How To: Forecast Seasonal Time Series Data

Using STATGRAPHICS Centurion

by

Dr. Neil W. Polhemus

October 31, 2005

Introduction

When data are recorded at equally spaced points in time, the data are commonly referred to as a *time series*. Typical examples include monthly sales, daily closing stock prices, weekly airline passenger miles, and automated samples taken from a manufacturing process. When the data show cyclical ups and downs at a fixed period, such as sales of candy that peak around Easter and Halloween, the data are said to have a *seasonal* component.

This guide examines procedures designed to analyze and forecast seasonal time series data. In STATGRAPHICS Centurion, there are several important procedures for handling such data:

> **Descriptive Methods** – for plotting the data and identifying the frequency of a seasonal component if it is not known already.
>
> **Seasonal Decomposition** – for splitting the time series into trend, cycle, seasonal, and irregular components.
>
> **User-Specified Model** – for fitting a forecasting model specified by the analyst.
>
> **Automatic Forecasting** – for automatic selection of the best fitting forecast model from amongst different candidates.

Sample Data

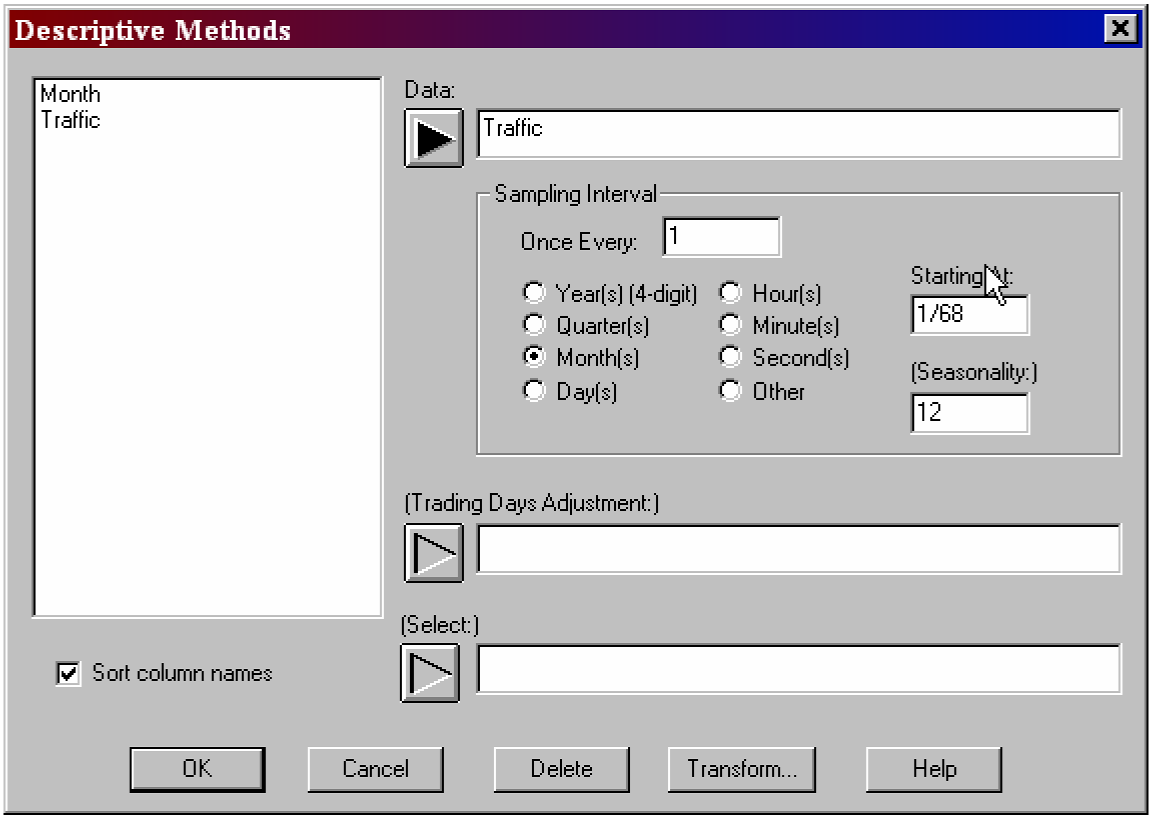
As an example, we will consider recorded monthly traffic volumes across the Golden Gate Bridge during the period from January, 1968 through December, 1981. This data was taken from the records of the Golden Gate Bridge Authority while the author was on sabbatical at the University of California at Berkeley. It covers a time period during which there were two major gasoline shortages, both of which had significant impact on traffic in the Bay Area. The data are contained in the file *Howto5.sf6*.

Step 1: Descriptive Methods

The first step in analyzing any data is to plot it. For time series data, this is best done by:

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

The procedure begins by displaying a dialog box on which to specify information about the time series:



Descriptive Methods

Month
Traffic

Data: Traffic

Sampling Interval

Once Every: 1

○ Year(s) (4-digit)
○ Quarter(s)
● Month(s)
○ Day(s)
○ Hour(s)
○ Minute(s)
○ Second(s)
○ Other

Starting At: 1/68

(Seasonality:) 12

(Trading Days Adjustment:)

(Select:)

☑ Sort column names

Figure 1: Data Input Dialog Box for Descriptive Methods Procedure

The data have been placed in a column of the datasheet named *Traffic*. The dialog box indicates that the data were sampled once every month, beginning in January of 1968. It also suggests that the data may have a *seasonality* of 12, which means that we expect to see a cyclical effect with a period of 12 months. If the seasonality is not known, this field may be left blank.

The analysis window displays several panes, including a *Time Sequence Plot* of the data:

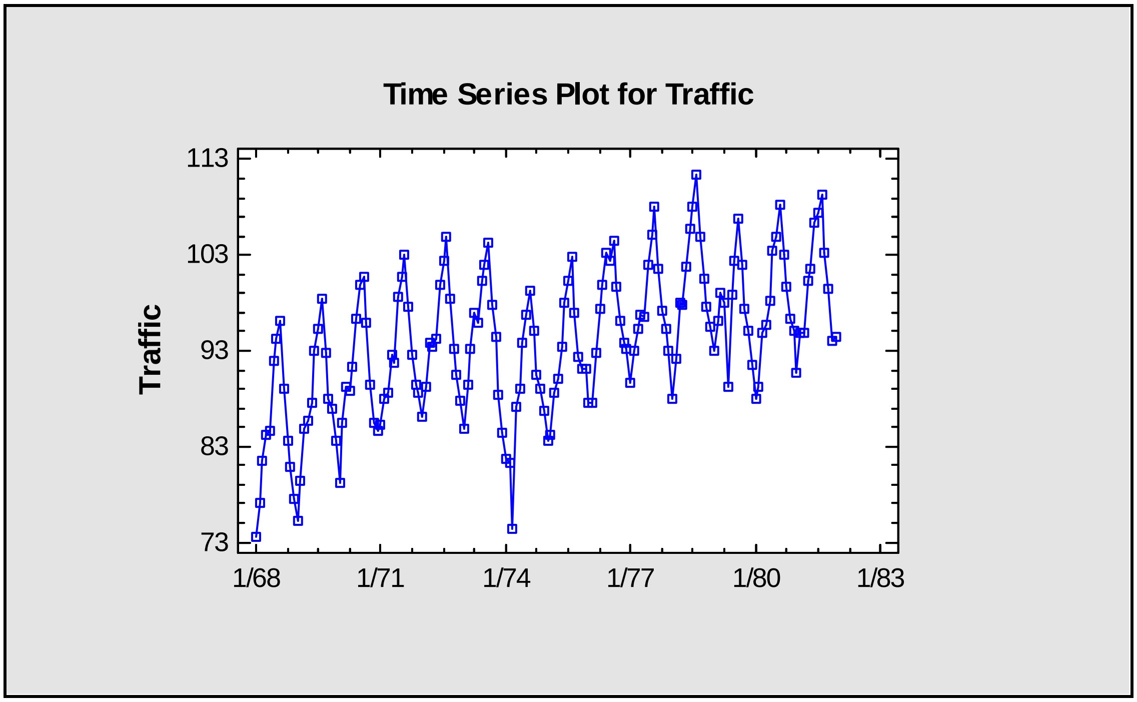


Figure 2: Time Sequence Plot of Monthly Traffic Volumes

Notice that the overall trend during the sampling period was up, although that trend was interrupted twice: in the fall of 1973, and in the spring of 1979, both times due to gasoline shortages which caused a drop in traffic across the bridge. Note also the very regular cycle of ups and downs every 12 months, rising to a peak during the summer months when tourist traffic in San Francisco is the heaviest.

A second important tool for examining seasonal data is the *Autocorrelation Function*, shown below:

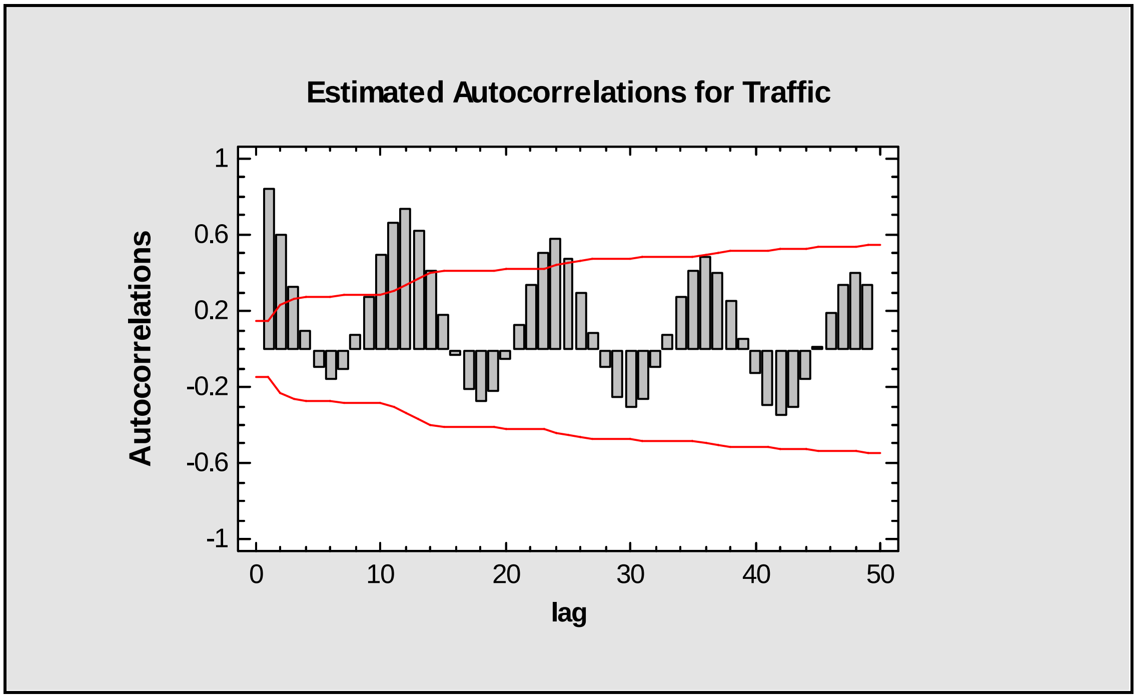


Figure 3: Sample Autocorrelation Function

The autocorrelation at lag *k* is defined as the correlation between observations separated by *k* time periods. For example, the relatively large positive correlation at lag 1 reflects the fact that observations one month apart tend to be similar in magnitude. The correlation falls off as the separation between observations increases, but rises again at lags 12, 24, 36, and 48, reflecting the fact that traffic in a given month tends to be similar to that observed during the same month in previous years. This is a clear indication of a seasonality of order 12.

One additional plot worth noting is the *Periodogram*:

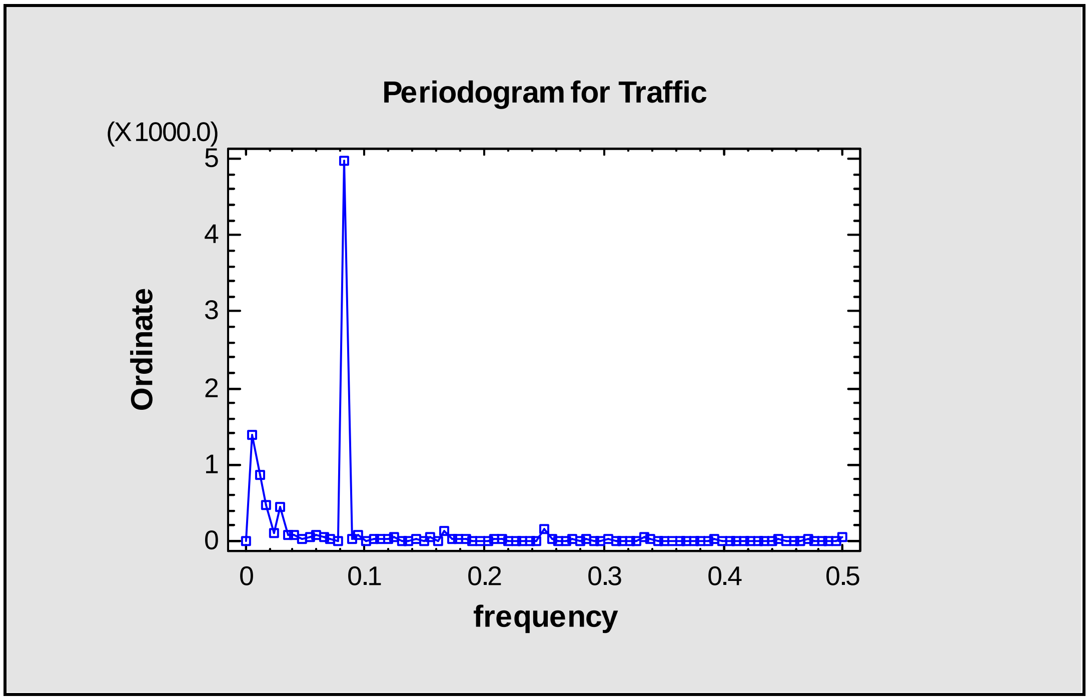


Figure 4: Sample Periodogram

The periodogram shows the contribution to the overall variance of the time series data attributable to oscillations at different frequencies. It is based on a Fourier decomposition of the data into a sum of sine waves. A strong seasonal component at a frequency such as 1/12 months will result in a large spike as shown above. You can also see very small peaks at multiples of the fundamental frequency (call harmonics), which reflect the fact that the seasonal oscillation is not very sinusoidal. The contributions at very low frequencies come from the overall trend in the series.

The periodogram is one of the best tools for identifying cyclical patterns at unknown frequencies. If a large peak is observed, it may well provide a clue to some important source of variability in the data. In a manufacturing process, removal of that source could reduce the overall variability in the process.

Step 2: Seasonal Decomposition

When analyzing time series data, it is common to view the data as consisting of four components:

1. *Trend* (T) – a general long-term pattern observed over the entire data set. For example, many economic time series tend to show an increasing trend when viewed over many years.

2. *Cycle (C)* – cyclical variations around the trend line. Unlike seasonal effects, these cycles do not have a fixed frequency. General up and downs of the world economy is a typical example.

3. *Seasonality (S)* – cyclical variations with a fixed frequency, such as yearly cycles in the sales of lawnmowers. Seasonal effects repeat on a regular and predictable basis.

4. *Random or Irregular (R)* – the residual component left behind after the other three components have been accounted for.

There are two basic models upon which a decomposition of a time series into its component parts may be based: a *multiplicative* model and an *additive* model. The multiplicative model assumes that the data at time *t* may be represented as the product of the four components according to:

$$Y_t = T_t C_t S_t R_t$$

The additive model assumes that the components add:

$$Y_t = T_t + C_t + S_t + R_t$$

To decompose an observed time series into its various components:

- If using the Classic menu, select: *Describe – Time Series – Seasonal Decomposition*.
- If using the Six Sigma menu, select: *Forecast – Seasonal Decomposition*.

The data input dialog box is the same as for *Descriptive Methods*. By default, a multiplicative model is assumed, although *Analysis Options* may be used to switch to an additive model. Three graphs are useful in illustrating the decomposition. The first displays the *Trend-Cycle*:

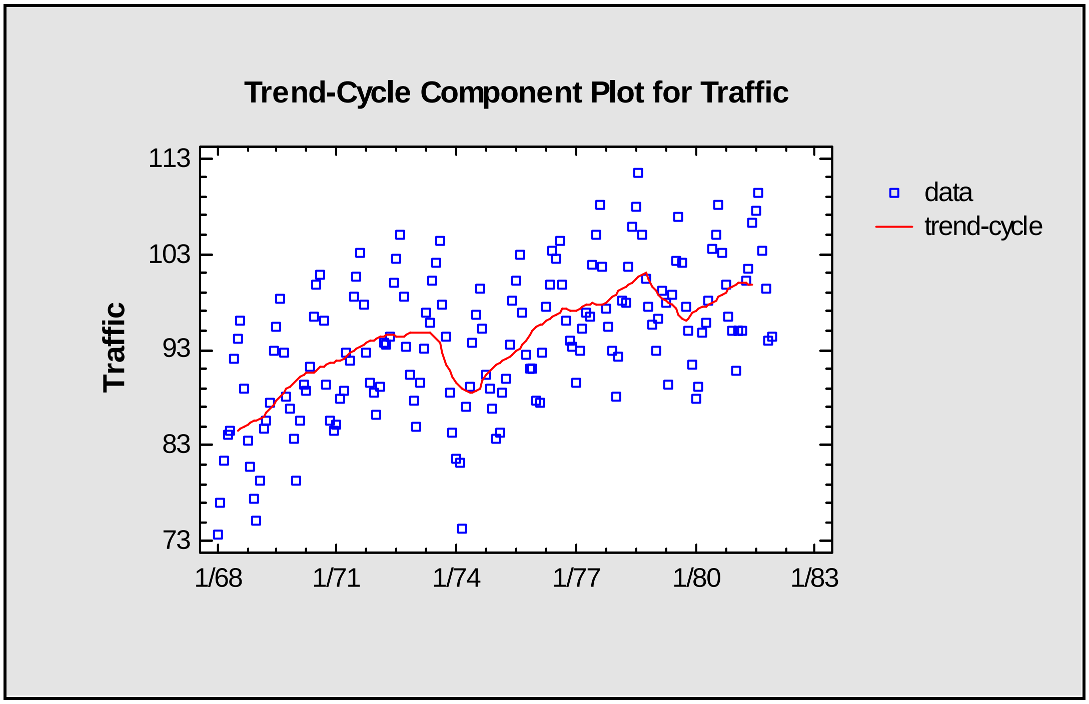


Figure 5: Plot of Trend-Cycle Component

This plot displays the original data, on which a moving average of length equal to the seasonal order has been added. The moving average estimates the combined trend and cycle components T_tC_t, which are not usually separated. The effect of the two gasoline shortages is easily seen.

The second plot displays the *Seasonal Indices*:

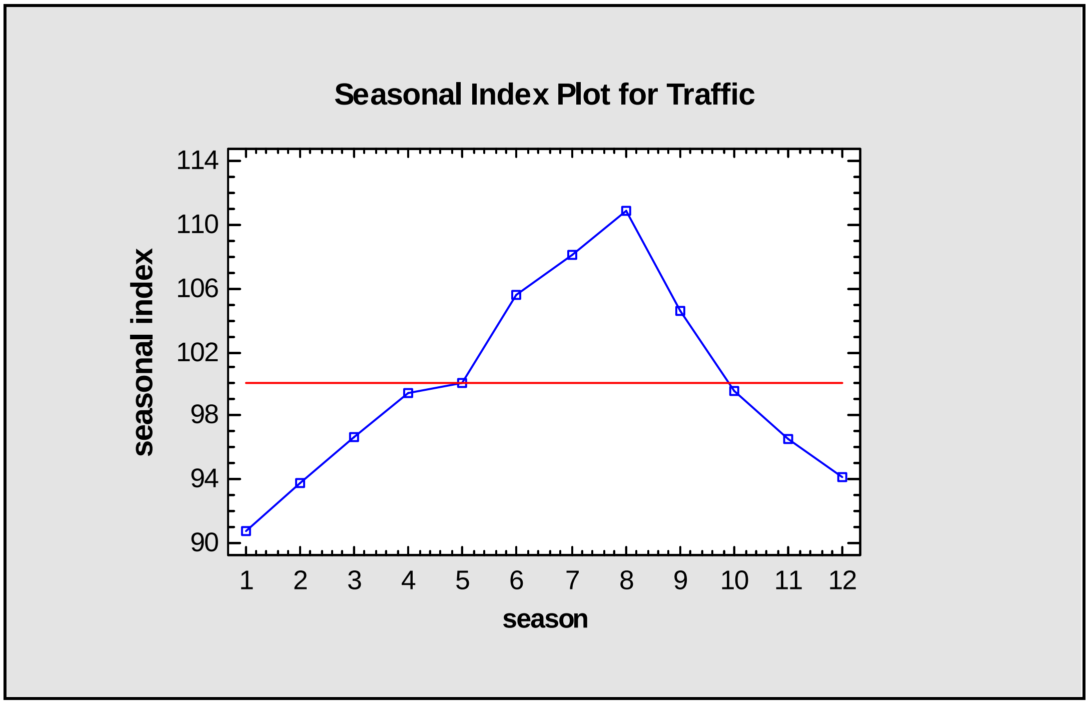


Figure 6: Plot of Seasonal Indices

The seasonal indices estimate the seasonal component S_t. When using a multiplicative model, the indices are expressed on a percentage basis, such that an index of 90 for a given month would indicate that such a month is only 90% of an average month. Note the strong seasonal effect for the traffic data, rising from a low in January to a peak in August and then falling off again.

The third graph displays the *Irregular Componen*t:

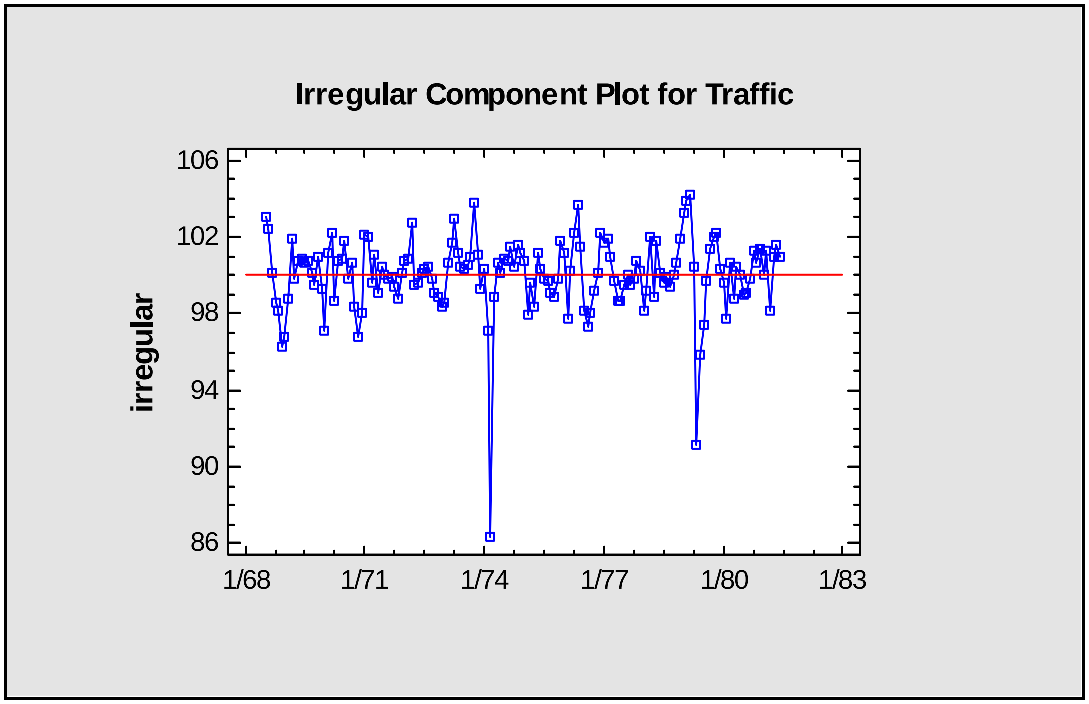


Figure 7: Plot of Irregular or Residual Component

For the multiplicative model, this component is also expressed on a percentage basis, with the average value scaled to equal 100. In March of 1974, the irregular component fell to approximately 86%, implying that traffic during that month was 14% less than expected, having already accounted for the trend-cycle and seasonal components.

Once the decomposition has been performed, we can take the original data and divide it by the estimated seasonal indices to obtain the seasonally adjusted data Y_t', defined by:

$$Y_t' = \frac{Y_t}{S_t}$$

The seasonally adjusted data are plotted below:

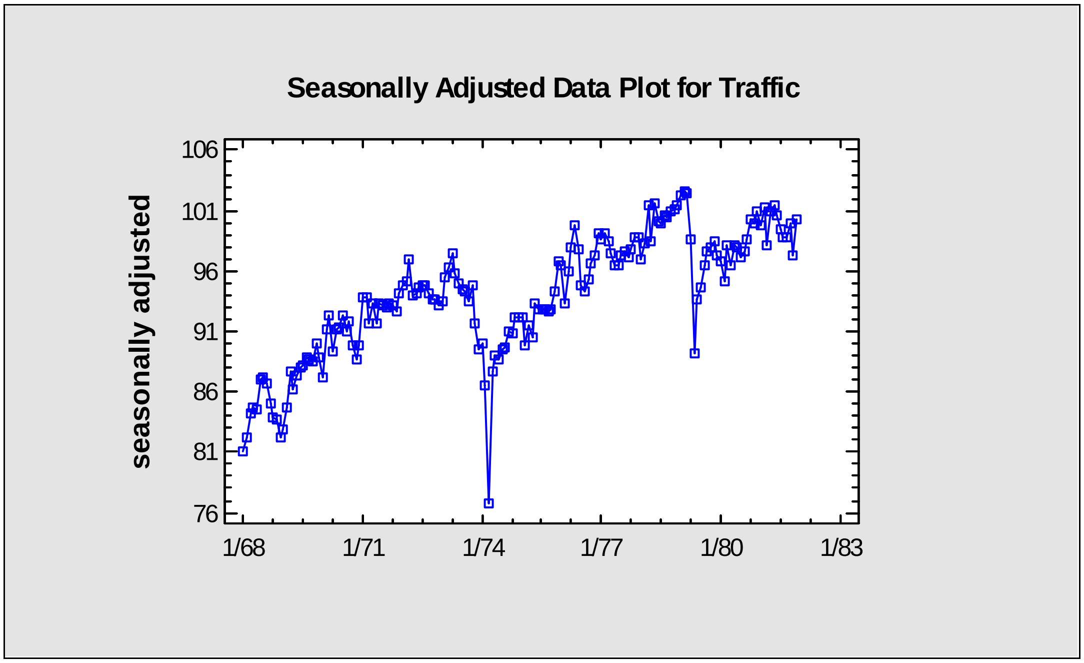


Figure 8: Seasonally Adjusted Data

The seasonally adjusted data retain the trend-cycle and irregular components. However, the seasonality has been removed.

Step 3: Forecasting

Let's suppose that we now wish to forecast the values of traffic across the Golden Gate Bridge in the months immediately following the end of the data sample. We can do this in two ways:

1. Use a seasonal forecasting model to forecast the traffic data directly.

2. Use a nonseasonal forecasting model to forecast the seasonally adjusted data, and then multiply the nonseasonal forecasts by the seasonal indices to put the seasonality back.

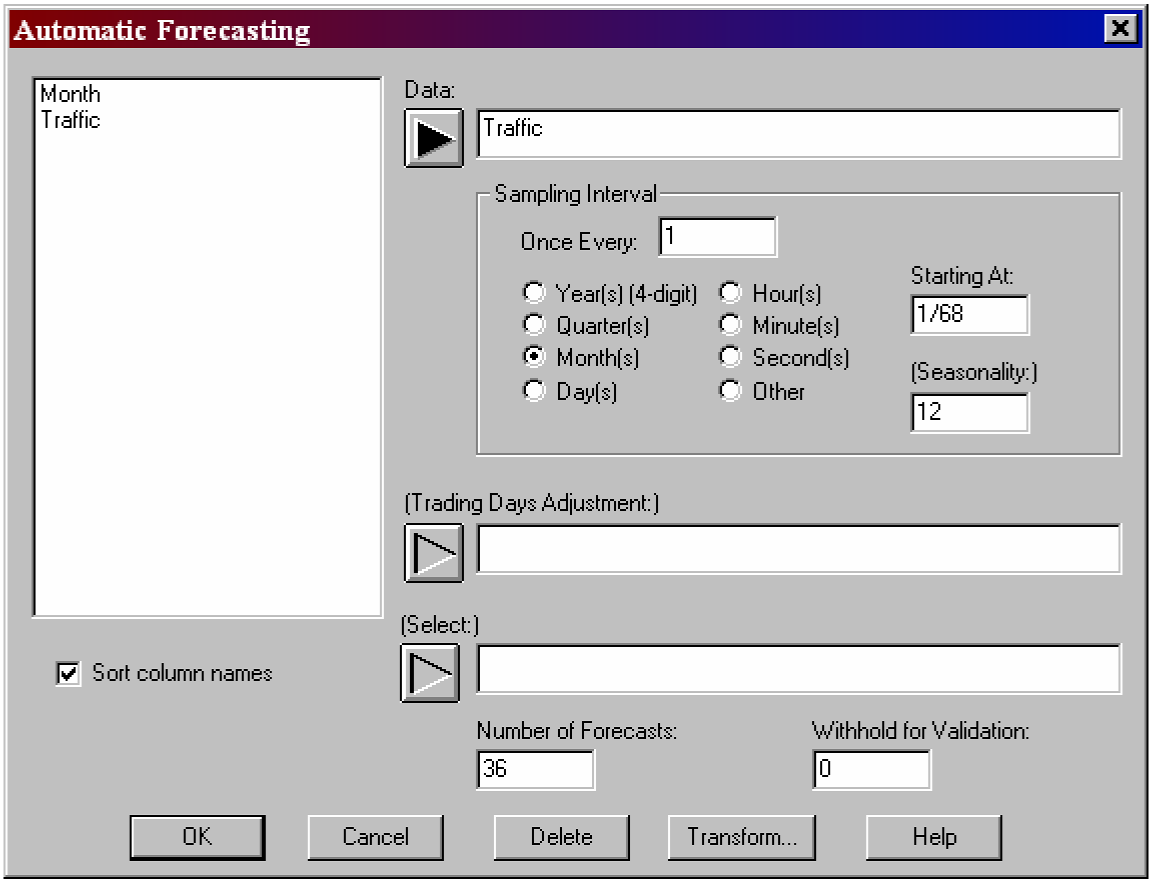
We will try both methods and compare the results.

Seasonal Forecasting Models
To fit a seasonal forecasting model to the Golden Gate Bridge time series, we will use the *Automatic Forecasting* procedure. To access this procedure:

- If using the Classic menu, select: *Describe – Time Series – Automatic Model Selection*.
- If using the Six Sigma menu, select: *Forecast – Forecasting - Automatic Model Selection*.

This procedure will generate forecasts using various procedures and select the method that works best according to a selected information criterion.

On the data input dialog box, be sure the *Seasonality* field is set to 12:



Automatic Forecasting

Month
Traffic

Data: Traffic

Sampling Interval

Once Every: 1

○ Year(s) (4-digit)
○ Quarter(s)
● Month(s)
○ Day(s)
○ Hour(s)
○ Minute(s)
○ Second(s)
○ Other

Starting At: 1/68

(Seasonality:) 12

(Trading Days Adjustment:)

(Select:)

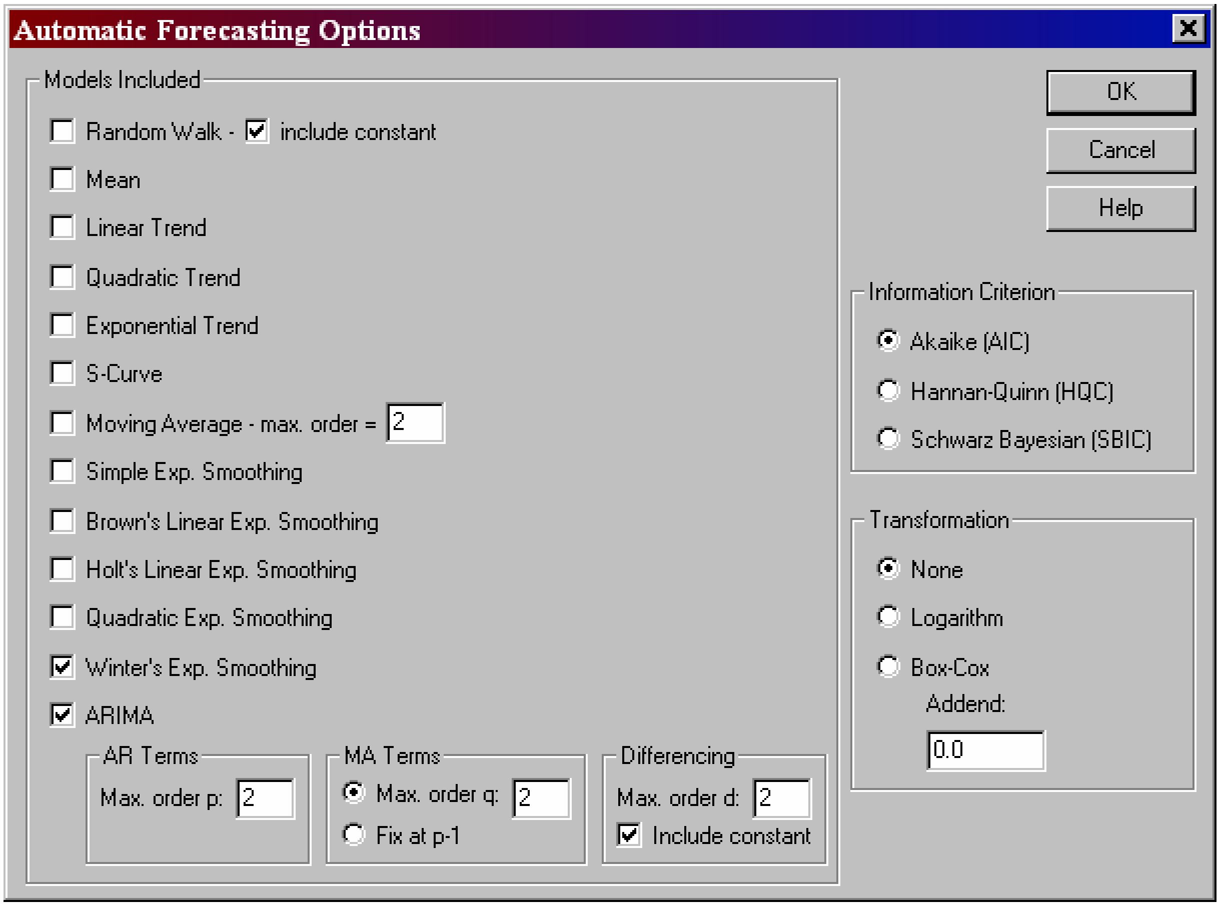
☑ Sort column names

Number of Forecasts: 36

Withhold for Validation: 0

Figure 9: Data Input Dialog Box for Automatic Forecasting

Press OK to create an analysis window. Then select *Analysis Options* to specify the models to be considered:



Models Included

☐ Random Walk - ☑ include constant
☐ Mean
☐ Linear Trend
☐ Quadratic Trend
☐ Exponential Trend
☐ S-Curve
☐ Moving Average - max. order = 2
☐ Simple Exp. Smoothing
☐ Brown's Linear Exp. Smoothing
☐ Holt's Linear Exp. Smoothing
☐ Quadratic Exp. Smoothing
☑ Winter's Exp. Smoothing
☑ ARIMA

AR Terms
Max. order p: 2

MA Terms
● Max. order q: 2
○ Fix at p-1

Differencing
Max. order d: 2
☑ Include constant

Information Criterion
● Akaike (AIC)
○ Hannan-Quinn (HQC)
○ Schwarz Bayesian (SBIC)

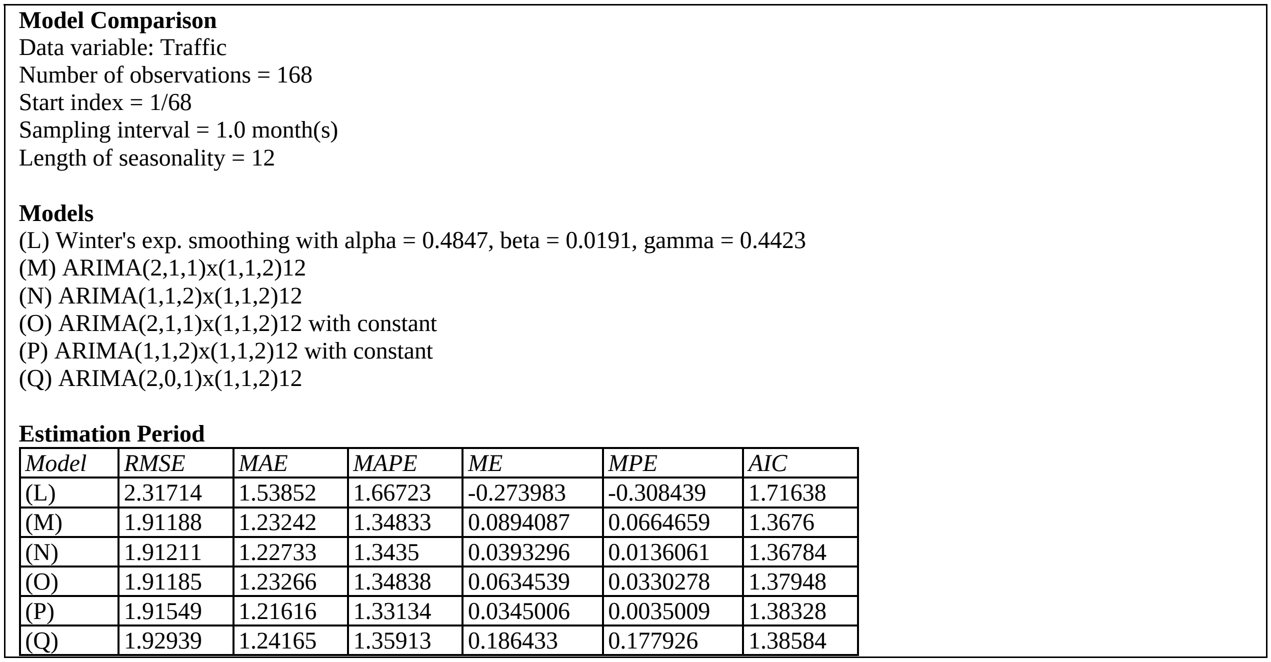
Transformation
● None
○ Logarithm
○ Box-Cox
Addend: 0.0

Figure 10: Analysis Options Dialog Box with Selection of Seasonal Models

There are two models that are useful for forecasting data with a strong seasonal component: Winter’s Exponential Smoothing, and the ARIMA models. Winter’s Exponential Smoothing assumes that the data follow a multiplicative model and uses triple exponential smoothing to estimate a linear trend and seasonal indices. The ARIMA models are parametric time series models that describe the observation at time *t* as a linear combination of “shocks” to the system at time *t* and earlier times. The details of both methods are described in the STATGRAPHICS documentation for the *Forecasting* procedure.

The *Automatic Forecasting* procedure will fit all requested models, optimizing any model parameters. In the case of the ARIMA models, it will try various combinations of (*p*,*d*,*q*) and (*P*,*D*,*Q*), where *p* is the order of the nonseasonal autoregressive (AR) component, *d* is the order of nonseasonal differencing, *q* is the order of the nonseasonal moving average (MA) component, *P* is the order of the seasonal autoregressive component, *D* is the order of seasonal differencing, and *Q* is the order of the seasonal moving average component.

The *Model Comparisons* table compares the fit of these two types of models:



Model Comparison
Data variable: Traffic
Number of observations = 168
Start index = 1/68
Sampling interval = 1.0 month(s)
Length of seasonality = 12

Models
(L) Winter's exp. smoothing with alpha = 0.4847, beta = 0.0191, gamma = 0.4423
(M) ARIMA(2,1,1)x(1,1,2)12
(N) ARIMA(1,1,2)x(1,1,2)12
(O) ARIMA(2,1,1)x(1,1,2)12 with constant
(P) ARIMA(1,1,2)x(1,1,2)12 with constant
(Q) ARIMA(2,0,1)x(1,1,2)12

Estimation Period

Model	*RMSE*	*MAE*	*MAPE*	*ME*	*MPE*	*AIC*
(L)	2.31714	1.53852	1.66723	-0.273983	-0.308439	1.71638
(M)	1.91188	1.23242	1.34833	0.0894087	0.0664659	1.3676
(N)	1.91211	1.22733	1.3435	0.0393296	0.0136061	1.36784
(O)	1.91185	1.23266	1.34838	0.0634539	0.0330278	1.37948
(P)	1.91549	1.21616	1.33134	0.0345006	0.0035009	1.38328
(Q)	1.92939	1.24165	1.35913	0.186433	0.177926	1.38584

Figure 11: Comparison of Seasonal Forecasting Models

Model (L) is Winter's procedure. The three exponential smoothing parameters, α, β, and γ, have been optimized by minimizing the mean squared forecasting error. Models (M) through (Q) represent the 5 best ARIMA models. To rank the goodness of fit of the models, Akaike's Information Criterion (AIC) has been computed. The AIC is based on the mean squared error in attempting to forecast one period beyond the end of the data, penalized by the number of parameters that need to be estimated. It is calculated from

$$AIC = 2\ln\left(RMSE\right) + \frac{2c}{n}$$

where *RMSE* is the root mean squared error during the estimation period, *c* is the number of estimated parameters in the fitted model, and *n* is the sample size. In general, the model will be selected that minimizes the mean squared error without using too many parameters (relative to the amount of data available).

According to the AIC, model (M) is the best. The forecasts from this model are shown on the *Time Sequence Plot*:

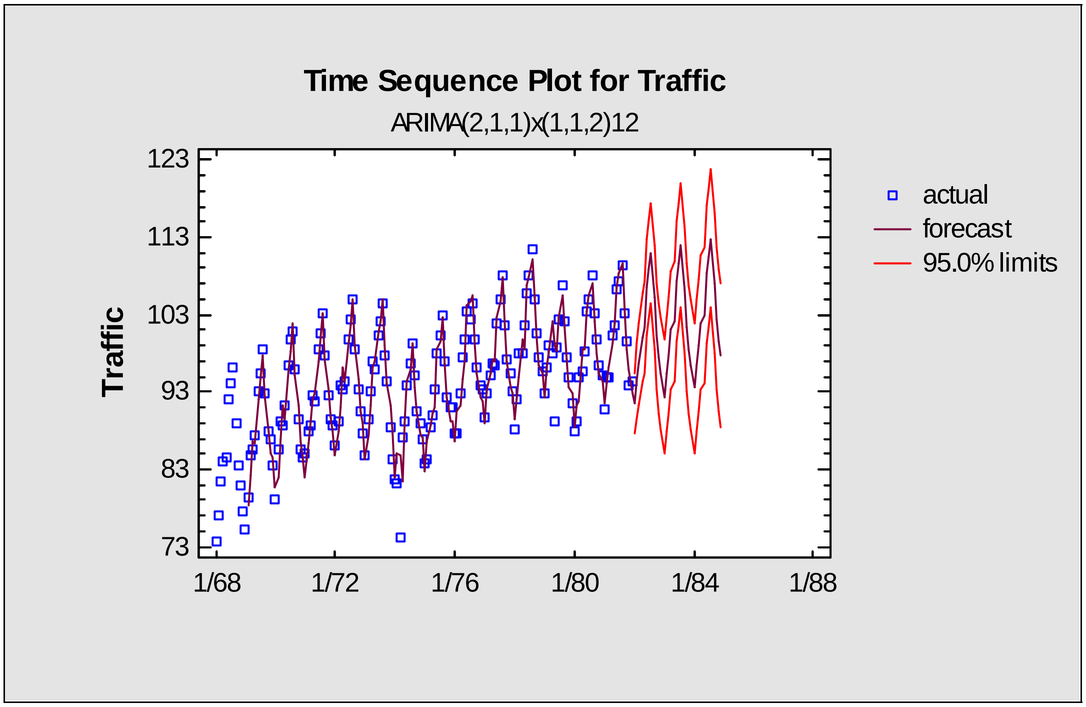
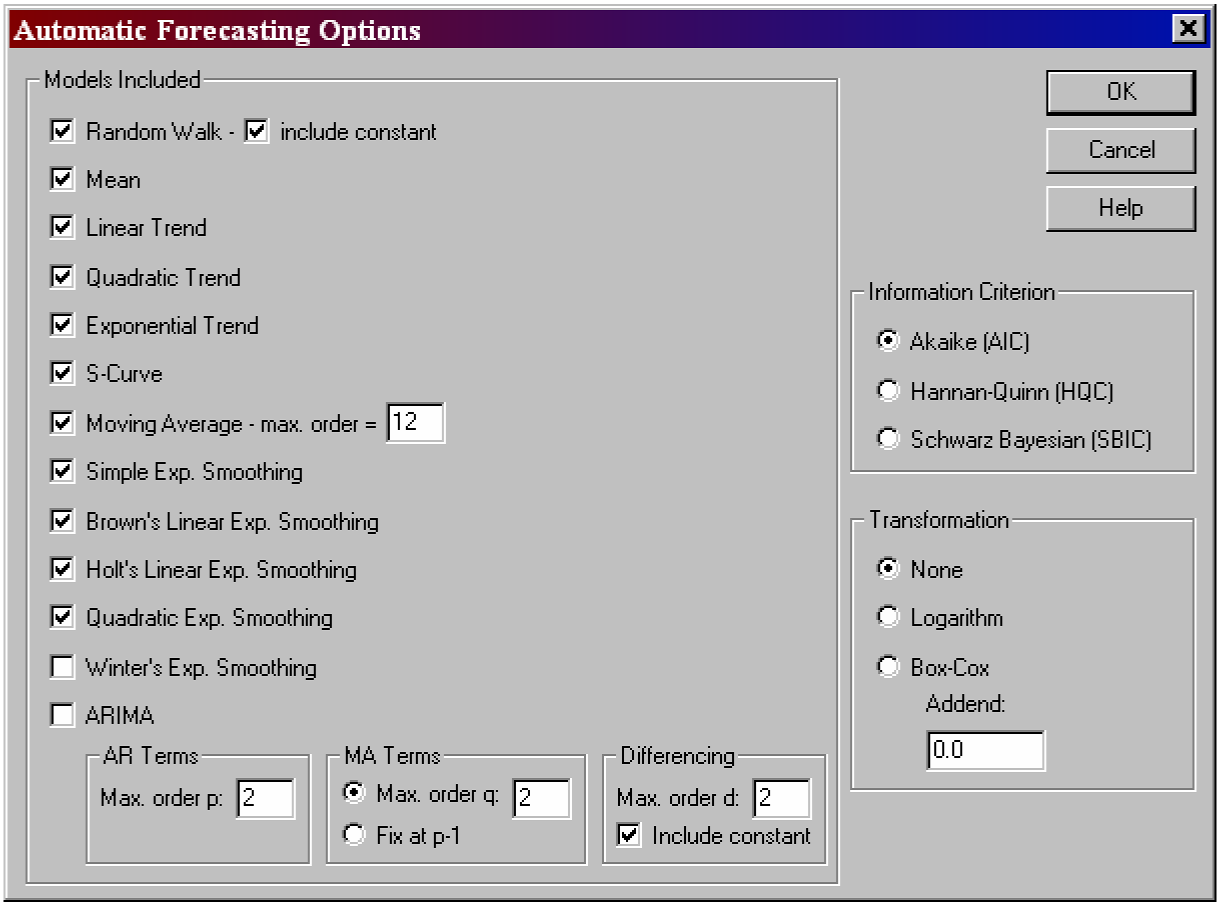


Figure 12: Forecasts from the Optimal Seasonal Forecasting Model

The solid brown line shows the point forecasts, while the red lines show the 95% forecast limits. The forecasts appear to extrapolate the recent behavior quite well. The forecast limits are wide, but this is to be expected given some of the large residuals in the historical data.

Nonseasonal Forecasting Models Applied to the Seasonally Adjusted Data

To fit nonseasonal models to the seasonally adjusted data, select *Analysis Options* again. Complete the dialog box as shown below:



Models Included

☑ Random Walk - ☑ include constant

☑ Mean

☑ Linear Trend

☑ Quadratic Trend

☑ Exponential Trend

☑ S-Curve

☑ Moving Average - max. order = 12

☑ Simple Exp. Smoothing

☑ Brown's Linear Exp. Smoothing

☑ Holt's Linear Exp. Smoothing

☑ Quadratic Exp. Smoothing

□ Winter's Exp. Smoothing

□ ARIMA

AR Terms
Max. order p: 2

MA Terms
● Max. order q: 2
○ Fix at p-1

Differencing
Max. order d: 2
☑ Include constant

Information Criterion
● Akaike (AIC)
○ Hannan-Quinn (HQC)
○ Schwarz Bayesian (SBIC)

Transformation
● None
○ Logarithm
○ Box-Cox
Addend: 0.0

Figure 13: Analysis Options Dialog Box with Selection of Nonseasonal Models

When the *Automatic Forecasting* procedure fits a nonseasonal model to data for which seasonality was specified on the data input dialog box, it:

1. First seasonally adjusts the data as illustrated for the *Seasonal Decomposition* procedure.

2. Applies the forecasting methods to the seasonally adjusted data and generates forecasts for it.

3. Multiplies the forecasts of the seasonally adjusted data by the seasonal indices to create forecasts for the original data.

The results are shown below:

Model Comparison
Data variable: Traffic
Number of observations = 168
Start index = 1/68
Sampling interval = 1.0 month(s)
Length of seasonality = 12

Models
(A) Random walk
(B) Constant mean = 93.9718
(C) Linear trend = 68.8458 + 0.0836141 t
(D) Quadratic trend = 40.8552 + 0.274889 t + -0.000318261 t^2
(E) Exponential trend = exp(4.27143 + 0.0008988 t)
(F) S-curve trend = exp(4.80816 + -77.9949 /t)
(G) Simple moving average of 2 terms
(H) Simple exponential smoothing with alpha = 0.7443
(I) Brown's linear exp. smoothing with alpha = 0.3269
(J) Holt's linear exp. smoothing with alpha = 0.7438 and beta = 0.0146
(K) Brown's quadratic exp. smoothing with alpha = 0.1923

Estimation Period

Model	*RMSE*	*MAE*	*MAPE*	*ME*	*MPE*	*AIC*
(A)	2.07408	1.31347	1.43666	-0.00214272	-0.0210771	1.58662
(B)	5.15304	4.02891	4.39876	0.00647326	-0.300213	3.42203
(C)	3.07679	2.32241	2.53535	0.00156945	-0.109794	2.40253
(D)	3.01293	2.19924	2.40124	0.00139851	-0.103337	2.37249
(E)	3.09741	2.35088	2.56383	0.0512234	-0.0557736	2.4159
(F)	2.99392	2.21169	2.4117	0.0479026	-0.052079	2.34793
(G)	2.72468	1.81704	1.9799	0.544652	0.530305	1.68104
(H)	2.02123	1.3171	1.44231	0.143187	0.130268	1.55027
(I)	2.22817	1.46057	1.59461	0.0280951	0.0131666	1.74521
(J)	2.04348	1.30953	1.43489	-0.167097	-0.204924	1.58407
(K)	2.35599	1.56361	1.70557	0.0039615	-0.015076	1.85678

Figure 14: Comparison of Nonseasonal Forecasting Models Applied to the Seasonally Adjusted Data

Simple Exponential Smoothing gives the lowest value of the AIC, resulting in the following forecasts:

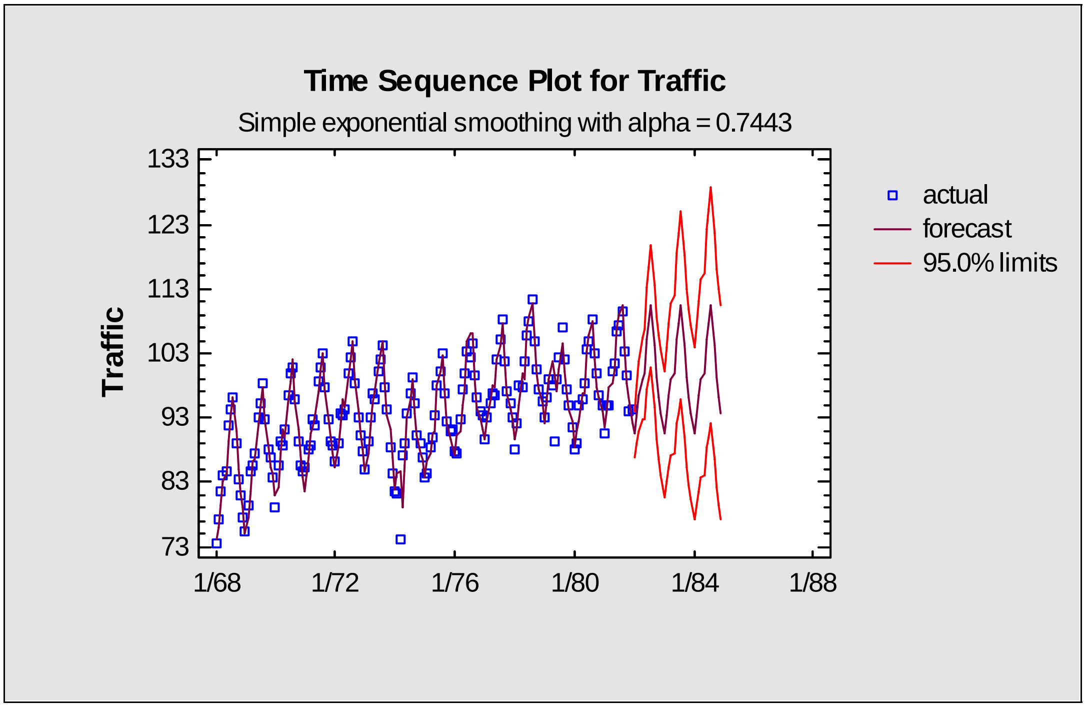


Figure 15: Forecasts from Simple Exponential Smoothing of the Seasonally Adjusted Data

While the *Simple Exponential Smoothing* forecasts pick up the seasonality well, they have no trend component. Note also that the AIC is well above that of the best seasonal ARIMA model.

Conclusion

STATGRAPHICS contains several procedures for analyzing time series data. The *Descriptive Methods* procedure is useful for plotting the data, and also for identifying seasonal effects through the autocorrelation function and periodogram. The *Seasonal Decomposition* procedure splits the time series into trend-cycle, seasonal, and irregular components. The *Automatic Forecasting* procedure fits many different forecasting models and selects the method that optimizes a specified information criterion.

The general approach described in this guide should work well on many time series. In some cases, you may need to transform the data using a square root or logarithm before fitting the forecasting models. As in all data analysis efforts, it is important to plot the data at each stage to be sure that the results make sense in the context of the application for which the results will be used.

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