Automatic Forecasting

Summary

The **Automatic Forecasting** procedure is designed to forecast future values of time series data. A *time series* consists of a set of sequential numeric data taken at equally spaced intervals, usually over a period of time or space. Unlike the *Forecasting* procedure that expects the user to select the forecasting model to use, this procedure tries many models and selects the one which performs best according to a specified criteria. The available criteria for selecting a model include the Akaike Information Criteria (AIC), the Hannan-Quinn Criterion (HQC), and the Schwarz-Bayesian Criteria (SBC). This criteria select the model with smallest mean squared error, subject to a penalty for the number of unknown parameters that need to be estimated.

Since the output of this procedure is similar to the *Forecasting* procedure, this document will highlight only the unique aspects of the *Automatic Forecasting* procedure. For a detailed discussion of all tables and graphs, refer to the *Forecasting* documentation.

Sample StatFolio: *autocast.sgp*

Sample Data:

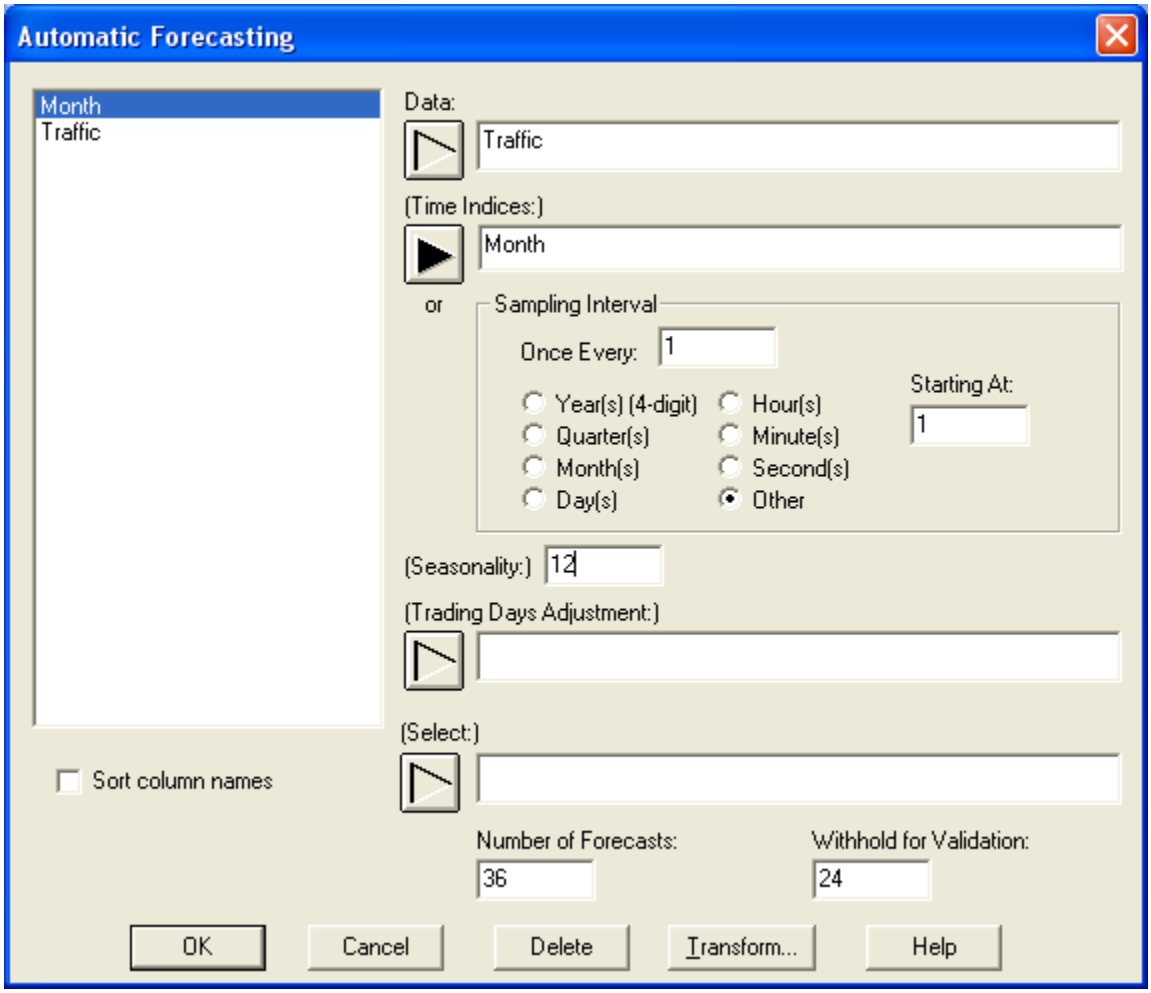
The file *golden gate.sgd* contains monthly traffic volumes on the Golden Gate Bridge in San Francisco for a period of $n = 168$ months from January, 1968 through December, 1981. The table below shows a partial list of the data from that file:

Month	*Traffic*
1/68	73.637
2/68	77.136
3/68	81.481
4/68	84.127
5/68	84.562
6/68	91.959
7/68	94.174
8/68	96.087
9/68	88.952
10/68	83.479
11/68	80.814
12/68	77.466
1/69	75.225
…	…

The data were obtained from a publication of the Golden Gate Bridge.

Data Input

The data input dialog box requests the name of the column containing the time series data:



- **Data:** numeric column containing *n* equally spaced numeric observations.

- **Time indices:** time, date or other index associated with each observation. Each value in this column must be unique and arranged in ascending order.

- **Sampling Interval:** If time indices are not provided, this defines the interval between successive observations. For example, the data from the Golden Gate Bridge were collected once every *month*, beginning in January, 1968.

- **Seasonality**: the length of seasonality *s*, if any. The data is seasonal if there is a pattern that repeats at a fixed period. For example, monthly data such as traffic on the Golden Gate Bridge have a seasonality of $s = 12$. Hourly data that repeat every day have a seasonality of $s = 24$. If no entry is made, the data is assumed to be nonseasonal ($s = 1$).

- **Trading Days Adjustment**: a numeric variable with *n* observations used to normalize the original observations, such as the number of working days in a month. The observations in the *Data* column will be divided by these values before being plotted or analyzed. There must be enough entries in this column to cover both the observed data and the number of periods for which forecasts are requested.

- **Select:** subset selection.

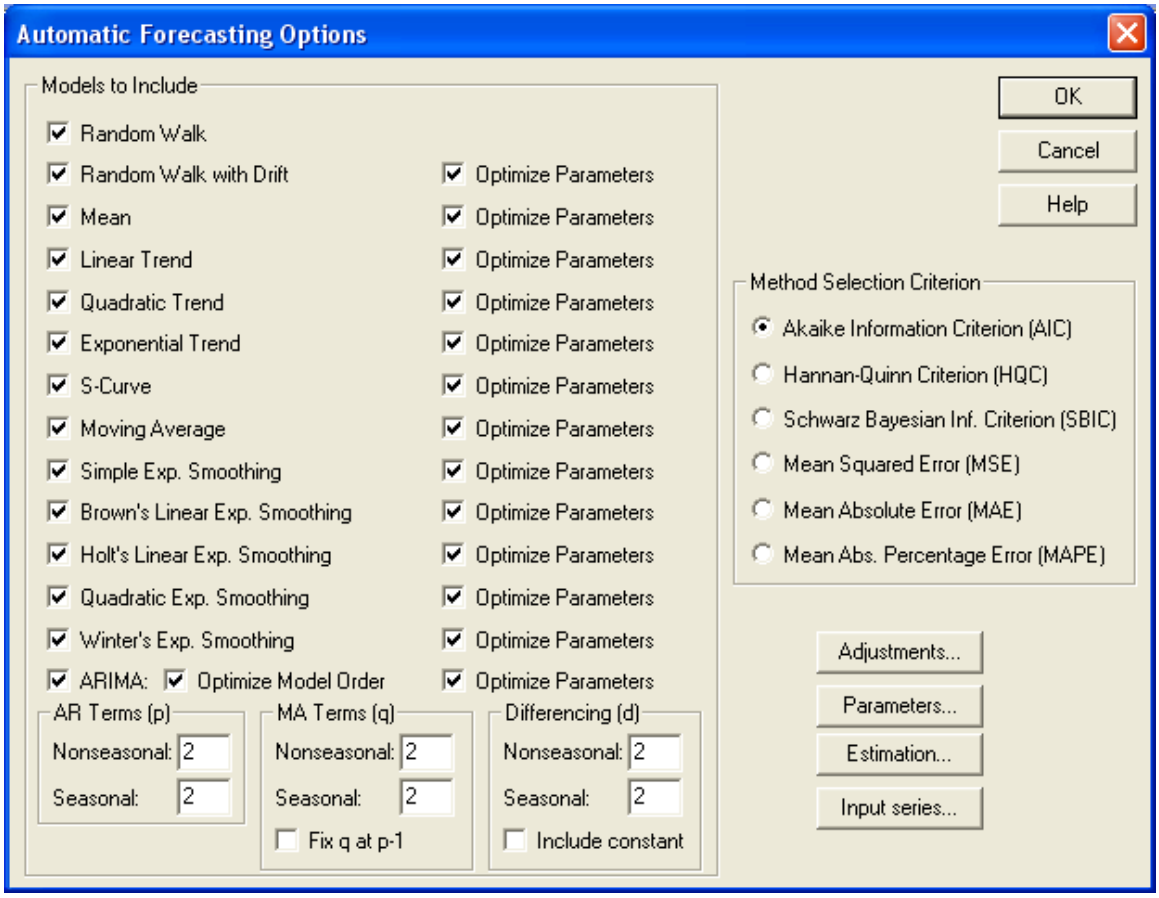
- **Number of Forecasts**: number of periods following the end of the data for which forecasts are desired.

- **Withhold for Validation**: number of periods *m* at the end of the series to withhold for validation purposes. The data in those periods will not be used to estimate the forecasting model. However, statistics will be calculated describing how well the estimated model is able to forecast those observations.

In the current example, the traffic data is monthly beginning in January, 1968, and has a seasonality of $s = 12$. $m = 24$ observations at the end of the series will be withheld for validation purpose, while forecasts will be generated for the next 36 months.

Analysis Options

The models fit by the *Automatic Forecasting* procedure are controlled by the *Analysis Options* dialog box:



Models to Include: specify the models that should be fit to the data. These are the models from which the "best" model will be selected. Descriptions of each of the models are given in the *Forecasting* documentation. For most of the models, additional information must are provided:

> *Optimize parameters*: If checked, unknown parameters in the model will be estimated so as to optimize the specified forecasting criterion. If not checked, specific values for the parameters may be entered by pressing the *Parameters* button.
>
> *ARIMA model: Optimize Model Order* – If checked, all models with terms of order up to those specified will be fit. If not checked, the only model fit will be the one with terms exactly equal to the specified order.
>
> *AR Terms (p)* – specify the maximum order *p* of the autoregressive terms in the ARIMA model.

MA Terms(q) – specify the maximum order *q* of the moving average terms in the ARIMA model. You may also elect to consider only models for which $q = p - 1$.

Differencing (d) – specify the maximum order of differencing *d* in the ARIMA model. Select *Include constant* to consider models that include a constant term when differencing is performed.

- **Method Selection Criterion**: the criterion used to select the best model.

The procedure fits each of the models indicated and selects the model that gives the smallest value of the selected criterion. They are six criteria to choose from:

Akaike Information Criterion
The Akaike Information Criterion (AIC) is calculated from

$$AIC = 2\ln(RMSE) + \frac{2c}{n} \qquad (1)$$

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

Hannan-Quinn Criterion
The Hannan Quinn Criterion (HQC) is calculated from

$$HQC = 2\ln(RMSE) + \frac{2p\ln(\ln(n))}{n} \qquad (2)$$

This criterion uses a different penalty for the number of estimated parameters.

Schwarz-Bayesian Information Criterion
The Schwarz-Bayesian Information Criterion (SBIC) is calculated from

$$SBIC = 2\ln(RMSE) + \frac{p\ln(n)}{n} \qquad (3)$$

Again, the penalty for the number of estimated parameters is different than for the other criteria.

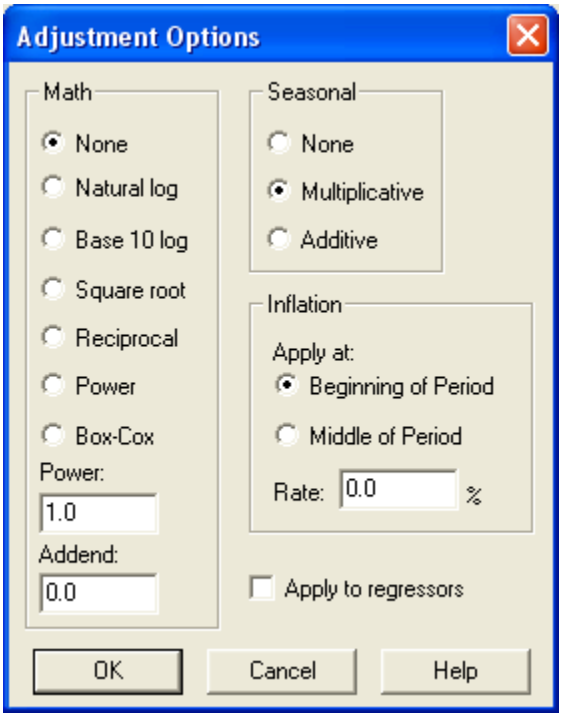
Mean Squared Error (MSE)
The selected model is the one with the smallest root mean squared error RMSE, with no penalty for the number of estimated model parameters.

Mean Absolute Error (MAE)
The selected model is the one with the smallest mean absolute error.

Mean Absolute Percentage Error (MAPE)
The selected model is the one with the smallest mean absolute percentage error.

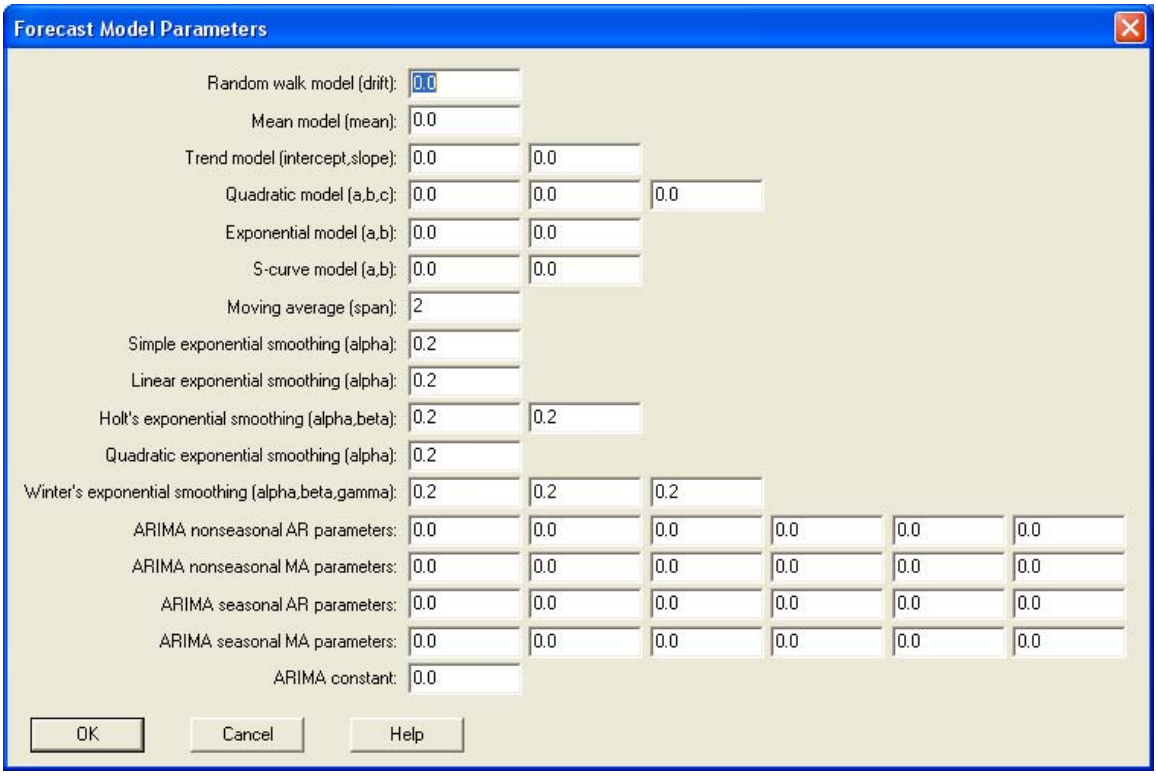
There are also four buttons that provide additional options:

- **Adjustments**: Press this button to specify adjustments to be made to the data before the forecasting models are fit:



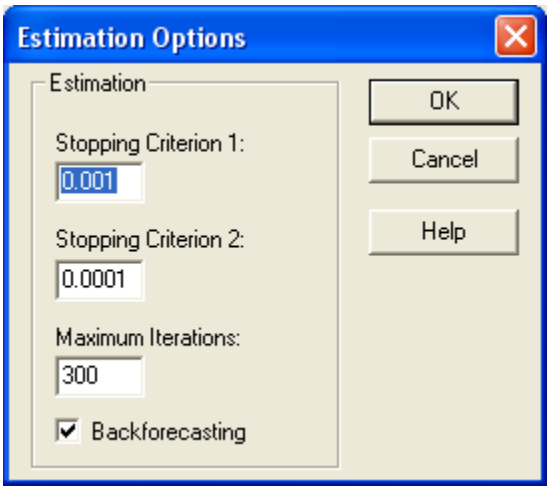
  After the models are fit and forecasts are made of the adjusted data values, the adjustments are reversed to provide the final forecasts. If *Apply to regressors* is checked, the same adjustments will be made to any regressor variables in the models.

- **Parameters**: Press this button to enter values for each of the model parameters:



The entries in this dialog box are only used for models in which the *Optimize Parameters* button is not checked.

- **Estimation**: Press this button to change the default values for certain estimation options:



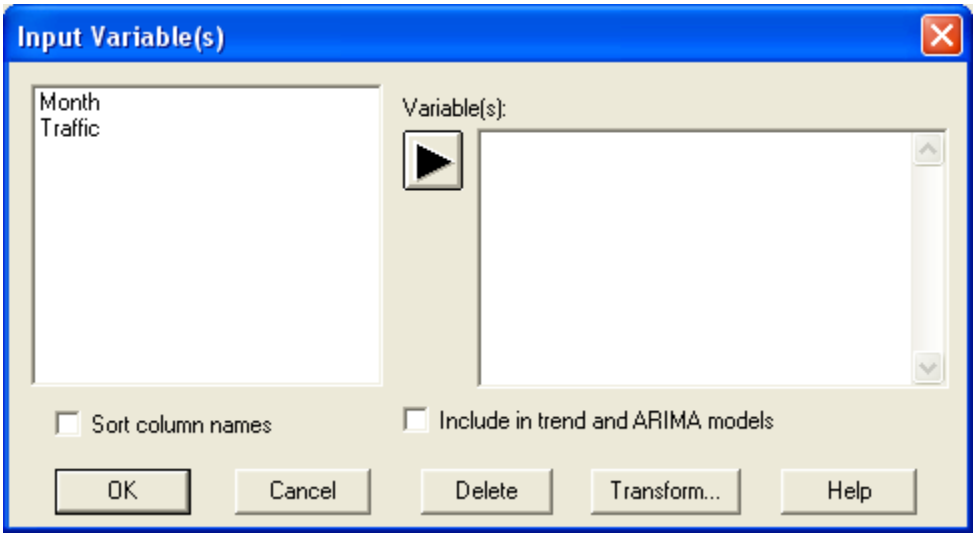
Stopping Criterion 1: The algorithm is assumed to have converged when the relative change in the residuals sums of squares from one iteration to the next is less than this value.

Stopping Criterion 2: The algorithm is assumed to have converged when the relative change in all parameter estimates from one iteration to the next is less than this value.

Maximum Iterations: Estimation stops if convergence is not achieved within this many iterations.

Backforecasting: Uses a method called *backforecasting* to forecast values prior to time $t = 1$. These values are used to generate the initial values which are needed to generate forecasts for small values of t. For details, see Box, Jenkins and Reinsel (1994).

- **Input Series**: Press this button to enter one or more input variables to act as regressors in the trend and ARIMA forecasting models:



Analysis Summary

The *Analysis Summary* gives the standard forecasting output for whichever model gives the smallest value of the specified criterion.

Automatic Forecasting - Traffic
Data variable: Traffic

Number of observations = 168
Time indices: Month
Length of seasonality = 12

Forecast Summary
Forecast model selected: ARIMA(0,1,2)x(2,1,2)12
Number of forecasts generated: 36
Number of periods withheld for validation: 24

	Estimation	*Validation*
Statistic	*Period*	*Period*
RMSE	2.0522	1.55727
MAE	1.35877	1.16503
MAPE	1.49719	1.18882
ME	-0.0402442	0.0221547
MPE	-0.0732127	0.00847592

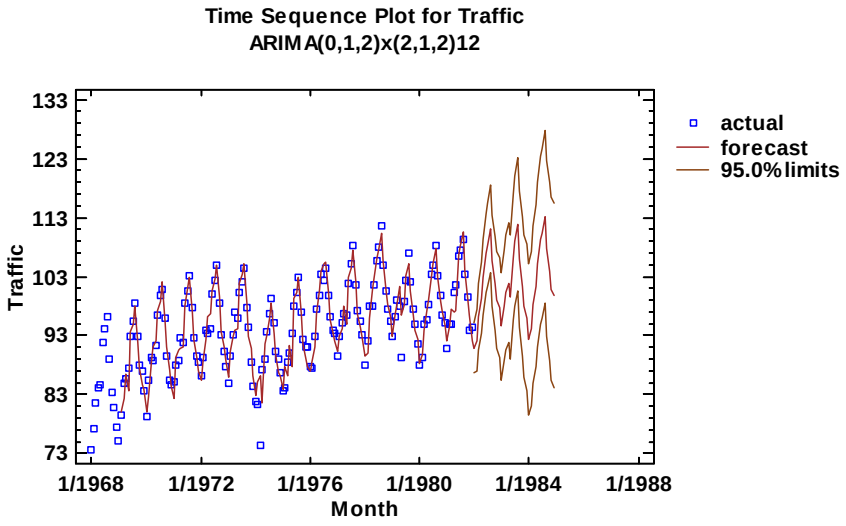
ARIMA Model Summary

Parameter	*Estimate*	*Stnd. Error*	*t*	*P-value*
MA(1)	0.242799	0.0877347	2.76742	0.006367
MA(2)	0.238687	0.0869116	2.74632	0.006770
SAR(1)	-1.04978	0.13349	-7.86407	0.000000
SAR(2)	-0.0556194	0.119539	-0.465283	0.642408
SMA(1)	-0.243651	0.0819685	-2.97249	0.003446
SMA(2)	0.899649	0.0464212	19.3801	0.000000

Backforecasting: yes
Estimated white noise variance = 4.37813 with 149 degrees of freedom
Estimated white noise standard deviation = 2.0924
Number of iterations: 17

When searching for models, the procedure tries all of the models checked on the *Analysis Options* dialog box. Note: except for *Winter's Exponential Smoothing* and the *ARIMA Models*, seasonal data will first be seasonally adjusted before the forecasts are applied. Once the forecasts have been generated, the seasonality will be put back to create the final forecasts.

For the Golden Gate Bridge data, the procedure selected an ARIMA(0,1,2)x(2,1,2)12 model. Note that all of the coefficients in the model are statistically significant. As can be seen from the plot of the fit and forecasts, the results are quite satisfactory:



Model Comparisons

The *Model Comparisons* pane displays information on the best-fitting models of each type requested. The top section summarizes the data and lists the fitted models:

Model Comparison
Data variable: Traffic
Number of observations = 168
Length of seasonality = 12
Number of periods withheld for validation: 24

Models
(A) Random walk with drift = 0.112976
(B) Constant mean = 93.1497
(C) Linear trend = 86.4665 + 0.0921817 t
(D) Quadratic trend = 85.384 + 0.136667 t + -0.000306797 t^2
(E) Exponential trend = exp(4.46059 + 0.000995511 t)
(F) S-curve trend = exp(4.54305 + -0.266804 /t)
(G) Simple moving average of 2 terms
(H) Simple exponential smoothing with alpha = 0.7784
(I) Brown's linear exp. smoothing with alpha = 0.3239
(J) Holt's linear exp. smoothing with alpha = 0.6519 and beta = 0.0136
(K) Brown's quadratic exp. smoothing with alpha = 0.2008
(L) ARIMA(0,1,2)x(2,1,2)12
(M) ARIMA(1,1,2)x(1,1,2)12
(N) ARIMA(2,1,1)x(1,1,2)12
(O) ARIMA(1,1,2)x(2,1,2)12
(P) ARIMA(1,0,2)x(2,1,2)12

The five ARIMA models in the list are those that fit best, among dozens that were fit.

The next section summarizes the performance of each model during the estimation period:

Estimation Period

Model	*RMSE*	*MAE*	*MAPE*	*ME*	*MPE*	*AIC*
(A)	2.15283	1.3111	1.44863	-0.00127742	-0.0222611	1.70023
(B)	5.07826	3.86041	4.2478	0.00872822	-0.291876	3.4166
(C)	3.24311	2.41585	2.65599	0.00332193	-0.120246	2.53362
(D)	3.21721	2.35022	2.58029	0.00315715	-0.116598	2.53148
(E)	3.25638	2.43744	2.67806	0.0578209	-0.0612388	2.54179
(F)	4.426	3.34494	3.66479	0.0953224	-0.111391	3.15555
(G)	2.27062	1.42557	1.57284	0.165409	0.148573	1.79288
(H)	2.11212	1.32564	1.46535	0.141871	0.127809	1.66205
(I)	2.34656	1.5048	1.65656	0.0433801	0.0267387	1.87256
(J)	2.15267	1.34019	1.48466	-0.248543	-0.299132	1.71397
(K)	2.48951	1.62759	1.78848	0.0382572	0.0175499	1.99084
(L)	2.0522	1.35877	1.49719	-0.0402442	-0.0732127	1.52116
(M)	2.06766	1.28876	1.42164	-0.0258518	-0.0589859	1.53617
(N)	2.06832	1.28222	1.41562	-0.0226957	-0.0556233	1.53681
(O)	2.0578	1.33689	1.47617	-0.0434688	-0.0761309	1.5405
(P)	2.058	1.35559	1.49442	-0.0109378	-0.0425823	1.5407

The column on the far right shows the value of the selected criterion for each of the models. In the sample data, the ARIMA(2,0,1)x(2,1,1)12 model (Model M) does best, though several other ARIMA models are very close.

The output also shows how well each model did during the validation period in forecasting the observations that were withheld from the estimation process:

Validation Period

Model	*RMSE*	*MAE*	*MAPE*	*ME*	*MPE*
(A)	2.72132	1.42365	1.46028	0.00416832	-0.00391219
(B)	35.7947	5.74478	5.79108	5.74478	5.79108
(C)	6.17679	2.06352	2.09452	-2.01854	-2.04853
(D)	2.49534	1.38226	1.41552	-0.368482	-0.386012
(E)	7.17915	2.2584	2.29013	-2.25692	-2.28857
(F)	2.80136	1.41307	1.4416	-0.748106	-0.767949
(G)	2.18546	1.30179	1.32752	0.136708	0.130939
(H)	2.56813	1.36359	1.40201	0.542393	0.537291
(I)	2.46638	1.39127	1.4176	0.374832	0.376192
(J)	4.15101	1.66548	1.70837	1.02078	1.01629
(K)	2.90303	1.44906	1.4631	0.424273	0.437374
(L)	4.15181	1.73076	1.76437	0.987824	0.953436
(M)	2.4251	1.16503	1.18882	0.0221547	0.00847592
(N)	2.09206	1.18016	1.22359	-0.3114	-0.330986
(O)	2.62911	1.13357	1.15323	0.301173	0.292195
(P)	2.0961	1.18316	1.22549	-0.0621341	-0.0776316
(Q)	2.26482	1.25482	1.29717	-0.260738	-0.278414

The selected ARIMA model performed well, especially on the MAPE at approximately 1.2%, although it was beaten by a couple of the other ARIMA models with respect to the RMSE.

