Time Series Modeling for Quality Control in Clinical Chemistry
Layth C. Alwan1 and Michael G. Bissel2,3

Autocorrelation of clinical chemistry quality-control (Q/C) measurements causes one of the basic assumptions underlying the use of Levey–Jennings control charts to be violated and performance to be degraded. This is the requirement that the observations be statistically independent. We present a proposal for a new approach to statistical quality control that removes this difficulty. We propose to replace the current single control chart of raw Q/C data with two charts: (a) a common cause chart, representing a Box–Jenkins ARIMA time-series model of any underlying persisting nonrandomness in the process, and (b) a special cause chart of the residuals from the above model, which, being free of such persisting nonrandomness, fulfills the criteria for use of the standard Levey–Jennings plotting format and standard control rules. We provide a comparison of the performance of our proposed approach with that of current practice.

One of the methods most commonly used to monitor the quality of a given process is the statistical control chart, as described initially by Walter Shewhart in 1931 (1). Shewhart’s method involved taking quality measurements on the products of an industrial process. Means and ranges of subsamples of these measurements, drawn successively over time, were then plotted. In the industrial use of these Shewhart mean and range charts, the typical subsample size ranged from two to six.

In 1950, Levey and Jennings (2) introduced Shewhart’s method into clinical laboratories and based their chart on quality measurements performed on specially prepared control materials, repetitively analyzed over time. Because this now represented quality control of a measurement process alone and not a product, these materials, of presumed constant composition, served as the benchmark. Today, of course, clinical laboratory quality control is nearly universally based on the use of such control materials. The original chart of Levey and Jennings was a Shewhart chart with subsample size of two. This has subsequently been reduced to one in most practices, and this form of statistical quality control has become an industry standard.

The purpose of the control chart is to identify statistically significant departures from a state of statistical control. Statistical control is identified with a process reflecting random analytical variation only; in mathematical statistics this is referred to as a “random process.” A random process is one that generates values that are all statistically independent of one another and represent samplings of the same underlying probability distribution. Such a process is said to be “independent and identically distributed” or “IID.”

Control charts usually make the further assumption that the underlying distribution is normal. So, in control-chart theory, statistical control is associated with “IID normal.” Departures from statistical control are usually assumed to be local rather than pervasive: that is, they are reflected in the behavior of individual observations (outliers) or short sequences (runs). Such departures are interpreted as being due to “assignable” or “special” causes. [“Assignable cause” is a term introduced by Shewhart; “special cause” is an alternative term suggested by Deming in 1982 (3).] The interpretation of local departures is usually based on the application of statistically derived criteria known as control rules. An example of a set of control rules in widespread use is the multirule method of Westgard et al. (4).

In practice, it may be difficult to attain statistical control. Indeed, in many applications, it appears that statistical control is never achieved, except possibly as a very crude approximation. Our examination of published and unpublished data from actual applications suggests that systematic nonrandom patterns—reflecting common causes—are present throughout data. [The term “common cause” was suggested by Deming (3).] In particular, substantial autocorrelation in the data is very common. In a classic paper on measurement processes, Eisenhart (5) notes:

Experience shows that in the case of measurement processes the ideal of strict statistical control that Shewhart prescribes is usually very difficult to attain, just as in the case of industrial production processes. Indeed, many measurement processes do not and, it would seem, cannot be made to conform to this ideal of producing successive measurements of a single quantity that can be “observed values” of independent identically distributed random variables.

When systematic nonrandom patterns are present, it is difficult to locate departures from control for which there are special causes because the effects of special causes may be masked by the systematic patterns. The detrimental effects of systematic behavior were observed by Montgomery (6):

The most important of these assumptions (required for control charts) is that of independence of the observations, for conventional control charts do not work well if the quality characteristic exhibits correlation over time. Specifically, these control charts will frequently give misleading results if the data are correlated. Unfortunately, the assumption of uncorrelated or independent observation is not even approximately satisfied in some manufacturing processes. Examples include chemical processes where the product characteristics are often highly correlated from batch to batch, or in automated test and inspection procedures, where every quality characteristic is measured on every unit in unit time order of production. [emphasis ours]

It should be pointed out that the conventional control charts considered by Montgomery include the Shewhart control chart, as well as others such as cusum (cumulative sum) charts and the exponentially weighted moving average (EWMa) chart. Johnson and Bagshaw (7) note in the case of the cusum method:

Our primary conclusion is that the cusum test is not robust with respect to departures from independence. The use of cusum tests is now widespread and the presence of serial correlation common so that attention should be drawn to the seriousness of this lack of robustness.
In addition to the potential masking of special causes, there is another equally hazardous side effect of systematic nonrandom behavior. Even if a systematic nonrandom process is void of special causes, the process tends to be viewed as a sequence of isolated episodes, each with its own ad hoc explanation and associated hints as to appropriate intervention. This view is often misleading and such preoccupation with nonexistent special causes diverts attention from the underlying common cause.

In light of the problem with serial correlation or autocorrelation, some investigators have assumed that autocorrelated behavior follows an underlying stochastic process other than the usual one, i.e., IID, and then have taken this into account in an explicit readjustment of control limits [see Alt and Deutsch (8)]. In our opinion, this approach lacks appeal, owing to its inflexibility to potential systematic behaviors other than the one assumed.

The purpose of this paper is to add a new dimension to statistical quality control by showing how time-series modeling can be used in process control. The basic theme of our proposal is described by Montgomery (6); "to directly model the correlative structure, use that model to remove the autocorrelation from the data, and apply control charts to the residuals." A more in-depth treatment of the modeling approach can be found in Alwan and Roberts (9). The methodology described here can (a) offer improved performance in terms of a lower false alarm rate and higher analytic error detection rate; (b) be easily adapted and integrated into existing quality-control practice, using existing control charts and control rules; (c) be easily implemented on microcomputers with readily available software packages; and (d) offer the opportunity of better understanding of the process, i.e., improving precision.

Materials and Methods

Time is an essential attribute of quality measurement series. Generally, a control chart can be looked at as a display of quality measurements over time. The time classification is always laid out on the horizontal axis, with the data scale on the vertical axis. When data occur in a time-ordered sequence, the data form a time series, denoted by \( x_t \). An observed series \( x_t \) can be considered as a realization from some underlying stochastic process. As mentioned earlier, the standard assumption of control charts is that data in statistical control follow a very special process, an IID process. Under statistical control, observations \( x_t \) are generated from the process

\[
x_t = \mu + \epsilon_t
\]

where \( \mu \) is a constant and \( \epsilon_t \) is a sequence of uncorrelated errors each normally distributed with mean 0 and variance \( \sigma^2 \). The model represented by equation 1 is also referred to as a "constant mean model," because series that follow this model are characterized by random variation around a constant mean \( \mu \). This assumption is rarely met in practice. Usually an observation on the quality characteristic in a given time period is correlated with (statistically dependent on) the value of the same variable in the previous time period(s). The term "autocorrelation" is used to describe such a situation.

A natural solution to this difficulty is to model the systematic behavior statistically by time-series models that go beyond the simple benchmark of the constant mean model or IID. Thus, models that capture the correlative structure have to be considered. In particular, we suggest a special class of time-series models called the autoregressive integrated moving average (ARIMA) models of Box and Jenkins (10), which can represent a variety of different correlative structures. For detailed elaboration on ARIMA models, the reader should review the Appendix.

Analogous to regression analysis, successful time-series modeling decomposes the original series as follows:

1. Actual observation = fitted value + residual (retrospective, i.e., "control chart as a judgement").
2. Actual observation = predicted value + error (prospective, i.e., "control chart as a process").

If the model is well fitted, the resulting residuals will exhibit no or little autocorrelation and be distributed approximately normal with mean 0 and constant variance; i.e., the residuals are IID normal. Except for local departures, they will be in control.

So, for purpose of process control, the conventional charts for actual values can be replaced by two charts:

1. Common cause chart: chart of fitted values (or one-step forecasts) based on the ARIMA model, without computation of control limits. This plot can be regarded as a series of point estimates of the conditional mean of the process—our best current guess, based on past data, of the location of the underlying process.
2. Special cause chart: chart of residuals (or one-step forecast errors) from the fitted ARIMA model. This chart can be used in traditional ways to detect any special causes, without the danger of confounding of special causes with common causes.

The common cause chart gives a view of the time-varying mean value of the process and its evolution through time. It allows us to see the systematic nonrandom behavior that is present throughout the data. This behavior may aid in real-time control or in better understanding of underlying common causes that make the process behave as it does. An important example of real-time control is determining a policy of when and how much should the process be adjusted so as to maintain a performance criterion.

The special cause chart allows us to identify any possible assignable causes that may have been potentially obscured by the systematic nonrandom behavior. The IID model guides the interpretation of this chart of residuals. Thus, in the case of the special cause chart, control-chart theory is appropriate and justified. In particular, we advocate the use of the multirule Shewhart chart based on its extensive use in clinical chemistry laboratories.

The data for the experimental portion of this study are routine quality-control results on 26 analytes generated by the General Clinical Chemistry Laboratory at the University of Chicago Hospitals from October 1986 through February 1987. The laboratory runs a 24-h operation, providing urgent ("stat"), routine, and pediatric automated chemistries in a 450-bed acute- and general-care urban setting, and performing approximately 1.5 to 2.0 million tests per year. The instrumentation consists of two Kodak Ektachem analyzers one-way interfaced via DEC 1103 computer to a laboratory information system run on a VAX 780 mainframe. Clinical chemistry results are transmitted and archived via Sunquest software, on which all quality-control data are routinely entered, logged, and plotted monthly in standard Levey–Jennings format with Westgard multirule control rules. The control materials used are Kodak I and II controls, which are lyophilized bovine serum products to which purified nonhuman enzymes have been added, to give two concentrations for each analyte.

The statistical analysis done in our study was performed
by the time-series routines found in the interactive IBM PC-version of Minitab. Similar routines can be found in IDA, SPSS, BMDP, SCA, SAS, and numerous other statistical software packages.

Results

To illustrate the time-series modeling approach, we consider data on Kodak II (high level) control for conjugated bilirubin (Bc), starting in October, 1986, and continuing to February, 1987. First, we show the sequence plot of the series in Figure 1. To simplify discussion, we consider only two control rules: Westgard's 1x and 3x limits. Applying those control rules to the original Bc series, we see that observations 3, 37, 65, 127, 150, and 154 are regarded as "warnings," while there are no "rejections."

From visual examination, we see that the series is obviously out of control. The series shows strong evidence of positively autocorrelated behavior; i.e., there is a meandering pattern to the observations. It is not clear that the process is even mean-stationary (see Appendix).

The runs test shown in Figure 2 confirms the strongly nonrandom visual impression. If the process is random, i.e., in control, we can expect on average 115.76 runs. The low runs count of 67 confirms the clustering tendency. The associated P-value of 0.0000 (i.e., <0.0001) clearly rejects the null hypothesis of randomness for any reasonable level of significance, typically chosen to be 0.05.

The autocorrelation function (ACF) and partial autocorrelation function (PACF) with 95% confidence limits are given in Figure 3. The ACF shows significant values at all 25 lags given. Furthermore, the decay in the autocorrelation as the lag k increases is quite slow, suggesting that the series could be nonstationary and that it therefore might be differentiated further the analysis. This confirms possible nonstationarity of mean observed visually.

The analysis of the ACF and PACF suggests that we should difference the series to achieve stationarity. We denote the differenced series by \( \Delta x_k \). The ACF and PACF, together with their probability limits, for the differenced series are given in Figure 4. The ACF clearly cuts off after the first lag. The PACF appears to decay in a steady fashion from the first lag. The indication is that the differenced series, \( \Delta x_k \), should be fitted with an MA(1) model. Thus, the original Bc series, \( x_k \), appears to follow an ARIMA(0,1,1).

An ARIMA(0,1,1) was fitted with results shown in Figure 5. The t-ratio of 23.77 for the moving average term is highly significant. Modified Box–Pierce (11) statistics, which test the null hypothesis that autocorrelations of the residuals
are zero up to a specified lag $k$, are all insignificant for the computer-provided lags of 12, 24, 36, and 48.

The runs test of the residuals in Figure 6 gives a $P$-value of 0.173; thus, we are unable to reject randomness of residuals. The ACF and PACF of the residuals, together with their probability limits, in Figure 7 show no significant correlations—further evidence that the residuals have no systematic nonrandom behavior to them; i.e., the residuals appear random.

The histogram and normal probability plot of the standardized residuals are given in Figure 8. The plots give no apparent indication of nonnormality; hence, it is reasonable to assume normality of the residuals.

Thus, the model diagnostics for the fitted ARIMA(0,1,1) indicate that the residuals resemble a normally distributed random series. The fitted model is satisfactory and reasonably accounts for the systematic nonrandom behavior found in the conjugated bilirubin control series. The estimated final model is given by

$$x_t = x_{t-1} + 0.8391 \varepsilon_{t-1} + \varepsilon_t$$

or equivalently,

$$\Delta x_t = -0.8391 \varepsilon_{t-1} + \varepsilon_t$$

where $x_t$ is the fitted observation at time $t$ and $\varepsilon_t$ is the estimated random shock at time $t$. The interpretation is that the difference between two successive time periods is a multiple, $-0.8391$, of the last period’s unexplainable random shock plus the current random shock.

Unlike the original series, the residuals are random and normal and, thus, they are in a state of statistical control, with the major exception of local departures or assignable causes. We can now apply control rules to the residuals to detect these possible assignable causes, without the danger of confounding of assignable causes with the systematic nonrandom pattern reflected in the fitted ARIMA(0,1,1) for our data. As before, we consider only the $I_{3a}$ and $I_{3b}$ control rules of Westgard. The residuals chart with control limits can be found in Figure 9. Observations 37, 60, 69, 93, 126, 160, 165, 170, and 222 are “warnings,” while observations 64 and 127 are “rejections.” Comparing this with the original series, we see that for the original series six warnings occur vs nine warnings for the residuals chart. More importantly, two rejections were revealed by the residuals chart that went undetected in the original series, one of which (observation 127) is near four standard deviations away from the mean! More careful analysis reveals that three warnings in the original series should be regarded as nonwarnings, seven nonwarnings should be regarded as warnings, and two warnings should be regarded as rejections. It is clear that nonrandom behavior of the data distorted the picture in two ways: higher rate of false alarms and lower rate of true warnings.

---

**Final Estimates of Parameters**

<table>
<thead>
<tr>
<th>Type</th>
<th>Estimate</th>
<th>St. Dev.</th>
<th>t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA 1</td>
<td>0.8391</td>
<td>0.0333</td>
<td>25.77</td>
</tr>
</tbody>
</table>

**Differencing: 1 regular difference**

<table>
<thead>
<tr>
<th>No. of obs.: Original series 222, after differencing 221</th>
<th>Residuals: 95 = 2.9501 (Backcasts excluded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS = 0.01109    DF = 230</td>
<td></td>
</tr>
</tbody>
</table>

**Modified Box-Pierce chi-square statistic**

<table>
<thead>
<tr>
<th>Lag</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>4.2 (DF=11)</td>
<td>13.2 (DF=23)</td>
<td>29.3 (DF=35)</td>
<td>41.1 (DF=47)</td>
</tr>
</tbody>
</table>

---

**Residuals Series**

<table>
<thead>
<tr>
<th>$K = -0.0062$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>THE OBSERVED NO. OF RUNS = 104</strong></td>
</tr>
<tr>
<td><strong>THE EXPECTED NO. OF RUNS = 116.3247</strong></td>
</tr>
<tr>
<td><strong>111 OBSERVATIONS ABOVE X (120) BELOW</strong></td>
</tr>
<tr>
<td><strong>THE TEST IS SIGNIFICANT AT 0.1750</strong></td>
</tr>
<tr>
<td><strong>CANNOT REJECT AT ALPHA = 0.05</strong></td>
</tr>
</tbody>
</table>

---

**Fig. 5. Parameters of the best-fit time-series model (an ARIMA(0,1,1)) for the conjugated bilirubin series**
Fig. 7. Plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the time-series plot of the residuals from the best-fit ARIMA(0,1,1) model for the conjugated bilirubin series shown in Fig. 9.

Midpoint Count

-2.5 3
-2.0 4
-1.5 18
-1.0 23
-0.5 48
0.0 23
0.5 51
1.0 22
1.5 14
2.0 3
2.5 1
3.0 1
3.5 1
4.0 1

Fig. 8. Histogram and normal probability plot for the residuals from the best-fit ARIMA(0,1,1) model for the conjugated bilirubin series shown in Fig. 9.

Correlation of scores and sres = 0.991

Fig. 9. Time-series plot of the residuals from the best-fit ARIMA(0,1,1) model for the conjugated bilirubin series shown in Fig. 1 (ordinate in days).

and lower rate of error detection.

The fitted chart or common cause chart can be seen in Figure 10. This plot gives us a view of the evolution of the process mean value through time. In other words, the common cause chart shows us the underlying systematic behavior of the process.

Similar analyses were performed for 24 analytes at low concentration and 23 analytes at high concentration. Tables 1 and 2 show the summary of fitted ARIMA models and the corresponding model equations. It is particularly noteworthy that none of the series fitted the IID model. We also observe that 29 of 49 analytes were fitted with a nonstationary model.

To compare the modeling approach with the present approach, we need to consider a measure that summarizes the difference between the two methods. In particular, we
consider a discordance ratio for the low- and high-concentration series. In our calculation of the ratio and comparison of the two approaches, we use all six control rules recommended by Westgard. For each analyte, we first sum the total number of warnings and rejections in both the residual and original series. Second, we count the number of warnings and rejections not in agreement in both cases. Third, for a given concentration, we sum across analytes to obtain the total number of rejections and warnings, denoted by T, and the total number of disagreements between the two approaches, denoted by D. Then, the discordance ratio is given by (D/T).

Applying the above calculations to our data, we find for the low-concentration series a total of 666 warnings and rejections and 426 of these signals to be in disagreement, giving a discordance ratio of 63.96%. For the high-concentration control, 630 warnings and rejections with 407 disagreements give a discordance ratio of 64.60%.

Discussion

One intention of our work is to clarify and delineate causes of quality problems found in clinical chemistry laboratories. Two distinct types of causes were identified and named "special" and "common."

Common causes are related to the system as a whole, inherently exist in the process over time, period after period, and affect every component of the process. In contrast, special causes are those that are related to an individual component of the system such as a specific reagent or instrument component or a particular worker. In order to improve any process, it is vitally important to know not only when to make an adjustment but also to know who needs to make it. Identification and removal of special causes can potentially be handled by the technologists or their supervisors at the bench. However, removal of common causes may be a system problem beyond the authority of a specific local operator to address. Only management has the resources and directive to implement a fundamental change in the process required to eliminate or reduce the common causes, i.e., change equipment, operating procedures, or the working environment.

In addition to outlining the separation of causes, we suggest the use of an established statistical methodology, ARIMA modeling, to detect the existence of common causes.
and to unweb the confounding effects of common causes from the special causes. This phenomenon of confounding effects can be quite dramatic, as we found in our application. The discordance ratio seems to provide a reasonable measure of this phenomenon in the sense that it summarizes the difference in warning and rejection signals between the original series and the modeled series. We found the discordance ratio to be as high as 64.60%, clearly reflecting more than a marginal impact of the common causes.

In developing a model it is pedagogically reasonable to illustrate the new approach retrospectively with the available data. However, it is imperative to have real-time control by prospectively implementing the time series approach. Analogous to all current statistical process control methods, the proposed method forecasts future performance of the monitored system based on a sample of its previous performance. For instance, in the case of the standard Levey–Jennings chart, the method assumes that a particular model ("IID normal" model) has been fit to previous data. In the real-time use of the Levey–Jennings method, one then projects a simple linear-forecast function of the grand mean and control limits of ±k standard deviations. This optimum forecast of the grand mean is based on a minimum mean-square error criterion. Our method differs only in having the more general assumption that a broader class of model has been fit.

In a similar fashion to the Levey–Jennings chart, once the correct time-series model has been identified, its parameters estimated, and diagnostics performed, the model may be used to generate forecasts that are optimal based on minimum mean-square error. If the current period is \( T \), we denote the optimum forecast of \( x_{T+1} \) by \( f_{T+1} \), which we can call the one-step forecast. The one-step forecast error is:

\[
epsilon_{T+1} = x_{T+1} - f_{T+1}
\]

Box and Jenkins (10) show that for a minimum mean-square error criterion, the one-step forecast errors, \( \epsilon_{T+1} \), are IID normal with mean zero and variance equaling the variance of the underlying error process, denoted by \( \sigma^2 \) (see Appendix). Thus, the prospective special cause chart would simply be a plotting of the one-step forecast errors, with control limits that are linear projections of ±ks around zero, where s is the estimated standard deviation of the error process. The estimated standard deviation s can be obtained from the standard deviation of the residuals from the retrospective fit. It is equally important to extend the common cause chart by plotting the one-step forecasts to track the future evolution of the process.

In Table 3 we give the one-step forecast function for four commonly encountered models (refer to the Appendix for model notation). To forecast more-general ARIMA models, Box and Jenkins (10) provide a comprehensive section on forecasting. We note that the forecast function \( f_{T+1} \) for three of the models found in Table 3 is recursively defined. It is common practice to start up the sequence of forecasts by setting \( f_{0,1} = 0 \), where period 0 is the period before the prospective charts are started.

The use in practice of the two charts in our proposed methodology is as follows: the common cause chart provides a visual representation of the persistence that affects the whole system. Thus, it is the common cause chart that should be used by management. Clinical chemists, biomedical engineers, and vendors of instrumentation and control materials can all potentially provide valuable information and feedback to management in the search for sources of the common causes.

On-line operators can use the special cause chart with standard control rules, such as the popular multirules of Westgard et al., to identify the sporadic special causes requiring on-line decisions. We think that this modeling approach can be implemented with little burden on the on-line operators. An automated package for time-series modeling can be implemented with existing software. Such an on-line system will produce standard Levey–Jennings plots of the residuals from the fitted model rather than from the original series. Instead of recognizing warnings and rejections in the original data, the on-line technologist will be asked to notice warnings and rejections in the residuals, an operation requiring no changes in interpretation or protocol and no greater statistical expertise. In fact, the model would be entirely "transparent" to the operator.

The interpretation of the fitted model, as well as the fitting process, requires some skill in the analysis of time series. However, we believe the ability to sift out common causes from special causes more than justifies the use of more sophisticated methods. With the availability of automatic fitting of ARIMA models, the implementation can be accomplished by people with limited statistical skills.

We summarize the fitting process in the following steps:

- Plot series to visually detect possible systematic nonrandomness.
- Apply runs test, ACF, and PACF to series, to test for nonrandomness.
- If nonstationarity is evident, difference the series d times till the series appears stationary. Reapply ACF and PACF to differenced series in order to identify orders of moving average and autoregressive components.
- Fit the identified model to the series.
- Check residuals for randomness by using visual inspection, runs test, modified Box–Jenkins statistics, ACF, and PACF applied to residuals. Also, check residuals for normality by using histograms and normal probability plots.
- If residuals pass diagnostics, establish the fitted model.

We would like to thank Professor Harry Roberts of the Graduate School of Business for his review of this manuscript, the editorial reviewer of Clinical Chemistry for helpful comments, and Mr. Shailk Hussein, Chief Technologist of the General Clinical Chemistry Laboratory, for his assistance with data acquisition.

### Appendix

A time series is a chronologically ordered sequence of observations on a particular variable, for example: \( x_1, x_2, \ldots \)
..., x_{t-1}, x_t, x_{t+1}, ... If these observations are statistically dependent upon each other, then ARIMA methodology is appropriate. The purpose of this appendix is to provide a brief explanation of the Box–Jenkins approach to univariate time series analysis. For greater detail on ARIMA modeling the reader should consult Box and Jenkins (10).

### Autoregressive Models

Autoregressive models are linear models accounting for correlation between adjacent observations. For example, a time series generated by the model

\[ x_t = \alpha x_{t-1} + \delta + \epsilon_t \]

where \( \epsilon_t \) is a zero-mean error process and \( \delta \) is a constant term, is referred to as a "first-order autoregressive model" and is denoted by AR(1). The error process, \( \epsilon_t \), is assumed to be independently normally distributed with a zero mean and constant variance, \( \sigma^2 \). The \( \alpha \) constant can be interpreted as a carryover effect of the last period to the current period. The model has two distinct interpretations, depending on whether \( |\alpha| < 1 \) or \( |\alpha| \geq 1 \). When \( |\alpha| < 1 \), then the process can be shown to be stationary. A stationary process is characterized by having a constant long-run mean and variance, and the correlation between two time points depends only on the time gap and not on time itself; that is, the correlation structure is time-invariant. In the second case, \( |\alpha| \geq 1 \), the mean and variance of \( x_t \) increases dramatically as \( t \) increases. A time series is referred to as nonstationary when it contains growth or decline in either the long-run mean or variance.

Let us assume that the series follows a stationary AR(1); i.e., \( |\alpha| < 1 \). It can be shown that the variance of the process is given by

\[ \text{var}(x_t) = \frac{\sigma^2}{1 - \alpha^2} \]

The process has a mean given by

\[ \mu = \frac{\delta}{1 - \alpha} \]

If we define \( \rho_k \) as the correlation between observations \( k \) periods apart, i.e.,

\[ \rho_k = \frac{\text{cov}(x_t, x_{t-k})}{\sqrt{\text{var}(x_t)} \sqrt{\text{var}(x_{t-k})}} \]

it can be shown that

\[ \rho_k = \alpha^k \quad \text{for} \quad k = 1, 2, \ldots \]

Thus, this process has infinite memory; i.e., the current observation depends on all past observations, with the degree of dependence declining with time. For \( \alpha > 0 \), the autocorrelations decay exponentially to zero, and for \( \alpha < 0 \), the autocorrelations decay in an oscillatory manner. So, there is always some correlation between all time points for any lag \( k \), in particular, \( \rho_1 = \alpha \). Series that have \( \alpha \) near \(+1\) will appear smooth (relative to a simple random process) as can be seen in Figure A1. If \( \alpha \) is near \(-1\) the process is less smooth and appears choppy, see Figure A2.

The general autoregressive process is one generated by

\[ x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + \delta + \epsilon_t \]

and is called an autoregressive series of order \( p \), denoted AR(p). Stationarity conditions for the \( \alpha_1, \alpha_2, \ldots, \alpha_p \) can be derived, but are beyond the scope of this Appendix [see Box and Jenkins (10)].

### Moving-Average Models

An alternative way of producing series with more structure than a simple random process is with a moving average; for example, a process generated by

\[ x_t = \mu + \epsilon_t - \beta \epsilon_{t-1} \]

where \( \epsilon_t \) is a zero-mean error process and \( \mu \) is a constant term, is referred to as a "first-order moving average model," denoted by MA(1). Here, the current observation, \( x_t \), is some fraction, \( \beta \), of last period’s error deviation, \( \epsilon_{t-1} \), plus the current error term, \( \epsilon_t \). The process variance is given by

\[ \text{var}(x_t) = \sigma^2 (1 + \beta^2) \]

The mean of the moving average process is \( \mu \). The autocorrelations are described by

\[ \rho_1 = -\beta \]

\[ \rho_k = 0 \quad \text{for} \quad k > 1 \]

This implies that observations one step apart are correlated. Observations more than one step apart are uncorrelated. The implication of this is that the process "forgets" what happened more than one period in the past; i.e., the process has a memory of one period.

The general moving average process is one generated by

\[ x_t = \mu + \epsilon_t - \beta_1 \epsilon_{t-1} - \ldots - \beta_q \epsilon_{t-q} \]

and is called a "moving average series of order \( q \)," denoted MA(q). It should be noted that a pure moving average process is always stationary.

### Autoregressive-Moving Average Models

A more general class of models can be generated by mixing autoregressive and moving average models. For example, a time series generated by the model...
\[ x_t = \alpha x_{t-1} - \beta e_{t-1} + \delta + \epsilon_t \]

is referred to as an "autoregressive-moving average model of order 1," denoted ARMA(1,1).

The stationarity condition for an ARMA(1,1) model is the same as that of an AR(1), i.e., \(|\alpha| < 1\). The autocorrelations are described by

\[ \rho_k = \frac{(1 - \alpha^2)}{1 + \beta^2 - 2\alpha \beta} \quad \text{and} \quad \rho_k = \alpha \rho_{k-1} \quad \text{for} \quad k > 1 \]

It should be noticed that the autocorrelation for a stationary ARMA(1,1) process is similar to that of an AR(1) process: it is characterized by an exponential decay. Again, the decay is smooth if \(\alpha > 0\) and oscillating if \(\alpha < 0\).

The general autoregressive-moving average process is one generated by

\[ x_t = \alpha_1 x_{t-1} + \ldots + \alpha_p x_{t-p} - \beta_1 e_{t-1} - \ldots - \beta_q e_{t-q} + \delta + \epsilon_t \]

and is called an "autoregressive-moving average series of order \(p,q\)," denoted ARMA(p,q).

Autoregressive Integrated Moving Average Models

In practice, many observed time series are not stationary. In particular, series often exhibit time-changing means and variances.

A series exhibiting a time-changing mean often can be well described by a low-order deterministic polynomial in time, for example, a deterministic linear trend model

\[ x_t = b_0 + b_1 t + \epsilon_t \]

The above model is simply a regression model with time as an independent variable.

The deterministic polynomial approach to model fitting is appropriate when the coefficients of the polynomial stay constant through time. However, it is the case that many series have trends for which the coefficients change stochastically over time.

In such cases, Box and Jenkins (10) strongly advocate transforming the nonstationary series into a stationary series by taking successive differences of the series. Upon transforming the nonstationary series into a stationary series, one fits the transformed series by the previously discussed ARMA models. If the series is differenced \(d\) times before it becomes stationary, and the resulting series is identified by an ARMA(p,q), the original series is said to be an "autoregressive-moving average integrated series of order \(p,d,q\)," denoted ARIMA(p,d,q).4

As an example, consider the following model for \(x_t\)

\[ x_t - x_{t-1} = \beta e_{t-1} + \epsilon_t \]

Here the first difference of the series \((x_t - x_{t-1})\), denoted by \(Vx_t\), follows an MA(1); thus, the process we are considering is an ARIMA(0,1,1). As reported by Alwan and Roberts (9), this model underlies the well-known exponentially weighted moving-average (EWMMA) model, and is quite flexible in reasonably fitting a wide range of applications. Harris (12) found the ARIMA(0,1,1) to be a reasonable model for tracking a patient's progress based on a short series of repeat tests of an analyte.

Autocorrelation and Partial Autocorrelation Functions

The autocorrelations \(\rho_k\) considered as a function of \(k\) are referred to as the "autocorrelation function" (ACF). A plot of the estimated autocorrelations vs \(k\) plays an important role in the identification stage in model building. As we have seen earlier, the ACF of an AR(1) declines steadily or in an oscillating manner, while the ACF of an MA(1) cuts off after \(k = 1\).

In general, the ACF of autoregressive processes damps out and it is often difficult to differentiate among processes of different orders. So, a second function is considered to aid in the discrimination of different processes. This function is known as the partial autocorrelation function (PACF). The partial autocorrelation function is defined as the coefficient of \(x_{t-k}\) when the following kth order regression is fitted

\[ x_t = a_{k1} x_{t-1} + \ldots + a_{kk} x_{t-k} + \epsilon_t \]

The partial autocorrelation coefficient, \(a_{kk}\), measures correlation between \(x_t\) and \(x_{t-k}\) after adjusting for \(x_{t-1}, \ldots, x_{t-k+1}\).

Model Building

Model building, as proposed by Box and Jenkins (10), consists of three stages:

(a) Identification (also referred to as "specification"), which is the procedure for obtaining an approximate idea of the structure of the model, i.e., the degree of \((p,d,q)\).

(b) Estimation of the parameters \(\delta, \alpha_1, \ldots, \alpha_p \beta_1, \ldots, \beta_q\).

(c) Diagnostic checking to ensure that the fitted model is appropriate.

If the fitted model does not pass the diagnostic checks, then a new set of \((p,d,q)\) is specified, and the three stages are repeated. This iterative process continues until the final model passes the diagnostic checks.

(a) Model identification. At the identification stage, we make a choice of one or of just a few models that are worthy of consideration. If the series is found to be nonstationary, Box and Jenkins strongly advocate differencing the series successively until a stationary series is produced. Nonstationarity can usually be detected visually from the time series plot of the series. The ACF is also useful in identifying nonstationarity. Generally, if the autocorrelations do not dampen rapidly, this is an indication of nonstationarity. If the series is found to be nonstationary, we take the first difference. If the first difference is stationary, an ARMA(p,q) model is identified and fitted to the differenced series. If the first difference is nonstationary, a second difference is taken, i.e., a first difference of the differenced series, and stationarity is checked. This procedure continues until stationarity is achieved. It should be noted that differencing a stationary series will result in a stationary series. However, overdifferencing a series can lead to serious problems. In particular, overdifferencing can introduce additional variation and the

| Table A1. Theoretical Patterns of ACF and PACF for General ARMA Processes |
|-----------------------------|-----------------------------|-----------------------------|
| Model                     | ACF                        | PACF                        |
| AR(p)                      | Exponential or oscillatory decay | Cuts off for \(k > p\). |
| MA(q)                      | Cuts off for \(k > q\).     | Dominated by damped exponentials and (or) sine waves. |
| ARMA(p,q)                  | Tails off after \(q - p\) lags. Exponential or oscillatory decay after \(q - p\) lags. | Tails off after \(p - q\) lags. Dominated by damped exponentials and (or) sine waves after \(p - q\) lags. |

4 Notationally, an ARIMA(p,0,q) is equivalent to an ARMA(p,q).
sample variances will increase. It has been suggested that changes in sample variances of successive differences be used to determine the degree of differencing (see Tinter (13)).

Once we have obtained a stationary series, we need to determine the orders of $p$ and $q$. Visual examination of the ACF and the PACF, coupled with the knowledge of theoretical patterns of known processes, provides us with an initial guess at the orders of the autoregressive and moving-average terms.

In Table A1, we summarize the rules for identification.

To illustrate the rules, we give ACF and PACF for simulated series following a pure autoregressive process, a pure moving-average process, and a mixed autoregressive moving-average process. For instance, we see in Figure A3, the autocorrelations in the ACF decay in a steady (exponential) fashion down to negligible correlations. The partial autocorrelations clearly cut off after lag 2. Thus, the process appears to be an AR(2).

In Figure A4, the autocorrelations in the ACF seem to cut off after lag 2, while the partial autocorrelations in the PACF show steady (oscillatory) decay. The indication is that the process follows an MA(2).

In Figure A5, the ACF is characterized by exponential decay from the first lag, i.e., an AR(1), but decay starts from lag 1 rather than lag 0. The PACF reveals a large initial value, which is followed by exponential decay similar to an MA(1) process. Thus, the indication is that the process is an ARMA(1,1).

(b) Model estimation. Once the model has been identified, the parameters of the model need to be estimated. The procedure by which the parameters of an ARIMA model are estimated involves the use of nonlinear estimation methods. These methods are packaged in time-series estimation rou-

---

**Fig. A3.** Autocorrelation (ACF) and partial autocorrelation function (PACF) for a simulated AR(2) process

**Fig. A4.** Autocorrelation (ACF) and partial autocorrelation function (PACF) for a simulated MA(2) process
(c) Diagnostics. After estimating the model, it is important to check its appropriateness. This is done by diagnostic checking.

If an ARMA model is fit well, we expect to see (a) the residuals approximately normally distributed, (b) the mean of the residuals near zero, (c) the variance of the residuals approximately constant, and (d) the autocorrelations of the residuals negligible. Again, the ACF and PACF are useful tools in suggesting if the residuals are an uncorrelated error process.

For large n, it can be shown that the standard error of the estimated autocorrelations and partial autocorrelations, when the residuals are an uncorrelated normal error process, is approximately $1/\sqrt{n}$. So, to check whether the autocorrelations and partial autocorrelations of the residuals are negligible, we can compare the estimated autocorrelations and partial autocorrelations with their standard error. Typically, a significance level of $\alpha = 0.05$ is chosen, which corresponds to confidence intervals of $\pm 1.96 (1/\sqrt{n})$. Correlations that fall beyond these confidence limits are considered significant at the 5% level.

There are other useful diagnostic checks for model appropriateness. Ljung and Box (11) give a chi-square test for the null hypothesis that all autocorrelations up to a specified lag are zero (modified Box–Pierce statistic). The nonparametric runs test is another test of randomness of a series [see Duncan (14)].

The normality of the residuals can be checked by constructing a histogram of the standardized residuals and seeing if gross departures from normality are revealed. Another useful graph is the normal probability plot of the residuals. If the plot is approximately a straight line, the distribution is approximately normal. Other tests for normality include goodness-of-fit tests, for instance the chi-square or Kolmogorov–Smirnov, that can be used for testing the normality of the error term. Based on experience and personal preference, we have found the histogram and probability plot of the residuals to suffice. It should be further noted that if nonnormality is severe, simple transformations, for example $\log(x)$, square root ($x$), etc., found in regression analysis will usually remedy the problem [see Box and Cox (15)].

All of the above-mentioned statistics (ACF, PACF, Box–Pierce, runs test, etc.) can be found in most statistical software packages, such as MINITAB, SCA, BMDP, SAS, or SPSS.

References