J. R. Statist. Soc. A,
(1973), **135**, *Part* 3, p. 337

Some Comments on a Paper by Chatfield and Prothero and on a Review by Kendall

By G. E. P. Box and G. M. Jenkins

University of Wisconsin, Madison *University of Lancaster*

In a recent paper read to a meeting of the Society, which neither of us was able to attend, Chatfield and Prothero (1973) applied the "Box–Jenkins" approach to the forecasting of a particular series and were dissatisfied with the results they obtained. They included in their paper a large number of general comments, remarks and conclusions on the topic of forecasting which they believed to be supported by extrapolation from their unsatisfactory experience with one particular series. In fact, as was pointed out at the Society's meeting, their difficulties arose because they applied the wrong transformation to the data. Their conclusions are not, therefore, supported by their example.

We have written this note because we believe many of their comments to be misleading. In particular, the authors thought that they found in their anomalous results support for views, expressed by M. G. Kendall (1971) in a review of our book, which appeared in this *Journal*. We take this opportunity to comment on these views.

Analysis of the Chatfield–Prothero Data

Chatfield and Prothero justify their use of the logarithmic transformation by saying that the "trend lines for the 12 different months turn out to be roughly linear and parallel indicating that the additive seasonal pattern of the transformed data is reasonably stable". Plots of this kind are indeed a valuable aid in preliminary model identification but if these plots are examined more closely a different conclusion is reached. At this stage there is no need for elaborate methods and in Fig. 1 the "trend" for each month has been estimated simply as the difference between the 1970 value (the last year for which there is a complete set of data) and the 1965 value. These differences have been plotted against the monthly average over the same period. The plot is made for x, $x^{\frac{1}{2}}$, $x^{\frac{1}{4}}$ and $\log x$. It is at once evident that the log over-transforms, the square root under-transforms, while $x^{\frac{1}{4}}$ is about right. That this is so may be confirmed more formally by the likelihood technique (Box and Cox, 1964) whose use was suggested in our book. However, our point here is that it is also obvious by examining the single plots which the authors have themselves suggested. Wilson (1973), using the more refined analysis, obtains the more exact transformation $z = x^{\cdot 37}$ for the Chatfield–Prothero data. However, the approximate transformation $z = x^{\cdot 25}$ suggested by Fig. 1 gives essentially similar results and if we confine ourselves to a model with the same structure as Chatfield and Prothero's, we obtain

$$(1 + \cdot 5B)\, w_t = (1 - \cdot 8B^{12})\, a_t, \tag{1}$$

following the procedure they describe where

$$w_t = \nabla\nabla_{12} z_t = \nabla\nabla_{12} x^{\cdot 25} \quad \text{and} \quad \hat{\sigma}_a^2 = \cdot 02336.$$

As was demonstrated by Wilson (1973), after suitable transformation a model of this kind gives very good forecasts which do not suffer from the disabilities which

afflicted the model obtained from the overtransformed data. In particular, it does not produce forecasts with a seasonal amplitude which is obviously too high, nor is it sensitive to the forecast origin, characteristics which were displayed by the Chatfield–Prothero model. Furthermore, much narrower probability limits are now found than those given by Chatfield and Prothero for each lead time from a fixed forecast origin.

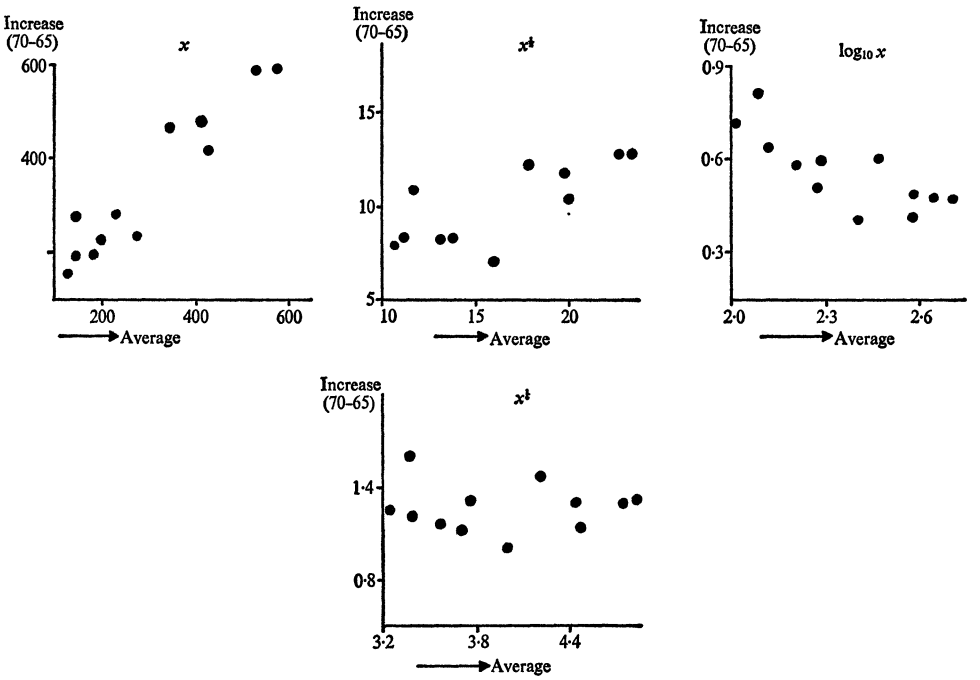

FIG. 1. Plots of increase between 1965 and 1970 against average value for each of the 12 months using various transformations.

Table 1 shows results obtained using model (1) for the origin at May 1971 which Chatfield and Prothero had found particularly troublesome. Table 1 should be compared with Table 6 in the Chatfield–Prothero paper. Also shown in Table 1 are the 50 per cent probability limits which sometimes more closely reflect what a business forecaster might regard as measuring the "probable" error of his forecast.

TABLE 1

Forecasts (with approximate probability limits) of x_{t+l} *made in May* 1971 *for lead times* 1 *to* 6

l:	1	2	3	4	5	6
Forecast	286	409	511	761	966	1091
95% Probability limits	±85	±128	±184	±280	±373	±420
50% Probability limits	±29	±44	±63	±96	±128	±145

What does the Model Mean?

One point made in the Chatfield–Prothero paper was that the ARIMA model was more difficult to understand than traditional models. On the contrary, we believe that these models are usually rather easy to understand and that it is always worth while to take the trouble to consider what they are trying to tell us. This is well illustrated by the model of equation (1) which may be written

$$(1+\cdot 5B)(1-B)(z_t-\bar{z}_{t-12}) = a_t, \tag{2}$$

where

$$z_t-\bar{z}_{t-12} = \frac{(1-B^{12})}{(1-\cdot 8B^{12})}z_t = \left(\frac{1-\cdot 8B^{12}-\cdot 2B^{12}}{1-\cdot 8B^{12}}\right)z_t = \{1-\cdot 2(B^{12}+\cdot 8B^{24}+\ldots)\}z_t$$

so that

$$\bar{z}_{t-12} = \cdot 20z_{t-12}+\cdot 16z_{t-24}+\cdot 13z_{t-36}+\cdot 10z_{t-48}+\ldots. \tag{3}$$

Thus if t refers say, to the month of May, $\bar{z}_{t-12}$ would be an average of all previous May results in which somewhat more weight† is given to recent history than to the past.

Now equation (2) may be written

$$(1-\cdot 5B-\cdot 5B^2)(z_t-\bar{z}_{t-12}) = a_t$$

or

$$z_t = \bar{z}_{t-12}+\tfrac{1}{2}\{(z_{t-1}-\bar{z}_{t-13})+(z_{t-2}-\bar{z}_{t-14})\}+a_t. \tag{4}$$

Equation (4) displays in a readily understandable form the forecasting rule which is tacitly being used here. It can be stated in words as follows: To forecast (say) next June's values, take an exponentially weighted average of previous June figures and adjust it by the average amount that this year's May and April figures (or their forecasts if we are forecasting more than one step ahead) exceeded last year's corresponding exponentially weighted moving averages.

This simple analysis makes it easy to see why it is important to choose the transformation carefully in this example. Chatfield and Prothero remark that the seasonal effect is remarkably regular. It is undoubtedly this regularity which is responsible for the choice of a value for $\Theta = \cdot 8$ in (1) which gives very slow decay of the weights in the average $\bar{z}_{t-12}$ in (3). For this particular series there is a great deal of currently useful information in the past and we are told to make use of it. However, (4) implies that the amplitude which is being projected into the forecast approximates a simple arithmetic average of the amplitude for the previous years. Over-transformation thus guarantees that the average will be too large at the "narrow end of the funnel".

Alternative Forecasting Methods

We are reminded in the Chatfield–Prothero paper (i) that the Box–Jenkins method is only one of many procedures (some of which were listed by the authors but none were tried and compared), (ii) that there is no such thing as "the best forecasting method" and (iii), quoting Kendall's (1971) statement, that we must decide for ourselves if the methods are an advance on more traditional approaches. We now discuss these comments.

† In the present circumstance where the estimate of the seasonal parameter is so close to unity, not one but at least three cycles of the forward and backward iteration discussed in Box and Jenkins (1970, p. 217) are needed to achieve convergence. A value of the seasonal parameter is then found close to ·9 yielding an even more uniform weighting.

It is difficult to imagine any systematic method of forecasting which does not imply belief in some kind of probability model for the data. Thus the man who wants to obtain forecasts by projecting regression lines fitted to the figures for each month implies belief† in a model of the form

$$(1-B^{12})^2 x_t = (1-\Theta B^{12})^2 a_t \tag{5}$$

with Θ slightly less than unity.

Again a decision to apply a simple exponentially weighted average with an arbitrarily chosen smoothing factor of say ·2 would imply the model

$$(1-B)\, x_t = (1-\cdot 8B)\, a_t. \tag{6}$$

Now equations (5) and (6) are both special cases of the class of seasonal ARIMA models we consider. This is also true of the traditional approach of "trend plus seasonal plus random component" model to be found in texts such as Croxton and Cowden (1955). Furthermore, it has recently been shown by Cleveland (1972) that even a process as complex as the Census Bureau X-11 method for Seasonal Adjustment due to Shiskin (1967) is implied very nearly by a not very complicated ARIMA model. Thus the alternative and traditional common-sense forecasting methods referred to in the Chatfield–Prothero paper are for the most part special cases of the ARIMA model. While they were candidates for selection, these other methods were not selected for this particular series. The forms of the model and imposed values of the parameters which these methods required were rejected for the very good reason that these models were inappropriate! Notice that the model (1) actually selected, although not empirically arrived at, yields the common-sense rule implied by equation (4) which might have been suggested empirically. However, it has the merit that among the multitude of empirical forecasting rules that make sense it will give good forecasts for the *particular* series under study.

The ARIMA Model

The ARIMA model derives from Yule, Slutsky and Yaglom. The reason for its success is probably that it makes sense theoretically to think of most time series z_t as behaving like the output from a dynamic system (not necessarily stable) subjected to random shocks a_t. Furthermore, it is natural and parsimonious to represent such a dynamic system in terms of a linear difference equation.

$$\left.\begin{aligned} \phi(B)\nabla^d z_t &= \theta(B)\, a_t, \\ \varphi(B)\, z_t &= \theta(B)\, a_t. \end{aligned}\right\} \tag{7}$$

The model (7) has extreme flexibility in that the projection (i.e. the forecast function) can be any polynomial, exponential, sine–cosine function or mixture of these functions. Furthermore, the manner in which the forecast uses past data is extremely flexible.

In general we can write $\theta^{-1}(B)\varphi(B)\, z_t = a_t$ as $\{1-\pi(B)\}\, z_t = a_t$. Thus

$$z_t = (\pi_1 z_{t-1} + \pi_2 z_{t-2} + \ldots) + a_t \tag{8}$$

and the forecast at lead l is

$$\hat{z}_t(l) = \sum_{j=1}^{\infty} \pi_j \hat{z}_{t+l-j} \tag{9}$$

† The adherent to B.F.E. (Bold Freehand Extrapolation) betrays a similar latent commitment.

with

$$\tilde{z}_{t+l-j} = \begin{cases} \hat{z}_t(l-j) & \text{if } l-j>0, \\ z_{t+l-j} & \text{if } l-j\leqslant 0. \end{cases} \tag{10}$$

Chatfield and Prothero refer to the "difficulty of distinguishing observationally between different autoregressive and moving average models", again referring to Kendall (1971) who was troubled lest "several models fit the data equally well". To illustrate, consider the first-order moving average model $z_t = (1-\cdot 20B)\,a_t$ which is equivalent to the infinite autoregressive model

$$(1-\cdot 20B)^{-1} z_t = a_t = (1+\cdot 20B+\cdot 04B^2+\ldots)\,z_t = a_t. \tag{11}$$

This model is very nearly the same as the first-order autoregressive model

$$(1+\cdot 20B)\,z_t = a_t. \tag{12}$$

How can the two models be distinguished? For series of lengths we are likely to examine, the answer is, of course, that they cannot. But then they do not need to be. Either model would fit the data equally well and it is clear that it does not matter which we use. These models, like every other, are approximations. When two approximations are essentially equivalent we can use either and obtain essentially the same result.

Following this kind of argument, however, Kendall (1971) suggests that "we might as well be content with autoregressive series...taking [their order] as far as is necessary to give approximate independence in the residuals...". We disagree. It is often the case (and was the case in the data of the Chatfield–Prothero paper) that the moving average parameters are not small. For simplicity let us use the first-order moving average example again but suppose now that the parameter has a value of ·8. Then

$$z_t = (1-\cdot 80B)\,a_t. \tag{13}$$

This is equivalent to the infinite autoregressive model

$$(1+\cdot 80B+\cdot 64B^2+\cdot 51B^3+\cdot 41B^4+\cdot 33B^5+\cdot 26B^6+\cdot 21B^7+\ldots)\,z_t = a_t. \tag{14}$$

We believe the practitioner will appreciate that to use a model based on (14) with, say, seven unknown parameters when (13) is available with only one parameter is to look for trouble.

Considering again equations (8) and (9), we see that for an autoregressive process of order p the forecast memory of past values abruptly stops after p terms. For example, a second-order autoregressive process cannot remember more than the last two terms of the series in forecasting the next term. Whereas this might be appropriate for special cases, it seems to be unreasonable in general. Usually, we would expect the forecast memory to die out slowly and if this is to be achieved with reasonably few parameters, we require to use moving average terms and to impose the condition of invertibility.

For an opinion of invertibility, Chatfield and Prothero again refer to Kendall (1971) who dislikes it and says that a non-invertible series can be "stationary and [therefore ?] thoroughly respectable".

It was surely the supposed primacy of autoregressive models and of stationarity that led the workers in traditional time series analysis for many years to overlook the very great possibility of such simple models as (6) which produce the extremely useful

exponentially weighted average as a best forecast for a model that contains only one parameter. Faced with a virtual vacuum if was left to operation research workers such as Winters and Holt to devise empirically useful methods for business and economic forecasting which imply models like (6) and which we have developed further. One can often find out how nature behaves by studying the methods that have evolved to deal with that behaviour. As we have pointed out earlier (1970), the modes of control (proportional, integral and differential) which have slowly evolved from James Watt's first governor again imply non-stationary disturbances in which the difference series is a moving average of some kind. Whether respectable or not, it is non-stationary and invertible models which are most often needed for forecasting and for representing disturbances in control problems.

There are other points in Kendall's review which are not specifically referred to in the Chatfield–Prothero paper but which we now discuss.

Differencing to induce Stationarity

When applied to non-stationarity series, differencing can often induce stationarity. This permits us to do our model building *via* the appropriately differenced stationary series. Kendall (1971) is disturbed by certain aspects of the differencing process. He points out that "the autocorrelation between members of a first difference, say σ_k are expressible in terms of those of the original series

$$\sigma_k = (\Delta^2 \rho_{k+1})/(1 - \rho_1). \tag{15}$$

Consequently irregularities in the ρ's are enhanced in the σ's and it becomes correspondingly more difficult to assign a type to observed correlograms. This may be the reason why Box and Jenkins say that d [the order of differencing] rarely exceeds 2." The sentence which begins with the word "Consequently" is a *non sequitur*. The series obtained after differencing have sampling properties which are the same as those for any other time series. Their approximate variances and covariances are, for example, given by Bartlett's well-known formulae. The existence of the relationship (15), which is not used, has nothing to do with the case.

Independently of the argument above, it is surely well known that for many economic series the popular assumption that errors are independent and identically distributed is much more likely to be approximately true for the first difference than for the original series. It is the *failing* to difference rather than the differencing which has sometimes led to error (see, for example, Coen *et al.*, 1969 and Box and Newbold, 1971).

The reason that d rarely exceeds 2 is that, for example, if d were 3, the resulting series would be the sum of the sum of the sum of a sequence of stationary random variables. Such series typically make vast excursions and are extremely smooth in a manner rarely characteristic of situations which are of interest.

Model Building

A statistical model of the form (7) is a transformation of data to white noise, that is a completely uncorrelated series. Such a model is justified on the grounds that since white noise is informationless the transformation itself must embody all the available information. We move towards such a model by the process of identification, fitting and diagnostic checking used *iteratively*. We make no claim of originality for the above for it is surely what any practising statistician has always done and must do to build realistic models. Identification is an informal technique of roughly matching

with the data a subclass of models which make some theoretical sense. Thus the plots in our Fig. 1 are part of the identification process. We emphasize in our book the part involving the matching of theoretical autocorrelations of models with the sample autocorrelations of the data. While, as Kendall (1971) has pointed out, sample autocorrelations can sometimes be misleading, we have not found this troublesome in the iterative context where no irrevocable decisions are being made at the identification stage.

There seems to be some misunderstanding about our proposal for the next stage of model building. After remarking on the problem of estimating the two sets ϕ (of autoregressive) and θ (of moving average) coefficients, Kendall (1971) says that "The procedure [we adopt in our book] essentially due to Durbin is to iterate, guessing values of one set and estimating the other, improving estimates of the first and so on (hopefully) to convergence". However, as any reader of our book will find, we do not use or even mention this method. The fitting procedure we employ directly maximizes the approximate likelihood *via* the Gauss least-squares iteration.

Spurious Correlations

In their paper, Chatfield and Prothero state that the high autocorrelation obtained in lags 10 and 11 of $\nabla\nabla_{12}(\log_{10} x_t)$ is "rather unexpected". In an attempt to throw some light on this phenomenon, they make the point in the Appendix to their paper (Case A) that if the original series $z_t = \log_{10} x_t$ consists of a white noise error $n_t = a_t$ superimposed on a linear trend plus a seasonal factor, then $w_t = \nabla\nabla_{12} z_t$ will have non-zero autocorrelations at lags 1, 11, 12 and 13. For reasons we are unable to understand, this exercise is supposed to explain the correlation at lag 11 which was found in their example and is said to be spurious. In fact, of course, if z_t had uncorrelated errors, the identification process we have suggested would lead precisely to the correct model. For we should find approximately that

$$(1-B)(1-B^{12})z_t = (1-B)(1-B^{12})a_t$$

which on integration yields the model assumed by Chatfield and Prothero, namely

$$z_t = a+bt+s_t+n_t.$$

Bias

Chatfield and Prothero have talked of the forecast and its possible bias. It is important to realize that given the assumptions, the procedures being discussed produce not just a forecast, but the complete conditional distribution $p(z_{t+l}|z_t, z_{t-1}, \ldots)$ of the future observation z_{t+l} (see, for example, Box and Jenkins, 1970, p. 137).

Now a data-based transformation, such as that employed by Wilson (1973) will find that metric which most nearly reproduces the (Normal) model form. The conditional distribution $p(z_{t+l}|z_t, z_{t-1}, \ldots)$ with $z = x^{.37}$ will thus be approximately Normal with known mean and known standard deviation. From this we can calculate, if we wish, the approximate distribution $p(x_{t+l}|x_t, x_{t-1}, \ldots)$ of the original $x_{t+l} = z_{t+l}^{2.7}$. As we might expect from the nature of the data this distribution will be rather skewed. The best way to tell our client what the data say about a future observation x_{t+l} would be to show him this whole distribution and, given that we are using a computer anyway, this is not difficult† to do. If he needs some summary values the latter may

† Complaints that methods are complicated, and expensive in time and money, must be balanced against the fact that while it does take some effort to learn new procedures, poor forecasting is usually too expensive to tolerate.

be computed in any way that he finds useful. For example, percentiles of the conditional distribution could be quoted. He may want the most probable value (which will, of course, be the transformed forecast $\{\hat{z}_t(l)\}^{2\cdot7}$) or he may want the conditional mean of x_{t+l} which, of course, will *not* be $\{\hat{z}_t(l)\}^{2\cdot7}$. This is so for the very good reason that the mean of a skewed distribution is not equal to its mode. It seems to us quaint to refer to this discrepancy as bias.

Other Applications

The authors refer to our methods as "a new, powerful but rather complicated procedure, which is as yet relatively untried" and state that the only published example for seasonal series they have seen concerns the airline data. To our knowledge there are a large number of applications of these methods. In addition to the extensive comparisons involving over a hundred time series made by Reid (1969, 1971) and Granger and Newbold (1972), we have applied seasonal methods to the forecasting of petroleum products, pharmaceuticals, newsprint, cement, foodstuffs, carpets, steel, tonnage chemicals, etc. Published accounts include:

(1) the forecasting of telephone installations and removals (Thompson and Tiao, 1971);
(2) the forecasting of U.S. Money Supply and Bank Deposits (Cramer and Miller, 1972);
(3) the forecasting of demand for telephones (Tomasek, 1972).

Conclusions

We have emphasized in our book the importance of model parsimony. While we would certainly not wish to claim that it will always be possible, there seem to be many examples where, provided that an appropriate transformation of the data is used, the number of model parameters needed is small.

The Chatfield–Prothero paper and its discussion serve a useful purpose in underlining the point that considerable care may be needed in choosing a metric when using such parsimonious models.

ACKNOWLEDGEMENT

The authors are indebted to Greta Ljung and Pat Ashworth for assistance with the calculations made in this paper. One of us (G.E.P.B.) was supported by Air Force Office of Scientific Research under Grant AFOSR72–2363.

REFERENCES

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc.* B, **26**, 211–243.

BOX, G. E. P. and JENKINS, G. M. (1970). *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day.

BOX, G. E. P. and NEWBOLD, P. (1971). Some comments on a paper of Coen, Gomme and Kendall. *J. R. Statist. Soc.* A, **134**, 229–240.

CHATFIELD, C. and PROTHERO, D. L. (1973). Box–Jenkins seasonal forecasting: problems in a case study (with Discussion). *J. R. Statist. Soc.* A, **136**, 295–336.

CLEVELAND, W. P. (1972). Analysis and forecasting of seasonal time series. Unpublished Ph.D. Thesis, University of Wisconsin, Madison.

COEN, P. G., GOMME, E. D. and KENDALL, M. G. (1969). Lagged relationships in economic forecasting. *J. R. Statist. Soc.* A, **132**, 133–152.

CRAMER, R. H. and MILLER, R. B. (1972). Development of a deposit forecasting procedure for use in banks. Tech. Report No. 312, Department of Statistics, Uuiversity of Wisconsin, Madison. (To appear in *J. Banking Res.*)

CROXTON, F. E. and COWDEN, D. J. (1955). *Applied General Statistics*, 2nd edn. Englewood Cliffs, New Jersey: Prentice-Hall.
GRANGER, C. W. J. and NEWBOLD, P. (1970). Economic forecasting—the atheist's viewpoint. Nottingham University Forecasting Project, Note 11.
KENDALL, M. G. (1971). Book review. *J. R. Statist. Soc.* A, **134**, 450–453.
REID, D. J. (1969). A comparative study of prediction techniques on economic data. Ph.D. Thesis, University of Nottingham.
—— (1971). A comparison of forecasting techniques on economic time series. Paper given at conference organized by the Society for Long Range Planning and Forecasting Study Group of the Operational Research Society.
SHISKIN, J., YOUNG, A. H. and MUSGROVE, J. C. (1967). X–II variant of Census Method II seasonal adjustment program. Technical report no. 15, Bureau of Census, U.S. Dept. of Commerce.
THOMPSON, H. E. and TIAO, G. C. (1971). Analysis of telephone data. *Bell J. Econ. Manag. Sci.*, **2**, 514–541.
TOMASEK, O. (1972). Statistical forecasting of telephone time series. *ITO Telecommunication Journal, Geneva*, December 1972, 1–7.
WILSON, G. T. (1973). Contribution to discussion of "Box–Jenkins seasonal forecasting: problems in a case study", by C. Chatfield and D. L. Prothero. *J. R. Statist. Soc.* A, **136**, 315–319.

A REPLY BY DR CHATFIELD AND DR PROTHERO

The paper by Box and Jenkins suggests that many of the comments in our paper are misleading. We disagree.

We still believe that the broad conclusions of our paper *are* justified. Some of their points have already been dealt with in our reply to the discussion. We will now deal with the remainder, indicating general areas of agreement and disagreement and we will show that some of their claims are unjustified. In fact consideration of their remarks has drawn our attention to further problems with the Box–Jenkins procedure which require investigation.

The Transformation

Box and Jenkins claim that our results are invalid because we have used the wrong transformation. In reply we would like to add the following points to those already made in our reply to the discussion.

(a) Most of the remarks we made in Section 12 of Chatfield and Prothero (1973) do *not* depend on whether or not we made the correct transformation in the paper. Most of the difficulties we described were also present in the second analysis which we discussed verbally at the meeting, even though no transformation was applied and "reasonable" forecasts were obtained.

(b) Box and Jenkins attempt to show that our transformation was inappropriate by constructing their Fig. 1 in which they effectively estimate trend by comparing the first year's data with the 1970 values. This is not an efficient way of estimating trend. If, for example, we use this sort of approach to compare the second year's (1966) data with the 1970 data, we obtain the results shown in our Fig. 1. This appears to indicate that the logarithmic transformation is better than the fourth-root transformation, but is, of course, no more reliable than the Box–Jenkins Fig. 1.

The complete graph of the monthly logarithmic values is shown in Fig. 2 and we think that most statisticians would agree that the lines are roughly parallel, except possibly for the first year's data, and would therefore have been satisfied with the logarithmic transformation.

(c) The approach of Box and Jenkins, as followed by us and by Thompson and Tiao (1971), is to choose the transformation so as to make the seasonal pattern additive.

But Box and Cox (1964) point out that other, possibly more important, purposes of a transformation are to make the error variance constant and to make the error distribution normal. Making the seasonal pattern additive will only make the error variance constant if the error standard deviation for different months is proportional to the size of the month's seasonal effect. Thus the above method, while much simpler than the Box–Cox procedure, is only an approximate way of making the error variance constant.

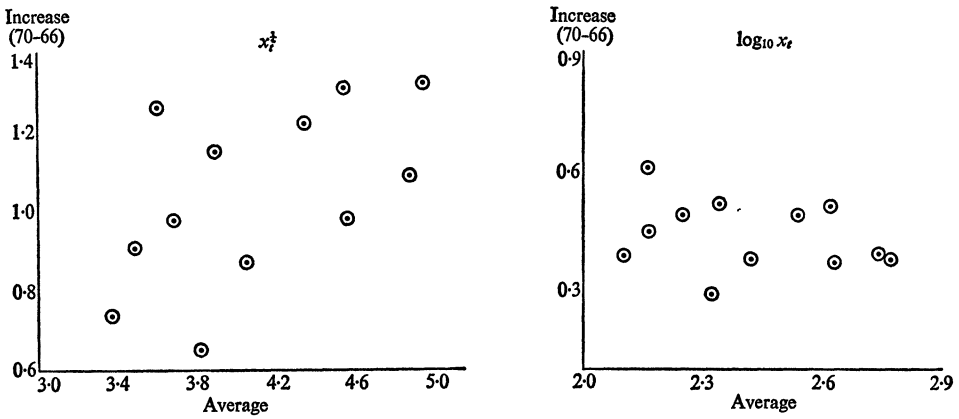

FIG. 1. Plots of increase between 1966 and 1970 against average value for each of the 12 months using fourth-root and logarithmic transformations.

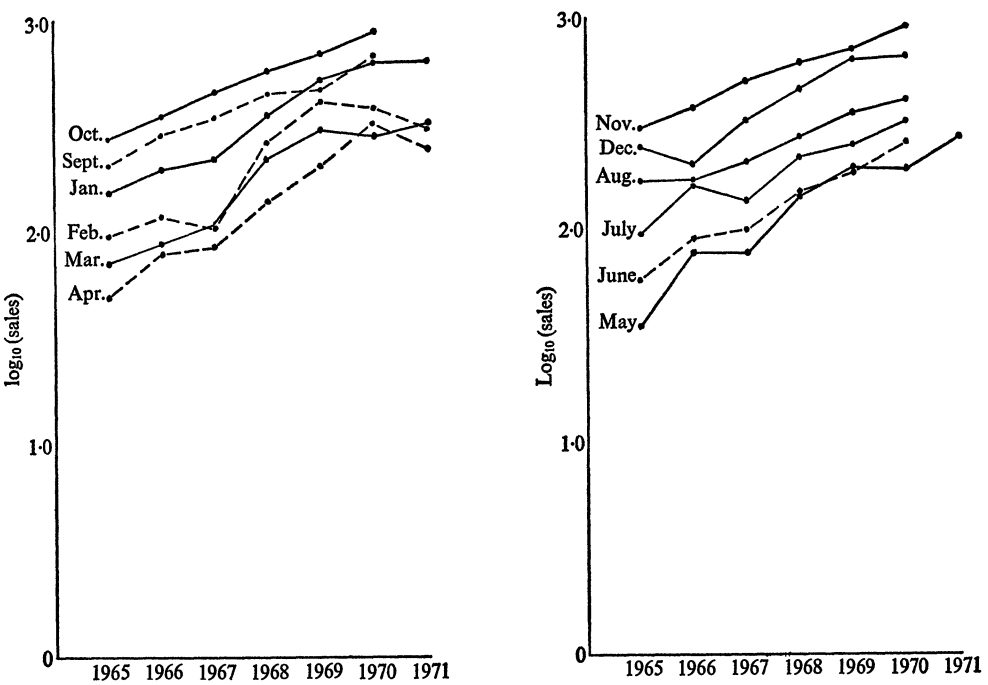

FIG. 2. Log_{10} (sales) for each of the 12 months between 1965 and May 1971.

(d) The more precise, but complicated, method of evaluating the transformation is by the method of Box and Cox. Using the first 60 observations, Dr Wilson estimated $\hat{\lambda} = 0{\cdot}34$ for our model *A*. We repeated this procedure using all the data and found $\hat{\lambda} = 0{\cdot}24$. Finally we excluded the first year's data, which seem somewhat "untypical", and found $\hat{\lambda} = 0{\cdot}16$ with a 95 per cent confidence interval which includes $\lambda = 0$, corresponding to the logarithmic transformation. In view of all these results, we still maintain that the logarithmic transformation was reasonable in the light of what is written in Box and Jenkins (1970). Nevertheless, as we agree that the fourth-root transformation gives more reasonable forecasts, it appears that our mistake was to give too little weight to data collected six years before the forecast was made!

(e) We also estimated the Box–Cox transformation parameter for Dr Wilson's model involving ∇^2_{12}, both using the first 60 observations, and also using all 77. The estimates we obtained (using over 40 minutes of computing time!) were $\hat{\lambda} = 0{\cdot}32$ and $\hat{\lambda} = 0{\cdot}24$, which are very close to those obtained for our model *A*. It therefore seems strange, at first sight, that Dr Wilson should analyse the *untransformed* observations. But we now take the view that this *is* the best approach and hope the following remarks will clarify the situation.

(f) We have seen that a "small" change in λ from 0 to 0·25 has a substantial effect on the resulting forecasts from model *A* even though the goodness of fit does not seem to be much affected. This reminds us that a model which fits well does not necessarily forecast well. Since small changes in λ close to zero produce marked changes in forecasts, it is obviously advisable to avoid "low" values of λ, since a procedure which depends critically on distinguishing between fourth-root and logarithmic transformations is fraught with peril.

(g) On the other hand a "large" change in λ from 0·25 to 1 appears to have relatively little effect on forecasts. So we conjecture that Box–Jenkins forecasts are robust to changes in the transformation parameter away from zero.

(h) If a transformation is used in conjunction with difference operators, then non-linear growth is obtained when forecasts are transformed back (cf. Professor Harrison's remarks on pp. 319–324).

(i) The Box–Cox procedure is very complicated when applied to ARIMA models.

(j) For all these reasons, but *not* for the reasons given by Box and Jenkins, we are now of the opinion that the logarithmic transformation should generally be avoided. Nor would we use the fourth-root transformation suggested by Box and Jenkins. Our approach now would always be to analyse the *untransformed* observations, except possibly in exceptional circumstances which need to be determined. If the seasonal effect is approximately multiplicative (as for our data), then we would use the operator ∇^2_{12}, since this removes a linear trend and multiplicative seasonal pattern. If the seasonal effect is approximately additive, we would use $\nabla\nabla_{12}$, since this removes a linear trend and additive seasonal pattern.

What does the Model Mean?

The reader must decide for himself if the explanation given by Box and Jenkins of their model (1) is "easy to understand". The reader must also decide if he would have thought of such an explanation by himself.

We were initially very puzzled by their equation (3) in the draft version available to us which can be written

$$\bar{z}_{t-12} = 0{\cdot}2(0{\cdot}8z_{t-12} + 0{\cdot}8^2 z_{t-24} + \ldots)$$

so that, although the weights are geometric, they do *not* sum to one. However, we then found an algebraic error in the previous line so that their equation (3) should read

$$\begin{aligned}\bar{z}_{t-12} &= 0{\cdot}2(z_{t-12} + 0{\cdot}8z_{t-24} + 0{\cdot}8^2 z_{t-36} + \ldots) \\ &= 0{\cdot}2z_{t-12} + 0{\cdot}16z_{t-24} + 0{\cdot}13z_{t-36} + \ldots . \end{aligned} \quad (1)$$

No doubt this mistake will be corrected by the time their paper appears in print. However, it does illustrate the difficulty of dealing with ARIMA-type equations.

It is not, in any case, clear how one would adapt their method to deal with some other ARIMA models.

It is also rather misleading to give their equation (3) in a form which implies that an infinite amount of data are available. If six years' data are available, then the updating equation

$$\bar{z}_{t-12} = 0{\cdot}2z_{t-12} + 0{\cdot}8\bar{z}_{t-24}$$

used recursively gives us

$$\bar{z}_{t-12} = 0{\cdot}2z_{t-12} + 0{\cdot}16z_{t-24} + 0{\cdot}13z_{t-36} + 0{\cdot}1z_{t-48} + 0{\cdot}41\bar{z}_{t-60}. \quad (2)$$

If backcasting is not used and we set $\bar{z}_{t-60} = z_{t-60}$, then this equation obviously gives too much weight to the first year's observation. Estimating $\bar{z}_{t-60}$ by backcasting we find

$$\bar{z}_{t-60} = 0{\cdot}2z_{t-60} + 0{\cdot}16z_{t-48} + 0{\cdot}13z_{t-36} + 0{\cdot}1z_{t-24} + 0{\cdot}41z_{t-12}$$

and when this is substituted into equation (2) we find

$$\bar{z}_{t-12} = 0{\cdot}37z_{t-12} + 0{\cdot}20z_{t-24} + 0{\cdot}18z_{t-36} + 0{\cdot}17z_{t-48} + 0{\cdot}08z_{t-60}.$$

This equation, using just one cycle of forward and backward iteration, bears little resemblance to equation (1) and can no longer be interpreted as an exponentially weighted moving average. This obviously affects the remarks made by Box and Jenkins following their equation (4).

Alternative Forecasting Methods

Box and Jenkins state that "alternative and traditional common-sense forecasting methods...are for the most part special cases of the ARIMA model". This may be true in one sense, but it does *not* follow that Box–Jenkins forecasts will necessarily be as good as other forecasts. For example the Holt–Winters method gives better forecasts than the Box–Jenkins method for about one-third of all time-series despite all the extra effort involved in the Box–Jenkins procedure (Granger and Newbold, 1972). And for regression models it is easy to demonstrate that the Box–Jenkins procedure is less efficient than traditional methods.

Let us consider the simplest possible regression model

$$x_t = a + bt + a_t$$

consisting of a linear trend plus random error. The traditional statistician would simply fit a straight line by least squares and extrapolate. The Box–Jenkins man

would try to fit the equivalent ARIMA model

$$(1-B)\,x_t = b+(1-B)\,a_t.$$

There are two difficulties with this model. Firstly it is *not* invertible. So one would have to try and fit the model

$$(1-B)\,x_t = b+(1-\theta B)\,a_t$$

with θ "slightly less than unity". Secondly the series

$$w_t = \nabla x_t$$

has a non-zero mean owing to the deterministic trend (Box and Jenkins, 1970, p. 92). So Box and Jenkins (1970, p. 210) suggest that

$$\bar{w} = \sum_{t=1}^{n} w_t/n$$

is substituted for the mean. Then the Box–Jenkins estimate of x_{n+k} made at time n is

$$\hat{x}_n(k) = x_n + k\bar{w} - \hat{\theta}\hat{a}_n$$

and the forecasts lie along a straight line with slope $\bar{w}$. Now if we have $(n+1)$ observations available, $x_0, \ldots, x_n$, then it is easy to show that

$$\bar{w} = (x_n - x_0)/n$$

so that the Box–Jenkins procedure is effectively estimating the trend by the inefficient procedure of drawing a straight line between the first and last observations.

For quadratic regression, the Box–Jenkins procedure gives even more surprising results. After differencing twice, it is easy to show that the estimate of the quadratic coefficient depends only on the first two and last two observations.

For linear regression with a multiplicative deterministic seasonal component we have

$$x_t = (a+bt)\,s_t + a_t$$

which, when differenced, gives

$$(1-B^{12})^2\,x_t = (1-B^{12})^2\,a_t. \tag{3}$$

This equation is not invertible so Box and Jenkins suggest their equation (5). But as $(1-B^{12})^2$ appears on both sides of the above equation we suspect that it is better not to difference at all. Using Box and Jenkins's equation (5), the estimate of say next June's observation would depend only on previous June figures whereas the regression forecast would involve all the observations.

An example of the sort of data where we expect differencing *will* be beneficial are the data given by Box and Jenkins (1970, p. 410) which show non-stationarity of the kind illustrated by Box and Jenkins (1970, p. 91). No regression model seems appropriate and we *would* expect the Box–Jenkins procedure to do well compared with other methods.

Invertibility

The discussion about invertibility does not directly concern our paper, but it may well be that some readers are still not sure why Box and Jenkins restrict themselves to invertible models. We will therefore try to clarify the situation by means of an example.

Consider the following first-order moving average processes, where $\{a_t\}$, $\{a'_t\}$, are white noise processes:

$$w_t = a_t + \tfrac{1}{2}a_{t-1}, \tag{4}$$

$$w_t = a'_t + 2a'_{t-1}. \tag{5}$$

It can easily be shown that both these processes have exactly the same autocorrelation function, given by

$$\rho_k = \begin{cases} 1, & k = 0, \\ 0{\cdot}4, & k = \pm 1, \\ 0, & \text{otherwise.} \end{cases}$$

If we further choose $V(a'_t) = \frac{1}{4}V(a_t)$, then the two processes have identical properties. For example, they both have the same spectral density function. If we generate a set of data from either model, then using the Box–Jenkins procedure we would calculate the observed autocorrelation function and find $r_1 \simeq 0{\cdot}4$ and $r_k \simeq 0$ for $k > 1$. This indicates that a first-order moving average process of the form

$$w_t = a_t + \theta a_{t-1} \tag{6}$$

is appropriate, but we are unable to distinguish between models (4) and (5). Box and Jenkins (1970) therefore suggest that we choose model (4) because it is invertible in that if we rewrite the equation as

$$w_t/(1 + \tfrac{1}{2}B) = a_t = w_t - \tfrac{1}{2}w_{t-1} + \tfrac{1}{4}w_{t-2} - \ldots,$$

we can express a_t as a convergent series of past values of w_t. However, for model (5) we get

$$a_t = w_t - 2w_{t-1} + 4w_{t-2} \ldots, \tag{7}$$

which gives progressively more weight to observations in the past. Kendall (1971) points out that even so model (5) is a perfectly respectable stationary series in which the $\{w_t\}$ are generated from the $\{a_t\}$ and not vice versa. So does it matter that (7) is a divergent series? The answer is yes, because the Box–Jenkins forecasting procedure requires one to generate estimated residuals, and, if equation (7) is used, an unstable forecasting procedure is obtained. This becomes clear when one tries to estimate θ in equation (6), using the Box–Jenkins estimation procedure. Setting

$$\hat{a}_1 = w_1,$$

$$\hat{a}_t = w_t - \theta \hat{a}_{t-1}$$

we choose θ so as to minimize $\sum \hat{a}_t^2$. For model (4) we expect $\hat{\theta}$ to be about 0·5, but for model (5) one might expect $\hat{\theta}$ to be about 2.

We generated some sets of observations from both models (4) and (5) and found that the estimated value of θ is close to 0·5 for both models. Thus the optimum predictions for model (5) are the same as for model (4). (If the starting value a_0 is known, and the data follow model (5) exactly without rounding errors, then a second lower minimum occurs on the sum of squares surface close to $\theta = 2$. But of course these conditions will never apply in practice.)

Thus the Box–Jenkins estimation procedure naturally leads to a value of θ less than one which corresponds to the invertible model.

A control engineer's view of invertibility is discussed, for example, by Astrom and Eykhoff (1971, p. 130).

Induced Correlation

Box and Jenkins are quite right to question our use of the term "spurious correlation". A much better term is "induced correlation".

However, we strongly disagree with their subsequent statement concerning a seasonal model with linear trend that if "z_t has uncorrelated errors, the identification process we have suggested would lead *precisely* to the correct model". The model

$$(1-B)(1-B^{12})z_t = (1-B)(1-B^{12})a_t$$

is not invertible and cannot be identified by the Box–Jenkins process.

Bias

Box and Jenkins suggest that it is "quaint" to refer to the discrepancy between the mode and mean of the conditional distribution of x_{t+l} as "bias". This terminology is not our own but is suggested by Granger and Newbold (1970), and seems to us to be perfectly reasonable.

Box and Jenkins go on to suggest that one can calculate the whole of the conditional distribution but, of course, like everybody else they simply give [unbiased ?] point forecasts in their Table 1 together with symmetric probability limits.

Later Observations

In our paper, we comment that there is little point in spending too much time and effort preparing univariate forecasts as conditions may change. This point is well illustrated by the forecasts which Box and Jenkins give for the next six months in their Table 1, and which we now compare with the actual observations.

TABLE 1

Box–Jenkins forecasts and actual observations for June–November, 1971

	June	*July*	*August*	*September*	*October*	*November*
Box–Jenkins forecasts	286	409	511	761	966	1091
Actual observations	260	304	390	614	783	872

We think that most readers will judge these forecasts to be rather poor. Admittedly no univariate procedure would have been able to predict the effects of the 1971 economic recession, but this highlights our view that univariate forecasts should be robust and simple and should be adjusted where necessary on a subjective basis.

Other Applications

Box and Jenkins state that they know of a large number of applications of their methods. They list three "published accounts" which we went to considerable trouble to track down. This is what we found.

(a) Thompson and Tiao. Due to an incorrect version of the journal's name in the draft of the Box–Jenkins Comments, (the correct name is "*Bell Journal of Economics and Management Science*"), we only managed to get a copy just before completing this reply. Their data show linear growth and stable multiplicative seasonality and we would expect Holt–Winters to do equally well.
(b) Cramer and Millar (1973). We have been unable to trace any such paper which is published at the time of writing.
(c) Tomasek (1972). This paper, in an obscure journal, describes the analysis of a set of data showing a high seasonal component and a steady trend. We found that Holt–Winters also explains over 98 per cent of the variation about the mean and seems to us perfectly adequate.

The only other published account we have seen is by Makridakis and Wheelwright (1972). Again we found that Holt–Winters gave equally good forecasts.

The work of Reid (1969, 1971) and Granger and Newbold (1972) is not published in a journal and does not in any case give individual examples of the method.

Conclusions

(a) More work is required to investigate the use of transformations in the Box–Jenkins procedure.
(b) We still maintain that ARIMA models are harder to understand than traditional models.
(c) We do not accept that alternative and traditional common-sense forecasting methods are for the most part special cases of the approach based on ARIMA models. In particular, if a regression model is appropriate, the Box–Jenkins procedure is likely to be inefficient.
(d) We accept that the term "spurious correlation" is inappropriate, and should be replaced by "induced correlation".
(e) We believe that univariate forecasts should generally be robust and simple.
(f) We would like to see more published applications of the Box–Jenkins method so as to better assess its potential.
(g) There are occasions when Box–Jenkins does substantially *better* than other techniques of time-series analysis, in which case the extra costs can certainly be justified. There are also many occasions when Box–Jenkins does little better or worse than alternative techniques, in which case the extra costs cannot be justified. One requirement of future research is to establish those conditions where Box–Jenkins is likely to be worth while on a cost–benefit basis.

ACKNOWLEDGEMENT

We would like to acknowledge some stimulating conversations with Professor K. V. Diprose, School of Electrical Engineering, University of Bath, who pointed out the effect of using the Box–Jenkins procedure on a linear regression model.

ADDITIONAL REFERENCE

ASTROM, K. J. and EYKHOFF, P. J. (1971). System identification—a survey. *Automatica*, **7**, 123–162.