

How To: Construct a Control Chart for Autocorrelated Data Using STATGRAPHICS Centurion

by

Dr. Neil W. Polhemus

July 21, 2005

Introduction

Standard SPC control charts are based on the assumption that the observations obtained at each time period are independent, at least when the system is in control. Unfortunately, many processes do not satisfy this assumption, particularly when sampled very frequently. Instead, the observations close together in time exhibit “autocorrelation”, i.e., large values tend to be followed by other large values and small values tend to be followed by other small values. Such dynamic swings are often unavoidable, particularly if one is measuring some characteristic of a continuous process, as when measuring viscosity every hour.

For data that exhibit autocorrelation, the proper analysis depends upon constructing a parametric time series model that captures the dynamic features of the process. Such models allow one to separate current shocks to the system from effects left over from earlier events. In this guide, we will examine a typical set of measurements and show how to develop such a model. The model will then be used to construct an ARIMA control chart that accounts for the observed autocorrelations.

Sample Data

As an example, we will consider Series A from the well-known book by Box, Jenkins, and Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd edition (Pearson Education, 1994). It consists of 197 readings of the concentration in a chemical process, taken once every two hours. The data have been stored in a STATGRAPHICS data file called *howto3.sfb*.

Step 1: Plot the Data

The first step in analyzing any new set of data is to plot it. For sequentially ordered data, a runs chart is usually very informative.

Procedure: Runs Chart

To create a runs chart in STATGRAPHICS Centurion:

- If using the Classic menu, select: *Plot – Time Sequence Plots – Run Charts – Individuals*.
- If using the Six Sigma menu, select: *Measure – Time Sequence Plots – Run Charts – Individuals*.

On the data input dialog box, indicate the name of the column containing the measurements:

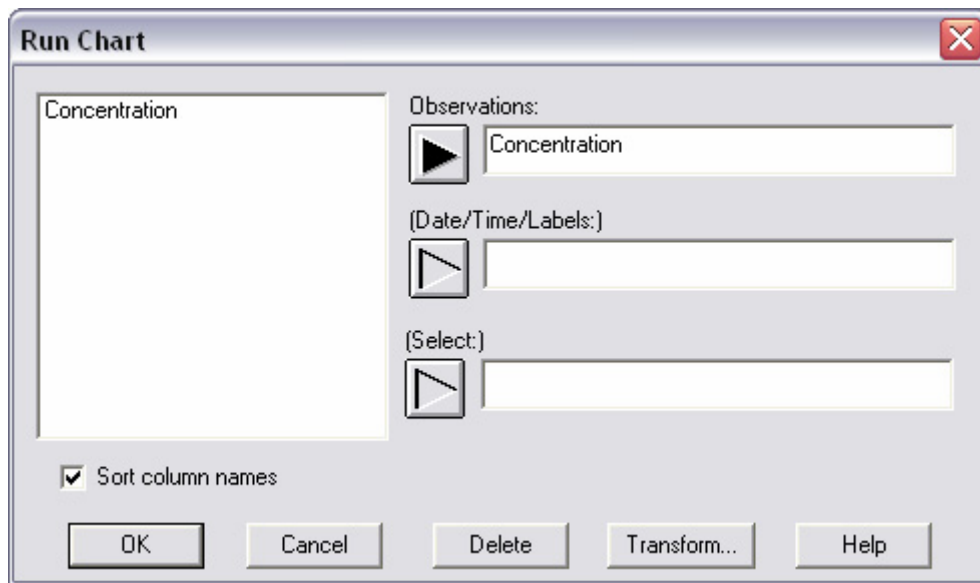


Figure 1: Data Input Dialog Box for Individuals Run Chart

The resulting plot shows strong swings around the central value:

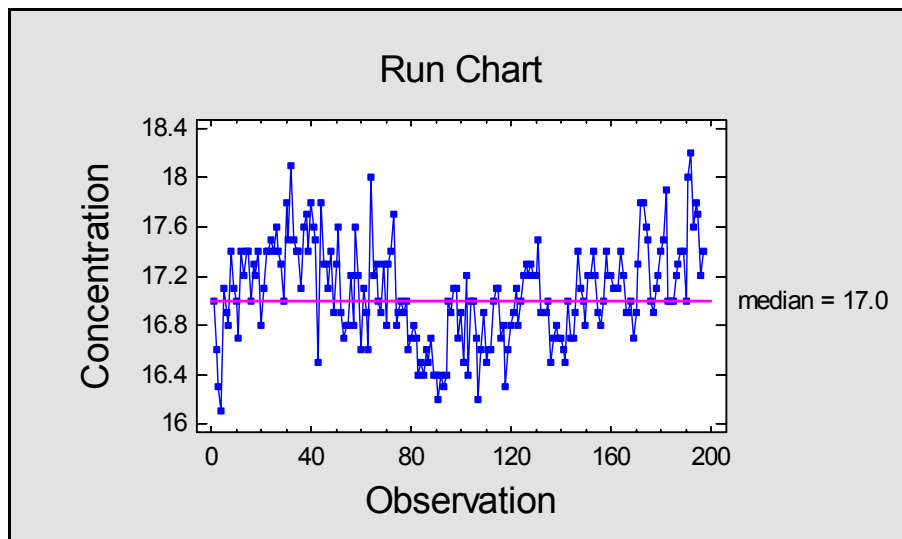


Figure 2: Run Chart for Chemical Process Concentrations

The *Analysis Summary* generated by the *Run Chart* procedures shows the results of several runs tests used to determine whether the observed data form a sequence of independent observations:

Run Chart (Grouped Data) - Concentration

Data variable: Concentration
197 values ranging from 16.1 to 18.2
Median = 17.0

Test	Observed	Expected	Longest	P(>=)	P(<=)
Runs above and below median	44	88.0795	23	1.0	1.38815E-11
Runs up and down	110	115.0	5	0.840614	0.207331

The StatAdvisor

This procedure is used to examine data for trends or other patterns over time. Four types of non-random patterns can sometimes be seen:

1. Mixing - too many runs above or below the median
2. Clustering - too few runs above or below the median
3. Oscillation - too many runs up and down
4. Trending - too few runs up and down

The P-values are used to determine whether any apparent patterns are statistically significant. Since the P(equal or less) value for the runs above and below the median is less than 0.025, there is statistically significant clustering at the 95% confidence level.

Figure 3: Run Chart Analysis Summary

Note the very small P-Value highlighted in red in the last column (1.388×10^{-11}). This corresponds to a test for clustering. It compares the observed number of runs above or below the median (44) to that expected if the observations were randomly sampled from any population (88.08). To reject the assumption of independence between consecutive observations at the 5% significance level, the P-Value had only to fall below 0.05.

Procedure: Individuals Chart

If we were not paying attention to the lack of independence between consecutive observations, we might blindly create an individuals control chart. To create an individuals control chart in STATGRAPHICS Centurion:

- If using the Classic menu, select: *SPC – Control Charts – Basic Variables Charts – Individuals.*
- If using the Six Sigma menu, select: *Control – Variables Control Charts – Basic Control Charts – Individuals.*

Complete the data input dialog box as shown below:

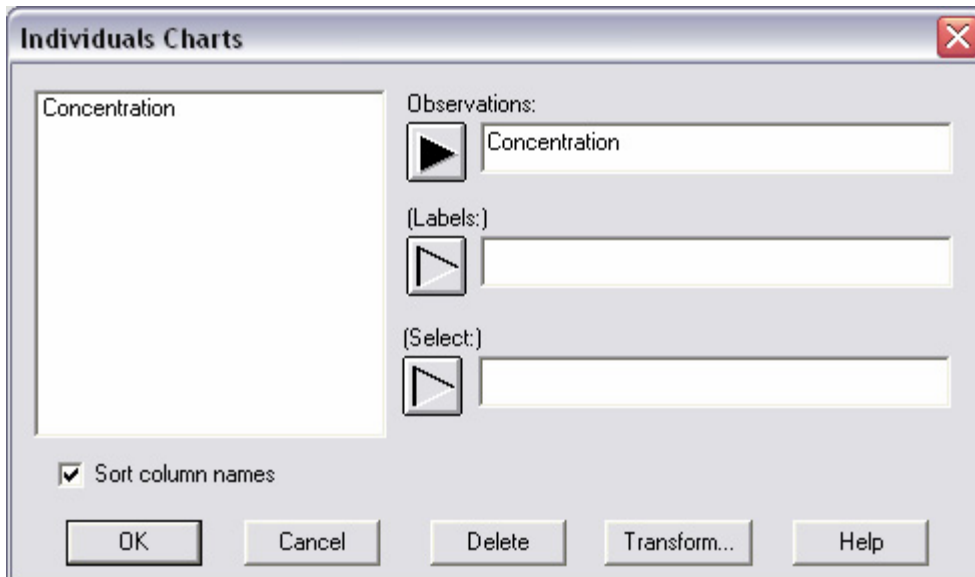


Figure 4: Data Input Dialog Box for Individuals

The resulting control chart shows many points beyond the 3-sigma control limits, and many violations of the usual runs rules:

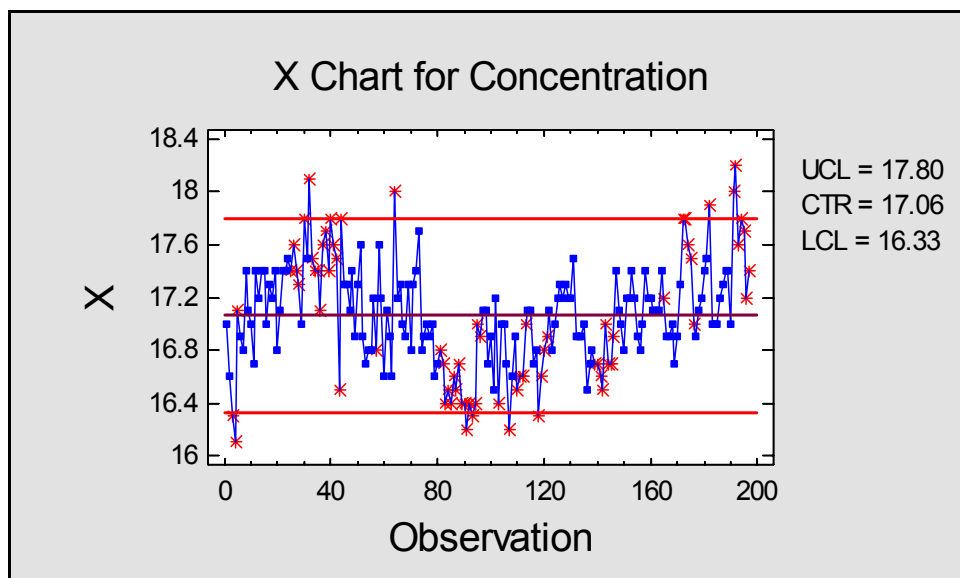


Figure 5: X Chart of Chemical Process Concentrations

At the top and bottom of each cycle, alarms are created. There are also long runs of points either above or below the mean. Such a control chart is useless for monitoring this process, where swings around the mean are an inherent part of the dynamics of the process and do not indicate that the process is “out of control”.

Step 2: Construct a Parametric Time Series Model

In order to monitor this process, we first need to understand the nature of its dynamics. For a stationary process (a process with a constant long-term mean and variance), a very useful class of models are the ARIMA (AutoRegressive, Integrated, Moving Average) models. The general model we will consider represents Y_t , the observed concentration at time period t , as a linear combination of the concentrations observed at the last p time periods, a random shock to the

system at the current time period a_t , and random shocks that occurred at the previous q time periods:

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

The parameters that define this model are:

θ_0 = a constant

$\phi_1, \phi_2, \dots, \phi_p$ = autoregressive parameters

$\theta_1, \theta_2, \dots, \theta_q$ = moving average parameters

This model is capable of representing different types of dynamic behavior and has been widely applied to many types of systems. In fact, it can be shown that when data is sampled from a system governed by a p -th order differential equation, the data should follow a model with p autoregressive parameters and $q = p - 1$ moving average parameters.

From a statistical viewpoint, the general problem is determining what order of model to use (what values to use for p and q) and estimation of the model parameters. In the section below, we will consider this problem, restricting our attention to models in which $q = p - 1$.

Procedure: Descriptive Time Series Methods

To determine which type of ARIMA model to use for a particular set of data, we will rely on two important tools: the autocorrelation function and the partial autocorrelation function. These functions may be created by selecting:

- If using the Classic menu: *Describe – Time Series – Descriptive Methods*.
- If using the Six Sigma menu: *Forecast – Descriptive Time Series Methods*.

The data input dialog box is shown below:

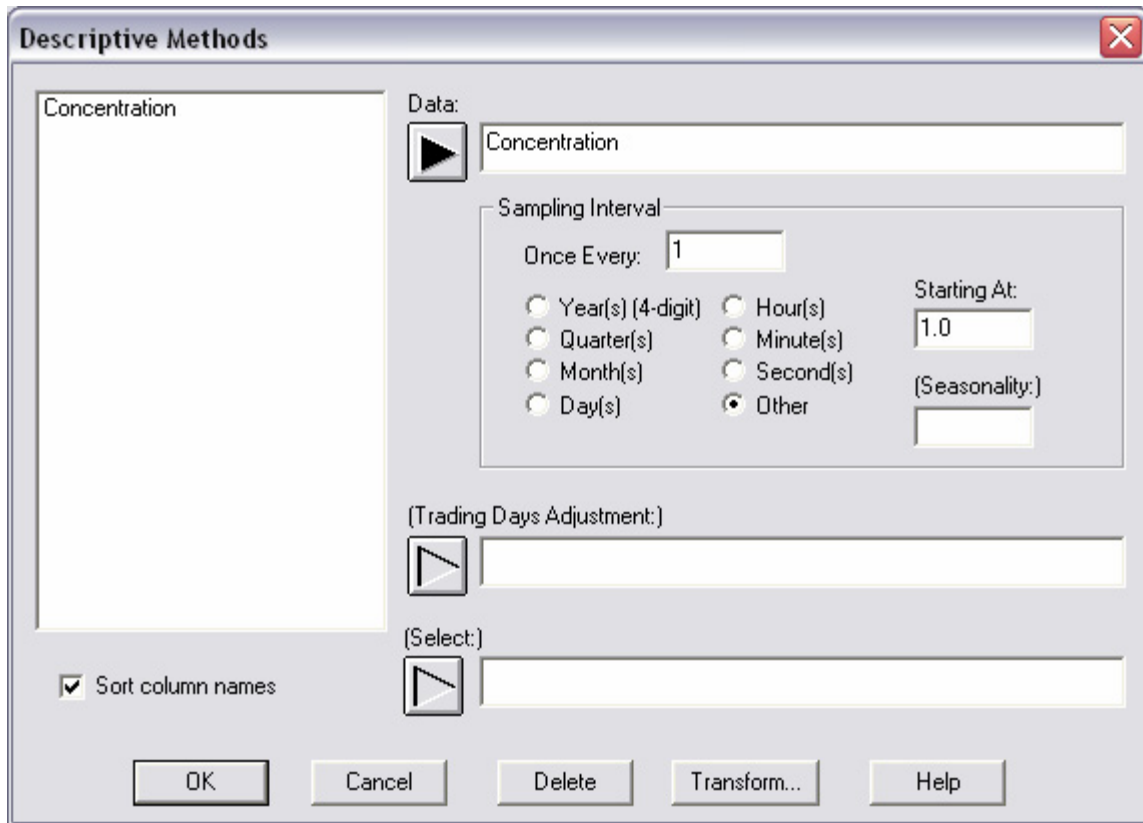


Figure 6: Data Input Dialog Box for Descriptive Time Series Methods

One of the plots generated by default is the *Autocorrelation Function*:

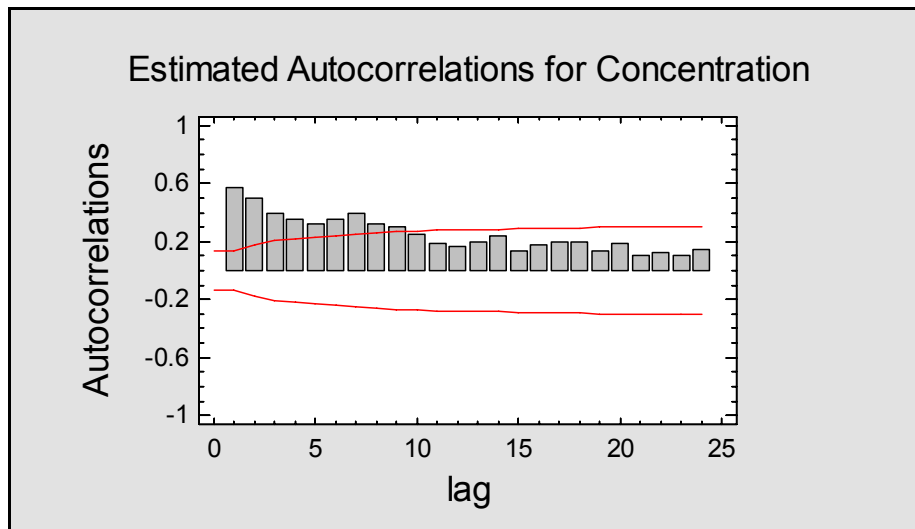


Figure 7: Plot of Sample Autocorrelation Function

The autocorrelation function displays estimates of the correlation coefficient between observations separated by k time periods, where k is called the *lag*. It is most easily thought of as showing how the effect of a shock to the system extends into the future. Think about a pendulum knocked away from its equilibrium position. It will return to its initial position over time, perhaps as a simple exponential (first-order system) or perhaps overshooting that position and exhibiting damped oscillations as it returns to equilibrium (second-order system). The shape of the autocorrelation function is like a thumbprint, identifying the type of dynamics that are present.

For autoregressive models of the type we wish to consider, a second useful function is the *Partial Autocorrelation Function*. This function may be plotted by pressing the *Graphs* button on the analysis toolbar. It plots the partial autocorrelation coefficients as a function of lag k :

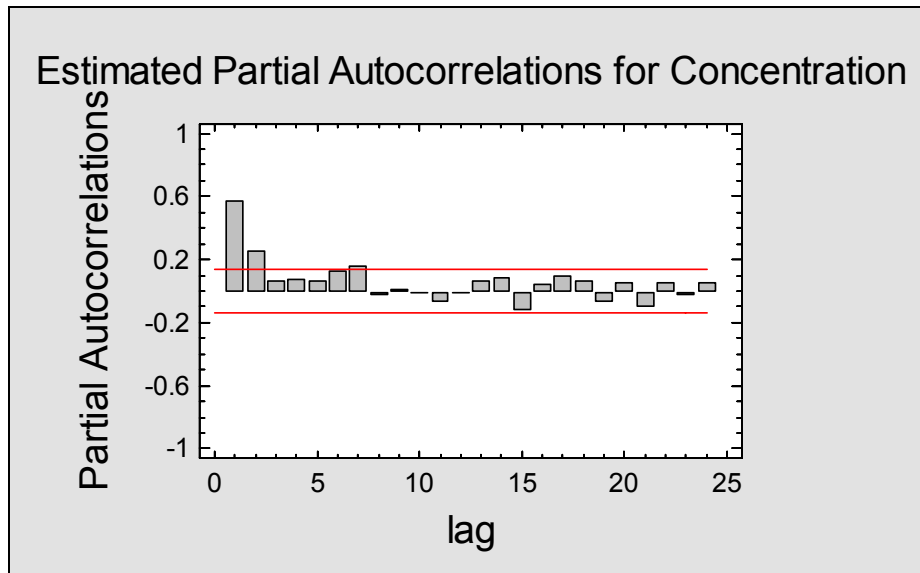


Figure 8: Plot of Sample Partial Autocorrelation Function

Any coefficients extending beyond the 95% probability limits would indicate the need for an autoregressive parameter of that order. For example, the above plot shows significant partial autocorrelation at lags 1 and 2, suggesting that an ARIMA model with $p = 2$ is likely to be needed for this data.

Procedure: Automatic Forecasting

To verify the proper type of model for this time series, it is useful to run the *Automatic Forecasting* procedure, accessed by selecting:

- If using the Classic menu: *Forecast –Automatic Forecasting*.
- If using the Six Sigma menu: *Forecast –Automatic Forecasting*.

This procedure can be set to try different types of ARIMA models, automatically selecting the model that is best by a specified information criterion. The data input dialog box is shown below:

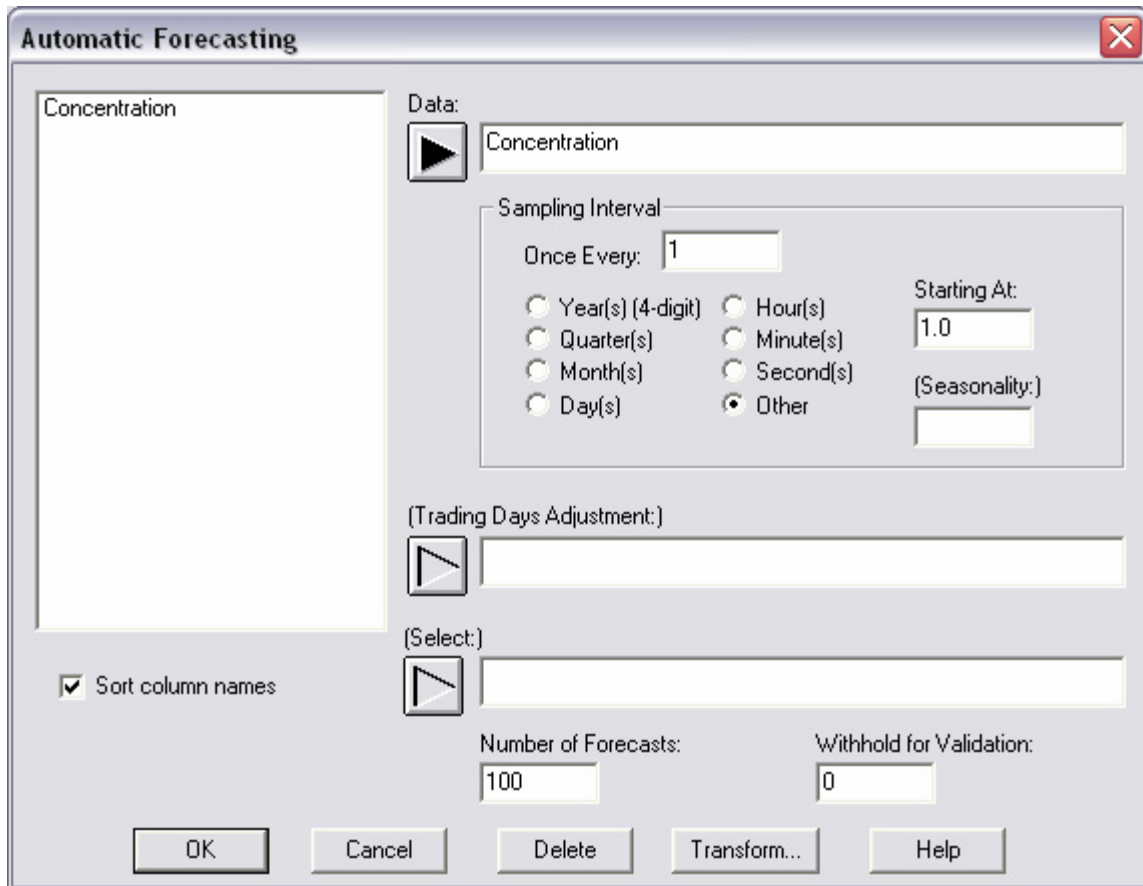


Figure 9: Automatic Forecasting Data Input Dialog Box

Note that we have entered both the name of the column containing the data and the number of forecasts we would like to generate.

As soon as the *Automatic Forecasting* analysis window is created, press the *Analysis Options* button and complete the dialog box as shown below:

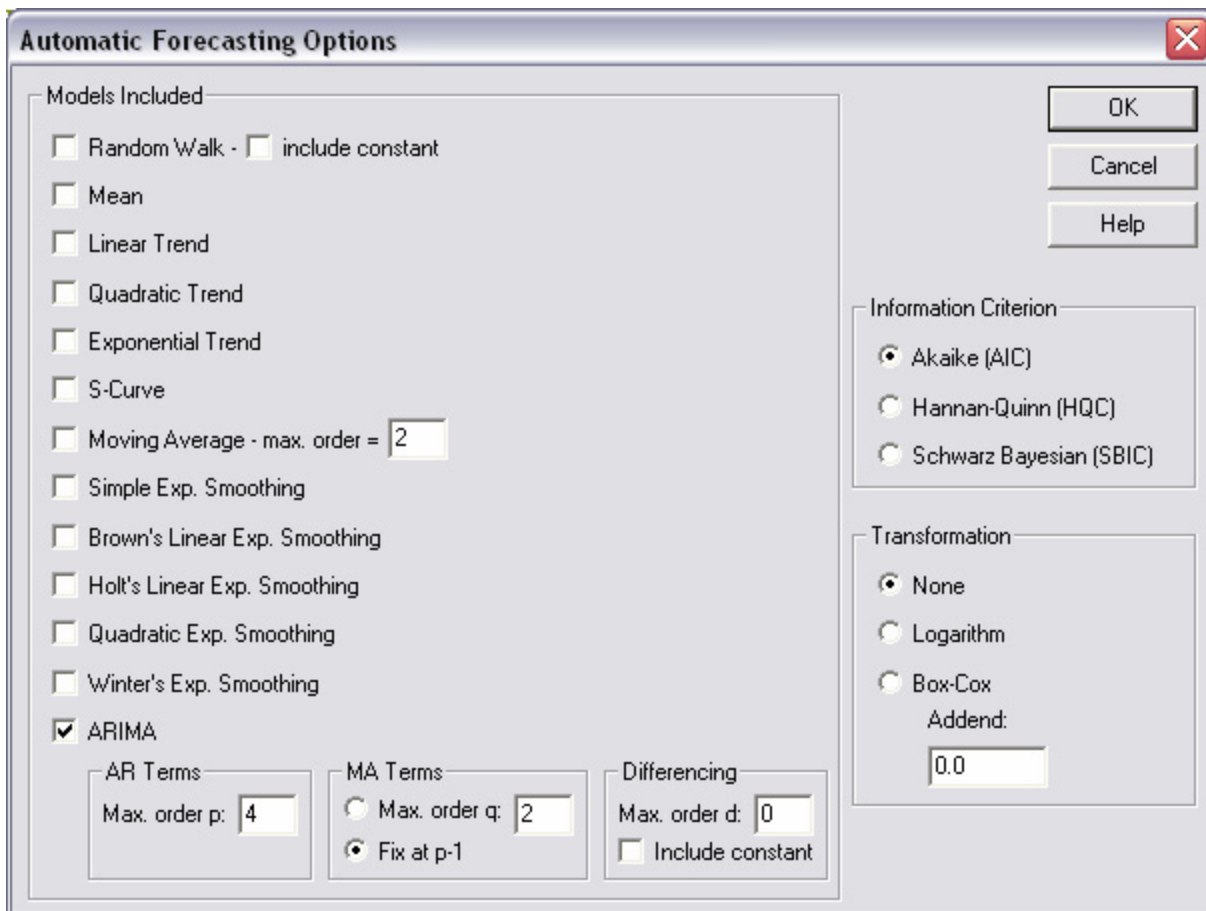


Figure 10: Automatic Forecasting Analysis Options Dialog Box

We have:

1. Turned off all models except for the ARIMA models.
2. Specified a maximum order of $p = 4$ for the *AR Terms*. The procedure will try all ARIMA models up to and including a 4-th order autoregressive model.
3. Asked to fix the *MA Terms* at $p - 1$. Only models for which $q = p - 1$ will therefore be considered.
4. Requested the use of Akaike's information criterion (AIC) to determine the best model. The AIC is basically a penalized mean squared error, with models having more parameters being penalized for their additional degrees of freedom.

When OK is pressed, the models will be fit. The best model will be selected and forecasts generated from it, as shown on the *Time Sequence Plot*:

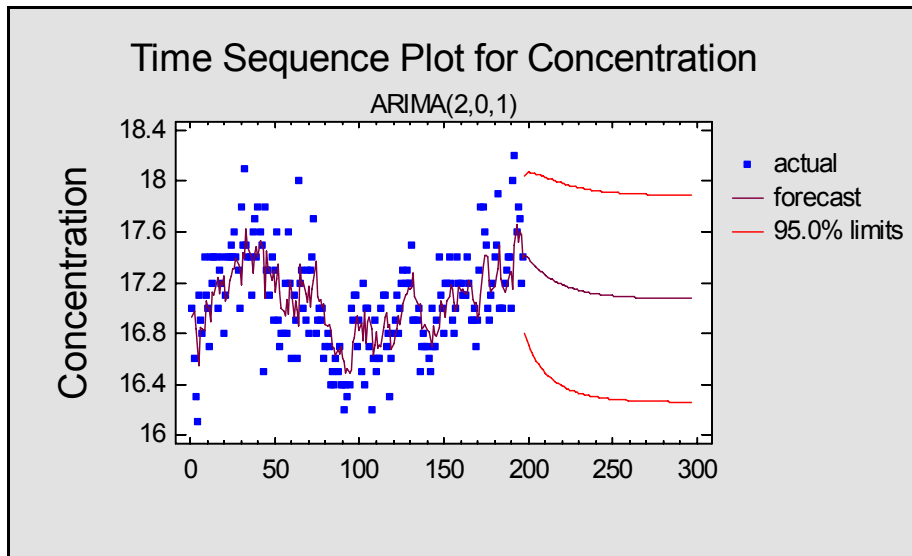


Figure 11: Forecast Function for Selected Model

The best-fitting model has $p = 2$ and $q = 1$. Each of the data values is shown, together with the “one-ahead” forecasts. The one-ahead forecasts are the forecasts that would be made at each time t for the following time period. At the end of the data, the forecasts for 100 future time periods are also displayed, with 95% probability limits. We can think of the forecast function as illustrating how the system would return to its equilibrium value if the shocks to it ceased suddenly at the end of the data collection period.

The *Analysis Summary* displays statistics for the selected model:

Automatic Forecasting - Concentration
 Data variable: Concentration

Number of observations = 197
 Start index = 1.0
 Sampling interval = 1.0

Forecast Summary
 Forecast model selected: ARIMA(2,0,1) with constant
 Number of forecasts generated: 100
 Number of periods withheld for validation: 0

ARIMA Model Summary

Parameter	Estimate	Std. Error	t	P-value
AR(1)	1.12018	0.141195	7.93353	0.000000
AR(2)	-0.162049	0.11433	-1.41738	0.157984
MA(1)	0.74416	0.116662	6.37879	0.000000
Mean	17.0722	0.126207	135.271	0.000000
Constant	0.714836			

Backforecasting: yes
 Estimated white noise variance = 0.0987145 with 193 degrees of freedom
 Estimated white noise standard deviation = 0.314189
 Number of iterations: 10

Figure 12: Summary of Fitted ARIMA Model

Of most interest here is the estimated process mean $\hat{\mu} = 17.07$ and the estimated standard deviation of the shocks to the system, $\hat{\sigma}_a = 0.3142$. Note that the latter is *not* equal to the process sigma $\hat{\sigma}_Y$, which is a function of both $\hat{\sigma}_a$ and the fitted model parameters.

Step 3: Construct an ARIMA Control Chart

Once we know the proper ARIMA model to use, we can then construct a control chart for the data. ARIMA control charts may be created by selecting:

- If using the Classic menu: *SPC – Control Charts – Special Purpose Control Charts – ARIMA Individuals Chart.*
- If using the Six Sigma menu: *Control – Variables Control Charts – Special Purpose Control Charts – ARIMA Individuals Chart.*

The data input dialog box is shown below:

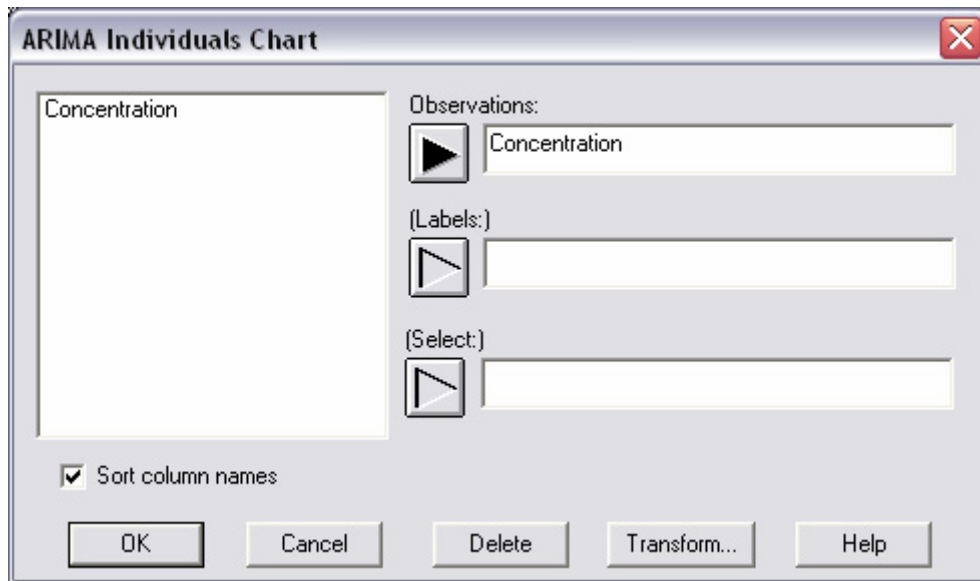


Figure 13: Data Input Dialog Box for ARIMA Individuals Chart

When the analysis window appears, use *Analysis Options* to specify the type of control chart desired:

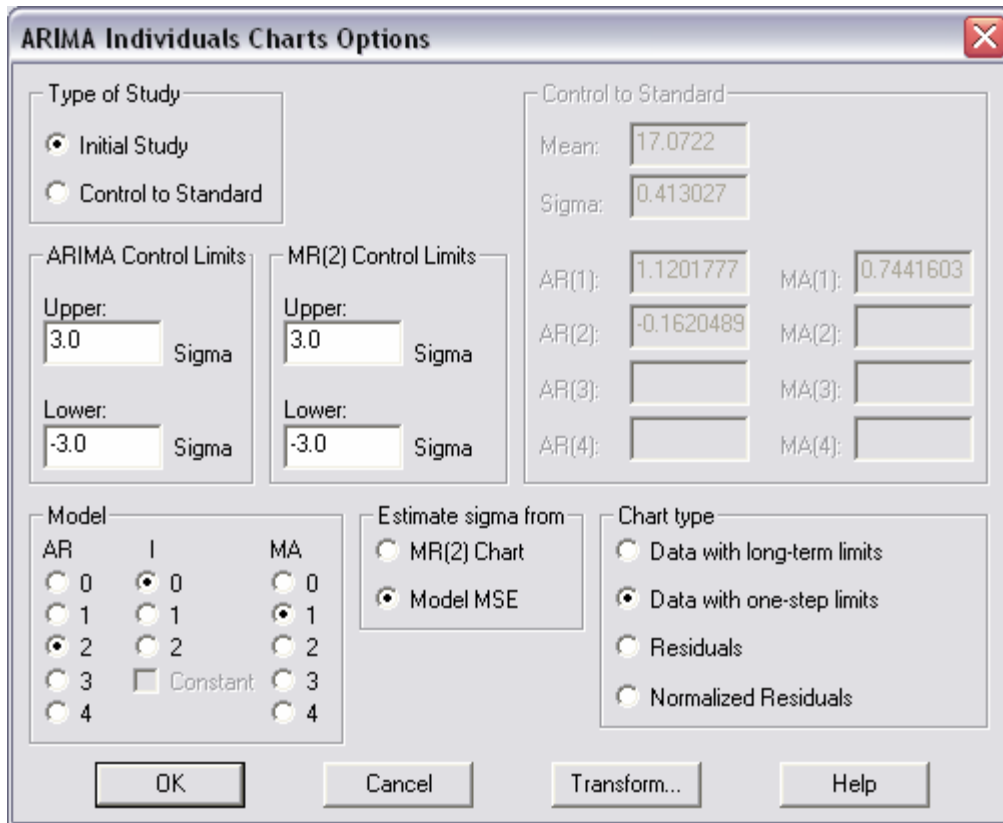


Figure 14: Analysis Options Dialog Box for ARIMA Individuals Chart

Note that we have selected order 2 for the AR term and order 1 for the MA term. We have also asked to *Estimate sigma from the Model MSE*, to be consistent with the time series model fit earlier, although the process sigma could be estimated instead using the average moving range of the model residuals, shown later in this guide. We have also set the *Chart type to Data with long-term limits*, which creates the chart shown below:

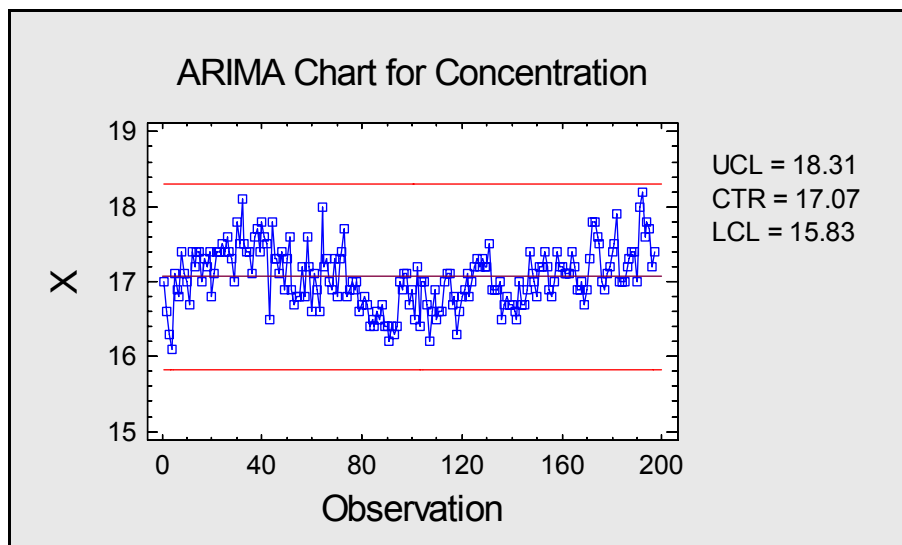


Figure 15: Control Chart Based on Long-Term Limits

On this chart, the centerline is located at the estimated process mean $\hat{\mu} = 17.07$ with control limits at plus and minus the estimated process sigma $\hat{\sigma}_y$. The estimated process sigma is displayed on the *Analysis Summary*:

ARIMA Individuals Chart - Concentration

Number of observations = 197

0 observations excluded

Distribution: Normal

Transformation: none

ARIMA Chart

Period	#1-197
UCL: +3.0 sigma	18.3113
Centerline	17.0722
LCL: -3.0 sigma	15.8332

0 beyond limits

MR(2) Chart

Period	#1-197
UCL: +3.0 sigma	1.08911
Centerline	0.333337
LCL: -3.0 sigma	0.0

5 beyond limits

Estimates

Period	#1-197
Process mean	17.0722
Process sigma	0.413027
Residual sigma	0.314189

Sigma estimated from MSE of fitted model

ARIMA Model Summary

Parameter	Estimate	Std. Error	t	P-value
AR(1)	1.12018	0.141195	7.93353	0.0000
AR(2)	-0.162049	0.11433	-1.41738	0.1580
MA(1)	0.74416	0.116662	6.37879	0.0000
Mean	17.0722	0.126207	135.271	0.0000
Constant	0.714836			

Backforecasting: yes

Estimated white noise variance = 0.0987145 with 193 degrees of freedom

Estimated white noise standard deviation = 0.314189

The StatAdvisor

This procedure creates an ARIMA individuals chart for Concentration. It is designed to allow you to determine whether the data come from a process which is in a state of statistical control. The control charts are constructed under the assumption that the data come from a normal distribution with a mean equal to 17.0722 and a standard deviation equal to 0.413027. These parameters were estimated from the data. Of the 197 nonexcluded points shown on the charts, 0 are beyond the control limits on the first chart while 5 are beyond the limits on the second chart.

Figure 16: Analysis Summary for ARIMA Individuals Chart

Notice that the estimated process sigma is $\hat{\sigma}_Y = 0.4130$. This results in considerably wider bounds than the individuals chart shown in Figure 5, since it allows for the autocorrelation in the data.

It is important to note that the above chart monitors the long-term behavior of the process. Individuals points on the chart are not independent, so that standard runs rules cannot be applied. Also, the chart does not monitor the shocks to the system that occur at each time period. Its use is primarily to detect when the process deviates from the long-term mean more than expected given the dynamics of the process.

To monitor the separate shocks, return to *Analysis Options* and select a different *Chart type*. To display the shocks directly, request a *Residuals* chart:

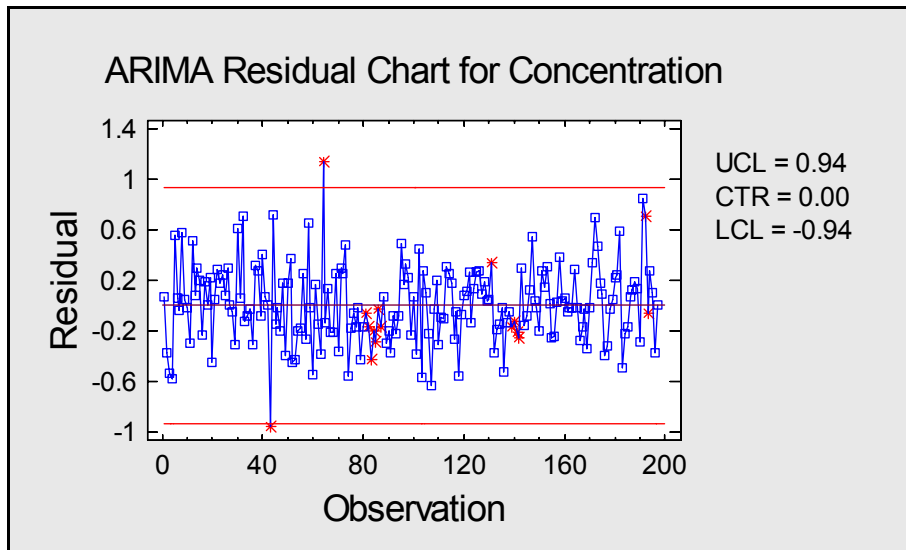


Figure 17: Control Chart of Model Residuals

This chart shows the estimated a_t 's in the ARIMA model, which are the shocks that affect the system at each time period. In the above chart, there are two shocks beyond the 3-sigma limits, indicating that twice during the sampling period, unusually large errors occurred. There are also at least two places where the shocks showed a significant run of negative values.

A moving range chart of the model residuals is also useful:

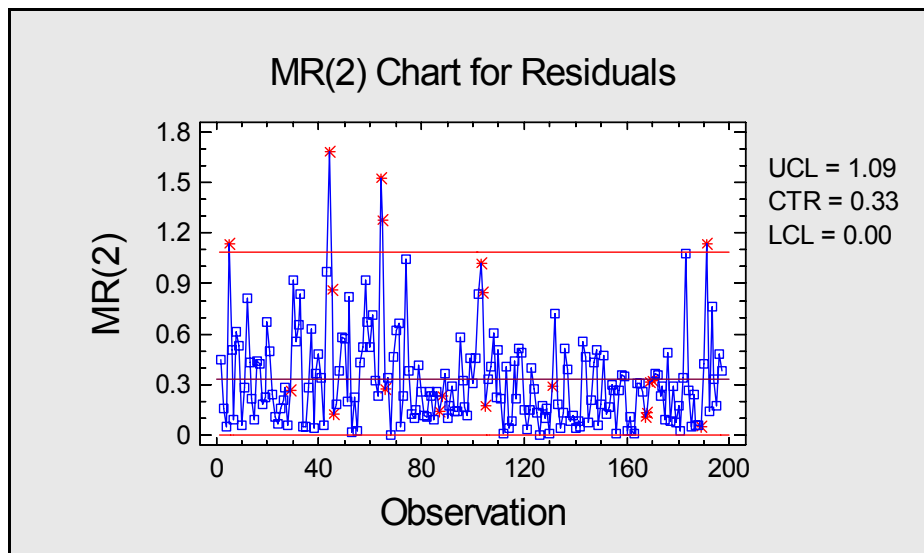


Figure 18: MR(2) Control Chart of Model Residuals

Several times, the moving range exceeds the upper control limits.

An indirect way of plotting the model residuals is preferred by some practitioners. If you return to *Analysis Options* and select *Data with one-step limits*, the following chart will be displayed:

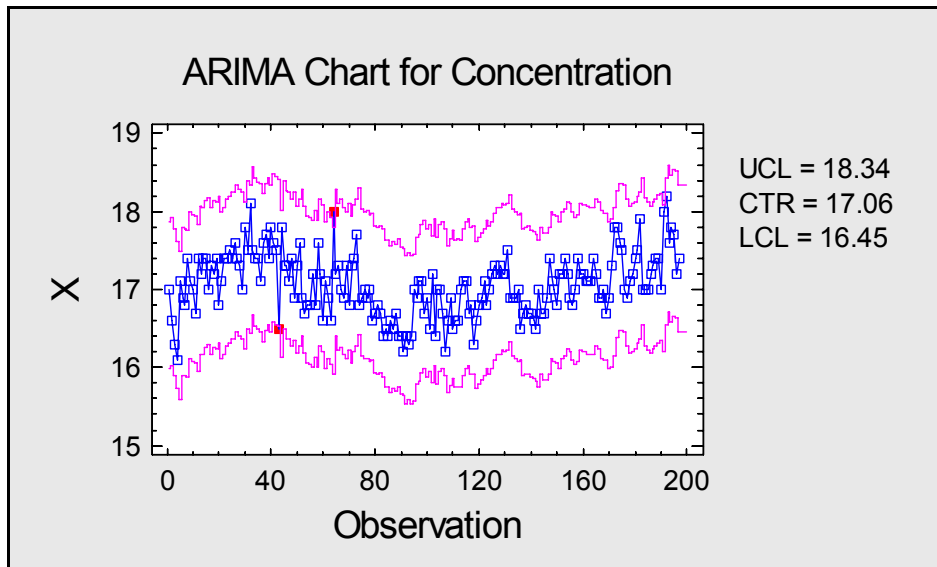



Figure 19: Control Chart with One-Step Limits

This chart plots the original data with moving control limits. At each point in time t , the control limits are centered around the forecasted value for Y_t made at time $t-1$, $\pm 3\hat{\sigma}_a$. This chart will give the same alarms with respect to exceeding the control limits as will the chart of the model residuals.

When using an ARIMA control chart, it is a good idea to check the *Residual Autocorrelation Function* to be sure that the type of ARIMA model used has resulted in residuals that are approximately independent. This chart is available by pressing the Graphs button  on the analysis toolbar:

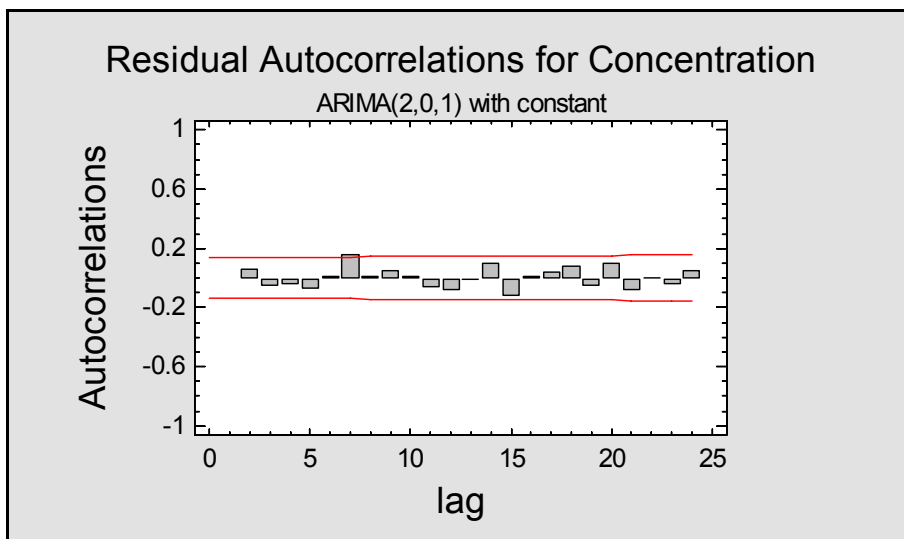


Figure 20: Residual Autocorrelation Function

Note that all the bars are within (or very close to) the 95% probability limits, indicating that no significant autocorrelation remains in the residuals.

NOTE: the *Analysis Options* dialog box shown in Figure 14 allows you to select either a Phase I (*Initial Studies*) chart or a Phase II (*Control to Standard*) chart. When constructing a Phase II chart, you must specify:

1. The process mean μ .
2. The process sigma σ_y .
3. The AR and MA parameters.

The last would usually be obtained from the *Analysis Summary* table generated by the *Automatic Forecasting* procedure (Figure 12), during a previous Phase I study.

Conclusion

When the data to be plotted on a control chart are not independent from one time period to the next, standard control charts should not be used. Instead, an ARIMA control chart can be applied which takes into account the dynamics of the process. Development of an ARIMA control chart requires the extra step of constructing a time series model for the data. The STATGRAPHICS Centurion *Automatic Forecasting* procedure can easily identify a good model for a time series. If desired, the models of interest may be restricted to ARMA models with $q = p - 1$, so that the problem reduces to determining the proper autoregressive model order p .

As we move into the 21st century, the ease of data collection has led more and more individuals to consider applying control charts to their processes. Practitioners must be prepared, however, to go beyond simple charts when assumptions such as independence are violated. Otherwise, the charts will soon be discarded, especially if they give too many false alarms.

Note: The author welcomes comments about this guide. Please address your responses to neil@statgraphics.com.