Chapter 1

Performing Descriptive Methods Analyses

Background Information

Statistical techniques for analyzing time-series data can range from straightforward to very complex; however, the first step in an analysis is always to identify the characteristics of the data. Identifying the characteristics, such as seasonality, provides a clear picture of the data and helps determine any seasonal effects, cyclic changes, trends, errors, outliers, or turning points. In turn, this helps in choosing appropriate smoothing and forecasting methods.

STATGRAPHICS *Plus* provides various techniques you can use to adjust or transform data to improve the results of an analysis. The Descriptive Methods Analysis allows you to perform statistical calculations that reveal correlations in the data, to test for nonrandomness of the data, and to produce various plots that reveal trends and cycles as well as data errors and outliers.

Some of the methods available to help you analyze time-series data are the following.

Autocorrelations

Autocorrelations help to determine whether the data are random or have a pattern. STATGRAPHICS *Plus* calculates the correlation coefficients in a time-series variable and values for earlier time periods as well. You can display a table of estimated autocorrelations, standard errors, and upper and lower probability limits for each lag. For a single time series, it is very useful to compare the observation at one time period with the observation at another time period; similarly, it is possible to compare the series with itself, lagged two periods, three periods, and so on. You can also display the autocorrelation as a graph, which makes it easier to determine if the data have a pattern. In the graph, the height of the bars represents the amount of correlation.

Tests for Randomness

You use tests for randomness to obtain metrics that indicate whether data are random or nonrandom. STATGRAPHICS *Plus* performs two runs tests and a Box-Pierce test. The Runs Above and Below the Median test counts the number of runs that are completely above or completely below the median and ignores the values that are equal to the median. If the test statistic is large (corresponds to a *p*-value less than .05), you can conclude that the values occur in a nonrandom order. The Runs Up and Down test counts the number of times a sequence rises or falls. This test is sensitive to long-term cycles. If the test statistic is small (corresponds to a *p*-value greater than .05), you can conclude that the values occur in a random order. The Box-Pierce test verifies whether the autocorrelation is equal to 0. If the test statistic is large (corresponds to a *p*-value less than .05), the autocorrelation is not equal to 0.

Time Sequence Plots

A Horizontal Time Sequence Plot creates a connected lineplot for the time-series variable against time. The plot shows important features such as trend, seasonality, discontinuity, and outliers. The existence of a trend in the data means that successive values will be positively correlated. Seasonality is a pattern that repeats itself over fixed intervals of time. Discontinuity is an abrupt change or shift in a pattern, and outliers are extreme values in the data.

Periodogram Table and Plot

A periodogram is a plot of frequency and ordinate pairs for a specific time period you use to construct a frequency spectrum. A periodogram can be helpful in identifying randomness and seasonality in time-series data, and in recognizing the predominance of negative or positive autocorrelation — help that you often need to identify an appropriate model for forecasting a given time series. If the periodogram contains one spike, the data may not be random and you may need to use a periodogram table to determine the seasonality. A periodogram table includes a column of ordinates and a column of periods. The table also shows the cumulative sum and integrated periodograms for each period.

To access the analysis, choose SPECIAL... TIME-SERIES ANALYSIS... DESCRIPTIVE METHODS... from the Menu bar to display the Descriptive Methods Analysis dialog box (see Figure 1-1).

Tabular Options

Analysis Summary

The Analysis Summary option displays a summary after the program constructs various statistics and plots for the variable. The summary includes the name of the variable, the number of observations, the start index, and the sampling interval (see Figure 1-2). If you provided estimates for missing values, the summary also displays those values.

If you made adjustments to the data using either the Descriptive Methods Analysis dialog box or the Adjustment Options dialog box, the summary lists those adjustments.

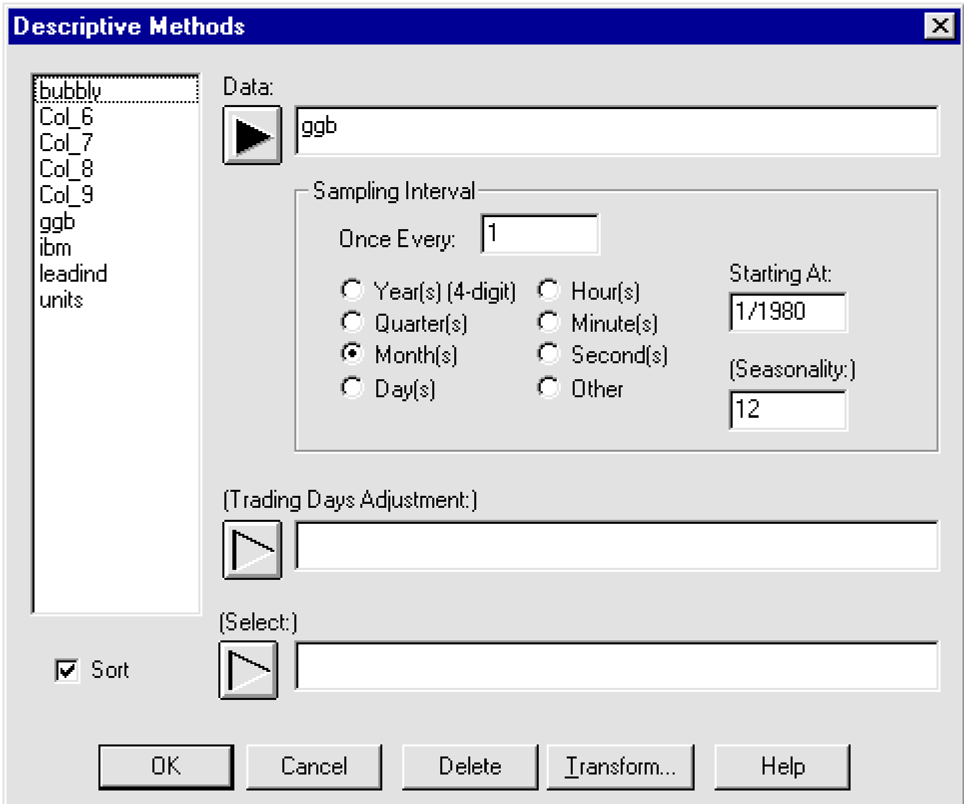


Figure 1-1. Descriptive Methods Analysis Dialog Box

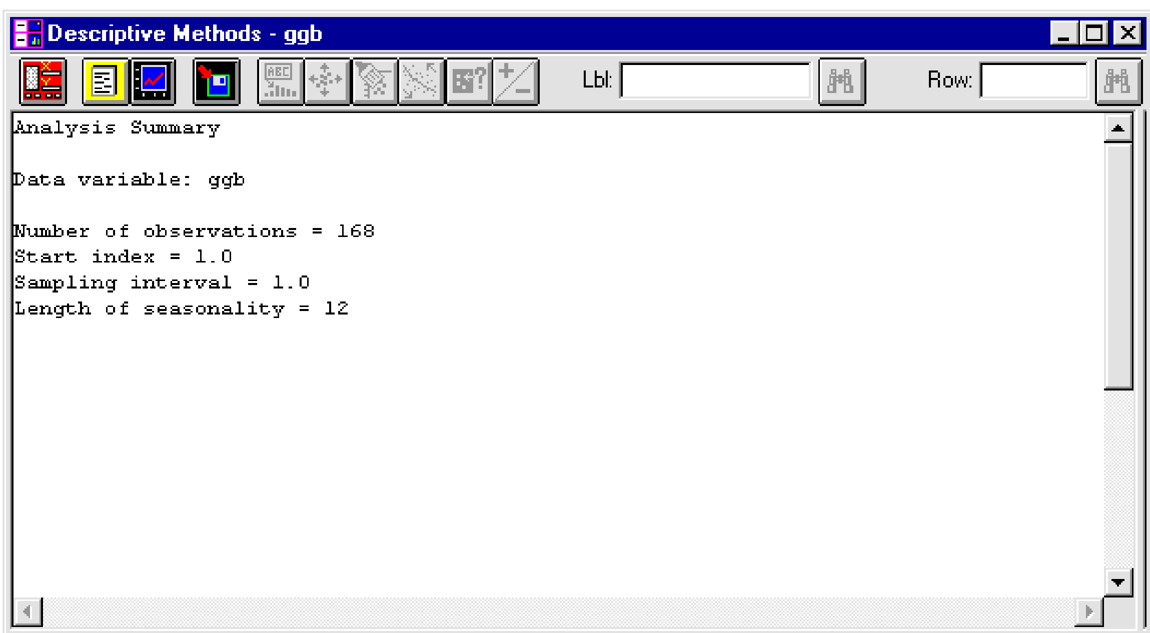


Figure 1-2. Analysis Summary

There are five types of data adjustments. If you choose more than one adjustment, the data are adjusted in this order: inflation, math, seasonal, trend, and differencing.

Use the *Adjustment Options* dialog box to choose options for each of the adjustments.

Data Table

The Data Table option displays a table that lists the adjusted values for each period and the values for the original data (see Figure 1-3). The values are adjusted according to the sampling interval and options you chose.

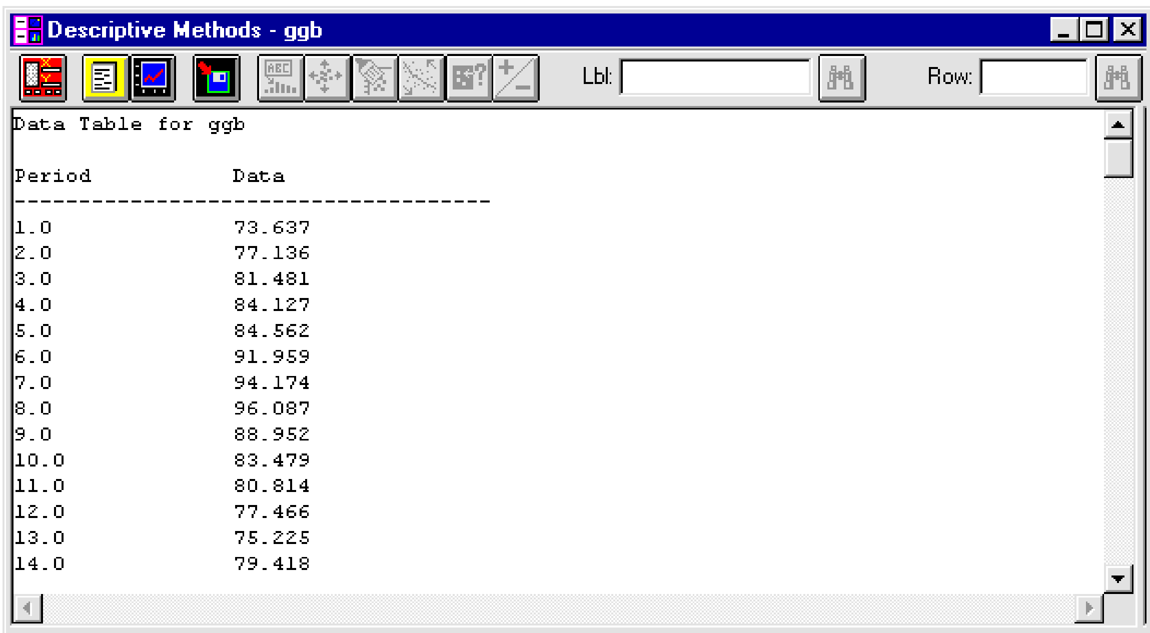
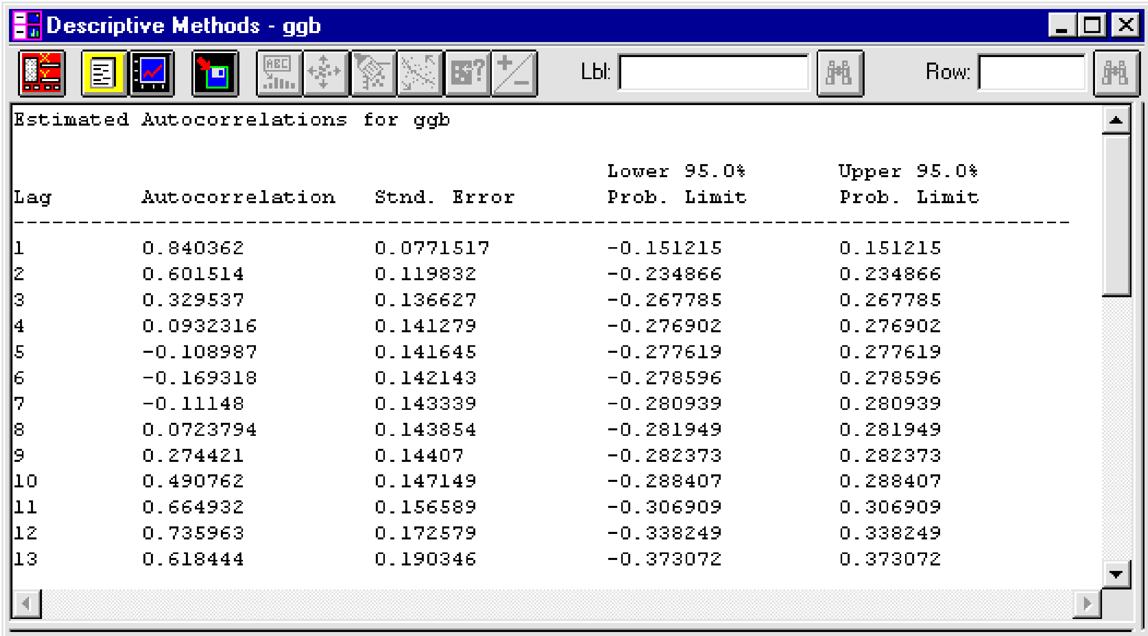


Figure 1-3. Data Table

Autocorrelations

The Autocorrelations option displays a table of the estimated autocorrelations between values of the variable at various time lags (see Figure 1-4). The lag k autocorrelation coefficient measures the correlation between values at time t and time $t - k$. If the probability limits at a particular lag does not contain the estimated coefficient, there is a statistically significant correlation at that lag. You can use the correlation coefficients to test for seasonal patterns or as a preliminary step in determining an appropriate model for the data.



Estimated Autocorrelations for ggb

Lag	Autocorrelation	Stnd. Error	Lower 95.0% Prob. Limit	Upper 95.0% Prob. Limit
1	0.840362	0.0771517	-0.151215	0.151215
2	0.601514	0.119832	-0.234866	0.234866
3	0.329537	0.136627	-0.267785	0.267785
4	0.0932316	0.141279	-0.276902	0.276902
5	-0.108987	0.141645	-0.277619	0.277619
6	-0.169318	0.142143	-0.278596	0.278596
7	-0.11148	0.143339	-0.280939	0.280939
8	0.0723794	0.143854	-0.281949	0.281949
9	0.274421	0.14407	-0.282373	0.282373
10	0.490762	0.147149	-0.288407	0.288407
11	0.664932	0.156589	-0.306909	0.306909
12	0.735963	0.172579	-0.338249	0.338249
13	0.618444	0.190346	-0.373072	0.373072

Figure 1-4. Autocorrelations

Use the *Autocorrelation Function Options* dialog box to enter values for the maximum number of lags for which the program should estimate autocorrelations and for the confidence level. Typically, the number of lags should not exceed the number of observations divided by 4.

Partial Autocorrelations

The Partial Autocorrelations option creates a table of the estimated partial autocorrelations between values at various lags (see Figure 1-5). The lag *k* partial autocorrelation coefficient measures the correlation between values at time *t* and time $t + k$, accounting for the correlations at all lower lags. The table is also useful in judging the order of an autoregressive model needed to fit the data. If the probability limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlation at that lag.

The table shows the partial autocorrelations, the standard errors, and the upper and lower probability limits for each lag. Use the table to test for seasonal patterns, or as a preliminary step in determining an appropriate ARIMA model for forecasting the data.

Use the *Partial Autocorrelation Function Options* dialog box to enter values for the maximum number of lags for which the program should estimate partial autocorrelations and for the confidence level. Typically, the number of lags should not exceed the number of observations divided by 4.

Estimated Partial Autocorrelations for ggb

Lag	Partial Autocorrelation	Stnd. Error	Lower 95.0% Prob. Limit	Upper 95.0% Prob. Limit
1	0.840362	0.0771517	-0.151215	0.151215
2	-0.356356	0.0771517	-0.151215	0.151215
3	-0.220746	0.0771517	-0.151215	0.151215
4	-0.0325883	0.0771517	-0.151215	0.151215
5	-0.132289	0.0771517	-0.151215	0.151215
6	0.294114	0.0771517	-0.151215	0.151215
7	0.174468	0.0771517	-0.151215	0.151215
8	0.309939	0.0771517	-0.151215	0.151215
9	0.0754963	0.0771517	-0.151215	0.151215
10	0.219082	0.0771517	-0.151215	0.151215
11	0.252541	0.0771517	-0.151215	0.151215
12	0.0188762	0.0771517	-0.151215	0.151215
13	-0.267034	0.0771517	-0.151215	0.151215

Figure 1-5. Partial Autocorrelations

Periodogram Table

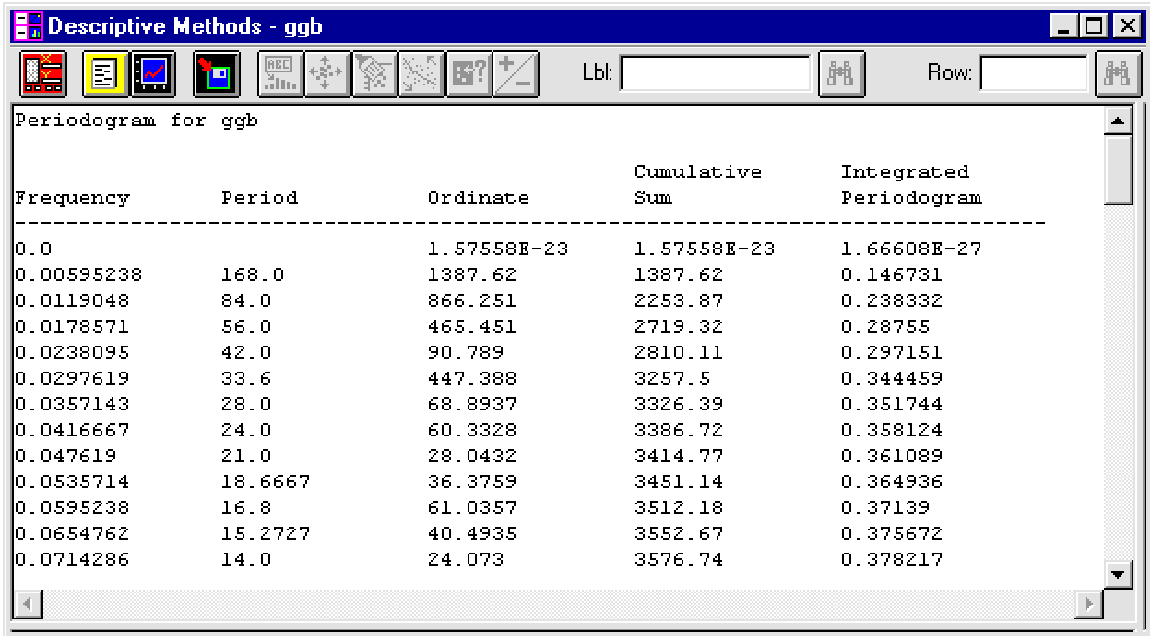
The Periodogram Table option creates a table that shows the frequency and ordinate pairs for a period, as well as the cumulative sum and integrated periodograms (see Figure 1-6). The table is often used to identify cycles of fixed frequency while the pairs are used to plot the amplitudes and frequencies estimated from the data.

A periodogram is constructed by fitting a series of sine functions at each frequency. The ordinates are equal to the squared amplitudes of the sine functions. Because the sum of the ordinates equals the total corrected sum of squares in an ANOVA Table, a periodogram is often thought of as an analysis of variance by frequency.

Use the *Periodogram Table Options* dialog box to indicate if the mean should be subtracted from each observation and if the beginning and end of the time-series data should be tapered by using a split cosine-bell window. For example, if you enter a value of 10, the first and last 5 percent of the data are tapered. See Bloomfield (1976) for a discussion of the tapering method.

Tests for Randomness

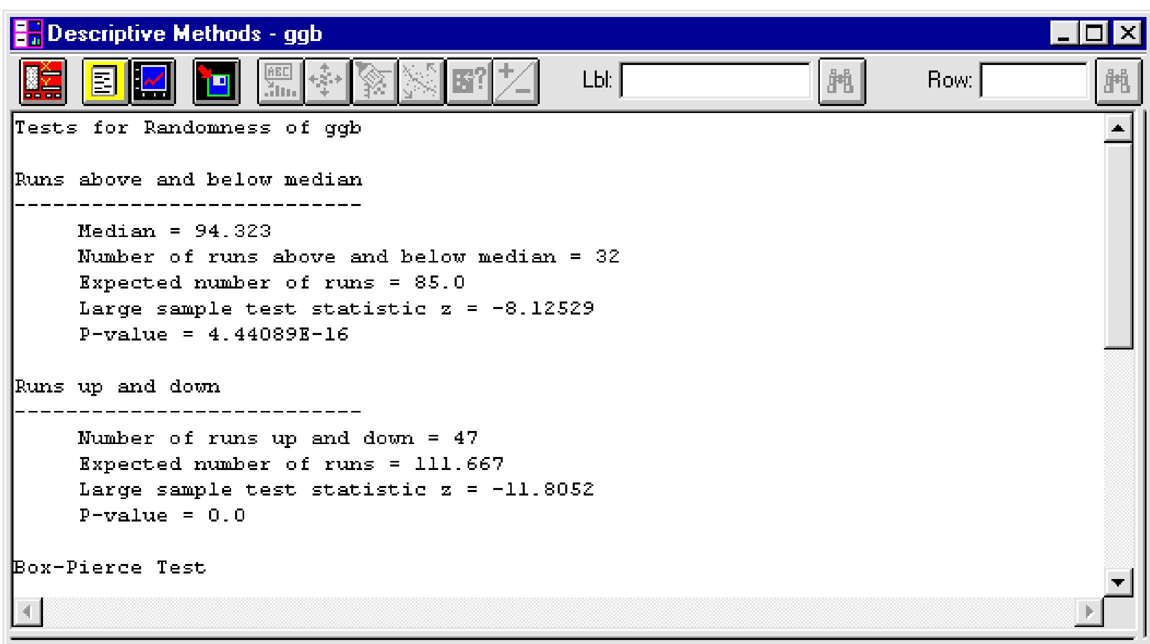
The Tests for Randomness option creates the results of two runs tests and a Box-Pierce test, which determine if the chosen variable is a random sequence of numbers (see Figure 1-7). A time series of random numbers is often called *white noise* because it contains equal contributions at many frequencies.



Periodogram for ggb

Frequency	Period	Ordinate	Cumulative Sum	Integrated Periodogram
0.0		1.57558E-23	1.57558E-23	1.66608E-27
0.00595238	168.0	1387.62	1387.62	0.146731
0.0119048	84.0	866.251	2253.87	0.238332
0.0178571	56.0	465.451	2719.32	0.28755
0.0238095	42.0	90.789	2810.11	0.297151
0.0297619	33.6	447.388	3257.5	0.344459
0.0357143	28.0	68.8937	3326.39	0.351744
0.0416667	24.0	60.3328	3386.72	0.358124
0.047619	21.0	28.0432	3414.77	0.361089
0.0535714	18.6667	36.3759	3451.14	0.364936
0.0595238	16.8	61.0357	3512.18	0.37139
0.0654762	15.2727	40.4935	3552.67	0.375672
0.0714286	14.0	24.073	3576.74	0.378217

Figure 1-6. Periodogram Table




```
Tests for Randomness of ggb

Runs above and below median
---------------------------
     Median = 94.323
     Number of runs above and below median = 32
     Expected number of runs = 85.0
     Large sample test statistic z = -8.12529
     P-value = 4.44089E-16

Runs up and down
---------------------------
     Number of runs up and down = 47
     Expected number of runs = 111.667
     Large sample test statistic z = -11.8052
     P-value = 0.0

Box-Pierce Test
```

Figure 1-7. Tests for Randomness

The first test counts the number of times the sequence was above or below the median; the second counts the number of times the sequence rose or fell; and the third test is based on the sum of squares of the first *n* autocorrelation coefficients. Because all three tests are sensitive to different types of departure from random behavior, failure to pass any of the tests indicates that the time series may not be completely random.

Crosscorrelations

The Crosscorrelations option creates a table of results used to estimate the crosscorrelation at lag *k*. The estimate is created by measuring the strength of the linear relationship between the value of the first time series at time *t* and the value of the other time series *k* periods earlier (see Figure 1-8). The table is used to determine if the second time series would help forecast the first.

Lag	Crosscorrelation
-24	0.244508
-23	0.243752
-22	0.251946
-21	0.263876
-20	0.275036
-19	0.287915
-18	0.304905
-17	0.318167
-16	0.330149
-15	0.338919
-14	0.345775
-13	0.351421
-12	0.367495
-11	0.381429

Figure 1-8. Crosscorrelations

Use the *Crosscorrelations Function* dialog box to choose the variable that contains the second set of time-series data and to enter the maximum number of lags necessary to calculate the crosscorrelations.

Graphical Options

Horizontal Time Sequence Plot
The Horizontal Time Sequence option creates a connected line plot of the time series against time (see Figure 1-9). The plot shows important features such as trend, seasonality, discontinuity, and outliers.

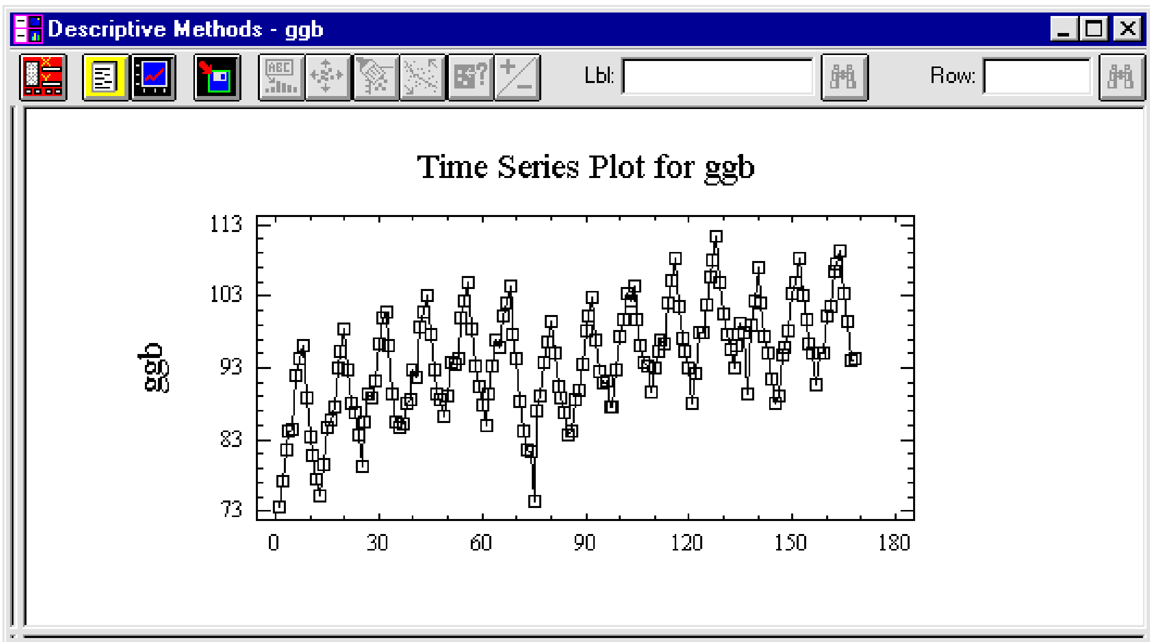


Figure 1-9. Horizontal Time Sequence Plot

Use the *Horizontal Time Sequence Plot Options* dialog box to indicate if points and lines will appear on the plot.

Vertical Time Sequence Plot
The Vertical Time Sequence option creates a plot of the time-series variable in relation to the baseline (see Figure 1-10). The plot contains a vertical line that extends from the baseline to each value. Use the plot to identify the values that are above and below the baseline.

Use the *Vertical Time Sequence Plot Options* dialog box to enter a value for the point at which the baseline is to be drawn.

Autocorrelation Function
The Autocorrelation Function option creates a plot of the estimated autocorrelations between values at various lags (see Figure 1-11). The lag *k* autocorrelation coefficient measures the correlation between values at time *t* and time *t* - *k*. Also

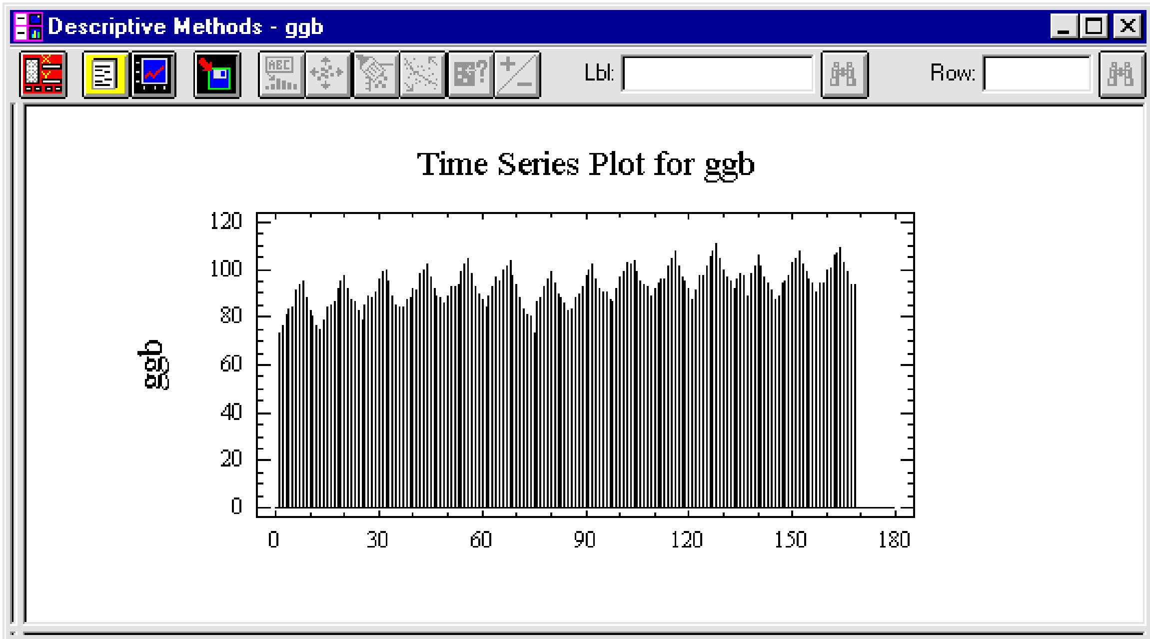


Figure 1-10. Vertical Time Sequence Plot

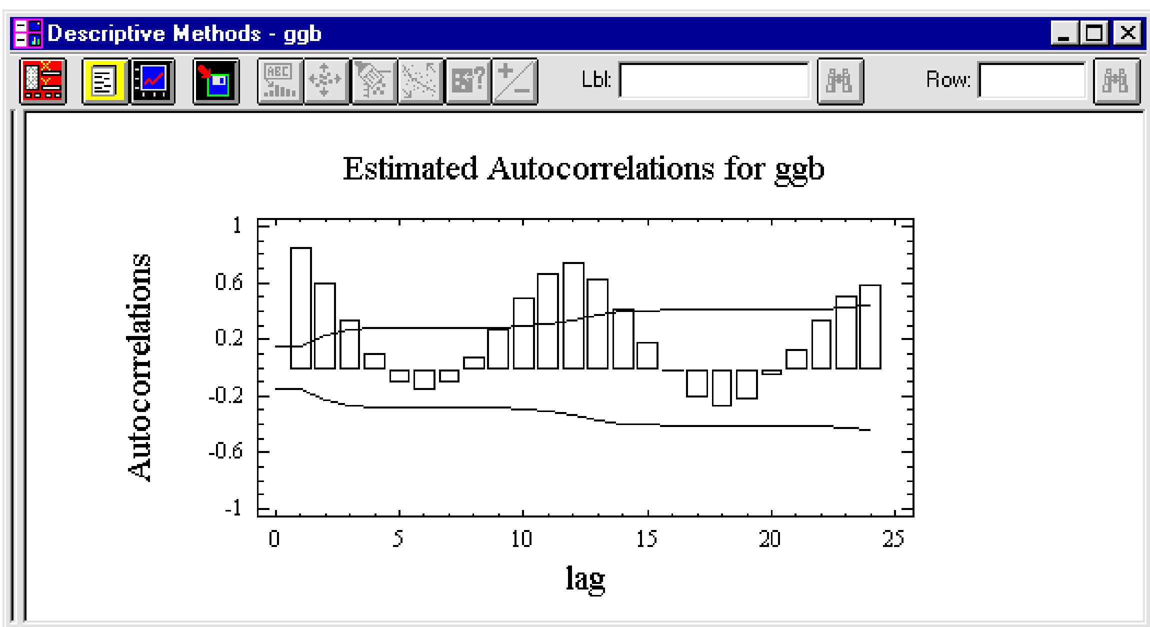


Figure 1-11. Autocorrelation Function Plot

shown in the plot are the probability limits. If the probability limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlation at that lag.

The plot contains vertical bars that represent the coefficient for each lag and a pair of dotted lines at a distance from the baseline that are a multiple of the standard error at each lag. Statistically significant coefficients appear as bars that extend beyond either line. The plot is useful for testing seasonality or other cycles.

Use the *Autocorrelation Function Options* dialog box to enter the maximum number of lags for which the autocorrelations will be estimated and to enter a value for the confidence level.

Partial Autocorrelation Function

The Partial Autocorrelation Function option creates a plot of the estimated partial autocorrelations among values at various lags (see Figure 1-12). The lag *k* partial autocorrelation coefficient measures the correlation among values of lags at time *t* and time *t* + *k* after accounting for the correlations at all the lower lags. If the probability limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlation at that lag.

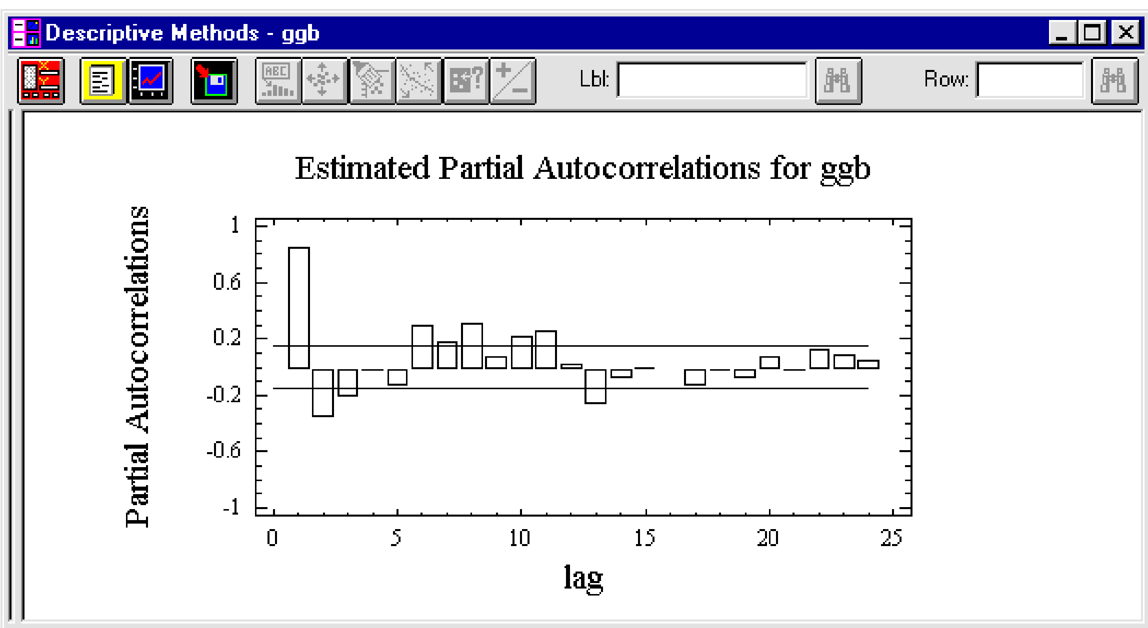


Figure 1-12. Partial Autocorrelation Function Plot

Each coefficient is plotted as a bar whose height is proportional to the value of the coefficient. By default, the probability limits are plotted as dashed lines at plus or minus 2, divided by the square root of the number of observations in the time series.

The bounds are useful for indicating partial autocorrelations that are significantly different from 0. Bars that extend beyond either line indicate significant correlations. The plot is useful for judging the order of autogressive model that is needed to fit the data.

Use the *Partial Autocorrelation Function Options* dialog box to enter the maximum number of lags for which the partial autocorrelations will be estimated and to enter a value for the confidence level.

Periodogram

The Periodogram option displays a plot that uses the ordinates to construct a frequency spectrum (see Figure 1-13). The plot is constructed by fitting a series of sine functions at each of the frequencies. The ordinates are equal to the squared amplitudes of the sine functions.

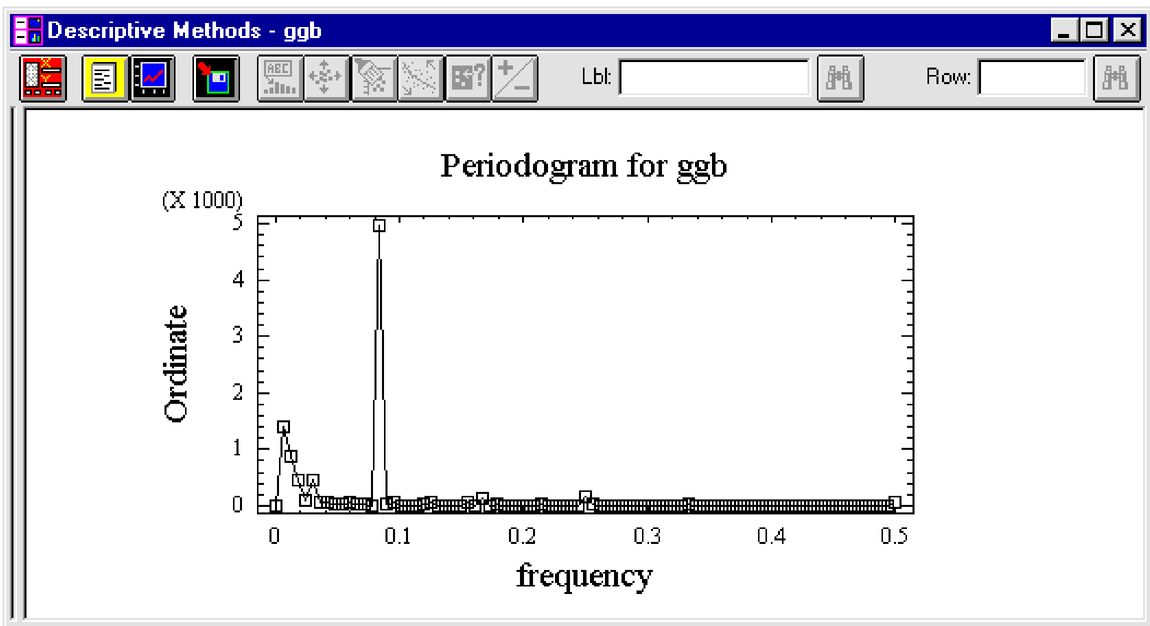


Figure 1-13. Periodogram

A Periodogram can also be thought of as an analysis of variance by frequency because the sum of the ordinates equals the total sum of squares in an ANOVA table. The plot is used to identify cycles of fixed frequency in the data.

Use the *Periodogram Options* dialog box to indicate if the mean should be subtracted from each observation, if the observations should be plotted as points, if lines should connect the observations, and if the beginning and end of the time-series data should be tapered using a split cosine-bell window. See Bloomfield (1976) for a discussion of tapering.

Integrated Periodogram

The Integrated Periodogram option creates a plot of the cumulative sum of the periodogram ordinates (see Figure 1-14). Also shown are the Kolmogorov-Smirnov bounds. If the variable is a purely random time series, the integrated periodogram should fall approximately along the diagonal line. If it crosses the outer 99 percent bounds, the hypothesis that the series is random can be rejected at the 99 percent confidence level.

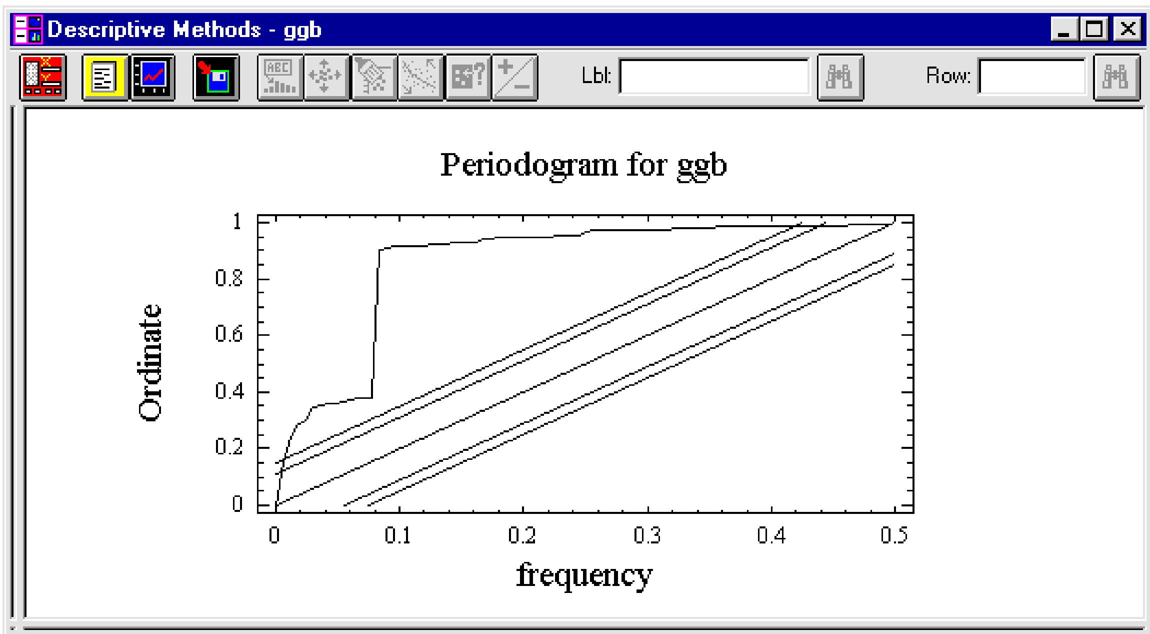


Figure 1-14. Integrated Periodogram

Use the *Integrated Periodogram Options* dialog box to indicate if the mean should be subtracted from each observation and if the beginning and end of the time-series data should be tapered using a split cosine-bell window. See Bloomfield (1976) for a discussion of tapering.

Crosscorrelation Function

The Crosscorrelation Function option creates a plot of the estimated correlations between a time-series variable at time *t* and a second time-series variable at time *t* + *k* as a function of the lag or time differential *k* (see Figure 1-15).

Use the *Crosscorrelation Function* dialog box to enter the second set of time-series data and to enter the maximum number of lags for which the crosscorrelation function should be calculated.

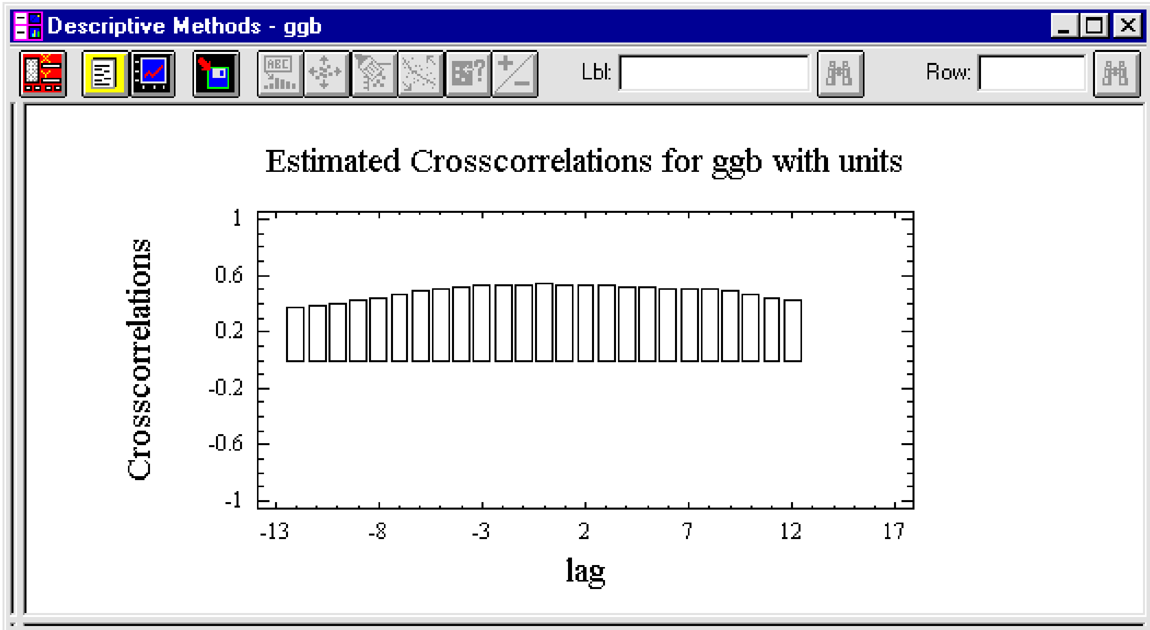


Figure 1-15. Crosscorrelation Function

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are six selections: Data, Adjusted Data, Autocorrelations, Partial Autocorrelations, Periodogram Ordinates, Fourier Frequencies.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Bloomfield, P. 1976. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.

Box, G. E. P. and Jenkins, G. M. 1976. *Time Series Analysis, Forecasting, and Control*, second edition. San Francisco: HoldenDay.

Guttman, L., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Makridakis, W., Wheelwright, S. C., and McGee, U. E. 1983. *Forecasting: Methods and Applications*, second edition. New York: Wiley.

Mendenhall, W. and Reinmuth, J. E. 1974. *Statistics for Management and Economics*, second edition. North Scituate, Massachusetts: Duxbury Press.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGrawHill.

Chapter 2

Performing Smoothing Analyses

Background Information

If time-series data involve a trend (an upward or downward direction) or a seasonal effect (for example, strong sales of disposable tissues in winter months), or both a trend and a seasonal effect, you need to use a smoothing method on the data. Smoothing methods move through the known data, period by period, as opposed to using all the past data in one model-fitting exercise.

When you smooth data you combine observations, which tends to reduce randomness by allowing the positive and negative random effects to partially offset each other. For example, when you analyze and plot time-series data, the values do not always follow a straight line. Often the scatterplot you produce will have points that are scattered more or less randomly, which makes it difficult to see cycles and trends. To eliminate the randomness in the data pattern, you can combine two or more observations from periods during which the causal factors were in effect to achieve a smoothed value. The plot you create after smoothing the data will produce a curve that more clearly shows the data's general pattern.

The Smoothing Analysis in STATGRAPHICS *Plus* applies up to two consecutive smoothing methods to time-series data. After smoothing has been applied, plots are created to reveal trends and cycles as well as data errors, outliers and residuals. The analysis also provides several methods used to adjust or transform data to make it more appropriate for an evaluation.

To access the analysis, choose SPECIAL... TIME-SERIES ANALYSIS... SMOOTHING... from the Menu bar to display the Smoothing Analysis dialog box shown in Figure 2-1.

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the name of the variable, the number of observations, the start index, the sampling interval, and the names of the two smoothers you choose on the Smoothing Options dialog box (see Figure 2-2). If you make adjustments to the data, the summary lists those adjustments.

There are five types of data adjustments. If you choose more than one adjustment, they are adjusted in this order: inflation, math, seasonal, trend, and differencing.

In a trend adjustment, the trend is estimated and removed using unweighted regression analysis. The analysis fits a linear or quadratic line, exponential power

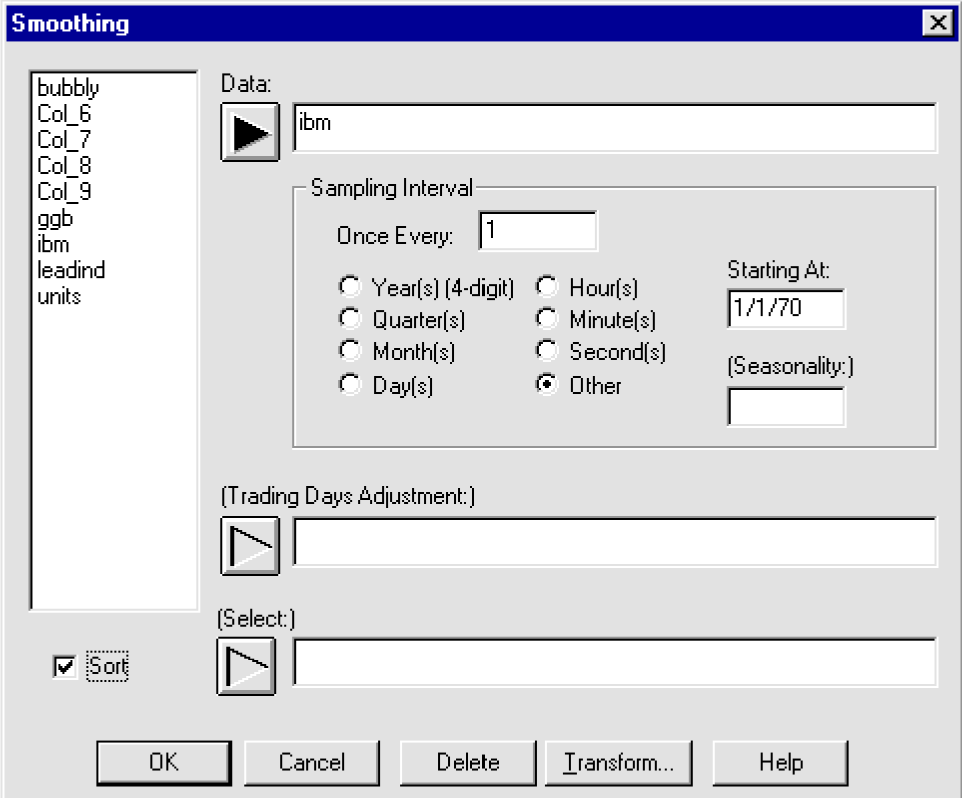


Figure 2-1. Smoothing Analysis Dialog Box

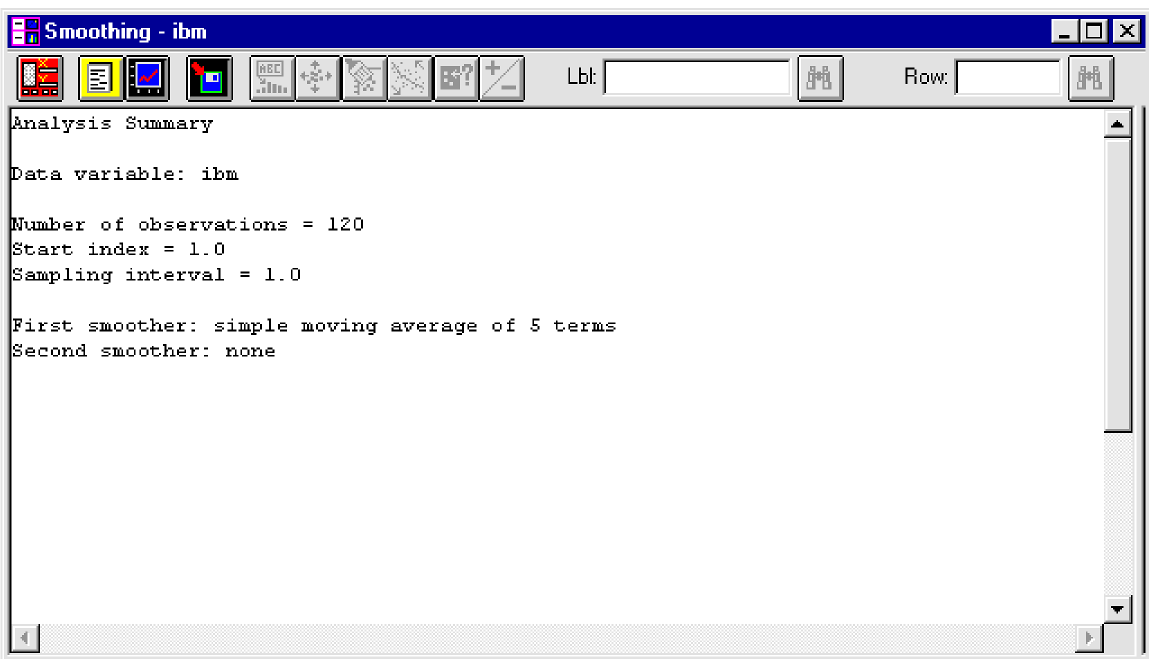


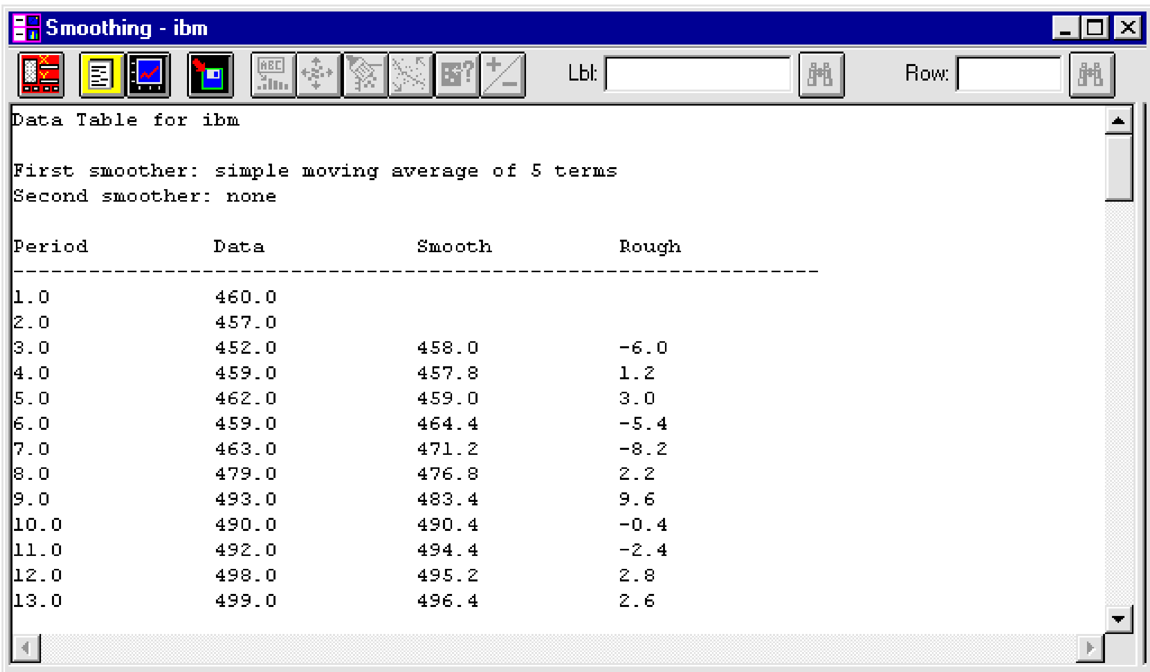
Figure 2-2. Analysis Summary

curve, or an S-curve through the time-series data. Each of the four fitted models estimates a function Z(*t*), where *t* represents the time index.

Use the *Adjustment Options* dialog box to choose the type of adjustment and adjustment options that will be applied to the data.

Data Table

The Data Table option creates a table that shows the results of applying the chosen smoother to the data. The table shows the values for the original data and the smoothed and rough residuals for each period (see Figure 2-3).



Data Table for ibm

First smoother: simple moving average of 5 terms
Second smoother: none

Period	Data	Smooth	Rough
1.0	460.0		
2.0	457.0		
3.0	452.0	458.0	-6.0
4.0	459.0	457.8	1.2
5.0	462.0	459.0	3.0
6.0	459.0	464.4	-5.4
7.0	463.0	471.2	-8.2
8.0	479.0	476.8	2.2
9.0	493.0	483.4	9.6
10.0	490.0	490.4	-0.4
11.0	492.0	494.4	-2.4
12.0	498.0	495.2	2.8
13.0	499.0	496.4	2.6

Figure 2-3. Data Table

Use the *Smoothing Options* dialog box to choose either Smoother 1 or 2 and its corresponding options, to enter values for the length of the moving average, and the EWMA smoothing constant for the chosen smoother.

Graphical Options

Time Sequence Plot

The Time Sequence Plot option creates a connected line plot of the chosen smoothers applied to the variable (see Figure 2-4). Important features such as trend, seasonality, discontinuity, and outliers are shown on the plot.

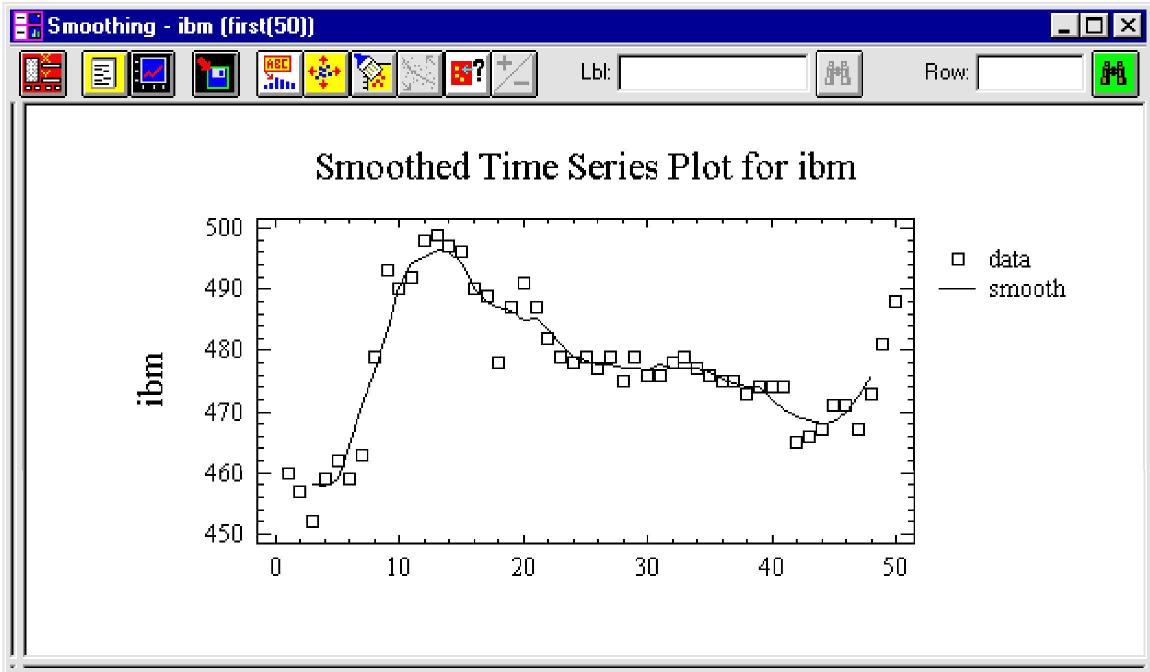


Figure 2-4. Time Sequence Plot

Residual Plot
The Residual Plot option creates a plot of the residuals that were calculated as a result of the smoothing (see Figure 2-5). The residuals are points that fall outside a smoothed curve (the roughs).

Saving the Results

The Save Results Options dialog box allows you to choose the results that will be saved. There are four selections: Data, Adjusted Data, Smooth, and Rough.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Kendall, M. G. and Stuart, A. 1961. *The Advanced Theory of Statistics*, Volume 2. New York: Hafner Publishing Company.

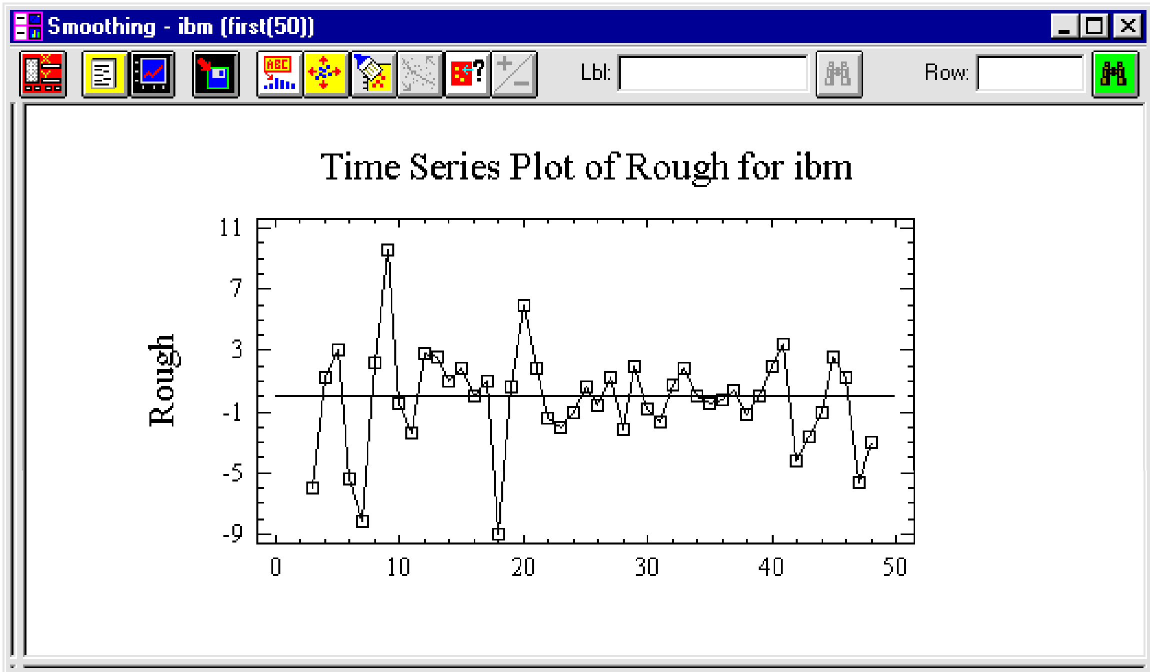


Figure 2-5. Residual Plot

Makridakis, W., Wheelwright, S. C., and McGee, U. E. 1983. *Forecasting: Methods and Applications*, second edition. New York: Wiley.

Tukey, J. W. 1977. *Exploratory Data Analysis.* Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston: PWS-Kent Publishing Company.

Chapter 3

Performing Seasonal Decomposition Analyses

Background Information

The time-series decomposition approach to forecasting is based on the principle of "breaking down" a time series into each of its components and then forecasting by predicting each component separately, except randomness, which you cannot predict (Makridakis et al., 1983).

The Seasonal Decomposition Analysis in STATGRAPHICS *Plus* allows you to perform a classical decomposition of the data. Classical decomposition breaks the time-series data into four components: trend, seasonality, cycle, and irregular.

- Trend, which is the long-term behavior of the data.

- Seasonality, which is the periodic fluctuation of constant lengths that usually repeats itself at fixed intervals.

- Cycle, which is a wavelike pattern of ups and downs that are similar to seasonality. Cycles can take place over a long time period such as several years. The length of time between peaks is not fixed as it is with seasonal movement.

- Irregular, which is movement in the data that cannot be attributed to trend, seasonality, or cycle. This component is often known as *noise*.

You can choose either a multiplicative or an additive model to adjust or transform the data to make it more appropriate for the analysis you are performing.

To access the analysis, choose SPECIAL... TIME-SERIES ANALYSIS... SEASONAL DECOMPOSITION... from the menu bar to display the Seasonal Decomposition Analysis dialog box shown in Figure 3-1.

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis whose purpose is to apply a multiplicative seasonal decomposition to the chosen variable (see Figure 3-2). The decomposition separates the variable into trend-cycle, seasonal, and random components.

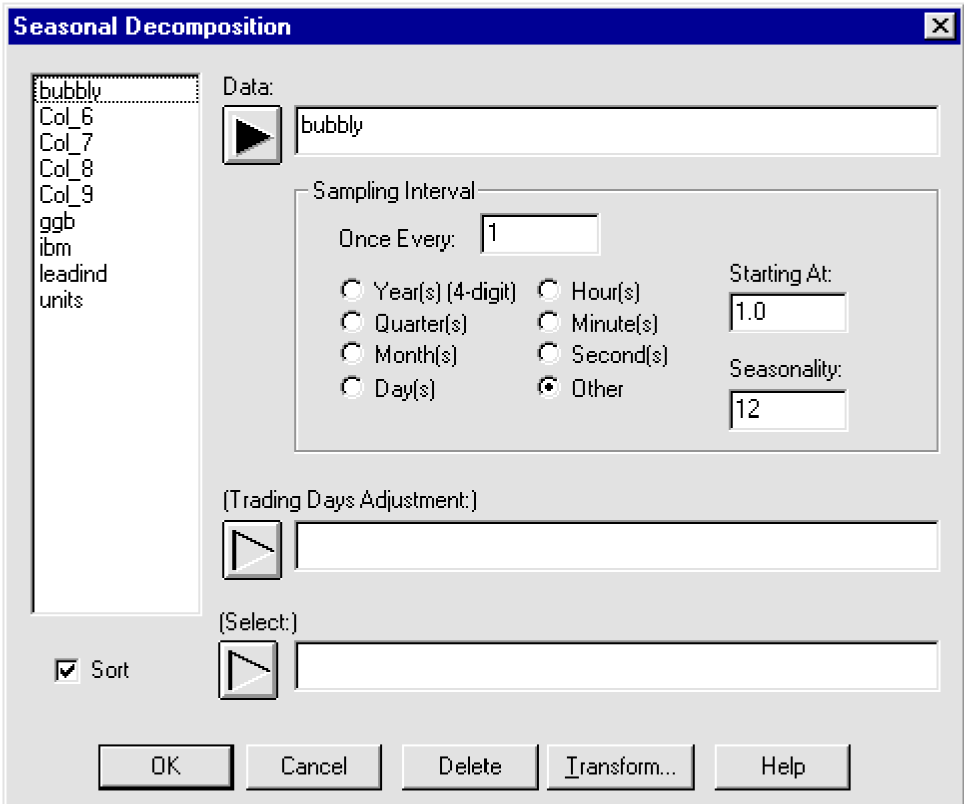


Figure 3-1. Seasonal Decomposition Analysis Dialog Box

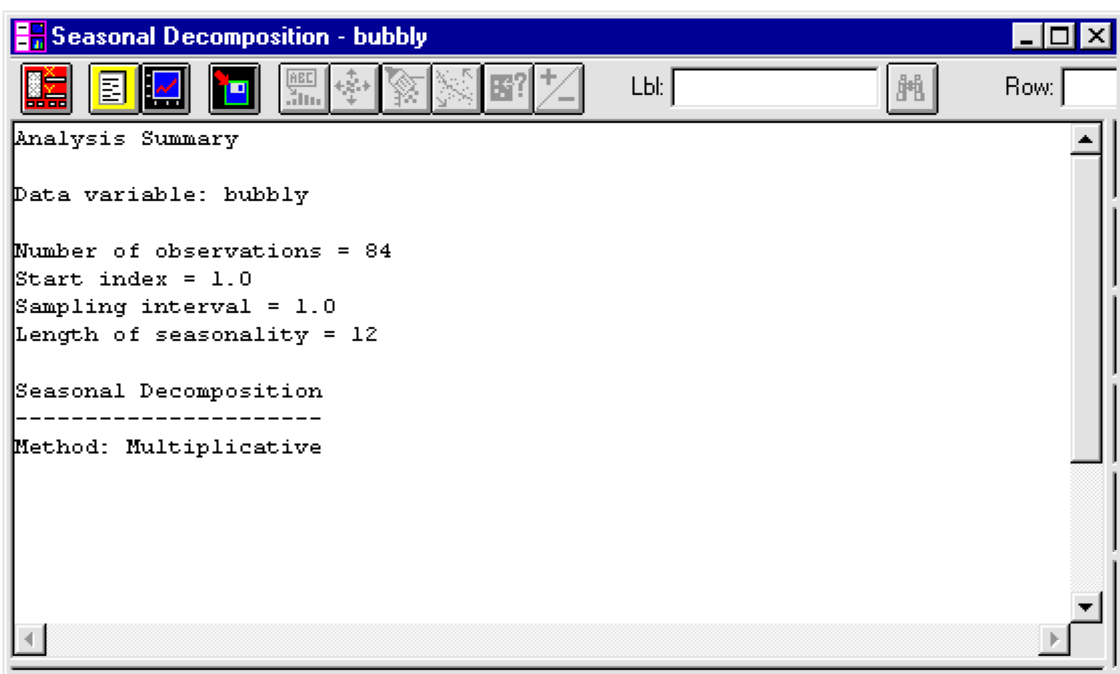


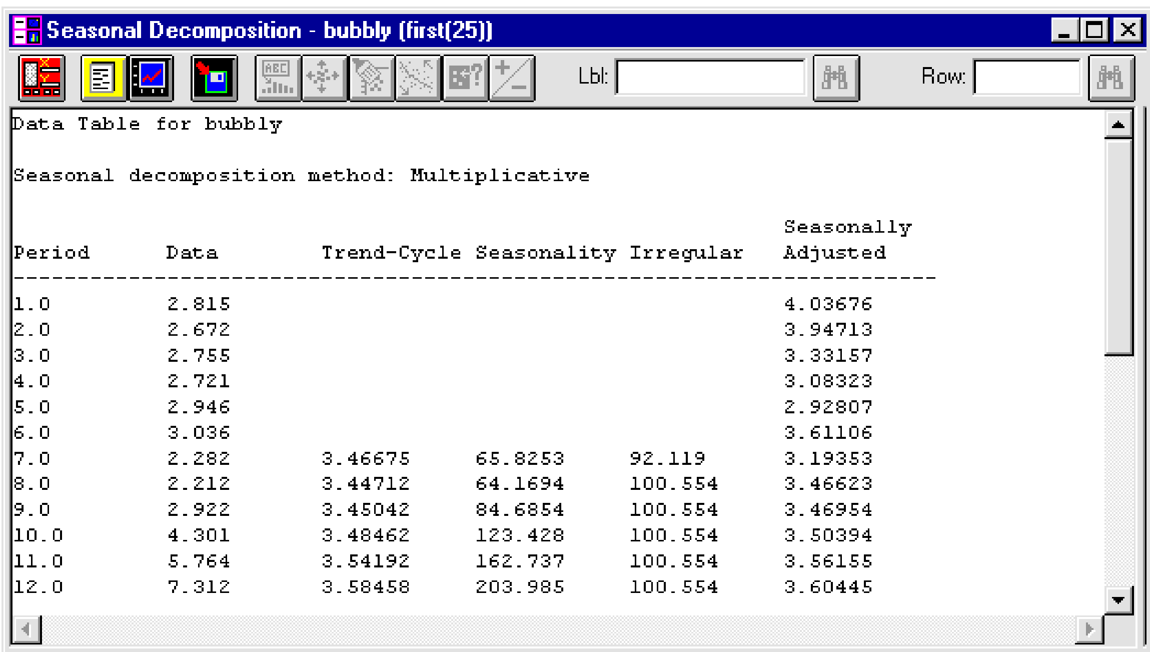
Figure 3-2. Analysis Summary

The summary shows the name of the variable, the number of observations, the start index, the sampling interval, and the method of decomposition you are using. If you make adjustments to the data, the summary lists those adjustments.

Use the *Adjustment Options* dialog box to choose an adjustment method and its options.

Data Table

The Data Table option creates a table that shows each step of the seasonal decomposition (see Figure 3-3). The Trend-Cycle column displays the results of a centered moving average of the seasonal length applied to the variable. The Seasonality column displays the data divided by the moving average and multiplied by 100. The program then calculates the indices for each season by averaging the ratios across all the observations in that season, then scaling the indices so an average season equals 100. The data are then divided by the trend-cycle and the seasonal estimates to give the irregular or residual component and the component is then multiplied by 100.



Data Table for bubbly

Seasonal decomposition method: Multiplicative

Period	Data	Trend-Cycle	Seasonality	Irregular	Seasonally Adjusted
1.0	2.815				4.03676
2.0	2.672				3.94713
3.0	2.755				3.33157
4.0	2.721				3.08323
5.0	2.946				2.92807
6.0	3.036				3.61106
7.0	2.282	3.46675	65.8253	92.119	3.19353
8.0	2.212	3.44712	64.1694	100.554	3.46623
9.0	2.922	3.45042	84.6854	100.554	3.46954
10.0	4.301	3.48462	123.428	100.554	3.50394
11.0	5.764	3.54192	162.737	100.554	3.56155
12.0	7.312	3.58458	203.985	100.554	3.60445

Figure 3-3. Data Table

Use the *Seasonal Decomposition Options* dialog box to choose either the multiplicative or the additive method, which will be used to combine the component factors.

Seasonal Indices

The Seasonal Indices option creates a table that shows the indices for each season scaled so an average season equals 100 (see Figure 3-4). The indices in the table are all the ranges in the season; they indicate the average seasonal swing throughout the course for one complete cycle. A seasonally adjusted value is the actual value divided by the seasonal index.

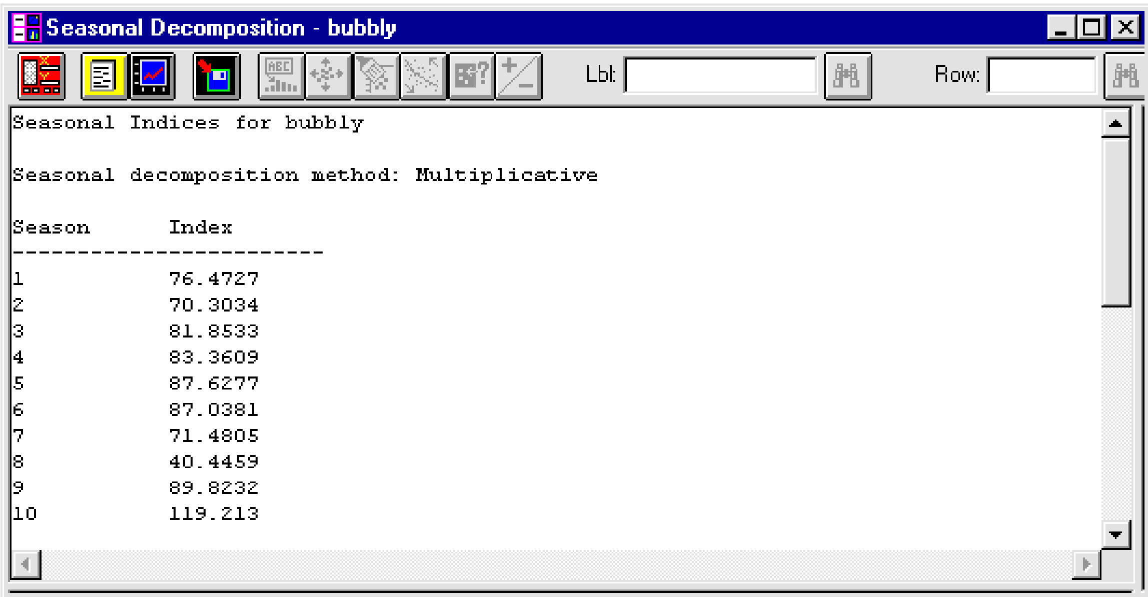


Figure 3-4. Seasonal Indices

Graphical Options

Trend-Cycle

The Trend-Cycle option creates a plot of the estimated trend-cycle, which is calculated from the centered moving average of the seasonality length applied to the chosen variable (see Figure 3-5). The trend-cycle represents the long-term fluctuations in the data.

Seasonal Indices

The Seasonal Indices option creates a plot of the seasonal indices for each season scaled so an average season equals 100 (see Figure 3-6). The indices indicate all the ranges in the season, which is the average seasonal swing throughout the course of one complete cycle.

If you choose the Multiplicative method from the Seasonal Decomposition Options dialog box, the indices are based on the average ratios of the moving average. If you choose the Additive method, they are based on the average differences.

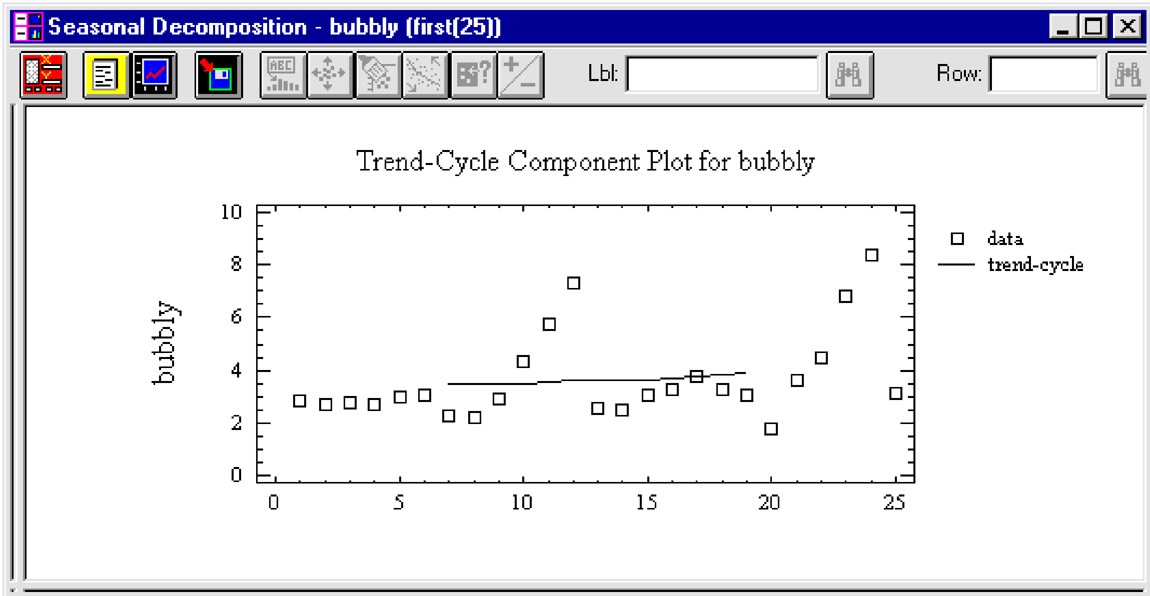


Figure 3-5. Trend-Cycle Plot

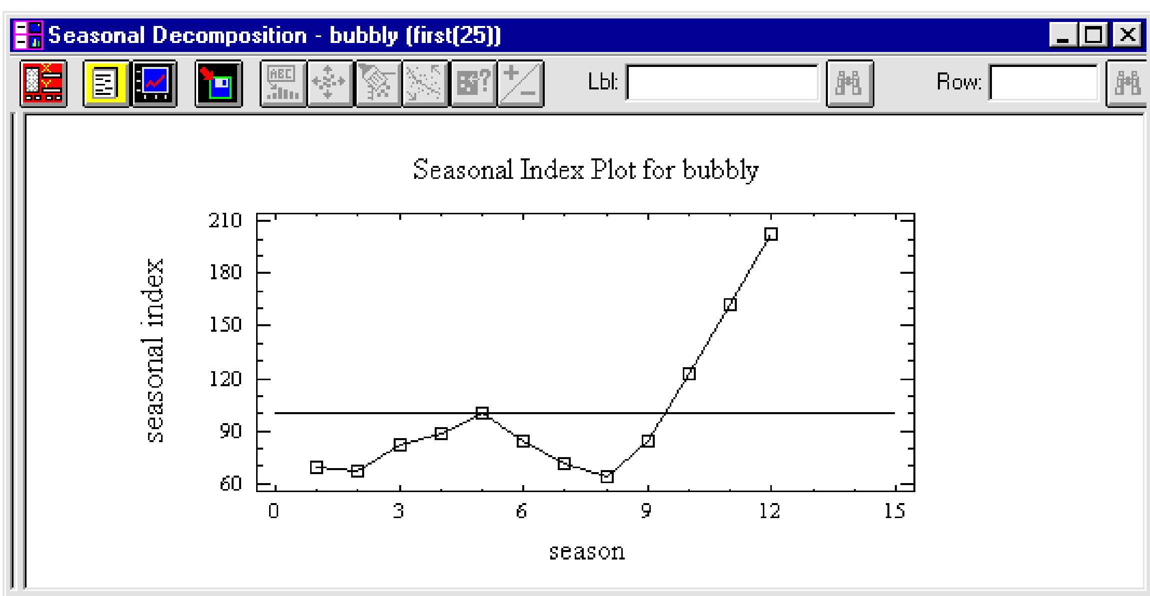


Figure 3-6. Seasonal Indices Plot

Irregular Component

The Irregular Component option creates a plot of the irregular or residual component that was left behind when the trend-cycle and seasonal components were removed (see Figure 3-7). The plot is scaled so the average residual equals 100.

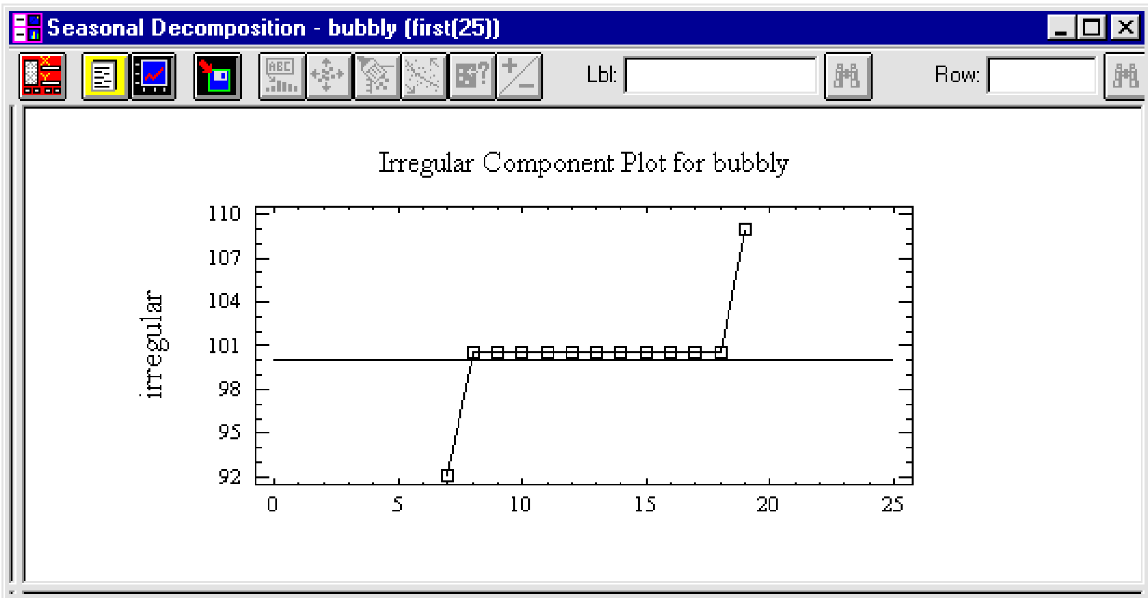


Figure 3-7. Irregular Component Plot

Depending on the method you choose from the *Seasonal Decomposition Options* dialog box, the program either divides each value by its index or subtracts the index from each value. The resulting spikes represent the irregular or residual component of the data; in other words, movement that cannot be attributed to trend, seasonality, or cycle.

Seasonally Adjusted Data

The Seasonally Adjusted Data option creates a plot of the seasonally adjusted values from the Data Table, which are the actual values divided by the seasonality indices (see Figure 3-8).

Use the *Seasonal Decomposition Options* dialog box to choose either the multiplicative or the additive method, which will be used to combine the component factors.

Seasonal Subseries Plot

The Seasonal Subseries option creates a plot of the time-series data arranged by season (see Figure 3-9). Horizontal lines represent the average values of the observations for each season. Vertical lines represent the actual values in the season.

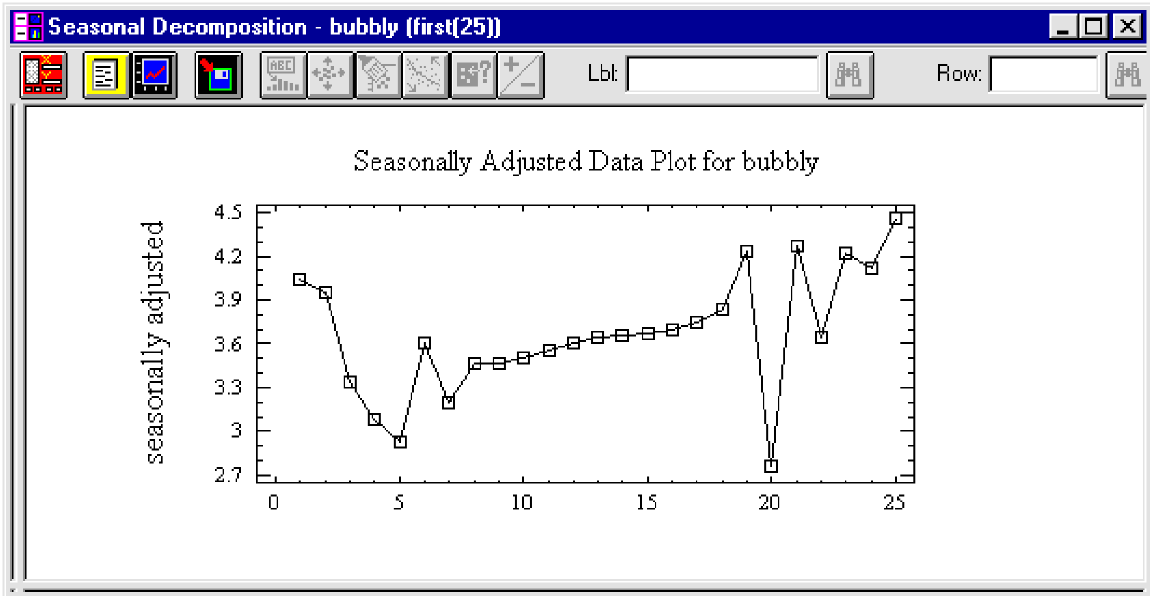


Figure 3-8. Seasonally Adjusted Data

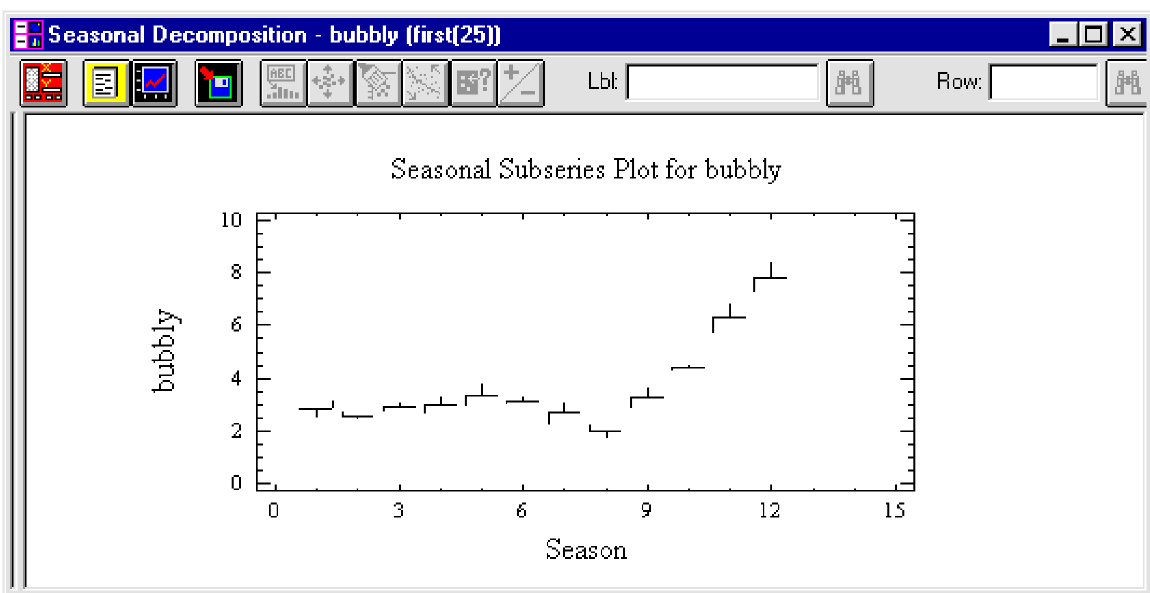


Figure 3-9. Seasonal Subseries Plot

Use the *Seasonal Subseries Plot Options* dialog box to indicate how the points will be plotted.

Annual Subseries Plot

The Annual Subseries option creates a plot of the values of the time-series data arranged by cycle (see Figure 3-10). If the data are monthly with seasonality length of 12, each cycle represents one year.

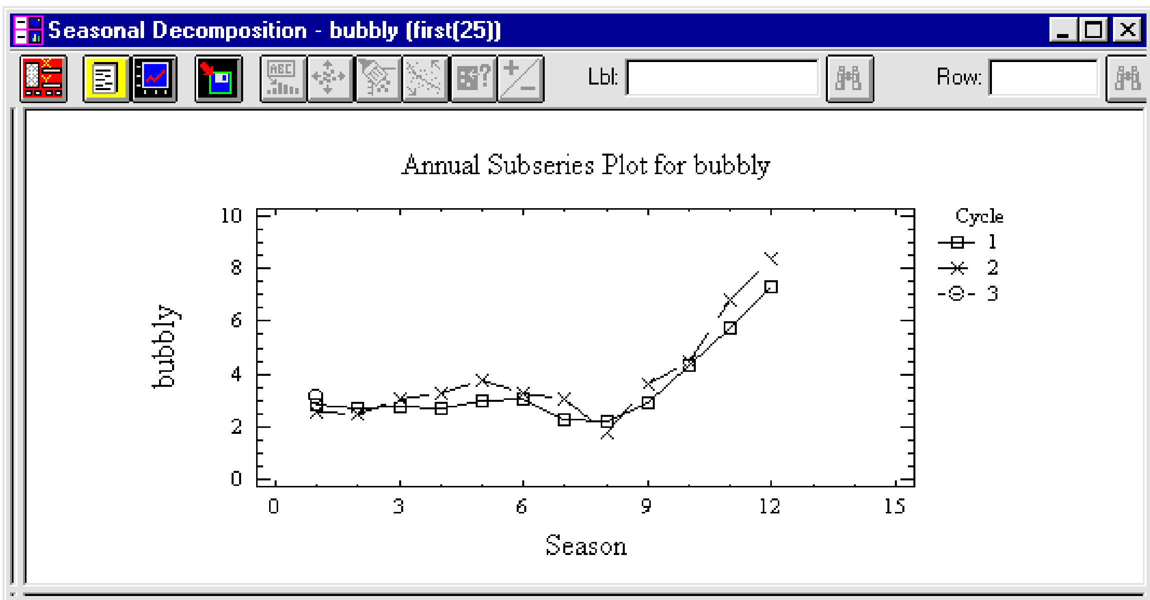


Figure 3-10. Annual Subseries Plot

Use the *Annual Subseries Plot Options* dialog box to indicate if the plot will contain cumulative values.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are five selections: Data, Trend-Cycle, Seasonal Indices, Irregular, and Seasonally Adjusted Data.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Cleveland, W. S. and Terpenning, I. J. 1982. "Graphical Methods for Seasonal Adjustment," *Journal of the American Statistical Association*, **77**:52-62.

Makridakis, W. and Wheelwright, S. C. 1989. *Forecasting Methods for Management*, fifth edition. New York: Wiley.

Makridakis, W., Wheelwright, S. C., and McGee, U. E. 1983. *Forecasting: Methods and Applications*, second edition. New York: Wiley.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: PWS-Kent Publishing Company.

Chapter 4

Forecasting Values

Background Information

Great quantities of data in business, economics, engineering, and the natural sciences occur in the form of time series where observations are dependent. Being able to forecast optimally is practical and very important for business planning, where forecasting is based on fitting a model to past observations in a given time series.

One objective in time-series analysis is to develop models that can be used to obtain optimal forecasts of future values. The model you choose depends on how the forecast will be used, the degree of accuracy required from the forecast, the amount of time and capital available, the amount and type of available data, and how far ahead you must forecast.

The fitted model determines if the forecast projections should follow a straight line, an exponential curve, and so on. The fitted model also allows you to see exactly how the forecasts use past data to determine the variation of the forecast errors and to calculate limits within which a future value of the series will lie with a given probability.

STATGRAPHICS *Plus* includes 13 forecasting models you can modify by changing the parameters. You can choose up to five models. With each model you can transform the data or apply a constant rate of inflation. Depending on the model you choose, you can also optimize parameters or include a constant term. If the data are seasonal, you can apply differencing and choose a seasonal adjustment.

Random Walk
Randomly forecasts the next observation based on the current observation and the mean and standard deviation of the difference of the values.

Mean
Forecasts by adding together the values of the items and dividing them by the number of items; more simply, by using the average.

Linear Trend
Fits a straight line through the data and into the forecasting periods.

Quadratic Trend
Fits a quadratic curve through the data and into the forecasting periods.

Exponential Trend
Fits an exponential curve through the data and into the forecasting periods.

S-Curve
Fits an S-shaped curve through the data and into the forecasting periods.

Moving Average
Uses the moving average to smooth the data and to predict future values.

Simple Exponential Smoothing
Smoothes the data and predicts future values by exponentially weighting the values in the time series.

Brown's Linear Exponential Smoothing
Smoothes the data and predicts future values by applying a double-smoothing formula to the data using one parameter, alpha.

Holt's Linear Exponential Smoothing
Applies a double-smoothing formula to the data using two smoothing parameters: alpha and beta.

Quadratic Exponential Smoothing
Assumes that the data follow a trend characterized as a second-order polynomial.

Winter's Exponential Smoothing
Applies three separate smoothing parameters to the data: alpha, beta, and gamma. It estimates the level of the series (the stationarity), the linear trend, and the seasonality.

ARIMA Model
Estimates and forecasts using the methods prescribed by Box and Jenkins (1976).

Using the Forecasting Analysis

To access the analysis, choose SPECIAL... TIME-SERIES ANALYSIS... FORECASTING... from the Menu bar to display the Forecasting Analysis dialog box shown in Figure 4-1.

Note: When you are entering data for time-series variables, data that contain missing or embedded values cannot be entered at the beginning or end of the series.

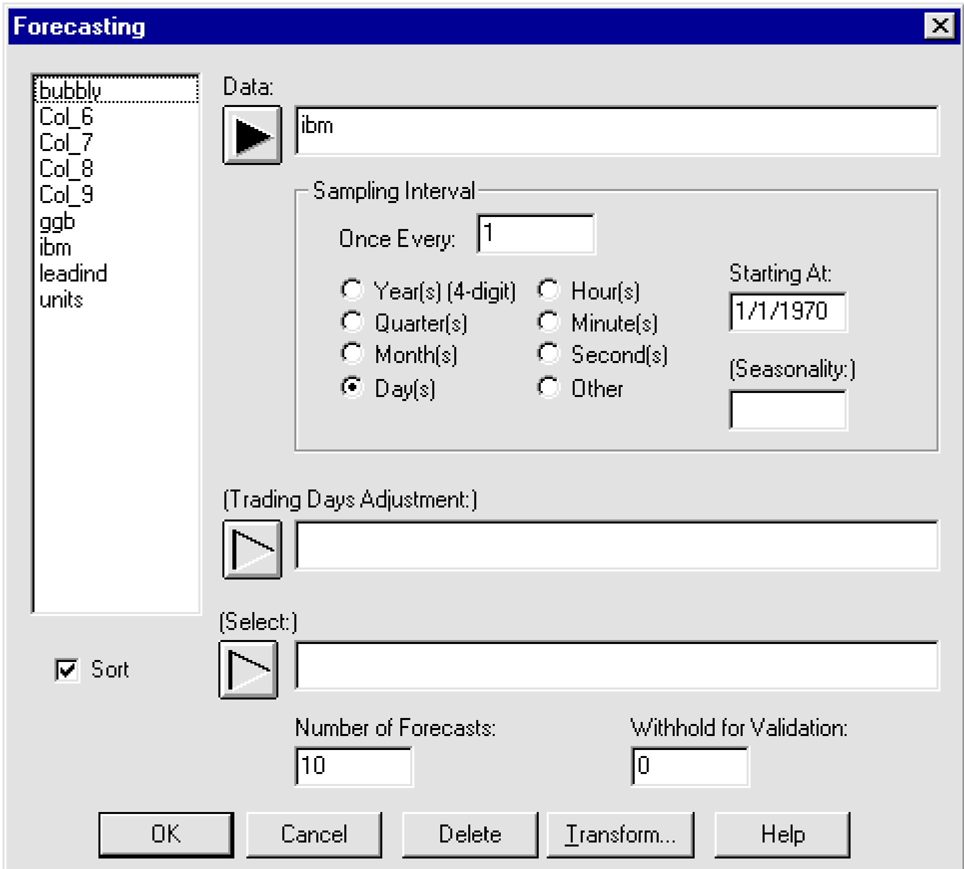


Figure 4-1. Forecasting Analysis Dialog Box

Tabular Options

Analysis Summary
The Analysis Summary option creates a summary of the analysis that displays the name of the variable, the number of observations, the start index, and the sampling interval (see Figure 4-2). The summary also displays a summary of the forecast statistics: the name of the chosen forecast model, the number of forecasts generated, and the number of periods withheld for validation.

A table shows the results of the validation, which are used to validate a model's accuracy. The statistics include:

MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in a dataset, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different forecasting model.

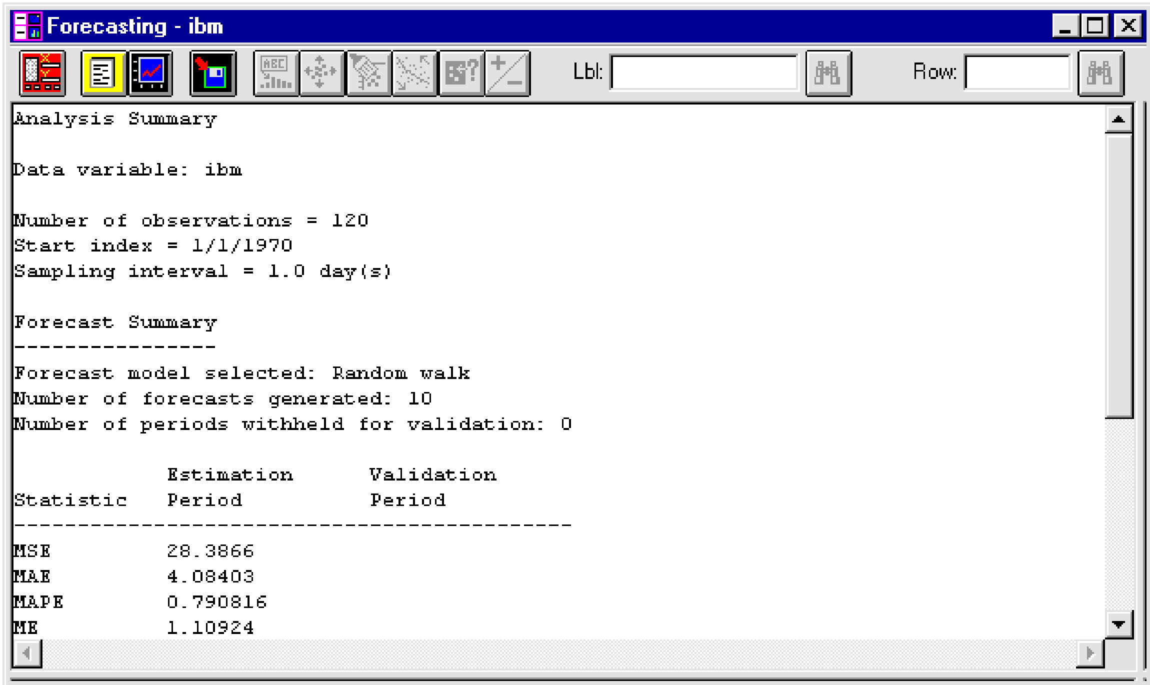


Figure 4-2. Analysis Summary

MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different forecasting model.

MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given dataset without regard to sign (that is, the absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the forecast.

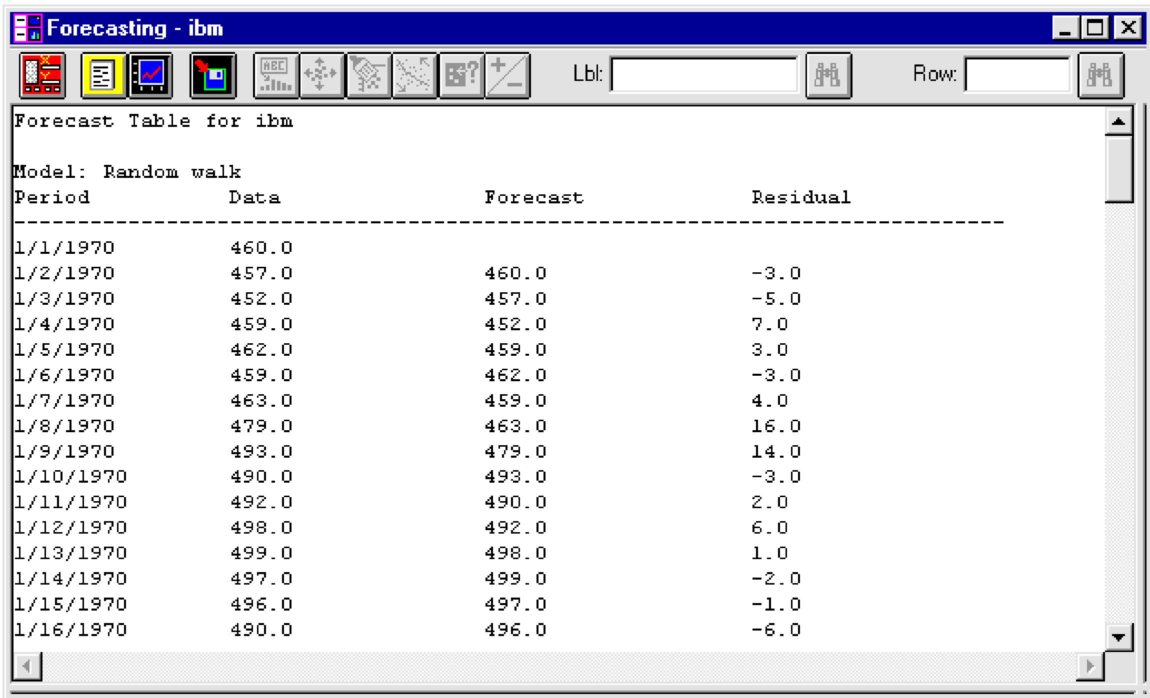
MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. The one-ahead forecast errors are divided by the actual values. Like MAPE, it is independent of scale.

If you choose certain model types, a summary of the statistics appears at the bottom of the Analysis Summary. The summary includes the parameter(s) for the model, the estimate, the standard error, and the *t*- and *p*-values. Each of the statistics is based on the one-ahead forecast errors, which are the differences between the value at time t and the forecast of that value made at time *t - 1*.

Use the *Model Specification Options* dialog box to choose the forecasting model that will be used to fit the data; to indicate the type of forecasting model that will be used to set the parameters and terms for the model, to indicate the method that will be used to perform the seasonal adjustment on the data, to indicate the order of differencing that will be applied to the data, to indicate the type of adjustment that will be applied to the data, and to indicate how the data will be adjusted for a constant rate of inflation. You can also use the Estimation... and Regression... commands to access other dialog boxes.

Forecast Table

The Forecast Table option creates a table of the forecasted values. During periods when actual data are available, the table displays the predicted values from the fitted model and the residuals (see Figure 4-3).



Forecast Table for ibm

Model: Random walk

Period	Data	Forecast	Residual
1/1/1970	460.0		
1/2/1970	457.0	460.0	-3.0
1/3/1970	452.0	457.0	-5.0
1/4/1970	459.0	452.0	7.0
1/5/1970	462.0	459.0	3.0
1/6/1970	459.0	462.0	-3.0
1/7/1970	463.0	459.0	4.0
1/8/1970	479.0	463.0	16.0
1/9/1970	493.0	479.0	14.0
1/10/1970	490.0	493.0	-3.0
1/11/1970	492.0	490.0	2.0
1/12/1970	498.0	492.0	6.0
1/13/1970	499.0	498.0	1.0
1/14/1970	497.0	499.0	-2.0
1/15/1970	496.0	497.0	-1.0
1/16/1970	490.0	496.0	-6.0

Figure 4-3. Forecast Table

During time periods beyond the end of the series, the table shows the prediction limits for the forecasts. Assuming the fitted model is appropriate for the data, the limits show, with 95 percent confidence, the location of the true data value at a chosen future time. Use the Vertical Scroll Bar to the right of the table to view the entire table. The future forecasts and their corresponding upper and lower confidence limits are displayed at the bottom of the table.

Model Comparisons

The Model Comparisons option creates a table of the results of comparing five different forecasting models (see Figure 4-4). In general, the better model has smaller MSE and MAE values and values of ME and MPE that are closer to 0. The table also summarizes the results of five tests run on the residuals to determine if each model is adequate for the data.

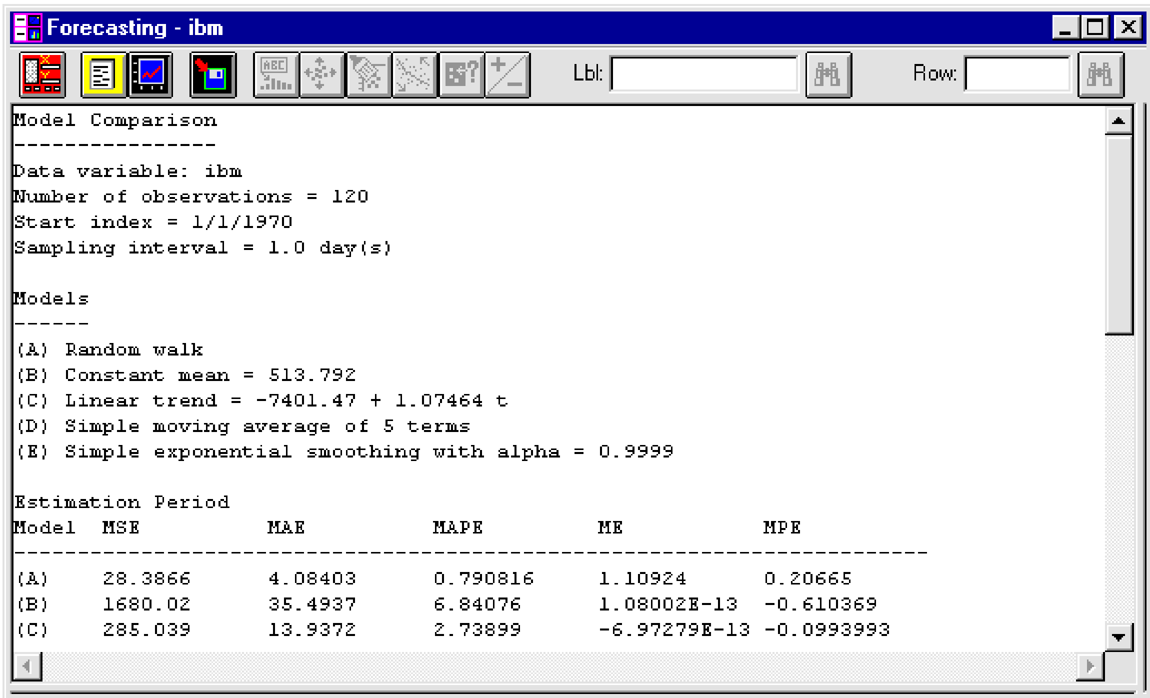


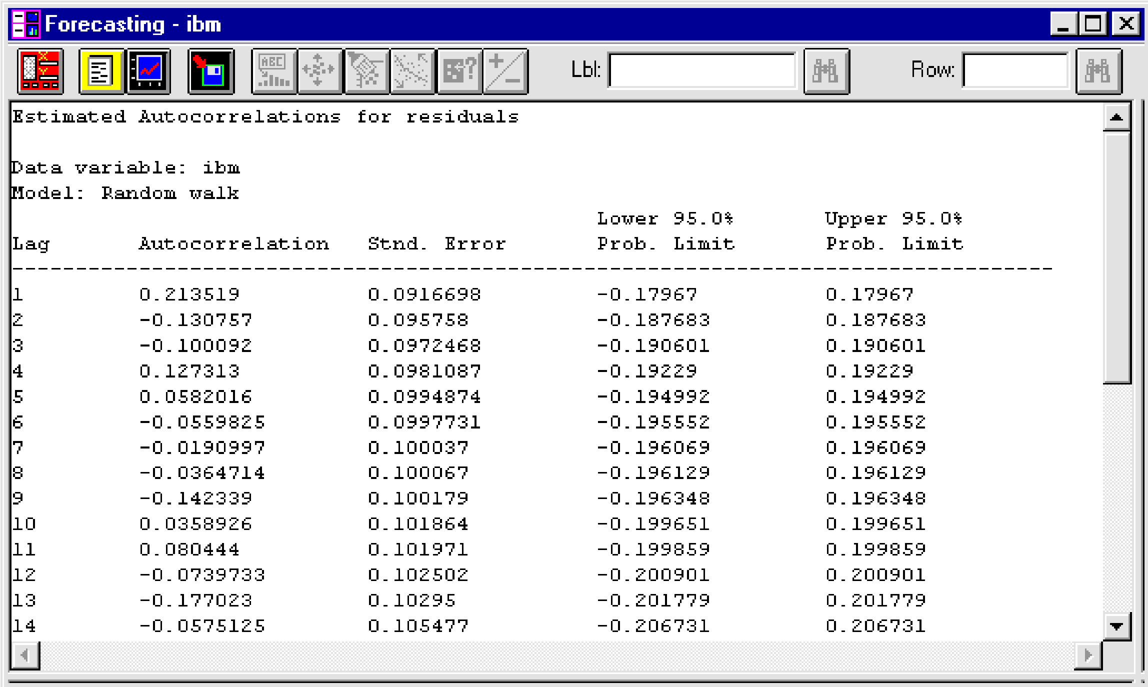
Figure 4-4. Model Comparisons

Residual Autocorrelations

The Residual Autocorrelations option creates a table that shows the estimated autocorrelations between the residuals at various lags (see Figure 4-5). The lag *k* autocorrelaton coefficient measures the correlation between the residuals at time *t* and time *t* - *k*. The probability limits are also shown. If the limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlaton at that lag.

The table also shows the standard errors, and the upper and lower probability limits for each lag. You can use the residual autocorrelations to determine if the chosen model is appropriate for the data.

Use the *Autocorrelation Function Options* dialog box to enter a value for the maximum number of lags that will be estimated for the residual autocorrelations, and to enter a value for the size of the probability limits that will be used to calculate the residual autocorrelations.



Estimated Autocorrelations for residuals

Data variable: ibm
Model: Random walk

Lag	Autocorrelation	Stnd. Error	Lower 95.0% Prob. Limit	Upper 95.0% Prob. Limit
1	0.213519	0.0916698	-0.17967	0.17967
2	-0.130757	0.095758	-0.187683	0.187683
3	-0.100092	0.0972468	-0.190601	0.190601
4	0.127313	0.0981087	-0.19229	0.19229
5	0.0582016	0.0994874	-0.194992	0.194992
6	-0.0559825	0.0997731	-0.195552	0.195552
7	-0.0190997	0.100037	-0.196069	0.196069
8	-0.0364714	0.100067	-0.196129	0.196129
9	-0.142339	0.100179	-0.196348	0.196348
10	0.0358926	0.101864	-0.199651	0.199651
11	0.080444	0.101971	-0.199859	0.199859
12	-0.0739733	0.102502	-0.200901	0.200901
13	-0.177023	0.10295	-0.201779	0.201779
14	-0.0575125	0.105477	-0.206731	0.206731

Figure 4-5. Residual Autocorrelations

Residual Partial Autocorrelations

The Residual Partial Autocorrelations option creates a table of the estimated partial autocorrelations between the residuals at various lags (see Figure 4-6). The lag *k* partial autocorrelation coefficient measures the correlation between the residuals at time t and time *t* + *k* after accounting for the correlations at all the lower lags.

Use the results to determine the order of autoregressive model that is needed to fit the data. The probability limits are also shown. If the limits do not contain the estimated coefficient at a particular lag, it is an indication that there is a statistically significant correlation at that lag at the 95 percent confidence level.

Use the *Partial Autocorrelation Function Options* dialog box to enter the maximum number of lags that will be estimated for the residual autocorrelations and to enter the size of the probability limits that will be used to calculate the residual autocorrelations.

Residual Periodogram Table

The Residual Periodogram Table option creates a table that shows the periodogram ordinates for the residuals that are used to identify cycles of fixed frequency of the data (see Figure 4-7). A periodogram is constructed by fitting a series of sine functions at each of the chosen frequencies. The ordinates are equal to the squared amplitudes of the sine functions.

Forecasting - ibm

Lbl: Row:

Estimated Partial Autocorrelations for residuals

Data variable: ibm
Model: Random walk

Lag	Partial Autocorrelation	Stnd. Error	Lower 95.0% Prob. Limit	Upper 95.0% Prob. Limit
1	0.213519	0.0916698	-0.17967	0.17967
2	-0.184772	0.0916698	-0.17967	0.17967
3	-0.0298993	0.0916698	-0.17967	0.17967
4	0.146898	0.0916698	-0.17967	0.17967
5	-0.0322048	0.0916698	-0.17967	0.17967
6	-0.0348484	0.0916698	-0.17967	0.17967
7	0.0362123	0.0916698	-0.17967	0.17967
8	-0.0778322	0.0916698	-0.17967	0.17967
9	-0.143798	0.0916698	-0.17967	0.17967
10	0.126759	0.0916698	-0.17967	0.17967
11	-0.00209736	0.0916698	-0.17967	0.17967
12	-0.119444	0.0916698	-0.17967	0.17967
13	-0.0641319	0.0916698	-0.17967	0.17967
14	-0.0238641	0.0916698	-0.17967	0.17967

Figure 4-6. Residual Partial Autocorrelations

Forecasting - ibm

Lbl: Row:

Periodogram for residuals

Data variable: ibm
Model: Random walk

Frequency	Period	Ordinate	Cumulative Sum	Integrated Periodogram
0.0		2.56678E-29	2.56678E-29	7.94281E-33
0.00840336	119.0	51.6828	51.6828	0.0159931
0.0168067	59.5	65.1404	116.823	0.0361505
0.0252101	39.6667	4.73365	121.557	0.0376153
0.0336134	29.75	128.274	249.83	0.0773091
0.0420168	23.8	186.222	436.053	0.134935
0.0504202	19.8333	43.6034	479.656	0.148428
0.0588235	17.0	69.4891	549.145	0.169931
0.0672269	14.875	91.9591	641.104	0.198387
0.0756303	13.2222	3.30172	644.406	0.199409
0.0840336	11.9	3.83304	648.239	0.200595
0.092437	10.8182	18.9301	667.169	0.206453
0.10084	9.91667	22.6504	689.82	0.213462
0.109244	9.15385	172.115	861.934	0.266722

Figure 4-7. Residual Periodogram Table

A periodogram can also be thought of as an analysis of variance by frequency because the sum of the ordinates equals the total corrected sum of squares in an ANOVA table.

Use the *Periodogram Table Options* dialog box to indicate that the mean should be subtracted from each observation and to indicate that the program should taper the beginning and end of the time-series data using a split cosine-bell window. See Bloomfield (1976) for a discussion of the tapering method.

Residual Tests for Randomness

The Residual Tests for Randomness option runs three tests to determine if the residuals form a random sequence of numbers (see Figure 4-8). A sequence of random numbers is often called white noise because it contains equal contributions at many frequencies.

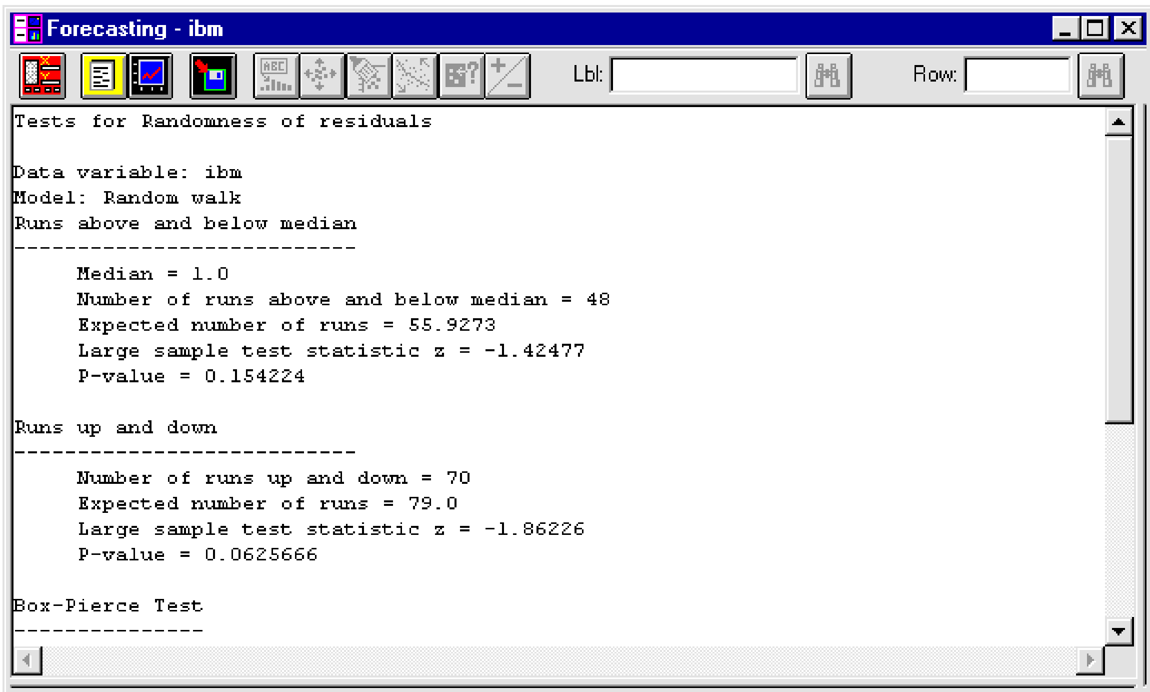


Figure 4-8. Residual Tests for Randomness

The tests are Runs Above and Below the Median, Runs Up and Down, and the Box-Pierce test. The Runs Above and Below the Median test counts the number of runs that are completely above or completely below the median; it ignores the values that are equal to the median. The test is sensitive to trends in the data. If the test statistic is large (corresponds to a *p*-value less than .05), it can be concluded that the values occur in nonrandom order.

The Runs Up and Down test counts the number of times a sequence rises or falls; it is sensitive to long-term cycles. If the test statistic is small (corresponds to a *p*-value greater than .05), it can be concluded that the values occur in random order.

The Box-Pierce test determines if the autocorrelation is equal to 0. If the test statistic is large (corresponds to a *p*-value less than .05), the autocorrelation is not equal to 0, which indicates that the model is not adequate.

Use the *Box-Pierce Test Options* dialog box to enter the maximum number of lags for which statistics will be estimated for the test.

Graphical Options

Time Sequence Plot
The Time Sequence Plot option creates a connected line plot for the forecasted values against time (see Figure 4-9). The plot also includes the prediction limits for the forecasts, which shows with 95 percent confidence, where the true value of the variable is likely to be at any point in the future.

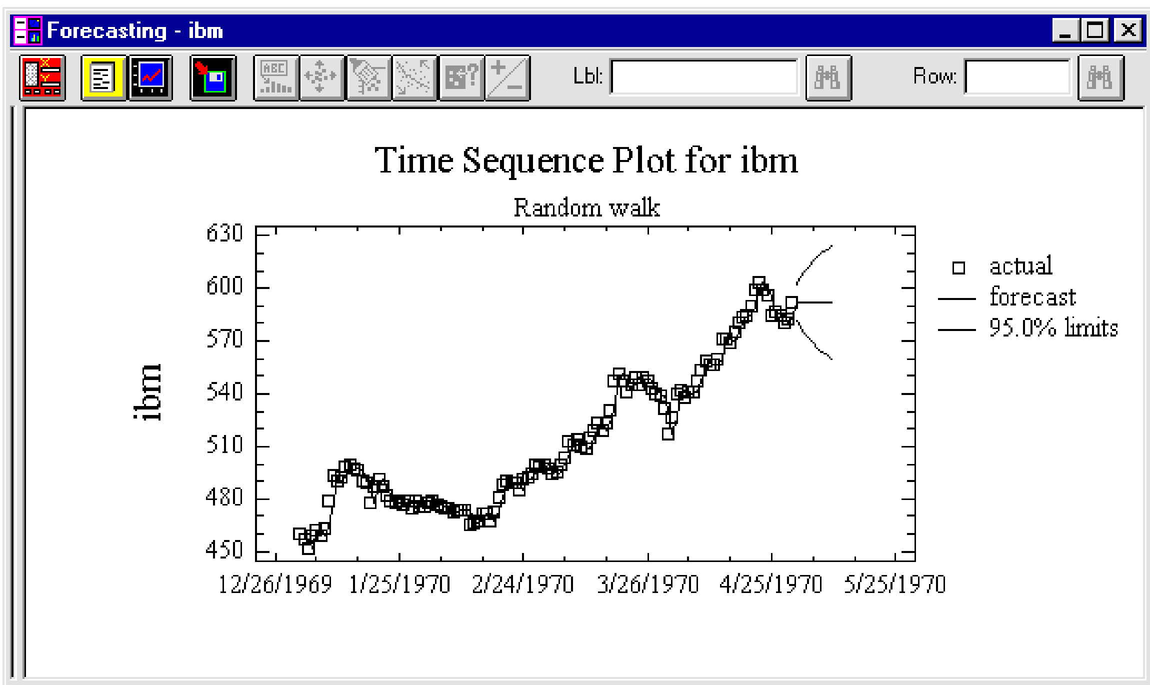


Figure 4-9. Time Sequence Plot

Use the Forecast Limits Options dialog box to enter the percentage that will be used for the calculations.

Forecast Plot
The Forecast Plot option creates a connected line plot of the forecasted values (see Figure 4-10). The plot also includes the prediction limits for the forecasts, which shows with 95 percent confidence, where the true value of the variable is likely to be at any point in the future.

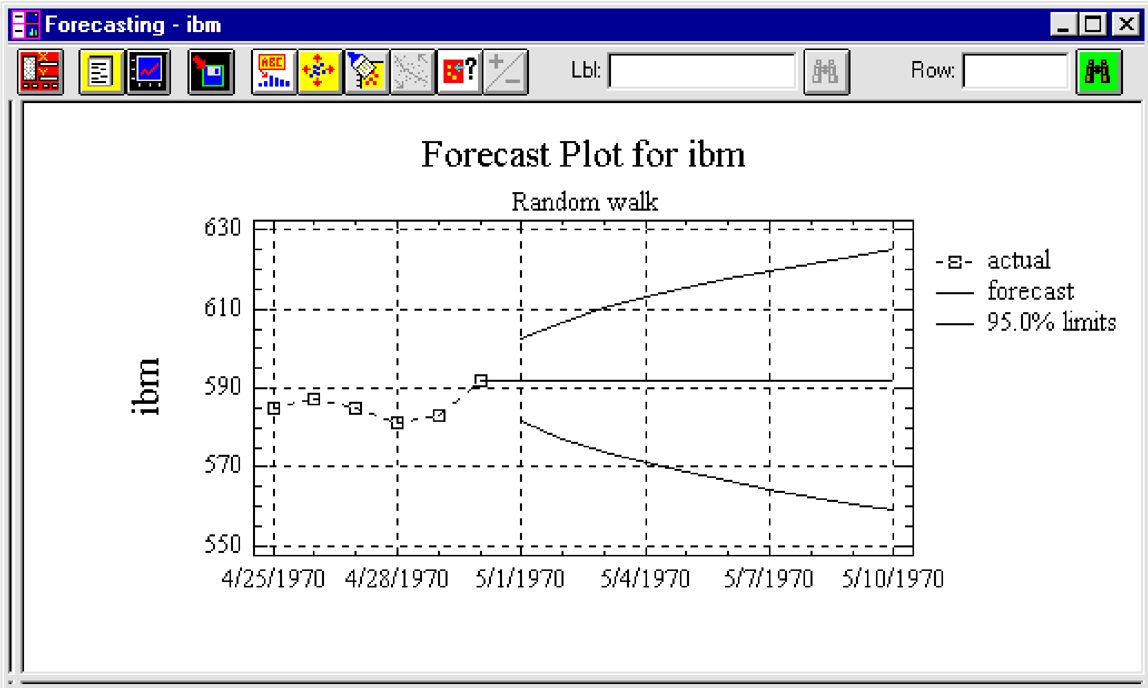


Figure 4-10. Forecast Plot

Use the Forecast Limits Options dialog box to enter the percentage that will be used for the calculations.

Residual Plots

The Residual Plots option creates plots of the residuals from the fitted model. If all the dynamic structure in the variable is captured, the residuals should be random (white noise).

Use the Forecasting Residual Plots Options dialog box to choose the type of plot that will be created.

Time Sequence Plot **-** a connected line plot of the residuals against time (see Figure 4-11).

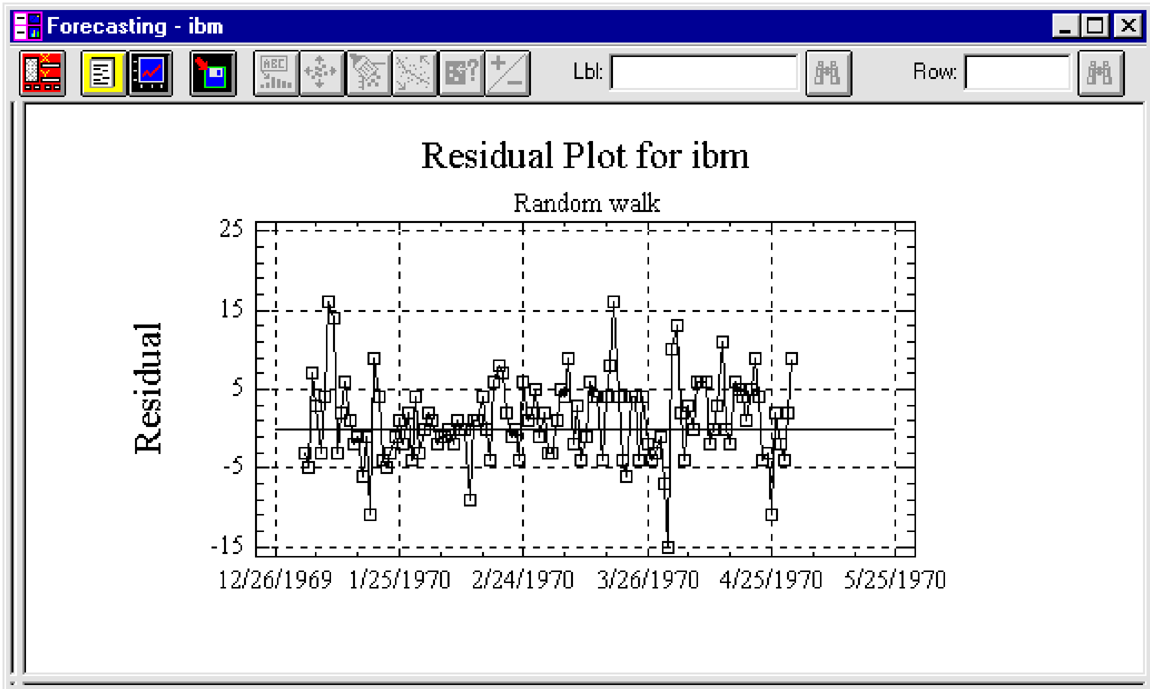


Figure 4-11. Time Sequence Plot of the Residuals

Probability Plot (Horizontal) **-** a plot of the residuals on the Y-axis from smallest to largest (see Figure 4-12). The X-axis is a normal probability scale (it "straightens out" the plot of a cumulative normal distribution). If the residuals do not fall approximately along a straight line, they are not from a normal distribution. This suggests that you may not have used the best model for the data.

Probability Plot (Vertical) - a plot of the residuals on the X-axis from smallest to largest. The Y-axis is a normal probability scale (it "straightens out" the plot of a cumulative normal distribution). If the residuals do not fall approximately along a straight line, they are not from a normal distribution. This suggests that you may not have used the best model for the data.

Residual Autocorrelation Function

The Residual Autocorrelation Function option creates a plot of the estimated autocorrelations between the residuals at various lags (see Figure 4-13). The lag k autocorrelation coefficient measures the correlations between the residuals at time t and time t - k. The probability limits are also shown. If the limits do not contain the estimated coefficient at a particular lag, it is an indication that there is a statistically significant correlation at that lag at the 95 percent confidence level.

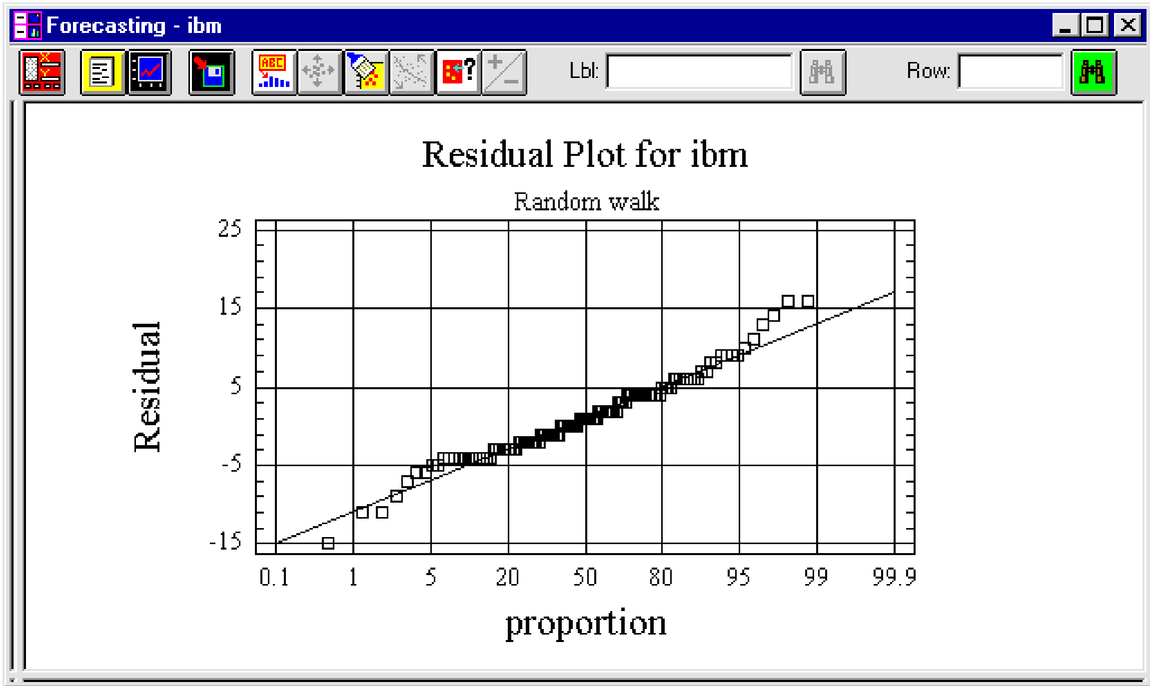


Figure 4-12. Probability Plot (Horizontal)

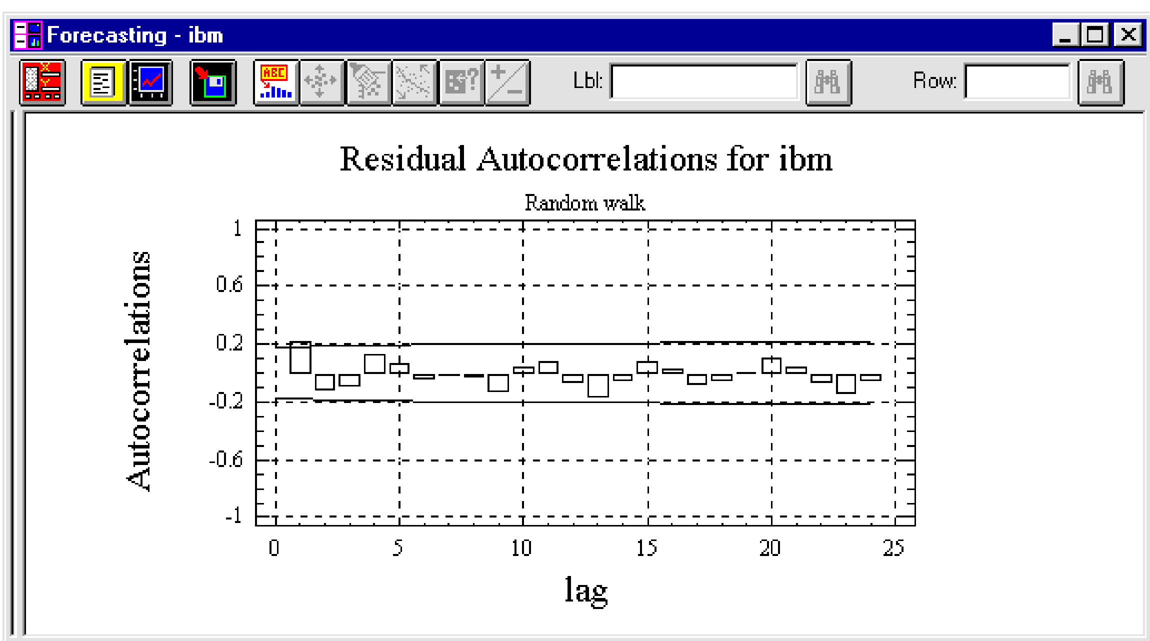


Figure 4-13. Residual Autocorrelation Function Plot

The plot contains vertical bars that represent the coefficient for each lag and a pair of dotted lines at a distance from the baseline that are a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits.

Use the Autocorrelation Function Options dialog box to enter the maximum number of lags that will be estimated for the residual autocorrelations and to enter the size of the probability limits that will be used to calculate the residual autocorrelations.

Residual Partial Autocorrelation Function

The Residual Partial Autocorrelation Function option creates a plot of the estimated partial autocorrelations between the residuals at various lags (see Figure 4-14).

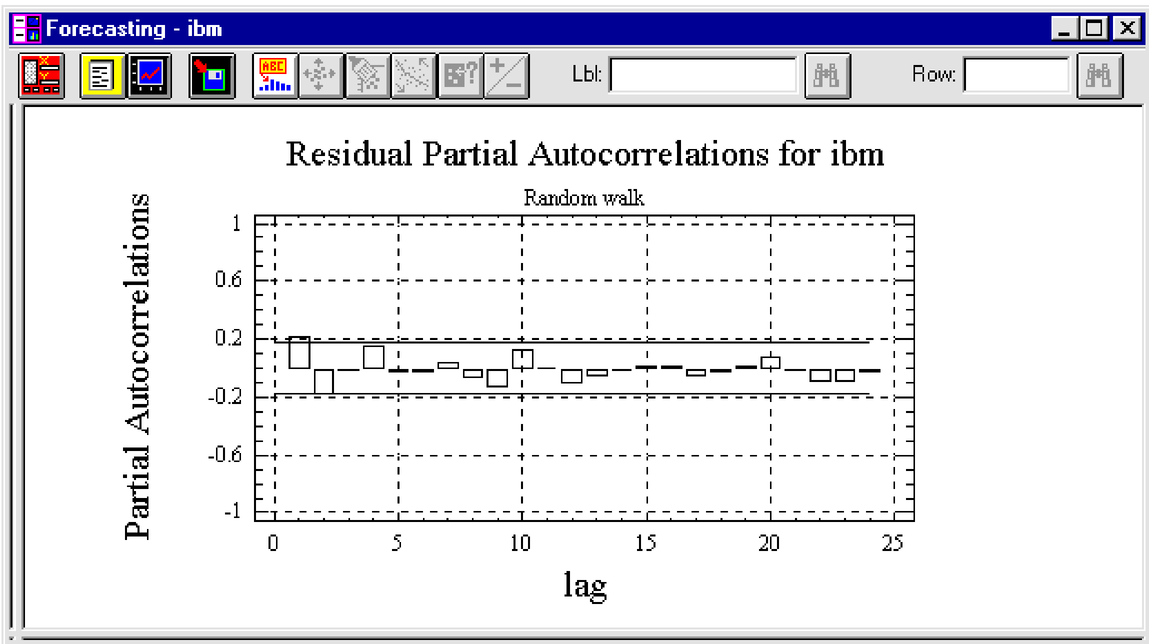


Figure 4-14. Residual Partial Autocorrelation Function Plot

The lag k partial autocorrelation coefficient measures the correlation between the residuals at time t and time t + k after accounting for the correlations at all the lower lags.

Each coefficient is plotted using a bar whose height is proportional to the value of the coefficient. By default, it plots probability limits as dashed lines at plus or minus 2 divided by the square root of the number of observations in the time series. The bounds are useful for indicating partial autocorrelations that are significantly different from 0. Bars that extend beyond either line indicate significant correlations.

Use the Partial Autocorrelation Function Options dialog box to enter the maximum number of lags that will be estimated for the residual autocorrelations and to enter the size of the probability limits that will be used to calculate the residual autocorrelations.

Residual Periodogram

The Residual Periodogram option creates plot that shows the periodogram ordinates for the residuals (see Figure 4-15). The periodogram is constructed by fitting a series of sine functions at each specified frequency. The ordinates are equal to the squared amplitudes of the sine functions.

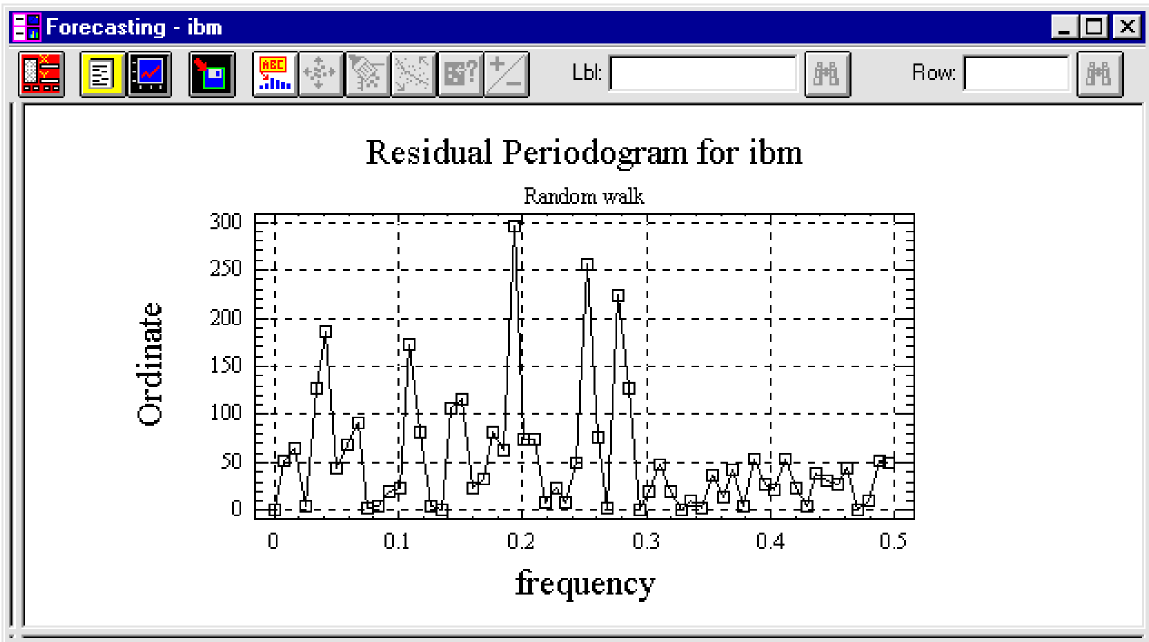


Figure 4-15. Residual Periodogram

A residual periodogram is similar to an analysis of variance by frequency because the sum of the ordinates equals the total sum of squares in an ANOVA table.

Use the Periodogram Options dialog box to indicate that the mean should be subtracted from each observation, to indicate if the observations should be plotted as points, to indicate if lines should connect the observations, and to indicate that the time-series data should be tapered at its beginning and end using a split cosine-bell window. See Bloomfield (1976) for a discussion of the tapering method.

Residual Crosscorrelation Function
The Residual Crosscorrelation Function option creates a plot that shows the crosscorrelations between the residuals and the chosen variable (see Figure 4-16).

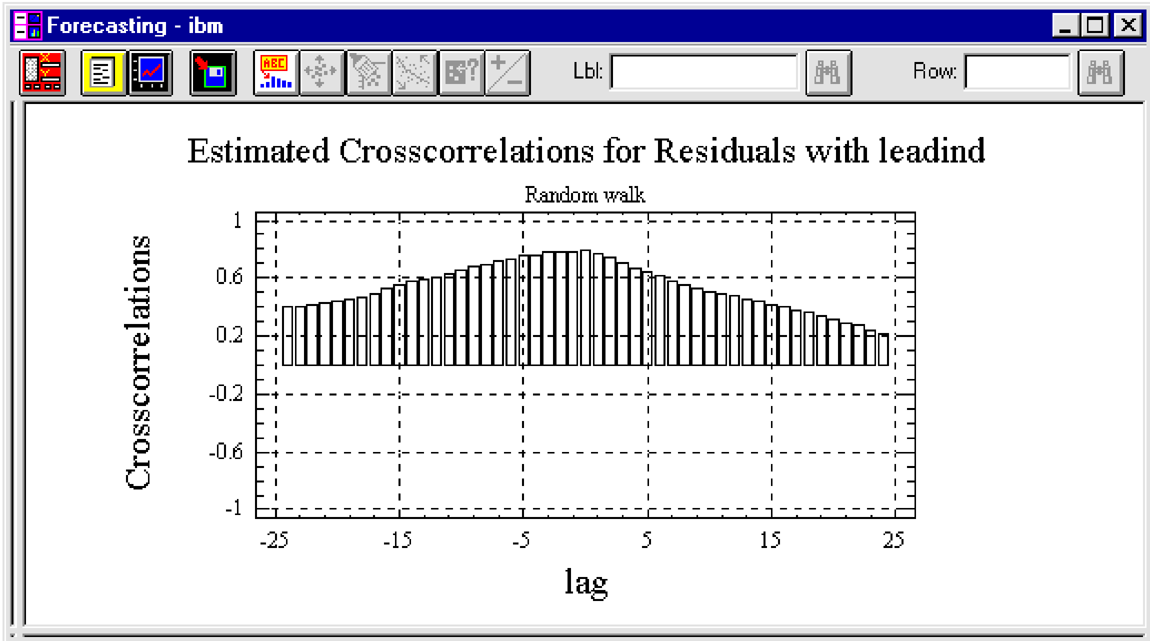


Figure 4-16. Residual Crosscorrelation Function

The crosscorrelation at lag k measures the strength of the linear relationship between the variable's value at time t and the value of the second variable k periods earlier. The plot helps determine if the second variable is helpful in forecasting the first variable.

Use the Crosscorrelation Function Analysis dialog box to choose the variable that contains the second set of time-series data, and to enter the maximum number of lags that will be calculated for the crosscorrelation function.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are ten selections: Data, Adjusted Data, Forecasts, Upper Forecast Limits, Lower Forecast Limits, Residuals, Autocorrelations, Partial Autocorrelations, Residual Periodogram Ordinates, and Fourier Frequencies.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Bloomfield, P. 1976. Fourier Analysis of Time Series: An Introduction. New York: Wiley.

Box, G. E. P. and Cox, D. R. 1964. "An Analysis of Transformations," Journal of the Royal Statistical Society, 26:211-243.

Box, G. E. P. and Jenkins, G. M. 1976. Time Series Analysis, Forecasting, and Control, second edition. San Francisco: Holden-Day.

Brown, R. G. 1977. Materials Management. New York: Wiley.

Brown, R. G. 1963. Smoothing, Forecasting, and Prediction of Discrete Time Series. Englewood Cliffs: Prentice-Hall.

Brown, R. G. and Meyer, R. F. 1983. "The Fundamental Theorem of Exponential Smoothing," Operations Research, 9:1961.

Kendall, M. G. and Stuart, A. 1961. The Advanced Theory of Statistics, Volume 2. New York: Hafner Publishing Company.

Makridakis, S. and Wheelwright, S. C. 1978. Interactive Forecasting: Univariate & Multivariate Methods. Oakland, California: Holden-Day.

Montgomery, D., Johnson, L., and Gardiner, J. 1990. Forecasting and Time Series Analysis, second edition. New York: McGraw-Hill.

Neter, J., Wasserman, W., and Kutner, M. 1985. Applied Linear Statistical Models, second edition. Homewood, Illinois: Richard E. Irwin, Inc.

Tukey, J. W. 1977. Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics and Computing of Exploratory Data Analysis*. Boston: PWS-Kent Publishing Company.

Winters, P. R. 1960. "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science*, 6(3):324-342.

Chapter 5

Using Automatic Forecasting

Background Information

Chapter 4 described in detail procedures for identification, estimation, and diagnostic checking of a wide variety of forecasting models. The *Time Series Forecasting* analysis described there is designed to give the analyst maximum flexibility in examining different types of models. The available models extend from classic exponential smoothing models to the class of ARIMA models popularized by Box and Jenkins (1970).

While maximum flexibility is often important, other situations exist where the forecaster might prefer that the computer automatically select a reasonable model without his or her active intervention. When many time series need to be forecast, it may not be practical to manually select a model type for each series. The *Automatic Forecasting* selection on the *Time Series Analysis* menu is designed for such situations. It automatically selects a reasonable model for any input time series and uses that model to generate short-term forecasts.
Most of the options and output for *Automatic Forecasting* are identical to those described in the previous chapter.

Automatic Forecasting in STATGRAPHICS *Plus*

To access the analysis, choose SPECIAL... TIME-SERIES ANALYSIS... AUTOMATIC FORECASTING... from the Menu bar to display the Automatic Forecasting Analysis dialog box shown in Figure 5-1.

Note: When you are entering data for time-series variables, data that contain missing or embedded values cannot be entered at the beginning or end of the series.

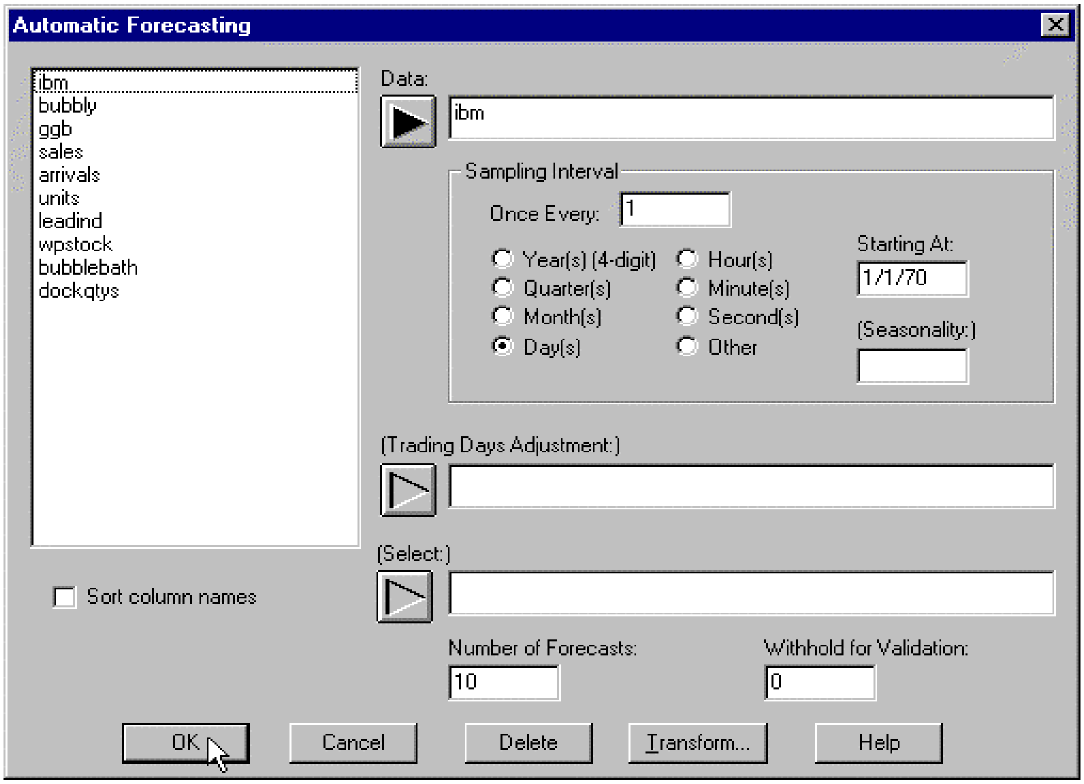


Figure 5-1. Automatic Forecasting Analysis Dialog Box

Tabular Options

Analysis Summary
The Analysis Summary option creates a summary of the analysis that displays the name of the variable, the number of observations, the start index, and the sampling interval (see Figure 5-2). The summary also displays a summary of the forecast statistics: the name of the chosen forecast model, the number of forecasts generated, and the number of periods withheld for validation.

Forecast Model Selected --- this line shows the type of model selected by the analysis. In this case, it is an autoregressive moving average (ARMA) model with second order seasonal and nonseasonal autoregressive terms and first order seasonal and nonseasonal moving average terms.
Math Adjustment --- the indicated model has been fit to the data after applying a Box-Cox transformation power equal to .282812.

A table shows the results of the validation, which are used to validate a model's accuracy. The statistics include:

MSE (Mean Square Error) — a measure of accuracy computed by squaring the individual error for each item in a dataset, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different forecasting model.

```
Analysis Summary

Data variable: bubbly

Number of observations = 84
Start index = 1/93
Sampling interval = 1.0 month(s)
Length of seasonality = 12

Forecast Summary
----------------
Math adjustment: Box-Cox with power = 0.282812 and addend = 0.0

Forecast model selected: ARMA(2,1) SARMA(2,1)
Number of forecasts generated: 12
Number of periods withheld for validation: 0

            Estimation      Validation
Statistic   Period          Period
--------------------------------------------
MSE         0.395949
MAE         0.426168
MAPE        10.6613
ME          0.0136617
MPE         -2.03669

                            ARIMA Model Summary
Parameter           Estimate        Stnd. Error     t               P-value
----------------------------------------------------------------------------
AR(1)               1.1016          0.148483        7.41903         0.000000
AR(2)               -0.111534       0.138057        -0.80788        0.421649
MA(1)               0.87202         0.0879382       9.91628         0.000000
SAR(1)              0.378681        0.238734        1.58621         0.116791
SAR(2)              0.654547        0.241697        2.70813         0.008333
SMA(1)              -0.242121       0.280623        -0.862798       0.390928
Mean                4.4829          15.3977         0.291141        0.771726
Constant            -0.00147985
----------------------------------------------------------------------------
Backforecasting: yes
Estimated white noise variance = 0.393778 with 77 degrees of freedom
Estimated white noise standard deviation = 0.627517
Number of iterations: 11
```

Figure 5-2. Analysis Summary

MAE (Mean Absolute Error) — the average of the absolute values of the residuals; appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different forecasting model.

MAPE (Mean Absolute Percentage Error) — the mean or average of the sum of all the percentage errors for a given dataset without regard to sign (that is, the absolute values are summed and the average is computed). Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

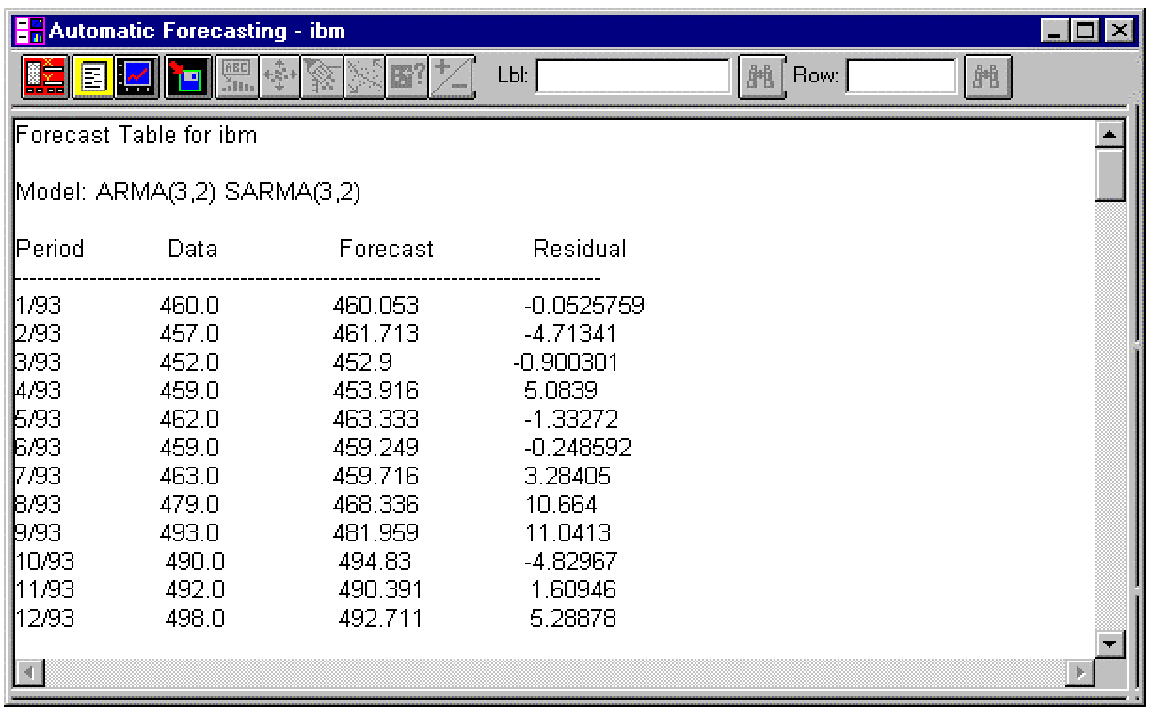
ME (Mean Error) — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the forecast.

MPE (Mean Percentage Error) — the average of the absolute values of the residuals divided by the corresponding estimates. The one-ahead forecast errors are divided by the actual values. Like MAPE, it is independent of scale.

If you choose certain model types, a summary of the statistics appears at the bottom of the Analysis Summary. The summary includes the parameter(s) for the model, the estimate, the standard error, and the *t*- and *p*-values. Each of the statistics is based on the one-ahead forecast errors, which are the differences between the value at time t and the forecast of that value made at time *t - 1*.

Forecast Table

The Forecast Table option creates a table of the forecasted values. During periods when actual data are available, the table displays the predicted values from the fitted model and the residuals (see Figure 5-3).



Forecast Table for ibm

Model: ARMA(3,2) SARMA(3,2)

Period	Data	Forecast	Residual
1/93	460.0	460.053	-0.0525759
2/93	457.0	461.713	-4.71341
3/93	452.0	452.9	-0.900301
4/93	459.0	453.916	5.0839
5/93	462.0	463.333	-1.33272
6/93	459.0	459.249	-0.248592
7/93	463.0	459.716	3.28405
8/93	479.0	468.336	10.664
9/93	493.0	481.959	11.0413
10/93	490.0	494.83	-4.82967
11/93	492.0	490.391	1.60946
12/93	498.0	492.711	5.28878

Figure 5-3. Forecast Table

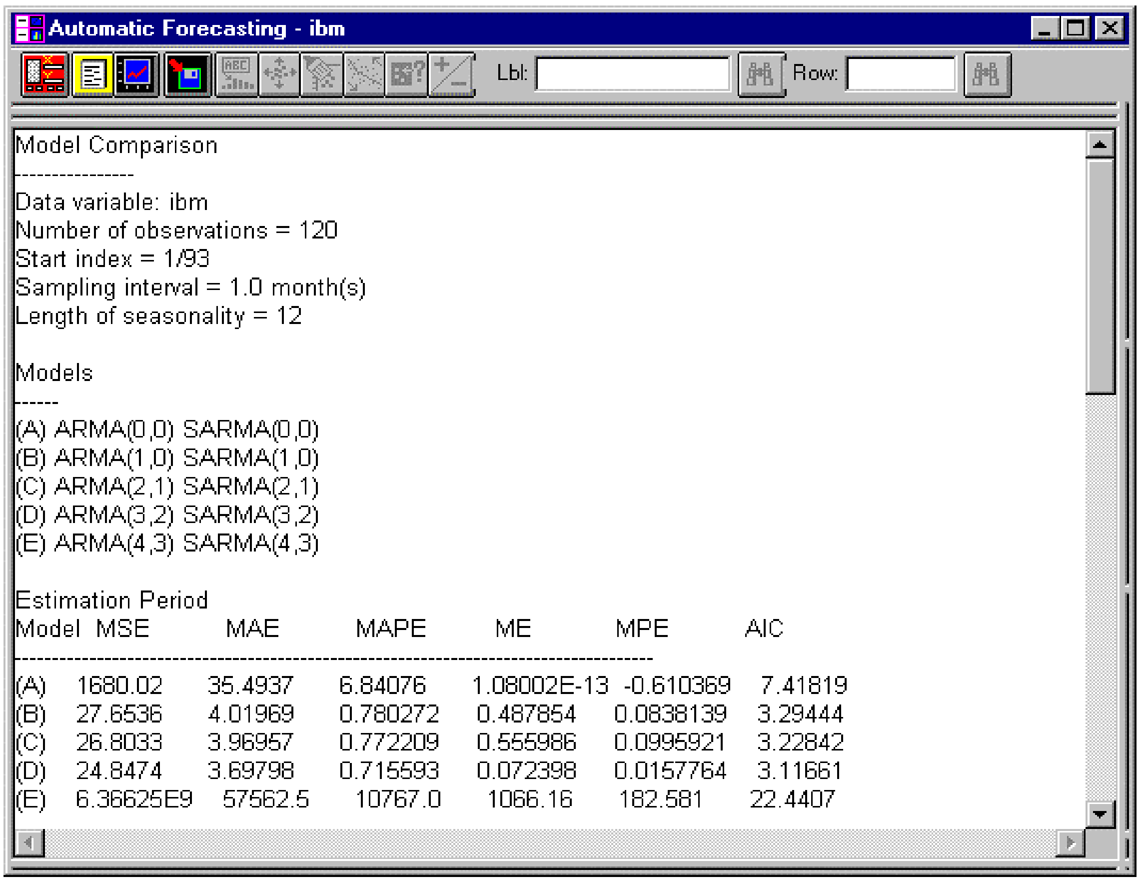
During time periods beyond the end of the series, the table shows the prediction limits for the forecasts. Assuming the fitted model is appropriate for the data, the limits show, with 95 percent confidence, the location of the true data value at a chosen future time. Use the Vertical Scroll Bar to the right of the table to view the entire table. The future forecasts and their corresponding upper and lower confidence limits are displayed at the bottom of the table.

Model Comparisons

The Model Comparisons option creates a table of the results of comparing five different forecasting models (see Figure 5-4). In general, the better model has smaller MSE and MAE values and values of ME and MPE that are closer to 0. The table also summarizes the results of five tests run on the residuals to determine if each model is adequate for the data.

The Model Comparisons option creates a table of the results of fitting different ARMA models to the data. The model with the lowest value of the Akaike Information Criterion (AIC) has been used to generate the forecasts.

The table also summarizes the results of five tests run on the
residuals to determine whether each model is adequate for the data.


Automatic Forecasting - ibm

Lbl:

Row:

Model Comparison

Data variable: ibm
Number of observations = 120
Start index = 1/93
Sampling interval = 1.0 month(s)
Length of seasonality = 12

Models

(A) ARMA(0,0) SARMA(0,0)
(B) ARMA(1,0) SARMA(1,0)
(C) ARMA(2,1) SARMA(2,1)
(D) ARMA(3,2) SARMA(3,2)
(E) ARMA(4,3) SARMA(4,3)

Estimation Period

Model	MSE	MAE	MAPE	ME	MPE	AIC
(A)	1680.02	35.4937	6.84076	1.08002E-13	-0.610369	7.41819
(B)	27.6536	4.01969	0.780272	0.487854	0.0838139	3.29444
(C)	26.8033	3.96957	0.772209	0.555986	0.0995921	3.22842
(D)	24.8474	3.69798	0.715593	0.072398	0.0157764	3.11661
(E)	6.36625E9	57562.5	10767.0	1066.16	182.581	22.4407

Figure 5-4. Model Comparisons

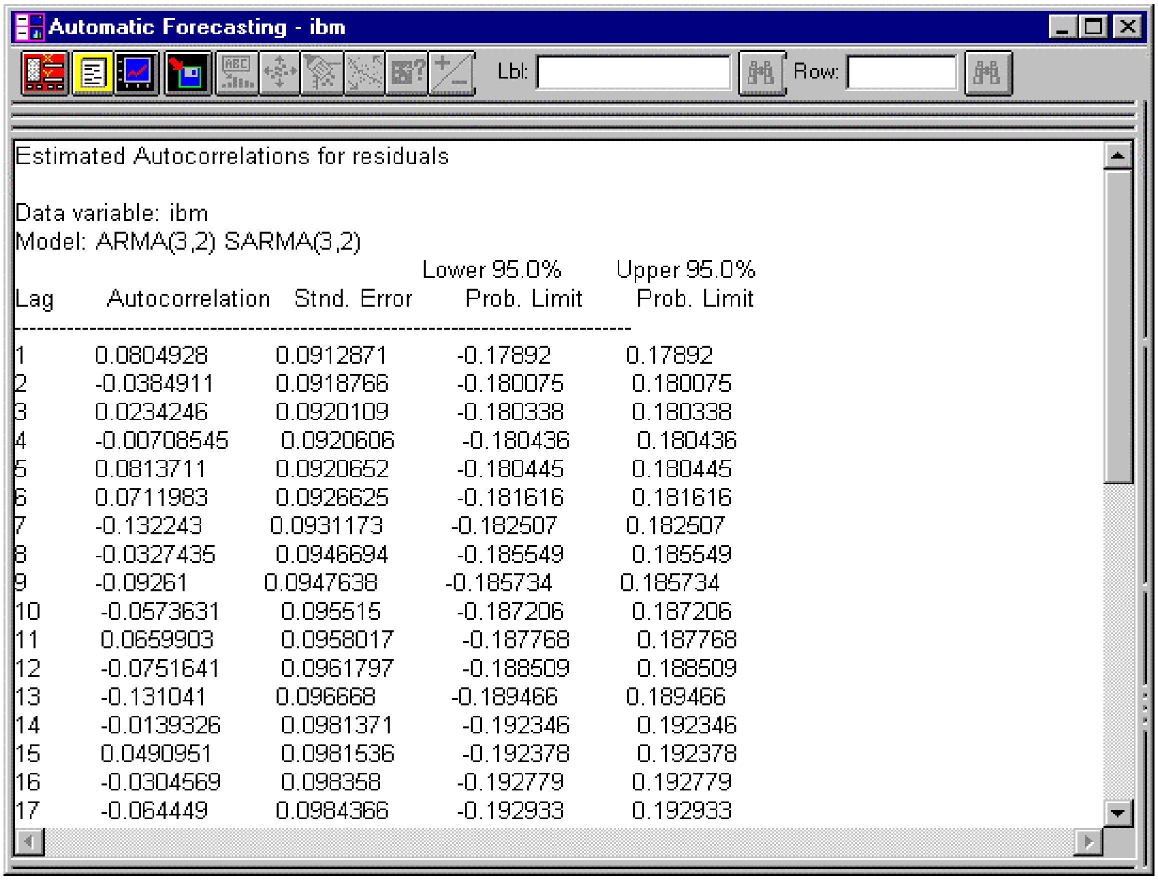
Residual Autocorrelations

The Residual Autocorrelations option creates a table that shows the estimated autocorrelations between the residuals at various lags (see Figure 5-5). The lag *k* autocorrelaton coefficient measures the correlation between the residuals at time *t* and time *t*

- *k*. The probability limits are also shown. If the limits at a particular lag do not contain the estimated coefficient, there is a statistically significant correlaton at that lag.

The table also shows the standard errors, and the upper and lower probability limits for each lag. You can use the residual autocorrelations to determine if the chosen model is appropriate for the data.

Use the *Autocorrelation Function Options* dialog box to enter a value for the maximum number of lags that will be estimated for the residual autocorrelations, and to enter a value for the size of the probability limits that will be used to calculate the residual autocorrelations.



Estimated Autocorrelations for residuals

Data variable: ibm
Model: ARMA(3,2) SARMA(3,2)

Lag	Autocorrelation	Stnd. Error	Lower 95.0% Prob. Limit	Upper 95.0% Prob. Limit
1	0.0804928	0.0912871	-0.17892	0.17892
2	-0.0384911	0.0918766	-0.180075	0.180075
3	0.0234246	0.0920109	-0.180338	0.180338
4	-0.00708545	0.0920606	-0.180436	0.180436
5	0.0813711	0.0920652	-0.180445	0.180445
6	0.0711983	0.0926625	-0.181616	0.181616
7	-0.132243	0.0931173	-0.182507	0.182507
8	-0.0327435	0.0946694	-0.185549	0.185549
9	-0.09261	0.0947638	-0.185734	0.185734
10	-0.0573631	0.095515	-0.187206	0.187206
11	0.0659903	0.0958017	-0.187768	0.187768
12	-0.0751641	0.0961797	-0.188509	0.188509
13	-0.131041	0.096668	-0.189466	0.189466
14	-0.0139326	0.0981371	-0.192346	0.192346
15	0.0490951	0.0981536	-0.192378	0.192378
16	-0.0304569	0.098358	-0.192779	0.192779
17	-0.064449	0.0984366	-0.192933	0.192933

Figure 5-5. Residual Autocorrelations

Residual Partial Autocorrelations

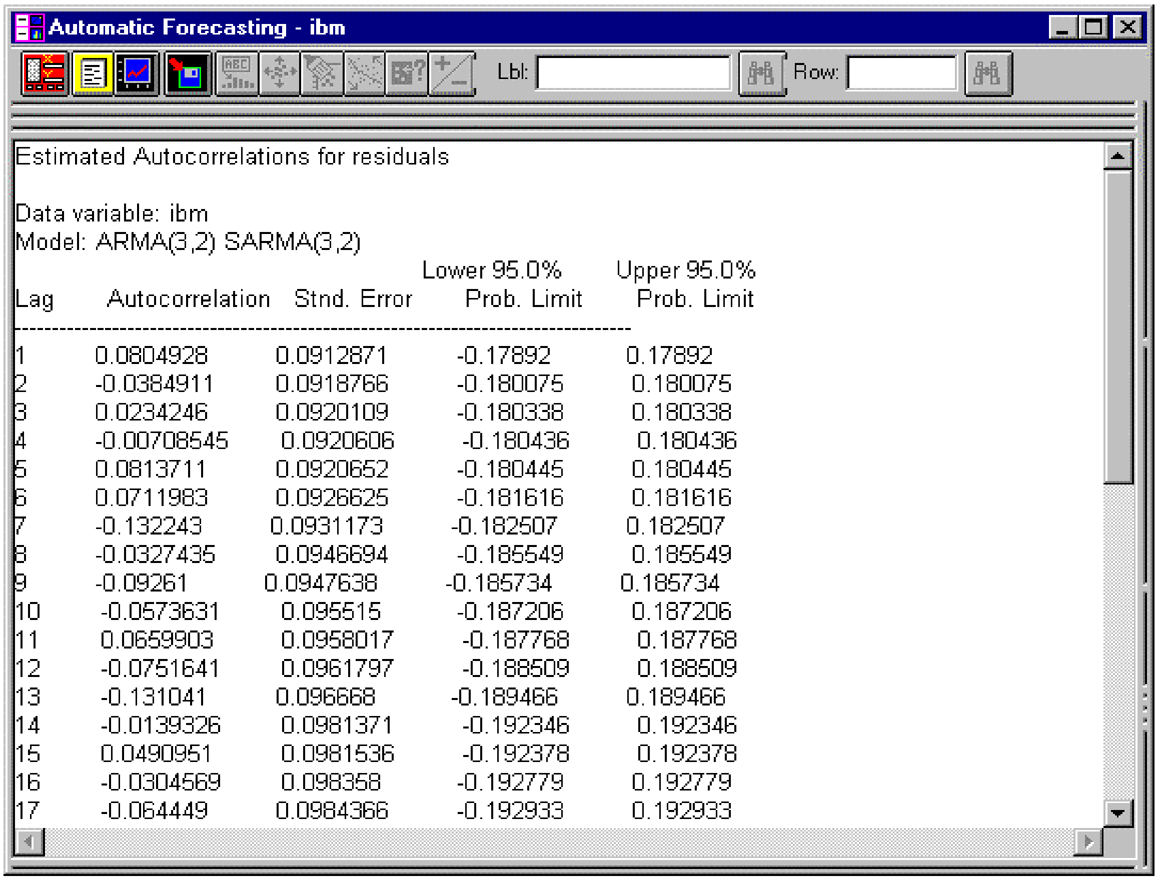
The Residual Partial Autocorrelations option creates a table of the estimated partial autocorrelations between the residuals at various lags (see Figure 5-6). The lag *k* partial autocorrelation coefficient measures the correlation between the residuals at time t and time *t* + *k* after accounting for the correlations at all the lower lags.

Use the results to determine the order of autoregressive model that is needed to fit the data. The probability limits are also shown. If the limits do not contain the estimated coefficient at a particular lag, it is an indication that there is a statistically significant correlation at that lag at the 95 percent confidence level.

Use the *Partial Autocorrelation Function Options* dialog box to enter the maximum number of lags that will be estimated for the residual autocorrelations and to enter the size of the probability limits that will be used to calculate the residual autocorrelations.

Residual Periodogram Table

The Residual Periodogram Table option creates a table that shows the periodogram ordinates for the residuals that are used to identify cycles of fixed frequency of the data (see Figure 5-7). A periodogram is constructed by fitting a series of sine functions at each of the chosen frequencies. The ordinates are equal to the squared amplitudes of the sine functions.

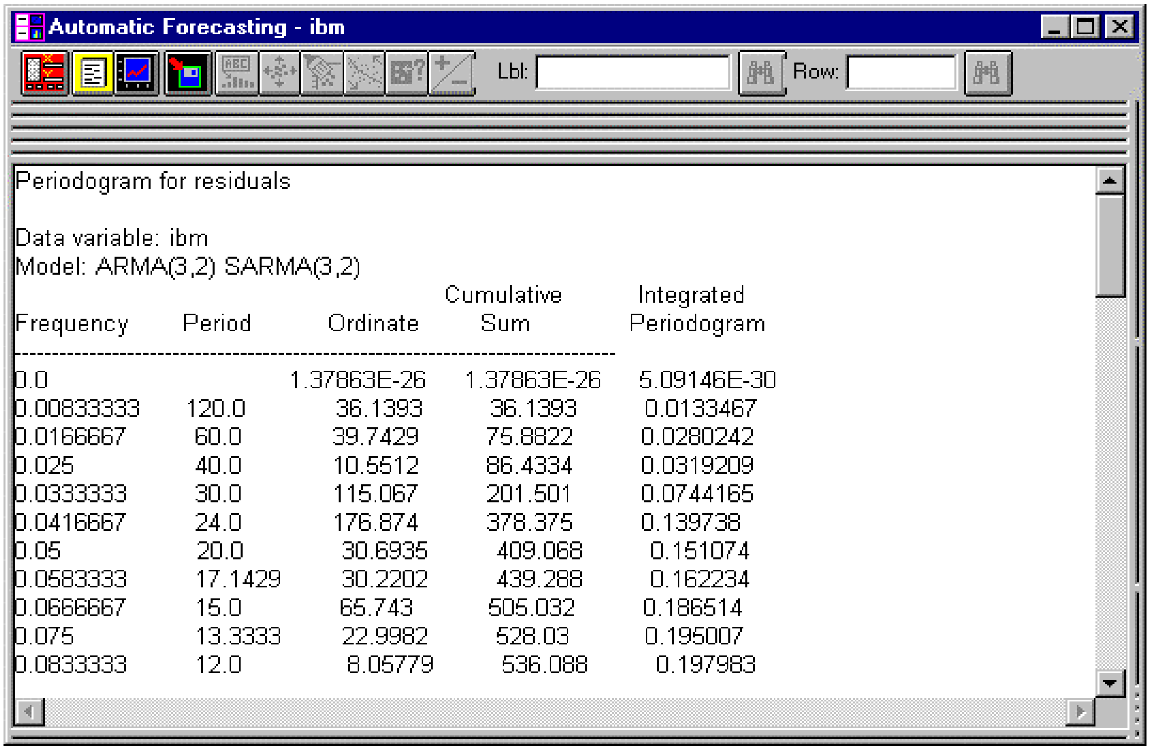


Estimated Autocorrelations for residuals

Data variable: ibm
Model: ARMA(3,2) SARMA(3,2)

Lag	Autocorrelation	Stnd. Error	Lower 95.0% Prob. Limit	Upper 95.0% Prob. Limit
1	0.0804928	0.0912871	-0.17892	0.17892
2	-0.0384911	0.0918766	-0.180075	0.180075
3	0.0234246	0.0920109	-0.180338	0.180338
4	-0.00708545	0.0920606	-0.180436	0.180436
5	0.0813711	0.0920652	-0.180445	0.180445
6	0.0711983	0.0926625	-0.181616	0.181616
7	-0.132243	0.0931173	-0.182507	0.182507
8	-0.0327435	0.0946694	-0.185549	0.185549
9	-0.09261	0.0947638	-0.185734	0.185734
10	-0.0573631	0.095515	-0.187206	0.187206
11	0.0659903	0.0958017	-0.187768	0.187768
12	-0.0751641	0.0961797	-0.188509	0.188509
13	-0.131041	0.096668	-0.189466	0.189466
14	-0.0139326	0.0981371	-0.192346	0.192346
15	0.0490951	0.0981536	-0.192378	0.192378
16	-0.0304569	0.098358	-0.192779	0.192779
17	-0.064449	0.0984366	-0.192933	0.192933

Figure 5-6. Residual Partial Autocorrelations



Periodogram for residuals

Data variable: ibm
Model: ARMA(3,2) SARMA(3,2)

Frequency	Period	Ordinate	Cumulative Sum	Integrated Periodogram
0.0		1.37863E-26	1.37863E-26	5.09146E-30
0.00833333	120.0	36.1393	36.1393	0.0133467
0.0166667	60.0	39.7429	75.8822	0.0280242
0.025	40.0	10.5512	86.4334	0.0319209
0.0333333	30.0	115.067	201.501	0.0744165
0.0416667	24.0	176.874	378.375	0.139738
0.05	20.0	30.6935	409.068	0.151074
0.0583333	17.1429	30.2202	439.288	0.162234
0.0666667	15.0	65.743	505.032	0.186514
0.075	13.3333	22.9982	528.03	0.195007
0.0833333	12.0	8.05779	536.088	0.197983

Figure 5-7. Residual Periodogram Table

A periodogram can also be thought of as an analysis of variance by frequency because the sum of the ordinates equals the total corrected sum of squares in an ANOVA table.

Use the *Periodogram Table Options* dialog box to indicate that the mean should be subtracted from each observation and to indicate that the program should taper the beginning and end of the time-series data using a split cosine-bell window. See Bloomfield (1976) for a discussion of the tapering method.

Residual Tests for Randomness

The Residual Tests for Randomness option runs three tests to determine if the residuals form a random sequence of numbers (see Figure 5-8). A sequence of random numbers is often called white noise because it contains equal contributions at many frequencies.

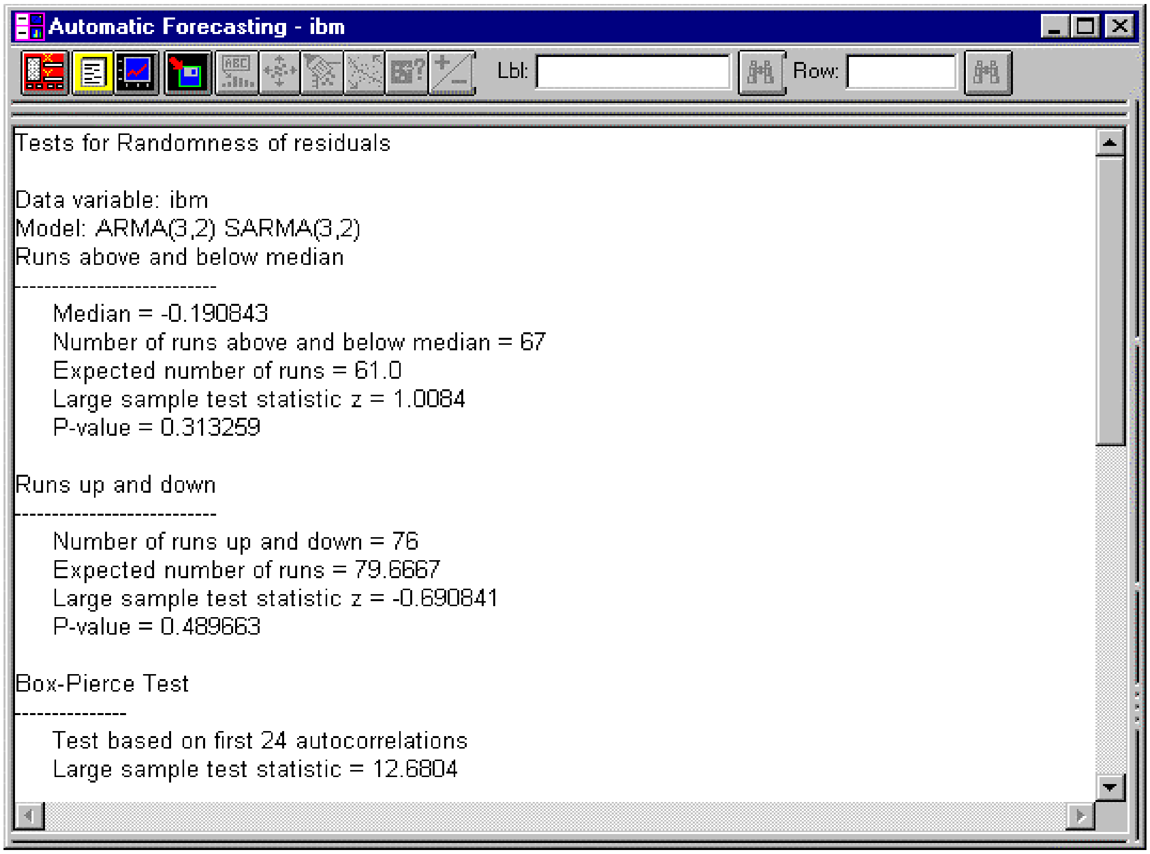


Figure 5-8. Residual Tests for Randomness

The tests are Runs Above and Below the Median, Runs Up and Down, and the Box-Pierce test. The Runs Above and Below the Median test counts the number of runs that are completely above or completely below the median; it ignores the values that are equal to the median. The test is sensitive to trends in the data. If the test statistic is large (corresponds to a *p*-value less than .05), it can be concluded that the values occur in nonrandom order.

The Runs Up and Down test counts the number of times a sequence rises or falls; it is sensitive to long-term cycles. If the test statistic is small (corresponds to a *p*-value greater than .05), it can be concluded that the values occur in random order.

The Box-Pierce test determines if the autocorrelation is equal to 0. If the test statistic is large (corresponds to a *p*-value less than .05), the autocorrelation is not equal to 0, which indicates that the model is not adequate.

Use the *Box-Pierce Test Options* dialog box to enter the maximum number of lags for which statistics will be estimated for the test.

Graphical Options

Time Sequence Plot

The Time Sequence Plot option creates a connected line plot for the forecasted values against time (see Figure 5-9). The plot also includes the prediction limits

for the forecasts, which shows with 95 percent confidence, where the true value of the variable is likely to be at any point in the future.

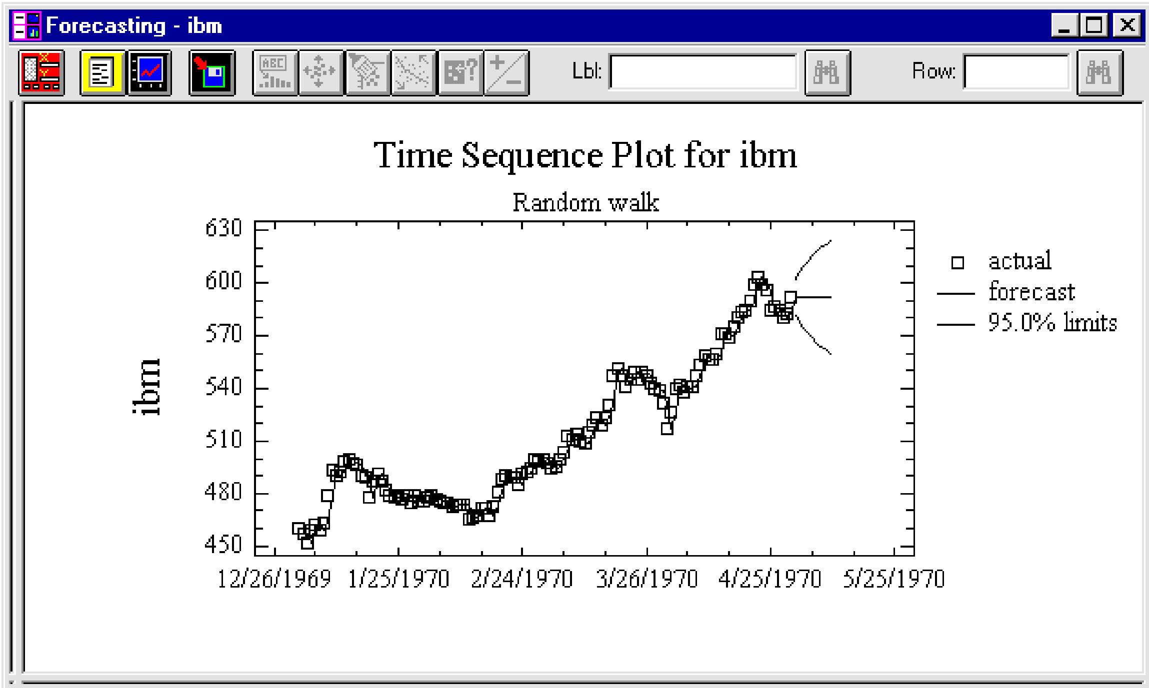


Figure 5-9. Time Sequence Plot

Use the *Forecast Limits Options* dialog box to enter the percentage that will be used for the calculations.

Forecast Plot
The Forecast Plot option creates a connected line plot of the forecasted values (see Figure 5-10). The plot also includes the prediction limits for the forecasts, which shows with 95 percent confidence, where the true value of the variable is likely to be at any point in the future.

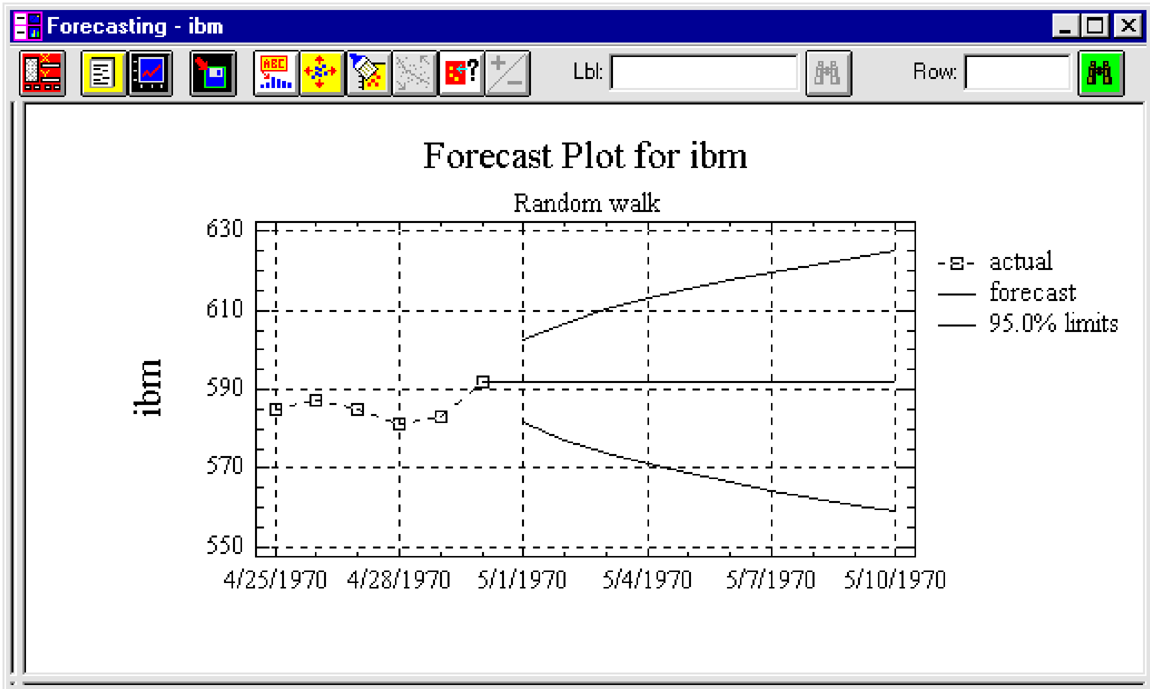


Figure 5-10. Forecast Plot

Use the *Forecast Limits Options* dialog box to enter the percentage that will be used for the calculations.

Residual Plots
The Residual Plots option creates plots of the residuals from the fitted model. If all the dynamic structure in the variable is captured, the residuals should be random (white noise).

Use the *Forecasting Residual Plots Options* dialog box to choose the type of plot that will be created.

Time Sequence Plot - a connected line plot of the residuals against time (see Figure 5-11).

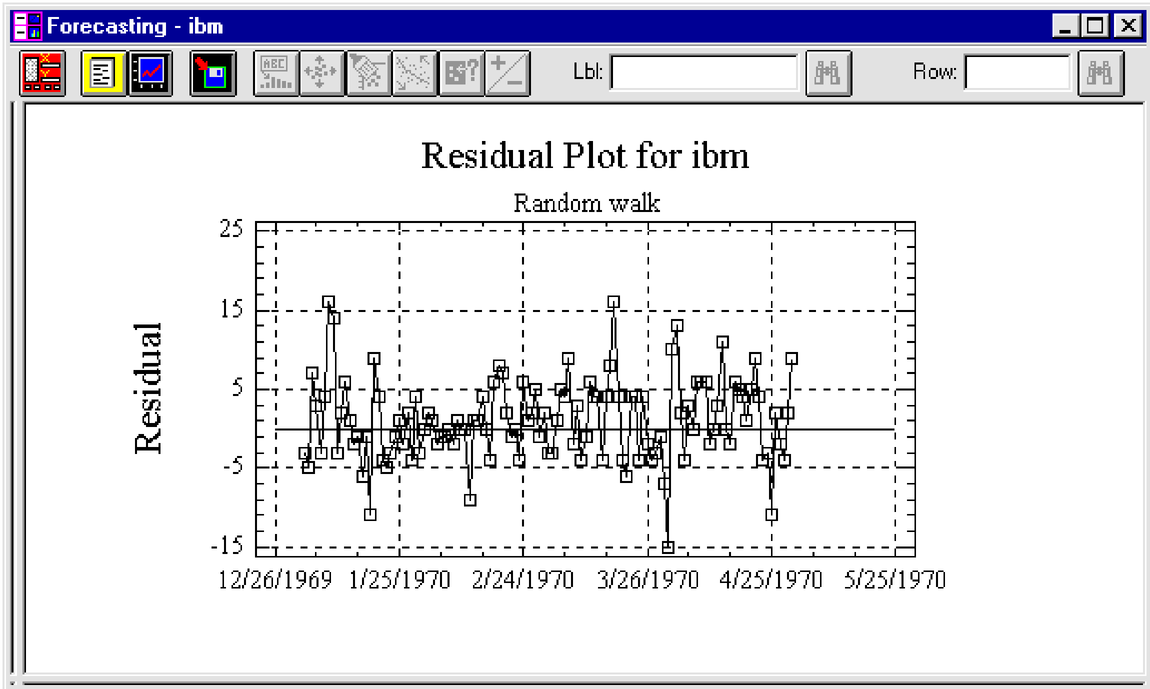


Figure 5-11. Time Sequence Plot of the Residuals

Probability Plot (Horizontal) - a plot of the residuals on the Y-axis from smallest to largest (see Figure 5-12). The X-axis is a normal probability scale (it "straightens out" the plot of a cumulative normal distribution). If the residuals do not fall approximately along a straight line, they are not from a normal distribution. This suggests that you may not have used the best model for the data.

Probability Plot (Vertical) - a plot of the residuals on the X-axis from smallest to largest. The Y-axis is a normal probability scale (it "straightens out" the plot of a cumulative normal distribution). If the residuals do not fall approximately along a straight line, they are not from a normal distribution. This suggests that you may not have used the best model for the data.

Residual Autocorrelation Function

The Residual Autocorrelation Function option creates a plot of the estimated autocorrelations between the residuals at various lags (see Figure 5-13). The lag *k* autocorrelation coefficient measures the correlations between the residuals at time *t* and time *t* - *k*. The probability limits are also shown. If the limits do not contain the estimated coefficient at a particular lag, it is an indication that there is a statistically significant correlation at that lag at the 95 percent confidence level.

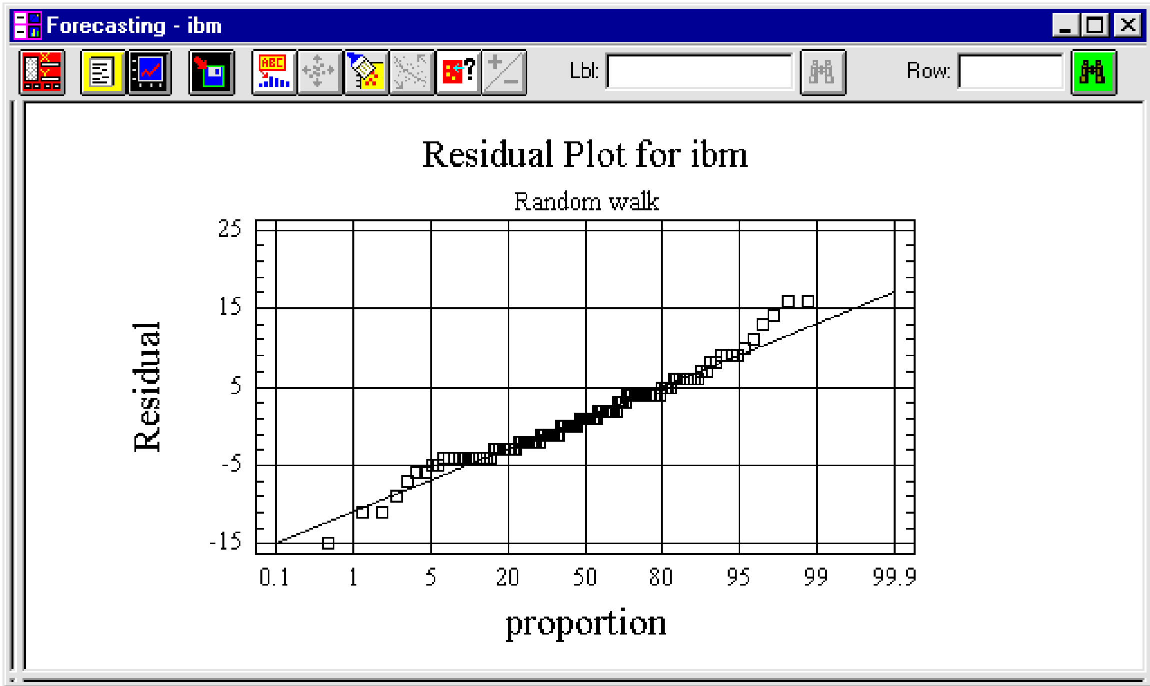


Figure 5-12. Probability Plot (Horizontal)

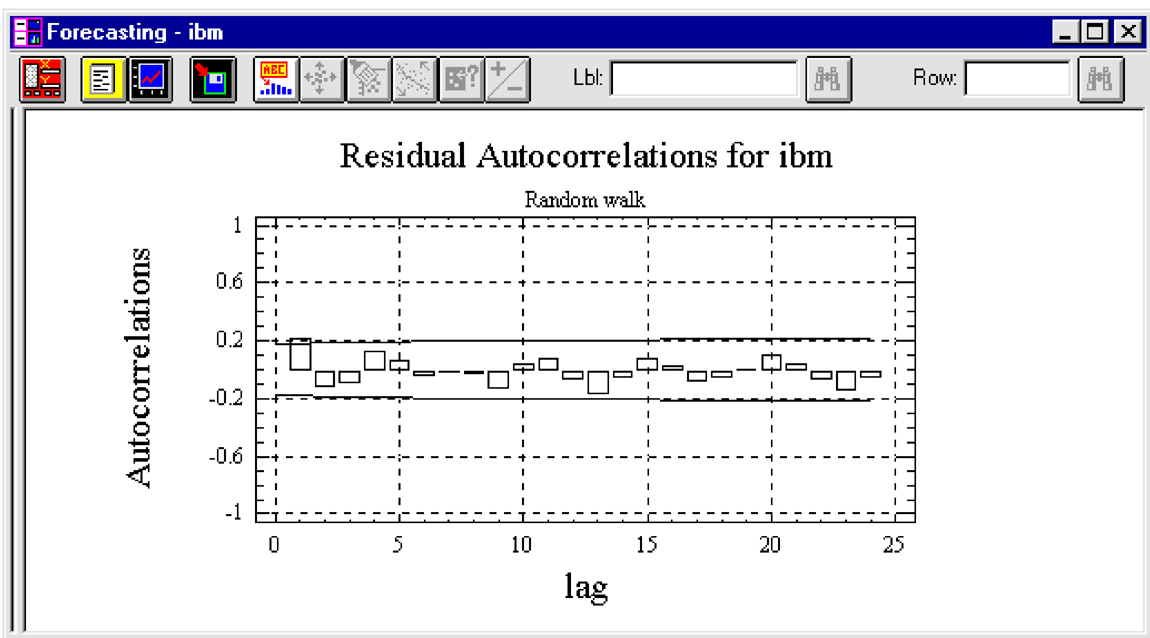


Figure 5-13. Residual Autocorrelation Function Plot

The plot contains vertical bars that represent the coefficient for each lag and a pair of dotted lines at a distance from the baseline that are a multiple of the standard error at each lag. Significant autocorrelations extend above or below the confidence limits.

Use the *Autocorrelation Function Options* dialog box to enter the maximum number of lags that will be estimated for the residual autocorrelations and to enter the size of the probability limits that will be used to calculate the residual autocorrelations.

Residual Partial Autocorrelation Function

The Residual Partial Autocorrelation Function option creates a plot of the estimated partial autocorrelations between the residuals at various lags (see Figure 5-14).

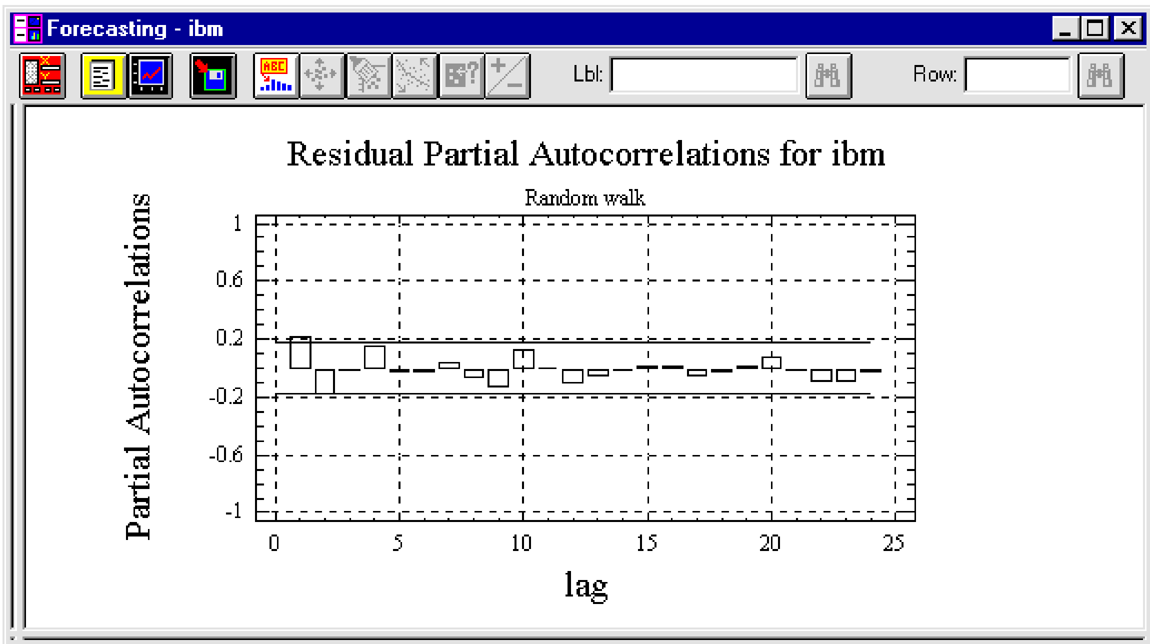


Figure 5-14. Residual Partial Autocorrelation Function Plot

The lag *k* partial autocorrelation coefficient measures the correlation between the residuals at time *t* and time *t* + *k* after accounting for the correlations at all the lower lags.

Each coefficient is plotted using a bar whose height is proportional to the value of the coefficient. By default, it plots probability limits as dashed lines at plus or minus 2 divided by the square root of the number of observations in the time series. The bounds are useful for indicating partial autocorrelations that are significantly different from 0. Bars that extend beyond either line indicate significant correlations.
Use the *Partial Autocorrelation Function Options* dialog box to enter the maximum number of lags that will be estimated for the residual autocorrelations and to enter the size of the probability limits that will be used to calculate the residual autocorrelations.

Residual Periodogram

The Residual Periodogram option creates plot that shows the periodogram ordinates for the residuals (see Figure 5-15). The periodogram is constructed by fitting a series of sine functions at each specified frequency. The ordinates are equal to the squared amplitudes of the sine functions.

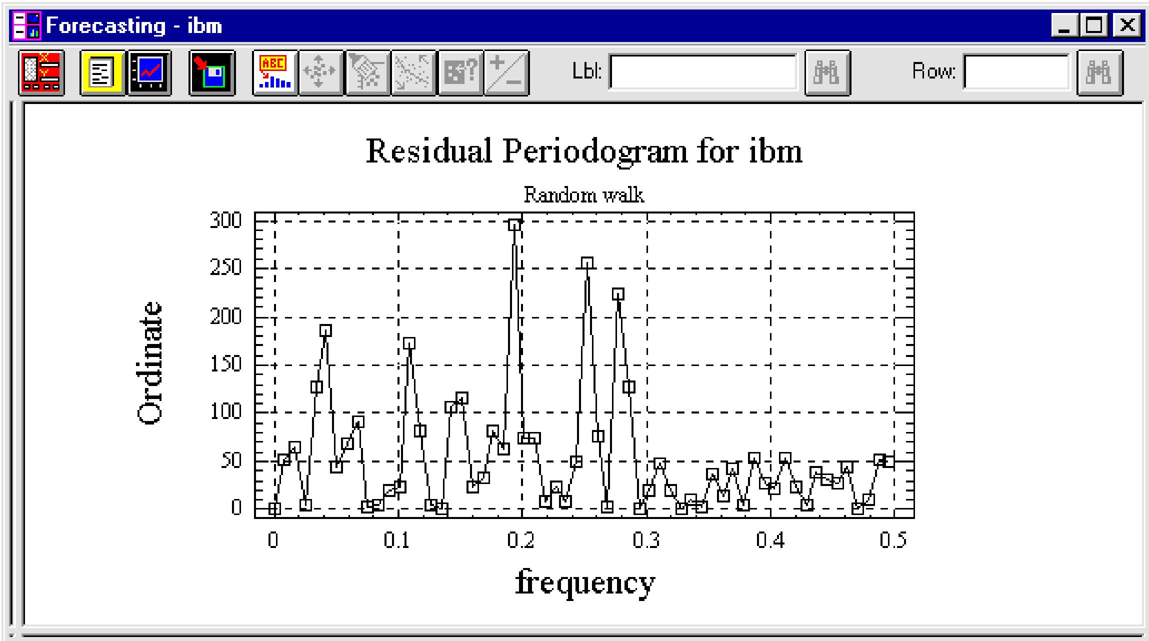


Figure 5-15. Residual Periodogram

A residual periodogram is similar to an analysis of variance by frequency because the sum of the ordinates equals the total sum of squares in an ANOVA table.

Use the Periodogram Options dialog box to indicate that the mean should be subtracted from each observation, to indicate if the observations should be plotted as points, to indicate if lines should connect the observations, and to indicate that the time-series data should be tapered at its beginning and end using a split cosine-bell window. See Bloomfield (1976) for a discussion of the tapering method.

Residual Crosscorrelation Function
The Residual Crosscorrelation Function option creates a plot that shows the crosscorrelations between the residuals and the chosen variable (see Figure 5-16).

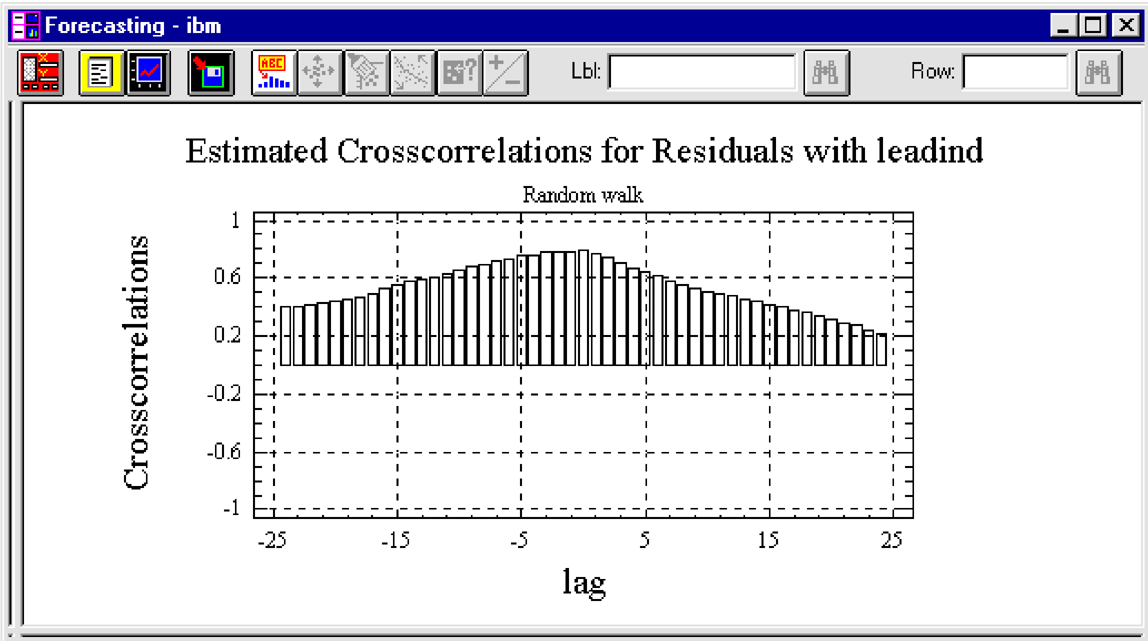


Figure 5-16. Residual Crosscorrelation Function

The crosscorrelation at lag k measures the strength of the linear relationship between the variable's value at time t and the value of the second variable k periods earlier. The plot helps determine if the second variable is helpful in forecasting the first variable.

Use the Crosscorrelation Function Analysis dialog box to choose the variable that contains the second set of time-series data, and to enter the maximum number of lags that will be calculated for the crosscorrelation function.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are ten selections: Data, Adjusted Data, Forecasts, Upper Forecast Limits, Lower Forecast Limits, Residuals, Autocorrelations, Partial Autocorrelations, Residual Periodogram Ordinates, and Fourier Frequencies.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Bloomfield, P. 1976. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.

Box, G. E. P. and Cox, D. R. 1964. "An Analysis of Transformations," *Journal of the Royal Statistical Society*, 26:211-243.

Box, G. E. P. and Jenkins, G. M. 1976. *Time Series Analysis, Forecasting, and Control*, second edition. San Francisco: Holden-Day.

Brown, R. G. 1977. *Materials Management*. New York: Wiley.

Brown, R. G. 1963. *Smoothing, Forecasting, and Prediction of Discrete Time Series*. Englewood Cliffs: Prentice-Hall.

Brown, R. G. and Meyer, R. F. 1983. "The Fundamental Theorem of Exponential Smoothing," *Operations Research*, 9:1961.

Kendall, M. G. and Stuart, A. 1961. *The Advanced Theory of Statistics*, Volume 2. New York: Hafner Publishing Company.

Makridakis, S. and Wheelwright, S. C. 1978. *Interactive Forecasting: Univariate & Multivariate Methods*. Oakland, California: Holden-Day.

Montgomery, D., Johnson, L., and Gardiner, J. 1990. *Forecasting and Time Series Analysis*, second edition. New York: McGraw-Hill.

Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Models*, second edition. Homewood, Illinois: Richard E. Irwin, Inc.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics and Computing of Exploratory Data Analysis*. Boston: PWS-Kent Publishing Company.

Winters, P. R. 1960. "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science*, 6(3):324-342.