_1^2}$	$\rho_1 = \dfrac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$
	$\phi_2 = \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2}$	$\rho_2 = \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2}$
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	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$

in the sample autocorrelation function which had no place 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).

Given the behavior of the theoretical autocorrelation and partial autocorrelation functions, as shown in Table 6.1, it is also important that there are some means to judge whether their sample counterparts are effectively zero after some specific lag. Box and Jenkins (1970, section 6.2.2) suggested using the Bartlett formula (3.16), 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 that the standard error of the kth sample partial autocorrelation, which we 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.

6.17 To illustrate the identification stage of ARMA model building, we shall again use the sunspot index from 1700 to 2011, which was fitted as an AR(2) process in **§2.21**, 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 6.5. 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 (1967) and Morris (1977), although a mixed model, such as an ARMA(2, 1) process, might be appropriate.

Series A, consisting of 197 two-hourly concentration readings on a chemical process, is plotted as Figure 6.6 and appears to be stationary. The sample autocorrelation and partial autocorrelation functions are shown in Figure 6.7 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.

The likelihood function of an ARMA model

6.18 Having selected a particular ARMA model or, often more likely, a small set of candidate models, these now need to be fitted to the data. The advances in both computing power and numerical algorithms during the 1960s (see, for example, Hartley, 1961, and Marquardt, 1963) meant that the estimation methods outlined in, for example, **§§4.3–4.9** were quickly superseded by non-linear estimation techniques based on the *likelihood principle*.[3]

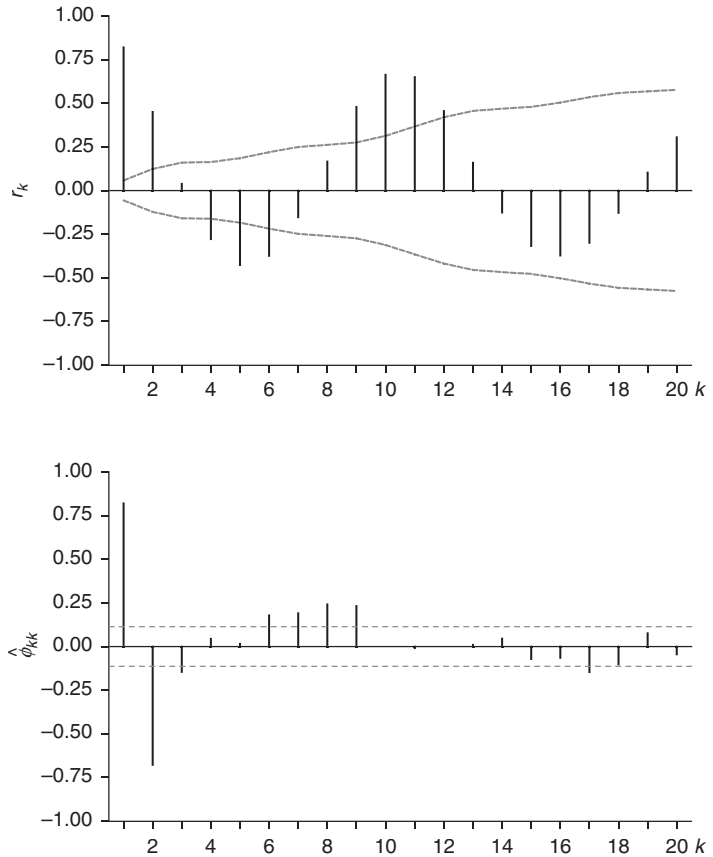


Figure 6.5 Sample autocorrelation and partial autocorrelation functions for the sunspot index with, respectively, one- and two-standard error bounds

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 which may be written

$$\begin{aligned} a_t &= x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} + \theta_1 a_{t-1} \\ &\quad + \cdots + \theta_q a_{t-q} \qquad t = 1, 2, \ldots, T \end{aligned} \tag{6.15}$$

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

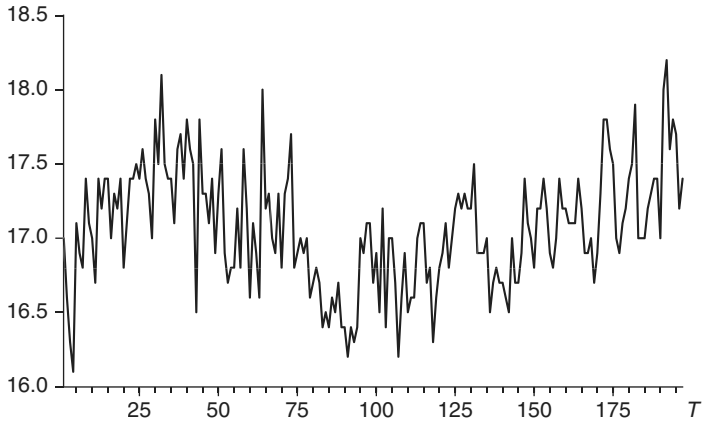


Figure 6.6 Series A from Box and Jenkins (1970): $T = 197$ two-hourly concentration readings of a chemical process

mean $\bar{X}$, but if desired it can be estimated along with the other parameters of (6.15), $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)'$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$ 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 (6.15) 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 (6.15) 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

$$\ell_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a) = -T \ln \sigma_a - \frac{S_*(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{6.16}$$

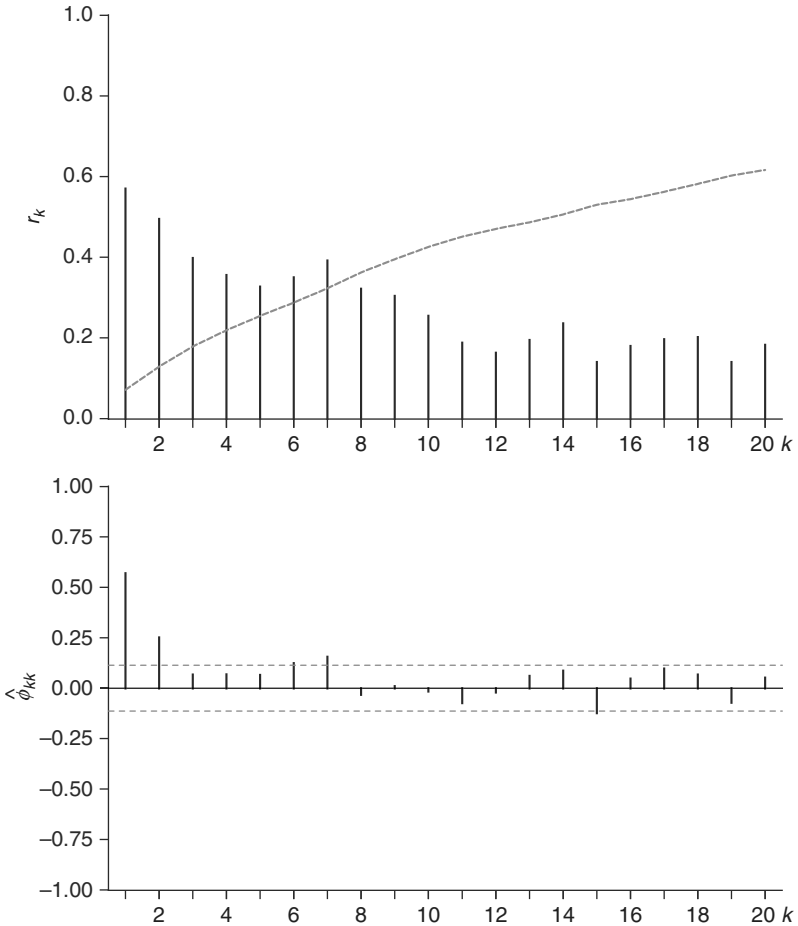


Figure 6.7 Sample autocorrelation and partial autocorrelation functions for Box and Jenkins' Series A with, respectively, one- and two-standard error bounds

where

$$S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{T} a_t^2(\boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{x}_*, \mathbf{a}_*, \mathbf{x}) \tag{6.17}$$

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 behaviour of the conditional likelihood can therefore be studied by examining the conditional sum of squares function.

6.19 Although the unconditional likelihood is strictly what is needed for parameter estimation, if T is reasonably large then a sufficiently close approximation to it is obtained by using the conditional likelihood with suitable values substituted for the elements of $\mathbf{x}_*$ and $\mathbf{a}_*$ in (6.17). 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 (cf. the condition in **§6.8**). In these circumstances the process is approaching non-stationarity 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 (6.15) 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 (6.17), 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.

6.20 The unconditional likelihood of (6.15) is given by (Box and Jenkins, 1970, chapter 7.1.4 and Appendix A7.4)

$$\ell(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a) = f(\boldsymbol{\phi}, \boldsymbol{\theta}) - T \ln \sigma_a - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{6.18}$$

where $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ is a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, and

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{T} E(a_t | \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x})^2 \tag{6.19}$$

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 (6.19), known as the least squares estimates, usually provide very close approximations to the maximum likelihood estimates.

Backward ARMA processes

6.21 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 (recall **§4.4**)

$$x_t = (1 - \theta B)a_t \qquad |\theta| < 1 \tag{6.20}$$

This has the dual, but not invertible, representation (ibid., section 6.4.2)

$$x_t = (1 - \theta^{-1}B)\alpha_t$$

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

$$x_t = (1 - \theta F)e_t \tag{6.21}$$

is then seen to be the dual of the forward process (6.20), 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:

$$\phi(B)x_t = \theta(B)a_t \tag{6.22}$$

$$\phi(F)x_t = \theta(F)e_t \tag{6.23}$$

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 (6.23), a procedure termed by Box and Jenkins as 'back forecasting' (or simply 'backcasting').

Calculating the sum of squares function

6.22 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 (6.23) to generate the backcasts

$$\phi(F)[x_t] = \theta(F)[e_t]$$

and then using (6.22) 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

$$\begin{aligned}(1 - \phi B)x_t &= (1 - \theta B)a_t \\ (1 - \phi F)x_t &= (1 - \theta F)e_t\end{aligned}$$

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

$$[e_t] = [x_t] - 0.3[x_{t+1}] + 0.7[e_{t+1}] \qquad t = 1, 2, \ldots, T - 1 = 11$$

on noting that $[e_{12}] = 0$ and $[e_t] = 0$ for $t \leq 0$. The backcasts of x_t are then generated from

$$[x_t] = 0.3[x_{t+1}] - 0.7[e_{t+1}] \qquad t = -Q, -Q + 1, \ldots, 0$$

With the starting value $[a_{-Q}] = [x_{-Q}]$, successive values of $[a_t]$ are then generated from

$$[a_t] = [x_t] - 0.3[x_{t-1}] + 0.7[a_{t-1}] \qquad t = -Q + 1, -Q + 2, \ldots, T = 12$$

and the unconditional sum of squares is calculated as

$$S(0.3, 0.7) = \sum_{t=-Q}^{T=12} [a_t^2]$$

The calculations are shown in Table 6.2 for $Q = 4$, from which we obtain $S(0.3, 0.7) = 89.16$. Box and Jenkins discussed 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 showed that little was gained by doing this and that, in general, the procedure converged very quickly.

The two conditional sums of squares suggested as approximations in **§6.19** 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 the unknown a's and e's. In the above example the unconditional expectation of x is

Table 6.2 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

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

$$\sum_{t=2}^{12} (e_t|0.3, 0.7, x_{12} = 1.1, e_{12} = 0, \mathbf{x})^2 = 82.44$$

respectively. The sum of squares using (i) is a poor approximation, although the discrepancy, which is over 10 per cent 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 eventually die out. The approximation (ii) is much more accurate and confirms Box and Jenkins' view in **§6.19** that this is the method to employ if a conditional approximation is to be used.

6.23 Figure 6.8 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

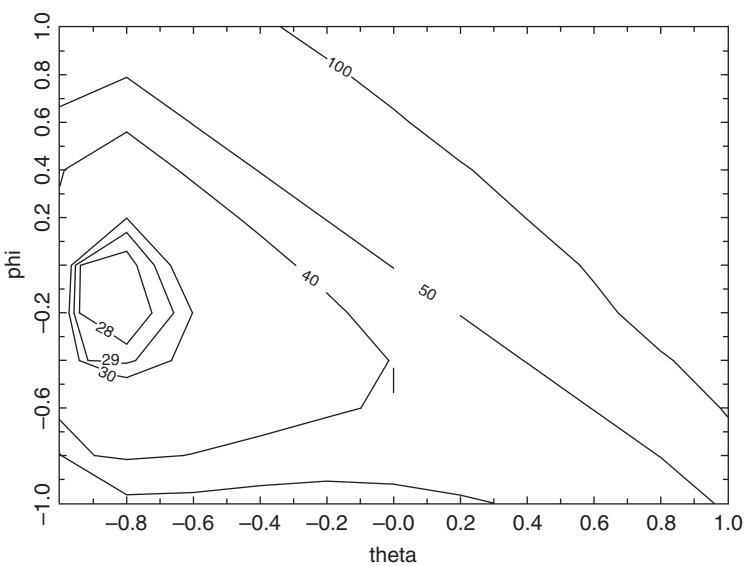


Figure 6.8 Contour plot of $S(\phi, \theta)$ calculated from the 12 values of x_t

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, would 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's value, 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 6.8 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 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

6.24 Following Box and Jenkins (1970, chapter 7.1.7), we define $\boldsymbol{\beta}$ to be a vector whose $k = p + q$ elements, β_i, $i = 1, \ldots, k$, 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

$$\ell(\boldsymbol{\xi}) = \ell(\boldsymbol{\beta}, \sigma_a) \cong \ell(\hat{\boldsymbol{\beta}}, \sigma_a) + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \ell_{ij} (\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_j)$$

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 (6.19) 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.

6.25 The *information matrix* for the $\boldsymbol{\beta}$ parameters is the $k \times k$ matrix defined by Whittle (1953) 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 (6.18),

$$\ell_{ij} \cong -\frac{S_{ij}}{2\sigma_a^2} = -\frac{1}{2\sigma_a^2} \frac{\partial^2 S(\boldsymbol{\beta}|\mathbf{x})}{\partial \beta_i \partial \beta_j}$$

so that

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \cong 2\sigma_a^2 \begin{bmatrix} \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial \beta_1^2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} \\ \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial \beta_2^2} \end{bmatrix}^{-1} = 2\sigma_a^2 \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}^{-1} \tag{6.24}$$

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

6.26 The square roots of the diagonal elements of (6.24) 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 6.8, 0.95 and 0.99 confidence regions are bounded by the contours given by

$$\begin{aligned} S_{0.95}(\phi, \theta) &= 26.05\left(1 + \frac{5.99}{12}\right) = 39.05 \\ S_{0.99}(\phi, \theta) &= 26.05\left(1 + \frac{9.21}{12}\right) = 46.04 \end{aligned}$$

These regions are shown in Figure 6.9 and, not surprisingly 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 $\mathrm{SE}(\hat{\phi}) = 0.313$ and $\mathrm{SE}(\hat{\theta}) = 0.137$.

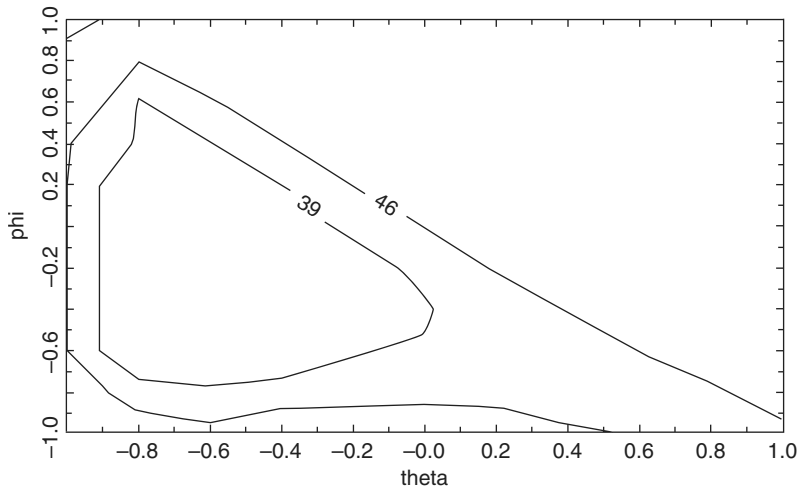


Figure 6.9 0.95 (labeled 39) and 0.99 (labeled 46) confidence regions for ϕ, θ around $(0.1, -0.9)$

Non-linear estimation

6.27 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, non-linear estimation algorithms may be applied. The need for a non-linear algorithm is seen by contrasting the autoregressive process $[a_t] = \phi(B)[x_t]$, for which

$$\frac{\partial[a_t]}{\partial\phi_i} = -[x_{t-i}] + \phi(B)\frac{\partial[x_t]}{\partial\phi_t} \tag{6.25}$$

with the moving average process $[a_t] = \theta^{-1}(B)[x_t]$, for which

$$\frac{\partial[a_t]}{\partial\theta_j} = \theta^{-2}(B)[x_{t-j}] + \theta^{-1}(B)\frac{\partial[x_t]}{\partial\theta_j} \tag{6.26}$$

In (6.25) $[x_t] = x_t$ and $\partial[x_t]/\partial\phi_j = 0$ 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 non-linear function of $\boldsymbol{\theta}$ in (6.26). 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 $\boldsymbol{\beta}_0' = (\beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0})$:

$$[a_t] = [a_{t,0}] - \sum_{i=1}^{k} (\beta_i - \beta_{i,0}) z_{i,t} \tag{6.27}$$

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 (6.27) 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 z's. 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 (6.25) and (6.26) 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 non-linear estimation of ARMA models to be carried out.

Estimated models for the sunspot index and Series A

6.28 Table 6.3 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

Table 6.3 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)
$\hat{\mu}$	50.03 (3.16)	51.31 (6.68)	51.59 (8.70)	50.03 (2.79)
$\hat{\phi}_1$	1.39 (0.04)	1.18 (0.06)	1.22 (0.04)	1.47 (0.05)
$\hat{\phi}_2$	−0.69 (0.04)	−0.40 (0.09)	−0.52 (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)	0.20 (0.03)	–
$\hat{\theta}_1$	–	–	–	0.16 (0.08)
$\hat{\sigma}_a$	16.65	15.17	15.15	16.55

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 some 16 per cent smaller) when the previously identified AR(9) model is fitted. Several autoregressive coefficients are found to be insignificant, however, and so a restricted autoregression was also estimated with the coefficients $\phi_3, \ldots, \phi_8$ set to zero, which further improves the fit.

For the ARMA(1, 1) model identified for Series A, initial estimates of ϕ and θ may be obtained by solving the expressions for ρ_1 and ρ_2 in Table 6.1 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 $\hat{\theta} \approx 0.48$. The estimated ARMA(1,1) model is

$$x_t - \underset{(\pm 0.04)}{0.92}\, x_{t-1} = 1.41 + \hat{a}_t - \underset{(\pm 0.08)}{0.61}\, a_{t-1} \qquad \hat{\sigma}_a^2 = 0.099$$

which accords well with the estimated model provided by Box and Jenkins (1970, Table 7.13).

Diagnostic checking of fitted ARMA models

6.29 The iterative model building procedure proposed by Box and Jenkins consists of three stages, the first two being 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)

6.30 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

$$\begin{array}{lll} x_t - 1.05x_{t-1} + 0.11x_{t-2} = 1.14 + a_t - 0.68a_{t-1} & & \hat{\sigma}_a^2 = 0.098 \\ \quad (\pm 0.15) \quad (\pm 0.12) \qquad\qquad (\pm 0.13) & & \\ x_t - 0.94x_{t-1} = 1.06 + a_t - 0.59a_{t-1} - 0.08a_{t-2} & & \hat{\sigma}_a^2 = 0.098 \\ \quad (\pm 0.04) \qquad\qquad (\pm 0.08) \quad (\pm 0.08) & & \end{array}$$

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 **§3.17**, 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 this estimate of the standard error 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, $Q(20) = 53.66 \sim \chi^2(18)$, 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 $Q(20) = 22.35 \sim \chi^2(17)$, 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 subsamples of the data.

6.31 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

$$\phi_0(B)x_t = \theta_0(B)b_t \tag{6.28}$$

are non-random and from their autocorrelation function the model

$$\bar{\phi}(B)b_t = \bar{\theta}(B)a_t \tag{6.29}$$

was identified. Eliminating b_t from (6.28) and (6.29) leads to the new model

$$\phi_0(B)\bar{\phi}(B)x_t = \theta_0(B)\bar{\theta}(B)a_t$$

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 $(1 - \bar{\phi}B) = b_t$, then the ARMA(2, 1) model

$$(1 - \phi B)(1 - \bar{\phi}B)x_t = (1 - \phi_1 B - \phi_2 B^2)x_t = (1 - \theta B)a_t$$

could then be fitted in a second iteration of the modelling strategy.

Forecasting using ARIMA models

6.32 Having obtained an adequate ARIMA model for a particular series, Box and Jenkins (1968, 1970, chapter 5) then proceeded to develop

a theory of forecasting for ARIMA(p, d, q) processes. They focused on the general model

$$\varphi(B)x_t = \theta(B)a_t \tag{6.30}$$

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 (6.30) can be expressed in three equivalent forms. First, it can be written directly as the difference equation

$$x_{t+l} = \varphi_1 x_{t+l-1} + \cdots + \varphi_{p+d} x_{t+l-p-d} - \theta_1 a_{t+l-1} - \cdots - \theta_q a_{t+l-q} + a_{t+l} \tag{6.31}$$

Second, it can be written as an infinite weighted sum of current and past shocks $a_{t+l}, a_{t+l-1}, \ldots$

$$x_{t+l} = \sum_{j=-\infty}^{t+l} \psi_{t+l-j} a_j = \sum_{j=0}^{\infty} \psi_j a_{t+l-j} \tag{6.32}$$

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

$$x_{t+l} = C_t(l) + a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{t-1} a_{t+1} \tag{6.33}$$

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

$$x_{t+l} = \sum_{j=1}^{\infty} \pi_j x_{t+l-j} + a_{t+l} \tag{6.34}$$

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 $d \geq 1$,

$$\bar{x}_{t+l-1}(\pi) = \sum_{j=1}^{\infty} \pi_j x_{t+l-j}$$

will be a weighted moving average, since $\sum_{j=1}^{\infty} \pi_j = 1$.

6.33 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

$$\begin{aligned} x_{t+l} &= (a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}) + (\psi_l a_t + \psi_{l+1} a_{t-1} + \cdots) \\ &= e_t(l) + \hat{x}_t(l) \end{aligned}$$

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. **§6.21**)

$$[x_{t+l}] = E[x_{t+l} | x_t, x_{t-1}, \ldots]$$

then

$$\hat{x}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots = [x_{t+l}] \tag{6.35}$$

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 refer 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.

6.34 The forecast error for lead time l is

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{l+1}$$

Since $[e_{t+l}] = 0$ the forecast is unbiased and the variance of the forecast error is

$$V(l) = Var[e_t(l)] = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \tag{6.36}$$

The one-step ahead forecast error is, from **§6.33**,

$$e_t(1) = x_{t+1} - \hat{x}_t(1) = a_{t+1}$$

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 $e_t(l + j)$, that is, 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

6.35 The forecasts from the ARIMA model (6.30) can be written down in three different ways, corresponding to the three equivalent expressions in **§6.32**. Taking conditional expectations of the difference equation (6.31) yields

$$\begin{aligned}[x_{t+l}] = \hat{x}_t(l) = {} & \varphi_1[x_{t+l-1}] + \cdots + \varphi_{p+d}[x_{t+l-p-d}] \\ & -\theta_1[a_{t+l-1}] - \cdots - \theta_q[a_{t+l-q}] + [a_{t+l}]\end{aligned}$$

while using (6.32) and (6.33), respectively, give

$$\begin{aligned}[x_{t+l}] = \hat{x}_t(l) = {} & [a_{t+l}] + \psi_1[a_{t+l-1}] + \cdots + \psi_{t-1}[a_{t+1}] + \psi_t[a_t] \\ & +\psi_{t+1}[a_{t-1}] + \cdots + [a_{t+l}]\end{aligned}$$

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 (6.34) yields

$$[x_{t+l}] = \hat{x}_{t+l} = \sum_{j=1}^{\infty} \pi_j[x_{t+l-j}] + [a_{t+l}] \tag{6.37}$$

Box and Jenkins (ibid., page 130) noted that, although the MMSE forecast was defined in terms of the conditional expectation $[x_{t+l}] = E[x_{t+l}|x_t, x_{t-1}, \dots]$, 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 [6.37] 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 $x_t, x_{t-1}, \dots, x_{t-k}$ is needed to calculate the forecasts to sufficient accuracy.

The conditional expectations can be calculated using the results

$$\begin{aligned}
[x_{t-j}] &= x_{t-j} && j = 0, 1, 2, \dots \\
[x_{t+j}] &= \hat{x}_t(j) && j = 1, 2, \dots \\
[a_{t-j}] &= a_{t-j} = x_{t-j} - \hat{x}_{t-j-1}(1) && j = 0, 1, 2, \dots \\
[a_{t+j}] &= 0 && j = 1, 2, \dots
\end{aligned}$$

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 $x_{t-j} (j = 0, 1, 2, \dots)$, which have already occurred at origin t, are left unchanged.
> The $x_{t+j} (j = 1, 2, \dots)$, which have yet to occur, are replaced by their forecasts $\hat{x}_t(j)$ at origin t.
> The $a_{t-j} (j = 0, 1, 2, \dots)$, which have occurred, are calculated as $x_{t-j} - \hat{x}_{t-j-1}(1)$.
> The $a_{t+j} (j = 1, 2, \dots)$, which have yet to occur, are replaced by their expectation of zero.

6.36 As an example of constructing ARIMA forecasts, consider Box and Jenkins' Series C, which in **§6.14** was suggested as being 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

$$\begin{aligned}
\hat{x}_t(1) &= 1.8x_t - 0.8x_{t-1} \\
\hat{x}_t(2) &= 1.8\hat{x}_t(1) - 0.8x_t \\
\hat{x}_t(l) &= 1.8\hat{x}_t(l-1) - 0.8\hat{x}_t(l-2) \qquad l = 3, 4, 5, \ldots
\end{aligned}$$

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

6.37 The ψ-weights are obtained by equating the coefficients of powers of B in

$$(1 - \varphi_1 B - \cdots - \varphi_{p+d} B^{p+d})(1 + \psi_1 B + \psi_2 B^2 + \cdots) = (1 - \theta_1 B - \cdots - \theta_q B^q)$$

that is, 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 (6.35) 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.

6.38 The expression (6.36) 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 **§§6.18–6.23**. 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 $\hat{x}_t(2)$[4]:

50% limits: $\hat{x}_t(2) \pm 0.674 \times (1 + 1.8^2)^{1/2} \times 0.134 = \hat{x}_t(2) \pm 0.19$
95% limits $\hat{x}_t(2) \pm 1.960 \times (1 + 1.8^2)^{1/2} \times 0.134 = \hat{x}_t(2) \pm 0.55$

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

6.39 At time $t + l$ the ARIMA model may be written

$$x_{t+l} - \varphi_1 x_{t+l-1} - \cdots - \varphi_{p+d} x_{t+l-p-d} = a_{t+l} - \theta_1 a_{t+l-1} - \cdots - \theta_q a_{t+l-q} \tag{6.38}$$

Taking conditional expectations at time t yields, for $l > q$,

$$\hat{x}_t(l) - \varphi_1 \hat{x}_t(l-1) - \cdots - \varphi_{p+d} \hat{x}_t(l-p-d) = 0 \qquad l > q$$

where it is understood that $\hat{x}_t(-j) = x_{t-j}$ for $j \geq 0$. This difference equation has the solution

$$\hat{x}_t(l) = b_1^{(t)} f_1(l) + b_2^{(t)} f_2(l) + \cdots + b_{p+d}^{(t)} f_{p+d}(l) \tag{6.39}$$

for $l > q - p - d$. Box and Jenkins referred to (6.39) 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 **§6.13** 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 (6.39). 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 (6.39). 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)

$$\mathbf{b}^{(t)} = (\mathbf{F}_l^{-1} \mathbf{F}_{l+1}) \mathbf{b}^{(t-1)} + (\mathbf{F}_l^{-1} \boldsymbol{\psi}_l) a_t \tag{6.40}$$

where

$$\mathbf{F}_l = \begin{bmatrix} f_1(l) & f_2(l) & \cdots & f_{p+d}(l) \\ f_1(l+1) & f_2(l+1) & \cdots & f_{p+d}(l+1) \\ \vdots & \vdots & & \vdots \\ f_1(l+p+d) & f_2(l+p+d) & \cdots & f_{p+d}(l+p+d) \end{bmatrix}$$

$$\mathbf{b}^{(t)} = \begin{bmatrix} b_1^{(t)} \\ b_2^{(t)} \\ \vdots \\ b_{p+d}^{(t)} \end{bmatrix} \qquad \boldsymbol{\psi}_l = \begin{bmatrix} \psi_l \\ \psi_{l+1} \\ \vdots \\ \psi_{l+p+d} \end{bmatrix}$$

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, that is, how the $b_j^{(t)}$ are to be calculated and updated.

> In general, since only one function of the form [6.39] 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)

6.40 Substituting for the conditional expectations in (6.37) 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

$$\pi_j^{(2)} = \pi_1 \pi_j + \pi_{j+1} \qquad j = 1, 2, \ldots$$

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

6.41 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

$$\begin{aligned} \hat{x}_t(1) &= x_t - \theta a_t \\ \hat{x}_t(l) &= \hat{x}_t(l-1) \qquad l \geq 2 \end{aligned}$$

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

$$\hat{x}_t(l) = \hat{x}_{t-1}(l) + \lambda a_t \tag{6.41}$$

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,

$$\hat{x}_t(l) = \lambda x_t + (1 - \lambda)\hat{x}_{t-1}(l) \tag{6.42}$$

> This implies that the new forecast is a linear interpolation at argument λ between old forecast and new observation. The form [6.42] 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 [6.41] 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 **§6.39**, $f_1(l) = 1$ 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 (6.40), 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

$$\pi(B) = \frac{1-B}{1-\theta B} = \frac{1-\theta B-(1-\theta)B}{1-\theta B}$$
$$= 1 - (1-\theta)(B + \theta B^2 + \theta^2 B^3 + \cdots)$$

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.

6.42 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

$$\begin{aligned}
\psi(B) &= \frac{1 - \theta_1 B - \theta_2 B^2}{(1 - B)^2} \\
&= 1 + (2 - \theta_1)B + (3 - 2\theta_1 - \theta_2)B^2 + \cdots \\
&\quad + (1 + \theta_2 + j(1 - \theta_1 - \theta_2))B^j + \cdots
\end{aligned}$$

The eventual forecast function is the solution of $(1 - B)^2 \hat{x}_t(l) = 0$, which from **§6.39** is

$$\hat{x}_t(l) = b_1^{(t)} + b_2^{(t)} l \qquad l > 0$$

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

$$\mathbf{F}_l = \mathbf{F}_{l+1} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \qquad \psi_l = \begin{bmatrix} 2 - \theta_1 \\ 3 - 2\theta_1 - \theta_2 \end{bmatrix}$$

so that (6.40) yields the following updating equations

$$\begin{aligned}
b_1^{(t+1)} &= b_1^{(t)} + b_2^{(t)} + (1 + \theta_2)a_{t+1} \\
b_2^{(t+1)} &= b_2^{(t)} + (1 - \theta_1 - \theta_2)a_{t+1}
\end{aligned}$$

The variance of the lead-l forecast is (ibid., page 149)

$$V(l) = \sigma_a^2 \left(\begin{array}{c} 1 + (l-1)(1 + \theta_2)^2 + \frac{1}{6} l(l-1)(2l-1)(1 - \theta_1 - \theta_2)^2 \\ + l(l-1)(1 + \theta_2)(1 - \theta_1 - \theta_2) \end{array} \right)$$

which again increases with l, although now in a rather complicated manner.

6.43 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

$$\hat{x}_t(l) = b_0 + b_1^{(t)} l = (l-1)\theta_0 + \frac{\theta_0}{1-\theta_1} + b_1^{(t)} l \qquad l > 0$$

where, as in **§6.41**,

$$b_1^{(t)} = b_1^{(t-1)} + (1 - \theta_1)a_t$$

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.,

$$\hat{x}_t(l) = l\theta_0 + x_t \qquad l > 0$$

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 (6.39).

6.44 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 $j \leq 0$.

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

$$\hat{x}_t(l) = b_0 + \sum_{j=1}^{p} b_j^{(t)} f_j(l) + \sum_{j=p+1}^{p+d} b_j^{(t)} l^{j-p-1} \tag{6.43}$$

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 (6.43) 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

$$\begin{aligned} \hat{x}_t(1) &= (1+\phi)x_t - \phi x_{t-1} - \theta a_t \\ \hat{x}_t(l) &= (1+\phi)\hat{x}_t(l-1) - \phi \hat{x}_t(l-2) \qquad l > 1 \end{aligned}$$

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

$$\mathbf{F}_l = \begin{bmatrix} 1 & \phi \\ 1 & \phi^2 \end{bmatrix} \qquad \mathbf{F}_{l+1} = \begin{bmatrix} 1 & \phi^2 \\ 1 & \phi^2 \end{bmatrix} \qquad \psi_l = \begin{bmatrix} \dfrac{1-\theta}{1-\phi} + \dfrac{\theta-\phi}{1-\phi}\phi \\ \dfrac{1-\theta}{1-\phi} + \dfrac{\theta-\phi}{1-\phi}\phi^2 \end{bmatrix}$$

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

$$b_1^{(t)} = x_t + \frac{\phi}{1-\phi}(x_t - x_{t-1}) - \frac{\theta}{1-\phi}a_t$$
$$b_2^{(t)} = \frac{\theta a_t - \phi(x_t - x_{t-1})}{1-\phi}$$

so that

$$\hat{x}_t(l) = x_t + \phi\frac{(1-\phi^l)}{1-\phi}(x_t - x_{t-1}) - \theta\frac{(1-\phi^l)}{1-\phi}a_t \to b_1^{(t)} \qquad \text{as } l \to \infty$$

The ARIMA(1, 1, 0) model used to forecast Series C in **§6.36** 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

$$b_1^{(t)} = b_1^{(t-1)} + 5a_t = x_t + 4(x_t - x_{t-1})$$
$$b_2^{(t)} = b_2^{(t-1)} - 4a_t = -4(x_t - x_{t-1})$$

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).

6.45 The research effort over the thirty-year period beginning with Kendall's work on oscillatory time series 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 program, it was much more than that, for it also extended the analysis to non-stationary time series and to the modeling of seasonal time series and to the relationships between series, and it is to these latter areas that we now turn to.

7

Box and Jenkins: Modelling Seasonal Time Series and Transfer Function Analysis

The Box–Jenkins approach to modelling seasonality

7.1 As was developed in some detail in Chapter 6, the Box–Jenkins approach to modelling time series revolves around the ARMA process

$$\varphi(B)x_t = \theta(B)a_t$$

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. **§6.39**). 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 non-stationary. 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 use of the differencing operator $\Delta^d = (1 - B)^d$ to effectively model homogenously non-stationary 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 (cf. **§6.11**), Box and Jenkins considered the seasonal difference operator $\Delta_s = 1 - B^s$, where s is the period of seasonality (for example, $s = 12$

for monthly data). Δ_s is a stable non-stationary 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 $(s-1)/2$ if s is odd.

7.2 When analyzing 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

$$\Phi(B^s)\Delta_s^D x_t = \Theta(B^s)\alpha_t \tag{7.1}$$

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 is introduced, this being an ARIMA(p, d, q) process for α_t

$$\phi(B)\Delta^d \alpha_t = \theta(B)a_t \tag{7.2}$$

Substituting (7.2) into (7.1) obtains the general *multiplicative* model

$$\phi(B)\Phi(B^s)\Delta^d \Delta_s^D x_t = \theta(B)\Theta(B^s)a_t \tag{7.3}$$

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'

7.3 Box and Jenkins focused their attention on the seasonal time series shown in Figure 7.1, which is Series G from Box and Jenkins (1970),

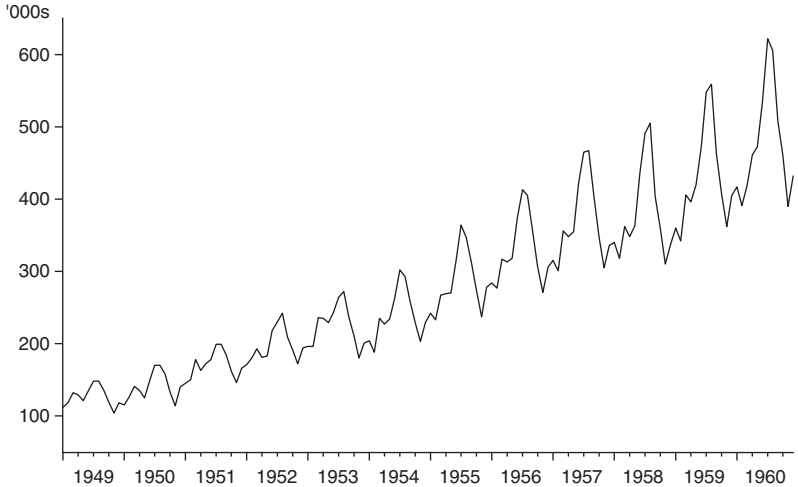


Figure 7.1 Series G from Box and Jenkins (1970): international airline passengers (in thousands), monthly, 1949–1960

originally provided by Brown (1963). These are monthly observations from 1949 to 1960 on international airline passengers and have since become a stock series for analyzing seasonality, often being referred to as the 'airline data'.

The general multiplicative model (7.3) contains a high level of generality and, in accord with their principle of parsimony, Box and Jenkins focused attention on generalizing a simple and widely applicable stochastic process for modelling non-stationary time series, the ARIMA(0, 1, 1) model, 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

$$\Delta \Delta_{12} x_t = (1 - \theta B)(1 - \Theta B^{12}) a_t \tag{7.4}$$

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 (7.4) 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'.[1]

7.4 Forecasts from (7.4) can be made directly by using the difference equation approach of **§6.35**. Thus, using the airline parameter estimates, the first three months-ahead forecasts are given by

$$\hat{x}_t(1) = x_t + x_{t-11} - x_{t-12} - 0.4\hat{a}_t - 0.6\hat{a}_{t-11} + 0.24\hat{a}_{t-12}$$
$$\hat{x}_t(2) = \hat{x}_t(1) + x_{t-10} - x_{t-11} - 0.6\hat{a}_{t-10} + 0.24\hat{a}_{t-11}$$
$$\hat{x}_t(3) = \hat{x}_t(2) + x_{t-9} - x_{t-10} - 0.6\hat{a}_{t-9} + 0.24\hat{a}_{t-10}$$

Figure 7.2 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).

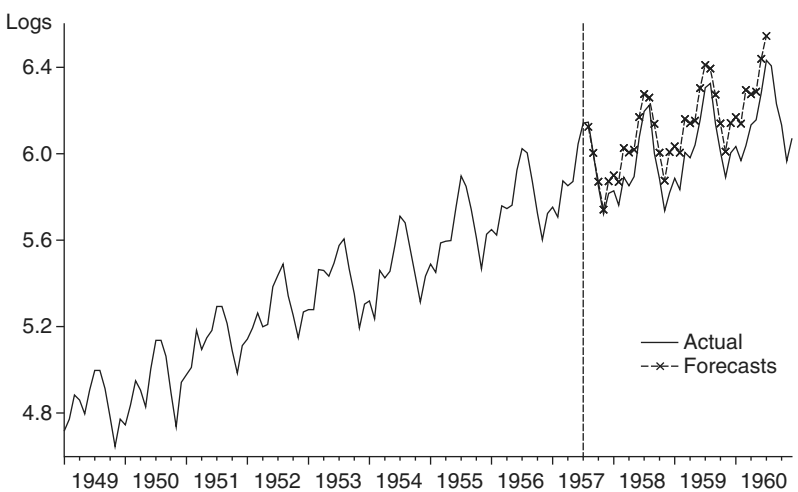


Figure 7.2 Logarithms of the airline data with forecasts for $1, 2, 3, \ldots, 36$ months ahead made from the origin July 1957

On defining $\lambda = 1 - \theta$ and $\Lambda = 1 - \Theta$, the ψ-weights of (7.4) (cf. **§6.32**) are given by

$$\psi_{12r+m} = \lambda(1 + r\Lambda) + \delta\Lambda \qquad r = 0, 1, 2, \ldots \qquad m = 1, 2, 3, \ldots, 12$$

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 (6.36) 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 $l = 36$.

7.5 The π-weights of the airline model are obtained by equating coefficients in

$$(1 - B)(1 - B^{12}) = (1 - \theta B)(1 - \Theta B^{12})(1 - \pi_1 B - \pi_2 B^2 - \cdots)$$

to give

$$\begin{aligned} \pi_j &= \theta^{j-1}(1 - \theta) && j = 1, 2, \ldots, 11 \\ \pi_{12} &= \theta^{11}(1 - \theta) + (1 - \Theta) \\ \pi_{13} &= \theta^{12}(1 - \theta) - (1 - \theta)(1 - \Theta) \\ (1 - \theta B - \Theta B^{12} + \theta\Theta B^{13})\pi_j &= 0 && j > 14 \end{aligned}$$

These are plotted in Figure 7.3 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 (7.4) can be written as

$$a_{t+1} = \left\{1 - \frac{\lambda B}{1 - \theta B}\right\}\left\{1 - \frac{\Lambda B^{12}}{1 - \Theta B^{12}}\right\} x_{t+1} \tag{7.5}$$

A useful way of rewriting (7.5) is to note that

$$\begin{aligned} \frac{\lambda}{1 - \theta B} x_t &= \lambda(1 + \theta B + \theta^2 B^2 + \cdots)x_t \\ &= \lambda(1 + (1 - \lambda)B + (1 - \lambda)^2 B^2 + \cdots)x_t \\ &= \mathrm{EWMA}_\lambda(x_t) \end{aligned}$$

where $\mathrm{EWMA}_\lambda(x_t)$ denotes an exponentially weighted moving average of x_t with parameter λ. Similarly,

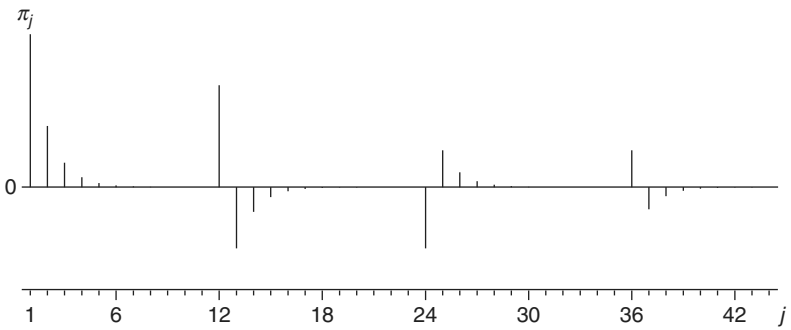


Figure 7.3 π-weights of the airline model for $\theta = 0.4$ and $\Theta = 0.6$

$$\frac{\Lambda}{1-\Theta B^{12}} x_t = \Lambda(1 + (1-\Lambda)B^{12} + (1-\Lambda)^2 B^{24} + \cdots)x_t = \text{EWMA}_\Lambda(x_t)$$

so that (7.5) can be written as

$$a_{t+1} = (1 - \text{EWMA}_\lambda(x_t))(1 - \text{EWMA}_\Lambda(x_t)B^{11})x_{t+1}$$

On substituting $\hat{x}_t(1) = x_{t+1} - a_{t+1}$, this becomes

$$\hat{x}_t(1) = \text{EWMA}_\lambda(x_t) + \text{EWMA}_\Lambda(x_{t-11} - \text{EWMA}_\lambda(x_{t-12})) \qquad (7.6)$$

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,

> 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 [7.6] would be an EWMA taken 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.

7.6 Recall from Table 6.1 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 $(0, 1, 1) \times (0, 1, 1)_{12}$ process (7.4) the only non-zero

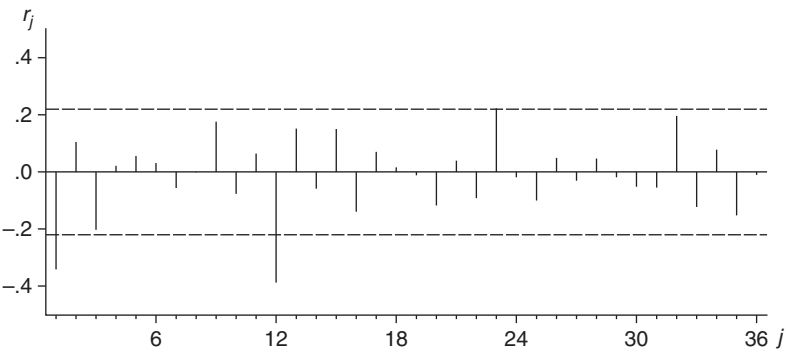


Figure 7.4 Sample autocorrelations of $\Delta\Delta_{12}x_t$ for the airline data with ±2-standard error bounds

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 7.4. On the assumption that the model is of the form (7.4), the variances for the higher-order sample autocorrelations are given by

$$V(r_j) \approx (T-13)^{-1}(1+2(\rho_1^2+\rho_{11}^2+\rho_{12}^2+\rho_{13}^2)) \qquad j > 13$$

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 7.4. 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 $(0,1,1)\times(0,1,1)_{12}$ process might provide an adequate fit to the airline data.

Seasonal ARMA models

7.7 More general seasonal ARMA models of the form (7.3) 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 6 (Box and Jenkins find no major inadequacies in the model fitted to the airline data).

Transfer function analysis

7.8 Although dynamic relationships between time series had initially been analyzed by Irving Fisher (1925) through the concept of a *distributed* lag and Kendall (1943, 1944) had discussed the cross-correlation between two time series (**§3.11–3.12**), Fisher's distributed lag concept resurfaced in the time series literature during the 1960s in the guise of the *linear transfer function model*, whose development formed chapters 10 and 11 of Box and Jenkins (1970).[2] 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*

$$\begin{aligned} Y_t &= \upsilon_0 X_t + \upsilon_1 X_{t-1} + \upsilon_2 X_{t-2} + \cdots \\ &= (\upsilon_0 + \upsilon_1 B + \upsilon_2 B^2 + \cdots) X_t \\ &= \upsilon(B) X_t \end{aligned} \tag{7.7}$$

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 (7.7), 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 (7.7) 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.

7.9 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

$$(1 + \xi_1\Delta + \cdots + \xi_r\Delta^r)Y_t = g(1 + \eta_1\Delta + \cdots + \eta_s\Delta^s)X_{t-b} \tag{7.8}$$

known as a transfer function model of order (r, s). This may also be written in terms of $B = 1 - \Delta$ as

$$(1 - \delta_1 B - \cdots - \delta_r B^r)Y_t = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s)X_{t-b} \tag{7.9}$$

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 (7.9), 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 (7.9) yields the identity

$$\begin{aligned}(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)(\upsilon_0 + \upsilon_1 B + \upsilon_2 B^2 + \cdots)\\ = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s)B^b\end{aligned}$$

On equating coefficients of B, the following relationships are obtained

$$\begin{aligned}
\upsilon_j &= 0 && j < b \\
\upsilon_j &= \delta_1\upsilon_{j-1} + \delta_2\upsilon_{j-2} + \cdots + \delta_r\upsilon_{j-r} + \omega_0 && j = b \\
\upsilon_j &= \delta_1\upsilon_{j-1} + \delta_2\upsilon_{j-2} + \cdots + \delta_r\upsilon_{j-r} - \omega_{j-b} && j = b+1, b+2, \ldots, b+s \\
\upsilon_j &= \delta_1\upsilon_{j-1} + \delta_2\upsilon_{j-2} + \cdots + \delta_r\upsilon_{j-r} && j > b+s
\end{aligned} \tag{7.10}$$

The weights $\upsilon_{b+s}, \upsilon_{b+s-1}, \ldots, \upsilon_{b+s-r+1}$ supply r starting values for the difference equation

$$\delta(B)\upsilon_j = 0 \qquad j > b+s$$

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.

7.10 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

$$\begin{aligned}
&(1 - \delta_1^*B - \delta_2^*B^2 - \cdots - \delta_{r+1}^*B^{r+1})(V_0 + V_1B + V_2B^2 + \cdots) \\
&\quad = (\omega_0 - \omega_1B - \cdots - \omega_sB^s)B^b
\end{aligned}$$

with

$$(1 - \delta_1^*B - \delta_2^*B^2 - \cdots - \delta_{r+1}^*B^{r+1}) = (1-B)(1 - \delta_1B - \delta_2B^2 - \cdots - \delta_rB^r)$$

Using (7.10), 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.

7.11 An example of the representations (7.9) and (7.10) 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 $y_t = \upsilon(B)x_t$ may be characterized thus:

Models with $r = 0$. With r and s both equal to zero, the impulse response consists of a single value $\upsilon_b = \omega_0 = g$, 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 $\upsilon_b = \omega_0, \upsilon_{b+1} = -\omega_1, \ldots, \upsilon_{b+s} = -\omega_s$. 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 $\upsilon_b = \omega_0 = g/(1+\xi_1) = g/(1-\delta_1)$. 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 $V_b = \upsilon_b$

and $V_{b-1} = 0$. For $s = 1$ the initial impulse response $\upsilon_b = \omega_0 = g(1 + \eta_1)/(1 + \xi_1)$ 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

$$\upsilon_j - \delta_1 \upsilon_{j-1} - \delta_2 \upsilon_{j-2} = 0 \qquad j > b + s \tag{7.11}$$

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 (7.11) 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 (7.11), 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.

7.12 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

7.13 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

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t \tag{7.12}$$

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 **§6.15–6.31**.

7.14 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. **§3.11**). It is assumed that, if X_t and Y_t are individually non-stationary, 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

$$\gamma_{xy}(k) = E[(x_t - \mu_x)(y_{t+k} - \mu_y)] \qquad k = 0, \pm 1, \pm 2, \pm \cdots$$

from which the *cross-correlation function* may be defined as

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \qquad k = 0, \pm 1, \pm 2, \pm \cdots$$

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

$$c_{xy}(k) = \begin{cases} \tau^{-1} \sum_{t=1}^{\tau-k} (x_t - \bar{x})(y_{t+k} - \bar{y}) & k = 0, 1, 2, \ldots \\ \tau^{-1} \sum_{t=1}^{\tau-k} (x_{t-k} - \bar{x})(y_t - \bar{y}) & k = 0, -1, -2, \ldots \end{cases}$$

from which the sample cross-correlation at lag k is defined as

$$r_{xy}(k) = \frac{c_{xy}(k)}{s_x s_y} \qquad k = 0, \pm 1, \pm 2, \pm \ldots$$

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.

7.15 Figure 7.5 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 7.6 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 be expected since an increase in the input X produces a decrease in the output Y, as can be seen from Figure 7.5.

7.16 Box and Jenkins used the following formula, originally obtained by Bartlett (1955) as an extension of the univariate formulae of (3.16), to obtain standard errors to attach to cross-correlations:

$$\begin{aligned} V[r_{xy}(k)] \approx (\tau - k)^{-1} \sum_{v=-\infty}^{+\infty} & \rho_{xx}(v)\rho_{yy}(v) + \rho_{xy}(k+v)\rho_{xy}(k-v) \\ & + \rho_{xy}^2(k)\left\{\rho_{xy}^2(v) + \tfrac{1}{2}\rho_{xx}^2(v) + \tfrac{1}{2}\rho_{yy}^2(v)\right\} \\ & - 2\rho_{xy}(k)\{\rho_{xx}(v)\rho_{xy}(v+k) + \rho_{xy}(-v)\rho_{yy}(v+k)\} \end{aligned} \tag{7.13}$$

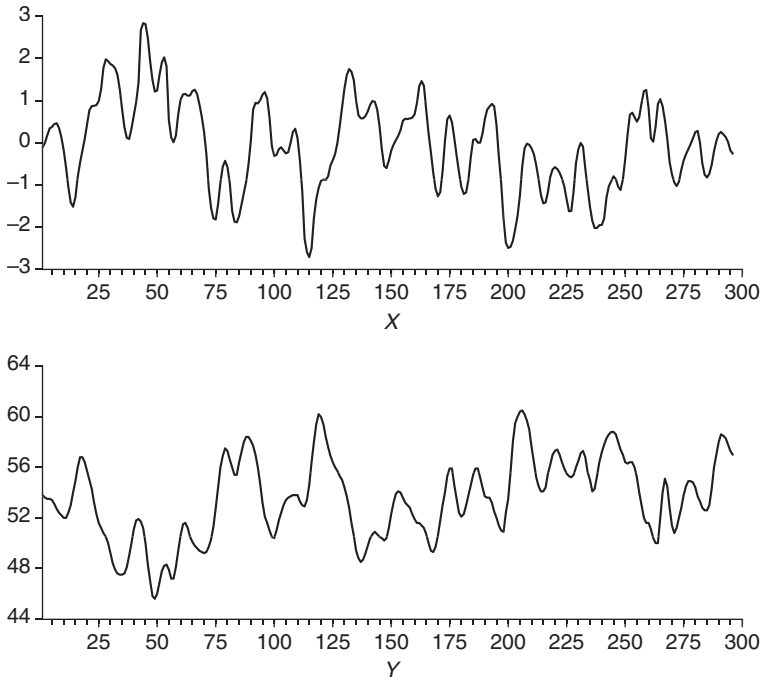


Figure 7.5 Series J from Box and Jenkins (1970): X is the input gas feed rate into a furnace; Y is the percentage output CO_2 concentration

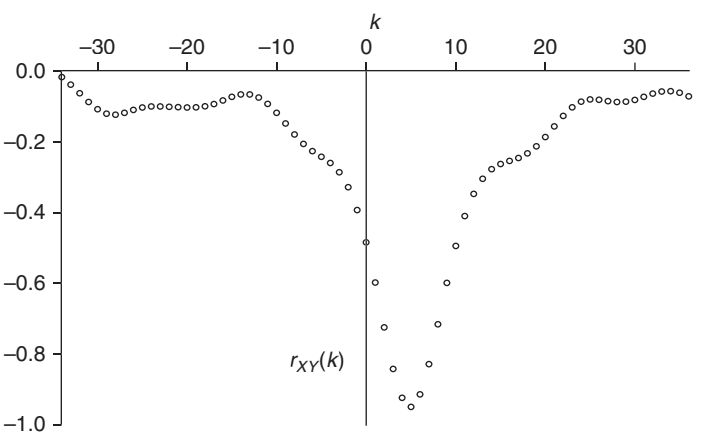


Figure 7.6 Cross-correlation function between X and Y of Figure 7.5

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 (7.13), an approximate standard error for a sample cross-correlation.

There are some interesting special cases of (7.13) that can be very useful in practical applications. For example, on the null hypothesis that x_t and y_t have *no cross-correlation*, (7.13) 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 $x_t = a_t$, 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.

7.17 The procedure for identifying a transfer function model of the form (7.12) consists of the following steps:

(i) deriving rough estimates $\hat{v}_j$ 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 $\hat{v}_j$ into equations (7.10) to obtain initial estimates of the parameters in $\delta(B)$ and $\omega(B)$.

The properties of the v_j implied by (7.10) and outlined in **§7.9** 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 **§§6.13–6.14**. Given this value of d, the model can be written as

$$y_t = v(B)x_t + n_t \tag{7.14}$$

where $n_t = \Delta^d N_t$.

Box and Jenkins argued that the identification procedure would be considerably simplified if the input series was white noise or, if x_t follows an ARMA process, if it was 'pre-whitened', i.e., 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.[3] If the same transformation is applied to y_t to obtain

$$\beta_t = \phi_x(B)\theta_x^{-1}(B)y_t$$

then (7.14) may be written

$$\beta_t = \upsilon(B)\alpha_t + \varepsilon_t \tag{7.15}$$

where $\varepsilon_t = \phi_x(B)\theta_x^{-1}(B)n_t$ is the transformed noise. Multiplying (7.15) 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 pre-whitening the input, the cross-correlation function between the pre-whitened 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

$$\hat{\upsilon}_k = \frac{r_{\alpha\beta}s_\beta}{s_\alpha}$$

7.18 To identify a transfer function model for the gas furnace data of Figure 7.5, an ARMA model for the input X_t was first obtained, this being the AR(3) process

$$(1 - 1.98B + 1.37B^2 - 0.34B^3)X_t = \alpha_t \qquad s_\alpha^2 = 0.0360$$

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 7.7, 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

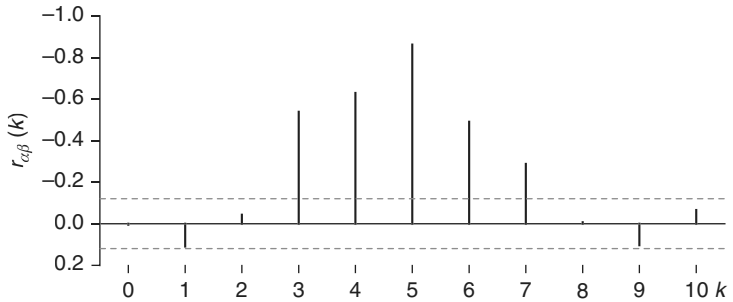


Figure 7.7 Estimated cross-correlation function for the gas furnace data

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 **§7.9**, the subsequent pattern of the $\hat{\upsilon}$'s might be accounted for by a model with (r, s, b) equal to either $(1, 2, 3)$ 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 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

$$(1 - \delta_1 B - \delta_2 B^2)Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)X_{t-3} \tag{7.16}$$

or some simplification of it. Assuming this model, the equations (7.10) become

$$\begin{aligned}
\upsilon_j &= 0 \qquad j < 3 \\
\upsilon_3 &= \omega_0 \\
\upsilon_4 &= \delta_1\upsilon_3 - \omega_1 \\
\upsilon_5 &= \delta_1\upsilon_4 + \delta_2\upsilon_3 - \omega_2 \\
\upsilon_6 &= \delta_1\upsilon_5 + \delta_2\upsilon_4 \\
\upsilon_7 &= \delta_1\upsilon_6 + \delta_2\upsilon_5
\end{aligned}$$

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

$$\hat{\omega}_0 = \hat{\upsilon}_3 = -0.54$$
$$\hat{\omega}_1 = \hat{\delta}_1\hat{\upsilon}_3 - \hat{\upsilon}_4 = (0.55)(-0.54) + 0.63 = 0.33$$
$$\hat{\omega}_2 = \hat{\delta}_1\hat{\upsilon}_4 + \hat{\delta}_2\hat{\upsilon}_3 - \hat{\upsilon}_5 = (0.55)(-0.63) + (0.02)(-0.54) + 0.86 = 0.50$$

Hence preliminary identification leads to the transfer function model

$$(1 - 0.55B - 0.02B^2)\, Y_t = -(0.54 + 0.33B + 0.50B^2)X_{t-3}$$

$\hat{\delta}_2$ is seen to be very small, suggesting that this parameter may be omitted from the model.

7.19 In general, given an estimate of the transfer function $\hat{\upsilon}(B)$, an estimate of the noise series is provided, from (7.14), by

$$\begin{aligned}\hat{n}_t &= y_t - \hat{\upsilon}(B)x_t = y_t - \hat{\delta}^{-1}(B)\hat{\omega}(B)x_{t-b} \\ &= y_t + \hat{\delta}_1(\hat{n}_{t-1} - y_{t-1}) + \hat{\delta}_2(\hat{n}_{t-2} - y_{t-2}) + \cdots + \hat{\delta}_r(\hat{n}_{t-r} - y_{t-r}) \\ &\quad - \hat{\omega}_0 x_{t-b} + \hat{\omega}_1 x_{t-b-1} + \cdots + \hat{\omega}_s x_{t-b-s}\end{aligned}$$

A straightforward approach to identifying the ARMA structure of the noise is to use the conventional identification procedure of **§§6.15–6.17** 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 $\hat{\varphi}_2 = -0.20$.[4] Thus the identified model for the gas furnace data is

$$Y_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1 - \delta_1 B - \delta_2 B^2} X_{t-3} + \frac{1}{1 - \varphi_1 B - \varphi_2 B^2} a_t$$

7.20 Box and Jenkins stressed the importance of using the rational form $\upsilon(B) = \delta^{-1}(B)\omega(B)$ for the transfer function in order 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 **§7.21** below) may still prove successful.

Box and Jenkins also emphasized the problems that may arise through *lack of uniqueness*. Since the model (7.12) 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

7.21 Box and Jenkins estimated the transfer function with noise model (7.12) using an extension of the non-linear least squares method outlined in **§§6.18–6.27**: 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 pre-whitened input α_t and the residuals $\hat{a}_t$.

The model (7.12) 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

$$y_t = \upsilon_0(B)x_t + \psi_0(B)a_{0t} \tag{7.17}$$

These residuals can be written as

$$a_{0t} = \psi_0^{-1}(B)\{\upsilon(B) - \upsilon_0(B)\}x_t + \psi_0^{-1}(B)\psi(B)a_t \tag{7.18}$$

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_t's and hence the α_t's which generate the x_t's. Two important cases need considering.

Transfer function model correct: noise model incorrect. If $\upsilon_0(B) = \upsilon(B)$ then (7.18) 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 (7.18), if the transfer function is incorrect then, as stated above, the a_{0t}'s will be autocorrelated and cross-correlated with both the x_t's and the α_t's, *even if the noise model were correct*, so that a cross-correlation analysis could indicate the modifications needed in the transfer function model.

7.22 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 (7.17) 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 discussed previously in **§6.30** 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 pre-whitened 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 distributed approximately 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.)

7.23 The transfer function model fitted to the gas furnace data was

$$\left(1 - \underset{(\pm 0.21)}{0.57}\,B - \underset{(\pm 0.14)}{0.01}\,B^2\right) Y_t = -\left(\underset{(\pm 0.08)}{0.53} + \underset{(\pm 0.15)}{0.37}\,B + \underset{(\pm 0.16)}{0.51}\,B^2\right) X_{t-3} + \frac{1}{\left(1 - \underset{(\pm 0.05)}{1.53}\,B + \underset{(\pm 0.05)}{0.63}\,B^2\right)} a_t$$

where $\pm$ 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.

The impulse response and step functions are shown in Figure 7.8: 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 and control using leading indicators

7.24 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

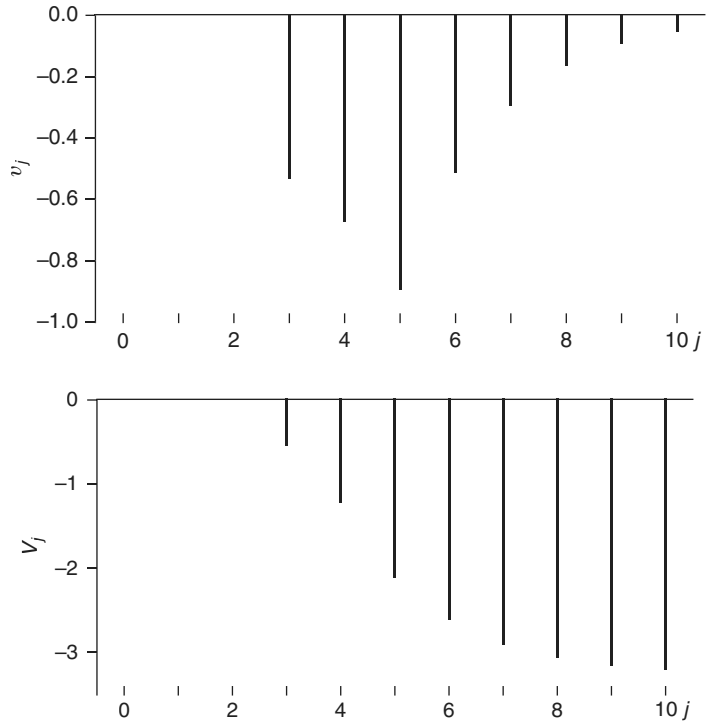


Figure 7.8 Impulse and step responses for the transfer function model $(1 - 0.57B)Y_t = -(0.53 + 0.57B + 0.51B^2)X_{t-3}$ fitted to the gas furnace data

time series, as discussed in **§6.32–6.38**, and will not be developed here: see Box and Jenkins (1970, chapter 11.5) for details. They also discussed the design of feedforward and feedback control schemes based on the transfer function–noise model: see (ibid., chapters 12 and 13).

Box and Jenkins' contribution to time series analysis

7.25 Box and Jenkins' book had an enormous impact on the time series community and its ideas quickly permeated many fields. As Mills, Tsay and Young (2011, pages 1–2) wrote in commemorating the fortieth anniversary of its publication

> after 40 years, many still regard this book as the bible of time series analysis. It not only popularizes time series applications, but also leads

> to many new developments in time series research. In an era without powerful computers and statistical software, the book provided a practical iterative procedure for modelling time series data. The concept of identification-estimation-checking continues to be as important and relevant as ever. …
>
> The airline seasonal time series model remains widely applicable for most business and economic data. Indeed, … students are equally impressed by the simplicity and applicability of the model in forecasting quarterly earnings of companies across many industrial sectors. The transfer function model and pre-whitening continue to be at the forefront of studying cross-dependence over time between time series. The spirit of motivating statistical research by solving real-world problems never fades with time. In the presence of so many time series books available, one often finds oneself drawn to Box and Jenkins when one looks for important concepts and ideas in time series analysis. The book continues to be inspirational for statisticians in general and for time series researchers in particular.

A revised edition was published in 1976, which was then followed in 1994 and 2008 by the third and fourth editions, now co-authored by Gregory Reinsel, which incorporate some of the more recent developments in time series analysis that will subsequently be covered in later chapters (see, for example, Box, Jenkins and Reinsel, 2008).

8
Box and Jenkins: Developments Post-1970

Forecasting issues

8.1 Box and Jenkins' book had an immediate impact on the practice of time series analysis and it was not long before Chatfield and Prothero (1973) presented an application of their seasonal modelling procedure to short-term sales forecasting, concluding that 'with our particular set of data, the Box–Jenkins procedure has not proved satisfactory. … Clearly one cannot generalize from one set of data, but discussion with other statisticians who have tried the Box–Jenkins approach has strengthened our view that the procedure is likely to be less useful in seasonal sales forecasting than in other time-series applications particularly when the data exhibit high multiplicative seasonal variation' (ibid., page 313). They suspected that, in part, this could be due to their choice of logarithms to transform the series prior to their ARIMA analysis, and this was also the view of several of the academic discussants of the paper, although others, typically from the perspective of industrial and sales forecasting, concurred with Chatfield and Prothero that a major drawback of the Box–Jenkins approach was that 'the subjective assessment involved in choosing a model means that considerable experience is required in interpreting sample correlation functions. In addition the procedure is expensive in terms of both computing and staff time' (ibid., page 313).

Box and Jenkins (1973) commented in detail on the Chatfield and Prothero study and, at the same time, took the opportunity to respond to Kendall's (1971) book review (recall **§3.24**). Box and Jenkins were able to confirm that most of the difficulties encountered by Chatfield and Prothero were indeed a consequence of taking logarithms, which 'over-transforms' the data, and that a cube-root or fourth-root transformation was more appropriate. They also provided an explanation of the ARIMA

$(1,1,0) \times (0,1,1)_{12}$ model $(1+0.5B)\Delta\Delta_{12}z_t = (1-0.8B^{12})a_t$, where $z_t = x_t^{0.25}$, that was identified and fitted to the transformed data, which was a concern of Chatfield and Prothero. The model can be written as

$$(1+0.5B)(1-B)(z_t - \bar{z}_{t-12}) = a_t \tag{8.1}$$

where

$$\begin{aligned} z_t - \bar{z}_{t-12} = \frac{1-B^{12}}{1-0.8B^{12}} z_t &= \frac{1-0.8B^{12}-0.2B^{12}}{1-0.8B^{12}} z_t \\ &= (1-0.2(B^{12}+0.8B^{24}+\cdots))z_t \end{aligned}$$

so that

$$\begin{aligned} \bar{z}_{t-12} &= 0.20(z_{t-12}+0.80z_{t-24}+0.80^2 z_{t-36}+\cdots) \\ &= 0.20z_{t-12}+0.16z_{t-24}+0.13z_{t-36}+\cdots \end{aligned}$$

that is, it is a EWMA of observations observed one year apart. Equation (8.1) may then be written

$$(1-0.5B-0.5B^2)(z_t - \bar{z}_{12}) = a_t$$

or

$$z_t = \bar{z}_{t-12} + 0.5((z_{t-1} - \bar{z}_{t-13}) + (z_{t-2} - \bar{z}_{t-14}))$$

Thus 'to forecast (say) next June's values, take an exponentially weighted moving average of previous June figures and adjust it by the average amount that this year's May and April figures (or their forecasts if we are forecasting more than one step ahead) exceeded last year's corresponding exponentially weighted moving averages' (Box and Jenkins, 1973, page 339).

In terms of forecasting, Chatfield and Prothero did suggest the improvement of using backcasts of residuals rather than setting initial values of a_t to zero and this was subsequently incorporated into Jenkins' programs for forecasting ARIMA models.

8.2 Kendall (1971) was concerned with several aspects of the Box–Jenkins procedure, particularly the possibility that it might be difficult to distinguish between several alternative ARIMA models, in which

case 'we might as well be content with autoregressive series … taking [their order] as far as is necessary to give approximate independence in residuals' (ibid., page 452). He was also disturbed by certain aspects of the differencing process and was of the opinion that non-invertibility should not be ruled out as inadmissible: 'it does not appear to me a valid reason for rejecting the model. The series is stationary and thoroughly respectable. … A man is not the less an individual because he has two parents, four grandparents, eight great grandparents and a divergent series of ancestors' (page 451).

Box and Jenkins (1973) responded to each of these concerns. Different models could easily fit a time series equally well, especially for short series, but this was simply a reflection of them being close approximations to each other and so were basically equivalent: 'we can use either and obtain essentially the same result' (ibid., page 341). They were quite clear, however, that they did not think that approximating a low order moving average with a higher order autoregression was a good idea when the moving average parameters were not small, as is often the case, pointing out that a first-order moving average parameter of 0.8 (as was found by Chatfield and Prothero and see equation (8.1)) has an autoregressive representation which still has a lag seven coefficient of 0.21, making any autoregressive approximation necessarily heavily parameterized compared to the moving average specification. The use of a moving average (plus the assumption of invertibility) also enables the forecast memory to die out slowly while using only a small number of parameters. For Box and Jenkins, it was models that were both non-stationary and invertible that were the most useful in forecasting and for control problems. With regard to Kendall's perceived problem with differencing, they emphasized that it was *failing* to difference rather than actual differencing which tended to lead to errors, pointing towards the problems found in the Coen, Gomme and Kendall (1969) paper by Box and Newbold (1971) as examples of this.

8.3 Recalling **§§3.22–3.23**, and in particular equation (3.17), Coen et al. had used a regression model to forecast the quarterly FT ordinary share price, Y_t, using the 'inputs' $X_{1,t-6}$ and $X_{2,t-7}$, UK car production lagged six quarters and commodity prices lagged seven quarters respectively. Since they had linearly detrended the series before cross correlating them to obtain their chosen specification, Box and Newbold pointed out that the regression could more usefully be written as

$$Y_t = \alpha + \beta_0 t + \beta_1 X_{1,t-6} + \beta_2 X_{2,t-7} + n_t \tag{8.2}$$

where Coen et al. had tacitly assumed that $n_t = a_t$, a white noise process. Box and Newbold suggested that a more realistic assumption was that the noise followed an ARIMA(0, 1, 1) process, the 'noisy random walk' $\Delta n_t = a_t - \theta a_{t-1}$. Introducing this noise process into (8.2) yields the transformed model

$$y_t = \beta_0 + \beta_1 x_{1,t-6} + \beta_2 x_{2,t-7} + a_t - \theta a_{t-1}$$

where

$$y_t = \Delta Y_t \qquad x_{1,t} = \Delta X_{1,t} \qquad x_{2,t} = \Delta X_{2,t}$$

Alternatively, Box and Newbold considered the AR(2) noise structure $n_t = \phi_1 n_{t-1} + \phi_2 n_{t-2} + a_t$, pointing out that the two noise models would 'intersect' at the random walk if $\theta = \phi_2 = 0$ and $\phi_1 = 1$. The two models have estimated coefficients of, respectively,[1]

$$\begin{aligned} \hat{\theta} &= 0.06 \pm 0.15, & \hat{\beta}_0 &= 1.78 \pm 2.74 \\ \hat{\beta}_1 &= 0.00016 \pm 0.00008 & \hat{\beta}_2 &= -1.16 \pm 1.20 \end{aligned}$$

and

$$\begin{aligned} \hat{\alpha} &= 2.47 \pm 122 & \hat{\beta}_0 &= 2.42 \pm 1.05 \\ \hat{\beta}_1 &= 0.00017 \pm 0.00009 & \hat{\beta}_2 &= -1.81 \pm 1.25 \\ \hat{\phi}_1 &= 0.93 \pm 0.16 & \hat{\phi}_2 &= -0.13 \pm 0.16 \end{aligned}$$

It is clear from the estimates of the parameters of the two noise models that n_t is very close to being a random walk and, on assuming this, the other parameter estimates become

$$\hat{\beta}_0 = 1.74 \pm 2.58 \qquad \hat{\beta}_1 = 0.00017 \pm 0.00008 \qquad \hat{\beta}_2 = -1.27 \pm 1.17$$

Although $\hat{\beta}_1$ is just significant at the 5% level it is very small in magnitude, so that Coen et al.'s regression model is effectively $\Delta Y_t = a_t$, the familiar random walk model of stock prices. Box and Newbold also pointed out that, whereas Coen et al.'s equation (3.17) has a highly significant Durbin–Watson statistic (cf. **§4.2**) of 0.98, the random walk model has a Durbin–Watson statistic of 1.86, which shows no sign of residual autocorrelation. It thus follows that, by specializing the example in **§6.41**, the MMSE forecast of Y_{t+l} made at origin t is simply $Y_t(l) = Y_t$.

8.4 Box and Newbold next turned their attention to the puzzling cross-correlations between the Coen et al. series. Figure 8.1(a) shows a plot of

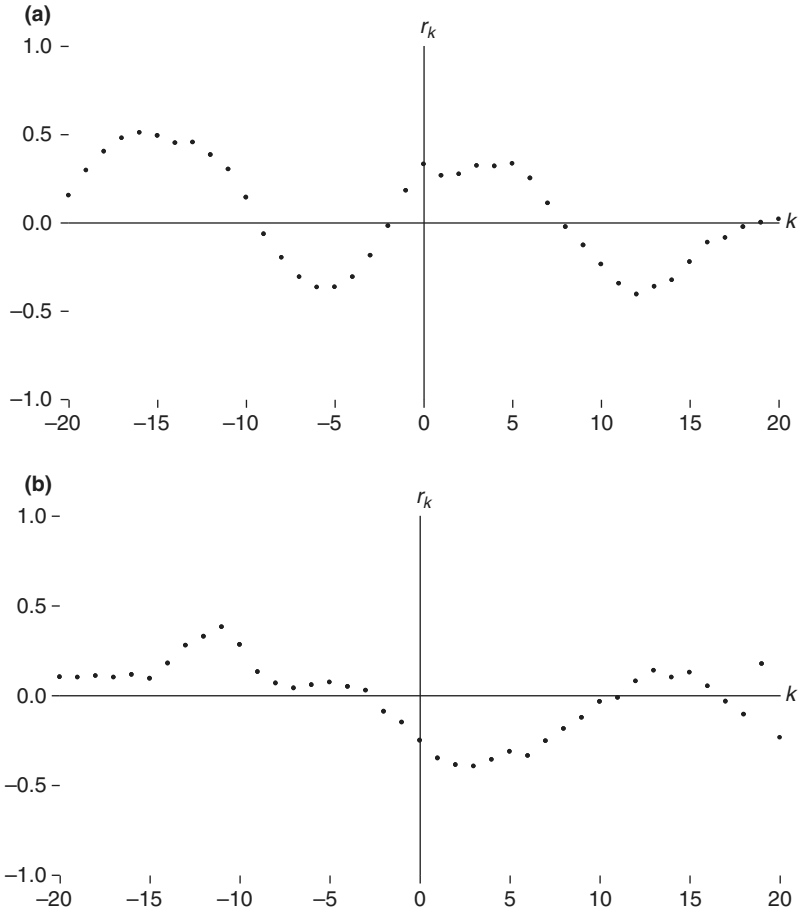


Figure 8.1 (a) Sample cross-correlations r_k between the FT share index (detrended) and lagged values of UK car production (detrended). Fifty-one pairs of observations. *(b)* Sample cross-correlations r_k between two unrelated detrended random walks. Fifty pairs of observations

the cross-correlation function between the share price (Y detrended) and car production (X_1 detrended):

> when cross-correlations with negative as well as positive lags are plotted one finds even larger cross-correlations existing at negative lags than those found in the [Coen et al.] paper at positive lags. This might suggest on the reasoning of that paper that the stock price might be

> used to forecast car production instead of vice versa. And *a priori* this seems at least equally plausible. (Box and Newbold, 1971, page 233)

To answer the question of how such cross-correlation patterns could arise, Box and Newbold considered what might happen if the individual series were not adequately represented by the linear trend model $X_t = \alpha + \beta t + a_t$, as is implied by linear detrending, but rather followed the random walk $\Delta X_t = \beta + a_t$ or, equivalently,

$$X_t = \alpha + \beta t + \sum_{j=0}^{\infty} a_{t-j}$$

To gain some insight into the behavior of cross-correlations between series generated by models of this kind, Box and Newbold conducted a small simulation experiment, which we recreate here. Five independent random walks were obtained by, in each case, cumulating 50 standard normal random deviates. These were then linearly detrended and the sample cross-correlations between each pair computed. Figure 8.1(b) shows the cross-correlations between a representative pair of these detrended random walks, while Table 8.1 reports the largest absolute values of each of the cross-correlation functions and the lag at which this correlation appeared.

> (P)ersuasive features of the cross-correlation patterns in the [Coen et al.] paper, to which the authors have drawn attention are:
>
> (i) their smoothness;
> (ii) the large absolute magnitude of the biggest cross-correlation.
>
> But it is exactly these features which are displayed by the cross-correlations of the random walks. (Box and Newbold, 1971, page 235)

Table 8.1 Largest cross-correlations (with lag in brackets) found between five detrended independent random walks

i / *j*	2	3	4	5
1	0.55 [0]	−0.41 [12]	−0.42 [18]	−0.71 [0]
2		−0.24 [0]	−0.40 [3]	−0.55 [0]
3			0.31 [−4]	−0.20 [−7]
4				0.43 [−18]

Box and Newbold then estimated regressions of the form

$$Y'_t = \alpha + \beta X'_{t-m} + a_t$$

where Y'_t and X'_t are a pair of detrended random walks and m is chosen to give the maximum cross-correlation from Table 8.1. By applying the standard t-test on β it is found that *every one* of the ten pairs of series yields a regression 'significantly different' from zero *at least* at the 10 per cent level (and eight at the 5 per cent level), even though the series are, by construction, independent. Figure 8.2 plots two of the random walks ($i = 1$, $j = 2$, having maximum cross-correlation of 0.55 at lag 0: see Table 8.1).

> The apparent relationship is partly due to the flexibility allowed in what is treated as similar – we can in effect adjust for location, spread, trend and lag before we need find similarity; partly due to the comparative smoothness of what is to be compared – to find a correlation only *a few* detrended rescaled and suitably lagged bumps have to roughly match; and partly due to the selection process – among n series there are $\frac{1}{2}n(n-1)$ pairs of series that could show such an apparent relationship. (*ibid.*, page 235)

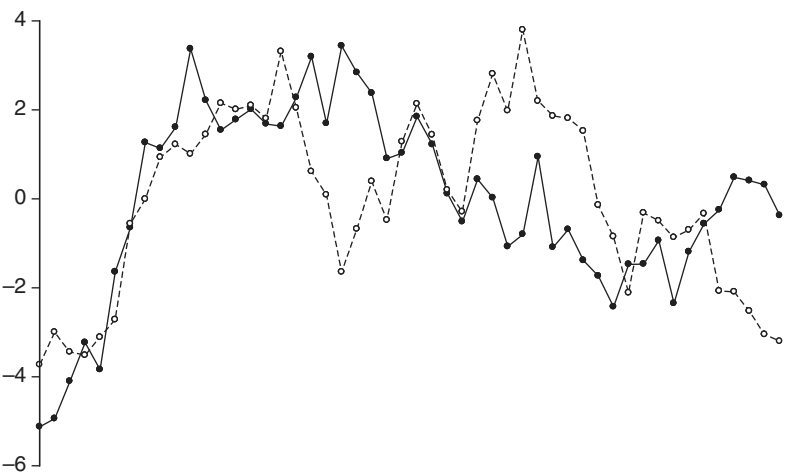


Figure 8.2 A plot of two detrended independent random walks whose maximum cross-correlation of 0.55 is at lag 0

The reason for the smoothness and large absolute magnitudes of the cross-correlations was explained by Box and Newbold as follows. Suppose we have T observations on two random walk processes

$$\Delta X_t - \beta_1 = x_t - \beta_1 = u_t$$

and

$$\Delta Y_t - \beta_2 = y_t - \beta_2 = v_t$$

where u_t and v_t are independent white noise processes, and we define the cross-covariance between the differences x_t and y_t in usual fashion as

$$C_k^* = (T-1)^{-1} \sum_{t=2}^{T-k} (x_t - \bar{x})(y_{t+k} - \bar{y})$$

Now consider the detrended series (any values or estimates of the detrending parameters will do)

$$X_t' = X_t - \alpha_1 - \beta_1 t \qquad Y_t' = Y_t - \alpha_2 - \beta_2 t$$

with cross-covariance

$$C_k = T^{-1} \sum_{t=1}^{T-k} (X_t' - \bar{X}')(Y_{t+k}' - \bar{Y}')$$

Box and Newbold showed that, for T moderate or large, to a close approximation

$$\Delta^2 C_{k+1} = -C_k^*$$

Since the C_k^* between two independent white noise processes are independently distributed about zero with constant variance, $e_k = -C_{k-1}^*$ forms a white noise process and the C_k's satisfy the difference equation $\Delta^2 C_k = e_k$, the solution of which may be written as

$$C_k = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} e_{k-i-j}$$

Hence the cross-covariances C_k follow a highly *non-stationary* stochastic process – the cumulative sum of the cumulative sum of random

deviates – so that the appearance of any particular series of cross-covariances, and hence the corresponding cross-correlations, must therefore be smooth.

> Thus, even though X and Y are generated by independent processes, their cross-correlations will wander about in a smooth pattern peculiar to each generating set of random numbers, in much the same way as was found for the economic series in the [Coen et al.] paper. This will be so irrespective of whether, or in which way, the series is detrended. (ibid., page 237)

The potentially large magnitudes of the cross-correlations follow from the result that, if X and Y are generated by unrelated first-order autoregressive processes each with parameter ϕ, then (cf. **§3.17**)

$$V(r_{XY}(k)) = T^{-1}\frac{1+\phi^2}{1-\phi^2}$$

The 'inflation factor' $(1+\phi^2)/(1-\phi^2)$ becomes large as ϕ approaches unity and X and Y approach random walks and hence large cross-correlations become very likely.

Box and Newbold were therefore able to conclude that

> Coen, Gomme and Kendall end their paper with the conclusion that their method deserves serious consideration for short-term forecasting. We have written this paper because on the contrary we believe this method should not be employed because of an innate and insidious capacity to mislead which we have discussed in some detail. (ibid., page 238)

8.5 Box and Tiao (1976) considered using forecasts to analyze a possible change in the system generating a time series.

> Suppose a system has been subjected to a change. A natural way to consider the possible effect of that change is to compare, with actuality, forecasts made from a stochastic model built on data prior to the suspected change with data actually occurring. (ibid., page 195)

Thus suppose that the general model $x_t = \psi(B)a_t$, or $\pi(B)x_t = a_t$ with $\pi(B)\psi(B) = 1$, has been fitted to data on x up to 'origin' t. If the MMSE forecast of x_{t+l} is $\hat{x}_t(l)$, $l = 1, 2, \ldots$, then from **§6.34** the lead l forecast

error $e_{t+l} = x_{t+l} - \hat{x}_t(l)$ is given by

$$e_{t+l} = \sum_{j=1}^{l} \psi_{l-j} a_{t+j}$$

where $\psi_0 = 1$. For the set of forecasts $\hat{x}_t(1), \hat{x}_t(2), \ldots, \hat{x}_t(m)$, and writing $\mathbf{a}' = (a_{t+1}, \ldots, a_{t+m})$ and $\mathbf{e}' = (e_{t+1}, \ldots, e_{t+m})$, the transformation from random shocks to forecast errors is $\mathbf{e} = \boldsymbol{\psi}\mathbf{a}$, where

$$\boldsymbol{\psi} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \psi_1 & 1 & \ddots & & & \vdots \\ \psi_2 & \psi_1 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & \vdots & & \psi_1 & 1 & 0 \\ \psi_{m-1} & \psi_{m-2} & \cdots & \psi_2 & \psi_1 & 1 \end{bmatrix}$$

Conversely, $\mathbf{a} = \boldsymbol{\pi}\mathbf{e}$, where

$$\boldsymbol{\pi} = \boldsymbol{\psi}^{-1} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -\pi_1 & 1 & \ddots & & & \vdots \\ -\pi_2 & -\pi_1 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & \vdots & & -\pi_1 & 1 & 0 \\ -\pi_{m-1} & -\pi_{m-2} & \cdots & -\pi_2 & -\pi_1 & 1 \end{bmatrix}$$

The covariance matrix of $\mathbf{e}$ is $\mathbf{V} = E(\mathbf{e}\mathbf{e}') = \boldsymbol{\psi}\boldsymbol{\psi}'\sigma_a^2$ so that it follows that, if the original model is appropriate during the period $l = 1, \ldots, m$, then $Q = \mathbf{e}'\mathbf{V}^{-1}\mathbf{e} \sim \chi^2(m)$. If, on the other hand, the model changes in some way then it might be expected that Q would be inflated to a value above some critical value of the distribution. Since

$$Q = \mathbf{e}'\mathbf{V}^{-1}\mathbf{e} = \mathbf{e}'\boldsymbol{\pi}'\boldsymbol{\pi}\mathbf{e}/\sigma_a^2 = \mathbf{a}'\mathbf{a}/\sigma_a^2 = \sigma_a^{-2}\sum_{l=1}^{m} a_{t+l}^2$$

the statistic can be regarded as the standardized sum of squared one-step ahead forecast errors. The use of Q implies that nothing specific is known

about how the model might have broken down in the forecast period. If such information is available, say in terms of a change in level of x or a shift in the parameters, then Box and Tiao suggest ways in which this information may be modeled and tested.

8.6 Abraham and Box (1978) considered the distinction between 'deterministic' and ARIMA models through the associated behaviour of the forecast functions (cf. **§6.39**) of the latter models. For example, suppose at time origin t we wish to use a model to obtain a forecast $\hat{x}_t(l)$ of the value x_{t+l} l periods in the future. One possibility is to consider the deterministic model

$$x_{t+l} = \beta_0^{(t)} + \beta_1^{(t)} l + a_{t+l} \tag{8.3}$$

which has the forecast function

$$\hat{x}_t(l) = \beta_0^{(t)} + \beta_1^{(t)} l$$

This model has the properties that the variance of the forecast error is constant at σ_a^2 and independent of the lead time l and that the function does not change as the forecast origin is changed. For example, when the forecast origin is changed from t to $t+1$ the coefficients change only so as to express the same function from the new origin. Thus, from (8.3) the forecast function from origin $t+1$ is

$$\hat{x}_{t+1}(l) = \beta_0^{(t+1)} + \beta_1^{(t+1)} l \tag{8.4}$$

and the *shift formulae* for the coefficients are

$$\beta_0^{(t+1)} = \beta_0^{(t)} + \beta_1^{(t)}, \qquad \beta_1^{(t+1)} = \beta_1^{(t)}$$

On the other hand, the ARIMA$(0, 2, 2)$ process $\Delta^2 x_{t+l} = (1 - \theta_1 B - \theta_2 B^2) a_t$ has, from **§6.42**, the same forecast function (8.4) but the coefficient *updating equations*

$$\beta_0^{(t+1)} = \beta_0^{(t)} + \beta_1^{(t)} + (1 + \theta_2) a_{t+1}, \quad \beta_1^{(t+1)} = \beta_1^{(t)} + (1 - \theta_1 - \theta_2) a_{t+1} \tag{8.5}$$

so that each updating formula consists of the appropriate shift formula plus an adjustment proportional to the one-step ahead forecast error $a_{t+1} = x_{t+1} - \hat{x}_t(1)$.

It can then be seen that if $\theta_1 \to 2$ and $\theta_2 \to -1$ the updating formulae approach the shift formulae, as do the solutions of the two models. In general, if the ARMA process $\phi(B)x_{t+l} = \theta(B)a_{t+l}$ has a moving average polynomial that cancels with the autoregressive polynomial then

$$x_{t+l} = f_t(l) + a_{t+l}$$

where $f_t(l)$ may include polynomials, exponentials, trigonometric functions or any mixture of these.

Abraham and Box presented a general analysis of the factorization of difference equations and provided examples to show how these ideas may be put to practical use.

8.7 In his last journal article, Jenkins (1982) attempted to place forecasting into the wider context of decision taking within systems, which had been a concern of his for the last 25 years of his life. While the paper makes for very interesting reading across a whole range of issues, it is perhaps the following statement that, within our present context, provides the most fitting epitaph to Gwilym Jenkins, representing as it does a synthesis of his views concerning the modelling and forecasting of time series that have been developed over the last three chapters.

> (T)he history of forecasting is littered with '*ad hoc*' methods, many of which have merit in the overall scheme of things. However, … the building of statistical models for forecasting requires a similar outlook to the building of 'models' in other areas of statistics and in science generally. In particular, the model building process requires 'experience' and 'craftsmanship' which in turn requires extensive practical application and frequent interaction between theory and practice. Model building also requires a method of approach, or a *methodology*, that eventually leads to models that contain no detectable inadequacies – at least for the time being. However, in the fullness of time they will be shown to be inadequate as more evidence and understanding becomes available. On the other hand, *ad hoc* forecasting methods are attractive to many practitioners because they can be applied with little thought – so that the forecaster can use the computer to proceed from data to forecasts with speed and with the minimum of 'pain and suffering'. In contrast, statistical model building requires thought and an *interactive and iterative dialogue* between those who know about the problem and the data, the forecaster and the computer. Since thinking is a painful process, it is

> not surprising that some will want to shy away from it and use *ad hoc* methods. Quite apart from their arbitrariness, *ad hoc* methods, mainly based on forecasting a time series from its past history, are not much use for policy making. To be effective in this area it is necessary to gain *understanding* of the relevant system before attempting to forecast the effect of continuing with present policies, or embarking on new policies. Although *properly constructed* models which forecast a time series from its past history have a very important role to play, ... for effective policy making it is necessary to introduce policy variables into a model – again in a systematic not an *ad hoc* manner. Thus we are against '*ad hoc*' methods both for 'past history forecasting' and for forecasting involving policy variables. (ibid., page 4: italics in original)

Several examples of this philosophy of forecasting may be found in Jenkins and Alavi (1981) and in two collections of case studies, Jenkins (1979) and Jenkins and McLeod (1983), to which we will return in **§8.31**.

Estimation and diagnostic checking

8.8 The likelihood function of an ARMA process and Box and Jenkins' (1970) approach to estimation via an approximation using non-linear least squares was discussed in **§§6.18–6.28**. While this approach and other closely related procedures (see, for example, Anderson, 1975, 1977, for developments and references) were usually quite accurate, difficulties began to be encountered when the sample was small or the moving average parameters were near or on the boundary of the invertibility region, for example, when θ was close to or equal to 1 in a first-order moving average. In such circumstances evaluation of the exact likelihood, and hence exact maximum likelihood estimation, tended to provide superior estimates. Expressions for the exact likelihood of an ARMA process and methods for evaluating it were given by several authors, notably Newbold (1974) and Ansley (1979).

Ljung and Box (1979) suggested writing the ARMA(p, q) process

$$x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad t = 1, 2, \ldots, T$$

in matrix form as, recalling the notation of **§6.18**,

$$\mathbf{L}_1 \mathbf{x} = \mathbf{L}_2 \mathbf{a} + \mathbf{V}\mathbf{u}_* \tag{8.6}$$

where $\mathbf{u}_* = (\mathbf{x}'_*, \mathbf{a}'_*)'$ is the vector of $p+q$ initial values and $\mathbf{L}_1$ and $\mathbf{L}_2$ are the $T \times T$ matrices

$$\mathbf{L}_1 = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -\phi_1 & 1 & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & \vdots \\ -\phi_p & & \ddots & 1 & \ddots & & \vdots \\ 0 & \ddots & & \ddots & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -\phi_p & \cdots & -\phi_1 & 1 \end{bmatrix}$$

$$\mathbf{L}_2 = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -\theta_1 & 1 & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & \vdots \\ -\theta_q & & \ddots & 1 & \ddots & & \vdots \\ 0 & \ddots & & \ddots & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -\theta_q & \cdots & -\theta_1 & 1 \end{bmatrix}$$

$\mathbf{V}$ is the $T \times (p+q)$ matrix

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{0} \end{bmatrix}$$

with

$$\mathbf{V}_1 = \begin{bmatrix} \phi_p & \phi_{p-1} & \cdots & \phi_1 & -\theta_q & -\theta_{q-1} & \cdots & -\theta_1 \\ 0 & \phi_p & \cdots & \phi_2 & 0 & -\theta_q & \cdots & -\theta_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \phi_p & 0 & \cdots & \cdots & -\theta_q \end{bmatrix}$$

being of order $m \times (p+q)$, $m = \max(p, q)$, and $\mathbf{0}$ is the null matrix of order $(n-m) \times (p+q)$.

The unconditional, or exact, likelihood function now takes the form, under the assumption that the a_t's are normally distributed,

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a | \mathbf{x}) = (2\pi\sigma_a^2)^{-\frac{T}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \exp\left(-(2\sigma_a^2)^{-1} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x}\right)$$

where $\boldsymbol{\Sigma} = \sigma_a^{-2}\boldsymbol{\Sigma}_{\mathbf{x}}$, $\boldsymbol{\Sigma}_{\mathbf{x}}$ being the covariance matrix of $\mathbf{x}$. This may be maximized with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by minimizing the function

$$\ell_0(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}) = (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})|\boldsymbol{\Sigma}|^{1/T} \tag{8.7}$$

On writing (8.6) as $\mathbf{a} = \mathbf{M}\mathbf{x} - \mathbf{N}\mathbf{u}_*$, where $\mathbf{M} = \mathbf{L}_2^{-1}\mathbf{L}_1$ and $\mathbf{N} = \mathbf{L}_2^{-1}\mathbf{V}$, Ljung and Box applied generalized least squares theory to show that the constituents of (8.7) can be written as

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \hat{\mathbf{a}}_0'\hat{\mathbf{a}}$$

and

$$\boldsymbol{\Sigma}^{-1} = \mathbf{M}(\mathbf{I} - \mathbf{N}(\boldsymbol{\Sigma}_*^{-1} + \mathbf{N}'\mathbf{N})^{-1}\mathbf{N}')\mathbf{M}$$

where

$$\begin{aligned} \hat{\mathbf{a}}_0 &= \mathbf{M}\mathbf{x}, \qquad \hat{\mathbf{a}} = \mathbf{M}\mathbf{x} - \mathbf{N}\hat{\mathbf{u}}_* \\ \mathbf{u}_* &= (\boldsymbol{\Sigma}_*^{-1} + \mathbf{N}'\mathbf{N})^{-1}\mathbf{N}'\mathbf{M}\mathbf{x} \end{aligned}$$

Evaluation of (8.7) thus requires the computation of $\boldsymbol{\Sigma}_*$ and $\mathbf{N}$ and Ljung and Box both provided a method for doing this and showed that their approach compared favourably, in terms of speed and efficiency, with that of Ansley (1979).

8.9 Ljung and Box (1978) revisited the portmanteau statistic of Box and Pierce (1970), which had been suggested as an overall test of lack of fit in ARMA models (see **§6.30**). Davies, Trigg and Newbold (1977), amongst others, had found that the $Q(K)$ statistic could deviate considerably from the assumed $\chi^2(K - p - q)$ distribution, with suspiciously low values of the statistic appearing frequently, so that the chance of incorrectly rejecting the null hypothesis of model adequacy could be much smaller than the chosen significance level. Ljung and Box therefore examined the performance of an alternative statistic proposed by Box and Pierce, viz.,

$$Q^*(K) = T(T+2)\sum_{k=1}^{K}(T-k)^{-1}r_k^2(\hat{a})$$

They found that this modified statistic offered a marked improvement in terms of approximating the underlying χ^2 distribution, particularly

for small sample sizes, and it has since become a widely used diagnostic test for the appropriateness of the noise specification in both ARMA and transfer function models, although Davies and Newbold (1979) and Godfrey (1979) both quickly showed that the power of this test was still quite low even in the presence of severe misspecification. This prompted the development of other types of diagnostic tests, most notably the class of Lagrange multiplier tests (see, for example, Godfrey and Tremayne, 1988).

Intervention analysis and outliers

8.10 Box and Tiao (1965) investigated the problem of making inferences about a possible shift in the level of a non-stationary time series at some particular point in time. Assuming that the process generating the series is a EWMA with known parameter γ_0 and pre-shift location μ, that a shift of size δ occurs after τ_1 observations and that there are τ_2 observations after the shift, then

$$
\begin{aligned}
x_1 &= \mu + a_1 \\
x_t &= \mu + \gamma_0 \sum_{i=1}^{t-1} a_{t-i} + a_t \qquad t = 2, \ldots, \tau_1 \\
x_t &= \mu + \delta + \gamma_0 \sum_{i=1}^{t-1} a_{t-i} + a_t \qquad t = \tau_1 + 1, \ldots, \tau_1 + \tau_2
\end{aligned}
$$

Assuming that τ_1 and τ_2 are large, Box and Tiao showed that estimates of μ and δ are given by

$$
\begin{aligned}
\hat{\mu} &= \gamma_0 \sum_{j=1}^{\tau_1} (1-\gamma_0)^{j-1} x_j \\
\hat{\delta} &= \gamma_0 \left(\sum_{j=1}^{\tau_2} (1-\gamma_0)^{j-1} x_{\tau_1+j} - \sum_{j=1}^{\tau_1} (1-\gamma_0)^{\tau_1-j} x_j \right)
\end{aligned}
$$

leading to the test statistic

$$
\frac{\hat{\delta} - \delta}{s_a \sqrt{\gamma_0(2-\gamma_0)}} \sim t(\tau_1 + \tau_2 - 2)
$$

where s_a^2 is an estimate of the noise variance σ_a^2 whose formula is given by Box and Tiao (1965, equation (3.11)). The estimate $\hat{\delta}$ is thus the

difference between two EWMAs, one having maximum weight immediately prior to the shift and the other having maximum weight immediately after.

Box and Tiao considered various extensions of the model, specifically the adjustments to the estimates and test statistics when τ_1 and τ_2 are small so that truncation of the sums in the estimator formulae are needed, an extension to a first-order autoregressive model, and how inferences can be made when γ_0 is unknown.

8.11 A decade later, Box and Tiao (1975) returned to the problem of detecting a shift in a time series, now couched in terms of the question 'given a known intervention, is there evidence that change in the series of the kind expected actually occurred, and, if so, what can be said of the nature and magnitude of the change?' (page 70).[2] Box and Tiao suggested modelling interventions by using the transfer function framework developed in **§§7.8–7.23**. Recalling (7.12), a simple model with one intervention can be written as

$$Y_t = R_t + N_t = \delta^{-1}(B)\omega(B)I_t + N_t \qquad N_t = \varphi^{-1}(B)\theta(B)a_t \tag{8.8}$$

I_t is the intervention variable and typically is a 'dummy' or 'indicator' sequence taking the values 1 and 0 to denote the occurrence or not of the exogenous intervention. Various dummy variables have been found to be useful for representing different forms of intervention, the two most popular being the following.

(i) A *pulse* variable, which models an intervention lasting only for the observation τ,

$$I_t = P_t^{(\tau)} = \begin{cases} 0 & t \neq \tau \\ 1 & t = \tau \end{cases}$$

(ii) A *step* variable, which models a step change in Y_t beginning at τ,

$$I_t = S_t^{(\tau)} = \begin{cases} 0 & t < \tau \\ 1 & t \geq \tau \end{cases}$$

Thus an output step change of unknown magnitude immediately following a known step change would be modelled as $R_t = \omega B S_t^{(\tau)}$. If, however, a step change is not expected to produce an immediate response but

rather a dynamic response, a 'first-order' intervention might then be defined as

$$R_t = (\omega B/(1 - \delta B))S_t^{(\tau)}$$

The steady-state gain is $\omega/(1 - \delta)$ and, as δ approaches unity, the transfer function becomes

$$R_t = (\omega B/(1 - B))S_t^{(\tau)}$$

so that a step change in the input produces a 'ramp' response in the output.

Since $(1 - B)S_t^{(\tau)} = P_t^{(\tau)}$ any of these transfer functions could equally well be used with a unit pulse intervention. An intervention which has no lasting effect could then be modelled as

$$R_t = (\omega B/(1 - \delta B))P_t^{(\tau)}$$

with ω being the initial increase immediately following the intervention and δ being the rate of decay of this increase. More complex responses can be achieved by combining interventions. For example

$$R_t = (\omega_1 B/(1 - \delta B) + \omega_2 B/(1 - B))P_t^{(\tau)} = (\omega_1 B/(1 - \delta B))P_t^{(\tau)} + \omega_2 B S_t^{(\tau)}$$

would allow an increase of ω_2 to persist, while an intervention extending over several time intervals could be represented by a series of pulses.

8.12 The identification and estimation of intervention models essentially follow that of transfer function modelling, although care needs to be taken when identifying intervention response functions in the presence of a possibly non-stationary noise structure.

One of the examples used by Box and Tiao was the examination of the effect of price controls on US inflation. Figure 8.3 shows the latter part of a record of the monthly rate of change in the consumer price index (CPI). The complete (July 1953 to December 1972) data set contains 234 successive values, 218 of which occurred prior to the institution of price controls in August 1971. As indicated in Figure 8.3, Phase I control was applied in the three months beginning September 1971 and after that Phase II was in effect to the end of the data set.

Inspection of the autocorrelation functions of the first 218 observations and their differences prior to Phase I suggested the noise model

$$(1 - B)N_t = (1 - \theta B)a_t$$

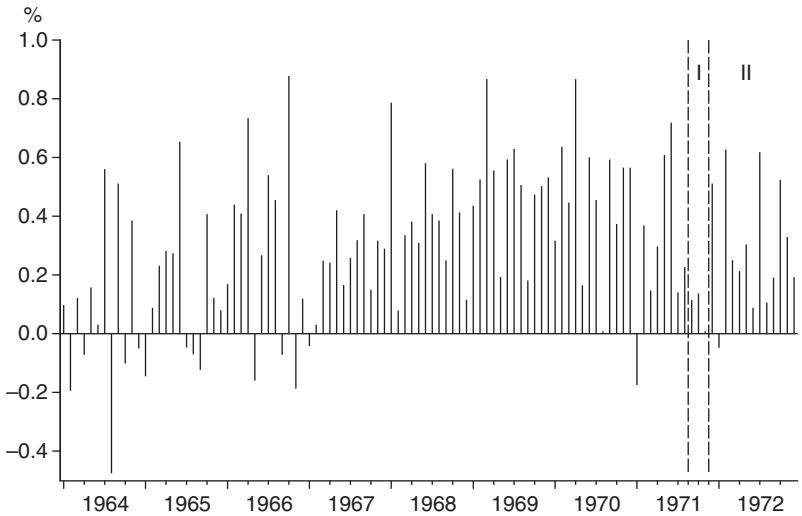


Figure 8.3 Monthly rate of inflation of the US Consumer Price Index: January 1964–December 1972. I denotes that Phase I price controls were in effect; II denotes that Phase II price controls were in effect

and estimating this model obtained $\hat{\theta} = 0.92 \pm 0.03$ and $\hat{\sigma}_a = 0.0026$. Diagnostic checks on the residuals revealed no obvious inadequacies of this model and it was thus used as a basis for answering the question 'what were the possible effects of Phases I and II?' To answer this, Box and Tiao assumed that: (i) Phases I and II could be expected to produce changes in the level of the rate of change of the CPI; and (ii) the form of the noise model remained essentially the same, thus leading to the overall model

$$Y_t = \omega_1 I_{1t} + \omega_2 I_{2t} + ((1 - \theta B)/(1 - B))a_t$$

where

$$\begin{aligned} I_{1t} &= P_t^{(219)} + P_t^{(220)} + P_t^{(221)} \\ &= \begin{cases} 1 & t = \text{September, October, November 1971} \\ 0 & \text{otherwise} \end{cases} \\ I_{2t} &= \begin{cases} 1 & t \geq \text{December 1971} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Since

$$(1-B)/(1-\theta B) = 1 - B(1-\theta)(1+\theta B+\theta^2 B^2+\cdots)$$

the model can be written as

$$z_t = \omega_1 x_{1t} + \omega_2 x_{2t} + a_t$$

where

$$z_t = Y_t - \tilde{Y}_{t-1} \qquad x_{1t} = I_{1t} - \tilde{I}_{1,t-1} \qquad x_{2t} = I_{2t} - \tilde{I}_{2,t-1}$$

$\tilde{Y}_{t-1}$, $\tilde{I}_{1,t-1}$ and $\tilde{I}_{2,t-1}$ being EWMAs of the form, for example,

$$\tilde{Y}_{t-1} = (1-\theta)(Y_{t-1} + \theta Y_{t-2} + \theta^2 Y_{t-3} + \cdots)$$

The estimated parameters of this intervention model are $\hat{\theta} = 0.93 \pm 0.02$, $\hat{\omega}_1 = 0.0030 \pm 0.0016$, $\hat{\omega}_2 = 0.0008 \pm 0.0011$ and $\hat{\sigma}_a = 0.0026$, enabling Box and Tiao to conclude that the 'analysis suggests that a real drop in the rate of increase of the CPI is associated with Phase I, but the effect of Phase II is less certain' (ibid., page 75).

8.13 Abraham and Box (1979) considered a related application of intervention variables. If the model $\varphi(B)Y_t = \theta(B)a_t$ is written $\pi(B)Y_t = a_t$, with $\pi(B) = \theta^{-1}(B)\varphi(B)$, and it is supposed that any given innovation has a small probability, say α, of being 'aberrant', then we can write

$$\pi(B)Y_t = \omega I_t + a_t$$

where $I_t = 1$ if there is an aberrant innovation at t and $I_t = 0$ otherwise. Abraham and Box refer to this as the *aberrant innovation model*. Alternatively, the aberration may affect the observation itself rather than the innovation, in which case we have

$$\pi(B)(Y_t + \omega I_t) = a_t$$

which they refer to as the *aberrant observation model*. While Abraham and Box offered a Bayesian analysis of these models under the purely autoregressive assumption $\theta(B) = 1$, the framework was subsequently extended by Hillmer, Bell and Tiao (1983) (see also Tsay, 1986, 1988, and Chang, Tiao and Chen, 1988) to allow the detection of *innovational* and *additive* outliers, respectively, occurring at unknown times.

Modelling multiple time series

8.14 The transfer function modelling strategy of **§7.8–7.23** relies on the assumption that there is no *feedback* from the output Y_t to the input X_t. Box and MacGregor (1974, 1976) analyzed the consequences of a breakdown of this assumption.[3] If both the output and input require differencing d times to induce stationarity, thus defining $y_t = \Delta^d Y_t$ and $x_t = \Delta^d X_t$, then suppose that the transfer function (cf. (7.14))

$$y_t = \upsilon(B)x_t + n_t \tag{8.8}$$

where $n_t = \phi^{-1}(B)\theta(B)a_t$ is a stationary noise, is supplemented by the feedback relationship

$$x_t = \omega(B)y_t + m_t \tag{8.9}$$

If x_t can be pre-whitened to $\alpha_t = \phi_x(B)\theta_x^{-1}(B)x_t$ and the same transformation is applied to y_t, n_t and m_t to obtain $\beta_t = \phi_x(B)\theta_x^{-1}(B)y_t$, $\varepsilon_t = \phi_x(B)\theta_x^{-1}(B)n_t$ and $\xi_t = \phi_x(B)\theta_x^{-1}(B)m_t$, then the closed-loop equations (8.8) and (8.9) can be written

$$\begin{aligned} \beta_t &= \upsilon(B)\alpha_t + \varepsilon_t \\ \alpha_t &= \omega(B)\beta_t + \xi_t \end{aligned} \tag{8.10}$$

Multiplying (8.10) by α_{t-k}, taking expectations and dividing by $\sigma_x\sigma_y$ yields

$$\begin{aligned} \rho_{\alpha\beta}(k) &= \upsilon_k\frac{\sigma_\alpha}{\sigma_\beta} + \rho_{\alpha\varepsilon}(k)\frac{\sigma_\varepsilon}{\sigma_\beta} \quad k \geq 0 \\ &= \rho_{\alpha\varepsilon}(k)\frac{\sigma_\varepsilon}{\sigma_\beta} \qquad\qquad k < 0 \end{aligned} \tag{8.11}$$

where $\rho_{\alpha\beta}(k)$ and $\rho_{\alpha\varepsilon}(k)$ are appropriate cross-correlations at lag k. If the input x_t is uncorrelated with the disturbance n_t then $\rho_{\alpha\varepsilon}(k) = 0$ and (8.11) reduces to

$$\rho_{\alpha\beta}(k) = \upsilon_k\frac{\sigma_\beta}{\sigma_\alpha} \qquad k \geq 0 \tag{8.12}$$

which is the relationship used in **§7.16** for identifying a transfer function from open-loop data. However, the relationship (8.12) will not hold under the feedback conditions (8.8) and (8.9) since x_t is related to n_t by

$$x_t = \omega(B)(1 - \upsilon(B)\omega(B))^{-1}n_t + (1 - \upsilon(B)\omega(B))^{-1}m_t$$

so that identifying the transfer function using (8.11) will almost certainly lead to an incorrect model identification.

8.15 To circumvent this problem, Haugh and Box (1977) proposed prewhitening y_t as well as x_t, i.e., obtaining the white noise series $\delta_t = \phi_y(B)\theta_y^{-1}(B)y_t$, which, of course, will typically differ from the generally autocorrelated series β_t. The series α_t and δ_t may then be cross-correlated in the usual way and a tentative model identified, which can then be recombined with the two univariate models for x_t and y_t to obtain a joint model for these two series.

The theoretical background to this approach begins with the covariance-generating function for x_t and y_t,

$$\mathbf{\Gamma}(B) = \sum_{k=-\infty}^{\infty} \mathbf{\Gamma}_k B^k = \sum_{k=-\infty}^{\infty} \begin{bmatrix} \gamma_x(k) & \gamma_{xy}(k) \\ \gamma_{yx}(k) & \gamma_y(k) \end{bmatrix} B^k$$

which is assumed to be rational (each element of $\mathbf{\Gamma}_k$ being a rational function of B). It is also assumed that $\det(\mathbf{\Gamma}(B))$ has zeros lying outside the unit circle. This ensures that the joint (x, y) process is invertible and of full rank, which eliminates the possibility, for example, of processes in which both series are transformations of the same white noise, and also allows $\mathbf{\Gamma}(B)$ to be uniquely factorized as $\mathbf{\Gamma}(B) = \mathbf{\Phi}(B^{-1})\mathbf{\Sigma}_\varepsilon \mathbf{\Phi}'(B)$, where $\mathbf{\Phi}(0) = \mathbf{I}$ and $\mathbf{\Sigma}_\varepsilon$ is the covariance matrix of MMSE one-step-ahead forecast errors.[4] Under these assumptions there will be a unique model generating the observed covariance structure. Haugh and Box showed that such a unique representation may be written as

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \mathbf{V}(B)\mathbf{a}_t = \begin{bmatrix} V_{xx}(B) & V_{xy}(B) \\ V_{yx}(B) & V_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix}$$

$$= \begin{bmatrix} \theta_{xx}(B)/\phi_{xx}(B) & C_{xy}\theta_{xy}(B)/\phi_{xy}(B) \\ C_{yx}\theta_{yx}(B)/\phi_{yx}(B) & \theta_{yy}(B)/\phi_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix}$$

where $\mathbf{V}(0) = \mathbf{V}_0$ is lower triangular with unit diagonal and

$$E(\mathbf{a}_t\mathbf{a}_t') = \mathbf{\Sigma}_a = \text{diag}(\sigma_{a_x}^2, \sigma_{a_y}^2).$$

The $\phi(B)$ polynomials have unit leading terms and roots outside the unit circle, $\theta_{xx}(B)$ and $\theta_{yy}(B)$ have unit leading terms, $\theta_{xy}(0) = 0$, and the roots of $\det(\mathbf{V}(B))$ lie outside the unit circle.

8.16 Each series can be modelled individually, leading to

$$\begin{aligned}\begin{bmatrix} x_t \\ y_t \end{bmatrix} &= \begin{bmatrix} V_x(B) & 0 \\ 0 & V_y(B) \end{bmatrix} \begin{bmatrix} u_{xt} \\ u_{yt} \end{bmatrix} \\ &= \begin{bmatrix} \theta_x(B)/\phi_x(B) & 0 \\ 0 & \theta_y(B)/\phi_y(B) \end{bmatrix} \begin{bmatrix} u_{xt} \\ u_{yt} \end{bmatrix}\end{aligned} \tag{8.13}$$

The joint process (u_x, u_y) will not, however, be bivariate white noise since u_x and u_y may be cross-correlated at non-zero lags:

$$\begin{aligned}\begin{bmatrix} u_{xt} \\ u_{yt} \end{bmatrix} = \mathbf{W}(B)\mathbf{a}_t &= \begin{bmatrix} W_{xx}(B) & W_{xy}(B) \\ W_{yx}(B) & W_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix} \\ &= \begin{bmatrix} \theta_{11}(B)/\phi_{11}(B) & C_{12}\theta_{12}(B)/\phi_{12}(B) \\ C_{21}\theta_{21}(B)/\phi_{21}(B) & \theta_{22}(B)/\phi_{22}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix}\end{aligned} \tag{8.14}$$

A complete model for x and y can then be formed by combining (8.13) and (8.14):

$$\begin{aligned}\begin{bmatrix} x_t \\ y_t \end{bmatrix} &= \begin{bmatrix} V_x(B) & 0 \\ 0 & V_y(B) \end{bmatrix} \begin{bmatrix} u_{xt} \\ u_{yt} \end{bmatrix} \\ &= \begin{bmatrix} V_x(B) & 0 \\ 0 & V_y(B) \end{bmatrix} \begin{bmatrix} W_{xx}(B) & W_{xy}(B) \\ W_{yx}(B) & W_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix} \\ &= \begin{bmatrix} V_x(B)W_{xx}(B) & V_x(B)W_{xy}(B) \\ V_y(B)W_{yx}(B) & V_y(B)W_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix}\end{aligned}$$

Haugh and Box made no claim that this factorization of $\mathbf{V}(B)$ would necessarily be parsimonious, either for identification or structural description, but they believed that this two-stage analysis would lead to more easily interpretable cross-correlation analysis, particularly as their experience, and of others at the time (for example, Pierce, 1977), was that the cross-correlation function of the univariate residual series u_x and u_y typically revealed a quite simple structure for $\mathbf{W}(B)$.

8.17 Haugh and Box's idea was then to compute the sample cross-correlation function between $\hat{u}_x$ and $\hat{u}_y$, the residual series obtained from fitting univariate models to x and y. By knowing the cross-correlation patterns appropriate to various bivariate models, a pattern

in $r_{\hat{u}_x\hat{u}_y}(\cdot)$ could then be matched to a 'true' cross-correlation pattern $\rho_{u_xu_y}(\cdot)$, which should then lead to the identification of a model of the form (8.14).

An important result for identifying models is that when there is no feedback from y to x then $V_{xy}(B) = 0$ and $\rho_{u_xu_y}(k) = 0$ for negative k. Thus if $r_{\hat{u}_x\hat{u}_y}(\cdot)$ seems to indicate no feedback effect, in that no significant cross-correlations occur at negative lags, a *dynamic regression* model may be identified as

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} V_x(B) & 0 \\ 0 & V_y(B) \end{bmatrix} \begin{bmatrix} W_{xx}(B) & 0 \\ W_{yx}(B) & W_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix}$$

$$= \begin{bmatrix} V_x(B)W_{xx}(B) & 0 \\ V_y(B)W_{yx}(B) & V_y(B)W_{yy}(B) \end{bmatrix} \begin{bmatrix} a_{xt} \\ a_{yt} \end{bmatrix}$$

In this case $W_{xx}(B) = 1$ and $a_{xt} = u_{xt}$, so that

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} V_x(B) & 0 \\ V_y(B)W_{yx}(B) & V_y(B)W_{yy}(B) \end{bmatrix} \begin{bmatrix} u_{xt} \\ a_{yt} \end{bmatrix}$$

and

$$y_t = V_y(B)W_{yx}(B)V_x^{-1}(B)x_t + V_y(B)W_{yy}(B)a_{yt}$$

is a rational distributed lag model with transfer function $\omega(B)/\delta(B) = V_y(B)W_{yx}(B)/V_x(B)$ and ARMA noise model $\theta(B)/\phi(B) = V_y(B)W_{yy}(B)$.

Haugh and Box considered four cases of primary interest.

(a) x and y are uncorrelated at all lags k, so that x will have no influence on forecasts of y and vice versa.
(b) $\rho_{u_xu_y}(0) \neq 0$ but x and y are uncorrelated at all other non-zero lags k. Although dynamic regression models for y on x and x on y can both be built, the past of x will be of no use in forecasting y beyond that already achieved by using the past of y, and again vice versa.
(c) There is at least one non-zero cross-correlation $\rho_{u_xu_y}(k)$ for positive k and there is no cross-correlation at negative lags. A dynamic regression model for y on x may then be built which will improve the forecastability of y.
(d) There exist non-zero cross-correlations $\rho_{u_xu_y}(k)$ for both positive and negative k so that there is feedback and it is not possible to build a dynamic regression model of y on present and past x.

In either (b) or (c) there will exist a dynamic regression model of the form

$$y_t = \delta^{-1}(B)\omega(B)x_t + \phi^{-1}(B)\theta(B)a_{yt}$$

The cross-correlation function $\rho_{u_x u_y}(k)$ will then have the following structure:

$$\begin{aligned}
\rho_{u_x u_y}(k) &= \sigma_{u_x}^{-1}\sigma_{u_y}^{-1}Eu_{xt}u_{yt} \\
&= \sigma_{u_x}^{-1}\sigma_{u_y}^{-1}Eu_{xt}(\theta_y^{-1}(B)\phi_y(B)y_{t+k}) \\
&= \sigma_{u_x}^{-1}\sigma_{u_y}^{-1}Eu_{xt}(\theta_y^{-1}(B)\phi_y(B)\delta^{-1}(B)\omega(B)x_{t+k} \\
&\quad + \theta_y^{-1}(B)\phi_y(B)\phi^{-1}(B)\theta(B)a_{yt}) \\
&= \sigma_{u_y}^{-1}\theta_y^{-1}(B)\phi_y(B)\delta^{-1}(B)\omega(B)\phi_x^{-1}(B)\theta_x(B)\rho_{u_x}(k)
\end{aligned}$$

from which it is clear that the cross-correlations are not proportional to the transfer function weights. However, if the analogous *dynamic shock model*,

$$u_{yt} = \delta'^{-1}(B)\omega'(B)u_{xt} + \phi'^{-1}(B)\theta'^{-1}(B)a_t = V'(B)u_{xt} + \psi'(B)a_t$$

is considered then $\rho_{u_x u_y}(k)$ takes on the simpler form

$$\rho_{u_x u_y}(k) = \sigma_{u_y}^{-1}V'(B)\rho_{u_x}(k) = \sigma_{u_y}^{-1}V'_k$$

and $\rho_{u_x u_y}(k)$ is then directly indicative of the transfer function $V'(B)$. The residual cross-correlation function $r_{\hat{u}_x\hat{u}_y}(\cdot)$ may then be used to identify $V'(B)$ using the approach of **§§7.13–7.20**.

As well as indicating the form of the transfer function $V'(B)$, the residual cross-correlation $r_{\hat{u}_x\hat{u}_y}(\cdot)$ can also be used to identify the form of the noise model $\psi'(B) = \theta'(B)/\phi'(B)$. Haugh and Box showed that, to ensure u_y is white noise, it must be the case that $\phi'(B) = \delta'(B)$ and $\theta'(B)$ is at most of order r' or s', these being the orders of $\delta'(B)$ and $\omega'(B)$ respectively.

8.18 One of the examples used by Haugh and Box to illustrate this model-building methodology employs two UK macroeconomic indicators analyzed by Bray (1971). These are the first differences of GDP (x_t) and the logarithms of unemployment (y_t), observed quarterly from 1955II to 1968IV ($T = 55$) and shown in Figure 8.4.

The AR(1) model $(1 - 0.62B)y_t = \hat{u}_{yt}$ gives a reasonable univariate fit to the y series, while the x series is adequately fitted by the simple white noise $x_t = 0.66 + \hat{u}_{xt}$. The cross-correlation function of the residual series

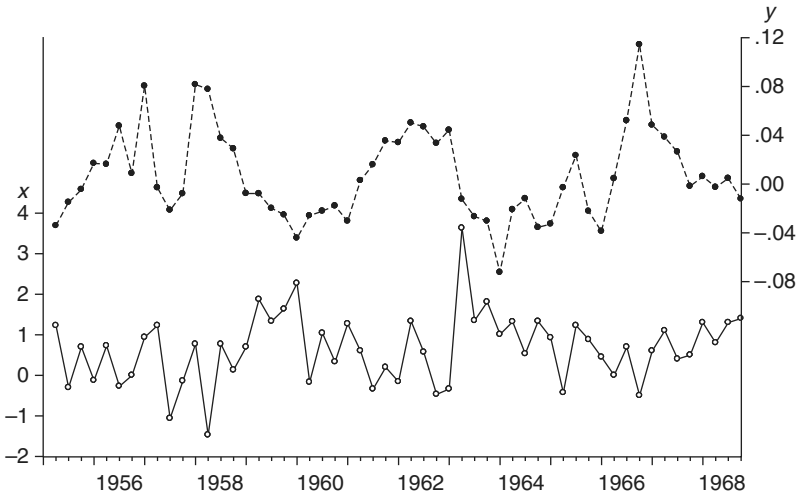


Figure 8.4 First differences of UK GDP (x) and logarithms of unemployment (y): 1955II–1968IV

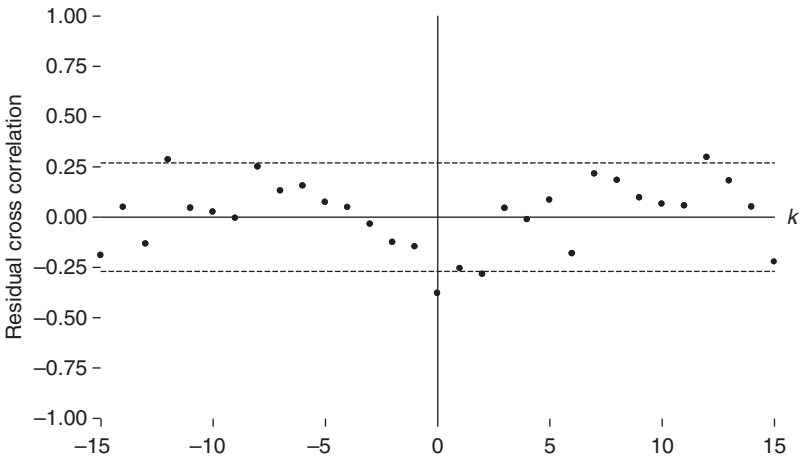


Figure 8.5 Estimated residual cross-correlation function with two standard error bounds at ±0.27

$\hat{u}_x$ and $\hat{u}_y$ is shown in Figure 8.5. Given the standard error of these cross-correlations under the assumption of independence ($1/\sqrt{T} = 0.135$), there is no evidence of feedback from $\hat{u}_y$ to $\hat{u}_x$ since no cross-correlation at negative k is significant, at least for low lags. The highest

correlation is at $k = 0$ and the cross-correlations at lags 1 and 2 are also significant. This leads to the identification of the dynamic shock model

$$u_{yt} = (\omega_0' - \omega_1' B - \omega_2' B^2) u_{xt} + (1 - \theta_1' B - \theta_2' B^2) a_{yt}$$

with the order of $\theta'(B)$ chosen to be the same as $\omega'(B)$ to insure that u_y can be specified as white noise.

Combining this dynamic shock model with the univariate models for $\hat{u}_x$ and $\hat{u}_y$ leads to the dynamic regression formulation

$$y_t = \theta_0 + \frac{(\omega_0' - \omega_1' B - \omega_2' B^2)}{(1 - \phi_1 B)} x_t + \frac{(1 - \theta_1 B - \theta_2 B^2)}{(1 - \phi_1 B)} a_{yt}$$

which was fitted by non-linear least squares to give

$$y_t = \underset{(\pm 0.007)}{0.041} - \frac{\left(\underset{(\pm 0.0044)}{0.0135} + \underset{(\pm 0.0044)}{0.0184} B + \underset{(\pm 0.0042)}{0.0182} B^2\right)}{\left(1 - \underset{(\pm 0.124)}{0.801} B\right)} x_t + \frac{\left(1 - \underset{(\pm 0.168)}{0.615} B - \underset{(\pm 0.156)}{0.354} B^2\right)}{\left(1 - \underset{(\pm 0.124)}{0.801} B\right)} a_{yt}$$

with standard errors shown in parentheses. The coefficients of the model are all precisely determined and, in addition, the estimate of the residual variance is 0.00061, which is some 30 per cent smaller than that of the univariate model for y. Usual residual auto- and cross-correlation checks (**§7.22**) indicate no model inadequacies and Haugh and Box (1977, page 128) thus concluded that 'the differenced GDP series seems to be of some value in explaining the stochastic structure of the differenced unemployment series'.

8.19 Box and Tiao (1977) explicitly considered an n-dimensional stationary autoregressive process in which there could be feedback. Denoting this $n \times 1$ vector process as $\mathbf{z}_t$, conveniently taken to have zero mean, then if $\mathbf{z}_t$ follows a pth-order autoregression it may be represented as

$$\mathbf{z}_t = \hat{\mathbf{z}}_{t-1}(1) + \mathbf{a}_t \tag{8.15}$$

where

$$\hat{\mathbf{z}}_{t-1}(1) = E(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots) = \sum_{j=1}^{p} \mathbf{\Pi}_j \mathbf{z}_{t-j}$$

is the expectation of $\mathbf{z}_t$ conditional on past history up to time $t-1$, the $\mathbf{\Pi}_j$ are $n \times n$ matrices and $\mathbf{a}_t$ is a sequence of independently and normally distributed $n \times 1$ vector random shocks, independent of $\hat{\mathbf{z}}_{t-1}(1)$ and having mean zero and positive-definite covariance matrix $\mathbf{\Sigma}$. Writing (8.15) as

$$\left(\mathbf{I} - \sum_{j=1}^{p} \mathbf{\Pi}_j B^j \right) \mathbf{z}_t = \mathbf{a}_t$$

stationarity is then ensured if $\det(\mathbf{I} - \sum \mathbf{\Pi}_j B^j)$ has all its zeros lying outside the unit circle. This model has come to be known as a vector autoregression of order p, or VAR(p). If $n = 1$, so that we have just the single series z_t, stationarity implies that

$$E(z_t^2) = E(\hat{z}_{t-1}(1))^2 + E(a_t^2)$$

or

$$\sigma_z^2 = \sigma_{\hat{z}}^2 + \sigma_a^2$$

from which can be defined the quantity $\lambda = \sigma_{\hat{z}}^2 / \sigma_z^2 = 1 - \sigma_a^2 / \sigma_z^2$, which measures the predictability of a stationary series from its past. Box and Tiao (1977, page 356) considered the multivariate generalization of this idea using the following 'thought experiment'.

> Suppose that we are considering n different stock market indicators such as the Dow Jones Average, the Standard and Poors index, etc., all of which exhibit dynamic growth. It is natural to conjecture that each might be represented as some aggregate of one or more common inputs which may be nearly nonstationary, together with other stationary or white noise components. This leads one to contemplate linear aggregates of the form $u_t = \mathbf{m}' \mathbf{z}_t$, where $\mathbf{m}$ is a $n \times 1$ vector. The aggregates which depend most heavily on the past, namely having large λ, may serve as useful composite indicators of the overall growth of the stock market. By contrast, the aggregates with λ nearly

> zero may reflect stable contemporaneous relationships among the original indicators. The analysis given [below] yields n 'canonical' components from least to most predictable. The most predictable components will often approach nonstationarity and the least predictable will be stationary or independent. Thus we may usefully decompose the n-dimensional space of the observations $\mathbf{z}_t$ into independent, stationary and nonstationary subspaces. Variables in the nonstationary space represent dynamic growth while those in the stationary and independent spaces can reflect relationships which remain stable over time. (Notation altered for consistency)

8.20 If $\mathbf{\Gamma}_j(\mathbf{z}) = E(\mathbf{z}_{t-j}\mathbf{z}_t')$ is the lag j autocovariance matrix of $\mathbf{z}_t$ then, from (8.15),

$$\mathbf{\Gamma}_0(\mathbf{z}) = \sum_{j=1}^{p} \mathbf{\Pi}_j \mathbf{\Gamma}_j(\mathbf{z}) + \mathbf{\Sigma} = \mathbf{\Gamma}_0(\hat{\mathbf{z}}) + \mathbf{\Sigma}$$

say, where $\mathbf{\Gamma}_0(\hat{\mathbf{z}})$ is the positive-definite covariance matrix of $\hat{\mathbf{z}}_{t-1}(1)$. The linear combination $u_t = \mathbf{m}'\mathbf{z}_t$ will have the property that $u_t = \hat{u}_{t-1}(1) + v_t$, where $\hat{u}_{t-1}(1) = \mathbf{m}'\hat{\mathbf{z}}_{t-1}(1)$ and $v_t = \mathbf{m}'\mathbf{a}_t$. The predictability of u_t from its past is then measured by

$$\lambda = \sigma_{\hat{u}}^2/\sigma_u^2 = (\mathbf{m}'\mathbf{\Gamma}_0(\hat{\mathbf{z}})\mathbf{m})/(\mathbf{m}'\mathbf{\Gamma}_0(\mathbf{z})\mathbf{m}) \tag{8.16}$$

from which it follows that, for maximum predictability, λ must be the largest eigenvalue of $\mathbf{\Gamma}_0^{-1}(\mathbf{z})\mathbf{\Gamma}_0(\hat{\mathbf{z}})$ and $\mathbf{m}$ the corresponding eigenvector. Similarly, the eigenvector corresponding to the smallest eigenvalue will yield the least predictable combination of $\mathbf{z}_t$.

If the n real eigenvalues are denoted $\lambda_1, \ldots, \lambda_n$ and are ordered with λ_1 being the smallest, and the corresponding linearly independent eigenvectors $\mathbf{m}_1', \ldots, \mathbf{m}_n'$ are gathered together to form the n rows of a matrix $\mathbf{M}$, then a transformed process $\mathbf{y}_t = \mathbf{M}\mathbf{z}_t$ can be constructed such that

$$\begin{aligned} \mathbf{y}_t &= \hat{\mathbf{y}}_{t-1}(1) + \mathbf{b}_t \\ \mathbf{\Gamma}_0(\mathbf{y}) &= \mathbf{\Gamma}_0(\hat{\mathbf{y}}) + \tilde{\mathbf{\Sigma}} \end{aligned} \tag{8.17}$$

where

$$\mathbf{b}_t = \mathbf{M}\mathbf{a}_t \qquad \hat{\mathbf{y}}_{t-1}(1) = \sum_{j=1}^{p} \tilde{\mathbf{\Pi}}_j \mathbf{y}_{t-j} \qquad \tilde{\mathbf{\Pi}}_j = \mathbf{M}\mathbf{\Pi}_j\mathbf{M}^{-1}$$

and

$$\boldsymbol{\Gamma}_0(\mathbf{y}) = \mathbf{M}\boldsymbol{\Gamma}_0(\mathbf{z})\mathbf{M}' \qquad \boldsymbol{\Gamma}_0(\hat{\mathbf{y}}) = \mathbf{M}\boldsymbol{\Gamma}_0(\hat{\mathbf{z}})\mathbf{M}' \qquad \tilde{\boldsymbol{\Sigma}} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}'$$

From (8.16) it then follows that

$$\mathbf{M}'^{-1}\boldsymbol{\Gamma}_0^{-1}(\mathbf{z})\boldsymbol{\Gamma}_0(\hat{\mathbf{z}})\mathbf{M}' = \boldsymbol{\Lambda} \qquad \mathbf{M}'^{-1}\boldsymbol{\Gamma}_0^{-1}(\mathbf{z})\boldsymbol{\Sigma}\mathbf{M}' = \mathbf{I} - \boldsymbol{\Lambda}$$

where $\boldsymbol{\Lambda}$ is the $n \times n$ diagonal matrix with elements $(\lambda_1, \ldots, \lambda_n)$. It is also the case that $0 < \lambda_i < 1$, $i = 1, \ldots, n$, and, for $i \neq j$, $\mathbf{m}_i'\boldsymbol{\Sigma}\mathbf{m}_j = \mathbf{m}_i'\boldsymbol{\Gamma}_0(\hat{\mathbf{z}})\mathbf{m}_j = 0$, so that $\mathbf{M}\boldsymbol{\Sigma}\mathbf{M}'$, $\mathbf{M}\boldsymbol{\Gamma}_0(\hat{\mathbf{z}})\mathbf{M}'$ and therefore $\mathbf{M}\boldsymbol{\Gamma}_0(\mathbf{z})\mathbf{M}'$ are all diagonal. Thus the transformation (8.17) produces n components $(y_{1t}, \ldots, y_{nt})$ which (i) are ordered from least predictable to most predictable; (ii) are contemporaneously independent; (iii) have predictable components $(\hat{y}_{1,t-1}(1), \ldots, \hat{y}_{n,t-1}(1))$ which are contemporaneously independent; and (iv) have unpredictable components $(b_{1t}, \ldots, b_{nt})$ which are contemporaneously and temporally independent.

8.21 If the first n_1 roots, $\lambda_1, \ldots, \lambda_{n_1}$, are zero then

$$\boldsymbol{\Gamma}_0(\hat{\mathbf{y}}) = \mathbf{M}\boldsymbol{\Gamma}_0(\hat{\mathbf{z}})\mathbf{M}' = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$$

where $\mathbf{D}$ is a $n_2 \times n_2$ diagonal matrix for $n_2 = n - n_1$. On defining the partitions $\mathbf{y}_t' = [\mathbf{y}_{1t}' \quad \mathbf{y}_{2t}']$, $\mathbf{b}_t = [\mathbf{b}_{1t}' \quad \mathbf{b}_{2t}']$ and

$$\tilde{\boldsymbol{\Pi}}_j = \begin{bmatrix} \tilde{\boldsymbol{\Pi}}_{11}^{(j)} & \tilde{\boldsymbol{\Pi}}_{12}^{(j)} \\ \tilde{\boldsymbol{\Pi}}_{21}^{(j)} & \tilde{\boldsymbol{\Pi}}_{22}^{(j)} \end{bmatrix}$$

where $\mathbf{y}_{1t}$ is $n_1 \times 1$ and $\tilde{\boldsymbol{\Pi}}_{11}^{(j)}$ is $n_1 \times n_1$, etc., $\mathbf{y}_t$ can then be written as

$$\begin{bmatrix} \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \end{bmatrix} = \sum_{j=1}^{p} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \tilde{\boldsymbol{\Pi}}_{21}^{(j)} & \tilde{\boldsymbol{\Pi}}_{22}^{(j)} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1,t-j} \\ \mathbf{y}_{2,t-j} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_{1t} \\ \mathbf{b}_{2t} \end{bmatrix} \tag{8.18}$$

Thus, the canonical transformation decomposes the original $n \times 1$ vector process $\mathbf{z}_t$ into two components: (i) a component $\mathbf{y}_{1t}$ which follows a n_1-dimensional white noise process, and (ii) a component $\mathbf{y}_{2t}$ which is stationary but whose predictable part depends on both $\mathbf{y}_{1,t-j}$ and $\mathbf{y}_{2,t-j}$ for $j = 1, \ldots, p$.

The practical importance of (8.18) is that it implies that there are n_1 relationships between the original variables having the 'static' form $m_{i1}z_{1t} + \cdots + m_{in}z_{nt} = b_{it}$, $i = 1, \ldots, n_1$, where the b_{it} are contemporaneously and temporally independent and $\mathbf{m}_i = (m_{i1}, \ldots, m_{in})$.

8.22 Box and Tiao focused attention on the VAR(1) process obtained by setting $\mathbf{\Pi}_1 = \mathbf{\Phi}$ and $\mathbf{\Pi}_j = \mathbf{0}$, $j > 1$:

$$\mathbf{z}_t = \hat{\mathbf{z}}_{t-1}(1) + \mathbf{a}_t = \mathbf{\Phi}\mathbf{z}_{t-1} + \mathbf{a}_t \tag{8.19}$$

Since now $\mathbf{\Gamma}_1'(\mathbf{z}) = \mathbf{\Phi}\mathbf{\Gamma}_0(\mathbf{z})$ it follows that $\mathbf{\Gamma}_0(\mathbf{z}) = \mathbf{\Phi}\mathbf{\Gamma}_0(\mathbf{z})\mathbf{\Phi}' + \mathbf{\Sigma}$ and the n eigenvalues λ_i and eigenvectors $\mathbf{m}_i$ are obtained from the matrix $\mathbf{\Gamma}_0^{-1}(\mathbf{z})\mathbf{\Phi}\mathbf{\Gamma}_0(\mathbf{z})\mathbf{\Phi}'$. On defining $\tilde{\mathbf{\Phi}} = \mathbf{M}\mathbf{\Phi}\mathbf{M}^{-1}$, the transformed process can be written

$$\mathbf{y}_t = \tilde{\mathbf{\Phi}}\mathbf{y}_{t-1} + \mathbf{b}_t \tag{8.20}$$

The general condition for stationarity hinges on the zeros of $\det(\mathbf{I} - \sum \mathbf{\Pi}_j B^j)$, which for (8.19) becomes

$$\det(\mathbf{I} - \mathbf{\Phi}B) = \prod_{i=1}^{n}(1 - \alpha_i)B$$

where $\alpha_1, \ldots, \alpha_n$ are the eigenvalues of $\mathbf{\Phi}$. If one or more of these eigenvalues lie on the unit circle then $\mathbf{\Gamma}_0(\mathbf{z})$ does not exist and the canonical transformation method breaks down. Suppose, however, that n_2 of the eigenvalues approach values on the unit circle. Box and Tiao showed that, under these circumstances, n_2 of the λ_i will approach unity and, in the limit, the transformed model (8.20) for $\mathbf{y}_t$ becomes

$$\begin{bmatrix} \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{\Phi}}_{11} & \mathbf{0} \\ \tilde{\mathbf{\Phi}}_{21} & \tilde{\mathbf{\Phi}}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{2,t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_{1t} \\ \mathbf{b}_{2t} \end{bmatrix}$$

The canonical transformation thus again decomposes $\mathbf{z}_t$ into two components: one, $\mathbf{y}_{1t}$, that follows a stationary first-order autoregressive process, and a second, $\mathbf{y}_{2t}$, which approaches non-stationarity and also depends upon $\mathbf{y}_{1,t-1}$.

These results can be generalized to the case where n_1 of the λ_i are zero, n_3 of them approach unity and the remaining $n_2 = n - n_1 - n_3$

are intermediate in size. Employing an obvious partitioning notation, (8.20) becomes

$$\begin{bmatrix} \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \\ \mathbf{y}_{3t} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \tilde{\mathbf{\Phi}}_{21} & \tilde{\mathbf{\Phi}}_{22} & \mathbf{0} \\ \tilde{\mathbf{\Phi}}_{31} & \tilde{\mathbf{\Phi}}_{32} & \hat{\mathbf{\Phi}}_{33} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{2,t-1} \\ \mathbf{y}_{3,t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_{1t} \\ \mathbf{b}_{2t} \\ \mathbf{b}_{3t} \end{bmatrix}$$

Thus there are three components: (i) a n_1-dimensional white noise process, $\mathbf{y}_{1t}$; a n_2-dimensional stationary process $\mathbf{y}_{2t}$ such that its predictable part depends only on $\mathbf{y}_{1,t-1}$ and $\mathbf{y}_{2,t-1}$; and (iii) a n_3-dimensional near non-stationary process $\mathbf{y}_{3t}$ such that its predictable part depends on all three lagged components.

8.23 For the ith element y_{it} of $\mathbf{y}_t$, its variance $\sigma^2_{y_i}$ can be written as

$$\sigma^2_{y_i} = \sum_{j=1}^{n} \tilde{\phi}^2_{ij} \sigma^2_{y_j} + \sigma^2_{b_i}$$

where $(\tilde{\phi}_{i1}, \ldots, \tilde{\phi}_{in})$ is the ith row of $\tilde{\mathbf{\Phi}}$. The proportional contributions of $y_{1,t-1}, \ldots, y_{n,t-1}$ and b_{it} to the variance of y_{it} are therefore $\tilde{\phi}^2_{ij}\sigma^2_{y_j}/\sigma^2_{y_i}$ and $\sigma^2_{b_i}/\sigma^2_{y_i} = 1 - \lambda_i$. A convenient scaling is one for which the variances of y_{it} are all unity. This can be achieved by choosing $\mathbf{M}$ such that $\mathbf{M}\mathbf{\Gamma}_0(\mathbf{z})\mathbf{M}' = \mathbf{I}$. In this scaling the transformed model becomes

$$\mathbf{x}_t = \hat{\mathbf{x}}_{t-1}(1) + \mathbf{d}_t = \bar{\mathbf{\Phi}}\mathbf{x}_{t-1} + \mathbf{d}_t \qquad \overline{\mathbf{\Phi}\mathbf{\Phi}}' = \mathbf{I}$$

with $\bar{\phi}^2_{ij} = \tilde{\phi}^2_{ij}\sigma^2_{y_j}/\sigma^2_{y_i}$, so that the rows of $\bar{\mathbf{\Phi}}$ are orthogonal and the sum of squares of the ith row is λ_i.

8.24 Box and Tiao used this framework to analyze Quenouille's (1957) 'hog data', which contains five series of 82 annual observations from 1867 to 1948, the data used by them being plotted in the left-hand column of Figure 8.6. On fitting a first-order model to the data, the estimated eigenvalues λ_i and eigenvectors $\mathbf{m}_i$ are shown in Table 8.2. These lead to the transformed series shown in the right-hand column of Figure 8.6, which are scaled so that all components of the transformed process $\mathbf{x}_t = \bar{\mathbf{\Phi}}\mathbf{x}_{t-1} + \mathbf{d}_t$ have unit estimated variances. The estimated $\bar{\mathbf{\Phi}}$ are shown in Table 8.3 and the estimated proportional contributions of $x_{1,t-1}, \ldots, x_{5,t-1}$ and d_{it} to x_{it} are shown in Table 8.4.

From these calculations it is seen that there is very little contribution to x_{1t} and x_{2t} from past history, so that these series are essentially white

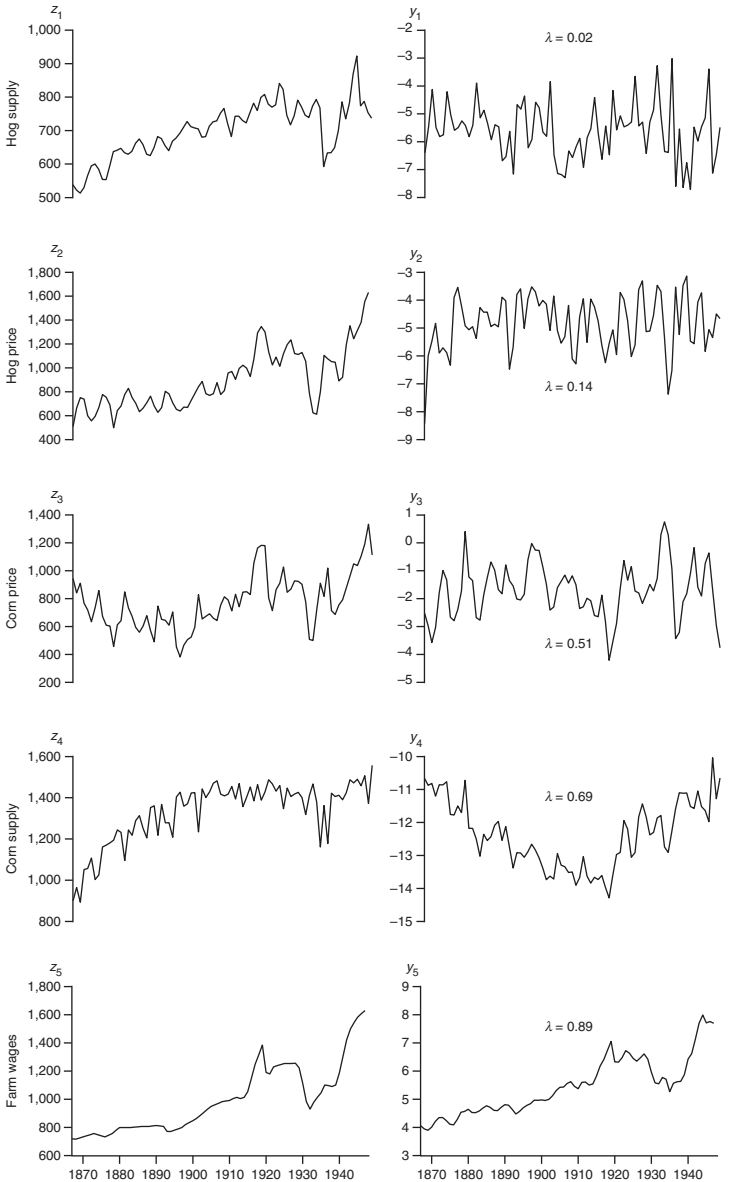


Figure 8.6 US hog data

Table 8.2 Estimated eigenvalues and eigenvectors for the hog data

i	λ_i	z_1	z_2	z_3	z_4	z_5	
1	0.0232	(1.0000	0.3876	−0.2524	−0.5896	−0.2665)	×0.0284
2	0.1421	(0.2080	1.0000	−0.8614	−0.3382	−0.3655)	×0.0111
3	0.5061	(0.8925	−0.6433	−0.8277	−0.4784	1.0000)	×0.0074
4	0.6901	(−0.9358	−0.2410	−0.4391	−0.5614	1.0000)	×0.0129
5	0.8868	(0.6687	−0.1206	−0.0134	−0.0396	1.0000)	×0.0039

Table 8.3 Estimates of the $\bar{\mathbf{\Phi}}$ matrix

$$\begin{bmatrix} 0.1213 & -0.0778 & 0.0465 & -0.0110 & 0.0113 \\ 0.2215 & 0.2766 & -0.1241 & -0.0309 & 0.0119 \\ -0.0321 & 0.3167 & 0.6334 & 0.0444 & -0.0404 \\ 0.0885 & -0.0025 & -0.0492 & 0.8235 & 0.0416 \\ -0.0801 & 0.0378 & 0.0396 & -0.0363 & 0.9360 \end{bmatrix}$$

Table 8.4 Component variances of the transformed series

	$x_{1,t-1}$	$x_{2,t-1}$	$x_{3,t-1}$	$x_{4,t-1}$	$x_{5,t-1}$	d_{it}
x_{1t}	0.015	0.006	0.002	0.000	0.000	0.977
x_{2t}	0.049	0.077	0.015	0.001	0.000	0.858
x_{3t}	0.001	0.100	0.401	0.002	0.002	0.494
x_{4t}	0.008	0.000	0.002	0.678	0.002	0.310
x_{5t}	0.006	0.001	0.002	0.001	0.876	0.113

noise. In contrast, x_{3t}, x_{4t} and x_{5t} are highly dependent on the past and, in fact, the latter pair of series can be expressed as two independent univariate first-order autoregressive processes:

$$x_{4t} = 0.82x_{4,t-1} + d_{4t} \qquad x_{5t} = 0.94x_{5,t-1} + d_{5t}$$

The model for x_{5t} implies $y_{5t} - 0.94y_{5,t-1} = 0.35 + d_{5t}$, so that the series is close to a random walk with a drift of 0.35. Since $x_{5t} = \mathbf{m}_5'\mathbf{z}_t$, with $\mathbf{m}_5'$ given in the last row of Table 8.2, it is essentially a linear combination of farm wages and hog supply,

$$x_{5t} \approx 0.0039(z_{5t} + 0.67z_{1t})$$

and serves as an indicator of the overall dynamic growth pattern in the data.

The components x_1 and x_2 both have small λ values and are nearly random. Their existence implies that any linear combination of the pair in the hyperplane

$$Z = \alpha x_1 + \beta x_2 = c_1 z_1 + c_2 z_2 + c_3 z_3 + c_4 z_4 + c_5 z_5$$

will vary nearly independently about fixed means. Noting that the data are in logarithms, Box and Tiao offered arguments that, on choosing α and β appropriately, enable the antilog of Z to be interpreted as either the ratio of return to expenditure or the ratio between hog supply and the hog price–corn price ratio, both of which are relatively stable through time.

8.25 Box and Tiao finally made the point that, if the five series were analyzed individually, all bar z_3, perhaps, would be candidates for differencing.

> However, in analyzing multiple time series of this kind, it is useful to entertain the possibility that the dynamic pattern in the data may be due to a small subset of nearly nonstationary components and that there may exist stable contemporaneous linear relationships among the variables. If this is so, then differencing all the original series could lead to complications in the analysis. (ibid., page 362)

As an example, consider the bivariate model

$$z_{1t} = z_{1,t-1} + a_{1t} \qquad z_{2t} = \beta z_{1t} + a_{2t}$$

so that each series is individually non-stationary and the model has the bivariate autoregressive representation

$$\begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 0 \end{bmatrix} \begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ \beta a_{1t} + a_{2t} \end{bmatrix}$$

If, on the other hand, the differenced series $w_{1t} = z_{1t} - z_{1,t-1}$ and $w_{2t} = z_{2t} - z_{2,t-1}$ are used then

$$w_{1t} = a_{1t} \qquad w_{2t} = \beta a_{1t} + a_{2t} - a_{2,t-1}$$

and these do not have an autoregressive representation, thus complicating the analysis.

8.26 Peña and Box (1987) considered a common factor representation of $\mathbf{z}_t$ in terms of an unobservable $r \leq n$-dimensional vector process $\mathbf{y}_t$ such that

$$\mathbf{z}_t = \mathbf{P}\mathbf{y}_t + \boldsymbol{\varepsilon}_t \tag{8.21}$$

where $\mathbf{P}$ is an $n \times r$ matrix of unknown parameters and $\boldsymbol{\varepsilon}_t$ is an n-dimensional white noise sequence with full-rank covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. If $r < n$ then the representation (8.21) leads to a reduction in dimensionality without any loss of information, while if $r = n$ (and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathbf{0}$) $\mathbf{P}$ provides a linear transformation of $\mathbf{z}_t$ that perhaps allows a simpler representation of the system. Peña and Box provided a canonical transformation of $\mathbf{z}_t$ that would do this and gave an example in which a first-order vector autoregression of five wheat price series was canonically transformed into two factors, one the mean of the five series, the second the ratio of two of them.

8.27 Tiao and Box (1981) focused attention on the vector ARMA (VARMA) model analyzed originally by Quenouille (1957):

$$\boldsymbol{\varphi}(B)\mathbf{z}_t = \boldsymbol{\theta}(B)\mathbf{a}_t$$

where

$$\boldsymbol{\varphi}(B) = \mathbf{I} - \boldsymbol{\varphi}_1 B - \cdots - \boldsymbol{\varphi}_p B^p$$

and

$$\boldsymbol{\theta}(B) = \mathbf{I} - \boldsymbol{\theta}_1 B - \cdots - \boldsymbol{\theta}_q B^q$$

are matrix polynomials in B, the $\boldsymbol{\varphi}$'s and $\boldsymbol{\theta}$'s are $n \times n$ matrices and $\mathbf{a}_t$ is an n-dimensional sequence of random shock vectors identically, independently and normally distributed with zero mean vector and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{a}}$. The zeros of the determinantal polynomials $|\boldsymbol{\varphi}(B)|$ and $|\boldsymbol{\theta}(B)|$ are assumed to lie on or outside the unit circle. If the zeros of $|\boldsymbol{\varphi}(B)|$ are all outside the unit circle then $\mathbf{z}_t$ will be stationary, while if those of $|\boldsymbol{\theta}(B)|$ are all outside the unit circle then $\mathbf{z}_t$ will be invertible. Non-stationarity may be modelled by allowing the zeros of $|\boldsymbol{\varphi}(B)|$ to lie on the unit circle, e.g. $(1 - B)\mathbf{z}_t = (\mathbf{I} - \boldsymbol{\theta}B)\mathbf{a}_t$, but Tiao and Box point out that, for vector time series, linear combinations of the elements of $\mathbf{z}_t$ may often be stationary and simultaneous differencing of all series

may lead to unnecessary complications during model fitting (recall **§8.25** and compare this with the concept of co-integration to be discussed in **§10.21–10.40**).

Under the assumption of stationarity, the lag k cross-covariance matrix is

$$\boldsymbol{\Gamma}(k) = E(\mathbf{z}_{t-k}\mathbf{z}_t') = \begin{cases} \displaystyle\sum_{j=k-r}^{k-1} \boldsymbol{\Gamma}(j)\boldsymbol{\varphi}_{k-j}' - \sum_{j=0}^{r-k} \boldsymbol{\psi}_j \boldsymbol{\Sigma}_{\mathbf{a}} \boldsymbol{\theta}_{j+k}', & k = 0, \ldots, r \\ \displaystyle\sum_{j=1}^{r} \boldsymbol{\Gamma}(k-j)\boldsymbol{\varphi}_j', & k > r \end{cases}$$

where the $\boldsymbol{\psi}_j$'s are obtained from the relationship

$$\boldsymbol{\psi}(B) = \boldsymbol{\varphi}^{-1}(B)\boldsymbol{\theta}(B) = (\mathbf{I} + \boldsymbol{\psi}_1 B + \cdots)$$

$\boldsymbol{\theta}_0 = -\mathbf{I}$, $r = \max(p, q)$ and it is understood that if $p < q$, $\boldsymbol{\varphi}_{p+1} = \cdots = \boldsymbol{\varphi}_r = \mathbf{0}$, while if $q < p$, $\boldsymbol{\theta}_{q+1} = \cdots = \boldsymbol{\theta}_r = \mathbf{0}$. The lag k cross-correlation matrix is then defined as $\boldsymbol{\rho}(k) = \boldsymbol{\Gamma}^{-1}(0)\boldsymbol{\Gamma}(k)$.

If $p = 0$, so that we have a vector MA(q) model,

$$\boldsymbol{\Gamma}(k) = \begin{cases} \displaystyle\sum_{j=0}^{q-k} \boldsymbol{\theta}_j \boldsymbol{\Sigma}_{\mathbf{a}} \boldsymbol{\theta}_{j+k}', & k = 0, \ldots, q \\ \mathbf{0} & k > q \end{cases}$$

so that all auto- and cross-correlations are zero for $k > q$. For a vector autoregressive model the auto- and cross-correlations will tend to decay gradually as $|k|$ increases.

Tiao and Box then defined the *partial autoregression matrix function* $\mathscr{P}(k)$ having the property that, if the model is VAR(p), then

$$\mathscr{P}(k) = \begin{cases} \boldsymbol{\varphi}_k, & k = p \\ \mathbf{0}, & k > p \end{cases}$$

Expressions for $\mathscr{P}(k)$ for $k < p$ are given by equation (3.11) of Tiao and Box (1981).

8.28 Tiao and Box proposed a method of identifying models of the form (8.21) using the *sample* cross-correlation and partial autoregressive matrices $\hat{\boldsymbol{\Gamma}}(k)$ and $\hat{\mathscr{P}}(k)$ in an analogous fashion to that of the univariate

case discussed in **§§6.15–6.17**. However, as plotting the sample auto- and cross-correlations against k quickly becomes cumbersome as the number of series under analysis increases, they suggested displaying the matrices with indicator symbols replacing numerical values: values in excess of $2T^{-1/2}$ being replaced by a '+', those with values less than $-2T^{-1/2}$ by a '−', and those in between these two extremes being replaced by '·'. Thus for a sample of size $T = 100$, the cross-correlation matrix

$$\begin{bmatrix} -0.28 & 0.37 & 0.06 \\ -0.21 & -0.19 & 0.12 \\ 0.46 & -0.03 & 0.15 \end{bmatrix}$$

would be represented as

$$\begin{bmatrix} - & + & \cdot \\ - & \cdot & \cdot \\ + & \cdot & \cdot \end{bmatrix}$$

Tiao and Box emphasized that, because the standard errors of the cross-correlations could be considerably greater than $T^{-1/2}$ when the series are highly autocorrelated, these symbols should not be interpreted as a formal significance test but rather as a rough guide to the general pattern of autocorrelation. A similar approach can be used for sample partial autoregression matrices, obtained by fitting vector autoregressive models of successively higher orders $k = 1, 2, \ldots$ by standard multivariate least squares.

The order of a vector autoregressive model may be tentatively identified by employing likelihood ratio statistics corresponding to testing the null hypotheses $\boldsymbol{\varphi}_k = \mathbf{0}$ against the alternative $\boldsymbol{\varphi}_k \neq \mathbf{0}$ when a VAR(k) model is fitted. If

$$\mathbf{S}(k) = \sum_{t=k+1}^{T} (\mathbf{z}_t - \hat{\boldsymbol{\varphi}}_1 \mathbf{z}_{t-1} - \cdots - \hat{\boldsymbol{\varphi}}_k \mathbf{z}_{t-k})(\mathbf{z}_t - \hat{\boldsymbol{\varphi}}_1 \mathbf{z}_{t-1} - \cdots - \hat{\boldsymbol{\varphi}}_k \mathbf{z}_{t-k})'$$

is the matrix of residual sums of squares and cross-products after fitting a VAR(k) model, then the likelihood ratio statistic is the ratio of determinants

$$U = |\mathbf{S}(k)|/|\mathbf{S}(k-1)|$$

The statistic

$$M(k) = -\left(T - p - \tfrac{1}{2} - kn\right) \ln U$$

is then asymptotically distributed as χ^2 with n^2 degrees of freedom.

8.29 Tiao and Box proposed that (8.22) could be estimated by maximizing the conditional likelihood function, particularly during the preliminary stages of model identification. This approach might, however, be inadequate with small samples when one or more zeros of $|\boldsymbol{\theta}(B)|$ lie on or close to the unit circle, in which case maximization of the exact likelihood function should be used.

To guard against model misspecification and to search for directions of improvement, Tiao and Box recommended that the residuals

$$\hat{\mathbf{a}}_t = \mathbf{z}_t - \hat{\boldsymbol{\varphi}}_1 \mathbf{z}_{t-1} - \cdots - \hat{\boldsymbol{\varphi}}_p \mathbf{z}_{t-p} + \hat{\boldsymbol{\theta}}_1 \hat{\mathbf{a}}_{t-1} + \cdots + \hat{\boldsymbol{\theta}}_q \hat{\mathbf{a}}_{t-q}$$

should be subjected to various diagnostic checks, such as plotting standardized residual series against time and/or other variables and investigating the cross-correlation matrices of the residuals. Although multivariate portmanteau-style statistics were available, Tiao and Box did not regard such tests as substitutes for more detailed study of the correlation structure.

8.30 To illustrate their model-building technique, Tiao and Box reconsidered the Coen, Gomme and Kendall (1971) data within a multivariate framework (recall **§§8.3–8.4**). Thus we denote $\mathbf{z}_t = (z_{1t}, z_{2t}, z_{3t})'$, where z_1 is the FT ordinary share price index, z_2 is UK car production and z_3 is the FT commodity price index. Using the sample period 1952III to 1967IV, Table 8.5 shows the sample cross-correlations between the three series using a second display device in which the sequence of cross-correlations between z_{it} and $z_{j,t-k}$ are shown using the indicator symbols introduced in **§8.28**. These series show high and persistent auto- and cross-correlations but examination of the partial autoregression matrices, $M(k)$ statistics and residual variances (the diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{a}}$) suggests that k is at most 2 (see Table 8.6).

Fitting both VAR(2) and VARMA(1,1) models revealed no major misspecifications and Tiao and Box chose the latter model to report, in which a vector of constants was included because the series were not de-meaned. The estimated full model

$$(\mathbf{I} - \hat{\boldsymbol{\varphi}} B)\mathbf{z}_t = \hat{\boldsymbol{\theta}}_0 + (\mathbf{I} - \hat{\boldsymbol{\theta}} B)\hat{\mathbf{a}}_t$$

Table 8.5 Pattern of sample cross-correlations for the Coen et al. data, $k = 1, \ldots, 20$

	$z_1(-k)$	$z_2(-k)$	$z_3(-k)$
z_1	++++++++++++ ++++++++++..	++++++++++++ +++++.....	---------- ----------
z_2	++++++++++++ ++++++++++..	++++++++++++ +++++++...	---------- ----------
z_3	-------.... ....++++++	....++++	+++++++++.. .....

Table 8.6 Partial autoregression matrices and related statistics

Lag	Indicator symbols for partials	$M(k) \sim \chi^2(9)$	Diagonal elements of Σ
1	$\begin{bmatrix} + & \cdot & \cdot \\ \cdot & + & \cdot \\ \cdot & \cdot & + \end{bmatrix}$	304.2	2.98×10^2 0.95×10^9 4.63
2	$\begin{bmatrix} - & \cdot & \cdot \\ \cdot & + & \cdot \\ - & + & - \end{bmatrix}$	18.8	3.92×10^2 0.96×10^9 3.73
3	$\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$	9.7	2.85×10^2 1.01×10^9 4.00
4	$\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$	3.6	3.09×10^2 1.08×10^9 3.97
5	$\begin{bmatrix} \cdot & + & \cdot \\ \cdot & + & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$	12.1	2.96×10^2 1.04×10^9 4.06

is shown in Table 8.7, along with a simpler model obtained by setting to zero those coefficients whose estimates were small compared to their standard errors. This simpler model implies that the system can be approximated by

$$
\begin{aligned}
(1 - 0.98B)z_{1t} &= a_{1t} \\
(1 - 0.93B)z_{2t} &= 0.2 + a_{2t} \\
(1 - 0.83B)z_{3t} &= 2.8 + 0.40a_{1,t-1} + (1 + 0.41B)a_{3t} \\
&= 2.8 + 0.40(1 - 0.98B)z_{1,t-1} + (1 + 0.41B)a_{3t}
\end{aligned}
$$

Table 8.7 Estimation results for the vector ARMA(1,1) model

	$\hat{\boldsymbol{\theta}}_0$	$\hat{\boldsymbol{\varphi}}$	$\hat{\boldsymbol{\theta}}$
Full model	1.11 (0.64) 1.74 (0.82) 4.08 (1.47)	0.81 0.15 −0.06 (0.08) (0.07) (0.04) −0.07 0.98 −0.09 (0.10) (0.10) (0.05) −0.32 0.30 0.76 (0.18) (0.17) (0.08)	−0.29 0.23 0.06 (0.15) (0.11) (0.07) −0.45 0.20 −0.15 (0.22) (0.17) (0.11) −0.79 0.57 −0.44 (0.28) (0.21) (0.13)
Restricted model	0.12 (0.08) 0.24 (0.10) 2.76 (1.07)	0.98 · · (0.03) · 0.93 · (0.04) · · 0.83 (0.06)	· · · · · · −0.40 · −0.41 (0.23) (0.12)

All three series behave approximately as random walks with slightly correlated innovations, since $\boldsymbol{\Sigma}_{\mathbf{a}}$ is estimated to have small but non-zero off-diagonal elements. The model thus implies that the ordinary share index is a leading indicator of the commodity share index, which is a very different model from that fitted by Coen et al. (1969).

8.31 Tiao and Box ended their paper by contrasting their procedure to some other approaches then extant in the literature and emphasizing the potential usefulness of including a moving average structure to model the dynamics of the data. In fact, contemporaneously to Tiao and Box, Jenkins and Alavi (1981) had also analyzed the vector ARMA model, providing a similar treatment and methodology for identifying, fitting and checking such models. It is fair to say, however, that the building of vector ARMA models has never really taken off in the thirty years since this paper was published, with the focus, particularly in economics, being on analyzing, possibly highly and overly parameterized, vector autoregressions: the ubiquitous VAR of (8.15), as exemplified by the contemporaneously published and hugely influential paper by Sims (1980).

Seasonal ARIMA models

8.32 Box and Jenkins' interpretation of the monthly airline model for seasonal data as a EWMA taken over previous months modified by a second EWMA of discrepancies found between similar monthly EWMAs

and actual observations in previous years (recall **§7.5**) was extended in Box, Hillmer and Tiao (1979) to more general models of the form

$$\Phi(B^{12})\varphi(B)x_t = \Theta(B^{12})\theta(B)a_t \tag{8.22}$$

This model may be written as

$$\frac{\varphi(B)}{\theta(B)}\frac{\Phi(B^{12})}{\Theta(B^{12})}x_t = R(B)Q(B^{12})x_t = a_t$$

where

$$R(B) = 1 - R_1B - R_2B^2 - \cdots$$

and

$$Q(B^{12}) = 1 - Q_1B^{12} - Q_2B^{24} - \cdots$$

On defining

$$x_t^{(R)} = (R_1 + R_2B + \cdots)x_t = (1 - R(B))B^{-1}x_t$$

and

$$x_t^{(Q)} = (Q_1 + Q_2B^{12} + \cdots)x_t = (1 - Q(B^{12}))B^{-12}x_t$$

we have

$$x_{t+1} = x_t^{(R)} + (x_{t-11} - x_{t-12}^{(R)})^{(Q)} + a_{t+1}$$

which is an extension of (7.6) with monthly and seasonal weights following more general, and not necessarily exponential, patterns.

8.33 The airline model

$$\begin{aligned}(1 - B)(1 - B^{12})x_t &= (1 - B)^2(1 + B + \cdots + B^{11})x_t \\ &= (1 - \theta B)(1 - \Theta B^{12})a_t\end{aligned} \tag{8.23}$$

has the forecast function (cf. **§6.39**)

$$\hat{x}_t(l) = b_0^{(t)} + b_1^{(t)}l + b_{0,m}^{(t)} \tag{8.24}$$

which takes the form of an updated straight line plus seasonal adjustment factors which automatically adjust as each new observation becomes available and are weighted averages of past data. An alternative form of (8.24) is

$$\hat{x}_t(l) = b_0^{(t)} + b_1^{(t)} l + \sum_{j=1}^{5} \left\{ b_{1j}^{(t)} \cos \frac{2\pi jl}{12} + b_{2j}^{(t)} \sin \frac{2\pi jl}{12} \right\}$$

in which the seasonal component contains a complete set of undamped sinusoids, adaptive in amplitude and phase with frequencies of $0, 1, \ldots, 6$ cycles per year.

Box, Pierce and Newbold (1987) showed that the parameters of (8.24) are given by

$$\begin{aligned} b_1^{(t)} &= \tfrac{1}{12}(\hat{x}_t(13) - \hat{x}_t(1)) \\ b_0^{(t)} &= \frac{1}{12} \sum_{l=1}^{12} \hat{x}_t(l) - \frac{13}{2} b_0^{(t)} \\ b_{0,m}^{(t)} &= \hat{x}_t(m) - b_0^{(t)} - b_1^{(t)} m \qquad m = 1, 2, \ldots, 12 \end{aligned}$$

If the observed series has the additive decomposition $x_t = p_t + s_t + e_t$ into independent trend, p_t, seasonal, s_t, and noise, e_t, components, then the optimal estimate of the trend at time $t + l$ is given by $b_0^{(t)} + b_1^{(t)} l$, while the optimal estimate of the seasonal is $b_{0,m}^{(t)}$ for $l = m + 12j$, j being the number of years ahead that are being forecast. In terms of available data on x_t, Box, Pierce and Newbold provided expressions for the parameters that involve functions of the π-weights as in **§6.40**.

Trend and seasonal extraction using model-based procedures

8.34 Box, Hillmer and Tiao (1979) also investigated the problem of how, given an ARIMA model for the observed series x_t, estimates of the trend and seasonal components could be obtained that were consistent with this model. They assumed that the observed series was generated as $\varphi(B)x_t = \eta(B)a_t$, where $\varphi(B)$ is of order p with zeros on or outside the unit circle, $\eta(B)$ is of order u with zeros outside the unit circle and $\varphi(B)$ and $\eta(B)$ have no common zeros. Assume for the moment the simple trend plus noise decomposition $x_t = p_t + e_t$, where the trend p_t and the

noise e_t are independent of each other, p_t follows some ARIMA model and e_t is white noise with variance σ_e^2. Box et al. showed that the model for the trend component will be $\varphi(B)p_t = \alpha(B)c_t$, where $\alpha(B)$ is of order $q \leq \max(p, u)$ and c_t is white noise with variance σ_c^2.

The various lag polynomials and variances are linked through the relationship

$$\sigma_a^2 \eta(B)\eta(B^{-1}) = \sigma_c^2 \alpha(B)\alpha(B^{-1}) + \sigma_e^2 \varphi(B)\varphi(B^{-1}) \tag{8.25}$$

Various combinations of σ_c^2, σ_e^2 and $\alpha(B)$ will satisfy this equation so some further conditions need to be placed on them, leading Box et al. to prove the following results. A model for p_t is said to be *acceptable* if $\alpha(B)$ has zeros on or outside the unit circle and satisfies (8.25) for some $\sigma_c^2 \geq 0$ and $\sigma_e^2 \geq 0$. Every model for x_t has at least one acceptable model for p_t and, given a model for x_t, every σ_e^2 in the range $0 \leq \sigma_e^2 \leq K^*$ determines a unique acceptable model for p_t. When $\sigma_e^2 = 0$, $x_t = p_t$, while if $\sigma_e^2 = K^*$ the variance of the added white noise is maximized. The bound K^* is attainable and occurs when $\alpha(B)$ has a zero on the unit circle. Thus for any model for x_t, the maximum value of σ_e^2 that is consistent with this model can be calculated.

Box et al. then go on to show that, when t is not close to the beginning or end of the observed series, an estimate of the trend, $\hat{p}_t$, is given by a symmetric moving average of p_t with the weights, ω_j, being given by the coefficients of B in the generating function

$$\omega(B) = \frac{\sigma_c^2}{\sigma_a^2} \frac{\alpha(B)\alpha(B^{-1})}{\eta(B)\eta(B^{-1})} = 1 - \frac{\sigma_e^2}{\sigma_a^2} \frac{\varphi(B)\varphi(B^{-1})}{\eta(B)\eta(B^{-1})} \tag{8.26}$$

so that knowledge of the model for p_t together with σ_e^2 will enable the trend to be estimated; a method by which the weight function $\omega(B)$ may be determined is given in the appendix of Box, Hillmer and Tiao (1979).

For values of $\hat{p}_t$ near the end of the observed series there will not be enough x_t values available for entering into the weight function $\omega(B)$. Box et al. proposed that the model for p_t should be used to provide enough forecasts to extend the observed series so that $\omega(B)$ can be calculated, with a similar procedure using backcasts employed at the beginning of the series.

8.35 As an example of these ideas, suppose that the model for the observed series is $(1 - B)x_t = (1 - \eta B)a_t$, so that the model for the trend

must be $(1-B)p_t = (1-\alpha B)c_t$ for some α and (8.25) becomes

$$\sigma_a^2(1-\eta B)(1-\eta B^{-1}) = \sigma_c^2(1-\alpha B)(1-\alpha B^{-1}) + \sigma_e^2(1-B)(1-B^{-1})$$

By setting $B = 1/\alpha$, solving for σ_e^2 obtains

$$\sigma_e^2 = -\frac{\sigma_a^2(\alpha-\eta)(1-\alpha\eta)}{(1-\alpha)^2}$$

Setting the derivative of this expression with respect to α to zero shows that the maximum value of σ_e^2 occurs when $\alpha = -1$, thus agreeing with the result that the bound K^* is attained when the zero of $(1-\alpha B)$ is on the unit circle. The trend model corresponding to the largest possible σ_e^2 is therefore $(1-B)p_t = (1+B)c_t$ with

$$\sigma_e^2 = \frac{(1+\eta)^2}{4}\sigma_a^2$$

From (8.26) the weight function is

$$\omega(B) = 1 - \frac{(1+\eta)^2}{4}\frac{(1-B)(1-B^{-1})}{(1-\eta B)(1-\eta B^{-1})}$$

and Box et al. derived the weights as

$$\omega_0 = \frac{1-\eta}{2}, \qquad \omega_1 = \omega_{-1} = \frac{1-\eta^2}{4}, \qquad \omega_j = \omega_{-j} = \eta\omega_{j-1}, \qquad j = 2, 3, \ldots$$

The estimate of the trend in the middle of the observed series is thus

$$\hat{p}_t = \frac{1-\eta}{2}x_t + \frac{1-\eta^2}{4}((x_{t+1}+x_{t-1}) + \eta(x_{t+2}+x_{t-2}) + \cdots) \quad (8.27)$$

while at the very end of the series the estimate becomes

$$\hat{p}_T = \frac{1-\eta}{2}x_T + \frac{1-\eta^2}{4}((\hat{x}_T(1)+x_{T-1}) + \eta(\hat{x}_T(2)+x_{T-2}) + \cdots)$$

Since

$$\begin{aligned}\hat{x}_T(1) &= (1-\eta)x_T + \eta(1-\eta)x_{T-1} + \eta^2(1-\eta)x_{T-2} + \cdots,\\ \hat{x}_T(l) &= \hat{x}_T(1), \quad l = 2, 3, \ldots\end{aligned}$$

this final estimate can be written as

$$\hat{p}_T = \frac{(3 - 2\eta - \eta^2)}{4} x_T + \frac{1 - \eta^2}{4}((1 - \eta)x_{T-1} + \eta(1 - \eta)x_{T-2} + \cdots)$$

with the trend for other observations near the end of the series being obtained in a similar manner.

The model fitted to the monthly rate of change of the US CPI (inflation) in **§8.12** was $(1 - B)x_t = (1 - 0.92B)a_t$ with $\sigma_a = 0.0026$. The trend component is shown superimposed upon inflation in Figure 8.7 for the period January 1955 to December 1971 and is seen to be a smooth, slowly varying function since (8.27) takes the form

$$\begin{aligned}\hat{p}_t = {} & 0.04x_t + 0.0384((x_{t+1} + x_{t-1}) + 0.92(x_{t+2} + x_{t-2}) \\ & + 0.846(x_{t+3} + x_{t-3}) + \cdots)\end{aligned}$$

so that the filter weights are all small and decline only slowly, a consequence, of course, of η being large and positive. The white noise error component will have $\sigma_e = 0.0025$.

8.36 Box et al. then extended this 'model-based decomposition' method to the seasonal model $x_t = p_t + s_t + e_t = s_t + T_t$, where T_t is now termed

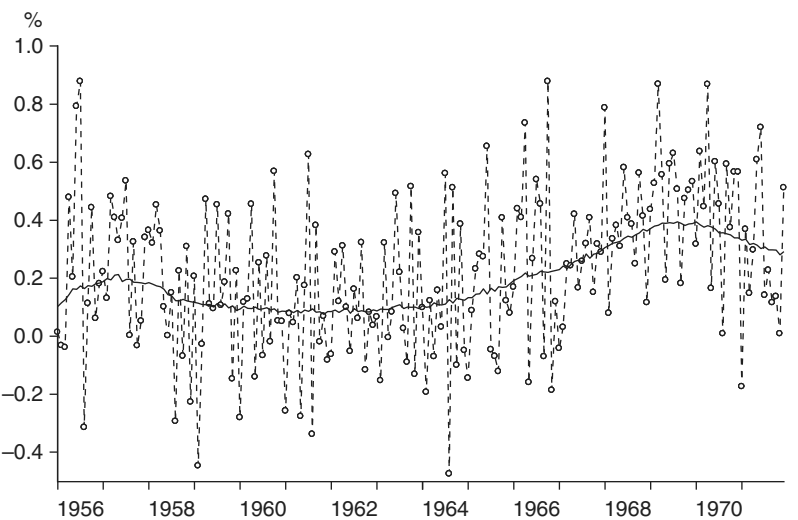


Figure 8.7 Monthly rate of inflation of the US consumer price index: January 1955 to December 1971 with fitted trend superimposed

the 'trend plus noise' component. They assumed that the seasonal component s_t was able to evolve over time so that the sum of twelve consecutive values varies about zero with minimum variance, arguing that if this was not the case then the excess variation should properly be reflected in p_t or e_t. Focusing on the airline model (8.23) (a natural seasonal extension of the model of **§8.34)**, Box et al. assumed that s_t and T_t followed independent ARIMA processes. Given the model for x_t, the product of the trend plus noise and seasonal autoregressive operators must be $(1-B)(1-B^{12}) = (1-B)^2(1+B+\cdots+B^{11}) = (1-B)^2U(B)$, so that appropriate models for the components will be

$$\begin{aligned} U(B)s_t &= (1-\psi_1 B-\cdots-\psi_{11}B^{11})b_t = \psi(B)b_t \\ (1-B)^2T_t &= (1-\eta_1 B-\eta_2 B^2)d_t = \eta(B)d_t \end{aligned}$$

where b_t and d_t are two independent white noise processes with zero means and variances σ_b^2 and σ_d^2 respectively. Letting $\theta(B) = (1-\theta B)(1-\Theta B^{12})$ and noting that

$$(1-B)(1-B^{12})x_t = (1-B)(1-B^{12})s_t + (1-B)(1-B^{12})T_t$$

it follows that

$$\theta(B)a_t = (1-B)^2\psi(B)b_t + U(B)\eta(B)d_t$$

and

$$\begin{aligned} \sigma_a^2\theta(B)\theta(B^{-1}) = {} & \sigma_b^2(1-B)^2\psi(B)(1-B^{-1})^2\psi(B^{-1}) \\ & + \sigma_d^2U(B)\eta(B)U(B^{-1})\eta(B^{-1}) \end{aligned}$$

Under the assumption that $\hat{s}_t = U(B)s_t$ has to have minimum variance, Box et al. showed that the components are estimated as $\hat{s}_t = w(B)x_t$ and $\hat{T}_t = h(B)x_t$, where

$$w(B) = \frac{\sigma_b^2}{\sigma_a^2}\frac{(1-B)^2\psi(B)(1-B^{-1})^2\psi(B^{-1})}{\theta(B)\theta(B^{-1})}$$

and

$$h(B) = \frac{\sigma_d^2}{\sigma_a^2}\frac{U(B)\eta(B)U(B^{-1})\eta(B^{-1})}{\theta(B)\theta(B^{-1})}$$

The trend component itself can then be estimated as

$$\hat{p}_t = \left(1 - \frac{\sigma_e^2}{\sigma_d^2} \frac{(1-B^2)(1-B^{-1})^2}{\eta(B)\eta(B^{-1})}\right) \hat{T}_t$$

Box et al. provided additional details on the computational aspects of this model based procedure and an example using a US unemployment series.

This model-based approach to seasonal adjustment, in which the seasonally adjusted series is defined as $x_t^{SA} = x_t - \hat{s}_t$, was extended in subsequent years, notable contributions being Burman (1980), Hillmer and Tiao (1982), Hillmer, Bell and Tiao (1983), Maravall and Pierce (1987), Pierce (1978) and Tiao and Hillmer (1978) and, in a further development, Maravall (2000). The optimality of the Bureau of the Census X-11 seasonal adjustment method, by now the 'industry standard' for empirically-based adjustment, was also investigated using similar model-based techniques: see Cleveland and Tiao (1976), Bell and Hillmer (1984) and Burridge and Wallis (1984). Box was thus instrumental in being at the forefront of developments in that most practical of time series areas, seasonal adjustment. This was to be his last major involvement in time series research, however, as his interests became increasingly focused on statistical issues of quality control, although even here the use of time series techniques in process control remained at the forefront, as shown by Box and Kramer (1992) and Box and Luceño (1995).

9

Granger: Spectral Analysis, Causality, Forecasting, Model Interpretation and Non-linearity

Clive Granger

9.1 Clive William John Granger was born on September 4, 1934 in Swansea, Wales, only a few months after the birth, and only a few miles from the birthplace, of Gwilym Jenkins. Unlike Jenkins, however, Granger's family was English, his father being a commercial traveller for Chivers, a then well-known company based near Cambridge, England, producing preserves such as jams and marmalades. His sojourn in Wales was very brief, with the family quickly relocating to Lincoln, England, and then, during the war while his father was serving in the RAF, to Cambridge, where both sets of grandparents lived. In 1946, after his return from war duties, Granger's father was once again relocated, this time to Nottingham, where they lived in the suburb of West Bridgford. After attending Nottingham Grammar School, Granger entered Nottingham University in the first intake of the joint mathematics and economics degree program, although he switched after the first year to single honours mathematics. On obtaining a first in 1955, he stayed on to do a PhD in statistics even though, by his own admission, he knew very little about the subject. Wanting to research in a subject related to economics, he fortuitously chose time series as an area that looked to be ripe for development!

After just six months of doctoral research, an opportunity arose to apply for a lectureship in statistics in the Department of Mathematics at Nottingham, for which he was successful, and he joined the academic staff at the age of 22 in 1956. On receiving his doctorate in 1959 for a thesis on testing for non-stationarity, Granger applied for a Harkness Fellowship and, on receiving the award, moved to Princeton for a year to work with Oscar Morgenstern on his new 'Time Series Project', which

essentially formed the basis for Granger's early publications on spectral analysis (see **§§9.2–9.8**). Although he returned to Princeton over the next two summers, primarily working on stock prices, Granger remained at Nottingham until 1974, having become a professor a decade earlier. At that point, and by now having a world class reputation in time series analysis and forecasting, Granger moved to the Department of Economics at the University of California, San Diego (UCSD), where he remained for the rest of his career, being instrumental in building up one of the best econometric groups in the world.

His research in time series and econometrics was honored in 2003, when he was awarded, jointly with Robert Engle, his long-time colleague and collaborator at UCSD, with the Nobel Prize in Economics, the citation being 'for methods of analyzing economic time series with common trends (co-integration)' (see **§§10.21–10.39**). Having retained his British nationality, Granger was further honored by being knighted in the 2005 New Year's Honours List. Although he had officially retired in the summer of 2003, the Nobel award ensured that he was in great demand throughout the world and he continued to be a very active researcher, lecturer and writer until his untimely death on 27 May 2009, from complications related to a brain tumor. Further biographical details on Sir Clive Granger may be found in a variety of sources, most notably Phillips (1997) and Frängsmyr (2004), with a curriculum vitae, including a full list of publications, being published as Granger (2010a).

Attempting to distill Granger's published research, even just those on time series (he also published widely on consumer attitudes to prices and various aspects of finance and statistics) would be a herculean task requiring a book of its own, so we focus in this and the next chapter on five major areas: (i) spectral analysis, causality and feedback, (ii) forecasting and evaluation, (iii) model interpretation and non-linearity, (iv) long memory, and (v) spurious regression and co-integration. A brief summary of Granger's other research and an appreciation of his impact on economics and statistical science in general is then provided.[1]

Spectral analysis, causality and feedback

9.2 Granger's first publication was on a statistical model for sunspot activity (Granger, 1957), so continuing a long line of research by time series analysts in this area (recall Yule's analysis of the sunspot index discussed in Chapter 2, while Moran, 1954, and Whittle, 1954b, had both published prior to Granger on the topic). His second publication

was on estimating the probability of flooding on a tidal river (Granger, 1959), so establishing early on his wide range of statistical interests. The time spent at Princeton on the Time Series Project quickly bore fruit and Granger's initial publications on spectral analysis were the book *Spectral Analysis of Economic Time Series*, written in association with Michio Hatanaka (Granger and Hatanaka, 1964), which introduced the technique to many economists and econometricians, and the article in the journal *Information and Control* (Granger, 1963), which was essentially chapter 7 of the book, in which the concepts of *feedback* and *causality* were introduced. Of course, as we saw in **§§5.4–5.18**, Gwilym Jenkins was contemporaneously attempting to introduce spectral concepts to statisticians, while Granger's concepts of feedback and causality were to be formalized some years later in what was to become a very influential article: Granger (1969a).[2]

9.3 Granger and Hatanaka (1964) covered a good deal of the same basic material as Jenkins but provided further interpretation of the concepts of cross-, co- and quadrature spectrum and coherence discussed in **§§5.16–5.17**. Here they considered the Cramér (1940) representation of the real time series X_t and Y_t:

$$X_t = \int_{-\pi}^{\pi} e^{it\omega} dz_x(\omega) = \int_0^{\pi} \cos t\omega du_x(\omega) + \int_0^{\pi} \sin t\omega dv_x(\omega)$$
$$Y_t = \int_{-\pi}^{\pi} e^{it\omega} dz_y(\omega) = \int_0^{\pi} \cos t\omega du_y(\omega) + \int_0^{\pi} \sin t\omega dv_y(\omega)$$

where the $dz_x(\omega)$, etc., are random and uncorrelated processes. X_t and Y_t can thus each 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 the components has a random amplitude, $du_x(\omega)$, etc., and Granger and Hatanaka showed that, for both processes, these 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

$$E(du_x(\omega)du_y(\omega)) = E(dv_x(\omega)dv_y(\omega)) = 2c_{xy}(\omega)d\omega$$

and

$$E(du_x(\omega)dv_y(\omega)) = 2q_{xy}(\omega)d\omega$$
$$E(du_y(\omega)dv_x(\omega)) = -2q_{xy}(\omega)d\omega$$

where c_{xy} and q_{xy} are the co-spectrum and quadrature spectrum introduced in **§5.16**. 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).

9.4 Granger and Hatanaka took the analysis of the cross-spectrum much further than Jenkins. 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 $(X_{1t}, X_{2t}, \ldots, X_{Mt})$. Each series has its own (auto) spectra, $f_{ii}(\omega)$, and there will be a set of cross-spectra, $f_{ij}(\omega)$, $i, j = 1, \ldots, M$, which will typically be complex quantities. The matrix of these spectra, $\mathbf{\Sigma}(\omega)$, 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

$$\mathbf{\Sigma}(\omega) = \left[\begin{array}{cc|ccc} f_{11}(\omega) & f_{12}(\omega) & f_{13}(\omega) & \ldots & f_{1M}(\omega) \\ f_{21}(\omega) & f_{22}(\omega) & f_{23}(\omega) & \ldots & f_{2M}(\omega) \\ \hline f_{31}(\omega) & f_{32}(\omega) & f_{33}(\omega) & \ldots & f_{3M}(\omega) \\ \vdots & \vdots & \vdots & & \vdots \\ f_{M1}(\omega) & f_{M2}(\omega) & f_{M3}(\omega) & \ldots & f_{MM}(\omega) \end{array}\right] = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

and define the matrix

$$\mathbf{\Sigma}_{12\cdot k}(\omega) = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} = \begin{bmatrix} f_{11\cdot k}(\omega) & f_{12\cdot k}(\omega) \\ f_{21\cdot k}(\omega) & f_{22\cdot k}(\omega) \end{bmatrix}$$

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:

$$C^2_{12\cdot k}(\omega) = \frac{|f_{12\cdot k}(\omega)|^2}{f_{11\cdot k}(\omega) f_{22\cdot k}(\omega)} \qquad \psi_{12\cdot k}(\omega) = \frac{\text{Imaginary part of } f_{12\cdot k}(\omega)}{\text{Real part of } f_{12\cdot k}(\omega)}$$

These concepts have the following interpretation. Suppose that a 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 $\hat{X}_{2t}$. $f_{11\cdot k}(\omega)$ will thus be the spectrum of $\hat{X}_{1t}$, and $C^2_{12\cdot k}(\omega)$ 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 92–93)

9.5 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

$$\begin{aligned} Y_t &= \int_{-\pi}^{\pi} e^{it\omega} a(\omega) e^{-i\Phi\omega} dz(\omega) \\ &= a(\omega) \int_0^{\pi} \cos t\omega\Phi(\omega) du_x(\omega) + a(\omega) \int_0^{\pi} \sin t\omega\Phi(\omega) dv_x(\omega) \end{aligned}$$

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

$$Y_t = X_t\{a(\omega), \phi(\omega)\} + U_t \tag{9.1}$$

where U_t is some stationary series such that $C^2_{xu}(\omega) = 0$, so that (cf. equation (5.15))

$$0 < C^2_{yx}(\omega) = \frac{a^2 f_x(\omega)}{f_y(\omega)} < 1$$

If, as well as (9.1),

$$X_t = Y_t\{b(\omega), \theta(\omega)\} + V_t$$

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 process continually lags the other.

To provide a formal definition of feedback from which tests may be developed, Granger (1969a) 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 $\bar{A} = A(1)$ and $\bar{\bar{A}} = A(0)$ 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 $P_t(X|\bar{X})$ 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 $\sigma^2(X|\bar{I}) < \sigma^2(X|\bar{I} - \bar{Y})$ then Y is said cause X, denoted $Y \Rightarrow X$. Thus 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 $\sigma^2(X|\bar{I}) < \sigma^2(X|\bar{I} - \bar{Y})$ and $\sigma^2(Y|\bar{I}) < \sigma^2(Y|\bar{I} - \bar{X})$ 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 $\sigma^2(X|\bar{I}, \bar{\bar{Y}}) < \sigma^2(X|\bar{I})$ then instantaneous causality is said to occur, 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 $\sigma^2(X|\bar{I} - Y(k)) < \sigma^2(X|\bar{I} - Y(k+1))$. Thus 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 non-stationarity was 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 $\sigma^2(X|\bar{X}) > \sigma^2(X|\bar{X}, \bar{Y})$. 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 (see **§9.11**). Spurious instantaneous causality is another possibility when the sampling interval is greater than the causality lag (see **§9.9**).

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.

9.6 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. Suppose these series are generated by the bivariate process

$$\begin{aligned} X_t &= \sum_{j=1}^{p} a_j X_{t-j} + \sum_{j=1}^{p} b_j Y_{t-j} + \varepsilon_t = a(B)X_t + b(B)Y_t + \varepsilon_t \\ Y_t &= \sum_{j=1}^{p} c_j X_{t-j} + \sum_{j=1}^{p} d_j Y_{t-j} + \eta_t = c(B)X_t + d(B)Y_t + \eta_t \end{aligned} \tag{9.2}$$

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 (9.2), 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 (9.2) may be represented as

$$\int_{-\pi}^{\pi} e^{it\omega}((1 - a(e^{-i\omega}))dz_x(\omega) - b(e^{-i\omega})dz_y(\omega) - dz_\varepsilon(\omega)) = 0$$
$$\int_{-\pi}^{\pi} e^{it\omega}(-c(e^{-i\omega})dz_x(\omega) + (1 - d(e^{-i\omega}))dz_y(\omega) - dz_\eta(\omega)) = 0$$

From this representation, Granger (1969a) showed that the spectra of X_t and Y_t are given by

$$f_x(\omega) = \frac{1}{2\pi\Delta}(|1 - d|^2\sigma_\varepsilon^2 + |b|^2\sigma_\eta^2)$$
$$f_y(\omega) = \frac{1}{2\pi\Delta}(|c|^2\sigma_\varepsilon^2 + |1 - a|^2\sigma_\eta^2)$$

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

$$f_{xy}(\omega) = \frac{1}{2\pi\Delta}((1 - d)c\sigma_\varepsilon^2 + (1 - a)b\sigma_\eta^2) = f_1(\omega) + f_2(\omega)$$

where

$$f_1(\omega) = \frac{\sigma_\varepsilon^2}{2\pi\Delta}(1 - d)c \qquad f_2(\omega) = \frac{\sigma_\eta^2}{2\pi\Delta}(1 - a)b$$

Thus, if Y_t is *not* causing X_t then $b = 0$ and $f_2(\omega)$ vanishes and, similarly, if X_t is *not* causing Y_t then $c = 0$ and $f_1(\omega)$ vanishes. Hence the cross-spectrum may be decomposed into the sum of two components: $f_1(\omega)$, depending upon the causality of Y by X, and $f_2(\omega)$, 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*,

$$C_{\vec{xy}}^2(\omega) = \frac{|f_1(\omega)|^2}{f_x(\omega)f_y(\omega)} = \frac{\sigma_\varepsilon^4|(1 - d)c|^2}{(\sigma_\varepsilon^2|1 - d|^2 + \sigma_\eta^2|b|^2)(\sigma_\varepsilon^2|c|^2 + \sigma_\eta^2|1 - a|^2)}$$

may be considered to be the strength of the causality $X \Rightarrow Y$ at frequency ω. Similarly,

$$\phi_{\overrightarrow{xy}}(\omega) = \tan^{-1} \frac{\text{Imaginary part of } f_1(\omega)}{\text{real part of } f_1(\omega)}$$

will measure the phase lag at frequency ω of $X \Rightarrow Y$. Similar functions can be defined for $Y \Rightarrow X$ using $f_2(\omega)$.

Instantaneous causality may be allowed for by including the terms $b_0 Y_t$ and $c_0 X_t$ in the respective equations in the representation (9.2). The cross-spectrum is then given by

$$\begin{aligned} f_{xy}(\omega) &= \frac{1}{2\pi\,\Delta'}((1-d)(c+c_0)\sigma_\varepsilon^2 + (1-a)(b+b_0)\sigma_\eta^2) \\ &= f_1'(\omega) + f_2'(\omega) + f_3'(\omega) \end{aligned}$$

where $\Delta' = |(1-a)(1-d) - (b+b_0)(c+c_0)|^2$, $f_1'(\omega)$ and $f_2'(\omega)$ are defined as $f_1(\omega)$ and $f_2(\omega)$ but using Δ' rather than Δ, and

$$f_3'(\omega) = \frac{1}{2\pi\,\Delta'}(c_0(1-d)\sigma_\varepsilon^2 + b_0(1-a)\sigma_\eta^2)$$

The presence of instantaneous causality clearly means that the measures of causal strength and phase lag lose their distinct interpretations.

Granger causality

9.7 Granger (1969a) 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.[3] Fortunately, this was not long in coming, with Christopher Sims providing both this and a very thought provoking example on the causal links between money and income, a 'hot' topic at the time and for some years after (Sims, 1972). Essentially, Sims recommended testing sets of coefficients in (9.2) directly in the time domain using group F-tests: for example, the hypothesis that Y *does not cause* X, $Y \not\Rightarrow X$, may be parameterized as $b_1 = \cdots = b_p = 0$, which may then be tested directly with rejection leading to $Y \Rightarrow X$. Within Sims' framework Granger causality became straightforward to test for and applications and theoretical extensions

quickly followed. Sims (1977) and Geweke (1984) were particularly useful contributions which also set out the links between Granger causality and econometric concepts of exogeneity.

Naturally, there were also several critiques of the concept, both from economists (notably Zellner, 1979) and from a wider philosophical and statistical perspective (for example, Holland, 1986). Granger had been quite clear from the outset about his definition of causality, it being

> based entirely upon the predictability of some series, say X_t. If some other series Y_t contains information in past terms that helps in the prediction of X_t and if this information is contained in no other series used in the predictor, then Y_t is said to cause X_t,

adding that '(t)he flow of time clearly plays a central role in these definitions' and that '(i)n the author's opinion there is little use in the practice of attempting to discuss causality without introducing time, although philosophers have tried to do so' (Granger, 1969a, page 430). A decade later, Granger (1980a) returned to this theme, placing his definition of causality given in **§9.5** within a wider setting based on the following three axioms.

Axiom A

The past and present may cause the future, but the future cannot cause the past.

Axiom B

I_t contains no redundant information, so that if some variable Z_t is functionally related to one or more other variables, in a deterministic fashion, then Z_t should be excluded from I_t.

Axiom C

All causal relationships remain constant in direction throughout time.

Given these axioms, Granger proposed the following

General Definition

Y_t is said to cause X_{t+1} if

$$P(X_{t+1} \in A|I_t) \neq P(X_{t+1} \in A|I_t - Y_t) \quad \text{for some } A$$

Thus, for causation to occur, Y_t must have some unique information about what value X_{t+1} will take in the immediate future.

9.8 This general definition is not operational, in the sense that it could not be used with actual data, so Granger re-stated it in terms of a vector $\boldsymbol{Y}_t$ causing another vector $\boldsymbol{X}_t$. Thus suppose that J_t is an information set defined to consist of the vector $\boldsymbol{Z}_t$, i.e. $J_t : \boldsymbol{Z}_{t-j}$, $j \geq 0$: J_t is said to be a *proper* information set with respect to $\boldsymbol{X}_t$ if $\boldsymbol{X}_t$ is included within $\boldsymbol{Z}_t$. Suppose further that $\boldsymbol{Z}_t$ does not include any components of $\boldsymbol{Y}_t$, so that the intersection of $\boldsymbol{Z}_t$ and $\boldsymbol{Y}_t$ is zero and we define $J'_t : \boldsymbol{Z}_{t-j}, \boldsymbol{Y}_{t-j}$, $j \geq 0$, as J_t plus the past and present values of $\boldsymbol{Y}_t$.

If $F(\boldsymbol{X}_{t+1}|J_t)$ is the conditional distribution function of $\boldsymbol{X}_{t+1}$ given J_t, with the mean of this distribution being $E[\boldsymbol{X}_{t+1}|J_t]$, then Granger introduced the following definitions.

Definition 1

$\boldsymbol{Y}_t$ *does not* cause $\boldsymbol{X}_{t+1}$ with respect to J'_t if $F(\boldsymbol{X}_{t+1}|J_t) = F(\boldsymbol{X}_{t+1}|J'_t)$, so that the extra information in J'_t does not affect the conditional distribution. A necessary condition is that $E[\boldsymbol{X}_{t+1}|J_t] = E[\boldsymbol{X}_{t+1}|J'_t]$.

Definition 2

If $J'_t = I_t$, the universal information set, and if $F(\boldsymbol{X}_{t+1}|I_t) \neq F(\boldsymbol{X}_{t+1}|I_t - \boldsymbol{Y}_t)$, then $\boldsymbol{Y}_t$ is said to *cause* $\boldsymbol{X}_{t+1}$. This is equivalent to the general definition of causality introduced in **§9.7**.

Definition 3

If $F(\boldsymbol{X}_{t+1}|J'_t) \neq F(\boldsymbol{X}_{t+1}|J_t)$ then $\boldsymbol{Y}_t$ is said to be a *prima facie cause* of $\boldsymbol{X}_{t+1}$ with respect to the information set J'_t.

Definition 4

$\boldsymbol{Y}_t$ is said *not to cause* $\boldsymbol{X}_{t+1}$ *in mean* with respect to J'_t if $\delta_{t+1}(J'_t) = E[\boldsymbol{X}_{t+1}|J'_t] - E[\boldsymbol{X}_{t+1}|J_t]$ is identically zero.

Definition 5

If $\delta_{t+1}(I_t)$ is not zero, then $\boldsymbol{Y}_t$ is said to *cause* $\boldsymbol{X}_{t+1}$ *in mean*.

Definition 6

If $\delta_{t+1}(J'_t)$ is not identically zero, then $\boldsymbol{Y}_t$ is said to be a *prima facie cause in mean* of $\boldsymbol{X}_{t+1}$ with respect to J'_t.

The final three definitions become relevant if just (one-step ahead) point forecasts obtained using a least squares criterion are employed, rather than the whole distribution of $\boldsymbol{X}_{t+1}$, which is often the case in practice. These forecasts will often be linear functions of the information set, although non-linear functions are not excluded by the definitions, and

attention is often focused on pairs of series, Y_t and X_{t+1}, rather than pairs of vectors, $\boldsymbol{Y}_t$ and $\boldsymbol{X}_{t+1}$. It is also necessary to assume that the series are stationary, or at least that they belong to some simple class of models with time varying parameters, for practical implementation of the definitions.

9.9 There are a number of important implications of this definition of causality that Granger addressed by way of a sequence of examples. He first made a general point using the simple model

$$X_t = \varepsilon_t + \eta_{t-1} \qquad Y_t = \eta_t + \varepsilon_{t-1}$$

where ε_t and η_t are a pair of independent white noises. Since these equations imply that $X_{t+1} = Y_t + \varepsilon_{t+1} - \varepsilon_{t-1}$ and $Y_{t+1} = X_t + \eta_{t+1} - \eta_{t-1}$ it is clear that feedback exists, with X causing Y and Y causing X.

Example 1

$$X_t = \varepsilon_t \qquad Y_t = \varepsilon_{t-1} + \eta_t \qquad Z_t = \eta_{t-1}$$

In this example there are four information sets to consider. If $J_t(X, Y, Z)$ denotes the information set containing the past and present values of X_{t-j}, Y_{t-j} and Z_{t-j}, $j \geq 0$, then the subsets $J_t(X, Y)$, $J_t(X, Z)$ and $J_t(Y, Z)$ may be defined analogously. Since the example implies $Y_{t+1} = X_t + \eta_{t+1} = X_t + Z_{t+1}$ and $Z_{t+1} = Y_t - \varepsilon_{t-1} = Y_t - X_{t-1}$, then clearly X causes Y with respect to either $J_t(X, Y)$ or $J_t(X, Y, Z)$ and Y causes Z with respect to $J_t(Y, Z)$ and $J_t(X, Y, Z)$. However, X does not cause Z with respect to $J_t(X, Z)$ but does cause Z with respect to $J_t(X, Y, Z)$, because Z_{t+1} is completely determined from Y_{t-j} and X_{t-j} but not from Y_{t-j} alone, thus demonstrating the importance of stating the information set being utilized. Granger's general point is that, although it may be the case that X causes Y and Y causes Z, it may not necessarily be true that X causes Z. This is further illustrated by

Example 2

$$X_t = \varepsilon_t + \omega_t \qquad Y_t = \varepsilon_{t-1} \qquad Z_t = \varepsilon_{t-2} + \eta_t$$

where ω_t is another independent white noise. Here $Z_{t+1} = X_{t-1} + \eta_{t+1} - \omega_{t-1} = Y_t + \eta_{t+1}$ so that X causes Z in $J_t(X, Z)$ but not in $J_t(X, Y, Z)$.

If Y_t causes X_{t+1} then $Y'_t = a(B)Y_t$ causes $X'_{t+1} = b(B)X_{t+1}$ if $a(B)$ and $b(B)$ are one-sided filters, but this may not be the case if the filters are two-sided, as might occur in some seasonal adjustment procedures, since Axiom A is disrupted.

It is impossible for a series that is *self-deterministic* (that is, perfectly forecastable from its past) to be caused by any other variable, as is demonstrated by

Example 3

$$X_t = a + bt + ct^2 \qquad Y_t = dX_{t+1}$$

As well as the quadratic in time, the following pair of equations will also generate X_t exactly:

$$X_t = d^{-1}Y_{t-1} \qquad X_t = 2X_{t-1} - X_{t-2} + 2c$$

so that it would appear that X_t is 'caused' by time, by Y_{t-1}, or by its own past. Since all three equations fit perfectly it is impossible to distinguish between them and a statistical test for causality is impossible.

A particular problem is that of missing variables, which can lead to apparent causation due to a common cause, as in

Example 4

$$Z_t = \eta_t \qquad X_t = \eta_{t-1} + \omega_t \qquad Y_t = \eta_{t-2} + \varepsilon_t$$

Since $X_{t+1} = Z_t + \omega_{t+1}$ and $Y_{t+1} = Z_{t-1} + \varepsilon_{t+1}$, Z will cause both X and Y with respect to $J_t(X, Z)$, $J_t(Y, Z)$ and $J_t(X, Y, Z)$. However, since $Y_{t+1} = X_t + \varepsilon_{t+1} - \omega_t = Z_{t-1} + \varepsilon_{t+1}$, X will cause Y in $J_t(X, Y)$ but not in $J_t(X, Y, Z)$. This apparent causation of Y by X in $J_t(X, Y)$ may be thought of as spurious because it vanishes when the information set is expanded. This situation is encountered when dealing with leading indicators: X is a leading indicator of Y but will cease to be a cause of Y when Z is observed.

A related problem occurs when one variable is measured with an error having some form of time structure, as in

Example 5

$$X_t = \eta_t \qquad Y_t = \delta_t \qquad Z_t = X_t + \varepsilon_t + \beta\varepsilon_{t-1}$$

where η_t and δ_t are white noises which are contemporaneously correlated. Since $Z_t = \eta_t + \varepsilon_t + \beta\varepsilon_{t-1}$, it can be written as $Z_t = e_t + \theta e_{t-1} = (1 + \theta B)e_t$, where e_t is a white noise, so that

$$e_t = (1 + \theta B)^{-1}\eta_t + (1 + \theta B)^{-1}(1 + \beta B)\varepsilon_t$$

The one-step ahead forecast of Z_{t+1} using Z_{t-j} $(j \geq 0)$ will be θe_t with forecast error e_{t+1}. This error will be a function of η_{t-j} $(j \geq 0)$, which itself is correlated with $\delta_{t-j} = Y_{t-j}$ $(j \geq 0)$. Thus Y_{t-j} can help forecast Z_{t+1} so that, apparently, Y causes Z with respect to $J_t(Y, Z)$, although this would not be the case if X_{t-j} were observable, so that $J_t(X, Y, Z)$ could be considered.

These definitions focus on one-step forecasts rather than h-step ahead forecasts for any h. It can be shown that if Y causes X with respect to $J_t(X, Y)$ when using an $h > 1$-step forecasting criterion then it will necessarily be found that Y causes X with a one-step criterion. This does not, however, appear to be true in the multivariate case, for consider

Example 6

$$X_t = \varepsilon_t \qquad Y_t = \varepsilon_{t-2} + \eta_t \qquad Z_t = \varepsilon_{t-1} + \omega_t$$

Since $Z_{t+1} = X_t + \omega_{t+1}$, X_t causes Z_{t+1} with respect to both $J_t(X, Z)$ and $J_t(X, Y, Z)$. However, since $Y_{t+2} = X_t + \eta_{t+2}$ but $Y_{t+1} = Z_t + \eta_{t+1} - \omega_t$, X_t causes Y_{t+2} with respect to both $J_t(X, Y)$ and $J_t(X, Y, Z)$ but only causes Y_{t+1} with respect to $J_t(X, Y)$.

9.10 The definitions given above do not allow for *instantaneous causality*. Granger (1988) provided a detailed discussion of this concept, although several of his earlier writings touch upon it (recall **§9.5**). Define the one-step ahead forecast errors to be

$$e_{X,t+1} = X_t - E[X_{t+1}|J_t(X, Y)] \qquad e_{Y,t+1} = Y_t - E[Y_{t+1}|J_t(X, Y)]$$

If $\rho = \text{corr}(e_{X,t+1}, e_{Y,t+1}) \neq 0$ then there is apparent instantaneous causality between X and Y. Such a definition might be regarded as unsatisfactory since no direction of causality may be deduced just from the data: what is required is some further knowledge, say that X cannot cause Y. This situation is essentially identical to the long-standing problem that correlation cannot be equated to causality unless an assumption is made concerning the structure of the relationship between X and Y.

Three possible explanations for apparent instantaneous causality were discussed by Granger:

(i) There actually is true instantaneous causality so that some variables react *without any measurable time delay* to changes in some other variables.

(ii) There is no true instantaneous causality but the finite time delay between cause and effect is small compared to the time interval over which the data is collected, so that the apparent causation is a consequence of temporal aggregation.
(iii) There is a jointly causal variable Z_t that causes both X_{t+1} and Y_{t+1} but is not included in the information set, possibly because it cannot be observed.

9.11 The bivariate relationship between the zero mean, jointly stationary series X_t and Y_t can be characterized in various ways (recall **§§8.14–8.17**). The moving average representation is

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \psi_{11}(B) & \psi_{12}(B) \\ \psi_{21}(B) & \psi_{22}(B) \end{bmatrix} \begin{bmatrix} a_t \\ b_t \end{bmatrix} \tag{9.3}$$

where $[a_t\ b_t]'$ is a two-element white noise vector with zero correlation between a_t and b_s except possibly when $t = s$. Assuming invertibility of the moving average matrix, the corresponding autoregressive model is

$$\begin{bmatrix} A(B) & H(B) \\ C(B) & D(B) \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} a_t \\ b_t \end{bmatrix} \tag{9.4}$$

As in **§8.17**, the series could be pre-whitened using the filters $F(B)X_t = u_t$ and $G(B)Y_t = v_t$, leading to the moving average and autoregressive models linking $[u_t\ v_t]'$ to $[a_t\ b_t]'$:

$$\begin{bmatrix} u_t \\ v_t \end{bmatrix} = \begin{bmatrix} \theta_{11}(B) & \theta_{12}(B) \\ \theta_{21}(B) & \theta_{22}(B) \end{bmatrix} \begin{bmatrix} a_t \\ b_t \end{bmatrix} \tag{9.5}$$

and

$$\begin{bmatrix} \alpha(B) & \beta(B) \\ \gamma(B) & \delta(B) \end{bmatrix} \begin{bmatrix} u_t \\ v_t \end{bmatrix} = \begin{bmatrix} a_t \\ b_t \end{bmatrix} \tag{9.6}$$

If $\rho_{uv}(k)$ is the cross-correlation between u_{t-k} and v_t then there exists the following regression

$$v_t = \sum_{j=-\infty}^{\infty} \omega_j u_{t-j} + f_t \tag{9.7}$$

where $\omega_k = (\sigma_v/\sigma_u)\rho_{uv}(k)$. Similarly, there will exist the regression

$$Y_t = V(B)X_t + h_t \tag{9.8}$$

in which $V(B) = (F(B)/G(B))\omega(B)$. The residuals f_t and h_t are uncorrelated with u_{t-j} and X_{t-j}, respectively, but are not necessarily white noise. It will then be the case that the following theorems, originally proved by Pierce and Haugh (1977) but set out by Granger (1980a), hold.

Theorem 1

Instantaneous (prima facie) causality (in mean) exists if and only if the following equivalent conditions hold:

(i) *at least one of* $\text{cov}(a_t, b_t)$, $\gamma(0)$, $\beta(0)$ *in* (9.6) *are non-zero, or*
(ii) *at least one of* $\text{cov}(a_t, b_t)$, $H(0)$, $C(0)$ *in* (9.4) *are non-zero.*

Theorem 2

Y is not a (prima facie) cause (in mean) of X if and only if the following equivalent conditions hold:

(1) $\psi_{12}(B)$ *in* (9.3) [*equivalently* $\theta_{12}(B)$ *in* (9.5)] *can be chosen to be zero.*
(2) $\theta_{12}(B)$ *in* (9.5) *is either* 0 *or a constant.*
(3) $\psi_{12}(B)$ *in* (9.3) *is either* 0 *or proportional to* $\psi_{11}(B)$.
(4) $V_j = 0$ $(j < 0)$ *in* (9.8).
(5) $\beta(B)$ *is either* 0 *or a constant.*
(6) $H(B)$ *in* (9.4) *is either* 0 *or proportional to* $A(B)$.
(7) $\rho_{uv}(k)$ *or, equivalently,* $\omega(k) = 0$ $(k < 0)$ *in* (9.7).

If any of these conditions do not hold then Y will be a prima facie cause of X in mean with respect to $J_t(X, Y)$. Multivariate generalizations of these conditions to the vectors $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ may also be obtained. Because of the variety of equivalent conditions, numerous statistical tests may be devised and Granger pointed out that the performance of such tests would clearly need investigating, either by using statistical theory or by Monte Carlo simulation, especially as some were suspected of being occasionally biased or to be lacking in power.

An approach that emphasizes the predictive implications of causality and which marries causality testing with forecasting (see **§§9.13–9.19**) is Ashley, Granger and Schmalensee (1980). Here a bivariate model is built for X and Y and a set of one-step ahead forecasts generated for a post-sample period. These are then compared to the one-step ahead

forecasts produced by univariate models for X and Y to see if the bivariate model actually forecasts better. Ashley et al. argued that focusing attention on post-sample forecasts avoids a number of difficulties inherent in basing causality conclusions entirely on within-sample performance, illustrating their procedure with an example investigating the causal links between advertising and consumption expenditure.

9.12 Over the last 40 years Granger's concept of causality has stimulated a remarkable amount of research, be it empirical (as well as being prevalent in economics and finance, it has found application in many other fields, such as meteorology (Mosedale et al,. 2006) and neuroscience (Seth and Edelman, 2007)), theoretical (for example, Chamberlain, 1982; Florens and Mouchart, 1982), related to economic policy issues (notably Buiter, 1984), or as an integral part of modern econometric theory and practice able to unify a wide range of disparate topics (see Engle, Hendry and Richard, 1983; Hendry and Mizon, 1999). The concept has also continued to attract controversy as to whether it provides a general definition of causality: see, for example, Jacobs, Leamer and Ward (1979), Hoover (2001, 2008) and, for his last words on this all pervasive topic, Granger (2008a).

Error functions and combining forecasts

9.13 On being asked by Box and Jenkins to comment on an advance copy of *Time Series Analysis: Forecasting and Control*, Granger became interested in forecasting issues and quickly published two papers on the topic in the same volume of *Operational Research Quarterly* (*ORQ*). Box and Jenkins' univariate forecasting framework (**§§6.32–6.45**) essentially focused on linear models and a MMSE forecast criterion. Granger (1969b) took a rather more general perspective on forecasting by considering the optimum point prediction of the purely non-deterministic stationary series x_{t+l} by some, possibly non-linear, function $f_t(l) = h(x_t, x_{t-1}, \dots)$ using a general 'cost of error' function $C(e_t(l))$, where $e_t(l) = x_t - f_t(l)$. This cost function has $C(0) = 0$ and $C(|e_1|) > C(|e_2|)$ for $|e_1| > |e_2|$, so that if a forecast is made without error, no cost arises, but if there is an error then the larger it is the larger the cost. $C(e)$ does not have to be symmetrical, however, so that $C(-e)$ will not necessarily equal $C(e)$. Granger showed that the optimum function $h(x_t, x_{t-1}, \dots)$ is that which minimizes

$$E[C(x_{t+l} - f_t(l))|x_t, x_{t-1}, \dots] = [C(x_{t+l} - f_t(l))] = \int_{-\infty}^{\infty} C(x - f_t(l)) f_{c,l}(x) dx$$

where the conditional expectation notation of **§6.33** is used, $f_{c,l}(x)$ is the conditional frequency function of x_{t+l} given $x_t, x_{t-1}, \ldots$, and it is assumed that the integral exists.

The MMSE criterion uses the quadratic cost of error function $C(e_t(l)) = e_t^2(l)$, in which case, on writing $M_l = [x_{t+l}]$,

$$[x_{t+l} - h]^2 = [x_{t+l} - M_l]^2 + [M_l - h]^2$$

and the optimum predictor is $h(x_t, x_{t-1}, \ldots) = M_l$. When x_t is Gaussian, so that every finite subset of the process is normally distributed, then this optimum least squares predictor is $M_l = \sum_{j=0}^{\infty} a_j x_{t-j}$, where the a_j's are fully determined by the covariance matrix and mean vector of the multivariate normal distribution of $x_{t+l}, x_t, x_{t-1}, \ldots$. It will also be the case that $f_{c,l}(x)$ is normal with mean M_l. If x_t is not Gaussian then M_l need not be a linear function of $x_t, x_{t-1}, \ldots$, in which case the restriction to linear predictors is sub-optimal, although Granger argued that the gains in computational simplicity generally made such a choice a reasonable one.

Granger next considered the asymmetric linear cost function

$$\begin{aligned} C(e) &= ae \quad e \geq 0 \quad a > 0 \\ &= be \quad e < 0 \quad b < 0 \end{aligned}$$

for which

$$[C(x_{t+l} - f_t(l))] = a \int_h^{\infty} C(x - f_t(l)) f_{c,l}(x) dx + b \int_{-\infty}^{h} C(x - f_t(l)) f_{c,l}(x) dx$$

Granger showed that this function is minimized when $F_{c,l}(f_t(l)) = a/(a-b)$, where $F_{c,l}(x)$ is the conditional cumulative distribution function of x_{t+l}. If $C(e)$ is symmetric, so that $a = -b$, $F_{c,l}(f_t(l)) = \frac{1}{2}$, and the optimum $f_t(l)$ is the median. In the asymmetric case, the optimal forecast will be of the form $M_l + \alpha$, where α is a constant, independent of $x_t, x_{t-1}, \ldots$, obtained from $F_{c,l}(f_t(l))$ under the assumption of normality.

For more general cost functions, Granger showed that, under symmetry and some simple conditions on $C(e)$ and $f_{c,l}(x)$, the optimal predictor will be $f_t(l) = M_l$. If x_t is Gaussian then the optimum predictor will again be of the form $M_l + \alpha$; if x_t is not Gaussian M_l will be non-linear and α will be a function of $x_t, x_{t-1}, \ldots$. It would therefore appear that a MMSE approach, corresponding to a quadratic cost function, is both more general and more defensible than might have first been thought. With Gaussian data the optimal least-squares predictor will also be optimal

for all symmetric cost functions. If the cost function is asymmetric then an appropriate procedure is to obtain the best least-squares predictor and then add the bias term α depending upon the cost function being used.[4]

9.14 The second paper in *ORQ*, Bates and Granger (1969), took as its lead two sets of one-step ahead forecast errors prepared by Barnard (1963) and obtained by forecasting the airline passenger data (recall **§7.3**) using the 'Box–Jenkins method' and the 'adaptive forecasting method' attributed to Brown (1959), although in neither case were any details of the modelling and estimation provided. Bates and Granger computed the variance of the forecast errors to be 177.7 for Brown's adaptive forecasting and 148.6 for Box–Jenkins, these therefore suggesting that the latter were a clearly superior set of forecasts. They also computed the variance of the errors of a third forecast, that of the averages of the two sets, and found it to be 130.2, so that 'even though Brown's forecasts had a larger variance than that of Box–Jenkins's forecasts, they were clearly of some value' (Bates and Granger, 1969, page 452).

This prompted Bates and Granger to consider whether combinations of forecasts could be constructed using optimal, rather than a priori assigned equal, weights that would prove superior to any of the individual forecasts. They began by assuming that the performance of the individual forecasts was consistent over time (a type of stationarity assumption), so that the variances of the two sets of forecasts errors could be regarded as constants, σ_1^2 and σ_2^2, say. Under the further assumption that the forecasts are unbiased, a combined forecast was defined as the linear combination $f_T^{(c)} = kf_T^{(1)} + (1-k)f_T^{(2)}$, where $f_T^{(1)}$ and $f_T^{(2)}$ are the individual forecasts of x_T. Hence, if the forecast errors are $e_T^{(j)} = x_T - f_T^{(j)}$, $j = 1, 2$, with $E(e_T^{(j)}) = 0$, $E(e_T^{(j)^2}) = \sigma_j^2$ and $E(e_T^{(1)}e_T^{(2)}) = \rho\sigma_1\sigma_2$, so that ρ is the correlation between the individual forecast errors, the forecast error of $f_T^{(c)}$ will be

$$e_T^{(c)} = x_T - f_T^{(c)} = ke_T^{(1)} + (1-k)e_T^{(2)}$$

with variance

$$\sigma_c^2 = k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2k(1-k)\rho\sigma_1\sigma_2$$

This variance will be minimized when k is given by

$$k_0 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{9.9}$$

so that the minimum achievable error variance is

$$\sigma_{c,0}^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2} \tag{9.10}$$

It then follows that

$$\sigma_1^2 - \sigma_{c,0}^2 = \frac{\sigma_1^2(\sigma_1-\rho\sigma_2)^2}{(\sigma_1-\rho\sigma_2)^2+\sigma_2^2(1-\rho^2)} \geq 0$$

and similarly $\sigma_2^2 - \sigma_{c,0}^2 \geq 0$, the equality occurring only if $\rho = \sigma_1/\sigma_2$ or σ_2/σ_1, in which case $\sigma_{c,0}^2 = \min(\sigma_1^2, \sigma_2^2)$. The best available combined forecast should therefore outperform the better individual forecast and, in any event, it cannot do worse. From (9.9) it can be seen that $k_0 \gtreqless 0$ if and only if $\sigma_2/\sigma_1 \gtreqless \rho$. It then follows that, if $f_T^{(2)}$ is the optimal forecast based on a particular information set, any other forecast $f_T^{(1)}$ based on the same information set must be such that $\rho = \sigma_2/\sigma_1$ exactly.

Note that it is possible for k_0 to be negative: an inferior forecast may still be worth including with negative weight if its relatively high error is outweighed by a large ρ value, which would be the case if the part of x_T that is left unexplained by the poorer forecast $f_T^{(1)}$ is sufficiently strongly related to the part left unexplained by the better forecast $f_T^{(2)}$.

The behavior of $\sigma_{c,0}^2$ as ρ approaches its limiting values of -1 and $+1$ is worthy of attention. In the former case $\sigma_{c,0}^2$ tends to zero so that a perfect forecast become obtainable. Interestingly, this also appears to happen as ρ approaches $+1$ except when $\sigma_1 = \sigma_2$, in which case $\sigma_{c,0}^2 = \frac{1}{2}\sigma_1^2(1+\rho)$ and its limit is σ_1^2. This, on first sight counter-intuitive, result may be explained by considering two forecasts producing perfectly positively correlated errors $e_T^{(1)}$ and $Ae_T^{(1)}$, where A is positive:

$$e_T^{(1)} = x_T - f_T^{(1)} \qquad e_T^{(2)} = x_T - f_T^{(2)} = A(x_T - f_T^{(1)})$$

Thus

$$f_T^{(2)} = x_T - e_T^{(2)} = (1-A)x_T + Af_T^{(1)}$$

which, for $A \neq 1$, contains x_T and implies the exact relationship

$$x_T = -\frac{A}{1-A}f_T^{(1)} + \frac{1}{1-A}f_T^{(2)}$$

9.15 Of course, as it stands (9.19) is not operational unless numerical values of σ_1^2, σ_2^2 and ρ are available. Suppose that sets of forecasts have been made over the previous $T-1$ periods, yielding errors $e_t^{(j)} = x_t - f_t^{(j)}$, $t = 1, 2, \ldots, T-1$. If the combining weights are allowed to alter through time as evidence accumulates about the relative performance of the two forecasts, then the combined forecast at time T can more correctly be written as $f_T^{(c)} = k_T f_T^{(1)} + (1 - k_T) f_T^{(2)}$. In order to provide estimates of the weight k_T, Bates and Granger offered three desirable properties of combining methods:

(a) The weight should approach the optimum value given by (9.10) as the number of forecasts increases.
(b) The weight should adapt quickly if there is a lasting change in the success of one of the forecasts.
(c) The weight should vary only a little about the optimum value.

In the spirit of these properties, they suggested several combining methods, most notably

$$\hat{k}_{1T} = \frac{\sum_{t=T-\nu}^{T-1} e_t^{(2)^2}}{\sum_{t=T-\nu}^{T-1} (e_t^{(1)^2} + e_t^{(2)^2})}$$

$$\hat{k}_{2T} = \alpha \hat{k}_{2,T-1} + (1-\alpha)\hat{k}_{1T} \qquad 0 < \alpha < 1$$

$$\hat{k}_{3T} = \frac{\sum_{t=1}^{T-1} w^t (e_t^{(2)^2} - e_t^{(1)} e_t^{(2)})}{\sum_{t=1}^{T-1} w^t (e_t^{(1)^2} + e_t^{(2)^2} - 2e_t^{(1)} e_t^{(2)})} \qquad w \geq 1$$

$$\hat{k}_{4T} = \frac{\sum_{t=1}^{T-1} w^t e_t^{(2)^2}}{\sum_{t=1}^{T-1} w^t (e_t^{(1)^2} + e_t^{(2)^2})} \qquad w \geq 1$$

Bates and Granger applied these combining formulae to the airline passenger data using pairs of univariate one-step ahead forecasts, and generally obtained successful results. This prompted a wider study by Newbold and Granger (1974), in which three one-step ahead forecasts (Box–Jenkins, Holt–Winters exponential smoothing and stepwise autoregression) were combined for 80 monthly economic series. A further combining method, based directly on (9.9), was also proposed:

$$\hat{k}_{5T} = \frac{\sum_{t=T-\nu}^{T-1} (e_t^{(2)^2} - e_t^{(1)} e_t^{(2)})}{\sum_{t=T-\nu}^{T-1} (e_t^{(1)^2} + e_t^{(2)^2} - 2e_t^{(1)} e_t^{(2)})}$$

This can be regarded as the ML estimator of k_0 if the forecast errors are assumed to be bivariate normally distributed or, alternatively, as a least squares estimator given that the combined forecast at time t can be written

$$f_t^{(c)} - f_t^{(2)} = k(f_t^{(1)} - f_t^{(2)}) \tag{9.11}$$

or

$$e_t^{(2)} = k(e_t^{(2)} - e_t^{(1)}) + e_t^{(c)}$$

Forecasts were combined in pairs and also with all three combined together (Newbold and Granger, 1974, equations (1) to (5), provided multivariate extensions of the various combining formulae), with settings of $v = 1, 3, 6, 9, 12$, $\alpha = 0.5, 0.7, 0.9$ and $w = 1, 1.5, 2, 2.5$, and with the weights constrained to lie between zero and unity. A concise summary of the results was provided in Granger and Newbold (1986, chapter 9.2). The general findings were that methods which ignored correlation between forecast errors ($\hat{k}_1$, $\hat{k}_2$ and $\hat{k}_4$) performed better than those ($\hat{k}_3$ and $\hat{k}_5$) that attempted to take it into account and that, overall, in pairwise combining, $\hat{k}_2$ with $v = 12$ produced the best forecasts in terms of mean square error. Combining stepwise autoregression and Holt–Winters exponential smoothing, both automatic forecasting procedures, was found to perform competitively with Box–Jenkins forecasting. Including a third forecast produced a marginal improvement in forecast accuracy.

Further results on combining were provided by Granger and Newbold (1975), in which econometric model forecasts and Box–Jenkins forecasts of inventory investment were combined. Although the Box–Jenkins forecast was on average considerably better than the econometric forecast, the combined forecast, in line with previous results, produced a notable further improvement in forecast accuracy.

9.16 Judging from the published discussion of Newbold and Granger (1974), the potential improvements in accuracy from forecast combining were initially regarded with some skepticism, particularly by Gwilym Jenkins (see pages 148–150 of the discussion). Nevertheless, the idea quickly took hold and twenty years after it was first introduced a special section on combining forecasts was published in the *International Journal of Forecasting* (volume 5, number 4), which included a major survey by Clemen (1989) containing an annotated bibliography of more than 200

works, followed by a special issue of the *Journal of Forecasting* (volume 8, number 3) on the topic.[5]

Granger's major contribution to this literature was to point out, in Granger and Ramanathan (1984), that although the least squares interpretation of (9.11) implies that $e_t^{(c)}$ is uncorrelated with $f_T^{(1)} - f_T^{(2)}$, it may not necessarily be uncorrelated with $f_T^{(1)}$ and $f_T^{(2)}$ individually, so that $e_t^{(c)}$ may be forecastable from them, in which case the combination will not be optimal. Relaxing the implied condition that the combining weights sum to unity loses the unbiasedness property, so Granger and Ramanathan suggested including the unconditional mean $E(x_T) = m$ in the combination:

$$f_T^{(c*)} = \alpha_1 f_T^{(1)} + \alpha_2 f_T^{(2)} + \alpha_3 m \qquad \alpha_1 + \alpha_2 + \alpha_3 = 1$$

These weights can be estimated from the unconstrained regression

$$x_t = \alpha_1 f_t^{(1)} + \alpha_2 f_t^{(2)} + a + e_t^{(c*)}$$

where, by construction, $e_t^{(c*)}$ is uncorrelated with $f_T^{(1)}$ and $f_T^{(2)}$.

If $f_T^{(1)}$ and $f_T^{(2)}$ are based on the same information set, finding $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$ would suggest that neither forecast can be considered to be optimal. For example, the two forecasts may be based on different assumptions about functional form, linear or logarithmic say. If a forecast combination successfully beats both individual forecasts then this suggests that the best functional form is neither of those originally selected. If the individual forecasts are based on different information sets then a combined forecast with non-zero weights would suggest that a model which combines the two information sets should be considered.

In his contribution to the *Journal of Forecasting*'s special issue, Granger (1989a) formalized these ideas. Suppose there are N forecasters, with the jth having the information set $I_{jT} : I_{0T}, J_{jT}$ at time T, where I_{0T} is the information available to everyone and J_{jT} is the information available only to the jth forecaster, the contents of which are assumed to be independent of I_{0T} and J_{kT}, $k \neq j$. Assuming, for convenience, that each forecaster has only a single series in their information set, then $I_{0T} : z_{T-s}$ and $J_{jT} : x_{j,T-s}$, $s \geq 0$. The universal information set, consisting of all the information available to all forecasters, is $U_T : I_{0T}, J_{1T}, \ldots, J_{NT}$. If y_t is the series being forecast, then the optimum linear least squares forecast of y_T if U_{T-1} was known will be

$$F_U = E[y_T | U_{T-1}] = a(B) z_{T-1} + \sum_{j=1}^{N} \beta_j(B) x_{j,T-1}$$

The optimal forecast of the jth forecaster, on the other hand, is

$$F_j = E[y_T|I_{j,T-1}] = a(B)z_{T-1} + \beta_j(B)x_{j,T-1}$$

Forming a simple average of these individual forecasts gives

$$\bar{F} = \frac{1}{N}\sum_{j=1}^{N} F_j = a(B)z_{T-1} + \frac{1}{N}\sum_{j=1}^{N}\beta_j(B)x_{j,T-1} \tag{9.12}$$

the second term of which is the sum of N independent components divided by N and so will have a standard deviation of order $N^{-1/2}$. Thus, if N is large, $\bar{F} = a(B)z_{T-1}$, which is the forecast made using just $I_{0,T-1}$. This result will generally hold for any other set of weights summing to one, so that F_U cannot be obtained by optimally combining the individual forecasts F_j. However, if the additional forecast

$$F_0 = E[y_T|I_{0,T-1}] = a(B)z_{T-1}$$

becomes available, then the optimal forecast can be achieved by setting

$$F_U = \sum_{j=1}^{N} F_j - (N-1)F_0 \tag{9.13}$$

Granger used this analysis to illustrate a number of general points.

(i) Aggregating forecasts is not the same as aggregating information sets: $\bar{F}$ is based on all available information but is not equal to F_U as the information is not being used efficiently.
(ii) Equal weight combinations, as in (9.12), will be useful if each information set contains both common and independent components. If the amount of shared information varies across forecasters, unequal weights will usually result.
(iii) A new forecast can improve the combined forecast even if it is not based on new information, e.g., F_0.
(iv) Negative weights can be useful, as in (9.13).
(v) It is also useful to include as many forecasts as is possible in the combination, again as in (9.13).

Granger (1989a) went on to discuss various extensions of the combining technique to time varying weights, possibly based on second

moments of the forecast errors, as in Engle, Granger and Kraft (1984), to non-stationary situations, and to combining forecasts of quantiles rather than just means (Granger, White and Kamstra, 1989). He also discussed the links between a dominant forecast (a forecast with a zero weight in the combination is said to be dominated by the other forecast) and the deeper econometric concept of 'encompassing' proposed in Mizon and Richard (1986). Forecast combining using changing weights derived from non-linear models was examined in Deutsch, Granger and Teräsvirta (1994).

Forecast evaluation

9.17 The early 1970s witnessed the building of the first generation of large-scale econometric models and, with the publication of the forecasts from those models, attention began to be focused on how such forecasts should be evaluated. Granger and Newbold (1973, page 35) were critical of the evaluation procedures then in use: '(m)ost of the available techniques … are almost entirely concerned with discovering the "best" forecasting methods from some set or, equivalently, in ranking methods. We … argue that much of this work has little or no operational significance and that a wider viewpoint is required'.

Granger and Newbold began by considering the forecasts from an autoregressive model. From **§6.35**, $f_t(l) = \pi(B)[x_{t+l-1}]$, where $[x_{t+j}] = f_t(j), j = 1, \ldots, l-1$ and $[x_{t-j}] = x_{t-j}, j = 0, 1, \ldots$. One-step ahead forecast errors are given by $e_{t,1} = x_{t+1} - f_t(1) = \varepsilon_{t+1}$ and constitute a zero-mean white noise process with variance $\sigma_\varepsilon^2 = \sigma_x^2 - V(f_t(1))$. Thus the variance of the optimum one-step ahead forecast will be less than the variance of the series that is being forecast. The spectrum of $f_t(1)$ will be $|\pi(e^{i\omega})|^2 f_x(\omega)$ and hence $f_t(1)$ will also have time series properties that are different to the series it is attempting to forecast. The cross-spectrum between $f_t(1)$ and x_{t+1} is $\pi(e^{i\omega}) f_x(\omega)$, so that the coherence is unity at every frequency but the phase diagram will generally be complicated and difficult to interpret. Granger and Newbold (1973, page 35) were thus led to conclude that

> the optimal predictor generally has different distributional and time series properties than the series being predicted. … It therefore follows that it is pointless to compare, as many practitioners do, the distributional or time series properties of the predictor and predicted series. … (R)ather than consider properties of predictor and actual series separately, the most fruitful approach is to consider the

distributional and time series properties of the forecast error series, particularly the one-step errors.

Thus, if there are T forecast errors $e_t = x_t - f_t$ (the l-step ahead nature of such forecasts being suppressed for clarity of notation) and a quadratic cost of error criterion is used, then the MSE is

$$D_T^2 = \frac{1}{T}\sum_{t=1}^{T} e_t^2$$

Theil's (1958) first U-statistic was an early attempt at evaluating a set of forecasts. Defined as

$$U_1 = \frac{D_T}{\left(\frac{1}{T}\sum f_t^2\right)^{1/2} + \left(\frac{1}{T}\sum x_t^2\right)^{1/2}}$$

this 'inequality coefficient' clearly takes the value zero if f_t is a perfect forecast of x_t, but, as Granger and Newbold demonstrated, in general its use could be rather problematic. To show this, suppose that x_t is generated by an AR(1) process with autoregressive coefficient α but that a set of one-step ahead suboptimal forecasts are made using $f_t = \beta x_{t-1}$, $0 \leq \beta \leq 1$. In the limit as $T \to \infty$,

$$\lim_{T\to\infty} D_T^2 = ((1-\alpha^2) + (\beta - \alpha)^2)V(x) \qquad V(f) = \beta^2 V(x)$$

so that, after some algebra,

$$\lim_{T\to\infty} U_1^2 = 1 - \frac{2\beta(1+\alpha)}{(1+\beta)^2}$$

This expression is minimized when $\beta = 1$ and not for the optimal forecast for which $\beta = \alpha$, so that U_1 can fail to select the optimum forecast from a group of forecasts which includes it! The reason for this is that the denominator of U_1 contains the variance of the forecasts, which will be highest (and U_1 lowest) when $\beta = 1$.

9.18 An important part of any forecast evaluation is an assessment of the quality of a set of forecasts. As comparisons with a competitor forecast are often impossible, Theil's (1966) second U-statistic compares the performance of a set of forecasts with that of a simple 'no change' rule,

in which forecasts are set at the most recent observed value:

$$U_2^2 = \frac{D_T^2}{\sum x_t^2}$$

This will be minimized when the expected MSE,

$$S = E(D_T^2) = E(x_T - f_T)^2 = (\mu_x - \mu_f)^2 + \sigma_x^2 + \sigma_f^2 - 2\rho_{xf}\sigma_x\sigma_f \quad (9.14)$$

is minimized, where μ_x and μ_f are the means of the observations and the forecasts, respectively, σ_x and σ_f are their standard deviations, and ρ_{xf} is the correlation between them. Since

$$\frac{\partial S}{\partial \mu_f} = -2(\mu_x - \mu_f) \qquad \frac{\partial S}{\partial \sigma_f} = -2(\sigma_f - \rho_{xf}\sigma_x) \qquad \frac{\partial S}{\partial \rho_{xf}} = -2\sigma_x\sigma_f$$

S will be minimized by taking ρ_{xf} as large as possible with $\mu_x = \mu_f$ and $\sigma_f = \rho_{xf}\sigma_x$, so that the two standard deviations should optimally not be equal except when $\rho_{xf} = 1$, i.e., for deterministic processes for which the forecasts are perfectly correlated with the actual series.

Theil (1958) proposed two MSE decompositions based on rewriting (9.14) as

$$S = (\mu_x - \mu_f)^2 + (\sigma_x - \sigma_f)^2 + 2(1 - \rho_{xf})\sigma_x\sigma_f \quad (9.15)$$

and

$$S = (\mu_x - \mu_f)^2 + (\sigma_f - \rho_{xf})^2 + (1 - \rho_{xf}^2)\sigma_x^2 \quad (9.16)$$

The decomposition (9.15) leads to the definition of the following quantities in terms of sample statistics:

$$U^M = (\bar{x} - \bar{f})/D_T^2 \qquad U^S = (s_x - s_f)^2/D_T^2 \qquad U^C = 2(1 - r_{xf})s_x s_f/D_T^2$$

Clearly $U^M + U^S + U^C = 1$ and Theil suggested that the three quantities had useful interpretations. Granger and Newbold doubted this, however, and pointed out that, if x_t was again generated by an AR(1) process with autoregressive parameter α, as in **§9.16**, the limiting values of the quantities would be

$$\lim_{T\to\infty} U^M = 0 \qquad \lim_{T\to\infty} U^S = \frac{1-\alpha}{1+\alpha} \qquad \lim_{T\to\infty} U^C = \frac{2\alpha}{1+\alpha}$$

Thus, as α varies from 0 to 1, U^S and U^C can take on any values subject only to the restrictions $0 \leq U^S, U^C \leq 1$ and $U^S + U^C = 1$, so that interpretation of these quantities is impossible. In general, the problem is that some series are inherently difficult to forecast so that a large value of ρ_{xf} may be difficult to achieve, in which case the standard deviation of the optimal forecasts will be much lower than that of the observed series, so that U^S differs substantially from zero. For more predictable series the value taken by U^S can be expected to be lower for optimal forecasts.

The decomposition (9.16) again leads to the quantity U^M, now accompanied by

$$U^R = \frac{(s_f - r_{xf}s_x)^2}{D_T^2} \qquad U^D = \frac{(1 - r_{xf}^2)s_x^2}{D_T^2}$$

U^R will tend to zero along with U^M for optimal forecasts and hence U^D should be close to unity. These requirements may be placed in an alternative context. Consider the regression of the actual values on the forecasts, i.e., $x_t = a + bf_t$. Mincer and Zarnowitz (1969) called a forecast 'efficient' if $a = 0$ and $b = 1$ and these restrictions are equivalent to U^M and U^R both being zero. Granger and Newbold raised several objections to this definition , pointing out that, if x_t is generated by a random walk, then the entire set of one-step ahead predictors $f_t(l) = x_{t-l}$, $l = 1, 2, \ldots$, will in theory lead to $a = 0$ and $b = 1$ so that all these forecasts would be deemed to be 'efficient' by this criterion. Granger and Newbold regarded the criterion as a necessary condition for forecast efficiency but by no means a sufficient one, although they suggested that such a regression could be examined within the context of Theil's 'prediction-realization diagram', in which a plot of forecasts against actuals is made, yielding a spread of points around the 'line of perfect forecasts' $x_t = f_t$.

Even this approach has drawbacks, for Granger and Newbold argued that such a plot would invariably look very impressive whenever the series being forecast follows an integrated process (recall **§8.4** – in particular Figure 8.2 – where Box and Newbold (1971) demonstrated that a random walk can give reasonable predictions of another *independent* random walk). Their solution was to plot predicted against actual *changes*, since these will be much less smooth than the levels.

9.19 Finally, Granger and Newbold argued that what was of primary importance in forecast evaluation was an examination of the forecast errors themselves, particularly the one-step ahead errors, which should be zero-mean white noise. More generally, optimal *l*-step ahead forecast

errors should have autocorrelations of order l or higher equal to zero, for otherwise the forecast error would be correlated with some information that was known at the time the forecast was made, and so the forecast could have been improved upon. As it is unlikely that a sufficient number of forecasts errors are available for detailed statistical analysis, Granger and Newbold suggested that, at the very least, the one-step ahead errors should be tested for randomness, against the alternative of first-order autocorrelation, by using the von Neumann ratio

$$\frac{T}{T-1}\frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T}(e_t - \bar{e})^2}$$

Forecasting transformed series

9.20 It is very often the case that, rather than model and subsequently forecast the observed series x_t, a function of the series is analyzed instead. The most obvious example of this is the differencing operation Δ^d employed to remove homogenous non-stationarity, but a second example that is regularly encountered is the logarithmic transformation $y_t = \log x_t$. More generally, Granger and Newbold (1976) considered the implications of forecasting the *instantaneously* transformed series $y_t = T(x_t)$, where $T(\ )$ is some well-behaved function, an example of which is the well-known Box and Cox (1964) transformation $y_t = ((x_t - m)^\theta - 1)/\theta$, which might be used to produce data that is nearer to Gaussianity.

Granger and Newbold assumed that x_t was a stationary, Gaussian series with mean μ, variance σ^2 and autocorrelation sequence $\rho_{x,k}$, from which it follows that $z_t = (x_t - \mu)/\sigma$ is a stationary, Gaussian series with zero mean, unit variance and the same autocorrelations $\rho_{x,k}$. They then considered the instantaneous transformation $y_t = T(z_t)$, where $T(\ )$ can be expanded in terms of *Hermite polynomials* in the form

$$T(z) = \sum_{j=0}^{M} \alpha_j H_j(z)$$

The jth Hermite polynomial $H_j(z)$ is a polynomial in z of order j with, for example, $H_0(z) = 1$, $H_1(z) = z$, $H_2(z) = z^2 - 1$, $H_3(z) = z^3 - 3z$ and, in general,

$$H_j(z) = j! \sum_{r=0}^{[j/2]} (-1)^r (2^r r!(j-2m)!)^{-1} z^{j-2r}$$

If X and Y are standard normal random variables with correlation ρ, Hermite polynomials have the following orthogonality properties

$$\begin{aligned} E[H_j(X)H_i(X)] &= 0 \qquad j \neq i \\ &= j! \qquad j = i \end{aligned}$$

and

$$\begin{aligned} E[H_j(X)H_i(Y)] &= 0 \qquad j \neq i \\ &= \rho^j j! \qquad j = i \end{aligned}$$

Using these properties, Granger and Newbold showed that $E(y_t) = \alpha_0$ and

$$Cov(y_t, y_{t-k}) = \sum_{j=1}^{M} \alpha_j^2 j! \rho_{x,k}^j \qquad Cov(x_t, y_{t-k}) = \alpha_1 \rho_{x,k} \sigma$$

so that the linear properties of the transformed series can be determined. For example, the autocorrelation sequence of y_t is

$$\rho_{y,k} = \sum_{j=1}^{M} \alpha_j^2 j! \rho_{x,k}^j \Big/ \sum_{j=1}^{M} \alpha_j^2 j! \tag{9.17}$$

and, if $\rho_{x,k} \neq 0$ for some k, it follows that $|\rho_{y,k}| < |\rho_{x,k}|$ for $M > 1$, so that the transformed series is 'closer' to white noise than the original series.

As an example, Granger and Newbold first considered the quadratic transformation

$$\begin{aligned} y_t &= a + bx_t + cx_t^2 = a + b(\sigma z_t + \mu) + c(\sigma z_t + \mu)^2 \\ &= a + b\mu + c\mu^2 + (b + 2c\mu)\sigma z_t + c\sigma^2 z_t^2 \end{aligned}$$

In terms of Hermite polynomials this can be written as

$$\begin{aligned} y_t &= T(z_t) = \alpha_0 H_0(z_t) + \alpha_1 H_1(z_t) + \alpha_2 H_2(z_t) \\ &= \alpha_0 + \alpha_1 z_t + \alpha_2 (z_t^2 - 1) \end{aligned}$$

where

$$\alpha_0 = a + b\mu + c(\mu^2 + \sigma^2) \qquad \alpha_1 = (b + 2c\mu)\sigma^2 \qquad \alpha_2 = c\sigma^2$$

The auto-covariance sequence of y_t is then

$$Cov(y_t, y_{t-k}) = \alpha_1^2 \rho_{x,k} + 2\alpha_2^2 \rho_{x,k}^2 = (b + 2c\mu)^2 \sigma^2 \rho_{x,k} + 2c^2 \sigma^2 \rho_{x,k}^2$$

Granger and Newbold used this result to show that the quadratic transformation of an AR(p) process will be ARMA($\frac{1}{2}p(p+3)$, $\frac{1}{2}p(p+1)$), so that a quadratic transformation of an AR(1) will be an ARMA(2, 1), for example. On the other hand, a quadratic transformation of an MA(q) will also be a moving average process of at most order q.

They then specialized these results to consider the square of the AR(1) process $(x_t - \mu) = \phi(x_{t-1} - \mu) + a_t$, for which $\sigma^2 = \sigma_a^2/(1 - \phi^2)$ and $\rho_{x,k} = \phi^k$. The transformation $y_t = x_t^2$ thus has, since $a = b = 0$ and $c = 1$, $E(y_t) = \mu^2 + \sigma^2$ and the auto-covariance structure

$$Cov(y_t, y_{t-k}) = \alpha_1^2 \rho_{x,k} + 2\alpha_2^2 \rho_{x,k}^2 = \sigma^2(4\mu^2 \phi^k + 2\sigma^2 \phi^{2k})$$

The transformed series y_t will therefore have the same auto-covariance structure as $y_{1,t} + y_{2,t}$, where $y_{1,t} = \phi y_{1,t-1} + a_{1,t}$ and $y_{2,t} = \phi^2 y_{2,t-1} + a_{2,t}$, where $a_{1,t}$ and $a_{2,t}$ are independent white noises with variances $4\mu^2\sigma^2(1 - \phi^2)$ and $2\sigma^4(1 - \phi^4)$, respectively. Hence y_t has the auto-covariance properties of the ARMA(2, 1) series

$$(1 - \phi B)(1 - \phi^2 B) y_t = (1 - \phi^2 B) e_{1,t} + (1 - \phi B) e_{2,t} = (1 - \Phi B) e_t$$

If the mean μ is very large compared to the standard deviation σ the behavior of the transformed series will be approximately AR(1), rather than ARMA(2, 1).

The exponential transformation can be written as

$$y_t = \exp(x_t) = \exp(\mu + \sigma z_t) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \sum_{j=0}^{\infty} \frac{\sigma^j}{j!} H_j(z_t)$$

Hence $E(y_t) = \exp(\mu + \frac{1}{2}\sigma^2)$ and the covariance sequence is

$$\begin{aligned} Cov(y_t, y_{t-k}) &= \exp(2\mu + \sigma^2) \sum_{j=1}^{\infty} \frac{(\sigma^2 \rho_{x,k})^j}{j!} \\ &= \exp(2\mu + \sigma^2)(\exp(\sigma^2 \rho_{x,k}) - 1) \end{aligned} \tag{9.18}$$

These results prompted Granger and Newbold (1976, page 192) to conclude that 'if the series to be fitted is subjected to instantaneous

transformation, it can be the case (except for moving average processes) that the autocorrelation function of the transformed series exhibits markedly different behavior patterns from that of the original series'.

Granger and Newbold also considered instantaneous transformations of integrated processes, showing that if Δx_t follows an ARMA process then so will Δy_t. In particular, if Δx_t is AR(1) then Δx_t^2 will be ARMA(2, 2).

9.21 Suppose now that a forecast $f_T(l)$ has been made for the Gaussian series x_t but, rather than x_{T+l}, a forecast is actually required for

$$y_{T+l} = T\left(\frac{x_{T+l} - \mu}{\sigma}\right)$$

A typical example would be where x_t represents the logarithm of the variable of interest but forecasts are required for the variable itself and not its logarithms.

Several forecasts of y_{T+l} were considered by Granger and Newbold. The first is the optimal quadratic loss forecast $g_T^{(1)}(l) = E[y_{T+l}|I_T]$, where $I_T : x_{T-j}$, $j \geq 0$. Defining $e_T^{(x)}(l) = x_{T+l} - f_T(l)$ to be the l-step ahead optimal forecast error of x_{T+l} and $s^2(l)$ to be the associated forecast error variance then, with $w_{t+l} = e_t^{(x)}(l)/s(l)$, y_{T+l} can be written as

$$y_{T+l} = \sum_{i=0}^{M} \gamma_i H_i(w_{T+l})$$

and it follows that $g_T^{(1)}(l) = \gamma_0$. The expected squared forecast error conditional on I_T is

$$V_c^{(1)}(l) = \sum_{j=1}^{M} \gamma_j^2 j!$$

The optimal, generally non-linear, forecast is given by

$$g_{T,o}^{(1)}(l) = \sum_{j=0}^{M} \alpha_j A_j H_j(P)$$

where

$$A = (1 - s^2(l)/\sigma^2)^{1/2} \qquad P = \frac{f_T(l) - \mu}{(\sigma^2 - s^2(l))^{1/2}}$$

Thus $E[g_{T,o}^{(1)}(l)] = E[y_{t+l}] = \alpha_0$ and the variance of the unconditional expected squared forecast error can be shown to be

$$V^{(1)}(l) = E[e_T^{(y)}(l)] = E[y_{T+l} - g_{T,o}^{(1)}(l)] = \sum_{j=1}^{M} \alpha_j^2 j!(1 - A^{2j})$$

Granger and Newbold then defined a measure of forecastability as the ratio of the variance of the optimal forecast of y_{T+l} to the unconditional variance:

$$R_{y,l}^2 = \frac{V(g_T^{(1)}(l))}{V(y_{T+l})} = \frac{\sum_{j=1}^{M} \alpha_j^2 j! A^{2j}}{\sum_{j=1}^{M} \alpha_j^2 j!}$$

Since $0 \leq R_{x,l}^2 = A^2 < 1$ it thus follows that $R_{y,l}^2 < R_{x,l}^2$ for $M > 1$, so that the transformed series y is always less forecastable than the original series x and, in this sense, is 'nearer white noise'. An interesting corollary was considered by Granger (1983): can a forecastable series be transformed completely to white noise? It is clear from (9.17) that, if any $\rho_{x,k}$ is positive, then the corresponding $\rho_{y,k}$ will also be positive and so y_t cannot be white noise: thus, for example, no AR(1) process can be transformed to white noise as $\rho_{x,2}$ must always be positive. However, consider the MA(1) process $x_t = \varepsilon_t + b\varepsilon_{t-1}$ and the transformation

$$y_t = \alpha_1 x_t + \alpha_2 (x_t^2 - 1)$$

If

$$\frac{\alpha_1^2}{2\alpha_2^2} = -\frac{b}{1 + b^2}$$

so that b and $\rho_{x,1} = b/(1 + b^2)$ must both be negative, then

$$\rho_{y,1} = \frac{\alpha_1^2 \rho_{x,1} + 2\alpha_2^2 \rho_{x,1}^2}{\alpha_1^2 + 2\alpha_2^2} = 0$$

and, as $\rho_{x,k} = 0$ for $k > 1$, it follows that $\rho_{y,k} = 0$ for all $k > 0$. Thus it is possible for a series that cannot be forecast linearly from its own past (here y_t) to be transformed into a series

$$x_t = \left(-\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2 (y_t + \alpha_2)} \right) / 2\alpha_2$$

that is forecastable. In fact, if x_t were known then y_t could be forecast from it, although Granger (1983) wondered how such 'hidden non-linear forecastability' would ever be detected.

For the transformation $y_{t+l} = \exp(x_{t+l})$ the optimal conditional forecast is given by

$$g_T^{(1)}(l) = \exp\left(f_T(l) + \tfrac{1}{2}s^2(l)\right)$$

with

$$\begin{aligned} V_c^{(1)}(l) &= \exp(2f_T(l) + s^2(l)) \sum_{j=1}^{\infty} \frac{s^{2j}(l)}{(j!)^2} j! \\ &= \exp(2f_T(l) + s^2(l))(\exp(s^2(l)) - 1) \end{aligned}$$

which may be compared with the unconditional variance

$$V^{(1)}(l) = \exp(2(\mu + \sigma^2))(1 - \exp(-s^2(l)))$$

9.22 The second forecast considered by Granger and Newbold was the 'naïve' forecast $g_T^{(2)}(l) = T((f_T(l) - \mu)/\sigma)$ obtained by substituting $f_T(l)$ for x_{T+l} in the transformation $T(\ )$. Since we can write

$$y_{T+l} - g_T^{(2)}(l) = (y_{T+l} - g_T^{(1)}(l)) + (g_T^{(1)}(l) - g_T^{(2)}(l))$$

the conditional expected squared forecast error has variance

$$V_c^{(2)}(l) = V_c^{(1)}(l) + (g_T^{(1)}(l) - g_T^{(2)}(l)) = \sum_{j=1}^{M} \gamma_j^2 j! + (g_T^{(1)}(l) - g_T^{(2)}(l))$$

while the variance of the unconditional expected squared error is

$$V^{(2)}(l) = V^{(1)}(l) + \sum_{j=0}^{M} A^{2j}(j!)^{-1} \left(\sum_{i=1}^{[\frac{1}{2}(M-j)]} \alpha_{j+2i} \frac{(j+2i)!}{i!} \left(-\tfrac{1}{2}(1-A^2)\right)^i \right)^2$$

with the second term representing the average amount lost in squared error through the use of the naïve predictor.

For the exponential transformation, the biased naïve forecast is $g_T^{(2)}(l) = \exp(f_T(l))$ and it can be shown that

$$V^{(2)}(l) = \exp(2(\mu+\sigma^2)) \times \left(1-\exp(-s^2(l)) + \left[\exp\left(-\tfrac{1}{2}s^2(l)\right) - \exp(-s^2(l))\right]^2\right)$$

so that use of the naïve predictor leads to a proportionate increase in expected squared forecast error of

$$\frac{V^{(2)}(l)-V^{(1)}(l)}{V^{(1)}(l)} = \frac{(\exp(-\frac{1}{2}s^2(l)) - \exp(-s^2(l)))^2}{1-\exp(-s^2(l))}$$

9.23 Granger and Newbold then considered forecasting the transformed variable y_t either as a linear combination of the elements of the information set I_T, that is, current and past values of x_t, or as a linear combination of current and past values of y_t itself. In the former case, they showed that the optimal forecast of y_{T+l} under quadratic loss is

$$g_T^{(3)}(l) = \alpha_0 + \alpha_1\left(\frac{f_T(l)-\mu}{\sigma}\right)$$

with unconditional and conditional expected squared forecast errors

$$V^{(3)}(l) = \sum_{j=2}^{M}\alpha_j^2 j! + \alpha_1^2 s^2(l)/\sigma^2$$

and

$$V_c^{(3)}(l) = V_c^{(1)}(l) + (g_T^{(1)}(l) - g_T^{(3)}(l))$$

In the latter case the forecast, $g_T^{(4)}(l)$, is, in principle, given by matching the known auto-covariance structure

$$Cov(y_t, y_{t-k}) = \sum_{j=1}^{M}\alpha_j^2 j!\rho_{x,k}^j$$

with a specific ARMA model $\phi(B)(y_t-\alpha_0) = \theta(B)a_t$ and forecasting in the usual manner (cf. **§§6.32–6.45**), so producing the forecast error variance

$$V^{(4)}(l) = \sigma_a^2(1+\psi_1^2+\cdots+\psi_{l-1}^2)$$

where $\psi(B) = \phi^{-1}(B)\theta(B)$ and σ_a^2 is the variance of the white noise a_t. This may not always be straightforward in practice and Granger and Newbold suggested that the ARMA model for y_t may need to be

approximated by a moving average and Wilson's (1969) algorithm for obtaining the ψ_j's and σ_a^2 from a given set of autocovariances used to enable $g_T^{(4)}(l)$ and $V^{(4)}(l)$ to be computed.

Granger and Newbold illustrated these results using the MA(1) process $x_t - \mu = \varepsilon_t + 0.5\varepsilon_{t-1}$, where the white noise ε_t will have variance $0.8\sigma^2$ and the autocorrelations of x_t are $\rho_{x,1} = 0.4$ and $\rho_{x,k} = 0$, $k > 1$. If $y_t = \exp(x_t)$, then using (9.18) yields

$$\rho_{y,k} = (\exp(\sigma^2 \rho_{x,k}) - 1)/(\exp(\sigma^2) - 1)$$

and so

$$\rho_{y,1} = (\exp(0.4\sigma^2) - 1)/(\exp(\sigma^2) - 1), \qquad \rho_{y,k} = 0, \quad k > 1$$

i.e., the transformed process is also MA(1). If $\sigma^2 = 1$ then $\rho_{y,1} = 0.286$ and y_t follows the process

$$y_t - \exp\left(\mu + \tfrac{1}{2}\right) = a_t + 0.31a_{t-1}$$

where $\sigma_a^2 = (1 + 0.31^2)^{-1} E(y_t^2) = 4.26 \exp(2\mu)$. Hence

$$V^{(4)}(1) = 4.26 \exp(2\mu), \qquad V^{(4)}(l) = 4.67 \exp(2\mu), \qquad l > 1$$

Since $s^2(1) = 0.8$ and $s^2(l) = 1$ for $l \geq 1$, the optimum forecast $g_T^{(1)}(l)$ will have $V^{(1)}(1) = 4.07 \exp(2\mu)$ and $V^{(1)}(l) = 4.67 \exp(2\mu) = V^{(4)}(l)$ for $l > 1$, so that there is no loss in using the linear model for forecasting more than one step ahead, although for one-step ahead forecasts the proportionate increase in MSE from using the linear model is approximately 5%:

$$\frac{V^{(4)}(1) - V^{(1)}(1)}{V^{(1)}(1)} = 0.047$$

This analysis led Granger and Newbold (1976, page 201) to conclude that

> one can frequently do much better than using $g_T^{(2)}(l)$, the naïve forecast, in which the optimum forecast of x_{T+l} is inserted into the transforming function. For many of the models and transformations met in practice, it is possible to find the optimum forecast for y_{T+l} and this is to be recommended. However, some extra effort is required to

> do this and for speed one could use the naïve forecast or the linear forecast of y_{T+l} based on x_{T-j}, $j \geq 0$, i.e., $g_T^{(3)}(l)$. Neither is necessarily near to the optimal forecast and one is not clearly always superior to the other, but both are easily obtained. A better compromise might be to combine these two simple forecasts, … but no exercise is yet available on combining this particular pair of forecasts. The methods … enable a wide class of transformations and models to be considered, although in some cases the amount of algebraic manipulation required to find the optimum forecast, or the loss from using a sub-optimal one, is rather large. (Notation altered for consistency)

9.24 Nelson and Granger (1979) applied this analysis to the forecasting performance of the Box–Cox transformation $y_t = x_t^{(\theta)} = (x_t^\theta - 1)/\theta$, for which $y_t = \log(x_t)$ as $\theta \to 0$.[6] The naïve forecast of x_{T+l} is then given by

$$\begin{aligned} g_T^{(\theta)}(l) &= (\theta f_T^{(\theta)}(l) + 1)^{1/\theta}, \qquad \theta \neq 0 \\ &= \exp(f_T^{(0)}(l)), \qquad \theta = 0 \end{aligned}$$

where $f_T^{(\theta)}(l)$ is the forecast of $y_{T+l} = x_{T+l}^{(\theta)}$. Unfortunately, there is no closed form for the optimal l-step ahead forecast, $g_{T,\sigma}^{(\theta)}(l)$, and it has to be obtained, under the assumption of normality, from the integral

$$\frac{1}{s_\theta(l)\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp -\tfrac{1}{2} \left(\frac{z - f_T^{(\theta)}}{s_\theta(l)} \right)^2 (\theta z - 1)^{1/\theta} dz$$

In an extensive empirical exercise using five alternative forecasts over twenty economic time series and up to ten forecasting horizons, Nelson and Granger (1979, pages 68, 69) found that 'when the necessary underlying assumptions are true, the Box-Cox transformation works well and does produce superior forecasts when a transformation is really justified'; they concluded, however, that 'the extra inconvenience, effort and cost is usually such as to make the use of these transformations not worthwhile', arguing that the 'main problem seems to be the extreme non-normality of actual economic data, and the use of the transformation does not dramatically reduce the problem'.

Later thoughts on forecast evaluation and related issues

9.25 Granger's research on forecasting up to the end of the 1980s was effectively distilled into the second editions of the monograph *Forecasting*

Economic Time Series (Granger and Newbold, 1986), which emphasized the technical aspects of the subject, and the more general and descriptive text, *Forecasting in Business and Economics* (Granger, 1989b). He did, however, continue to research into various aspects of forecasting, some of which will be discussed in later sections of this and the next chapter. We shall primarily concentrate here on a theme that Granger continually returned to, that of the evaluation of forecasts.

9.26 Granger and Pesaran (2000a, 2000b: see also Granger and Machina, 2006) developed a 'decision theoretic' approach to forecast evaluation because, in their view, '(i)n the real, non-academic world forecasts are made for a purpose ... (which is) to help decision makers improve their decisions. It follows that the correct way to evaluate forecasts is to consider and compare realized values of different decisions made from alternative sets of forecasts' (Granger and Pesaran, 2000b, page 537). This approach focuses on predictive distributions rather than point forecasts and on the evaluation of probability forecasts using the concept of 'economic value', rather than on cost functions using forecast errors.

The set-up in Granger and Pesaran (2000a) is deceptively simple. There are two states of the world, say 'good' and 'bad'. A forecaster provides a probability forecast $\hat{\pi}$ that the good state will occur (so that the probability of the bad state occurring is $1 - \hat{\pi}$). A decision maker can decide whether or not to take some action on the basis of this forecast, leading to the set of payoffs shown in Table 9.1, where the Y_{ij}'s are the utilities or profit payoffs under each state and action net of any costs of taking the action. The action to ensure that the good state prevails should be undertaken if

$$\frac{\hat{\pi}}{1-\hat{\pi}} > \frac{Y_{22} - Y_{12}}{Y_{11} - Y_{21}}$$

An alternative type of forecast, the 'event forecast', consists of the forecaster announcing the event that is judged to have the highest

Table 9.1 Payoff matrix

Action	State	
	Good	Bad
Yes	Y_{11}	Y_{12}
No	Y_{21}	Y_{22}

probability. Granger and Pesaran showed that using an event forecast will be suboptimal compared to using a predictive distribution but, in general, their analysis clearly illustrated the advantages of using an economic cost function along with a decision-theoretic approach, rather than some statistical evaluation measure.

Granger and Pesaran (2000b) used this type of model to establish links between simple decision problems and measures of forecast accuracy, not just those based on quadratic loss but also the *Kuipers score*, defined as $KS = H - F$, where H is the fraction over time that bad events were correctly forecast to occur and F is the fraction of good events that had been incorrectly forecast to have come out bad, often referred to as the 'false alarm rate'. The Kuipers score was originally designed to be used with meteorological forecasts and has the desirable feature that both random forecasts and forecasts that consistently predict good or bad events will produce an average *KS* score of zero. It does, however, have several undesirable features and Granger and Pesaran suggested that a better evaluation measure would be simply the overall fraction of events (be they good *or* bad) that were correctly forecast.

They also showed that the Kuipers score is related to the market timing test statistic proposed by Pesaran and Timmermann (1992). A stock market timing example was analyzed in some detail, showing that the use of a decision theoretic economic measure, terminal wealth, which incorporates transactions costs, produces a better evaluation of the accuracy of market timing forecasts than the statistical measures, although at the cost of some considerable effort of analysis and computation.

Granger and Machina (2006) explored the links between decision problems and their associated loss functions, asking such questions as whether every statistical loss function can be derived from some well-specified decision problem (there is a close but not unique link between the two) and what the use of squared-error loss reveals or implies about the underlying decision problem (the utility or profit functions of the decision problem are then non-standard and may have some unrealistic properties, such as 'location independence', in which profit shortfalls are the same for all forecast errors of a given size irrespective of the price level).

9.27 An important property of forecasts from a stationary series or a random walk is that optimum forecasts lag the actual observations. This is most easily seen in the case of a random walk, $x_t = x_{t-1} + a_t$, where the optimum least squares forecast of x_{T+1} given the information set x_{T-j}, $j \geq 0$, is $f_T(1) = x_T$ and so the forecast, x_T, lags the actual, x_{T+1}, by one

time unit. Granger and Jeon (2003a, 2003b) introduced the concept of *time-distance*, implicit in the above property, as an alternative way of evaluating forecasts.

If there are a pair of time series, y_t and z_t, $t = 1, 2, \ldots, T$, say, their 'nearness' is typically measured in terms of their *vertical difference* $y_t - z_t$, so providing measures such as the 'mean absolute deviation' $\frac{1}{T}\sum |y_t - z_t|$ and 'mean squared error' $\frac{1}{T}\sum (y_t - z_t)^2$, which, if the series are the actual and forecast values of a particular variable, are standard measures of forecast evaluation. Rather than the vertical difference, Granger and Jeon focused attention on the *horizontal distance*, so defining the notion of time-distance in the following way.

A pair of *adjacent* points y_{t+k} and y_{t+k+1} is said to *include* z_t if either $y_{t+k} \leq z_t \leq y_{t+k+1}$ or $y_{t+k} \geq z_t \geq y_{t+k+1}$. $[S_t^+]$ is then defined to be the smallest value of $k \geq 0$ such that y_{t+k} and y_{t+k+1} include z_t. Thus if y_t and y_{t+1} include z_t then $[S_t^+] = 0$; if $[S_t^+]$ is not zero but y_{t+1} and y_{t+2} include z_t then $[S_t^+] = 1$, and so on. A similar quantity $[S_t^-]$ may be defined for $k \leq 0$ and measures the number of time units one has to move through the negative time periods until z_t is included.

Now suppose that the discrete series y_t is interpolated using straight lines between adjacent points. Fractional time distances, $\bar{S}_t^+$ and $\bar{S}_t^-$, can then be defined on noting that $\bar{S}_t^+ y_{t+k+1} + (1 - \bar{S}_t^+) y_{t+k} = z_t$, so that

$$\bar{S}_t^+ = \begin{cases} \dfrac{z_t - y_{t+k}}{y_{t+k+1} - y_{t+k}} & \text{if } y_{t+k} \neq y_{t+k+1} \\ 0 & \text{if } y_{t+k} = y_{t+k+1} \end{cases}$$

where $k = [S_t^+]$. Clearly, as z_t gets near to y_{t+k}, $\bar{S}_t^+ \to 0$, while as z_t gets near to y_{t+k+1}, $\bar{S}_t^+ \to 1$. Similarly,

$$\bar{S}_t^- = \begin{cases} \dfrac{z_t - y_{t-k}}{y_{t-k-1} - y_{t-k}} & \text{if } y_{t-k} \neq y_{t-k-1} \\ 0 & \text{if } y_{t-k} = y_{t-k-1} \end{cases}$$

where now $k = [S_t^-]$. Both of these quantities are positive fractions and enable the complete time-distances to be defined as

$$S_t^+ = [S_t^+] + \bar{S}_t^+ \qquad S_t^- = [S_t^-] + \bar{S}_t^-$$

for any z_t. If the interpolated series is regarded as the continuous time process $y(t)$ then, since $y_{t+k} + (y_{t+k+1} - y_{t+k})S_t^+ = z_t$, S_t^+ is then the shortest time-distance for which $y(t + S_t^+) = z_t$ for the nearest z_t on the

positive side. Similarly, $y(t - S_t^-) = z_t$ then defines S_t^- for the nearest z_t on the negative side. These time-distances lead to two further measures:

$$S_t = \min(S_t^+, S_t^-)$$

and

$$S_t^{\text{sign}} = \begin{cases} S_t^+ & \text{if } S_t^+ \leq S_t^- \\ -S_t^- & \text{otherwise} \end{cases}$$

S_t may be considered to be the time-distance version of the mean absolute error, while, unlike the other measures, S_t^{sign} can take negative values, which would suggest that z_t leads y_t.

9.28 Granger and Jeon (2003a) investigated the performance of time-distance measures by considering first the simple bivariate model

$$z_t = \alpha z_{t-1} + \beta y_{t-1} + e_t \qquad y_t = y_{t-1} + \varepsilon_t$$

where e_t and ε_t are independent white noises with variances σ_e^2 and σ_ε^2 respectively. Thus

$$(1 - \alpha B)(1 - B)z_t = \beta\varepsilon_{t-1} + (1 - B)e_t = w_t \tag{9.19}$$

for which

$$E(w_t w_{t-j}) = \begin{cases} \beta^2\sigma_\varepsilon^2 + 2\sigma_e^2 & \text{for } j = 0 \\ -\sigma_e^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

Since w_t can be written as $w_t = a_t + \theta a_{t-1}$, where a_t is white noise with variance σ_a^2, it will have autocovariances

$$E(w_t w_{t-j}) = \begin{cases} (1 + \theta^2)\sigma_a^2 & \text{for } j = 0 \\ \theta\sigma_a^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus it must be the case that

$$\begin{aligned} \beta^2\sigma_\varepsilon^2 + 2\sigma_e^2 &= (1 + \theta^2)\sigma_a^2 \\ -\sigma_e^2 &= \theta\sigma_a^2 \end{aligned}$$

from which θ and σ_a^2 can be obtained as functions of β, σ_e^2 and σ_ε^2. Taking the invertible solution $|\theta| < 1$, then (9.19) can be written as

$$(1 - \alpha B)(1 - B)z_t = (1 + \theta B)a_t$$

or

$$(1 - \alpha B)(1 - B)(1 - \theta B + \theta^2 B^2 - \theta^3 B^3 + \cdots)z_t = a_t$$

This can be written as the autoregression

$$z_t = \sum A_j z_{t-j} = a_t$$

where

$$A_1 = 1 + \alpha + \theta \qquad A_j = -(-\theta)^{j-2}(1 + \theta)(\alpha + \theta), \qquad j \geq 1$$

Thus

$$E[z_{t+1}|z_t] = A_1 z_t + A_2 z_{t-1} + \cdots$$

as compared to

$$E[z_{t+1}|z_t, y_t] = \alpha z_t + \beta y_t$$

Through simulation, Granger and Jeon (2003a) showed that, under full information, that is, using $E[z_{t+1}|z_t, y_t]$, the usual MSE measure will provide an accurate estimate of the variance σ_e^2, but with only partial information, so that $E[z_{t+1}|z_t]$ is used, the MSE can be considerably upward biased.

When $\beta = 0$, so that $z_t = \alpha z_{t-1} + e_t$, $\theta = -1$, $A_1 = \alpha$ and $A_j = 0$, $j \geq 1$, S_t and S_t^{sign} both have median values of unity, as the model suggests. For $\beta > 0$, the value of S_t^{sign} is often very small, although values near one can occur for $\beta \leq 0.5$ when σ_e^2 is large. In general, using univariate misspecified models can produce poor forecasts, in terms of time-distance, compared to bivariate models based on full information.

When the set-up is extended to allow for feedback between y_t and z_t, both S_t^- and S_t^+ can be large and the overall simulation results suggest that misspecification can produce substantial leads and lags.

Granger and Jeon (2003a) suggested that the distributions of S_t^+, S_t^- and S_t, but not S_t^{sign}, could be modelled using survival, or duration,

analysis, although in a study of inflation forecasting in Granger and Jeon (2003b) the distributions were obtained by bootstrap simulation.

9.29 In applying this concept to inflation forecasting, Granger and Jeon (2003b) returned to the result in **§9.17** that $\sigma_x^2 = V(f_t(1)) + \sigma_e^2$, where σ_e^2 is the variance of the one-step ahead forecast error $e_{t,1} = x_{t+1} - f_t(1)$, and hence $\sigma_x^2 \geq V(f_t(1))$, so that MSE forecasts have *inhibited* variability in the sense that the variability of the one-step ahead realizations is greater than that of the one-step ahead forecasts. Granger and Jeon suggested that the forecasts be multiplied by a 'timing-varying' scaling factor so as to equate the two variances.

9.30 Granger presented an extended discussion of his views on model evaluation in general in his 1998 Marshall Lectures at Cambridge, published as Granger (1999b), and later amplified these in Granger (2005, 2007), both of which also looked forward to predict how the subject was likely to advance during the first decades of the twenty-first century.

A long-standing interest of Granger's was the forecasting of financial variables and he published numerous articles on this topic, notable ones being Granger (1992a), which was chosen by the editors of the *International Journal of Forecasting* as the best paper published by the journal in the years 1992 and 1993, and Granger and Poon (2003), a comprehensive and influential survey on forecasting volatility.

Model interpretation

9.31 In their discussion of the Box–Jenkins approach (recall **§8.1**), Chatfield and Prothero (1973, page 311) expressed some concern over the difficulty of interpreting ARIMA models, arguing that they 'are generally more difficult to understand conceptually than traditional models involving trend and seasonal terms together with an error term which may or may not be autocorrelated. … (T)he Box–Jenkins procedure does not provide a simple *description* of the data although it does provide straightforward forecasts' (italics in original). They did, though, mention a 1972 conference paper by Granger in which he 'suggested several ways of generating mixed moving average autoregressive models'. This paper eventually became Granger and Morris (1976), whose theme was indeed to suggest a number of ways in which an ARMA model could arise from simpler processes.

9.32 Suppose that X_t and Y_t are independent, zero mean, stationary series following ARMA processes, so that we can write $X_t \sim ARMA(p, m)$ and $Y_t \sim ARMA(q, n)$. If $Z_t = X_t + Y_t$ then $Z_t \sim ARMA(x, y)$, a statement

which Granger and Morris denoted as

$$ARMA(p, m) + ARMA(q, n) = ARMA(x, y)$$

They first provided the following

Lemma

$$MA(m) + MA(n) = MA(y)$$

where

$$y \leq \max(m, n)$$

and then the

Basic Theorem

$$ARMA(p, m) + ARMA(q, n) = ARMA(x, y)$$

where

$$x \leq p + q \qquad y \leq \max(p + n, q + m)$$

This theorem is straightforward to prove. Let the ARMA representations for X_t and Y_t be $a(B)X_t = c(B)\varepsilon_t$ and $b(B)Y_t = d(B)\eta_t$, where ε_t and η_t are independent white noises. Then, since $Z_t = X_t + Y_t$, it follows that

$$a(B)b(B)Z_t = a(B)b(B)X_t + a(B)b(B)Y_t = b(B)c(B)\varepsilon_t + a(B)d(B)\eta_t$$

The right-hand side of this equation is of the form $MA(q + m) + MA(p + n) = MA(y)$, where, from the lemma, $y \leq \max(p + n, q + m)$. The order of the autoregressive polynomial $a(B)b(B)$ cannot be more than $p + q$, so establishing the theorem.

The inequalities in the expressions for x and y arise primarily because $a(B)$ and $b(B)$ may contain common roots: for example, if $X_t \sim AR(1)$ and $Y_t \sim AR(2)$ then, from the basic theorem, $Z_t \sim ARMA(x, y)$ where $x \leq 3$ and $y \leq 2$. However, if $(1 - aB)X_t = \varepsilon_t$ and $(1 - aB)(1 - bB)Y_t = \eta_t$, so that $a(B)$ and $b(B)$ contain a common root, then

$$(1 - aB)(1 - bB)Z_t = (1 - bB)\varepsilon_t + \eta_t \sim ARMA(2, 1)$$

In general, if $a(B)$ and $b(B)$ have k roots in common then the inequalities in the basic theorem become $x = p + q - k$ and $y \leq \max(p + n - k, q + m - k)$. The continuing need for the inequality on y may be seen

from the following example. Suppose that again $(1 - aB)X_t = \varepsilon_t$ but now $(1 + aB)Y_t = \eta_t$, and suppose further that ε_t and η_t have common variance σ^2. Now

$$(1 - aB)(1 + aB)Z_t = (1 + aB)\varepsilon_t + \eta_t = \zeta_t$$

say. It then follows that $E(\zeta_t^2) = 2(1 + a^2)\sigma^2$ and $E(\zeta_t \zeta_{t-j}) = 0$ for all $j > 0$, so that ζ_t is white noise and $Z_t \sim AR(2)$ rather than ARMA(2, 1), as would generally occur when two independent AR(1) processes are added together. Granger and Morris called such a case a 'coincidental situation'.

As Granger and Morris remarked, if X_t and Y_t are both stationary, all cases of common roots and other situations where x and y take less than their maximum values might be considered coincidental. For ARIMA models, however, the presence of unit roots in the autoregressive polynomials will naturally lead to common roots. Thus, if $X_t \sim ARIMA(p, d_1, m)$ and $Y_t \sim ARIMA(q, d_2, n)$ then $Z_t \sim ARIMA(x, d, y)$, where $x \leq p + q$ and $d = \max(d_1, d_2)$. If $d_1 \geq d_2$, $y \leq \max(p + n + d_1 - d_2, q + m)$, while if $d_2 \geq d_1$, $y \leq \max(p + n, q + m + d_2 - d_1)$.

The basic theorem may be generalized to cover the sum of any number of independent series, so that, using an obvious extension of notation,

$$\sum_{i=1}^{N} ARMA(p_i, m_i) = ARMA(x, y)$$

where

$$x \leq \sum_{i=1}^{N} p_i \qquad y \leq \max\,(x - p_i + m, \quad i = 1, \ldots, N)$$

9.33 Granger and Morris went on to consider a number of special cases of the basic theorem, continuing to assume that independent components are being aggregated and ruling out coincidental reductions of parameters.

(i) $AR(p) + \textit{white noise} = ARMA(p, p)$
This corresponds to an AR(p) signal plus a simple white noise observational error.

(ii) $AR(p) + AR(q) = ARMA(p + q, \max(p, q))$, e.g., $AR(1) + AR(1) = ARMA(2, 1)$

This might correspond to a series which is the aggregate of two independent AR series: note that the sum of k AR(1) series will be ARMA($k, k-1$).

(iii) $MA(p) + MA(q) = MA(\max(p,q))$, e.g., $MA(p) + \textit{white noise} = MA(p)$
The addition of a white noise error to an MA process will not alter the form of the process, although the parameter values will change.

(iv) $ARMA(p,m) + \textit{white noise} = ARMA(p,p)$ if $p > m$, but $= ARMA(p,m)$ if $p < m$
The addition of a white noise error may alter the order of an ARMA model but need not do so.

(v) $AR(p) + MA(n) = ARMA(p, p+n)$

Cases (ii) and (v) suggest that a series that is an aggregate of several series, some of which are AR, will very likely be an ARMA process, as will the addition of white noise to either an AR or ARMA process (cases (i) and (iv)). Only if all the individual series are MA processes or white noise will the aggregate be MA. These conclusions can also generally be reached if the assumption of independence is relaxed in a realistic fashion.

9.34 Granger and Morris considered other situations that may occur in practice and which would lead to ARMA models. These included time aggregation, in which a variable obeys a simple model such as an AR(1) when it is recorded at an interval of K units but an ARMA model when it is actually observed at an interval of $M > K$ units, and situations where, for example, a variable obeys the model $X_t - aX_{t-b} = \varepsilon_t$, where b is non-integer. This model may be shown to have the AR(∞) representation $\sum h_j X_{t-j} = \varepsilon_t$, where $h_j = (\sin(j-b)\pi)/(j-b)\pi$, which may be approximated by an ARMA process.

If there is a bivariate autoregressive scheme with feedback in operation, so that

$$a(B)X_t + b(B)Y_t = \varepsilon_t \qquad c(B)X_t + d(B)Y_t = \eta_t \qquad b(0) = c(0) = 0$$

then eliminating Y_t leads to the univariate model

$$(a(B)d(B) - c(B)b(B))X_t = d(B)\varepsilon_t + b(B)\eta_t$$

for X_t, which is ARMA(p, q) with, generally, $p > q$. Granger and Morris thus concluded that many real data situations could give rise to ARMA

models and hence that these were the most likely to be found in practice.

9.35 They then asked whether a given ARMA(p, q) model could have arisen from aggregating some simpler processes: this is referred to as *realizability*. Sometimes this question can be answered immediately: from the bivariate example above, if $p < q$ then a feedback model is not appropriate. In other cases simplifications may not always be possible, as certain conditions on the coefficients of the ARMA model may need to be satisfied for a simpler model to be realizable. As an example, Granger and Morris considered whether an observed ARMA(1, 1) model could equal an AR(1) plus white noise. To investigate this, suppose that $(1 + aB)X_t = \varepsilon_t$ and $Y_t = \eta_t$. Then, from case (i) of **§9.33**, $Z_t = X_t + Y_t \sim ARMA(1, 1)$, given by

$$(1 + aB)Z_t = \varepsilon_t + (1 + aB)\eta_t$$

If the observed ARMA(1, 1) process is

$$(1 + cB)Z_t = (1 + dB)\zeta_t$$

then for these processes to be equivalent we clearly require that $c = a$ but we also need some further realizability conditions to ensure that the right-hand sides of the two processes are equivalent. These are obtained by equating variances and lag one autocovariances,

$$\begin{aligned} d\sigma_\zeta^2 &= a\sigma_\eta^2 \\ (1 + d^2)\sigma_\zeta^2 &= \sigma_\varepsilon^2 + (1 + a^2)\sigma_\eta^2 \end{aligned}$$

and defining the lag one autocorrelation $\rho_1 = d/(1 + d^2)$. It then follows that

$$\rho_1 = \frac{c}{(1 + c^2) + \sigma_\varepsilon^2/\sigma_\eta^2}$$

and so the realizability conditions are

$$\frac{1}{1 + c^2} > \frac{\rho_1}{c} \geq 0$$

By a generalization of this approach, the ARMA(2, 2) model

$$(1 + c_1B + c_2B^2)Z_t = (1 + d_1B + d_2B^2)\zeta_t$$

can only be written as the sum of an AR(2) and white noise if the following realizability conditions are satisfied:

$$\frac{1}{1+c_1^2+c_2^2} > \frac{\rho_2}{c_2} \geq 0 \qquad \frac{\rho_1}{c_1(1+c_2)} = \frac{\rho_2}{c_2}$$

where

$$\rho_1 = \frac{d_1(1+d_2)}{1+d_1^2+d_2^2} \qquad \rho_2 = \frac{d_2}{1+d_1^2+d_2^2}$$

If only the first of these conditions hold then the ARMA(2, 2) can be written as the sum of an ARMA(2, 1) plus white noise. Granger and Morris concluded from these examples that some models are not capable of simplification and that the realizability conditions will typically be rather complicated.

Invertibility and non-linearity

9.36 In (linear) ARMA models the requirement of invertibility ensures that the π–weights of the autoregressive representation form a convergent series: 'in general, the linear process $\pi(B)x_t = a_t$ is invertible if the weights π_j are such that the series $\pi(B)$ converges on, or within the unit circle' (Box and Jenkins, 1970, page 51). Granger and Andersen (1978a) investigated the concept of invertibility when the assumption of linearity has been relaxed by considering a general class of univariate models defined by

$$x_t = f(x_{t-j}, \varepsilon_{t-j}; j = 1, \ldots, P) \tag{9.20}$$

where ε_t is an unobserved input into the system which is assumed to be *pure* white noise, so that ε_t and ε_s are *independent* for $t \neq s$ rather than just being uncorrelated (see, for example, Granger, 1983, and **§9.41** below, for more on the distinctions between white noise, pure white noise and empirical white noise). For models of this type to be useful for forecasting, and provided the function $f(\,)$ actually contains some ε_{t-j}, $j \geq 1$, we must be able to estimate the ε_t sequence from the observed x's. A way of doing this is to assume values for as many initial ε's as are needed, $\hat{\varepsilon}_{-j}$, $j = 0, \ldots, P-1$, say, and then, assuming that x_{-j}, $j = 0, \ldots, P-1$, are also available, estimate ε_1 directly from (9.20) as

$$\hat{\varepsilon}_1 = x_1 - f(x_{1-j}, \hat{\varepsilon}_{1-j}; j = 1, \ldots, P)$$

after which the sequence $\hat{\varepsilon}_2, \hat{\varepsilon}_3, \ldots$ can be obtained recursively. Denoting the 'estimation error' as $e_t = \varepsilon_t - \hat{\varepsilon}_t$, Granger and Andersen defined (9.20) to be invertible if $E[e_t^2] \to 0$ as $t \to \infty$, so that the variance of the error involved in estimating ε_t from a finite number of past and present x's tends to zero as that number tends to infinity, unconditional on the initial values. If the parameter values of (9.20) are not known exactly but have been estimated, the invertibility condition can be replaced by $E[e_t^2] \to c$, where c is some finite constant.

For the simple MA(1) model $x_t = \varepsilon_t + b\varepsilon_{t-1}$, the MMSE forecast of x_{T+1} is $f_T(1) = b\varepsilon_T$, but this needs to be made operational by using an estimate $\hat{\varepsilon}_T$ obtained by setting $\hat{\varepsilon}_0 = 0$ and recursively generating $\hat{\varepsilon}_t = x_t - b\hat{\varepsilon}_{t-1}$. The use of $\hat{\varepsilon}_t$ will yield the error series

$$e_t = x_t - b\varepsilon_{t-1} - (x_{t-1} - b\hat{\varepsilon}_{t-1}) = -be_{t-1}$$

which will have the solution $e_t = (-b)^t e_0$. Clearly $e_t \to 0$, and hence $E[e_t^2] \to 0$, if $|b| < 1$, which is the standard condition that the MA(1) model is invertible (recall **§4.4**). The ARMA(p, 1) model will have exactly the same invertibility condition.

If the moving average parameter is replaced by an estimate $\hat{b}$ then the error series becomes

$$e_t = (b - \hat{b})\varepsilon_{t-1} - \hat{b}e_{t-1}$$

which will have the solution

$$e_t = (-\hat{b})^t e_0 + (b - \hat{b}) \sum_{j=1}^{t} \hat{b}^{j-1} \varepsilon_{t-j}$$

If $|\hat{b}| < 1$ then

$$E[e_t^2] = \hat{b}^{2t} e_0^2 + \frac{(b - \hat{b})^2}{1 - \hat{b}^2} \sigma_\varepsilon^2$$

where $\sigma_\varepsilon^2 = E[\varepsilon_t^2]$, will tend to a finite value given by the second term.

9.37 For the MA(1) process the only non-zero autocorrelation is $\rho_1 = b/(1 + b^2)$, which may be solved to obtain b as a function of ρ_1. It is easily shown, however, that there will be two solutions, b and $1/b$, thus leading to the choice of the invertible solution $|b| < 1$ to ensure a convergent AR(∞) representation. The extension of this idea to MA(q) processes,

or indeed to ARMA(p, q) processes, leads to the general definition used by Box and Jenkins (1970, page 74) that such models are invertible if the roots of the characteristic equation associated with the moving average polynomial $\theta(B)$ all lie outside the unit circle, so that the AR(∞) representation is convergent. It is clear that this definition is only relevant to linear models and Granger and Andersen went on to show that their definition is operational for non-linear as well as linear processes.

As an example of a non-invertible model they consider the non-linear moving average

$$x_t = \varepsilon_t + \alpha\varepsilon_{t-1}^2 \tag{9.21}$$

for which the $\hat{\varepsilon}$'s will be obtained by recursively solving

$$\hat{\varepsilon}_t = x_t - \alpha\hat{\varepsilon}_{t-1}^2$$

The error series is then

$$\begin{aligned} e_t &= x_t - \alpha\varepsilon_{t-1}^2 - (x_t - \alpha\hat{\varepsilon}_{t-1}^2) = -\alpha(\varepsilon_{t-1}^2 - \hat{\varepsilon}_{t-1}^2) \\ &= -\alpha(\varepsilon_{t-1} - \hat{\varepsilon}_{t-1})^2 - 2\alpha\hat{\varepsilon}_{t-1}(\varepsilon_{t-1} - \hat{\varepsilon}_{t-1}) \\ &= -\alpha e_{t-1}^2 - 2\alpha\hat{\varepsilon}_{t-1}e_{t-1} \end{aligned}$$

The solution to this equation has two components, one of which is also the solution to $z_t = -\alpha z_{t-1}^2$, a difference equation that Granger and Andersen showed was unstable and whose only non-trivial solution is explosive. Thus $\hat{\varepsilon}_t$ will diverge from ε_t and (9.21) must therefore be non-invertible. A similar argument shows that the non-linear models analyzed by, for example, Robinson (1977) and which take the form

$$x_t = \varepsilon_t + \alpha\varepsilon_{t-1}\varepsilon_{t-2}$$

are also never invertible.

Bilinear models

9.38 A related non-linear model is the *bilinear form*

$$x_t = \varepsilon_t + \alpha x_{t-1}\varepsilon_{t-1} \tag{9.22}$$

for which

$$\hat{\varepsilon}_t = x_t - \alpha x_{t-1}\hat{\varepsilon}_{t-1}$$

and

$$e_t = -\alpha x_{t-1} e_{t-1}$$

This model has the solution

$$e_t = (-\alpha)^t \left(\prod_{j=1}^{t} x_{j-1} \right) e_0$$

Granger and Andersen (1978b) proved that a sufficient condition for invertibility is that

$$E[\alpha^2 x_t^2] = \frac{\lambda^2(2\lambda^2 + 1)}{1 - \lambda^2} < 1$$

where $\lambda = \alpha\sigma_\varepsilon$, provided $\lambda^2 < 1$. Solving the quadratic (in λ^2) $2\lambda^4 + 2\lambda^2 - 1 = 0$ implies that a sufficient condition for invertibility is that $|\lambda| < 0.605$.

Granger and Andersen (1978b) proved several interesting properties concerning the bilinear model (9.22). They showed that x_t is stationary if $|\lambda| < 1$ and that its ACF is

$$\rho_1 = \frac{\lambda^2(1 - \lambda^2)}{1 + \lambda^2 + \lambda^4} \qquad \rho_k = 0, \quad k > 1$$

ρ_1 increases in value as $|\lambda|$ increases from zero, reaching a maximum of 0.155 at $|\lambda| = 0.605$ and then decreasing. The model is found to be non-invertible for $|\lambda| > 0.707$, but in the interval from approximately 0.6 to 0.7 the issue is less clear-cut. The distribution of x_t is generally skewed, with the third moment increasing with $\lambda > 0$, reaching a maximum at $\lambda = 0.8$ and decreasing thereafter. The fourth moment does not exist for $\lambda > 0.75$ and some higher moments do not exist for non-zero λ.

These very complicated properties, being highly dependent upon λ, contrast sharply with the simple properties of the linear AR(1) and MA(1) models. Given the behavior of the ACF, the bilinear form could very easily be mistaken for an MA(1), with corresponding loss of forecast accuracy: for example, if $\sigma_\varepsilon^2 = 1$ and $\alpha = \lambda = 0.55$, the error variance from the MA(1) model is nearly double that from using the true bilinear form (9.22). Can the bilinear model actually be identified from an observed series or is it completely indistinguishable from an MA(1)? From the results of **§9.20**, if $x_t \sim MA(1)$ then $x_t^2 \sim MA(1)$, but Granger and Andersen showed that if x_t is of the form (9.22) then the ACF of x_t^2 is

the same as that of an ARMA(1, 1) process, so that identification of the correct bilinear form is possible, at least in principle.

9.39 The bilinear form (9.22) is a special case of the general *Bilinear ARMA* (BARMA) model of order (p, q, P, Q)

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=0}^{Q} \sum_{l=1}^{P} \alpha_{kl} \varepsilon_{t-k} x_{t-l},$$

This model is linear in the x and ε variables separately but not in both. If $p = 0$ ($q = 0$) the model is said to be homogenous in the output (input), although Granger and Andersen (1978b, 1978c) concentrated on the simpler model that is homogenous in both variables:

$$x_t = \varepsilon_t + \sum_{k=0}^{Q} \sum_{l=1}^{P} \alpha_{kl} \varepsilon_{t-k} x_{t-l} \tag{9.23}$$

If $\alpha_{kl} = 0$ for all $l < k$ then the model is said to be *superdiagonal* and the multiplicative terms with non-zero coefficients are such that the input ε_{t-k} occurs *after* the output x_{t-l}, so that these terms are independent. Correspondingly, the model is *subdiagonal* if $\alpha_{kl} = 0$ for all $l \geq k$, while the special case when $\alpha_{kl} = 0$ for all $l \neq k$ is said to be *diagonal*: (9.23) is thus BARMA(0, 0, 1, 1) with the additional restriction $\alpha_{01} = 0$. This restriction is a useful one because, if $\alpha_{0l} = 0$ for all l, then (9.23) can always be written as

$$x_t = \varepsilon_t + H(\varepsilon_{t-1}, \varepsilon_{t-2}, \dots)$$

where $H(\ )$ is some non-linear function, and forecasting becomes particularly easy if the model is invertible.

A simple example of a superdiagonal model is

$$x_t = \alpha x_{t-2} \varepsilon_{t-1} + \varepsilon_t \tag{9.24}$$

which will be stationary if $|\lambda| = |\alpha \sigma_\varepsilon| < 1$ and invertible for $|\lambda| < 0.707$. Since $\rho_k = 0$ for all non-zero k, the model thus has a 'white noise' property that suggests that x_t cannot be forecast from its past when clearly it can be, although in a non-linear fashion, as the ACF of x_t^2 from (9.24) can be shown to be that of an ARMA(2,1) process.

9.40 A particularly simple way in which a bilinear form might arise is to suppose that the proportionate change (or rate of return) of a series is

generated by an MA(1) process, so that

$$\frac{x_t - x_{t-1}}{x_{t-1}} = \varepsilon_t + b\varepsilon_{t-1}$$

This can be written as

$$x_t = x_{t-1} + \varepsilon_t x_{t-1} + b\varepsilon_{t-1}x_{t-1}$$

which is a bilinear form, albeit a non-stationary one. If x_t can only be observed with added noise, so that the available data is $y_t = x_t + n_t$, then y_t will follow a non-homogenous bilinear model.

Chaotic processes

9.41 Granger (1983) discussed a variety of models that appear to be white noise but have the potential of being forecast non-linearly. To make his development more precise, he defined three types of white noise: 'standard' white noise, where x_t is *uncorrelated* with its past values; 'pure' white noise, where x_t is *independent* of its past values; and 'empirical' white noise, where the sample covariances tend to zero as the sample size increases, so that the SACF will, in the limit, be the same as white noise.

One particular class of models that created considerable interest during the 1980s was that of 'white chaos'. Three examples of this class are the logistic

$$x_t = 4x_{t-1}(1 - x_{t-1})$$

the triangular

$$x_t = 1 - 2\left|x_{t-1} - \tfrac{1}{2}\right|$$

and the cubic

$$x_t = x_{t-1} + 4x_{t-1}(x_{t-1}^2 - 1)$$

All three processes produce sequences (at least for most values) that are empirical white noise and lie in the region 0 to 1, even though they are deterministic and x_t is perfectly forecastable from x_{t-1}. Each process has a pair of non-stable equilibrium points (values for which $x_t = x_{t-1}$),

these being 0 and $\frac{3}{4}$ for the logistic, 0 and $\frac{2}{3}$ for the triangular and 0 and 1 for the cubic, with the triangular model producing a sequence that appears to be rectangularly distributed over 0 to 1.

A further example is the *tent map*

$$\begin{aligned} x_t &= a^{-1}x_{t-1} && 0 \le x_{t-1} < a \\ &= (1-a)^{-1}(1-x_{t-1}) && a \le x_{t-1} \le 1 \end{aligned}$$

for which a sequence generated from the map will display an SACF consistent with an AR(1) process. When the constant a is close to 0.5, the autocorrelations will be close to empirical white noise. A final example is $x_t = x_{t-1}^2 - 2$ with starting value $-2 < x_0 < 2$. This has equilibrium points at 2 and -1, although unless $x_0 = 2$ the first equilibrium point is never reached. For this process both x_t and x_t^2 display empirical white noise even though a regression of x_t on x_{t-1}^2 will give perfect forecasts.

9.42 Granger was rather skeptical about the practical usefulness of chaotic models and, in particular, as to whether there was any evidence of such processes actually occurring in reality rather than just as computer simulations or in experiments in physics laboratories. Three quotes from Granger's comments on the Chatterjee and Yilmaz (1992) and Berlinger (1992) reviews of the links between chaos and statistics make this skepticism abundantly clear.

> There is a great deal to be admired in the extensive work on chaos that has appeared in recent years, including some startling but simple theorems, and also the best art work produced by mathematics. However, in my opinion, it is often surrounded by an unnecessary amount of hype, considerable zeal and possibly some illogical arguments and confusion. Granger (1992b, page 102)

> Chatterjee and Yilmaz take the position that [chaos] is ubiquitous, finding examples in 'such diverse fields as physiology, geology, ..., economics' and 'theoretical models of population biology'. There are also theoretical models in economics that produce chaos, but that does not imply that it occurs in practice. I would prefer to suggest that there is *no* evidence of chaos outside of laboratories. My reason is that there exists no statistical test, that I know of, that has chaos as its null hypothesis. There are plenty of tests that have as a null H_0 : iid (or linear) and are designed to have power against chaos. However, as is well known by statisticians, if one rejects the null a specific alternative hypothesis cannot be accepted. If a null of linearity or iid

> is rejected, one cannot accept (white) chaos, as nonlinear stochastic models are also appropriate. … Until a property *P* can be found that holds *only* for chaos and not for stochastic series, and a test is based on *P* with chaos as the null, can there be a suggestion that chaos is found in the real world. (ibid., page 104: italics in original)

> I think that scientists working in this area are doing a disservice to this important area of research by overselling its relevance, by trying to equate it with randomness and by using concepts (such as probability) that are unnecessary and only lead to confusion. The techniques being developed for analysis of chaotic processes … are potentially powerful and useful when applied to truly stochastic, real-world series. There is a need for statistical methods to investigate the properties of these techniques … and this, in my opinion, is the natural link between chaos and statistics. (ibid., page 104)

Taking up the theme of the latter quotes, Granger focused his own attention on testing both for chaos and also for more general forms of non-linearity: see Lui, Granger and Heller (1992), Lee, White and Granger (1993) and Granger, Teräsvirta and Lin (1993). The general conclusion from these exercises confirmed Granger in his view 'that probabilistic methods are … the most appropriate technique for analyzing economic time-series data. We suspect that this conclusion also applies to much data where chaos has been "found" in the behavioral sciences, biology, health sciences and education' (Lui, Granger and Heller, 1992, page S39).

9.43 Granger continued to investigate models of stochastic non-linearity, whether in terms of forecasting ability (Granger and Lin, 1994a), theoretical issues (Granger and Lin, 1994b; Granger, Inoue and Morin, 1997; Granger, 2008b), or modelling methodology (Granger, 1991, 1993), as well as publishing various applied contributions. He also drew together the non-linear models extant in the early 1990s in the monograph *Modelling Nonlinear Economic Relationships* (Granger and Teräsvirta, 1993), an updated version of which was published posthumously as Teräsvirta, Tjostheim and Granger (2011).

10

Granger: Long Memory, Fractional Differencing, Spurious Regressions and Co-integration

Long memory and fractional differencing

10.1 Granger's research on bilinear models had left him dissatisfied as he did not feel that they, or indeed many other forms of non-linear models, were of much practical use. He was also struck by the limitations imposed by ARIMA(p, d, q) models on the behaviour of the ACF, which declines either geometrically, when $d = 0$, or linearly when $d = 1$. In Granger (1979) he suggested the possibility of models displaying *long memory*, taking his cue from the water resources literature (as surveyed by Lawrence and Kottegoda, 1977) and, in particular, the models proposed by Mandelbrot and Van Ness (1968). Rather than having a spectrum taking the form ω^{-2} for small frequencies ω, as would be the case for a process that required first differencing for it to be rendered stationary (see **§10.2** below), such models would have spectra proportional to $\omega^{-\alpha}$ for $0 < \alpha < 2$. If long memory models should prove useful then 'ordinary integer differencing is inappropriate, yet the series would have an infinite variance and its correlogram would suggest differencing according to the Box–Jenkins rules. If such series arise in practice, they could be of considerable importance and "fractional differencing" should become a standard component of analysis' (Granger, 1979, page 251). An example given by Granger of a long-memory process was the infinite moving average

$$x_t = \sum_{j=0}^{\infty} b^{\log j} a_{t-j}$$

where a_t is white noise, although he presented no analysis of it.

10.2 The idea of long memory was formally developed by Granger in two papers published in the subsequent year: Granger and Joyeux (1980) and Granger (1980b).[1] The starting point was to consider $y_t = \Delta^d x_t$, where, adapting the terminology of **§9.34**, $y_t \sim I(0)$ and $x_t \sim I(d)$: y_t and x_t are said to be integrated processes of order zero and d respectively. If y_t has spectrum $f_y(\omega)$ then x_t, although it does not strictly possess a spectrum, can be thought to have the 'pseudo-spectrum'

$$f_x(\omega) = |1 - z|^{-2d} f_y(\omega) = 2^{-d}(1 - \cos\omega)^{-d} f_y(\omega) \qquad z = e^{-i\omega} \quad \omega \neq 0$$

If y_t has the ARMA representation $\phi(B)y_t = \theta(B)a_t$ then $\lim_{\omega\to 0} f_y(\omega) = c > 0$ and it then follows that, for ω small, $f_x(\omega) = c\omega^{-2d}$, where

$$c = \frac{\sigma_a^2}{2\pi}\left(\frac{\theta(1)}{\phi(1)}\right)^2$$

More generally, suppose that x_t has the pseudo-spectrum

$$f_x(\omega) = \alpha(1 - \cos\omega)^{-d} \qquad \alpha > 0 \quad \omega \neq 0 \tag{10.1}$$

When differenced d times x_t will produce white noise but, for $-1 < d < \frac{1}{2}$, $d \neq 0$, the ACF of x_t will be of the form

$$\rho_k = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(k+1-d)}$$

where $\Gamma(n) = (n-1)!$ is the gamma function. Using the standard approximation that, for k large, $\Gamma(k+a)/\Gamma(k+b) \approx k^{a-b}$, these autocorrelations may be approximated as

$$\rho_k \approx \frac{\Gamma(1-d)}{\Gamma(d)} k^{2d-1} = A(d)k^{2d-1} \tag{10.2}$$

This may be contrasted with the autocorrelations from a stationary ARMA model which, for large k, are approximately of the form $A\theta^k$ with $|\theta| < 1$. These tend to zero at an exponential rate and thus decay quicker than the hyperbolic decline of the ρ_k given in (10.2), which thus display a 'long memory' property.

10.3 The AR(∞) representation $\pi(B)x_t = a_t$ has π-weights given by

$$\pi_k = \frac{\Gamma(k-d)}{\Gamma(-d)\Gamma(k+1)} = \frac{(k-d-1)!}{(-d-1)!k!} \approx Ak^{-(1+d)}$$

while the MA(∞) representation $x_t = \psi(B)a_t$ has ψ-weights given by

$$\psi_k = \frac{\Gamma(k+d)}{\Gamma(d)\Gamma(k+1)} = \frac{(k+d-1)!}{(d-1)!k!} \approx Ck^{d-1}$$

Thus the ψ_k and $|\pi_k|$ also tend to zero hyperbolically and hence decay more slowly than the exponential decline associated with a stationary process, so that no ARMA(p, q) process with finite p and q would provide an adequate approximation for large k. If d is positive then the π-weights are negative and the ψ-weights positive, with these signs being reversed for negative d. The partial correlations are given by $\phi_{kk} = d/(k-d)$ and hence decay as k^{-1} independently of d.

Writing the MA(∞) representation as

$$x_t = C\sum_{k=1}^{\infty} k^{d-1}a_{t-k} + a_t$$

shows that x_t has variance

$$V(x_t) = C^2\sigma_a^2\left(1 + \sum_{k=1}^{\infty} k^{2(d-1)}\right)$$

Since $\sum_{k=1}^{\infty} k^{-s}$ converges for $s > 1$, the variance of x_t will be finite provided $d < \frac{1}{2}$ but will be infinite for $d \geq \frac{1}{2}$.

10.4 The model represented by the spectrum in (10.1) is

$$\Delta^d x_t = \left(1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots\right)x_t = a_t \quad (10.3)$$

and may be regarded as *fractional white noise*. The ARIMA(p, d, q) process with p and q integer but d real, $\phi(B)\Delta^d x_t = \theta(B)a_t$, will exhibit long-run behaviour that is similar to the ARIMA(0, d, 0) process (10.3) because the ϕ and θ parameters model the correlation structure at low lags and will thus have negligible influence on very distant observations, whose long-run correlation structure is modelled by d. Such processes have been given the acronym ARFIMA, with 'FI' standing for *fractional integration* (see Baillie, 1996, for a later survey of ARFIMA processes). In general, x_t will be stationary if $d < \frac{1}{2}$ and invertible if $d > -\frac{1}{2}$.

10.5 How might fractionally integrated models arise? Granger suggested that aggregation was a possible mechanism. Suppose there are

N independent AR(1) processes

$$x_{j,t} = \phi_j x_{j,t-1} + a_{j,t} \qquad j = 1, 2, \ldots, N \tag{10.4}$$

From **§9.35** the aggregate $\bar{x}_t = x_{1,t} + \cdots + x_{N,t}$ will then be ARMA(N, $N-1$) unless there is some cancellation of roots in the autoregressive and moving average polynomials. Since many macroeconomic variables, for example, are aggregates of a large number of micro-variables, this would suggest that ARMA models fitted to such aggregates would have to be of high order, which is not found to be the case in practice. Consider, then, the spectrum of $\bar{x}_t$,

$$\bar{f}(\omega) = \sum_{j=1}^{N} f_j(\omega)$$

where $f_j(\omega)$ is the spectrum of $x_{j,t}$ which, from (10.4), is

$$f_j(\omega) = \frac{\sigma_j^2}{2\pi} \frac{1}{|1 - \phi_j z|^2}$$

with σ_j^2 being the variance of $a_{j,t}$. If the ϕ_j are assumed to be random variables drawn from a population with distribution function $F(\phi)$ and the σ_j^2 are drawn from a population independent of the ϕ_j then, approximately,

$$\bar{f}(\omega) \approx \frac{N}{2\pi} E[\sigma_j^2] \cdot \int \frac{1}{|1 - \phi z|^2} dF(\phi)$$

If $F(\phi)$ is the distribution function of a discrete random variable on $[-1, 1]$, so that ϕ can take just m, say, values in this range then $\bar{f}(\omega)$ will be the spectrum of an ARMA($m, m-1$) process. However, if ϕ is continuous then $\bar{f}(\omega)$ will not correspond to any finite order ARMA process. Suppose that ϕ is beta distributed over the range $(0, 1)$:

$$\begin{aligned} dF(\phi) &= \frac{2}{B(p,q)} \phi^{2p-1} (1 - \phi^2)^{q-1} d\phi, \qquad 0 \le \phi \le 1 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{elsewhere} \end{aligned}$$

where $p > 0$, $q > 0$ and $B(p, q)$ is the beta function. Granger showed that, provided $q < 1$, the kth autocovariance of $\bar{x}_t$ could be written as

$$\bar{\gamma}_k = \frac{\Gamma(q-1)}{B(p,q)} \frac{\Gamma(p + k/2)}{\Gamma(p + k/2 + q - 1)} \approx D k^{1-q}$$

for large k. On comparing this with (10.2) it follows that $\bar{x}_t \sim I(1 - q/2)$ with $1 - q/2 > \frac{1}{2}$, so that $\bar{x}_t$ is fractionally integrated with infinite variance. Granger relaxed some of the conditions required for this approximation and showed that it was possible for $\bar{x}_t$ to be stationary (i.e., $I(0)$) if the ϕ_j's were constrained to be less than some quantity which is itself strictly less than one.

If the component series are now generated as $x_{j,t} = \phi_j x_{j,t-1} + y_{j,t}$, where the $y_{j,t}$ are independent of each other but serially correlated, then $\bar{x}_t \sim I(d_y + 1 - q)$, where d_y is the order of integration of $\bar{y}_t = y_{1,t} + \cdots + y_{N,t}$. Granger then considered more general models in which dependence and feedback were allowed and found that aggregation was often likely to lead to aggregate series that were fractionally integrated.

10.6 Granger thought that the practical usefulness of long-memory models lay in long-run forecasting, where the shape of the spectrum at low frequencies becomes paramount. It was also clear that a realistic method for estimating the fractional differencing parameter d needed to be found. A rudimentary method was proposed in Granger and Joyeux (1980) but much was left undeveloped, a gap that was quickly filled over the next few years by various researchers: for example, Geweke and Porter-Hudak (1983) proposed a method based on log-periodogram regression, Sowell (1992) introduced ML methods of estimating d jointly with the other parameters in an ARFIMA model, and a huge literature on semi-parametrically estimating d was subsequently developed, this being surveyed by Velasco (2006).

10.7 Granger returned to long memory processes in the mid-1990s, publishing a series of papers, primarily co-authored with Zhuanxin Ding, investigating a variety of models using a very long run of US stock prices. This series was the daily price of the S&P 500 index from January 1928 to August 1991, totaling over 17,000 observations. Figure 10.1 shows the daily price (p_t), daily (compounded) return, calculated as the log-differences $r_t = \ln p_t - \ln p_{t-1}$, and the absolute daily return ($|r_t|$). As Ding, Granger and Engle (1993) pointed out, there is an upward trend for p_t while r_t is fairly stable around a mean of 0.00018. Large values of $|r_t|$ are more likely to be followed by large values than small values and vice versa, so that market volatility appears to be changing over time. Volatility was particularly high during the Great Crash of 1929 and the subsequent Great Depression of the early 1930s and there was also a large drop in prices and a short burst of volatility at the Crash of October 1987, but otherwise the market has been relatively stable.

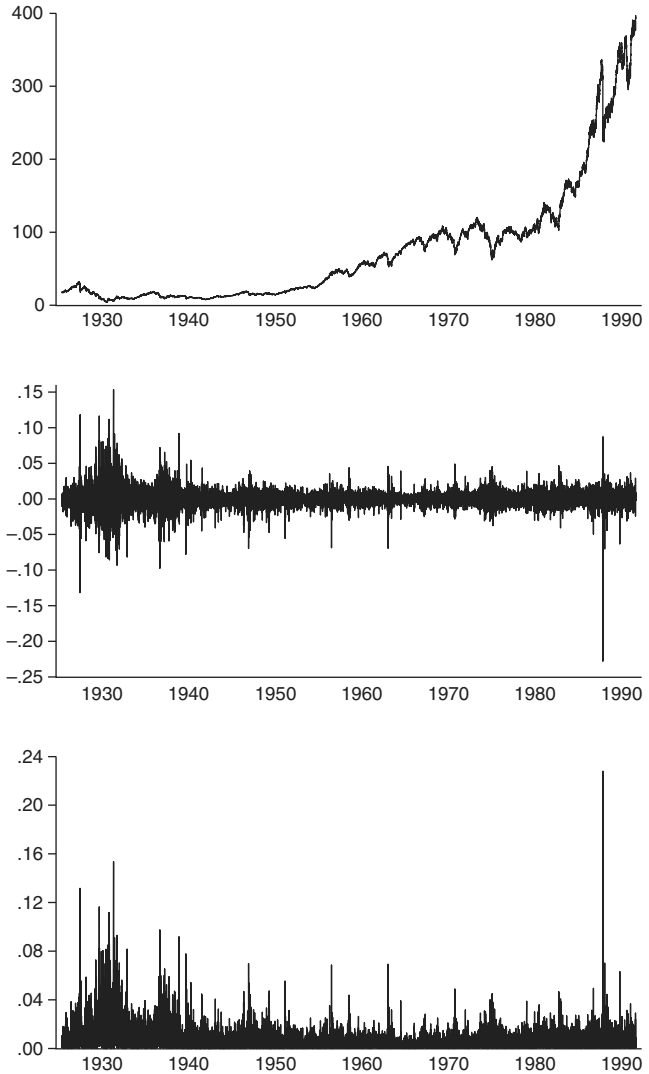


Figure 10.1 S&P 500 daily price index (top); daily returns (middle); absolute daily returns (bottom)

Figure 10.2 shows the SACFs for r_t, r_t^2 and $|r_t|$ for the first 200 lags along with $\pm 1.96/\sqrt{T} = \pm 0.015$ bounds, which correspond to a 95% confidence interval for the estimated sample autocorrelations if r_t is independently and identically distributed (i.i.d.). A considerable number of sample autocorrelations lie outside these bounds, particularly noticeable

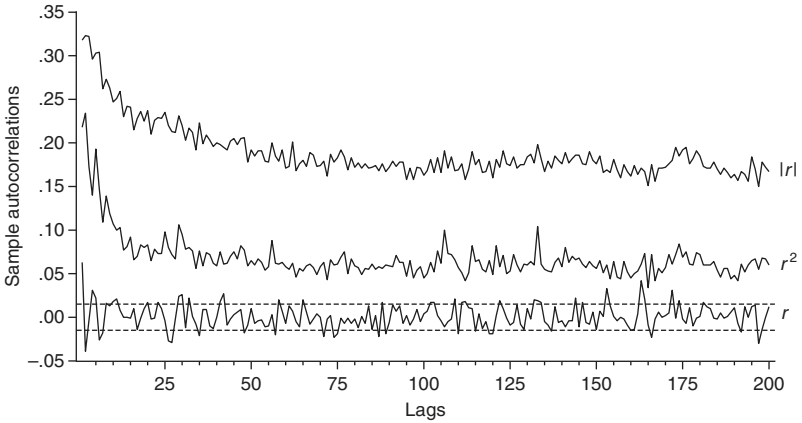


Figure 10.2 SACFs of daily returns (r), squared daily returns (r^2) and absolute daily returns ($|r|$) with 95% confidence bands under the i.i.d. hypothesis

ones being the first, estimated to be 0.063, and the second, −0.039, so that returns cannot be considered to be a realization from an i.i.d. process.

If r_t was an i.i.d. process then any transformation of r_t, such as r_t^2 and $|r_t|$, would also be i.i.d. These transformations would then also have sample autocorrelations with standard errors $1/\sqrt{T}$ under the i.i.d null as long as r_t^2 has finite variance and $|r_t|$ finite kurtosis. From Figure 10.2 it is seen that all sample autocorrelations for these transformations fall well outside the i.i.d. 95% confidence bands and, moreover, that they are all positive, with the sample autocorrelations for absolute returns always being greater than those for squared returns for every one of the first 200 lags: the daily S&P 500 return is clearly not an i.i.d. process.

Figures 10.3 and 10.4 show the SACFs of $|r_t|^d$ for various values of d. These power transformations of absolute returns have significant positive autocorrelations at least up to lag 200 for $d \geq 0.25$. The autocorrelations decrease relatively quickly during the first month or so, and then decrease very slowly. The largest autocorrelations are found for $d = 1$ and they decline almost monotonically as d moves away from 1 in either direction. This phenomenon, whereby the autocorrelations of power transformed absolute stock returns are greatest for $d = 1$ and exhibit the slow decline of a long-memory process, was termed the 'Taylor effect' by Ding and Granger (1995), as it was first reported in Taylor (1986).

Figure 10.5 shows the sample autocorrelations at lags 1, 2, 5 and 10 as a function of d. These autocorrelations are seen to be smooth functions of d, having a maximum in the region of $d = 1$ and a saddle point

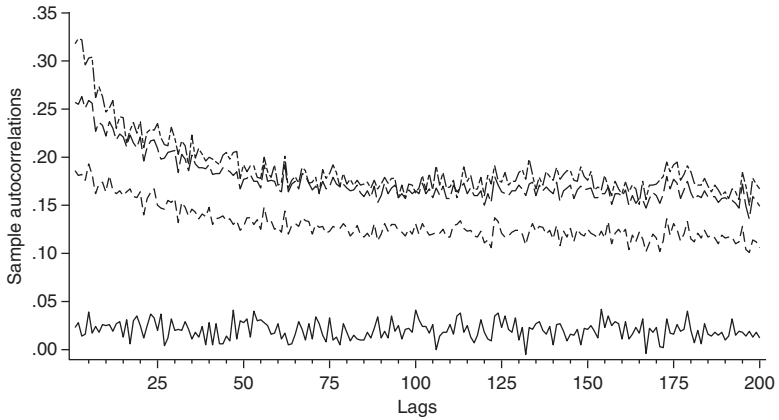


Figure 10.3 SACFs of $|r|^d$ for $d = 1, 0.5, 0.25, 0.125$ from high to low

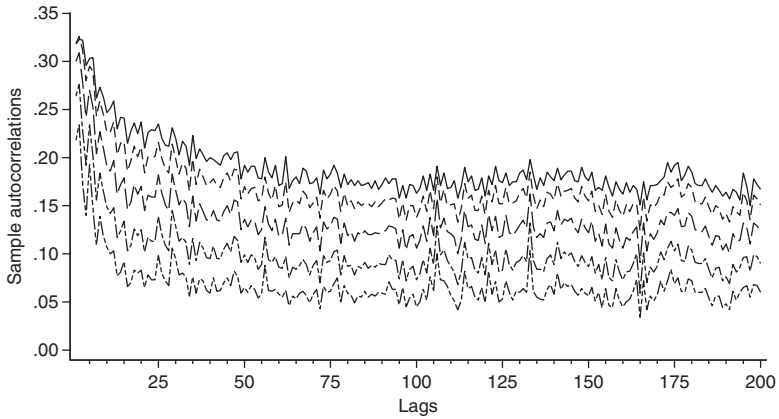


Figure 10.4 SACFs of $|r|^d$ for $d = 1, 1.25, 1.50, 1.75, 2$ from low to high

between $d = 2$ and 3, although they remain positive for all lags and values of d.

Shown below is the lag k^* at which the first negative autocorrelation appears for $|r_t|^d$ for various values of d. In most cases $|r_t|^d$ has positive autocorrelations over more than 2500 lags, i.e., over ten years!

d	0.125	0.25	0.5	0.75	1	1.25	1.5	1.75	2	3
k^*	2028	2534	2704	2705	2705	2705	2705	2685	2598	520

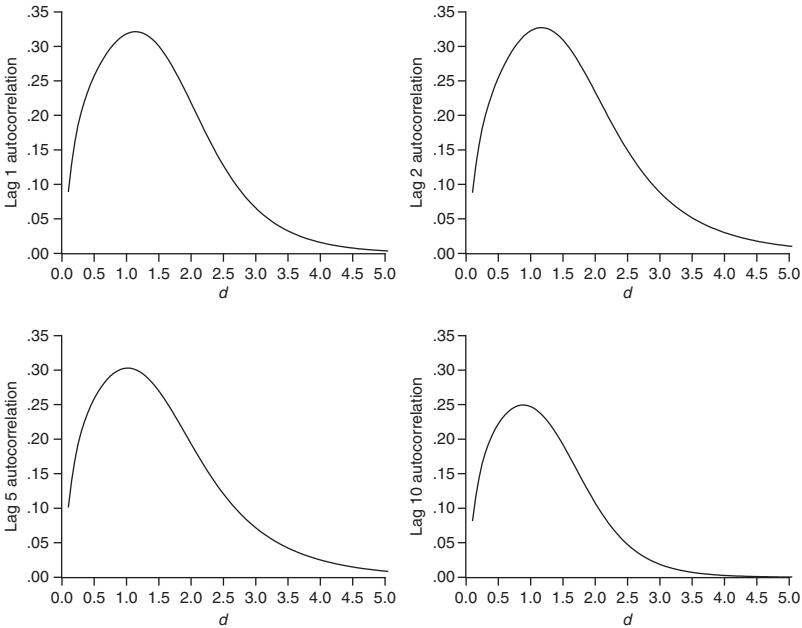


Figure 10.5 Autocorrelation of $|r|^d$ at lags 1, 2, 5 and 10

The sample autocorrelations for $|r_t|$ are shown in Figure 10.6 for the first 2500 lags. Not only are they all positive but they all lie outside the i.i.d. 95% confidence interval. Ding et al. (1993) fitted several models to these autocorrelations. The first assumes $\rho_k = \alpha\beta^k$, so that the autocorrelations decrease exponentially with k, similar to the ACF of an ARMA process. The second allows the autocorrelations to decline in a manner consistent with a fractionally integrated process, so that, from **§10.2**,

$$\begin{aligned}
\rho_k &= \frac{\Gamma(1-\beta)}{\Gamma(\beta)} \frac{\Gamma(k+\beta)}{\Gamma(k+1-\beta)} \\
&= \frac{\Gamma(1-\beta)}{\Gamma(\beta)} \frac{(k+\beta-1)\cdots\beta}{(k-\beta)\cdots(1-\beta)} \frac{\Gamma(\beta)}{\Gamma(1-\beta)} \\
&= \frac{(k+\beta-1)\cdots\beta}{(k-\beta)\cdots(1-\beta)} \\
&= \rho_{k-1} \frac{(k+\beta-1)}{(k-\beta)}
\end{aligned}$$

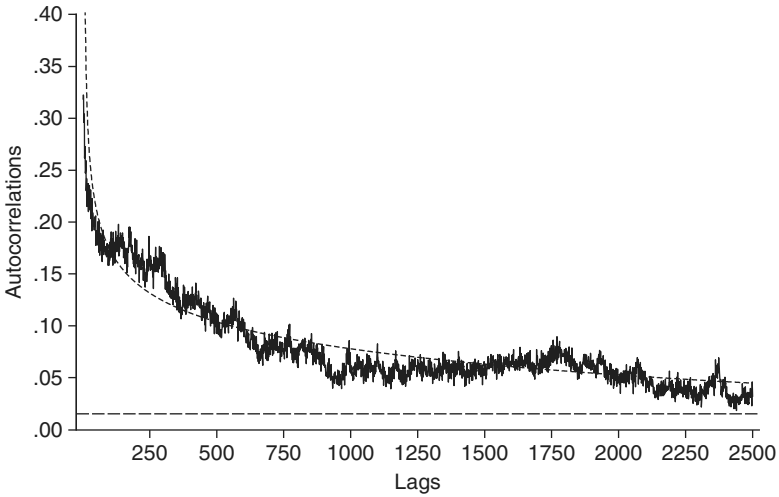


Figure 10.6 Sample and theoretical autocorrelations for absolute daily returns for 2500 lags

The third, and preferred, model is a combination of these with the ACF being specified as

$$\rho_k = \frac{\alpha \rho_{k-1} \beta_2^k}{k^{\beta_3}} \tag{10.5}$$

The parameters can easily be estimated by least squares applied to a log transformation of (10.5), leading to

$$\rho_k = 0.893 \rho_{k-1}^{0.784} (0.999955)^k / k^{0.057}$$

These theoretical autocorrelations are also shown in Figure 10.6 and are seen to fit the sample autocorrelations quite well except at very low lags.

Ding et al. considered both temporal aggregation of returns and subperiods of the data. Temporal aggregation did not alter the long memory property but long memory was found to be much more prevalent in the pre-1946 period than in the postwar period, in the sense that the later period was characterized by autocorrelations that were smaller and decreased quicker.

10.8 Ding and Granger (1995, 1996) proposed a formal model for the long-memory property of absolute returns. They began by assuming that

the return is given by the *product process* $r_t = \sigma_t e_t$, where e_t is i.i.d. with zero mean and unit variance. The model for the conditional standard deviation σ_t is, for $0 < d < \frac{1}{2}$,

$$\sigma_t = (1 - \Delta^d)\frac{|r_t|}{E|e_t|} = \sum_{j=1}^{\infty} \frac{d}{\Gamma(1-d)} \frac{\Gamma(j-d)}{\Gamma(j+1)} \frac{|r_{t-j}|}{E|e_t|} \tag{10.6}$$

For this model $\rho_k = 0$ for all $k > 0$ but the correlation between $|r_t|$ and $|r_{t-k}|$ is

$$\rho_k(|r|) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(k+1-d)}$$

which, from **§10.2**, is the same as the autocorrelation from a fractionally integrated series. This can also be seen by rewriting (10.6) as

$$\Delta^d |r_t| = \sigma_t(|e_t| - E|e_t|) = \varepsilon_t$$

where $\varepsilon_t = \sigma_t(|e_t - E|e_t||)$ is a mean zero short memory process with conditional heteroscedasticity.

Using the Geweke and Porter-Hudak (1983) log periodogram regression approach (known as the GPH estimator), the fractional differencing parameter is estimated to be $\hat{d} = 0.474$, which proves to be too large an estimate as the fitted autocorrelations are much larger than the sample autocorrelations (see Ding and Granger, 1996, Figure 4). A better fit is obtained by directly minimizing the squared differences between $\rho_k(|r|)$ and the sample autocorrelations of $|r_t|$, which produces $\hat{d} = 0.358$. Even this value of d does not fit the first 20 or so autocorrelations particularly well, so that the model (10.5), which includes a short-run component, performs better.

Further results on the absolute returns of a wide variety of speculative assets were provided in Granger, Spear and Ding (2000), who also extended (10.6) to

$$\sigma_t = \lambda\sigma_{t-1} + (1-\lambda)\sum_{j=1}^{\infty} \frac{d\Gamma(p+d)\Gamma(p+j-1)}{\Gamma(p)\Gamma(p+d+j)} \frac{|r_{t-j}|}{E|e_t|}$$

where $0 \leq \lambda \leq 1$. The specification (10.6) is recovered if $p + d = 1$ and $\lambda = 0$. This specification also allows the mean and standard deviation of $|r_t|$ to be equal, as was found for a number of series by Granger et al.,

particularly after some outlier reduction, when the marginal distribution of $|r_t|$ could be taken to be exponential (see also Granger and Jeon, 2002). Granger and Sin (2000) considered various aspects of forecasting absolute returns.

An important finding in these studies (confirmed by Mills, 1997, using other financial time series) was that d appeared to be time-varying, with estimates of the parameter altering markedly across subperiods of the sample: for example, splitting the S&P 500 sample into ten subperiods of approximately seven years produced estimates of d ranging from 0.156 for the period 1954–60 to 0.714 for 1974 to 1979. Ding and Granger (1996) offered some possible models for explaining this time-varying long memory.

10.9 Ding and Granger (1996) also discussed *generalized integrated* (GI) processes, a particular example of which is the GI(d, q) model

$$\Delta^d x_t = (\ln \Delta^{-1})^{-q} \varepsilon_t$$

where ε_t is white noise. Provided $d > 0$ or if $d = 0$ but $q > 0$, the process will have long memory. The GI(d, q) series has the spectrum

$$|1 - e^{i\omega}|^{-2d} |\ln(1 - e^{iw})^{-1}|^{-2q} \sigma_\varepsilon^2$$

which, for small ω, will be proportional to $\omega^{-2d}(\ln \omega)^{2q}$. The $d = 0$, $q > 0$ case, denoted GI(0^+), will have autocorrelations that, for large lags, are close to being proportional to the inverse of the lag.

Non-linearity and long memory

10.10 Using the Hermite polynomial approach of Granger and Newbold (1976) (cf. **§9.22**), Ermini and Granger (1993) considered non-linear transformations of integrated processes. Polynomial transformations of order m of a random walk with drift will contain polynomial time trends of order m and drifts (defined as the unconditional mean of the first differences of the transformed series) of order $m - 1$. If the original series has no drift the transformed series will exhibit a polynomial time trend of order $[m/2]$ and a drift of order $[m/2] - 1$, which will therefore be a constant for $m \leq 3$. The autocorrelations of the transformed process will approximate the autocorrelations of an $I(1)$ process irrespective of the order m as long as the sample size is large but, if the sample is small and m is large, the autocorrelations may appear to be those of an $I(0)$ process. The exponential transformation of a random walk will,

in general, contain exponential trends in both mean and variance even if the random walk does not contain a drift. It will also have autocorrelations that decline in a similar fashion to those of an AR(1) process, as will its changes. Exponential transformations of more general $I(1)$ processes will also contain exponential trends and stationary geometric decays of their autocorrelations. Periodic transformations of $I(1)$ processes behave, in large samples, as stationary, zero-mean, homoscedastic AR(1) processes.

10.11 Dittman and Granger (2002) subsequently extended these results to fractionally integrated processes. Their findings may be summarized thus.

(i) Non-linear transformations of $I(d)$ processes for which $0 < d < \frac{1}{2}$ remain fractionally integrated processes but with a reduced value of d, although the larger d is, the smaller will be this reduction. As a special case, the square of a Gaussian $I(d)$ process will be $I(2d - 0.5)$ for $\frac{1}{4} < d < \frac{1}{2}$ but $I(0)$ for $0 < d \leq \frac{1}{4}$.

(ii) Processes for which $-1 < d < 0$ are said to be *anti-persistent*. Although, in theory, non-linear transformations of such processes become short-memory, 'odd' transformations, such as the cubic, might still appear anti-persistent in finite samples.

(iii) For $d > \frac{1}{2}$, power transformations will contain trends in mean and variance: for a power of m the trend will be of order $t^{m(d-0.5)}$ while the variance will be of order $t^{m(2d-1)}$, which will therefore dominate the trend in mean for all m.

(iv) $I(d)$ processes for which $d > \frac{1}{2}$ will have time-dependent autocorrelations which individually converge to unity, a property that is maintained for a Gaussian process under any power transformation. Thus a power transformation of such a process will be $I(d')$ for $d' > \frac{1}{2}$ and so will still exhibit non-stationary long memory. In fact, the square of a non-stationary $I(d)$ process will also be $I(d)$ as long as $\frac{1}{2} < d < 1$, in contrast to (i), where squaring a stationary long-memory series *reduces* the amount of long memory (the $d = 1$ case is omitted as Granger, 1995, earlier showed that the square of a random walk is a random walk with drift having a variance that is quadratic in t).

(v) Anti-persistence disappears under cosine or exponential transformations, as these are even transformations, but is partly preserved under odd transformations such as the sine and logistic. All the transcendental transformations tend to reduce the extent of long

memory in the stationary case, $0 < d < \frac{1}{2}$, but their effect in the non-stationary case is varied. As d increases from 0.5 the long memory in the sine and cosine transformations decreases to such an extent that, for $d = 1$, the sine and cosine of a random walk become AR(1) processes with heteroscedastic errors. This behaviour is also shown by the exponential transformation, where the long-memory parameter of the transformed series decreases as d approaches unity. Indeed, the exponential transformation of a random walk behaves like an AR(1) process but with an exponentially increasing variance. The logistic function exactly retains the long memory of stationary long-memory processes while, for non-stationary long-memory processes, the non-stationary long memory is reduced but still retained. Since the logistic transformation is bounded, this implies that non-stationary long-memory processes can therefore also be bounded. The logistic transformation of a random walk is still a random walk but with a constant variance (Granger, 1995).

10.12 Can long-memory models be mistaken for non-linear models and vice versa? Granger and Teräsvirta (1999) considered the simple non-linear model

$$x_t = \text{sgn}(x_{t-1}) + \varepsilon_t$$

where ε_t is a zero mean Gaussian i.i.d. process and the 'sign function' is

$$\begin{aligned} \text{sgn}(x) &= 1 && \text{if } x > 0 \\ &= 0 && \text{if } x = 0 \\ &= -1 && \text{if } x < 0 \end{aligned}$$

If

$$p = P(\varepsilon_t < -1) = P(\varepsilon_t > 1)$$

then, as shown by Rydén, Teräsvirta and Åsbrik (1998), the theoretical autocorrelations of x_t are given by $\rho_k = (1 - 2p)^k$ and so will decline exponentially as in a linear, stationary AR(1) process.

If p is small then a plot of x_t will show that it is essentially a regime-switching process, taking the value 1 plus a small random error for a considerable period of time until a low enough value of ε_t occurs, with probability p, in which case x_t switches to -1 plus errors until another switch occurs (see Figure 10.7, where 2000 values of x_t are simulated with $p = 0.01$). The sample autocorrelations from generated values of the

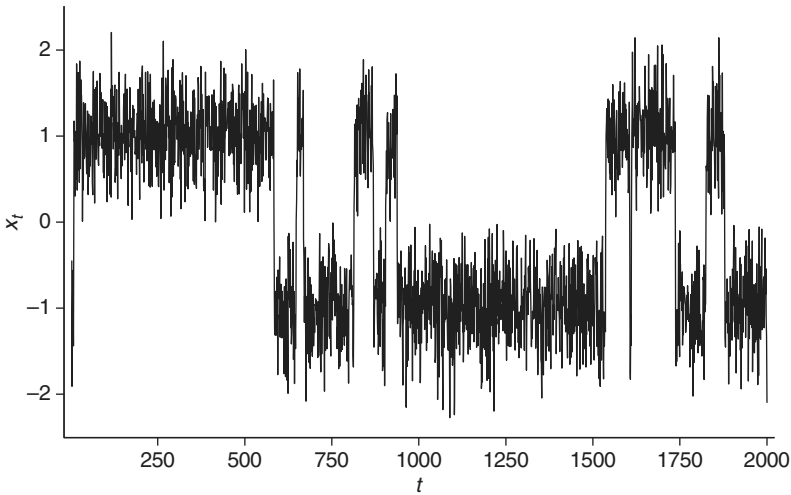


Figure 10.7 Simulated $x_t = \text{sgn}(x_{t-1}) + \varepsilon_t$ with $p = 0.01$

series are found to decline at a rate that is slower than exponential but consistent with a fractionally integrated process. For the series shown in Figure 10.7 the theoretical autocorrelations should decline as $\rho_k = 0.98^k$, thus reaching 0.603 by $k = 25$ and 0.133 by $k = 100$. The lag one sample autocorrelation, however, is only 0.850, but the subsequent decline in the sample autocorrelations is much slower than exponential, with the $k = 25$ and $k = 100$ values being 0.606 and 0.387 respectively. Estimates of d obtained using the GPH estimator will increase as p gets smaller, as there will be less regime switching: for the series in Figure 10.7 d is estimated to be 0.727. It is thus quite plausible that, by focusing on just the linear autocorrelation properties, it might be concluded that x_t is long memory rather than regime switching.

10.13 A useful definition of long memory, related to the hyperbolic decline of the autocorrelations found in **§10.2**, is that the quantity $\lim_{T\to\infty} \sum_{k=-T}^{T} |\rho_k|$ is non-finite. Granger and Hyung (2004) used this definition, in conjunction with an 'occasional break' model, to show that infrequent level shifts in mean can indeed give rise to observed long memory. This occasional break model represents the observed series x_t as

$$x_t = \mu_t + \varepsilon_t \qquad t = 1, \ldots, T \tag{10.7}$$

where ε_t is a noise variable and the level μ_t is defined as

$$\mu_t = \mu_{t-1} + q_t \eta_t \tag{10.8}$$

η_t is a zero mean i.i.d. process with variance σ_η^2, measuring the size of the shift in the mean, while q_t follows an i.i.d. binomial distribution such that $q_t = 0$ with probability $1 - p$ and $q_t = 1$ with probability p.

Granger and Hyung made the following assumption: the probability of a break converges to zero slowly as the sample size increases, i.e., $p \to 0$ as $T \to \infty$, but $\lim_{T\to\infty} Tp$ is a non-zero finite constant. This implies that the expected number of breaks, Tp, is bounded from above even in the extreme case that T increases to infinity, so that, regardless of sample size, realizations from this process have a finite number of breaks. Combining (10.7) and (10.8) allows x_t to be written as

$$x_t = (\mu_0 + q_1\eta_1 + \cdots + q_t\eta_t) + \varepsilon_t$$

The time-varying mean $\mu_0 + q_1\eta_1 + \cdots + q_t\eta_t$ thus shows infrequent level shifts depending on the size of p.

Granger and Hyung showed that the autocorrelations of x_t are such that

$$\rho_k = \frac{Tp\sigma_\eta^2}{\sqrt{Tp\sigma_\eta^2 + \sigma_\varepsilon^2}\sqrt{(T-k)p\sigma_\eta^2 + \sigma_\varepsilon^2}} \to \left(1 + \frac{\sigma_\varepsilon^2}{Tp\sigma_\eta^2}\right)^{-1} > 0$$

as $T \to \infty$ for all k and thus possess the long memory property. They also showed that the sample autocorrelations, $\hat{\rho}_{k,T}$, converge to non-zero values for any k such that $k/T \to 0$ as $T \to \infty$:

$$\hat{\rho}_{k,T} \to \left(1 + \frac{6\sigma_\varepsilon^2}{Tp\sigma_\eta^2}\right)^{-1}$$

These sample autocorrelations do not decline exponentially but decay very slowly as k increases, approaching a non-zero constant. As Tp increases there are more breaks and the sample autocorrelations increase in magnitude. A similar effect is found for an increase in σ_η^2, which produces breaks with larger magnitude, so that increases in Tp and σ_η^2 make the occasional-break process closer to a random walk.

The log-periodogram GPH estimator of the fractional differencing parameter d was shown by Granger and Hyung to be seriously biased

away from zero if x_t is generated by the occasional-break model. These results were confirmed by simulations, allowing Granger and Hyung to conclude that, if just linear properties of the data, such as autocorrelations, are considered, then an occasional break model will exhibit long-memory and that disentangling this model from an $I(d)$ process will become difficult as the size and number of breaks increase.

10.14 Granger and Hyung also discussed the converse problem, that a fractionally integrated series may exhibit spurious breaks. They were able to show that, as the sample size goes to infinity, a positive number of breaks in a fractionally integrated series will be detected, the actual number depending upon the value taken by d. On analyzing the absolute daily returns of the S&P 500, they found that, on analyzing subperiods, those with large estimates of d tended to exhibit the largest number of breaks, although this was not an exact relationship because the size of the breaks also needs to be taken into account. There was also little difference in the forecasting performance of the two models, thus pointing to the fact that it may be very difficult to discriminate between the two processes in many situations.

The links between long-memory processes, structural breaks and regime shifts have since been investigated by, for example, Diebold and Inoue (2001) and Banerjee and Urga (2005).

Extended memory

10.15 Granger and Hallman (1991a) and Granger (1995) extended the definitions of long memory in various ways. These extensions are based on the conditional probability density function of x_{t+h} given the information set $I_t : x_{t-j}, \mathbf{q}_{t-j}$, $j \geq 0$, where $\mathbf{q}_t$ is a vector of other explanatory variables: x_t is said to be *short memory in distribution* (SMD) with respect to I_t if

$$|P(x_{t+h} \text{ in } A | I_t \text{ in } B) - P(x_{t+h} \text{ in } A)| \to 0 \tag{10.9}$$

as $h \to \infty$ for all appropriate sets A and B such that $P(I_t \text{ in } B) > 0$. If (10.9) does not hold then x_t is said to have *long memory in distribution* (LMD). A narrower definition of memory focuses on the conditional mean

$$E(x_{t+h}|I_t) = f_{t,h}$$

so that $f_{t,h}$ is the optimum least squares forecast of x_{t+h} using I_t. x_t is said to be *short memory in mean* (SMM) if $\lim_{h\to\infty} f_{t,h} = F$, where F

is a random variable with a distribution which does not depend on I_t. If $f_{t,h}$ depends on I_t for all h then x_t is *extended memory in mean* (EMM).[2] If the optimum forecast is linear, so that, for example,

$$f_{t,h} = \sum_{j=0}^{t} \beta_{h,j} x_{t-j}$$

and the sequence $\beta_{h,j}$ does not tend to zero as h increases for all j, then x_t is called *linear EMM* (LEMM).

If x_t is SMD then it will also be SMM, as will any function of x_t provided that the unconditional mean of the function exists. If x_t is EMM then it must be LMD but not necessarily vice versa, although if this is the case then many functions $g(x_t)$ will be EMM. If x_t is EMM then any monotonic non-decreasing function of x_t will also be EMM. However, if x_t is EMM then a function of x_t that is not monotonic non-decreasing may be SMM, examples being the sine and cosine functions.

10.16 Although the concept of extended memory has made few inroads into applied time series analysis, the general use of fractional differencing and long memory models has become widespread across many disciplines. Recent surveys of this very popular and important concept are Velasco (2006), who concentrates on estimation of d, and Gil-Alana and Hualde (2009), who provide a wide-ranging discussion of the many applications of fractional differencing and long memory across subject areas. The legacy of Granger's initial foray into the topic and his subsequent extensions and applications is thus seen to be all pervading.

Spurious regressions

10.17 Granger and Newbold (1974) returned to the question considered by Box and Newbold (1971) and discussed in **§8.4**, that of explaining why puzzling cross-correlations may be found between detrended random walks (in fact, one can trace this puzzle back to Yule's nonsense correlations of **§§2.9–2.15**). More specifically, they focused attention on the then common practice in the applied econometrics literature of reporting time series regressions with an apparently high degree of fit, as measured by the coefficient of multiple correlation, R^2, accompanied by extremely low values of the Durbin–Watson statistic (recall **§4.2** but now denoted dw to avoid confusing it with the differencing parameter d).

> We find it very curious that whereas virtually every textbook on econometric methodology contains explicit warnings of the dangers of autocorrelated errors, this phenomenon crops up so frequently in well-respected applied work. ... The most extreme example we have met is an equation for which $R^2 = 0.99$ and $dw = 0.093$. However, we shall suggest that cases with much less extreme values may well be entirely spurious. (Granger and Newbold, 1974, page 111)

10.18 Granger and Newbold noted that two of the major consequences of autocorrelated regression errors, that estimates of the regression coefficients were inefficient and that forecasts based on the fitted regressions were suboptimal, were both well documented. They therefore focused on a third consequence, that the usual significance tests on the coefficients were invalid. To do this, they considered the usual linear regression model with stochastic regressors

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_{kt} X_{kt} + \varepsilon_t \qquad t = 1, \ldots, T \tag{10.10}$$

where $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$ and $E(\varepsilon_t \varepsilon_s) = 0$, $s \neq t$. A test of the null hypothesis that the 'independent' variables contribute nothing towards explaining the variation in the dependent variable, i.e., $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, is given by the statistic

$$F = \frac{T-k-1}{k} \frac{R^2}{1-R^2} \sim F(k, T-k-1) \tag{10.11}$$

Granger and Newbold made the important point that, although it is always possible, whatever the properties of the individual time series in (10.10), that there could exist some set of β's such that the assumptions on ε_t were satisfied, to the extent that the Y_t's did not constitute a white noise process, the null hypothesis H_0 *could not be true* and tests of it were therefore inappropriate.

If H_0 is correct and (10.10) is fitted to the levels of a set of non-stationary (or, at best, highly autocorrelated) series then the quantity F in (10.11) *will not* follow an F-distribution since under H_0 the residuals from (10.10), $\varepsilon_t = Y_t - \beta_0$, will have the same autocorrelation structure as Y_t itself.

Suppose that $k = 1$ in (10.10) and the regression is written $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$. If Y_t and X_t are *independent* AR(1) processes,

$$Y_t = \phi Y_{t-1} + a_t \qquad X_t = \phi^* X_{t-1} + b_t$$

then R^2 will be the square of the sample correlation between Y_t and X_t. Granger and Newbold showed that if ϕ and ϕ^* are large, in the region of 0.9 say, then the expected value of R^2 will be around 0.5, so that a high value of this statistic should not be regarded as evidence of a significant relationship between autocorrelated series. Furthermore, a low value of *dw* would suggest that there does not exist a set of β's such that the assumptions placed on ε_t are satisfied. Granger and Newbold thus argued that the phenomenon whereby R^2 exceeds *dw* might well arise from an attempt to fit regression equations relating the levels of independent time series.

10.19 To investigate this phenomenon in more detail, Granger and Newbold conducted a number of important simulation experiments which we recreate here, albeit using a greater number of simulations than they were able to.[3] Granger and Newbold began by considering the bivariate regression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, where Y_t and X_t are generated as *independent* random walks. Table 10.1 shows the frequency distribution of the t-statistic

$$t = \frac{|\hat{\beta}_1|}{\widehat{SE}(\hat{\beta}_1)}$$

which is customarily used to test the significance of $\hat{\beta}_1$, obtained from 1000 simulations of pairs of independent random walks, each of length $T = 50$, with starting values $Y_0 = X_0 = 100$ and each with standard normal innovations.

Using the traditional t-test at the 5% significance level (so that the critical t value is approximately 2), the null hypothesis of no relationship between the two series ($\beta_1 = 0$) would be incorrectly rejected two-thirds of the time. If $\hat{\beta}_1/\widehat{SE}(\hat{\beta}_1)$ was distributed as standard normal, then the expected value of t would be $\sqrt{22}/\pi = 0.8$, but the average value of the observed t-statistics is 4.13, suggesting that the standard deviation of $\hat{\beta}_1$

Table 10.1 t-statistics obtained from regressing two independent random walks

t	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9	9–10
Frequency	172	161	138	112	85	79	66	49	37	35
t	10–11	11–12	12–13	13–14	14–15	15–16	16–17	17–18	18–19	19–20
Frequency	19	17	8	10	5	3	0	0	3	0

is being underestimated by a factor of over 5. Thus, instead of using a t-value of 2, a value in excess of 10 should be used when deciding whether an estimated coefficient is significant or not at the 5% level (observe that there are 65 t-statistics greater than 10 reported in Table 10.1).

Granger and Newbold's second simulation allowed for up to $k = 5$ regressors and considered series generated both as random walks and as ARIMA(0, 1, 1) processes, along with their changes (that is, white noise and MA(1) processes). The ARIMA(0, 1, 1) processes were generated as the sum of a random walk and independent standard normal white noise, again using starting values of 100 and standard normal innovations for the random walks.[4] The results from 1000 simulations with $T = 50$ are shown in Table 10.2, where the proportion of times the null of no relationship is rejected when levels are used increases with the number of regressors, being in excess of 85 per cent for $k \geq 3$, although

Table 10.2 Regressions of a series on k independent 'explanatory' variables. R^2 is corrected for degrees of freedom

		% times H_0 rejected at 5% level	Average *dw*	Average R^2	% $R^2 > 0.7$
		Random walks			
Levels	$k = 1$	67	0.33	0.23	5
	$k = 2$	86	0.46	0.38	14
	$k = 3$	93	0.58	0.49	23
	$k = 4$	96	0.70	0.55	31
	$k = 5$	97	0.81	0.61	40
Changes	$k = 1$	5	2.01	−0.001	0
	$k = 2$	4	2.01	−0.002	0
	$k = 3$	5	2.01	0.000	0
	$k = 4$	6	2.01	0.000	0
	$k = 5$	6	2.01	−0.000	0
		ARIMA(0, 1, 1)			
Levels	$k = 1$	59	0.69	0.18	1
	$k = 2$	78	0.86	0.28	4
	$k = 3$	85	1.01	0.37	9
	$k = 4$	90	1.13	0.42	13
	$k = 5$	91	1.24	0.46	16
Changes	$k = 1$	7	2.59	0.004	0
	$k = 2$	8	2.57	0.007	0
	$k = 3$	8	2.54	0.011	0
	$k = 4$	8	2.52	0.013	0
	$k = 5$	9	2.51	0.018	0

matters considerably improve when changes are used.[5] These findings led Granger and Newbold (1974, page 116–17) to report that

> the probability of accepting H_0, the hypothesis of no relationship, becomes very small indeed for $k \geq 3$ when regressions involve independent random walks. The average R^2 steadily rises with k, as does the average *dw*, in this case. Similar conclusions hold for the ARIMA(0, 1, 1) process. When white noise series, i.e., changes in random walks, are related, classical regression yields satisfactory results since the error series will be white noise and least squares fully efficient. However, in the case where changes in the ARIMA(0, 1, 1) series are considered – that is, first order moving average processes – the null hypothesis is rejected, on average, twice as often as it should be.
>
> It is quite clear from these simulations that if one's variables are random walks, or near random walks, and one includes in regression equations variables which should not in fact be included, then *it will be the rule* rather than the exception to find spurious relationships. It is also clear that a high value of R^2, combined with a low value of *dw*, is *no indication of a true relationship*. (italics in original: notation altered for consistency)

In a subsequent paper, Granger and Newbold (1977) provided further simulation results in which two ARIMA(0, 1, 1) processes were regressed together. These processes were defined as

$$Y_t = Y_{t-1} + a_t + \theta a_{t-1} \qquad X_t = X_{t-1} + b_t + \theta^* b_{t-1}$$

The regression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ was computed for various values of θ and θ^* using 1,000 simulations (again with $T = 50$), producing the results shown in Table 10.3.

> The main conclusion from this table is that employment of the decision procedure, 'reject the null hypothesis of no relationship between two series only if t differs significantly from zero and *dw* does not differ significantly from two', will generally not lead one astray (although, of course, neglect of the second condition will do so). The exception is for moderately large values of $-\theta$, with $-\theta^*$ not too large. Given this combination of circumstances, a significant regression coupled with an absence of warning signals from the Durbin–Watson statistic will be found on about 20 percent of occasions. (Granger and Newbold, 1977, page 11: notation altered for consistency)

Granger and Newbold (1977) then 'corrected' for autocorrelated errors using the Cochrane–Orcutt (1949) iterative procedure, a then common approach for dealing with autocorrelation. Table 10.4 shows the percentage of times for which significant estimates of β_1 were obtained after this correction, with Granger and Newbold concluding that the null hypothesis $\beta_1 = 0$ was still incorrectly rejected for a wide range of wholly reasonable moving average parameter values.

10.20 What conclusions did Granger and Newbold draw for econometric practice from these simulation results? They first emphasized a supposedly well-known implication that they thought had perhaps been stated insufficiently strongly:

> if a regression equation relating economic variables is found to have strongly related residuals, equivalent to a low Durbin–Watson value, the *only conclusion that can be reached is that the equation is mis-specified*, whatever the value of R^2 observed. (Granger and Newbold, 1974, page 117: italics in original)

They then considered the question of what to do about such a misspecification. The usual solutions are threefold: include a lagged dependent variable as an additional regressor, take first-differences of the variables, or assume a simple first-order autoregressive form for the residual. As can be seen from Table 10.4, the third option does not appear to produce satisfactory results. Granger and Newbold were also not convinced by the first solution as they thought that estimation bias could be substantial, particularly with the short samples then typically available to econometricians. They therefore recommended the taking of first differences as, although it may not completely remove the problem, it would considerably improve the interpretability of the coefficients.

Indeed, Granger and Newbold stressed that they were not advocating first differencing as a 'sure-fire universal solution', but they did think that it would be useful for a class of time series that occurs frequently in practice:

> many economic series are rather smooth, in that the first serial correlation coefficient is very near unity and the other low-order serial correlations are also positive and large. Thus, if one has a small sample, of say twenty terms, the addition of a further term adds very little to the information available, as this term is so highly correlated with its predecessor. It follows that the total information available is

Table 10.3 Percentage of times the t and *dw* statistics are significant at 5% level for a regression of an ARIMA(0, 1, 1) series on an independent ARIMA(0, 1, 1) series

			$\theta^* = 0$		$\theta^* = -0.2$		$\theta^* = -0.4$		$\theta^* = -0.6$		$\theta^* = -0.8$	
			t		t		t		t		t	
			N.Sig	Sig	N.Sig	Sig	N.Sig	Sig	N.Sig	Sig	N.Sig	Sig
$\theta = 0$	*dw*	N.Sig	0	0	0	0	0	0	0	0	0	0
		Inconc.	0	0	0	0	0	0	0	0.1	0	0
		Sig	34.1	65.9	36.9	63.1	36.3	63.7	44.0	55.9	62.3	37.7
		Mean *dw*	0.33		0.35		0.38		0.42		0.35	
		Mean t	3.85		3.78		3.46		3.01		1.86	
$\theta = -0.2$	*dw*	N.Sig	0.1	0	0.1	0.2	0	0	0.1	0	0	0.1
		Inconc.	0	0.1	0	0.2	0	0.1	0	0	0	0.1
		Sig	35.8	64.0	34.4	65.1	36.6	63.3	44.3	55.6	62.8	37.0
		Mean *dw*	0.46		0.49		0.50		0.52		0.47	
		Mean t	3.74		3.81		3.40		2.85		1.87	
$\theta = -0.4$	*dw*	N.Sig	0.3	0.8	0.3	0.8	0.5	1.0	0.5	0.6	0.5	0.2
		Inconc.	0.2	0.6	0.3	0.9	0.1	0.8	0.2	0.9	0.3	0.3
		Sig	39.3	58.8	37.6	60.1	40.2	57.4	48.1	49.7	67.3	31.4
		Mean *dw*	0.69		0.35		0.71		0.72		0.65	
		Mean t	3.48		3.78		3.21		2.62		1.71	

$\theta = -0.6$	dw	N.Sig	5.1	7.0	6.8	7.5	6.1	6.7	5.6	5.7	6.4	2.4
		Inconc.	2.3	2.9	1.9	3.2	2.4	1.6	1.7	2.2	3.6	2.0
		Sig	33.0	49.7	39.1	41.5	41.2	42.0	45.9	38.9	59.6	26.0
		Mean dw	1.10		1.09		1.09		0.42		1.03	
		Mean t	3.05		2.73		2.59		3.01		1.58	
$\theta = -0.8$	dw	N.Sig	40.5	20.6	38.0	20.1	40.7	21.5	41.5	17.4	42.4	9.2
		Inconc.	7.0	3.6	4.7	4.3	5.2	2.6	5.2	2.5	6.8	1.3
		Sig	16.2	12.1	20.4	12.5	20.0	10.0	23.7	9.7	33.7	7.6
		Mean dw	1.69		1.66		1.68		1.65		1.59	
		Mean t	1.87		1.85		1.72		1.57		1.18	

Table 10.4 Percentage of times the t-statistic is significant at 5% level in a regression of an ARIMA(0, 1, 1) series on an independent ARIMA(0, 1, 1) series 'allowing' for first order serial correlation in residuals by Cochrane–Orcutt iterative estimation technique

	$\theta^* = 0$	$\theta^* = -0.2$	$\theta^* = -0.4$	$\theta^* = -0.6$	$\theta^* = -0.8$
$\theta = 0$	16.2	16.5	15.1	12.7	5.5
$\theta = -0.2$	19.4	20.3	19.6	16.5	7.5
$\theta = -0.4$	23.7	26.0	23.2	17.9	9.6
$\theta = -0.6$	31.1	27.8	26.0	21.9	11.7
$\theta = -0.8$	27.5	28.3	24.6	21.4	13.2

> very limited and the estimates of parameters associated with this data will have high variance values. However, a simple calculation shows that the first differences of such a series will necessarily have serial correlations that are small in magnitude, so that a new term of the differenced series adds information that is almost uncorrelated to that already available and this means that estimates are more efficient. One is much less likely to be misled by efficient estimates. (*ibid.*, page 118)

Moreover, Granger and Newbold thought that differencing would prove beneficial when testing economic theories, because

> if one does obtain a very high R^2 value from a fitted equation, one is forced to rely on the correctness of the underlying theory, as testing the significance of adding further variables becomes impossible. It is one of the strengths of using changes, or some similar transformations, that typically lower R^2 values result and so more experimentation and testing can be contemplated. In any case, if a "good" theory holds for levels, but is unspecific about the time-series properties of the residuals, then an equivalent theory holds for changes so that nothing is lost by model building with both levels and changes. However, much would be gained from this strategy as it may prevent the presentation in econometric literature of possible spurious regressions, which we feel is still prevalent despite the warnings given in the text books about this possibility. (ibid., page 120)

Error correction and co-integration

10.21 Granger and Newbold's advocacy of differencing for alleviating the spurious regression problem did not find universal support.

The econometricians from the 'LSE group' (recall **§1.2**) were particularly sceptical, as was shown in David Hendry's (1977) comments on, inter alia, Granger and Newbold (1977), described by Christopher Sims in the introduction to the volume in which these papers appeared as 'somewhat acerbic'!

Hendry reconsidered the setup of Table 10.3 and noted that the regression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, with errors falsely assumed to follow the process $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$ for $|\rho| < 1$, was equivalent to

$$Y_t = \beta_0(1-\rho) + \beta_1 X_t - \rho\beta_1 X_{t-1} + \rho Y_{t-1} + \nu_t \tag{10.12}$$

which itself is a restricted version of

$$Y_t = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 Y_{t-1} + w_t \tag{10.13}$$

The restriction is $\gamma_2 + \gamma_1\gamma_3 = 0$, which Sargan (1964) had shown could be tested by the likelihood ratio statistic $\ell = T\ln\left(\sum \hat{v}_t / \sum \hat{w}_t\right) \overset{a}{\sim} \chi^2(1)$ if (10.13) *really is* the unrestricted version of (10.12). If the apparent autocorrelation arises, however, because the regression is a misspecified approximation to (10.13) then the latter would provide a better fit than (10.12) and hence a large value of ℓ. Hendry suggested that ℓ would reject (10.12) reasonably frequently relative to the number of cases of spurious significance found in Table 10.4 and also suggested that X_t and X_{t-1} would now rarely have a significant effect in (10.13). Consequently, he argued that, if (10.13) was used as a new baseline model with w_t allowed to be autocorrelated, 'it is hard to see why an approximately correct model could not be detected even for the paradigm used by Granger and Newbold' (Hendry, 1977, page 184).

10.22 Hendry then offered two distinct interpretations of differencing an equation such as (10.13). The first was the 'operator form', implicitly considered by Granger and Newbold, which transforms (10.13) to

$$\Delta Y_t = \gamma_1 \Delta X_t + \gamma_2 \Delta X_{t-1} + \gamma_3 \Delta Y_{t-1} + \Delta w_t \tag{10.14}$$

If an intercept had been included then this would correspond to a trend term appearing in (10.13). The autocorrelation properties of the error term are completely altered since Δw_t will only be white noise if w_t is a random walk.

On the other hand, an equation in first differences could be obtained from (10.13) by imposing the parameter restrictions $\gamma_1 + \gamma_2 = 0$ and

$\gamma_3 = 1$, producing the 'restriction form'

$$\Delta Y_t = \gamma_0 + \gamma_1 \Delta X_t + w_t \tag{10.15}$$

If the restrictions are valid then the interpretation of the intercept and the error term now remain unaltered and (10.15) implies the exclusion of the regressors ΔX_{t-1} and ΔY_{t-1} compared to (10.14). Indeed, if (10.15) *is* the true data generation process such that w_t is white noise then so must be the error term in the 'levels' equation (10.13). The regression (10.14) is then an incorrect specification as it falsely includes ΔX_{t-1} and ΔY_{t-1}, excludes the intercept, and has a moving average error with a coefficient of -1.[6]

Hendry thought that distinguishing between (10.14) and (10.15) should not present difficulties (even though the original spurious regression problem, that of obtaining nonsense results if $\gamma_1 = 0$ but Y_t was regressed on X_t without including Y_{t-1}, still 'lurked in the background'). Yet he continued to feel that differencing remained problematic, as both (10.14) and (10.15) have unacceptable features in terms of being universally valid formulations for economic systems, although these features may not be as problematic for other environments. In particular, (10.15) either has *no* equilibrium solution in terms of Y_t and X_t, or one that collapses to zero if $\gamma_1 = 0$. Moreover, the time paths that Y_t can describe are independent of the states of disequilibrium existing in prior periods. Hendry was particularly keen to emphasize that there were more ways of transforming to stationarity than by just differencing and that the choice of which transformation to adopt should be based on considerations from economic theory. While marginal adjustments might favour differencing, long-run considerations could suggest specifications of the form

$$\Delta Y_t = \gamma_0 + \gamma_1 \Delta X_t + (\gamma_2 - 1)(Y_{t-1} - X_{t-1}) + w_t \tag{10.16}$$

which is obtained from (10.13) by imposing the restriction that $\gamma_1 + \gamma_2 + \gamma_3 = 1$. If the variables are measured in logarithms then the term $Y_{t-1} - X_{t-1}$ can be regarded as the logarithm of the ratio of the lagged levels and implies that the two variables are related in the long-run by having a unit elasticity, with their individual non-stationarities being 'cancelled' by the act of taking log-differences. The specification (10.16) thus embodies a long-run equilibrium condition in levels in an equation otherwise containing differences and is thus able to model both short and long-run relationships between Y_t and X_t. Equations of this type

quickly became known as 'error correction' models, although they had been suggested earlier by Sargan (1964) and could also be interpreted as having derivative and proportional control mechanisms, corresponding to γ_1 and $\gamma_2 - 1$ respectively (cf. Box and Jenkins', 1962, three term predictor of **§6.9**, for which (10.16) may be regarded as a multivariate generalization).

10.23 This critique of differencing and the emphasis on error correction models forced Granger to consider more carefully the implications of the orders of integration of variables appearing in a regression equation. Granger (1981, page 121) termed such a regression a *generating equation* if 'a simulation of the explanatory side should produce the major properties of the variable being explained', calling an equation which had this property *consistent*, although it has subsequently become known as 'balanced' (see Banerjee et al., 1993, pages 164–8, and Granger, 1999b, pages 18–22). He gave as a simple example of a non-consistent equation a regression in which Y_t was positive but X_t was unbounded in both directions, a specific example of which is when Y_t is exponentially distributed and X_t is normally distributed.

In terms of integrated series, suppose that $x_t \sim I(d_x)$, $y_t \sim I(d_y)$, where d_x and d_y may be non-integer, and $a(B)$ is an integrating filter of order d', that is, a filter of infinite order such that $a(B) = \Delta^{-d'} a'(B)$, where $a'(z)$ has no roots at $z = 0$. From Granger (1981) it then follows that $a(B)x_t \sim I(d_x + d')$ and, in general, that $z_t = bx_t + cy_t \sim I(\max(d_x, d_y))$. This result follows from noting that the spectrum of z_t is

$$f_z(\omega) = b^2 f_x(\omega) + c^2 f_y(\omega) + 2bc f_{xy}(\omega)$$

where $|f_{xy}(\omega)|^2 \leq f_x(\omega) f_y(\omega)$ (extending the results in **§5.16**). For small ω, $f_x(\omega)$ and $f_y(\omega)$ are proportional to ω^{-2d_x} and ω^{-2d_y}, respectively, so that the variable with the largest d value will dominate at low frequencies.

More generally, consider the equation

$$b(B)y_t = c(B)x_t + h(B)\varepsilon_t \tag{10.17}$$

where all the polynomials are of finite order and ε_t is white noise with finite variance and independent of x_t. This equation will only be consistent if $d_x = d_y$: if $d_x < d_y$, for example, then, for it to be consistent, either $c(B)$ must be an integrating filter of order $d_y - d_x$ or $h(B)$ must be an integrating filter of order d_y (or indeed both), so that in neither case can the polynomials be of finite order. As an example, suppose $d_x < \frac{1}{2}$ and

$1 > d_y > \frac{1}{2}$, so that x_t has finite variance, but the variance of y_t is infinite. Clearly y_t cannot be explained by x_t using just finite polynomials and (10.17) cannot be consistent.

For the more general model

$$b(B)y_t = c(B)x_t + g(B)z_t + h(B)\varepsilon_t \tag{10.18}$$

the relevant condition must be that $d_y = \max(d_x, d_z)$ unless one of the polynomials in (10.18) corresponds to an integrating filter and is hence of infinite order.

10.24 Of crucial importance, however, was that Granger found a special case in which these rules did not hold.[7] For simplicity, suppose that $c(B) = c$ and $g(B) = g$ in (10.18) and that ε_t has unit variance with $h(B)\varepsilon_t \sim I(d_y)$ for $d_y > 0$. The spectrum of the right-hand side of (10.18) will then be

$$(c^2 f_x(\omega) + g^2 f_z(\omega) + gcf_{xz}(\omega)) + |h(z)|^2/2\pi \tag{10.19}$$

Granger's special case has

(i) $f_x(\omega) = \alpha^2 f_z(\omega)$ for small ω, so that $d_x = d_z$;
(ii) $f_{xz}(\omega) = \alpha f_z(\omega)$ for small ω, which implies that the coherence between x_t and z_t will be unity and the phase angle zero for small ω.

Any pair of series obeying (i) and (ii) was termed *co-integrated* by Granger. If, as well, $g = -c\alpha$ then the spectrum (10.19) will reduce to just $|h(z)|^2/2\pi$ at low frequencies, so that a model of the form (10.18) would be appropriate even when $d_y < \max(d_x, d_z)$: indeed, since now

$$b(B)y_t = c(x_t - \alpha z_t) + h(B)\varepsilon_t \tag{10.20}$$

the difference between two co-integrated series can result in an $I(0)$ series.

More generally, consider $x_t = z_t + q_t$, where $d_z = d_x$, $d_q < d_x$ and z_t and q_t are independent. It then follows that x_t and z_t will be co-integrated but the difference $q_t = x_t - z_t$ will be $I(d_q)$. It will also follow that $\alpha(B)x_t$ and $\beta(B)z_t$ will also be co-integrated for $\alpha(B)$ and $\beta(B)$ of finite order, so that x_t and z_{t-k} will also be co-integrated for all k (although the approximation that the phase is zero at low frequencies may become untenable for large values of k). Two co-integrated series will therefore move in a similar way over long periods and, although they may be unequal in the short term, they will be 'tied together' in the long run.

10.25 Granger (1981) noted that the appearance of $x_t - \alpha z_t$ in (10.20) was akin to the error correction term in (10.16), pointing out that if $d_X > d_w$ in the latter then X_t and Y_t would be co-integrated and hence they would move closely in the long run, so explaining why such models had proved to be empirically popular: as well as Sargan (1964), influential applications of the error correction model had been provided by, amongst others, Davidson et al. (1978), Hendry and von Ungern-Sternberg (1981) and Currie (1981). If, however, $d_X = d_Y = d_w$ then co-integration would not hold as then the coherence between X_t and Y_t would not necessarily be high at low frequencies: in this case (10.16) would essentially just be an algebraic rearrangement of a levels specification between X_t and Y_t.

These ideas began to be formalized in Granger and Weiss (1983), who considered first the bivariate model

$$a_1(B)\Delta^d y_t = \beta(y_{t-1} - Ax_{t-1}) + b_1(B)\Delta^d x_t + c_1(B)\varepsilon_{1t} \quad (10.21a)$$

$$a_2(B)\Delta^d x_t = c_2(B)\varepsilon_{2t} \quad (10.21b)$$

where ε_{1t} and ε_{2t} are independent zero mean white noises and d is either 0 or 1, with $\Delta^d x_t$ and $\Delta^d y_t$ assumed to be stationary. The polynomials in B are all finite with the property that $a_1(1) \neq 0$, $b_1(1) \neq 0$, etc., and $a_1(0) = a_2(0) = c_1(0) = c_2(0) = 1$. Note that (10.21) has a one-way causal structure, with x_t causing y_{t+1} but y_t not causing x_{t+1}, while allowing $b_1(0)$ to be non-zero allows for the possibility of simultaneity between x_t and y_t. Only a single error correction term is included in (10.21a) since additional terms such as $\beta_2(y_{t-2} - Ax_{t-2})$ can always be incorporated without altering the structure of the model. For example, to take a simple case,

$$\begin{aligned}\Delta y_t &= \beta_1(y_{t-1} - Ax_{t-1}) + \beta_2(y_{t-2} - Ax_{t-2}) + a\Delta y_{t-1} + b_1\Delta x_t + \varepsilon_{1t} \\ &= (\beta_1 + \beta_2)(y_{t-1} - Ax_{t-1}) + (a - \beta_2)\Delta y_{t-1} + b\Delta x_t + \beta_2 A\Delta x_{t-1} + \varepsilon_{1t}\end{aligned}$$

Equation (10.21a) may be written as

$$\alpha_1(B)y_t = \alpha_2(B)x_t + c_1(B)\varepsilon_{1t} \quad (10.22)$$

on defining $\alpha_1(B) = \Delta^d a_1(B) - \beta B$ and $\alpha_2(B) = \Delta^d b_1(B) - \beta AB$. Eliminating x_t from (10.22) using (10.21b) gives

$$a_2(B)\alpha_1(B)\Delta^d y_t = \alpha_2(B)c_2(B)\varepsilon_{2t} + c_1(B)a_2(B)\Delta^d\varepsilon_{1t} \quad (10.23)$$

It is clear that, irrespective of whether d is 0 or 1, the right-hand side of this equation can always be written as a finite moving average (recall case (iii) of **§9.33**). It then follows that $y_t \sim I(d)$ regardless of the value of β. However, if $d = 1$ the value of the error correction coefficient has a dramatic impact on the low-frequency component of y_t. If $\beta \neq 0$ then replacing B by $e^{i\omega}$ in (10.23) and letting ω become small such that $1 - e^{i\omega}$ is negligible ensures that, when considered in the frequency domain, the term involving ε_{1t} on the right-hand side is essentially zero. The low-frequency component of y_t will then be largely determined by the low frequency component of ε_{2t}, which in turn also determines the low-frequency component of x_t through (10.21b). If $\beta = 0$, on the other hand, it is clear that the low-frequency components of both ε_{1t} and ε_{2t} will jointly determine the low-frequency component of y_t.

On defining $z_t = y_t - Ax_t$, this is seen to have the univariate model

$$a_2(B)\alpha_1(B)z_t = c_2(B)(b_1(B) - Aa_1(B))\varepsilon_{2t} + c_1(B)a_2(B)\varepsilon_{1t}$$

and it follows immediately that $z_t \sim I(0)$ even if x_t and y_t are both $I(1)$, which leads to Granger's definition of co-integration: if $x_t \sim I(d)$ and $y_t \sim I(d)$ and there exists a constant A such that $z_t = y_t - Ax_t \sim I(0)$, then x_t and y_t are said to be co-integrated. A will be unique. When $d = 1$ both x_t and y_t will have infinite variance but there will exist a unique and constant A such that z_t has finite variance, which will not in general be the case, as a linear combination of infinite variance series will typically also have infinite variance. Moreover, if x_t and y_t are generated by (10.21) with $d = 1$ then they will necessarily be co-integrated, whereas if they are not co-integrated the error correction model will be inappropriate. This is because, while the left-hand side of (10.21a), Δy_t, will have finite variance, the right-hand side will contain the infinite variance term $y_{t-1} - Ax_{t-1}$ and the equation is clearly not consistent. The differenced series Δy_t and Δx_t will have a coherence of unity and a phase of zero at low frequencies so that y_t and Ax_t will have identical low-frequency components (that is, they are said to have 'common stochastic trends'), but they can differ substantially at high frequencies.

If x_t and y_t are co-integrated then so will be any linear transformations and series obtained by applying finite-length filters: for example, $x'_t = a + bx_{t-s}$ and $y'_t = c + fy_{t-k}$ will be co-integrated for any finite (although not too large) values of s and k and for any constants a, b, c and f.

When $d = 0$ it is clear that $y_t - Ax_t$ will be $I(0)$ for *any* A. The model (10.22) can then *always* be written as (10.21a) with $d = 1$ but x_t will

be given by (10.21b) with $d = 0$. Consequently, for $I(0)$ series the error correction model has no special implications.

10.26 Suppose that x_t and y_t are co-integrated but that $x_t = x_{1t} + \gamma x_{2t}$. The error correction term in (10.21a) now becomes $\beta(y_{t-1} - A_1 x_{1,t-1} - A_2 x_{2,t-1})$ and, if $y_t \sim I(1)$, a necessary condition for both x_{1t} and x_{2t} to enter the error correction is that they are both $I(1)$. If, say, $x_{1t} \sim I(d)$, $d > 1$, then the error correction term cannot be $I(0)$, while if $x_{1t} \sim I(0)$ it cannot contribute to the coherence, at low frequencies, between Δy_t and Δx_t and therefore should not appear in the error correction. Assuming this condition, Granger and Weiss (1983) obtain the following relationship between coherences at low frequencies:

$$1 - C_{12}^2 - C_{1y}^2 - C_{2y}^2 + 2C_{12}C_{1y}C_{2y} = 0$$

where C_{12} is the coherence between x_{1t} and x_{2t} at low frequencies and C_{1y} and C_{2y} are the coherencies between these series and y_t. Some consequences of this relationship are

(i) If $C_a = 1$ then $C_b = C_c$: if any pair of series are co-integrated then the remaining pairs must be equally related at low frequencies.
(ii) If $C_b = C_c = 1$ then $C_a = 1$: if any two pairs are co-integrated then the remaining pair must also be co-integrated.
(iii) If $C_a = 0$ then $C_b + C_c = 1$: even if x_t and y_t are co-integrated it does not necessarily mean that x_{1t} and x_{2t} will be co-integrated with y_t.

This last property implies that a search for co-integrated series should not be done in pairs when more than two series are being considered.

10.27 The model (10.21) allows only for causality running from x_t to y_{t+1}. Feedback may be allowed by extending the model (with $d = 1$) to

$$a_1(B)\Delta y_t = \beta_1(y_{t-1} - A_1 x_{t-1}) + b_1(B)\Delta x_t + c_1(B)\varepsilon_{1t} \quad (10.24a)$$

$$a_2(B)\Delta x_t = \beta_2(y_{t-1} - A_2 x_{t-1}) + b_2(B)\Delta y_t + c_2(B)\varepsilon_{2t} \quad (10.24b)$$

or, equivalently, as

$$\alpha_1(B)y_t = \alpha_2(B)x_t + c_1(B)\varepsilon_{1t} \quad (10.25a)$$

$$\alpha_3(B)x_t = \alpha_4(B)y_t + c_2(B)\varepsilon_{1t} \quad (10.25b)$$

where

$$\alpha_1(B) = \Delta a_1(B) - \beta_1 B \qquad \alpha_2(B) = \Delta b_1(B) - \beta_1 A_1 B$$
$$\alpha_3(B) = \Delta a_2(B) + A_2\beta_2 B \qquad \alpha_4(B) = \Delta b_2(B) + \beta_2 B$$

To identify the model a recursive scheme may be assumed, so that $b_2(0) = 0$ but $b_1(0) \neq 0$. The univariate models for y_t and x_t take the form

$$D(B)y_t = c_1(B)\alpha_3(B)\varepsilon_{1t} + c_2(B)\alpha_2(B)\varepsilon_{2t}$$
$$D(B)x_t = c_1(B)\alpha_4(B)\varepsilon_{1t} + c_2(B)\alpha_1(B)\varepsilon_{2t}$$

where

$$D(B) = \alpha_1(B)\alpha_3(B) - \alpha_2(B)\alpha_4(B)$$

To ensure that y_t and x_t are both $I(1)$, $D(B)$ must contain the factor Δ, which will be the case if either $\beta_1\beta_2 = 0$ or $A_1 = A_2$. The model for $z_t = y_t - Ax_t$ takes the form

$$\begin{aligned} D(B)z_t &= (c_1(B)\alpha_3(B) - Ac_1(B)\alpha_4(B))\varepsilon_{1t} + (c_2(B)\alpha_2(B) - Ac_2(B)\alpha_1(B))\varepsilon_{2t} \\ &= f_1(B)\varepsilon_{it} + f_2(B)\varepsilon_{2t} \end{aligned}$$

If $\beta_1\beta_2 = 0$ or $A_1 = A_2 = A$ then $f_1(B)$ and $f_2(B)$ will both contain the factor Δ, which then cancels with the same factor in $D(B)$, giving $z_t \sim I(0)$ as required. It must therefore be the case that, for y_t and x_t to be co-integrated and for an error-correction term to appear in both equations of (10.24), $A_1 = A_2 = A$. If only one error-correction term appears in (10.24), say if $\beta_2 = 0$ and $\beta_1 \neq 0$, then y_t and x_t will be co-integrated, with the low-frequency component of ε_{2t} driving the low-frequency components of both y_t and x_t. If both β_1 and β_2 are non-zero then the low-frequency components of y_t and x_t are driven by a mixture of the low-frequency components of ε_{1t} and ε_{2t}.

10.28 Given the potential importance of co-integrating relationships and, consequently, of error-correction representations, it was clearly important to discover whether such relationships could be uncovered if they existed. Granger and Weiss pointed out that, if both y_t and x_t are $I(1)$, then the typical approach to model building would be to focus on bivariate models of the differenced series Δy_t and Δx_t, consequently identifying models such as

$$\alpha_1(B)\Delta y_t = \alpha_2(B)\Delta x_t + c_1(B)\Delta\varepsilon_{1t}$$

Although this model is over-differenced, it is very likely that the unit root in the moving average term $c_1(B)\Delta\varepsilon_{1t}$ would not be found, especially if it was not being looked for. Granger and Weiss thus suggested that, prior to a full analysis, simple tests of the error-correction specification should be developed, basing these on least squares estimation of the 'levels' regression

$$y_t = m + Ax_t + e_t$$

Their first suggestion was to try to determine whether the residuals $\hat{e}_t = y_t - \hat{m} - \hat{A}x_t$ were $I(0)$ or $I(1)$ by standard identification methods such as whether the sample autocorrelation function declines fast enough for the residuals to be considered stationary (cf. **§§6.13–6.17**). There are clearly two major difficulties with this identification procedure: the estimate of A will generally be inefficient, as there is no reason to suppose that e_t will be white noise, and no formal *test* of $e_t \sim I(0)$ is being used. Granger and Weiss thus proposed an 'efficient test' based on the residuals from fitting the extended models

$$y_t = m + Ax_t + \sum_{i=1}^{p} \alpha_j \Delta y_{t-i} + \sum_{j=0}^{q} \beta_j \Delta x_{t-j} + \varepsilon_t$$

and

$$y_t = m + Ax_t + \gamma(y_{t-1} - Ax_{t-1}) + \sum_{i=1}^{p} \alpha_j \Delta y_{t-i} + \sum_{j=0}^{q} \beta_j \Delta x_{t-j} + \varepsilon_t$$

Under the hypothesis that y_t and x_t are co-integrated ε_t should be white noise if the lag lengths p and q are chosen appropriately. Note that asking if the estimate of A is significant in these regressions does not provide a test for the presence of error correction because, in its absence, a spurious regression may well occur.

10.29 Although Granger and Weiss presented several examples of this approach to testing for co-integration, it quickly became apparent that it faced many difficulties and that providing useful tests would need further research. Moreover, a more general formulation of co-integration and error correction was also clearly required. Both of these extensions were provided in Engle and Granger (1987), which has since become one of the most heavily cited papers in the history of econometrics and certainly the most heavily cited paper in time series econometrics

(see Boswijk, Franses and van Dijk, 2010, and, for personal reminiscences on the genesis of the paper, Granger, 2010b).

Engle and Granger first offered a broader definition of co-integration: the components of a vector $\mathbf{x}_t$ are said to be *co-integrated of order d, b*, denoted $\mathbf{x}_t \sim CI(d, b)$, if (i) all components of $\mathbf{x}_t$ are $I(d)$; (ii) there exists a vector $\boldsymbol{\alpha}(\neq 0)$ such that $z_t = \boldsymbol{\alpha}'\mathbf{x}_t \sim I(d-b)$, $b > 0$. The vector $\boldsymbol{\alpha}$ is called the *co-integrating vector* and z_t is known as the equilibrium error, using the idea that, if there is an equilibrium condition $\boldsymbol{\alpha}'\mathbf{x}_t = 0$, $z_t = \boldsymbol{\alpha}'\mathbf{x}_t$ represents the extent to which the system diverges from this equilibrium. As Engle and Granger (1987, pages 253–4) went on to explain, in the typical case where $d = b = 1$,

> co-integration would mean that if the components of $\mathbf{x}_t$ were all $I(1)$, then the equilibrium error would be $I(0)$ and z_t will rarely drift far from zero if it has zero mean and z_t will often cross the zero line. Putting this another way, it means that equilibrium will occasionally occur, at least to a close approximation, whereas if $\mathbf{x}_t$ was not co-integrated, then z_t can wander widely and zero-crossings would be very rare, suggesting that in this case the equilibrium concept has no practical implications. The reduction in the order of integration implies a special kind of relationship with interpretable and testable consequences. If however all the elements of $\mathbf{x}_t$ are already stationary so that they are $I(0)$, then the equilibrium error z_t has no distinctive property if it is $I(0)$.

This move to a multivariate setup has several intriguing consequences. If $\mathbf{x}_t$ has $N > 2$ components then there may be more than one co-integrating vector, for it is clearly possible that several equilibrium relations might govern the joint behaviour of the variables. Engle and Granger assumed that there are exactly $r \leq N-1$ linearly independent co-integrating vectors, and that these are gathered together in the $N \times r$ matrix $\mathbf{A}$ which, by construction, has rank r, known as the 'co-integrating rank' of $\mathbf{x}_t$.

Engle and Granger then defined the error correction representation of $\mathbf{x}_t$ as

$$\boldsymbol{A}(B)\Delta\mathbf{x}_t = -\boldsymbol{\Gamma}\mathbf{z}_{t-1} + \mathbf{u}_t \tag{10.26}$$

where $\mathbf{u}_t$ is a stationary multivariate disturbance, $\boldsymbol{A}(0) = \mathbf{I}$, all elements of $\boldsymbol{A}(1)$ are finite, $\mathbf{z} = \mathbf{A}'\mathbf{x}_t$ is an $r \times 1$ vector of error corrections and $\boldsymbol{\Gamma} \neq \mathbf{0}$ is an $N \times r$ matrix of coefficients.

10.30 If each component of $\mathbf{x}_t$ is $I(1)$ then there will always exist a multivariate Wold representation of the form

$$\Delta \mathbf{x}_t = \boldsymbol{C}(B)\boldsymbol{\varepsilon}_t$$

where $\boldsymbol{C}(0) = \mathbf{I}$ and $\boldsymbol{\varepsilon}_t$ is a zero mean white noise vector. The moving average polynomial $\boldsymbol{C}(B)$ can always be expressed as

$$\boldsymbol{C}(B) = \boldsymbol{C}(1) + (1 - B)\boldsymbol{C}^*(B)$$

and if $\boldsymbol{C}(B)$ is of finite order then so will be $\boldsymbol{C}^*(B)$. With this representation Engle and Granger proved and stated the

Granger Representation Theorem

If the $N \times 1$ vector $\mathbf{x}_t$ is co-integrated with $d = b = 1$ and has co-integrating rank r, then

(1) $\boldsymbol{C}(1)$ *is of rank* $N - r$;
(2) *There exists a vector ARMA representation*

$$\boldsymbol{A}(B)\mathbf{x}_t = d(B)\boldsymbol{\varepsilon}_t \tag{10.27}$$

with the properties that $\boldsymbol{A}(1)$ *has rank* r, $d(B)$ *is a scalar lag polynomial with* $d(1)$ *finite, and* $\boldsymbol{A}(0) = \mathbf{I}$. *When* $d(B) = 1$, *this is a vector autoregression.*
(3) *There exist* $N \times r$ *matrices* $\boldsymbol{A}, \boldsymbol{\Gamma}$ *of rank* r *such that*

$$\begin{aligned} \mathbf{A}'\boldsymbol{C}(1) &= \mathbf{0} \\ \boldsymbol{C}(1)\boldsymbol{\Gamma} &= \mathbf{0} \\ \boldsymbol{A}(1) &= \boldsymbol{\Gamma}\mathbf{A}' \end{aligned}$$

(4) *There exists an error correction representation with* $\mathbf{z}_t = \mathbf{A}'\mathbf{x}_t$, *an* $r \times 1$ *vector of stationary random variables:*

$$\boldsymbol{A}^*(B)\Delta \mathbf{x}_t = -\boldsymbol{\Gamma}\mathbf{z}_{t-1} + d(B)\boldsymbol{\varepsilon}_t \tag{10.28}$$

with $\boldsymbol{A}^*(0) = \mathbf{I}$.
(5) *The vector* $\mathbf{z}_t$ *is given by*

$$\begin{aligned} \mathbf{z}_t &= \boldsymbol{K}(B)\boldsymbol{\varepsilon}_t \\ \Delta \mathbf{z}_t &= -\boldsymbol{\alpha}'\boldsymbol{\gamma}\,\mathbf{z}_{t-1} + \boldsymbol{J}(B)\boldsymbol{\varepsilon}_t \end{aligned}$$

where $\boldsymbol{K}(B) = \mathbf{A}'\boldsymbol{C}^*(B)$ *is an* $r \times N$ *matrix of lag polynomials with all elements of* $\boldsymbol{K}(1)$ *finite with rank r, and* $det(\mathbf{A}'\boldsymbol{\Gamma}) > 0$.

(6) *If a finite vector autoregressive representation is possible, it will have the form given by* (10.27) *and* (10.28) *above with* $d(B) = 1$ *and both* $\boldsymbol{A}(B)$ *and* $\boldsymbol{A}^*(B)$ *as matrices of finite polynomials.*

Engle and Granger pointed out that, when $d(B) = 1$, (10.27) and (10.28) were akin to standard VARs (cf. **§8.19**) for $\mathbf{x}_t$ and $\Delta\mathbf{x}_t$ respectively. However, there are some crucial differences. In (10.27) co-integration imposes constraints on $\boldsymbol{A}(1)$ such that it has reduced rank and hence is singular, making estimation of the unrestricted VAR inefficient and analyses that make use of the moving average representation treacherous, while the lagged levels embodied in $\mathbf{z}_t$ imply that a VAR fitted to $\Delta\mathbf{x}_t$ will be mis-specified as the error correction term will have been omitted. From Engle and Granger's Lemma 1 it follows that $\boldsymbol{A}(B) = \text{adj}\,\boldsymbol{C}(B)/\Delta^{r-1}$ and $d(B) = \det \boldsymbol{C}(B)/\Delta^r$.

The matrices $\mathbf{A}$ and $\boldsymbol{\Gamma}$ are not uniquely defined since, if $\boldsymbol{\Theta}$ is an $r \times r$ matrix of full rank, $\boldsymbol{\Gamma}$ can be replaced by $\boldsymbol{\Gamma\Theta}$ and $\mathbf{A}$ by $\boldsymbol{\Theta}^{-1}\mathbf{A}'$ and the equation in part (3) of the Theorem will still hold. As an illustration, consider $N = 3$ and $r = 2$, so that $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and there are a pair of $I(0)$ error corrections which can be written as

$$z_t(\boldsymbol{\alpha}_1) = \alpha_{11}x_{1t} + \alpha_{12}x_{2t} + \alpha_{13}x_{3t}$$
$$z_t(\boldsymbol{\alpha}_2) = \alpha_{21}x_{1t} + \alpha_{22}x_{2t} + \alpha_{23}x_{3t}$$

As any linear combination of a pair of $I(0)$ variables will also be $I(0)$, then

$$z_t(\lambda) = (1 - \lambda)z_t(\boldsymbol{\alpha}_1) + \lambda z_t(\boldsymbol{\alpha}_2)$$

will be $I(0)$ and the equilibrium relations $\boldsymbol{\alpha}_1'\mathbf{x}_t = \boldsymbol{\alpha}_2'\mathbf{x}_t = 0$ will not be uniquely identified and the error-correction representation (10.28) cannot strictly be interpreted as 'correcting' for deviations from a particular pair of equilibrium relationships. The only invariant relationship is the line in (x_1, x_2, x_3) space defined by $z_t(\boldsymbol{\alpha}_1) = z_t(\boldsymbol{\alpha}_2) = 0$ or, equivalently, by $z_t(\lambda_1) = z_t(\lambda_2) = 0$ for any $\lambda_1 \neq \lambda_2$. This is termed the 'equilibrium sub-space' and the error-correction representation can then be interpreted as Δx_t being influenced by the distance the system is from the equilibrium sub-space. For general N and r the equilibrium sub-space will be a hyperplane of dimension $N - r$.

For the $N = 3$ and $r = 2$ case, the λ's can be chosen to give $z_{1t} = \alpha_1 x_{1t} + \alpha_2 x_{2t}$ and $z_{2t} = \alpha_3 x_{1t} + \alpha_4 x_{3t}$, say, and this seems to provide a natural way of 'normalising' the co-integrating relationships. For more general N and r the number of possible normalising combinations increases and alternative identifying conditions become possible.

10.31 Engle and Granger next focused attention on the estimation of co-integrated systems, noting that the error correction form (10.28) appears to be the most convenient (particularly if $d(B) = 1$, so that no moving average terms are involved). However, since the error-correction term is, essentially, $-\boldsymbol{\Gamma}\boldsymbol{\Lambda}'\mathbf{x}_{t-1}$, there are cross-equation restrictions, involving the coefficients in the co-integrating vector, that need to be taken into account, so that maximum likelihood estimation would require an iterative procedure.

Engle and Granger thus proposed a 'two-step' estimator for the case when $r = 1$, so that there is a single co-integrating vector. The first step is to estimate by least squares the static regression $y_t = \beta_0 + \boldsymbol{\beta}'\mathbf{x}^*_{t-1} + e_t$, where the partition $\mathbf{x}_t = (y_t, \mathbf{x}^*_{t-1})$ has been made (essentially the (non-unique) co-integrating vector $\boldsymbol{\alpha}$ has been normalized). This is known as the 'co-integrating regression', from which the estimate of the co-integrating vector is obtained as $\hat{\boldsymbol{\alpha}} = (1, -\hat{\boldsymbol{\beta}})$. Under the $I(1)$ assumption the co-integrating regression will be spurious, in the sense of **§§10.17–10.20**, but Stock (1987) established the consistency of $\hat{\boldsymbol{\alpha}}$ under co-integration, with convergence to probability limits being very rapid (convergence is at a rate of the sample size T, rather than the rate $\sqrt{T}$ associated with stationary processes: this is known as the 'super-consistency' theorem of co-integrating regressions). This estimate can then be substituted into (10.28), which then becomes linear in its unknown parameters and can then also be estimated by least squares. Engle and Granger then proved the following theorem:

> *The two-step estimator of a single equation of an error-correction system with one co-integrating vector, obtained by taking the estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ from the static regression in place of the true value for estimation of the error correction form at a second stage, will have the same limiting distribution as the maximum likelihood estimator using the true value of $\boldsymbol{\alpha}$. Least-squares standard errors in the second stage will provide consistent estimates of the true standard errors.*

They provided the following simple example to illustrate this approach. Suppose that $N = 2$ and the components of $\mathbf{x}_t$ are jointly generated

according to the following model:

$$\begin{aligned} x_{1t} + \beta x_{2t} &= u_{1t}, \qquad u_{1t} = u_{1,t-1} + \varepsilon_{1t} \\ x_{1t} + \alpha x_{2t} &= u_{2t} \qquad u_{2t} = \rho u_{2,t-1} + \varepsilon_{2t} \qquad |\rho| < 1 \end{aligned} \tag{10.29}$$

where ε_{1t} and ε_{2t} are possibly correlated white noises. The parameters α and β are clearly unidentified and the reduced form of the system is, with $\alpha \neq \beta$,

$$\begin{aligned} x_{1t} &= \alpha(\alpha - \beta)^{-1} u_{1t} - \beta(\alpha - \beta)^{-1} u_{2t} \\ x_{2t} &= -(\alpha - \beta)^{-1} u_{1t} + (\alpha - \beta)^{-1} u_{2t} \end{aligned}$$

Since both x_{1t} and x_{2t} depend linearly on u_{1t}, which is a random walk, they must be $I(1)$. However, $x_{1t} + \alpha x_{2t}$ must be $I(0)$ because u_{2t} is stationary and hence x_{1t} and x_{2t} are $CI(1, 1)$.

Regressing x_{1t} on x_{2t} will produce an excellent estimate of α because all linear combinations of x_{1t} and x_{2t}, except that defined by the co-integrating regression, will have infinite variance. The error-correction representation is obtained from the VAR representation

$$\begin{aligned} \Delta x_{1t} + \beta \Delta x_{2t} &= \varepsilon_{1t} \\ \Delta x_{1t} + \alpha \Delta x_{2t} &= \varepsilon_{2t} - (1 - \rho) x_{1,t-1} - (1 - \rho) x_{2,t-1} \end{aligned}$$

as

$$\begin{aligned} \Delta x_{1t} &= \beta\delta \, z_{t-1} + \eta_{1t} \\ \Delta x_{2t} &= -\delta \, z_{t-1} + \eta_{2t} \end{aligned} \tag{10.30}$$

where $\delta = (1 - \rho)/(\alpha - \beta)$, $\eta_{1t} = (\alpha - \beta)^{-1}(\alpha\varepsilon_{1t} - \beta\varepsilon_{2t})$ and $\eta_{2t} = (\alpha - \beta)^{-1}$ $(\varepsilon_{2t} - \varepsilon_{1t})$. From this representation it is seen that $\delta \neq 0$ if and only if $\rho \neq 1$, but $\rho = 1$ is exactly the condition that makes both u_{1t} and u_{2t} random walks and leads to a non-co-integrated system, i.e., if $\rho = 1$ the levels variables vanish in the VAR and the error correction reduces to that restricted form of the VAR.

Testing for co-integration

10.32 Engle and Granger's next task was to provide tests for co-integration. As they pointed out, the setup is nonstandard and closely related to tests for unit roots in observed series, as initially formulated

by Fuller (1976) and Dickey and Fuller (1979, 1981) and subsequently developed by many others (see Patterson, 2011, for a recent exposition and survey of this enormous literature).

To appreciate the problems inherent in testing for co-integration, consider again the simple model (10.29). The null hypothesis is taken to be that of no co-integration, or $\rho = 1$. If α is known then a test of this null hypothesis could be constructed as a Dickey–Fuller type unit root test by taking $z_t = x_{1t} + \alpha x_{2t}$ as the observed series, which has a unit root under the null. Although the distribution of the test statistic, the t-ratio on the slope coefficient of the regression of Δz_t on z_{t-1} (known as the Dickey–Fuller regression), is nonstandard, critical values are available via simulation. However, when α is unknown it must be estimated from the co-integrating regression but, if $\rho = 1$ is true, α is not identified, which has the implication that only if the series are co-integrated can α be estimated from the co-integrating regression. Yet a test must be based upon the distribution of a statistic when the null of no co-integration is true. Since OLS estimation seeks that α which minimizes the residual variance and is therefore most likely to be stationary, the usual critical values of the Dickey–Fuller test will reject the null too often when an estimated α is used to construct z_t.

To attack this problem, Engle and Granger proposed a range of test statistics for testing the null of non-co-integration against the alternative of co-integration. The basic setup is that the data are generated by (10.30) when $\rho = 1$ and so $\delta = 0$. This 'first-order' system is then extended to a stationary linear system in the Δx's, so that the null is defined over a full set of stationary autoregressive and moving average coefficients, leading to 'augmented' tests analogous to the augmented Dickey–Fuller univariate tests.

Using arguments concerning test similarity and the results of size and power simulations, Engle and Granger recommended that, if the first-order system was tenable, then the preferred test was to estimate the co-integrating regression and compare the Durbin–Watson *dw* statistic to a critical value of 0.386 for a 5% test: if the *dw* exceeds this critical value then the null should be rejected in favour of co-integration, utilising the results of Bhargava (1986) and Sargan and Bhargava (1983). However, a second test, that of comparing the Dickey–Fuller unit root test of the residuals of the co-integrating regression to a 5% critical value of -3.37 and rejecting the null if the statistic was less than this value, was almost as good. If the system was of higher order, Engle and Granger recommended using the augmented form of the Dickey–Fuller test on the co-integrating regression residuals.

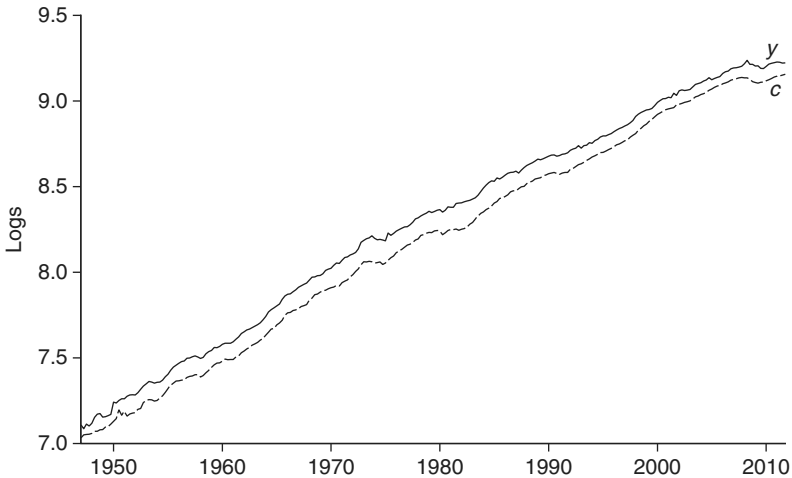


Figure 10.8 Real private consumption (c) and real personal disposable income (y) for the USA from 1947I to 2011IV

10.33 To illustrate these tests, two of Engle and Granger's examples are updated. Figure 10.8 shows plots of the logarithms of real private consumption (c) and real personal disposable income (y) for the USA from 1947I to 2011IV and the distinctive 'common trend' in the two variables is, at the very least, indicative of co-integration potentially existing between them. Standard univariate unit root tests confirmed that both series are $I(1)$ and a series of regressions are reported in Table 10.5 that establish, first, that there is evidence of co-integration between the two variables at approximately the 10% significance level, the critical value for this level being -2.91. For example, the co-integrating regression of c on y is estimated to be $c_t = -0.17 + 1.01y_t + z_t$ with $R^2 = 0.999$ and a t-statistic of 444 on $\hat{\alpha}$. As $dw = 0.21$ the 'co-integrating regression dw' test does not reject the null hypothesis of no co-integration but this statistic is only appropriate for a first order system, which from the regressions reported in Table 10.5 does not appear to be the case here. An augmented Dickey–Fuller regression on the co-integrating residuals produces a t-ratio of -2.93 on z_{t-1}, which is just significant at the 10% level, and a slightly stronger result is obtained from regressing y on c.

The remaining regressions in Table 10.5 build error-correction models for the two variables. Interestingly, there is only weak evidence that a lagged error correction appears in the equation for Δc, as it has an accompanying t-statistic of just 1.2. The lagged error correction is more

Table 10.5 Regressions of consumption and income for the USA, 1948I to 2011IV. *ec* and *ey* are the residuals from the regressions of *c* on *y* and *y* on *c* respectively. Absolute t-ratios are shown in parentheses; $\hat{\sigma}$ is the regression standard error.

Dep. Var	c	Δec	Δec	Δc	Δc
y	1.01 (444)				
$c(-1)$				−0.030 (1.3)	
$y(-1)$				0.028 (1.3)	
$ec(-1)$		−0.07 (2.7)	−0.08 (2.9)		−0.03 (1.2)
$\Delta c(-1)$				−0.00 (0.0)	
$\Delta c(-2)$				0.35 (4.9)	0.36 (5.7)
$\Delta c(-3)$				0.11 (1.5)	
$\Delta c(-4)$				−0.10 (1.5)	−0.13 (2.1)
$\Delta y(-1)$				0.12 (2.0)	0.14 (2.6)
$\Delta y(-2)$				0.03 (0.5)	
$\Delta y(-3)$				−0.17 (2.9)	−0.12 (2.3)
$\Delta y(-4)$				−0.06 (1.1)	
$\Delta ec(-1)$		−0.28 (4.5)	−0.28 (4.8)		
$\Delta ec(-2)$		−0.02 (0.2)			
$\Delta ec(-3)$		−0.08 (1.3)			
$\Delta ec(-4)$		−0.23 (3.8)	−0.20 (3.6)		
Constant	−0.17 (9.1)			0.013 (1.6)	0.006 (7.5)
$\hat{\sigma}$	0.02352	0.00944	0.00943	0.00770	0.00774
dw	0.21	2.0	2.0	2.0	2.0
Dep. Var	y	Δey	Δey	Δy	Δy
c	0.99 (444)				
$c(-1)$				−0.04 (1.6)	
$y(-1)$				0.05 (1.7)	
$ey(-1)$		−0.08 (2.8)	−0.08 (3.0)		−0.05 (2.0)
$\Delta c(-1)$				0.29 (3.6)	0.25 (3.3)
$\Delta c(-2)$				0.13 (1.6)	
$\Delta c(-3)$				0.28 (3.3)	0.25 (3.0)
$\Delta c(-4)$				0.19 (3.4)	0.19 (2.4)
$\Delta y(-1)$				−0.13 (1.8)	
$\Delta y(-2)$				−0.10 (1.5)	
$\Delta y(-3)$				−0.22 (3.2)	−0.16 (2.4)
$\Delta y(-4)$				−0.25 (3.9)	−0.21 (3.3)
$\Delta ey(-1)$		−0.28 (4.5)	−0.28 (4.8)		
$\Delta ey(-2)$		−0.02 (0.2)			
$\Delta ey(-3)$		−0.08 (1.3)			
$\Delta ey(-4)$		−0.23 (3.8)	−0.20 (3.6)		
Constant	0.18 (9.8)			0.040 (4.2)	0.006 (5.3)
$\hat{\sigma}$	0.02332	0.00935	0.00934	0.00907	0.00925
dw	0.21	2.0	2.0	2.1	2.2

significant in the equation for Δy, which suggests that consumption may be weakly exogenous even though the variables are co-integrated, which is the converse of what Engle and Granger found in their example, although they used levels of per capita consumption and income over a much shorter sample period ending in 1981.

The co-integrating regressions may be interpreted as providing a long-run equilibrium in which $c_t = k + y_t$ and the plot of the logarithm of the consumption-income ratio, $c - y$, is shown in Figure 10.9. This ratio takes long swings around its mean value but its stationarity enables the equilibrium condition $c_t - y_t = -0.102$, or a long-run average propensity to consume of $\exp(-0.102) = 0.903$, to assert itself in the long-run.

The second example looks for co-integration between short and long interest rates, as suggested by the efficient markets hypothesis of the term structure of interest rates. Figure 10.10 shows monthly observations from January 1953 to December 2011 on US three month Treasury Bill rates (r_t) and 20 year bond yields (R_t).

Both interest rates are found to be (driftless) $I(1)$ processes and, while there is some indication from Figure 10.10 that they are bound together over the long run, there have been several episodes when the 'spread' between the rates has been rather large for a considerable period of time, particularly since the financial crisis of 2008, after which the Federal Reserve Board has kept short rates at almost zero. This example would thus appear to be a challenging examination of co-integration.

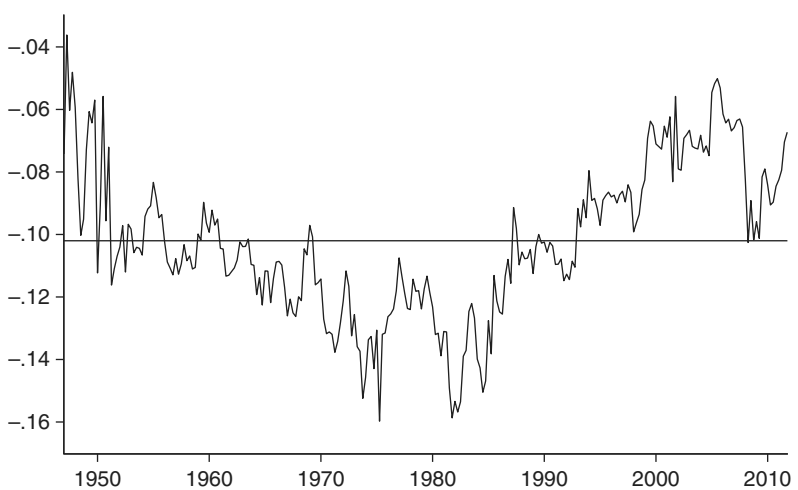


Figure 10.9 Logarithms of the consumption-income ratio, $c - y$

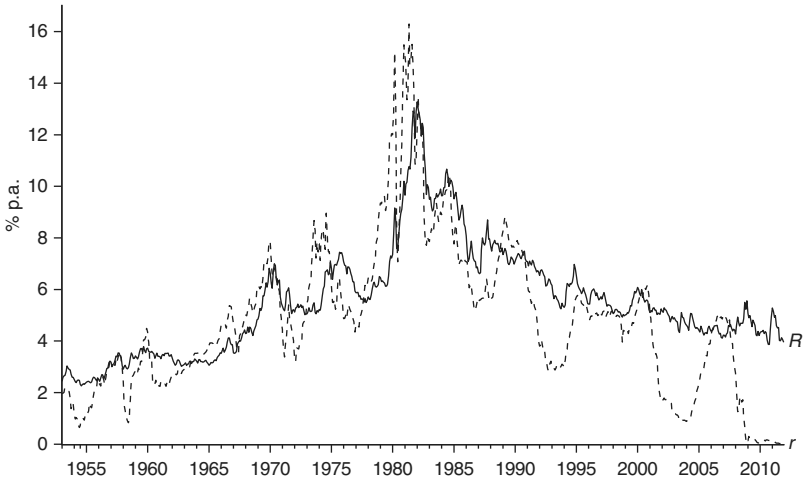


Figure 10.10 US short (r) and long (R) interest rates: January 1953 to December 2011

The co-integrating regression with the long rate as the dependent variable is estimated to be $R_t = 2.80 + 0.567r_t + z_t$ with $R^2 = 0.63$ and a t-statistic of 35 on $\hat{\alpha}$. The *dw* statistic is only 0.04 but a unit root test on z_t, using nine lagged differences, produces a test statistic of -3.39, which is clearly significant at the 5% level, the critical value being -3.17. From the reverse regression a unit root test on the error correction produces a statistic of -3.75, again with nine lagged differences. It therefore does appear that long and short interest rates are co-integrated and, once again, the *dw* test gives a false indication of non-co-integration in a system that is of a much higher order than one.

Generalizations and extensions of co-integration

10.34 In a companion piece to Engle and Granger (1987), Granger (1986) developed some extensions to the basic concept of co-integration. The Representation Theorem focuses on the 'typical' case when $d = b = 1$, but for any values of these parameters the error-correction model becomes

$$\boldsymbol{A}^*(B)\Delta^d \mathbf{x}_t = -\boldsymbol{\Gamma}(1 - \Delta^b)\Delta^{d-b}\mathbf{z}_t + d(B)\boldsymbol{\varepsilon}_t \tag{10.31}$$

Although $\mathbf{z}_t$ appears in this model, Granger noted that $1 - (1 - B)^b$, when expanded in powers of B, has no term in B^0 and so only lagged $\mathbf{z}$ actually

occur on the right-hand side of (10.31). Again, every term in the model is $I(0)$ when co-integration is present and Granger noted that it is also possible to have fractional differencing here as well, leading to the possibility of *fractional co-integration*.

Granger then considered a more general concept of co-integration. Suppose $\boldsymbol{\alpha}(B)$ is an $N \times 1$ vector of functions of B such that each component, say $\alpha_j(B)$, has the property that $\alpha_j(1) \neq 0$. If $\mathbf{x}_t$ is a vector of $I(d)$ processes such that $z_t = \boldsymbol{\alpha}'(B)\mathbf{x}_t \sim I(d-b)$ then $\mathbf{x}_t$ may again be called co-integrated. If a co-integrating vector $\boldsymbol{\alpha}$, as previously defined, occurs then there may be many $\boldsymbol{\alpha}(B)$ that also co-integrate, thus losing the property of uniqueness but possibly gaining extra flexibility. For example, suppose $N = 2$ and $\boldsymbol{\alpha}' = (1, \alpha)$: α will be unique if it does not depend on B so that $r = 1$. However, there may also exist another co-integrating vector of a quite different form, this being

$$\boldsymbol{\alpha}'(B) = (1 + \alpha'\Delta^{-1}, \alpha\alpha'\Delta^{-1})$$

This would occur in the following situation. Since the error correction $z_t = x_{1t} + \alpha x_{2t}$ will be $I(0)$, it follows that $y_t = \sum_{j=0}^{t} z_{t-j} = \Delta^{-1} z_t$ must be $I(1)$: x_{1t} and x_{2t} are then said to be *multi-cointegrated* if y_t and x_{1t} are co-integrated; if they are then y_t and x_{2t} will also be co-integrated. If this is the case then $w_t = x_{1t} + \alpha' y_t \sim I(0)$, from which it follows that

$$\begin{aligned} w_t &= x_{1t} + \alpha'\Delta^{-1} z_t = x_{1t} + \alpha'\Delta^{-1}(x_{1t} + \alpha x_{2t}) \\ &= (1 + \alpha'\Delta^{-1})x_{1t} + \alpha\alpha'\Delta^{-1} x_{2t} \\ &= \boldsymbol{\alpha}'(B)\mathbf{x}_t \end{aligned}$$

The concept of multi-cointegration was analysed and further developed by Granger and Lee (1989, 1990), where applications to production, sales and inventory relationships were provided.

10.35 Granger (1986) also considered relaxing the linearity with time-invariant parameters assumption implicit in the development so far. He began by considering the *time-varying parameter* (TVP) process

$$x_t = \beta(t) x_{t-1} + \varepsilon_t$$

where $\beta(t)$ is a deterministic function of time such that $|\beta(t)| < 1$. This will have a time-varying spectrum that is always finite and positive and will be called a TVP-$I(0)$ process. If the change in x_t is TVP-$I(0)$,

then x_t itself will be TVP-$I(1)$. For a vector process $\mathbf{x}_t$ that is TVP-$I(d)$ and has no deterministic components, there will exist a generalized Wold representation $\Delta^d \mathbf{x}_t = \boldsymbol{C}_t(B)\boldsymbol{\varepsilon}_t$. If $\boldsymbol{C}_t(1)$ has reduced rank $N-1$ for all t, then there will exist $N \times 1$ vectors $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\gamma}(t)$ such that

$$\begin{aligned} \boldsymbol{\alpha}'(t)\boldsymbol{C}_t(1) &= 0 \\ \boldsymbol{C}_t(1)\boldsymbol{\gamma}(t) &= 0 \end{aligned}$$

and the TVP equilibrium error process will be $z_t = \boldsymbol{\alpha}'(t)\mathbf{x}_t$. The corresponding error-correction model will be as (10.28) but with $\boldsymbol{A}^*(B)$, $\boldsymbol{\Gamma}$ and $d(B)$ all being functions of time.

Another route by which non-linearity may appear is through the error correction itself. In the model (10.28) z_{t-1} appears linearly so that changes in the variables are related to the lagged error correction, whatever its size. More realistic behaviour might be to ignore small equilibrium errors but to react substantially to large errors, thus suggesting a non-linear relationship. An error-correction model that captures this idea would be, say,

$$\begin{aligned} \Delta x_t &= f_1(z_{t-1}) + \text{lagged}(\Delta x_t, \Delta y_t) + \varepsilon_{1t} \\ \Delta y_t &= f_2(z_{t-1}) + \text{lagged}(\Delta x_t, \Delta y_t) + \varepsilon_{2t} \end{aligned}$$

Granger argued that generally z_t and $f(z_t)$ would be integrated of the same order so that a finding of co-integration between two variables might suggest that a non-linear error-correction model was a possibility.

Granger's bequest of co-integration to time series econometrics

10.36 As we remarked earlier, Engle and Granger (1987) has since become the most heavily cited paper in time series econometrics and it is interesting to ask why that is. Clearly, part of the answer must be that its success was, to a large extent, attributable to the concept of co-integration itself, which combines both elegance and usefulness to become readily applicable to many areas of economic modelling. Boswijk, Franses and van Dijk (2010) also advance several external factors for why co-integration appeared just at the right time. Essentially, theoretical and applied econometricians alike needed such a concept, as both large-scale macroeconomic systems and univariate models, such as ARIMA processes, were thought to be deficient in forecasting ability and

economic substance respectively. Co-integration was able to bridge the gap between them, especially as there was a sense of urgency in finding appropriate statistical tools to analyse trending economic data. It was also no doubt helped by the increased availability of long enough samples of data to allow long-run equilibrium relationships to be explored and by the rapid expansion of computing power and the dissemination of software that enabled many researchers to apply co-integration methods.

In the next few sections we briefly discuss some of the extensions to the co-integration framework that have been developed as a result of Granger's introduction of the concept. Little detail is given as much of this is now standard material in advanced econometric textbooks, thus confirming the importance of Granger's contribution of co-integration to time series econometrics.

10.37 Within the single co-integrating relationship framework, two major extensions have been undertaken. Although superconsistent, the least squares estimate of the co-integrating vector was shown to be adversely affected by second-order biases, which result in the asymptotic distributions of the estimators being biased and non-normal. Phillips and Hansen (1990) thus proposed *fully modified estimation*, in which the least squares estimate is modified by subtracting an estimate of the bias. A second approach, originally due to Saikkonen (1991) and Phillips and Loretan (1991), is to augment the static co-integrating regression by including leads and lags of $\Delta \mathbf{x}_t$, which counteract simultaneity bias, and also to include lags of Δy_t or the error correction z_t to deal with issues of autocorrelation.

With regard to testing for a single co-integrating vector, much more extensive tables of critical values of the test for a unit root in the co-integrating regression residuals were obtained using response surface methodology (see MacKinnon, 1991, 1996), and these quickly became available in most econometric software packages. Tests have also been derived using the error-correction representation, with the coefficient on the error-correction term being tested for significance, although the distribution of the statistic is again nonstandard (see, for example, Ericsson and MacKinnon, 2002).

10.38 The interest in co-integrated systems really started to take off when dealing with the general case of r co-integrating vectors. Johansen (1988a, 1988b, 1991) examined the mathematical structure of multivariate error-correction models and showed that ML estimation could be recast as an exercise in *reduced rank regression*. Within this framework,

tests of the value of r could be constructed using likelihood ratio principles: see Johansen (1995a) for extended development and discussion. Johansen also analysed the role of deterministic terms, such as time trends, in the error-correction model and their influence on hypothesis testing (Johansen, 1994).

The identification of co-integrated systems was considered by Phillips (1991) and Pesaran and Shin (2002), while co-integration in $I(2)$ systems was analysed in Johansen (1995b, 1997). Fractional co-integration was investigated by Baillie and Bollerslev (1994) and Cheung and Lai (1993), while seasonal co-integration was discussed by Hylleberg et al. (1990). The links between co-integration and structural change are surveyed in Perron (2006). Recent surveys of many of the issues mentioned in this section are Johansen (2006), Juselius (2006, 2009) and Gil-Alana and Hualde (2009).[8]

10.39 Granger's subsequent work in co-integration was often on non-linear extensions. Granger and Hallman (1991a) considered the representation of non-linear co-integration as a bivariate 'attractor' between variables that are individually EMM but have an SMM non-linear combination (recall the definitions of these concepts in **§10.15**). They suggested that non-linear equilibrium relationships could emerge between, say, prices of commodities traded at spatially separated markets due to the existence of varying marginal costs and profits. Granger and Hallman (1991b) found that variables are, in general, not co-integrated with non-linear transformations of themselves, but the same transformation applied to a pair of co-integrated series can result in co-integration between the transformed series.

Granger and Swanson (1996) suggested asymmetric error-correction models in which positive and negative errors have different effects, while Siklos and Granger (1997) argued that co-integrating relationships may switch according to a policy regime, proposing the concept of temporal co-integration to allow variables to be co-integrated in one regime and non-co-integrated in another. Granger and Yoon (2002) considered 'hidden co-integration', where co-integrated variables respond only to certain kinds of shocks, say positive or negative. Granger and Hyung (2006) developed an innovative regime-switching non-linear co-integration process using '*M*-*M*' models. Here x_t and y_t vary according to a switching regime process that allows a mixture of integration and co-integration. In each step a *max* or *min* operator is used to choose between integration (for example, $x_{t+1} = x_t + \varepsilon_t$) or co-integration (for example, $x_{t+1} = y_t + \varepsilon_t$) for each variable. Although in simple cases

M-M processes imply linear co-integrating relationships, they always have threshold-type non-linear error-correction relationships. Threshold effects in multivariate error-correction models are discussed more generally by Gonzalo and Pitarakis (2006).

Another area of co-integration that Granger worked in was that of interpreting such systems in terms of their underlying, but unobserved, components: typically the permanent factors that capture the long-memory or low-frequency variability in the observed series and the transitory factors that explain the shorter memory or high-frequency variation. Gonzalo and Granger (1995) proposed a decomposition having two important characteristics: first, both components are a function only of the current values of the series and, second, innovations in the persistent component are uncorrelated with the innovations in the transitory component. Granger and Haldrup (1997) took up the issue of how these components could be estimated in large systems, investigating whether the components might be computed separately for groups of series, thus avoiding having to model the entire system.

10.40 Regarding co-integration as that property of a set of time series which cancels out the common 'permanent' component that drives them in the long run leads to the 'common stochastic trends' interpretation of the concept (see, for example, Stock and Watson, 1988). Such was the impact of co-integration that Granger was awarded the Sveriges Riksbank Prize in Economic Science in memory of Alfred Nobel, jointly with Robert Engle, in October 2003 in 'recognition of his achievements in developing methods of analysing economic time series with common trends (co-integration)'. Hendry (2004) provided a formal appreciation of this award to Granger, whose own acceptance lecture was published as Granger (2004). In a memorial to Granger, Hendry's (2010) concluding paragraph, although couched in terms of Granger's contribution to econometrics, is nevertheless a fitting epitaph to a great time series analyst.[9]

> Clive's contributions have combined to implement his long-term research agenda of improving the quality of econometric model building by a better match with empirical evidence. His research has shaped the agenda of many econometricians, and given rise to a large number of applications, from sun-spot activity and land use in the Brazilian Amazon, through gold, silver and stock market prices, and eighteenth century wheat prices, to electricity demand, exchange rates, volatility clustering, and yield curves. His unravelling of cointegration and

common trends and their properties was a major development, buttressed by many later important insights. He was a major generator of new ideas, yet always in the context of solving a real problem, not just for its own sake: the wealth of further advances and applications of his ideas bear witness to his fecundity. Sir Clive Granger's was one of the most successful research programmes in the history of econometrics as his total citations of more than 45000 corroborates, and will be a lasting contribution to our discipline.

11
The End of the Affair?

In praise of British pragmatism

11.1 An overriding theme running through this book is one of British pragmatism, in that time and time again method and theory have been developed from attempts to solve practical problems.[1,2] Yule's interest in analysing time series stemmed, at least in part, from his desire to understand the nature of the nonsense correlation problem which bedevilled early empirical work, while Kendall's research was prompted by his work at the Ministry of Agriculture on analysing detrended and oscillatory agricultural time series and later on practical forecasting issues emanating from his consultancy contracts. Durbin had long associations with governmental statistical agencies from which his research on seasonal adjustment and regression stability over time was a natural development. Jenkins first encountered spectral analysis whilst tackling aircraft design problems at the Royal Aircraft Establishment and later developed his own consultancy company specializing in complete forecasting and decision systems for industry and government. Box cut his statistical teeth on wartime chemical experiments and later worked on control problems for ICI, maintaining his interest in this area throughout his career and setting up the Centre for Quality and Productivity Improvement at Madison. Box and Jenkins' joint research on forecasting and control was aimed at providing a practical method for understanding and solving such problems with observed time series. Granger's research, across numerous areas, was often sparked by the desire to understand and solve real, practical problems in time series.

Two quotes from opposite ends of our story reflect these preoccupations with practical problem solving rather than abstract technical theorizing. Maurice Kendall (1952, page 158), in his obituary of Yule,

remarked that he had 'the legitimate scepticism of a practical statistician for the monstrous regiment of mathematicians', while David Hendry (2010, page 168) said of Granger that 'he was a major generator of new ideas, yet always in the context of solving a real problem'.

11.2 But with the passing of Clive Granger and with George Box and James Durbin in their early nineties and late eighties respectively, it is apposite to ask whether the British love of pragmatism in time series analysis might now be fading away. I ask this because academic journals seem nowadays to be full of theoretical extensions of co-integration and volatility models and of ever deeper analyses of non-linearity, but few of these papers attempt to provide any serious practical applications.

Alongside co-integration, volatility modelling has engendered enormous interest since the original publication of Engle (1982) looking at the variability of inflation, undoubtedly a consequence of the attraction of empirically analysing financial markets, with their massive data sets just waiting to be attacked by the huge computing power now available, but there does not appear to be have been many major conceptual breakthroughs since those of Granger and Engle in the mid-1980s.[3]

A future for British time series analysis?

11.3 Will we see the likes of these statistical giants again? From a purely British perspective, it is difficult not to be pessimistic, even though there are pockets of excellence, notably the 'Nottingham group' of David Harvey, Steven Leybourne and Robert Taylor associated with the Granger Centre for Time Series Econometrics. Two factors conspire in this, I feel. Fewer students are being trained in statistics and econometrics in Britain and those that are taking postgraduate degrees are invariably not British and tend to concentrate on the minutiae of theoretical 'tweaking' of assumptions and models. This is compounded by the perceptions of academics working in Britain concerning the research assessment exercises undertaken over the last twenty years: the latest, now known as the Research Excellence Framework (REF), being due in 2014. It is felt by many that only articles published in a small range of journals will score highly in these assessments and that those journals prefer theoretical to applied research: hence the incentive structure by which academics perceive that the only way to secure tenure and then promotion is to play the game and concentrate their research efforts on technical but relatively minor advances that are easier to publish, rather than on engaging with

practical problems that require deep understanding of the context, data and robustness of techniques and models.[4]

But this is important. As greater amounts of data are made available, and computing requirements become ever cheaper and powerful, it surely becomes equally important to have methods available that can detect relationships existing both within and between observed time series. Having time series analysts trained in *both* theory and applied aspects of the subject must certainly be a key aim of future education and training in this increasingly useful and important area of statistics, with its obvious applications to the areas of economics, finance and meteorology – areas, it could be argued, that are key to the well-being of civilisation today and in the future.

Brown and Kass (2009, page 109), in their commentary on the current state of statistical training, concluded that 'many of today's big challenges throughout society are tackled by large teams, and these teams are in desperate need of statistical thinking at the very top levels of management. We suggest that a way forward begins with a focus on the fundamental goals of training, combined with a broad vision of the discipline of statistics.' As with the wider subject of statistics, we feel that such a view also pertains to the current state of time series analysis as well and we would like to think that these suggestions will take hold in the training of British time series analysts, although it is difficult, in the present environment, to be overly optimistic that this will be the case.

Notes

2 Yule: The Time–Correlation Problem, Nonsense Correlations, Periodicity and Autoregressions

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 Student's t-distribution. For further biographical details see Pearson (1950, 1990).
2. 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)).
3. 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 2.6.
4. The calculations required to construct Figure 2.7 are outlined in Yule (1926, page 56).
5. For a derivation of this result in the more general context of calculating 'intraclass' correlation, see Yule and Kendall (1950, §§11.40–11.41).
6. A small Gauss program was written to simulate Yule's sampling procedure and to compute the results shown in Table 2.1 and later in Tables 2.4 and 2.7. Necessarily, the results differ numerically from those of Yule because of the sampling process.
7. 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 2.6, which retains just two to maintain consistency with earlier tables.
8. The index numbers themselves are used, rather than the smoothed Index of Fluctuations, which are often analyzed (see Mills, 2011a, §§3.8–3.9). 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.
9. Yule (1927, pages 284–6) discussed further features of the sunspot numbers and their related disturbances estimated from equation (2.19). These features do not seem to appear in the extended series and are therefore not discussed here.
10. Indeed, the unusual behavior of sunspots over the last decades of the twentieth century has been the subject of great interest: see, for example, Solanki et al. (2004).

3 Kendall: Generalizations and Extensions of Stationary Autoregressive Models

1. 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'.
2. The use of the term autocorrelation follows Wold, 1938, with 'serial correlation' being reserved for the sample counterpart, r_k. Note also the simplification of the notation used by Kendall from that employed in Chapter 2.
3. Hall (1925) had earlier 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 **§6.9**), by some four decades.
4. A recent application of the approach is to be found in Mills (2007), where it is used to obtain recent trends in temperatures. An extension was suggested by Quenouille (1949), 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!
5. The data are provided in Table 1 of Coen et al. (1969). We have used EViews 6 to compute this regression, which leads to some very minor differences in the estimates as reported in Coen et al.'s Table 2 equation (7).
6. The estimates differ somewhat from Coen et al. (1969, Table 3, equation (10)) as some of the early observations on the lagged regressors are not provided there, so that the regression is here estimated over a slightly truncated sample period.

4 Durbin: Inference, Estimation, Seasonal Adjustment and Structural Modelling

1. 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.
2. Durbin recounts that the government of the time was becoming increasingly worried about the rising unemployment figures. The Prime Minister, Harold Wilson, 'had worked as an economic statistician in the government service during the war, and he was really rather good at interpretation of numerical data … and … was very interested in looking at the figures himself. He got the idea … that maybe the reason why the unemployment series appeared to

be behaving in a somewhat strange way was due to the seasonal adjustment procedure that was being used … and asked the CSO to look into this. … It turned out that Wilson was right and there was something wrong with the seasonal adjustment. I think it is remarkable that a point like this should be spotted by a prime minister' (Phillips, 1988, pages 140–1). To be fair, Wilson had a long-standing interest in economics and statistics, having been a lecturer in economic history at New College, Oxford (from the age of 21) and a Research Fellow at University College, as well as holding a variety of government posts that dealt with economic data.

3. The paper represented the work of a team of statisticians at the Research Branch of the CSO, with the authors being responsible for the supervision of the work and the writing of the paper. The Bureau of the Census X-11 seasonal adjustment package has been used extensively by statistical agencies across the world for adjusting economic time series. Mills (2011a, §§14.7–14.10) discusses its development in some detail. The Henderson moving average is based on the requirement that it should follow a cubic polynomial trend without distortion. The filter weights are conveniently derived in Kenny and Durbin (1982, Appendix) and also in Mills (2011a, §10.6).
4. Extensions to autoregressive models were later provided by Kulperger (1985) and Horváth (1993) and then to ARMA processes by Bai (1994).
5. 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 (2011, Appendix A) provides a simple vector-matrix derivation of the RLS algorithm.
6. 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. Durbin was aware of this historical link, for he remarked that it was 'of great interest to note that the Kalman approach to time series modelling was anticipated in a remarkable way for a particular problem by … Theile in 1880, as has been pointed out by Lauritzen (1981)' (Durbin, 1984, page 170).
7. The wearing of seatbelts by front seat occupants of cars and light goods vehicles had been made compulsory in the UK on 31 January 1983 for an experimental period of three years with the intention that Parliament would consider extending the legislation before the expiry of this period. The Department of Transport undertook to monitor the effect of the law on road casualties and, as part of this monitoring exercise, invited Durbin and Harvey to conduct an independent technical assessment of the statistical evidence.
8. This 'opportunity for public discussion' was facilitated by reading the paper at a meeting of the RSS. The resulting discussion, along with the author's rejoinder, was then published along with the paper. While we do not refer to the discussion here, it does make for highly entertaining and, in a couple of places, rather astonishing, reading, with non-statisticians being somewhat perplexed with the findings, to say the least! More generally, a number of Durbin's papers published in RSS journals are accompanied with discussants

remarks and author rejoinders and these typically make fascinating, if often tangential, reading.
9. Only Durbin and Koopman's classical solution to the estimation problem is discussed here, although they also provide formulae for undertaking Bayesian inference (see Durbin and Koopman, 2000).

5 Jenkins: Inference in Autoregressive Models and the Development of Spectral Analysis

1. The name white noise was coined by physicists and engineers because of its resemblance to the optical spectrum of white light, which consists of very narrow lines close together.

6 Box and Jenkins: Time Series Analysis, Forecasting and Control

1. Following Fisher, several others had considered differencing as a means of inducing stationarity (recall **§2.3–2.6**), most notably Tintner (1940), in his advocacy of the variate differencing method, and Yaglom (1955). 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.
2. 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).
3. 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.
4. 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.

7 Box and Jenkins: Modelling Seasonal Time Series and Transfer Function Analysis

1. Estimation of the model in EViews 6 obtained the estimates $\hat{\theta} = 0.403$, $\hat{\Theta} = 0.636$ and $\hat{\sigma}_a^2 = 1.332 \times 10^{-3}$.
2. Distributed lags had also made a considerable impact in econometrics through the work of, most notably, Koyck (1954), Jorgensen (1963) and Almon (1965). Further developments in cross-correlation and multiple time series modeling made during the 1940s and 1950s are discussed in Mills (2011a, Chapter 12).
3. Transforming to white noise was also advocated some two decades earlier by Orcutt and James (1948) in a static regression setting.
4. Box and Jenkins (1970, pages 384–6) suggested an alternative procedure for identifying the noise through the prewhitened input and output.

8 Box and Jenkins: Developments post-1970

1. Again estimated using EViews 6, so that there are minor differences to the estimates reported by Box and Newbold (1971).
2. The hypothesised change in the level of the series is now referred to as an *intervention*, a term attributed by Box and Tiao to Glass (1972), which led to the phrase *intervention analysis* to describe the modelling of such impacts.
3. Box and MacGregor (see also Box, Jenkins and MacGregor, 1974) were typically concerned with situations arising with process industries data where feedback control is being applied, so that the set-up is one of *closed-loop* operation, in which the feedback control scheme is deterministic, rather than the transfer function assumption of *open-loop* operation. The discussion here will concentrate on the more general framework in which both output and input have noise components.
4. Relaxation of the full rank assumption is considered in Granger's Representation Theorem of co-integration: see **§10.30**.

9 Granger: Spectral Analysis, Causality, Forecasting, Model Interpretation and Non-linearity

1. Granger's published output contains almost 300 items. It has to be said that, on close inspection, one is struck by the number of typographical errors that remain in, particularly, the journal articles, both in the text and in the mathematics. This sometimes becomes a distraction from the highly novel and innovative ideas that Granger is typically presenting in his published research.
2. Although this section focuses on the analysis of the cross-spectrum between two stationary series and its interpretation in terms of the concept of 'causality', Granger also published several other papers on spectral analysis, both on theoretical issues (Granger, 1966; Granger and Hughes, 1968) and on applied applications (Granger and Morgenstern, 1963; Godfrey, Granger and Morgenstern, 1964; Granger and Rees, 1968).
3. Granger fully recognized that a precursor of his causality framework had been proposed by Norbert Wiener, referencing Wiener (1956) from the very

beginning (see Granger and Hatanaka, 1964, page 114) and preferring to use the term Wiener-Granger causality.
4. Granger returned to the topic of generalized cost functions some thirty years later in Granger (1999a), where he developed the theory in much greater depth and generality.
5. Newbold and Granger (1974) also focused attention on comparisons between the individual forecasting methods themselves, so providing the first 'forecasting competition'. Such competitions subsequently became very popular, beginning with Makridakis and Hibon (1979). The next, Makridakis et al. (1982), quickly became known as the 'M-competition', and was followed by the M2-competition (Makridakis et al., 1992), in which the author himself took part (see Mills, 1992), and the M3-competition (Makridakis and Hibon, 2000). Other competitions were undertaken by Meese and Geweke (1984) and Fildes et al. (1998). These competitions spawned a vast number of papers reanalyzing the series contained in them and, although not universally favored (interestingly, Newbold, 1983, was particularly critical, although the 'forecast' contained in the title of his comments on the M2 competition clearly proved to be some way off the mark!), nevertheless provided a focus and much impetus for research and practice in forecasting. Fildes and Ord (2002) provide a detailed survey of them.
6. The use of Box–Cox transformations in ARMA modelling had been the source of some controversy between Chatfield and Prothero (1973) and Box and Jenkins (1973) (recall **§8.1**). For Nelson and Granger (1979, page 63), the 'main reason for using the Box-Cox transformation, according to Box and Jenkins (1973), is to produce improved forecasts'.

10 Granger: Long Memory, Fractional Differencing, Spurious Regressions and Co-integration

1. Joyeux (2010) offers Roselyne Joyeux's reflections on the writing of the joint paper, whose genesis was an invitation from Maurice Priestley to contribute to the first issue of his new *Journal of Time Series Analysis*. It should also be pointed out that, contemporaneously and independently, J.R.M. Hosking, a statistician then working at the Institute of Hydrology in Oxfordshire, England, also developed much of the material on fractional differenced processes: see Hosking (1981, 1982).
2. Granger and Hallman (1991a) referred to this property, perhaps more naturally, as *long memory in mean*. Granger (1995) preferred to use 'extended memory in mean' because 'the term "long memory in mean" … has a technical meaning in the time series literature, applying to processes whose spectrum tends to infinity as frequency goes to zero which essentially need not apply here and is too linear for our purposes' (page 272).
3. Granger recounted that when he presented these simulation results during a seminar presentation at the LSE they were 'met with total disbelief. Their reaction was that we must have gotten the Monte Carlo wrong – we must have done the programming incorrectly' (Phillips, 1997, page 262). Paul Newbold, who had actually done the programming, confirmed this story to me some years ago over a couple of pints in the Nottingham University Staff Club.

4. It is easily shown that this method of generating an ARIMA(0,1,1) process implies that the MA coefficient is forced to be −0.382 using the notation of Table 10.3. The use of 1000 simulations rather than the 100 used by Granger and Newbold obviously produces different results to those contained in their Tables 1 and 2, but the two sets are very similar and lead to identical conclusions.
5. It might be thought that these findings are simply a consequence of using a rather small sample by current standards. In fact, with five regressors the rejection rate increases to 99.2 per cent and then to 100 per cent as the sample size is increased to 100 and then to 1000. These simulation findings were later placed and explained within a formal theoretical framework by Phillips (1986).
6. These different interpretations and implications of differencing had, in fact, been discovered half a century earlier by Bradford Smith (1926) in a remarkably prescient article that quickly disappeared almost without trace until being rediscovered by Mills (2011b): see also Mills (2011a, §12.4).
7. In Phillips (1997, page 25) Granger recalls discussing the issue with Hendry: 'he was saying that he had a case where he had two $I(1)$ variables, but their difference was $I(0)$, and I said that is not possible, speaking as a theorist. He said he thought it was. So I went away to prove that I was right, and I managed to prove that he was right. Once I realized that this was possible, then I immediately saw how it was related to the formulation of an error correction model and their balancing. So, in a sense, all the main results of cointegration came together within a few minutes. I mean, without any proof, at least not any deep proof, I just sort of saw what was needed. … Then I had to go away and prove it. That was another thing. But I could see immediately what the main results were going to be.' See also Granger (2010b) for further recollections of this episode.
8. An interesting development was that of Bewley et al. (1994), who established that there were close links between Johansen's reduced rank vector autoregression estimator and Box and Tiao's (1977) canonical decomposition of a VAR discussed in **§§8.19–8.25**.
9. The previous year, after I, as the then head of the Department of Economics, had made the proposal, Clive Granger was awarded an Honorary Doctorate by Loughborough University. I had the honour of making the oration:

   Public Orator, Professor Terence Mills, presented the Honorary Graduand at the Degree Congregation held on the afternoon of Friday 12 July 2002.

   Chancellor,

   Clive Granger was born in Swansea, but completed his high school education at West Bridgford Grammar School, some 12 miles north of Loughborough on the southern outskirts of Nottingham. He then went to Nottingham University, becoming one of the initial intake into the first-ever joint degree in economics and mathematics. On graduating in 1955, Clive stayed on at Nottingham, becoming a lecturer in statistics in the Mathematics Department in 1956, publishing his first academic paper, 'A statistical model for sunspot activity', in the *Astrophysical Journal* in 1957, and obtaining a PhD in statistics in 1959.

   In the early 1960s Clive obtained a Harkness Fellowship and visited Princeton, working on spectral techniques with such famous scholars as John Tukey

and Oscar Morgenstern. Although he came back to Nottingham to become a Professor of Economics and Statistics, this visit to the United States had whetted his appetite for life in an American university. Clive eventually returned in 1974 to a professorship in the Economics Department at the University of California at San Diego, where he has remained and has been instrumental in building up the econometrics section to become one of the finest in the world.

From that early visit to Princeton, Clive Granger has been one of the most influential scholars in time series econometrics. His writings encompass all of the major developments over the last 40 years, and he is personally responsible for some of the most exciting ideas and methods of analysis that have occurred during this time. Indeed, it is now virtually impossible to do empirical work in time series econometrics without using some of his methods or being influenced by his ideas. I am thinking here particularly of the concepts of 'Granger-causality' and co-integration, which our final year students studying econometrics now employ as a matter of course in their project work. However, his research on spurious regression, long memory, and all aspects of forecasting has also had a profound and lasting influence. Most scholars would deem it the accomplishment of a lifetime if their work were to have the impact of a single one of these contributions. To have had repeated instances of such extraordinarily influential research is surely testimony to Clive Granger's special talent as a researcher and writer.

One of the most defining characteristics of Clive's work is his concern for the empirical relevance of his ideas. Another hallmark is the accessibility of his work, which stems from his unusually rich capacity to write highly readable papers and books, many of which have gone on to become citation classics. These successes in communication show us the vital role that good writing plays in the transmission of scientific knowledge. To me, it is no coincidence that the much improved performance of the Bank of England and the Treasury in monitoring and forecasting the UK economy has come about since Clive's ideas and research techniques have been promulgated by his many 'disciples' to recent generations of economics graduates, for it is they who form the research teams in these organisations.

Clive Granger's research has been an inspiration to all time series econometricians, and has been recognised internationally with many awards. Currently he is President of the Western Economic Association and has just become a Distinguished Fellow of the American Economic Association.

It is thus with great pleasure and honour, Chancellor, that I present Clive Granger to you and the University for the Degree of Doctor of Science, *honoris causa*.

11 The End of the Affair?

1. I use 'pragmatism' here for both its vernacular meaning of indicating a 'practical, matter-of-fact way of solving problems' and as a philosophical tradition that is centred on the linking of practice and theory, describing a process whereby theory is extracted from practice and then applied back to practice. Pragmatism as a philosophical movement began in the United States during the 1870s and is most closely associated with Charles Sanders Peirce

and William James (see Haack and Lane, 2006). Although often referred to as 'American pragmatism', it was heavily influenced by Charles Darwin and the earlier 'British empiricists' Locke and Hume (although 'British sceptical realists' might be a better term: see Buckle, 1999). Box (1984) emphasises the importance of theory-practice interaction using many examples from the development of statistical thinking.

2. 'Statistical pragmatism' has recently been proposed by Kass (2011) as a foundation for an eclectic statistical inference that goes beyond narrow frequentist and Bayesian positions and emphasises the 'identification of models with data' and 'the assumptions that connect statistical models to observed data', recognising 'that all forms of statistical inference make assumptions, assumptions that can only be tested very crudely and can almost never be verified'. As Box (1979) memorably stated, 'all models are wrong, but some are useful' and I think that this sums up the approach of our 'British pragmatists' admirably.
3. Interestingly, and as Engle (2004) recounts, the ARCH model of volatility, which led to Engle's joint Nobel Prize with Granger in 2003, was invented while Engle was on sabbatical at the LSE in 1979. Indeed, the term which produced the acronym, *autoregressive conditional heteroskedasticity*, was coined by David Hendry. Some of the major theoretical developments in co-integration and unit roots have been made by Peter Phillips, a New Zealander whose postgraduate work was undertaken at the LSE and whose first two academic appointments were in the UK at Essex and Birmingham.
4. It is certainly true that a book like this would not be recognised in the REF but then I'm far too long in the tooth to be unduly bothered by such things!

References

Abraham, B. and Box, G.E.P. (1978). 'Deterministic and forecast-adaptive time-dependent models'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 27, 120–30.

Abraham, B. and Box, G.E.P. (1979). 'Bayesian analysis of some outlier problems in time series'. *Biometrika* 66, 229–36.

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.

Almon, S. (1965). 'The distributed lag between capital appropriations and expenditures'. *Econometrica* 33, 178–96.

Anderson, R.L. (1942). 'Distribution of the serial correlation coefficient'. *Annals of Mathematical Statistics* 13, 1–13.

Anderson, T.W. (1975). 'Maximum likelihood estimation of parameters of autoregressive processes with moving average residuals and other covariance matrices with linear structures'. *Annals of Statistics* 3, 1283–304.

Anderson, T.W. (1977). 'Estimation for autoregressive moving average models in time and frequency domains'. *Annals of Statistics* 5, 842–65.

Ansley, C.F. (1979). 'An algorithm for the exact likelihood of a mixed autoregressive moving average process'. *Biometrika* 66, 59–65.

Ashley, R., Granger, C.W.J. and Schmalensee, R. (1980). 'Advertising and aggregate consumption: an analysis of causality'. *Econometrica* 48, 1149–67.

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.

Bai, J. (1994). 'Least squares estimation of a shift in linear processes'. *Journal of Time Series Analysis* 15, 435–72.

Baillie, R.T. (1996). 'Long memory processes and fractional integration in econometrics'. *Journal of Econometrics* 73, 5–59.

Baillie, R.T. and Bollerslev, T. (1994). 'Cointegration, fractional cointegration and exchange rate dynamics'. *Journal of Finance* 49, 737–45.

Banerjee, A., Dolado, J., Galbraith, J.W. and Hendry, D.F. (1993). *Co-integration, Error-correction, and the Econometric Analysis of Non-stationary Data*. Oxford: Oxford University Press.

Banerjee, A. and Urga, G. (2005). 'Modelling structural breaks, long memory and stock market volatility'. *Journal of Econometrics* 129, 1–34.

Barnard, G.A. (1963). 'New methods of quality control'. *Journal of the Royal Statistical Society, Series A* 126, 255–58.

Barnard, G.A. (1997). 'Kendall, Maurice George', in N.L. Johnson and S. Kotz (eds), *Leading Personalities in Statistical Sciences*. New York: Wiley, pp. 130–2.

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.
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, 686–7.
Bartlett, M.S. (1950). 'Periodogram analysis and continuous spectra'. *Biometrika* 37, 1–16.
Bartlett, M.S. (1955). *Stochastic Processes*. Cambridge: Cambridge University Press.
Bates, J.M. and Granger, C.W.J. (1969). 'The combination of forecasts'. *Operational Research Quarterly* 20, 451–68.
Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). 'The discarding of variables in multivariate analysis'. *Biometrika* 54, 357–66.
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.
Berlinger, L.M. (1992). 'Statistics, probability and chaos'. *Statistical Science* 7, 69–90.
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.
Bewley, R., Orden, D., Yang, M. and Fisher, L.A. (1994). 'Comparison of Box-Tiao and Johansen canonical estimators of cointegrating vectors in VEC(1) models'. *Journal of Econometrics* 64, 3–27.
Bhargava, A. (1986). 'On the theory of testing for unit roots in observed time series'. *Review of Economic Studies* 53, 369–84.
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
Boswijk, H.P., Franses, P.H. and van Dijk, D. (2010). 'Cointegration in a historical perspective'. *Journal of Econometrics* 158, 156–9.
Box, G.E.P. (1979). 'Robustness in the strategy of scientific model building', in R.L. Laumer and G.N. Wilkinson (eds), *Robustness in Statistics*. New York: Academic Press, pp. 201–36.
Box, G.E.P. (1983a). 'G.M. Jenkins, 1933–1982'. *Journal of the Royal Statistical Society, Series A* 146, 205–6.
Box, G.E.P. (1983b). 'Gwilym Jenkins, experimental design and time series'. *Questió* 7, 515–25.
Box, G.E.P. (1984). 'The importance of practice in the development of statistics'. *Technometrics* 26, 1–8.
Box, G.E.P. (1989). 'The R.A. Fisher Memorial Lecture, 1988. Quality improvement: an expanding domain for the application of scientific method'. *Philosophical Transactions of the Royal Society of London, Series A* 327, 617–30.
Box, G.E.P. and Cox, D.R. (1964). 'An analysis of transformations'. *Journal of the Royal Statistical Society, Series B* 26, 211–43.
Box, G.E.P., Hillmer, S.C. and Tiao, G.C. (1979). 'Analysis and modeling of seasonal time series', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*. Washington, DC: US Department of 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., Jenkins, G.M. and MacGregor, J.F. (1974). 'Some recent advances in forecasting and control: Part II'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 23, 158–79.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008). *Time Series Analysis: Forecasting and Control*, 4th edition. New York: Wiley.

Box, G.E.P. and Kramer, T. (1992). 'Statistical process monitoring and feedback adjustment: a discussion'. *Technometrics* 34, 251–85.

Box, G.E.P. and Luceño, A. (1995). 'Discrete proportional-integral control with constrained adjustment'. *Journal of the Royal Statistical Society, Series D (The Statistician)* 44, 479–95.

Box, G.E.P. and MacGregor, J.F. (1974). 'The analysis of closed-loop dynamic-stochastic systems'. *Technometrics* 16, 391–8.

Box, G.E.P. and MacGregor, J.F. (1976). 'Parameter estimation with closed-loop operating data'. *Technometrics* 18, 371–80.

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. (1965). 'A change in level of a non-stationary time series'. *Biometrika* 52, 181–92.

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.

Box, G.E.P. and Tiao, G.C. (1976). 'Comparison of forecast and actuality'. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 25, 195–200.

Box, G.E.P. and Tiao, G.C. (1977). 'A canonical analysis of multiple time series'. *Biometrika* 64, 355–65.

Bray, J. (1971). 'Dynamic equations for economic forecasting with the GDP–unemployment relation and the growth of G.D.P. as an example'. *Journal of the Royal Statistical Society, Series A* 134, 167–227.

Brown, E.N. and Kass, R.E. (2009). 'What is statistics?' *American Statistician* 63, 105–10.

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.L. and Durbin, J. (1968). 'Methods of investigating whether a regression relationship is constant over time'. *Selected Statistical Papers, European Meeting*. Mathematical Centre Tracts No. 26, Amsterdam.
Brown, R.L., Durbin, J. and Evans, J.M. (1975). 'Techniques of testing for the constancy of regression relationships over time'. *Journal of the Royal Statistical Society, Series B* 37, 141–92.
Buckle, S. (1999). 'British skeptical realism: a fresh look at the British tradition'. *European Journal of Philosophy* 7, 1–29.
Buiter, W.H. (1984). 'Granger-causality and policy effectiveness'. *Economica* 51, 151–62.
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.
Cave, B.M. and Pearson, K. (1914). 'Numerical illustrations of the variate-difference correlation method'. *Biometrika* 10, 340–55.
Chamberlain, G. (1982). 'The general equivalence of Granger and Sims causality'. *Econometrica* 50, 569–81.
Chang, I., Tiao, G.C. and Chen, C. (1988). 'Estimation of time series parameters in the presence of outliers'. *Technometrics* 30, 193–204.
Chatfield, C. and Prothero, D.L. (1973). 'Box–Jenkins seasonal forecasting: problems in a case-study'. *Journal of the Royal Statistical Society, Series A* 136, 295–336.
Chatterjee, S. and Yilmaz, M.R. (1992). 'Chaos, fractals and statistics'. *Statistical Science* 7, 49–68.
Cheung, Y.W. and Lai, K. (1993). 'A fractional cointegration analysis of purchasing power parity'. *Journal of Business and Economic Statistics* 11, 103–12.
Clemen, R.T. (1989). 'Combining forecasts: a review and annotated bibliography'. *International Journal of Forecasting* 5, 559–81.
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.
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.
Coen, P.J., Gomme, E.D. and Kendall, M.G. (1969). 'Lagged relationships in economic forecasting'. *Journal of the Royal Statistical Society, Series A* 132, 133–63.
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.
Currie, D. (1981). 'Some long-run features of dynamic time-series models'. *Economic Journal* 91, 704–15.
Daniels, H.E. (1956). 'The approximate distribution of serial correlation coefficients'. *Biometrika* 43, 169–85.

David, H.A. and Fuller, W.A. (2007). 'Sir Maurice Kendall (1907–1983): a centenary appreciation'. *American Statistician* 61, 41–6.
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.
De Groot, M.H. (1987). 'A conversation with George Box'. *Statistical Science* 2, 239–58.
Deutsch, M., Granger, C.W.J. and Teräsvirta, T. (1994). 'The combination of forecasts using changing weights'. *International Journal of Forecasting* 10, 47–57.
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.
Diebold, F.X. and Inoue, A. (2001). 'Long memory and regime switching'. *Journal of Econometrics* 101, 131–59.
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.
Ding, Z. and Granger, C.W.J. (1995). 'Some properties of absolute return, an alternative measure of risk'. *Annales d'Economie et de Statistique* 40, 67–91.
Ding, Z. and Granger, C.W.J. (1996). 'Varieties of long memory models'. *Journal of Econometrics* 73, 61–77.
Ding, Z., Granger, C.W.J. and Engle, R.F. (1993). 'A long memory property of stock market returns and a new model'. *Journal of Empirical Finance* 1, 83–116.
Dittman, I. and Granger, C.W.J. (2002). 'Properties of nonlinear transformations of fractionally integrated processes'. *Journal of Econometrics* 110, 113–33.
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.
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. (1960a). 'The fitting of time-series models'. *Review of the International Statistical Institute* 28, 233–44.
Durbin, J. (1960b). 'Estimation of parameters in time-series regression models'. *Journal of the Royal Statistical Society, Series B* 22, 139–53.

Durbin, J. (1961). 'Efficient fitting of linear models for continuous stationary time-series from discrete data'. *Bulletin of the International Statistical Institute* 33, 273–82.
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: Wiley, pp. 3–16.
Durbin, J. (1969). 'Tests of serial correlation in regression analysis based on the periodogram of least squares residuals'. *Biometrika* 56, 1–15.
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. (1984). 'Present position and potential developments: some personal views. Time series analysis'. *Journal of the Royal Statistical Society, Series A* 147, 161–73.
Durbin, J. (2000). 'The Foreman Lecture: The state space approach to time series analysis and its potential for official statistics'. *Australian and New Zealand Journal of Statistics* 42, 1–23.
Durbin, J. (2004). 'Introduction to state space time series analysis', in A.C. Harvey, S.J. Koopman and N. Shephard (eds), *State Space and Unobserved Component Models*. Cambridge: Cambridge University Press, pp. 3–24.
Durbin, J. and Harvey, A.C. (1985). 'The effects of seat belt legislation on road casualties in Great Britain: Report on Assessment of Statistical Evidence'. Annex to *Compulsory Seat Belt Wearing Report by the Department of Transport*. London: HMSO.
Durbin, J. and Kenny, P.B. (1979). 'Seasonal adjustment when the seasonal component behaves neither purely multiplicatively nor purely additively', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*. Washington, DC: US Department of Commerce, Bureau of the Census, pp. 173–200.
Durbin, J. and Koopman, S.J. (1997). 'Monte Carlo maximum likelihood estimation for non-Gaussian state space models'. *Biometrika* 84, 669–84.
Durbin, J. and Koopman, S.J. (2000). 'Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives'. *Journal of the Royal Statistical Society, Series B* 62, 3–56.
Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
Durbin, J. and Koopman, S.J. (2002). 'A simple and efficient simulation smoother for state space time series analysis'. *Biometrika* 89, 603–15.
Durbin, J. and Murphy, M.J. (1975). 'Seasonal adjustment based on a mixed additive-multiplicative model'. *Journal of the Royal Statistical Society, Series A* 138, 385–410.
Durbin, J. and Quenneville, B. (1997). 'Benchmarking by state space models'. *International Statistical Review* 65, 23–48.
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.
Elderton, E.M. and Pearson, K. (1915). 'Further evidence of natural selection in men'. *Biometrika* 10, 488–506.
Engle, R.F. (1982). 'Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation'. *Econometrica* 50, 987–1008.
Engle, R.F. (2004). 'Risk and volatility: econometric models and financial practice'. *American Economic Review* 94, 405–20.
Engle, R.F. and Granger, C.W.J. (1987). 'Cointegration and error correction: representation, estimation and testing'. *Econometrica* 55, 251–76.
Engle, R.F., Granger, C.W.J. and Kraft, D. (1984). 'Combining competing forecasts of inflation using a bivariate ARCH model'. *Journal of Economic Dynamics and Control* 8, 151–65.
Engle, R.F., Hendry, D.F. and Richard, J.-F. (1983). 'Exogeneity'. *Econometrica* 51, 277–304.
Enns, P.G., Machak, J.A., Spivey, W.A. and Wrobleski, W.J. (1982). 'Forecasting applications of an adaptive multiple exponential smoothing model'. *Management Science* 28, 1035–44.
Epstein, R. (1987). *A History of Econometrics*. Amsterdam: North-Holland.
Ermini, L. and Granger, C.W.J. (1993). 'Some generalizations on the algebra of I(1) processes'. *Journal of Econometrics* 58, 369–84.
Ericsson, N.R. and MacKinnon, J.G. (2002). 'Distributions of error correction tests for cointegration'. *Econometrics Journal* 5, 285–318.
Farebrother, R.W. (2006). 'Early explorations in econometrics', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*. Basingstoke: Palgrave Macmillan, pp. 88–116.
Fildes, R., Hibon, M., Makridakis, S. and Meade, N. (1998). 'Generalising about univariate forecasting methods: further empirical evidence'. *International Journal of Forecasting* 14, 339–58.
Fildes, R. and Ord, K. (2002). 'Forecasting competitions: their role in improving forecasting practice and research', in M.P. Clements and D.F. Hendry (eds), *A Companion to Economic Forecasting*. Oxford: Oxford University Press, pp. 322–53.
Fisher, I. (1925). 'Our unstable dollar and the so-called business cycle'. *Journal of the American Statistical Association* 20, 179–202.
Fisher, R.A. (1929). 'Tests of significance in harmonic analysis'. *Proceedings of the Royal Society of London, Series A* 125, 54–9.
Florens, J.P. and Mouchart, M. (1982). 'A note on noncausality'. *Econometrica* 50, 583–91.
Frängsmyr, T. (2004). *The Nobel Prize 2003*. Stockholm: Nobel Foundation.
Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.
Geweke, J. (1984). 'Inference and causality in economic time series models', in Z. Griliches and M.D. Intriligator (ed.), *Handbook of Econometrics, Volume II*. Amsterdam: North-Holland, pp. 1101–44.
Geweke, J. and Porter-Hudak, S. (1983). 'The estimation and application of long memory time series models'. *Journal of Time Series Analysis* 4, 221–38.

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.

Gilbert, C.L. (1989). 'LSE and the British approach to time series econometrics'. *Oxford Economic Papers* 41, 108–28.

Gilbert, C.L. and Qin, D. (2006). 'The first fifty years of modern econometrics', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 117–55.

Glass, G.V. (1972). 'Estimating the effects of intervention into a nonstationary time series'. *American Educational Research Journal* 9, 463–77.

Godfrey, L.G. (1979). 'Testing the adequacy of a time series model'. *Biometrika* 66, 67–72.

Godfrey, L.G. and Tremayne, A.R. (1988). 'Misspecification tests for time series and their application in econometrics'. *Econometric Reviews* 7, 1–42.

Godfrey, M.D., Granger, C.W.J. and Morgenstern, O. (1964). 'The random-walk hypothesis of stock market behavior'. *Kyklos* 17, 1–29.

Gonzalo, J. and Granger, C.W.J. (1995). 'Estimation of common long-memory components in cointegrated systems'. *Journal of Business and Economic Statistics* 13, 27–35.

Gonzalo, J. and Pitarakis, J.-Y. (2006). 'Threshold effects in multivariate error correction models', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory.* Basingstoke: Palgrave Macmillan, pp. 578–609.

Granger, C.W.J. (1957). 'A statistical model of sunspot activity'. *Astrophysical Journal* 126, 152–8.

Granger, C.W.J. (1959). 'Estimating the probability of flooding on a tidal river'. *Journal of the Institution of Water Engineers* 13, 165–74.

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. (1969a). 'Investigating causal relations by econometric methods and cross-spectral methods'. *Econometrica* 37, 424–38.

Granger, C.W.J. (1969b). 'Prediction with a generalized cost of error function'. *Operational Research Quarterly* 20, 199–207.

Granger, C.W.J. (1979). 'New classes of time series models'. *Journal of the Royal Statistical Society, Series D (The Statistician)* 27, 237–53.

Granger, C.W.J. (1980a). 'Testing for causality: a personal viewpoint'. *Journal of Economic Dynamics and Control* 2, 329–52.

Granger, C.W.J. (1980b). 'Long memory relationships and the aggregation of dynamic models'. *Journal of Econometrics* 14, 227–38.

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. (1983). 'Forecasting white noise', in A. Zellner (ed.), *Proceedings of the Conference on Applied Time Series Analysis of Economic Data*. Washington, DC: US Department of Commerce, Bureau of the Census, pp. 308–14.

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. (1988). 'Some recent developments in a concept of causality'. *Journal of Econometrics* 39, 199–211.
Granger, C.W.J. (1989a). 'Combining forecasts – twenty years later'. *Journal of Forecasting* 8, 167–73.
Granger, C.W.J. (1989b). *Forecasting in Business and Economics*, 2nd edition. San Diego: Academic Press.
Granger, C.W.J. (1991). 'Developments in the nonlinear analysis of economic series'. *Scandinavian Journal of Economics* 93, 263–76.
Granger, C.W.J. (1992a). 'Forecasting stock market prices – lessons for forecasters'. *International Journal of Forecasting* 8, 3–13.
Granger, C.W.J. (1992b). 'Comment'. *Statistical Science* 7, 102–4.
Granger, C.W.J. (1993). 'Strategies for modeling nonlinear time series relationships'. *Economic Record* 60, 233–8.
Granger, C.W.J. (1995). 'Modelling nonlinear relationships between extended-memory variables'. *Econometrica* 63, 265–79.
Granger, C.W.J. (1999a). 'Outline of forecast theory using generalized cost functions'. *Spanish Economic Review* 1, 161–73.
Granger, C.W.J. (1999b). *Empirical Modeling in Economics: Specification and Evaluation*. Cambridge: Cambridge University Press.
Granger, C.W.J. (2004). 'Time series analysis, cointegration and applications'. *American Economic Review* 94, 421–5.
Granger, C.W.J. (2005). 'Modeling, evaluation and methodology in the new century'. *Economic Inquiry* 43, 1–12.
Granger, C.W.J. (2007). 'Forecasting – looking back and forward. Paper to celebrate the 50th anniversary of the Econometrics Institute at the Erasmus University, Rotterdam'. *Journal of Econometrics* 138, 3–13.
Granger, C.W.J. (2008a). 'Causality in economics', in P. Machamel and G. Wolters (eds), *Thinking about Causes: from Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press, pp. 284–96.
Granger, C.W.J. (2008b). 'Non-linear models: where do we go next – time varying parameter-models?' *Studies in Nonlinear Dynamics and Econometrics* 12, 1–9.
Granger, C.W.J. (2010a). 'Curriculum vitae'. *Journal of Financial Econometrics* 8, 244–64.
Granger, C.W.J. (2010b). 'Some thoughts on the development of cointegration'. *Journal of Econometrics* 158, 3–6.
Granger, C.W.J. and Andersen, A.P. (1978a). 'On the invertibility of time series models'. *Stochastic Processes and Their Applications* 8, 87–92.
Granger, C.W.J. and Andersen, A.P. (1978b). *An Introduction to Bilinear Time Series Models*. Gottingen: Vandenhoeck and Ruprecht.
Granger, C.W.J. and Andersen, A.P. (1978c). 'Non-linear time series modeling', in D.F. Findley (ed.), *Applied Time Series Analysis*. San Diego: Academic Press, pp. 25–38.
Granger, C.W.J. and Haldrup, N. (1997). 'Separation in cointegrated systems and persistent-transitory decompositions'. *Oxford Bulletin of Economics and Statistics* 59, 449–64.
Granger, C.W.J. and Hallman, J. (1991a). 'Long memory series with attractors'. *Oxford Bulletin of Economics and Statistics* 53, 11–26.

Granger, C.W.J. and Hallman, J. (1991b). 'Nonlinear transformations of integrated time series'. *Journal of Time Series Analysis* 12, 207–24.
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. (1968). 'Spectral analysis of short series – a simulation study'. *Journal of the Royal Statistical Society, Series A* 131, 83–99.
Granger, C.W.J. and Hyung, N. (2004). 'Occasional structural breaks and long memory with an application to the S& P 500 absolute stock returns'. *Journal of Empirical Finance* 11, 399–421.
Granger, C.W.J. and Hyung, N. (2006). 'Introduction to M-M processes'. *Journal of Econometrics* 130, 143–64.
Granger, C.W.J., Inoue, T. and Morin, N. (1997). 'Non-linear stochastic trends'. *Journal of Econometrics* 81, 65–92.
Granger, C.W.J. and Jeon, Y. (2002). 'The distributional properties of shocks to a fractional I(d) process having a marginal exponential distribution'. *Applied Financial Economics* 11, 469–74.
Granger, C.W.J. and Jeon, Y. (2003a). 'A time-distance criterion for evaluating forecasting models'. *International Journal of Forecasting* 19, 199–215.
Granger, C.W.J. and Jeon, Y. (2003b). 'Comparing forecasts of inflation using time distance'. *International Journal of Forecasting* 19, 339–49.
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 Lee, T.H. (1989). 'Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models'. *Journal of Applied Econometrics* 4, S145–S159.
Granger, C.W.J. and Lee, T.H. (1990). 'Multicointegration', in G.F. Rhodes and T.B. Fomby (eds), *Advances in Econometrics: Cointegration, Spurious Regressions, and Unit Roots*. Greenwich, CT: JAI Press, pp. 71–84.
Granger, C.W.J. and Leybourne, S.J. (2009). 'The research interests of Paul Newbold'. *Econometric Theory* 25, 1460–5.
Granger, C.W.J. and Lin, J.-L. (1994a). 'Forecasting from non-linear models in practice'. *Journal of Forecasting* 13, 1–10.
Granger, C.W.J. and Lin, J.-L. (1994b). 'Using the mutual information coefficient to identify lags in non-linear models'. *Journal of Time Series Analysis* 15, 371–84.
Granger, C.W.J. and Machina, M.J. (2006). 'Forecasting and decision theory', in G. Elliott, C.W.J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting*. Amsterdam: Elsevier, pp. 81–98.
Granger, C.W.J. and Morgenstern, O. (1963). 'Spectral analysis of New York stock market prices'. *Kyklos* 16, 1–27.
Granger, C.W.J. and Morris, M.J. (1976). 'Time series modeling and interpretation'. *Journal of the Royal Statistical Society, Series A* 139, 246–57.
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. (1975). 'Forecasting economic series – the atheist's viewpoint', in G. Renton (ed.), *Modelling the Economy*. London: Heinemann, pp. 135–47.

Granger, C.W.J. and Newbold, P. (1976). 'Forecasting transformed series'. *Journal of the Royal Statistical Society, Series B* 38, 189–203.
Granger, C.W.J. and Newbold, P. (1977). 'The time series approach to econometric model building', in C.A. Sims (ed.), *New Methods in Business Cycle Research*. Minneapolis: Federal Reserve Bank of Minneapolis, pp. 7–22.
Granger, C.W.J. and Newbold, P. (1986). *Forecasting Economic Time Series*, 2nd edition. San Diego: Academic Press.
Granger, C.W.J. and Pesaran, M.H. (2000a). 'A decision theoretic approach to forecast evaluation', in W.-S. Chan, W.K. Li and H. Tong (eds), *Statistics and Finance: an Interface*. London: Imperial College Press, pp. 261–78.
Granger, C.W.J. and Pesaran, M.H. (2000b). 'Economic and statistical measures of forecast accuracy'. *Journal of Forecasting* 19, 537–60.
Granger, C.W.J. and Poon, S. (2003). 'Forecasting volatility in financial markets'. *Journal of Economic Literature* 41, 478–539.
Granger, C.W.J. and Ramanathan, R. (1984). 'Improved methods of combining forecasts'. *Journal of Forecasting* 3, 197–204.
Granger, C.W.J. and Rees, H. (1968). 'Spectral analysis of the term structure of interest rates'. *Review of Economic Studies* 35, 67–76.
Granger, C.W.J. and Sin, C.-Y. (2000). 'Modelling the absolute returns of different stock market indices: exploring the forecastability of an alternative measure of risk'. *Journal of Forecasting* 19, 277–98.
Granger, C.W.J., Spear, S. and Ding, Z. (2000). 'Stylized facts on the temporal and distributional properties of absolute returns: an update', in W.-S. Chan, W.K. Li and H. Tong (eds), *Statistics and Finance: an Interface*. London: Imperial College Press, pp. 97–120.
Granger, C.W.J. and Swanson, N.R. (1996). 'Further developments in the study of cointegrated variables'. *Oxford Bulletin of Economics and Statistics* 58, 374–86.
Granger, C.W.J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
Granger, C.W.J. and Teräsvirta, T. (1999). 'A simple non-linear model with misleading linear properties'. *Economics Letters* 62, 161–5.
Granger, C.W.J., Teräsvirta, T. and Lin, C.F. (1993). 'The power of the neural network linearity test'. *Journal of Time Series Analysis* 14, 209–20.
Granger, C.W.J. and Weiss, A.A. (1983). 'Time series analysis of error correcting models', in S. Karlin, T. Amemiya and L.A. Goodman (eds), *Studies in Econometrics, Time Series and Multivariate Statistics*. New York: Academic Press, pp. 255–78.
Granger, C.W.J., White, H. and Kamstra, M. (1989). 'Interval forecasting: an analysis based upon ARCH-quantile estimates'. *Journal of Econometrics* 40, 87–96.
Granger, C.W.J. and Yoon, G. (2002). 'Hidden co-integration', Economics Working Paper 2002-02, University of California San Diego.
Grenander, U. and Rosenblatt, M. (1953). 'Statistical spectral analysis of time series arising from stationary stochastic processes'. *Annals of Mathematical Statistics* 24, 537–58.
Haack, S. and Lane, R.E. (2006). *Pragmatism, Old & New: Selected Writings*. Amherst, NY: Prometheus Books.
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.
Harrison, P.J. and Stevens, C.F. (1976). 'Bayesian forecasting'. *Journal of the Royal Statistical Society, Series B* 38, 205–47.
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. (1981). *Time Series Models*. Oxford: Phillip Allan.
Harvey, A.C. (1984). 'A unified view of statistical forecasting procedures'. *Journal of Forecasting* 3, 245–75.
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.
Haugh, L.D. and Box, G.E.P. (1977). 'Identification of dynamic regression (distributed lag) models connecting two time series'. *Journal of the American Statistical Association* 72, 121–30.
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: Federal Reserve Bank of Minneapolis, pp. 183–202.
Hendry, D.F. (2004). 'The Nobel Memorial Prize for Clive W.J. Granger'. *Scandinavian Journal of Economics* 106, 187–213.
Hendry, D.F. (2010). 'Professor Sir Clive W.J. Granger and cointegration'. *Journal of Financial Econometrics* 8, 162–8.
Hendry, D.F. and Mizon, G.E. (1999). 'The pervasiveness of Granger causality in econometrics', in R.F. Engle and H. White (eds), *Cointegration, Causality and Forecasting*. Oxford: Oxford University Press, pp. 102–34.
Hendry, D.F. and Morgan, M.S. (1995). *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press.
Hendry, D.F. and von Ungern-Sternberg, T. (1981), 'Liquidity and inflation effects on consumers' behaviour', in A.S. Deaton (ed.), *Essays in the Theory and Measurement of Consumers' Behaviour*. Cambridge: Cambridge University Press, pp. 237–60.
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.
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 Department 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.
Holland, P.W. (1986). 'Statistics and causal inference'. *Journal of the American Statistical Association* 81, 945–60.
Hoover, K.D. (2001). *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K.D. (2008). 'Causality in economics and econometrics', in S.N. Durlauf and L.F. Blume (eds), *The New Palgrave Dictionary of Economics*, 2nd edition. Basingstoke: Palgrave Macmillan.
Hooker, R.H. (1901). 'Correlation of the marriage-rate with trade'. *Journal of the Royal Statistical Society* 64, 485–92.
Horváth, L. (1993). 'Change in autoregressive processes'. *Stochastic Processes and their Applications* 44, 221–42.
Hosking, J.R.M. (1981). 'Fractional differencing'. *Biometrika* 68, 165–76.
Hosking, J.R.M. (1982). 'Some models of persistence in time series', in O.D. Anderson (ed.), *Time Series Analysis: Theory and Practice 1*. Amsterdam: North-Holland, pp. 643–54.
Hylleberg, S., Engle, R.F., Granger, C.W.J. and Yoo, B.S. (1990). 'Seasonal integration and cointegration'. *Journal of Econometrics* 44, 215–38.
Jacobs, R.L., Leamer, E.E. and Ward, M.P. (1979). 'Difficulties in testing for causation'. *Economic Inquiry* 17, 401–13.
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 (1963a). 'Cross-spectral analysis and the estimation of linear open loop transfer functions', in M. Rosenblatt (ed.), *Proceedings of the Symposium on Time Series Analysis*. New York: Wiley, pp. 267–76.
Jenkins, G.M. (1963b). 'An example of the estimation of a linear open loop transfer function'. *Technometrics* 5, 227–45.
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. (1979). *Practical Experiences with Modelling and Forecasting Time Series*. Jersey: GJP.
Jenkins, G.M. (1982). 'Some practical aspects of forecasting in organizations'. *Journal of Forecasting* 1, 3–21.
Jenkins, G.M. and Alavi, A.S. (1981). 'Some aspects of modeling and forecasting multiple time series'. *Journal of Time Series Analysis* 2, 1–47.
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.
Jenkins, G.M. and Youle, P.V. (1971). *Systems Engineering*. London: Everyman's Library.
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. (1991). 'Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models'. *Econometrica* 59, 1551–80.
Johansen, S. (1994). 'The role of the constant and linear terms in cointegration analysis of non-stationary variables'. *Econometric Reviews* 13, 205–29.
Johansen, S. (1995a). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.
Johansen, S. (1995b). 'A statistical analysis of cointegration for $I(2)$ variables'. *Econometric Theory* 11, 25–59.
Johansen, S. (1997). 'Likelihood analysis of the $I(2)$ model'. *Scandinavian Journal of Statistics* 24, 433–62.
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.
Jones, R.H. (1966). 'Exponential smoothing for multivariate time series'. *Journal of the Royal Statistical Society, Series B* 28, 241–51.
de Jong, P. and Shephard, N. (1995). 'The simulation smoother for time series models'. *Biometrika* 82, 339–50.
Jorgenson, D.W. (1963). 'Capital theory and investment behavior'. *American Economic Review* 53, 247–59.
Joyeux, R. (2010). 'Long memory processes; a joint paper with Clive Granger'. *Journal of Financial Econometrics* 8, 184–6.
Juselius, K. (2006). *The Cointegrated VAR Model: Methodology and Applications*. Oxford: Oxford University Press.
Juselius, K. (2009). 'The long swings puzzle: what the data tell when allowed to speak freely', in T.C. Mills and K. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*. Basingstoke: Palgrave Macmillan, pp. 349–84.
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.
Kass, R.E. (2011). 'Statistical inference: the big picture'. *Statistical Science* 26, 1–9.
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. (1943a). 'Oscillatory movements in English agriculture'. *Journal of the Royal Statistical Society* 106, 43–52.
Kendall, M.G. (1943b). *The Advanced Theory of Statistics, Volume I*. London: Griffin.
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. (1973a). 'Techniques in spectral analysis'. *Journal of the Royal Statistical Society, Series D (The Statistician)* 21, 129–31.
Kendall, M.G. (1973b). *Time Series*. London: Griffin.
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.
Klein, J.L. (1997). *Statistical Visions in Time: A History of Time Series Analysis 1662–1938*. Cambridge: Cambridge University Press.
Koopman, S.J. and Durbin, J. (2000). 'Fast filtering and smoothing for multivariate state space models'. *Journal of Time Series Analysis* 21, 281–96.
Koopman, S.J. and Durbin, J. (2003). 'Filtering and smoothing of state vector for diffuse state space models'. *Journal of Time Series Analysis* 25, 85–98.
Koyck, L.M. (1954). *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland.
Kulperger, R.J. (1985). 'On the residuals of autoregressive processes and polynomial regression'. *Stochastic Processes and their Applications* 21, 107–18.
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.
Lawrence, A.J. and Kottegoda, N.T. (1977). 'Stochastic modeling of riverflow time series'. *Journal of the Royal Statistical Society, Series A* 140, 1–47.
Lee, T.-H., White, H. and Granger, C.W.J. (1993). 'Testing for neglected nonlinearity in time series models'. *Journal of Econometrics* 56, 269–90.
Leipnik, R.B. (1947). 'Distribution of the serial correlation coefficient in a circularly correlated universe'. *Annals of Mathematical Statistics* 18, 80–7.
Ljung, G.M. and Box, G.E.P. (1978). 'On a measure of lack of fit in time series'. *Biometrika* 65, 297–303.
Ljung, G.M. and Box, G.E.P. (1979). 'The likelihood function of stationary autoregressive-moving average models'. *Biometrika* 66, 265–70.
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.
Lui, T., Granger, C.W.J. and Heller, W.P. (1992). 'Using the correlation exponent to decide if an economic time series is chaotic'. *Journal of Applied Econometrics* 7, S25–S40.
Macaulay, F.R. (1931). *The Smoothing of Time Series*. New York: NBER.

MacKinnon, J.G. (1991). 'Critical values for cointegration tests', in R.F. Engle and C.W.J. Granger (eds), *Long-run Equilibrium Relationships*. Oxford: Oxford University Press, pp. 267–76.
MacKinnon, J.G. (1996). 'Numerical distribution functions for unit root and cointegration tests'. *Journal of Applied Econometrics* 11, 601–18.
Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982). 'The accuracy of extrapolation (time series) methods: results of a forecasting competition'. *Journal of Forecasting* 1, 111–53.
Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T.C., Ord, J.K. and Simmons, L. (1992). 'The M-2 competition: a real life judgmentally based forecasting study'. *International Journal of Forecasting* 9, 5–22.
Makridakis, S. and Hibon, M. (1979). 'Accuracy of forecasting: an empirical investigation'. *Journal of the Royal Statistical Society, Series A* 142, 97–145.
Makridakis, S. and Hibon, M. (2000). 'The M3-competition: results, conclusion and implications'. *International Journal of Forecasting* 16, 451–76.
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 Van Ness, J.W. (1968). 'Fractional Brownian motions, fractional noises and applications'. *SIAM Review* 10, 422–37.
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.
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.
Meese, R. and Geweke, J. (1984). 'A comparison of autoregressive univariate forecasting procedures for macroeconomic time series'. *Journal of Business and Economic Statistics* 2, 191–200.
Mills, T.C. (1982). 'The use of unobserved component and signal extraction techniques in modeling economic time series'. *Bulletin of Economic Research* 34, 92–108.
Mills, T.C. (1992). 'The M-2 competition: some personal reflections'. *International Journal of Forecasting* 9, 26.
Mills, T.C. (1997). 'Stylized facts on the temporal and distributional properties of daily FT-SE returns'. *Applied Financial Economics* 7, 599–604.
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. (2011a). *The Foundations of Modern Time Series Analysis*. Basingstoke: Palgrave Macmillan.
Mills, T.C. (2011b). 'Bradford Smith: an econometrician decades ahead of his time'. *Oxford Bulletin of Economics and Statistics* 73, 276–85.

Mills, T.C., Tsay, R.S. and Young, P.C. (2011). 'Introduction to special issue commemorating the 50th anniversary of the Kalman filter and 40th anniversary of Box and Jenkins'. *Journal of Forecasting* 30, 1–5.
Mincer, J. and Zarnowitz, V. (1969). 'The evaluation of economic forecasts', in J. Mincer (ed.), *Economic Forecasts and Expectations*. New York: NBER, pp. 1–46.
Mizon, G.E. and Richard, J.F. (1986). 'The encompassing principle and its application to non-nested hypotheses'. *Econometrica* 54, 657–78.
Moran, P.A.P. (1954). 'Some experiments on the prediction of sunspot numbers'. *Journal of the Royal Statistical Society, Series B* 16, 112–17.
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.
Mosedale, T.J., Stephenson, D.B., Collins, M. and Mills, T.C. (2006). 'Granger causality of coupled climate processes: ocean feedback on the North Atlantic Oscillation'. *Journal of Climate* 19, 1182–94.
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.
Nelson, H.L., Jr. and Granger, C.W.J. (1979). 'Experience with using the Box-Cox transformation when forecasting economic time series'. *Journal of Econometrics* 10, 57–69.
Nerlove, M., Grether, D.M. and Carvalho, J.L. (1979). *Analysis of Economic Time Series: A Synthesis*. New York: Academic Press.
Newbold, P. (1974). 'The exact likelihood function for a mixed autoregressive-moving average process'. *Biometrika* 61, 423–6.
Newbold, P. (1983). 'The competition to end all competitions: commentary on the Makridakis time series competition'. *Journal of Forecasting* 2, 276–9.
Newbold, P. and Granger, C.W.J. (1974). 'Experience with forecasting univariate time series and the combination of forecasts (with discussion)'. *Journal of the Royal Statistical Society, Series A* 131–65.
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'. *American Statistician* 38, 36–7.
Parzen, E. (1957). 'On consistent estimates of the spectrum of a stationary series'. *Annals of Mathematical Statistics* 28, 329–48.
Parzen, E. (1961). 'Mathematical considerations in the estimation of spectra'. *Technometrics* 3, 167–90
Patterson, K.D. (2011). *Unit Root Tests in Time Series*. Basingstoke: Palgrave Macmillan.
Pearson, E.S. (1922). 'On the variations in personal equation and the correlation of successive judgments'. *Biometrika* 14, 23–102.
Pearson, E.S. (1950). "Student' as statistician'. *Biometrika* 37, 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. and Elderton, E.M. (1923). 'On the variate difference method'. *Biometrika* 14, 281–310.

Pearson, K. and Rayleigh, Lord (1905). 'The problem of the random walk', *Nature* 72, 294, 318, 342.
Peña, D. (2001). 'George Box: an interview with the International Journal of Forecasting', *International Journal of Forecasting* 17, 1–9.
Peña, D. and Box, G.E.P. (1987). 'Identifying a simplifying structure in time series'. *Journal of the American Statistical Association* 82, 836–43.
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 (1917). 'On the variate difference correlation method and curve-fitting'. *Publications of the American Statistical Association* 15, 602–42.
Pesaran, M.H. and Shin, Y. (2002). 'Long run structural modelling'. *Econometric Reviews* 21, 49–87.
Pesaran, M.H. and Timmermann, A. (1992). 'A simple nonparametric test of predictive performance'. *Journal of Business and Economics Statistics* 10, 461–5.
Phillips, P.C.B. (1986). 'Understanding spurious regressions in econometrics'. *Journal of Econometrics* 33, 311–40.
Phillips, P.C.B. (1988). 'The ET interview: Professor James Durbin', *Econometric Theory* 4, 125–57.
Phillips, P.C.B. (1991). 'Optimal inference in co-integrated systems'. *Econometrica* 59, 282–306.
Phillips, P.C.B. (1997). 'The ET interview: Professor Clive Granger'. *Econometric Theory* 13, 253–303.
Phillips, P.C.B. and Hansen, B.E. (1990). 'Statistical inference in instrumental variables regression with $I(1)$ processes'. *Review of Economic Studies* 57, 99–125.
Phillips, P.C.B. and Loretan, M. (1991). 'Estimating long-run economic equilibria'. *Review of Economic Studies* 58, 407–36.
Pierce, D.A. (1977). 'Relationships – and the 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. (1978). 'Seasonal adjustment when both deterministic and stochastic seasonality are present', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*. Washington, DC: US Department of Commerce, Bureau of the Census, pp. 365–97.
Pierce, D.A. and Haugh, L.D. (1977). 'Causality in temporal systems: characterizations and a survey'. *Journal of Econometrics* 5, 265–93.
Plackett, R.L. (1950). 'Some theorems in least squares'. *Biometrika* 37, 149–57.
Priestley, M.B. (1981). *Spectral Analysis and Time Series*. San Diego: Academic Press.
Qin, D. (1993). *The Formation of Econometrics: a Historical Perspective*. Oxford: Oxford University Press.
Quenouille, M.H. (1947). '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. (1949). 'On a method of trend elimination'. *Biometrika* 36, 75–91.
Quenouille, M.H. (1957). *The Analysis of Multiple Time-Series*. London: Griffin.
Quenouille, M.H. (1958). 'Discrete autoregressive schemes with varying time-intervals'. *Metrika* 1, 21–7.
Robinson, P.M. (1977). 'The estimation of a non-linear moving average model'. *Stochastic Processes and Their Applications* 5, 81–90.

Rudra, A. (1955). 'A method of discrimination in time-series analysis I'. *Sankhyā: The Indian Journal of Statistics* 15, 9–34.
Rydén, T., Teräsvirta, T. and Åsbrik, S. (1998). 'Stylized facts of daily return series and the hidden Markov model'. *Journal of Applied Econometrics* 13, 217–44.
Saikkonen, P. (1991). 'Asymptotically efficient estimation of cointegrating regressions'. *Econometric Theory* 7, 1–21.
Sargan, J.D. (1964). 'Wages and prices in the United Kingdom: a study in econometric methodology', in P.E. Hart, G. Mills and J.K. Whitaker (eds), *Econometric Analysis for National Economic Planning*. London: Butterworths, pp. 25–63.
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.
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.
Schweppe, F.C. (1965). 'Evaluation of likelihood functions for Gaussian signals'. *IEEE Transactions on Information Theory* 11, 61–70.
Seth, A.K. and Edelman, G.M. (2007). 'Distinguishing causal interactions in neural populations'. *Neural Computing* 19, 910–33.
Siklos, P.L. and Granger, C.W.J. (1997). 'Temporary cointegration with an application to interest rate parity'. *Macroeconomic Dynamics* 1, 640–57.
Sims, C.A. (1972). 'Money, income and causality'. *American Economic Review* 62, 540–52.
Sims, C.A. (1977). 'Exogeneity and causal ordering in macroeconomic models', in C.A. Sims (ed.), *New Methods in Business Cycle Research*. Minneapolis: Federal Reserve Bank of Minneapolis, pp. 23–42.
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. (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.
Sowell, F.B. (1992). 'Maximum likelihood estimation of stationary univariate fractionally integrated time series models'. *Journal of Econometrics* 53, 165–88.
Spanos, A. (2006). 'Econometrics in retrospect and prospect', in T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*. Basingstoke: Palgrave Macmillan, pp. 3–58.
Spencer-Smith, J.L. (1947). 'The oscillatory properties of the moving average'. *Journal of the Royal Statistical Society, Series B* 9, 104–13.
Stock, J.H. (1987). 'Asymptotic properties of least squares estimators of cointegrating vectors'. *Econometrica* 55, 1035–56.

Stock, J.H. and Watson, M.W. (1988). 'Variable trends in economic time series'. *Journal of Economic Perspectives* 2, 147–74.
Stuart, A. (1984). 'Sir Maurice Kendall, 1907–1983'. *Journal of the Royal Statistical Society, Series A* 147, 120–2.
Stuart, A. and Kendall, M.G. (1971). *Statistical Papers of George Udny Yule*. London: Griffin.
Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, 6th edition. New York: Wiley-Blackwell.
Student (W.S. Gosset) (1914). 'The elimination of spurious correlation due to positions in time and space'. *Biometrika* 10, 179–80.
Tabery, J.G. (2004). 'The "Evolutionary Synthesis" of George Udny Yule'. *Journal of the History of Biology* 37, 73–101.
Taylor, S.J. (1986). *Modelling Financial Time Series*. Chichester: Wiley.
Teräsvirta, T., Tjostheim, D. and Granger, C.W.J. (2011). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
Theil, H. (1958). *Economic Forecasts and Policy*. Amsterdam: North-Holland.
Theil, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.
Theil, H. and Wage, S. (1964). 'Some observations on adaptive forecasting'. *Management Science* 10, 198–206.
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.
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'. *Sankhyā: The Indian Journal of Statistics, Series B* 35, 268–77.
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.
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.
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. (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. (1931). 'On periodicity in series of related terms'. *Proceedings of the Royal Society of London, Series A* 131, 518–32.
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.
Weiss, G. (1986). 'Random walks', in *Encyclopedia of Statistical Sciences*, *Volume 7*. New York: Wiley, pp. 574–80.

Whittle, P. (1953). 'Estimation and information in stationary time series analysis'. *Arkiv för Matematik* 2, 423–34.
Whittle, P. (1954a). 'Some recent contributions to the theory of stationary processes'. Appendix 2 of Wold (1954).
Whittle, P. (1954b). 'A statistical investigation of sunspot observations with special reference to H. Alfren's sunspot model'. *Astrophysical Journal* 120, 251–60.
Wiener, N. (1956). 'The theory of prediction', in E.F. Beckenback (ed.), *Modern Mathematics for Engineers*. New York: McGraw-Hill, pp. 165–90.
Williams, R.H. (2004). 'George Udny Yule: Statistical Scientist'. *Human Nature Review* 4, 31–7.
Wilson, G.T. (1969). 'Factorization of the generating function of a pure moving average process'. *SIAM Journal of Numerical Analysis* 6, 1–7.
Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell.
Wold, H. (1954). *A Study in the Analysis of Stationary Time Series*, 2nd edition. Stockholm: Almqvist and Wiksell.
Working, H. (1934). 'A random-difference series for use in the analysis of time series'. *Journal of the American Statistical Association* 29, 11–24.
Yaglom, A.M. (1955). 'The correlation theory of processes whose *n*th differences constitute a stationary process'. *Matem. Sbornik* 37, 141–96.
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. (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* 62, 249–95.
Yule, G.U. (1902). 'Mendel's Laws and their probable relations to intra-racial heredity'. *The New Phytologist* 1, 193–207, 222–38.
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. (1910). 'On the interpretation of correlations between indices or ratios'. *Journal of the Royal Statistical Society* 73, 644–7.
Yule, G.U. (1914). 'Fluctuations in sampling in Mendelian ratios'. *Proceedings of the Cambridge Philosophical Society* 17, 425–32.
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. (1944). *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. (1946). 'Cumulative sampling: a speculation as to what happens in copying manuscripts'. *Journal of the Royal Statistical Society* 109, 44–52.
Yule, G.U. and Kendall, M.G. (1950). *An Introduction to the Theory of Statistics*, 14th edition. London: Griffin.
Zellner, A. (1979). 'Causality and econometrics', in K. Brunner and A.H. Meltzer (eds), *Three Aspects of Policy and Policymaking: Knowledge, Data and Institutions*. Carnegie-Rochester Conference Series on Public Policy, 10, 9–54.

Index

Page numbers followed by 'n' denote references to notes.