PART VII

MULTIPLE INPUT - SINGLE OUTPUT MODELS

In many environmental systems, a single output or response variable is "caused" by one or more input or covariate series. For example, riverflows are caused by physical variables such as precipitation and temperature. To formally model the dynamic relationships which exist between a single output variable and the multiple input variables, a **transfer function-noise (TFN) model** can be employed. Qualitatively, a TFN model can be written as

$$\text{single output} = \text{dynamic component} + \text{noise}$$

where the dynamic component models the manner in which each input or covariate series affects the dynamic response of the output and the noise accounts for the stochastic disturbance in the system which cannot be modelled by the dynamic component. Because the behaviour of the output is dependent upon the way the input series affect the output over time, the overall TFN model is often referred to as a **dynamic model.**

An array of useful tools are available for **constructing TFN models** when following the identification, estimation and diagnostic check stages of model development. At the **identification** stage, a transfer function can be designed for mathematically describing the dynamic relationship over time which exists between each input and the output. An appropriate ARMA or ARIMA model can be identified as the autocorrelated noise component in the overall TFN model. Following the **estimation** of the model parameters and **checking** that the fitted model adequately describes the dynamic system being modelled, the calibrated TFN model can be used for applications such as forecasting and simulation. As is demonstrated in Part VII, the presence of the input variables in the model allows for a more accurate description of the physical system which in turn means more accurate forecasts (Chapter 18) and realistic simulated values can be produced by the model. Furthermore, TFN models can be built for either seasonal or nonseasonal time series for which the data points are evenly spaced over time.

In certain situations it may not be obvious if one physical variable causes another. For instance, do sunspot numbers cause riverflows? Consequently, in Chapter 16 statistical procedures are presented as **exploratory data analysis** tools for investigating possible **causal relationships** between two variables. When meaningful relationships are detected between two series using what is called the **residual cross-correlation function,** a TFN model can be constructed as a **confirmatory data analysis** procedure for rigorously describing the mathematical relationship between the input and output. In Chapter 17, comprehensive methods for constructing TFN models with a single output and multiple inputs are explained for both seasonal and nonseasonal time series using a number of interesting hydrological applications. Subsequent to calibrating a TFN model, the fitted model can be employed for **forecasting** by following the procedures of Chapter 18.

Sometimes the dynamic characteristics of a system may be changed by the imposition of one or more external interventions. For example, in environmental engineering, pollution abatement facilities are built to reduce the levels of certain pollutants. The stochastic effects upon the mean level of the output can be rigorously modelled using **intervention analysis.** As will be

thoroughly explained in Chapters 19 and 22, the intervention model is in fact a special type of TFN model. An extensive description of exploratory and confirmatory data analysis procedures for use in intervention analysis is presented in these chapters. Subsequent to calibrating a TFN model, the fitted model can be employed for **forecasting** by following the procedures of Chapter 18.

CHAPTER 16

CAUSALITY

16.1 INTRODUCTION

Is it possible to substantiate the claim of a Soviet hydrologist that yearly sunspot numbers have a significant affect upon the annual flows of the Volga River? What is the influence of temperature upon the price of wheat? In other words, how and when can one say that one phenomenon definitely causes another?

The foregoing kinds of questions have been baffling scientists for decades and previously some research had been carried out in an attempt to answer them. For example, Brillinger (1969) and Rodriguez-Iturbe and Yevjevich (1968) employed cross-spectral and other statistical methods to investigate relationships between natural time series. However, comprehensive statistical tools are now available to assist in solving causality problems and these useful techniques have yet to be applied to a large variety of environmental data sets. Consequently, the purpose of this chapter is to present flexible statistical procedures for formally answering causality questions and then to apply the methodologies to a wide range of natural time series. In particular, Granger's (1969) definition of *causality* is first defined and then it is explained how a *cross-correlation analysis of the residuals* from the stochastic models fitted to two series, can be employed to detect causal relationships (Pierce and Haugh, 1977). In the section on *applications*, a large number of interesting cross-correlation studies are carried out to detect possible causal relationships between many different phenomena. The time series studied include sunspot numbers, annual and monthly temperatures, seven annual riverflow series, Beveridge wheat price indices, and tree ring widths. Contrary to the suggestion of Smirnov (1969), it is found that annual sunspot numbers do not significantly affect the yearly flows of the Volga River in Russia. Other causality studies demonstrate that temperatures for certain months of the year can significantly affect the annual flows of rivers and also the price of wheat.

Upon detecting significant causal connections between two phenomena, the information from the cross-correlation analysis can be used to design a *transfer function-noise (TFN) model* to describe explicitly the mathematical relationship between the two data sets (Haugh and Box, 1977; Box and Jenkins, 1976, Ch. 11). In Chapter 17, the *construction* of TFN models which can handle a single output series and one or more input series, is thoroughly explained for the identification, estimation and diagnostic check stages of model development. Moreover, in Chapter 18, it is explained how one can calculate optimal *forecasts* using a TFN model. As would be expected, the information contained in the input or covariate series in a TFN model allows one to obtain more accurate forecasts for the output series. Finally, for the original presentation of the main contents of Chapter 16, the reader can refer to the paper of Hipel et al. (1985).

16.2 CAUSALITY

16.2.1 Definition

Wiener (1956) originally formulated a definition of causality between two time series, which is suitable for empirical detection and verification of meaningful relationships. More recently, Granger (1969) presented a formal definition of causality while Pierce and Haugh (1977) expanded upon the work of Granger (1969) and gave a comprehensive survey regarding research on causality in temporal systems. Other research which is related to Granger's (1969) definition of causality can be found by referring to the appropriate statistical literature (see for example Jenkins and Watts (1968), Haugh (1972, 1976), Haugh and Box (1977), and McLeod (1979)).

Granger (1969) defines *causality* between two time series in terms of predictability. A variable X causes another variable Y, with respect to a given universe or information set that includes X and Y, if the present Y can be better predicted by using past values of X than by not doing so (all other relevant information (including the past of Y) being used in either case). This definition of causality does not require the system to be linear but when it is, linear predictions are compared. To be more specific, let X_t and Y_t be two time series and let A_t for $t = 0,\pm1,\pm2,\ldots,$ be the given information set that includes at least X_t and Y_t . Allow $\bar{A}_t = \{A_s : s < t\}$, $\tilde{A}_t = \{A_s : s \le t\}$ and in a similar fashion define $\bar{X}_t$, $\tilde{X}_t$, $\bar{Y}_t$, and $\tilde{Y}_t$. Given the information set A_t, let $P_t(Y|A_t)$ be the minimum mean square error one step ahead predictor of Y_t and denote the resulting mean square error by $\sigma^2(Y|A_t)$. According to Granger (1969), X causes Y if

$$\sigma^2(Y|\bar{A}_t,\bar{X}_t) < \sigma^2(Y|\bar{A}_t) \qquad [16.2.1]$$

while X causes Y instantaneously if

$$\sigma^2(Y|\bar{A}_t,\tilde{X}_t) < \sigma^2(Y|\bar{A}_t) \qquad [16.2.2]$$

Causality from Y to X can be defined in the same way. *Feedback* occurs when X causes Y and Y also causes X.

16.2.2 Residual Cross-Correlation Function

To ascertain the type of causality relationship that exists between X and Y, the properties of the cross-correlations are examined for the prewhitened series. When *prewhitening* discrete time series such as X_t and Y_t , the first step is to consider suitable *transformations* to form the transformed series, x_t and y_t . The reasons for transforming the series include stabilizing the variance, improving the normality assumption, eliminating trends, removing seasonality, and getting rid of nonstationarity. The selected transformations should allow x and y to be related causally in the same manner as X and Y when considering Granger's (1969) definition of causality. In practice, causality is preserved by many of the common types of transformations. For example, often the given series may be transformed by the *Box-Cox transformation* (Box and Cox, 1964) given in [3.4.30] to remove non-normality and heteroscedasticity in the model residuals and following this the data may be differenced as in [4.3.3] to render the data stationary. As is explained in Section 13.2.2, when dealing with seasonal geophysical series the data may be

transformed using a Box-Cox transformation and subsequent to this the seasonality may be removed by invoking an appropriate *deseasonalization* technique. For instance, when modelling an average monthly riverflow series, often the series is first transformed by taking natural logarithms and then each data point is deseasonalized by subtracting out the monthly mean and dividing this by the monthly standard deviation as in [13.2.3]. A Box-Cox transformation such as natural logarithms should not alter causality relationships for series consisting of all positive values, since the manner in which one series affects the predictability of another will not be changed by a strictly monotonic transformation that preserves the same relative position of every data point in the series. Deseasonalizing each time series is equivalent to removing a periodic component to eliminate seasonality where the periodic component is ultimately due to hydrologic factors such as precipitation and temperature. Because the deseasonalization parameters are estimated from the historical data and are assumed to be the same in the future, the deseasonalization should not alter the causality relationship existing in the original series when entertaining Granger causality. However, the periodic portion still constitutes one of the components needed to form the overall seasonal series.

The second step in the prewhitening procedure is to fit appropriate stochastic models to the x_t and y_t series in order to obtain white noise residuals. For instance, when the transformed series are nonseasonal, it may be suitable to fit the ARMA model in [3.4.4] to x_t and y_t such that

$$\phi_x(B)(x_t - \mu_x) = \theta_x(B)u_t \qquad [16.2.3]$$

and

$$\phi_y(B)(y_t - \mu_y) = \theta_y(B)v_t \qquad [16.2.4]$$

where μ_x is the theoretical mean of the x_t series; B is the backward shift operator defined by $Bx_t = x_{t-1}$ and $B^k x_t = x_{t-k}$ where k is a positive integer; $\phi_x(B) = 1 - \phi_{x,1}B - \phi_{x,2}B^2 - \cdots - \phi_{x,p_x}B^{p_x}$, is the nonseasonal AR operator of order p_x such that the roots of the characteristic equation $\phi_x(B) = 0$ lie outside the unit circle for nonseasonal stationarity and the $\phi_{x,i}, i = 1,2,\cdots,p_x$, are the nonseasonal AR parameters; $\theta_x(B) = 1 - \theta_{x,1}B - \theta_{x,2}B^2 - \cdots - \theta_{x,q_x}B^{q_x}$, is the nonseasonal MA operator of order q_x such that the roots of $\theta_x(B) = 0$ lie outside the unit circle for invertibility and $\theta_{x,i}, i = 1,2,\cdots,q_x$, are the nonseasonal MA parameters; u_t is white noise (also called innovation or disturbance) that has a mean of zero and variance of σ_u^2; and similar definitions to μ_x, $\phi_x(B)$, $\theta_x(B)$, and u_t hold for μ_y, $\phi_y(B)$, $\theta_y(B)$, and v_t, respectively. As mentioned in Section 3.4.2, to indicate the orders of the AR and MA operators of the models in [16.2.3] or [16.2.4], the notation ARMA(p,q) is employed. Because of the linear nature of the operators in [16.2.3] and [16.2.4], this insures that u and v are causally related in the same way as X and Y. Of course, if the data were seasonal an appropriate seasonal model, such as one of those given in Chapters 12 to 15, could be used to prewhiten each series.

Subsequent to prewhitening of the time series, the *cross-correlation function (CCF)*, at lag k between the u_t and v_t series in [16.2.3] and [16.2.4], respectively, can be considered using

$$\rho_{uv}(k) = E[u_t v_{t+k}] / (E[u_t^2]E[v_t^2])^{1/2} \qquad [16.2.5]$$

Due to the form of [16.2.5], the values of the CCF can range from negative one to positive one. Unlike the ACF, the CCF is not usually symmetric about lag zero and therefore the properties of $\rho_{uv}(k)$ must be examined for $k = 0, \pm 1, \pm 2, \cdots$. In addition to reflecting the type of linear dependence between u and v and consequently between X and Y, $\rho_{uv}(k)$ gives the kind of causality relationship between these variables for linear systems.

As explained by Pierce and Haugh (1977), there are many possible types of causal interactions between X and Y which can be characterized by the properties of $\rho_{uv}(k)$. Using the results of Pierce and Haugh (1977, p. 276, Table 3), some of the important causal relationships are categorized according to the restrictions on $\rho_{uv}(k)$ in Table 16.2.1. Due to the findings of Price (1979) and also Pierce and Haugh (1979), any of the relationships in Table 16.2.1 which involve instantaneous causality are only valid when there is no feedback. The entries in Table 16.2.1 are self explanatory. For example, when there is unidirectional causality from X to Y, $\rho_{uv}(k) \neq 0$ for the $k > 0$, $\rho_{uv}(k) = 0$ for all $k < 0$, and $\rho_{uv}(0)$ may either be zero or else have some real non-zero value. For the case where Y does not cause X at all, there is no instantaneous causality between X and Y since $\rho_{uv}(0) = 0$.

When there is *feedback* between two variables, one variable can cause the other and vice versa. Although feedback is not too common in many natural problems, in economics, for example, inflation can cause unemployment which in turn affects inflation. As indicated in Table 16.2.1, $\rho_{uv}(k)$ is nonzero at both positive and negative lags if there is feedback between X and Y.

When checking for the type of causality between two given time series the estimated CCF of the model residuals must be examined to ascertain which values are significantly different from zero. Suppose that two sequences x_t and y_t are given for $t = 1,2, \ldots, n$. By utilizing [16.2.3] and [16.2.4] or other appropriate linear models, the two series can be prewhitened to obtain the estimated innovation series or residuals, $\hat{u}_t$ and $\hat{v}_t$, respectively. The residual CCF at lag k between $\hat{u}_t$ and $\hat{v}_t$ is estimated using

$$r_{\hat{u}\hat{v}}(k) = c_{\hat{u}\hat{v}}(k) / [c_{\hat{u}}(0)c_{\hat{v}}(0)]^{1/2} \qquad [16.2.6]$$

where

$$c_{\hat{u}\hat{v}}(k) = \begin{cases} n^{-1}\sum_{t=1}^{n-k} \hat{u}_t \hat{v}_{t+k} & k \geq 0 \\ n^{-1}\sum_{t=1-k}^{n} \hat{u}_t \hat{v}_{t+k} & k < 0 \end{cases}$$

is the *estimated cross-covariance function* at lag k between the residual series; $c_{\hat{u}}(0) = n^{-1}\sum_{t=1}^{n} \hat{u}_t^2$ is the sample variance of the $\hat{u}_t$ sequence; and $c_{\hat{v}}(0) = n^{-1}\sum_{t=1}^{n} \hat{v}_t^2$ is the estimated variance of the $\hat{v}_t$ series.

Table 16.2.1. Causal relationships between two variables.

RELATIONSHIPS	RESTRICTIONS ON $\rho_{uv}(k)$
X causes Y	$\rho_{uv}(k) \neq 0$ for *some* $k > 0$
Y causes X	$\rho_{uv}(k) \neq 0$ for *some* $k < 0$
Instantaneous Causality	$\rho_{uv}(0) \neq 0$
Feedback	$\rho_{uv}(k) \neq 0$ for *some* $k > 0$ and for some $k < 0$
X causes Y but not instantaneously	$\rho_{uv}(k) \neq 0$ for *some* $k > 0$ and $\rho_{uv}(0) = 0$
Y does not cause X	$\rho_{uv}(k) = 0$ for *all* $k < 0$
Y does not cause X at all	$\rho_{uv}(k) = 0$ for *all* $k \leq 0$
Unidirectional causality from X to Y	$\rho_{uv}(k) \neq 0$ for some $k > 0$ and $\rho_{uv}(k) = 0$ for either (a) all $k < 0$ or (b) all $k \leq 0$
X and Y are only related instantaneously	$\rho_{uv}(0) \neq 0$ and $\rho_{uv}(k) = 0$ for all $k \neq 0$
X and Y are independent	$\rho_{uv}(k) = 0$ for *all* k

The residual CCF can be plotted against lag k for $k \approx -n/4$ to $k \approx n/4$. In order to plot confidence limits, the distribution of the residual CCF must be known. Assuming that the x_t and y_t series are independent (so $\rho_{uv}(k) = 0$ for all k), Haugh (1972, 1976) shows that for large samples $\hat{r}_{\hat{u}\hat{v}}(k)$ is normally independently distributed with a mean of zero and variance of $1/n$. Consequently, to obtain the approximate 95% confidence limits a line equal to $1.96\, n^{-1/2}$ can be plotted above and below the zero level for the residual CCF. McLeod (1979) presents the asymptotic distribution of the residual CCF for the general case where the x_t and y_t series do not have to be independent of each other and, consequently, more accurate confidence limits can be obtained by utilizing his results.

One reason why the residual CCF is examined rather than the CCF for the x_t and y_t series, is that it is much easier to interpret the results from a plot of $r_{\hat{u}\hat{v}}(k)$. This is because when both the x_t and y_t series are autocorrelated, the estimates of the CCF for x_t and y_t can have high variances and the estimates at different lags can be highly correlated with one another (Bartlett, 1935). In other words, the distribution of the estimated CCF for x_t and y_t is more complex than the distribution of $r_{\hat{u}\hat{v}}(k)$. Monte Carlo studies executed by Stedinger (1981), demonstrate the advantages of prewhitening two series before calculating their CCF. Additionally, from an intuitive point of view it makes sense to examine the residual CCF. Certainly, if the *driving mechanisms* or residuals of two series are significantly correlated, then meaningful relationships would exist between the original series.

From an examination of the residual CCF, the type of relationship existing between X and Y can be ascertained by referring to the results in Table 16.2.1. Suppose, for example, the X variable is precipitation and the Y variable is riverflow. From a physical understanding of hydrology, it is obvious that precipitation causes riverflow. This knowledge would be mirrored in a plot of the residual CCF for these two series. For $k \geq 0$ there would be at least one value of $r_{\hat{u}\hat{v}}(k)$ which is significantly different from zero. However, all values of the residual CCF for $k < 0$ would not be significantly different from zero. In situations where the type of causality between two series is not known (for instance, do sunspots cause riverflows), an examination of the residual CCF can provide valuable insight into the problem (see Section 16.3).

Formal tests of significance may also be derived when examining causal relationships (see, for example, McLeod (1979) and Pierce (1977)). Suppose that it is known a priori that Y does not cause X so that $\rho_{uv}(k) = 0$ for $k < 0$ (for instance riverflows do not cause precipitation). Consequently, one may wish to test the null hypothesis that X does not cause Y and hence $\rho_{uv}(k) = 0$ for $k = 1,2,\ldots,L$, where L is a suitably chosen lag such that after L time periods it would be expected there would not be a relationship between the x_t and y_t series. The statistic

$$Q_L = n^2 \sum_{k=0}^{L} \frac{r_{uv}^2(k)}{n-k} \qquad [16.2.7]$$

is then approximately distributed as $\chi^2(L+1)$. A significantly large value for Q_L would mean that the hypothesis should be rejected and, therefore, X causes Y.

A limitation of the methods explained in this section is that they are only useful when describing the relationships between two time series. If three or more time series are mutually related, then analyzing them only two at a time may lead to finding spurious relationships among them. Consequently, further research on causality between linear systems is still required. Nevertheless, as shown by the applications in the next section, in many situations bivariate causality studies are of direct interest to the practitioner.

When sufficient data are available, an alternative approach for detecting causal linear relationships is to work in the frequency domain rather than the time domain by employing the *coherence function*. An advantage of this procedure is that it can be extended for handling multiple-input and multiple-output systems (Bendat and Piersol, 1980).

16.3 APPLICATIONS

16.3.1 Data

For a long time, hydrologists have been attempting to ascertain the impact of exogenous forces upon specific hydrological and meteorological phenomena. In many instances, the great complexity of the physical problem at hand has precluded the development of suitable physical or statistical models to describe realistically the situation. Consequently, a wide range of phenomena are now studied in order to detect and model meaningful dynamic relationships.

The time series investigated are listed in Table 16.3.1. Except for monthly temperatures from the English Midlands, all of the data sets consist of annual values. The sunspot numbers, annual and monthly temperatures, seven riverflow series in m^3/s where each average yearly flow is calculated for the water year from October 1st of one year to September 30th of the next year, and Beveridge wheat price indices, are obtained from articles by Waldemeier (1961), Manley (1953, pp. 255-260), Yevjevich (1963), and Beveridge (1921), respectively. The tree ring widths given in units of 0.01 mm are for Bristlecone Pine and were received directly from V.C. LaMarche of the Laboratory of Tree Ring Research, University of Arizona, Tuscon, Arizona. The length and accuracy of the tree ring series make it a valuable asset in cross-correlation studies for determining the effects of external variables such as temperature and the amount of sunlight. The reason for considering the Beveridge wheat price index data is that the series could be closely related to climatic conditions and, therefore, may be of interest to hydrologists and climatologists. For example, during years when the weather is not suitable for abundant grain production the price of wheat may be quite high.

16.3.2 Prewhitening

When checking for causality, the time series under investigation must first be prewhitened. Table 16.3.2 describes the types of models which were used to prewhiten the series from Table 16.3.1. In all cases, the models were determined by following the three stages of model construction in conjunction with the AIC (see Section 6.3) and in some instances the most appropriate models are constrained models for which some of the model parameters are omitted. For example, as explained in Sections 3.4.4 and 5.4.3, the best ARMA model for the sunspot series is a constrained ARMA (9,0) model where ϕ_3 to ϕ_8 are left out of the model and the original data are transformed by a square root transformation for which $\lambda = 0.5$ in [3.4.30] where x_t replaces z_t, and $c = 1$ due to some zero values in the series. Using the format in [16.2.3] or [3.4.4], the estimated sunspot model is written in difference equation form in [6.4.4] as

$$(1 - 1.245B + 0.524B^2 - 0.192B^9)(x_t - 10.673) = u_t \qquad [16.3.1]$$

where

$$x_t = (1/0.5)[(X_t + 1.0)^{0.5} - 1.0]$$

Notice for the Beveridge wheat price indices that the data are transformed using a natural logarithmic transformation where $\lambda = 0$ and $c = 0$ in [3.4.30]. The transformed data are then differenced once to remove nonstationarity by using [4.3.3] which is written as

Table 16.3.1. Time series used in the causality studies.

DATA SET	LOCATION	PERIOD	LENGTH
Sunspots	Sun	1700-1960	261
Annual Temperatures	English Midlands	1723-1970	248
12 Monthly Temperature Sequences	English Midlands	1723-1970	248 per month
St. Lawrence River	Ogdensburg, New York, USA	1860-1957	97
Volga River	Gorkii, USSR	1877-1935	58
Neumunas River	Smalininkai, USSR	1811-1943	132
Rhine River	Basle, Switzerland	1807-1957	150
Gota River	Sjotorp-Vanersburg, Sweden	1807-1957	150
Danube River	Orshava, Romania	1837-1957	120
Mississippi River	St. Louis, Missouri, USA	1861-1957	96
Beveridge Wheat Price Index	England	1500-1869	370
Tree Ring Widths	Campito Mountain, California, USA	1500-1969	470

$$y_t = \ln Y_{t+1} - \ln Y_t$$

for $t = 1,2,3,\ldots,n-1$. Following this, identification results explained in detail in Section 4.3.3 reveal that an ARMA (8,1) without ϕ_3 to ϕ_7 should be fitted to y_t where the estimated model is written using the notation of [16.2.4] as

$$(1 - 0.729B + 0.364B^2 + 0.119B^8)y_t = (1 - 0.783B)v_t \qquad [16.3.2]$$

The reader should bear in mind that only the family of ARMA models are entertained when selecting the best model to describe each data set in Table 16.3.2. In certain instances, it may be appropriate to also consider other types of models. For example, Akaike (1978) noted that because of the nature of sunspot activity a model based on some physical consideration of the generating mechanism may produce a better fit to the sunspot series than an ARMA model. For

Table 16.3.2. Models used to get residuals for the CCF studies.

DATA SET	ARMA (p,q) MODEL	$\hat{u}_t$	$\hat{v}_t$
Sunspots	(9,0) without ϕ_3 to ϕ_8, $\lambda = 0.5$ and $c = 1$	*	
Annual Temperature	(2,0) without ϕ_1	*	*
12 Monthly Temperature Sequences	(0,0) for all months	*	*
St. Lawrence River	(3,0) without ϕ_2 as in [6.4.2]		*
Volga River	(0,0)		*
Neumunas River	(0,1), $\lambda = 0$		*
Rhine River	(0,0)		*
Gota River	(2,0)		*
Danube River	(0,0)		*
Mississippi River	(0,1)		*
Beveridge Wheat Price Indices	(8,1) without ϕ_3 to ϕ_7, $\lambda = 0$, and series is differenced once		*
Tree Ring Widths	(4,0) without ϕ_2		*

modelling the sunspot series, Granger and Andersen (1978) utilized a bilinear model. Whatever the case, for each time series in Table 16.3.2 extensive diagnostic checking was executed to ensure that the best ARMA model was ultimately chosen.

16.3.3 Causality Studies

Following prewhitening, [16.2.6] is employed to calculate the residual CCF for two specified residual series. In the third and fourth columns of Table 16.3.2, *'s indicate when the residuals of a given series are used as $\hat{u}_t$ and/or $\hat{v}_t$, respectively, in [16.2.6]. Whenever two series are cross-correlated, the residual values are used for the time period during which the $\hat{u}_t$ and $\hat{v}_t$ data sets overlap. The sunspot residuals could possibly affect all the other series in Table 16.3.2 and, therefore, the sunspot residuals are separately cross-correlated with each of the remaining

series in Table 16.3.2. For the monthly temperature data, each monthly sequence is considered as a separate sample when the residual CCF is calculated between the sunspot series as $\hat{u}_t$ and a given monthly temperature data set as $\hat{v}_t$. However, it is also possible that temperature can affect the phenomena listed below the temperature series in Table 16.3.2. For example, April temperatures may influence tree ring growth in the Northern Hemisphere since the month of April is when the growing season begins after the winter months. Consequently, the residual series for the annual temperature data set and the 12 monthly temperature sequences are each cross-correlated with the residual series of each of the data sets given below the temperatures.

In many situations, it may not be known whether or not one phenomenon definitely causes another. Although the direction of suspected causality is often known a priori due to a physical understanding of the problem, proper statistical methods must be employed to ascertain if the available evidence confirms or denies the presence of a significant causal relationship. Consider, for example, determining whether or not sunspots and riverflows are causally related. Obviously, it is only physically possible for sunspots to cause riverflows and not vice versa. Based upon ad hoc graphical procedures comparing annual flows of the Volga River in the USSR with yearly sunspot numbers, Smirnov (1969) postulated that sunspots unequivocally affect riverflows. However, when the residual CCF is used to detect scientifically causality, the results do not support Smirnov's strong claim. In Figure 16.3.1, the residual CCF along with the 95% confidence limits are presented for the residuals from the ARMA model fitted to the annual flows of the Volga River at Gorkii, USSR, and the residuals from the ARMA model fitted to the annual sunspot numbers (refer to Table 16.3.1 for a description of these data sets and to Table 16.3.2 for the types of models fitted to the two time series). As can be seen, there are no significant values of the residual CCF at lag zero and the smaller positive lags. If sunspot activity did affect the Volga flows, it would be expected that this would happen well within the time span of a few years. Therefore, the absence of significant values of the CCF from lags 0 to 2 or 3 indicates that the current information does not support the hypothesis that sunspots cause the Volga riverflows. The slightly large magnitudes at lags 5 and 11 are probably due to chance. Nevertheless, it is possible, but highly unlikely, that the value at lag 11 could be due to the fact that the best ARMA model could not completely remove the periodicity present in the sunspot series. Previously, Granger (1957) found that the periodicity of sunspot data follows a uniform distribution with a mean of about 11 years. However, the constrained sunspot model in [16.3.1] is designed to account for this. Moreover, the fitted model is subjected to rigorous diagnostic checks to demonstrate that the periodicity is not present in the model residuals and none of the values of the residual ACF are significantly different from zero, even at lag 11.

Besides the annual flows of the Volga River, no meaningful causality relationships are detected when the sunspot residuals are separately cross-correlated with the other riverflow residuals and also the remaining residuals series which are considered as $\hat{v}_t$ in Table 16.3.2. As emphasized earlier, if correct statistical procedures are not followed it would not be possible to reach the aforesaid conclusions regarding the causality relationships between the sunspots and the other phenomena. For example, in Figure 16.3.2 it can be seen that the values of the CCF calculated for the given annual sunspot and Gota riverflows series are large in magnitude at negative and positive lags (recall that the 95% confidence limits in Figure 16.3.2 are derived for independent series). Furthermore, the cyclic nature of the sunspot data is portrayed by the sinusoidal characteristics in the graph. To uncover the underlying causal relationship between the series it is necessary to examine the residual CCF. As just noted, the residual CCF for the

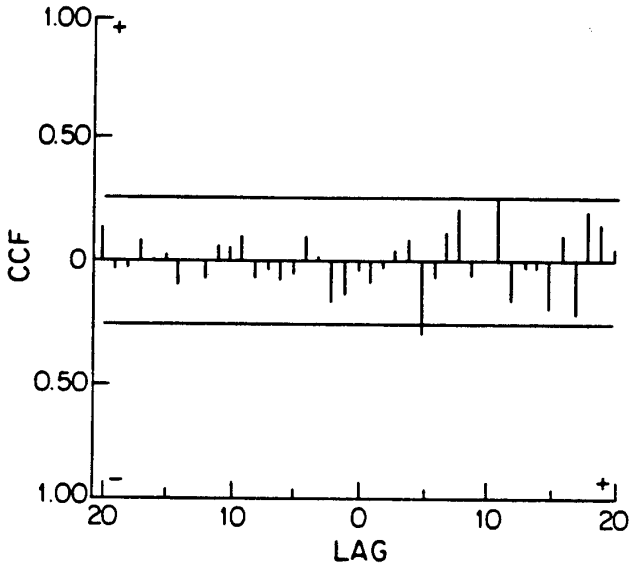

Figure 16.3.1. Residual CCF for the sunspot numbers and the Volga riverflows.

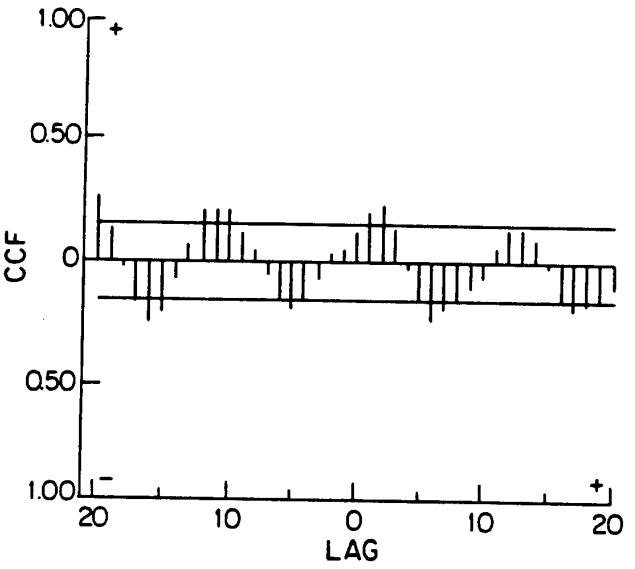

Figure 16.3.2. CCF for the given sunspot numbers and the Gota riverflows.

sunspot and Gota River series does not reveal that sunspot numbers affect the flows of the Gota River.

For the case where the $\hat{u}_t$ sequence, as represented by the residuals of the annual temperature data, is cross-correlated with each of the last nine $\hat{v}_t$ series in Table 16.3.2, no meaningful relationships are found. Nonetheless, some significant values of the residual CCF are discovered when each monthly temperature series is cross-correlated separately with each residual sequence for the riverflows and also the Beveridge wheat price indices. Table 16.3.3 shows the lags at which the residual CCF possesses large values when $\hat{u}_t$ is a designated monthly temperature series and $\hat{v}_t$ is either the annual riverflow or Beveridge wheat price index residuals. Since it would be expected from a physical viewpoint that a given monthly temperature data set would have the most effect upon the other time series in the current year or perhaps one or two years into the future, large values of the residual CCF are only indicated in Table 16.3.3 when they occur at lags 0 to 2. As an illustrative example, consider the graph of the residual CCF for the August temperatures and the Gota River residuals which is shown in Figure 16.3.3. As can be seen, the large negative correlation at lag zero extends well beyond the 95% confidence limits. When the Q_L statistic in [16.2.7] is calculated for lags 0 to 2, the estimated value for the residual CCF in Figure 16.3.3 is 26.6. Because this value is much larger than the tabulated $\chi^2(3)$ value of 7.8 for the 5% significance level, one must reject the null hypothesis that the August temperatures do not affect the annual flows of the Gota River.

It would be expected that temperature could significantly affect tree ring growth. As noted by La Marche (1974), because Bristlecone Pines are located at the upper treeline on mountains, temperature is a key factor in controlling growth. However, this growth would only be sensitive to local temperature conditions and the temperatures recorded in the English Midlands are probably not representative of the temperatures at Campito Mountain in California. If local temperatures were available, the residual CCF between the local temperatures and tree ring widths could be calculated to ascertain the type of causality which is present.

16.4 CONCLUSIONS

Comprehensive procedures are now available for detecting causal relationships between two time series. The results of Table 16.3.3 demonstrate that monthly temperatures can significantly affect annual riverflows and also the price of wheat. However, no meaningful links are found between the annual sunspot numbers and the other phenomena designated by $\hat{v}_t$ in Table 16.3.2. In particular, the statistical evidence from Figure 16.3.1 cannot support the claim (Smirnov, 1969) that sunspots significantly affect the annual flows of the Volga River. While some of the findings of Section 16.3 may be somewhat interesting, it is also informative to note the types of results that Pierce (1977) discovered in the field of econometrics. Using residual CCF studies, Pierce found that numerous economic variables which were generally regarded by economists as being strongly interrelated were in fact independent or else only weakly correlated. These conclusions are of course based upon the information included in the time series which Pierce analyzed. If it were possible to improve the design of the data collection scheme for a causality study, this would of course enhance the conclusions reached at the analysis stage. Certainly, it is necessary that a sufficiently wide range of values of the relevant variables appear in the sample in order to increase the probability of detecting relationships which do actually exist in the real world. However, as is the case in economics and also in the natural sciences, the experimenter

Table 16.3.3. Residual CCF results of monthly temperature and other series.

$\hat{v}_t$	MONTHLY TEMPERATURES $\hat{u}_t$	LAGS FOR LARGER VALUES OF RESIDUAL CCF
St. Lawrence River	February	1
Volga River	February April May July	2 2 0 and 1 2
Neumunas River	May July December	0 2 2
Rhine River	October	0
Gota River	June July August September	0 0 0 0
Danube River	September October	0 0
Mississippi River	December	1
Beveridge Wheat Price Index	February November December	0 0 0

has little control over the phenomena which produce the observations and must therefore be content with the data that can be realistically collected. Perhaps *God* may have a switch that can greatly vary the number of sunspots that appear on the sun so that *mortal man* can assess beyond a shadow of a doubt whether or not sunspots can significantly affect riverflows.

Given the available information, it is essential that the data be properly analyzed. For example, if a sample CCF were calculated for the x_t and y_t series, spurious correlations may seem to indicate that the variables are causally related (see Figure 16.3.2, for instance).

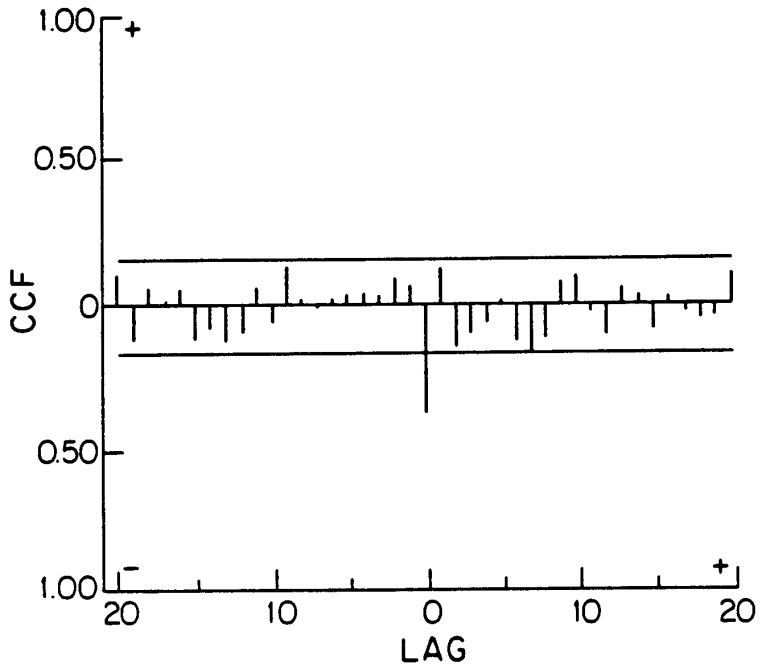

Figure 16.3.3. Residual CCF for the August temperatures and the Gota riverflows.

However, an examination of the residual CCF for the two series may clearly reveal that based upon the given data no meaningful relationships do in fact exist between the two phenomena. It is of course possible that no significant correlations may appear in the residual CCF even though two variables are functionally related. This is because correlation is only a measure of linear association and nonlinear relationships that contain no linear component, may be missed. To minimize the occurrence of this type of error, the fitted ARMA models that are used to prewhiten the series are subjected to stringent diagnostic checks. In this way, any problems that arise due to the use of these linear models will be detected prior to examining the residual CCF.

Subsequent to the revelation of causality using the residual CCF, a dynamic model can be built to describe mathematically the formal connections between the x_t and y_t series. In most hydrological and other geophysical applications, usually one variable causes another and there is no feedback. For instance, precipitation causes riverflows and this unidirectional causality cannot be reversed. In terms of the residual CCF, for unidirectional causality from X to Y, the residual CCF is nonzero at one or more lags for $k > 0$, $\rho_{uv}(0)$ may be either zero or have some

nonzero value, and the value of the residual CCF at all negative lags must be zero (see Table 16.2.1). To describe mathematically the formal connections between the x_t and y_t series, the TFN model described in the next chapter constitutes a flexible dynamic model which can be utilized. An inherent advantage of TFN models is that well developed methodologies are available for use at the three stages of model construction. For instance, at the identification step the results of the residual CCF study that detected the causal relationship in the first place, can be utilized to design the dynamic model (Haugh and Box, 1977). When the y_t series has been altered by one or more external interventions, then intervention components can be introduced into the TFN model to account for possible changes in the mean level (see Chapters 19 and 22).

When there is feedback between X and Y, Table 16.2.1 shows that $\rho_{uv}(k)$ is nonzero at both positive and negative lags. The multivariate models in Chapters 20 and 21 are the type of dynamic models which can be used to model rigorously the dynamical characteristics of the feedback. Nevertheless, the reader should keep in mind that TFN models are used much more than multivariate models in hydrology and environmental engineering, since most natural systems do not possess feedback. Consequently, TFN models are described in more detail than multivariate models within this text.

PROBLEMS

16.1 Granger causality is defined in Section 16.2.1. Explain at least one other way in which scientists define causality between two phenomena. You may, for instance, wish to examine the path analysis procedure for studying relationships among variables which Kaplan and Thode (1981) apply to water resources data. Another procedure to consider for investigating causality is the coherence function (Bendat and Piersol, 1980) mentioned at the end of Section 6.2.2. Compare the residual CCF method to the other techniques for causality detection in terms of similarities and differences in the basic procedures, as well as advantages and drawbacks.

16.2 As is illustrated in Figure 16.3.2, spurious relationships between two variables can be found by improperly comparing the two variables. One way to overcoming spurious statistical connections between two time series is to employ the residual CCF approach of section 16.2.2. Find a statistical study in a field which is of interest to you where you think that scientists may have discovered spurious causal connections between two variables which do not really exist. Point out where the authors followed an improper procedure and explain how it can be corrected.

16.3 Select two annual time series for which you suspect one variable causes the other. For instance, you may have a representative yearly regional precipitation series which causes average annual riverflows in a river falling within the region. For these two data sets, carry out the following tasks:

(a) Prewhiten each series by fitting an ARMA to the series and thereby obtaining the model residuals.

(b) Calculate and plot the residual CCF for the two series along with the 95% confidence limits.

(c) Describe the stochastic relationships that you find in part (b). Explain why your findings make sense by linking them with the physical characteristics of the system under study. If, for example, you are examining a hydrological system, include aspects of the hydrological cycle described in Section 1.5.2 in your explanation.

16.4 Design and calibrate a TFN model for formally describing the dynamic relationships between the two time series examined in problem 16.3. Perform diagnostic checks to ensure that your fitted model adequately links the two data sets.

16.5 Choose a set of 6 to 10 time series in a field which you are working. Following the approach employed for the time series in Section 16.3, carry out a systematic causality study among your data sets. Comment upon the interesting results that you discover.

16.6 Select two seasonal time series, such as average monthly precipitation and river-flows, for which it makes sense to remove the seasonality by employing a suitable deseasonalization technique from Section 13.2.2. After fitting an ARMA model to each of the deseasonalized series, carry out a causality study to examine the relationships among these series.

REFERENCES

DATA SETS

Beveridge, W. H. (1921). Weather and harvest cycles. *Economics Journal*, 31:429-552.

La Marche, Jr., V. C. (1974). Paleoclimactic inferences from long tree-ring records. *Science*, 183:1042-1048.

Manley, G. (1953). The mean temperatures of Central England (1698-1952). *Quarterly Journal of the Royal Meteorological Society*, 79:242-261.

Waldmeier, M. (1961). *The Sunspot Activity in the Years 1610-1960*. Schulthas and Company, Zurich, Switzerland.

Yevjevich, V. M. (1963). Fluctuation of wet and dry years, 1, Research data assembly and mathematical models. Hydrology Paper No. 1, Colorado State University, Fort Collins, Colorado.

CAUSALITY

Bartlett, M. S. (1935). *Stochastic Processes*. Cambridge University Press, London.

Bendat, J. S. and Piersol, A. G. (1980). *Engineering Applications of Correlation and Spectral Analysis*. Wiley, New York.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26:211-252.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424-438.

Haugh, L. D. (1972). *The Identification of Time Series Interrelationships with Special Reference to Dynamic Regression*. Ph.D. thesis, Department of Statistics, University of Wisconsin, Madison, Wisconsin.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71(354):378-385.

Hipel, K. W., McLeod, A. I. and Li, W. K. (1985). Causal and dynamic relationships between natural phenomena. In Anderson, O. D., Ord, J. K. and Robinson, E. A., Editors, *Time Series Analysis: Theory and Practice*, pages 13-34. North-Holland, Amsterdam.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

Kaplan, E. and Thode Jr., H. C. (1981). Water quality, energy and socioeconomics: path analyses for studies of causality. *Water Resources Research*, 17(3):491-503.

McLeod, A. I. (1979). Distribution of the residual cross-correlation in univariate ARMA time series models. *Journal of the American Statistical Association*, 74(368):849-855.

Pierce, D. A. (1977). Relationships - and the lack thereof - between economic time series, with special reference to money and interest rates. *Journal of the American Statistical Association*, 72(357):11-21.

Pierce, D. A. and Haugh, L. D. (1977). Causality in temporal systems. *Journal of Econometrics*, 5:265-293.

Pierce, D. A. and Haugh, L. D. (1979). The characterization of instantaneous causality, A comment. *Journal of Econometrics*, 10:257-259.

Price, J. M. (1979). The characterization of instantaneous causality, A correction. *Journal of Econometrics*, 10:253-256.

Stedinger, J. R. (1981). Estimating correlations in multivariate streamflow models. *Water Resources Research*, 17(1):200-208.

Wiener, N. (1956). The theory of prediction. In Beckenback, E., Editor, *Modern Mathematics for Engineers*, Series 1, Chapter 8, McGraw-Hill, New York.

SUNSPOT NUMBERS

Akaike, H. (1978). On the likelihood of a time series model. Paper presented at the Institute of Statisticians 1978 Conference on Time Series Analysis and Forecasting, Cambridge University.

Brillinger, D. R. (1969). A search for a relationship between monthly sunspot numbers and certain climatic series. *Bulletin of the International Statistical Institute*, 43:293-307.

Granger, C. W. J. (1957). A statistical model for sunspot activity. *Astrophysics Journal*, 126:152-158.

Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models.* Vandenhoeck and Ruprecht, Gottingen.

Rodriquez-Iturbe, I. and Yevjevich, V. M. (1968). The investigation of relationship between hydrologic time series and sunspot numbers, hydrology paper no. 26. Technical report, Colorado State University, Fort Collins, Colorado.

Smirnov, N. P. (1969). Causes of long-period streamflow fluctuations. *Bulletin of the All-Union Geographic Society (Izvestiya VGO),* 101(5):443-440.

TRANSFER FUNCTION NOISE MODELLING

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, Oakland, California, revised edition.

Haugh, L. D. and Box, G. E. P. (1977). Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association,* 72(357):121-130.