5 Multivariate ARIMA Models

For many readers, this may be the most interesting chapter of the volume. Whereas we were concerned only with *uni*variate time series analysis in preceding chapters, in this chapter, we will generalize the Box-Jenkins philosophy to *multi*variate time series analysis, that is, to the modeling of relationships between two or more time series.[1] There are many ways to view multivariate time series analysis. In *Design and Analysis of Time Series Experiments,* for example, Glass et al. (1975; Chapter 8) develop multivariate ARIMA modeling under the rubric of "concomitant variation." From this perspective, independent variable time series are introduced only for the purpose of reducing background noise (or unexplained variance) in the dependent-variable time series.

Noise reduction is not an unimportant consideration. Many of the problems addressed (or sidestepped) in preceding chapters can be seen as problems of background noise. Trend, for example, is a bothersome topic which raises philosophical dilemmas of the most complex nature. In time series analysis, trend must be equated with change. If a social process changes systematically throughout a finite realization, however, can it be assumed that the process will continue to change in the same systematic manner? The exogenous forces which underlie trend may be relatively constant during a finite period of time, so during that period, the constant term of an ARIMA model may adequately represent these forces. As the time frame grows larger, however, these forces may change subtly and gradually; their representation by a constant term may weaken.

Seasonality presents a similar dilemma. A seasonal ARIMA model mimics the effects of excluded (and often unknown) periodic exogenous forces. While the model performs remarkably well in this role, the analyst must always remember that the *essence* of seasonality has not really been captured in the model. In relatively short time frames, seasonality may be adequately explained as structured, periodic noise. Over longer time frames, however, exogenous seasonal forces may change gradually and the imitative power of the model may wane.

Finally, viewing outliers as background noise, the same dilemma arises. The input to a univariate ARIMA model is white noise and, in theory, a white noise process can generate an infinitely large random shock. A one-in-a-million random shock nevertheless complicates the analysis unless the time series being modeled is a million observations long.

These three specific problems of background noise can be mitigated by incorporating an independent-variable time series into the ARIMA model. If the same set of exogenous forces (which are responsible for trend, seasonality, and outliers) underlie two time series, then a bivariate model of the relationship may incorporate these forces indirectly. As multivariate ARIMA models "solve" all of these dilemmas, we will state unequivocally that a "fair" multivariate model is always preferred to a "good" univariate model. The analyst must remember nonetheless that a "good" or even "excellent" multivariate model gives only an approximate representation of the excluded exogenous forces.

From a concomitant variation perspective (as outlined by Glass et al.), trend, seasonality, and outliers are seen as background noise which can be reduced by incorporating an independent-variable time series into the model. The concomitant variation perspective misses the most important facet of multivariate ARIMA models, however. Multivariate ARIMA models are inherently causal. Although we acknowledge the importance of noise reduction, we will develop multivariate ARIMA modeling from a *causal* modeling perspective in this chapter.

The jump from univariate to multivariate time series analysis will not be difficult. The impact assessment models developed in Chapter 3, in fact, are multivariate models with the step function I_t as an independent variable. For a set of n independent variables, $X_{1t}, \ldots, X_{nt}$, the general multivariate ARIMA model may be written as

$$Y_t = f(X_{1t}, \ldots, X_{nt}) + N_t.$$

The functions of the several independent variables in this model are transfer functions such as those described in Chapter 3. The difference here is that these transfer functions will be identified *empirically*.

Our discussion begins with the cross-correlation function which may be used to identify a transfer function relationship between two time series. We will then illustrate the multivariate model-building strategy with one forecasting and one causal modeling example. Needless to say, our development of this material leans heavily on the principles developed in Chapters 2 and 3. The reader who is unsure of this material is thus advised to review those chapters before proceeding.

5.1 The Cross-Correlation Function

It is sometimes useful to think of autocorrelation as *within-series* correlation. In the same way that the ACF is used to identify within-series correlation, the cross-correlation function (CCF) is used to identify *between-series* correlation. Patterns of between-series correlation are used to identify a transfer function relationship *between* two time series in much the same way that the ACF is used to identify an ARIMA relationship *within* the time series.

As a first principle, we note that two nonstationary time series will always be correlated due to common patterns of drift or trend. *This correlation must always be regarded as spurious.* To eliminate between-series correlations due only to drift or trend, the time series are made stationary prior to estimation of the CCF. After an appropriate differencing,

$$x_t = (1 - B)^d(1 - B^S)^D X_t$$
$$z_t = (1 - B)^d(1 - B^S)^D Y_t,$$

the CCF may be estimated. By convention, x_t is referred to as the *input* series, or causor, and z_t is referred to as the *output* series, or effector. This terminology reflects the input-output relationship

which is explicitly causal. Given two stationary time series, the CCF for lags $\pm k$ is given by the formulae

$$\text{CCF}(+k) = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(z_{t+k} - \bar{z})}{\sqrt{\sum_{t=1}^{N} (x_t - \bar{x})^2 \sum_{t=1}^{N} (z_{t+k} - \bar{z})^2}}$$

$$\text{CCF}(-k) = \frac{\sum_{t=1}^{N+k} (x_{t-k} - \bar{x})(z_t - \bar{z})}{\sqrt{\sum_{t=1}^{N} (x_{t-k} - \bar{x})^2 \sum_{t=1}^{N} (z_t - \bar{z})^2}}$$

These formulae give the familiar Pearson product-moment correlation coefficient (approximately) between two time series separated by $\pm k$ observations. When $k = 0$, the formulae are identical. When $k \neq 0$, the first formula gives the *positive* half of the CCF by lagging the z_t series forward in time. The second formula gives the *negative* half of the CCF by lagging the x_t series forward in time.

A major difference between the CCF and the ACF (as noted in Section 2.8) is that the CCF need not be symmetrical about lag-zero. In other words, CCF $(+k) \neq$ CCF$(-k)$ generally. When the ACF is used to identify an ARIMA model, only one half of the ACF need be examined. ACF$(-k)$ is a mirror image of ACF$(+k)$ but this is not true of the CCF.

We are always reluctant to introduce tedious arithmetic into our argument. The relationship between CCF$(+k)$ and CCF$(-k)$ is one that cannot ordinarily be grasped without a basic demonstration, however. Apologies given, we present 10 pairs of numbers:

x_t	t	z_t
.665	1	−.160
−1.630	2	−.058
− .298	3	.333
.225	4	−.815
1.222	5	−.149
.531	6	.113
− .957	7	.611
.676	8	.266
− .723	9	−.479
.289	10	.338

These numbers were generated so that $\bar{x} = \bar{z} = 0$; the first nine values of x_t are random Normal numbers and the 10th was selected to ensure that $\bar{x} = 0$. The values of z_t were generated as

$$z_t = \frac{x_{t-2}}{2}.$$

For 8 of these 10 pairs of numbers, then, there is a perfect causal relationship:

$$x_t \dashrightarrow z_{t+2}$$

Applying the formulae for CCF$(\pm k)$ to these 10 pairs of numbers, starting with $k = -3$,

$$\text{CCF}(-3) = \frac{(-.160)(.225) + \ldots + (.611)(.289)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = .25$$

$$\text{CCF}(-2) = \frac{(-.160)(-.298) + \ldots + (.266)(.289)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = -.04$$

$$\text{CCF}(-1) = \frac{(-.160)(-1.630) + \ldots + (-.479)(.289)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = -.22$$

$$\text{CCF}(0) = \frac{(-.160)(.665) + \ldots + (.338)(.289)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = -.11$$

$$\text{CCF}(+1) = \frac{(-.058)(.665) + \ldots + (.338)(-.723)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = -.22$$

$$\text{CCF}(+2) = \frac{(.333)(.665) + \ldots + (.338)(.676)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = .95$$

$$\text{CCF}(+3) = \frac{(-.815)(.665) + \ldots + (.338)(-.957)}{\sqrt{[(-.160)^2 + \ldots + (.338)^2][(.665)^2 + \ldots + (.289)^2]}} = -.13.$$

CCF(+2) is the largest of these seven numbers. The value of CCF(+2) ≈ 1 indicates the perfect causal relationship built into these numbers. There are many other "large" correlations among the seven, however. The statistical significance of any CCF(±k) estimate can be assessed by comparing it with its standard error. One unit of standard error for CCF(±k) is given by

$$\text{SE}\,[\text{CCF}\,(\pm k)] = \frac{1}{\sqrt{N-k}}.$$

For these seven estimates, then, the standard errors are:

SE [CCF (±3)]	$= \sqrt{1/7}$	$= .378$
SE [CCF (±2)]	$= \sqrt{1/8}$	$= .354$
SE [CCF (±1)]	$= \sqrt{1/9}$	$= .333$
SE [CCF (0)]	$= \sqrt{1/10}$	$= .316.$

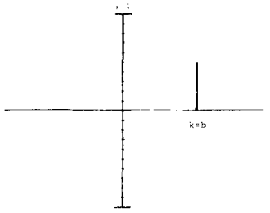
Dividing each of these estimated CCF(±k) by its standard error, the standardized estimates are:

CCF (−3)	$= .25/.378$	$= .66$ SE
CCF (−2)	$= -.04/.354$	$= -.11$ SE
CCF (−1)	$= -.22/.333$	$= -.66$ SE
CCF (0)	$= -.11/.316$	$= -.35$ SE
CCF (+1)	$= -.22/.333$	$= .66$ SE
CCF (+2)	$= .95/.354$	$= 2.68$ SE
CCF (+3)	$= -.13/.378$	$= -.34$ SE.

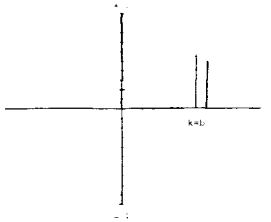
As a convention, the analyst may assume that any estimate of CCF(±k) smaller in absolute value than 2 SE is zero. By this rule, only the estimate of CCF(+2) is statistically different than zero.

Among other things, this exercise illustrates the interpretation of asymmetry in the CCF. *The CCF measures not only the strength of a relationship but also the direction.* When "x_t causes z_{t+b}," evidence of the relationship is found at CCF(+b), in the positive half of the CCF, that is. When "z_t causes x_{t+b}," on the other hand, evidence of the relationship is found at CCF(−b), in the negative half of the CCF. Asymmetry in the estimated ACF is thus interpreted on the basis the causal relationship specified a priori. When one

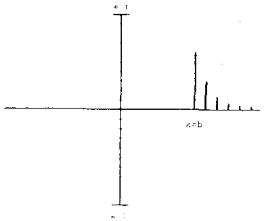
$z_t = \omega_0 x_{t-b}$ $\omega_0 > 0$



$z_t = (\omega_0 + \omega_1 B)x_{t-b}$ $\omega_0, \omega_1 > 0$



$z_t = (1 - \delta_1 B)^{-1}\omega_0 x_{t-b}$ $\delta_1, \omega_0 > 0$



$z_t = (1 - \delta_1 B)^{-1}(\omega_0 + \omega_1 B)x_{t-b}$ $\delta_1, \omega_0, \omega_1 > 0$

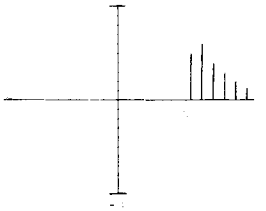


FIGURE 5.1 Expected CCFs for Several Bivariate Relationships

of the two time series has been specified as the causor, evidence of the relationship is expected in the positive CCF. But more important, the CCF(+b) estimated under the assumption that x_t is the causor will be identical to the CCF(−b) estimated under the assumption that x_t is the effector.

What has been demonstrated in the specific case for 10 pairs of numbers must now be demonstrated in the general case. Figure 5.1 shows the *expected* CCFs for several common transfer function relationships. First, the zero-order relationship

$$z_t = \omega_0 x_{t-b} + N_t$$

is expected to have a nonzero value of CCF (+b). All other lags of the CCF are expected to be zero. To demonstrate this, we define the expected CCF (+k) as

$$\text{CCF}(+k) = \frac{\text{COV}(x_{t-k}z_t)}{\sqrt{\text{VAR}(x_t)\,\text{VAR}(z_t)}}.$$

Then,

$$\begin{aligned}\text{COV}(x_{t-k}z_t) &= \text{E}[(x_{t-k})(\omega_0 x_{t-b} + N_t)] \\ &= \text{E}(\omega_0 x_{t-k}x_{t-b} + x_{t-k}N_t).\end{aligned}$$

Now in all cases, $Ex_{t-k}N_t = 0$, so

$$\text{COV}(x_{t-k}z_t) = \omega_0 Ex_{t-k}x_{t-b}.$$

Then assuming that x_t is a white noise process (we will cover the case in which x_t is not white noise in Section 5.3),

$$\begin{aligned}\text{COV}(x_{t-k}z_t) &= \omega_0\sigma_x^2 \text{ whenever } b = k \\ &= 0 \text{ otherwise.}\end{aligned}$$

Dividing this term by $\sigma_x\sigma_z$, the expected CCF is

$$\begin{aligned}\text{CCF}(k) &= \omega_0\frac{\sigma_x}{\sigma_z} \text{ whenever } b = k \\ &= 0 \text{ otherwise.}\end{aligned}$$

Following this same procedure, the reader may demonstrate that the converse relationship

$$x_t = \omega_0 z_{t-b} + N_t$$

is expected to have a nonzero value of CCF (−b). All other lags of the CCF are expected to be zero.

The first-order transfer function relationship

$$z_t = \frac{\omega_0}{1 - \delta_1 B}x_{t-b} + N_t$$

describes a *dynamic* causal relationship between the two time series. Rewriting the relationship as the infinite series,

$$z_t = \omega_0 \sum_{i=0}^{\infty} \delta_1^i x_{t-b-i} + N_t.$$

The covariance between x_{t-k} and z_t is:

$$\begin{aligned}\text{COV}(x_{t-k}z_t) &= \text{E}[(x_{t-k})(\omega_0 \sum_{i=0}^{\infty} \delta_1^i x_{t-b-i} + N_t)] \\ &= \text{E}(\omega_0 \sum_{i=0}^{\infty} \delta_1^i x_{t-k}x_{t-b-i}) \\ &= \omega_0 Ex_{t-k}x_{t-b} + \omega_0\delta_1 Ex_{t-k}x_{t-b-1} \\ &\quad + \omega_0\delta_1^2 Ex_{t-k}x_{t-b-2} + \ldots + \omega_0\delta_1^n Ex_{t-k}x_{t-b-n} + \ldots.\end{aligned}$$

When $b < k$, all terms of this expression are zero. When $b = k$, however, the first term of the infinite series is nonzero, so

$$\text{COV}(x_{t-b}z_t) = \omega_0\sigma_x^2.$$

When $k = b + 1$, the second term of the infinite series is nonzero:

$$\text{COV}(x_{t-b-1}z_t) = \omega_0\delta_1\sigma_x^2.$$

And when $k = b + n$, the $n + 1$st term of the infinite series is nonzero:

$$\text{COV}(x_{t-b-n}z_t) = \omega_0\delta_1^n\sigma_x^2.$$

Dividing these covariances by $\sigma_x\sigma_z$, the expected CCF is:

$$\text{CCF}(k) = 0 \text{ for } k < b$$

$$\text{CCF (b)} = \omega_0 \frac{\sigma_x}{\sigma_z}$$

$$\text{CCF (b+n)} = \omega_0 \delta_1^n \frac{\sigma_x}{\sigma_z}.$$

So the CCF for a dynamic first-order transfer function relationship is expected to be zero until CCF (b). Successive positive lags, CCF (b+1), CCF (b+2), . . . , CCF (b+n), decay exponentially back to zero. The expected CCF is thus identical with the ACF expected of an ARIMA (1,0,0) process.

The CCFs shown in Figure 5.1 all suggest a causal relationship between x_t and z_{t+b}.[2] A single spike in the positive CCF is interpreted as an ω parameter. Decay from a spike is interpreted as a δ parameter. These are *expected* CCFs, of course, and, in practice, identification of a transfer function relationship may be complicated by ambiguous identification statistics. Nevertheless, this first step in multivariate model building must produce some evidence of relationship before the next step in the procedure can begin. We will now demonstrate the multivariate model-building procedure with an example.

5.2 A Forecasting Example

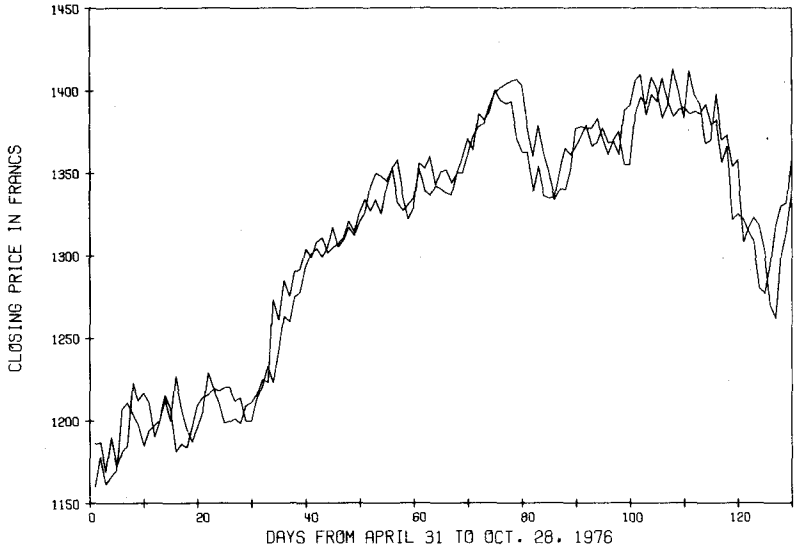


FIGURE 5.2(a) Paris and New York IBM Common Stock Prices

In Chapter 4, we noted that univariate forecasts often tend to be trivial. This is not true of multivariate forecasts. The two time series shown in Figure 5.2(a) are daily closing prices (in francs) of IBM common stock on the New York and Paris exchanges (see, Makridakis and Wheelwright, 1978: 487–488). Both series drift upward during the 130 days presented here. While one might conclude from this common pattern of drift that there is a causal relationship between these two series, it would be unwise to leap to this conclusion. Many stock price time series (perhaps even the majority) follow similar patterns of drift during this same period. Common patterns of drift or trend by themselves say nothing about the causal relationship among time series.

As a first step in building a bivariate ARIMA model, univariate models are built for both series. Univariate analysis shows that both time series are well represented by ARIMA (0,1,0) models. For the New York series,

$$(1 - B)X_t = a_t$$

```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS     =    129.

INPUT SERIES..  NYIBM     NEW YORK IBM STOCK PRICE (DIFFERENCED)
OUTPUT SERIES.. PARISIBM  PARIS IBM STOCK PRICE (DIFFERENCED)

 LAG   CORR    SE   -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                    +----+----+----+----+----+----+----+----+----+----+----+
 -15  -.013  .094                      (    I    )
 -14   .002  .093                      (    I    )
 -13  -.066  .093                      (  XXI    )
 -12   .130  .092                      (    IXXX )
 -11  -.034  .092                      (   XI    )
 -10   .007  .092                      (    I    )
  -9  -.067  .091                      (  XXI    )
  -8  -.114  .091                      ( XXXI    )
  -7   .070  .091                      (    IXX  )
  -6  -.153  .090                      (XXXXI    )
  -5   .036  .090                       (   IX   )
  -4  -.004  .089                       (   I    )
  -3   .053  .089                       (   IX   )
  -2  -.017  .089                       (   I    )
  -1  -.040  .088                       (  XI    )
   0   .151  .088                       (   IXXX*
   1   .064  .088                       (   IXX )
   2   .659  .089                       (   IXXX)XXXXXXXXXXXX
   3  -.004  .089                       (   I    )
   4   .041  .089                       (   IX   )
   5  -.041  .090                       (  XI    )
   6   .043  .090                      (    IX    )
   7   .097  .091                      (    IXX   )
   8  -.087  .091                      (  XXI     )
   9  -.025  .091                      (   XI     )
  10   .068  .092                      (    IXX   )
  11  -.110  .092                      ( XXXI     )
  12  -.056  .092                      (   XI     )
  13  -.129  .093                      ( XXXI     )
  14   .151  .093                      (    IXXXX)
  15  -.055  .094                      (   XI     )
                                     -2SE      +2SE
```

FIGURE 5.2(b) *Identification:* CCF Estimated from the Differenced Series

```
SERIES.. RESIDUAL (NOBS= 127)  ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =   127.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 30, 127) =  41.847      SIG =      .074

 LAG    CORR     SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                     +----+----+----+----+----+----+----+----+----+----+----+
   1  -.310  .090                     XXX(XXXXI     )
   2  -.140  .100                        (XXXXI     )
   3  -.080  .100                        (  XXI     )
   4   .160  .100                        (     IXXXX)
   5  -.120  .100                        ( XXXI     )
   6   .000  .100                        (     I    )
   7   .040  .100                        (     IX   )
   8   .090  .100                        (     IXX  )
   9  -.100  .100                        ( XXXI     )
  10  -.060  .100                        (  XXI     )
  11   .080  .100                        (     IXX  )
  12   .000  .100                        (     I    )
  13  -.070  .100                        (  XXI     )
  14   .070  .100                        (     IXX  )
  15   .100  .100                        (     IXXX )
  16  -.230  .110                       *XXXXXI      )
  17   .040  .110                       (     IX     )
  18   .040  .110                       (     IX     )
  19   .070  .110                       (     IXX    )
  20  -.160  .110                       ( XXXXI      )
  21   .130  .110                       (     IXXX   )
  22   .030  .110                       (     IX     )
  23  -.060  .110                       (   XXI      )
  24  -.020  .110                       (    XI      )
  25  -.020  .110                       (    XI      )
  26  -.030  .110                       (    XI      )
  27   .040  .110                       (     IX     )
  28   .010  .110                       (     I      )
  29   .070  .110                       (     IXX    )
  30   .040  .110                       (     IX     )
                                      -2SE         +2SE
```

FIGURE 5.2(c) *Identification:* ACF for the Residuals of the Model

$$Y_t = .767\,X_{t-2} + \frac{a_t}{1 - B}$$

and for the Paris series,

$$(1 + B)Y_t = a_t.$$

The major purpose of a preliminary univariate analysis is to make sure that both time series are stationary. As noted, a CCF estimated when one or both series are nonstationary will be overwhelmed by spurious correlations. Bivariate identification will thus require that both of these time series be differenced.

In this case, we have no a priori theory about the relationship between these series. The possibilities include the case in which a change in the New York series causes a change in the Paris series,

$$X_t \text{----------} \rightarrow Y_{t+b},$$


```
SERIES.. RESIDUAL (NOBS= 127)  ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =   127.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 29, 127) =  25.997      SIG =      .675

 LAG    CORR     SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                     +----+----+----+----+----+----+----+----+----+----+----+
   1   .120  .090                        (    IXXX )
   2  -.070  .090                        (  XXI    )
   3  -.110  .090                        ( XXXI    )
   4   .040  .090                        (    IX   )
   5  -.060  .090                        (  XXI    )
   6   .010  .090                        (    I    )
   7   .050  .090                        (    IX   )
   8   .070  .090                        (    IXX  )
   9  -.090  .090                        (  XXI    )
  10  -.100  .090                        ( XXXI    )
  11   .030  .090                        (    IX   )
  12  -.040  .090                        (   XI    )
  13  -.020  .090                        (   XI    )
  14   .070  .090                        (    IXX  )
  15   .020  .090                        (    IX   )
  16  -.170  .090                        (XXXXI    )
  17   .010  .100                        (    I    )
  18   .040  .100                        (    IX   )
  19   .130  .100                        (    IXXX )
  20  -.030  .100                        (   XI    )
  21   .200  .100                        (    IXXXX*
  22   .090  .100                        (    IXX  )
  23  -.030  .100                        (   XI    )
  24  -.090  .100                        (  XXI    )
  25  -.090  .100                        (  XXI    )
  26  -.080  .100                        (  XXI    )
  27   .050  .100                        (    IX   )
  28   .060  .100                        (    IXX  )
  29   .090  .100                        (    IXX  )
  30  -.010  .110                      (     I     )
                                      -2SE         +2SE
```

FIGURE 5.2(d) *Diagnosis:* ACF for the Residuals of the Model

$$Y_t = .987\,X_{t-2} + \frac{1 - .88B}{1 - B}\,a_t$$

the case in which a change in the Paris series causes a change in the New York series,

$$Y_t \text{----------} \rightarrow X_{t+b},$$

and the noncausal case in which, perhaps as a result of some underlying common variable, the two series appear to be causing each other. We will operate under the assumption that the New York series is the causor but this assumption is arbitrary.

Figure 5.2(b) shows the CCF estimated from the differenced time series. The lone spike at CCF (+2) suggests that the New York series leads the Paris series by exactly two days. Had the spike instead appeared at CCF (−2), the opposite inference would have been supported. Although the assumption that the New York series was the causor was arbitrary, it is supported

```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS     =    129.

INPUT SERIES..  NYIBM     PREWHITENED NEW YORK IBM STOCK PRICE
OUTPUT SERIES.. RESIDUAL  ESTIMATED MODEL RESIDUALS

 LAG   CORR    SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                  +----+----+----+----+----+----+----+----+----+----+----+
 -15  -.070  .094                        (  XXI    )
 -14  -.067  .093                        (  XXI    )
 -13  -.062  .093                        (  XXI    )
 -12  -.053  .092                        (  XI     )
 -11  -.041  .092                        (  XI     )
 -10  -.032  .092                        (  XI     )
  -9  -.024  .091                        (  XI     )
  -8  -.017  .091                        (   I     )
  -7  -.013  .091                        (   I     )
  -6  -.035  .090                        (  XI     )
  -5  -.012  .090                         (   I    )
  -4  -.012  .089                         (   I    )
  -3  -.030  .089                         (  XI    )
  -2  -.022  .089                         (  XI    )
  -1  -.014  .088                         (   I    )
   0  -.012  .088                         (   I    )
   1   .003  .088                         (   I    )
   2  -.046  .089                         (  XI    )
   3  -.021  .089                         (  XI    )
   4  -.036  .089                         (  XI    )
   5  -.005  .090                         (   I    )
   6  -.021  .090                        (  XI     )
   7  -.027  .091                        (  XI     )
   8  -.031  .091                        (  XI     )
   9  -.030  .091                        (  XI     )
  10  -.025  .092                        (  XI     )
  11  -.024  .092                        (  XI     )
  12  -.028  .092                        (  XI     )
  13  -.033  .093                        (  XI     )
  14  -.032  .093                        (  XI     )
  15  -.048  .094                        (  XI     )
                                       -2SE     +2SE
```

FIGURE 5.2(e) *Diagnosis:* CCF for the Differenced X_t Series and the Model Residuals

empirically by this CCF. In general, the analyst need not specify which series is the causor, but instead may make the specification empirically. If the positive CCF is statistically significant, then the X_t series is the causor; and if the negative CCF is statistically significant, then the Y_t series is the causor.

To be sure, there are many nonzero values in both the positive and negative halves of the CCF. The only statistically significant value is at CCF (+2), however, and this suggests the model

$$(1 - B)Y_t = (1 - B)\omega_0 X_{t-2} + N_t.$$

A change in the price of IBM common stock on the New York exchange is followed by an analogous change in price on the Paris exchange two days later. Finally, the CCF suggests that the effect is not distributed over successive days, that is, there is no dynamic transfer of effect which would be indicated by a pattern of decay from CCF (+2) to CCF (+3) to CCF (+4) and so forth.

The next step in the bivariate model-building procedure is to identify an ARIMA noise model for the N_t component. There are a number of ways in which this identification can be made. In our experience, however, the most satisfactory way is the straightforward one: Identify the N_t component from the transfer function residuals. To do this, the analyst first assumes that N_t is white noise. The tentative bivariate model is thus

$$(1 - B)Y_t = (1 - B)\omega_0 X_{t-2} + a_t$$

$$Y_t = \omega_0 X_{t-2} + \frac{a_t}{1 - B}.$$

This tentative model has only one parameter, whose value is estimated as

$$\hat{\omega}_0 = .7674 \text{ with t statistic} = 13.16.$$

The residual ACF for this model, shown in Figure 5.2(c), suggests that an ARIMA (0,0,1) model will adequately reflect the structure of autocorrelation in these residuals. This leads to the tentative model

$$(1 - B)\, Y_t = (1 - B)\, \omega_0 X_{t-2} + (1 - \Theta_1 B)\, a_t$$

$$Y_t = \omega_0 X_{t-2} + \frac{1 - \Theta_1 B}{1 - B}\, a_t.$$

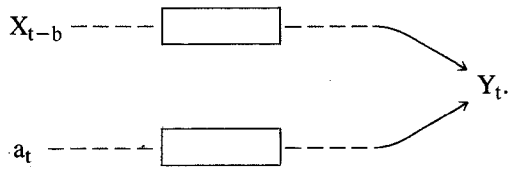
Parameter estimates for this model are:

$$\hat{\omega}_0 = .987 \text{ with t statistic} = 32.12$$

$$\hat{\Theta}_1 = .88 \text{ with t statistic} = 20.17.$$

Both estimates are statistically significant and otherwise acceptable.

Bivariate ARIMA models of the sort we have tentatively selected for these two time series may be diagramed as



The sense of this diagram is that two *distinct* input processes (the time series observation, X_{t-b}, and the random shock, a_t) pass through two *distinct* filters (a transfer function and an ARIMA structure) and are then combined additively into an output process (the time series observation, Y_t). The difference between this diagram and the input-output diagrams we drew for univariate processes in Chapter 2 is that there are two inputs here and this hints at the special problems of diagnosing a bivariate ARIMA model. In effect, the statistical adequacy of both the transfer function component *and* the noise component must be diagnosed.

First, as one might suspect, the model residuals must not be different than white noise. The residual ACF, shown in Figure 5.2(d), indicates that the noise component of this tentative model is adequate. If model residuals are different than white noise, a new noise component must be identified.

Second, the model residuals must be independent of the causor time series. To test the hypothesis of independence, a CCF is estimated from the model residuals and the input time series (the differenced New York IBM series in this case). This CCF, shown in Figure 5.2(e), has no significant values, indicating that the causor series and the model residuals are uncorrelated. As the hypothesis of independence stands, the tentative model is acceptable. Had this CCF indicated that the input series and the model residuals were not independent, a new transfer function component would have to be identified.

Because this model satisfies both diagnostic criteria, it is accepted. There nevertheless may be other acceptable bivariate models which through meta-diagnosis, could be compared with this one. The reader is invited to explore this possibility as an exercise.

Before commenting on the relative utility of this analysis, we must point out that our only purpose was to illustrate the procedures of bivariate modeling. In particular, we do not intend to endorse international stock speculations. While we have no experience here, we have been advised (by an investment analyst who wishes to remain anonymous) that it is not an easy matter to turn a profit on foreign stock exchanges, especially on blue chip shares. The hidden costs of buying and selling stock (not to mention buying francs with dollars) make speculation a generally risky enterprise.

Still, a foreign investor will realize the incremental utility of a bivariate forecasting model over a univariate forecasting model. The Paris IBM series is well represented by an ARIMA (0,1,0) model, so forecasts of the series are:

$$Y_t(n) = Y_t.$$

Because this series follows a random walk, the best univariate forecast of a future price is the current price. Using the New York series as a lead indicator, however, a bivariate forecast is possible. To compare the two models, we use the first 125 observations to forecast the 126th, the first 126 observations to forecast the 127th, and so forth. The results of this exercise are:

Day	Observation	Univariate Forecast	Bivariate Forecast
126	1270.000	1302.000	1277.735
127	1262.000	1270.000	1273.791
128	1300.000	1262.000	1288.419
129	1315.000	1300.000	1313.757
130	1337.000	1315.000	1326.185

Bivariate forecasts are clearly superior to the univariate forecasts. The MSFEs for these two models based on only these five forecasts are 648.20 and 87.65. This should not be surprising. One would also expect a bivariate model to have a lower residual variance than a univariate model. In this case, the univariate model has an RMS = 212.89 while the bivariate model has an RMS = 81.30.

In closing, a comment on the interpretation of the bivariate model is called for. The interpretation is explicitly causal. A rise or drop in the New York series is followed by an analogous rise or drop in the Paris series two days later. While a time-lagged correlation does not imply causation, causation does imply a time-lagged correlation. The null hypothesis is thus

$$H_0: X_t \dashrightarrow Y_{t+2}; \hat{\omega}_0 = 0.$$

Due to the relatively large t statistic for the estimate of ω_0, the null hypothesis must be rejected.

5.3 Prewhitening

The Paris-New York IBM model was identified with relatively little trouble because both series were well represented by ARIMA (0,1,0) models. After differencing, both series were white noise. Our concern now is with the problem of modeling bivariate relationships when the series are not white noise. When we derived the expected CCFs for zero- and first-order transfer functions in Section 5.1, we assumed that the x_t series was white noise. When this assumption is unsatisfied, as is usually the case with social science time series, the estimated CCF is uninterpretable. If the x_t and z_t series are *prewhitened,* however, an interpretable CCF can be estimated.

Figure 5.3(a) shows a time series of annual Swedish population increases (increase per thousand population) for the 1750–1849 century as reported by Thomas (1940). The rate increase in a given year is defined as

$$p_t = \text{birth rate} - \text{death rate in the } t^{th} \text{ year.}$$

The relationship between this time series and a time series of total population is a straightforward one. The total population of Sweden in 1749 was 1,760,000. Starting with this value, the total population (in thousands) at the end of 1750 is given by the expression

$$\begin{aligned} P_{1750} &= P_{1749} + P_{1749}(p_{1750}) \\ &= 1{,}760 + 1{,}760(p_{1750}) \\ &= 1{,}760(1 + p_{1750}). \end{aligned}$$

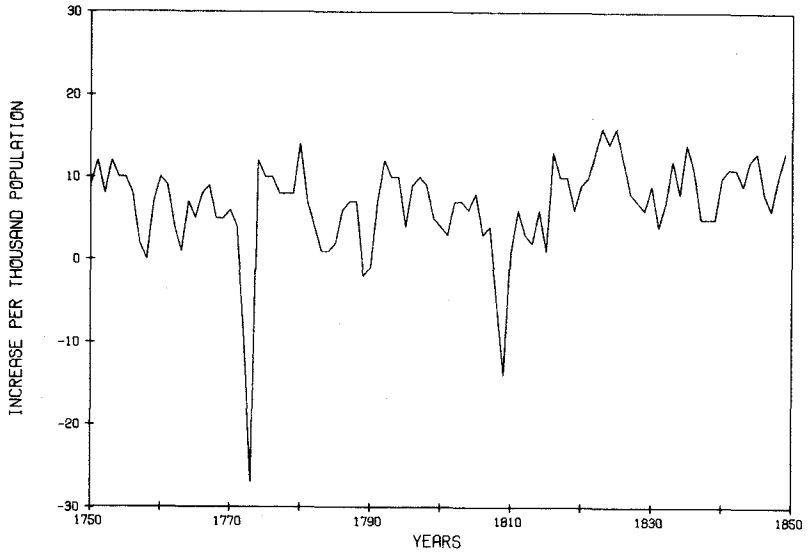


FIGURE 5.3(a) Swedish Population Rates, 1750–1849

The total population at the end of 1750, then, is given by the product of the total population at the end of 1749 and the 1750 rate increase. In the general case, total population at the end of the t^{th} year is:

$$\begin{aligned} P_t &= P_{t-1}(1 + p_t) \\ \frac{P_t}{P_{t-1}} &= 1 + p_t. \end{aligned}$$

Taking the natural logarithm of this expression,

$$\begin{aligned} \text{Ln}\left[\frac{P_t}{P_{t-1}}\right] &= \text{Ln}(1 + p_t) \\ \text{Ln}(P_t) - \text{Ln}(P_{t-1}) &= \text{Ln}(1 + p_t). \end{aligned}$$

The log-transformed total population series is thus the integration (though not necessarily a random walk) of the log-transformed rate increase series.

In Section 2.12.3, we analyzed the Swedish Harvest Index time series for the years 1749–1850. The Harvest Index is a crude measure of food production wherein a value of zero indicated a total crop failure and a value of nine indicated a superabundant crop. Our analysis demonstrated that the series could be well represented by an ARIMA (0,0,1) model:

$$h_t = (1 + .39B)a_t.$$

According to Gustav Sundbärg, an early demographer quoted by Thomas, the growth of Swedish population during the 1750–1849 century could be explained almost entirely as a function of agricultural production:

> Irrespective of which party had gained control, or whether the King himself was on the throne, if the harvest was good, marriage and birth rates were high and death rates comparatively low, that is, the bulk of the population flourished. On the contrary, when the harvest failed, marriage and birth rates declined and death devastated the land, bearing witness to need and privation and at times even to starvation. Whether the factories fared well or badly or whether the bank-rate rose or fell—all these things *at this time,* were scarcely more than ripples on the surface [1940: 82].

This hypothesis (which we will now call Sundbärg's hypothesis) has not been adequately tested. When Dorothy S. Thomas, Gunnar Myrdal, and others investigated the relationships claimed by Sundbärg, the methods we

have described here were unavailable; social scientists did not yet have access to computers. Early analyses of these data consisted largely of visual analyses which tend to mislead. The only satisfactory method for testing a causal hypothesis is to identify, estimate, and diagnose a bivariate time series model as we will do now.

As a first step in the analysis, we make sure that both series are stationary. The Harvest Index series, well represented by an ARIMA (0,0,1) model, is already stationary. An analysis of the population rate increase series leads also to an ARIMA (0,0,1) model:

$$p_t = (1 + .44B)a_t.$$

As both series are stationary, the CCF can be estimated from the undifferenced series.

The estimated CCF, shown in Figure 5.3(b), is not as clear as the CCF estimated in the previous example. There is no evidence of a strong, unambiguous causal relationship. Some analysts might see a two-way relationship here while others might see no relationship whatsoever. The argument is moot because, as we have noted, *the CCF can be interpreted only when the causor variable is a white noise process*. Even if this CCF did indicate a strong, unambiguous causal relationship, the evidence could not be accepted because the Harvest Index series is not white noise.

The problem here is that the CCF estimate is contaminated by within-series correlations, or autocorrelation. The hypothetical relationship between these two series is:

$$h_t \text{ - - - - - - - - - - - - } \rightarrow p_{t+b}.$$

Any change in the harvest causes a change in the birth and death rates b years later. There are other causal factors, however, namely,

$$h_{t-j} \text{ - - - - - - - - - - - - } \rightarrow h_t \text{ - - - - - - - - - - - - } \rightarrow p_{t+b}.$$

When the causor series is not white noise, the CCF will reflect both between-series dependencies and within-series dependencies.

To illustrate the confounding of within- and between-series correlation, we represent the relationship between two time series as

$$z_t = v_0x_t + v_1x_{t-1} + \ldots + v_kx_{t-k} + N_t.$$


```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS     =    100.

INPUT SERIES..  HARVEST   SWEDISH HARVEST INDEX 1750-1849
OUTPUT SERIES.. POP       SWEDISH POPULATION CHANGE 1750-1849

 LAG   CORR    SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                   +----+----+----+----+----+----+----+----+----+----+----+
 -15  -.076  .108                        (  XXI     )
 -14  -.073  .108                        (  XXI     )
 -13   .077  .107                        (     IXX  )
 -12   .064  .107                        (     IXX  )
 -11  -.075  .106                        (  XXI     )
 -10   .128  .105                        (     IXXX )
  -9   .165  .105                        (     IXXXX)
  -8   .116  .104                        (     IXXX )
  -7   .038  .104                        (     IX   )
  -6  -.212  .103                        *XXXXI     )
  -5  -.120  .103                        ( XXXI     )
  -4  -.044  .102                        (   XI     )
  -3   .114  .102                        (     IXXX )
  -2   .084  .101                        (     IXX  )
  -1  -.119  .101                        ( XXXI     )
   0   .057  .100                        (     IX   )
   1   .391  .101                        (     IXXXX)XXXXX
   2   .265  .101                        (     IXXXX)XX
   3   .006  .102                        (     I    )
   4  -.134  .102                        ( XXXI     )
   5  -.144  .103                        (XXXXI     )
   6  -.022  .103                        (   XI     )
   7   .026  .104                        (     IX   )
   8   .108  .104                        (     IXXX )
   9   .245  .105                        (     IXXXX)X
  10   .215  .105                        (     IXXXX*
  11   .143  .106                        (     IXXXX)
  12   .010  .107                        (     I    )
  13  -.087  .107                        (  XXI     )
  14  -.105  .108                        ( XXXI     )
  15   .005  .108                        (     I    )
                                       -2SE      +2SE
```

FIGURE 5.3(b) *Identification:* CCF for the h_t and p_t Series

Now in the general case, only a finite number of the v-weights will be nonzero. To derive the expected CCF, the v-weight model is multiplied by $x_t, x_{t-1}, \ldots, x_{t-k}$. The result is a set of k equations:

$$\begin{aligned}
x_tz_t &= v_0x_tx_t + \ldots + v_kx_tx_{t-k} + x_tN_t \\
x_{t-1}z_t &= v_0x_{t-1}x_t + \ldots + v_kx_{t-1}x_{t-k} + x_{t-1}N_t \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
x_{t-k}z_t &= v_0x_{t-k}x_t + \ldots + v_kx_{t-k}x_{t-k} + x_{t-k}N_t.
\end{aligned}$$

Taking the expectation of this equation system,

$$\begin{aligned}
\text{COV}(x_t z_t) &= v_0\sigma_x^2 &&+ \ldots + v_k\text{COV}(x_t x_{t-k}) \\
\text{COV}(x_{t-1} z_t) &= v_0\text{COV}(x_{t-1}x_t) &&+ \ldots + v_k\text{COV}(x_{t-1}x_{t-k}) \\
&\vdots &&\vdots \\
\text{COV}(x_{t-k} z_t) &= v_0\text{COV}(x_{t-k}x_t) &&+ \ldots + v_k\sigma_x^2.
\end{aligned}$$

Finally, dividing the system by $\sigma_x\sigma_z$,

$$\begin{aligned}
\text{CCF}(0) &= v_0\frac{\sigma_x}{\sigma_z} &&+ \ldots + v_k\frac{\sigma_x}{\sigma_z}\text{ACF}(k) \\
\text{CCF}(+1) &= v_0\frac{\sigma_x}{\sigma_z}\text{ACF}(1) &&+ \ldots + v_k\frac{\sigma_x}{\sigma_z}\text{ACF}(k-1) \\
&\vdots \\
\text{CCF}(+k) &= v_0\frac{\sigma_x}{\sigma_z}\text{ACF}(k) &&+ \ldots + v_k\frac{\sigma_x}{\sigma_z}.
\end{aligned}$$

The positive CCF thus is determined by the v-weight relationship between the two time series, by the variance of the two time series, *and by the ACF of the causor time series*.

Now in the New York-Paris IBM example of the previous section, both series were well represented by ARIMA (0,1,0) models. After differencing, both series were white noise. As the ACF of a white process is uniformly zero, the CCF is expected to be:

$$\begin{aligned}
\text{CCF}(0) &= v_0\frac{\sigma_x}{\sigma_z} \\
\text{CCF}(+1) &= v_1\frac{\sigma_x}{\sigma_z} \\
&\vdots \\
\text{CCF}(+k) &= v_k\frac{\sigma_x}{\sigma_z}
\end{aligned}$$

that is, the CCF will be uncontaminated by within-series correlation. In the Swedish population example, however, the Harvest Index series is *not* white noise and, as a result, no bivariate transfer function relationship can be identified from the CCF shown in Figure 5.3(b).

In theory, within-series correlation can be removed from the CCF by solving the k-equation system directly. In practice, however, this is nearly impossible. The components of the k equations, especially the ACF of the x_t series, must be estimated for the direct solution. Rounding error alone would make a direct solution inefficient and inaccurate. A more efficient and practical method of removing within-series correlation from the CCF is to *prewhiten* both series. Noting that the causor time series is well represented by an ARIMA model,

$$x_t = (1 - \phi_1 B - \ldots - \phi_p B^p)^{-1}(1 - \Theta_1 B - \ldots - \Theta_q B^q)a_t,$$

the x_t series can be prewhitened, *turned into white noise*, that is, by inverting the model.

$$a_t = (1 - \phi_1 B - \ldots - \phi_p B^p)(1 - \Theta_1 B - \ldots - \Theta_q B^q)^{-1} x_t.$$

Starting again with the v-weight relationship between the two time series,

$$z_t = v_0 x_t + v_1 x_{t-1} + \ldots + v_k x_{t-k} + N_t.$$

Applying the inverted ARIMA $(p,d,q)\ (P,D,Q)_S$ model to both sides of the equation,

$$z_t^* = v_0 a_t + v_1 a_{t-1} + \ldots + v_k a_{t-k} + N_t^*,$$

where

$$z_t^* = (1 - \phi_1 B - \ldots - \phi_p B^p)(1 - \Theta_1 B - \ldots - \Theta_q B^q)^{-1} z_t$$

$$N_t^* = (1 - \phi_1 B - \ldots - \phi_p B^p)(1 - \Theta_1 B - \ldots - \Theta_q B^q)^{-1} N_t.$$

To derive the expected CCF between z_t^* and a_t, the v-weight model is multiplied through by $a_t, a_{t-1}, \ldots a_{t-k}$. The k-equation system obtained by this procedure is:

$$\begin{aligned}
a_t z_t^* &= v_0 a_t a_t &&+ \ldots + v_k a_t a_{t-k} &&+ a_t N_t^* \\
a_{t-1} z_t^* &= v_0 a_{t-1} a_t &&+ \ldots + v_k a_{t-1} a_{t-k} &&+ a_{t-1} N_t^* \\
&\vdots &&\vdots &&\vdots \\
a_{t-k} z_t^* &= v_0 a_{t-k} a_t &&+ \ldots + v_k a_{t-k} a_{t-k} &&+ a_{t-k} N_t^*,
\end{aligned}$$

whose expectation is:

$$\begin{aligned}
\text{COV}\,(a_t z_t^*) &= v_0\sigma_a^2 \\
\text{COV}\,(a_{t-1} z_t^*) &= v_1\sigma_a^2 \\
&\vdots \\
\text{COV}\,(a_{t-k} z_t^*) &= v_k\sigma_a^2.
\end{aligned}$$

Finally, dividing these equations by $\sigma_a\sigma_{z^*}$

$$\begin{aligned}
\text{CCF}\,(0) &= v_0\frac{\sigma_a}{\sigma_{z^*}} \\
\text{CCF}\,(+1) &= v_1\frac{\sigma_a}{\sigma_{z^*}} \\
&\vdots \\
\text{CCF}\,(+k) &= v_k\frac{\sigma_a}{\sigma_{z^*}}
\end{aligned}$$

which may be a surprising result. *The CCF between the a_t and z_t^* series is proportional to the v-weights which define the bivariate relationship between x_t and z_t.* By prewhitening the time series prior to analysis, the effects of within series correlation (autocorrelation in the causor series) can be removed from the CCF.

In Figure 5.3(c), we show a bivariate model-building strategy. Like the univariate modeling strategy outlined in Section 2.11, the bivariate strategy is an iterative procedure whereby a parsimonious but statistically adequate ARIMA model is constructed. Because the strategy deals with two time series, of course, it has many more steps than its univariate analogue. The logic nonetheless is identical with the logic of the univariate model-building strategy.

As a first step, univariate models are constructed for both series. The results of these analyses will indicate whether either series must be differenced or transformed.

The univariate ARIMA model for x_t is inverted and applied to both series: *prewhitening*. A CCF is then estimated from the a_t and z_t^* series and used to identify a transfer function model for the relationship between the x_t and z_t time series.

The parameters of the transfer function are estimated. The residuals of

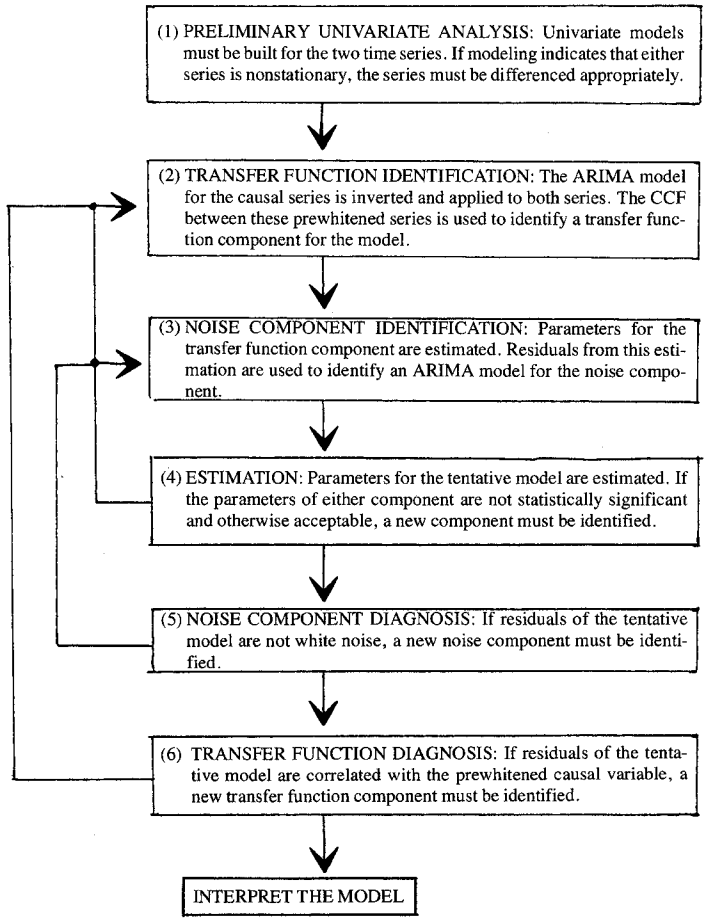


FIGURE 5.3(c) The Bivariate ARIMA Model-Building Strategy

this estimation are used to identify an ARIMA model for the N_t component.

Parameters of the fully identified model are estimated. All estimates must be statistically significant and otherwise acceptable. By "otherwise acceptable," we mean that noise component parameters must lie within the bounds of stationarity-invertibility; transfer function parameters must lie within the bounds of system stability. If the parameter estimates of either component are unacceptable, a new component must be identified.

The tentative model has two components,

$$z_t = f(x_{t-b}) + N_t,$$

and both components must pass diagnostic checks. The statistical adequacy of the noise component is diagnosed in the same way that a univariate ARIMA model is diagnosed: The model residuals must not be different than white noise.

The transfer function component has been specified so as to account for all process variance common to the x_t and z_t series. If the transfer function component is statistically inadequate, a portion of this common variance will show up as model residuals. A statistically adequate transfer function component will be independent of the noise component. To test the null hypothesis of independence, a CCF is estimated from the prewhitened x_t series and the model residuals. If the transfer function and noise components are not independent, there will be spikes at the low-order lags of the CCF. If the transfer function proves statistically inadequate by this criterion, a new transfer function must be identified, estimated, and diagnosed.

We can now apply this model-building strategy to theSwedish population growth example. Stated simply, Sundbärg's hypothesis is:

harvest in year t $\longrightarrow$ population growth in year t+b.

The analysis begins with prewhitening. As the Harvest Index series is well represented by the ARIMA (0,0,1) model

$$h_t = (1 + .39B)a_t.$$

It is prewhitened as

$$a_t = (1 + .39B)^{-1}h_t.$$

The same inverse operator is used to prewhiten the population rate increase series

$$z_t^* = (1 + .39B)^{-1}p_t.$$

The $a_t z_t^*$ CCF, shown in Figure 5.3(d), gives a clear picture of the relationship between the Harvest Index and population rate increase time series. The large spike at CCF (+1) suggests the transfer function model

$$p_t = \omega_0 h_{t-1} + N_t,$$


```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS     =    100.

INPUT SERIES..  HARVEST   PREWHITENED SWEDISH HARVEST INDEX
OUTPUT SERIES.. POP       PREWHITENED SWEDISH POPULATION CHANGE

 LAG   CORR    SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                   +----+----+----+----+----+----+----+----+----+----+----+
-15  -.053  .108                        (   XI    )
-14  -.082  .108                        (  XXI    )
-13   .078  .107                        (    IXX  )
-12   .095  .107                        (    IXX  )
-11  -.162  .106                        (XXXXI    )
-10   .152  .105                        (    IXXXX)
 -9   .105  .105                        (    IXXX )
 -8   .050  .104                        (    IX   )
 -7   .111  .104                        (    IXXX )
 -6  -.254  .103                       X(XXXXI    )
 -5  -.011  .103                        (    I    )
 -4  -.076  .102                        (  XXI    )
 -3   .111  .102                        (    IXXX )
 -2   .103  .101                        (    IXXX )
 -1  -.156  .101                        (XXXXI    )
  0  -.006  .100                        (    I    )
  1   .355  .101                        (    IXXXX)XXXX
  2   .165  .101                        (    IXXXX)
  3  -.018  .102                        (    I    )
  4  -.090  .102                        (  XXI    )
  5  -.130  .103                        ( XXXI    )
  6   .019  .103                        (    I    )
  7   .004  .104                        (    I    )
  8   .044  .104                        (    IX   )
  9   .195  .105                        (    IXXXX*
 10   .123  .105                        (    IXXX )
 11   .111  .106                        (    IXXX )
 12  -.005  .107                        (    I    )
 13  -.056  .107                        (   XI    )
 14  -.105  .108                        (  XXXI   )
 15   .043  .108                        (    IX   )
                                      -2SE     +2SE
```

FIGURE 5.3(d) *Identification:* CCF for the Prewhitened p_t and h_t Series

which implies that the current year's harvest determines the next year's population growth. Our estimate of ω_0 for this relationship is:

$$\hat{\omega}_0 = .87.$$

The residual ACF for this estimate, shown in Figure 5.3(e), suggests an ARIMA (1,0,0) model for the N_t component. This leads to the full model

$$p_t = \omega_0 h_{t-1} + \frac{a_t}{1 - \phi_1 B}.$$

```
SERIES.. RESIDUAL (NOBS= 100) ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =  100.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 29, 100) =  43.730      SIG =      .039

 LAG    CORR     SE   -1    -.8   -.6   -.4   -.2    0     .2    .4    .6    .8   +1
                      +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
   1    .450   .100                          (     IXXXX)XXXXXX
   2    .200   .120                        (       IXXXXX)
   3    .050   .120                        (       IX     )
   4    .080   .120                        (       IXX    )
   5    .130   .120                        (       IXXX   )
   6    .090   .120                        (       IXX    )
   7   -.050   .120                        (      XI      )
   8   -.170   .130                      (  XXXXI         )
   9   -.130   .130                      (   XXXI         )
  10   -.020   .130                      (     XI         )
  11   -.010   .130                      (      I         )
  12    .020   .130                      (      IX        )
  13   -.070   .130                      (    XXI         )
  14   -.090   .130                      (    XXI         )
  15   -.050   .130                      (     XI         )
  16    .040   .130                      (      IX        )
  17    .090   .130                      (      IXX       )
  18    .110   .130                      (      IXXX      )
  19    .160   .130                      (      IXXXX     )
  20    .100   .130                      (      IXXX      )
  21    .080   .130                      (      IXX       )
  22    .080   .130                      (      IXX       )
  23    .000   .140                      (      I         )
  24    .060   .140                      (      IXX       )
  25    .050   .140                      (      IX        )
  26    .020   .140                      (      IX        )
  27   -.070   .140                      (    XXI         )
  28   -.090   .140                      (    XXI         )
  29   -.080   .140                      (    XXI         )
  30   -.010   .140                      (      I         )
                                        -2SE             +2SE
```

FIGURE 5.3(e) *Identification:* ACF for the Residuals of the Model
$p_t = .866\, h_{t-1} + a_t$

Parameter estimates for this full model are:

$$\hat{\omega}_0 = .826 \text{ with t statistic} = 4.18$$
$$\hat{\phi}_1 = .460 \text{ with t statistic} = 4.94.$$

Both parameter estimates are statistically significant and otherwise acceptable.

To diagnose the tentative model, we first require that the model residuals are not different than white noise. The residual ACF, shown in Figure 5.3(f), has no statistically significant spikes and the Q-statistic is not significant. Second, the model residuals and the prewhitened Harvest Index series must be uncorrelated. The CCF for the residuals and the prewhitened h_t series, shown in Figure 5.3(g), has one statistically significant spike. As this tentative model satisfies both diagnostic criteria, we accept it.


```
SERIES.. RESIDUAL (NOBS=  99) ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =   99.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 28,  99) =  12.088      SIG =      .996

 LAG    CORR     SE   -1    -.8   -.6   -.4   -.2    0     .2    .4    .6    .8   +1
                      +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
   1    .000   .100                          (     I     )
   2    .010   .100                          (     I     )
   3   -.100   .100                          ( XXXI      )
   4    .020   .100                          (     IX    )
   5    .100   .100                          (     IXXX  )
   6    .090   .100                          (     IXX   )
   7   -.030   .100                          (    XI     )
   8   -.140   .100                          (XXXXI      )
   9   -.090   .110                        (    XXI      )
  10    .060   .110                        (      IXX    )
  11   -.010   .110                        (      I      )
  12    .070   .110                        (      IXX    )
  13   -.070   .110                        (    XXI      )
  14   -.080   .110                        (    XXI      )
  15   -.040   .110                        (     XI      )
  16    .040   .110                        (      IX     )
  17    .060   .110                        (      IXX    )
  18    .030   .110                        (      IX     )
  19    .120   .110                        (      IXXX   )
  20    .030   .110                        (      IX     )
  21    .010   .110                        (      I      )
  22    .090   .110                        (      IXX    )
  23   -.020   .110                        (     XI      )
  24    .020   .110                        (      IX     )
  25    .040   .110                        (      IX     )
  26    .030   .110                        (      IX     )
  27   -.060   .110                        (    XXI      )
  28   -.040   .110                        (     XI      )
  29   -.070   .110                        (    XXI      )
  30    .020   .110                        (      IX     )
                                        -2SE             +2SE
```

FIGURE 5.3(f) *Diagnosis:* ACF for the Residuals of the Model
$$p_t = .828\, h_{t-1} + \frac{a_t}{1 - .459B}$$

In years following a crop failure (a value of Harvest Index = 0 indicates crop failure), Swedish population was expected to increase by only .661% (the

$$p_t = .826 h_{t-1} + \frac{a_t}{1 - .46B}.$$

```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS     =     99.

INPUT SERIES..  HARVEST    PREWHITENED SWEDISH HARVEST INDEX
OUTPUT SERIES.. RESIDUAL   ESTIMATED MODEL RESIDUALS

 LAG    CORR     SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                     +----+----+----+----+----+----+----+----+----+----+
 -15  -.084   .109                        (  XXI     )
 -14   .092   .108                        (    IXX   )
 -13   .136   .108                        (    IXXX  )
 -12  -.110   .107                        ( XXXI     )
 -11   .163   .107                        (    IXXXX )
 -10   .104   .106                        (    IXXX  )
  -9   .006   .105                        (    I     )
  -8   .090   .105                        (    IXX   )
  -7  -.251   .104                       X(XXXI     )
  -6   .045   .104                        (    IX    )
  -5  -.043   .103                        (   XI     )
  -4   .132   .103                        (    IXXX  )
  -3   .165   .102                        (    IXXXX )
  -2  -.201   .102                        *XXXI      )
  -1  -.049   .101                        (   XI     )
   0   .008   .101                        (    I     )
   1   .185   .101                        (    IXXXX*
   2  -.022   .102                        (   XI     )
   3  -.067   .102                        (  XXI     )
   4  -.117   .103                        ( XXXI     )
   5   .042   .103                        (    IX    )
   6   .084   .104                        (    IXX   )
   7   .063   .104                        (    IXX   )
   8   .170   .105                        (    IXXXX )
   9   .108   .105                        (    IXXX  )
  10   .073   .106                        (    IXX   )
  11  -.028   .107                        (   XI     )
  12  -.009   .107                        (    I     )
  13  -.121   .108                        ( XXXI     )
  14   .051   .108                        (    IX    )
  15   .010   .109                        (    I     )
                                        -2SE      +2SE
```

FIGURE 5.3(g) *Diagnosis:* CCF for Prewhitened h_t Series and the Model Residuals

mean of the p_t series). In years following a superabundant crop (Harvest Index = 9), population was expected to increase by more than 1.4%.

Overall, the construction *and* interpretation of this model are simplified because it is a *bi*variate model. *In all cases,* bivariate ARIMA models (such as the Paris-New York IBM model of the previous section and population-harvest model of this section) are constructed by the routine outlined in Figure 5.3(c). In our experiences, students will have little trouble with bivariate analysis if this strategy is followed mechanically. The interpretation of a bivariate model is also straightforward: "x causes y." In both of the bivariate example analyses of this chapter, changes in one variable are *followed* by changes in the other. While this finding does not "prove" causation, the causal null hypothesis is rejected and the causal hypothesis is consequently more plausible than it was prior to the analysis. Model construction and interpretation are not so simple in the *multi*variate case, however. When a model has more than one input (or independent-variable) time series, the model-building strategy becomes less mechanical and interpretations of the model become more complicated. Nevertheless, it is only in the multivariate case that ARIMA models and methods achieve their full potential as tools of social research.

5.4 Multivariate Population Growth Model[3]

The bivariate ARIMA analysis supports Sundbärg's hypothesis. No matter how crucial agricultural production may have been during that century, however, we cannot conclude that the harvest was the sole determinant of population growth. A crop failure no doubt effected an increase in the death rate but the affect on the birth rate must have been less substantial. To better explain Swedish population growth, then, we can add other independent variables to the model.

A likely predictor of birth rates is shown in Figure 5.4(a). These are annual fertility rates (per thousand) for the 1750–1849 century. The fertility rate in the t^{th} year is defined as

$$f_t = \text{births per 1000 female population.}$$

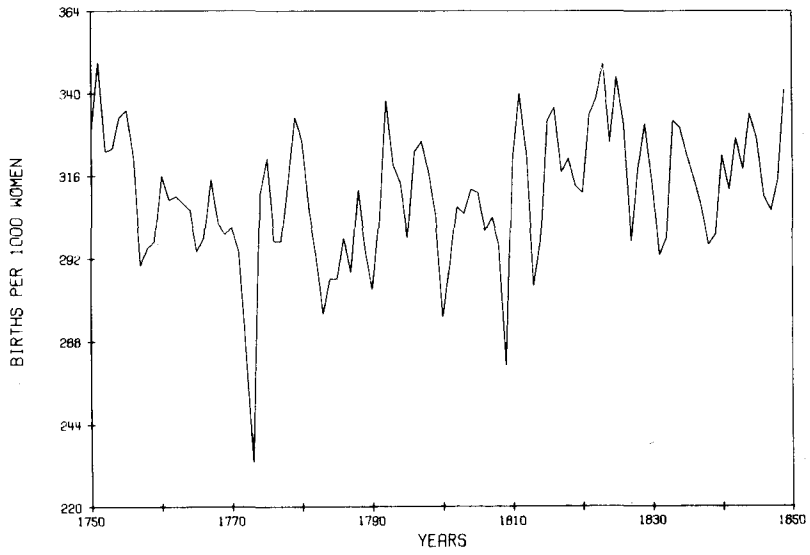


FIGURE 5.4(a) Swedish Fertility Rates, 1750–1849

```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS     =     98.

INPUT SERIES..  FERTILE   PREWHITENED SWEDISH FERTILITY RATE
OUTPUT SERIES.. POP       PREWHITENED SWEDISH POPULATION CHANGE

 LAG   CORR    SE   -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                   +----+----+----+----+----+----+----+----+----+----+----+
 -15  -.070  .110                      (  XXI    )
 -14   .129  .109                      (    IXXX )
 -13  -.123  .108                      ( XXXI    )
 -12   .059  .108                      (    IX   )
 -11   .002  .107                      (    I    )
 -10   .107  .107                      (    IXXX )
  -9  -.053  .106                      (   XI    )
  -8   .039  .105                      (    IX   )
  -7  -.075  .105                      (  XXI    )
  -6  -.132  .104                      ( XXXI    )
  -5   .160  .104                      (    IXXXX)
  -4   .098  .103                      (    IXX  )
  -3   .027  .103                      (    IX   )
  -2  -.092  .102                      (  XXI    )
  -1   .093  .102                      (    IXX  )
   0   .712  .101                      (    IXXXX)XXXXXXXXXXXXX
   1   .028  .102                      (    IX   )
   2   .095  .102                      (    IXX  )
   3  -.009  .103                      (    I    )
   4   .085  .103                      (    IXX  )
   5   .037  .104                      (    IX   )
   6   .141  .104                      (    IXXXX)
   7   .010  .105                      (    I    )
   8   .082  .105                      (    IXX  )
   9   .005  .106                      (    I    )
  10   .048  .107                      (    IX   )
  11   .011  .107                      (    I    )
  12   .087  .108                      (    IXX  )
  13  -.010  .108                      (    I    )
  14   .097  .109                      (    IXX  )
  15  -.011  .110                      (    I    )
                                     -2SE      +2SE
```

FIGURE 5.4(b) *Identification:* CCF for the Prewhitened p_t and h_t Series

The fertility rate is thus a type of birth rate. Assuming no effective birth control methods, fertility rate is a measure of the number of females of child-bearing age (15 to 45 years old) in the population. An analysis of the f_t series shows that it is well represented by the ARIMA (2,0,0) model

$$(1 - .62B + .23B^2)f_t = a_t.$$

To prewhiten this series, the ARIMA (2,0,0) model is applied:

$$\begin{aligned} a_t &= (1 - .62B + .23B^2)f_t \\ &= f_t - .62f_{t-1} + .23f_{t-2}. \end{aligned}$$


```
SERIES.. RESIDUAL (NOBS= 100) ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =  100.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 29, 100) =  26.740      SIG =      .586

 LAG   CORR    SE   -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                   +----+----+----+----+----+----+----+----+----+----+----+
   1   .250  .100                      (    IXXXX)X
   2   .030  .110                      (    IX   )
   3   .050  .110                      (    IX   )
   4   .050  .110                      (    IX   )
   5   .050  .110                      (    IX   )
   6   .150  .110                      (    IXXXX )
   7   .020  .110                      (    IX   )
   8  -.090  .110                      (   XXI   )
   9  -.080  .110                      (   XXI   )
  10   .050  .110                      (    IX   )
  11   .050  .110                      (    IX   )
  12  -.140  .110                      ( XXXXI   )
  13  -.040  .110                      (    XI   )
  14  -.110  .110                      (  XXXI   )
  15  -.100  .110                      (  XXXI   )
  16   .010  .110                      (    I    )
  17   .070  .110                      (    IXX  )
  18   .030  .120                      (    IX   )
  19   .070  .120                      (    IXX  )
  20  -.030  .120                      (   XI    )
  21  -.030  .120                      (   XI    )
  22   .010  .120                      (    I    )
  23  -.030  .120                      (   XI    )
  24  -.110  .120                      (  XXXI   )
  25  -.110  .120                      (  XXXI   )
  26  -.040  .120                      (   XI    )
  27  -.100  .120                      (  XXXI   )
  28  -.180  .120                      (XXXXXI   )
  29  -.130  .120                      (  XXXI   )
  30  -.110  .120                      (  XXXI   )
                                     -2SE      +2SE
```

FIGURE 5.4(c) *Identification:* ACF for the Residuals of the Model $p_t = .239\, f_t + a_t$

The same operator is then applied to the population rate increase time series

$$\begin{aligned} z_t^* &= (1 - .62B + .23B^2)p_t \\ &= p_t - .62p_{t-1} + .23p_{t-2}. \end{aligned}$$

The CCF for a_t and z_t^*, shown in Figure 5.4(b), suggests a zero-order transfer function for $f_t$and p_t with no time lag.

$$p_t = \omega_0 f_t + N_t.$$

```
SERIES.. RESIDUAL (NOBS= 100) ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =  100.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 29, 100) =  27.040      SIG =      .570

 LAG   CORR    SE  -1   -.8  -.6  -.4  -.2   0    .2   .4   .6   .8  +1
                   +----+----+----+----+----+----+----+----+----+----+----+
   1   .250  .100                        (    IXXXX)X
   2   .030  .110                        (     IX    )
   3   .050  .110                        (     IX    )
   4   .050  .110                        (     IX    )
   5   .050  .110                        (     IX    )
   6   .150  .110                        (     IXXXX )
   7   .020  .110                        (     IX    )
   8  -.090  .110                        (   XXI     )
   9  -.080  .110                        (   XXI     )
  10   .050  .110                        (     IX    )
  11   .050  .110                        (     IX    )
  12  -.140  .110                        ( XXXXI     )
  13  -.040  .110                        (    XI     )
  14  -.110  .110                        (  XXXI     )
  15  -.110  .110                        (  XXXI     )
  16   .010  .110                        (     I     )
  17   .070  .110                        (     IXX   )
  18   .030  .120                        (     IX    )
  19   .060  .120                        (     IXX   )
  20  -.030  .120                        (    XI     )
  21  -.030  .120                        (    XI     )
  22   .000  .120                        (     I     )
  23  -.030  .120                        (    XI     )
  24  -.120  .120                        (  XXXI     )
  25  -.110  .120                        (  XXXI     )
  26  -.040  .120                        (    XI     )
  27  -.100  .120                        (  XXXI     )
  28  -.180  .120                        (XXXXXI     )
  29  -.130  .120                        (  XXXI     )
  30  -.110  .120                        (  XXXI     )
                                      -2SE        +2SE
```

FIGURE 5.4(d) *Identification:* ACF for the Residuals of the Model $p_t = .239\, f_t - .01\, h_{t-1} + a_t$

Population growth is determined by fertility rates in the same year. As the fertility rate is a type of birth rate, the zero time lag, indicated by a spike at CCF (0), makes good sense. The estimate of ω_0 is:

$$\hat{\omega}_0 = .239.$$

The residual ACF for this estimate, shown in Figure 5.4(c), suggests an ARIMA (1,0,0) model for N_t. The fully specified bivariate model is thus

$$p_t = \omega_0 f_t + \frac{a_t}{1 - \phi_1 B}.$$

Parameter estimates for this model are:

$$\hat{\omega}_0 = .234 \text{ with t statistic} = 11.44$$
$$\hat{\phi}_1 = .26 \text{ with t statistic} = 2.55.$$

Both parameter estimates are statistically significant and otherwise acceptable. Diagnosis of the model residuals indicates that the noise and transfer function components are independent and that the model residuals are white noise. We thus accept this tentative model.

So far, the analysis of population rate increases leads us to conclude that the growth of Swedish population during the 1750–1849 century was due to (or caused by) the effects of *two* exogenous variables: agricultural production and fertility. The bivariate models for these two relationships are:

$$p_t = .826 h_{t-1} + \frac{a_t}{1 - .46B}$$

and

$$p_t = .234 f_t + \frac{a_t}{1 - .26B}.$$

A logical next step would be to incorporate both exogenous variables into a single multivariate ARIMA model. On the basis of the bivariate models, the multivariate model is specified as

$$p_t = \omega_0 f_t + \omega_0^* h_{t-1} + N_t.$$

Parameter estimates for the transfer function component are:

$$\hat{\omega}_0 = .239$$
$$\hat{\omega}_0^* = -.010.$$

Note that the ω-parameter for the effect of the Harvest Index series has dropped substantially in absolute value. This estimate could change dramatically when noise parameters are estimated, however. The residual ACF for these estimates, shown in Figure 5.4(d), suggests an ARIMA (1,0,0) model for the N_t component. The full model is thus

$$p_t = \omega_0 f_t + \omega_0^* h_{t-1} + \frac{a_t}{1 - \phi_1 B}$$

Parameter estimates for this tentative model are:

$$\hat{\omega}_0 = .231 \text{ with t statistic} = 2.64$$
$$\hat{\omega}_0^* = .046 \text{ with t statistic} = .28$$
$$\hat{\phi}_1 = .26 \text{ with t statistic} = 2.64.$$

There is no need to diagnose this tentative model. The estimate of ω_0^* is not statistically significant and must be dropped from the model.

This analysis would seem to disconfirm Sundbärg's hypothesis. When fertility rates are considered, the Harvest Index time series accounts for only a statistically insignificant proportion of the variance in the population change time series.

There is a simpler explanation for the finding of this analysis, however. A multivariate ARIMA model of the sort

$$y_t = f(x_{1t}) + \ldots + f(x_{nt}) + N_t$$

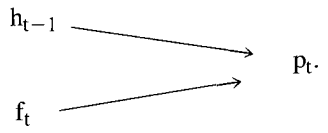
is generally nonlinear but is linear in terms of its components. To estimate parameters for the full model, *all components must be independent*.

As it turns out (and as a diagnosis of this model would have indicated), the fertility rate and harvest time series are highly correlated. The harvest in effect determines the values of future fertility rates.

$$h_t \longrightarrow f_{t+b}.$$

Thomas (1940) cites a number of plausible mechanisms for this relationship. First, in years following a crop failure, marriage rates (and hence, fertility rates) drop. Second, and more important, in years following a crop failure, young women who might otherwise bear children in Sweden are likely to emigrate (primarily to Finland and the United States during this period). As a result of emigration, the average age of the female population rises dramatically in years following a crop failure and fertility drops accordingly

If there is indeed a causal relationship between harvests and fertility rates, the two transfer function components of the multivariate ARIMA model are not independent. The model we built implies that



If there is a causal relationship between the harvest and fertility rates, however, the true model is:

$$h_{t-1} \longrightarrow f_t \longrightarrow p_t.$$

Our finding of no effect for the Harvest Index time series could thus be due only to a misspecification of the model.

As a first step in building a multivariate population growth model, the hypothesized causal relationship between harvests and fertility rates must be tested. The time series are prewhitened as

$$a_t = (1 + .39B)^{-1} h_t$$
$$z_t^* = (1 + .39B)^{-1} f_t.$$

CROSS-CORRELATIONS OF LAGS -15 TO 15.
NO. OF VALID OBSERVATIONS = 100.

INPUT SERIES.. HARVEST PREWHITENED SWEDISH HARVEST INDEX
OUTPUT SERIES.. FERTILE PREWHITENED SWEDISH FERTILITY RATE

LAG	CORR	SE	-1 -.8 -.6 -.4 -.2 0 .2 .4 .6 .8 +1
-15	-.006	.108	(I)
-14	-.024	.108	(XI)
-13	.121	.107	(IXXX)
-12	.093	.107	(IXX)
-11	-.110	.106	(XXXI)
-10	.039	.105	(IX)
-9	.146	.105	(IXXXX)
-8	-.045	.104	(XI)
-7	.053	.104	(IX)
-6	-.220	.103	X(XXXXI)
-5	.029	.103	(IX)
-4	.005	.102	(I)
-3	.024	.102	(IX)
-2	.096	.101	(IXX)
-1	-.269	.101	XX(XXXXI)
0	.017	.100	(I)
1	.417	.101	(IXXXX)XXXXX
2	.347	.101	(IXXXX)XXXX
3	-.079	.102	(XXI)
4	-.086	.102	(XXI)
5	.013	.103	(I)
6	.024	.103	(IX)
7	-.091	.104	(XXI)
8	0.000	.104	(I)
9	.185	.105	(IXXXX*
10	.099	.105	(IXX)
11	.017	.106	(I)
12	-.031	.107	(XI)
13	-.050	.107	(XI)
14	-.051	.108	(XI)
15	.118	.108	(IXXX)

-2SE +2SE

FIGURE 5.4(e) *Identification:* CCF for the Prewhitened f_t and h_t Series

```
SERIES.. RESIDUAL (NOBS=  99)  ESTIMATED MODEL RESIDUALS
NO. OF VALID OBSERVATIONS  =     99.

AUTOCORRELATIONS OF LAGS 1 - 30.
Q( 29,  99) =  52.133      SIG =      .005

 LAG    CORR     SE   -1    -.8   -.6   -.4   -.2    0     .2    .4    .6    .8   +1
                      +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
       .480   .100                          (     IXXXX)XXXXXXX
   2   .260   .120                           (      IXXXXX)X
   3   .140   .130                          (       IXXXX  )
   4   .230   .130                          (       IXXXXXX)
   5   .160   .130                          (       IXXXX  )
   6   .000   .130                          (       I      )
   7   .060   .130                          (       IXX    )
   8  -.650   .140                          (     XI       )
   9  -.030   .140                          (      XI      )
  10   .010   .140                          (       I      )
  11   .090   .140                          (       IXX    )
  12   .110   .140                          (       IXXX   )
  13   .060   .140                          (       IXX    )
  14   .030   .140                          (       IX     )
  15  -.010   .140                          (       I      )
  16   .000   .140                          (       I      )
  17   .100   .140                          (       IXXX   )
  18   .180   .140                          (       IXXXXX )
  19   .160   .140                          (       IXXXX  )
  20   .050   .140                          (       IX     )
  21   .020   .140                          (       IX     )
  22   .060   .140                          (       IXX    )
  23   .090   .140                          (       IXX    )
  24   .080   .140                          (       IXX    )
  25  -.020   .140                          (     XI       )
  26  -.010   .140                          (       I      )
  27  -.050   .140                          (     XI       )
  28  -.070   .140                          (    XXI       )
  29  -.040   .140                          (     XI       )
  30   .010   .140                          (       I      )
                                          -2SE          +2SE
```

FIGURE 5.4(f) *Identification:* ACF for the Residuals of the Model

$$f_t = \frac{3.498}{1 - .372B} h_{t-1} + a_t$$

The CCF for a_t and z_t^*, shown in Figure 5.4(e), shows a strong and unambiguous casual relationship between the Harvest Index and fertility rates time series. Statistically significant spikes at CCF (+1) and CCF (+2) imply the model

$$f_t = \frac{\omega_0}{1 - \delta_1 B} h_{t-1} + N_t.$$

Parameter estimates for this model are:

$$\hat{\omega}_0 = 3.498$$
$$\hat{\delta}_1 = .372.$$


```
CROSS-CORRELATIONS OF LAGS  -15 TO  15.
NO. OF VALID OBSERVATIONS      =     99.

INPUT SERIES..  POP        SWEDISH POPULATION CHANGE
OUTPUT SERIES.. RESIDUAL   ESTIMATED MODEL RESIDUALS

 LAG    CORR     SE   -1    -.8   -.6   -.4   -.2    0     .2    .4    .6    .8   +1
                      +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 -15  -.036   .109                          (    XI     )
 -14  -.006   .108                          (     I     )
 -13  -.130   .108                          ( XXXI     )
 -12   .045   .107                          (     IX     )
 -11  -.043   .107                          (    XI      )
 -10   .034   .106                          (     IX     )
  -9  -.023   .105                          (    XI      )
  -8  -.007   .105                          (     I      )
  -7   .011   .104                          (     I      )
  -6  -.141   .104                          (XXXXI       )
  -5   .116   .103                          (     IXXX   )
  -4   .121   .103                          (     IXXX   )
  -3  -.034   .102                          (    XI      )
  -2  -.145   .102                          (XXXXI       )
  -1  -.006   .101                          (     I      )
   0   .528   .101                          (     IXXXX)XXXXXXXX
   1   .308   .101                          (     IXXXX)XXX
   2   .196   .102                          (     IXXXX*
   3   .075   .102                          (     IXX    )
   4   .122   .103                          (     IXXX   )
   5   .074   .103                          (     IXX    )
   6   .134   .104                          (     IXXX   )
   7   .033   .104                          (     IX     )
   8  -.053   .105                          (    XI      )
   9  -.095   .105                          (   XXI      )
  10  -.070   .106                          (   XXI      )
  11   .004   .107                          (     I      )
  12   .136   .107                          (     IXXX   )
  13   .127   .108                          (     IXXX   )
  14   .100   .108                          (     IXXX   )
  15   .035   .109                          (     IX     )
                                          -2SE          +2SE
```

FIGURE 5.4(g) *Identification:* CCF for the f_t^* Series

The residual ACF for these estimates, shown in Figure 5.4(f), suggest an ARIMA (1,0,0) model for the N_t component. The full model is thus

$$f_t = \frac{\omega_0}{1 - \delta_1 B} h_{t-1} + \frac{a_t}{1 - \phi_1 B}.$$

Parameter estimates for the full model are:

$$\hat{\omega}_0 = 3.572 \text{ with t statistic} = 6.17$$

$$\hat{\delta}_1 = .439 \text{ with t statistic} = 3.34$$
$$\hat{\phi}_1 = .50 \text{ with t statistic} = 5.45.$$

All parameter estimates are statistically significant and otherwise acceptable. Diagnosis of this model indicates that the noise and transfer function components are independent and that the model residuals are white noise.

The analysis confirms the hypothesis that there is a strong causal relationship between harvests and fertility rates and the implications of this finding for a multivariate population growth model are clear. The multivariate model must reflect the structure

$$h_{t-1} \longrightarrow f_t \longrightarrow p_t.$$

To build this structure, we require a fertility rate time series that has been purged of the harvest effect. The residuals of the harvest-fertility bivariate model will prove adequate for this purpose. As this bivariate model is:

$$f_t = \frac{3.6}{1 - .44B} h_{t-1} + \frac{a_t}{1 - .50B},$$

we define a new fertility rate time series, f_t^*, as

$$f_t^* = a_t = (1 - .50B) f_t - \frac{1 - .50B}{1 - .44B} 3.6 h_{t-1}.$$

The f_t^* series is white noise and, by definition, uncorrelated with the h_t series. These two time series may thus be incorporated directly into the multivariate population growth model.

The transfer function relationship between the h_t and p_t series has already been determined. The transfer function relationship between f_t^* and p_t must be identified, however. We would ordinarily begin by prewhitening the series. But f_t^* is hypothesized to be the causor and, as f_t^* is already white noise, prewhitening is not required. The CCF estimated from f_t^* and p_t, shown in Figure 5.4(g), has significant spikes at CCF (0) and CCF (+1), suggesting the transfer function relationship

$$p_t = \frac{\omega_0}{1 - \delta_1 B} f_t^* + N_t.$$

Adding this structure to the previously identified relationship between p_t and

h_t, the full model is:

$$p_t = \frac{\omega_0}{1 - \delta_1 B} f_t^* + \omega_0^* h_{t-1} + N_t.$$

Parameter estimates for this transfer function are:

$$\hat{\omega}_0 = .234$$
$$\hat{\omega}_0^* = 1.017$$
$$\hat{\delta}_1 = .616.$$

The residual ACF for these estimates, shown in Figure 5.4(h), suggests an ARIMA (1,0,0) model for the N_t component. The fully specified model is thus

$$p_t = \frac{\omega_0}{1 - \delta_1 B} f_t^* + \omega_0^* h_{t-1} + \frac{a_t}{1 - \phi_1 B}.$$

Parameter estimates for this full model are:

$$\hat{\omega}_0 = .234 \text{ with t statistic} = 8.94$$
$$\hat{\omega}_0^* = .952 \text{ with t statistic} = 6.44$$
$$\hat{\delta}_1 = .620 \text{ with t statistic} = 9.70$$
$$\hat{\phi}_1 = .255 \text{ with t statistic} = 2.48.$$

All parameter estimates are statistically significant and otherwise acceptable. Diagnosis indicates that all components are independent of each other and that the model residuals are white noise. We accept this tentative model.

At this point, we must review our progress by comparing three models of Swedish population growth. First, the univariate model

$$p_t = (1 + .44B)a_t$$

has an RMS = 27.8. Second, the bivariate model

$$p_t = .826h_{t-1} + \frac{a_t}{1 - .46B}$$

has an RMS = 23.6. By incorporating the effects of harvests, the predictive

power of the model increases substantially. Finally, the multivariate model

$$p_t = \frac{.23}{1 - .62B} f_t^* + .952h_{t-1} + \frac{a_t}{1 - .26B}$$

has an RMS = 12.3. Again, by incorporating fertility rates, the predictive power of the model increases substantially.

The bivariate and multivariate models have increased our understanding of Swedish population growth in another, more important sense. While the univariate model does an adequate job of *predicting*, it does nothing to *explain* the substantive phenomenon. The bivariate and multivariate models, on the other hand, explain population growth in *causal* terms.

Figure 5.4(h) diagrams the multivariate population growth model as an input-output system. According to this diagram, population changes are due to the effects of three exogenous forces. Two of the three exogenous forces are white noise processes, a_t and f_t^*. The third exogenous force is h_{t-1}, the Harvest Index time series. Fertility rates, denoted by f_t in the diagram, play

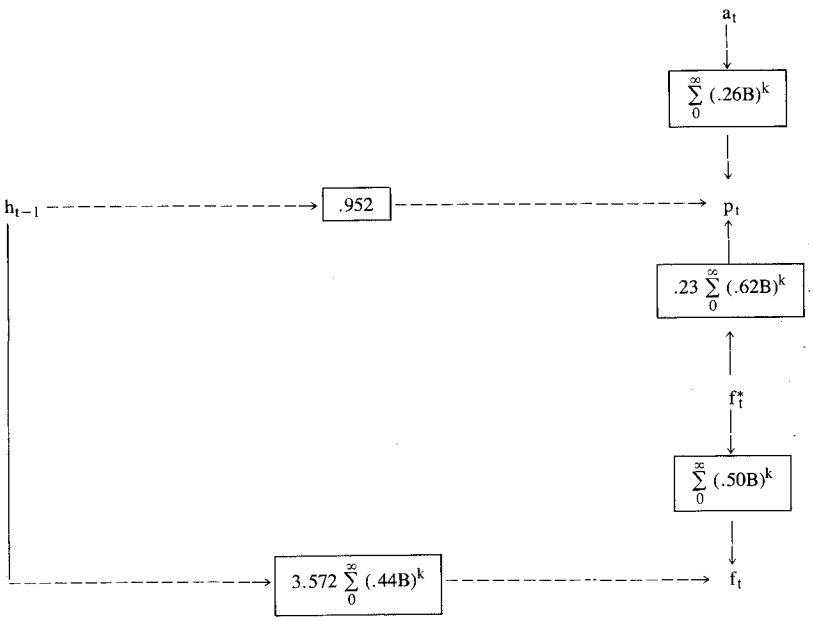


FIGURE 5.4(h) An Input-Output Diagram of the Multivariate Model

no causal role in this system. The f_t series is included in the model only because it shares two sources of variance with the p_t series. Causal arrows in the diagram lead *to* the f_t variable but not *from* it.

The reader who is familiar with structural equations models (see, e.g., Blalock, 1971; Goldberger and Duncan, 1973; Heise, 1975) or econometric models (see, e.g., Kmenta, 1971; Johnston, 1972) will immediately recognize Figure 5.4(h) for what it is: a structural model of Swedish population growth. The reader may surmise from this that the structural equations or econometric approaches to time series analysis and the ARIMA approach are substantially the same. In fact, this multivariate ARIMA model can be used for any purpose that a structural equations or econometric model would be used for; and of course, the ARIMA model is subject to any and all criticisms that could be made of a structural equations or econometric models.

The major difference between econometric or structural equations approaches to time series analysis and the ARIMA approach is that econometric models are ordinarily identified theoretically. ARIMA models are identified empirically, of course, and because of this, ARIMA models require relatively long time series. Beyond this practical point, there is no difference.

A structural equations model ordinarily begins with a set of "structural equations" deduced from theory. These elementary equations are then manipulated until a "reduced form equation" emerges. The reduced form equation includes all variables of the system and, under ideal conditions, parameters can be estimated directly from the reduced form. This is also true of the multivariate ARIMA model. The empirically identified ARIMA model for Swedish population growth is:

$$p_t = \frac{.23}{1 - .62B} f_t^* + .95h_{t-1} + \frac{a_t}{1 - .26B}.$$

However, noting that f_t^* is related to f_t as

$$f_t = \frac{3.6}{1 - .44B} h_{t-1} + \frac{f_t^*}{1 - .50B}$$

so

$$f_t^* = (1 - .50) f_t - \frac{(3.6)(1 - .50B)}{1 - .44B} h_{t-1}.$$

This expression for f_t^* may be substituted into the multivariate model to obtain

$$p_t = \frac{(.23)(1 - .50B)}{1 - .62B} f_t$$

$$- \left[\frac{(.23)(3.6)(1 - .50B)}{(1 - .62B)(1 - .44B)} - .952\right] h_{t-1}$$

$$+ \frac{a_t}{1 - .26B}.$$

This formidable model is the reduced form equation that would have resulted from the elementary structural equations. In this case, and in the general case, structural equations or econometric approaches to time series analysis and the ARIMA approach lead to the same model. The reader will best understand the ARIMA approach to time series analysis by viewing it as a special case of the structural equations or econometric approach. Multivariate ARIMA models are structural equations or econometric models in which the relationships among variables have been identified empirically. Empirical identification implies that relatively long time series are available, of course.

The final step in multivariate ARIMA analysis is model interpretation. As this relates to Sundbärg's hypothesis, the model shows clearly that, during the 1750–1849 century, *the harvest had a profound influence on population growth.*

5.5 Conclusion and Recommendations

So far, there has been little published research in the social sciences using multivariate ARIMA methods. One reason often cited for this situation is that the computer software required for multivariate ARIMA analysis is not widely available. We will address this point directly in the next chapter. For the time being, we note only that while this has been true in the past, there are now many suitable computer programs readily available in academic computing centers.

Another reason often cited for this situation is that the time series data required for ARIMA models are not available. An ARIMA model often requires a time series of 100 observations or more and data of this quantity are seldom found in the social sciences. Data availability varies from substantive area to substantive area, of course. While long time series may indeed be rare, we suspect that they are not so rare as people think. In the course of writing this volume, we encountered many data sets long enough for ARIMA analysis. Moreover, as the use of computers spreads, we suspect

that long time series will become increasingly common in the social sciences.

Finally, there is a popular misconception about the nature of ARIMA methods which might explain why these methods are not widely used. Regarding structural equations or econometric approaches to time series analysis, for example, the popular notion is that ARIMA models are empirical and *a*theoretical (or mindless) while econometric models are sometimes empirical but always theoretical. On this point, Hibbs writes.

> Box-Tiao or Box-Jenkins methods are essentially models for "ignorance" that are not based on theory and, in this sense, are void of explanatory power. Although these models are in many situations likely to yield good estimates of endogenous reponses to external interventions, they provide no insight into the causal structure underlying the transmission of exogenous impulses through a dynamic system of interdependent social, economic, or political relationships [1977: 172].

In this oft-quoted passage, however, Hibbs refers only to *univariate* ARIMA models. When multivariate models are considered, the differences between these two approaches to time series analysis are largely practical differences which have nothing to do with the quality of the models.

When ARIMA models were introduced to the social sciences a decade ago, an unproductive debate ensued over the relative merits of ARIMA and econometric approaches to time series analysis. More recently, these two approaches have converged. An eclectic approach to time series analysis recognizes that multivariate ARIMA models and econometric models are identical in every substantial respect.

The advantages of ARIMA models over econometric or structural equations models are obvious. Lag structures among variables are identified more precisely; seasonal variance is accounted for in a systematic manner; model parameters are estimated with a high degree of reliability; and so forth. But all of these advantages follow from the quantity of data required for the ARIMA model. ARIMA models require relatively long time series and, in this sense, multivariate ARIMA models are "better" than econometric models only because long time series are "better" than short time series.

The advantages of econometric or structural equations models over ARIMA models are also a consequence of this data requirement. In macroeconomic applications, an econometric model may incorporate *hundreds* of time series variables. An ARIMA model based on this same number of time series would have *thousands* of parameters to be estimated and, thus,

would require time series of a thousand or more observations. Assuming that time series of more than 200 or 300 observations are unavailable, multivariate ARIMA models are ordinarily limited to a dozen or fewer time series variables. Thus, when a rich body of theory points to many variables and many structural relationships, the econometric or structural equations approach to time series analysis will be superior to the ARIMA approach.

In our opinion, multivariate ARIMA models will become an important research tool for the social sciences. ARIMA models will never *replace* structural models as a research tool, however. These two approaches to time series analysis each have strengths and weaknesses. The particular approach taken should be determined by the quality and quantity of data available and by the problem itself.

For Further Reading

There is a considerable "transfer of learning" between the ARIMA and regression approaches to time series analysis. Hibbs (1974) and Ostrom (1978) both develop several useful regression models for time series analysis. The Hibbs work has become a classic in its field and should not be missed. Structural equations or econometric approaches to time series analysis differ from the more general regression approaches in that a model based on *several* regression equations is posited. Kmenta (1971) develops much of econometric theory without linear algebra. Johnston (1972) develops the same material but from a linear algebra basis; in addition, Johnston (1972: Chapter 4) includes an excellent introduction to linear algebra. The Kmenta book is widely used as an undergraduate text while the Johnston book is widely used as a graduate text. Dhrymes (1974), Malinvaud (1970), and Theil (1971) are generally thought of as "advanced" econometrics textbooks and are not suitable introductions. Goldberger and Duncan (1973) or Heise (1975) present this material from a more eclectic basis and thus may be more suitable than an econometric text. Hibbs (1977) outlines the relationship between structural equations and ARIMA approaches. While Hibbs's argument concerns only impact assessment models, it is general to any multivariate time series model. New advances in this field generally appear in such journals as *Econometrica, Journal of the American Statistical Association, Political Methodology,* and *Sociological Methods and Research.* The reader who wishes to keep abreast of new developments is advised to watch these journals.

NOTES TO CHAPTER 5

1. We use "multivariate" to mean time series models in which one output (dependent) series is explained as a function of several input (independent) series, that is,

$$Y_t = f(X_{1t}, X_{2t}, \ldots X_{nt}) + N_t.$$

Other authors may use this term to mean a time series model that has several output series, each a distinct measure of a single underlying concept.

2. The modeling strategy we develop in this chapter is only for the simplest case of *uni*directional cause, that is, for systems in which "x causes y":

$$x_t \dashrightarrow y_{t+b}.$$

*Bi*directional causal structures

$$x_t \dashleftarrow\dashrightarrow y_t$$

must be modeled through a more complicated strategy. The more complicated causal systems are beyond the scope of this introductory volume. The interested reader is directed to Granger and Newbold (1977: Chapter 7) for a discussion of advanced multivariate modeling topics.

3. We use the Thomas (1940) data only to demonstrate a strategy for multivariate ARIMA modeling. We do not generally endorse these methods for demographic accounting. Demography may be the only social science field in which sophisticated, proven mathematical models are routinely available. Keyfitz (1977: Chapter 8) gives a readable introduction to these models; see also Land (1980) for a comprehensive system of models.