Forecasting Non-Normal Time Series

A. L. SWIFT and G. J. JANACEK
University of East Anglia, U.K.

ABSTRACT

We look at the problem of forecasting time series which are not normally distributed. An overall approach is suggested which works both on simulated data and on real data sets. The idea is intuitively attractive and has the considerable advantage that it can readily be understood by non-specialists. Our approach is based on ARMA methodology and our models are estimated via a likelihood procedure which takes into account the non-normality of the data. We examine in some detail the circumstances in which taking explicit account of the nonnormality improves the forecasting process in a significant way. Results from several simulated and real series are included.

KEY WORDS   ARMA models   Marginal distributions
Hermite polynomials   Likelihood   Non-normal

Suppose we have a series of sequential data $\{Y_t\}$ which, on examination, is clearly autocorrelated and not normally distributed. Such data arise as, for example, the correlated inter-arrival times of queuing systems, river flow over short time periods, wind velocities or speech waves.

It is obviously desirable to model the data so that forecasts can be made. If the data were normally distributed Gaussian ARMA models could be used, or if they were uncorrelated and non-normal then the parameters of a class of marginal distributions could be estimated using likelihood. No standard procedure is available which can cope with both non-normality *and* serial correlation. Of course, for long series one could fit a linear model using least squares and rely on asymptotic results. This, however, provides no information on the distribution of the errors (or, equivalently, the marginal distribution) so that prediction intervals of the forecasts cannot be calculated and the usefulness of such a procedure is questionable.

Several explicit models have been developed to simulate data with special types of autocorrelation structure and marginal distribution. Lawrence and Lewis (1981, 1985, 1987) have considered linear models with non-normal noise inputs. The models allow the third and higher cross moments of the series to be represented while limiting the correlation structure, usually to an AR(1) or AR(2) type. These can be regarded as model structures for appropriate data series. Very little work, however, has been done on model fitting. Lawrence and Lewis (1985) fit a model to a very long wind-velocity series, but use moment estimates and a rather *ad hoc* procedure to estimate the parameters. Smith (1986) considers the likelihood of one of the Lawrence and Lewis models NEAR(2), and describes difficulties in maximization which

renders it impractical. A further study by Karlsen and Tjostheim (1988) obtains consistent and asymptotically normal estimates for the NEAR(2) and NLAR(2) models but concludes that the standard errors are very large.

In this paper we describe a model for autocorrelated, non-normal data which allows:

(1)   A general class of correlation structures, and
(2)   Any marginal distribution,

and for which maximum likelihood estimates (MLEs) can be found. Full mathematical details are described in Janacek and Swift (1990), but here our primary aim is to guide the forecaster in the use and suitability of this model for forecasting.

The model is introduced in the next section and we discuss methods of obtaining forecasts from it in the fourth section. A necessary 'tool'—the Hermite representation of a function—is described in the third section. In the fifth section we briefly review the estimation procedure. The sixth section establishes some criteria to enable the forecaster to ascertain whether it is likely to be beneficial to model a particular marginal/autocorrelation function combination using our model as opposed to a linear least squares fit. Values of these criteria for a range of models are reported in the seventh section. In the eighth section we give results obtained from model fitting and forecasting simulated series of varied descriptions and some real data. The model is easily extended to incorporate seasonality in the data and this is explained in the ninth section.

THE MODEL

In this section we introduce our model, henceforth referred to as the JS model. We regard our non-normal series $\{Y_t\}$ as an instantaneous transformation $T(Z_t)$ of an underlying Gaussian series $\{Z_t\}$. We assume that the series $\{Z_t\}$ has a standard normal marginal distribution and that all joint distributions are normal (i.e. the $\{Z_t\}$ series considered as a whole is multivariate normally distributed).

If the marginal distribution has (known) cumulative distribution function $F(y)$, the transformation $T(z)$ is easily uniquely identified as $F^{-1}(\Phi(Z_t))$, where $\Phi(z)$ is the standard normal cdf. Such a transformation implies the underlying $\{Z_t\}$ series. As the latter is Gaussian by assumption its autocorrelation structure can easily be represented using a normal ARMA model (see, for example, Priestley, 1981; Harvey, 1981). In consequence, our model is parameterized by the type and parameters of the marginal distribution and the ARMA parameters of the underlying linear model. Analysis of the underlying $\{Z_t\}$ model is quite standard and the whole range of techniques for analysis of normal time series is available. The properties of the non-normal series then follow through the transformation.

AN ALTERNATIVE REPRESENTATION OF $T(z)$

In this section we consider the implications of forecasting using the JS model. This requires the well-known technique of representing the transformation $T(z)$ in terms of Hermite polynomials. Details of these and how they are calculated are given in the appendix of Granger and Newbold (1976). Briefly, any function $T(z)$ can be represented as

$$T(z) = \sum_{j=0}^{\infty} \alpha_j H_j(z) \tag{1}$$

where the α_j are constants and each $H_j(z)$ is a particular polynomial in z of order j. The formulae for these polynomials are well known, the first few being

$$H_0(z) = 1, \quad H_1(z) = z, \quad H_2(z) = z^2 - 1, \quad H_3(z) = z^3 - 3z, \quad H_4 = z^4 - 6z^2 + 3, \ldots$$

Given any function $T(z)$, the coefficients α_j can be evaluated using suitable numerical techniques without difficulty. As the $\{Y_t\}$ series is instantaneous transformation of the $\{Z_t\}$ series it is not unnatural to expect that the autocovariances of $\{Y_t\}$ (say, $R_0, R_1, R_2, \ldots$, where R_k is the autocovariance at lag k) are related to those of the $\{Z_t\}$ series (ρ_k, say). Granger and Newbold show that

$$R_k = \sum_{j=1}^{\infty} \alpha_j^2 j! \rho_k^j \tag{2}$$

i.e. there is a relation between R_k and ρ_k and this can be expressed in terms of the Hermite coefficients α_j. It is not difficult to show that for $-1 < \rho_k < 1$ this relationship is one to one. This will be useful to us later and is one reason for our mention of the Hermite representation of $T(z)$.

FORECASTING

Suppose we have a known model of the type of described above. The $\{Z_t\}$ series has, by design, a simple (linear) Gaussian structure so that optimal (i.e. conditional expectations) forecasts and confidence limits for these are easily calculated (see Box and Jenkins, 1970). We denote such a forecast of Z_{t+n} given data up to time t, as $Z_{t+n/t}$, where $n > 0$.

An intuitively obvious way of forecasting the $\{Y_t\}$ series is to use the relation between the $\{Y_t\}$ and $\{Z_t\}$ series to calculate forecasts of the $\{Y_t\}$, i.e. $Y_{t+n/t} = T(z_{t+n/t})$ and so on. It can be shown that this forecast minimizes the mean absolute error (MAE) and we will refer to this as the 'naive' forecast. Granger and Newbold (1976) discuss several forecasts of a series $\{Y_t\}$ which is a transformation of another series $\{Z_t\}$. They show that the linear function of Z_t, Z_{t-1}, Z_{t-2}, ..., which gives the minimum expected squared forecast error of a forecast of Y_{t+n}, is just $\alpha_0 + \alpha_1 Z_{t+n/t}$, where α_1 is the first Hermite coefficient and so depends only on the transformation $T(z)$ and therefore in terms of our model only on the marginal distribution. Here $Z_{t+n/t}$ is the optimal z forecast mentioned earlier. This gives another way of forecasting using our model and we well refer to it as the linear z forecast.

Granger and Newbold derive expressions for the theoretical mean square of both the naive and the linear z forecasts. They also calculate the mean square error (MSE) of the optimal conditional expectation forecasts. In general, we cannot calculate these forecasts but the mean square error is useful in assessing how close (in the MSE sense) the naive and linear z forecasts come to optimality. We point out that when one is dealing with non-normal marginals it is not at all clear that mean square error is a sensible figure of merit for assessing forecasts. We shall use it, however, since we wish to make comparisons with normally based techniques for which it is appropriate. As stated earlier, it is also possible to fit a linear model to the $\{Y_t\}$ series using least squares. Such a model would not enable prediction intervals (other than normally based ones) to be calculated but would give forecasts. It is of interest to compare the MSE of such forecasts with those of the forecasts based on our model.

We can calculate the correlations of the $\{Y_t\}$ series implied by the model (using relation (2)) and find a linear model which produces the corresponding correlations. Such a procedure is not trivial for a finite order linear model, so we use Wilson's (1969) spectral factorization

technique to fit a large-order moving average (impulse response) process. From this the theoretical MSE of the forecasts can be calculated.

It should be noted that the MSEs obtained in this way merely provide a lower bound on those which would be found by fitting a linear model to data produced by our model. Details are given in Janacek and Swift (1990).

ESTIMATION

We do not wish to dwell on the details of estimation in this paper, so the reader is again referred to Janacek and Swift (1990, section 4) for comprehensive details. In essence, the likelihood function of our model is the likelihood of the ARMA model fitted to the $\{Z_t\}$ series multiplied by the Jacobian of the transformation of all the $\{Y_t\}$ series to all the $\{Z_t\}$. Software is readily available to calculate the likelihood of a Gaussian ARMA model (e.g. Gardner *et al.*, 1980) so this presents no problems. For practical likelihood maximization we use a two-stage procedure whereby in stage one the marginal parameters are estimated for fixed ARMA parameters and in stage two the situation is reversed. The stages are repeated until the parameter estimates converge. In this way the maximum likelihood estimates (MLEs) of the parameters of our model can be obtained. No new techniques are required, merely a merger of existing software.

If the model is appropriate the residuals from the ARMA model fitted to the underlying series should be asymptotically normally distributed. This enables conventional diagnostic checking procedures to be used (e.g. Box and Pierce, 1970) so that no new methodology is required.

WHEN TO USE THE MODEL

In this section we describe means of assessing types of data for which our model is advantageous. First, we consider the directionality of our model. A stochastic process $\{Y_t\}$ is *time reversible* if for every j, s, and t, $\{Y_t, Y_{t-1}, ..., Y_{t-j}\}$ and $\{Y_s, Y_{s+1}, ..., Y_{s+j}\}$ have the same joint probability density function. This is equivalent to saying that the process has the same probabilistic structure, even when it is reversed in time, i.e. is non-directional. Clearly, as Gaussian series are defined solely in terms of their autocorrelation structure and mean, our underlying $\{Z_t\}$ series is time reversible. Weiss (1975) says that any instantaneous transformation of a time-reversible series is also time reversible. It follows therefore that our model for the non-normal series $\{Y_t\}$ is also time reversible.

This can be restrictive, as many real-life series exhibit irreversibility. Models have been suggested to cope with this (for instance, see Tong, 1983; or Priestley 1989). Weiss (Theorem 2) also shows that if an ARMA process (with exclusion of a particular MA case) with identical and independently distributed noises is also time reversible then the noises (and so the series) must be Gaussian. As our model is non-normal this implies that an equivalent linear representation cannot exist. In practice, we assume that a series which is not obviously directional can be modeled in either way. This justifies our use of the linear model as a benchmark for the forecasting performance of our model.

From the discussion above it follows that data that are stationary and autocorrelated, that have a fixed non-normal marginal and that do not exhibit strong directionality may be represented using our model. One can test for normality using the tests developed by Lin and

Mudhoekar (1980) or Epps (1987) while some idea of reversibility or otherwise can be obtained by looking at the past history and examining runs, as in Tong (1983, Chapter 5). As stated earlier, if accurate prediction intervals are not required forecasts can be obtained by fitting a linear model to such data using least squares. In our experience we have found that data exhibiting some combinations of marginal distribution and autocorrelation structure can be forecasted just as well or only slightly better with our model than with such a linear model. Weiss's results imply that if data are time reversible, testing whether those data are linear (Luutkomen *et al.*, 1988) is equivalent to testing for normality (e.g. Gasser, 1975).

This is reflected in our model. Recall that the series $\{Y_t\}$ is regarded as an instantaneous transformation $Y_t = T(Z_t)$, so if this transformation is linear then the linear model for the $\{Z_t\}$ transforms to a linear model for $\{Y_t\}$. Also, Z_t for any t is (standard) normally distributed and a linear function of a normal distribution is still normal, so $\{Y_t\}$ must be normal. The converse (if the transformation is non-linear $\{Y_t\}$ must be non-normal *and* non-linear) follows easily. It therefore seems reasonable to look at the transformation $T(z)$ to give an indication of the linearity/normality of our model. We consider the 'best' linear approximation to a transformation $T(z)$ and fit the straight line which has the minimum expected squared deviation from the transformation. That is, we find an α and β which minimize

$$\int_{-\infty}^{\infty} [T(z) - (\alpha + \beta z)]^2 \phi(z)\, dz \qquad (3)$$

where $\phi(z)$ is the standard normal. By making use of the Hermite expansion for $T(z)$ and the orthogonality property of Hermite polynomials (Granger and Newbold, 1976) we can show that $\alpha = \alpha_0$ and $\beta = \alpha_1$ and we see that the first two terms in the Hermite expansion give the 'best' linear approximation to the transformation. This result extends to a polynomial approximation of any order.

The mean square deviation from linearity of $T(z)$ (i.e. the value of equation (3)) for this straight line is

$$\sum_{j=2}^{\infty} j! \alpha_j^2 \qquad (4)$$

From equation (2) this is equivalent to

$$\mathrm{var}(Y_t) - \alpha_1^2$$

To compare the deviation from linearity of several transformations it seems reasonable to scale this by $\mathrm{var}(Y_t)$, i.e.

$$1 - \frac{\alpha_1^2}{\mathrm{var}(Y_t)} \qquad (5)$$

Therefore the more non-normal (and non-linear) the model, the larger the value of equation (5). We will call $\alpha_1^2/\mathrm{var}(Y_t)$ the linearity factor (LF) of a transformation (and implied marginal). The closer this factor is to 1, the nearer our transformation is to linearity and the marginal distribution to normality. If the LF is close to 1 and our model is appropriate there is little advantage in using our model over a linear one. Note, however, that we have not taken the autocorrelation structure into account.

We have already established relation (2) between any ρ_x and the corresponding R_k (autocovariance of $\{Y_t\}$ at lag k). It will be convenient to consider the correlations $\{C_k\}$ of the $\{Y_t\}$ series, i.e. $C_k = R_k/R_0$.

The one-to-one relation between C_k and ρ_k will obviously pass through $(0,0)$ and $(1,1)$ and
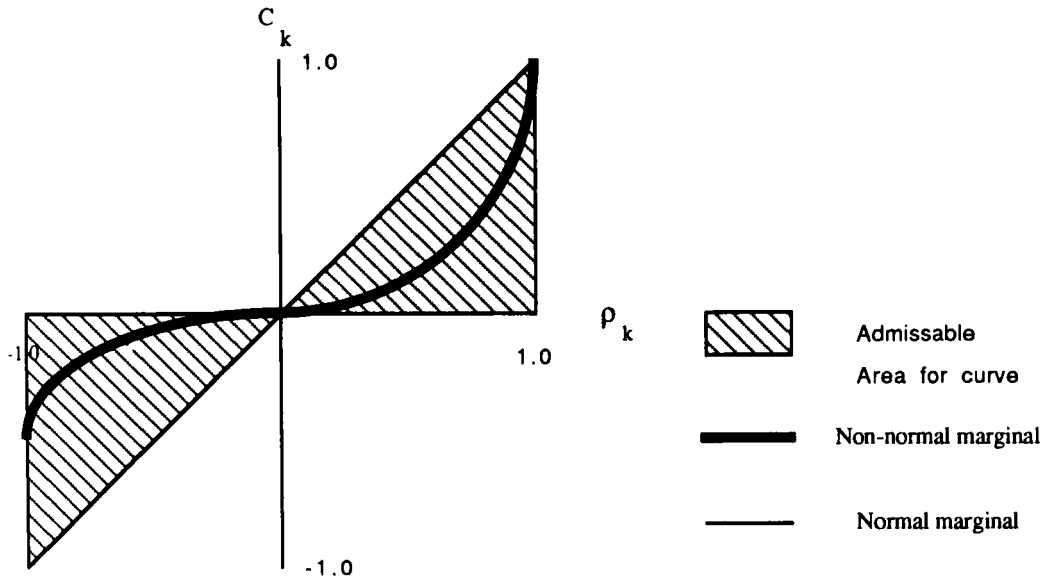
Figure 1

for any normal series $\{Y_t\}$ will be a straight line (see Figure 1). Note that $\{Z_t\}$ is always more predictable in the sense that it has higher correlations than $\{Y_t\}$, so $C_k \leqslant \rho_k$ if $\rho_k > 0$ and $C_k \geqslant \rho_k$ if $\rho_k < 0$. The ρ_k-C_k relation must therefore lie in the shaded area in Figure 1. A typical shape of the relation for a non-normal marginal for positive ρ_k is also shown.

For some classes of marginal distribution it is known that there is a minimum attainable correlation coefficient between two variates. For instance, Moran (1967) shows that the minimum for exponential variates is $1 - (\pi^2/6) \approx -0.64493$. As the ρ_k-C_k relation is one to one, setting $\rho_k = -1$ in equation (2) allows us to calculate this numerically for any marginal distribution.

Clearly, if this minimum correlation differs greatly from -1 the ρ_k-C_k curve for negative values of ρ_k must deviate appreciably from that of the linear transformation. The ρ_k-C_k curve therefore gives an indication of the distortion of the $\{Z_t\}$ correlations by the transformation. This depends only on the type and parameters of the marginal distribution.

The distortion for a particular model, however, depends whether the particular values of the actual $\{\rho_k\}$ or $\{C_k\}$ are distorted greatly, i.e. where on the curve the actual correlations lie.

For example, the exponential distribution has a linearity factor of 0.8158. Suppose the underlying ARMA model were AR(1):

$$Y_t = A Y_{t-1} + e_t$$

If $A = -0.6$ then $\rho_1 = -0.6$, $\rho_2 = (0.6)^2$, $\rho_3 = (0.6)^3 = -0.216$, etc. As the minimum correlation attainable is -0.64493, negative values of ρ are severely distorted, resulting in $C_1 = -0.4271$, $C_2 = 0.3170$, $C_3 = -0.1680$. A linear model fitted to these autocorrelations would differ considerably from the underlying ARMA model and therefore produce very different forecast mean square errors. These cannot be better than those of the underlying ARMA model (using equation (3)) as the underlying ARMA model is more correlated.

Consider the exponential marginal but with $A = 0.6$, so $\rho_1 = 0.6$, $\rho_2 = 0.6^2$, $\rho_3 = 0.6^3 = 0.216$, etc. Then $C_1 = 0.5548$, $C_2 = 0.3170$, $C_3 = 0.1846$. As the ρ_k's now lie in a

range of the $\rho_k - C_k$ curve where they are not distorted as much a linear model fitted to the implied C_k's will do almost as well.

We therefore suggest that the JS model is most likely to be useful for data where:

(1) The marginal distribution has a low linearity factor; and
(2) If there are negative correlations the minimum autocorrelation is substantially different from -1.

Note that for a symmetric marginal distribution the minimum correlation is always -1 as shown by Moran (1967). In the next section we give some results for various marginal distributions and their repercussions.

SOME RESULTS FOR PARTICULAR MARGINAL DISTRIBUTIONS AND THEIR IMPLICATIONS

In Tables I–III we list some common marginal distributions, their linearity factors and minimum correlations. We will call a parameter of the marginal distribution which merely

Table I. Beta marginals

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \qquad 0 \leqslant x \leqslant 1$$

a	b	Linearity factor	Minimum correlation
0.5	0.5	0.9000	-1
0.5	2.0	0.8436	-0.6973
1.0	1.0	0.9549	-1
2.0	0.5	0.8436	0.6973
2.0	0.8	0.9211	-0.8735
2.0	0.9	0.9360	-0.9064
2.0	6.0	0.9618	-0.9293
3.0	9.0	0.9746	-0.9520
8.0	2.0	0.9515	-0.9054
10.0	2.0	0.9438	-0.8888
10.0	5.0	0.9913	-0.9853
12.0	2.0	0.9380	-0.8765
50.0	1.0	0.8250	-0.6609

Table II. Gamma marginals

$$f(x) = \frac{a^b x^{b-1} e^{-ax}}{\Gamma(b)} \qquad x \geqslant 0$$

b	Linearity factor	Minimum correlation	
10	0.9782	-0.9564	
5	0.9571	-0.9146	
1	0.8158	-0.6450	(includes exponential)
0.5	0.6930	-0.4394	

Table III. Other marginals

Distribution	pdf/cdf	Linearity factor	Minimum correlation
Double exponential	$f(x) = (1/2b)\exp(-|x - a|/b)$	0.9631	−1
Extreme value	$F(x) = \exp(-e^{-(x-a)/b})$	0.9398	−1
Cauchy	$f(x) = (\pi a(1 + [(x-b)/a]^2)^{-1}$	0.0000	−1
Lognormal	$\log(x)$ is $N(\mu, 0.7^2)$	0.7749	−0.6130
Lognormal	$\log(x)$ is $N(\mu, 0.5^2)$	0.8802	−0.7789

scales the distribution a scaling parameter. Thus for a two-parameter marginal, suppose y has pdf $f(A, B)$ and sy has pdf $f(A^*, B)$, where s is a positive constant. Then the first parameter is a scaling parameter (the second is not). In this case the implied transformation $T(z)$ and the Hermite coefficients of these two marginal distributions will only differ by a factor of s, so that the linearity factor is *independent* of the scaling parameter.

As might be expected, the linearity factor gives a clear indication of the non-normality of the distribution. For instance, the double exponential distribution has LF = 0.9631 whereas Gamma (A, 0.25) (an extremely skew distribution) has LF = 0.5427.

Note that the exponential, beta and double exponential marginal distributions with positive autocorrelations are quite linear. This may explain in part why some models suggested by Lawrence and Lewis (1985) perform well (e.g. the NEAR(2) and NLAR models which have exponential and double exponential (Laplace) marginals, respectively, but an AR(2) type correlation structure). They are 'nearly linear' in our sense. The parameter of the exponential distribution is a scaling parameter so does not affect the LF. Both parameters, however, of the beta distribution affect the LF. An increase in the first parameter, a, decreases the LF (for fixed b), and a decrease in b decreases the LF. The extreme value distribution also exhibits near linearity (both parameters are scaling parameters). The first gamma parameter, a, is a scaling parameter, and, as might be expected because of the asymptotic normality of the distribution, as the second parameter b increases the LF becomes larger. For small values of b less than one the transformation is highly non-linear and the minimum correlation is much nearer 0 than −1. Similarly, the smaller the parameter of the chi-squared distribution, the greater the non-linearity.

As the variance of a Cauchy distribution is infinite, the LF is 0. (In practice, as we use a finite number of terms for our Hermite expansion the LF is calculated as a very small number.) The Cauchy distribution is an obvious candidate for our model. While correlations cannot be calculated, correlated data are easily found by plotting the data, giving a means of modelling data with infinite variance.

It is straightforward to compute the linearity factor for the lognormal distribution analytically. It is

$$\frac{\sigma^2}{(\exp(\sigma^2) - 1)}$$

where σ^2 is the variance of the logged variable. This will be small for values of σ much greater than 0. The $\{Z_t\}$ are, of course, the standardized $\{\log Y_t\}$. This explains the effectiveness of the log transformation commonly used and in part the prominent role of the lognormal distribution in the literature.

As can be seen from the discussion above, the JS model does allow one to take into account

the non-normality of a series. For some marginal distributions (Cauchy, for example) it may be very important to do so. As one might expect, however, for some marginal distributions the 'closeness' to normality is such that the additional complexity of the JS model is not necessary. The decision as to whether one should use JS or a more conventional technique is made reasonably straightforward by the use of the linearity factor, and Tables I–III can be used as a guide. The indications they give are borne out in practice, as we now show.

EXPERIENCES WITH SIMULATED AND REAL DATA

In this section we give some examples of the whole modelling and forecasting procedure on both simulated and real data which will illustrate the points made above. Many more series have been modeled.

For each simulated series of 400 items we fitted the class of marginals used to generate the

Table IV. Simulated series 1

Generating model						
Marginal type and parameters	Gamma (0.16, 0.5)					
AR parameters	-0.4, 0.7					
MA parameters	0					
Correlations of data	0.182	-0.314	-0.175	0.290	0.239	-0.051
Model fitted to independent data						
Marginal parameters	0.176707	0.514896				
Fitted model						
Marginal parameters	0.184316	0.525352				
AR parameters	-0.382660	0.713125				
MA parameters	-0.196626	0.023650				
Linearity factor	0.7030					
Minimum correlation	-0.4546					
Variance of model	15.4616					

Forecasts	*One-step*	*Two-step*		*Five-step*		
Mean square errors (empirical)						
Naive	8.13	8.78	⋮	11.27		
Linear z	8.55	8.94	⋮	10.60		
Linear y	11.08	11.34	⋮	11.66		
Mean absolute errors (empirical)						
Naive	1.64	1.71	⋮	2.14		
Linear z	2.04	2.08	⋮	2.48		
Linear y	2.39	2.46	⋮	2.59		
Mean square errors (theoretical) of fitted model						
Optimal	8.54	8.79	⋮	13.41		
Naive	8.56	8.81	⋮	13.45		
Linear z	9.76	9.92	⋮	13.55		
Linear y	11.84	12.00	⋮	13.64		
Correlations of fitted model						
ρ (of $\{Z_t\}$)	0.1668	-0.6605		-0.3317	0.3288	0.3909
C (of $\{Y_t\}$)	0.1250	-0.3524		-0.2249	0.2614	0.3179

This is an extreme example of how forecasts can be improved using our model. The series has a gamma marginal distribution with a small second parameter and some negative correlations.

data to the first 200 observations. A forecast comparison program COMPFOR enables us to use this model to forecast at each time threshold 1 to 5 steps ahead using each forecast type described above (naive, linear z and linear y) and calculate the empirical MSE and MAE of these for the remaining 200 data points. The linear model was fitted using least squares procedure. Obviously, errors in parameter estimation influence these empirical forecasts.

COMPFOR also takes the fitted model and calculates the theoretical mean square errors (as in Granger and Newbold) of the optimal, naive, linear z and linear y forecasts. These offer a comparison of the forecasting abilities of each forecast type for particular models assuming that the parameter values are known.

Tables IV–VIII give numerical results from the analysis of simulated data series. The ARMA coefficients given are those of the model

$$Y_t + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} = e_t + B_1 e_{t-1} + \cdots + B_q e_{t-q}, \text{ for given } p \text{ and } q$$

Table V. Simulated series 2

Generating model				
Marginal type and parameters	Gamma (0.16, 0.5)			
AR parameters	0.6			
MA parameters	0			
Correlations of data	0.562	0.340	0.248	0.122	0.035

Model fitted to independent data		
Marginal parameters	0.16238	0.543808

Fitted model		
Marginal parameters	0.164089	0.546799
AR parameters	−0.501389	−0.121582
MA parameters	0.178831	−0.104330
Linearity factor	0.7109	
Minimum correlation	−0.4670	
Variance of model	20.3050	

Forecasts	One-step	Two-step	:	Five-step
Mean square errors (empirical)				
Naive	15.52	22.98	:	28.09
Linear z	16.51	20.61	:	24.83
Linear y	13.99	20.02	:	25.03
Mean absolute errors (empirical)				
Naive	2.28	2.72	:	2.97
Linear z	2.70	2.94	:	3.28
Linear y	2.48	2.99	:	3.33
Mean square errors (theoretical) of fitted model				
Optimal	13.30	18.05	:	20.08
Naive	13.34	18.09	:	20.09
Linear z	14.28	18.17	:	20.08
Linear y	13.59	18.28	:	20.13

Correlations of fitted model				
ρ (of $\{Z_t\}$)	0.6423	0.3829	0.2700	0.1819	0.1124
C (of $\{Y_t\}$)	0.5724	0.3125	0.2118	0.1383	0.0923

The same gamma marginal distribution as series 1 but here all the correlations are positive. Theoretically, the naive forecast should do better than the linear y but empirically this is not the case, presumably due to estimation error. No great advantage in fitting JS model.

Table VI. Simulated series 3

Generating model					
Marginal type and parameters exponential (0.5)					
AR parameters	0				
MA parameters	-0.6				

Correlations of data	-0.316	-0.023	0.087	-0.072	-0.045

Model fitted to independent data
Marginal parameters   0.515334

Fitted model
Marginal parameters	0.007806	0.509974
AR parameters	0.007806	
MA parameters	-0.551769	
Linearity factor	0.8158	
Minimum correlation	-0.6450	

Forecasts	One-step	Two-step	:	Five-step
Mean square errors (empirical)				
Naive	3.88	4.99	:	5.08
Linear z	3.53	4.49	:	4.54
Linear y	3.98	4.49	:	4.55
Mean absolute errors (empirical)				
Naive	1.25	1.45	:	1.47
Linear z	1.27	1.49	:	1.51
Linear y	1.35	1.49	:	1.51
Mean square errors (theoretical) of fitted model				
Optimal	3.06	3.84	:	3.84
Naive	3.06	3.85	:	3.85
Linear z	3.10	3.85	:	3.85
Linear y	3.41	3.85	:	3.85

Correlations of fitted model					
ρ (of $\{Z_t\}$)	-0.4280	0.0033	0.0000	0.0000	0.0000
C (of $\{Y_t\}$)	-0.3172	0.0027	0.0000	0.0000	0.0000

This is an underlying MA model with exponential marginal.

We have also fitted our model to several real data sets and in this paper we give results from two. The first is a typical model fitting to some non-normal data, whereas the second is included as it demonstrates successful model fitting with a seasonal component and because it has a marginal distribution which exhibits no forecasting advantage in terms of MSE or MAE.

Daily riverflow data for the River Tees for a number of years was made available to us. The monthly means vary considerably from year to year, so we did not deseasonalize. We arbitrarily chose a year's data from July to June inclusive and used this to forecast 181 days ahead.

A histogram of the 365 observations (Figure 2) was clearly very skew, suggesting a gamma distribution, and a plot (Figure 3) showed the data to be correlated. The autocorrelations are all positive and exhibit a fading sinusoidal pattern. Details of the model fitted, the empirical forecast errors and results for the theoretical forecast errors are shown in Table IX. The gamma distribution fitted has a second parameter of 1.4623 and so is more linear than an exponential distribution. The naive and linear z forecast perform considerably better than the

Table VII. Simulated series 4

Generating model					
Marginal type and parameters double exponential $(2, 2.5)$					
AR parameters	0.6				
MA parameters	0				

Correlations of data	0.487	0.319	0.147	0.002	-0.089

Model fitted to independent data		
Marginal parameters	2.557947	2.236645

Fitted model		
Marginal parameters		
AR parameters	-1.360098	0.617119
MA parameters	-0.890985	0.324837
Linearity factor	0.9631	
Minimum correlation	-1	
Variance of model	10.0134	

Forecasts	*One-step*	*Two-step*	:	*Five-step*
Mean square errors (empirical)				
Naive	8.49	10.51	:	11.60
Linear z	8.39	10.68	:	12.04
Linear y	8.70	11.03	:	12.57
Mean absolute errors (empirical)				
Naive	2.16	2.42	:	2.56
Linear z	2.18	2.46	:	2.61
Linear y	2.23	2.51	:	2.69
Mean square errors (theoretical) of fitted model				
Optimal	7.19	8.69	:	9.74
Naive	7.23	8.74	:	9.76
Linear z	7.19	8.70	:	9.74
Linear y	7.24	8.74	:	9.75

Correlations of fitted model					
ρ (of $\{Z_t\}$)	0.5179	0.3171	0.1117	-0.0438	0.1285
C (of $\{Y_t\}$)	0.50369	0.3065	0.1076	-0.0422	0.1238

This is a series with a double exponential marginal. Linearity factor is borne out, as linear y forecasts do just as well as naive.

linear y forecasts in terms of both MSE (naive one-step MSE is 118.2, linear y MSE is 159.3) and MAE (5.26 against 7.29 for one-step). As might be expected, the linear z forecast does not perform as well as the naive in MAE terms.

The residuals from the model fitting had mean -0.0241 with standard error of mean 0.0222 and their autocorrelations were very small, indicating independence. A Q-Q plot gave a reasonably straight line.

As might be expected, the 95% prediction limits using our model and the naive forecast (Figure 4) gave a much narrower range than that obtained using the linear y forecasts and assuming normal errors (Figure 5). Ten out of 177 observed values corresponding to the one-step-ahead forecasts lay outside the naive limits and 15 outside the linear y limits, so there is nothing to suggest that the narrow range using our model is less appropriate. (Although 181 time periods were available for estimation, only 177 forecasts were made at each number of steps ahead, so the same number of forecasts were used to calculate empirical MSE and MAE for each.) We conclude that it is definitely advantageous to use the JS model for these data.

From a totally different source we obtained some monthly data from the St John River. The

Table VIII. Simulated series 5

Generating model
Marginal type and parameters   Cauchy (0.6, 0.6)
AR parameters   0.6
MA parameters   0

Correlations of data not applicable

Model fitted to independent data
Marginal parameters   0.574069   0.629051

Fitted model
Marginal parameters   5.600081   0.636315
AR parameters -1.143521 0.246908
MA parameters -0.515750 -0.105837
Linearity factor   0
Minimum correlation -1
Variance of model infinite

Forecasts	One-step	Two-step	:	Five-step
Mean square errors (empirical)				
Naive	1.49	1.67	:	1.75
Linear z	65.95	29.83	:	10.77
Linear y	not available			
Mean absolute errors (empirical)				
Naive	0.73	0.78	:	0.80
Linear z	65.9	4.45	:	2.60
Linear y	not available			

Correlations of fitted model

ρ (of $\{Z_t\}$)	0.6504	0.4362	0.3382	0.2790	0.2355
C (of $\{Y_t\}$)	—	—	—	—	—

This is a Cauchy series. The time-series plot clearly shows a cyclic pattern but autocorrelation is meaningless due to the infinite variance. We could not obtain a successful linear fit for this, or any of the other Cauchy data we simulated, and the spectral factorization fails for the theoretical linear y forecast MSEs.
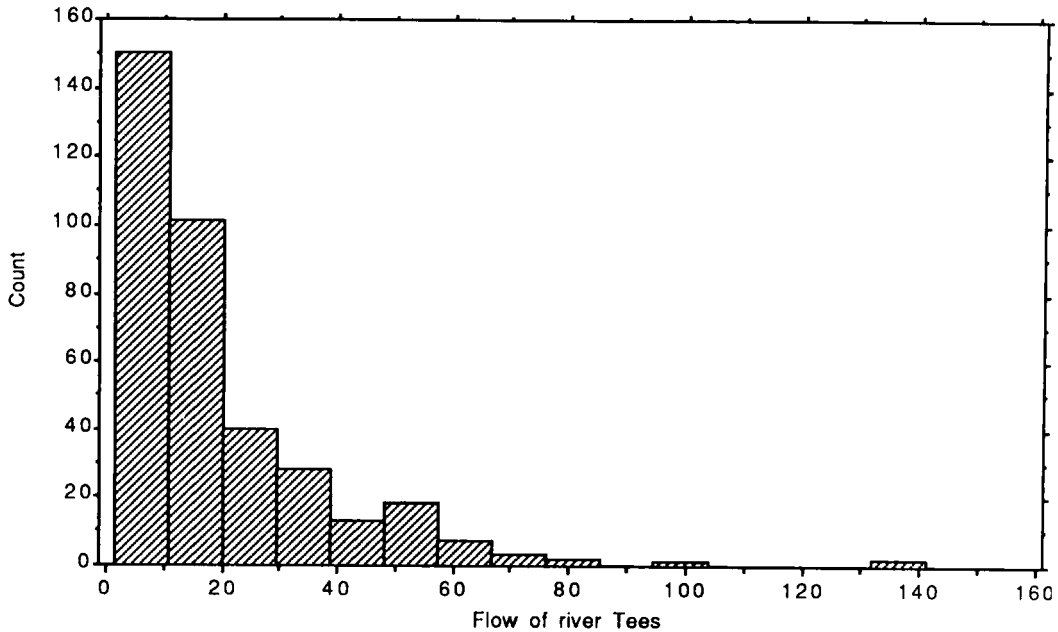

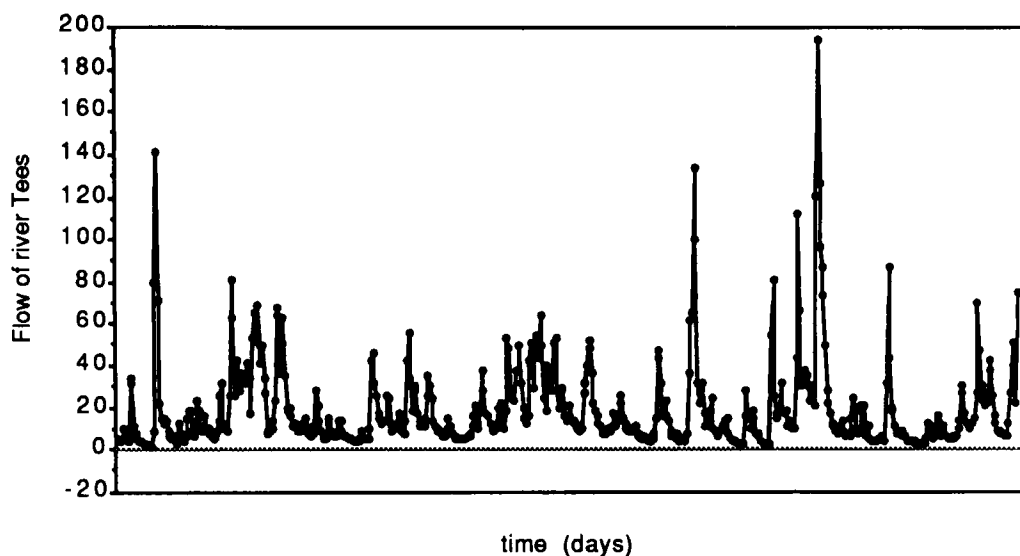Figure 2. Histogram of the daily flow of the River Tees over one year

time (days)

Figure 3. Flow of the River Tees

Table IX. Results for River Tees data

Correlations of data

0.787	0.511	0.418	0.357	0.273	0.214	0.206	:	decaying sinusoidal

Model fitted to independent data

Marginal parameters	Gamma	0.07745	1.46232

Fitted model

Marginal parameters	0.07210346	1.52248468
AR parameters	− 0.716068	
MA parameters	0.766700	
Linearity factor	0.8707	
Minimum correlation	− 0.7468	
Variance of model	292.833	

Forecasts	*One-step*	*Two-step*	:	*Five-step*
Mean square errors (empirical)				
Naive	118.2	322.0	:	402.6
Linear z	121.8	305.5	:	399.4
Linear y	159.3	326.6	:	404.0
Mean absolute errors (empirical)				
Naive	5.26	9.47	:	12.39
Linear z	6.51	10.59	:	14.15
Linear y	7.29	11.16	:	13.71
Mean square errors (theoretical) of fitted model				
Optimal	58.8	179.2	:	278.3
Naive	58.9	179.3	:	278.3
Linear z	84.1	185.8	:	278.4
Linear y	63.6	183.4	:	279.7

Correlations of fitted model

ρ (of $\{Z_t\}$)	0.8552	0.6124	0.4385	0.3140	0.2248
C (of $\{Y_t\}$)	0.8388	0.5812	0.4063	0.2859	0.2022

Naive one step prediction intervals

···· lower   —— actual   ······ upper
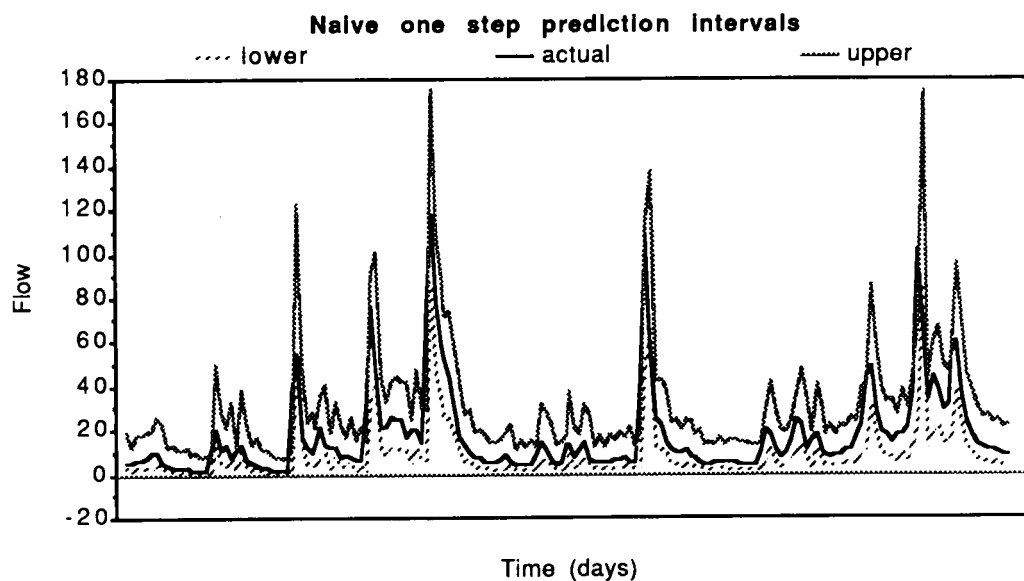

Flow

Time (days)

Figure 4. One-step prediction limits (95%) and actual values for the forecasting comparison period using the naive forecast

Linear one step prediction intervals
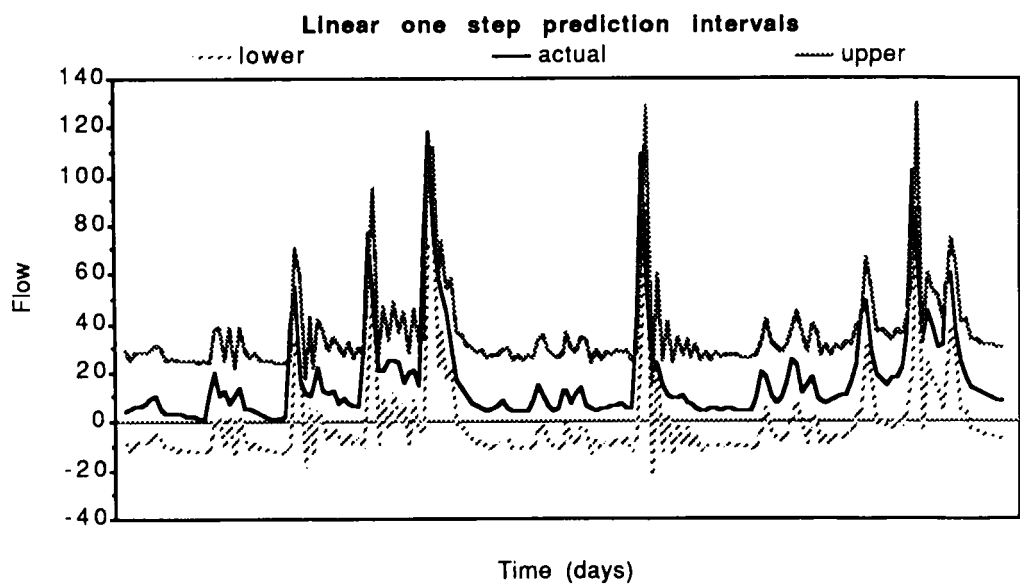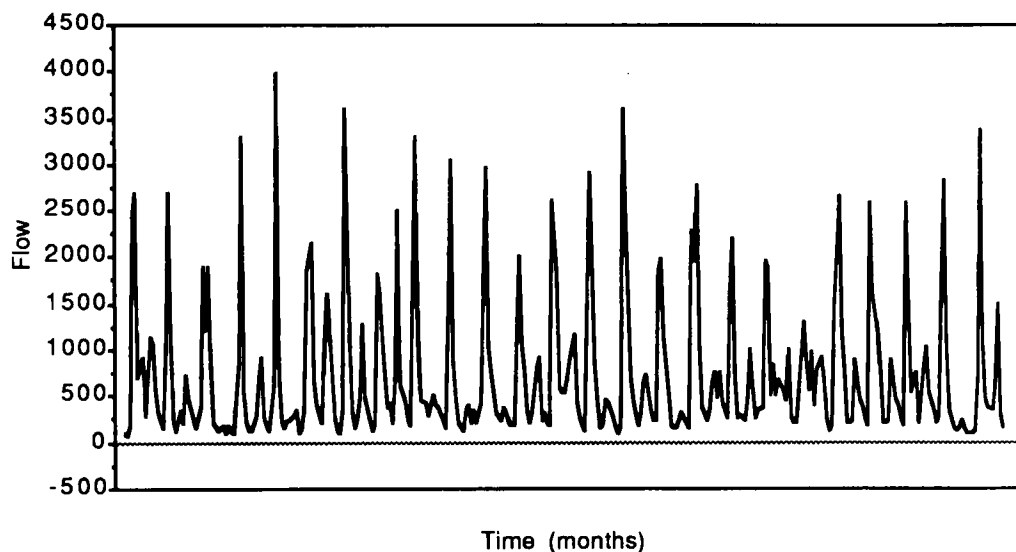
····· lower   —— actual   ······ upper

Flow

Time (days)

Figure 5. One-step prediction limits (95%) and actual values for the forecasting comparison period using the linear y forecast

Figure 6. The St John River flow data

data (Figure 6) are clearly seasonal, so they were deseasonalized assuming a multiplicative factor as described in the next section. We fitted 300 observations and used the succeeding 100 for forecasting comparison. As the magnitude of the data was large they were divided by 6000 (slightly above the maximum) for convenience. A histogram (Figure 7) showed the resulting data to be slightly skew and still slightly correlated. A beta distribution was fitted. The results from the fitted model are shown in Table X.

The naive and linear z forecasts had only a slightly smaller MSE or MAE than those of the linear y forecast. This is to be expected, as the linearity factor of the beta marginal is large
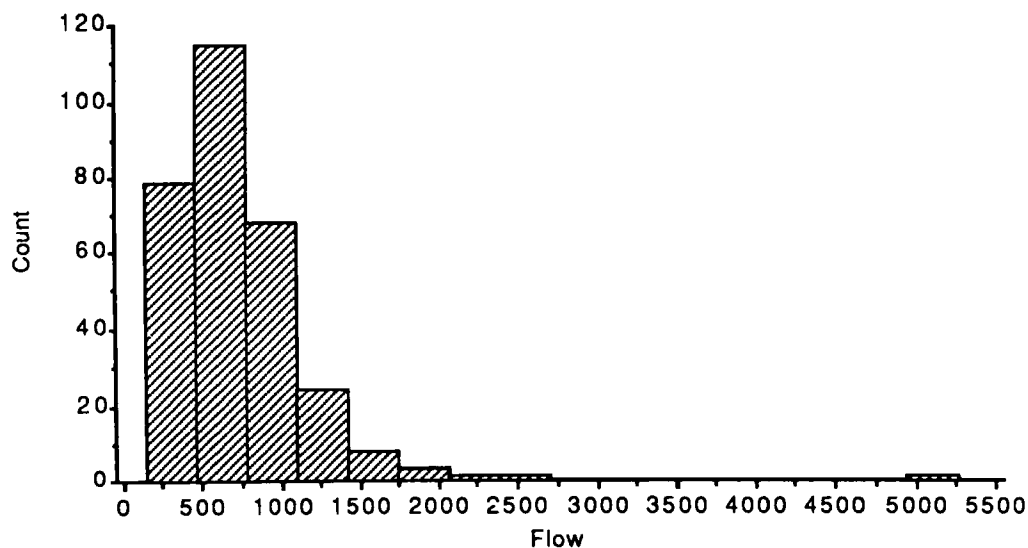

Figure 7. Histogram of the deseasonalized St John River data

Table X. Results for St John data

Correlations of data	0.220	0.082	0.015	−0.102	−0.025
Fitted model					
Marginal parameters	2.96	21.0772	1.0		
AR parameters	−0.413398				
MA parameters	−0.110416				
Linearity factor	0.9538				
Minimum correlation	−0.90760				
Variance of model	0.004329				

Forecasts	One-step	Two-step	:	Five-step
Mean square errors (empirical) ($\times$ 1000)				
Naive	3.86	4.54	:	4.66
Linear z	3.80	4.46	:	4.50
Linear y	3.96	4.51	:	4.53
Mean absolute errors (empirical) ($\times$ 1000)				
Naive	4.27	4.76	:	4.83
Linear z	4.39	4.92	:	4.94
Linear y	4.45	4.91	:	4.89
Mean square errors (theoretical) of fitted model ($\times$ 1000)				
Optimal	3.91	4.25	:	4.32
Naive	3.91	4.25	:	4.32
Linear z	3.91	4.25	:	4.32
Linear y	3.92	4.25	:	4.32

Correlations of fitted model					
ρ (of $\{Z_t\}$)	0.3140	0.1298	0.0537	0.0222	0.0092
C (of $\{Y_t\}$)	0.3040	0.1246	0.0513	0.0212	0.0088

(0.9538) and the linear y MSE is very close to the optimal. The residuals look reasonably normal (mean $-0\cdot189$, standard error $0\cdot0549$) although there is one large outlier which corresponds to an observation which looks like an outlier itself. While no forecasting advantage is found using the JS model as for the River Tees data, the 95% prediction intervals for the linear y forecasts are much wider than those of the naive forecasts, and yet a similar number of observed values lies outside for both types of forecast.

SEASONALITY

Up to this point we have made no mention of seasonality, and it is quite clear that the JS model must be flexible to take this into account. Several approaches could be taken. The natural one appears to us to be that which gives rise to a multiplicative seasonal adjustment process. This also fits well with the idea of a non-normal marginal.

Suppose our seasonal series $\{Y_t\}$ has a seasonal pattern of period p, then the magnitude of the series will vary over time. If we wish to specify the marginal then we need to build into the parameters of the distribution some mechanism for allowing for this variation. We thus have a situation in which the type of marginal distribution remains the same but one or more parameters varies. This approach has been followed by other authors. For example, Fernandez and Salez (1987) consider a model where each seasonal has a different three parameters in the distribution.

Writing the seasonal effects $s_1, s_2, ..., s_p$, we can think of our series as being made up of the seasonal effects and a non-seasonal component Y_t^*, i.e.

$$Y_t = s_{t \bmod p} \, Y_t^*$$

For many distributions (the exponential family, for instance) multiplication by a scale parameter is the same as changing one of the distributional parameters. Thus if X has the Gamma (a, b) distribution, sX is Gamma (sa, b). Given estimates of the seasonal components, the retrieval of the non-seasonal model and the estimation of the JS model is straightforward, since $Y_t^* = Y_t / s_{t \bmod p}$. Without loss of generality, we can assume that the sum of the seasonal effects is N, the number of observations. To estimate the seasonal effects we take the simple estimator based on the ratio of the monthly seasonal and the overall mean. Thus for the jth seasonal effect we estimate s_j by $\hat{s}_j = \hat{y}_j / \hat{y}$, where $\hat{y}_j$ is the mean of the data items in season j and $\hat{y}$ the overall mean. We justify this since

$$E(\hat{y}) = E\left(\frac{1}{N} \Sigma \, s_{t \bmod p} \, y_t^*\right) = \mu \quad \text{and} \quad E(\hat{y}_j) = E(s_j y_t^*) = \mu s_j$$

For the Gamma (a, b) case above $\mu = b/a$, so the mean of the seasonal series will be $s_j b / a$ at period j.

When the marginal does not belong to the class of distributions discussed above and does not a have a natural scaling parameter we suggest the introduction of a new scaling parameter c. As an example, we consider the Beta (a, b) distribution. Suppose, as before, that Y_t is seasonal and that the marginal of Y_t^* is beta distributed but with *three* parameters a, b and c. By this we mean that Y_t^*/c is Beta (a, b). We now recall that $Y_t = s_{t \bmod p} Y_t^*$, so Y_t is Beta $(a, b, (c/s_t))$. The new parameter c can now be estimated along with a and b, taking care that it is constrained to exceed the maximum data value.

NON-STATIONARY SERIES

To date, we have fixed our attention on non-stationary, possibly seasonal series and have ignored the problems involved with trends and other kinds of non-stationarity. This is not to say that the problem is intractable: the methods outlined in the section above extend fairly easily to series with trends of various types. What has to be realized is that one cannot approach the problem using differencing techniques. Rather, one has to take explicit account of the trend and work with a finite sample problem. We hope to expand these ideas in a forthcoming paper.

CONCLUSIONS

The JS model provides a straightforward and conceptually simple model for series with non-normal marginal distributions. Experience shows that when the appropriate conditions are satisfied the technique works well and that in some cases (for instance, stable distributions) there is no competitor. By using the linearity factor we can predict the circumstances where the additional complexity involved in the JS model is not warranted. This gives new insights on the application of transformations. Much work remains to be done. More practical experience of using the model needs to be obtained, the current methods of estimation need refinement and we would like to investigate the selection of the appropriate marginal

distribution. One possibility would be to consider families of marginals with many parameters such as the Generalized Betas of Bookstaber and MacDonald (1987).

REFERENCES

Bookstaber, R. M. and MacDonald, J. B. A., 'A general distribution for describing security price returns', *Journal of Business*, **60** (1987), No. 3, 401–24.

Box, G. E. P. and Pierce, D., 'Distribution of residual autocorrelations in ARIMA models', *JASA*, **65**, (1970), 1509–26.

Epps, T. W., 'Testing that a stationary time series is Gaussian', *Ann. Math. Statist.*, (1987), 1683–9.

Fernandez, B. and Salas, J. D., 'Periodic gamma AR processes for operational hydrology', *Water Resources Research*, **22** (1986), No. 10, 1385–96.

Gardner, G., Harvey, A. C. and Phillips, G. D. A., 'An algorithm for the exact likelihood estimation of ARMA models by means of Kalman filtering', *Applied Stats*, **29** (1980), 311–22.

Gasser, T., 'The goodness of fit tests for correlated data', *Biometrika*, **62** (1975), 563–70.

Granger, C. W. J. and Newbold, P., 'Forecasting transformed series', *J. Roy. Statist. Soc.*, **38** (1976), 189–203.

Granger, C. W. J. and Newbold, P., *Forecasting Economic Time Series*, Academic Press, New York, 1987.

Harvey, A. C., *Time Series Models*. Phillip Allan, Oxford, 1981.

Janacek, G. J. and Swift, A. L., 'A class of models for non-normal time series', *J. Time Series Analysis*, (1990), 1, 19–32.

Karlsen, H. and Tjostheim, D., 'Consistent estimates for the NEAR(2) and NLAR(2) time series models', *J. Roy. Statist. Soc. B*, **50** (1988), No. 2, 313–20.

Lawrance, A. J. and Kottegoda, N. T., 'Stochastic modelling of riverflow times series', *J. Roy. Statist. Soc. A*, **140** (1977), No. 1, 1–47.

Lawrance, A. J. and Lewis, P. A. W., 'A new autoregressive time series model in exponential variables', *Adv. Appl. Prob.*, **13** (1981), 826–45.

Lawrance, A. J. and Lewis, P. A. W., 'Modelling and residual analysis of nonlinear autoregressive time series in exponential variables', *J. Roy. Statist. Soc. B*, **47** (1985), 165–202.

Lawrance, A. J. and Lewis, P. A. W., 'Residual analysis for nonlinear time series', *ISI Review*, **55** (1987), 1, 21–38.

Lin, C. C. and Mudhoekar, G. S., 'A simple test for normality against asymmetric alternatives', *Biometrika*, **67** (1980), 455–61.

Luutkonen, R. Saikkonen, P. and Terasvirta, T., 'Testing linearity against smooth transition autoregressive models', *Biometrika*, **75** (1988), 491–9.

Moran, P. A. P., 'Testing for correlation between non-negative variates', *Biometrika*, **54** (1967), 385–94.

Priestley, M. B., *Spectral Analysis and Time Series*, Academic Press, London, 1981.

Priestley, M. B., *Non-linear and Non-stationary Time Series Analysis* Academic Press, London, 1988.

Smith, R. L., 'Maximum likelihood estimation for the NEAR(2) model', *J. Roy. Statist. Soc.*, **48** (1986), 251–7.

Tong, H., *Threshold Models in Nonlinear Time Series*, Springer Notes in Statistics 21, Springer-Verlag, Berlin, 1983.

Weiss, G., 'Time reversibility of linear stochastic processes', *J. Applied Prob.*, **12** (1975), 831–6.

Wilson, G. T., 'Factorization of the generating function of a pure moving average process', *SIAM J. Numerical Anal.*, **6** (1969), 1–7.

Authors' biographies:

Louise Swift is currently a Lecturer in Statistics in the School of Mathematics at the University of East Anglia, Norwich. After her first degree she worked as an Investment Analyst before going to UEA to obtain her doctorate in Statistics and then to a lectureship in the School of Information Systems.

Gareth Janacek is currently a Lecturer in Statistics in the School of Mathematics at the University of East Anglia, Norwich. After gaining his MSc at Manchester he worked in Genetics before going to

Nottingham to obtain a PhD in Time Series Analysis. Publications include *Mathematics for Seismic Processing*, with A. Camina.

Authors' address:
Louise Swift and **Gareth Janacek**, Centre for Statistics, School of Mathematics, University of East Anglia, University Plain, Norwich NR4 7TJ, UK.