An Introduction to Forecasting with Time Series Models

William R. Bell

Time series models have become popular in recent years since the publication of the book by Box and Jenkins (1970), and the subsequent development of computer software for applying these models. Our purpose here is to review the use of time series models in forecasting. We will emphasize several important points about forecasting:

1. Forecasting by the fitting and extrapolation of a deterministic function of time is generally not a good approach.
2. Providing reasonable measures of forecast accuracy is essential - sometimes it is more important to find out that a series cannot be forecast than to obtain the "best" forecast.
3. Subject matter knowledge should not be thrown out the window when doing time series modelling and forecasting.

We shall demonstrate that the main difficulty with forecasting by fitting and extrapolating a deterministic function is that such an approach does not generally provide reasonable measures of forecast accuracy. The main advantage to time series models is not that they necessarily provide better (more accurate) forecasts, but that they do provide a means for obtaining reasonable measures of forecast accuracy. The route to better forecasts does not lie through time series models alone, but through the combination of time series models with subject matter

knowledge about the series being forecast. This can be done via regression plus time series models which we discuss briefly (or more generally, through multivariate time series models, which we will not cover here).

1. Difficulties With Using Deterministic Functions To Do Forecasting

A natural approach to forecasting would seem to be to view the observed time series as a function of time observed with error, specify a function of time, f(t), that looks appropriate for the data, fit f(t) to the data by least squares (although other fitting criteria could be used), and forecast by extrapolating f(t) beyond the observed data. One could also try to use regression theory to produce confidence intervals for the future observations. However, there are a number of difficulties with this approach, that we shall discuss in turn. We shall illustrate these difficulties by using this approach to forecast the time series of daily IBM stock prices, taking as observations the data from May 17, 1961 through September 3, 1961.

The IBM stock price data plotted in Figure 1 illustrate the first difficulty with fitting a deterministic function:

1. It is often difficult to find a suitable function of time.

Although Figure 1 does not suggest any obvious function, as a first attempt we might try fitting a straight line, $f(t) = \alpha + \beta t$, to the data. The resulting fit is quite poor, as is also shown in Figure 1 (with the fitted line extrapolated 20 time periods (days) beyond the last (110th) observation). The

quadratic, $f(t) = \alpha + \beta t + \gamma t^2$, shown in Figure 2, might be regarded as a better fit, though there are stretches of the data over which the quadratic also fits poorly.

The difficulty in finding a suitable f(t) to fit to data for forecasting is analogous to the same problem in graduation of data, which is well known to actuaries (see Miller 1942). The problem is more severe in forecasting, since it is easier to find a suitable function to fit for interpolation within the range of the observed data, than for extrapolation beyond the range of the data. This problem in graduation led to the development of graduation methods such as Whittaker-Henderson (see Whittaker and Robinson (1944)), and that of Kimeldorf and Jones (1967), which make use of local smoothness of an assumed underlying function, without requiring an explicit form for the function. These graduation methods can be thought of as analogous to the ARIMA time series models we shall discuss later.

Figure 2 illustrates another problem with the deterministic function approach, which is

2. The forecasts can exhibit unreasonable long-run behavior. The fitted quadratic in Figure 2 approaches $+\infty$ at an increasing rate as t increases. This is a problem in any given situation will depend on the length of the forecast period and how fast the fitted function deviates from reasonable behavior.

A third problem that can arise with the deterministic function approach is the following:

3. If the fitted values differ much from the data at the last few time points, short-run forecasts can be poor.

Another way of saying this, is that if the fit is bad at the end of the series the first few forecasts are likely to be bad. Figure 1 shows the straight line fits poorly at the end of the 110 observations used. Figure 3.a shows the last 31 observations of the data we are using (t = 80 to t = 110) along with the next 20 observations to be forecast (t = 111 to t = 130). We see the initial straight line forecasts are indeed poor, although the series eventually wanders down closer to the forecasts. The problem here is that in fitting the linear function (or any other function) by (ordinary) least squares, all the observations are given equal weight, so there is no guarantee that the fit will be good near the end of the series. Generally, time series models make use of the last few observations in a way that gives the model a much better chance to produce good short-run forecasts.

One way around the above problem is to only fit to data at the end of the series. For Figure 1, the stretch of data from, say, t = 91 (August 15, 1961) to t = 110 would seem to be more amenable to the fitting of functions than any longer stretch at the end of the series. A straight line provides a good fit to this part of the series as shown in Figure 4. Further analysis of the stock price data will use this straight line fit to the last 20 observations.

In addition to forecasting the stock prices, we would like to estimate forecast error variances, and produce forecast intervals for the future values (assuming normality). This may

be easily done using standard regression theory (see Miller and Wichern 1977, chapter 5). Figure 3.b shows the resulting 95 percent forecast band about the least squares prediction line for forecasting 20 future observations from t = 110, the forecast period covering the dates September 4, 1961 through September 23, 1961. We notice that the forecasts are rather poor beyond the first four. More importantly, the first two future observations lie near the boundary of the 95% forecast band, and the fifth through the twentieth observations lie well outside the band. For this example standard regression theory does not provide reasonable measures of forecast accuracy. An investor using this approach to forecast future IBM stock prices from September 3, 1961 would have been given an unreasonable degree of confidence in the projected future linear increase in the stock price - an increase which failed to occur.

These results illustrate the fourth, and most important, problem with the deterministic function approach to forecasting:

4. Variances of forecast errors from regression theory are usually highly unrealistic.

The general regression model underlying the deterministic function approach is $Y_t = f(t) + e_t$ for $t = 1,\ldots,n$, where the Y_t are the n observations, and the e_t are assumed to be random (uncorrelated) error terms. The primary problem with forecast error variances from regression theory is not the difficulty in finding a suitable f(t), but rather the assumption that the errors, e_t, are uncorrelated. Time series observations are rarely uncorrelated, and are typically nonstationary in a way

that implies very high correlation in the observations. Such high correlation can easily result in grossly understated prediction error variances.

The goal of time series models is to provide a reasonable approximation to the correlation structure of the data via a model with a small number of parameters (in relation to the length of the series). When this is done it will often be seen that observed patterns in the data were in fact not due to the presence of some underlying smooth function, but merely to the high degree of correlation in the data, which is accounted for by the time series model.

The preceeding treatment was elementary, but was deliberately so in an effort to make clear some difficulties with fitting a deterministic function to a time series for the purpose of forecasting. Of the difficulties mentioned, we regard the problem of obtaining reasonable forecast error variances (so that probability statements about the future can be made), as the most important. In the constant search for forecasting methods to produce "better", i.e., more accurate, forecasts, the problem of producing good (or just reasonable) estimates of forecast error variances has frequently been overlooked by forecasters. We regard the problem of estimating forecast error variances as just as important as that of estimating future values. Sometimes it is more important to learn that you cannot forecast a series than to get the "best" forecast of it.

In the next section we discuss the use of ARIMA time series models in forecastng. While these models will not necessarily

lead to more accurate forecasts, they will almost certainly help the forecaster estimate forecast error variances, something some other approaches to forecasting cannot do at all.

2. ARIMA Time Series Models and Forecasting

As noted earlier, time series typically feature correlation between the observations. Time series models attempt to account for this correlation over time through a parametric model. Here we shall discuss the use of ARIMA (autoregressive - integrated - moving average) time series models in forecasting. We shall not provide the rationale behind these models, or discuss approaches to modeling, but refer the reader to the books by Box and Jenkins (1970) and Miller and Wichern (1977). We will assume the time series has been modelled and the model is known.

ARIMA models include the (purely) autoregressive (AR) model

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + a_t \qquad (2.1)$$

where $\phi_1, \ldots, \phi_p$ are parameters, the a_t's are independent, identically distributed $N(0,\sigma_a^2)$, and we assume, for now, $E(Y_t)=0$. Letting B be the backshift operator ($BY_t = Y_{t-1}$) we can write (2.1) as

$$(1 - \phi_1 B - \ldots - \phi_p B^p) Y_t = a_t \qquad (2.2)$$

or $\phi(B)Y_t = a_t$ where $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$. The (purely) moving average (MA) model is

$$Y_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad (2.3)$$

or

$$Y_t = (1 - \theta_1 B - \cdots - \theta_q B^q) a_t \qquad (2.4)$$

or $Y_t = \theta(B) a_t$. Including both autoregressive and moving average operators gives the ARMA(p,q) model

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad (2.5)$$

which we write as

$$(1-\phi_1 B - \cdots - \phi_p B^p) Y_t = (1-\theta_1 B - \cdots - \theta_q B^q) a_t \qquad (2.6)$$

or $\phi(B) Y_t = \theta(B) a_t$. For reasons we shall not go into fully here, we shall assume the zeroes of the polynomials $1 - \phi_1 x - \cdots - \phi_p x^p$ and $1 - \theta_1 x - \cdots - \theta_q x^q$ are greater than one in absolute value. The first of these conditions implies that the series Y_t following (2.5) is stationary. In practice, Y_t may well be nonstationary, but with stationary first difference, $Y_t - Y_{t-1} = (1-B) Y_t$. If $(1-B) Y_t$ is nonstationary we may need to take the second difference, $Y_t - 2Y_{t-1} + Y_{t-2} = (1-B)[(1-B) Y_t] = (1-B)^2 Y_t$. In general, we may need to take the dth difference $(1-B)^d Y_t$ (although rarely is d larger than 2). Substituting $(1-B)^d Y_t$ for Y_t in (2.6) yields the ARIMA(p,d,q) model

$$(1-\phi_1 B - \cdots - \phi_p B^p)(1-B)^d Y_t = (1-\theta_1 B - \cdots - \theta_q B^q) a_t \qquad (2.7)$$

or $\phi(B)(1-B)^d Y_t = \theta(B) a_t$. We shall also write this as

$$(1 - \Phi_1 B - \cdots - \Phi_{p+d} B^{p+d}) Y_t = (1 - \Theta_1 B - \cdots - \Theta_q B^q) a_t \qquad (2.8)$$

where $1 - \Phi_1 B - \cdots - \Phi_{p+d} B^{p+d} = (1 - \phi_1 B - \cdots - \phi_p B^p)(1-B)^d$.

If Y_t is stationary ($d=0$) it is usually inappropriate to assume $E(Y_t) = 0$, thus, Y_t in (2.1) - (2.6) should be replaced by $Y_t - \mu_y$ ($\mu_y = E(Y_t)$). For (2.6) this gives $\phi(B)(Y_t - \mu_y) = \Theta(B)a_t$. If $d > 0$ then $(1-B)^d(Y_t - \mu_y) = (1-B)^d Y_t$ since $(1-B)\mu_y = 0$, so we do not use $Y_t - \mu_y$ in (2.7). If it turns out that $(1-B)^d Y_t$ has a nonzero mean, this can be allowed for by using the model

$$\phi(B)(1-B)^d Y_t = \Theta_0 + \Theta(B)a_t \quad . \qquad (2.9)$$

In any of the above models Y_t could be some transformation of the original data, such as a power transformation (see Miller 1984).

2.1 Forecasting With ARIMA Models

To illustrate forecasting with ARIMA models, we shall use (2.8) written as

$$Y_{n+\ell} = \Phi_1 Y_{n+\ell-1} + \cdots + \Phi_{p+d} Y_{n+\ell-p-d} + a_{n+\ell} - \Theta_1 a_{n+\ell-1} - \cdots - \Theta_q a_{n+\ell-q} \qquad (2.10)$$

for $t = n+\ell$. We shall assume we want to forecast $Y_{n+\ell}$ for $\ell = 1,2,\ldots$ using data $Y_n, Y_{n-1},\ldots$ For simplicity, we are assuming for now that the data set is long enough so that we may

effectively assume it extends into the infinite past. (In practice, given the model, this assumption is typically innocuous, and it can be dispensed with if necessary - see Ansley and Newbold (1981).) The best (minimum mean squared error) forecast is given by the conditional expectation $E(Y_{n+\ell} \mid Y_n, Y_{n-1},\ldots) = \hat{Y}_n(\ell)$. From (2.10), $\hat{Y}_n(\ell)$ satisfies

$$\hat{Y}_n(\ell) = \phi_1\hat{Y}_n(\ell-1) + \ldots + \phi_{p+d}\hat{Y}_n(\ell-p-d) + \hat{a}_n(\ell) - \theta_1\hat{a}_n(\ell-1) - \ldots - \theta_q\hat{a}_n(\ell-q). \tag{2.11}$$

$\hat{Y}_n(\ell)$ can be computed recursively from (2.11) for $\ell = 1,2,3,\ldots$ using

$$\hat{Y}_n(j) = \begin{cases} Y_{n+j} & j \leq 0 \\ \hat{Y}_n(j) & j > 0 \end{cases} \qquad \hat{a}_n(j) = \begin{cases} a_{n+j} & j \leq 0 \\ 0 & j > 0 \end{cases}$$

Since a_t is independent of $Y_{t-1}, Y_{t-2},\ldots$, $\hat{a}_n(j) = 0$ for $j > 0$. The a_t's can be computed from the model using $Y_t, Y_{t-1},\ldots$, as discussed in Box and Jenkins (1970) (basically $a_t = \theta(B)^{-1}\phi(B)(1-B)^d Y_t$). Notice that for $\ell > q$ we get

$$\hat{Y}_n(\ell) = \phi_1\hat{Y}_n(\ell-1) + \ldots + \phi_{p+d}\hat{Y}_n(\ell-p-d) \qquad \ell > q \tag{2.12}$$

which can also be written $\phi(B)\hat{Y}_n(\ell) = 0$ with B operating on ℓ. Thus, $\hat{Y}_n(\ell)$ as a function of ℓ (called the forecast function)

satisfies a homogeneous difference equation of order p+d for $\ell > q$, with starting values $\hat{Y}_n(q), \hat{Y}_n(q-1),\ldots, \hat{Y}_n(q-p-d+1)$.

We are also interested in properties of the forecast error $Y_{n+\ell} - \hat{Y}_n(\ell)$. Box and Jenkins (1970) observe that

$$Y_{n+\ell} - \hat{Y}_n(\ell) = a_{n+\ell} + \psi_1 a_{n+\ell-1} +\ldots+ \psi_{\ell-1} a_{n+1} \tag{2.13}$$

where $\psi_1,\psi_2,\ldots$ are solved for by equating coefficients of $x, x^2, x^3,\ldots$ in

$$(1-\Phi_1 x-\ldots-\Phi_{p+d}x^{p+d})(1+\psi_1 x+\psi_2 x^2 +\ldots) = 1-\Theta_1 x-\ldots-\Theta_q x^q \tag{2.14}$$

so $\psi_1 = \Phi_1 - \Theta_1$

$\psi_2 = \Phi_2 + \Phi_1\psi_1 - \Theta_2$

etc.

For $j > q$ $\psi_j = \Phi_1\psi_{j-1} +\ldots+ \Phi_{p+d}\psi_{j-p-d}$ so the ψ_j's satisfy the same homogeneous difference equation as the forecast function. The variance of the ℓ-step ahead forecast error, $V(\ell)$, is easily seen from (2.13) to be

$$V(\ell) = \mathrm{Var}(Y_{n+\ell}-\hat{Y}_n(\ell)) = (1+\psi_1^2+\ldots+\psi_{\ell-1}^2)\sigma_a^2 \tag{2.15}$$

Observe that a_{n+1} is the one-step ahead forecast error with variance σ_a^2.

In practice, we substitute estimates of the parameters Φ_j, Θ_j, and σ_a^2 in (2.11) and (2.15) to estimate $\hat{Y}_n(\ell)$ and $V(\ell)$. Assuming normality, we can use these to make probability

statements about $Y_{n+\ell}$ given the data through time n. For example, a 95 percent forecast interval for $Y_{n+\ell}$ is

$$\hat{Y}_n(\ell) - 1.96\,(V(\ell))^{1/2} < Y_{n+\ell} < \hat{Y}_n(\ell) + 1.96(V(\ell))^{1/2} \qquad (2.16)$$

2.2 Forecasting for Some Particular Models

(1,0,0) Model: AR(1)

For the AR(1) model, $Y_t - \mu_y = \phi_1(Y_{t-1}-\mu_y) + a_t$ with $|\phi_1| < 1$, we see from (2.11) that (replacing $\hat{Y}_n(j)$ by $\hat{Y}_n(j) - \mu_y$)

$$\hat{Y}_n(1) = \mu_y + \phi_1(Y_n - \mu_y)$$

$$\hat{Y}_n(\ell) = \mu_y + \phi_1(\hat{Y}_n(\ell-1) - \mu_y) \quad \ell > 1.$$

Using these results, it is easy to show that

$$\hat{Y}_n(\ell) = \mu_y + \phi_1^{\ell}(Y_n - \mu_y) \quad \ell \geq 1.$$

Using (2.14) it can be shown that $\psi_j = \phi_1^j$, so

$$V(\ell) = (1 + \phi_1^2 + \ldots + \phi_1^{2(\ell-1)})\sigma_a^2.$$

Notice that as $\ell \to \infty$, $\hat{Y}_n(\ell) \to \mu_y$ and $V(\ell) \to \sigma_a^2/(1-\phi_1^2)$, which is $\mathrm{Var}(Y_t)$. For the stationary AR(1) model, the forecast function damps out exponentially to the series mean, and the forecast variance converges to the variance of the series.

(0,1,0) Model: "Random Walk"

For the random walk model, $Y_t = Y_{t-1} + a_t$, (2.11) yields

$$\hat{Y}_n(1) = Y_n \qquad \hat{Y}_n(\ell) = Y_n(\ell-1) = Y_n \quad \ell > 1.$$

Here the forecast for all lead times is simply the last observation. Also, $\psi_j = 1 \quad j \geq 1$ so that $V(\ell) = (\ell-1)\sigma_a^2$. Notice these results are analogous to those for the AR(1) model but with $\phi_1 = 1$ and $\mu_y = 0$. However, unlike the stationary AR(1) model, as $\ell \to \infty$ $\hat{Y}_n(\ell) \to Y_n$ and $V(\ell) \to \infty$.

(0,1,1) Model: "Exponential Smoothing"

For the (0,1,1) model, $Y_t = Y_{t-1} + a_t - \theta_1 a_{t-1}$, (2.11) leads to

$$\hat{Y}_n(1) = Y_n - \theta_1 a_n \qquad \hat{Y}_n(\ell) = \hat{Y}_n(1) \quad \ell > 1.$$

It can be shown that

$$\hat{Y}_n(\ell) = (1-\theta_1)[Y_n + \theta_1 Y_{n-1} + \theta_1^2 Y_{n-2} + \ldots]$$

so that the forecast is an "exponentially weighted moving average" of the past observations (the weights on the past observations sum to 1). Forecasting with this model is referred to as "exponential smoothing". (2.14) leads to $\psi_j = 1-\theta_1$ for all j so that $V(1) = \sigma_a^2$ and

$$V(\ell) = [1 + (1-\theta_1)^2(\ell-2)]\sigma_a^2 \qquad \ell \geq 2 \bullet$$

2.3 Properties of Forecast Functions

If Y_t follows (2.7) and (2.8), then $\hat{Y}_n(\ell)$ satisfies the homogeneous difference equation (2.12) with starting values $\hat{Y}_n(q),\ldots,\hat{Y}_n(q-p-d+1)$. Fuller (1976, section 2.4) gives properties of solutions to difference equations, which may be used to show that $\hat{Y}_n(\ell) = \sum_{i=1}^{p} \beta_i \xi_i^{\ell} + (\alpha_0 + \alpha_1\ell + \ldots + \alpha_{d-1}\ell^{d-1})$ where $\xi_1^{-1},\ldots,\xi_p^{-1}$ are the zeroes of $1 - \phi_1 x - \ldots - \phi_p x^p = \phi(x)$ (for simplicity, we assume these are distinct), and the coefficients $\beta_1,\ldots,\beta_p,\alpha_0,\ldots\alpha_{d-1}$ are determined by the starting values. If Y_t follows the model (2.9), then $\hat{Y}_n(\ell)$ satisfies the non-homogeneous difference equation obtained by adding θ_0 to the right hand side of (2.12). The effect of this on the solution for $\hat{Y}_n(\ell)$ is to add a term $\alpha_d\ell^d$, where $\alpha_d = \theta_0/(1-\phi_1 - \cdots - \phi_p)d!$.

Using (2.12) - (2.15), and properties of solutions to difference equations, one can establish the following general results.

(i) If $d=0$, so Y_t is stationary,
$\hat{Y}_n(\ell) \to \mu_y$ and $V(\ell) \to \mathrm{Var}(Y_t)$ as $\ell \to \infty$
($\mu_y = \theta_0/(1-\phi_1 - \ldots - \phi_p)$ in (2.9) and is 0 in (2.7))

(ii) If $d>0$, so Y_t is nonstationary, $\hat{Y}_n(\ell)$ is eventually dominated by a polynomial of degree $d-1$ if $\theta_0 = 0$, and of degree d if $\theta_0 \neq 0$, and
$V(\ell) \to \infty$ as $\ell \to \infty$.

For the particular case of the (0,d,q) model in (2.7), (2.12) becomes $(1-B)^d \hat{Y}_n(\ell) = 0$, so that $\hat{Y}_n(\ell)$ exactly follows a polynomial of degree d-1 for $\ell > q$.[1] The coefficients of the polynomial are determined by the starting values, $\hat{Y}_n(q),\ldots,\hat{Y}_n(q-p-d-1)$, which in turn depend on $Y_n, Y_{n-1},\ldots$. The polynomial is adaptive and need only apply locally, i.e., its coefficients are redetermined as each new data point is added. This contrasts with simply fitting a single polynomial over the entire range of the data.

For the model $(1-B)^d Y_t = \Theta_0 + \Theta(B)a_t$ ((2.9) with $\phi(B) = 1$), $\hat{Y}_n(\ell)$ is a polynomial in ℓ of degree d, with the coefficient of ℓ^d equal to $\Theta_0/d!$. The forecast function here is non-adaptive in that the same Θ_0 is used at each time point. If a "polynomial plus error" model, $Y_t = \alpha_0+\alpha_1 t+\ldots+\alpha_d t^d+a_t$, is really appropriate, then the time series modelling process should lead to the model

$$(1-B)^d Y_t = \Theta_0 + (1-B)^d a_t \qquad (\Theta_0 = d!\alpha_d).$$

Solving this difference equation for Y_t leads back to the polynomial plus error model (see Box and Abraham (1978)). Thus, ARIMA models allow for polynomial projection when appropriate.

---

[1] Keyfitz (1972) has suggested one way demographic projections might be done is by passing a polynomial of some degree d through the last d data points. This in fact corresponds to forecasting with an ARIMA(0,d+1,0) model.

2.4 Example: IBM Stock Price Series

For the IBM stock price data, two models were fitted to the full stretch of data from May 17, 1961 through September 3, 1961. These were the (0,1,1) model and the (0,1,1) model with trend:

$$(1-B)Y_t = (1-\theta_1 B)a_t \qquad \hat{\theta}_1 = -.29 \quad \hat{\sigma}_a^2 = 26.0$$

$$(1-B)Y_t = \theta_0 + (1-\theta_1 B)a_t. \qquad \hat{\theta}_1 = -.26 \quad \hat{\theta}_0 = 1.20 \quad \hat{\sigma}_a^2 = 25.3$$

Twenty forecasts from September 3, 1961 are shown for these models in Figures 5 and 6. We notice either of these models produces better forecasts than the straight line fits in Figures 3a and 3b. However, this is partly due to the fact that we selected the stretch of data we are using to illustrate the dangers of fitting and extrapolating a straight line. The important difference is in the forecast intervals. The intervals for the time series models are quite wide and increase substantially with increasing ℓ, allowing for a wide range of behavior for the future stock prices. The interval from the straight line model in Figure 5 is clearly too narrow. The message from the time series models is quite clear: the IBM stock price series is difficult to forecast. It is much more important to learn this from the model, than to get the "best" forecast, which is likely to be inaccurate anyway.

2.5 Seasonal Models

If the series exhibits periodic behavior to some degree (such as an annual period in monthly or quarterly data) then the ARIMA

models discussed above need to be enhanced. For a seasonal series with period s, we can use the seasonal ARIMA(p,d,q) x $(P,D,Q)_s$ model as discussed in Box and Jenkins (1970). For example, for monthly data one useful model is the $(0,1,1) \times (0,1,1)_{12}$ model

$$(1-B)(1-B^{12})Y_t = (1-\theta_1 B)(1-\theta_{12}B^{12})a_t$$

($B^{12}Y_t = Y_{t-12}$). Models such as this produce forecast functions with periodic behavior.

2.6 Weak Points in the ARIMA Approach

There are some difficulties with using ARIMA models in forecasting that users should be aware of, especially since research may suggest improved procedures to deal with these problems. Since there is no difficulty with the forecasting mathematics once we know the ARIMA model, the problems have to do with the fact that we never really know the model.

Even if we know the orders (p,d,q) of the ARIMA model, the parameters can only be estimated from the data. This introduces additional error into the forecasts which is not accounted for in $V(\ell)$. Fortunately, for long series (large n) the effect of parameter estimation error on forecasts and forecast error variances can be shown to be negligible (Fuller 1976, section 8.6). The problem is more important for short series. It has been investigated by Ansley and Newbold (1981) who suggest a means of inflating $V(\ell)$ to allow for parameter estimation error. Another approach to this problem is to use the bootstrap

technique to assess the forecast accuracy (see Freedman and Peters (1982)).

In practice the true model is never known, and certainly need not be an ARIMA model. However, ARIMA models are sufficiently flexible to well-approximate the correlation structure of many time series. For forecasting the most important part of an ARIMA model to get right is the differencing order. Even if we do not get this right at the identification stage, fitting a model with AR terms may tell us that differencing is needed. To illustrate, the (1,1,0) model for the birth rates analyzed in section 3.4 can be written

$$(1-B)(1-\phi_1 B)Y_t = (1 - (1+\phi_1)B + \phi_1 B^2)Y_t = a_t.$$

So we could have fit the AR(2) model $(1-\Phi_1 B-\Phi_2 B^2)Y_t = a_t$ and examined the estimated $1 - \Phi_1 B - \Phi_2 B^2$ to see if it contained a factor (1-B). In doing this our estimates, $\hat{\Phi}_1$ and $\hat{\Phi}_2$, converge rapidly in probability to values producing a "unit root" (a (1-B) factor) in $1 - \hat{\Phi}_1 B - \hat{\Phi}_2 B^2$ (Fuller 1976).

2.7 Summary and Demographic Applications

Forecasters have traditionally developed new forecasting methods in an attempt to produce more accurate forecasts. While this is important, it is also crucial to provide good estimates of forecast error variability. Some series are inherently difficult to forecast, and finding this out is more important than refining the point forecast. ARIMA time series models are a

flexible class of models that can be used in many situations to produce both reasonable forecasts and reasonable estimates of forecast error variance.

Keyfitz (1972, 1981) has also argued that providing measures of the expected size of forecast errors is essential, and he notes that population forecasters have virtually unanimously failed to do this. Keyfitz (1981) presents empirical measures of forecast accuracy for historical population forecasts as a guide to accuracy of current and future forecasts. Stoto (1983) also analyzes the accuracy of historical population forecasts. He analyzes the forecast errors to produce estimates of forecast error variance, and then develops confidence intervals for United States population through the year 2000. McDonald (1981) uses ARIMA models to forecast an Australian births series.

3. Use of Subject Matter Knowledge in Forecasting

The forecaster should not discard his or her subject matter knowledge when using time series models in forecasting. ARIMA models attempt to account for the correlation over time in time series data, and then use this correlation in forecasting. They cannot deal with other forecasting problems that may require interaction of subject matter knowledge about the series being forecast with time series models.

3.1 Deciding What Time Series to Forecast

As pointed out by Long (1984), this is a traditional problem faced by demographers doing population projections. For example,

consider the basic demographic accounting relation (P_t = population at time t)

$$P_t = P_{t-1} + \text{Births}_t - \text{Deaths}_t + \text{Inmigration}_t - \text{Outmigration}_t.$$

In forecasting P_t we must decide whether to forecast P_t directly, or indirectly by forecasting the components. We must also decide whether to break down the series by age, sex, race or other factors (see Long (1984) for a discussion). Once the series to be forecast have been decided upon, time series models can be useful in doing the forecasting.

Another aspect of this is the selection of a transformation to be used, if any, on the series. To an extent this is a statistical problem (see Miller 1984), but transformations involve a rescaling of the data, the implications of which should be considered. For example, if Y_t is a series of proportions the logistic transformation, $z_t = \ln(Y_t/(1-Y_t))$, can be useful. In logistically transforming the interval (0,1) to $(-\infty,\infty)$, the variation in Y_t when it is near 0 or 1 is enhanced relative to the variation when Y_t is not near the boundary. Forecasting z_t and transforming back via $Y_t = \exp(z_t)/(1+\exp(z_t))$, will produce forecasts and confidence intervals for Y_t that do not stray outside the interval (0,1).

3.2 Deciding What Part of the Data to Use

Time series methods, like other statistical methods, work better when more observations are available. However, this

assumes that all portions of the series follow the same model. Real series may, for example, be affected by structural changes, unusual events, or changes of definition. While it is best to avoid these difficulties, this conflicts with the need to use as long a series as possible when modelling. Knowledge about the series being forecast can help in deciding how much of the past to use in modelling.

3.3 Regression Plus Time Series Models

In some cases it is possible to explicitly incorporate subject matter knowledge into the forecast model. A useful means of doing this is to use regression plus time series models. These are closely related to transfer function models (also called distributed lag models). McDonald (1981) fits models of this type to an Australian births series. More generally, multivariate time series models might be used - see Tiao and Box (1981) for a general discussion, and Miller (1984) for a demographic example.

The regression plus ARIMA(p,d,q) time series model is

$$\phi(B)(1-B)^d[Y_t - (\beta_1 X_{1t}+\ldots+\beta_k X_{kt})] = \Theta(B)a_t \qquad (3.1)$$

where $X_{1t},\ldots,X_{kt}$ are the independent variables and $\beta_1,\ldots,\beta_k$ the regression parameters. Inference results for models of the form (3.1) are given in Pierce (1971). Forecasts can be obtained for $\ell = 1,2,\ldots$ by writing (3.1) as

$$Y_{n+\ell} = \Phi(B)[\beta_1 X_{1,n+\ell} + \ldots + \beta_k X_{k,n+\ell}] + \Phi_1 Y_{n+\ell-1} + \ldots + \Phi_{p+d} Y_{n+\ell-p-d} + \Theta(B) a_{n+\ell}. \quad (3.2)$$

To produce forecasts of $Y_{n+\ell}$ from (3.2) requires future values of the X_{it} series. The accuracy of the forecast for any ℓ will depend on the extent to which the X_{it} are "known" through time $n+\ell$.

The ideal situation is where the X_{it} are known exactly for all t. This can happen in practice - Bell and Hillmer (1983) discuss the use of regression plus time series models for economic series exhibiting calendar variation, where the X_{it} are functions of the calendar and thus are known for all time. When the future X_{it}'s are not known, they must be forecast as well (this is really what distinguishes a transfer function model from a regression plus time series model), and the accuracy of the resulting forecast of $Y_{n+\ell}$ will obviously depend on the accuracy of the $X_{i,n+\ell}$ forecasts. Also, the forecast error variance should include the additional error variance due to forecasting the $X_{i,n+\ell}$ (see Box and Jenkins (1970, section 11.5)).

An intermediate situation is where Y_t depends on the value of another series, W_t, at an earlier time point. A simple example would be the model

$$(1-B)[Y_t - \beta_1 W_{t-r}] = (1-\Theta_1 B) a_t.$$

In this case W_t is a leading indicator for Y_t. It will be known

exactly when forecasting $Y_{n+\ell}$ for $\ell = 1,\ldots,r$, but must be forecast after that.

To clarify the roles of the regression and time series parts of the model, let the observed data $\underset{\sim}{Y} = (Y_1,\ldots,Y_n)'$ have mean vector $\underset{\sim}{\mu} = (\mu_1,\ldots,\mu_n)'$, and nxn covariance matrix $\Sigma = (\mathrm{Cov}(Y_i,Y_j))$. Notice μ_t is not constant over time here. To forecast $Y_{n+\ell}$, we use the covariances $\underset{\sim}{g}' = (\mathrm{cov}(Y_{n+\ell},Y_1),\ldots,\mathrm{Cov}(Y_{n+\ell},Y_n))$. The best (minimum mean squared error) forecast of $Y_{n+\ell}$ given $\underset{\sim}{Y}$ is the conditional expectation, which under the multivariate normal distribution, is

$$E(Y_{n+\ell}|\underset{\sim}{Y}) = \mu_{n+\ell} + \underset{\sim}{g}'\Sigma^{-1}(\underset{\sim}{Y}-\underset{\sim}{\mu}) \quad . \tag{3.3}$$

The objective of the regression part of the model is to model μ_t as $\beta_1 X_{1t}+\ldots+\beta_k X_{kt}$, thus getting at $\mu_{n+\ell}$ and $\underset{\sim}{\mu}$ in (3.3). Time series models, on the other hand, seek a parametric model to describe $\mathrm{Cov}(Y_t,Y_{t+j})$, and hence $\underset{\sim}{g}'$ and Σ in (3.3). Thus, regression models and time series models are complementary and should not be viewed as competitors. Just as it is unwise to use pure regression models with correlated data (as was illustrated in section 1), it is also unwise to blindly apply pure time series models to a series known to be affected by certain independent variables.

3.4 Example: Forecasting Birth Rates and Births

We shall illustrate some of the considerations mentioned in the previous sections by using data through 1975 to forecast time

series of live births to women in age groups 20 to 24 and 25 to 29 (all races) in the U.S.[2] The data are plotted in Figure 7. In forecasting these series we will try to point out some places where subject matter expertise can play a role. However, since demography is not our area of expertise, we caution the reader that this exercise is for illustrative purposes only.

The first question to address is what time series to forecast. We have chosen to use data on 5-year age groups and all races for illustration, although it might be better to use single year of age data broken down by race, as is done by the Census Bureau (Long 1984). We also will be doing a period analysis, looking at the data for successive calendar years, whereas it might be better to proceed on a cohort basis. In these respects, subject matter expertise might suggest choices other than those we have made here for simplicity in our illustration.

Rather than forecasting the births series, B_t, directly, we shall forecast the birth rates, R_t, and apply these to Census Bureau population projections for women 20-24 and 25-29 to forecast births. (We shall let B_t, R_t, and Y_t refer to either age group.) For simplicity, we shall assume the Census Bureau projections introduce no additional error into our forecasts (actually, for 1976 - 1981 the observed population estimates were used so there is no additional error for these years). This can

---

[2] The birth rate data are given in Miller (1984). Census Bureau population estimates and projections were taken from Bureau of the Census (1982a,b).

be partly justified since we shall only forecast up through 20 years ahead (1995), so the women 20-24 and 25-29 in the forecast period were all alive in 1975. Thus, the errors in the Census Bureau projections for these groups through 1995 are due entirely to errors in forecasting deaths and migration - errors far less serious than those due to forecasting fertility (Long 1984).

Another decision we must make is whether to use some transformation of the birth rates. Miller (1984) investigates the use of power transformations for these and the 15-19, 30-34, and 35-39 birth rates, finding some evidence, though weak, for use of the reciprocal transformation. For this reason, and for another reason to be mentioned later, we will directly forecast $Y_t = R_t^{-1}$. We might even try to make a demographic interpretation of this transformation. Since annual birth rates are defined as

$$R_t = \frac{\text{Number of Births}}{(\text{Number of Women}) \times (\text{Number of Years})}$$

(in this case Number of Years = 1 since the births were tabulated annually), the units on $Y_t = R_t^{-1}$ are woman-years per birth. So Y_t represents the average waiting time to birth in year t for the given age group.

The next question to address is what part of these series to use in modelling and forecasting. Figure 7 shows the birth rate series for 1917 through 1980. There are sharp drops in both birth rate series during World War II. To get around this problem we could either use only the data after the dips (roughly 1948 and beyond), or we could put regression terms in our models

involving indicator variables for the affected years. However, Miller and Hickman (1981) found significant evidence of model change when comparing models for the pre-war and the post-war "baby boom" data. Therefore, we shall fit the (1,1,0) model considered by Miller (1984) to the post-war data. One could speculate that the "baby boom" was an aberration and that birth rates have returned to normal, which would suggest fitting the (1,1,0) models to the pre-war data and then applying them to the end of the series to forecast.

The (1,1,0) models were fitted to the 1948 - 1975 data with the following results:

Age 20-24	$(1-.72B)(1-B)Y_t = a_t$	$\hat{\sigma}_a^2 = .591\times10^{-7}$
Age 25-29	$(1-.70B)(1-B)Y_t = a_t$	$\hat{\sigma}_a^2 = .568\times10^{-7}$.

These models were used to produce forecasts $\hat{Y}_n(\ell)$, for 1976 - 1995, leaving the five data points for 1976 - 1980 for comparing forecasts to actual data. Upper and lower 95 percent forecast limits were obtained from $U(\ell) = \hat{Y}_n(\ell) + 1.96(V(\ell))^{1/2}$ and $L(\ell) = \hat{Y}_n(\ell) - 1.96\ (V(\ell))^{1/2}$ for $\ell = 1,\ldots,20$. These were inverted to point forecasts and forecast limits for the birth rates:

$$\hat{R}_n(\ell) = \hat{Y}_n(\ell)^{-1} \qquad \ell = 1,\ldots,20$$

$$LR(\ell) = U(\ell)^{-1} \qquad UR(\ell) = L(\ell)^{-1}$$

(Notice $.95 = \Pr(L(\ell) < Y_{n+\ell} < U(\ell)) = \Pr(U(\ell)^{-1} < Y_{n+\ell}^{-1} < L(\ell)^{-1})$). The results are shown in Figures 8 and 9, which show the 1948 - 1980 data, the forecasted birth rates, and the forecast intervals. We notice the first five forecasts for the 20-24 group are rather accurate, and those for the 25-29 are less so, the first two there falling on the upper 95 percent limit. In both cases the forecast intervals widen rapidly with increasing ℓ, reflecting considerable uncertainty when forecasting much more than 5 steps ahead. Long-run forecast accuracy will not come from the data used here. It will require other knowledge about the series.

Notice that the forecast intervals in Figures 8 and 9 are highly asymmetric, widening much faster above the forecasts than below them. The reciprocal transformation is responsible for this. It in fact prevents the nonsensical situation of the lower limit becoming negative, which is the reason for using it alluded to earlier. Had we chosen to forecast R_t directly using a(1,1,0) model, the results would be as shown in Figure 10 (for age group 20-24). While the point forecasts are little affected by the reciprocal transformation, the forecast limits are considerably different depending on whether or not we transform.

Finally, point forecasts and forecast limits for the births were obtained as

$$\hat{B}_n(\ell) = P_{n+\ell}\hat{R}_n(\ell)$$

$$LB(\ell) = P_{n+\ell}LR(\ell) \qquad UB(\ell) = P_{n+\ell}UR(\ell)$$

where $P_{n+\ell}$ is the Census Bureau population projection (assumed error free). Figures 11 and 12 show the actual births, forecasts, and forecast intervals for age groups 20-24 and 25-29. The point forecasts and limits are modulated by the fluctuations in $P_{n+\ell}$, producing behavior we would not have obtained by forecasting births directly.

The next step in trying to improve on our forecasts of 20-24 and 25-29 births might be to look for other variables to include in a regression plus time series model for Y_t. We might try using the 20-24 birth rates as a 5-year leading indicator in a model for the 25-29 birth rates. This was tried with no success. While there is a strong contemporaneous linear relationship between the series (Miller 1984), this will not help in long-run forecasting since neither of these series is easy to forecast far ahead. Inclusion of economic variables might improve short-run forecasts of the birth rates; long-run forecasts of the birth rates would require long-run forecasts of the economic variables, which are likely to be quite inaccurate.

3.5 Automatic Forecasting Procedures

Many automatic forecasting procedures have been proposed, some of which involve the automatic selection and fitting of time series models, and computer programs have been marketed for their use. While such procedures may provide reasonable forecasts in many cases, they have some important disadvantages. One is that some of the procedures are ad-hoc and do not provide estimates of forecast error variance. Also, automatic approaches necessarily

dissociate the forecaster from her or his data, making it difficult to include subject matter expertise in the forcasting process, and restricting the forecaster's ability to deal with unusual problems that arise. They also tend to reduce what the forecaster learns from the data in the forecasting process.

Figure 1

IBM STOCK PRICE -- 5/17/61 TO 9/3/61
(WITH STRAIGHT LINE FIT)

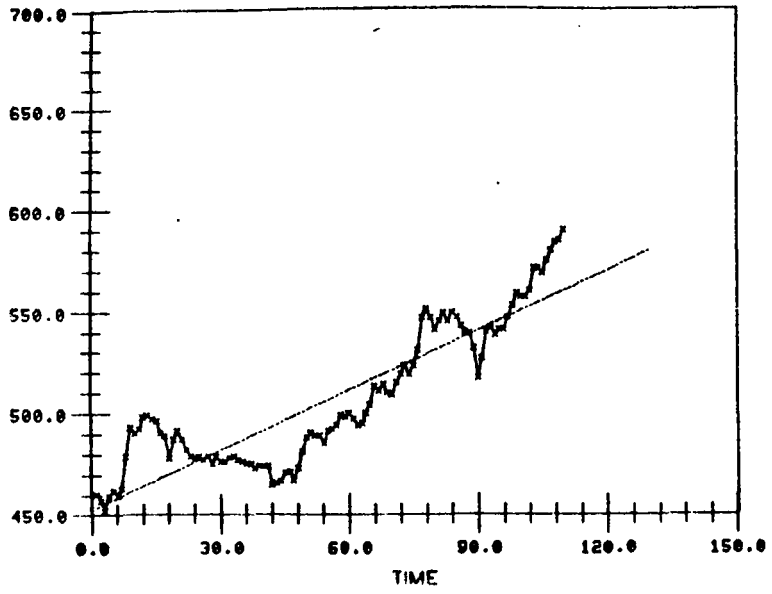


Figure 2

IBM STOCK PRICE -- QUADRATIC FIT

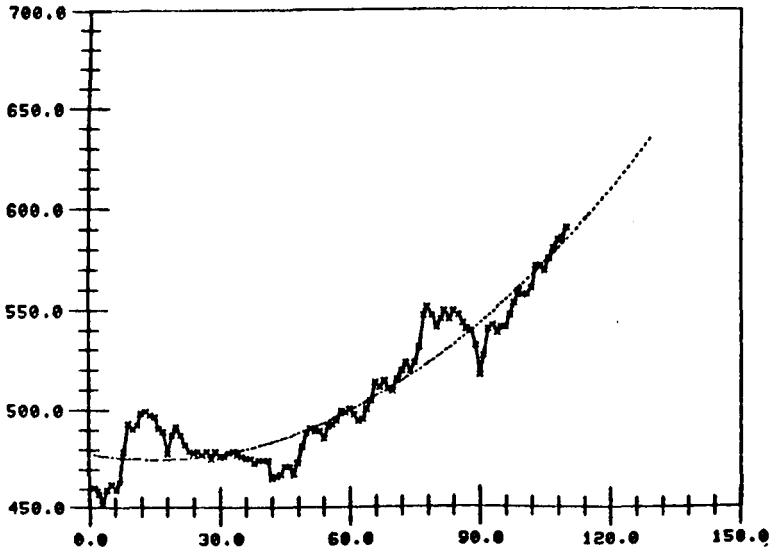


Figure 3.a

IBM STOCK PRICE -- STRAIGHT LINE FORECAST USING ENTIRE SERIES

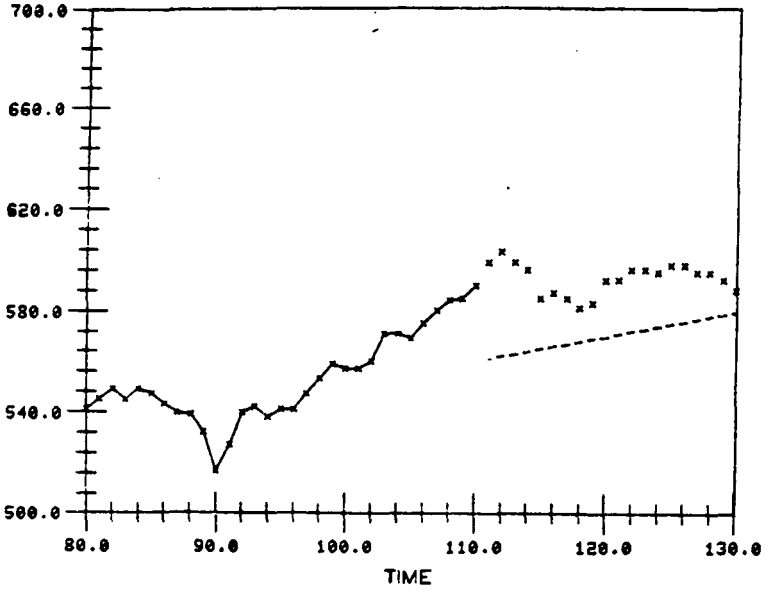


Figure 3.b

IBM STOCK PRICE -- STRAIGHT LINE FORECAST
(USING LAST 20 OBSERVATIONS)

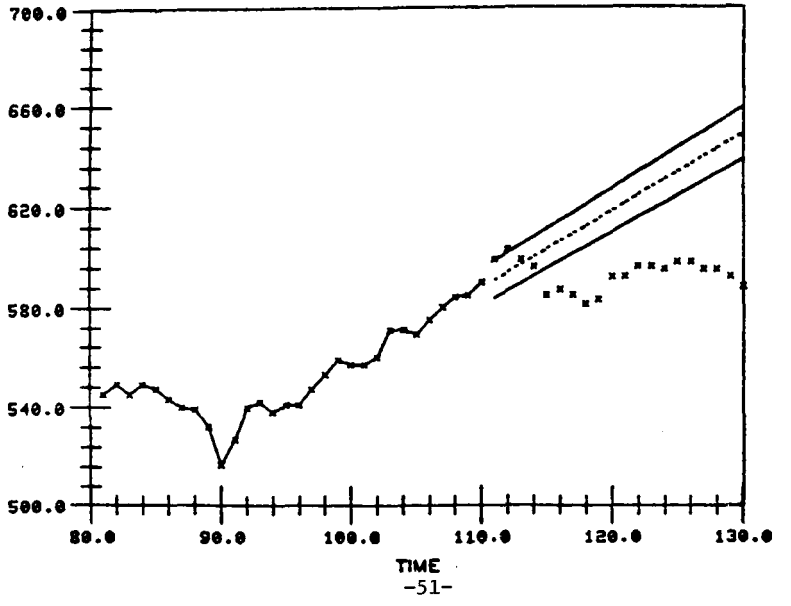


Figure 4

IBM STOCK PRICE +-- STRAIGHT LINE FIT
TO LAST 20 OBSERVATIONS

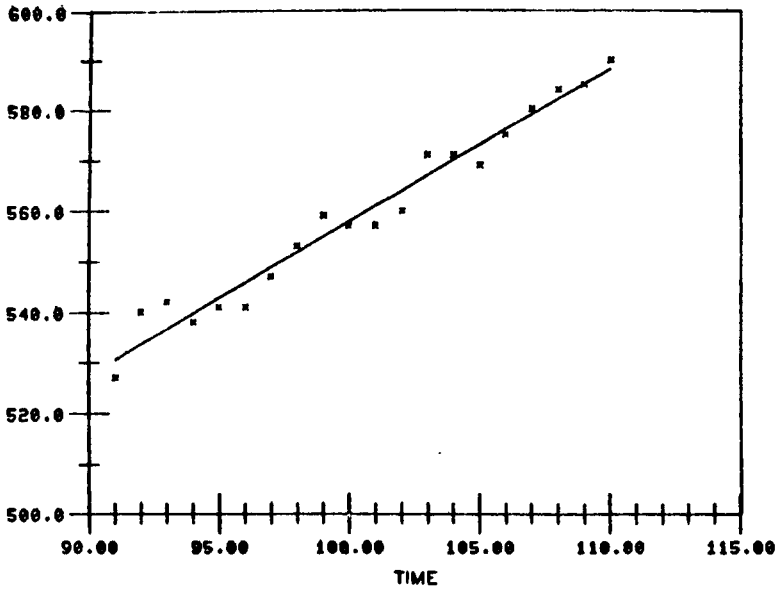


Figure 5
IBM STOCK PRICE -- (0,1,1) FORECAST

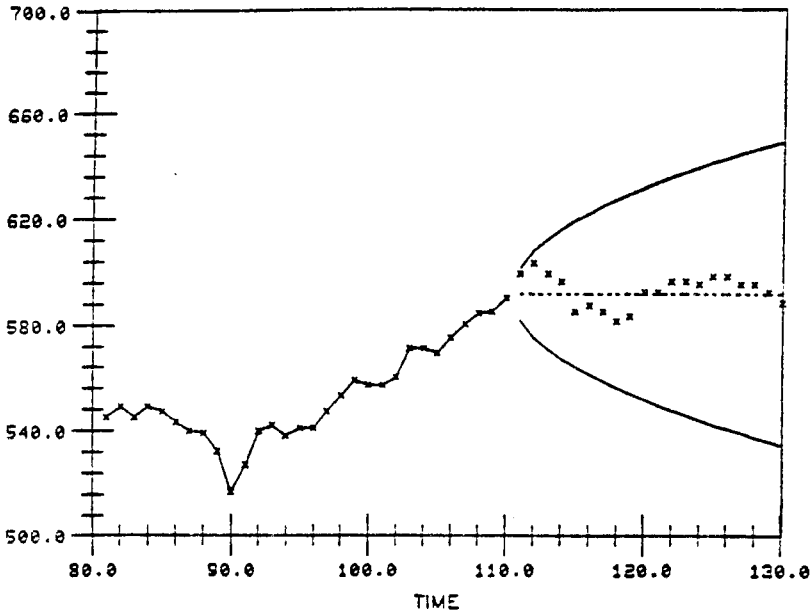


Figure 6
IBM STOCK PRICE -- (0,1,1) WITH TREND FORECAST

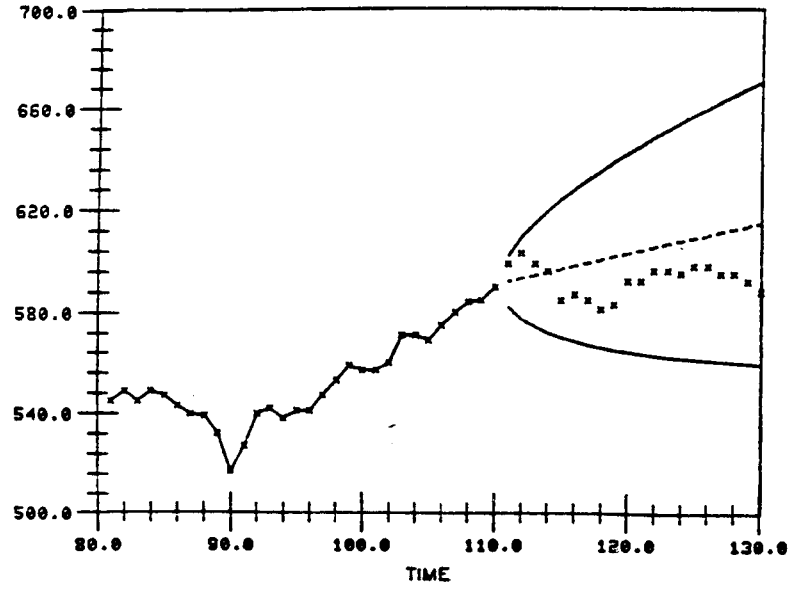


Figure 7

ANNUAL U.S. AGE SPECIFIC BIRTH RATES 1917 - 1980

(PER THOUSAND)

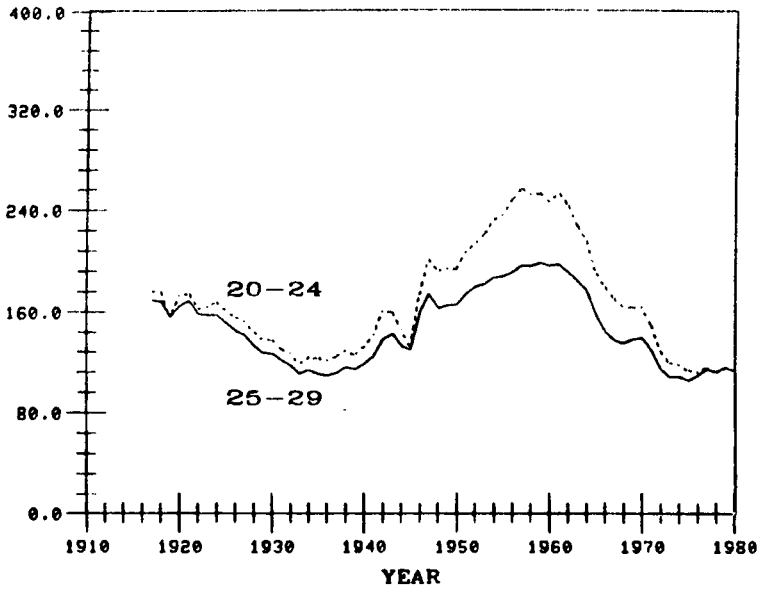


Figure 8
FORECASTING 20-24 BIRTH RATES FROM 1975

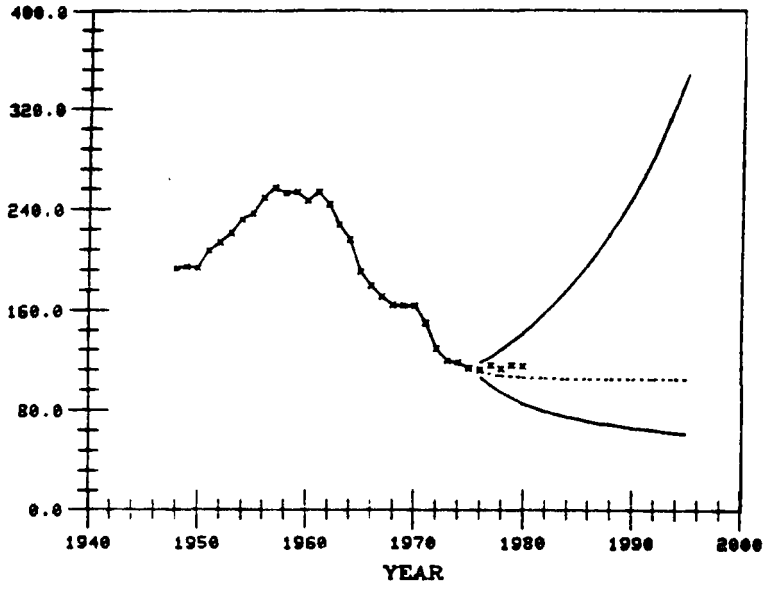


Figure 9
FORECASTING 25-29 BIRTH RATES FROM 1975

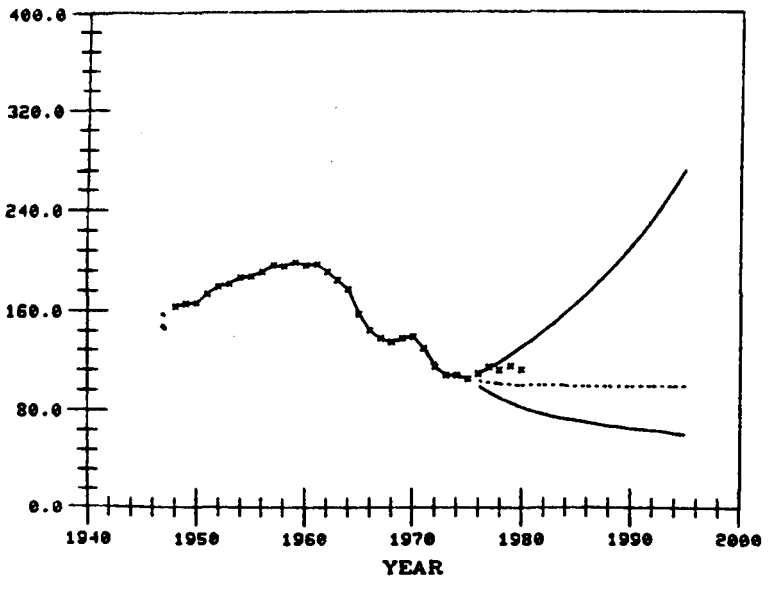


Figure 10

FORECASTING 20-24 BIRTH RATES WITH NO TRANSFORMATION

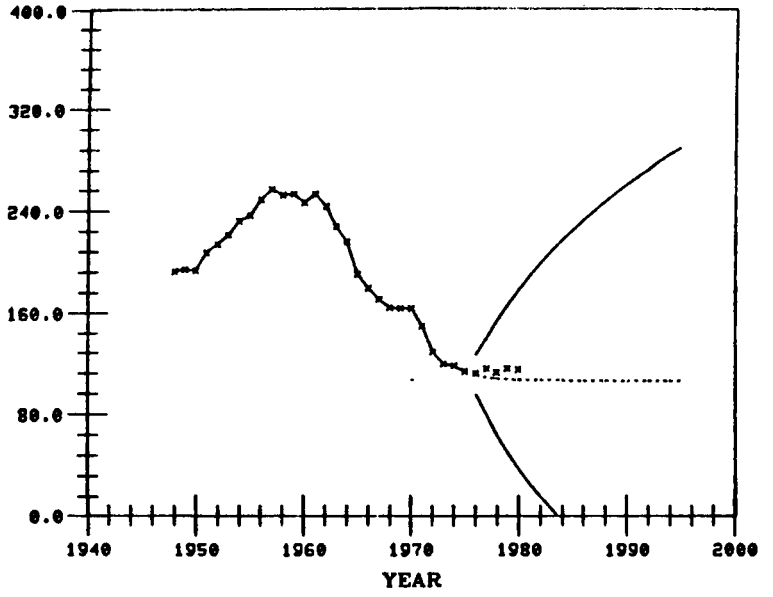


Figure 11

FORECASTING 20-24 BIRTHS (IN 1000'S) FROM 1975

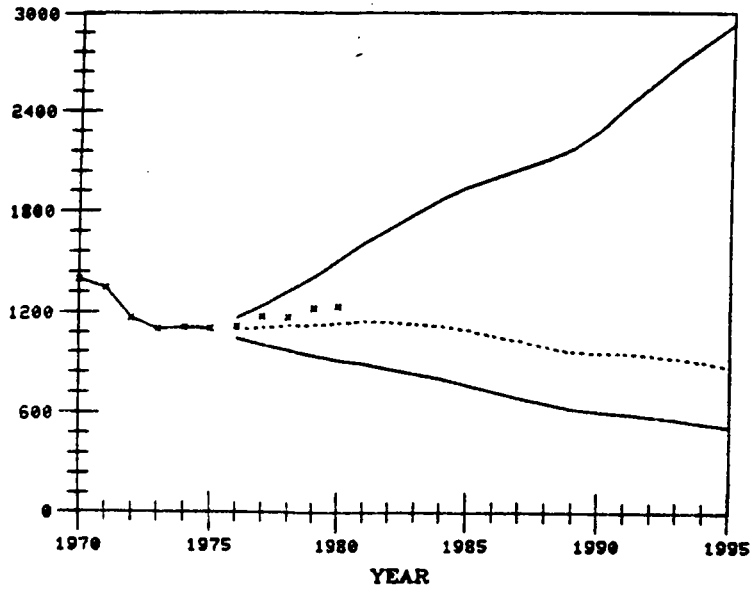
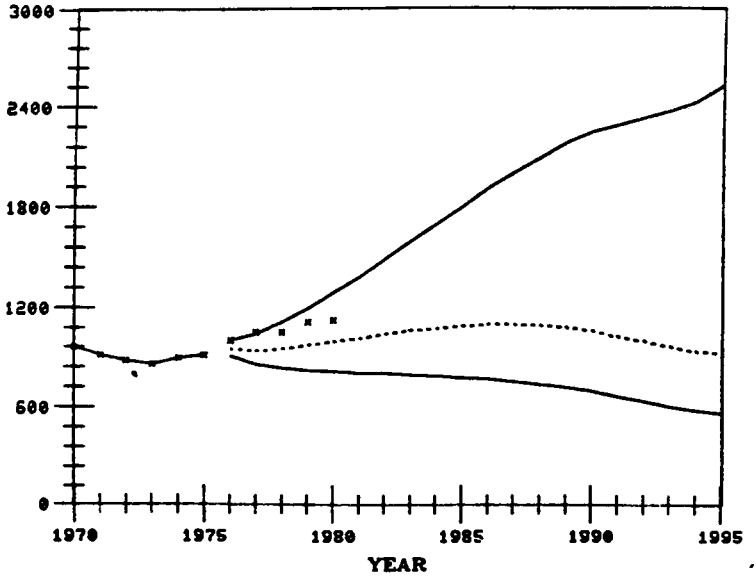


Figure 12

FORECASTING 25-29 BIRTHS (IN 1000'S) FROM 1975



REFERENCES

Ansley, C.F. and Newbold, P. (1981) "On the Bias in Estimation of Forecast Mean Squared Error," Journal of the American Statistical Association, 76, 569-578.

Bell, W.R. and Hillmer, S.C. (1983) "Modeling Time Series With Calendar Variation," Journal of the American Statistical Association, 78, 526-534.

Box, G.E.P. and Abraham, B. (1978) "Deterministic and Forecast - Adaptive Time - Dependent Models," Applied Statistics, 27, 120-130.

Box, G.E.P., and Jenkins, G.M. (1970), Time Series Analysis: Forecasting and Control, San Francisco: Holden Day.

Bureau of the Census (1982a) "Preliminary Estimates of the Population of the United States, by Age, Sex, and Race: 1970 to 1981" Current Population Reports, Series P-25, No. 917, G.P.O., Washington.

_____ (1982b) Unpublished data consistent with middle series of "Projections of the Population of the United States: 1982 to 2050 (Advance Report)," Current Population Reports, Series P-25, No. 922, G.P.O., Washington.

Freedman, D.A. and Peters, S.C. (1982) "Bootstrapping a Regression Equation: Some Empirical Results" Technical Report No. 10, Department of Statistics, University of California, Berkeley.

Fuller, W.A. (1976), Introduction to Statistical Time Series, New York: Wiley.

Granger, C.W.J. and Joyeux, R. (1980) "Introduction to Long - Memory Time Series Models and Fractional Differencing," Journal of Time Series Analysis, 1, 15-30.

Keyfitz, N. (1972) "On Future Population," Journal of the American Statistical Association, 67, 347-363.

_____ (1981) "The Limits of Population Forecasting", Population and Development Review, 7, 579-593.

Kimeldorf, G. and Jones, D. (1967), "Bayesian Graduation", Transactions of the Society of Actuaries, 19, 66-112.

Long, J.F. (1984) "U.S. National Population Projection Methods: A View From Four Forecasting Traditions," included in this volume.

McDonald, J. (1981) "Modeling Demographic Relationships: An Analysis of Forecast Functions for Australian Births," Journal of the American Statistical Association, 76, 782-792.

Miller, M. (1942), Elements of Graduation, New York, Actuarial Society of America.

Miller, R.B. (1984) "Evaluation of Transformations in Forecasting Age Specific Birth Rates," included in this volume.

Miller, R.B. and Hickman, J.C. (1981), "Time Series Modeling of Births and Birth Rates" Working Paper 8-81-21, Graduate School of Business, University of Wisconsin, Madison.

Miller, R.B. and Wichern, D.W. (1977) Intermediate Business Statistics: Analysis of Variance, Regression, and Time Series, New York: Holt, Rinehart and Winston.

Pierce, D.A. (1971), "Least Squares Estimation in the Regression Model With Autoregressive - Moving Average Errors," Biometrika, 58, 299-312.

Stoto, M.A. (1983) "The Accuracy of Population Projections," Journal of the American Statistical Association, 78, 13-20.

Tiao, G.C. and Box, G.E.P. (1981) "Modeling Multiple Time Series With Applications" Journal of the American Statistical Association, 76, 802-816.

Whittaker, E.T. and Robinson, G. (1944), The Calculus of Observations, London: Blackie and Sons.