RAIRO. RECHERCHE OPÉRATIONNELLE

OLIVER D. ANDERSON

Analysing time series for forecasting (a personal view)

RAIRO. Recherche opérationnelle, tome 23, n° 2 (1989), p. 113-150

<http://www.numdam.org/item?id=RO_1989__23_2_113_0>

Recherche opérationnelle/Operations Research
(vol. 23, n° 2, 1989, p. 113 à 150)

ANALYSING TIME SERIES FOR FORECASTING (A PERSONAL VIEW) (*)

by Oliver D. ANDERSON (1)

Abstract. — *This paper discusses the author's views on orthodox time-domain modelling practice, together with suggested modifications and extensions to the now well-established Box-Jenkins methodology for analysing and forecasting time series. We concentrate on improving interpretation of the serial correlation structure for the purpose of enhancing model identification. For instance, by more sensitive discrimination between stationary and nonstationary situations, we may decrease expected forecast error when dealing with long-memory processes.*

Keywords : Autocorrelation, ARIMA models, General unit-circle nonstationarity, Simplifying operators, Wichern behaviour.

Résumé. — *Cet article porte sur les conceptions qu'a son auteur quant aux choix conventionnels de modèles reliés aux temps, ainsi que sur ses suggestions quant aux modifications et aux extensions de la méthodologie désormais éprouvée de Box-Jenkins pour l'analyse des séries chronologiques et leurs prévisions. Nous préconisons en particulier une meilleure interprétation de la structure de corrélation des séries afin de faciliter l'identification du modèle qui convient. Par exemple, nous pouvons réduire l'espérance de l'erreur associée à une prévision dans le cas des processus à mémoire prolongée, en distinguant de façon plus précise les situations stationnaires de celles qui ne le sont pas.*

1. INTRODUCTION

As a time series analyst, I do not regard myself as a statistician. Time series is both narrower in conception and broader in scope than statistics, and I would regard it as a discipline in its own right. Narrower, because we are only concerned with data that develop through time (or along some other dimension) — unlike mainstream statistics, where good design carefully arranges that the data are individualistically independent. Indeed, with time series, we welcome the interdependence of data, which makes sequentially recorded observations interesting and useful for making predictive inferences.

(*) Received in November 1988.
(1) Department of Statistical and Actuarial Sciences, Room 3005 EMSc, University of Western Ontario, London, Canada N6A 5B9.

Recherche opérationnelle/Operations Research, 0399-0559/89/02 113 18/$ 5.80

(It could perhaps even be argued that classical statistics in a sub-discipline of time series analysis, which restricts itself to the study of data that is serially uncorrelated – but we do not wish to become too contentious.)

Broader, as we are really concerned with applications, and these occur in every area where quantitative measurements are made consecutively, and inferences are to be drawn as to what will happen farther down the line. Thus, time series analysis becomes virtually the sole branch of statistics that is relevant to many applied areas in the physical sciences and technology, for instance in oil exploration where the dominant technique is to analyse observer initiated shock waves as time series traces.

Time series tools then dominate the work of scientists using monitoring instruments, either as diagnostic or early warning devices, be they surgeons in hospitals or experts keeping an eye on earth tremors and volcanic activity. And fast, effective, automatic, on-line procedures now promise a rich harvest of microprocessor applications in technology, where a single chip will be able to monitor instruments, processing the information to predict impending problems, and almost instantaneously apply appropriate remedial control to steer the system out of difficulty.

In the social and management sciences, time series techniques are less widespread, because no longer do we have such enormously long runs of data. Rather than thousands or even millions of consecutive observations, we may be lucky to have 30 or 40 which can be rightly regarded as having been more-or-less consistently generated. For, of course, in the world of human interactions, change is the only really constant condition.

Thus the time series analysis of business data is much more difficult than that, say, along a coal-seam. However, the greater intrinsic difficulties do not reduce the relevance of time series thinking when dealing with data in the Management Sciences. And clearly we should not be content with explaining temporally dependent data by merely regressing it against time-which, unfortunately, is what is most generally done in practice.

For, if one wishes to fit a "z on t" regression to data, the implicit assumption is that one has a model of the form

$$z_t = f(t) + \varepsilon_t \tag{1}$$

where the t are precisely and accurately fixed, and the error terms, the ε_t, are independently and identically distributed as normal random variables with zero mean and constant variance. These conditions on the ε_t imply that

the stochastic (statistical) parts of the z_t are not serially related; although, of course, the deterministic components follow the fixed functional form, $f(t)$.

Now, it is true that Econometricians, for instance, have not been slow to recognize this failing of regression models to satisfactorily represent serially correlated data; and their treatment of autocorrelated residuals by successive least squares operations does, to some extent, eliminate the problem (or, as I would prefer to have them say, extracts most of the extra available information). But this is neither the optimal approach, nor a natural one; and, anyhow, is rarely employed by other management scientists.

We therefore advocate that, although difficult to handle due to their inherent shortness, business time series must be treated as temporally interdependent data and analysed accordingly. It is worth investing the effort to master the appropriate methods, because time series are by far the most common form of data found in business, be the series a relevant economic indicator, a record of company sales, or production line figures – say.

In the physical sciences, where external conditions can be controlled, and long runs of data may consequently be realistically considered, two complementary (and mathematically equivalent) approaches are feasible, deriving from respectively what are called the Time and Frequency Domains – although many physical scientists have preferred to work in the latter.

This preference can be attributed to the way most physical scientists and engineers have been trained to think in terms of periodic wave forms; and, historically, frequency domain tools were developed earlier. Even social scientists, in particular economists, used the Frequency Domain and Spectral Analysis. However, the last twenty years has seen a marked swing in favour of Time Domain methods in all areas dealing with shorter time series, where it is believed that spectra are far harder to interpret than the corresponding time domain results (serial correlations).

It is certainly true that the spectrum is difficult to treat for short series, when all sorts of unavoidable aberrations occur (bias and leakage) which make the true message hard to disentangle. But, even as a confirmed time domainer, I am not convinced that these problems are in fact any less present in the time domain. All that we observe is that they are less apparent, so we just do not notice them. This does not mean that they are any the less marked in their consequences, or that our overlooking them will be any the less serious an oversight.

If short series are likened to extreme short-sightedness, then spectral analysis is like wearing spectacles. One is immediately aware that one is stretching the capabilities of simple optics – with various distortions in the image quite

apparent. It is virtually impossible to decide whether a long line is straight or slightly curved; and, by turning one's head, a grid of squares can look like a mesh of rhombuses.

Analogous to Time Domain methods, wearing contact lenses removes many of the visual clues that one is not seeing too well. For instance, one can not look over contact lenses, or past the sides of them. But this lack of reference points, constantly reminding one of one's handicap, does not mean that one actually sees any better through contact lenses. (Although everyone I know, who wears the things, seems to think it does—and that includes some physicists!)

Be that as it may, my experience has mostly been in the time domain; so, for me, there is a considerable advantage in working there. Also, for social and management scientists, the concepts needed to develop ideas in the time domain seem more natural—so we can justify, as opposed to merely rationalise, our decision to only discuss this side of the coin.

The enormous range of specialisations, where time series analysis is a major applications component, makes time series either a multidisciplinary topic—or, as I would have it, a discipline in its own right. And it is hoped that the University of Western Ontario will be the first to formally recognise this, by the creation of a centre of excellence in the subject.

Finally, before we get fully involved with the serious content of this paper, I should perhaps explain how I became a time series analyst. I can assure you I was neither born nor made one, rather I arrived here from apprehension of the alternatives finding me out.

I started as a bridge engineer, but could not quite comprehend how my structures ever stayed up. I thought it best to beat a dignified advance retreat, before one actually fell down. I then worked through a range of disciplines, before hitting on Statistics—with the almost ideal blanket of sample error and random variation to hide one's uncertainties (dare we say incompetence?) behind.

From there, it was only a short step to realising that long-term forecasting was the safe field, with one no longer being around when the predictions proved false. However, this did not seem to provide a meal ticket—people were also wary of paying for information which (even if correct) would be beyond their experience. So, the only way forward was to become involved with producing tools for others to risk in action. (I discovered quite early on in Whitehall, that decision makers there really only needed forecasts that would support the actions that they wanted to take. So, forecasting was not a matter of predicting the future, but rather of reading the current power

struggle correctly. Thus, warnings in the mid-seventies that the UK was heading for three to four million unemployed by the early eighties were dismissed as irrelevant. That sort of scenario did not fit into the future plans on any ambitious official's desk. However, civil servants are conservative, and the name of the game is to play safe and follow the pack. In business, on the other hand, there are people who are prepared to take risks if the expected gain is high enough—and so forecasts can result in substantial action. It is this possibility of influencing policy which gives forecasting its rewards, indeed its only *raison d'être*, so really all predictions should be seen as, say, "Forecasting for Planning". It is only when we are developing tools for our trade, that we ever consider forecasting as an end in itself.)

The rest of this paper falls mainly into three sections, which give the author's views as to: (i) where the area of the time domain, that particularly interests him, is coming from; (ii) where it has currently got to; and (iii) where he would like to see it heading. The acronym, TSA&F, associated with his professional (as opposed to academic) activity, denotes Time Series Analysis and Forecasting.

The way this article is structured thus symbolises an important feature of successful forecasting: we must look back before we can peer forward. To predict the future, one must first adequately analyse and interpret the present and the past.

2. WHERE TSA&F IS COMING FROM

2.1 What Time Series Are

A time series is a sequence of observations recorded at equispaced instants and denoted by, say, $\{z_1, z_2, \ldots, z_n\}$, where n is the number of data points, conveniently referred to as the "length" of the series. For instance, the $z_t (t=1, \ldots, n)$ might be the weekly wage bills for a large company during a calendar year, with $n=52$ in that case.

The series can be considered as a realisation of some underlying stochastic process $\{Z_t: t=1, \ldots, n\}$; and, then, we are evidently interested in the statistical structure of this process, should we wish to obtain forecasts for future values of the series.

2.2. Time Domain Modelling

One way of describing such structure is to obtain a parametric model for the process; and the method of time series analysis, associated with and advocated by Box and Jenkins (1976), first postulates a plausible class of models for initial investigation. It then proceeds to tentatively choose, or "identify", a promising member from this class, whose particular parameters are next efficiently estimated; and, finally, the success of the resulting fit is assessed. The now precisely defined model (identified and estimated) is either accepted by this verification stage, or the diagnostic checks carried out will find it wanting in certain respects and should then suggest a sensible modified identification – after which further estimation and checking takes place; and the cycle of identification, estimation and verification is repeated until a satisfactory fit obtains.

It is important not to overparameterize the model since, although this might improve the goodness of fit for the series history at hand, it is likely to result in the model portraying spurious features of the sampled data, which may detract from the usefulness of the achieved fit. For instance, unnecessarily poor forecasts for future observations on the series are a typical outcome of ignoring this principle of parsimonious parameterization, and then over-mining the data to discover structure in the past record which is purely fortuitous and will not be reflected in the future observations.

Although it is frequently not realised, part of the verification should be to ensure that the fitted model does make sense in the context from which the data were drawn. This model interpretation is sometimes thought of as an extra fourth stage to the Box-Jenkins cycle.

(Of course, this whole cycle of events is directly analogous to the way any pragmatically orientated statistician or management scientist approaches data. And the real impact of Box and Jenkins, in the academic world, is to have put actual data analysis at least on a par with the development of abstract theory, in the eyes of appointments and promotions committees. The mathematical statisticans may be none-too-happy with the lack of rigour which allows data to speak for themselves, but they have been wise enough not to be too openly hostile to the new wave – which really does promise to be the future of statistics. Theory must serve to clarify the obscurities experienced by practitioners, and provide them with effective tools to overcome the difficulties they encounter when analysing the various sorts of data that actually occur.)

2.3. Linear Processes

The univariate models, entertained by Box and Jenkins (1976), are a very general class of linear process which may be taken as being driven by some completely random sequence of unknown "shocks", denoted by $\{A_t\}$ say. These A_t are assumed to be uncorrelated and identically distributed zero-mean random variables, all with the same variance, σ_A^2 say; and $\{A_t\}$ is then referred to as a "white noise" process. It is frequently convenient to assume that the shocks are in fact normally distributed, but this assumption then needs to be justified in applications.

The process of interest, $\{Z_t\}$, is considered to be obtained by applying a linear filter to the shocks $\{A_t\}$, according to

$$Z_t = A_t + \psi_1 A_{t-1} + \psi_2 A_{t-2} + \ldots \tag{2}$$

for suitable choices of $\psi_1, \psi_2, \ldots$. In this representation, each Z_t is taken as being formed from a weighted sum of the current and previous shocks, with psi weights $\psi_0 = 1, \psi_1, \psi_2, \ldots$.

The simplest case is when Z_t is itself completely random, giving for all t

$$Z_t = A_t$$

which is just white noise.

Next, we have the so-called Moving Average models of general order $q \geqq 0$, denoted by MA (q) and satisfying

$$Z_t = A_t + \theta_1 A_{t-1} + \ldots + \theta_q A_{t-q} \tag{3}$$

which of course reduces to white noise in the special case with $q = 0$. One usually restricts the theta parameters $\theta_1, \ldots, \theta_q$ to values such that the polynomial $\theta_q(\zeta) \equiv 1 + \theta_1 \zeta + \ldots + \theta_q \zeta^q$, in the complex variable ζ, has no zeros within the unit circle. This is the "invertibility" condition. (Many practitioners would also exclude the possibility of zeros occurring on the unit circle, giving what we will term "strict invertibility".)

It is convenient to introduce the backshift operator B such that, for any function $f(t)$ (for example, z_t or A_t) and all integers s and t, $B^s f(t) \equiv f(t-s)$. Then, for instance, (3) can be written as

$$Z_t = \theta_q(B) A_t \tag{4}$$

where $\theta_q(B)$, a polynomial in B of degree q, is an operator obtained by writing B in place of ζ in $\theta_q(\zeta)$.

If θ_s is put at α^s for $s=1, \ldots, q$ and q is allowed to go to infinity, $\theta_q(B)$ becomes $(1-\alpha B)^{-1}$ and we can then rewrite (4) as

$$(1-\alpha B) Z_t = A_t \tag{5}$$

or, alternatively, as

$$Z_t = \alpha Z_{t-1} + A_t. \tag{6}$$

Expressions (5) and (6) represent the first order AutoRegressive model, or AR (1), in which each Z_t is "autoregressed" on its previous Z_{t-1} value.

This type of model generalises to AR $(p+d)$, with $p+d \geqq 0$, the $(p+d)th$ order autoregressive model

$$Z_t = \alpha_1 Z_{t-1} + \ldots + \alpha_{p+d} Z_{t-p-d} + A_t \tag{7}$$

or, in backshift notation,

$$\alpha_{p+d}(B) Z_t = A_t$$

where $\alpha_{p+d}(B) = (1-\alpha_1 B - \ldots - \alpha_{p+d} B^{p+d})$—and, again, none of the zeros of $\alpha_{p+d}(\zeta)$ must lie within the unit circle. Once more, when $p+d=0$, (7) reduces to white noise.

2.4. Nonstationary Models

Box and Jenkins in fact distinguish between those zeros of $\alpha_{p+d}(\zeta)$ which lie on the unit circle, say d of them, as opposed to those lying outside it, the remaining p zeros. They would then rewrite $\alpha_{p+d}(B)$ in the associated factored form $\varphi_p(B) S_d(B)$, where $\varphi_p(\zeta)$, corresponding to the "stationary" autoregressive part, has no zeros on the unit circle. If $d=0$, this condition ensures the "stationarity" of the process. Otherwise, the process is homogeneously non-stationary. (That is, it possesses a type of non-stationarity which can be removed by an appropriate unit-circle operator transformation.)

The simplest case of such a homogeneously nonstationary process is the random walk $Z_t = Z_{t-1} + A_t$. For instance, one might hypothesise, as a first approximation, that an unemployment series followed a random walk. (We might reason that this month's unemployment total equalled the sum of last month's together with an "innovation" value for this month, consisting of the number of freshly unemployed less those who have now found work and so have left the unemployed register.)

2.5. Simplifying Operators

A typical example of $S_d(B)$ is $(1-B)^d$, where the operator $(1-B)$ effects a (unit) differencing of the series under study. Thus, $(1-B)z_t = z_t - z_{t-1}$, and $(1-B)^d$ results in d successive (unit) differences being taken. (In practice, it is very rare that a degree of differencing d needs to be greater than one or two.)

Operators of the form $S_d(B)$ are termed "simplifying operators". They represent ways in which the raw series should be transformed initially before detailed statistical analysis is begun. As they yield linear transformations of the data, they do not lead to inefficient analysis in the ways that non-linear transformations do.

2.6. Other Initial Transformations

However, in certain instances, notably for the purposes of stabilising the variance of a series and obtaining more nearly Gaussian behaviour, non-linear transformations of the raw data might be made initially. These are usually of the Box-Cox (1964) form, namely $z_t \to z_t^{(\lambda, m)}$, where

$$z_t^{(\lambda, m)} = \begin{cases} (z_t + m)^\lambda, & \lambda \neq 0 \\ \ln(z_t + m), & \lambda = 0 \end{cases}$$

and m and λ need to be chosen by the analyst. But it should be noted that, unless the resulting change in metric appears desirable from other considerations, such transformation tends to be controversial. (For instance, good forecasts might be obtained for the transformed series, but the advantage lost on transforming back). To avoid cumbrous notation, we will suppress the transformation superscripts in what follows.

2.7. ARMA and ARIMA Models

Note that the general stationary AR (p) model

$$Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + A_t \tag{8}$$

can also be written as an MA (∞), $Z_t = \varphi_p^{-1}(B) A_t$, which is a special case of (2), with $\psi(B) = (1 + \psi_1 B + \psi_2 B^2 + \ldots) = \varphi_p^{-1}(B)$. A further generalisation to (8) and (3) is the mixed (stationary) AutoRegressive Moving Average model of order (p, q), or ARMA (p, q),

$$Z_t = \varphi_1 Z_{t-1} + \ldots + \varphi_p Z_{t-p} + A_t + \theta_1 A_{t-1} + \ldots + \theta_q A_{t-q}$$

which, in operator notation, becomes

$$\varphi_p(B)\, Z_t = \theta_q(B)\, A_t.$$

Introducing unit circle nonstationary zeros into the autoregressive part, we first get models of the form

$$\varphi_p(B)\,(1-B)^d\, Z_t = \theta_q(B)\, A_t \tag{9}$$

which are termed AutoRegressive Integrated Moving Average models of order (p, d, q), or ARIMA (p, d, q); and, when $(1-B)^d$ is replaced by more general $S_d(B)$, the models have been described as ARUMA (p, d, q)—for instance, see Anderson (1980).

2.8. Seasonal Models

Should B be replaced by B^T in (9), where T is some integer greater than unity, we get a purely seasonal model of period T. Such models are usually denoted by

$$\Phi_P(\mathrm{B}^T)\,(1-B^T)^D\, Z_t = \Theta_Q(B^T)\, A_t \tag{10}$$

where $(1-B^T)$ effects a seasonal differencing, according to

$$(1-B^T)\, z_t = z_t - z_{t-T},$$

and capital letters help to distinguish (10) from the earlier discussed non-seasonal models. Thus, the purely seasonal first order stationary autoregressive process of period 12, conveniently written as $\mathrm{AR}_{12}(1)$, would have the form $Z_t = \Phi\, Z_{t-12} + A_t$.

Mixed non-seasonal seasonal models can occur. These may be expressed as

$$\varphi_p(B)\,\Phi_P(B^T)\,(1-B)^d\,(1-B^T)^D\, Z_t = \theta_q(B)\,\Theta_Q(B)\, A_t \tag{11}$$

and, indeed, models with more than one seasonal period are possible. For instance, hourly electricity demand over the year would be likely to depend, not only on B and B^{24} (daily), but on B^{168} (weekly) as well. Also, note that multiplicative factors such as

$$(1-\varphi B)\,(1-\Phi B^T) = (1-\varphi B-\Phi B^T+\varphi\Phi B^{1+T})$$

may be generalised to, say

$$(1-\varphi B-\Phi B^T+\chi B^{1+T}).$$

2.9. Models with Deterministic Trend

Finally [and rather as with (1), the denigrated regression], univariate models can be written with a deterministic trend $f(t)$ on the right hand side, although this is frequently removed (at least in part) by the unit and seasonal differencing operators which are often employed. For instance, if a series contains a linear trend ct, with slope c, simple differencing of the raw data will reduce this to just a constant c, since

$$(1-B)\,ct = c\,[t-(t-1)] = c.$$

The simplest "trend" occurs when $E[Z_t]$ exists and is a non-zero constant μ. Then $f(t)=\mu$; and, to achieve models such as (2), Z_t is replaced by $Z_t-\mu$. So when $\bar{z}$, the series mean, is significantly different from zero (as would be the case in our wages example), the z_t are considered to be replaced by the mean-corrected series, $\{\tilde{z}_t = z_t - \bar{z}\}$, which is a (linear) Box-Cox transform of the original data. (Again, the "tilde" is usually suppressed in practice, when "mean correcting" is understood to have taken place).

2.10. Box and Jenkins

The family of linear models of the types described above are commonly referred to as Box-Jenkins models. Although they were mostly originally due to earlier workers, such as Kolmogorov, Wold, Yaglom and Yule.

Box and Jenkins, however, deserve the credit for bringing together, developing and popularising an extensive methodology (rightly known as the Box-Jenkins approach) which has been highly successful as a means of analysing time series met with in a very wide range of application areas. This success is founded on the fact that the various Box-Jenkins models can, between them, mimic the behaviours of diverse types of series – and do so adequately without usually requiring very many parameters to be estimated in the final choice of model. The disadvantage, however, is that successful analysis generally requires a modicum of skill – although some quite promising automatic modelling computer packages are now beginning to prove themselves. (For instance, see the recent study by Texter and Ord, 1986; and Shumway's enthusiastic review, 1986.)

The formal objective of a Box-Jenkins analysis may be considered as discovering that parsimoniously parameterized filter which satisfactorily reduces the original series to a residual white noise series $\{a_t\}$, with small variance. What is satisfactory will depend on the context from which the data were drawn, and on the purpose of the analysis, as well as purely statistical criteria.

2.11. Serial Correlation

The main analytical tool for series identification is the sequence of sampled serial correlations, $\{r_1, \ldots, r_{n-1}\}$, where

$$r_k = \sum_{t=k}^{n} \tilde{z}_t \tilde{z}_{t-k} \Big/ \sum_{t=1}^{n} \tilde{z}_t^2,$$

although frequently only about the first quarter of them are computed. These $\{r_k : k=1, \ldots, n-1\}$ are taken to mimic the theoretical autocorrelations $\{\rho_k\}$, defined as $\mathrm{Cov}[Z_t Z_{t-k}]/\mathrm{Var}[Z_t]$. So the task of orthodox identification is, given the observed sample correlation pattern for the series, to try to match it with the known population one for some particular process.

For instance, a proper (strictly invertible) MA(1) model is characterized by $\rho_1 \neq 0$ (in fact, $0 < |\rho_1| < 1/2$) and $\rho_k = 0 (k > 1)$. So a set of serials with r_1 substantial, but later r_k negligible, would suggest that an MA(1) should be tentatively tried. What count as substantial or negligible serial "spikes" depend (a little) on the particular model being considered and (more) on the length of the observed series. Given these facts, significance tests are available.

Certain structure in the sampled correlations can suggest that a simplifying operator should first be applied to the raw data. For example, a slow roughly linear declining sequence of positive values for the early r_k is often taken as an indication that unit differencing is necessary.

Note that the theoretical autocorrelation patterns, for mixed non-seasonal seasonal models, may be quite complicated. However, for purely seasonal processes, the results are directly analogous to those for the appropriate corresponding non-seasonal models. For instance, for the AR(1) model, $\rho_k = \varphi^k$; whilst, for the $\mathrm{AR}_{12}(1)$ process, $\rho_{12k} = \Phi^k$ with all other intermediate ρ's zero (alternatively written as $\{\rho_k \equiv \delta_k \Phi^{k/12}\}$, where δ_k is one or zero, depending on whether k is, or is not, an exact multiple of 12). And, in general, a purely seasonal model (of any type and complexity) has autocorrelations all zero, at lags which are not multiples of the seasonal period T, and autocorrelations at lags kT which are just those at lags k, for the corresponding non-seasonal model (obtained by replacing B^T everywhere by B).

So, for instance, if we know the $\{\rho_k\}$ pattern for

$$\Phi_P(B)(1-B)^D Z_t = \Theta_Q(B) A_t,$$

say $\{\rho_0^* (=1), \rho_1^*, \rho_2^*, \rho_3^*, \ldots\}$; then those, for model (10), are obtained by just inserting $T-1$ zeros between successive ρ_k^*, to give the pure-seasonal

model's autocorrelations as

$$\{\rho_0^*(=1), 0, \ldots, 0, \rho_1^*, 0, \ldots, 0, \rho_2^*, 0, \ldots, 0, \rho_3^*, \ldots\}.$$

2.12. Partial Correlation

Another useful tool for identification is the sequence of partial autocorrelations – sampled $\{p_k\}$, theoretical $\{\pi_k\}$. Thus, an AR(1) is characterized by $\pi_1 \neq 0$ (in fact, $|\pi_1| < 1$), $\pi_k = 0\,(k > 1)$. So, if p_1 is significant, but none of the later p_k are, an AR(1) model would be indicated.

One can consider π_k to be the conditional correlation between Z_i and Z_{i-k}, given all the intermediate Z's, namely $Z_{i-1}, \ldots, Z_{i-k+1}$. So, evidently, the partial autocorrelations are associated with the autocorrelations; the specific relation being conveniently written in terms of the autocorrelation matrix

$$\mathbf{P}_k = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & & \rho_{k-2} \\ \vdots & & & & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{pmatrix},$$

as $\pi_k = |\mathbf{P}_k^*| / |\mathbf{P}_k|$, where $\mathbf{P}_k^*$ is $\mathbf{P}_k$ with every (r, k)th element replaced by ρ_r.

2.13. Forecasting

A frequent purpose for analysing time series is to obtain good forecasts. Given a series $\{z_1, \ldots, z_n\}$ running up to time n=now, the aim then, typically, is to forecast z_{h+n} at h time intervals hence. If we assume that the generating process has the "random shock" form (2), it can be shown that the optimal least squares forecast, ${}_hf_n$ say, is the expected value of Z_{h+n} conditional on the information available at time n. Now

$$Z_{h+n} = A_{h+n} + \psi_1 A_{h+n-1} + \ldots + \psi_{h-1} A_{1+n} + \psi_h A_n + \ldots.$$

So

$${}_h f_n = \psi_h a_n + \ldots$$

(since the expectations of future shocks are all zero, whereas those for past and present ones take the actual values which have already occurred). And the forecast error, ${}_h\varepsilon_n = z_{h+n} - {}_hf_n$, is given by

$${}_h\varepsilon_n = a_{h+n} + \psi_1 a_{h+n-1} + \ldots + \psi_{h-1} a_{1+n},$$

with variance

$$_hV_n = (1 + \psi_1^2 + \ldots + \psi_{h-1}^2)\,\sigma_A^2;$$

from which probability limits for the forecasts can be obtained, on replacing σ_A^2 by its estimate, the sample variance of the residual shock series.

2.14. Some Recent Model Extensions

Currently there is much interest in extensions of the linear models discussed so far, to cater for at least part of the non-linearity common in some applications areas. One such extension gives the bilinear model, which is achieved by introducing additional product terms $Z_{t-u}A_{t-v}$ into the right of the linear model. Other popular generalisations involve substituting time varying parameters for the constants in (11).

2.15. Transfer Function Models

However, Box and Jenkins themselves (in conjunction with their co-workers) developed certain extensions to the univariate modelling described above. First they considered building transfer function models, which would perhaps improve the forecasts obtained for the series of interest, $\{z_t\}$, by extracting relevant information contained in some appropriate leading indicator series, $\{y_t\}$ say. This is done by relating the current z to some of the previous y's, according to a model

$$Z_t = \omega(B)\,B^{\delta}\,Y_t + E_t$$

where $\omega(B)$ is a linear filter $(\omega_0 + \omega_1 B + \ldots)$, with $\omega_0 \neq 0$, and the B^{δ} factor indicates that there is a delay of δ units before a y value can begin to affect the observed z. $\{E_t\}$ is a sequence of error terms, which is assumed to follow some ARMA (p, q) process. In general, parsimonious parameterization can be achieved by writing $\omega(B)$ as the quotient of two finite lengthed operators, $\alpha(B)$ and $\beta(B)$ say, so that the model fitted has the form

$$Z_t = \frac{\alpha(B)}{\beta(B)}\,Y_{t-\delta} + \frac{\theta_q(B)}{\varphi_p(B)}\,A_t. \tag{12}$$

Box and Jenkins (1976) provide a well-defined iterative model-building procedure for estimating these transfer function noise processes, which is analogous to that for the univariate case. Identification of the relationship

between the "input" y and "output" z series relies heavily on the cross-correlations between the two series, preferably after a procedure called "prewhitening" has first been effected. Here, the filter needed to reduce the y_t to white noise is initially determined, and then this prewhitening filter is applied to both the input and the output series before the cross-correlations are computed.

2.16. Multivariate Modelling

Equation (12) can be simply generalised to cater for several input series. Again, a deterministic trend can be incorporated on the right, the Z_t and Y_t can be transformed initially, seasonal factors can be introduced into the various filters, and differencing can be employed.

The univariate models can also be generalised to multivariate ones with the basic form

$$\boldsymbol{\varphi}(B)\,\mathbf{Z}_t = \boldsymbol{\theta}(B)\,\mathbf{A}_t \tag{13}$$

where the matrix operators $\boldsymbol{\varphi}(B)$ and $\boldsymbol{\theta}(B)$ have elements which are polynomials in B of general finite order, with the restrictions that those along the leading diagonals start with unity, whilst the rest start with powers of B, and the stationarity and (strict) invertibility conditions are, respectively, that all the zeros of the determinants $|\boldsymbol{\varphi}(B)|$ and $|\boldsymbol{\theta}(B)|$ lie outside the unit circle.

A further extension is to consider multivariate transfer function models whose basic structure is

$$\mathbf{Z}_t = \boldsymbol{\Omega}(B) \otimes \mathbf{Y}_t + \mathbf{E}_t \tag{14}$$

where the transfer function matrix, $\boldsymbol{\Omega}(B)$, has elements of the form $\omega(B)\,B^{\delta}$, $\otimes$denotes the Kronecker product and $\mathbf{E}_t$ follows a multivariate model, like (13).

Again, for models (13) and (14), as previously, there is a three-stroke cycle of identification, estimation and diagnostic checks for obtaining satisfactory fits; and, as well as cross-correlations, partial cross-correlations between individual pairs of residual series are also used. When building any of the models (12), (13) or (14), univariate stochastic modelling – of the type that was discussed earlier – is required for the individual series. So skilful univariate analysis is a prerequisite for all the more advanced Box-Jenkins methodology.

2.17. Intervention Modelling

Finally, in all the processes mentioned so far, it may be necessary to take account of "abnormal" events, such as strikes, changes in the law, or freak weather conditions. Box-Jenkins "intervention" models allow such effects to be represented by dummy variables, typically introduced as a filtered pulse on the input side. For instance, to model a step change in level of magnitude Λ occurring at time I, one needs to include a term $(1-B)^{-1}\Lambda\Pi_t^{(I)}$ on the right, where

$$\Pi_t^{(I)} = \begin{cases} 1 & t = I \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

It should be clear that pure "Intervention Analysis" (Box and Tiao, 1975), is merely a special case of Transfer Function modelling. Thus, using (12), the step change model described by (15) would give

$$Z_t = \frac{\Lambda}{(1-B)}\Pi_t^{(I)} + \frac{\theta_q(B)}{\varphi_p(B)}A_t.$$

(Outliers and missing observations can also be dealt with by interpolating realistic values which are estimated from the remaining data).

2.18. Some Further Reading

We conclude this discussion of "where we are coming from in the time domain", with a brief (if biased) look at the literature. First, some readable articles: Anderson (1976 *a*) provides a formula-free introduction, and a more mathematical treatment of the same material is found in Anderson (1976 *b*). Newbold (1975) gives an excellent treatment at about the same level, and Anderson (1989 *a*) states a later updated view.

As for books: Anderson (1975) and Nelson (1973) supply simple introductions, Box and Jenkins (1976) gives the best work for reference purposes, whilst Jenkins (in Anderson, 1979 *a*) provides an authoritative account of the more advanced topics – although Granger and Newbold (1977) is of related interest for multivariate modelling, as is Newbold and Reed (in Anderson, 1979 *a*). Ledolter (1978) also provides an excellent introductory coverage of multivariate Box-Jenkins, with a good application discussed in Melicher, Ledolter and D'Antonio (1981). Vandaele (1983) is possibly the best of the more recent texts covering the subject as far as transfer function modelling and intervention analysis.

In the rest of this paper we will restrict ourselves to univariate modelling. This is not because we regard multivariate data as unimportant or uninteresting, far from it. But there still remains a lot to be done in improving methods for analysing single series; and advancing, prematurely, into the multi-series arena seems very much like attempting to run, before one can really walk (with similar painful consequences).

It is reasonable to believe that an understanding of the ways single series behave will be the easier to come by; and many of the important univariate ideas will have generalizations, in multiple series systems, which should then be easier to establish. And, as we have already indicated, current multivariate methods anyhow largely depend for their effectiveness on the building blocks of satisfactory univariate modelling. So it really does seem wise to try to master univariate analysis first. Be this reason or rationalisation, it is what we intend to do.

3. WHERE UNIVARIATE TSA&F HAS GOT TO

3.1. Serial Correlation

Let us first look at a couple of examples from recent working papers by Wood, McInish and Ord (1983), and Koot and Young (1985): our figures 1 and 2, respectively.

Both sets of authors took these serial correlation plots to indicate a first order autoregressive model, but neither set managed to interpret the run of negative values (following the initial positive ones) – although Wood, McInish and Ord (1983) attributed theirs to perhaps "some degree of non-stationarity". The extreme commonness of this type of pattern, makes it incredible that virtually no one has thought to either comment on it or attempt to explain it.

As we will show, this phenomenon is a more obvious manifestation of the largely unnoticed aberration which occurs in the time domain. The point being that even the expected serial correlations, computed for finite series realisations, are a distortion of the population autocorrelations that correspond to the underlying process generating the data. The $\{ r_k \}$ do not measure the $\{ \rho_k \}$ with sampling error, but provide a distorted version of the $\{ \rho_k \}$ with sampling error then superimposed.

Of course, it has long been known that (*a*) r_k was, in general, a biased estimator of ρ_k, and (*b*) adjacent r_k, r_j were generally not independent but correlated. (See, for instance, Kendall (1954) and Barlett (1948), respectively.)

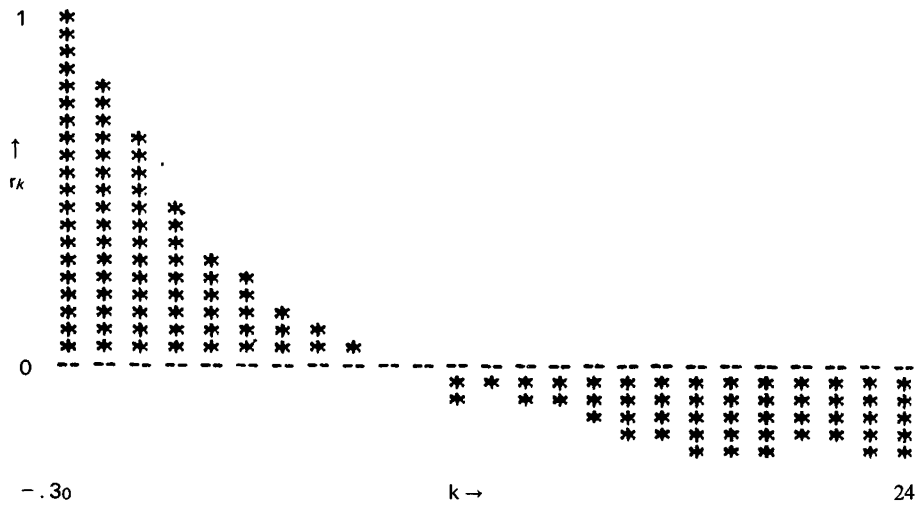

Figure 1. – Serial correlations of Global Market Return Averages for first thirty minutes of trading (adapted from figure 7, Wood, McInish and Ord, 1983; $n=30$).

However, almost no one has seen fit to study how observed serial correlation sequences for series relate to (or, rather, systematically depart from) their parent process autocorrelation functions. And, for instance, it turns out that the $\{r_k\}$ characteristic of an AR(1) process is not a geometric decay (like the associated $\{\rho_k\}$), but follows a smooth curve downwards which dips below the zero axis – at a value of k smaller than $n/3.42$ – and then proceeds to a minimum, before curving smoothly up towards zero again (reaching the axis, theoretically, at $k=n$). The "cross-over" from positive r_k to negative is more precisely given by the solution of equation 32 from Anderson (1979*c*), namely

$$n\varphi^k \simeq (1+\varphi)/(1-\varphi). \tag{16}$$

Now Wood, McInish and Ord (1983) reported $r_1 = .82481$; and, as we know, this estimate of $\rho_1 = \varphi$ is likely to be biased down substantially. However, for all values of $\varphi \in [.8, .9]$, the solution to (16) has k ranging from 4.3 to 5.4. The larger observed cross-over ($k \simeq 9.5$), together with the relatively

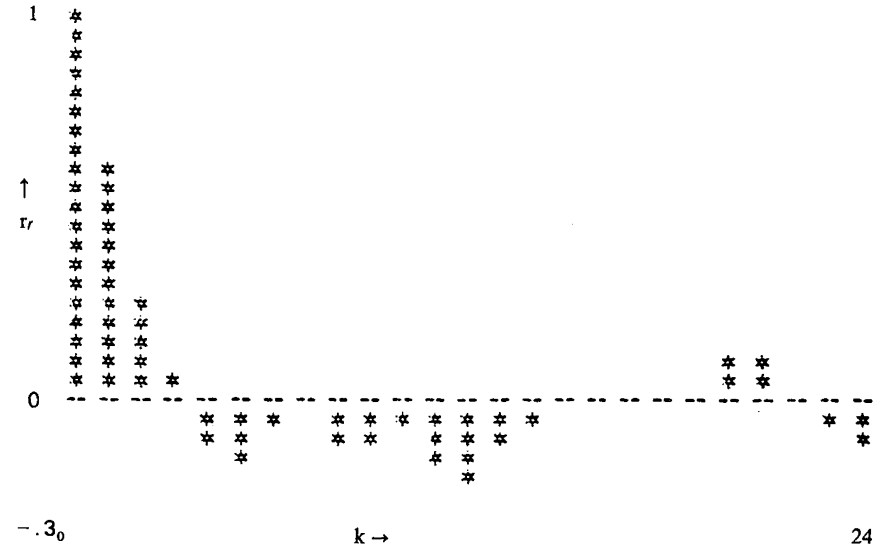

Figure 2. — Serial correlations for first differenced unemployment rate series (adapted from Koot and Young, 1985; $n=86$).

smooth appearance of the serial correlation pattern, suggests that a considerably higher φ is needed—perhaps even indicating a situation close to non-stationarity (as indeed was speculated by Wood, McInish and Ord).

However, $\varphi = .9$ is about the upper limit of validity for (16) when n is as small as 30, so we can not go further with it. Indeed, at $\varphi = 29/31$ ($\simeq .935$), (16) reduces to the nonsense $\varphi^k \simeq 1$. And, as a typical well-known estimate [2] for the bias in r_1, for an AR (1), is given by

$$E[r_1] \simeq \varphi - (1+3\varphi)/(n-1) \tag{17}$$

([2]) This result seems preferable to the other one Kendall derives, namely $E[r_1] \simeq \varphi - (1+4\varphi)/n$ (yielding the even greater $\hat{\varphi} = .990$), which is based on a mathematically convenient (but rather unnatural) condition that the series is "circular", having extended (unobserved) values $z_{n+j} \equiv z_j$. However, (17) also has a flaw. It assumes a random choice of a starting value, z_0, from $N(0, \sigma_z^2)$—which does not seem very realistic (given that, for any realisation, we will in fact know z_1; and backcasting then yields $E[Z_0] = z_1/\varphi$). It is interesting to note that these two estimates of φ correspond very closely to the least squares and Yule-Walker estimates, respectively, for which Tjøstheim and Paulsen (1983) drew the same conclusion as to which was preferable.

(see Kendall, 1954), which implies an estimator

$$\hat{\varphi} \simeq \left(r_1 + \frac{1}{n-1}\right) \Big/ \left(1 - \frac{3}{n-1}\right), \tag{18}$$

that gives $\hat{\varphi} = .958$ here, it is clear that we could have a situation with $\varphi > 29/31$. Moreover, for an ARIMA(0,1, q) or (more briefly) IMA(1, q) model, a result similar to (16) (Anderson, 1979 *b*, section 6) gives the cross-over at approximately $n/(2+\sqrt{2})$, $=8.8$ here, which is considerably closer to the $k = 9.5$ that is observed in figure 1.

Similarly, for the Koot and Young series (for which $n = 89$, and $r_1 = .59832$), (18) gives $\hat{\varphi} = .631616$, whence (16) yields $k \simeq 6.5$. Then the observed cross-over, at $k = 3.4$, is only $.0\dot{3}5\,n$ from the theoretical value for the roughly estimated AR(1) model. Given the greater sampling variability for the serial correlations from this far more stationary process (even considering the extra length of series), we would again appear to have a plausible explanation for the observed correlation structure.

In orthodox linear time domain modelling, the generally advocated approach is to attempt a match between the observed serial correlations and the known pattern of theoretical autocorrelations for some appropriate candidate parent process. However, this implicity assumes that the $\{r_k\}$ does in fact mimic the $\{\rho_k\}$ closely, except for sampling error. We regard such an assumption as false. For thoroughly stable processes, we tend to get a lot of sampling error, relatively, and so the systematic departure of the $\{r_k\}$-characteristic from the $\{\rho_k\}$-pattern (which is anyway less marked) is none too obvious. But it still usually exists; $\{r_k\}$ generally([3]) only gives us a distorted view of $\{\rho_k\}$, irrespective of sampling error.

For instance, no realisation from any AR(1) process can ever yield positive serial correlations following a geometric decay; no matter how one tries to fabricate the series. Indeed, for any series whatsoever, the totality of the serial correlations from r_1 to r_{n-1} has to be negative (as opposed to $\rho_1 + \ldots + \rho_{n-1} > 0$, for every $\{\rho_k \equiv \varphi^k, \varphi > 0\}$). But, the amount, by which the characteristic r_k-decline from unity is initially steeper than the corresponding ρ_k-decay, depends on both φ and n.

However, it is steeper. Although the ρ_k-decay can be as slow as we please, for suitable choice of φ sufficiently close to unity. Thus Wood, McInish and

([3]) That we are able to create counter-examples is demonstrated by choosing any length-2 realisation of $\{Z_t = (1-B)A_t\}$. Then $\rho_1 = -1/2 = r_1$.

Ord (1983) may be mistaken when they write for the serial correlations shown in figure 3: "The decay rate of the... correlation function... is much slower than the exponential decay which would be expected from a low order autoregressive process...".

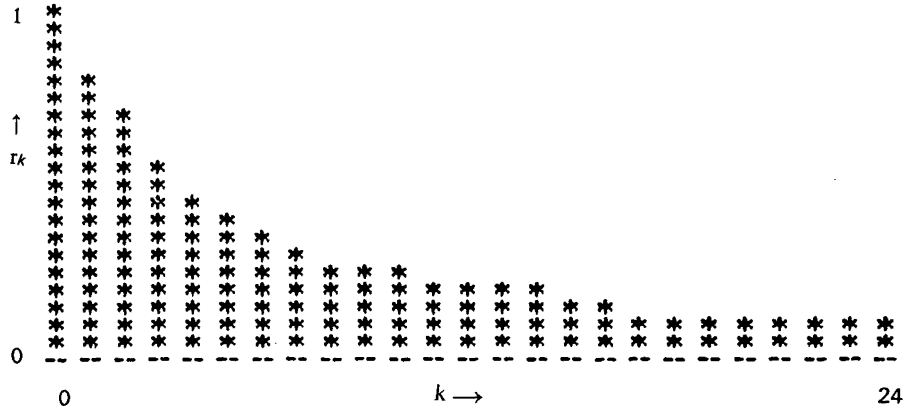

Figure 3. – Serial correlations of global market return averages for 330 trading minutes (adapted from figure 6 of Wood, McInish and Ord, 1983).

For an AR(1) series realisation of length 330 with $\hat{\varphi} \simeq (.80969 + (1/329))/(1 - (3/329)) = .8202086$ (using 18), we get from (16) that the cross-over $k \simeq 17.6$. Extrapolating the recorded serials, we might expect the actual cross-over to have occurred at about lag 39.4. So the observed decay is indeed slower than that expected from our roughly estimated AR(1), with the cross-over occurring 21.8 lags later – that is a discrepancy of $.073\,n$. But we imagine that Wood, McInish and Ord were thinking of the comparison of figure 3 with $\rho_k = \varphi^k$, φ something like .80969 or .82021.

However, if we had φ rather larger([4]), .9 say, then (16) would give $k \simeq 27.1$; while if $\varphi = .95$, $k \simeq 41.6$, and with $\varphi = .99$ (again about the limit for (16), given $n = 330$), $k \simeq 50.3$. (In the $\varphi = .95$ case, the observed decay would be slightly *faster* than the characteristic one, with a discrepancy between cross-overs of just 2.2 lags, or $.007\,n$.) Whilst for an IMA(1, q) model([5]), when $n = 330$, $k \simeq 96.7$, which seems far too great (a discrepancy of $.191\,n$).

([4]) This would be plausible if there were a moving average factor, $(1 - \theta B)$, partially cancelling the autoregressive $(1 - \varphi B)$.

([5]) With, again, a partially cancelling MA(1) factor, $(1 - \theta B)$.

When processes approach being homogeneously nonstationary, the sampling variability of the serial correlations decreases dramatically, and the differences between $\{r_k\}$ and $\{\rho_k\}$ will be emphasised. The discrepancy between the two is also correspondingly greater, so that the distinctness of the two patterns should be very clear. However, again people have been slow to latch onto this point.

In particular, the serial correlation behaviour is often modified considerably as we approach non-stationarity, so that it should frequently be possible to discriminate between parameters giving "nearly nonstationary", but nevertheless still stationary processes, and ones (perhaps only slightly different in magnitude) which do imply nonstationarity. Such a distinction has long been recognised as important for its implications for forecasting; but, again, few people have considered the possibility of the serial correlation pattern being able to provide a fairly sensitive test of this point in many instances—a possibility we will discuss in more detail below.

Basically, then, what we have been concerned with to date is establishing reliable results for the moments and distributions of the serial correlations, given any linear process and any length of series. And, as n is a crucial parameter in all formulae, we emphasise its role by rewriting r_k as $r_k^{(n)}$, which underlines the fact that (in our view) it is relatively pointless to look at, say, the first 20 serial correlations, unless we know the length of the series from which they were derived.

One of the things that has militated against cross-over and related aberrant phenomena being more widely noticed is the practice of only considering at most the first $n/4$, say, of the $r_k^{(n)}$. Now that computational costs are so much reduced, it is strongly recommended that one almost always outputs plots of the complete serial (and partial) correlation functions in the social, business and management sciences (where series lengths are usually never very great, anyway).

3.2. Cautionary Tales

Rather than give detailed formulae and results concerning the moments and distributions of serial correlations (which can be found in the two survey papers Anderson (1980, 1989 *b*), and the references therein), we will restrict ourselves to presenting a few illustrations that warn against thinking in terms of the theoretical ρ_k-patterns, instead of the more practically illuminating characteristic $r_k^{(n)}$ behaviours.

3.3. A Misconception Concerning the Serial Correlations for Series Realisations from ARIMA Processes

Consider a random walk as the limit of an AR (1) process with parameter $\varphi \to 1$. Then the walk's theoretical autocorrelations may be taken as the limits as $\varphi \to 1$ of $\{\varphi^k\}$; that is, all may be put at unity. Similarly, the $\{\rho_k\}$ for general ARIMA processes can be taken as $\{\rho_k \equiv 1\}$.

However, the $\{r_k^{(n)}\}$ for series realisations from such processes will in no way resemble this $\{\rho_k\}$-pattern. The difference in behaviour being indicated for the random walk by figure 4.

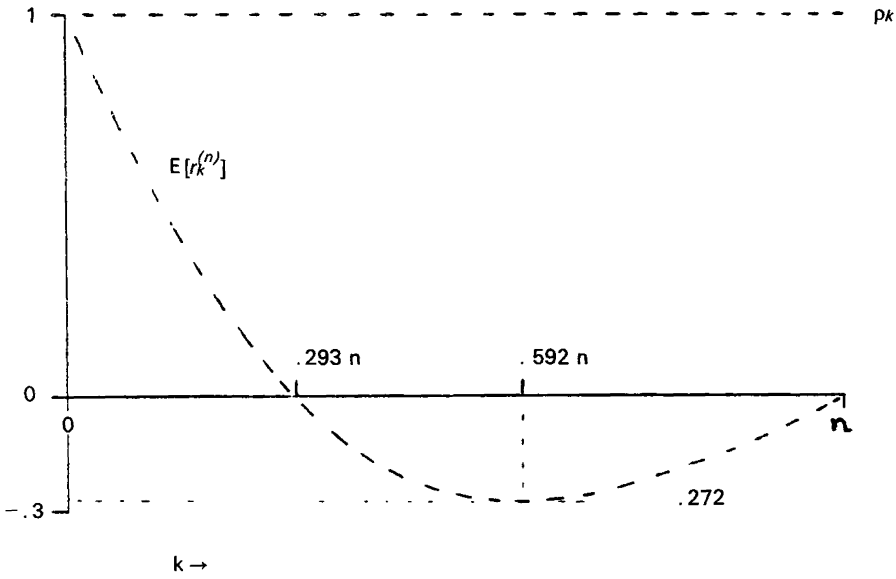

Figure 4. — Divergence, for the random walk model, between: (*a*) expected serial correlations, $\{E[r_k^{(n)}]\}$, from series realisations; and (*b*) theoretical process autocorrelations, $\{\rho_k\}$.

Box and Jenkins (1976) noticed that, in practice, the serial correlations at low lags, for many ARIMA (*p*, *d*, *q*) processes, tended to follow slow roughly linear declines from a zero-lag value of unity. However, rather than trying to explain this as being due to the serials following the $\{E[r_k^{(n)}]\}$ and not the $\{\rho_k\}$, they considered it as a result of the generating process not being quite nonstationary. For instance, with the AR (1) model, instead of letting $\varphi \to 1$, write $\varphi = 1 - \varepsilon$ where ε takes a small positive value. Then $\rho_k = \varphi^k \simeq 1 - k\,\varepsilon$, for low lags k.

As we have seen, this is not the correct explanation. It derives from attempting to rationalise inferences made from the suboptimal procedure of trying to identify models by matching observed serial correlations with the $\{\rho_k\}$, as opposed to the $\{E[r_k^{(n)}]\}$ patterns. What might have alerted researchers, to recognizing the error, is the peculiar tendency of ε to be frequently approximately either $5/n$ (corresponding to ARIMA $(p, 1, q)$ models) or $3/n$ (for all other ARIMA (p, d, q) processes, with $d>1$ ([6])). But, of course, this is not such an obvious point, when one is not always bearing the value of n in mind. (Quite a helpful procedure, here, is to routinely "standardise" serial correlation plots, by arranging the (horizontal) scale so that n lags always occupy the same length whilst maintaining a fixed (vertical) scale for the actual serial correlation "spikes".)

3.4. The Thompson and Tiao Serial Correlations

Thompson and Tiao (1971) reported a case-study with serial correlations as given in figure 5.

Note that no classical Box-Jenkins ARIMA model has a ρ_k-pattern even remotely resembling this plot. However, Wichern (1973) recorded one of the

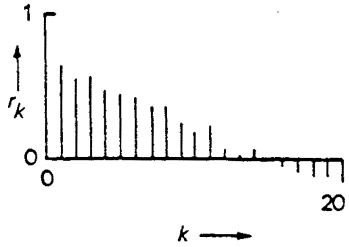

Figure 5. – Serial correlations for telephone data reported by Thompson and Tiao (1971).

few (now) relatively well-known situations where the serial correlations could depart substantially from the theoretical autocorrelations, that of the IMA (1,1) model

$$(1-B)\,Z_t=(1-\theta\,\mathrm{B})\,A_t \tag{19}$$

with $0<\theta<1$.

([6]) Also for models of the form (1), where $f(t)$ represents either a linear trend or a "staircase" step function – *see* Anderson (1988).

Wichern argued that, although the nonstationary $1-B$ operator would, by itself, give rise to serial correlations which followed a slow linear decline from unity; should $\theta=1$, this operator would be cancelled on both sides of (19), yielding white noise which is characterised by $\rho_k \equiv 0$ $(k>0)$. Then, for θ relatively close to 1, a partial cancelling of operators would effectively occur, and the typical pattern for the serial correlations would be intermediate between those for a random walk and those for white noise. Wichern then showed that the serials would be expected to follow a decline from a value less than unity, and would pass into the negative region, attaining a minimum before increasing again (to reach zero at the hypothetical lag $k=n$).

Anderson (1975, p. 110) simulated such a process with $\theta = .6$, and recorded the serial correlations for a series realisation of length 50. See *figure* 6.

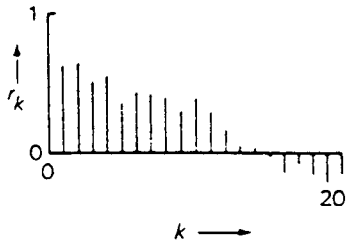

Figure 6. – Serial correlations for a length 50 simulation of $(1-B)Z_t=(1-.6B)A_t$.

As we note, the similarity between figures 5 and 6 is quite remarkable; and a reasonable inference might be that the Thompson and Tiao data were generated by a model close to

$$(1-B)Z_t=(1-.6B)A_t. \tag{20}$$

However, we have forgotten n. Figure 6 was based on 50 observations, and the theoretical cross-over for the IMA(1,1), described by (20), is thus $50/(2+\sqrt{2})=14.6$ (which agrees very well with what is established by eye in figure 6, namely a cross-over at about $k=14.7$). But, for the Thompson and Tiao data, $n=190$, for which an IMA(1,1) cross-over would be at $k=55.6$; not the 15 indicated by figures 5.

This compares with an expected cross-over of 13.7 obtained from (16) for an AR(1) realisation with $n=190$ and $\varphi = .8$ (and the "slow linear decline" of the r_k, from a value less than unity, actually suggests a larger φ, with a partially cancelling $(1-\theta B)$ factor on the MA side).

In fact, Thompson and Tiao eventually fitted an ARMA (1,1) model

$$(1-.92\,\mathrm{B})\,\mathrm{Z}_t=(1-.66\,\mathrm{B})\,A_t$$

for which the expected cross-over would be at $k=24.3$, invoking a deduction from equation (32) of Anderson (1979 *c*), namely:

$$n\,\varphi^{k-1}\simeq\frac{1-2\,\theta\,\varphi+\theta^2}{(\varphi-\theta)\,(1-\theta\,\varphi)}+\frac{2}{1-\varphi}.$$

Actually there is a hint, in the displayed serial correlations themselves, that figure 5 is based on a greater length of series than is figure 6. The serials for figure 5 give a rather smoother plot than those in figure 6, suggesting perhaps the smaller sampling variability associated with larger n.

3.5. When Intuition Can Fail

In the last subsection, we considered models of the form $(1-B)\,Z_t=(1-\theta\,B)\,A_t$ $(0<\theta<1)$, and discovered that the typical shape for the serial correlations was a smooth curve downwards, from a value below unity, which crossed into the negative region, decreased until a minimum was reached and then curved back upwards towards zero.

I would like the reader, then, to consider the expected behaviour for the serials of the purely seasonal process

$$(1-B^2)\,Z_t=(1-\theta\,B^2)\,A_t \qquad (0<\theta<1). \tag{21}$$

One would expect this to be basically the same as for model (19), except that the serials will be zero at odd lags and take the "Wichern-like" pattern just for their even values, the seasonal period of the process being two. Thus, if we considered a particular case of (21), say

$$(1-B^2)\,Z_t=(1-.64\,B^2)\,A_t \tag{22}$$

and $n=100$, we might expect something like figure 7 for a typical series realisation, where the dots indicate (schematically) the supposed $E\,[r_k^{(n)}]$. And, I do not think anyone will quibble very much with that.

Figure 8 shows the serial correlations that were in fact obtained for a (typical) actual simulation. Evidently we have a counter-intuitive result, although this behaviour is exactly as predicted by results in Anderson (1979 *d*); and, with a little thought, figure 8 does appear reasonable.

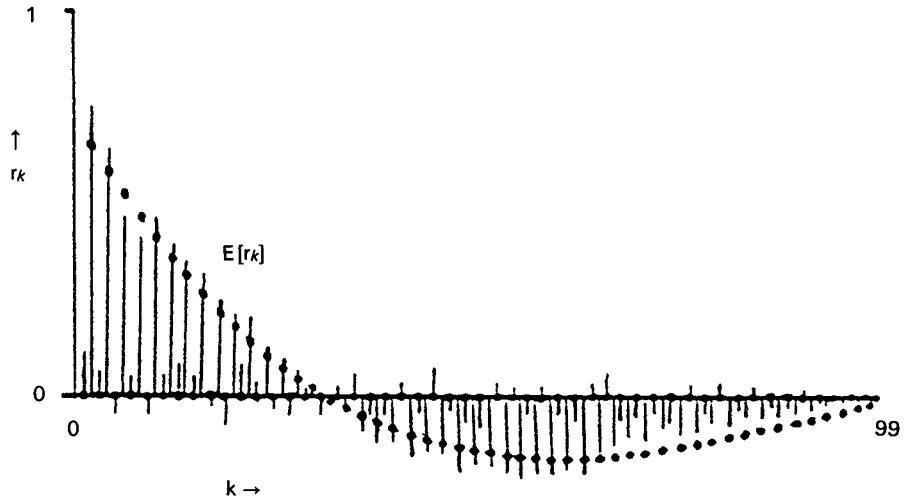

Figure 7. — Suggested schematic forms for serial correlations and $E[r_k^{(n)}]$ for a length-100 simulation from the model $(1-B^2)Z_t=(1-64B^2)A_t$.

Consider, for simplicity, the model

$$(1-B^2)Z_t=A_t. \tag{23}$$

Here the odd and even Z_t are clearly independent, and the two subprocesses formed by them both follow (independent) random walks.

Imagine what happens when simulating. Perhaps one starts by writing $z_{-1}=z_0=0$. Then, initially, the pair of walks are likely to intertwine somewhat. But, after a reasonable induction period, they are very likely to be separate, one above the other, as far as short sections of the combined series are concerned. Basically, the two walks are able to wander freely; and, as their paths are not dependent on each other, they will usually be exploring disjoint (or nearly disjoint) fragments of $(-\infty, \infty)$. Thus, schematically, we can consider them as in figure 9—where, arbitrarily, the odds are placed above the evens.

Clearly, there, the mean $\bar{z}$ lies between the two separated walks and so $(z_{\text{odd}}-\bar{z})$ is always positive, whereas $(z_{\text{even}}-\bar{z})$ is always negative, and both deviations from the mean are large compared with the standard deviation of the A_t. Thus, from the definition of serial correlations, two things occur, the $r_{\text{odd}}^{(100)}$ will all be negative and the $r_{\text{even}}^{(100)}$ positive, and they will tend to quite

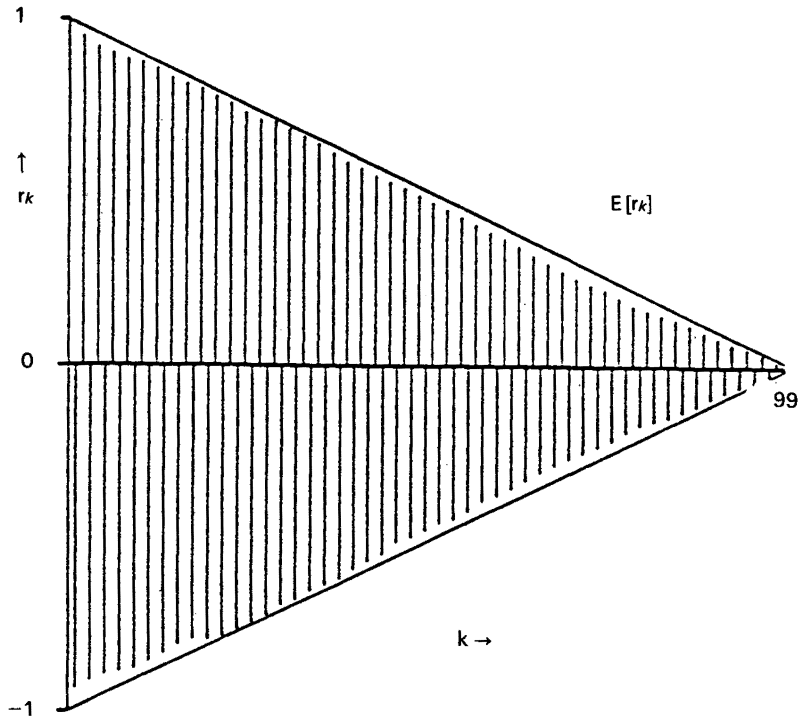

Figure 8. — Actual $E[r_k^{(n)}]$ lines and serial correlations obtained from a length-100 simulation from the model $(1-B^2)Z_t=(1-.64B^2)A_t$.

closely follow the $E[r_k^{(100)}]$ lines, as shown in figure 8, declining roughly linearly from values with large magnitude at low lag to those of small magnitude at high lag.

Of course, had a real process (23) been running infinitely long, its two walks would be expected to generally have an infinite separation. However, if it is to be observed, an actual realisation would have to have its halves finitely separated (which would also occur if the process has only been running for a finite time). Evidently, we could get a situation where figure 9 does not approximate to the sort of behaviour displayed by a realisation from process (23), but this will be a relatively rare event for observed realisations other than extremely young ones.

What is happening is that, as we have already seen with the Wichern-type behaviour (where the nonstationary operator of an appropriate IMA(1,1) can be to some extent annulled by "partial cancellation" with the moving

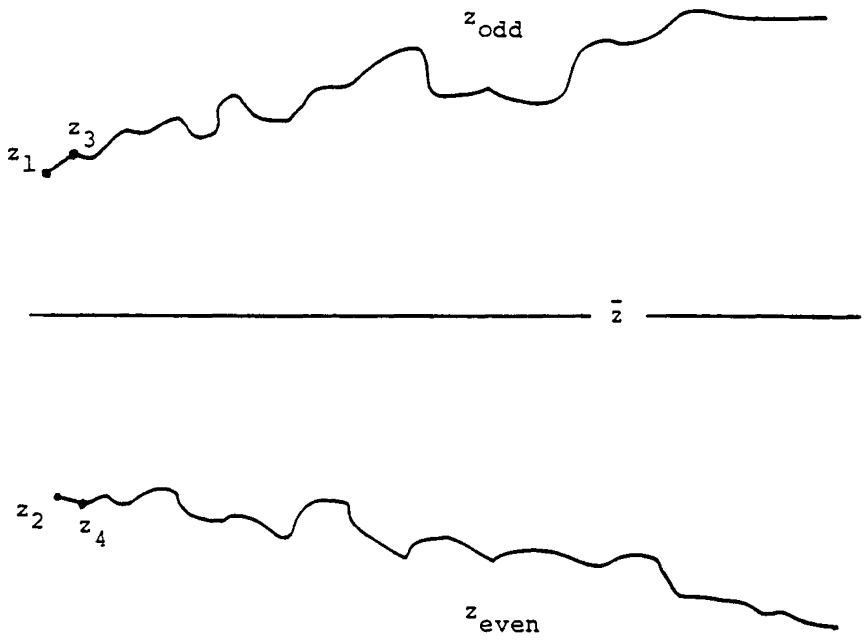

Figure 9. — Schematic representation of a typical short (but warmed up) realisation from the process $(1-B^2)Z_t = A_t$.

average "echo" operator), $(1-B)$ is an anomalously weak nonstationarity operator; and in $(1-B^2)$, which factorises into $(1-B)$ $(1+B)$, the $(1+B)$ part dominates. So ARUMA $(p, 2, q)$ models with $S_2(B)=(1-B^2)$ give rise to serial correlation structure whose form is dictated by the $(1+B)$ factor. On applying the $(1+B)$ simplifying operator — as is suggested by figure 8 (for instance, see equation 6 in Anderson, 1980) — to the simulation from model (22), serial correlations similar to figure 10([7]) would result, leading to a further simplification by differencing before the remaining MA (2) part would be identified. (And note that we have the full simplification going through in stages.)

([7]) Another example of Wichern-type behaviour.

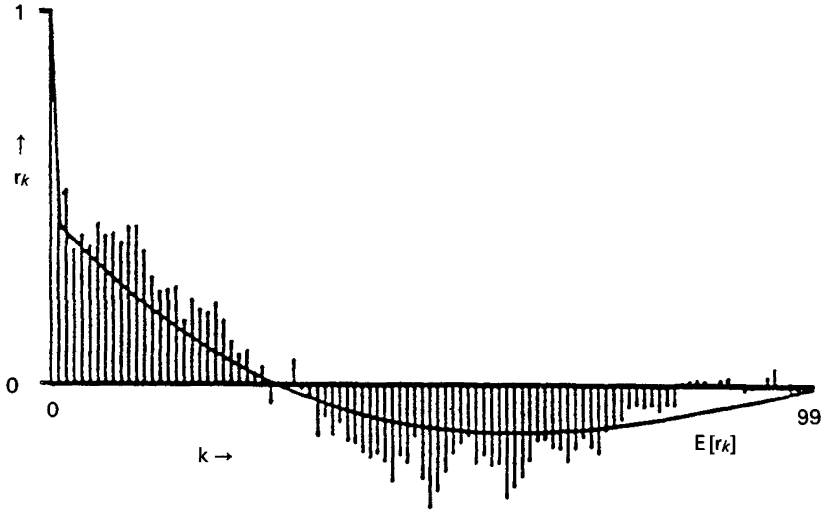

Figure 10. — Serial correlation plot and $E[r_k^{(n)}]$ curve for a single length-100 simulation from the model $(1-B)Z_t=(1-.64B^2)A_t$; where the seasonal period-2 MA-operator factorises into $(1-.8B)(1+.8B)$, with the $(1-.8B)$ partially cancelling out the nonstationary $(1-B)$.

3.6. A Final Test of Identification Skill

Table I gives the first five pairs of serial and partial correlations for an actual series of length 37. Can you identify the model?

TABLE I

Serial and partial correlations for a series of length 37.

k	1	2	3	4	5
r_k	0.498	0.245	0.118	0.053	0.019
p_k	0.498	−0.003	−0.004	−0.005	−0.006

It is clear that the $\{r_k\}$ closely follows a geometric decay, from a value of r_1 near 0.5, and the only "large" partial autocorrelation is at lag 1. The immediate conclusion might well be that we have a realisation from something close to $Z_t=(1/2)Z_{t-1}+A_t$.

In fact the series, giving rise to the table, was generated by a process which violates the stationarity conditions, namely

$$\left.\begin{aligned} Z_1&=1 \\ Z_t&=2Z_{t-1}+A_t \qquad (t=2,\ \ldots,\ 37)\end{aligned}\right\} \tag{24}$$

with $A_t=0$. A bit of a cheat, you might say, it is not even stochastic. But note that, whatever moderate value one gives to σ_A^2, a realisation of (24) is virtually certain to "break loose" almost immediately and then closely follow the deterministic process, $Z_t=2Z_{t-1}$. And, for even a short length of 37 terms, its r_k and p_k will differ very little from the values shown in table I.

But, of course, for as short a history from a stationary process, one would expect much more distortion of the population patterns in the sampled table. We repeat the moral that one must constantly bear in mind the length n of the realisation, when analysing a time series.

Box and Jenkins (1976) gave a mere 10-term simulation of the same explosive process, starting with $z_1=0$ and putting $\sigma_A^2=1$. Even there we see the same type of behaviour (table II).

TABLE II

Correlations for the Box and Jenkins (1976) explosive simulation [from $Z_t=2Z_{t-1}+A_t$, $A_t\sim$IN(0,1)].

k	1	2	3
r_k	0.458	0.165	−0.001
p_k	0.458	−0.057	−0.070

Evidently, a glance at the actual series values in this case would again, and more emphatically, prevent one from inferring a stable AR (1) generator. But what if the process had been only slightly explosive? The second moral is that one should not omit closely scrutinizing the raw data, at the very start of an analysis.

Heuristically, the behaviour of the correlations for explosive series, such as (24), can be explained by considering the reversed realisation, which will have identical correlation properties, but which now represents a decay, with parameter 1/2, from some high initial value [of order 10^{11} for (24)].

More formally, when given $\{z_t=2z_{t-1}+a_t : t=1, \ldots, n\}$ and n of moderate size or large, it is straightforward to show that:

(*a*) $r_k\simeq 2^{-k}$ for $k\ll n$.

(*b*) Starting from lag one, $\{r_k\}$ tends to decrease monotonically, with the initial positive values being followed by a negative run, until a minimum is attained. Then the values start increasing again, but remain negative, and finally satisfy

(*c*) $r_{n-j}\simeq(2^{-j}-1)/((1/3)n-1)$ for $j\ll n$.

To illustrate all this, figures 11 and 12 show the complete sets of r_k, for the Box-Jenkins simulation and our "cheat" series, respectively. Note that the general $\{r_k\}$ patterns are again of the cross-over type.

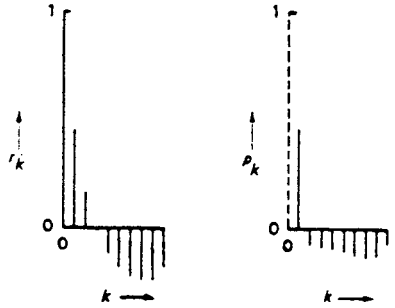

Figure 11. – Complete correlation patterns for the Box and Jenkins simulation.

Also note the patterns of the partial correlations in figures 11 and 12. Although we have not the space, here, to explain what is happening, such runs of small negative values are indicative that we have a strongly nonstationary process (not a mere ARIMA $(p, 1, q)$ model, with the weak $(1-B)$ nonstationarity operator). Indeed, looking at the $\{p_k\}$ in table I, the smooth run of negative partials should surely have warned an analyst that the series was not a thoroughly stationary one, such as a AR(1) with $\varphi = 1/2$, where considerable sampling error would be expected.

3.7. Errors Arising from Interpreting Stationary Data as Nonstationary and *Vice Versa*

The motivation for this subsection derives from remarks made by Box and Jenkins in their book (1976) and earlier publications, and substantiated by the subsequent pratice of both themselves and many other practitioners working in the time domain. These are that, when faced with a situation in which a series realisation could perhaps be modelled by a linear ARMA process, having an autoregressive operator factor $(1-\varphi B)$ with φ apparently a little less, but not much less than unity ([8]), then a recommended prudent

([8]) The sort of largest value for $1-\varphi=\varepsilon$, say, with which we are concerned depends marginally on the rest of the model but very heavily on the series length n. For shorter n, ε can be larger, the maximum $\varepsilon(n)$ being roughly inversely proportional to n.

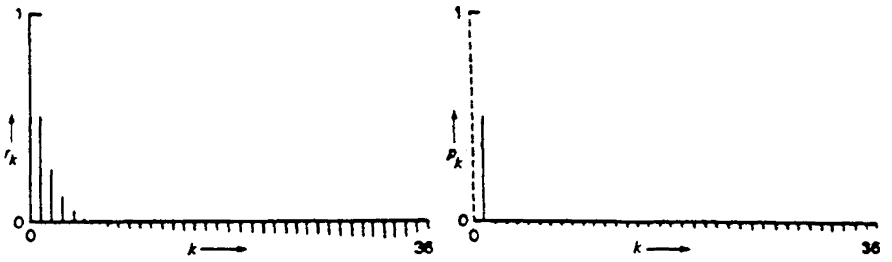

Figure 12. – Complete correlation patterns for the "cheat" series.

strategy is to replace the $(1-\varphi B)$ in the model by the "differencing" operator $(1-B)$ – which, when applied to an observed realisation, has the effect of transforming the raw series to its sequence of first differences.

Such statements are supported by arguments that the $(1-\varphi B)$ case causes forecasts for future values of a time series to be tied to the mean value, $\bar{z}$, of the past observations, whereas the $(1-B)$ choice allows future predictions to wander freely from wherever the series has got to at its last observed point. See figure 13.

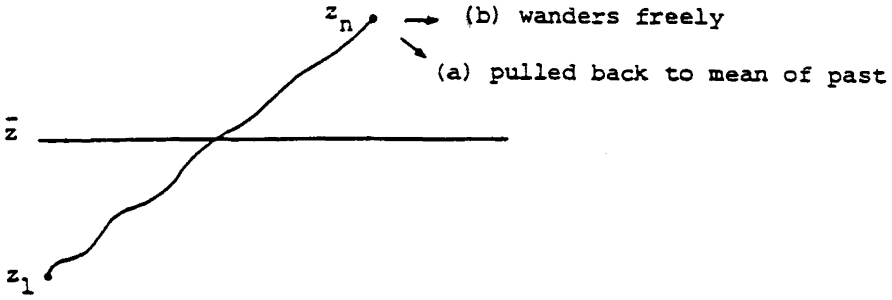

Figure 13. – Schematic representation of Forecasts from a series modelled with factor *(a)* $(1-\varphi B)$ **and** *(b)* $(1-B)$.

As commented by Anderson (1975, p. 122), when φ is near to unity, there is not a very marked difference, in practice, between the behaviour of short term forecasts obtained from the two choices. For, as φ gets closer to 1, although there is still a mathematical tendency for the future values of the series to revert to the mean of their past, the pull towards this mean becomes weaker and the behaviour approaches that of an unconstrained meander around the latest local level.

Also, of course, there are reasons for avoiding unnecessary differencing, which tends to increase the residual variance of the random shock component that is achieved by the Box-Jenkins modelling (and so leads to unnecessarily high predicted forecast error variances), and which frequently gives rise to problems in estimating the (overdifferenced) model's parameters. Such considerations have led many other analysts to oppose routine differencing.

However, we intend here to give fresh arguments which question the logic of what we will call the "play-safe" strategy of automatically replacing a possible $(1-\varphi B)$ autoregressive factor, having φ near 1, with $(1-B)$. We believe our studies indicate that, when we generalise these ideas, for the purpose of distinguishing between "nearly nonstationary" ARMA models of the form

$$\varphi(B) S_d(\alpha B) Z_t = \theta(B) A_t \tag{25}$$

with α less than but not much less than 1, and homogeneous nonstationary ARUMA models of the form

$$\varphi(B) S_d(B) Z_t = \theta(B) A_t, \tag{26}$$

then we obtain insight and methodological extensions that can lead to much swifter, simpler and cheaper model identification, given series from these types of long-memory processes, which will consequently often give improved modelling and an expectation of better forecasts in practice.

If one knows, for certain, that the series was in fact generated by a model with $\varphi = \varphi^*$ and not $\varphi = 1$, it would appear intuitively obvious that the φ^* choice should be modelled, if optimum forecasts are required. However, evidently in practice, the right model is never known, and the "play-safe" strategy would then seem to be based on two implicit assumptions:

(i) It is not possible to discriminate between φ^* and 1.

(ii) Due to this, it is wiser to choose $\varphi = 1$, as resulting costs from forecast errors are then expected to be smaller in the long run.

To focus attention, we first considered the pair of models

$$(1-.95B) Z_t = (1-.74B) A_t \tag{27}$$

$$(1-B) Z_t = (1-.8B) A_t \tag{28}$$

and simulated 100 realisations, each of length $n = 100$, from them. The model for each realisation was chosen at random from the two choices and was unknown until after the subsequent analysis had been completed.

We looked at (27) and (28) because Wichern (1973) had previously fitted an ARMA (1,1) to a simulated realisation of 100 terms from (28) and come up with model (27), which he claimed was indistinguishable from the true model, given the series length considered—a point with which we feel that most practitioners would agree.

Our discrimination experiment (described in Anderson and de Gooijer, 1979), however, although only based on some very approximate theory (Anderson, 1977), gave us a success rate of 4 : 1, which we offer as *sample* evidence to refute (i). (See, also, Anderson and de Gooijer, 1980 *a*).

If we next consider table III which shows the possible choices that the analyst can make, given the true model options, we can easily demonstrate that: (a) (from a population viewpoint) the costs, in terms of mean square forecast errors, associated with the two possible types of misspecification are indeed not symmetric; and (*b*) that the expected costs, incurred by mistaking $\varphi = 1$ for $\varphi = .95$ (the case ringed), are greater than those from picking $\varphi = 1$, when $\varphi = .95$ in fact. This is discussed in Anderson (1981 *a*).

TABLE III

Possible choices open to analyst given the true model, with $\varphi = .95$ *or* 1.

		Model Picked by Analyst	
		$\varphi = .95$	1
True Model φ =	.95	✓	x
	1	(x)	✓

Assuming that our discrimination procedure has a long run efficiency of 80% correctly specified series (which we believe is not unreasonable, given our greater current insight), it can thus be shown that the approach leads to an improvement in forecasting performance at all leads. For instance, at lead-ten, a reduction of 1.2% in mean square forecast error is expected.

This improvement is only small, because it can be shown that the ringed misidentification in table III is almost four times as costly, in terms of the

resultant deterioration in expected mean square forecast error, as is the unringed (play-safe) one.

However, for all other $S_d(B)$ (that is, for all but the "weak" $(1-B)$ nonstationary operator), correct discrimination between (25) and (26) is far more certain; and the gains from applying the approach much higher in consequence. (See Anderson and de Gooijer, 1980 *b*, for some examples; and Anderson, 1981 *b*, for an introductory discussion.)

A word of warning for those wishing to run simulation experiments on general homogeneous nonstationary processes and their nearly nonstationary analogues. Long induction periods for warming-up the simulated processes are generally needed. For instance, correctly simulating series, of length 100, from any such nonstationary process [excluding the case of weak nonstationarity, with $S_d(B)=(1-B)$] requires a warm-up period of around 400 discarded initial terms. See Anderson (1979 *e*), for the only known published treatment of this problem, although this restricts itself to the very simplest case of an AR (1) simulation.

4. Where We Are Going To

Frankly, we do not know. It would seem wise to first formally put all our material together in a coherent and rigorous thesis – which will be done in a proposed book for Wiley. We shall then proceed along the most promising avenues, seeking perhaps to do less, but establish rather more. (It is not the results one achieves that matter, but how well you succeed in disseminating them.) Less effort will be given to preaching – instead we will concern ourselves more with getting people to pay attention.

More work is needed in three areas:

(1) We need to set the inference and discrimination ideas in a rigorous framework.

(2) We must provide an unequivocal demonstration of the gains to forecasting, from using the serial correlation characteristic patterns rather than the theoretical autocorrelations, for purposes of process identification.

(3) A full investigation into the warm-up problem is required, so as to ensure efficient and effective simulation experiments.

ACKNOWLEDGMENTS

I am most grateful to the following for helpful criticism: Neville Davies, Jan de Gooijer, Jim Durbin, Norman Johnson, Paul Newbold, Keith Ord and Morris Walker.

REFERENCES

O. D. ANDERSON, *Time Series Analysis and Forecasting: The Box-Jenkins Approach*, Butterworths, London, 1975.

O. D. ANDERSON, *Box-Jenkins in Government: A Development in Official Forecasting*, Statistical News, Vol. 32, 1976 *a*, pp. 14-20.

O. D. ANDERSON, *Some Methods in Time Series Analysis*, Mathematical Scientist, Vol. 1, 1976 *b*, pp. 27-41.

O. D. ANDERSON, *Time Series Analysis and Forecasting: A Further Look at the Box-Jenkins Approach*, Cahiers du CERO, Vol. 19, 1977, pp. 223-256.

O. D. ANDERSON, Editor *Forecasting*, North-Holland, Amsterdam & New York. 1979 *a*.

O. D. ANDERSON, *Some Sample Autovariance Results for a Once Integrated qth-order Moving Average Process*, Statistica, Vol. 39, 1979 *b*, pp. 287-299.

O. D. ANDERSON, *Formulae for the Expected Values of the Sampled Variance and Covariances from Series Generated by General Autoregressive Integrated Moving Average Processes of Order (p, d, q)*, *Sankhya B*, Vol. 41, 1979 *c*, pp. 177-195.

O. D. ANDERSON, *The Autocovariance Structures Associated with General Unit Circle Nonstationary Factors in the Autoregressive Operators of otherwise Stationary ARMA Time Series Models*, Cahiers du CERO, Vol. 21, 1979 *d*, pp. 221-237.

O. D. ANDERSON, *On Warming-up Time Series Simulations Generated by Box-Jenkins Models*, J. Opl Res. Soc., Vol. 30, 1979 *e*, pp. 587-589.

O. D. ANDERSON, *Serial Dependence Properties of Linear Processes*, J. Opl Res. Soc., Vol. 31, 1980, pp. 905-917. Correction, Vol. 35, 1984, p. 171.

O. D. ANDERSON, *On Forecasting Certain Misspecified Models*, J. Information & Optimization Sciences, Vol. 2, 1981 *a*, pp. 236-240.

O. D. ANDERSON, *Covariance Structure of Sampled Correlations from ARUMA Models*, In *Time Series Analysis*, O. D. ANDERSON and M. R. PERRYMAN Eds., North-Holland, Amsterdam & New York, 1981 *b*, pp. 3-26.

O. D. ANDERSON, *The Serial Correlation Structure for a Random Process with Steps*, Metrika, Vol. 36, 1988 pp. 349-376.

O. D. ANDERSON, *Time Series Analysis and Forecasting: A Fresh Look at Time Domain Modelling*, 1989 *a*, in preparation.

O. D. ANDERSON, *More Serial Dependence Properties for Linear Processes*, 1989 *b*, in preparation.

O. D. ANDERSON and J. G. DE GOOIJER, *On Discriminating Between IMA* (1,1) *and ARMA* (1,1) *Processes: Some Extensions to a Paper by Wichern*, Statistician, Vol. 28, 1979, pp. 119-133.

O. D. ANDERSON and J. G. DE GOOIJER, *Distinguishing Between IMA* (1,1) *and ARMA* (1,1) *Models: A Large Scale Simulation Study of Two Particular Box-Jenkins Time Processes*, in *Time Series*, O. D. ANDERSON Ed., North-Holland, Amsterdam & New York, 1980 *a*, pp. 15-40.

O. D. ANDERSON and J. G. DE GOOIJER, *Distinguishing certain Stationary Time Series Models from their Nonstationary Approximations and Improved Box-Jenkins Forecasting*, in *Analysing Time Series*, O. D. ANDERSON Ed., North-Holland, Amsterdam & New York, 1980 *b*, pp. 21-42.

M. S. BARTLETT, *On the Theoretical Specification and Sampling Properties of Autocorrelated Time Series*, Supplement J. Roy. Statist. Soc. B, Vol. 8, 1946, pp. 27-41. Correction, Vol. 10, 1948, pp. 1.

G. E. P. Box and D. R. Cox, *An Analysis of Transformations*, J. Roy. Statist. Soc. B, Vol. 26, 1964, p. 211-243.

G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*, Revised Edition, Holden-Day, San Franciso, 1976.

G. E. P. Box and G. C. Tiao, *Intervention Analysis with Applications to Economic and Environmental Problems*, J. Amer. Statist. Assoc., Vol. 70, 1975, pp. 70-90.

C. W. J. Granger and P. Newbold, *Forecasting Economic Time Series*, Academic Press, New York, 1977, Second Edition, 1986.

M. G. Kendall, *Note on Bias in the Estimation of Autocorrelation*, Biometrika, Vol. 41, 1954, pp. 403-404.

R. S. Koot and P. Young, *Some Empirical Tests of the Political Business Cycle*, Working Paper No 85-10, Division of Management Science, The Pennsylvania State University, 1985.

J. Ledolter, *The Analysis of Multivariate Time Series Applied to Problems in Hydrology*, J. Hydrology, Vol. 36, 1978, pp. 327-352.

R. W. Melicher, J. Ledolter and L. J. D'Antonio, *A Time Series Analysis of Aggregate Merger Activity*, Review of Economics and Statistics Vol. 65, 1983, pp. 423-430.

C. R. Nelson, *Applied Time Series Analysis for Managerial Forecasting*. Holden-Day, San Francisco, 1973.

P. Newbold, *The Principles of the Box-Jenkins Approach*, Opl Res. Quarterly, Vol. 26, 1975, pp. 397-412.

R. H. Shumway, *Statistical Computing Software Review on "Autobox (Version 1.02)"*, American Statistician, Vol. 40, 1986, pp. 299-301.

P. A. Texter and J. K. Ord, *Forecasting Using Automatic Identification Procedures: A Comparative Analysis*, Working Paper No 86-3, Division of Management Science, The Pennsylvania State University, 1986.

H. E. Thompson and G. C. Tiao, *Analysis of Telephone Data: A Case Study of Forecasting Seasonal Time Series*, Bell J. Econ. Man. Sci., Vol. 2, 1971, pp. 515-541.

D. Tjøstheim and J. Paulsen, *Bias of Some Commonly-Used Time Series Estimates*, Biometrika, Vol. 70, 1983, pp. 389-399.

W. Vandaele, *Applied Time Series and Box-Jenkins Models*. Academic Press, New York, 1983.

D. W. Wichern, *The Behaviour of the Sample Autocorrelation Function for an Integrated Moving Average Process*, Biometrika, Vol. 60, 1973, pp. 235-239.

R. A. Wood. T. M. McInish and J. K. Ord, *An Investigation of Nonstationarity in the Market Index Using Minute-by-Minute Returns*, Working Paper No 83-3, Division of Management Science, The Pennsylvania State University, 1983.