REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

OLIVER D. ANDERSON
JAN G. DE GOOIJER

Discrimination between nonstationary and nearly nonstationary processes, and its effect on forecasting

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 24, n° 1 (1990), p. 67-91.

<http://www.numdam.org/item?id=RO_1990__24_1_67_0>

Recherche opérationnelle/Operations Research
(vol. 24, n° 1, 1990, p. 67 à 91)

DISCRIMINATION BETWEEN NONSTATIONARY AND NEARLY NONSTATIONARY PROCESSES, AND ITS EFFECT ON FORECASTING (*)

by Oliver D. ANDERSON (1) and Jan G. De GOOIJER (2)

Abstract. – *We present theoretical and empirical evidence to show that the structure, for the observed serial dependence between the values of a series realisation, is quite sensitive to the distinction between a near-nonstationary model and a just nonstationary approximation to it. Reliable discrimination between the two may well be possible then, in practice, and this implies that improved modelling, as judged by increased forecasting effectiveness, can perhaps be achieved.*

We study exact and approximate measures of serial covariance and serial correlation, respectively, for a wide class of non-explosive linear time processes, including the ARMA and ARIMA models.

Keywords : Forecasting misspecified models; identifying time series; serial correlation distributional properties.

Résumé. – *Nous présentons des arguments théoriques et empiriques qui montrent que la structure, pour la dépendance sérielle observée entre les valeurs d'une suite de réalisations, est très sensible à la distinction entre un modèle approximativement non stationnaire et une approximation exactement non stationnaire de ce dernier. Une discrimination fiable entre les deux est donc possible, en pratique. Ceci entraîne qu'une modélisation améliorée (l'amélioration étant jugée par une efficacité accrue de la prévision) pourrait être réalisée.*

Nous étudions des mesures exactes ou approximatives de covariance et de corrélation sérielles pour une large classe de processus temporels linéaires non explosifs, qui incluent les modèles ARMA et ARIMA.

1. INTRODUCTION

Consider any time series of length $n \geqslant 2$, $\{z_1, \ldots, z_n\}$ with mean $\bar{z} = (z_1 + \ldots + z_n)/n$ (where n denotes the latest observation available to the analyst), and let the associated generating process $\{Z_i\}$ be defined by the

(*) Received April 1989.

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

(2) Department of Economic Statistics, University of Amsterdam, Jodenbreestraat 23, 1011 NH Amsterdam, The Netherlands.

Recherche opérationnelle/Operations Research, 0399-0559/90/01 67 25/$ 4.50

linear model

$$(1-\varphi_1 B-\ldots-\varphi_p B^p)(1-\alpha\psi_1 B-\ldots-\alpha^d\psi_d B^d)Z_i = (1-\theta_1 B-\ldots-\theta_q B^q)A_i. \quad (1)$$

Here B is the backshift operator such that $B^j f(i)\equiv f(i-j)$, for any $f(.)$ and all integers i and j; $\{A_i\}$ is a white noise sequence of uncorrelated but identically distributed zero-mean random variables, all with variance σ^2 say; and $\varphi_1, \ldots, \varphi_p, \alpha, \psi_1, \ldots, \psi_d, \theta_1, \ldots, \theta_q$ are real parameters satisfying the conditions (2) to (5) that follow:

$$0<\alpha\leqslant 1; \quad (2)$$

and, defining ζ as a complex variable,

$$\Phi(\zeta)=(1-\varphi_1\zeta-\ldots-\varphi_p\zeta^p) \quad (3)$$

has all its zeros outside the unit circle,

$$\Psi(\zeta)=(1-\psi_1\zeta-\ldots-\psi_d\zeta^d) \quad (4)$$

has all its zeros on the unit circle, and

$$\Theta(\zeta)=(1-\theta_1\zeta-\ldots-\theta_q\zeta^q) \quad (5)$$

has none of its zeros within the unit circle.

Evidently (1) can be written more concisely as

$$\Phi(B)\Psi(\alpha B)Z_i=\Theta(B)A_i \quad (6)$$

and we will be frequently interested in cases where $\alpha=1$, namely

$$\Phi(B)\Psi(B)Z_i=\Theta(B)A_i \quad (7)$$

which we refer to as the ARUMA (p, d, q) class of *all* homogeneous nonstationary models. Of course, when $\Psi(B)=(1-B)^d$, the ARUMA process reduces to the ARIMA model of Box and Jenkins. In models such as (6), $\Phi(B)\Psi(\alpha B)$ is referred to as the autoregressive operator of the process.

For our time series, we define the k-th serial covariance by

$$c_{k,\alpha}^{(n)}=\frac{1}{n}\sum_{i=1}^{n-k}(z_i-\bar{z})(z_{i+k}-\bar{z}) \qquad (k=0, 1, \ldots, n-1). \quad (8)$$

Then the serial correlation, at lag k, is given as $r_{k,\alpha}^{(n)} = c_{k,\alpha}^{(n)}/c_{0,\alpha}^{(n)}$ when $k=1, \ldots, n-1$. For a stationary model, such as (6) with $\alpha < 1$, where $E[Z_i]\,(=\mu_\alpha$, say) exists and is not dependent on i, we have these sample quantities corresponding to the theoretical autocovariances and autocorrelations, $\{\gamma_{k,\alpha}\}$ and $\{\rho_{k,\alpha}\}$ respectively, which are defined by

$$\gamma_{k,\alpha} = E[(Z_i - \mu_\alpha)(Z_{i+k} - \mu_\alpha)]$$

and

$$\rho_{k,\alpha} = \gamma_{k,\alpha}/\gamma_{0,\alpha}.$$

In cases where $d > 0$ and $\alpha = 1$, $E[Z_i]$ no longer exists and so we define

$$\rho_{k,1} = \lim_{\alpha \nearrow 1} (\gamma_{k,\alpha}/\gamma_{0,\alpha})$$

as $\gamma_{k,1}$ is not then defined for any k. Note that $\rho_{k,1}$, so defined, always exists and is unique for any model (7).

When $d=0$, (7) reduces to the stationary ARMA (p, q) model whose theoretical autocovariances and autocorrelations will always be denoted by $\{\gamma_k\}$ and $\{\rho_k\}$, respectively. Similarly, the corresponding sample quantities are then written as $c_k^{(n)}$ and $r_k^{(n)}$.

Note that for $\Psi(\alpha B) = (1 - \alpha B)$

$$\gamma_k = -\alpha\gamma_{k-1,\alpha} + (1+\alpha^2)\gamma_{k,\alpha} - \alpha\gamma_{k+1,\alpha} \tag{9}$$

from which it follows that

$$r_k^{(n-1)} \simeq -\alpha r_{k-1,\alpha}^{(n)} + (1+\alpha^2) r_{k,\alpha}^{(n)} - \alpha r_{k+1,\alpha}^{(n)} \tag{10}$$

and when α needs to be estimated

$$\hat{r}_k^{(n-1)} \simeq -\hat{\alpha} r_{k-1,\alpha}^{(n)} + (1+\hat{\alpha}^2) r_{k,\alpha}^{(n)} - \hat{\alpha} r_{k+1,\alpha}^{(n)}. \tag{11}$$

Relations analogous to (9), (10) and (11) can easily be obtained for other possible $\Psi(\alpha B)$, and we believe that the degree of approximation involved is frequently small.

Finally, as an approximation to $E[r_{k,\alpha}^{(n)}]$, we will use the quantity $E_k^{(n)}$ defined by $E[c_{k,\alpha}^{(n)}]/E[c_{0,\alpha}^{(n)}]$, for any stationary ARMA $(p+d, q)$ model, and as $\lim_{\alpha \nearrow 1} E[c_{k,\alpha}^{(n)} \mid (6)]/E[c_{0,\alpha}^{(n)} \mid (6)]$, for any homogeneous nonstationary model (7).

2. BACKGROUND

The motivation for the study which we shall describe in this paper derives from remarks made by Box and Jenkins in their book (1976) and earlier publications, and substantiated by the subsequent practice 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 ARIMA process, having an autoregressive operator factor $(1 - \varphi B)$ with φ apparently a little less, but not much less than unity ([3]), then a recommended prudent 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 1.

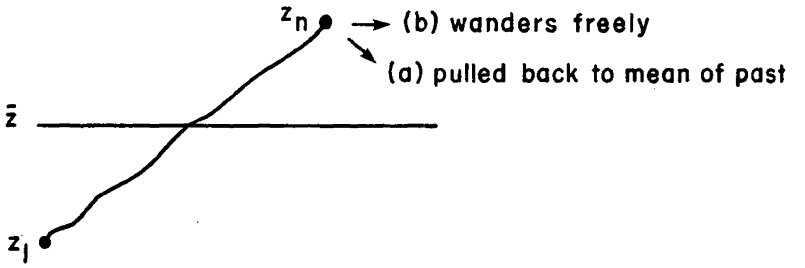

Figure 1. — Schematic representation of forecasts from a series modelled with factor (*a*) $(1 - \varphi B)$ **and** (*b*) $(1 - B)$.

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

([3]) 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.

weaker and the behaviour approaches that of an unconstrained meander around the latest local level.

Also, of course, there are good 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, with φ near 1, by $(1-B)$. We believe our studies indicate that when we generalise these ideas, for the purpose of distinguishing between "nearly nonstationary" ARMA models of the form (6), with α less than but not much less than 1, and homogeneous nonstationary ARUMA models of the form (7), then we obtain insight and methodological extensions that can lead to much swifter, simpler and cheaper model identification, for certain types of series, which will consequently often give improved modelling and an expectation of better forecasts in practice.

3. DISCRIMINATING BETWEEN A CHOICE OF $(1-\varphi^* B)$ AND $(1-B)$

If one knows 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

$$(1-.95B)Z_i=(1-.74B)A_i \tag{12}$$

$$(1-B)Z_i=(1-.8B)A_i \tag{13}$$

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 (12) and (13) because Wichern (1973) had previously fitted an ARMA (1, 1) to a simulated realisation of 100 terms from (13) and come up with model (12), 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*a*), gave us a success rate of 4:1, which we offer as *sample* evidence to refute (i).

If we next consider table I which shows the possible choices that the analyst can make, given the true model options, we can easily demonstrate that (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 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 the next section.

TABLE I

Possible choices open to analyst given the true model.

	Model picked by analyst	
	$\varphi = .95$	1
True model $\varphi =$ .95	√	×
True model $\varphi =$ 1	⊗	√

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 a rather unspectacular improvement in forecasting performance at all leads. For instance, at lead-ten, a reduction of 1.2% in mean square forecast error is expected.

However, this analysis is not quite relevant to the problem in practice, as we are actually only interested in the first column of table I. For, if we choose $\varphi = 1$, we are then just doing the same as someone employing the "play-safe" strategy – and we are evidently only concerned with the *different* identifications achieved from our discrimination. That is, we are just interested in those particular realisations which are identified as stationary.

But the calculation does provide us with some *a priori* reason for believing that discrimination may, in fact, lead to improved expectations for mean

square forecast error. This is because, we now know that the ringed error is expected to be a little less than 4 times as costly as the unringed one. Also, in our discrimination paper, we noted that 84 of the simulations gave clear cut identification – the remaining 16 cases being doubtful. For these 84, we were in fact correct 85% of the time (and, with our greater insight, we feel this could be nearer to 90% now), but we were not at all successful in the doubtful cases (7 right and 9 wrong). It would therefore seem sensible to propose the use of the "play-safe" choice whenever $\varphi = 1$ is indicated or the situation does not appear clear cut, and to only go for $\varphi = .95$ when this possibility is strongly suggested. This might then be expected to yield an improved forecasting strategy – that will refute assumption (ii).

However, it does seem difficult to ever fully justify this belief theoretically, as what is really relevant, of course, is not the whole population of series realisations from the two models, but just those cases where the analyst (using our proposed discrimination) will correctly pick out the model $\varphi = .95$ (instead of playing safe with $\varphi = 1$) and incorrectly mistake what is, in fact, $\varphi = 1$ for $\varphi = .95$. So we would need to look at the forecasting characteristics associated with sub-populations which will be hard to pin down mathematically, being those which cause the analyst to think he is dealing with the $\varphi = .95$ case. (Incidentally, again *a priori*, these realisations would appear to be ones for which the forecasts obtained from using the two factors are not too desperately different, as one might expect the latest local level not to be very far from the realisation's mean, for otherwise the stationary model would be unlikely to have been identified with conviction.)

We therefore propose, instead, to conduct a second discrimination experiment where, this time, realisations of 100 length-120 series will be simulated, at random, from models (12) and (13). Again we will use the first 100 terms of each (in isolation) for discriminating, but then restrict our attention to just those cases where $\varphi = .95$ is strongly indicated. For these cases we will identify, estimate and verify an ARIMA $(p+1, d, q)$ fit, as appropriate, and also the "play-safe" ARIMA $(p, d+1, q)$ fit (p, d and q not necessarily being the same for both models). The two fits will then be used to predict the next 20 terms of the series, already simulated, and hence the actual sample averaged square forecast errors for the two cases can be computed and compared.

It is believed that this second experiment may result in significantly better forecasts for our strategy. Although, even if this should not be so, we feel that the generalisations of our approach from $(1-B)$ to $\Psi(B)$ models (and their close stationary approximations) will still be found extremely useful.

To complete this section, we conjecture the situation where our forecasting experiment, as described above, proves successful. Then, of course, it will be necessary to take the discrimination approach out of the controlled but unrealistic environment of simulation experiments and try it on real series.

The problem which then arises is that, since gains in forecasting performance are not expected to be dramatic, any improvements which we report are likely to be summarily dismissed, by sceptics, as resulting from our having just selected data that support our proposed procedure. Likewise, any contrary study refuting our claims can be similarly disregarded, and the only way that individual analysts may be persuaded of the approach's usefulness would seem to come from their giving it an impartial trial and finding that it works.

However, we do believe from our experience that anyone trying the idea will definitely find it useful for discrimination and, hopefully, also for forecasting.

Finally, it should be pointed out that our approach contains an element of subjectivity – although this does not appear crucial. Typically, everyone should score between, say, 75% and 85% correct in the basic discrimination. Some completely objective (modified) method could easily be devised, but it would involve extra computation and we think it would score rather worse than 80% right – that is, less than we can get from immediate visual inspection. There is, in consequence, little motivation, at present, for us to remove the small degree of subjectivity.

4. FORECASTING MISSPECIFIED MODELS

An obvious question arises as to whether being able to distinguish between models (12) and (13) with 80% certainty leads to improved forecasts – and the answer to this is not immediately apparent. For it could happen that the occasional wrong inference of a stationary model, when the actual process was (13), caused losses which more than cancelled the gains from now being able to frequently identify (12) correctly.

Throughout what follows in this section, we will assume that we are dealing with some specific process of interest, $\{Z_i\}$, which is generated from a particular white noise process, $\{A_i\}$, according to either model (12) or (13), but that this model (whichever it is) has been misspecified as the incorrect one.

It is well-known that, when the correct model is used to form forecasts, the one-step ahead forecast errors have a variance of σ^2. However, the

expected magnitudes of the errors arising from forecasting (12) by (13) are not the same as those occurring when forecasting (13) by (12). In the first case, the lead-1 forecast error variance is simply shown to be $1.00123\,\sigma^2$; whereas, in the second, it is infinite – and this might be considered as sufficient reason for always preferring model (13).

However, in general, an actually observed series will possess an overall level significantly different from zero, so the raw data needs to be mean-corrected, before fitting a model. So, in practice, one is interested in models such as (12) and (13) where z_i is replaced by $\tilde{z}_i = z_i - \bar{z}$, with $\bar{z}$ the mean of the observed series from which forecasts are to be obtained.

We then get, for the case of misspecifying (12) by (13),

$$\tilde{Z}_{n+1} = .95\,\tilde{z}_n + A_{n+1} - .74\,a_n$$

and

$$F_{n,1} = \tilde{z}_n - .8\,a_n$$

where $F_{n,j}$ denotes the forecast made at lead-j from time n. So the forecast error, $E_{n,1} = \tilde{Z}_{n+1} - F_{n,1}$, is given by

$$\begin{aligned} E_{n,1} &= -.05\,\tilde{z}_n + A_{n+1} + .06\,a_n \\ &= -.05\,(1 + .95\,B + .95^2\,B^2 + \ldots)(1 - .74\,B)\,a_n + A_{n+1} + .06\,a_n \\ &= A_{n+1} + .01\,a_n - .0105\,(1 + .95\,B + .95^2\,B^2 + \ldots)\,a_{n-1} \end{aligned}$$

with variance $V_{n+1} = 1.001\,23\,\sigma^2$ as before, when n is fairly large – in particular, for $n = 100$. (Indeed, the analysis is unchanged on replacing Z_i by $\tilde{Z}_i$.)

However, when (13) is mistaken for (12), the same argument leads to

$$E_{n,1} = .05\,\tilde{z}_n + A_{n+1} - .06\,a_n, \tag{13 a}$$

and $\tilde{Z}_n$ replacing Z_n means that this evidently now has finite variance. In fact, noting that $(1 - B)\,Z_i \equiv (1 - B)\,\tilde{Z}_i$, $\{A_i\}$ is seen not to change, given model (13), when Z_i is replaced by $\tilde{Z}_i$. Thus, we have

$$\begin{aligned} \tilde{z}_n &= \{1 - n^{-1}(1 + B + \ldots + B^{n-1})\}\,z_n \\ &= n^{-1}\{(n-1) + (n-2)\,B + \ldots + B^{n-2}\}(1 - B)\,z_n \\ &= n^{-1}\{(n-1) + (n-2)\,B + \ldots + B^{n-2}\}(1 - .8\,B)\,a_n. \end{aligned}$$

So (13 *a*) yields

$$E_{n,1} = A_{n+1} - .01\,(1+5\,n^{-1})\,a_n + .01\,n^{-1}\{(n-6) + (n-7)\,B + \ldots - 4\,B^{n-2}\}\,a_{n-1},$$

and this then gives $V_{100,1} = 1.002\,92\,\sigma^2$.

We conclude that mistaken identification of the nonstationary model as stationary will indeed be the more costly misspecification; but, given equal chances of having each of the two models, our discrimination method should result in an average lead-1 forecast error variance of something like $\sigma^2 \times$

$$\{80 + 10\,(1.001\,23 + 1.002\,92)\}/100 = 1.000\,42$$

as opposed to

$$(1 + 1.001\,23)/2 = 1.000\,62$$

which is obtained when playing safe.

Of course, this difference would often be negligible in practice – although the cost of implementation will usually be even less. However, for longer leads, the gain could become more substantial. For instance, lead-10 forecast error variances are given in table II.

TABLE II

Ratios of lead-10 forecast error variances to driving shock variances (σ^2)

		Actual model of process	
		12	13
Model used for	12	1.272 6	1.545 2
Forecasting	13	1.350 0	1.36

So, perfect discrimination would give an average lead-10 forecast error variance/σ^2 of $(1.272\,6 + 1.36)/2 = 1.316\,3$, as opposed to $(1.350\,0 + 1.36)/2 = 1.355\,0$ from "playing safe". An improvement of nearly 3%; whilst, in the lead-1 case, these would be 1 and 1.000 6 respectively, a decrease of only .06%.

Our discrimination should give an average lead-10 forecast error variance ratio of something like

$$.4\,(1.272\,6 + 1.36) + .1\,(1.545\,2 + 1.350\,0) = 1.342\,6$$

a gain of .92% as opposed to only .02% for lead-1.

Finally, denote lead-j forecasts of model r using model s by $F_{n,j}(r, s)$, with error variance written as $V_{n,j}(r, s)$. Then

$$V_{n,j}(r, s: r \neq s) = V_{n,j}(r, r) + \mathrm{Var}\{F_{n,j}(r, r) - F_{n,j}(r, s)\},$$

since the difference between $\tilde{Z}_{n+j}$ and $F_{n,j}(r, r)$ is clearly uncorrelated with $F_{n,j}(r, r) - F_{n,j}(r, s)$. So

$$\lim_{j \to \infty} \{V_{n,j}(r, s: r \neq s) - V_{n,j}(r, r)\} = \mathrm{Var}\{Z_n - .8 A_n\}$$

which, for $r = 12$ and 13 respectively, approaches $.492\,3\,\sigma^2$ and $1.161\,4\,\sigma^2$ from below. Thus the average absolute gain from our discrimination is expected to be as shown in table III.

TABLE III

Average absolute decrease in forecast error variance that might be expected from 80% *correct discrimination.*

Lead	1	10	$\to \infty$
Decrease	$.000\,2\,\sigma^2$	$.012\,4\,\sigma^2$	$.080\,8\,\sigma^2$

Note that, if the reader does not wish (as a routine procedure) to mean-correct series whose means are not significant, it is quite straight-forward to extend the analysis of this section to cover such cases. The results do not change substantially.

5. THEORETICAL RESULTS

(Throughout this section formulae hold for $0 \leqslant k \leqslant n-1$)

For any ARMA (p, q) model, given by (7) with $\Psi(B) = 1$, we get that

$$E[n c_k^{(n)}] = \frac{1}{n^2}\left\{n^2 (n-k)\gamma_k - kn\gamma_0 - 2k \sum_{j=1}^{n-1} (n-j)\gamma_j \right.$$
$$\left. + n\left(\sum_{j=1}^{k} j\gamma_{n-j} + \sum_{j=1}^{k-1} j\gamma_{n-2k+j} - \sum_{j=1}^{n-k-1} j\gamma_{2k-n+j} - \sum_{j=1}^{n-k} j\gamma_{n-j} \right)\right\}. \quad (14)$$

Again, for any ARIMA $(p, 1, q)$ process, obtained by putting $\Psi(B)=(1-B)$ in (7), we have that

$$E[nc_{k,1}^{(n)}]=\frac{1}{6n^2}\Bigg[n(n-k)\{(n^2-4kn+2k^2-1)\gamma_0-6n\sum_{j=1}^{k-1}(k-j)\gamma_j\}$$
$$-2n\sum_{j=1}^{k-2}(k-j)\{(k-j)^2-1)\}\gamma_j+2n\sum_{j=1}^{n-k-2}(n-k-j)\{(n-k-j)^2-1)\}\gamma_j$$
$$+2k\sum_{j=1}^{n-2}(n-j)\{(n-j)^2-1)\}\gamma_j\Bigg] \quad (15)$$

where the $\{\gamma_k\}$ refer to the ARMA (p, q) that results from taking first differences. Finally, for any ARIMA (p, d, q) with $d>1$, we find that

$$E_k^{(n)}=(n-k)(n^2-2nk-k^2-1)/\{n(n^2-1)\}. \quad (16)$$

These formulae are all derived in Anderson (1979 *a*).

Extending the ideas to general ARUMA models, Anderson (1979 *b*) showed that:

(*a*) For (7) with $\Psi(B)=(1+B)$,

$$E_k^{(n)}=(-1)^k(1-k/n)+R(k, n) \quad (17)$$

where $R(k, n)$ is known and is of order n^{-2} or less.

(*b*) For $\Psi(B)=(12\ B\cos\omega+B^2)$ in (7), to the same degree of approximation,

$$E_k^{(n)}=(1-k/n)\cos k\omega. \quad (18)$$

(*c*) $E_k^{(n)}$ formulae corresponding to all other ARUMA models can be easily written down and in every case, provided $\Psi(B)\neq 1$ or $(1-B)^d$, the $E_k^{(n)}$ pattern "touches" the lines $\pm(1-k/n)$.

Similar results can be obtained for the variability of the sampled variances and covariances from all these models - see Andeson and De Gooijer (1983, 1988). For instance, corresponding to (14), which can be written in the form

$$E[nc_k^{(n)}]=\sum_{j=0}^{n-1}g_j\gamma_j,$$

we have, for the ARMA model, that

$$\mathrm{Var}\,[n^2\, c_k^{(n)}] = \sum_{i,\,j=0}^{n-1}\sum g_{i,\,j}\gamma_i\gamma_j,$$

where the $g_{i,\,j}$ are all known.

From formulae given in Anderson (1979 *b*, p. 290), we can then obtain approximate expressions for the means, variances and covariances of the serial correlations for ARUMA models. Alternatively, given a particular process, the moments and distributions for these serials can be obtained (to any required degree of accuracy), either "approximately" by empirical simulation (for instance, see Anderson and De Gooijer 1987) or "exactly" by computation, using the numerical integration approach due to Imhof (1961). Attempts made several years ago to obtain analytic expressions via Saddlepoint methods (following Daniels 1956) have however (for the present) been abandoned in favour of the other approaches, which, from a pragmatic point of view at any rate, appear to be preferable.

Note that (14) updates an earlier formula given by Kendall (1954), who used an alternative definition to (8) for the serial covariance.

It has also been suggested that explicit expressions, like (14), are of limited value since, for any particular model, it would be as easy to compute $E[nc_k^{(n)}]$ from the quickly derived formula of a matrix product trace [*see* Anderson 1979 *a*, equation (19)]. However, although we sympathise with this view to some extent, we believe our explicit formulae throw considerably more light, than do the intermediate trace results, on how the expected serial covariances behave for general models, lengths n and lags k, and they also avoid the need for rather elaborate programming before cheap numerical results can be achieved.

6. DISCRIMINATION PROCEDURE

For models (12) and (13), the serial correlation functions obtained from sample realisations of length 100 are typically fairly smooth with a characteristic shape that follows a sequence of decreasing terms which starts off positive, changes to negative and reaches a minimum before increasing again towards zero, although sampling error does frequently introduce a certain raggedness into this fundamental pattern. This basic shape can be considered as that of the $E[r_k^{(100)}]$ (or $E[r_{k,\,1}^{(100)}]$), given the particular generating process and series

length, since Anderson and De Gooijer (1987) have shown that these $r_k^{(100)}$ distributions are not very different from Gaussian ones.

It is suggested that $E_k^{(n)}$ is quite a reasonable approximation to the corresponding expected serial correlation, for many homogeneous nonstationary or nearly nonstationary stationary models, and so the theoretical characteristic patterns for our two models, (12) and (13), are assumed to be approximately known from (14) and (15). In particular, these shapes cross the axes at, say, points $k_{12,n}$ and $k_{13,n}$, respectively, where $k_{12,100}=20.46$ and $k_{13,100}=28.90$.

The feature we concentrate on, when dealing with an appropriate series, is where (from visual inspection) the sequence of serial correlations appears to cross from positive to negative. If the "cross-over" is nearer to $k_{12,n}$ than $k_{13,n}$, we choose model (12), otherwise we pick (13). Evidently this is a very naive decision rule; but, in Anderson and De Gooijer (1979), it was shown to give 80% correct discrimination. For instance, figure 2 shows a pair of such serial correlation plots together with the resulting inferences – both of which are correct.

7. AN INDICATION OF HOW THE APPROACH GENERALISES

The "autoregressive" factors $(1-\varphi B)$ and $(1-B)$ form, in fact, the most difficult basic pair of near-nonstationary and homogeneous nonstationary operators to distinguish. The other first order duo, $(1+\varphi B)$ and $(1+B)$, is generally extremely easy to correctly discriminate between. Amongst the quadratic factors, $(1-\alpha B)^2$ and $(1-B)^2$ form an easy pair to distinguish as indeed do other $(1+2\alpha B\cos\omega+\alpha^2 B^2)$ and $(1+2B\cos\omega+B^2)$, in general, including $(1+\alpha B)^2$ and $(1+B)^2$.

However, period-2 "seasonal" quadratic operators behave anomalously. $(1+B^2)$ behaves rather like a period-2 version of $(1+B)$, but $(1-B^2)=(1-B)(1+B)$ tends to reflect the behaviour of the more "dominant" $(1+B)$, *without seasonality*. Often the $(1-B^2)$ serials will quite closely resemble those from just a $(1+B)$; but also, frequently (if one does not warm up the simulation adequately), the $(1-B)$ component succeeds in reducing the (positive) even-lagged serials substantially, and the magnitudes of the (negative) odd lagged ones rather more, to give a serial correlation pattern intermediate between that characteristic of $1+B$ and seasonal period-2 Wichern-like behaviour analogous to that of $1-B$. (Evidently the two components reinforce, for even lags, but are in opposition at odd ones.)

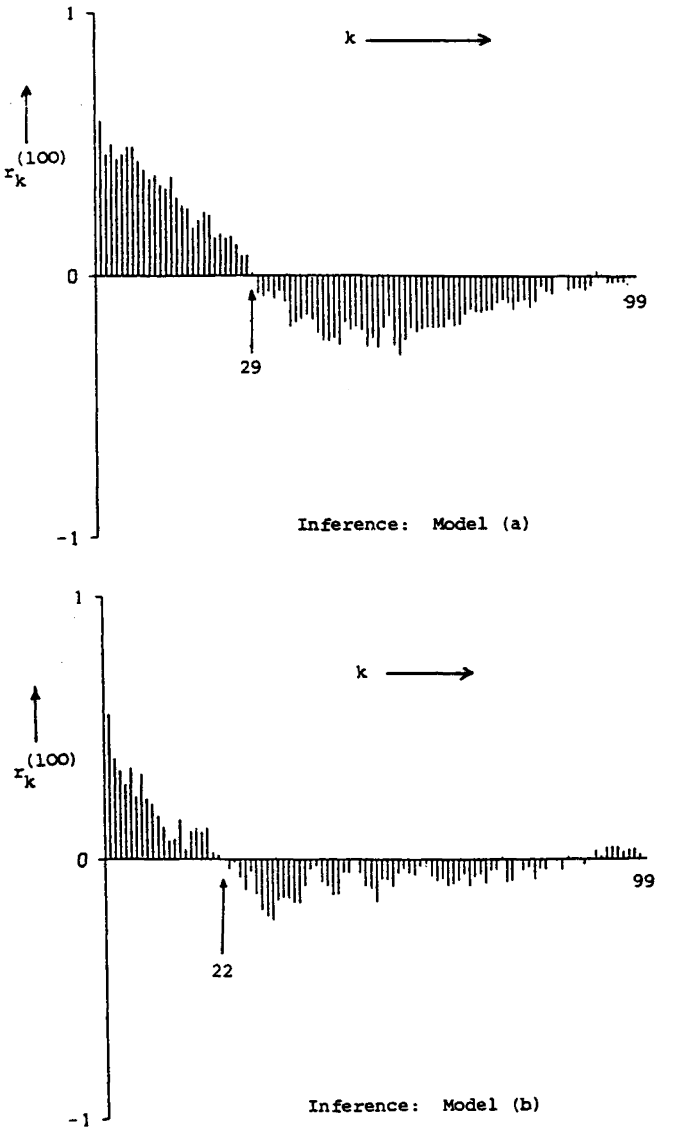

Figure 2. — Serial correlation plots for two length-100 simulations, each from an unseen choice of either model (*a*) $(1-B)\,Z_i=(1-.8\,B)\,A_i$ **or model** (*b*) $(1-.95\,B)\,Z_i=(1-.74\,B)\,A_i$.

In figures 3 to 9, we give some results for series realisations of length 100 from various models. These show that, for the nonstationary cases, the $E_k^{(n)}$ patterns are closely followed and that the nearly nonstationary models should

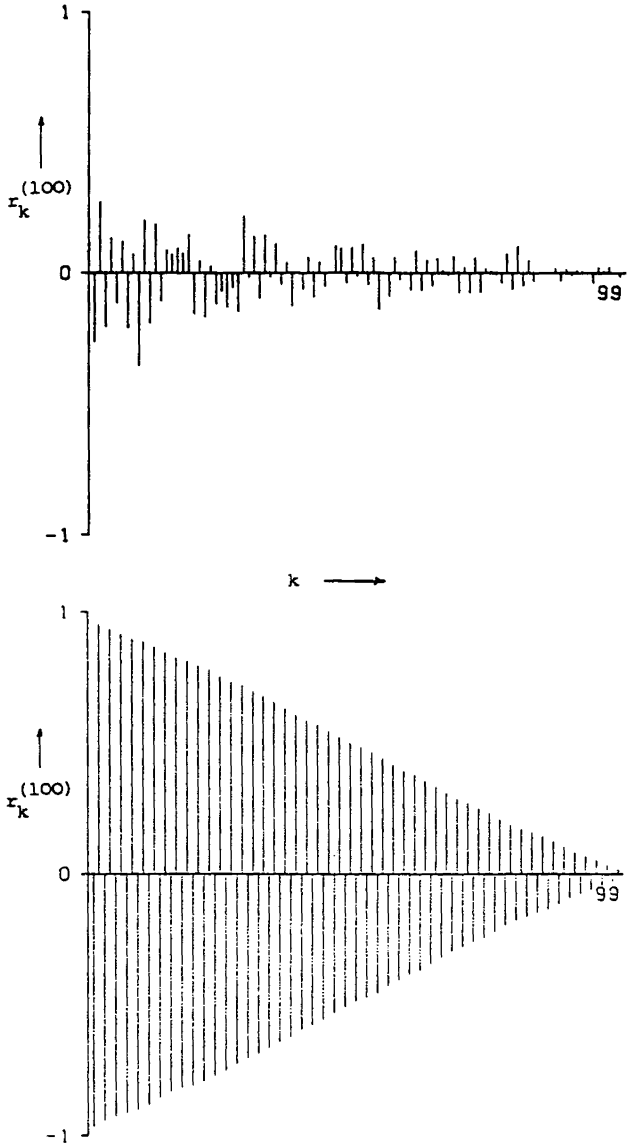

Figure 3. — Serial correlation plots for two length-100 simulations from models (top) $(1+.95B)Z_i=(1+.74B)A_i$ **and (bottom)** $(1+B)Z_i=(1+.8B)A_i$.

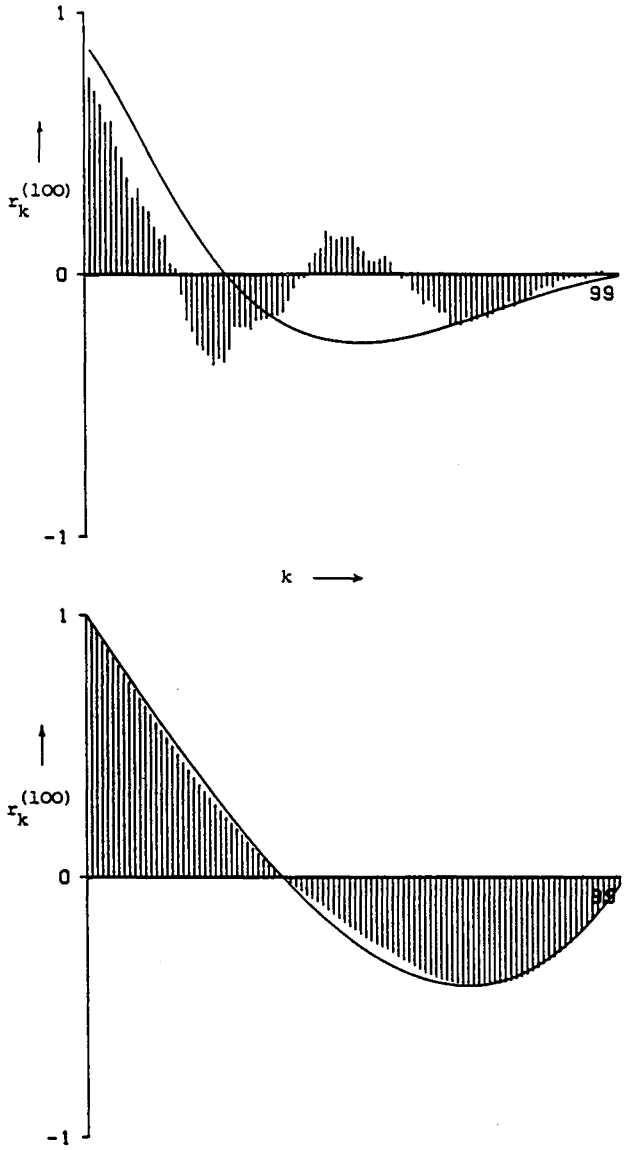

Figure 4. — Serial correlation plots for two length-100 simulations from models (top) $(1-.95\,B)^2\,Z_i=(1-.74\,B)^2\,A_i$ and (bottom) $(1-B)^2\,Z_i=(1-.8\,B)^2\,A_i$, with $E_k^{(100)}$ lines drawn in.

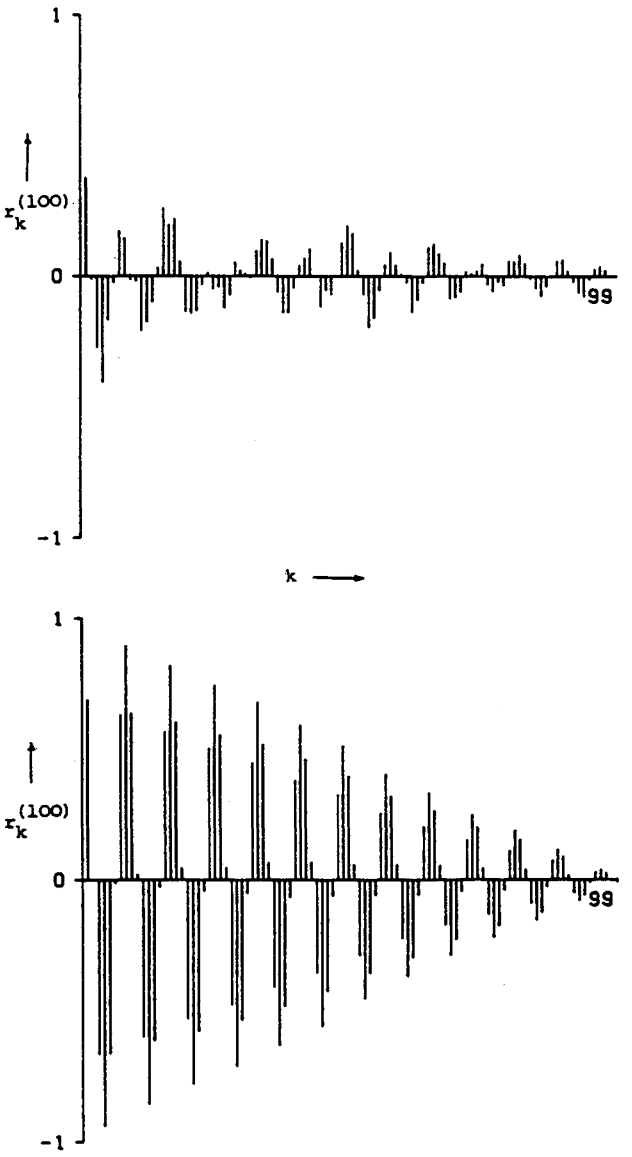

Figure 5. — Serial correlation plots for two length-100 simulations from models (top) $(1-1.9B\cos 45°+.902\,5B^2)Z_i=(1-1.48B\cos 45°+.547\,6B^2)A_i$ **and (bottom)** $(1-2B\cos 45°+B^2)\ Z_i=(1-1.6B\cos 45°+.64B^2)A_i$.

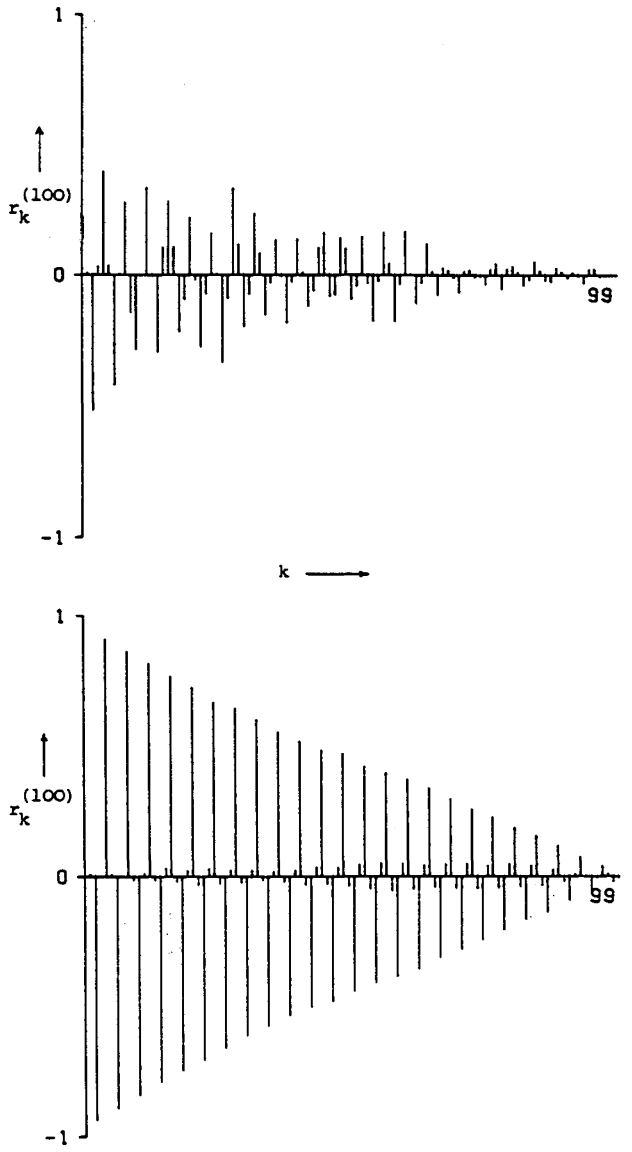

Figure 6. — Serial correlation plots for two length-100 simulations from models (top) $(1+.902\,5\,B^2)\,Z_i=(1+.547\,6\,B^2)\,A_i$ **and (bottom)** $(1+B^2)\,Z_i=(1+.64\,B^2)\,A_i$.

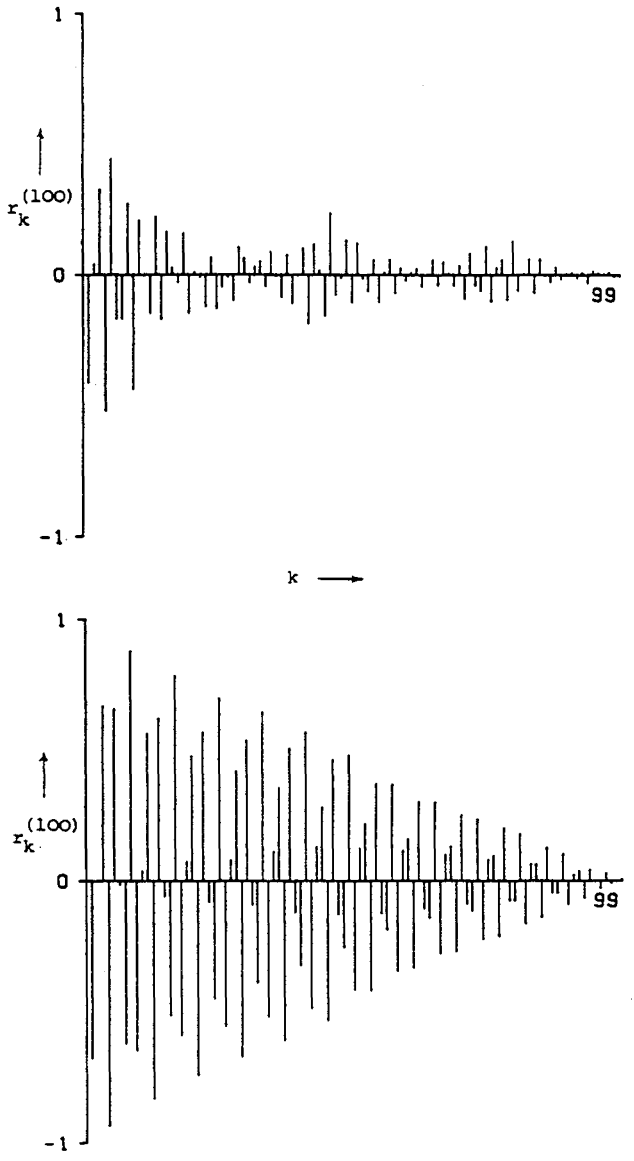

Figure 7. — Serial correlation plots for two length-100 simulations from models (top) $(1+1.9B\cos 45°+.902\,5B^2)\,Z_i=(1+1.48B\cos 45°+.547\,6B^2)\,A_i$ **and (bottom)** $(1+2B\cos 45°+B^2)\,Z_i=(1+1.6B\cos 45°+.64B^2)\,A_i$.

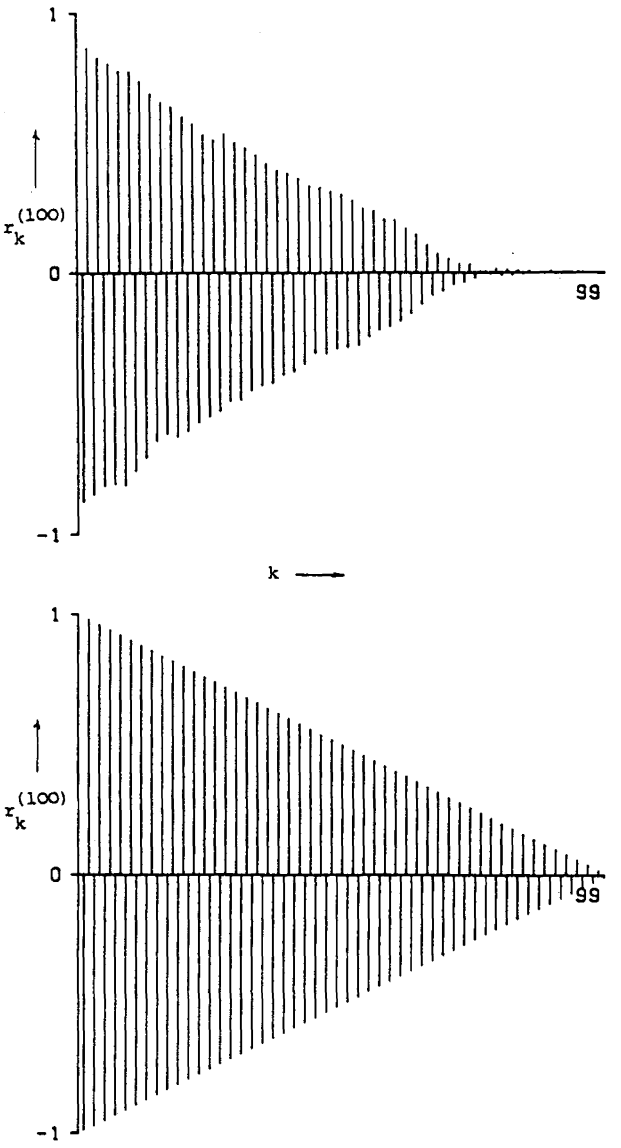

Figure 8. — Serial correlation plots for two length-100 simulations from models (top) $(1+.95\,B)^2\,Z_i=(1+.74\,B)^2\,A_i$ **and (bottom)** $(1+B)^2\,Z_i=(1+.8\,B)^2\,A_i$.

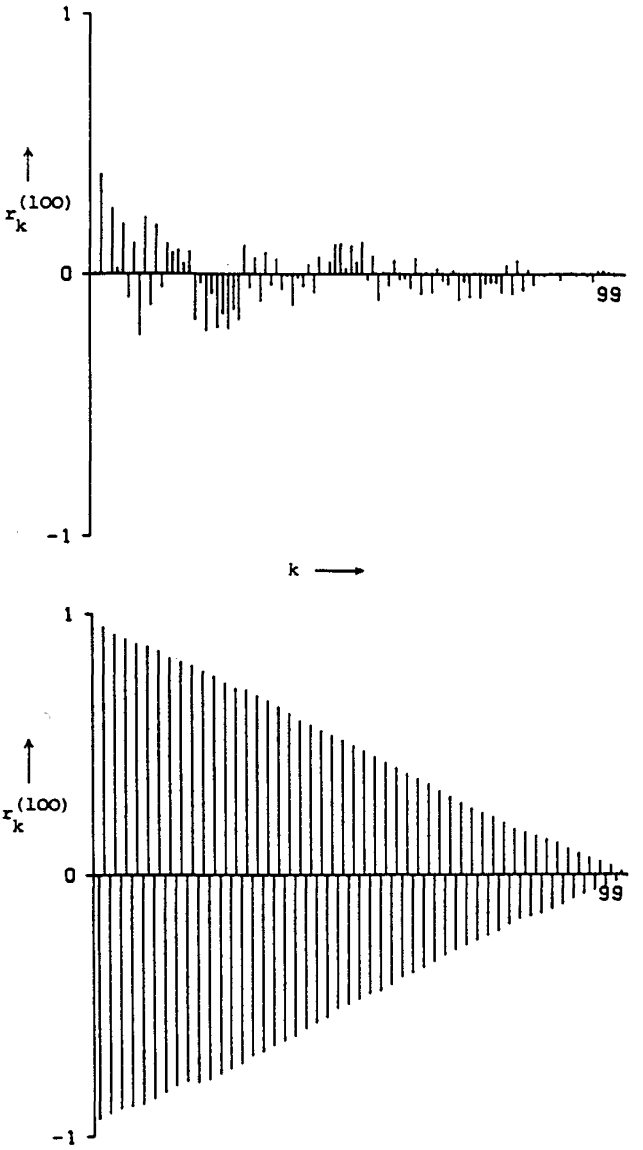

Figure 9. — Serial correlation plots for two length-100 simulations from models (top) $(1-.902\,5\,B^2)\,Z_i=(1-.547\,6\,B^2)\,A_i$ **and (bottom)** $(1-B^2)\,Z_i=(1-.64\,B^2)\,A_i$.

not be confused with them. With the stationary processes, the $E_k^{(n)}$ pattern is not necessarily so close to that of the $E[r_k^{(n)}]$, but the very much increased sampling variability makes this less important anyway. Basically, the difference in behaviour is: for a nonstationary model, a sample serial correlation plot in very close agreement with $E_k^{(n)}$ (hardly any bias or scatter); whilst, for a near-nonstationary one, the sampling fluctuations are very substantial and, also, about values considerably displaced from those of the nonstationary case.

It is important to realise that, when simulating nonstationary models, considerable warm up periods may be necessary. For this reason, we started up from cold and simulated 1,000 terms for a variety of models. We then split these runs into ten consecutive sub-series, each of length 100, and produced the corresponding serial correlation plots.

This enabled us to get some idea, empirically, of how long a warm up period is necessary. 400 terms appeared enough for all our runs (and the plots of figures 3 to 9 correspond to the fifth subseries from each run). We also used the same shocks for each model, so as to ensure that marked differences in behaviour were indeed due to the different models and not to peculiarities of the sampled driving shocks.

Of course, this type of design has severe drawbacks as far as arriving at conclusions (on how easy discrimination will be in practice) is concerned. For we have restricted ourselves to just ten sets of independent shocks, and so our results could evidently appear good by luck. Further (and perhaps worse), if a long warm up period is needed for a particular nonstationary model, then consecutive sub-runs (once warming up has been achieved) may be highly dependent.

Thus, if we judge warm up as having occurred when the serial correlation structure appears to have got close to the $E_k^{(n)}$ pattern, then perhaps staying close to this pattern is fairly likely in the short term (for the next few sub-runs), but the possibility of the structure slowly wandering away again is not precluded. However, we do believe that we shall eventually have all the apparatus necessary for answering such questions – and we are confident that our ideas will then be vindicated.

8. IMPROVED IDENTIFICATION

Of course, the situation of discriminating between just a pair of appropriate models is not very relevant, in practice, But we believe the ideas can be very usefully employed at the identification stage of Box-Jenkins analysis.

It is well-known that, when a process contains a $\Psi(B)=(1-B)$, little can be learnt (by conventional methods) from studying the serial correlations, except for the fact that some degree of series differencing is needed. We believe that $(1-\varphi B)$ factors are also frequently mistaken for $(1-B)$ ones; and that, by considering the cross-over and drawing on our experience, we can often spot when this has been done – although our work does not yet provide a formal test procedure. When a $(1-\varphi B)$ is indicated, we believe that the "cross-over" in the serial correlations can be used to obtain a fairly efficient rough estimate of φ, and we would then propose employing $(1-\hat{\varphi} B)$ as a simplifying operator – rather than $(1-B)$ – for purposes of investigating the remaining structure. To save time and computation, we suggest that the serial correlations for the simplified series be obtained approximately from those of the original series, through relation (11).

The same approach can be used to simplify models for which general operators $\Psi(\alpha B)$ or $\Psi(B)$ have been identified, these identifications following from observing a fairly smooth serial correlation function (taking into account possible alternating sign) not too far away from the lines $\pm(1-k/n)$ or a very smooth function virtually touching these lines, respectively. The degree of smoothness to be expected evidently depends on n, and decreases with shorter series.

We conclude our contribution with a reference to Anderson (1990) which gives a large number of examples of this augmented identification approach.

ACKNOWLEDGEMENT

The basic material of this paper was first presented at the series of International Time Series Meetings (ITSM). The authors are most grateful for helpful comments from Neville Davies, Jim Durbin, and Morris Walker.

REFERENCES

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

O. D. ANDERSON, Time series analysis and forecasting: a further look at the Box-Jenkins approach, *Cahiers du CERO*, Vol. 19, 1977 *a*, pp. 223-256.

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 *a*, 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 *b*, pp. 221-237.

O. D. ANDERSON, Some sample autocovariance function results for a once integrated q-th order moving average process, *Statistica*, Vol. 39, 1979 *c*, pp. 287-299.

O. D. ANDERSON, Identification of linear time processes: an augmented Box-Jenkins approach, Working Paper, 1990.

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, *The Statistician*, Vol. 28, 1979, pp. 119-133.

O. D. ANDERSON and J. G. DE GOOIJER, Formulae for the covariance structure of the sampled autocovariances from series generated by general autoregressive moving average processes of order (p, d, q), $d=0$ or 1. *Sankya B*, Vol. 45, 1983, pp. 249-256.

O. D. ANDERSON and J. G. DE GOOIJER, Discriminating between IMA (1,1) and ARMA (1, 1) models: a simulation study; Working Paper, 1987.

O. D. ANDERSON and J. G. DE GOOIJER, Sampled autocovariance and autocorrelation results for linear time processes, *Communications in Statistics* (Simulation and Computation), Vol. 17, 1988, pp. 489-513.

G. E. P. BOX and G. M. JENKINS, *Time Series Analysis, Forecasting and Control*, Holden Day, San Francisco (Revised Edition), 1976.

H. E. DANIELS, The approximate distribution of serial correlation coefficients, *Biometrika*, Vol. 43, 1956, pp. 169-185.

J. P. IMHOF, Computing the distribution of quadratic forms in normal variables, *Biometrika*, Vol. 48, 1961, pp. 419-426.

M. G. KENDALL, Note on bias in the estimation of autocorrelation, *Biometrika*, vol. 41, 1954, pp. 403-404.

D. W. WICHERN, The behaviour of the sample autocorrelation function for an integrated moving average process, *Biometrika*, Vol. 60, 1973, pp. 235-239.