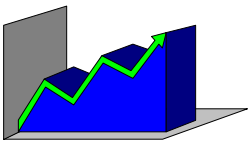
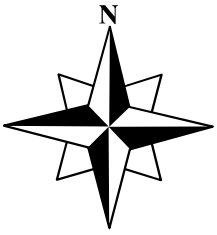
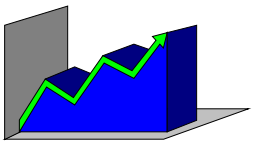


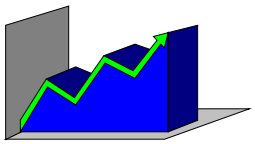
# User's Guide
# Autobox – Interactive Version



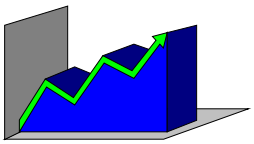

This manual is published by

Automatic Forecasting Systems©, Inc.





# Table of Contents



# Introduction

Thank you for selecting Autobox. If you are having any questions or issues call us at 215-675-0652.

Autobox must be installed as the "administrator" on Vista machines. You download Autobox and save it to disk then right click on the downloaded and choose "run as administrator" to install. You then right click on the icon to run as administrator whenever you go to run Autobox.

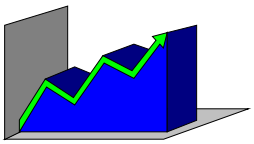
Autobox can be used with any "time series" data in 5 broad ways:

- Data Cleansing (i.e. Identify outliers and correct the data for it's errors)
- Modeling past behavior (i.e. Did the Promotion Coupon work?)
- Forecasting (i.e. Extrapolate a series of numbers into the future)
- Exception Reporting (i.e. Which series are out of control? What time period has the most outliers across my different SKUs?)
- Simulation/Scenario Analysis (i.e. What would happen if I lowered the price down to $xx?)

Autobox uses automatic modeling heuristics (not pick best) with intervention detection. It tailors the forecast model to the problem at hand including selecting the best lead and lag structures for each input series. It corrects for omitted variables (e.g., holidays or price changes that have affected the historical data that the system has no knowledge of) by identifying pulses, seasonal pulses, level shifts and local time trends, and then adding the needed structure through surrogate variables.

Autobox has a set of graphing tools that help present complex statistical information in a way that is easy and clear at every stage of the forecasting process and even a simple narrative description of the final model. Graphs of autocorrelation, partial-autocorrelation and cross-correlation functions are all available.

We include over 700 example data sets taken from textbooks with the installation (look in the folders in the installation directory). This includes the classic "Airline Passenger series" which Box-Jenkins studied and is the most studied time series. It is located in the "Box-j" folder.



When you go to model, you can include causal variables, retain future observations for error analysis, provide future values of the causal variables or tweak the modeling process that Autobox uses.

Autobox will **automatically** aid the modeling process for weekly, daily, hourly and semi-hourly data. If you have weekly, daily or hourly data, Autobox will add 51 dummy variables for the different weeks of the year. You need at least 1 1/2 year of historical data for this to happen. If you have daily, hourly or semi-hourly data Autobox will add 6 dummy variables for day of the week. If you have hourly data Autobox will add 23 dummy variables for hour of the day. If you have semi-hourly data Autobox will add 47 dummy variables for each half-hour of the day.

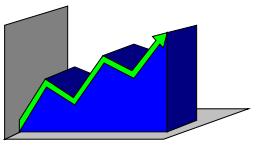
For daily data that covers all 7 days(Monday to Sunday), Autobox will enact 4 different modeling approaches. You trigger Autobox to do this by providing a series name like this "__040106Y11". To tell Autobox to look for these daily effects, just add two "_" before the date and the name of the series where 040106 represents April 1, 2006 and the series(SKU) name is "Y11", for example. If a holiday lands on a weekend, Autobox will look for a "Friday before" and "Monday after" effect automatically. Autobox will look for a day of the month effect. Autobox will add in U.S. holidays. Note that you can always create variables like this yourself and add them in as a causal variable.

Our competitors will NOT let you pre-evaluate their software. They often will only let you see a slideshow. If they do let you "see" their software you will have to buy it with a 30 day refund policy. We would rather let you try before buying.

Note: If you have data that is very different in scale, we **strongly** recommend scaling your data(by dividing or multiplying) when you have small values and large values. If your Y is 10,000,000 and your causal is .075 then you should scale. You should keep a gap of 6 digits or smaller between the size of the variables (ie 1,000 in sales and causal variable .07 is ok). This is not a "quirk" of Autobox, but rather a common issue for everyone trying to estimate.

Contact us for any questions:

Afs Inc.
P.o. Box 563
Hatboro PA 19040
sales@autobox.com
Phone 215-675-0652 Fax 215-672-2534



# About AFS

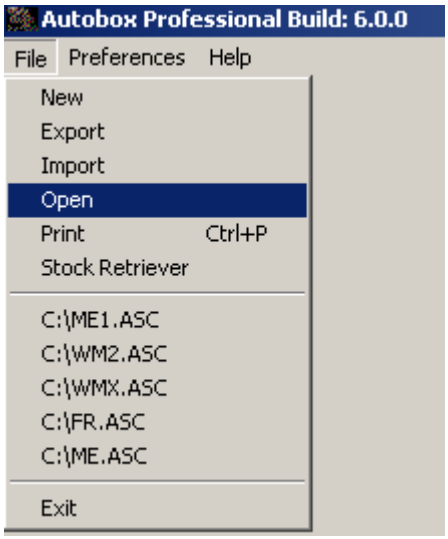
AFS has been in business since 1976 and has many business and academic customers.

AFS was given the award in the textbook “Principles of Forecasting” as the 'BEST DEDICATED FORECASTING PACKAGE".

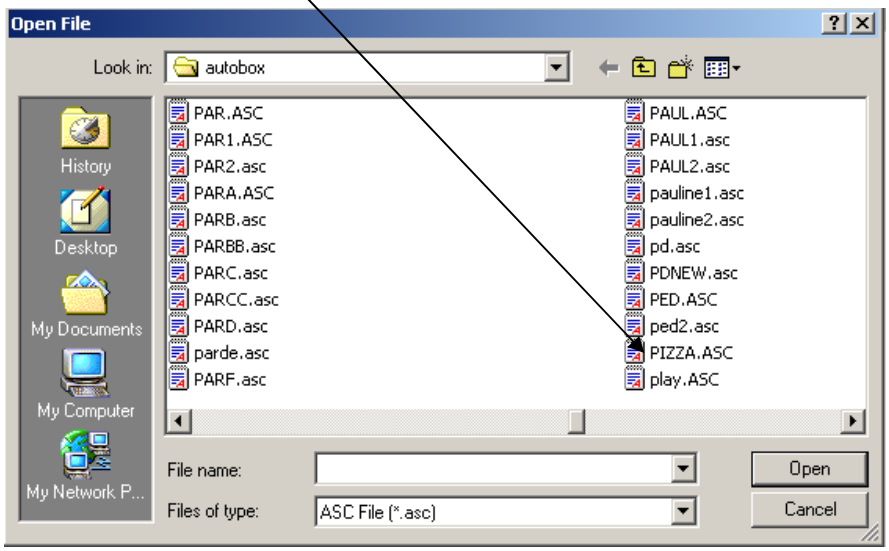
Autobox stands **alone** at the top of **Automated** Forecasting Software in the "Daily Data" 2008 International Society of Forecasters Forecasting Competition **(NN5)**

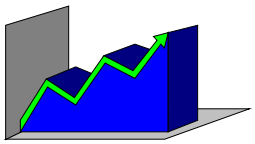
# A Quick Tour Through Autobox

Let's bring in some data that already exists. Click on "File/Open" and choose the file "Pizza.asc"

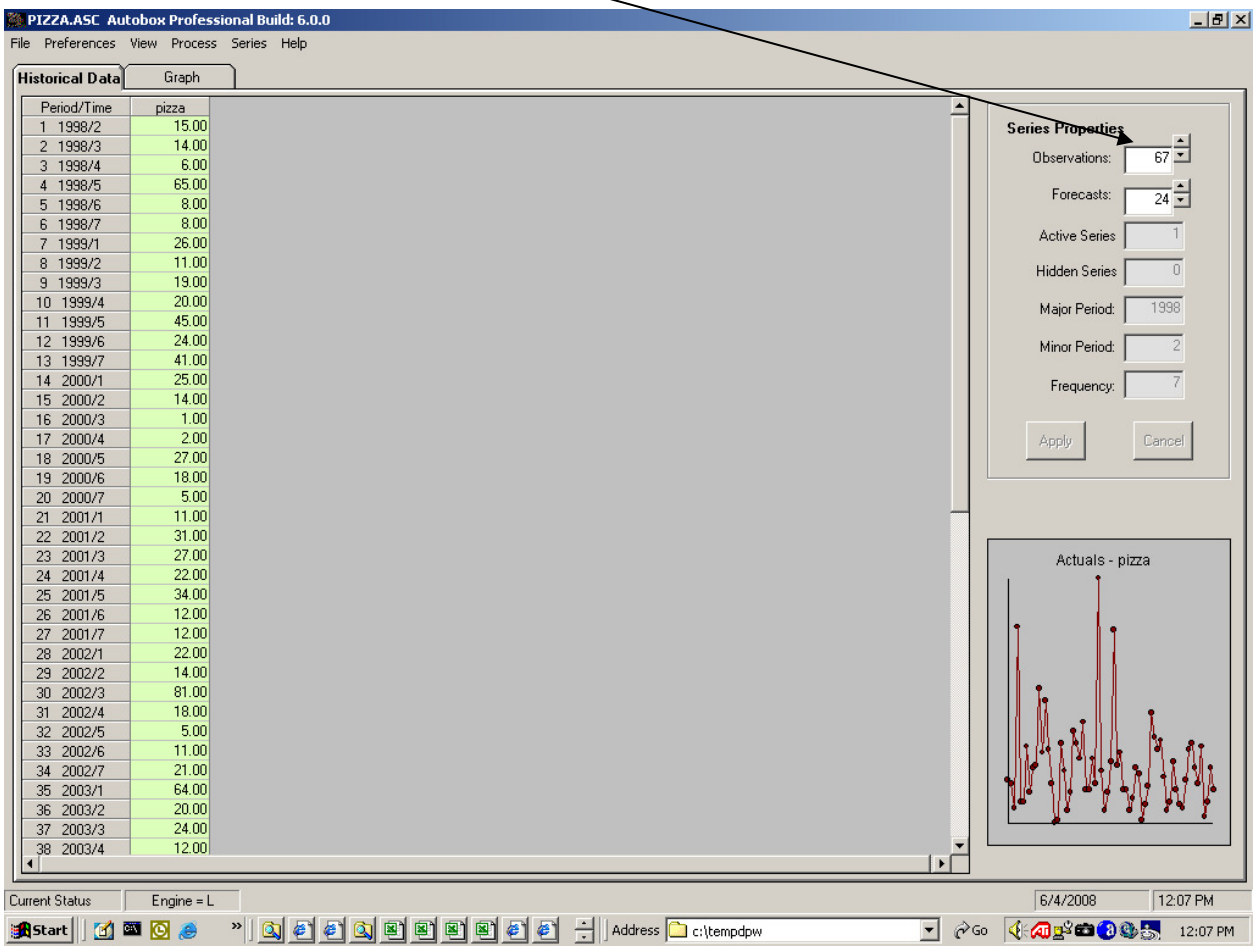


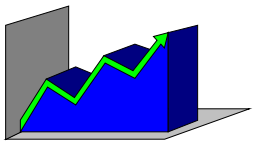

Double click on pizza.asc





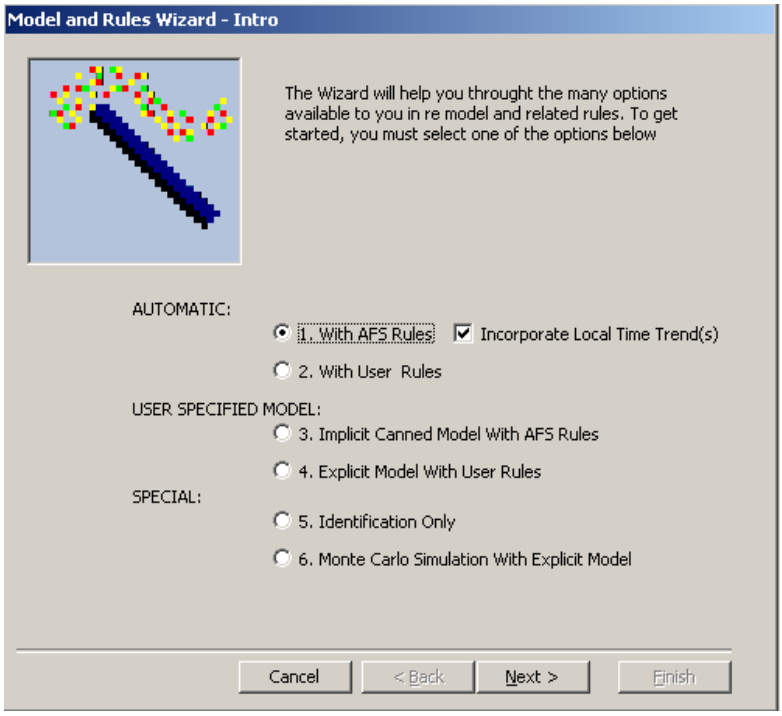
You can change the length of the forecast or the number of observations by editing the box on the right then click on “Apply”.

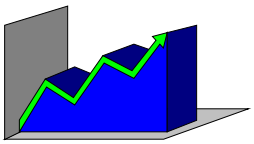




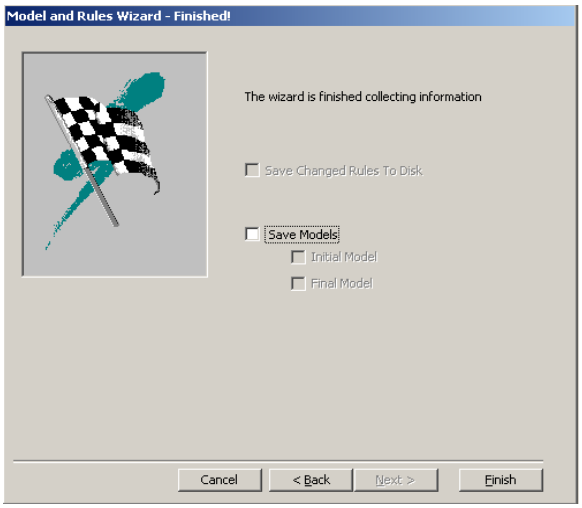
To run, choose “Process/Run Autobox. We are going to run using option “1” and click “Next” to run using the “AFS Rules” or otherwise known as the “Expert system”. Option ‘1’ has **exclusive** conditions that are not disclosed.

We will walk through options ‘2’ through ‘6’ later in the “Models and Rules Wizard” section.

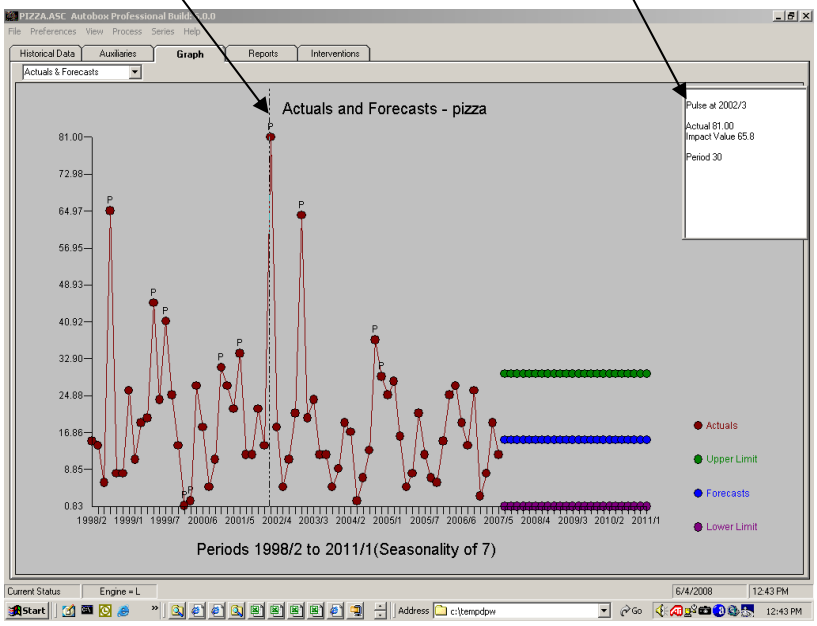


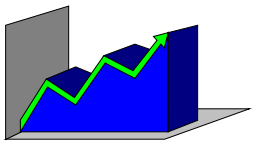


Click “Finish” to run. Notice that you “saved changed rules to disk” is greyed out as no adjustments were made. You can save the final model, but not the initial model as there were no initial model specified which is reserved for option ‘4’ where the user can specify the model.

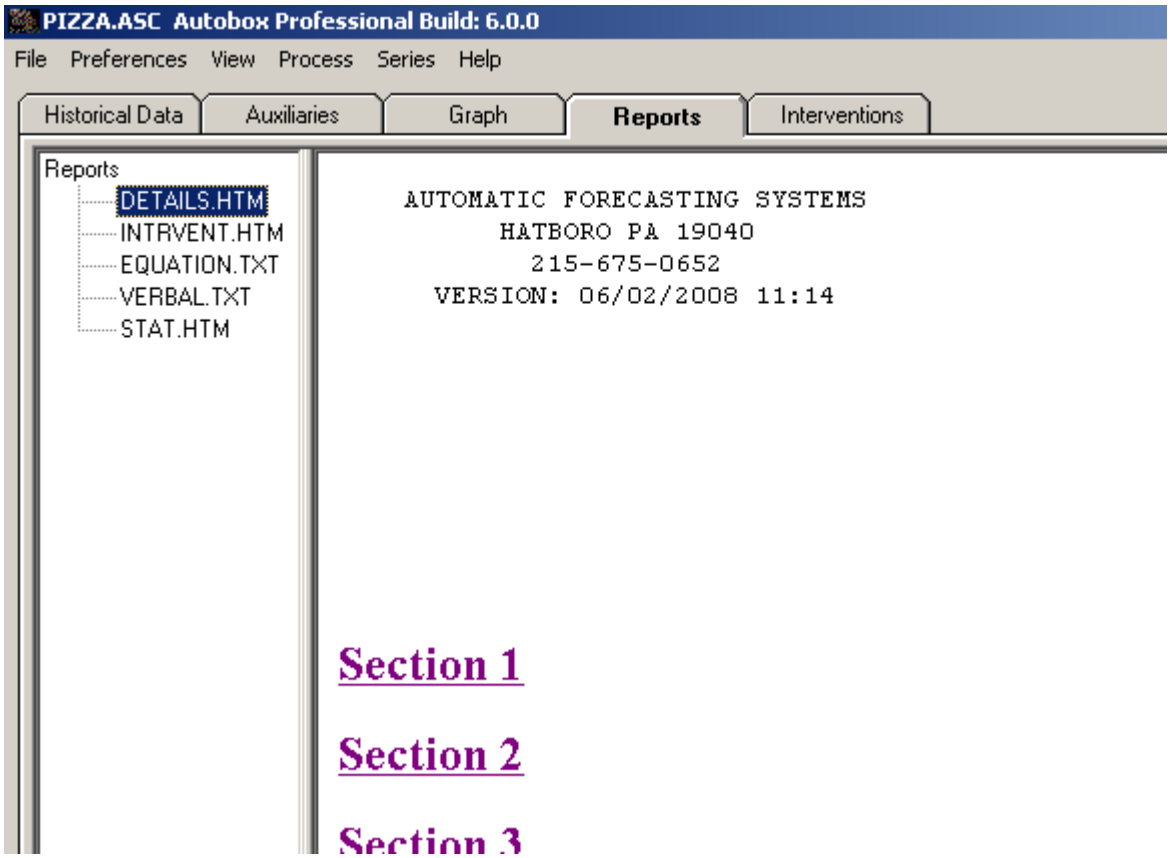


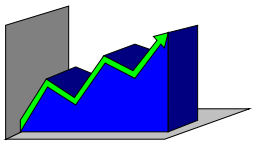

Upon completion, the “Actuals and Forecasts” graph will automatically appear. Note that as the vertical bar passes over a data point information is displayed. When an outlier is indicated, the text box at the top right describes the type of outlier and magnitude of its effect.



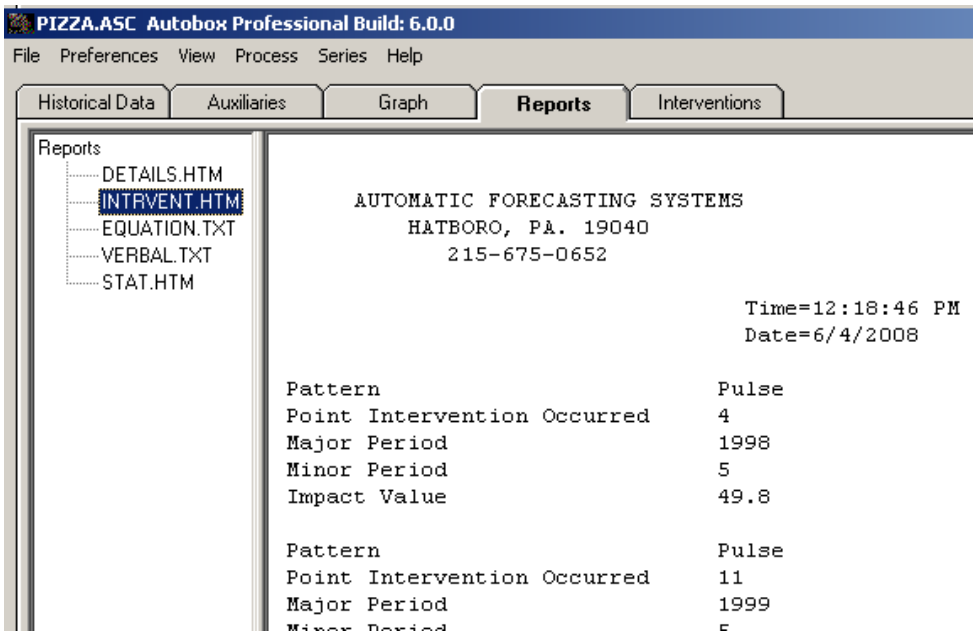


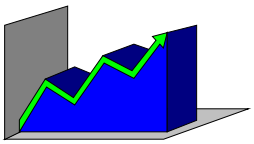
Click on “REPORTS” and you will see a list of reports to choose from on the left hand side of the screen.



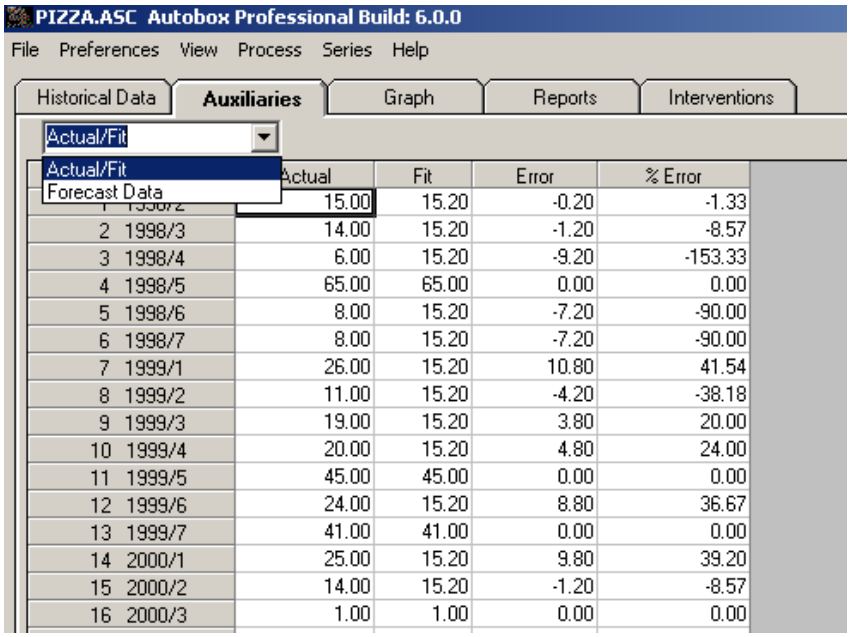


5)There is a report generated that shows how the model was built and the forecast generated (DETAILS.HTM), a report showing(shown here)at what periods interventions were identified (INTRVENT.HTM), a report that shows the equation used to forecast(EQUATION.TXT), a report that provides of an executive summary of what was identified during the modeling process(VERBAL.HTM) plus breakdown of what each numerical part of the model contributed to the forecasts to get a full picture of the dynamics, a report showing the summary statistics (STAT.HTM), a report showing the equation as a pure regression equation; and so on. For example, if you click on "INTRVENT.HTM" you will see a list of all the interventions with the type of intervention (i.e. P – Pulse, S – Seasonal Pulse, L –Level), the date they took place and their magnitude.



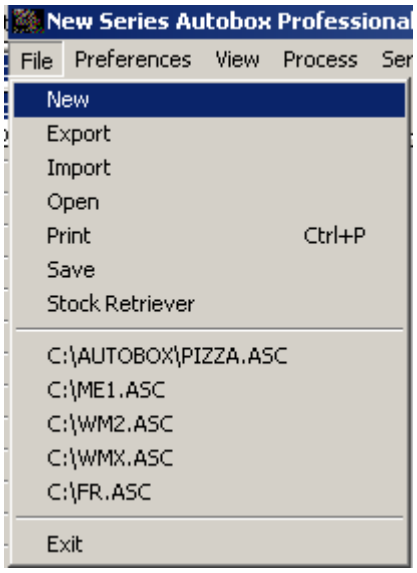


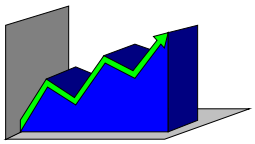
6)If you choose the “Auxiliaries” TAB you will have the option of viewing the forecast, actuals and fit in a spreadsheets. Here we show the forecast and fit.



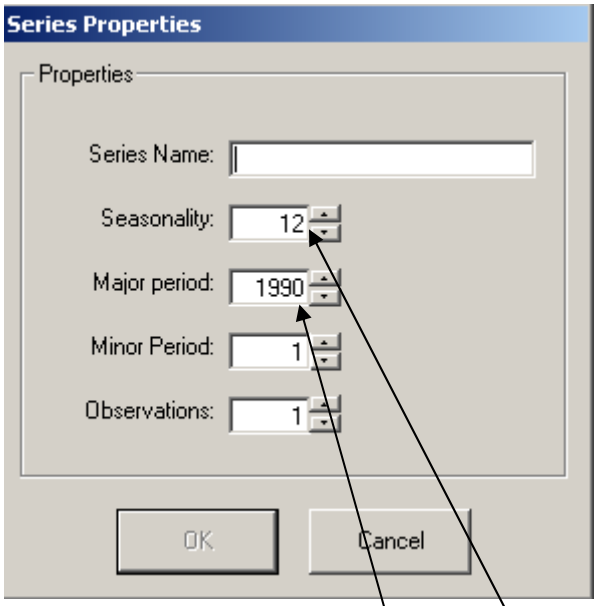

| | | Actual | Fit | Error | % Error |
|---|---|---|---|---|---|
| 1 | 1998/2 | 15.00 | 15.20 | -0.20 | -1.33 |
| 2 | 1998/3 | 14.00 | 15.20 | -1.20 | -8.57 |
| 3 | 1998/4 | 6.00 | 15.20 | -9.20 | -153.33 |
| 4 | 1998/5 | 65.00 | 65.00 | 0.00 | 0.00 |
| 5 | 1998/6 | 8.00 | 15.20 | -7.20 | -90.00 |
| 6 | 1998/7 | 8.00 | 15.20 | -7.20 | -90.00 |
| 7 | 1999/1 | 26.00 | 15.20 | 10.80 | 41.54 |
| 8 | 1999/2 | 11.00 | 15.20 | -4.20 | -38.18 |
| 9 | 1999/3 | 19.00 | 15.20 | 3.80 | 20.00 |
| 10 | 1999/4 | 20.00 | 15.20 | 4.80 | 24.00 |
| 11 | 1999/5 | 45.00 | 45.00 | 0.00 | 0.00 |
| 12 | 1999/6 | 24.00 | 15.20 | 8.80 | 36.67 |
| 13 | 1999/7 | 41.00 | 41.00 | 0.00 | 0.00 |
| 14 | 2000/1 | 25.00 | 15.20 | 9.80 | 39.20 |
| 15 | 2000/2 | 14.00 | 15.20 | -1.20 | -8.57 |
| 16 | 2000/3 | 1.00 | 1.00 | 0.00 | 0.00 |

If you decide that you want to enter in or cut and paste in your own data then you would click on “File/New” from the menu.



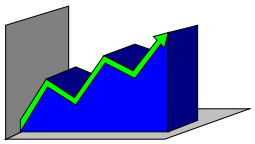


You then need to fill in the properties of the series. When the spreadsheet appears, you can enter your data by cutting and pasting from Excel or entering it by hand, and then click "File/Save" to store the file for the future.




Here are some examples on the type of seasonality you could enter – Choose 1 for annual, 4 for quarterly, 12 for monthly, 52 for weekly, 7 for daily(7 days in a week), 5 for daily(5 days in a week) and 24 for hourly.

Here are some examples on the Major/Minor Period – For January 1990 monthly data, choose 1990 for the major period and 1 for the minor period. For January 2, 2006 daily data, choose 1 for the major period and 2 for the minor period.



# Three Ways To Get Data Into Autobox

These are three ways to get data into Autobox:

1)Import series from Excel

Select File/Import. The data ***MUST*** start in either row 1 or row 2 (row 2 if there is a header). You will be prompted for the series header information (i.e. what is the seasonality?)

2)Cut and paste into the data sheet or Data entry.

Select File\New. You will be prompted to enter the series name and properties. You can then enter the data by hand or cut and paste into the Historical Data spreadsheet.

3) Create you own Autobox ASC file.

You need to create a txt file with an *.ASC extension. You store the data and series header information in a specified format in a column. Select File\Open and the ASC file with the data. The properties and data will be automatically inserted into the proper spreadsheet(s). We have many ASC files that were created in the installation directory which can be used as reference.

# Import Series From Excel

Our Excel import is used for importing data from spreadsheets that are vertically oriented. You can transpose the data in Excel if need be. This means that series are in columns and observations are in rows extending down. ***You must have the data beginning in row 1 or 2!*** Let's take, for example, the spreadsheet as seen below as an example of vertical data.

It is important to note that the causal variables will have a data type (which defines for Autobox how the future values are created and used in Autobox – more on that on page 35). The important thing to note is that if the series has no future values then the data type is defaulted to '0'. If the causal variable does have future values then the data type is defaulted to '2'. You can change these by clicking on "Series/Series Information". Note that if you have a promotion that if you have daily data and a promotion that goes on for many periods you would want to use a data type of '1' as you can't really look for lead/lag from multiple time periods.

demoBook.xls

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SKU | Qty | Adv | Date | | | | | |
| 2 | 1000 | 95623 | 0 | 3/31/2006 | | | | | |
| 3 | 1000 | 83721 | 0 | 6/30/2006 | | | | | |
| 4 | 1000 | 104953 | 1 | 9/31/2006 | | | | | |
| 5 | 1000 | 123721 | 0 | 12/31/2006 | | | | | |
| 6 | 1001 | 95623 | 0 | 3/31/2006 | | | | | |
| 7 | 1001 | 83721 | 0 | 6/30/2006 | | | | | |
| 8 | 1001 | 104953 | 1 | 9/31/2006 | | | | | |
| 9 | 1001 | 123721 | 0 | 12/31/2006 | | | | | |
| 10 | 1002 | 95623 | 0 | 3/31/2006 | | | | | |
| 11 | 1002 | 83721 | 0 | 6/30/2006 | | | | | |
| 12 | 1002 | 104953 | 1 | 9/31/2006 | | | | | |
| 13 | 1002 | 123721 | 0 | 12/31/2006 | | | | | |
| 14 | 1004 | 95623 | 0 | 3/31/2006 | | | | | |
| 15 | 1004 | 83721 | 0 | 6/30/2006 | | | | | |
| 16 | 1004 | 104953 | 1 | 9/31/2006 | | | | | |
| 17 | 1004 | 123721 | 0 | 12/31/2006 | | | | | |
| 18 | | | | | | | | | |
| 19 | | | | | | | | | |
| 20 | | | | | | | | | |
| 21 | | | | | | | | | |
| 22 | | | | | | | | | |
| 23 | | | | | | | | | |

Fig. 1

Notice how each column has a heading identifying what is contained in the column and that each row contains and observation extending through time. When taking a close look at the first column you notice that the same number is repeated several times indicating observations for a particular SKU and then repeating with different SKU numbers. When importing a spreadsheet like this many SKU numbers as one series won't make much sense to Autobox, so we need some way to filter out just the items that we need. This can be accomplished by using the filter wizard step shown in Fig. 2.

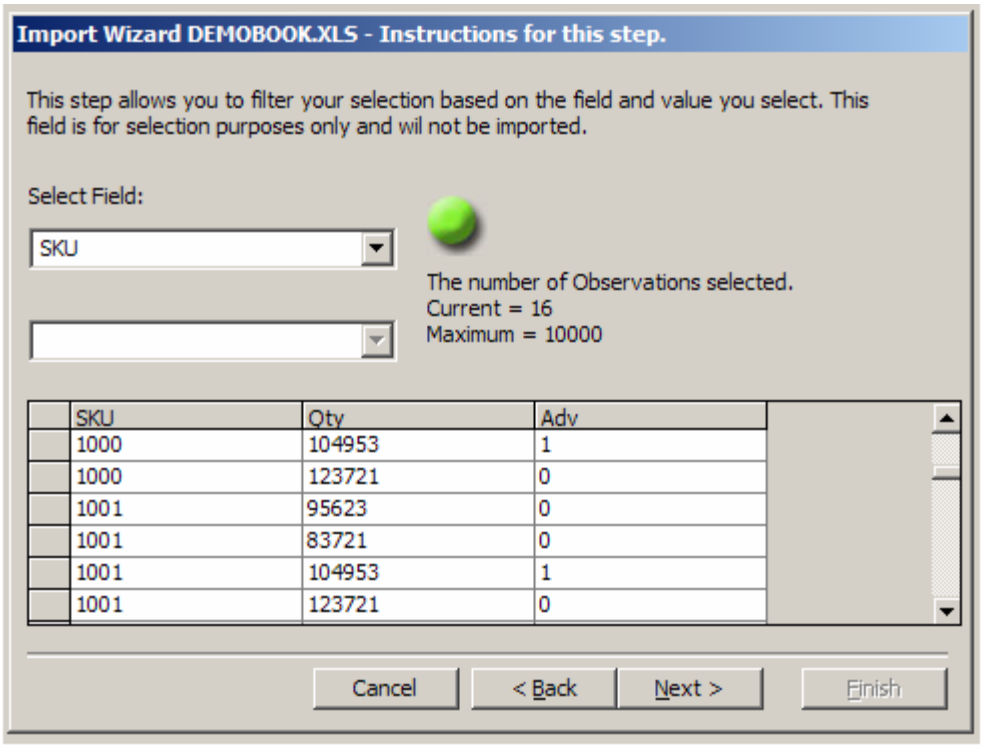



Fig. 2

To filter use the select field pull down to choose the field you would like use to filter your Data. Please note that only fields in the current data set will be available for selection. For this example we will be using the SKU field to filter our data. After selecting the SKU field, you can then select the value you would like to use as the filter. This is displayed in Fig. 3.

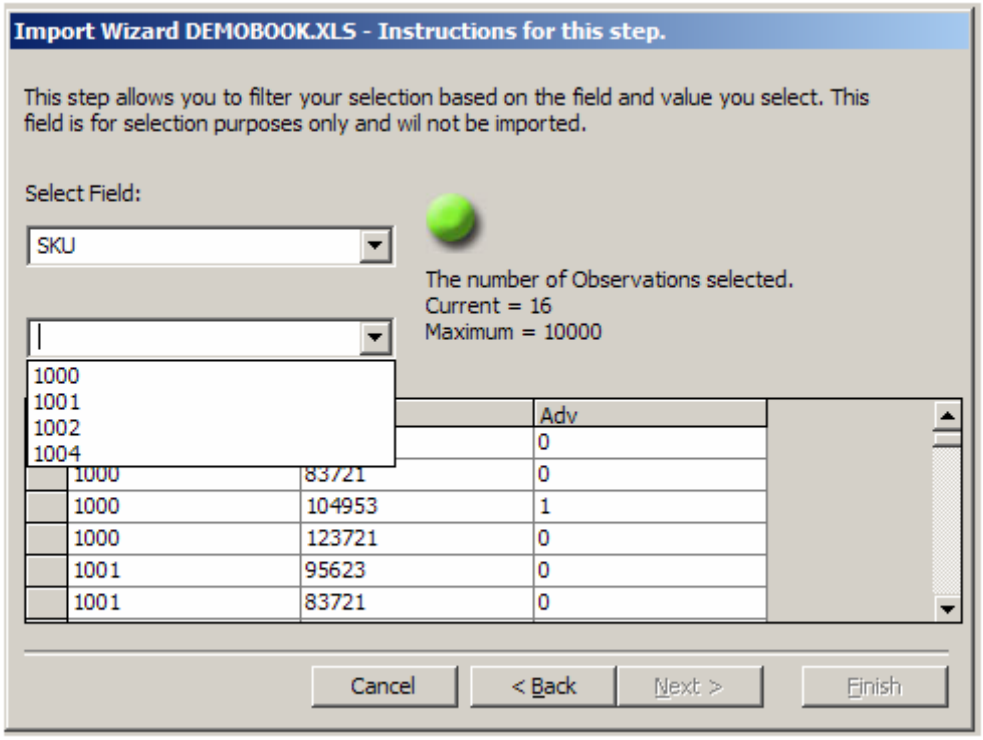



Fig 3.

When the pull down for the value is clicked, only the unique values that pertain to the SKU field are available for selection. By choosing the value 1000 we have selected the 1000 SKU as our filter; and only data that pertain to that SKU is available for import. An example of the data is displayed in Fig. 4.

**NOTE:** Please keep in mind when using the filter function that the field that you use to select your data will not be imported. The reason for this is that part numbers or SKU's are not used in modeling----just the data is needed.

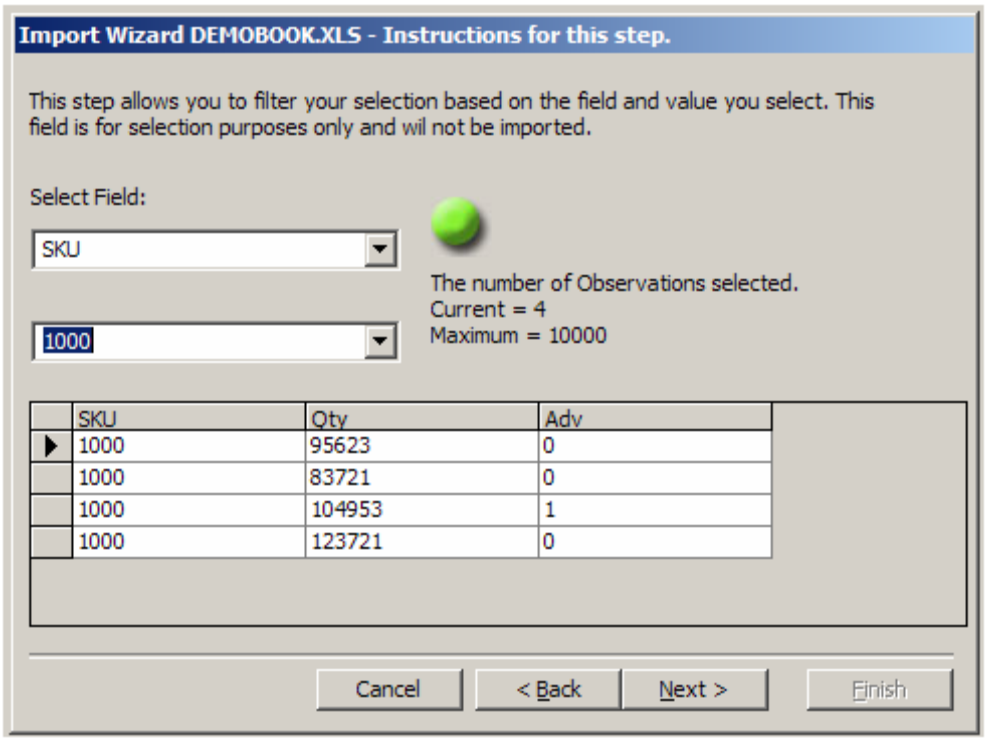
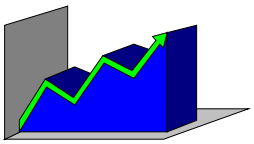



Fig. 4



The next step is to setup the properties for the Series you are importing. Fig 5 shows the Series properties step of the import wizard. If your data is not a causal problem, just pick any field as the output series field. Autobox will save the ASC file(s) into the folder where it imported the XLS file from. Autobox will also automatically generate an ASC file with a column for causal and an ASC file for each and every one of the columns in anticipation of the data NOT being a causal dataset.

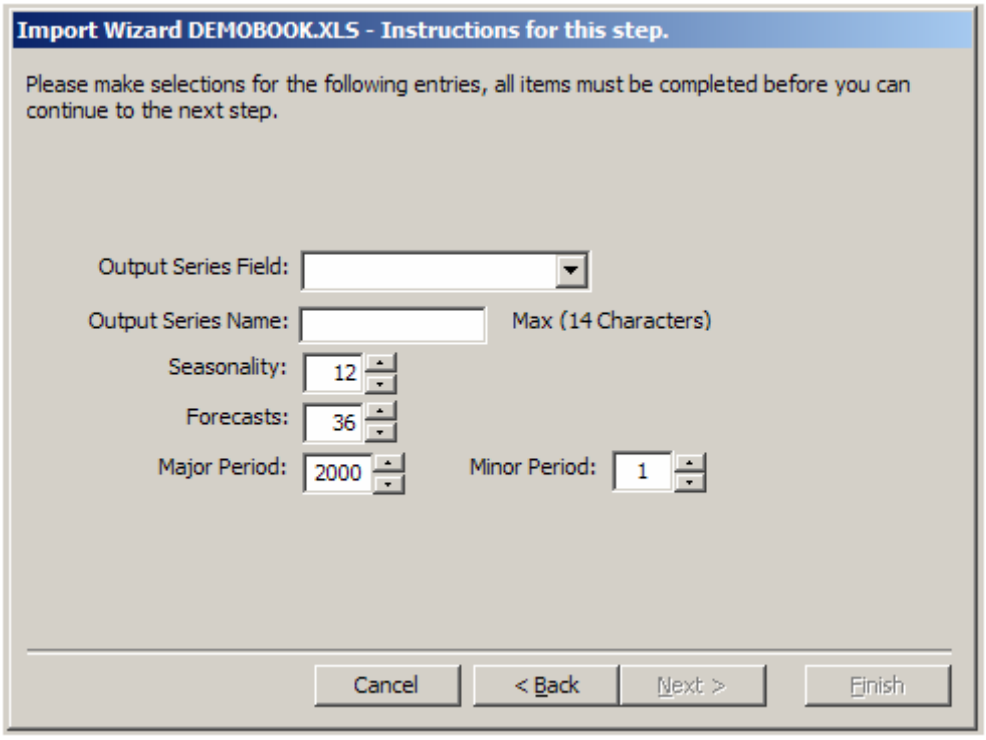
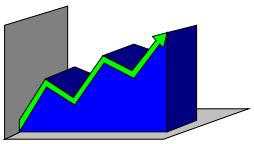



Fig 5.



The following explains the individual properties

| | |
|---|---|
| Output Series Field: | Use this field to select the series you would like to forecast in this case it would be QTY. |
| Output Series Name: | This is just a name for the series that can be any alphanumeric characters we will use SKU1000 |
| Seasonality | How often the observations are taken |
| Forecasts | The number of forecasts you would like Autobox to generate |
| Major Period | In this case it would be the year 2006 |
| Minor Period | Enter the period the first observation had taken place, this number cannot exceed the seasonality. Example: Seasonality = 4 the maximum value for Minor Period would be 4 |

Known issues and limitations:

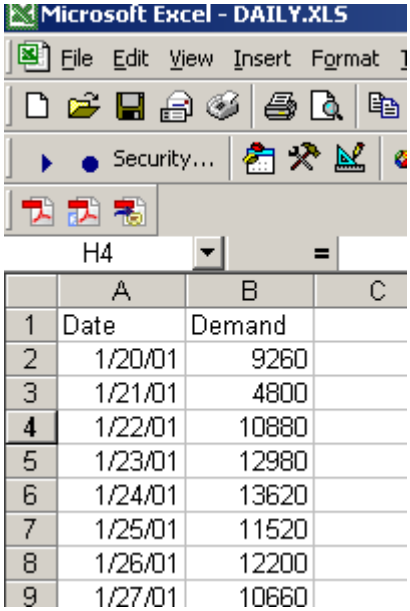
- The spreadsheet must be closed in order for the import to work correctly. This means it cannot be open in Excel while you are trying to import.
- Horizontal data cannot be imported at this time, this means series that are listed in rows 1 – X and observations listed in columns A – Z…. will not import.
- Series names are limited to 14 characters.
- Vista XLS files will not import

# Cut and Paste or “Data Enter” to get Data into Autobox

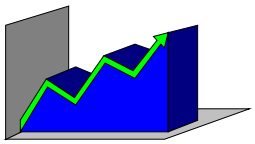
For these examples, we provide you with the Excel file and a tutorial on how to take that data and get it into Autobox in order to get started.

## Daily Sales

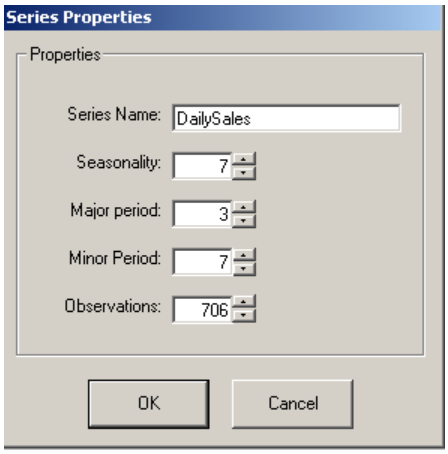
This example is daily data and our data is in a file named “daily.xls” which is installed when Autobox is installed. Here is a partial view of the data.



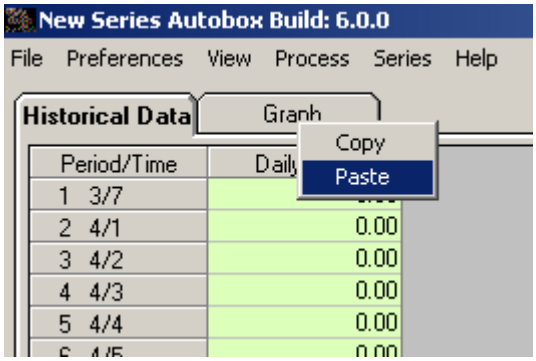

| | A | B | C |
|---|---|---|---|
| 1 | Date | Demand | |
| 2 | 1/20/01 | 9260 | |
| 3 | 1/21/01 | 4800 | |
| 4 | 1/22/01 | 10880 | |
| 5 | 1/23/01 | 12980 | |
| 6 | 1/24/01 | 13620 | |
| 7 | 1/25/01 | 11520 | |
| 8 | 1/26/01 | 12200 | |
| 9 | 1/27/01 | 10660 | |

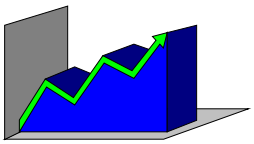


We enter Autobox and choose “File/New” and then add the series properties. We choose 7 as the seasonality as the data is daily observations. The major period of ‘3’ is week 3 and minor period of ‘7’ is the 7th day of the week. Monthly data would have seasonality of “12”.




After you click “OK”, you can go into Excel and highlight your data (only the data, no need to highlight any text or dates!), copy it (CTRL-C or Edit/Paste) and paste it into Autobox’s Historical Data spreadsheet.





# Create Your Own .ASC File

Instead of entering your own data or cutting and pasting you can create a file that can be read directly into Autobox. You can open an ASC file to look at and you will quickly see that the first ~13 rows are "header information" that Autobox needs to know before running. You can't read in more than one output series to be forecast since we don't allow Autobox to run in batch mode (we can't give everything away!).

This is really an easy process, but the information and data must be entered in a text file in a very specific order in a single column:

Objectives (all are required)

Data properties (all are required)

Data names (in the order of 1st input series to nth input series, if any; and then the output series)

Data type (in the same order as the data names)

Historical Data (in the same order as the data names)

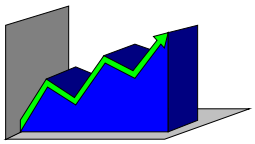
Future Values (for all input series which have a data type of 1, 2, or 3, if any, in the same order as data names)

Retained Data (If any, for all series in the same order as the data names)

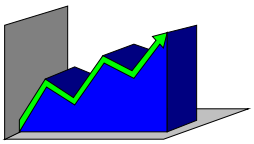
The following structure tables indicate the parameters and/or limitations for each of the above categories.

**DATAPROP Structure:**

| Name | Description |
|---|---|
| OBJECTIVE(1) | Sets forth the model conditions as indicated by the following:<br><br>0 = Totally Automatic<br><br>NONCAUSAL MODELS IN AUTOBOX MEMORY<br>1 = MEAN<br>2 = AUTOREGRESSIVE (1) WITH CONSTANT |



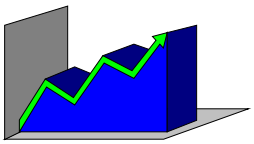
| | |
|---|---|
| | 3 = AUTOREGRESSIVE(2) WITH CONSTANT<br>4 = SIMPLE EXPONENTIAL SMOOTHING NO CONSTANT<br>5 = LINEAR (HOLT) EXPONENTIAL SMOOTHING NO CONSTANT<br>6 = RANDOM WALK NO CONSTANT<br>7 = RANDOM WALK WITH CONSTANT<br>8 = TIME TREND<br>9 = TIME TREND PLUS AR(1) CORRECTION<br>10 = FOURIER<br>11 = HOLT LINEAR TREND PLUS ADDITIVE SEASONAL FACTORS (TREND FORM)<br>12 = DAMPED TREND LINEAR EXPONENTIAL SMOOTHING NO CONSTANT<br>13= SEASONAL EXPONENTIAL SMOOTHING NO CONSTANT<br>14 = HOLT LINEAR TREND PLUS ADDITIVE SEASONAL FACTORS ( ARIMA FORM)<br>15 = AIRLINE<br>98 = HOLT-WINTERS TREND PLUS MULTIPLICATIVE SEASONAL FACTORS (TREND FORM)<br><br>CAUSAL MODELS IN AUTOBOX MEMORY<br>51 = REGRESSION<br>52 = REGRESSION WITH AR(1) CORRECTION<br><br>COMMON<br>97 = IDENTIFICATION ONLY<br>99 = STARTMOD.123<br>199 = STARTMOD.123 + SIM<br>200 = Totally Automatic + ABOXLITE model is developed |



| | |
|---|---|
| OBJECTIVE (2) | Sets source of conditions for processing:<br><br>0 = Use default conditions in memory |
| OBJECTIVE (3) | 0 = save reports |

**DATAPROP Structure:**

| Name | Description |
|---|---|
| DATAPROP(1) | Number of series in the problem |
| DATAPROP(2) | Not used |
| DATAPROP(3) | Not used. |
| DATAPROP(4) | Beginning period. the starting point of the data. (i.e. 1 for the 1st week in the year)<br><br>Please note that all series in the model must have the same Beginning Period. If you wish to use series whose original Beginning Period are different, you must determine the common matrix for the series and use that starting point as the Beginning Period. |
| DATAPROP(5) | Number of historical values in each of the time series in the model |
| DATAPROP(6) | Number of future values to be included for each applicable input series.<br><br>If a causal model (includes a dependent and independent series) and the DATATYPE of any the input(independent) series is 1 , 2, or 3, enter DATAPROP(7) + Number of Future Values (this must equal the number of forecasts to be calculated) to be supplied by the user.<br><br>If DATATYPE of all input series is 0, or if a noncausal |



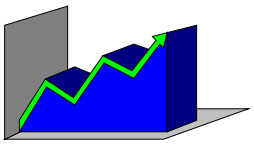
| | |
|---|---|
| | model, this must show a 0. |
| DATAPROP(7) | The number of values retained from the end of the series to be used to evaluate prior forecasts (enter 0 if none) |
| DATAPROP(8) | Number of forecast values to be calculated |

**DATANAME Structure:**

| Name | Description |
|---|---|
| DATANAME | Actual name of each series in model in the order 1st Input series, 2d input series, …N input series, output series<br><br>These names must be limited to 22 characters for Input series and 14 characters for the output series; and they cannot contain space(s), period(.), exclamation point(!), backquote(‘), brackets([]), wild card characters such as * or ?, and control characters(ASCII values 0 through 31). |

**DATATYPE Structure:**

| Name | Description |
|---|---|
| 0 | Future values are self-projected; contemporaneous and lag effects allowed |
| 1 | Future values are user specified; contemporaneous effects allowed |
| 2 | Future values are user specified; contemporaneous and lag effects allowed |
| 3 | Future values are user specified; contemporaneous and lag and lead effects allowed |



The following is an example of an .ASC file for a noncausal (single) series{annotations are not included in the file}:

| | |
|---|---|
| 0 | (objective(1) indicates totally automatic modeling) |
| 0 | (objective(2) indicates use default rules in memory) |
| 0 | (objective(3) indicates full output) |
| 1 | (DataProp(1) number of series in the problem set) |
| 52 | (DataProp(2) seasonality of the series) |
| 1998 | (DataProp(3) beginning year or major period) |
| 2 | (DataProp(4) beginnng or minor period) |
| 67 | (Dataprop(5) number of historical data in series) |
| 0 | (DataProp(6) number of future values ) |
| 0 | (DataProp(7) number of retained data |
| 24 | (DataProp(8) number of forecasts to be calculated |
| pizza | (output series name – causal variables would be added before this line, Also, by specifying the date (January 1, 2006) with the name in this format "__010106Y11" and line 5 has a '7' then U.S. holidays are automatically generated, day of the month effect is analyzed and the effect of "Friday before" and "Monday after" a weekend holiday is analyzed) |
| 0 | (Data type) |
| 15 | (historical data – 67 observations - causal data would be added before this series) |
| 14 | |
| 6 | |
| . | |
| 12 | We recommend that you open an ASC file that already exists (such as 116x573.asc in \Kitchsnk directory) as a good example of how a file should look if it is to be a casual model. |

future values would be added here - note only future causal data!

Retained future values would be added here - causal and series to predicted!

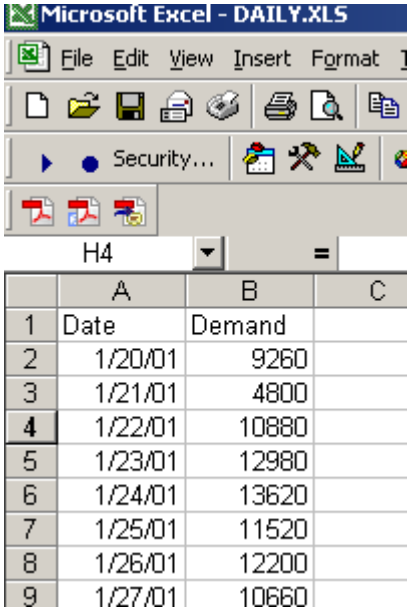
# Some Examples To Show You How To Get Started

We will **repeat** some of the steps reading in the data **on the next two pages** as we did before and then introduce some new concepts.

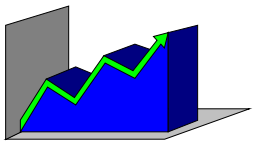
For these examples, we provide you with the Excel file and a tutorial on how to take that data and get it into Autobox in order to get started. Again, a reminder that you must have the data in row 1 or row 2 in the Excel file.

## Daily Sales

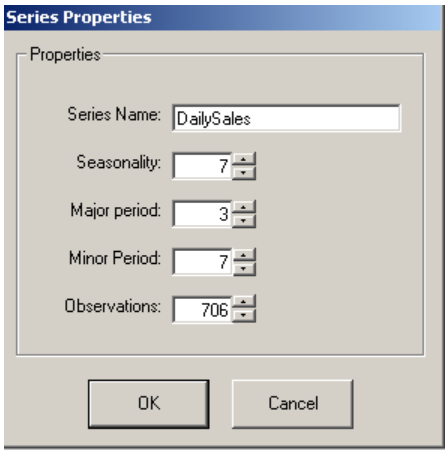
This example is daily data and our data is in a file named "daily.xls" which is installed when Autobox is installed. Here is a partial view of the data.



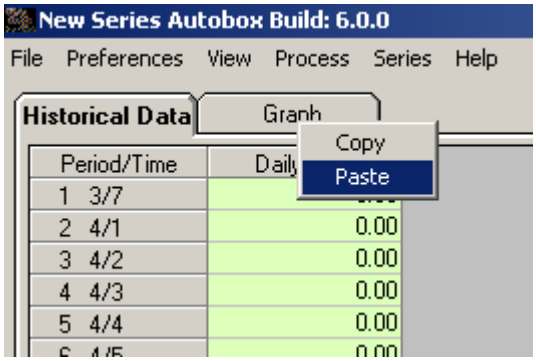

| | A | B | C |
|---|---|---|---|
| 1 | Date | Demand | |
| 2 | 1/20/01 | 9260 | |
| 3 | 1/21/01 | 4800 | |
| 4 | 1/22/01 | 10880 | |
| 5 | 1/23/01 | 12980 | |
| 6 | 1/24/01 | 13620 | |
| 7 | 1/25/01 | 11520 | |
| 8 | 1/26/01 | 12200 | |
| 9 | 1/27/01 | 10660 | |

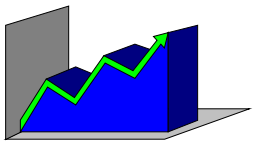


We enter Autobox and choose “File/New” and then add the series properties. We choose 7 as the seasonality as the data is daily observations. The major period of ‘3’ is week 3 and minor period of ‘7’ is the 7th day of the week. Monthly data would have seasonality of “12”.

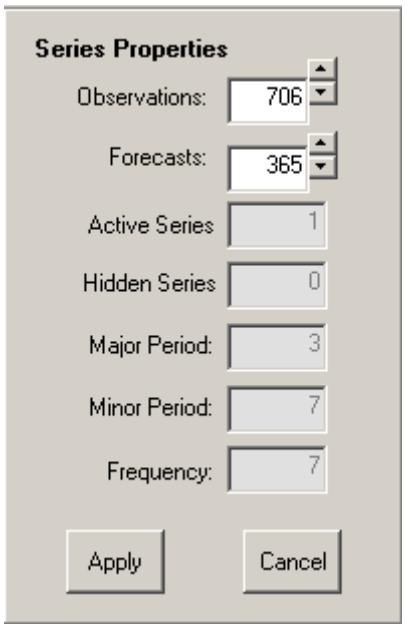



After you click “OK”, you can go into Excel and highlight your data (only the data, no need to highlight any text or dates!), copy it (CTRL-V or Edit/Paste) and paste it into Autobox’s Historical Data spreadsheet.

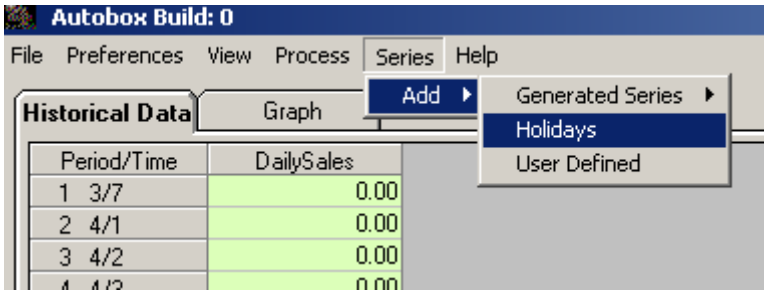


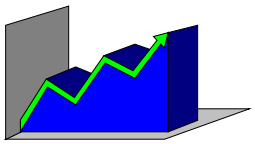


Change the number of forecasts to 365 as seen in the series properties on the right hand side of the screen and the click “APPLY”. Note that the period & frequency can’t be changed at all.

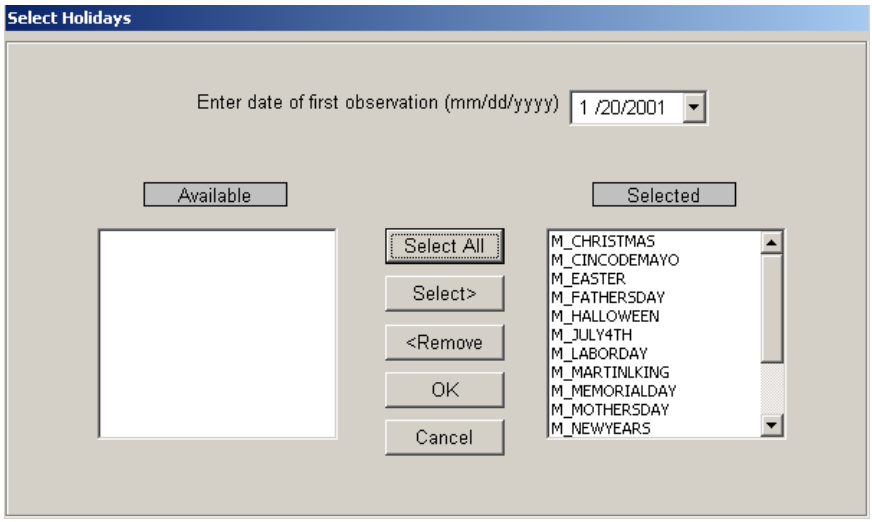



Choose Series/Add/Holidays.

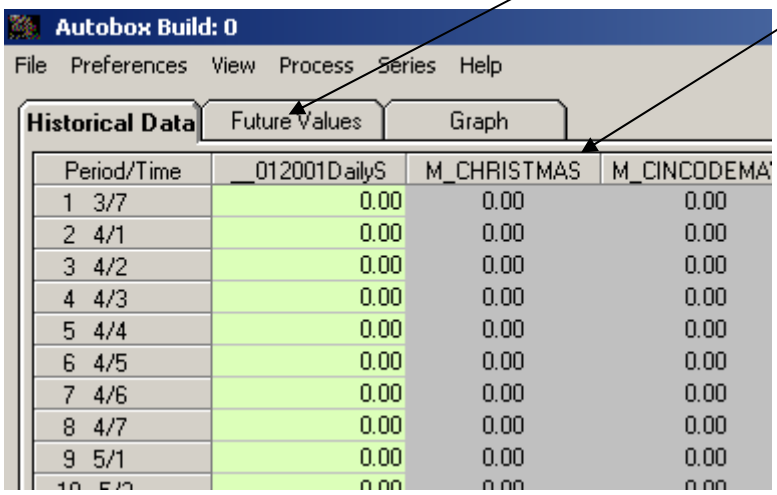


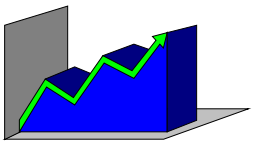


Change the date to the start of the series 01/20/2001 and “then choose “Select All” and then click “OK’. This will let Autobox evaluate all of the different holidays in the year for any unusual behavior and account for it in the model/forecast. In addition, Autobox has created 365 future observations of these holiday variables so if they are significant they will be used to forecast the series.

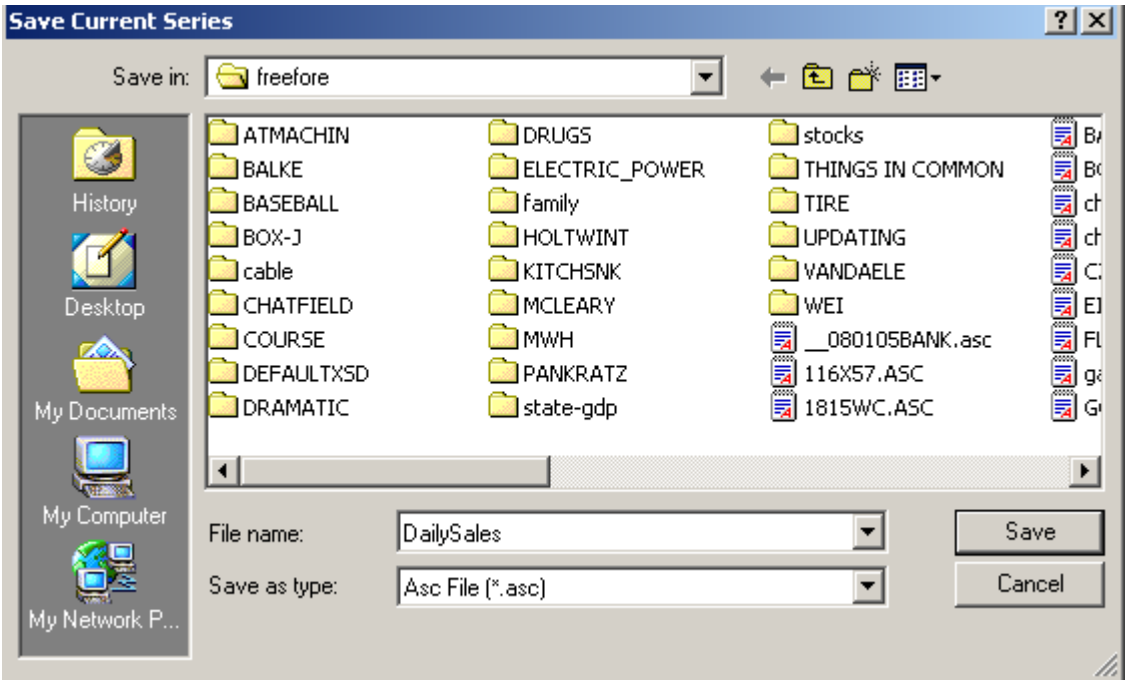



The output series was renamed to include the first date and various holiday variables have been generated.





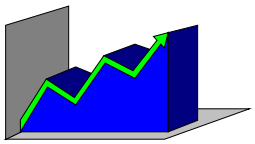
You can then choose “Process/Run” and get your results. You can save your example by choosing “File/Save” so that you will have your example saved.



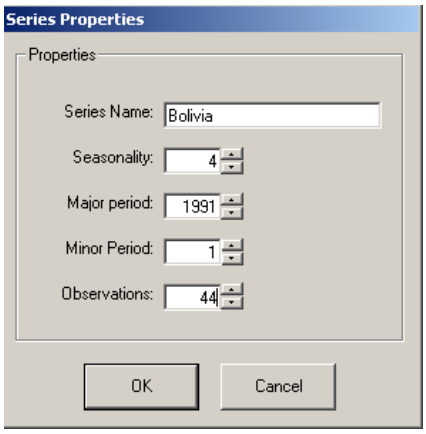

Univariate example

We have one dataset that we are looking to forecast. The series, which contains sales in Bolivia, starts on 1/1/1991, and has a seasonality of 4. We have 44 observations that we are looking to forecast 4 periods out. Our data is in Excel file named Ex2.xls.

| | A | B | C |
|---|---|---|---|
| 1 | Date | Sales | |
| 2 | 1/1/1991 | 15602 | |
| 3 | 4/1/1991 | 17115 | |
| 4 | 7/1/1991 | 18795 | |
| 5 | 10/1/1991 | 19002 | |
| 6 | 1/1/1992 | 17851 | |
| 7 | 4/1/1992 | 18296 | |
| 8 | 7/1/1992 | 19980 | |
| 9 | 10/1/1992 | 20097 | |
| 10 | 1/1/1993 | 19381 | |
| 11 | 4/1/1993 | 19902 | |
| 12 | 7/1/1993 | 21030 | |
| 13 | 10/1/1993 | 22590 | |



We enter Autobox and choose “File/New” and then add the series name and properties.

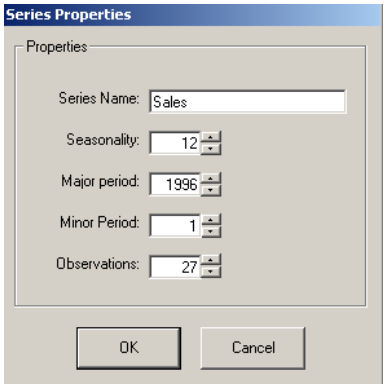


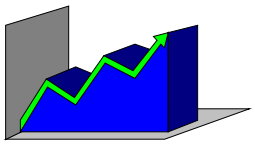

After you click “OK”, you can go into Excel and highlight your data and paste it into Autobox. Then change the number of forecasts to 4. Choose “Process/Run” to get your results. You can save your example by choosing “File/Save”.

Lead Effect example

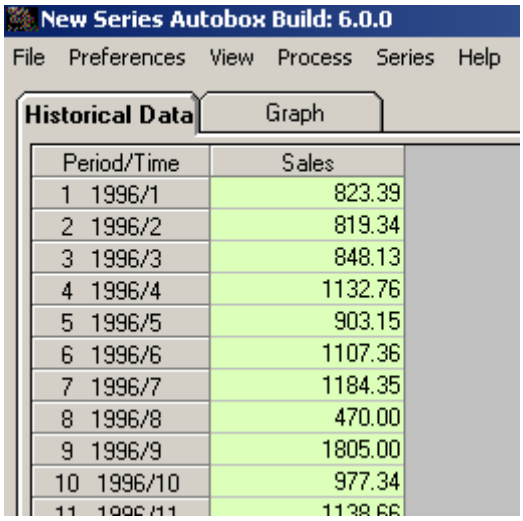
There is going to be a promotion on 9/1996 and 7/1997. We would like to account for that in the modeling process. This example will detect an effect BEFORE the promotion as people may not end up buying the product before the sale and buy more when the sale is in effect. Autobox will look up to 4 periods before an event, but that can be overridden using engine.afs (search for engine.afs in the manual for more on how to edit engine.afs). You shouldn’t create causal variables with a data type of ‘3’ within 4 periods of one another as this would conflict the search for lead/lag. Autobox will search for these instances and correct for this situation.

We enter Autobox and choose “File/New” and then add these series properties.

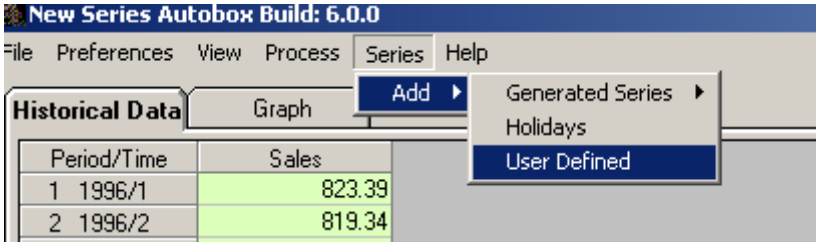


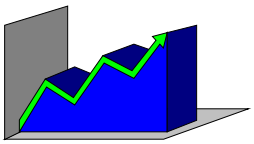


Open the Excel file named “Ex1.xls” and cut and paste the data into Autobox. Then change the number of forecasts to 36.

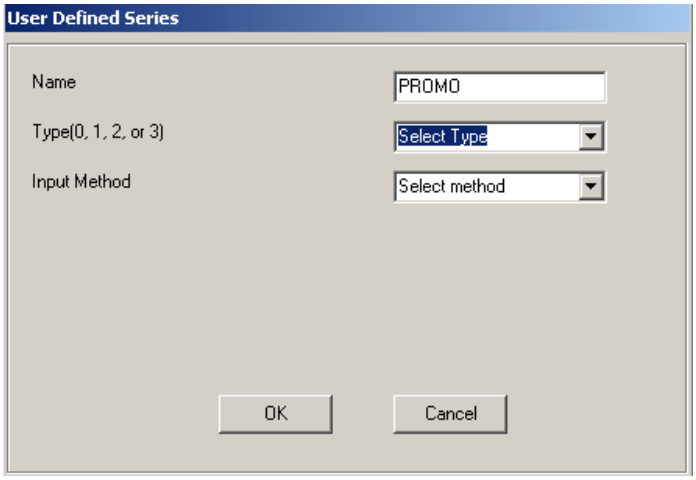


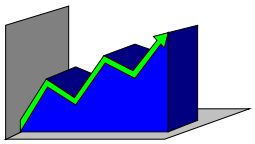

Let’s add the promotion variable by choosing “Series/Add/User Defined”





Enter “Promo” for the name and choose “data types”





Let's discuss the ***data types*** as it is very important how Autobox treats your causals. If you know there is no way for the causal variables can't have a lag then you wouldn't want to choose option '2' or '3'. Bad choices here can have bad effects in your results.

## Data Types:

0 – Autobox will forecast the future values of the causal series and try and identify current and lag effects during the analysis

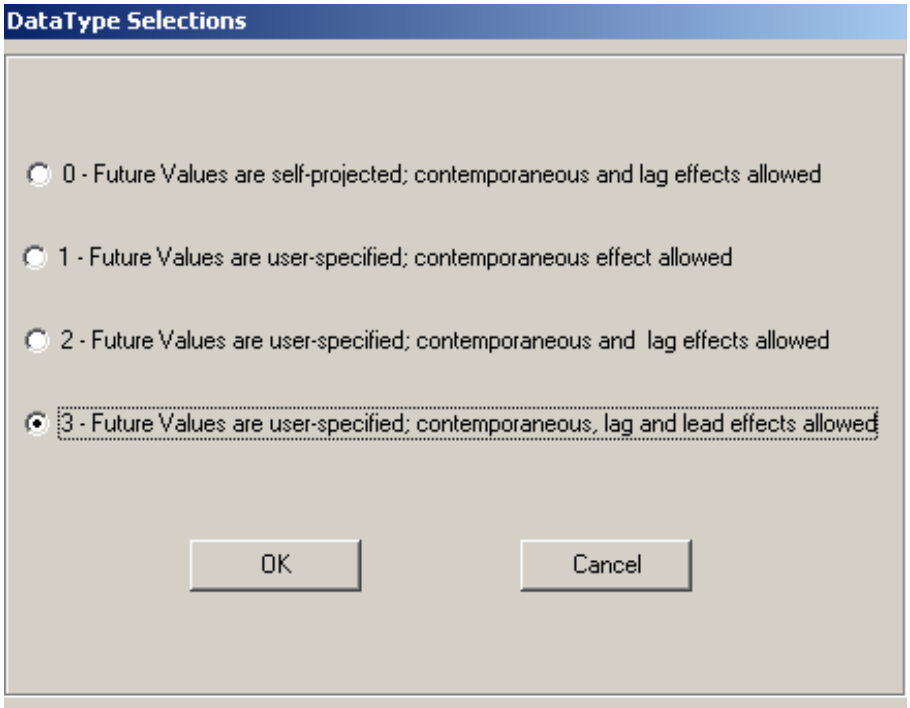
The next 3 have the future values defined by the user AND

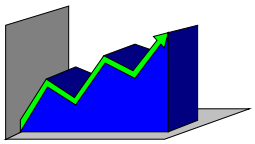
1 – current time period effects (use this for something like "seasonal dummies" or you know there is no lag/lead effect) are analyzed

2 – current time period and lag effects (same as data type '0' except the user is specifying the future values) are analyzed

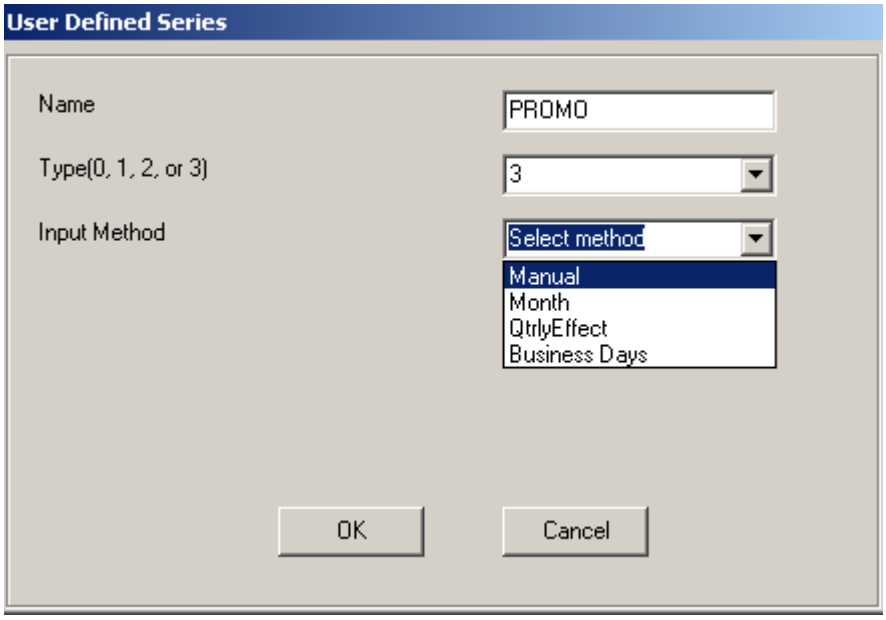
3 – current time period and lag and lead effects are analyzed

We choose 3 for this example.

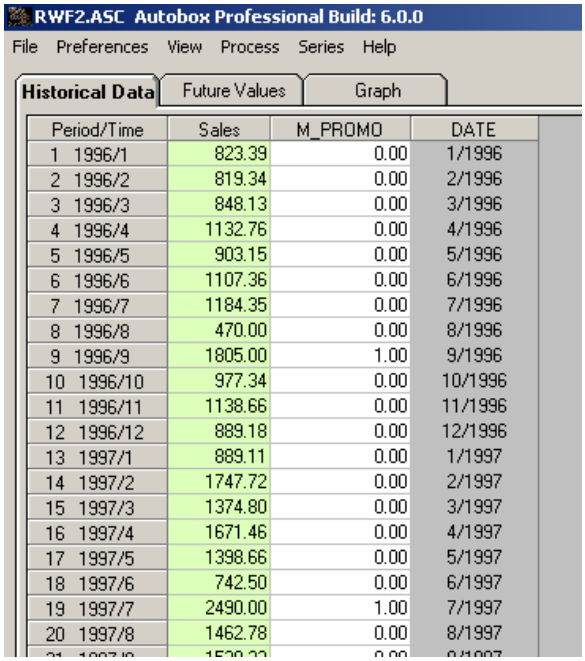




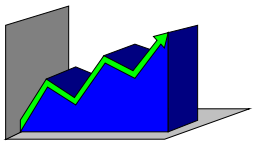
Choose “Manual” as the input method.



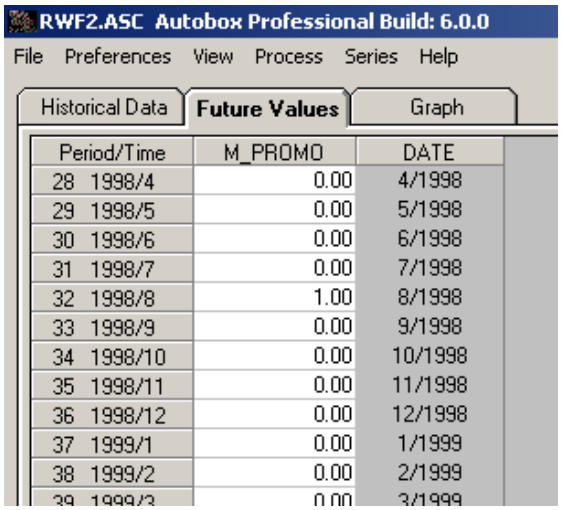

So, now we enter a “1” for 9/96 and 7/97 for the historical values of promotion,




| Period/Time | Sales | M_PROMO | DATE |
|---|---|---|---|
| 1 1996/1 | 823.39 | 0.00 | 1/1996 |
| 2 1996/2 | 819.34 | 0.00 | 2/1996 |
| 3 1996/3 | 848.13 | 0.00 | 3/1996 |
| 4 1996/4 | 1132.76 | 0.00 | 4/1996 |
| 5 1996/5 | 903.15 | 0.00 | 5/1996 |
| 6 1996/6 | 1107.36 | 0.00 | 6/1996 |
| 7 1996/7 | 1184.35 | 0.00 | 7/1996 |
| 8 1996/8 | 470.00 | 0.00 | 8/1996 |
| 9 1996/9 | 1805.00 | 1.00 | 9/1996 |
| 10 1996/10 | 977.34 | 0.00 | 10/1996 |
| 11 1996/11 | 1138.66 | 0.00 | 11/1996 |
| 12 1996/12 | 889.18 | 0.00 | 12/1996 |
| 13 1997/1 | 889.11 | 0.00 | 1/1997 |
| 14 1997/2 | 1747.72 | 0.00 | 2/1997 |
| 15 1997/3 | 1374.80 | 0.00 | 3/1997 |
| 16 1997/4 | 1671.46 | 0.00 | 4/1997 |
| 17 1997/5 | 1398.66 | 0.00 | 5/1997 |
| 18 1997/6 | 742.50 | 0.00 | 6/1997 |
| 19 1997/7 | 2490.00 | 1.00 | 7/1997 |
| 20 1997/8 | 1462.78 | 0.00 | 8/1997 |



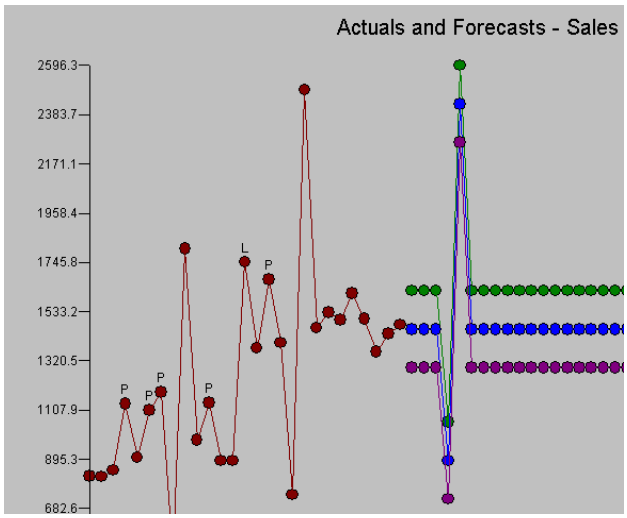
We then click on “future values” and enter our future promotion for 8/98.

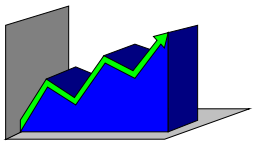



| Period/Time | M_PROMO | DATE |
|---|---|---|
| 28 1998/4 | 0.00 | 4/1998 |
| 29 1998/5 | 0.00 | 5/1998 |
| 30 1998/6 | 0.00 | 6/1998 |
| 31 1998/7 | 0.00 | 7/1998 |
| 32 1998/8 | 1.00 | 8/1998 |
| 33 1998/9 | 0.00 | 9/1998 |
| 34 1998/10 | 0.00 | 10/1998 |
| 35 1998/11 | 0.00 | 11/1998 |
| 36 1998/12 | 0.00 | 12/1998 |
| 37 1999/1 | 0.00 | 1/1999 |
| 38 1999/2 | 0.00 | 2/1999 |
| 39 1999/3 | 0.00 | 3/1999 |

Now back to our example, you can enter a “1” at 9/1996 and 7/1997 to show that there was a promotion.

You can then choose “Process/Run” and get your results. You can see that the promotion did have an effect on the sales and the forecast reflects that. You will notice that before the promotion there is a reduction in demand and on the promotion it is increased. You can save your example by choosing “File/Save”.





We will show you other examples if you choose the other "Input Methods" you assign a monthly effect using "Monthly" or a business days effect "Business Days" reflecting the number of trading days in a month or a quarterly effect.

RWF2.ASC Autobox Professional Build: 6.0.0

File Preferences View Process Series Help

Historical Data | Future Values | Graph

| Period/Time | Sales | M_PROMO | G_GF | G_BUSDAYS | G_MONTH | DATE |
|---|---|---|---|---|---|---|
| 1 1996/1 | 823.39 | 0.00 | 4.00 | 21.00 | 0.00 | 1/1996 |
| 2 1996/2 | 819.34 | 0.00 | 4.00 | 19.00 | 0.00 | 2/1996 |
| 3 1996/3 | 848.13 | 0.00 | 5.00 | 20.00 | 0.00 | 3/1996 |
| 4 1996/4 | 1132.76 | 0.00 | 4.00 | 22.00 | 0.00 | 4/1996 |
| 5 1996/5 | 903.15 | 0.00 | 4.00 | 21.00 | 0.00 | 5/1996 |
| 6 1996/6 | 1107.36 | 0.00 | 5.00 | 19.00 | 0.00 | 6/1996 |
| 7 1996/7 | 1184.35 | 0.00 | 4.00 | 22.00 | 0.00 | 7/1996 |
| 8 1996/8 | 470.00 | 0.00 | 4.00 | 22.00 | 1.00 | 8/1996 |
| 9 1996/9 | 1805.00 | 1.00 | 5.00 | 20.00 | 0.00 | 9/1996 |
| 10 1996/10 | 977.34 | 0.00 | 4.00 | 22.00 | 0.00 | 10/1996 |
| 11 1996/11 | 1138.66 | 0.00 | 4.00 | 19.00 | 0.00 | 11/1996 |
| 12 1996/12 | 889.18 | 0.00 | 5.00 | 21.00 | 0.00 | 12/1996 |
| 13 1997/1 | 889.11 | 0.00 | 4.00 | 21.00 | 0.00 | 1/1997 |

And one more follow up on this topic. When you have daily data with a periodicity of 5 (i.e. Mon-Fri) or 7 (Mon-Sun) you must select the date of the beginning observation.

User Defined Series

Name: PROMO

Type(0, 1, 2, or 3): 1

Enter date of 1st observation(mm/dd/yyyy): 7 / 6 /2007

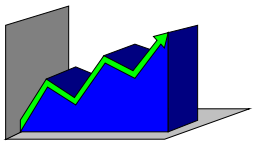
Input Method

July 2007

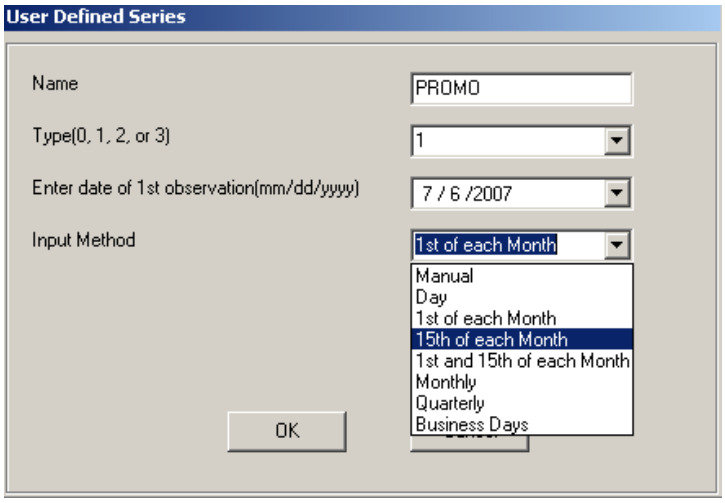
| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|---|---|---|---|---|---|---|
| 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| 29 | 30 | 31 | 1 | 2 | 3 | 4 |

Today: 6/6/2008

OK



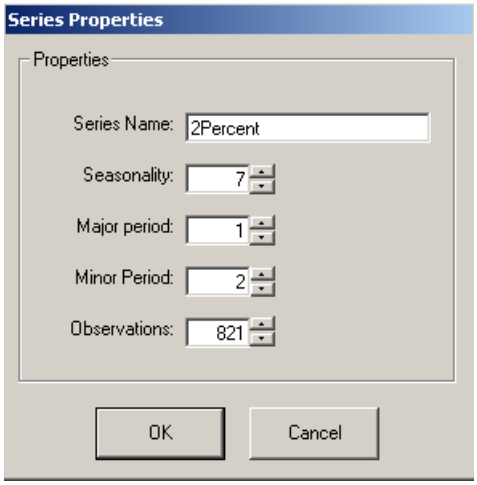
Also, you can choose some "day of the month" effects like 1st day of the month, 15th of the month or 1st and 15th.

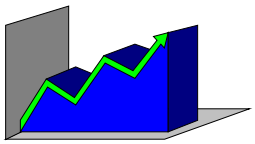



Using Causal Variables for Daily Prediction

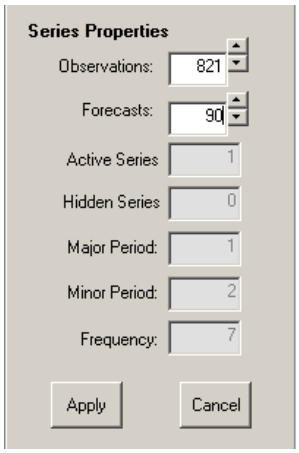
We will create a new series, by choosing "File/New". In this example, we will include not only holiday variables, but other variables that will help predict the series.

Enter in these properties



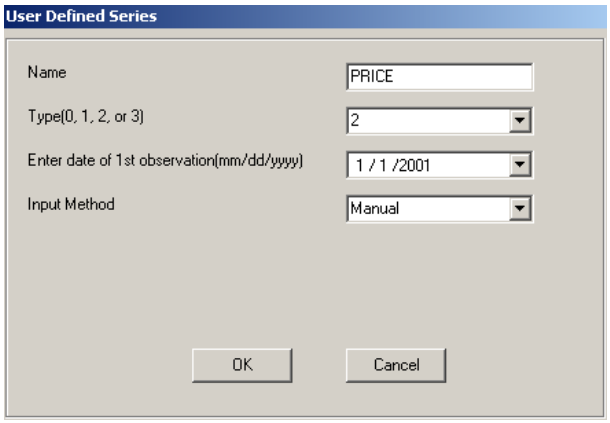


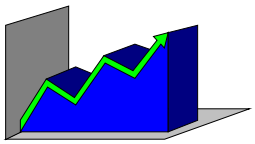
Since the actual start was January 1, 2001, a Monday, that would be a minor period of 2 since Sunday is considered the beginning of the week. Before we do anything else, we will need to set the number of forecasts on the right hand side to 90.




We then open Ex3.XLS and paste the data from column "A" into Autobox.

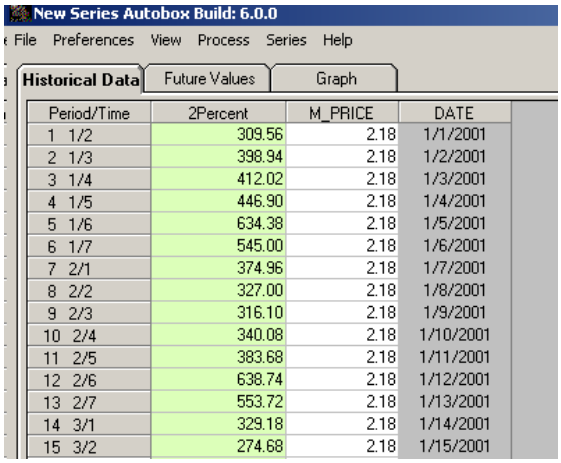
Choose "Series/Add/User Defined" with the following properties:



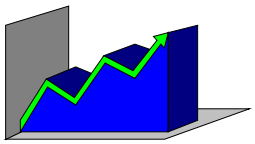


It should be pointed that daily data like this will take MUCH longer to run. Series with a periodicity of 5 will be evaluated for "day of the week" and "week of the year" effects. In addition to those two checks, if you have a periodicity of 7 then a "day of the month" effect will also be evaluated (i.e. Social Security checks always come on the 3rd of the month).

Column B of Ex3.XLS is the price of 2% milk and it contains 821 historical data and 90 future values. Cut and paste the 821 observations into the Price column in the Historical Data spreadsheet.

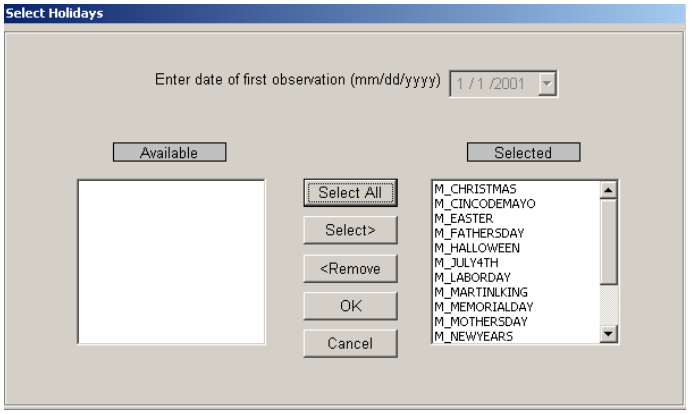



| Period/Time | 2Percent | M_PRICE | DATE |
|---|---|---|---|
| 1 1/2 | 309.56 | 2.18 | 1/1/2001 |
| 2 1/3 | 398.94 | 2.18 | 1/2/2001 |
| 3 1/4 | 412.02 | 2.18 | 1/3/2001 |
| 4 1/5 | 446.90 | 2.18 | 1/4/2001 |
| 5 1/6 | 634.38 | 2.18 | 1/5/2001 |
| 6 1/7 | 545.00 | 2.18 | 1/6/2001 |
| 7 2/1 | 374.96 | 2.18 | 1/7/2001 |
| 8 2/2 | 327.00 | 2.18 | 1/8/2001 |
| 9 2/3 | 316.10 | 2.18 | 1/9/2001 |
| 10 2/4 | 340.08 | 2.18 | 1/10/2001 |
| 11 2/5 | 383.68 | 2.18 | 1/11/2001 |
| 12 2/6 | 638.74 | 2.18 | 1/12/2001 |
| 13 2/7 | 553.72 | 2.18 | 1/13/2001 |
| 14 3/1 | 329.18 | 2.18 | 1/14/2001 |
| 15 3/2 | 274.68 | 2.18 | 1/15/2001 |

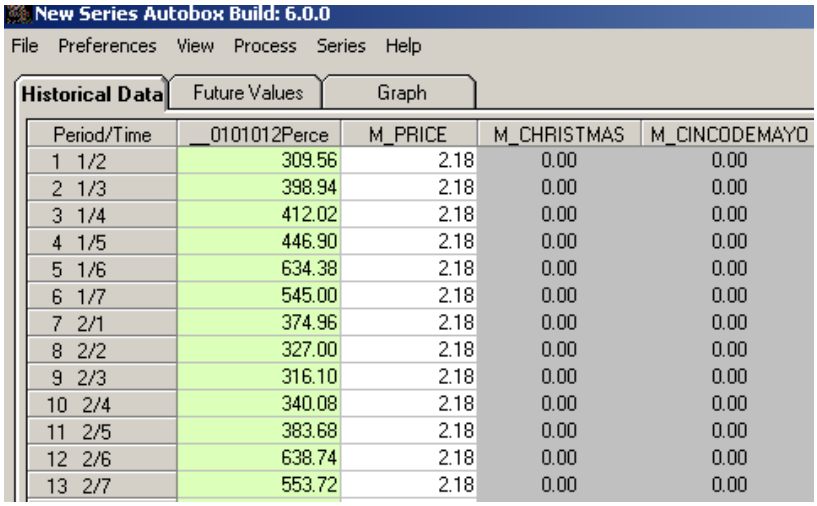


You then need to click on the “Future Values” tab so that you can cut and paste in the 90 future values of the price variable into the Price column. You need to go to row 822 in the Excel file to find the future values to paste into Autobox.

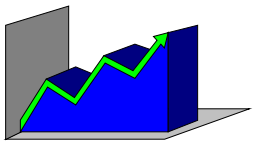
Choose “Series/Add/Holidays”. Select the date of 1/1/2001 and “select all” holidays and choose “ok”.



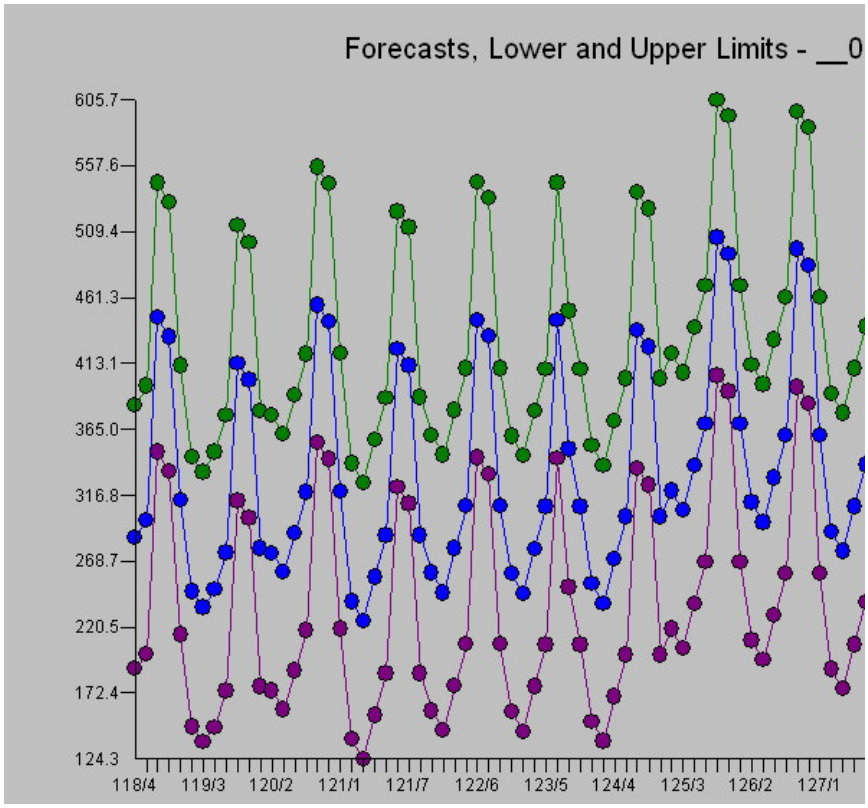

Press OK and Autobox will create dummy variables for each of the holidays… along with future values for the prediction of these events.

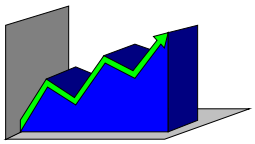



| Period/Time | _0101012Perce | M_PRICE | M_CHRISTMAS | M_CINCODEMAYO |
|---|---|---|---|---|
| 1 1/2 | 309.56 | 2.18 | 0.00 | 0.00 |
| 2 1/3 | 398.94 | 2.18 | 0.00 | 0.00 |
| 3 1/4 | 412.02 | 2.18 | 0.00 | 0.00 |
| 4 1/5 | 446.90 | 2.18 | 0.00 | 0.00 |
| 5 1/6 | 634.38 | 2.18 | 0.00 | 0.00 |
| 6 1/7 | 545.00 | 2.18 | 0.00 | 0.00 |
| 7 2/1 | 374.96 | 2.18 | 0.00 | 0.00 |
| 8 2/2 | 327.00 | 2.18 | 0.00 | 0.00 |
| 9 2/3 | 316.10 | 2.18 | 0.00 | 0.00 |
| 10 2/4 | 340.08 | 2.18 | 0.00 | 0.00 |
| 11 2/5 | 383.68 | 2.18 | 0.00 | 0.00 |
| 12 2/6 | 638.74 | 2.18 | 0.00 | 0.00 |
| 13 2/7 | 553.72 | 2.18 | 0.00 | 0.00 |



You can then choose “Process/Run” and get your results.



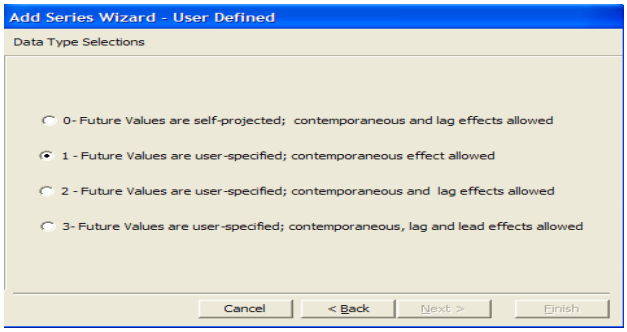


Let's talk about how to build your own dummy variable a little bit more. If you had a promotion starting in February and it ended in April, you would want to put a '1' for those three months to indicate the event.

Historical Data | Future Values | Graph

| Period/Time | 3 | M_PROMO |
|---|---|---|
| 1 2008/1 | 4186.00 | 0.00 |
| 2 2008/2 | 3462.00 | 0.00 |
| 3 2008/3 | 2552.00 | 0.00 |
| 4 2008/4 | 3045.00 | 0.00 |
| 5 2008/5 | 1287.00 | 0.00 |
| 6 2008/6 | 1395.00 | 0.00 |
| 7 2008/7 | 1290.00 | 0.00 |
| 8 2008/8 | 1681.00 | 0.00 |
| 9 2008/9 | 1435.00 | 0.00 |
| 10 2008/10 | 931.00 | 0.00 |
| 11 2008/11 | 401.00 | 0.00 |
| 12 2008/12 | 588.00 | 0.00 |
| 13 2009/1 | 2225.00 | 0.00 |
| 14 2009/2 | 15399.00 | 1.00 |
| 15 2009/3 | 9840.00 | 1.00 |
| 16 2009/4 | 22822.00 | 1.00 |
| 17 2009/5 | 6356.00 | 0.00 |
| 18 2009/6 | 3070.00 | 0.00 |
| 19 2009/7 | 2709.00 | 0.00 |

The question now becomes what type of "Data Type" is it? If you think that there is a "lead effect" to this promotion where demand is shifted before the event then this would suggest that the type should be set to '3', but it would be not be logical to have three dummy 1's consecutively when you are looking for a lead and lag relationship. There should just be 1 dummy set in March and the February and April should be a '0'. Autobox looks for lead effects up to 6 periods before the '1' dummy is specified. As a protective measure, Autobox will change your type from a '3' to a '1' if you have 1's that are within 6 periods of one another so be careful to follow these previous words or call us to discuss.

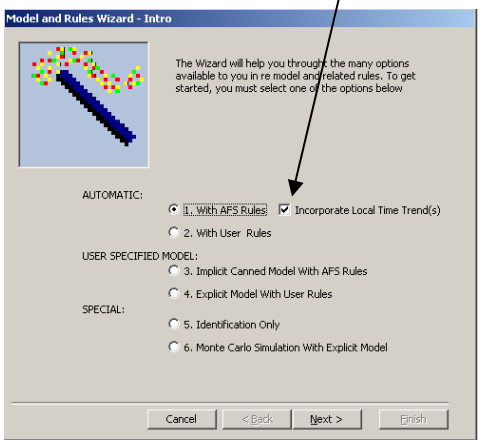


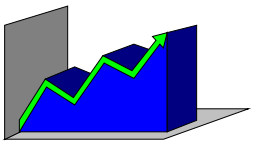
# Overriding the Expert System

We do allow you to override the default conditions for option '1' for some particular items so that you can customize our expert system for your needs. You need to create trigger files in the installation directory and there is existence will cause Autobox to react. Some overrides require data to be in the file. The trigger files need to be named "*.AFS".

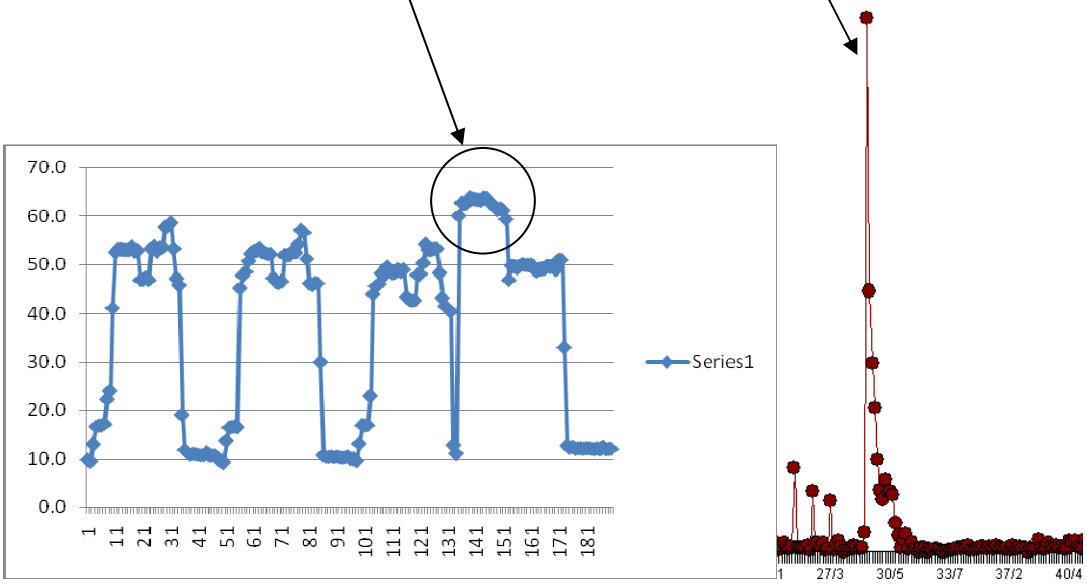
Here are the trigger files in alphabetical order:

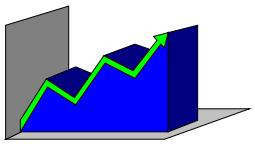
1)Allin.afs – Keeps all causal variables in the model whether they are significant or not.
2)Foreconf.afs – THIS IS NOT JUST A TRIGGER FILE, BUT ALSO HAS CONTENT. This determines the confidence level for the confidence limits (For 80% confidence limits put '80.0')
3)Integer.afs – Converts forecasts to integers
4)Noparcon.afs – Stops the testing & adjustment for constancy of parameters (ie. Removes older data that has a different pattern than the more recent data)
5)Novarcon.afs – Stops the testing & adjustment for constancy of variance (ie. weight the observations based on their variance)
6)Numbfore.afs – THIS IS NOT JUST A TRIGGER FILE, BUT ALSO HAS CONTENT. This defines the number of forecasts (For 24 forecasts put a '24' in this file)
7)Positive.afs – Converts forecasts to positive values
8)Stepupde.afs – THIS IS NOT JUST A TRIGGER FILE, BUT ALSO HAS CONTENT. This defines the number of interventions to be used in the model. (For 5 interventions put a '5' in this file)
9)Additionally, you can uncheck "the local time trend" box if you want Autobox to not look for "local time trends".





There is one trigger based on the name of the series: There is one very special Autobox trick that we want to discuss up front. For example, if you have a "dynamic" promotion over a period of two weeks (and you had daily data) that causes demand to shoot way up and it slowly ramps down back to the mean (decays). If you specify the causal series name with the words "DYN14" for example, Autobox will react by modeling the promotion to decay over the next 14 periods the promotion was running. Note that the data type(SEE THE SECTION ON CREATING YOUR OWN ASC FILE FOR MORE ON DATA TYPE) MUST be '3' for this to work. Also, if you have a "patch of outliers" that are in the same range (ie all zeroes) then you can use a '1' indicator during that patch. If you have a "patch of outliers" that varies wildly (ie high, low, etc.) then use "DYN" and the length and again data type must be equal to '3'.





# The Menu System

File

New-Allows you create a new dataset by entering or pasting the data into a spreadsheet

Export-Save the current spreadsheet into Excel

Import-Import series from Excel worksheet. See “Import Series From Excel” below

Open-Allows you to directly import a dataset from an existing .TXT or .ASC file

Print-Allows you to print spreadsheets and graphs

Save-Save your newly created dataset for permanent storage

Stock Retriever-Allows you to retrieve Stock Quotes to Forecast (this is just for fun ☺)

Recent files-Allows you to open a recently used series

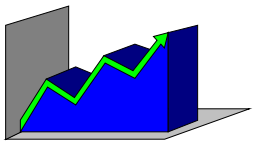
Exit-Exit Autobox

View

Show DataPoint Information-This gives the user the option to show/hide this information on graphs

Preferences

Allows you to change the number of decimals used during the process



Process

Run Autobox - Opens the "Model And Rules Wizard" that helps the user through the selection of execution options, model and rules required to process the data. Allows user to select automatic modeling, user specified model or special processing. In automatic modeling the program will determine the model based on the data supplied; and the user can choose to process with default rules held in memory or provide his own. In user specified modeling the user can select from a list of offered models with default rules held in memory or choose to provide his own model with rules he provides or specified default rules. The user can select to save changed rules, final model or reports, where applicable.

RunWhatIf-Generates forecast data only based on user adjusted future values

Series

Add

Generated Series-Allows user to select intervention series to be included

Level Intervention – Adds a variable with a series of 1's to account for a temporary change in the mean

Pulse Intervention – Adds a variable with one 1 to account for an outlier

Seasonal Pulse Intervention – Adds a variable with a 1 at regular intervals to account for a seasonal event

Trend Intervention – Adds a variable with 1,2,3,4, etc. to account for a trend in the data

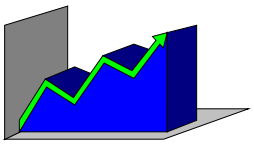
Holidays-Allows user to select holiday series to be included in the dataset as dummy variables

User Defined-Allows user to include his specified type of series

Delete-Allows user to delete any input (Independent) series

Hide columns-Allows user to "hide" rather than delete an input series; and process the remaining dataset (you **can't** change series information in this mode)

Series Information – Allows you to change the data type (0,1,2,3) of a series



Select New Output Series- Allows user to select an input series to replace the original output series and process the dataset to determine effect

Restore Columns-restore any columns that were hidden

Restore Original Output Series

Help

Contents

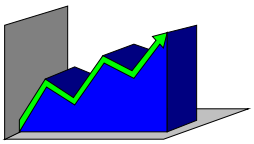
Search for Help on-Allows you use Index or Find options

User Guide-Open's this document

Autobox.com-Sends you to our website

About Professional-Sends you to website to see overview, model description and pricing

About Autobox -Tells you the version of the program



# These Are The Tabs Seen In The Program

There are menu items which we have already reviewed and there are folder "tabs" that allow you to view the data, graphs, reports, etc.

Historical Data - Shows you the past of the series to be forecasted. You can right click on the spreadsheet which generates a pop-up menu with Copy, Paste, Hide Column, and Restore Columns options. The Copy and Paste options work just like a clipboard. The Hide Column and Restore Columns are unique features that allows the user to hide a column(s) and thus excluding it from the process without having to delete it. Any Input (independent) series can be hidden and restored; however, the Output (dependent) Series cannot. To hide a column(s) left click on the header(name)cell while holding down the "Ctrl" key; and, when you have finished selecting columns, right click on the spreadsheet and select the Hide Column option. To restore the columns, merely right click on the spreadsheet and select the Restore Columns option. Please note that if any of the hidden columns have related columns in the Future Values and Retained Data spreadsheets, they will automatically be hidden/restored.

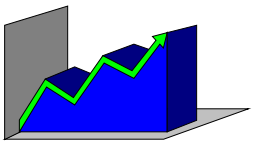
Future Values-Shows the anticipated future values for the specified series. You can manually enter in or paste future values.

Retained Data-Shows the data retained from the end of the specified series. If you haven't already loaded retained data via an .ASC file, you need to reduce the number of observations in the series properties box to generate them. This is for testing "out of sample error" such as MAPE comparisons using different lead time periods.

Auxiliaries - Shows the Actual/Fit data, Actual/Forecast data and the Forecasts if generated for the series that was modeled.

Graph- As you open each graph in Autobox, it will then be generated and saved in the installation directory as a jpg file. Note that as the vertical bar passes over a data point the information is indicated. When an outlier is indicated, the text describes the type of outlier and the magnitude of it's effect. We use a "P" for a pulse which is a one-time intervention which Autobox identifies and corrects for in the model itself (i.e. dummy variable). A "S" indicates a pulse that occurs every season. A "L" indicates that there has been a level change in the mean of the series a "SP" indicates a seasonal pulse.

Please note that we have added a new graph called "Adjustable Forecasts" which allows the user



to “tweak” the forecasts for presentation purposes based on judgment outside the modeling process. Just left click on a data point and while holding down the mouse button move the point up or down as desired. When finished, select the “Save Changes” button and the changes will be made to the Forecasts spreadsheet and the other graphs. You must click on each of the graphs to save the new graphs to a jpg file. Click the Restore Original Forecasts button and the Forecasts spreadsheet and graphs will be reverted. Again you would have to click on each graph to save them to a jpg file. If you select to use this graph and change the forecasts, you will be precluded from using the WhatIf function of the program.

Reports

Details.htm – Tracks the steps and decisions to create the model

Intrvent.htm – Lists all of the detailed information on interventions

Rhside.txt – Lists how each of the variables numerically contribute to the forecast.

Stat.htm – Shows the statistical fitting statistics (RMSE, AIC, etc.) and model with P-values

Verbal.txt – The model explained in “English”

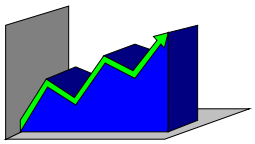
Ab50pro.123 – The forecasts with upper and lower limits

Forecast.csv – The forecasts in Excel format

Equation.txt – Reporting only the equation

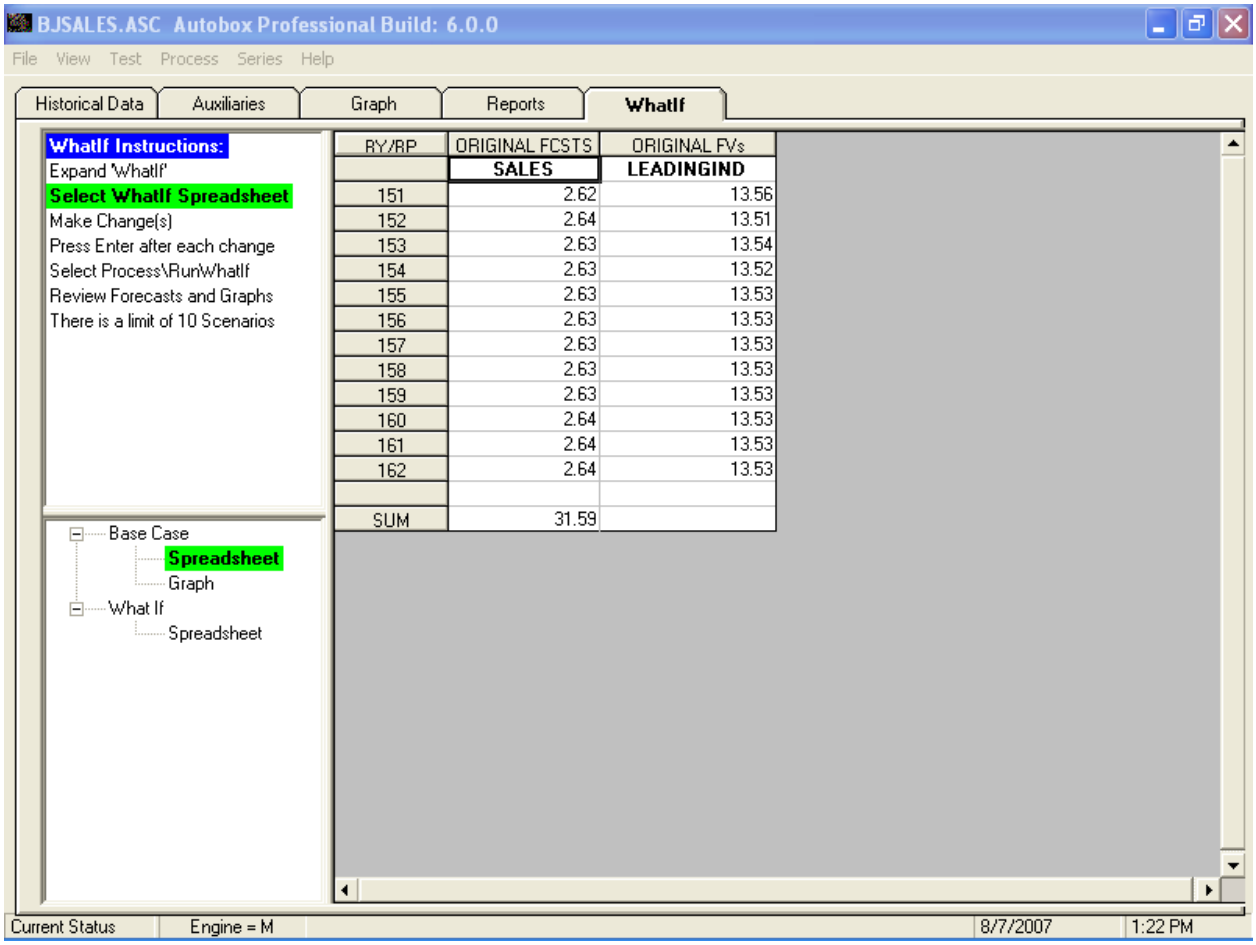
Equation.csv – The equation ion Excel format

Fitted.csv – The fitted values in Excel format

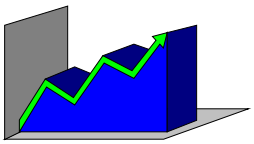


WhatIf

This Tab appears for Causal Models only. It allows the user to change values to determine their effect. When the Tab is selected, the Base Case spreadsheet is shown which indicates the original forecasts generated and the original future values for each input series included in the model. The upper left pane includes the basic instructions and, as you proceed, the next step is highlighted to assist you. The lower left pane includes the options to be selected. The options expand as the process continues.




To begin the WhatIf process, you must select the WhatIf spreadsheet, click on the cell you wish to edit, make your change and press Enter. When you finished making desired changes, select Process/RunWhatIf from the main menu. When processing is completed, the Forecasts spreadsheet is shown indicating the original forecasts and the forecasts for the current scenario. A



graph of the current WhatIf forecasts is also available.

At this time you have the option to select the WhatIf spreadsheet again and create another scenario or select Restore Values to revert to the original future values. Please note that there is a limit of 10 scenarios.

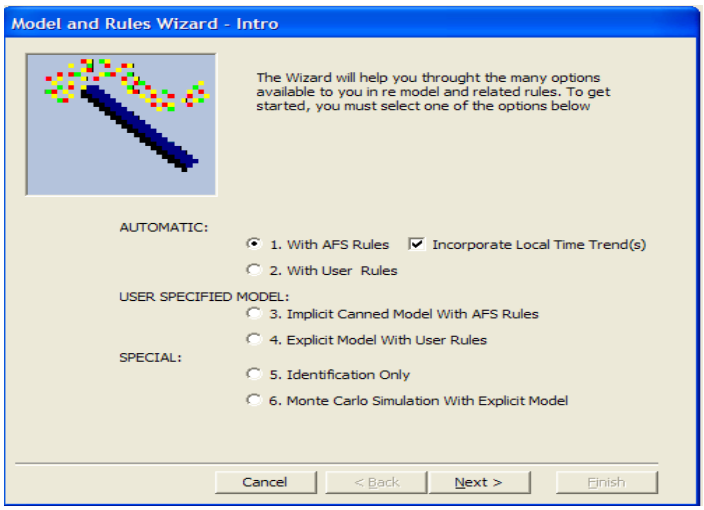
After each scenario is processed, the forecasts, graph data and specified values(changed values) are retained; and you can review the combined scenarios by expanding the Review All Scenarios option.

NOTE: If you have a series that has had a change in variance then no "What-if" option is displayed. It is not available in this release.

# Model And Rules Wizard

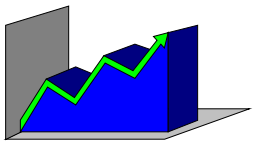
As we indicated above, select "Process\Run Autobox" and the wizard will appear as follows with the default set to option 1. This is because option '1' uses our heuristics. Option '2' is also automatic, but you can make many adjustments to the process.

Option '3' lets you pick a typical model (i.e. holt winters, etc.) and run, but we use this more for more benchmarking than for analysis. Option '4' allows you to specify your model for the dataset. This again could be a good comparison to what Autobox would report using option '1'. Option '5' allows you to run the initial step of the modeling process which is helpful for those trying to learn Box-Jenkins. Option '6' allows you to specify a model and simulate a data series. Again, this can again be a great tool for beginners who can then see if they can identify the model using option '5' or run option '1' and see if Autobox can do it too!



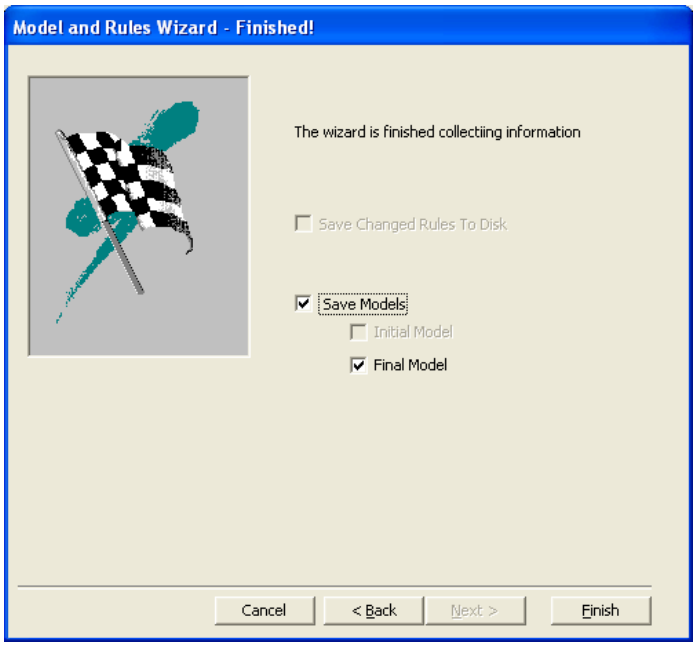

As the wizard indicates, your first need to choose what you want to do.

Option 1 – Run using proprietary AFS rules
Option 2 – Run using User's overrides of AFS rules
Option 3 – Choose a specific model like Holt Winters – it's only there for comparison in our opinion!
Option 4 – Choose your own model – this is for experts…be careful here
Option 5 – Runs only the identification process
Option 6 – Simulation - Create a series based on a model – Good for learning Box-Jenkins because you can then run Option 1 and see if Autobox can identify the model. Option 5 would

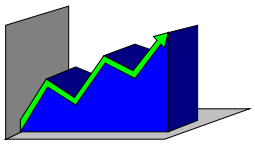


also be good to run with as you can for yourself try and review the Autocorrelation Function and Partial Autocorrelation Function for lag relationships. If you select optional and press “Next”, you will have the option to save the final model as indicated on the following:

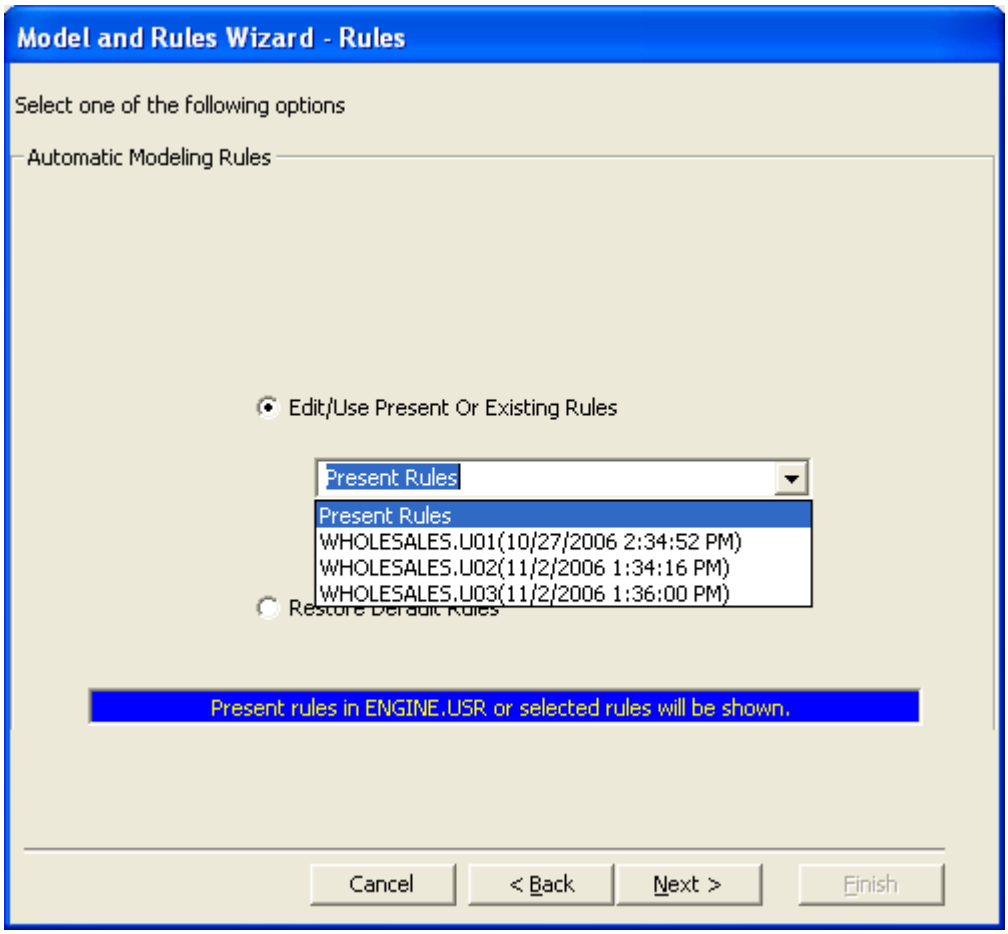
You can choose to save your rules and models for later use.




Click “Finish” and your dataset will be processed automatically.



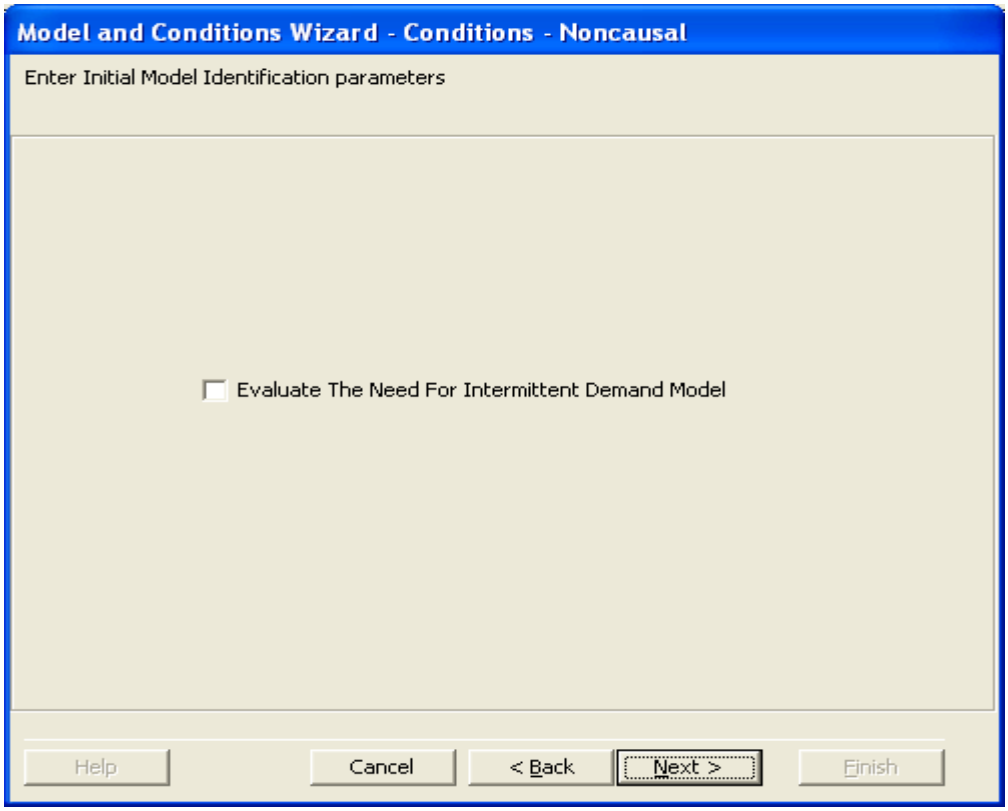
If you select option ‘2’ and press “Next”, you will be given the option to use the “Edit/Use Present or Existing Rules” (press the down arrow to be shown what is available). This will use the current rules or you select rules you had saved previously. If you want to reset the conditions back to the original factory installed defaults used choose “Restore Default Rules” and click “Next”.

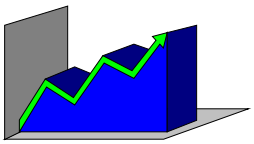


When you select the rules and press “Next”, they will be shown in the following order for your review and/or editing (please note that the rules vary between Noncausal(single) series and Causal(dependent series with independent series)series –we will show the causal screens after this section.

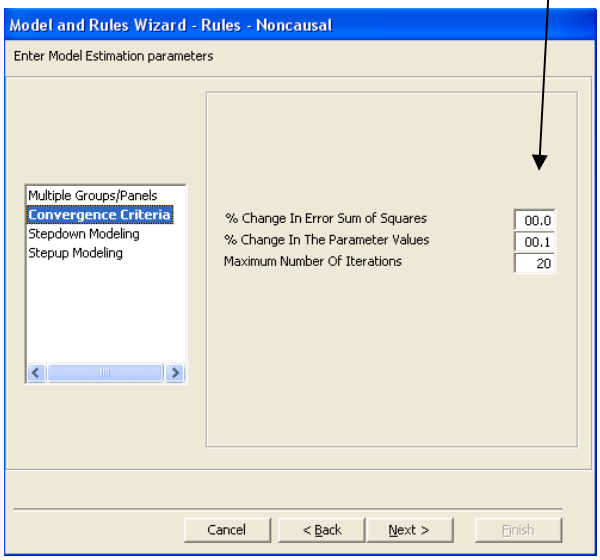
If you select this option, Autobox will check to see if more than 50% of the observations are zero and then run the following scheme to predict the intermittent demand.

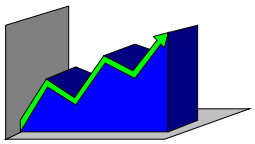
Two new time series will be created: the interval and the rate. The interval series will be the number of periods with zero demand between periods where there is demand (i.e. If there is demand at every time period then the interval would be 1). The rate is the interval divided by the demand. We find this as a better alternative to Croston’s method.





We will skip “Multiple Groups/Panels for now. It is discussed later. If you choose to change the “convergence criteria”, you really won’t need to make a change to the defaults. You can change the way Autobox iteratively goes about identifying the model. The % change in error sum of squares is set to 0 so you can’t be more precise. The % change in parameter values are set to .1% and you can increase the iterations to a maximum of 200, but it doesn’t make much of a difference.

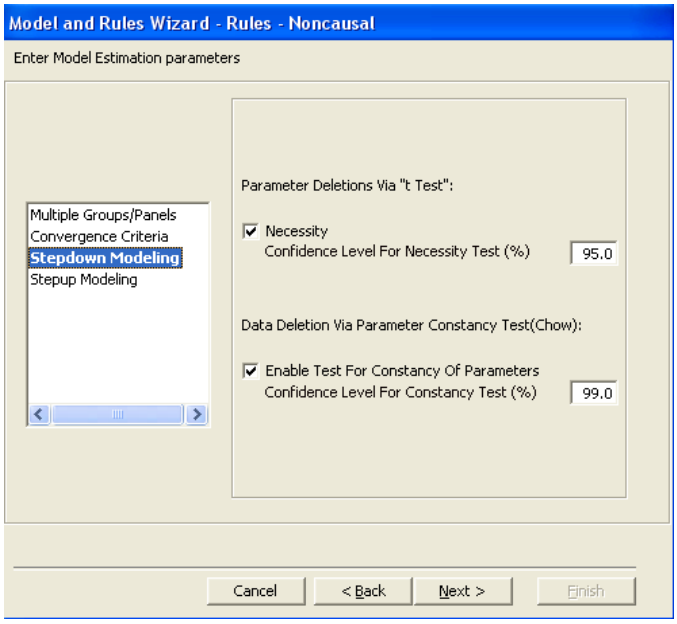


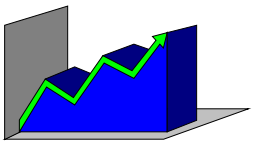


For stepdown modeling (where variables are removed or weighted), you can either adjust the confidence level for the option or "uncheck" the box so that the test is disabled. If these are checked then:

**Necessity -** This will drop all variables that are introduced by Autobox (i.e. outliers). (The opposite case is that it keeps all variables)

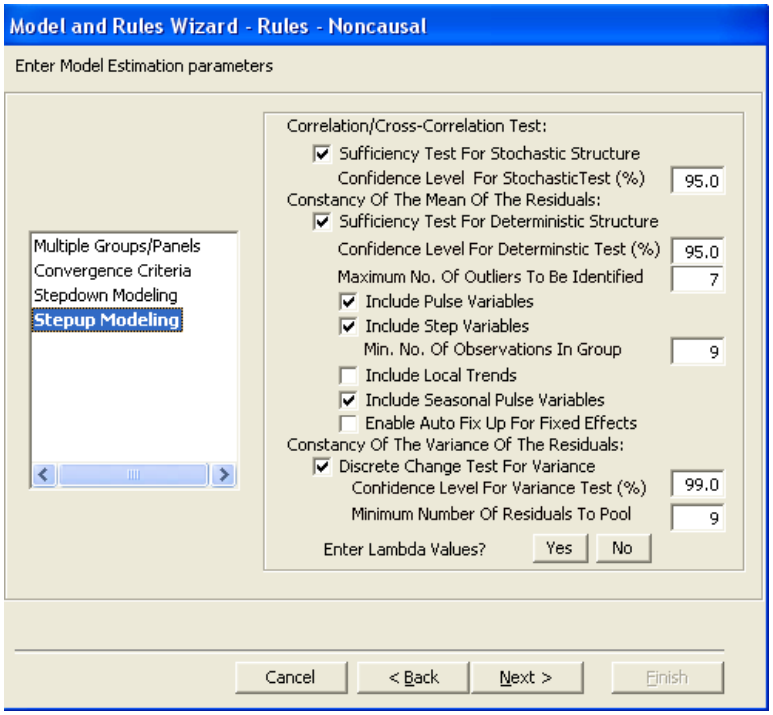
**Test for Constancy of Parameters -** Autobox will determine if you have too much historical data that does not match the more recent data and delete older data.

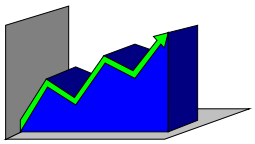




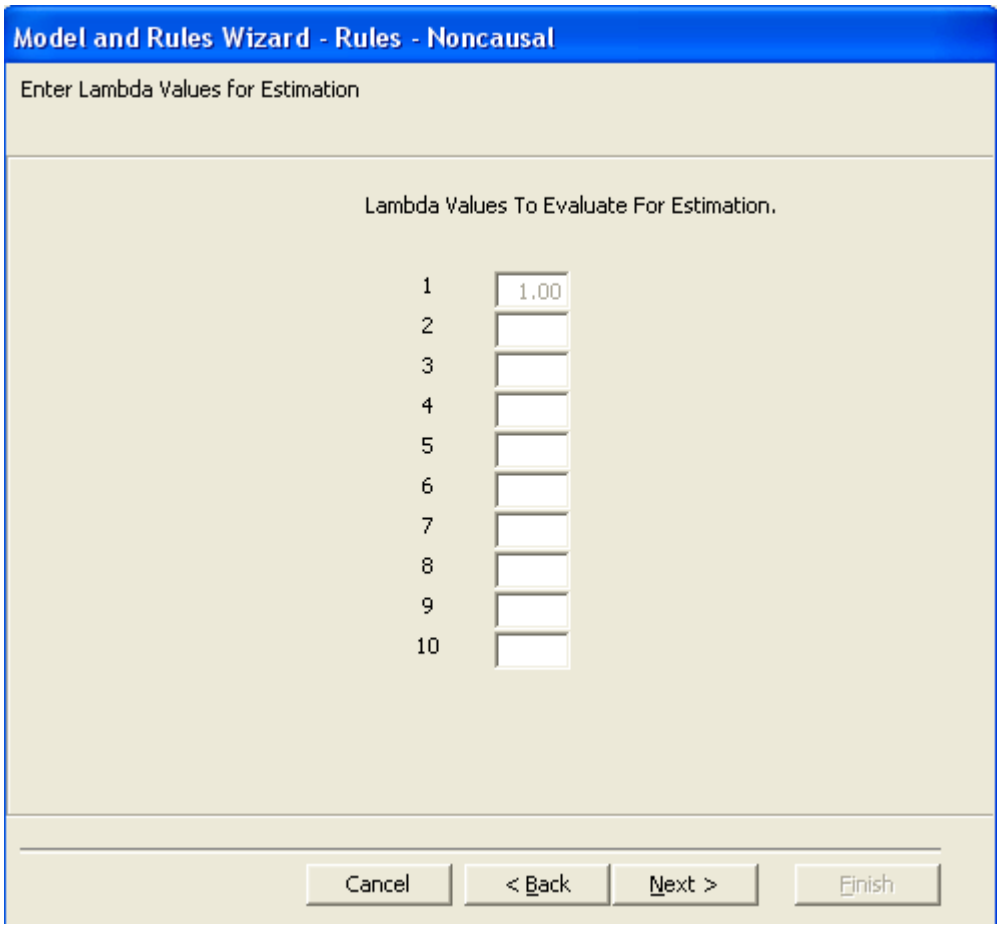
For stepup modeling (where variables are added into the model by Autobox), you can either adjust the confidence level for the option, "check" or "uncheck" the box, or modify the counts.

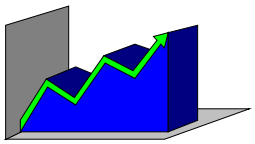
**Sufficiency Test for Stochastic Structure-** This will add in ARIMA variables if significant.
**Sufficiency Test for Determinstic Structure –** This will add in Intervention variables.
**Maximum No. Of Outliers To Be Identified –** This will limit the number of outliers that can be introduced.
**Include Pulse Variables –** This will allow one-time unusual variables to be accounted for.
**Include Step Variables –** This will allow sudden changes in mean to be accounted for.
**Min. no. Of Observations In Group –** This determines how many observations you need before you can create a step variable.
**Include Local Trends –** This looks for changes in intercept.
**Include Seasonal Pulse Variables –** This looks for pulses that occur every season.
**Enable Auto Fix Up For Fixed Effects –** For example, for monthly data, Autobox would include 11 dummies for monthly data to account for fixed effects.
**Discrete Change Test For Variance –** This will search for changes in variance and weight the history based on the observations.
**Minimum number of Residuals to Pool –** There are two groups created to compare the variance and this determines the smallest size of one of the groups.



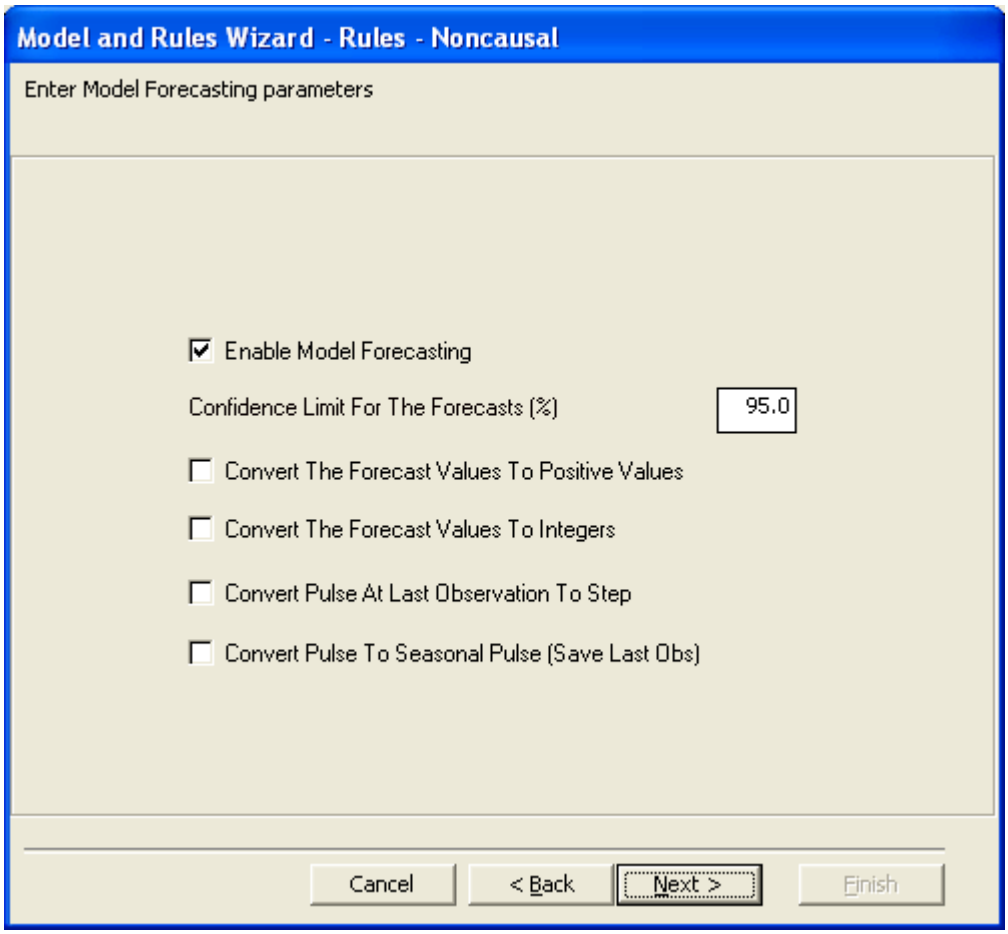


The bottom option is the “Enter Lamda Values”. This allows you to determine the form of the Box-Cox Transformation. The first observation is always 1 as a rule.



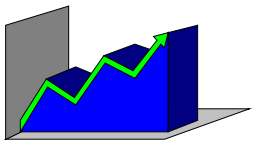


Click “Next” and the “Model Forecasting” parameters will be shown as follows(Please note that to get an explanation/definition of a particular checkbox or textbox merely click on it and press F1):

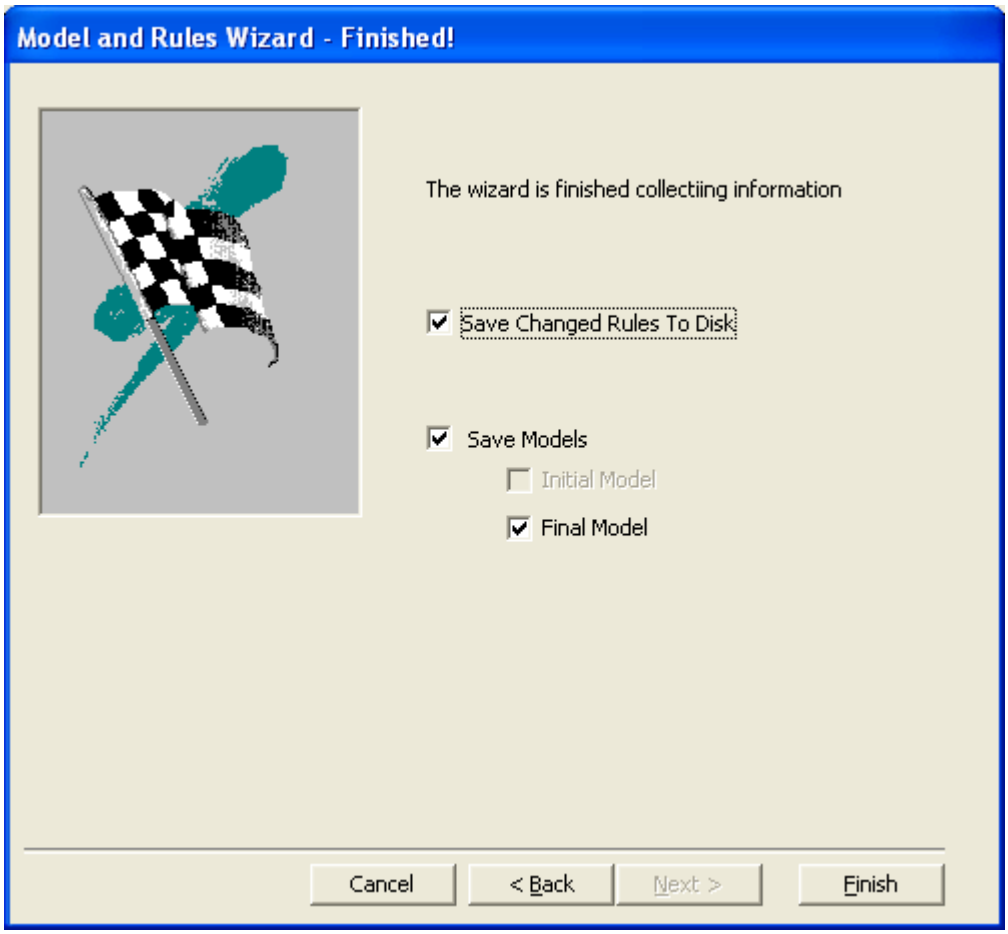



This is the process of using the model that best describes the past observations (to extrapolate the pattern of the data) to predict Confidence Levels into the future.

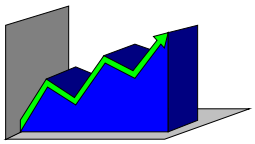
If this option is not enabled, then no forecasts will be made.



Press “Next” and you will be given the option to save changed rule and/or the final model as indicated in the following:



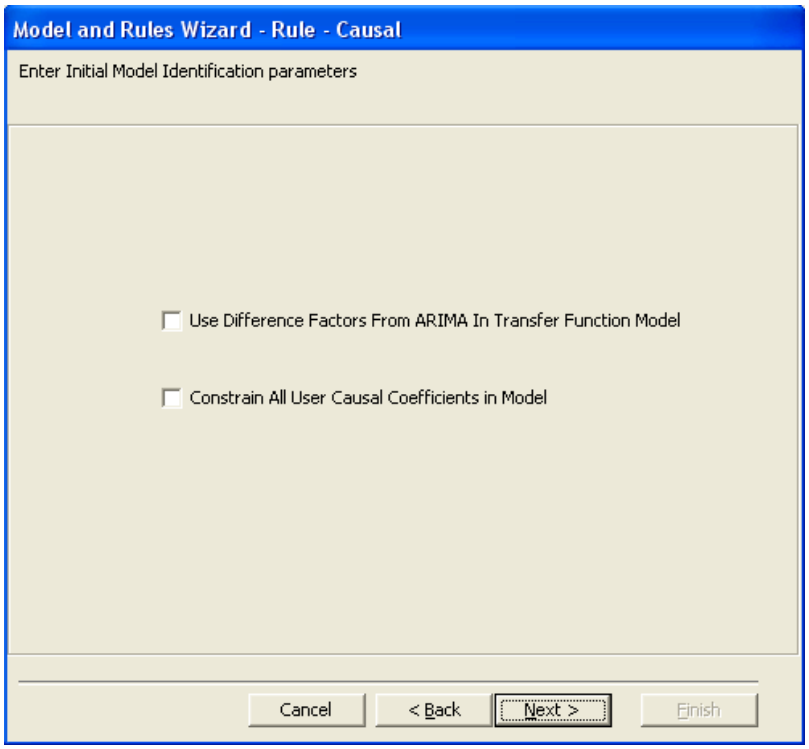

Again, press “Finish” and the dataset will be processed automatically.

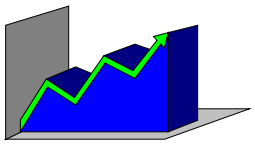


If you selected option ‘2’ and had causal variables then this additional screen would have been presented to you at the beginning.

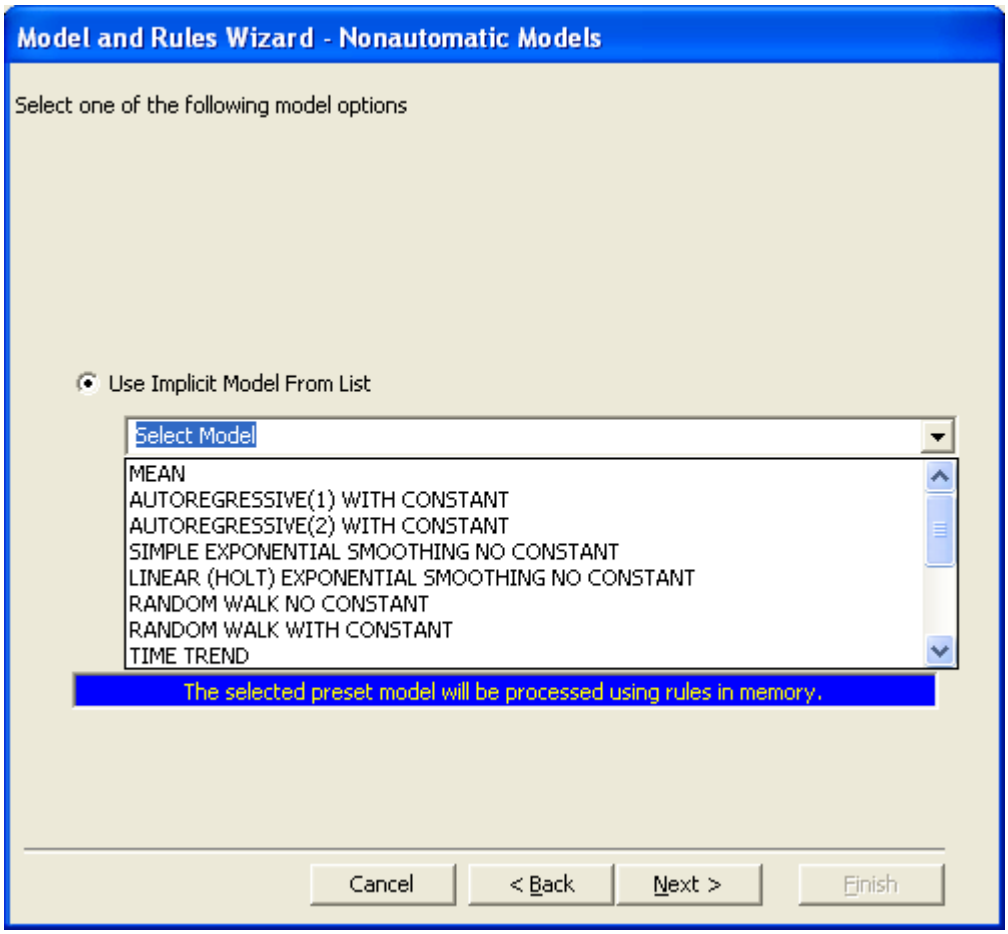
Use Difference Factors From ARIMA Model In Causal Model – This might be helpful in identification and problematic in estimation.

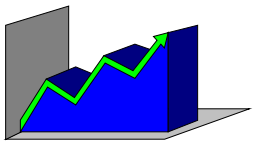
Constrain All User Causal Coefficients in Model – This keeps all user causal variables in model regardless of their significance.



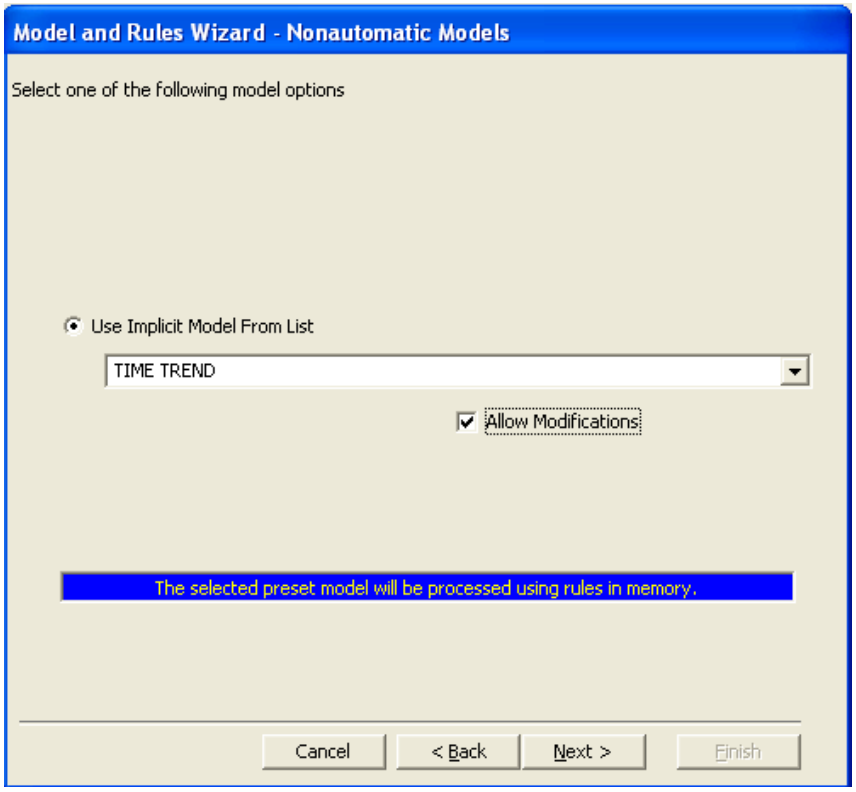


If you select option ‘3’ and press “Next”, you must select one of the offered models as indicated in the following:

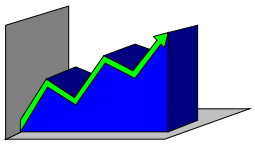




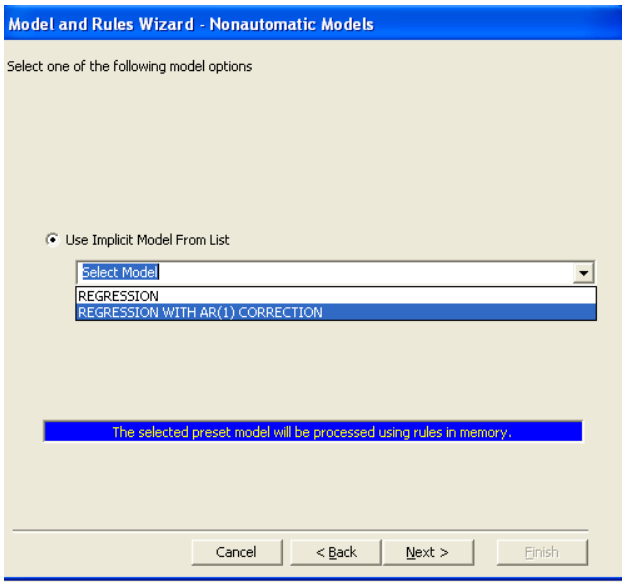
If you click the box “Allow Modifications” then Autobox will use all of its “expert system” approaches to try and create a better model given you picked a starting model like “Time Trend” as seen here. Note that we believe this is dangerous territory as all of the models in this area are typically a bad place to start.



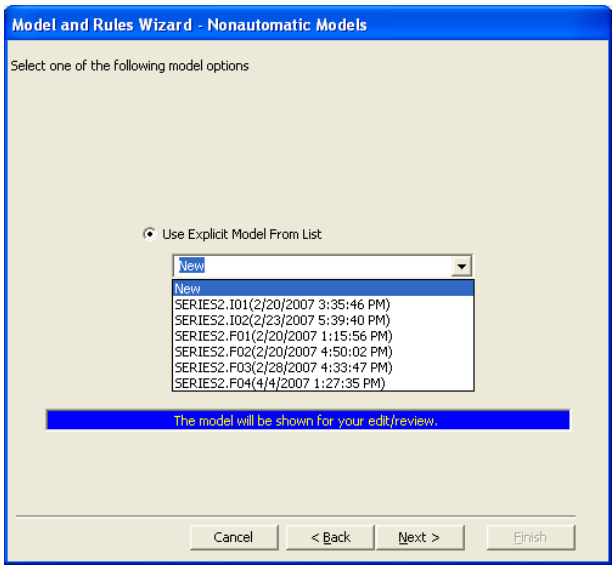

Press “Next” and you will have the option to save the final rules and model, then press “Finish” to process the dataset.

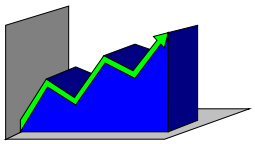


If you have causals then you have two “pre-selected” models to choose from along with the ability to choose “Allow Modifications”. Again, this is dangerous territory as opposed to using option ‘1’.



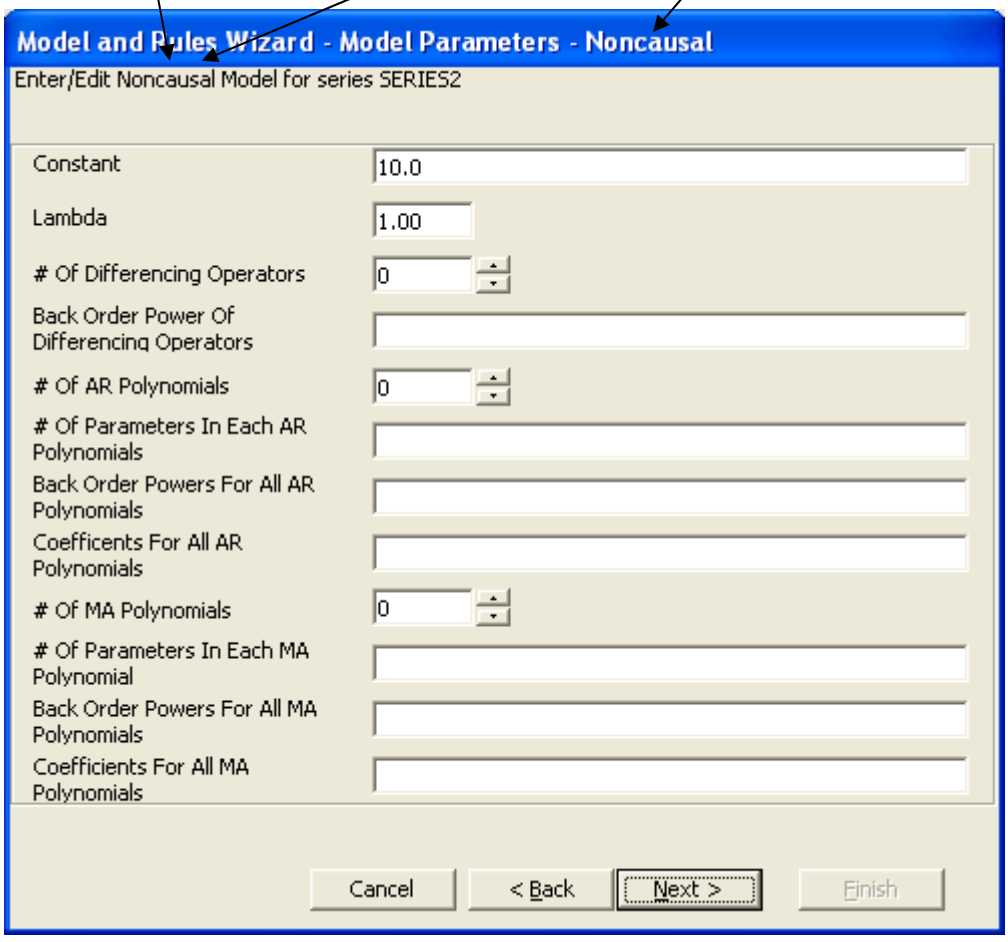

If you select option ‘4’ and press “Next”, you can choose to make a “New” model or use one of the saved models indicated in the following. Click “Next”

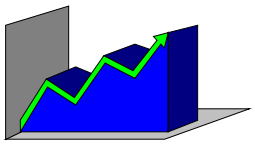




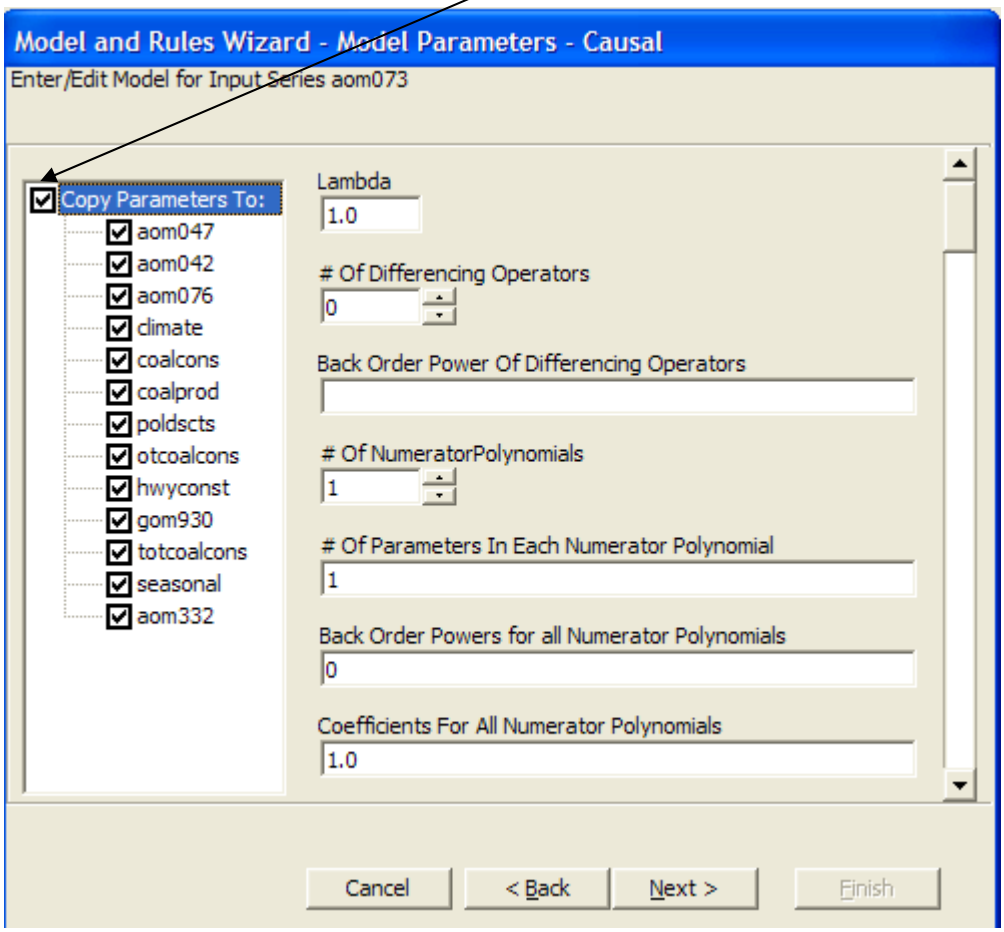
If you only have one series in your dataset you see the words “Noncausal” twice at the top of the screen.

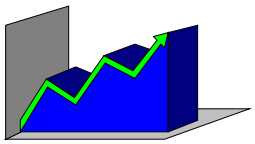
If you only have multiple series(not shown) in your dataset you see the words “Causal” at the top and the word “noise” here. The noise refers to the ARIMA structure in the transfer function model. We will show you a causal example on the next screen.



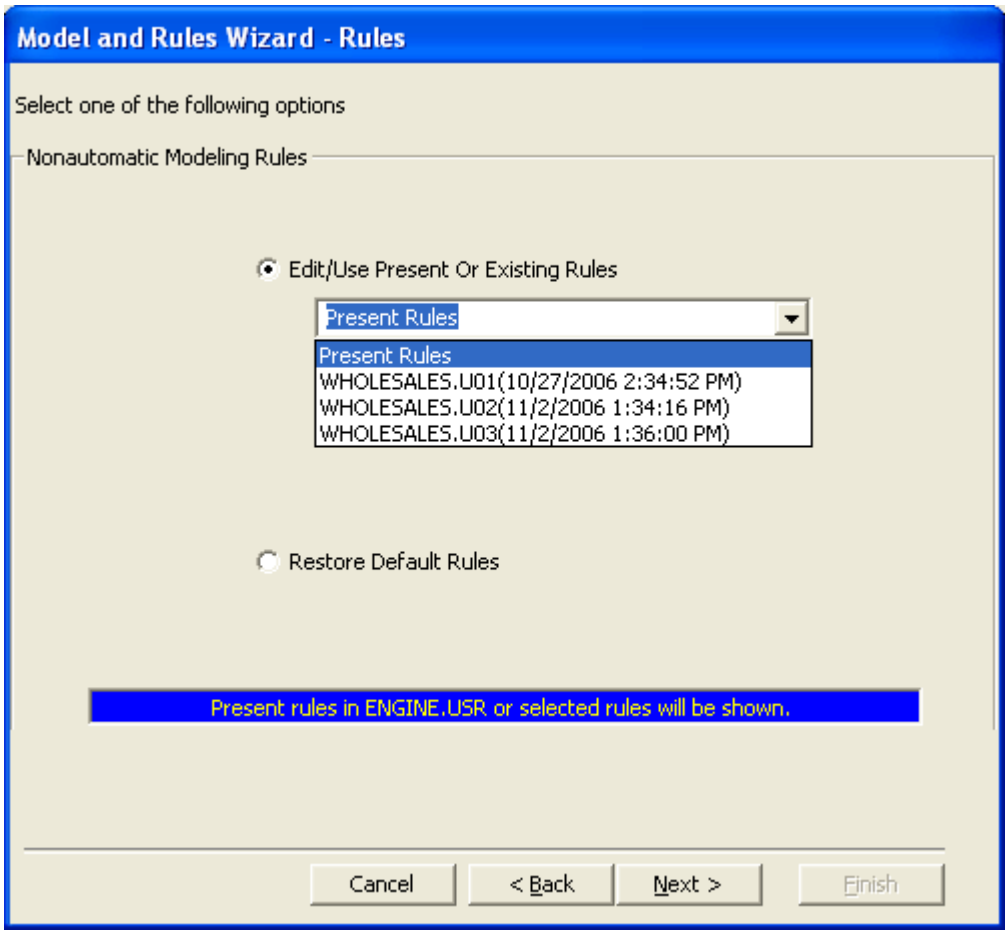


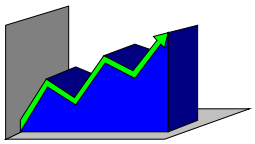
Here we specify the model for aom73, but we click here and then click on all of the other causal series so that they have the same model applied.



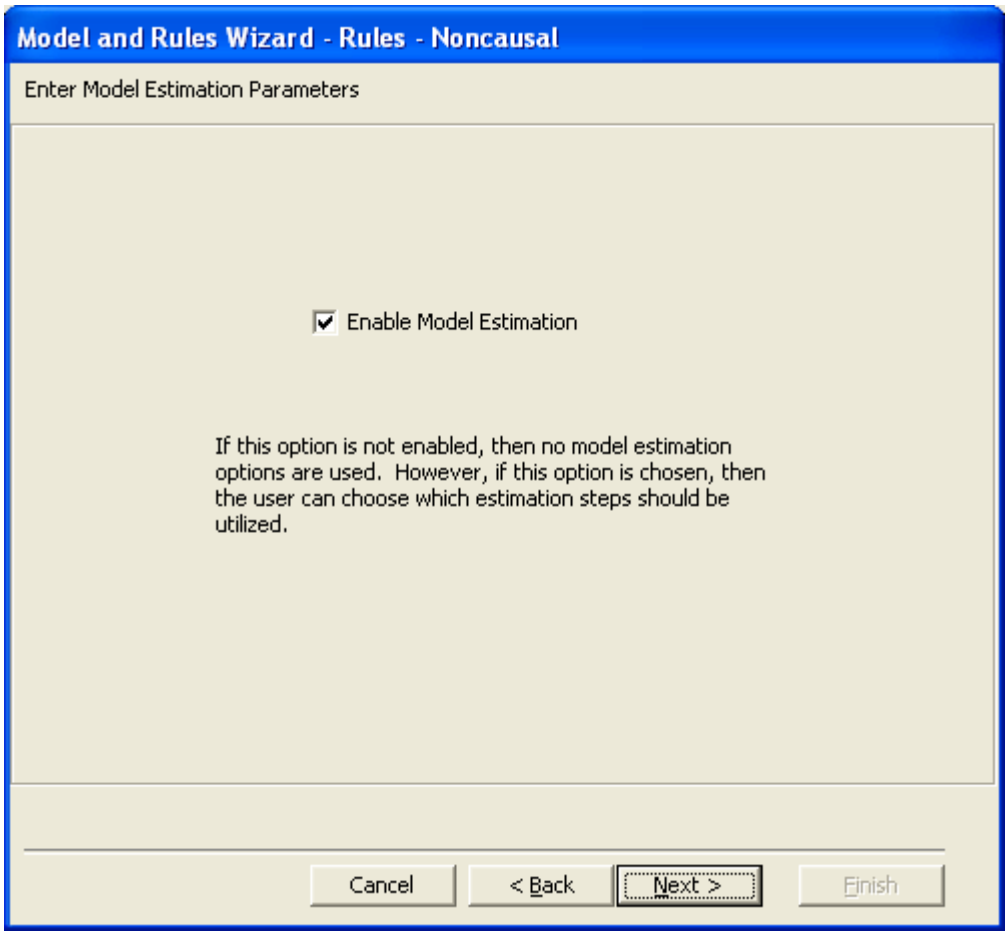


Press “Next” and you will be given the option to use the “Edit/Use Present Or Existing Rules” (press the down arrow to be shown what is available) or “Restore Default Rules” and as follows:

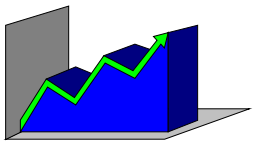




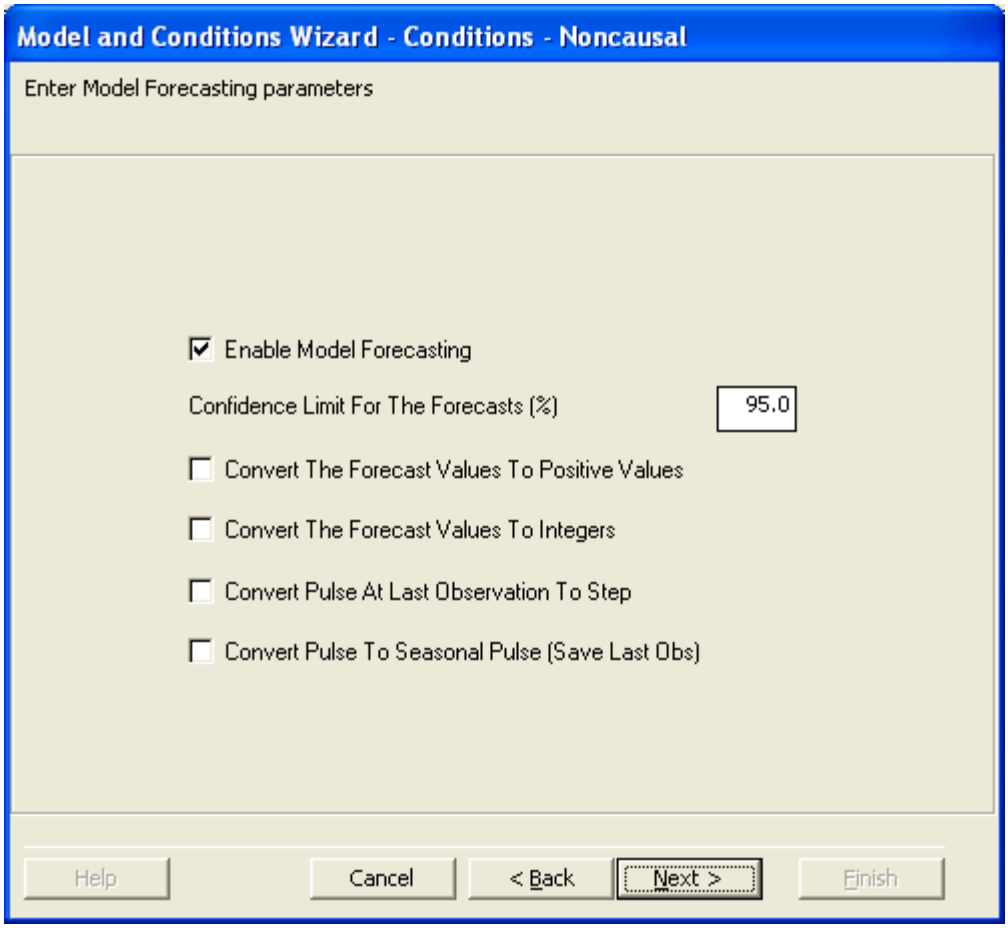
Press “Next” and the rules will be shown as follows for your review and/or edit:




Estimation and diagnostic checking represent the second phase of the Box-Jenkins modeling procedure. The estimation option computes the model coefficients via nonlinear least squares and generates residuals. Various statistics are computed for both the estimated parameters and the residuals from the model.



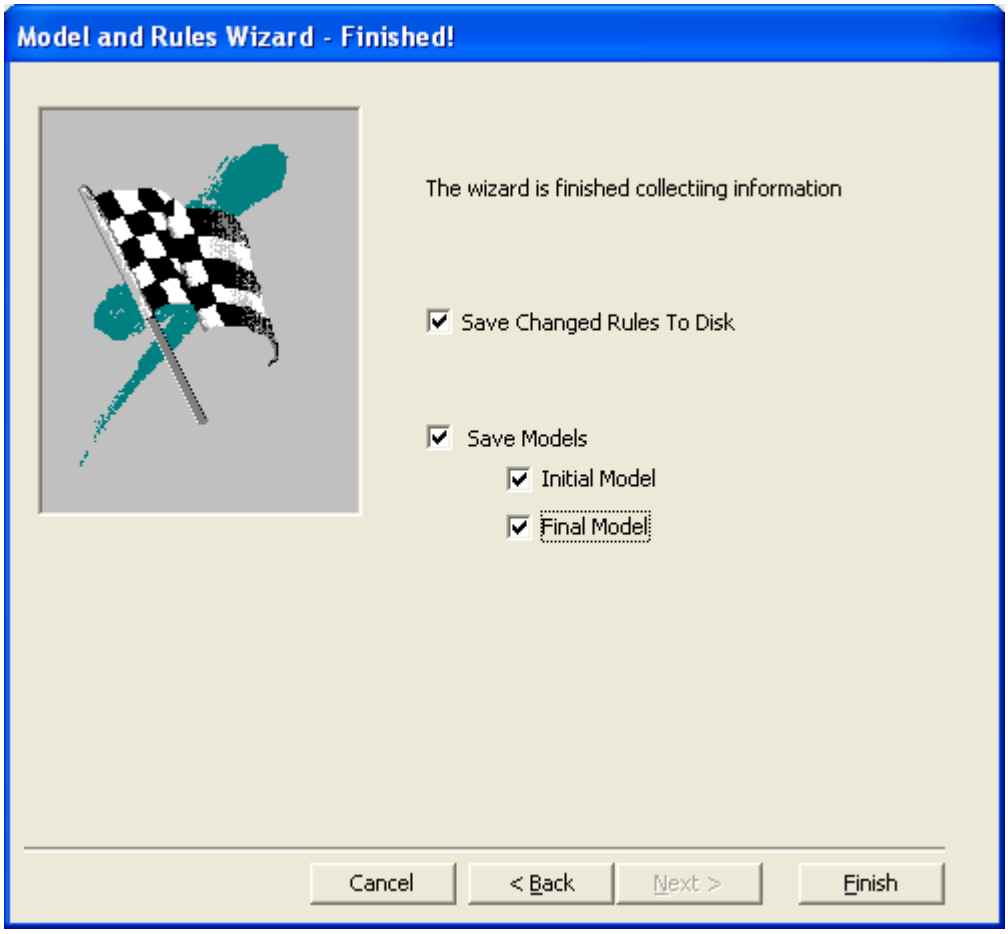
Press “Next” and the “Model Forecasting” parameters will be shown as follows:




This is the process of using the model that best describes the past observations (to extrapolate the pattern of the data) to predict Confidence Levels into the future.

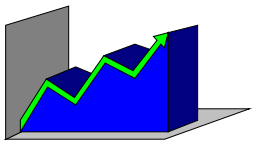
If this option is not enabled, then no forecasts will be made.

Press “Next” and you may elect to save changed conditions and/or the “Initial Model” and/or the “Final Model” as shown in the following”

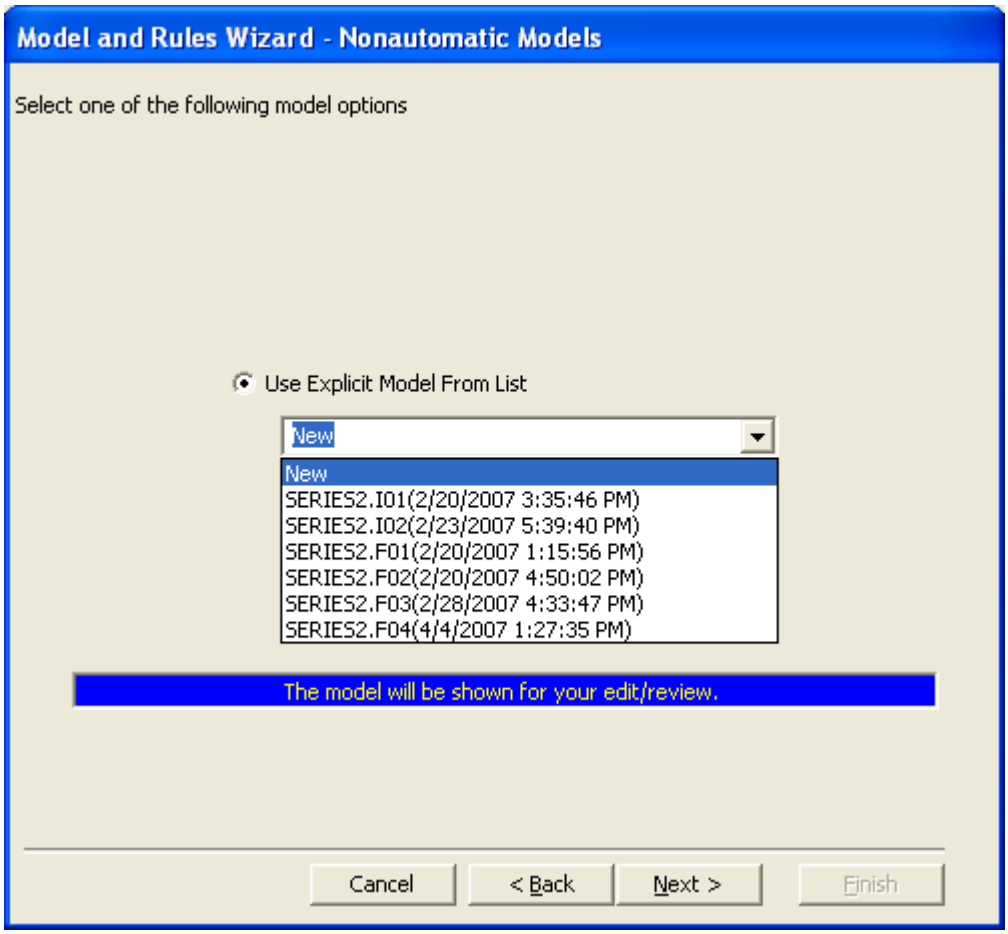


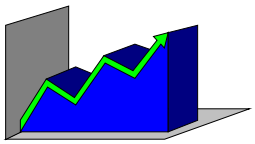

Again, press “Finish” to process the dataset.

If you select option5, press “Next” then “Finish” to process the dataset. Autobox will only perform identification. This is good for beginners just trying to learn Box-Jenkins to see if they can identify the model on their own.

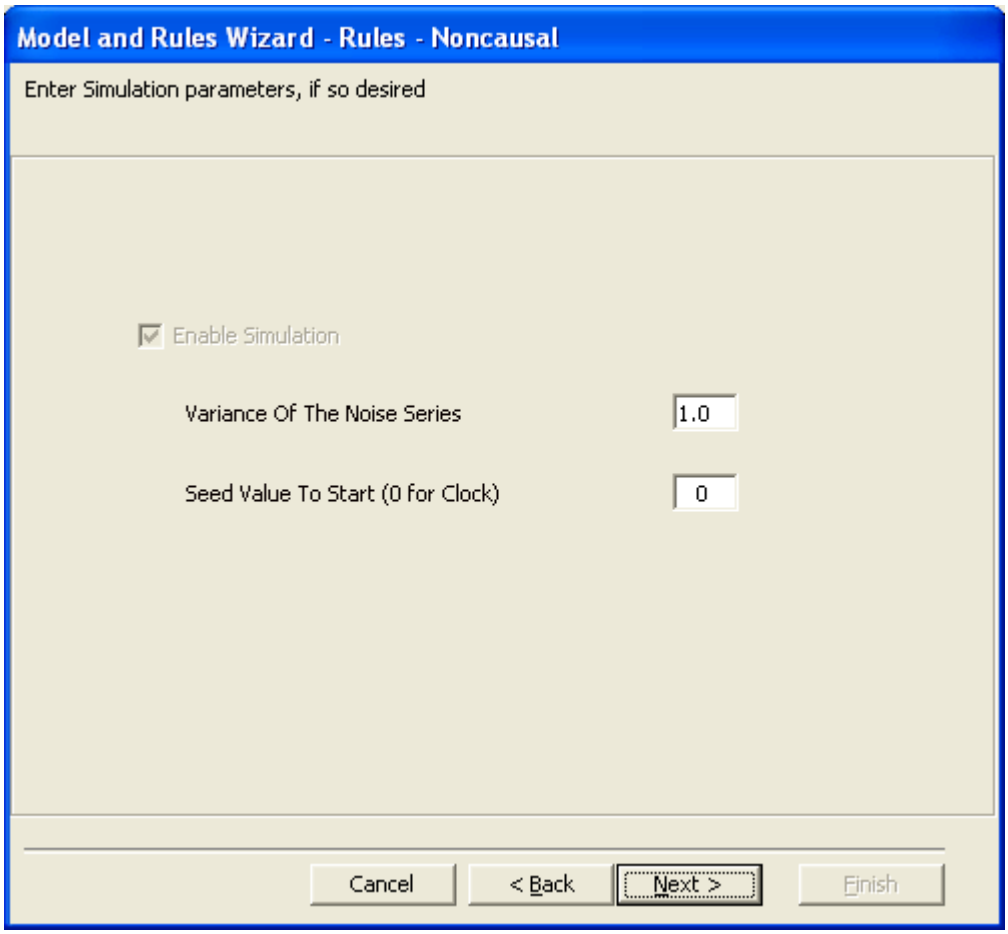


If you want to simulate choose option ‘6’ and press “Next”, you will be given the option to provide a “New” model or select a saved model as indicated in the following. You can create your own model:



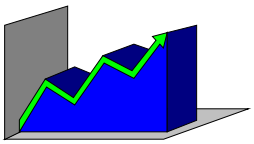


Press “Next” and you will be shown the model parameters for you review or edit. Press “Next” again and you will be given the option to edit the simulation parameters as follows:




Press “Next” and you may elect to save the changed rules then press “Finish” to process the dataset. A file named “SIMDATA.ASC” is created in the program directory (i.e. c:\autobox). Open this series to view the series that was created. The simulated series will have 300 observations.

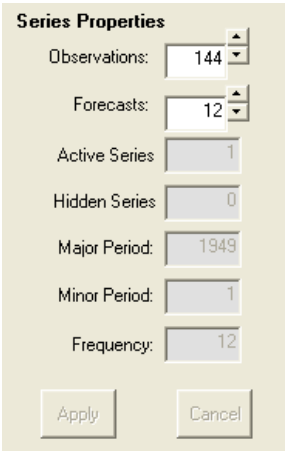
Simulation is the Autobox component that allows for the simulation or artificial creation of one or more stochastic time series. The time series are generated from a particular model form. The model form is specified by the user. The model forms allowed are subsets of the general class of time series models described by Box and Jenkins (1976) and Tiao and Box (1981). The simulation starts by generating a random process for each series in the equation. These random



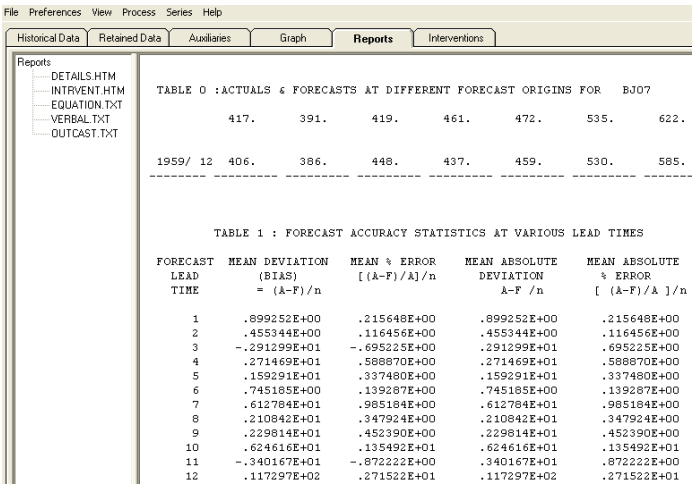
numbers are generated from a normal distribution with mean zero and variance. The larger the underlying variance of the noise process, the greater the background noise of the simulated data. The program allows for the simulation of stochastic time series that are modeled by univariate BOX-JENKINS (ARIMA) equations. It also allows for the simulation of stochastic time series that are part of a Causal Model equation. In addition, the Causal Model equations may contain deterministic variables. These are the explicit model forms that are available. What this means is that you can specify model forms that are not usually stated as Box-Jenkins models; you need only to re-state the model in a BOX-JENKINS format. In other words, since the general class of models described as Box-Jenkins models implicitly includes the more "popular" time series model forms, you can indirectly simulate these models also. For example, a simple moving average, a weighted moving average, an exponential smoothing, a regression model and many more models can be re-stated as Box-Jenkins models.

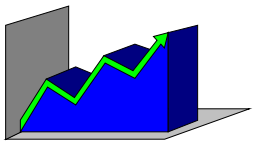
# Perform Error Analysis

If you are interested in analyzing forecasting performance you need to change the number of observations in your series properties (change it from 144 to 132 to see a 12 period out analysis for example).




Click “apply” and then choose “Delete observations from the end of your series” and click “OK” and choose “Use the Deleted Observations as “Retained Data”. The Number of Forecasts will be changed to Equal the number of retained data. Then choose “OK” and “OK” and then “Process/Run” and a report will be generated named “outcast.txt”.





# Case Study with Causals– Let's show you some data issues that we can have had to deal with some of our clients!

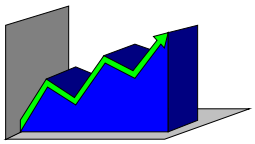
We have experienced clients who make mistakes when trying to perform analysis. The goal of this exercise twofold in that we want to show you how to import the data and get Autobox to do what you want it to using proper analysis.

For example, we get clients who send us data that have no data (see the variable X1 on the next page) for the first half of the dataset as they did not have this data and then last half is real data. This is incorrect as if you use all of the data then you are sure to have a variable that won't be found to be significant as your data preparation was messy. So, let's walk you through our example to show you how to do analysis correctly.

Let's import the file the client sent us (with errors and all!- AFS20.XLS…you can find it in your installation directory if you want to follow along).

We will first discuss the errors and then import it into Autobox and show you how to fix them!

Note that the last 4 observations have the Y withheld to test the accuracy of Autobox.



Problems:

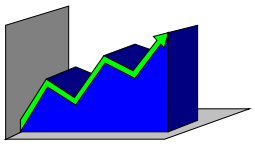
There are 80 months of data (we show only an excerpt here). The causal variable X1 has missing data. All of the historical data from Y1M01 to Y3M12 should be dropped.

Seasonal dummies (11 of them for each month of the year) are delivered (see series M1 through M11). If you run Autobox and only a few seasonal dummies are in the final model, then maybe you shouldn't have used seasonal dummies. This is an example of "MSB" Model Specification Bias where the user makes assumptions that may be incorrect. So, if you don't include seasonal dummies then Autobox might include an AR12 variable. The point is that you should run it with and without and compare the "Adjusted Variance" reported by Autobox to determine which is the better model.

X4 is a linearly redundant variable as it is all zeros and should be removed.

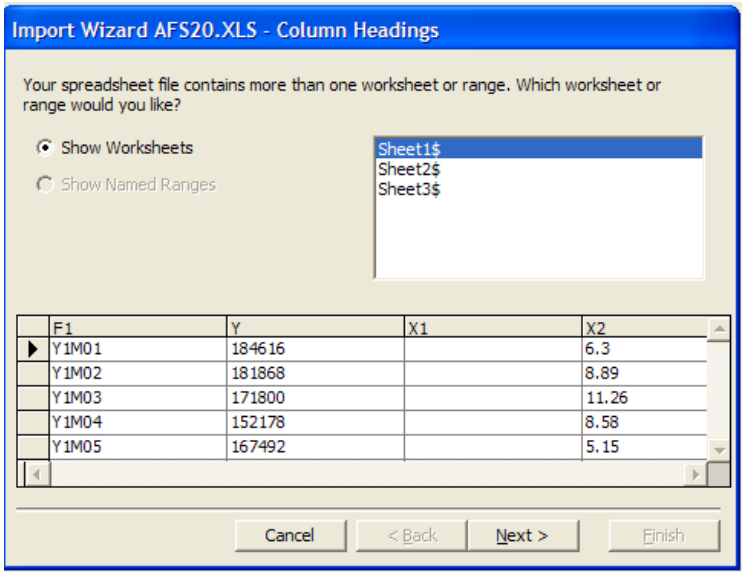
Y is large and might need to be scaled by dividing by 100,000. (We tested this out and did not show it, but it after scaling returned the same results). This can be important as estimation of the model is based on % change and large numbers return small % changes.

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Y | X1 | X2 | X3 | X4 | X5 | X6 | M1 | M2 | M3 | M4 | M5 |
| Y1M01 | 184616 | | 6.3 | 110.32 | 0 | 121.69 | 163.47 | 1 | 0 | 0 | 0 | 0 |
| Y1M02 | 181868 | | 8.89 | 175.95 | 0 | 131.56 | 223.04 | 0 | 1 | 0 | 0 | 0 |
| Y1M03 | 171800 | | 11.26 | 17.2 | 0 | 129.11 | 65.83 | 0 | 0 | 1 | 0 | 0 |
| Y1M04 | 152178 | | 8.58 | 36.21 | 0 | 142.39 | 27.38 | 0 | 0 | 0 | 1 | 0 |
| Y1M05 | 167492 | | 5.15 | 48.08 | 0 | 138.8 | 31.32 | 0 | 0 | 0 | 0 | 1 |
| Y1M06 | 120123 | | 8.87 | 42.67 | 0 | 131.85 | 15.69 | 0 | 0 | 0 | 0 | 0 |
| Y1M07 | 164718 | | 6.3 | 82.65 | 0 | 154.07 | 11.31 | 0 | 0 | 0 | 0 | 0 |
| Y1M08 | 156417 | | 8.89 | 46.21 | 0 | 179.84 | 14.18 | 0 | 0 | 0 | 0 | 0 |
| Y1M09 | 108093 | | 11.26 | 0 | 0 | 202.45 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y1M10 | 152192 | | 3.62 | 118.28 | 0 | 193.68 | 91.4 | 0 | 0 | 0 | 0 | 0 |
| Y1M11 | 121089 | | 0 | 161.32 | 0 | 190.07 | 115.48 | 0 | 0 | 0 | 0 | 0 |
| Y1M12 | 156689 | | 0 | 131.01 | 0 | 191.12 | 68.68 | 0 | 0 | 0 | 0 | 0 |
| Y2M01 | 157451 | | 0 | 27.38 | | 220.38 | 110.32 | 1 | 0 | 0 | 0 | 0 |
| Y2M02 | 162644 | | 0 | 31.32 | | 165.84 | 175.95 | 0 | 1 | 0 | 0 | 0 |
| Y2M03 | 160602 | | 0 | 15.69 | | 173.97 | 17.2 | 0 | 0 | 1 | 0 | 0 |
| Y2M04 | 249942 | | 4.09 | 11.31 | | 153.23 | 36.21 | 0 | 0 | 0 | 1 | 0 |
| Y2M05 | 154513 | | 12.37 | 14.18 | | 158.57 | 48.08 | 0 | 0 | 0 | 0 | 1 |
| Y2M06 | 159142 | | 7.47 | 0 | | 150.93 | 42.67 | 0 | 0 | 0 | 0 | 0 |
| Y2M07 | 159728 | | 6.63 | 91.4 | | 145.34 | 82.65 | 0 | 0 | 0 | 0 | 0 |
| Y2M08 | 184616 | | 13.59 | 115.48 | | 161.17 | 46.21 | 0 | 0 | 0 | 0 | 0 |
| Y2M09 | 181868 | | 8.48 | 68.68 | | 175.33 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y2M10 | 171800 | | 10.06 | 110.32 | | 167.17 | 118.28 | 0 | 0 | 0 | 0 | 0 |
| Y2M11 | 152178 | | 14.46 | 175.95 | | 191.29 | 161.32 | 0 | 0 | 0 | 0 | 0 |
| Y2M12 | 167492 | | 9.8 | 17.2 | | 194.38 | 131.01 | 0 | 0 | 0 | 0 | 0 |
| Y3M01 | 120123 | | 14.24 | 36.21 | | 186.17 | 146.54 | 1 | 0 | 0 | 0 | 0 |
| Y3M02 | 164718 | | 9.27 | 48.08 | | 184.23 | 202 | 0 | 1 | 0 | 0 | 0 |
| Y3M03 | 156417 | | 9.17 | 42.67 | | 201.48 | 53.94 | 0 | 0 | 1 | 0 | 0 |
| Y3M04 | 108093 | | 8.93 | 82.65 | | 165.84 | 44 | 0 | 0 | 0 | 1 | 0 |
| Y3M05 | 152192 | | 10.13 | 46.21 | | 173.97 | 0 | 0 | 0 | 0 | 0 | 1 |
| Y3M06 | 121089 | | 8.44 | 0 | | 153.23 | 8.27 | 0 | 0 | 0 | 0 | 0 |
| Y3M07 | 156689 | | 0 | 118.28 | | 158.57 | 11.2 | 0 | 0 | 0 | 0 | 0 |
| Y3M08 | 157451 | | 0 | 161.32 | | 150.93 | 8.22 | 0 | 0 | 0 | 0 | 0 |
| Y3M09 | 162644 | | 0 | 131.01 | | 145.34 | 7.25 | 0 | 0 | 0 | 0 | 0 |
| Y3M10 | 160602 | | 0 | 146.54 | | 161.17 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y3M11 | 249942 | | 0 | 202 | | 165.84 | 8.2 | 0 | 0 | 0 | 0 | 0 |
| Y3M12 | 154513 | | 0 | 53.94 | | 173.97 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y4M01 | 159142 | 5.99 | 0 | 32.55 | | 153.23 | 10.37 | 1 | 0 | 0 | 0 | 0 |
| Y4M02 | 159728 | 6.06 | 0.58 | 8.63 | | 158.57 | 12.67 | 0 | 1 | 0 | 0 | 0 |
| Y4M03 | 136379 | 6.25 | 0.72 | 10.18 | | 150.93 | 6.41 | 0 | 0 | 1 | 0 | 0 |

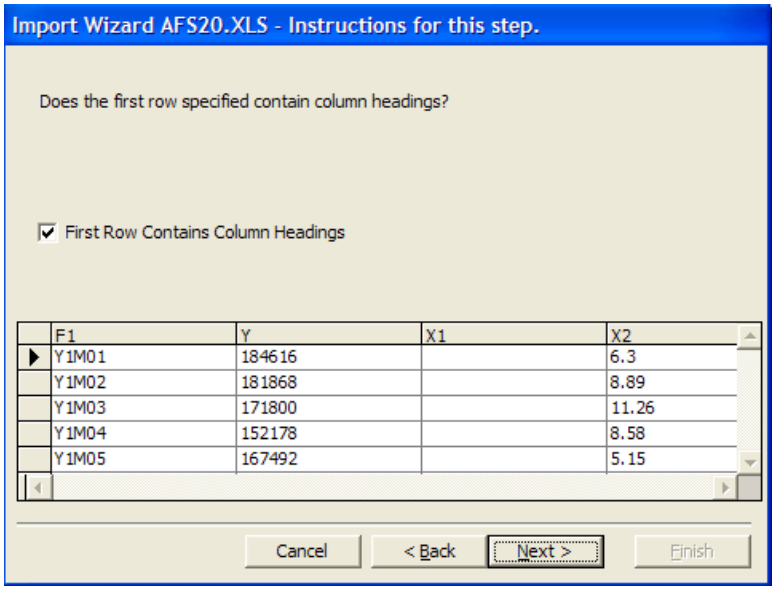


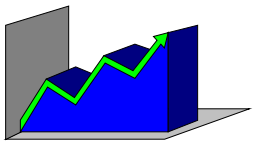
Solutions:

Let's import the file.  Just click "Next".

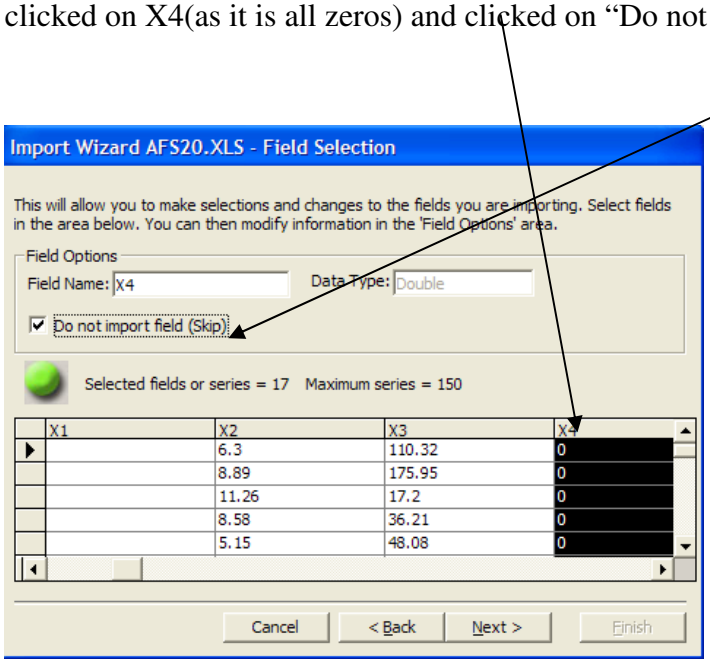



The first row does have column headings so just click "Next".

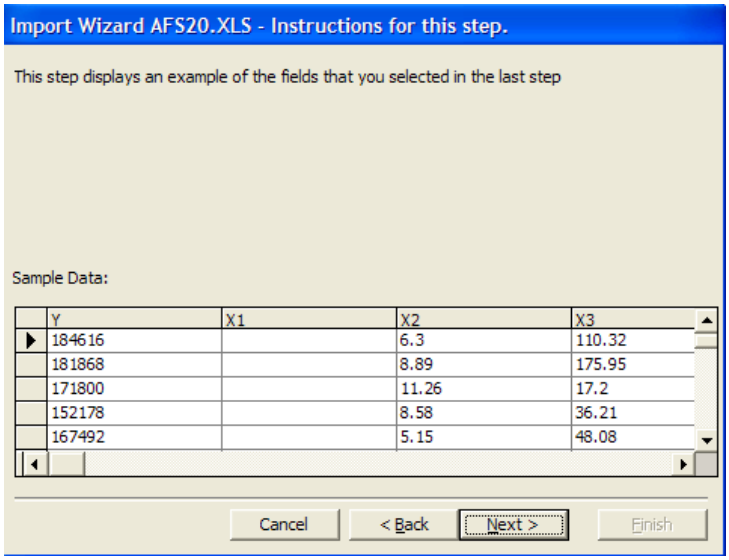


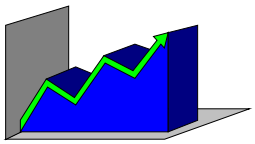


The first column was identified by Autobox as a non-number field and highlighted it. We also clicked on X4(as it is all zeros) and clicked on “Do not import” and then “Next”.

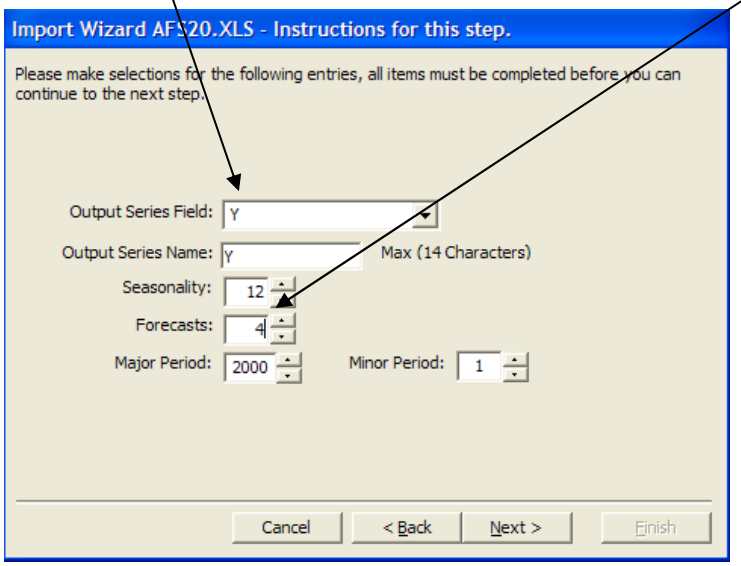



This shows you the proof that certain fields will not be imported. Click “Next” and also “Next” on the following screen (not shown here)

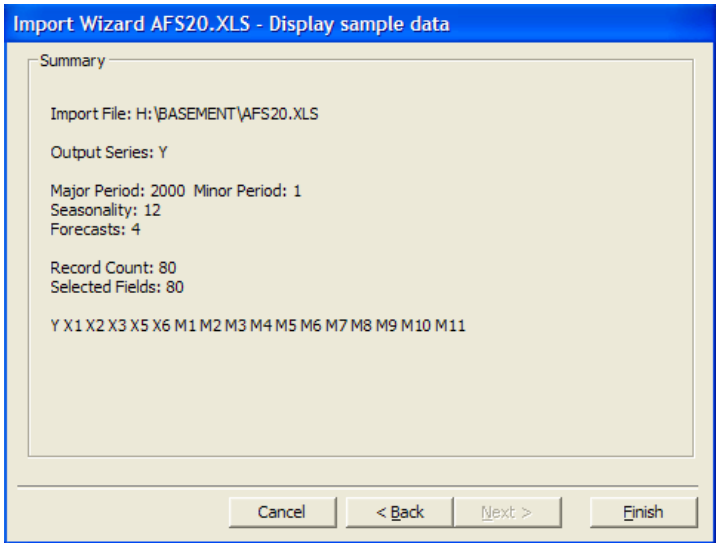


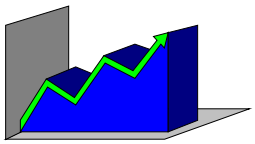


We select Y as the output series to be forecasted and change the number of forecasts to be 4 and click “Next”.

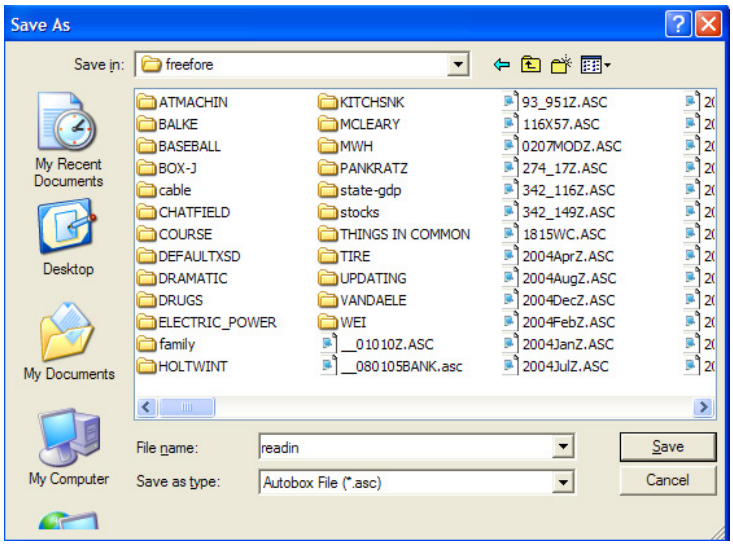



Here is a summary showing what has been imported. Click “Finish”.

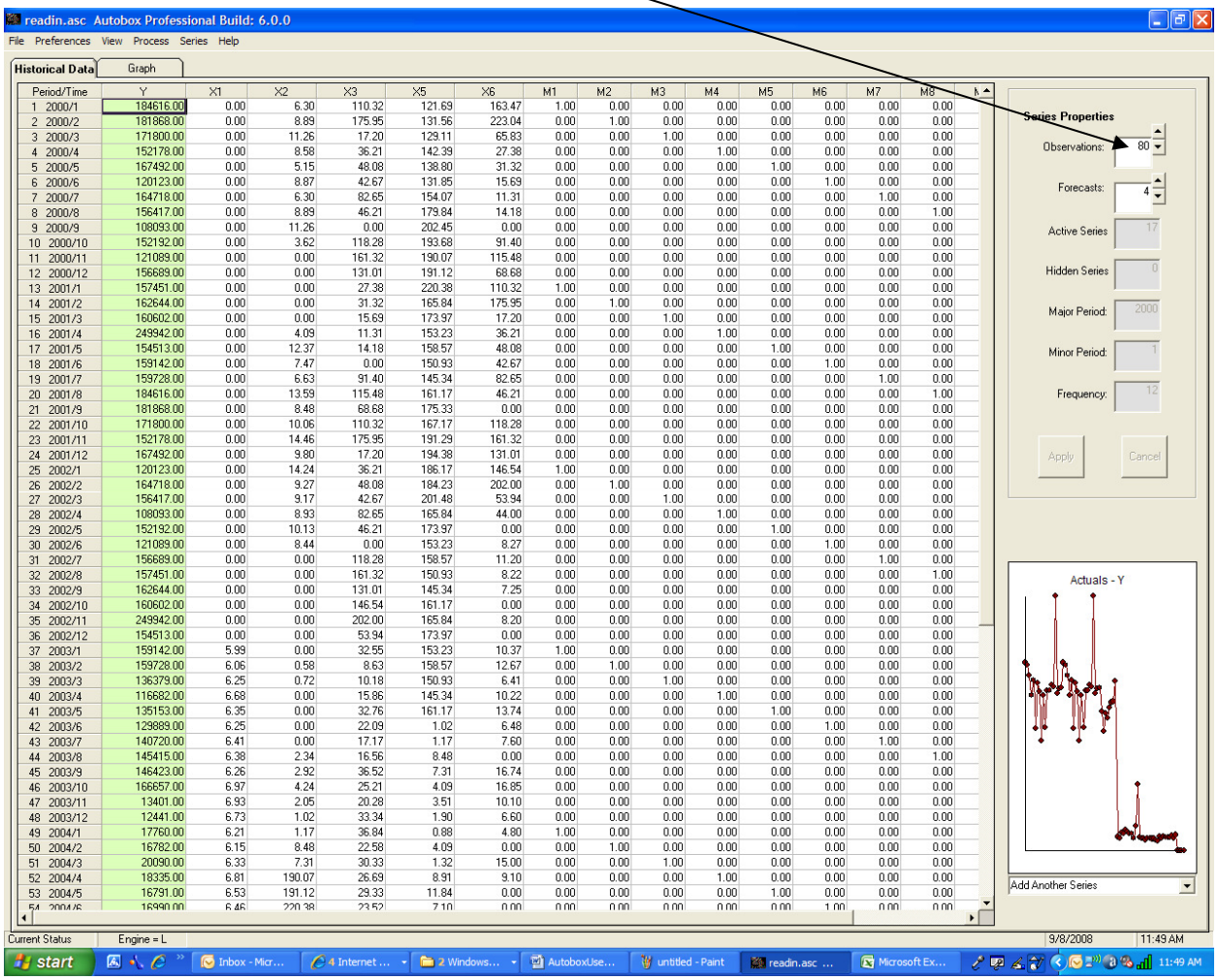


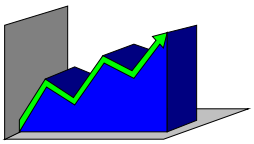


Save the file in anywhere on your computer, but it is best to keep in the “Autobox” installation directory or a subfolder.  We named it “readin.asc” and click “Save”.

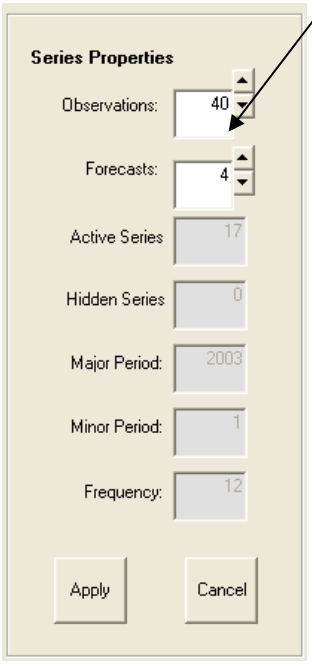


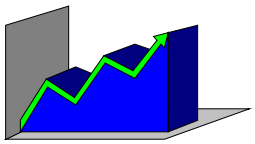
We can change the observations to 44 to get rid of the first 36 observations due to the problem with X1 had with missing data. (We could have deleted the column and ran the analysis without that causal variable which would be a thorough thing to do and we recommend doing just that on your own time!) We will change 80 to 44 so do that and click "Apply" and click on "Delete series from beginning of the series"(not shown below) which will shrink our dataset to 44 observations.





Change the observations now from 44 to 40 so that you can withhold 4 observations for evaluation. Click “apply” and then choose “Delete observations at the end of the series” and click “Next” and then choose “Use the Deleted Observations as “Retained Data”. The Number of Forecasts will be changed to Equal the number of retained data.” and click “Next”. Click “Ok” and then select all of the series to use for retained. Choose “select all” and then “ok”.

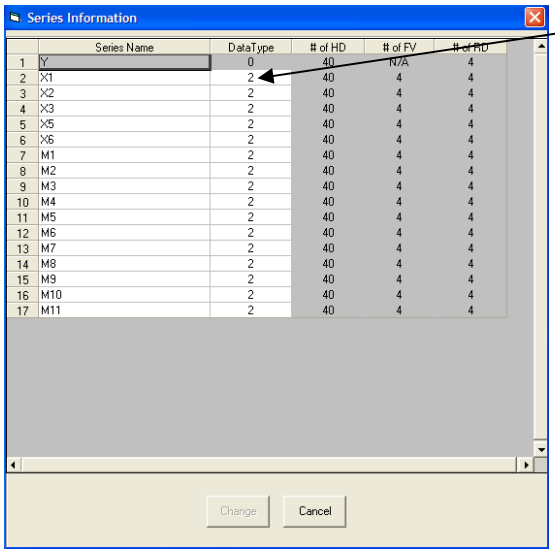


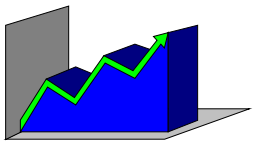


We need to go a little in depth on data types. When you import data from excel you are in essence asking Autobox to determine the data type wheras when you create an Autobox ASC file you are telling Autobox how your data should be treated up front. Autobox needs to default the data type to something when it goes through the import. This default may not be the best case for your situation. If you have no future values, the default data type is set to '0' which means Autobox will project the causal variable into the future to be used in the forecast process. If you have future values for causals then the default is set to '2'.

Let's explain why this is important as there are two cases where you would want to override the default. If you have future values that you would like to provide **or** if you have a causal variable like 11 seasonal dummies (for the 12 months) it makes no sense for the data type to be a '2' as with a '2' Autobox will look for a lag relationship and there would be no lag relationship that could possibly exist for a given seasonal dummy as the next month's dummy variable would control for that. For a situation where you have future values of the causal variable, you might know that there logically could be a lead effect so you would want to change it to '3'.

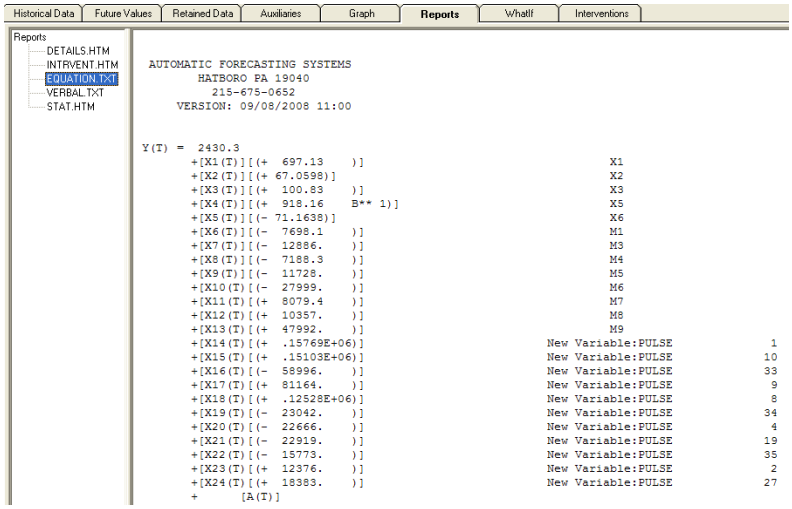
To change the default you need to have retained data. In this example, we have 4 future values of the causal variables. We created them on the previous page which now allows us to now change them by choosing " Series/Series Information" and change all of the seasonal dummies from a '2' to a '1'.



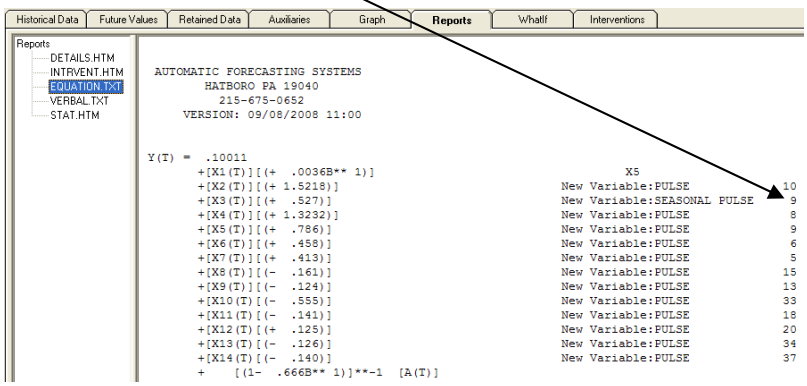


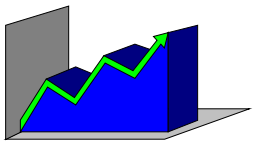
Now let's run Autobox and see what happens.

How interesting. Some of the seasonal dummies were not significant. The adjusted variance (otherwise known as the Mean Square Error) found in the stat.htm report was .001881.




If we run the same problem, but remove the seasonal dummies we only get one indication of any seasonal bump at period 9 and onward. The causal variables are all now **not** significant (except X5), but we do have an autoregressive relationship of 1 period lag. The adjusted variance is .001623 which is better by 14% than the seasonal dummy model.





# Pooled Time Series Cross-Sectional with Autobox

This is for those interested in comparing groups of data to see if the subsets are different than the whole. You may find that the states are different than the national level. Let's assume we have 10 states and the national level data in our example. We want to know which states have higher mortality, but first we need to first follow these steps:

Summary of Steps:

1)Run each of the groups individually

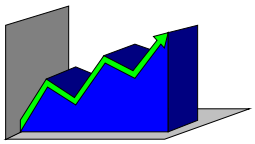
2)Identify common variables among the different groups

3)Run each of the groups individually, but now only using the common variables

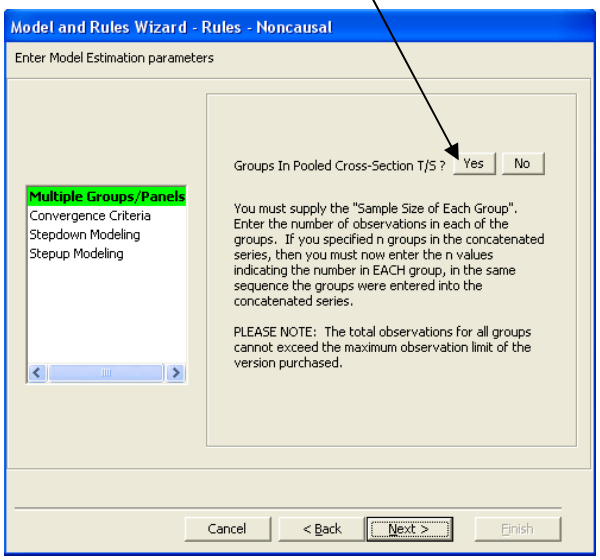
4)Run the Global model using the common variables

5)Compare Sum of Squares to get an F-test

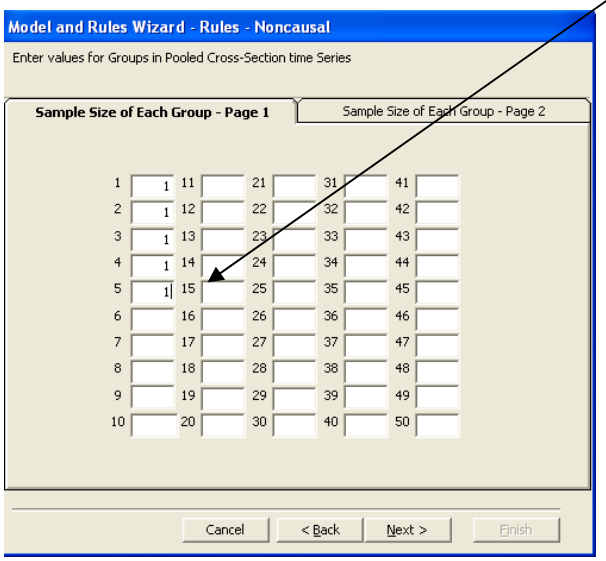
You run Autobox 10 times using option '1' for each of the 10 states. This means you would have 10 different ASC files. You save the model each time you run. You then review the saved models( choose variables that are common amongst the models and rerun all 10 states). Open the output file(s) "Stat.htm" and get the "Sum of Squares" for each of the state models. Sum all the Sum of Squares and let's call that SOSA.



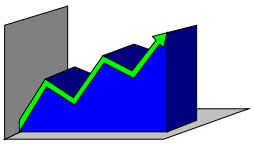
You then run Autobox using option '4' and specify your own model. You need to review the 10 models and establish a common model which is a superset of the 10 models (no interventions in the model). You then choose "Yes" on the "Pooled Cross-Sectional Time Series" option on the "Multiple Groups/Panels" screen.




You would specify the 5 states by typing a '1' here"

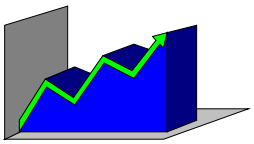



You would run Autobox option '1' time on an ASC file that has all of the states concatenated in



one ASC file. Make sure to disable all 'step up' conditions as want to keep our model in tact. Note the Error Sum of Squares for the national model. Let's call this SOSB.

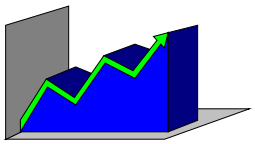
Take the difference of SOSA and SOSB to get SOSC. Divide SOSC by the number of groups(5 in our case) to get MS1 (Mean Square Error). MS2 is calculated by taking SOSA divided by N*K – K*NUMPAR to get MS2 where N = observations in a group and K = # of groups NUMPAR = number of parameters in a model). The f value is then calculated from MS1/MS2.



# Test Data That Comes With Autobox

We provide some classic data sets found and analyzed time and time over. We also provide some datasets that we have found to have caused us to rethink the way we approach time series analysis. Each of these sets of data has its own Directory under the Autobox installation directory. We did not describe every folder and problem set as there are ~720.

```
\BOX-J
Box, G.E.P. and Jenkins, G.M. (1976).  Time Series Analysis:
   Forecasting and Control, 2nd ed., San Francisco: Holden Day.
-----------------------------------------------------------------
bj01  Chemical Process Concentration
bj02  IBM Stock Prices 5/17/61-11/2/62
bj03  IBM Stock Prices 6/29/59-6/30/60
bj04  Chemical Process Temperatures
bj05  Chemical Process Viscosity
bj06  Wolfer Sunspot Numbers
bj07  Batch Chemical Process
bj08  International Airline Passengers
bj09  Methane Gas Input Feed
bj10  Carbon Dioxide Gas Output
bj11  Coded Dynamic Output
bj12  Coded Input
bj13  Coded Gas
bj14  Leading Indicator
bj15  Sales
```



\Mcleary

McLeary, R. and Hay, R. (1980). Applied Time Series Analysis for the Social Sciences, Los Angeles: Sage.

---

BOSTON

DIRECTOR

GERMAN

HYDEPARK

MINNDRUN

NONFATAL

NYIBM

PARISIBM

SPEED

SUTTER

USSUICID

\Pankratz

Pankratz, A. (1983). Forecasting with Univariate Box-Jenkins Models New York: Wiley.

---

PP01 Change in Business Inventories

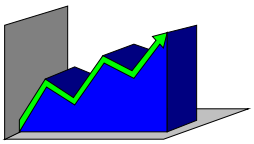
PP01 Saving Rate

PP03 Coal Production

PP04 Housing Permits

PP05 Rail Freight

PP06 AT&T Stock Prices



PP07  Real-Estate Loans

PP08  Parts Availability

PP09  Air-Carrier Freight

PP10  Profit Margin

PP11

PP12  Machine-Tool Shipments

PP13  Cigar Consumption

PP14  College Enrollment

PP15  Exports

\Wei

wei text  http://www.bestwebbuys.com/books/search?t=Author&q=wei+reilly

---

ww01  Truck Manufacturing Defects

ww02  Blowfly Data

ww03  Unemployed Females

ww04  Accidental Death Rate in PA

ww05  US Tobacco Production

ww06  Lynx Pelts Sold

ww07  Simulated model

ww08  Employed Males

ww09  US Beer Production

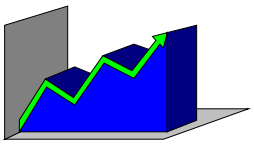
ww10  Series for Exercise 6.2(a)

ww11  Series for Exercise 6.2(b)

ww12  Series for Exercise 6.2(c)

ww13  Series for Exercise 6.2(d)

ww14  US Lumber Production



ww15  Ski Resort Visitors

ww16  US Liquor Sales

ww17  US Gasoline & Oil Consumption

ww18  Series for Exercise 12.1

\TIRE

1000D.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

1004D.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

1007D.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

1008D.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

etc

XL170A.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

XL250A.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

XPM12A.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

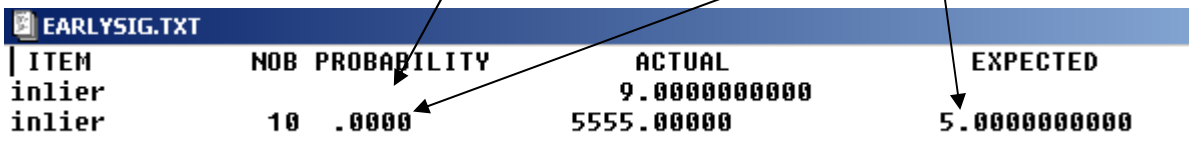
XWM12A.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

XWM16A.ASC  EUROPEAN MONTHLY SALES OF A PARTICULAR TIRE (coded data)

# Early Warning System Report

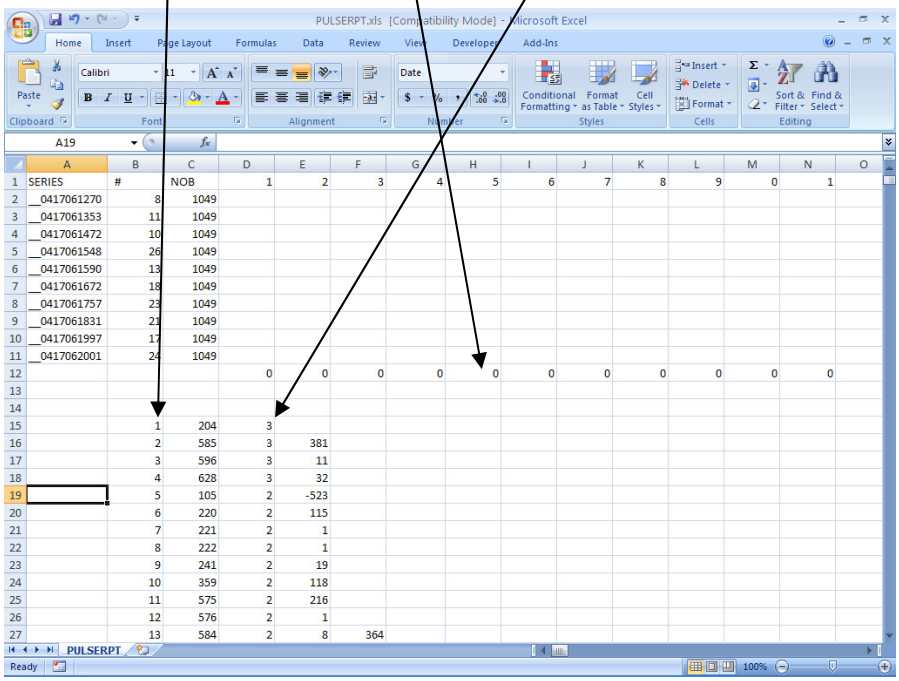
The report "earlysig.txt" is created to help find out if the last observation is "out of control". The report tells you the name of the series, the last observations number, the probability of out being out of control, the observation, and what the observation was expected to be. There is one record added to this file every time (up to 20 series and then the file is purged to avoid a large file—the batch version will continue to write out to this file so if you have 50,000 series this file will have 50,000 records).

You can bring this file into Excel and sort on probability (ascending) to find the series that seem to be "out of control". I ran the series inlier and there was nothing found to be "out of control in the last observation" as you can see here. However, I went and I changed the last observation from a 9 to 5,555 and then reran Autobox. The second row shows a low p-value to show that there is something wrong. It prints out what the value should have been here.

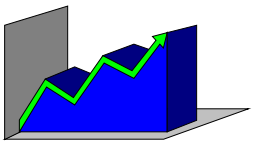


# Pulse Report

Pulserpt.csv – Log file showing a Table of a pulse outlier at different time periods(see the first 11 rows to see what it looks like in the picture below). If you have 200 series and you find that 150 have an outlier at time period 02 then it might cause you to think about what happened at this point in time that you failed to include as a variable in the model from the beginning for these 150 series(possibly for all 200 series?). In a couple of steps you can find if these occurrences also occurred annually suggesting that it was a holiday that was omitted in the modeling process. Open the file in Excel and sum each of the columns. Copy and transpose that summed row to a column. Create the counting numbers next to this column (1,2,3,4, etc). Sort the two columns by the count column largest to smallest. Now you have the count of the time period with the most outliers at the top. Below is an example with 10 SKU's with 1,049 daily observations. We did some investigation by subtracting different time periods to identify a missing holiday variable but we didn't find any differentials of 365 so given that we conclude that these are just interventions and not a systematic pattern since 3 out of 10 could randomly occur at a given time period by chance. Note that the series need to all start at the same time period so that the data is aligned!

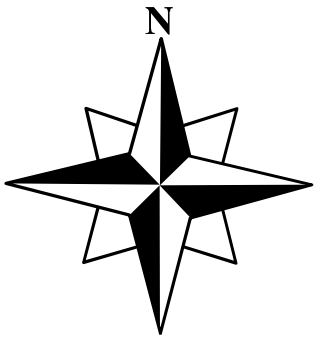


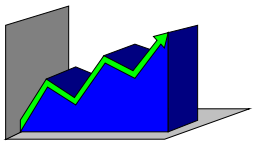
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SERIES | # | NOB | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | |
| 2 | _0417061270 | 8 | 1049 | | | | | | | | | | | | |
| 3 | _0417061353 | 11 | 1049 | | | | | | | | | | | | |
| 4 | _0417061472 | 10 | 1049 | | | | | | | | | | | | |
| 5 | _0417061548 | 26 | 1049 | | | | | | | | | | | | |
| 6 | _0417061590 | 13 | 1049 | | | | | | | | | | | | |
| 7 | _0417061672 | 18 | 1049 | | | | | | | | | | | | |
| 8 | _0417061757 | 23 | 1049 | | | | | | | | | | | | |
| 9 | _0417061831 | 21 | 1049 | | | | | | | | | | | | |
| 10 | _0417061997 | 17 | 1049 | | | | | | | | | | | | |
| 11 | _0417062001 | 24 | 1049 | | | | | | | | | | | | |
| 12 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 13 | | | | | | | | | | | | | | | |
| 14 | | | | | | | | | | | | | | | |
| 15 | | 1 | 204 | 3 | | | | | | | | | | | |
| 16 | | 2 | 585 | 3 | 381 | | | | | | | | | | |
| 17 | | 3 | 596 | 3 | 11 | | | | | | | | | | |
| 18 | | 4 | 628 | 3 | 32 | | | | | | | | | | |
| 19 | | 5 | 105 | 2 | -523 | | | | | | | | | | |
| 20 | | 6 | 220 | 2 | 115 | | | | | | | | | | |
| 21 | | 7 | 221 | 2 | 1 | | | | | | | | | | |
| 22 | | 8 | 222 | 2 | 1 | | | | | | | | | | |
| 23 | | 9 | 241 | 2 | 19 | | | | | | | | | | |
| 24 | | 10 | 359 | 2 | 118 | | | | | | | | | | |
| 25 | | 11 | 575 | 2 | 216 | | | | | | | | | | |
| 26 | | 12 | 576 | 2 | 1 | | | | | | | | | | |
| 27 | | 13 | 584 | 2 | 8 | 364 | | | | | | | | | |



# Autobox

# Reference Guide





# Contents

# 1

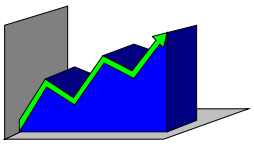
# Overview

### 1. General Remarks

Autobox from Automatic Forecasting Systems (AFS) is an expert system which can be used to model and forecast both univariate and multivariate time series. It's methodology is based on the Box-Jenkin's ground breaking body of work. Outliers or intervention variables can be detected and incorporated into both the non-causal (univariate) and the causal (transfer functions) models. In addition, variance constancy and parameter constancy can be tested and remedies developed. The Autobox system has been under development (see references) for the past 35 years by AFS.

Autobox is attractive to both the expert and the non-expert. The expert can apply his knowledge and control the flow of modeling and estimation. The non-expert can use it as a "statistician-in-the-box" or as a "productivity aid" as it delivers a cost-effective solution. Autobox offers sophisticated analytical tools which can be recommended to teachers, academic researchers and forecasters who usually analyze a handful of time series with great care. Autobox offers a total automation of the modeling and forecasting tasks. Therefore, it can also be recommended to practitioners who must process large amounts of data automatically.

AFS was the first company to automate the Box-Jenkins model building process. Our approach is to program the model identification, estimation and diagnostic feedback loop as originally described by Box and Jenkins. This is implemented for both ARIMA (univariate) modeling and transfer function (multivariate or regression) modeling. What this means is that the user from novice to expert can feed Autobox any number of series and the program's powerful modeling heuristic can do the work for you. This option is implemented in such a way that it can be turned on at *any* stage of the modeling process. There is complete control over the statistical sensitivities for the inclusion/exclusion of model parameters and structures. These features allow the user complete control over the modeling process. The user can let Autobox do as much or as little of the model building process as you or the complexity of the problem dictates.



## 2. Modeling Overview

Autobox is a forecasting engine built on the central modeling steps of the Box-Jenkins paradigm. This core is extended by several useful modules: intervention detection, a simulation option and numerous others, all of which expedite the forecasting practitioner's tasks.

Autobox allows a single endogenous equation incorporating either user specified candidate causal series or empirically identified dummy series. The set of user specified candidate causal series can be either stochastic or deterministic (dummy) in form. During the search for the most appropriate model form and the optimal set of parameters the program can either be:

1. Purely empirical or
2. A starting model could be used.

A final model may require one or more of the following structures:

1. Power transformation like Log, Square Root, Reciprocal, etc.

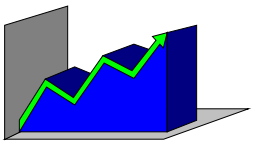
2. Variance stabilization due to deterministic changes in the background error variance.

3. Data segmentation or splitting as evidenced by a statistically significant changes in either model form or parameters.

Enroute to its tour de force, Autobox will evaluate numerous possible models/parameters that have been suggested by the data itself. In practice, a realistic limit is set on the maximum number of model form iterations. The exact specifics of each tentative model is not pre-set thus the power of Autobox emerges. The kind and form of the tentative models may never before been tried. Each data set speaks for itself and suggests the iterative process. The final model could be as simple as:

1. A simple trend model or a simple ordinary least squares model.

2. An exponential smoothing model.



3. A simple weighted average where the weights are either equal or unequal.

4. A Cochrane-Orcutt or ordinary least squares with a first order AR fixup.

5. A simple ordinary least squares model in differences containing some needed lags.

6. A spline-like set of local trends superimposed with an arbitrary ARIMA model with possible pulses.

The number of possible final models that Autobox could find is infinite and only discoverable via a true expert system.

A final model may require one or more of the following seasonal structures:

1. Seasonal ARIMA structure where the prediction depends on some previous reading S periods ago.

2. Seasonal structure via a ***complete*** set of seasonal dummy's reflecting a fixed response based upon the particular period.

3. Seasonal structure via a ***partial*** set of seasonal dummy's reflecting a fixed response based upon the particular period.

The final model will satisfy both:

1. Necessity tests that guarantee the estimated coefficients are statistically significant.

2. Sufficiency tests that guarantee that the error process is:

-unpredictable on itself .

-not predictable from the set of causals or their lags (leads).

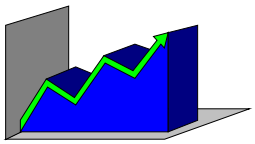
-has a constant mean of zero.

The final model will contain one or more of the following structures:

1. CAUSAL with correct lead/lag specification.

2. MEMORY with correct "autoregressive model components".

3. DUMMY with correct pulses, seasonal pulses, level shifts or spline time trends.



### 3. Statistical Features

Autobox considers the same model class, namely the seasonal autoregressive integrated moving average model with dynamic regressors (transfer components), abbreviated as SARIMAX(p,d,q)(P,D,Q)s[X]. The identification, estimation and outlier diagnosis issues are attacked in a rather elegant way via maximum likelihood estimation.

The user is responsible for selecting the dependent and independent series, and the time range so that Autobox will produce automatic forecasts. The way this is done is controlled by 155 switches specified by the user in various sub-menus.

Switches control the extent of automatic identification steps such as the use of Box-Cox transformation for variance stabilization, setting the confidence levels, the way how to handle differences to achieve stationarity and so on.

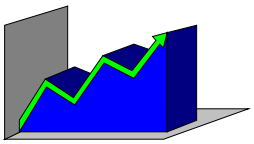
The results of the different modeling alternatives can be reviewed by setting all 45 switches governing "output report options" to a "YES". This yields huge amounts of output, but reveals at least some part of the inherent modeling wisdom of the program. Autobox uses several advanced and clever techniques to identify suitable forecast models, including necessity and sufficiency tests for the parameters, a stability check of the ARMA polynomials (stationarity and invertibility). In keeping with the latest methodologies, we have applied a recently developed test by Franses to discriminate between seasonal unit roots and seasonal dummy model components.

Outlier handling in Autobox recognizes several other types of outliers which **cannot** be handled by the other programs. One of these is the seasonal pulse, which can be very useful to model seasonal patterns which occur only in a few months (like Christmas effects in December). Another is the step or level shift variable which characterizes a permanent change in the level of the series. A recent advancement is the incorporation of trend detection where the model can

have a number of trends each with their own slope.

Autobox can detect dynamic patterns such as transient changes in the data (i.e. level shifts in the data can be detected in both systems as additive outliers in the first differences). It is also able to detect deterministic trend changes which are increasingly used by econometricians.

Different types of exponential smoothing schemes also can be estimated. Moreover, the parameters of the smoothing procedures are automatically optimized. More significantly for



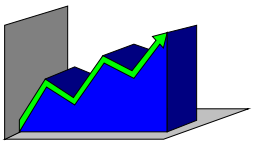
ARIMA users, Autobox extends its automatic modeling to causal modeling.

It offers two causal model identification options. The first technique is the well known prewhitening method developed by Box and Jenkins. As an alternative, the user can select the common filter/least squares technique which is useful when the input causal series are cross-correlated.

Autobox uses the theoretical covariances for alternative candidates as the scheme for both ARIMA and TRANSFER FUNCTION identification. This rather sophisticated pattern recognition selects the numerator/denominator polynomial orders for the rational transfer function weights.

Autobox permits the user to easily evaluate the forecast accuracy of the identified model. First, the user can withhold some observations for forecast evaluation using an out-of-sample analysis. Second, different forecast evaluation measures including the mean error (bias) and mean absolute percentage error (MAPE) can be computed for different forecast horizons without any additional programming.

Autobox comes with a complete set of identification and modeling tools for use in the BJ framework. This means that you have the ability to transform or prewhiten the chosen series for identification purposes. Autobox handles both ARIMA (univariate) modeling and Transfer Function (multivariate) modeling allowing for the inclusion of interventions (see below for more information). Tests for interventions, need for transformations, need to add or delete model parameters are all available. Autocorrelation, partial autocorrelation and cross-correlation functions and their respective tests of significance are calculated as needed. Model fit statistics, including $R^2$, SSE, variance of errors, adjusted variance of errors all reported. Information criteria statistics for alternate model identification approaches are provided.



**Intervention Detection**

One of the most powerful features of Autobox is the inclusion of Automatic Intervention detection capabilities in both ARIMA and Transfer Function models. Almost all forecasting packages allow for interventions to be included in a regression model. What these packages don't tell you is how sensitive **all** forecasting methodologies are to the impact of interventions or missing variables. These packages don't tell you if your series may be influenced by missing variables or changes that are outside the current model.

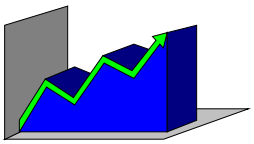
If a data series is impacted by changes in the underlying process at discrete points in time, both ARIMA models and Transfer Function models will produce poor results. For example, a competitors price change changes the level of demand for your product. Without a variable to account for this change your forecast model might perform poorly. Autobox implements ground breaking techniques which quickly and accurately identify potential interventions (level shifts, season pulses, single point outliers and changes in the variance of the series). These variables can then be included in your model *at your discretion*. The result is more robust models and greater forecast accuracy.

**Graphical Analysis Tools**

Autobox has a set of graphing tools that help present complex statistical information in a way that is easy and clear at every stage of the forecasting process. For example, graphs of autocorrelation, partial-autocorrelation and cross-correlation functions are all available. Even more incredibly these can be compared to theoretical values for various models forms.

**Forecasting and Diagnostics**

All forecast packages allow for you to produce forecasts using the models you have constructed. Autobox presents the critical information you need to determine if those forecasts are acceptable. Autobox has options that allow you to analyze the stability and forecasting ability of your forecast model. This is achieved through a series of ex-post forecast analyses. You can *automatically* withhold any number of observations, re-estimate the model form and forecast. Observations are then added back one at a time and the model is re-estimated and reforecast. Forecast accuracy statistics, including Mean Absolute Percent Error (MAPE) and Bias, are calculated at each forecast end point. Thus the stability of the model and its ability to forecast from various end points can be analyzed. Finally, you can optionally allow Autobox to actually re-identify the model form at each level of withheld data to see if the *model form* is unduly influenced by recent observations.

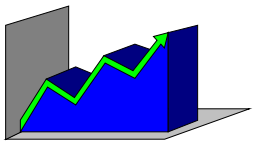


## 4. Conclusions

Academic researchers who usually analyze a handful of time series with great care and in great detail will be able to accomplish their tasks readily. For this group, the expert system component should be part of a data analysis support tool as it can be a “productivity aid”. Quite possibly, they might even find it a replacement (at times) for some of the repetitive tasks and allow them to focus on the “difficult series”.

Due to the high integration of the different time series analysis tools into an integrated framework, Autobox can be strongly recommended to practitioners who must handle many data sets. Practitioners with only little knowledge about Box-Jenkins models can use the program as a pure black-box tool, assuming some knowledgeable user with experience in forecasting methods has configured the system to match the specific features of the data and the specific needs of the end user.

Without the support of an experienced forecaster, a novice may unwittingly run into pitfalls. This is probably true for any sophisticated forecasting program not just this one.



# 2

# ANALYTICAL CAPABILITIES

Autobox is a comprehensive package for the statistical analysis of time series data. It incorporates a variety of techniques for time series modeling and forecasting, ranging from the simple to the complex. The user chooses the type of modeling from a list that includes:

1) Exponential Smoothing
2) Trend Curves
3) Simple Regression
4) Multiple Regression
5) ARIMA (Univariate Box-Jenkins)
6) Transfer Function (Multiple Input Box-Jenkins)

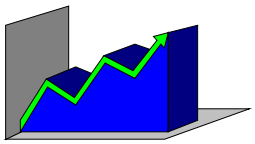
The user can choose to let Autobox automatically identify the model for you or you can do it yourself nonautomatically.

AUTOMATIC Identifies the model for you that best fits the data

NONAUTOMATIC The user initiates the modeling process and chooses the model to use.

IDENTIFICATION

Autobox provides the necessary statistics for model identification. This requires knowledge of the Box-Jenkins identification process. If estimation is enabled, then the initial model is estimated and potentially re-defined by either augmentation or simplification strategies. Augmentation is accomplished by carefully examining the residuals. The errors in the residuals are corrected by adding intervention variables into the model for evaluation. Model augmentation



is done considering both stochastic and deterministic components. Model simplification is done in a straight-forward way, deleting one coefficient at a time.

ESTIMATION Nonlinear least squares estimation of the user specified model parameters. Also computes the necessary statistics for diagnostic checking.

FORECASTING Generates the forecast values from the current model.

UNIVARIATE MODELING

When you build a model by only using the past of the time series.

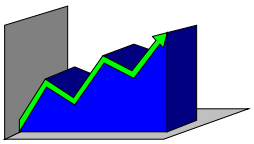
TRANSFER FUNCTION MODELING

This is causal modeling where there are series that help to predict our series of interest. You must identify if the input series need to have an ARIMA model built to filter them of their "within" relationship so that a true relationship can be identified "between" the input and output time series. This process is known as "filtering" or "**prewhitening**" the input series so as to clean them. Cross-correlations then can be computed to help identify what relationship might exist between the input and output to help suggest a model form. An alternative filtering process was developed by Liu and Hanssens named the "**common filter method**" which builds the same filter for each of the input series since they share similar ARIMA model forms.

ARIMA models

`Rational expectations model' - where a weighted average of the past will best fit the data. The $A_t$ represents the causal variables that were **omitted**.

$$Y_t1 = VO + V1\ Y_{t-1}1 + V2\ Y_{t-2}1 + ... + A_t$$



In general form,

$$Y_t = \left\{ \frac{\theta\ (B)}{\Phi\ (B)} \right\} A_t$$

where $Y_t$ = the discrete time series,

$\underline{\theta}$ = autoregressive factors

$\Phi$ = moving average factors,

$A_t$ = the noise series,

and B = the backshift operator which is interpreted as "a lag of…N".

Transfer function models

The model uses a weighted average of the past of both the input and output series to best fit the data.

$$Y_t1 = VO + V1\ Y_{t-1}1 + V2\ Y_{t-2}1 + \ldots + A_t$$
$$+ W0X_t1 + W0X_{t-1}1 + W0X_{t-2}1$$

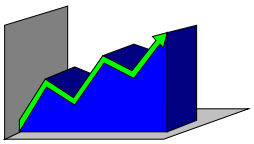
In general form,

$$Y_t = \left\{ \frac{\omega\ (B)}{\delta\ (B)} \right\} X_t + \left\{ \frac{\theta\ (B)}{\Phi\ (B)} \right\} A_t$$

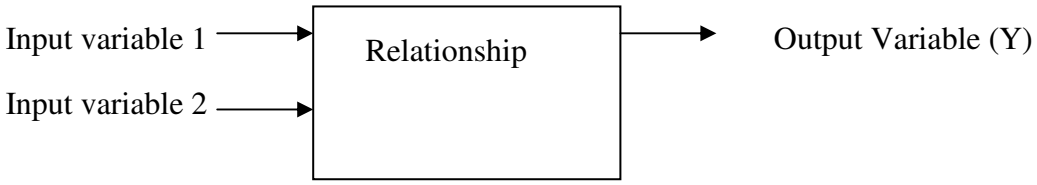
where

$\underline{\omega}$ = input lag factors (lag variables on the X)

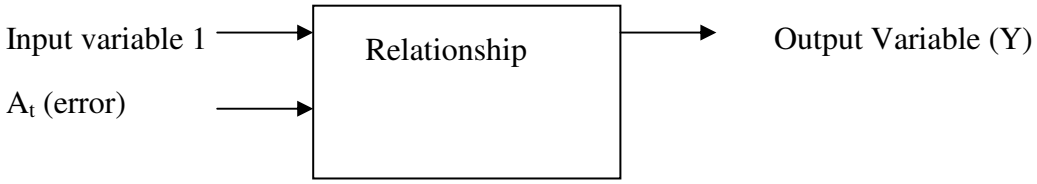
$\delta$ = output lag factors (lag variables on the Y)

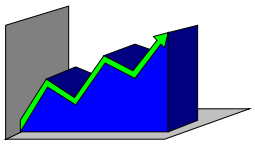


If we include two input variables to model the behavior of the time series Y, then the model will look as follows:




If we don't include variable 2, then we will have more error and $A_t$ represents what is "**omitted**" in a model.





# 3

# Learning Box-Jenkins

This explains the Box-Jenkins methodology and certain fine points in the process. We explain Univariate, and Transfer Functions, impact of time, impact of interventions, impact of trends, impact of changes in variance. Overall, this discussion tries to explain these concepts, but at the same time try to show how these points are handled by Autobox and eventually how you can stand to benefit by using Autobox.

UNIVARIATE BOX-JENKINS - INTRODUCTION

Univariate Box-Jenkins is a time series modeling process which describes a single series as a function of its own past values. The purpose of the B-J process is to find the equation (or filter) that reduces a time series with underlying structure to white noise. Since the filter accounts for the predictable portion of the time series, it can be used to forecast future values of the series.

The modeling procedure itself is a three stage iterative process of:

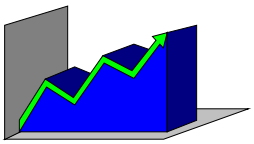
1) Identification - Choose a tentative model form by examining a plot of the series and several key statistics.

2) Estimation & Diagnostic Checks - Estimate the parameters in the identified model. Examine the residuals for model sufficiency and necessity.

3) Forecasting - Use the model to generate forecasts for future values of the time series.

Before proceeding to the discussion concerning the "how to" of modeling, it is necessary to define some terms. Understanding the terminology is frequently the most confusing aspect of understanding Box-Jenkins modeling. This truth is further complicated by the fact that different publications use interchangeable terms to describe B-J models.

The modeler examines statistical information generated from the data in order to choose a time series model from this general class. Formally defining this class of models and outlining the process of finding the correct model is the contribution that Box and Jenkins (1976) made to time series analysis.



A series may need to be differenced to remove the trend in its data. The goal is to transform the data so that is stationary and not trending. A model may have from 0 to ? differencing factors, where ? indicates as many as are necessary. Each differencing factor is a polynomial of the form $(1 - B^{o})^{d}$, where o is the order of the differencing factor, d is the degree of the differencing factor and B is the backshift (lag) operator.

$\Phi_p$ (B) autoregressive factors:

A model may have from 0 to ? autoregressive factors, where ? indicates as many as are necessary. Each autoregressive factor is a polynomial of the form :

$(1 - \Phi_1 B^1 - \Phi_2 B^2 - \Phi_3 B^3 - \ldots\ldots\ldots\ldots\ldots \Phi_p B^p)$,

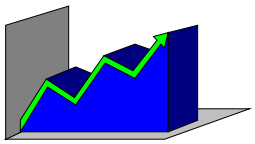
where $\Phi_1 \cdots\cdots\cdots \Phi_p$ are the parameter values of the polynomial, and B is the backshift operator. The values of the autoregressive factors ( $\Phi_1 \ldots\ldots \Phi_p$ ) need not all be nonzero. A zero parameter value indicates that the parameter is not included in the polynomial.

$\theta_q$ Moving average factors:

A model may have from 0 to ? moving average factors, where ? indicates as many as are necessary. Each moving average factor is a polynomial of the form :

$(1 - \theta_1 B^1 - \Phi_2 B^2 - \Phi_3 B^3 - \ldots\ldots\ldots\ldots\ldots \Phi_q B^q)$,

where $\theta_1 \ldots \Phi_q$ are the parameter values of the polynomial and B is the backshift operator. The values of the $\theta_p \ldots \Phi_q$ need not all be nonzero. A zero parameter value indicates that the parameter is not included in the polynomial.



B (backshift operator):

The backshift operator is a special notation used to simplify the representation of lag values. $B^J X_t$ is defined to be $X_{t-J}$. So, $(B^1)X_t = X_{t-1}$ which means a 1 period lag of X.

To summarize, we will define the terms of several example models. Please note that for illustrative purposes these sample models may be more complicated than the average model.

Example 1: $(1-.5B^{12}) (1-B^1)^2 \ Y_t = (1-.2B^1) (1+.3B^6) A_t$

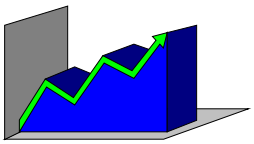
This B-J (ARIMA) model has one autoregressive factor, one differencing factor and two moving average factors. The autoregressive factor has one parameter, with a backorder (**lag**) value of 12 and a parameter value of .5. The differencing factor is of order 1, degree 2. Each of the two MA factors has one parameter. The first MA factor has a single parameter with a backorder power of 1 and a parameter value of .2 . The second factor has a single parameter with a backorder power of 6 and a parameter value of -.3. This model does not contain a trend parameter. Since there is at least one differencing factor, the mean parameter is zero; therefore, it is not explicitly included in the model.

Example 2: $(1-.9B^1 +.3B^2) (1-.7B^6) (Y_t - \mu) = A_t$

Example 2 has two autoregressive factors, zero differencing factors and zero moving average factors. Since there aren't any differencing factors, the mean parameter ($\mu$) must be included explicitly. The first AR factor has two parameters, with backorder powers of 1 and 2 respectively. Their associated parameter values are .9 and -.3 . The second AR factor has one parameter with a lag value of 6 and a parameter value of .7.

Example 3: $(1-B^1)^2 \ (1-B^{12})^1 \ Y_t = 1.7 + (1-.6B^1 -.2B^{12}) \ A_t$

The third example has two differencing factors, a deterministic trend parameter and one MA factor. The differencing factors are of order 1, degree 2 and order 12, degree 1, respectively. The value of the trend parameter is 1.7. The single MA factor has two parameters, with backorder powers of 1 and 12 and associated parameter values of .6 and .2.

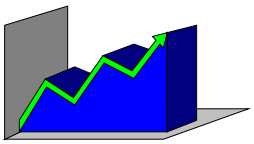


## UNIVARIATE BOX-JENKINS - ARIMA MODEL

The Box-Jenkins modeling process has become popular because the data itself is used to determine the appropriate model form, whereas many other time series modeling methods assume a given model form a priori. It turns out that most of these "assumed models" are, in actuality, subsets of the general ARIMA model. However, they are not necessarily the best for a given series. Using the Box-Jenkins process ensures that the subset model is optimal. This does not mean that an analyst can always perfectly describe the underlying process of a time series. It means that, given a time series with an underlying structure, the modeler can find the best model for that series. The question at hand is: "How does one determine the model form?" Answering that exact question is the purpose of this section.

## MODEL IDENTIFICATION

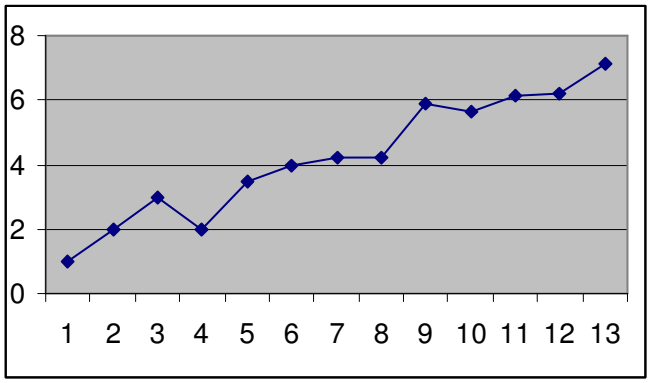
The identification phase entails examining the time series in order to choose a tentative model form. There are several key statistical tools used during this phase. These "tools" are the autocorrelations and the partial autocorrelations. The first step in identification is to make the time series stationary. In a stationary series, the mean and the variance are constant over time. Since there are two types of nonstationarity, there are two methods of inducing stationarity. Applying the appropriate differencing factors to a series creates a mean stationary series. Applying the correct power transformation (lambda value) creates a variance stationary series.



An Example of Stationary vs Nonstationary

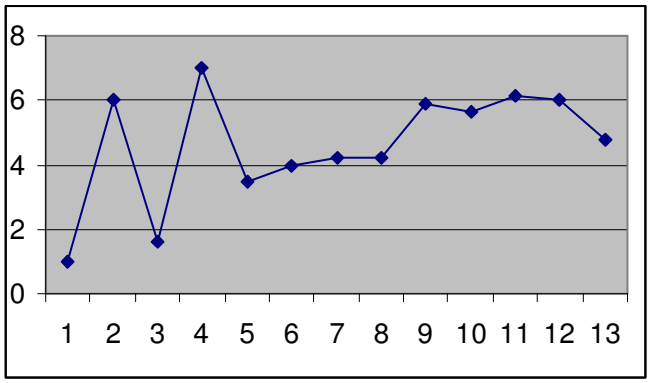
NONSTATIONARY

The variance is constant, but the mean changes over time



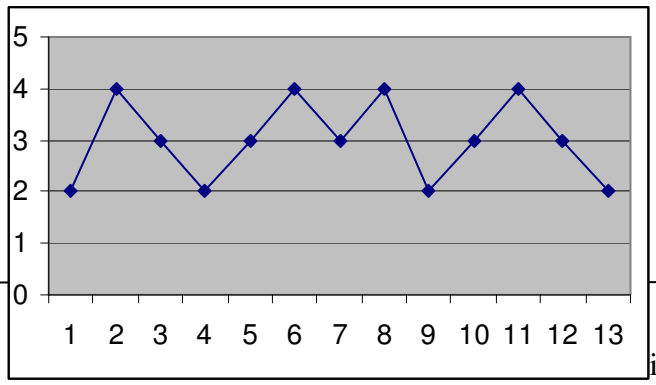

NONSTATIONARY

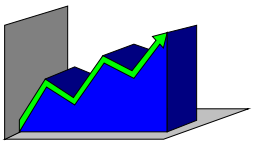
The mean is constant, but the variance changes over time. This illustrates a case where the variance change is not related to the mean change.




STATIONARY

The mean and the variance are constant over time

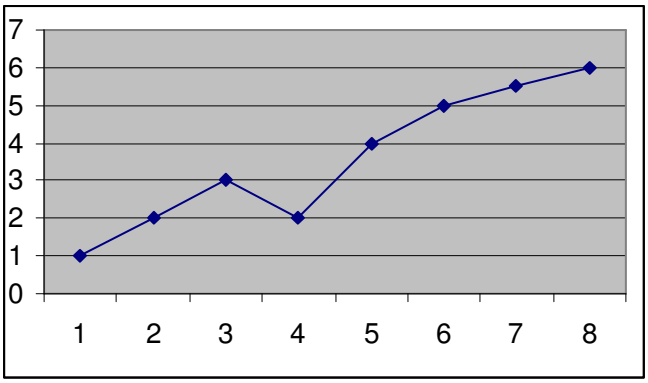




The autocorrelations are clues for what the appropriate level of differencing is needed. The need for a power transformation can be ascertained by examining plots of both the original series and the transformed series. Alternatively, the Box-Cox error sum of squares test can find the optimal power transformation. AFS's modeling software enables both tests for variance stationarity.

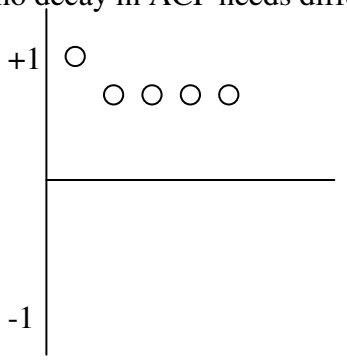
An autocorrelation function that starts out high (.9 or above) and decays slowly indicates the need for differencing. The order of the differencing is determined by the number of time periods between the relatively high autocorrelations. This is perhaps best explained with an example.

Nonstationary Time Series (trend)

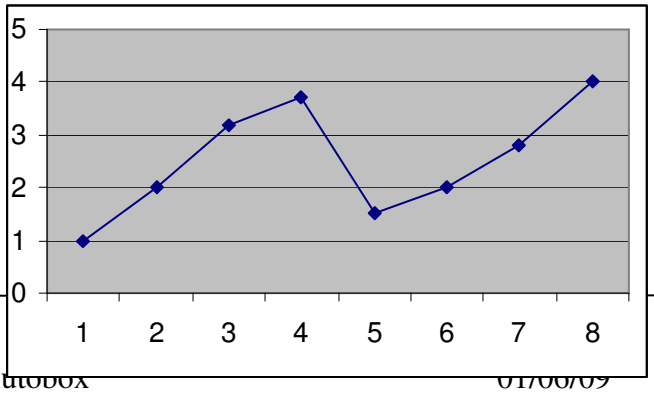



Example ACF

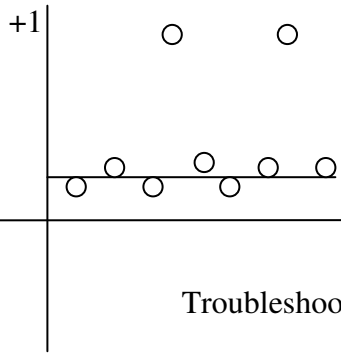
(no decay in ACF needs differencing)

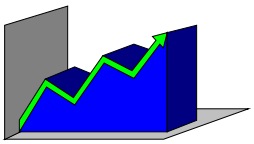



Nonstationary Time Series (seasonal)




(lags 4, 8, etc. are all high)needs seasonal differencing of order 4)



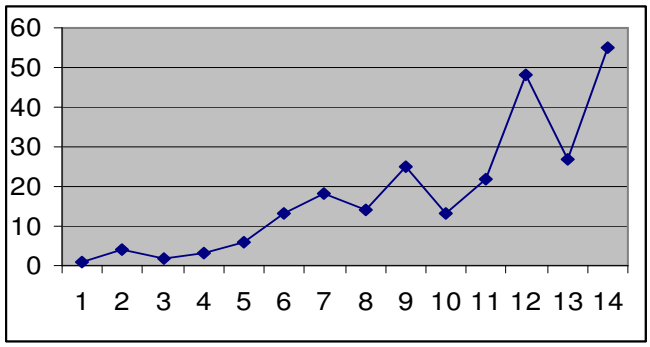




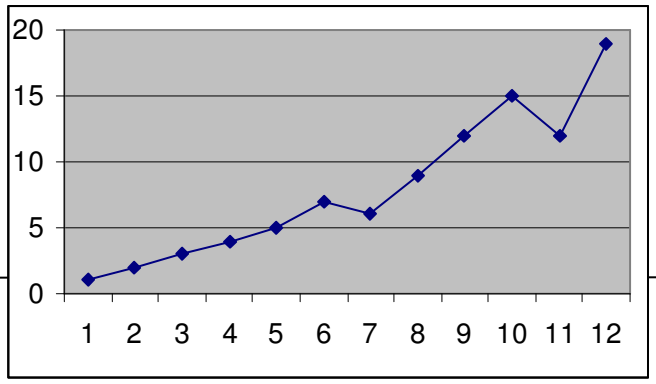
If the variance increases and the mean also increases, then this probably would require a power transformation. This also applies if the variance and mean decreases. This particular violation of the constant variance requirement requires a power transform (lambda value) in order to make it stationary. The lambda value may be anywhere from +1 to -1, where +1 indicates the original series, 0 indicates natural logs of the series, -1 indicates inverse of the series, etc. (See figure below).

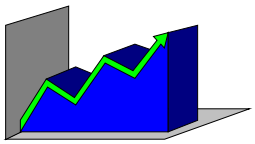
The Box-Cox test for determining the optimal lambda is more accurate than simply eyeballing a graph, however, it is an estimation based procedure and is therefore tied to a given model form. The test is to estimate a model using different lambda values. The lambda that produces the lowest error sum of squares is optimal for that model. The user can vary or evaluate alternative values of lambda in both the Automatic Initial model identification stage or the estimation stage.

An Example of Variance Instability due to Mean Change

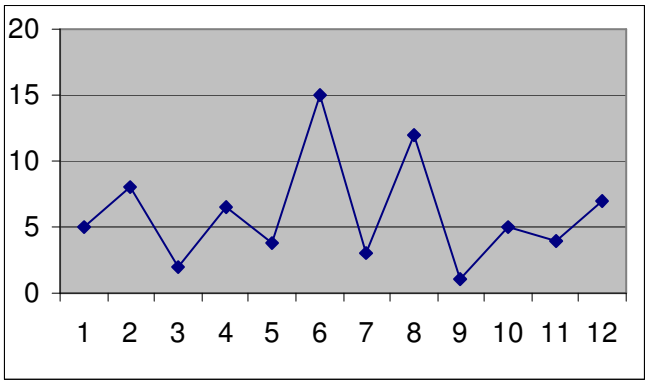



After transforming the series (natural log scale), the variance has become stable.

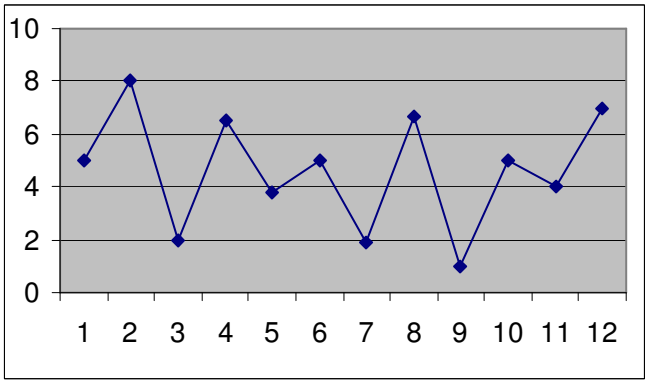




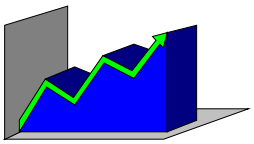
An Example of Variance Instability Not due to Mean Change




Transformed to remove the heterogenous variability




The step succeeding the inducement of stationarity is tentative identification of the autoregressive/moving average structure. The autocorrelations and the partial autocorrelations of

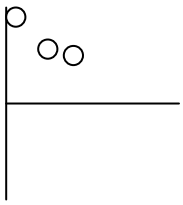
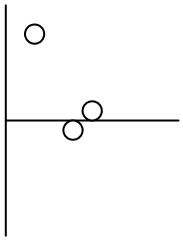
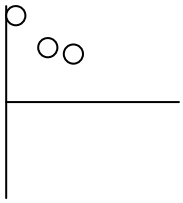
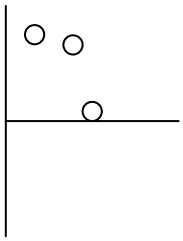
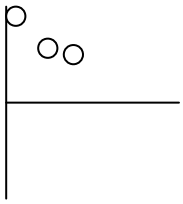
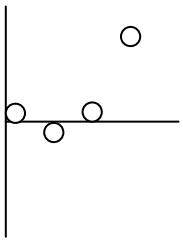


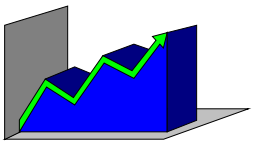
the appropriately differenced and transformed series have patterns which can be associated with a particular model form. The analyst can make a first pass conjecture about model form by examining the information in the ACF and PACF. This is the sole purpose of the univariate identification phase. Given the appropriate stationarity factors and a tentative model form, the analyst can proceed to phase two - model estimation.

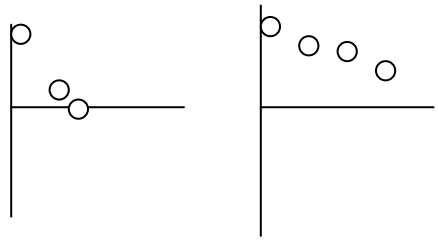
Our presentation of the above mentioned "patterns" will once again take the form of examples.

The next page contains a column of sample ACF's and PACF's and a column of tentative model forms. We suggest that you look over this chart and familiarize yourself with the model forms that are appropriate for a given pattern in the ACF and in its corresponding PACF. In general, decaying autocorrelations indicate autoregressive structure and decaying partial autocorrelations indicate moving average structure.

Here are some example ACF and PACF Patterns

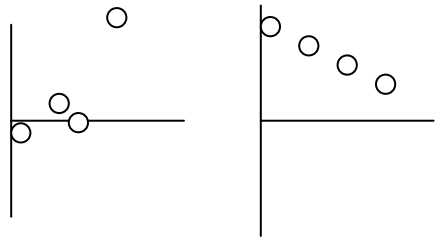
| EXAMPLE ACF | EXAMPLE PACF | SUGGESTED MODEL FORM |
|---|---|---|
|  |  | AR 1<br>$(1 - \Phi_1 B^1)(Y_t - \mu) = A_t$ |
|  |  | AR 2<br>$(1 - \Phi_1 B^1 - \Phi_2 B^2)(Y_t - \mu) = A_t$ |
|  |  | AR 4<br>$(1 - \Phi_1 B^4)(Y_t - \mu) = A_t$ |





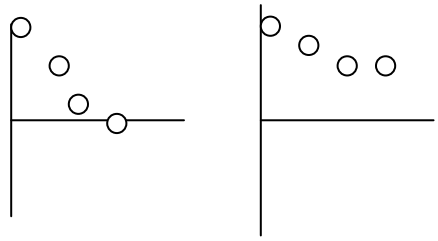
MA1

$(Y_t - \mu) = (1 - \theta_1 B^1) A_t$



MA4

$(Y_t - \mu) = (1 - \theta_1 B^4) A_t$



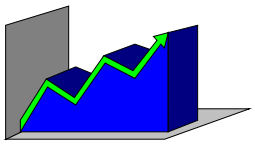
AR1 MA4

$(1 - \Phi_1 B^1)(Y_t - \mu) = (1 - \theta_1 B^4) A_t$

As illustrated above, matching up patterns in observed sample ACF's and PACF's with theoretical models can sometimes be a bit of a challenge. Autobox implements a search–and-capture heuristic which evaluates alternatives and then selects the best model using decision rules based upon the AIC criteria and the error sum of squares. This rule-based system then, at the user's request, is used to automatically identify the initial model. Trained users may identify and specify this initial model themselves and rate themselves vis-a-vis the internal expert within Autobox.

## MODEL ESTIMATION AND DIAGNOSTIC CHECKING

The second phase of the model building process is estimation of the coefficients in the tentatively



identified model. Autobox uses a conditional nonlinear least squares estimation procedure that is based on the Marquardt algorithm. Chapter 7 and pages 500-505 of the Box-Jenkins text (1976) cover this subject in detail.

There are three basic diagnostic checks that must be performed on the estimated model. These tests are for necessity, invertibility and sufficiency. Each parameter included in the model should be statistically significant (necessary) and each factor must be invertible. In addition, the residuals from the estimated model should be white noise (model sufficiency).

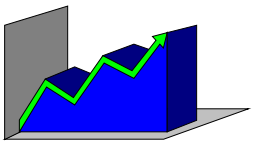
The test for necessity is performed by examining the T-ratios for the individual parameter estimates. Parameters with nonsignificant coefficients should be deleted from the model. Invertibility is determined by extracting the roots from each factor in the model. All of the roots must li.e. outside of the unit circle. If one of the factors is noninvertible, then the model must be adjusted. The appropriate adjustment is dictated by the type of the factor that is noninvertible. For example, a noninvertible autoregressive factor usually indicates under-differencing, while a noninvertible moving average factor may indicate over-differencing. A noninvertible moving average factor could also represent the presence of a deterministic factor. Since the model fixup is not really clear-cut, the analyst must consider the overall model when adjusting for noninvertibility.

The residuals are tested for white noise in much the same way as model identification is performed. If there are patterns in the residual autocorrelations and partial autocorrelations, then the analyst may need to add parameters to the model. One follows the pattern recognition rules described above when adding parameters to the model. Autobox's automatic procedures can be a great time saver for this step.

If all of the parameters are necessary, if each factor is invertible and if the model is sufficient, then the ARIMA model is adequate and it can be used for forecasting.

## MODEL FORECASTING

Model forecasting with the properly identified and estimated model is simply an algebraic process of applying the model form to the actual time series data and computing the forecast values from a given time origin. The confidence intervals give a measure of the uncertainty in the point forecasts.



TRANSFER FUNCTION MODELING - INTRODUCTION

The process of identifying an appropriate model between a time series and it's past and other series and their past can be confusing to modelers.

Multiple input Box-Jenkins is a time series modeling process which describes a single dependent series as a function of its own past values and the values of one or more independent input series. As with univariate modeling, the purpose of multiple input modeling is to find the model which accounts for the predictable portion of the dependent series. Such a model can then be used for both forecasting or control.

For the sake of definition, output series refers to the dependent series (Y) which is a function of one or more independent input series (X). The input series may be either stochastic (nondeterministic) or intervention (deterministic). The transfer function accounts for the measurable relationship between the output and each of the input series. The difference between the output series and that part of it which can be accounted for by the input series is called the noise. The noise itself is stochastic and therefore may be represented by an ARIMA model. The combined transfer function and noise model comprise the multiple input Box-Jenkins model.

In very general terms, this relationship is portrayed by the following equation:

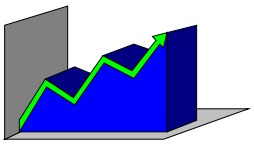
Although the term "transfer function" technically represents the X-Y relationship, in practice the term is frequently employed in reference to the entire B-J multiple input model. We also adhere to this practice, and many times we refer to a transfer function when we are actually discussing a multiple input Box-Jenkins model. The multiple input modeling procedure is a three stage iterative process:

1) Identification -> Choose a tentative model form by examining a plot of the series and several key statistics.

2) Estimation & Diagnostic Checks -> Estimate the parameters in the identified model. Examine the residuals for model sufficiency and necessity.

3) Forecasting -> Use the model to generate forecasts for future values of the dependent time series.

TRANSFER FUNCTION MODELING - MODELING PROCESS DESCRIBED



The time series analyst must identify the various components of the model. The first step is to decide which independent input variables are driving the output series. The next step is to develop the appropriate model. The initial step of choosing tentative input variables is a responsibility beyond the scope of this statistical method. However, we can use the Box-Jenkins method to ascertain which of the candidate input variables should be included in the final model. Hence, armed with an output series to be modeled and a list of potential input variables, one can proceed to the first phase of modeling - model identification.

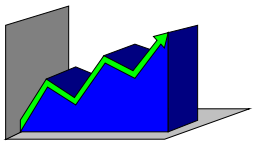
## MODEL IDENTIFICATION

The purpose of model identification is to select an appropriate preliminary model form for the given set of time series variables. Box and Jenkins (1976) outlined a procedure for determining this tentative model form. As in the univariate case, their procedure utilizes the information in the data itself. The analyst must "identify" a model as opposed to "assuming" a pre-defined model form. This section covers the basics of multiple input model identification.

The transfer function identification phase is in itself a multiple step process. The very first step is to develop a univariate model for each of the stochastic time series variables. Univariate model building has already been covered so it will not be repeated here. The next step is to compute certain key statistics which are, in turn, examined for clues as to the model form, much like univariate identification. After tentatively identifying the model form, one can proceed to phase two - model estimation.

Here we must digress a bit. Box and Jenkins discussed the single input - single output model identification problem in their original text (1970). Since then, Box and Tiao (1975), Liu and Hanssens (1982) and several others have expanded on the original theory to discuss the identification of a wider range of transfer function models. The works that we refer to include intervention modeling and multiple input transfer functions with input series noise processes that are not necessarily uncorrelated (uncorrelated input series noise processes is an assumption of the original transfer function discussions). We will cover the basic model identification procedures for these various "types" of transfer function models; however, we do recommend that you consult the following references for a more detailed study : (Bell, 1983), (Liu-Hanssens, 1982), (Pack, 1978), (Tsay, 1986).

In the second paragraph, we alluded to the fact that certain key statistics give clues as to the appropriate model form. Now we must discuss which statistics are important to model identification and how these statistics are to be interpreted. To do this, we have separated



multiple input model identification into four distinct methods. Each method is useful for a particular kind of transfer function model. The software package is programmed for all four methods. Please remember that this is not an all-inclusive coverage of transfer function models, we are simply covering the basics of model identification.

PREWHITENING/CROSS CORRELATION METHOD - OVERVIEW

Source : Box and Jenkins (1976)

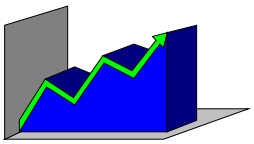
Use : Transfer function models with one or more stochastic input series, where the input series noise processes are not cross correlated.

The procedure for transfer function model identification outlined by Box and Jenkins uses the cross correlations between two prewhitened series to tentatively identify model form. The first step to this process, as was mentioned above, is to develop an ARIMA model for each time series in the equation. Each series must then be made stationary by applying the appropriate differencing and transformation parameters from its ARIMA model.

The stationary time series are, in turn, "prewhitened". Prewhitening refers to the process of applying a given set of autoregressive and moving average factors to a stationary series. Each input series is prewhitened by its own ARIMA model AR (autoregressive) and MA (moving average) factors. The output series is also prewhitened by the input series AR and MA factors. If there is more than one input series, then the stationary output series is prewhitened once for each different input. Prewhitening is necessary because it removes the intrarelationship in the individual series. This allows you to more accurately assess the interrelationship between the input and the output series.

The cross correlations between the prewhitened input and output reveal the extent of this interrelationship. The cross correlations can be converted to estimates of the impulse response weights. The patterns in the impulse response weights indicate the tentative transfer function model form. By applying these impulse response weights to the output series, one can generate a preliminary estimate of the noise series.

Following the rules for ARIMA model identification, the patterns in the autocorrelations and partial autocorrelations of the tentative noise process give clues as to the initial form of the noise model. Given the identified transfer function and noise model, one can proceed to the model estimation/diagnostic checking phase.



CROSS CORRELATION METHOD

This option performs the entire three-step iterative Box-Jenkins procedure for developing a transfer function model. The user must specify the time series variables that are to be included in the transfer function equation.

The process starts with the automatic (this is an option) development of each time series variable's prewhitening model, using the Autobox automatic ARIMA modeling algorithm. It then moves on to the initial identification of the transfer function - noise model. This is accomplished by appropriately prewhitening each time series and computing the estimated impulse response weights via the cross correlations (as described in the Box Jenkins 1976 text). The program then estimates the tentative model and performs all of the diagnostic checks for sufficiency, necessity and invertibility. The model is updated as needed, and the diagnostic checking stage ends when the criteria for an acceptable model are met.

COMMON FILTER/LEAST SQUARES METHOD - OVERVIEW

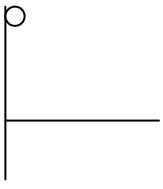
Source : Liu and Hanssens (1982)

Use : Transfer function models with one or more input series, where the input series are not stochastic and/or the input series noise processes are not necessarily free from cross correlation.

The impulse response function measures the correlation between X and Y. This function is similar to the ACF and PACF in that it measures correlation (in their case the correlation was on it self), but has a lag of 0 which the ACF and PACF do not since anything correlated on itself is 1. The lag of 0 for the impulse response function would signify a relationship between the X and Y variable with no lag.

Impulse Response Function Suggested Model Form(assume that $X_t$ and $Y_t$ are stationary)

+1

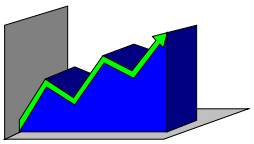


-1

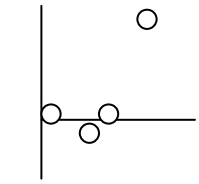
Numerator model with delay of 0

$Y_t = \omega 0 X_{t-0}$

Impulse Response Function Suggested Model Form(assume that $X_t$ and $Y_t$ are stationary)



+1

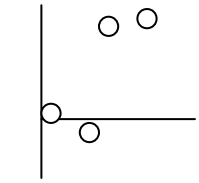


-1

Numerator model with delay of 3

$Y_t = \omega 0 X_{t-3}$

+1

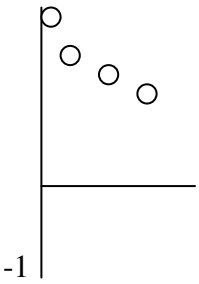


-1

Numerator model with delay of 2

$Y_t = (\omega 0 - \omega 1 B) X_{t-2}$

+1

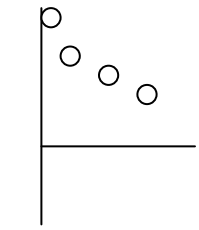


-1

Numerator model with denominator model with Delay of 0

$$Y_t = \frac{(\omega 0)}{(1 - \delta 1B)} X_{t-0}$$

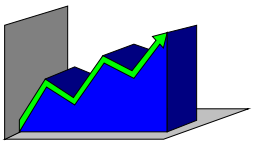
+1



-1

Numerator model with denominator model with Delay of 0

$$Y_t = \frac{(\omega 0)}{(1 - \delta 1B)} X_{t-0}$$



The above described prewhitening/cross correlation transfer function identification method works quite well for the single input - single output case. It also applies for multiple input models when the input series noise processes are uncorrelated. However, as Liu and Hanssens (1982) point out, generalized identification for the multiple input case is more difficult. They recommend a simultaneous identification procedure that uses common filters and least squares to estimate the impulse response weights. Given the estimates of the impulse response weights, one can identify the form of the combined transfer function - noise model. The process of model identification from a given set of impulse response weights was discussed in the section above and it will not be repeated here. What this section does cover is the Liu-Hanssens procedure for estimating these impulse response weights.

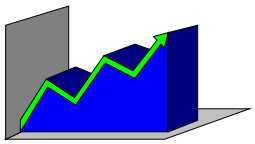
The first step is to build an ARIMA model for each of the stochastic time series. When we refer to each, we imply both the input and the output series. Each series should then be made stationary by applying the appropriate differencing and transformation parameters. After inducing stationarity, a first pass common filter must be identified. This common filter is obtained by examining the roots of the AR factors in each of the ARIMA models. If there are AR factors with roots close to 1, a common filter can be chosen from these factors. This filter is then applied to each of the stationary series.

A first pass estimate of the impulse response weights is computed via least squares. The developers suggest deleting nonsignificant lags from the regression. The residuals from the least squares estimation are then examined for structure. The ARIMA model for these residuals can be used as a second common filter. Better estimates of the impulse response weights can be obtained if a second common filter is applied to the filtered series. A final pass of least squares estimation for the appropriately filtered series produces the impulse response weights needed for model identification. Those weights can then be analyzed for clues as to the tentative model form.

## COMMON FILTER/LEAST SQUARES METHOD

This option to the program performs the entire-three step iterative Box-Jenkins modeling procedure for developing a transfer function model. The user must specify the time series variables that are to be included in the transfer function equation.

The program starts with the automatic (this is an option) development of each time series



variable's prewhitening model, using the Autobox automatic ARIMA modeling algorithm. It then moves on to the initial identification of the transfer function - noise model. This is accomplished by choosing the common filter, filtering each series and estimating the impulse response weights via the least squares method (described by Liu and Hanssens (1982)). The program then estimates the tentative model and performs all of the diagnostic checks for sufficiency, necessity and invertibility. The model is updated as needed, and the diagnostic checking stage ends when all of the criteria for an acceptable model are met. The final step is to generate the forecast values. The user controls the level of detail that the output report is to contain, as well some key options for modeling precision (lambda search and convergence criteria for example).

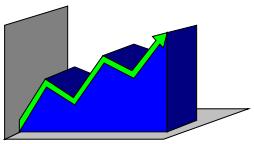
## MEASURING THE EFFECT OF UNSPECIFIED AND UNKNOWN ISOLATED EVENTS

Outliers and structure changes are commonly encountered in time series data analysis. The presence of the extraordinary events could and have misled conventional time series analysts resulting in erroneous conclusion. The impact of these events is often overlooked, however for the lack of a simple yet effective means to incorporate these isolated events. Several approaches have been considered in the literature for handling outliers in a time series.

We will first illustrate the effect of unknown events which cause simple model identification to go awry. We will then illustrate what to do in the case when one knows a priori about the date and nature of the isolated event. We will also point out a major flaw when one assumes an incorrect model specification. Then we introduce the notion of finding the intervention variables through a sequence of alternative regression models, yielding maximum likelihood estimates of both the form and the effect of the isolated event.

## SOME EXAMPLES OF THE EFFECT OF ISOLATED EVENTS ON IDENTIFICATION

Standard identification of ARIMA models uses the sample ACF as one of the two vehicles for model identification. The ACF is computed using the covariance and the variance. An outlier distorts both of these and in effect dampens the ACF by inflating both measures. Another problem with outliers is that they can distort the sample ACF and PACF by introducing spurious structure or correlations. For example, consider the circumstance where the outlier dampens the ACF:



$$\text{ACF} = \frac{\text{COVARIANCE}}{\text{VARIANCE}}$$

The net effect is to conclude that the ACF is flat; and the resulting conclusion is that no information from the past is useful. These are the results of incorrectly using statistics without validating the parametric requirements. It is necessary to check that no isolated event has inflated either of these measures leading to an "Alice in Wonderland" conclusion. Various researches have concluded that the history of stock market prices is information-less. Perhaps the conclusion should have been that the analysts were statistic-less.

Another way to understand this is to derive the estimator of the coefficient from a simple model and to evaluate the effect of a distortion. Consider the true model as an AR(1) with the following familiar forms:

$$\phi(B)\, Y_t = A_t \quad \text{or} \quad Y_t = \frac{A_t}{\phi(B)}$$

or

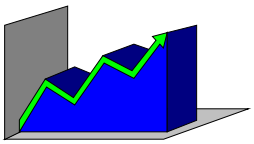
$$(1-\phi B)\, Y_t = A_t \quad \text{or} \quad Y_t = \phi Y_{t-1} \; A_t$$

the variance of Y can be derived as:

$$\sigma^2 Y_t = \phi^2 \sigma^2 Y_t + \sigma^2 A_t \text{ thus } \phi = \sqrt{\left[\frac{(1 - \sigma^2 A_t)}{\sigma^2 Y_t}\right]}$$

Now if the true state of nature is where an intervention of form $I_t$ occurs at time period t with a magnitude of W we have:

$$Y_t = \frac{A_t}{\phi(B)} + WI_t \quad \text{with}$$



true variance of Y plus it's intervention distortion

$$\sigma^2 Y_t = \phi^2 \sigma^2 Y_t + \sigma^2 A_t + (WI_t)^2$$

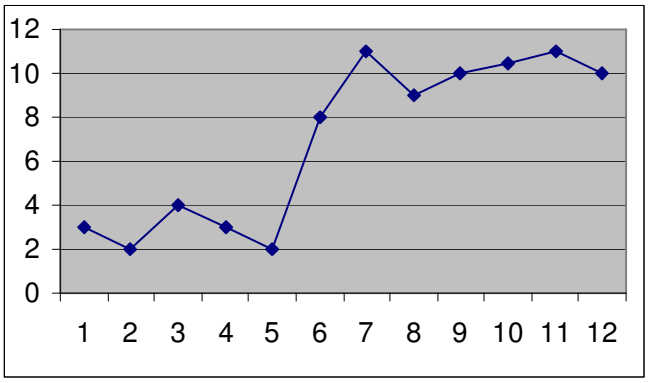
thus…

$$\phi = \sqrt{\left[\frac{(1 - (\sigma^2 A_t + (WI_t)^2)}{\sigma^2 Y_t}\right]}$$

The inaccuracy or bias due to the intervention is not predictable due to the complex nature of the relationship. At one extreme the addition of the squared bias to $\sigma^2 A_t$ would increase the numerator and drive the ratio to 1 and the estimate of $\phi$ to zero. The rate at which this happens depends on the relative size of the variances and the magnitude and duration of the isolated event. Thus the presence of an outlier could hide the true model.

Now consider another option where the $\sigma^2 Y_t$ is large relative to $\sigma^2 A_t$. The effect of the bias is to drive the ratio to zero and the estimate of $\phi$ to unity.

For example:

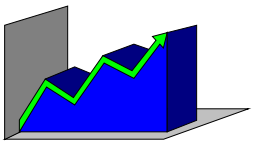



Step intervention

(permanent change in the mean)

would generate an ACF that did not die out slowly thus leading to a misidentified first difference model.

In conclusion, the effects of the outlier depend on the true state of nature. It can both incorrectly hide model form and incorrectly generate "evidence" of a bogus model.



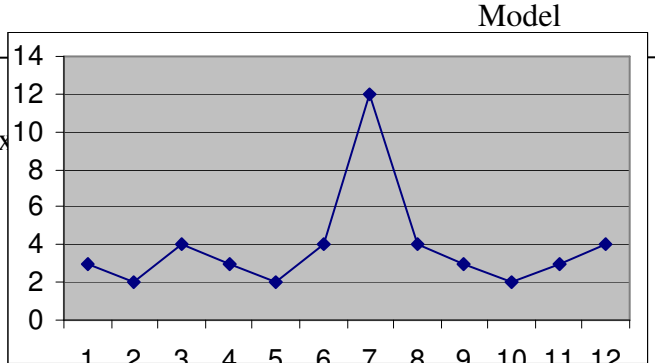
INTERVENTION MODELING OF KNOWN ISOLATED EVENTS

Pack's paper (1978) discussed a class of models and a model development process for time series influenced by identifiable isolated events. More commonly known as intervention analysis. This type of modeling is transfer function modeling with a stochastic output series and a deterministic input variable. The values of the input variable are usually set to either zero or one, indicating off or on. For instance, a time series disturbed by a single event, such as a strike or a price change, could be modeled as a function of its own past values and a "dummy" variable. The dummy input would be represented by a series of zeroes, with a value of one at the time period of the intervention. A model of this form may better represent the time series under study. Intervention modeling can also be useful for "what if" analysis - that is, assessing the effect of possible deterministic changes to a stochastic time series.

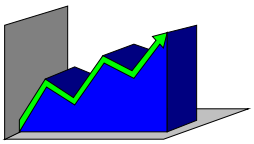
The identification procedure for models of this type is as follows. The first step is to develop an ARIMA model for the output series. This model should be developed for the pre-intervention series. Forecasts from this model can then be plotted against the actual values in order to determine the nature of the effect of the intervention. The form of the transfer function between the output series and the intervention variable is suggested by this effect. The figure shown below outlines the possible model forms for the various responses the output series makes to the given intervention variable. The analyst chooses the tentative transfer model form from this chart. The tentative noise model form is obtained from the ARIMA model developed for the output series. Given the combined multiple input model form, one can proceed to the estimation/diagnostic checking phase.

Example Intervention Models

Series Model




Autobox Troubleshooting

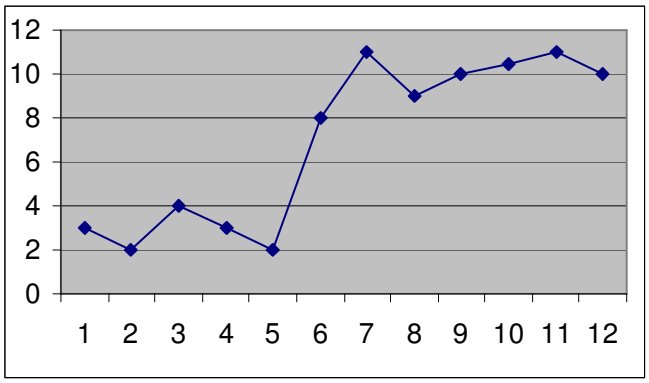


$$Y_t = \omega_0 P_t$$

$P_t$ is a pulse intervention

With dummy variable pattern

0,0,0,0,1,0,0,0,0,0,0,0,0

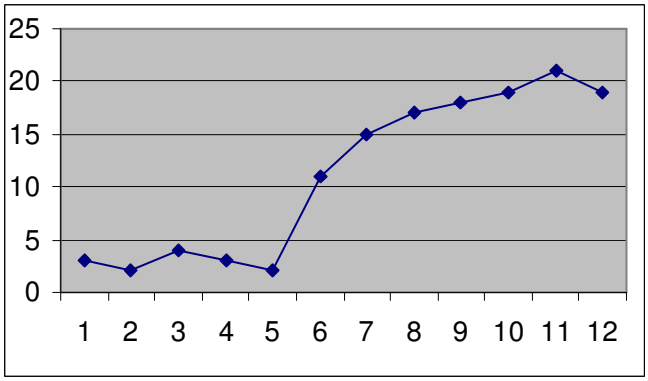



$$Y_t = \omega_0 L_t$$

$L_t$ is a level intervention

With dummy variable pattern

0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1



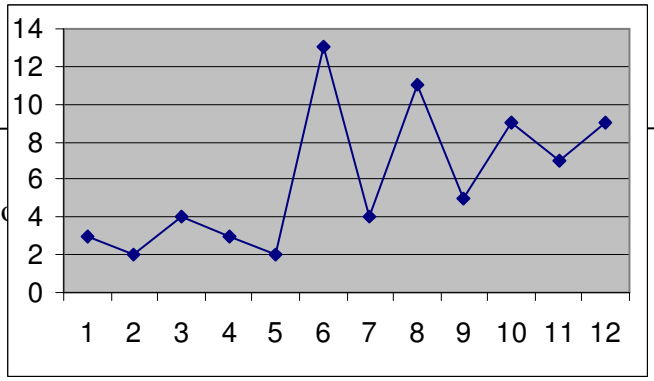

$$Y_t = \left[\frac{\omega_0 P_t}{1-cB}\right]$$

$P_t$ is a pulse intervention

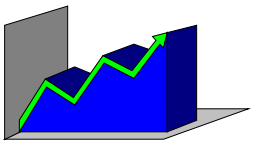
With dummy variable pattern

0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0

$\delta > 0$ and $\delta < 1$




$$Y_t = \left[\frac{\omega_0 P_t}{1-\delta B}\right]$$



$P_t$ is a pulse intervention

With dummy variable pattern

0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0

$\delta < 0$ and $\delta > -1$

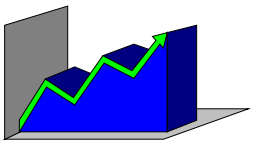
There is a major flaw associated with theory based models. It is called specification bias and the models on the last page suffer from it. Consider the assumption of a level shift variable starting at time period T. The modeler knows that this is the de jure date of the intervention. For example, the date that a gun law went into effect. If the true state of nature is that it took a few periods for the existence of the law to affect the behavior then no noticeable effect could be measured during this period of delay.

If you split the data into two mutually exclusive groups based upon theory, the test results will be biased towards no difference or no effect. This is because observations that the user is placing in the second group rightfully belong in the first group, thus the means then are closer than otherwise and a false conclusion can arise. This specification error has to be traded off with the potential for finding spurious significance when faced with testing literally thousands and thousands of hypothesis.

## INTERVENTION MODELING OF UNKNOWN ISOLATED EVENTS

Bell's paper (1983) describes a computer program for the routine identification of three types of outliers in a time series. These outliers can be represented as intervention variables of the forms: pulse, level shifts and seasonal pulses. The procedure for detecting the outlier variables is as follows. Develop the appropriate ARIMA model for the series. Test the hypothesis that there is an outlier via a series of regressions at each time period. Modify the residuals for any potential outlier and repeat the search until all possible outliers are discovered. These outliers can then be included as intervention variables in a multiple input B-J model.

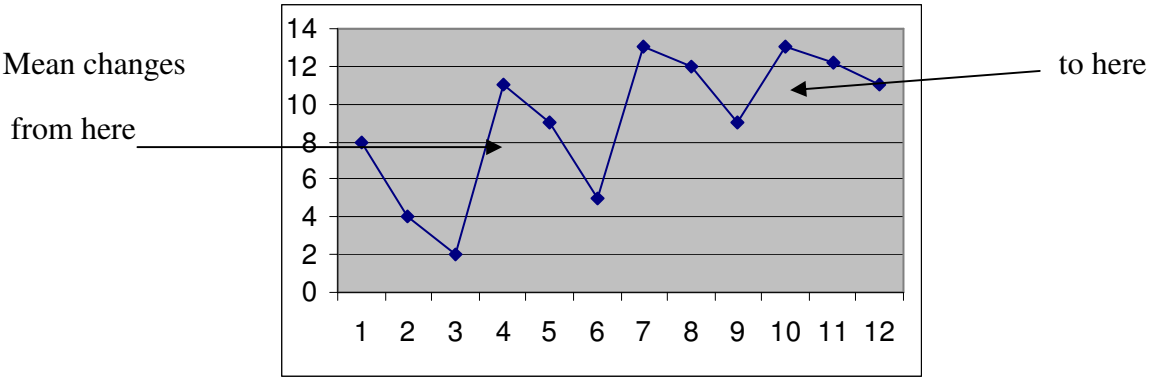
The noise model can be identified from the original series modified for the outliers. This option to the program provides a more complete method for the development of a model to forecast a univariate time series. The basic premise is that a univariate time series may not be homogeneous and therefore the modeling procedure should account for this. By homogeneous, we mean that



the underlying noise process of a univariate time series is random about a constant mean. If a series is not homogeneous, then the process driving the series has undergone a change in structure and an ARIMA model is not sufficient.

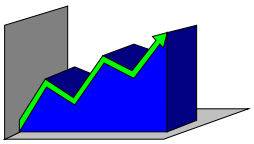
The Autobox heuristic that is in place checks the series for homogeneity and modifies the model if it finds any such changes in structure. The point is that it is necessary for the mean of the residuals to be close enough to zero so that it can be assumed to be zero for all intents and purposes. That requirement is necessary but it is not sufficient. The mean of the errors (residuals) must be near zero for all time slices or sections. This is a more stringent requirement for model adequacy and is at the heart of intervention detection. Note that some inferior forecasting programs use standardized residuals as the vehicle for identifying outliers. This is inadequate when the ARIMA model is non-null.

Consider the case where the observed series exhibits a change in level at a particular point in time.

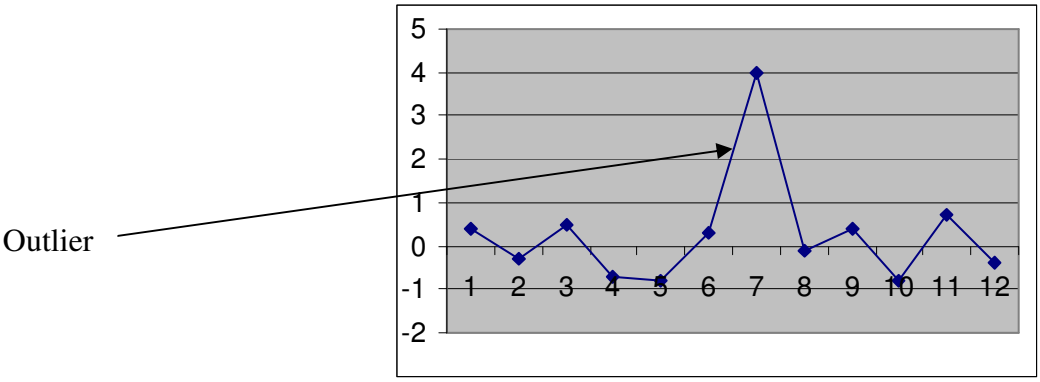



If you try to identify outliers or interventions in this series via classical standardized residuals you get:

$$A_t = \left\{ \frac{\Phi(B)}{\theta(B)} \right\} Y_t$$



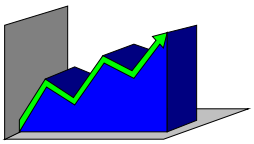
You would get a set of residuals which would look like:




thus identifying one outlier. The problem is that if you "fix" the bad observation at the identified time point, the subsequent value is identified as an outlier due to the recursive process. The simple-minded approach of utilizing standardized residuals is in effect identification of innovative outliers and not additive outliers.

The logic behind the automatic intervention procedure has its roots in the technique proposed by Chang and Tiao (1983) and programmed by Bell (1983). It starts by developing an ARIMA model for the univariate time series (using the automatic ARIMA algorithm). A series of regressions on the residuals from the ARIMA model checks for any underlying changes in structure. If the series is found to be homogeneous, then the ARIMA model is used to forecast. If the series is found to be nonhomogeneous, then the various changes in structure are represented in a transfer function model by dummy (intervention) input variables and the ARIMA model becomes the tentative noise model.

The program then estimates the transfer function-noise model and performs all of the diagnostic checks for sufficiency, necessity and invertibility. The model is updated as needed, and the diagnostic checking stage ends when all of the criteria for an acceptable model are met. The final step is to generate the forecast values. The user controls the level of detail that the output report is to contain, as well as some key options for modeling precision (lambda search and backcasting, for example). The user can also elect to have this process start with an examination of the original time series. This may be necessary for those cases where the series is



overwhelmingly influenced by outlier variables.

We now present a summary of the mathematical properties underlying this procedure. This is taken from the Downing and McLaughlin (1986) paper (with permission!). For purposes of this discussion, we present the following equation, which is the general ARIMA model:

$$\nabla \Phi_p (B) (N_t - \mu) = \theta_0 + \theta_q (B) A_t \qquad \text{(equation 1)}$$

where $N_t$ = the discrete time series,

$\mu$ = the mean of the stationary series,

$\nabla$ = the differencing factor(s),

$\Phi_p$ = the autoregressive factor(s),

$\theta_0$ = the deterministic trend,

$\theta_q$ = the moving average factor(s),

$A_t$ = the noise series,

and B = the backshift operator.

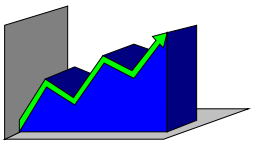
Outliers can occur in many ways. They may be the result of an error (i.e. a recording error). They may also occur by the effect of some exogenous intervention. These can be described by two different, but related, generating models discussed by Chang and Tiao (1983) and by Tsay (1986). They are termed the innovational outlier (IO) and additive outlier (AO) models.

An additive defined as, $t_o$

$$Y_t = N_t + W E_t \qquad \text{(equation 2)}$$

An innovational outlier defined as, $t_o$

$$Y_t = N_t + \frac{\theta(B)}{\Phi(B)} W E_t \qquad \text{(equation 3)}$$



where Y = the observed time series, t in length

W = the magnitude of the outlier,

$t_o$

$E_t$ = 1 if $t = t_o$,

= 0 if $t = t_o$,

that is, $E_t$ is a time indicator signifying the time occurrence of the outlier, and $N_t$ is an unobservable outlier free time series that follows the model given by equation 1. Expressing equation 2 in terms of white noise series $A_t$ in equation 1, we find that for the AO model

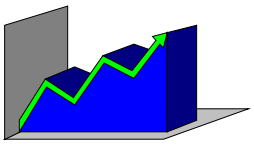
$$Y_t = \frac{\Theta(B)}{\Phi(B)} A_t \, W \, E_{t_o} \qquad \text{(equation 4)}$$

while for the the IO model

$$Y_t = \frac{\Theta(B)}{\Phi(B)} (A_t \, W \, E_{t_o}) \qquad \text{(equation 5)}$$

Equation 4 indicates that the additive outlier appears as simply a level change in the $t_o^{th}$ observation and is described as a o "gross error" model by Tiao (1985). The innovational outlier represents an extraordinary shock at time period to since it influences observations $Y_{t0}$ , $Y_{t0+1}$ ,... through the memory of the system described by:

$$\frac{\Theta(B)}{\Phi(B)}$$



The reader should note that the residual outlier analysis as conducted in the course of diagnostic checking is an AO type. Also note that AO and IO models are related. In other words, a single IO model is equivalent to a potentially infinite AO model and vice versa. To demonstrate this, we expand equation 5 to

$$Y_t = \frac{\Theta(B)}{\Phi(B)} A_t + \frac{\Theta(B)}{\Phi(B)} W\, E_{t_o} \qquad \text{(equation 6)}$$

and then express equation 6 in terms of equation 4

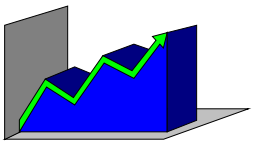
$$Y_t = \frac{\Theta(B)}{\Phi(B)} A_t + W^* E_{t_o} \qquad \text{(equation 7)}$$

$$W^* = \frac{\Theta(B)}{\Phi(B)} W$$

Due to estimation considerations, the following discussion will be concerned with the additive outlier case only. Those interested in the estimation, testing, and subsequent adjustment for innovative outliers should read Tsay (1986). Note that while the above models indicate a single outlier, in practice several outliers may be present.

The estimation of the AO can be obtained by forming:

$$II(B) = \frac{\Theta(B)}{\Phi(B)} = (1 - II_1B - II_2B - \ldots..) \qquad \text{(equation 8)}$$



and calculating the residuals $E_t$ by

$$E_t = \Pi(B)\, Y_t \qquad \text{(equation 9)}$$

$$= \Pi(B)\, \frac{(\Theta B)}{\Phi B}\, A_t + W\, E^{to}_t$$

$$= A_t + W\, \Pi(B)\, E^{to}_t$$

By least squares theory, the magnitude W of the additive outlier can be estimated by

$$\hat{W}_{t0} = n^2 \Pi(F) E_{t0} \qquad \text{(equation 10)}$$

$$= n^2 (1 - \Pi_1 F^2 - \Pi_2 F^2 - \Pi_{n-t0} F^{n-t0}) E_{t0}$$

$$n^2 = (1 + \Pi_1{}^2 + \Pi_2 F^2 + \Pi_{n-t0} F^{n-t0})\text{-}1$$

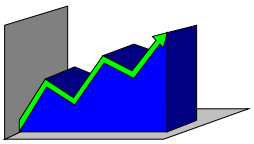
and F is the forward shift operator such that $Fe_t = E_{t+1}$ . The variance of $\hat{W}$ is given by:

$$\mathrm{Var}(\hat{W}_{t0}) = n^2 \sigma^2 \qquad \text{(equation 11)}$$

where $\sigma^2$ is the variance of the white noise(random) process $A_t$ .

Based on the above results, Chang and Tiao (1983) proposed the following test statistic for outlier detection:

$$\Upsilon_{t0} = \hat{W}_{t0} / n\, \sigma . \qquad \text{(equation 12)}$$



If the null hypothesis of no outlier is true, then $\Upsilon_{t0}$ has the to standard normal distribution. Usually, in practice the true parameters II and $\sigma^2$ are unknown, but consistent estimates exist. Even more important is the fact that to, the time of the outlier, is unknown, but every time point may be checked. In this case one uses the statistic:

$$\Upsilon = \max |\Upsilon_{t0}| \{t0:1 \leq t0 \leq n\} \qquad \text{(equation 13)}$$

and declares an outlier at time to if the maximum occurs at $t_o$ and is greater than some critical value C. Chang and Tiao (1983) suggest values of 3.0, 3.5 and 4.0 for C.

The outlier model given by Equation 4 indicates a pulse change in the series at time to. A step change can also be modeled simply by replacing $E_t^{to}$ with $S_t^{to}$ where:

$$S_t^{to} = 1 \text{ if } t \geqslant t_o \qquad \text{(equation 14)}$$

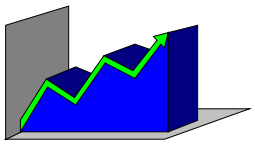
$$0 \text{ if not}$$

We note that $(1\text{-}B)S_t^{to} = E^{to}$. Using $S^{to}$ one can apply least squares to estimate the step change and perform the same tests of hypothesis reflected in Equations 12 and 13. In this way, significant pulse and/or step changes in the time series can be detected.

A straightforward extension of this approach to transfer functions has also been introduced in this version of Autobox with equation 9 being generalized to:

$$E_t = II(B)\left(\left[E_t - \Sigma \frac{\omega(B)}{\Delta(B)} X_t\right]\right)$$

this, of course, implies that the outliers or interventions are not only identified on the basis of the noise filter but the form and nature of the individual transfer functions.



HOW TO DETECT CHANGES IN VARIANCE

One of the critical assumptions underlying the tests of significance performed by Autobox is the "homogeneity of variance assumption". Essentially this requirement is met when the variance at each time point is constant. Since only one realization of the error is observed, standard practice is to pool errors into local groups and compare said variances.

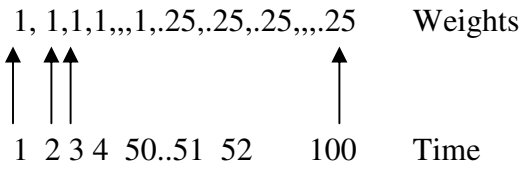
So, if an estimate of a model has 100 errors it would be possible to collect these errors into a number of groups and then perform F tests for these variances. An improvement over this approach is to sequentially alter the break points and then to identify which contrast provided the maximum F-value thus pointing to the optimal classification.

Consider the following set of errors where 100 observations were available:

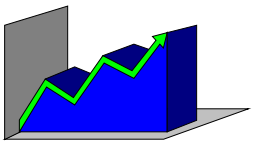
A(1),A(2),,,,,A(100) and a minimum interval is set at 10.

Compute a variance for the first 10 and for the last 90. Compute the ratio of these variances and store as F(11) where 11 represents the pointer to the second group. Now add the 11th error to group1 and delete it from group2. Now recompute the 2 variances and store the resultant F value as F(12). Perform this 88 more times finishing with 88 separately computed F-values. Select the maximum and determine it's associated probability value. If significant, then adjust the errors via the square root of the largest F value and perform the test again.

As an example, say the optimal break point was time period 51 where the variance in group 1 (1,,,50) was 1 and the variance in group 2 (51,52,,,100) was 4, leading to an F value of 4.0.




These weights are empirically constructed and may not be optimal insofar as they were identified based upon a specific model. That may or may not be critical. AFS does not think so, although individual cases might prove to the contrary.



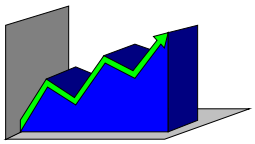
The program now re-estimates the model based upon these weights. If the user had started the process with a set of weights via the "Weights.In" file then the new weights will be used to refine the old weights.

Thus weighted least squares is the answer to variance heterogeneity that is not caused by mean level changes. The Box-Cox procedure (power transformations) normalizes the variance of the errors when the variance is proportional to shifts in the mean. The various power transformations test for linear dependence between dispersion and central tendency (e.g. Square root implies linear dependence between the standard deviation and the mean while logarithms imply a linear relationship between the variance and the mean). These kinds of fixups are clearly inadequate when variance heterogeneity exists but is not caused by mean shifts.

The iterative process of detecting multiple change points in the variance uses standard linear transformation techniques to obtain "adjusted data".

For example, if it has been proven that the variance in group 2 is 4 times that of group 1 then this implies that the observations in group 2 should be divided by the square root of 4. In this way, a set of residuals is obtained for which it can not be proved that there exists a statistically significant variance in any two consecutive groups.

Tsay, in his 1988 paper in the Journal of Forecasting, used a similar procedure to obtain an adjusted time series which had a homogeneous error structure. An adjusted series was then used for modeling purposes. The results are identical. We prefer the explicit incorporation of weights as they account for the lack of variance homogeneity in a way that allows cross-comparisons for a different time series.



4

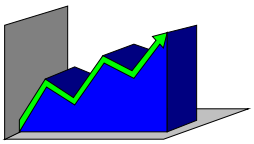
# Statistical Tests and Their Role in Autobox

Autobox allows the user to run a statistical test and to display that test for review.

1. Run the test, but don't display it's results. This is where you want to use the test, but you don’t want to see the details of the test in the output.
2. Run the test and display it's results. The test in the modeling process and you see the report so you can review it.
3. Don't run the test, but display it's results. This would be helpful to know whether it would have been useful to use the test in the analysis.
4. Don't run the test and don't display it's results. You get no output of the test and the modeling process does not use the test.

The user controls the output reporting & testing options )

| | Take action | Don't take action |
|---|---|---|
| Report results | Test done<br>Test reported<br>Take action | Test done<br>Test reported |
| Don't report results | Test done<br>Take action | Test not done |



These are the tests available:

1 model necessity with respect to estimated parameters

2 model necessity with respect to invertibility

3 automatic fixup with respect to detecting need for

a) deterministic linear component

4 b) deterministic seasonal dummies

5 model sufficiency m-level fixup w/respect to adding ARIMA structure

6 model sufficiency m-level fixup with respect to adding outliers

7 model sufficiency v-level fixup w/respect to incorporating weights

* If a test is not to be reported and no action is to be taken then no test is done.

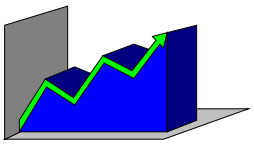
The user can select the particular specific path of tests desired and the level of confidence for each test, thus customizing the "expert system":

1 Enable the automatic fixup for necessity (model necessity with respect to estimated parameters) - STEPDOWN

2 Enable the automatic fixup for invertibility (model necessity with respect to invertibility) - SIMPLIFICATION

3 Enable automatic fixup for Deterministic Linear (is the model a linear trend?) - SIMPLIFICATION

4 Enable automatic fixup for Seasonal Dummies (should the seasonal difference be replaced by seasonal dummies) - SIMPLIFICATION



5 Enable the automatic fixup for sufficiency (model sufficiency m-level fixup with respect to adding ARIMA structure) - STEPUP

6 Enable the automatic fixup for outliers (model sufficiency m-level fixup with respect to adding outliers) - STEPUP

7 Enable the automatic fixup for variance change - STEPUP

When Box & Jenkins developed their integrated, self-checking procedure for constructing a model, computers were slow and cranky. In particular, estimation procedures required massive CPU's which were not available like they are today. This being true, they sub-divided model "construction" into three steps:

1. Initial identification
2. Estimation of the suggested model
3. Diagnostic checking

The third step was really a re-identification and focused on various tests of necessity, sufficiency and invertibility, thus we can simplify it to:

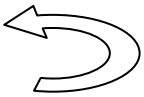
3 Steps: 1. Initial identification

2. Estimation of the suggested model

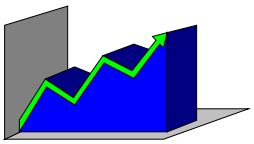
3. Re-identification

Steps 2 and 3 were repeated until the model met all the requirements thus the iterative aspect was explicitly introduced.

3 Steps: 1. Initial identification

2. Estimation of the suggested model 

3. Re-identification



In initial identification, they suggested a procedure which was based on capturing the pattern suggested by the sample ACF's and PACF's thus no model was assumed. These could be computed quickly. However, under closer inspection these ACF's and PACF's could be viewed as the residuals (a) from an assumed model of the form:

$$Y_t - \mu = A_t$$

Thus there were really 4 steps:

0. Estimation of an initially assumed model
1. Initial identification
2. Estimation of the suggested model
3. Re-identification

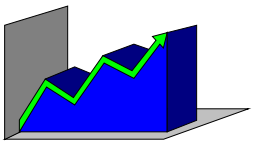
The initially assumed model only included the mean and they argued implicitly that this was a good place which to improve upon. This is an idea that has stood the test of time. It is always possible and, more often than not, incorrect to a priori assume more complex model forms because of the inability to recover from the poor start. There are, however, many cases where an astute selection of an initially assumed model can improve the overall performance of statistical procedures. It suffices to say that there is always an assumed starting model.

The 4 step procedure can be simplified into a repetitive 2 step procedure:

a. Estimation of a model via some efficient procedure

b. Re-identification via model testing

which is exactly what all modelers do and should do in their work. Let us now focus on the second step which speaks to the question, "Is this model good enough?".



MODEL TESTING

We can test a model in two different ways:

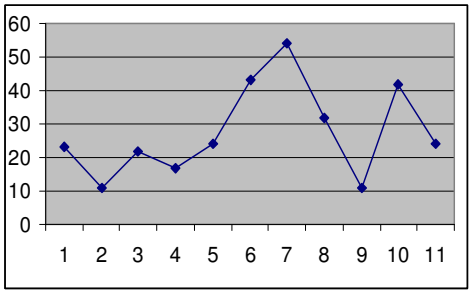
1) Within sample

2) Out of sample

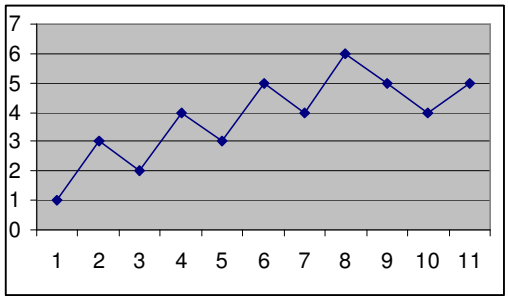
Philosophically, the proof that a model converts a realized or observed series into a predictable component and a random component is strategy 1 while the less objective, "How well does it forecast?", approach is consistent with strategy 2.

With strategy 1, we attempt to prove that:

| RANDOM COMPONENT | + | PREDICTABLE COMPONENT | = | OBSERVED COMPONENT |
|---|---|---|---|---|




THE MODEL

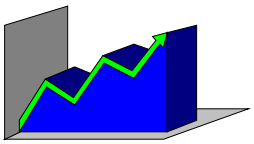



residuals which can be described by a N ($\mu$,$\sigma^2$)

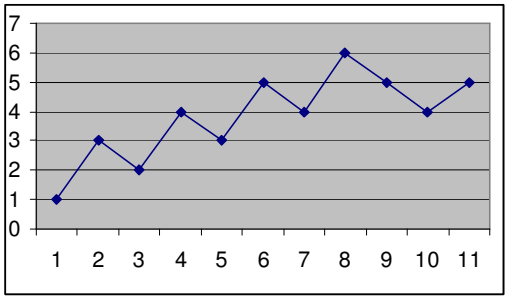
where $\mu$= 0 and $\sigma^2$ is i.i.d

observed time series created by perturbing by the random numbers by the model form

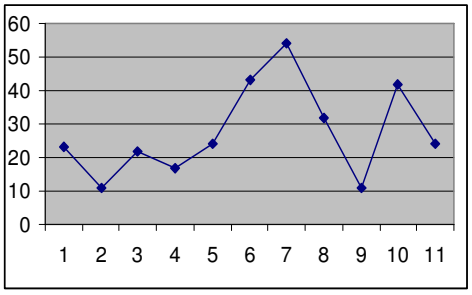
**or**



| OBSERVED SERIES | = | PREDICTABLE COMPONENT | + | RANDOM COMPONENT |
|---|---|---|---|---|




THE MODEL




observed time series created by perturbing by the random numbers by the model form

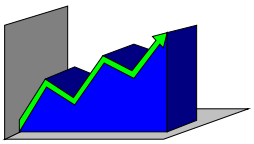
residuals which can be described by a N ($\mu$,$\sigma^2$) where $\mu$= 0 and $\sigma^2$ is i.i.d

The observed time series spawns two derivatives each of which offer opportunities for assessing validity using the within sample approach:

a) The model

b) The residuals

The model asks if we need (necessity) all of the coefficients in the equation or can we simplify it by elimination or deleting (step down) one or more of them? This is essentially a one dimensional test. Is the model invertible or should it be re-parameterized to be so. Lack of invertibility can usually be diagnosed as deficiencies in either included stochastic structure or omitted deterministic structure.



The residuals ask is the model sufficient or complete enough? Are there any indications that something has been omitted from the model? This test is then a multi-dimensional test insofar as there can be a number of ways to improve or add needed structure (step forward). The current known set (ever-expanding) of possible model improvements can be classified as classified as augmentations due to three kinds of violations.

$\mu$ violations known as m-level fixups

$\sigma^2$ violations known as m-level fixups

$\theta(B)$ Dynamics either in form or parameters for a constant form

$\phi(B)$

TESTING FOR THE PRESENCE OF LINEAR TREND AND/OR SEASONAL DUMMIES

If the true underlying model is:

$$Y_t = W0 + W1T + A_t$$

where T=1,2,3,etc.

then it can appear to have the same characteristics as:

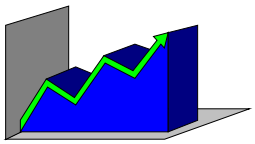
$$(1 - B)\, Y_t = W0_t + (1 - \theta B)\, A_t \quad \text{with } \theta = 1.$$

Let's walk through the use of how to incorporate time to understand how these two models are related:

Traditional time oriented regression analysis is usually presented in the following form:

$$Y_t = B0 + B1 * T + A_t$$

where B0 is the intercept and B1 is the trend or slope.



3 Possibilities:

1 Deterministically $Y_t = B0 + B1*T$
2 Stochastically $Y_t = B0 + B1*Y_{t-1}$
3 Both $Y_t = B0 + B1*T + B2*Y_{t-1}$

Deterministic model

It is conventional in econometric model building to use polynomials and dummy variables to describe "trends". Such methods are unsatisfactory since, it is unlikely that such deterministic trends are adequate to describe the development of observed time series, implying as they do that "growth rates remain constant indefinitely".

Stochastic model

By incorporating differences into the model one can capture stochastic or adaptive trends into the model.

Example of a deterministic model

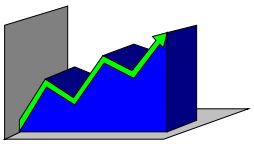
$$Y_t = B0 + B1_t \ X_t + A_t$$
where $X_t = 1, 2, 3, \ldots t$

The forecast for time period t+k is independent of the most recent observation. Even if one recomputed the model coefficients, the impact on these could be quite insignificant as every observation is equally unimportant.

Example of a stochastic model
( that identifies a deterministic model )

$$Y_t = B0 + Y_{t-1} + A_t - A_{t-1}$$
$$Y_t - Y_{t-1} = B0 + + A_t - A_{t-1}$$
$$Y_t (1-B) = B0 + + A_t (1-B)$$

The stochastic model identifies the need for the deterministic model relationships, both implicit and explicit.



Our purpose here will be to highlight the relationships, both implicit and explicit, in time series models. The issue here is to show how a deterministic trend model can be expressed as particular ARIMA model, with a root on the unit circle:

consider $Y_t = W0 + W1\ X_t + A_t$

where X is the set of natural numbers 1,2,...t, thus we can write:

$$Y_t = W0 + W1\ t + A_t$$

Thus, if we use a 1 period lag then:

$$Y_{t-1} = W0 + W1\ [t-1] + A_t$$

and then we get: $t = \frac{[\ Y_t - A_t - W0 + W1]}{W1}$

substituting for t we get:

$$Y_t - Y_{t-1} = W0 + A_t\ A_{t-1}$$

or $(1 - B)\ Y_t = W0 + (1 - \theta B)\ A_t$

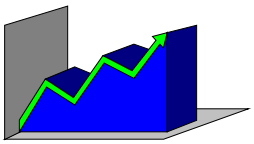
or in general as an ARIMA model where V is a generalized function

$Y_t = V [\ Y_{t-1}\ ]$ or $Y_t = \phi 1(B)\ Y$ or $\phi 1(B) = 1 - \phi(B)$

M-LEVEL FIXUPS

Here are the two types of level shifts that can be identified and corrected for:

- ARMA or stochastic model improvements as indicated by the sample ACF and PACF of the residuals.



- Deterministic model improvements are indicated by a fixed pattern in the residuals. A simple example is when a local mean of the residuals is statistically significantly different from another local mean (i.e. the residuals mean has shifted). This leads to the augmentation to the model of a step intervention starting at time period i through time period j. Intervention variables can also be pulse or be seasonal pulses and may start and at different points in time. An overall test for the residual series mean versus 0, is formerly a test of a deterministic structure, we can incorporate that test into the ARMA procedure.

V-LEVEL FIXUPS

Here are the two types of variance that can be identified and corrected for:

- Power transformations which might eliminate a possible linear relationship between the local residual $\mu$ and $\sigma$. As an example: a logarithmic transform to homogenize the variance of the residuals. Improvements in this are termed fractional models.

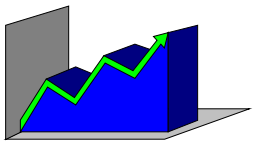
- Variance heterogeneity caused by a time varying but a locally predictable $\sigma^2$. A simple example of this might be excessive volatility in a series due to an identified change in the variance. F tests can be used to identify the change point.

$\frac{\theta(B)}{\phi(B)}$ Dynamics either in form or parameters for a constant form

Tests and procedures for identifying and fixing models for these issues are unfinished and represent thorny research problems.

In summary, there are three basic diagnostic checks that must be performed on the estimated model. These tests are for necessity, invertibility and sufficiency. Each parameter included in the model should be statistically significant (necessary) and each factor must be invertible. In addition, the residuals from the estimated model should be white noise (model sufficiency).

The test for necessity is performed by examining the T-ratios for the individual parameter estimates. Parameters with nonsignificant coefficients should be deleted from the model. Invertibility is determined by extracting the roots from each factor in the model. All of the roots must li.e. outside of the unit circle. If one of the factors is noninvertible, then the model must be adjusted. The appropriate adjustment is dictated by the type of the factor that is noninvertible. For example, a noninvertible autoregressive factor usually indicates under-differencing, while a noninvertible moving average factor may indicate over-differencing. The model fixup for noninvertibility is not always clear-cut. Careful re-analysis of each modeling step may be



necessary.

If all of the parameters are necessary and invertible, then the model is sufficient and can be used for forecasting.

SAMPLE VALIDATION

One can retain a set of observations from the analysis and assess various measures of forecast accuracy. This approach to model validation is very subjective and falls prey to the dictatorial effect of the number retained. If one retains n1 observations this usually leads to model "A" while if you retain n2 observations you get model "B". In some cases this approach is desired and we offer ways to measure the accuracy:

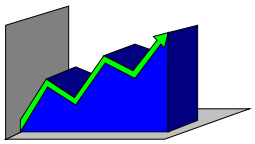
- Forecast mean deviation (mean)
- Forecast mean percent error
- Forecast mean absolute deviation
- Forecast mean absolute % error
- Forecast variance (precision)
- Forecast bias squared (reliability)
- Forecast mean square error (Accuracy)

The following two tables show the table of forecasts and an analysis of the forecast errors from origin 26.

Model Forecasting for Time Series Variable Y = PU

ORIGIN 26 1998/ 2

| TIME (T) | DATE | LOWER 80% LIMIT | UPPER 80% LIMIT | FORECAST | ACTUAL (IF KNOWN) | RESIDUAL | % ERROR |
|---|---|---|---|---|---|---|---|
| 27 | 1998/3 | .2827E+05 | .3336E+05 | .3081E+05 | .3967E+05 | .886E+04 | 22.33 |
| 28 | 1998/4 | .3016E+05 | .3538E+05 | .3277E+05 | .3783E+05 | .507E+04 | 13.39 |
| 29 | 1998/5 | .3655E+05 | .4325E+05 | .3990E+05 | .4724E+05 | .734E+04 | 15.54 |
| 30 | 1998/6 | .3203E+05 | .3898E+05 | .3551E+05 | .3698E+05 | .147E+04 | 3.98 |
| AGGREGATE | | .1356E+06 | .1424E+06 | .1390E+06 | .1617E+06 | .227E+05 | 16.36 |



ACCURACY STATISTICS FOR ORIGIN 26 1998/ 2

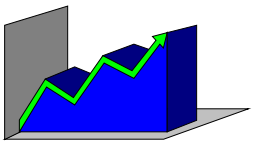
VALUES ARE IN TERMS OF THE ORIGINAL METRIC

| | | |
|---|---|---|
| | # of Actuals (N) | 4 |
| a) | Forecast Mean Deviation (Bias) | 5683.53 |
| b) | Forecast Mean Percent Error | 13.8080 |
| c) | Forecast Mean Absolute Deviation | 5683.53 |
| d) | Forecast Mean Absolute % Error | 13.8080 |
| e) | Forecast Variance (Precision) | .773586E+07 |
| f) | Forecast Bias Squared (Reliability) | .323026E+08 |
| g) | Forecast Mean Square Error (Accuracy) | .400384E+08 |
| h) | Relative Absolute Error | .549626 |

We will now define all of these terms so that you can know how they were computed.

R = A - F and N = NAIVE FORECAST

| | | | |
|---|---|---|---|
| a) | Forecast Mean Deviation (Bias) | | $[\Sigma(R)/N]=B$ |
| b) | Forecast Mean Percent Error | | $\Sigma(R/A)/N$ |
| c) | Forecast Mean Absolute Deviation | | $\Sigma(ABS(R))/N$ |
| d) | Forecast Mean Absolute % Error | | $\Sigma(ABS(R/A))/N$ |
| e) | Forecast Variance | (Precision) | $\Sigma(R-B)^2/N$ |
| f) | Forecast Bias Squared | (Reliability) | $B^2$ |
| g) | Forecast Mean Square Error | (Accuracy) | $\Sigma(R)^2/N$ |
| h) | Relative Absolute Error | | $\Sigma\lvert(A-F)\rvert/\Sigma(A-N)\rvert$ |



ACCURACY = PRECISION + RELIABILITY

THE PROGRAM THEN REPORTS THESE TWO TABLES FOR ORIGIN 27

ORIGIN 27 1998/ 3

| TIME (T) | DATE | LOWER 80% LIMIT | UPPER 80% LIMIT | FORECAST | ACTUAL (IF KNOWN) | RESIDUAL | % ERROR |
|---|---|---|---|---|---|---|---|
| 28 | 1998/4 | .3137E+05 | .3812E+05 | .3474E+05 | .3783E+05 | .309E+04 | 8.17 |
| 29 | 1998/5 | .4377E+05 | .5068E+05 | .4722E+05 | .4724E+05 | 16.6 | .04 |
| 30 | 1988/6 | .3423E+05 | .4311E+05 | .3867E+05 | .3698E+05 | -.170E+04 | -4.59 |
| 31 | 1988/7 | .3846E+05 | .4767E+05 | .4307E+05 | | | |
| AGGREGATE | | .1592E+06 | .1682E+06 | .1637E+06 | | | |

ACCURACY STATISTICS FOR ORIGIN 27 1998/ 3

VALUES ARE IN TERMS OF THE ORIGINAL METRIC

| | | |
|---|---|---|
| | # of Actuals (N) | 3 |
| a) | Forecast Mean Deviation (Bias) | 469.327 |
| b) | Forecast Mean Percent Error | 1.20294 |
| c) | Forecast Mean Absolute Deviation | 1601.47 |
| d) | Forecast Mean Absolute % Error | 4.26474 |
| e) | Forecast Variance (Precision) | .392311E+07 |
| f) | Forecast Bias Squared (Reliability) | 220268. |
| g) | Forecast Mean Square Error (Accuracy) | .414338E+07 |
| h) | Relative Absolute Error | .396938 |

These two tables are repeated for time periods 28 and 29 providing full information regarding forecasting accuracy. Autobox now summarizes the forecasting accuracy by lead time.

TABLE 1 : FORECAST ACCURACY STATISTICS AT VARIOUS LEAD TIMES

| LEAD TIME | MEAN DEVIATION (BIAS) $\mu=\Sigma(A-F)/n$ | MEAN % ERROR $\Sigma[(A-F)/A]/n$ | MEAN ABSOLUTE DEVIATION $\Sigma\|A-F\|/n$ | MEAN ABSOLUTE % ERROR $\Sigma[\|(A-F)/A\|]/n$ |
|---|---|---|---|---|
| 1 | .179314E+04 | .449392E+01 | .418065E+04 | .107531E+02 |
| 2 | .276783E+03 | .641870E+00 | .311165E+04 | .830854E+01 |
| 3 | .282041E+04 | .547135E+01 | .451862E+04 | .100640E+02 |
| 4 | .147108E+04 | .397842E+01 | .147108E+04 | .397842E+01 |

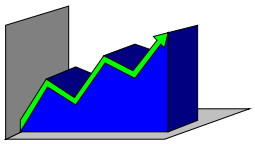


TABLE 2 : FORECAST ACCURACY STATISTICS AT VARIOUS LEAD TIMES

| LEAD TIME | VARIANCE (PRECISION) $\sigma^2=\Sigma[(A-F)-\mu]^2/n$ | BIAS SQUARED (RELIABILITY) $\mu^2$ | MEAN SQUARE ERR (ACCURACY) $\Sigma(A-F)^2/n$ | RELATIVE ABSOLUTE ERROR $\Sigma\|(A-F)\|/\Sigma\|(A-N)\|$ |
|---|---|---|---|---|
| 1 | .231070E+08 | .321536E+07 | .263223E+08 | .537811E+00 |
| 2 | .145060E+08 | .766090E+05 | .145826E+08 | .577375E+00 |
| 3 | .204180E+08 | .795469E+07 | .283726E+08 | .455348E+00 |
| 4 | .000000E+00 | .216408E+07 | .216408E+07 | .213621E+00 |

WHY WE FILTER

Univariate Box-Jenkins is a time series modeling process which describes a single series as a function of its own past values. The purpose of the B-J process is to find the equation (or filter) that reduces a time series with underlying structure to white noise. Since the filter accounts for the predictable portion of the time series, it can be used to forecast future values of the series.

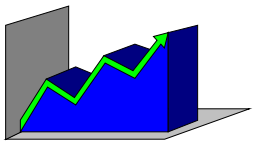
Problems that arise:

$$y = k + \sum_{j=1}^{N} \sum_{k=0}^{lj} v_{jk} \, x_{j,t-k} + n_t$$

→ 1) Dependence in the $n_t$

→ 2) Multicollinearity affecting coefficient estimates

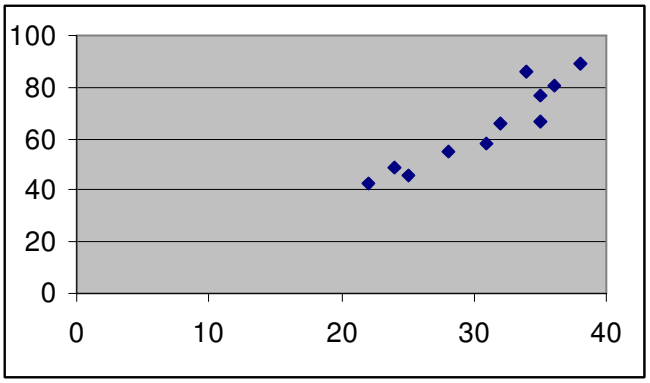
→ 3) Which $v_{jk}$ are initially presumed to be potentially significant



Ordinary correlation tests are misleading when using time series

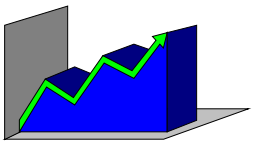
- Standard regression or OLS procedures fall prey to within correlation inducing the among correlation.
- Transfer function methodology would eliminate the within before testing the among concluding that although x and y are highly correlated there is no additional statistical information in x that does not exist in the past of y. Thus a simple ARIMA model (self-extrapolating) can be a better alternative. As Bartlett said in his 1935 paper, "correlations between time series can be largely due to chance alone". A common filter that we use everyday is a pair of glasses. The glasses represent an equation which permit you to see the real world as it exists. The primary purpose of a prewhitening filter is to eliminate the distortion brought by concomitant variables thus identifying the important variables in a transfer function model.

Consider the following analysis of two time series. We have plotted y versus x and have computed the cross-correlation between the two series. The relationship is strong and in some quarters would be used as the basis for some new "truths". Let us unravel how these series were generated, thus exposing the hidden flaw underlying regression or ordinary least squares.




Cross-corrlation
x lag

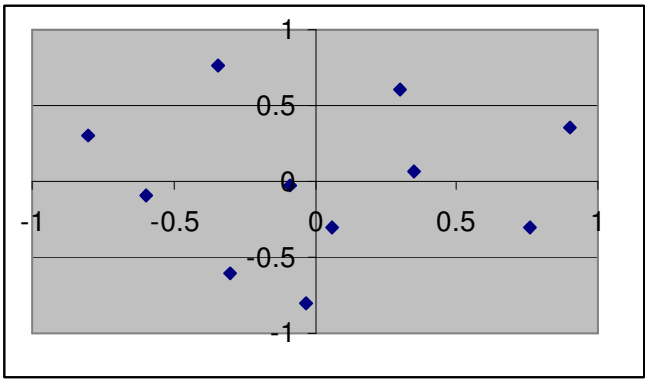
| | |
|---|---|
| 0 | .9831 |
| 1 | .9279 |
| 2 | .8784 |



We drew **two** sets of random numbers and called them Series1 and Series2 from a distribution with a mean of 0 and a standard deviation of 1.

| Series1 | Series2 |
|---|---|
| .03844 | .44343 |
| -.17814 | -.27347 |
| -.12518 | .29268 |
| -.97036 | -.08337 |
| -.01707 | .02067 |
| -.65135 | .11646 |
| .33108 | .32120 |
| .28320 | .10612 |
| .09822 | -.44640 |
| -.22822 | -.38033 |

To confirm the independence of the draw, we show the following scatter plot and statistics.

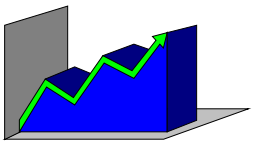



Cross-corrlation

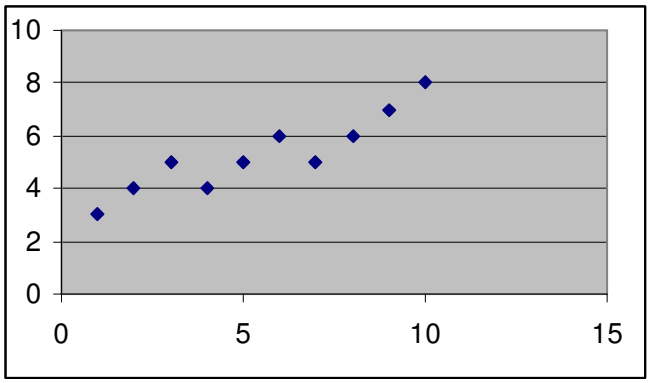
| | y lag | x lag |
|---|---|---|
| 0 | .0528 | .0528 |
| 1 | -.1943 | -.0252 |
| 2 | -.0026 | -.1014 |

We now construct a series called "x" and "y" using Series1 and Series2 with a trend added to each.

| Series1 | Series2 | trend | y | x |
|---|---|---|---|---|
| | | | Series1 + trend | Series2 + trend |
| .05988 | .44343 | 1 | .94012 | 1.44343 |
| -.17814 | -.27347 | 2 | 1.82186 | 1.72653 |
| -.12518 | .29268 | 3 | 2.87482 | 3.29268 |
| -.97036 | -.08337 | 4 | 3.02964 | 3.91663 |
| -.01707 | .02067 | 5 | 4.98293 | 5.02067 |
| -.65135 | .11646 | 6 | 5.34865 | 6.11646 |
| | | . | . | |
| .28320 | .10612 | 58 | 58.28320 | 58.10612 |
| .09822 | -.44640 | 59 | 59.09822 | 58.55360 |
| -.22822 | -.38033 | 60 | 59.77168 | 59.61967 |



If we plot Series x versus the trend we see a strong relationship and the same holds for true for Series y. The point of this exercise is to show that other variables (like time/trend) can cause two series to be seemingly related when it is actually another variable that is driving the relationship. If you misdiagnose that there is a time trend that is the cause, you could mistakenly conclude a real relationship when non exists or otherwise known as a "false positive".




INDEPENDENCE TO DEPENDENCE

We generated two uncorrelated random normal processes and called them Series1 and Series2. Their correlation structure confirmed the independence of the draw.

We created a series called tren by generating the counting numbers 1,2,3,,,60 . We then computed two series (called y and x) by adding the trend to each of the respective series.

Y = trend + Series1 x = trend + Series2

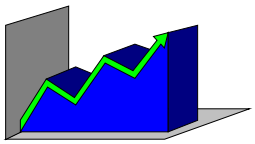
The correlation between these two series was computed to be .9831. In the absence of knowledge as to how y and x were generated, how would you have analyzed the "effect" that one had on the other? The answer to that dictates the need for more powerful methods. This is why you filter:

$Y_t$ ..... $X_t$

.

.

$X_{t-j}$



A common filter that we use everyday is a pair of glasses. The glasses represent an equation which permit you to see the real world as it exists. The primary purpose of a prewhitening filter is to eliminate the distortion brought by concomitant variables thus identifying the important variables in a transfer function model

Consider a transfer function

$$Y_t = V0\, X_t + V1\, X_{t-1} + V2\, X_{t-2} .... + Vj\, X_{t-j} + N_t$$

$$Y_t = \left\{ \frac{\omega(B)}{\delta(B)} \right\} X_t + \left\{ \frac{\theta(B)}{\Phi(B)} \right\} A_t$$

where an ARIMA model for X exists such that

$$\frac{\alpha(b)}{\beta(b)} X_t = \Upsilon_t$$

or

$$X_t = \frac{\beta(b)}{\alpha(b)} \Upsilon_t$$

or more generally with a pure delay of b periods

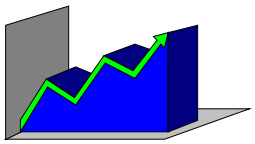
$$Y_t = \left\{ \frac{\omega(B)}{\delta(B)} \right\} X_{t-b} + \left\{ \frac{\theta(B)}{\Phi(B)} \right\} A_t$$

where b = the number of periods of pure delay before y responds to

$\{\theta(b)/\Phi(b)\}$ = ARMA model for unobserved series $A_t$

$\omega(b)$ = input lag structure reflecting static relationship of Y to X and is a polynomial of order s-1

$\delta(b)$ = output lag structure reflecting dynamic relationship of Y to X and is a polynomial of order r



and substituting the stationary Y' and X' for Y and X we get:

$$Y'_t = \left\{ \frac{\omega(B)}{\delta(B)} \right\} X'_{t-b} + \left\{ \frac{\theta(B)}{\Phi(B)} \right\} A_t$$

If we now identify the following ARMA filter for X:

$$\frac{\alpha(b)}{\beta(b)} X'_t = \Upsilon_t$$

If we now multiply the equation above by $\frac{\alpha(b)}{\beta(b)}$

$$\frac{\alpha(b)}{\beta(b)} Y_t = \left\{ \frac{\omega(B)\alpha(b)}{\delta(B)\ \beta(b))} \right\} X_{t-b} + \left\{ \frac{\theta\alpha(b)(B)}{\Phi\beta(b)\ (B)} \right\} A_t$$

and

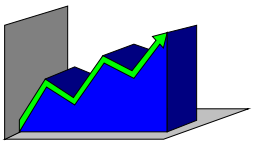
$$\frac{\alpha(b)}{\beta(b)} X'_{t-b} = \Upsilon_t$$

to yield:

$$\frac{\alpha(b)}{\beta(b)} Y'_t = \left\{ \frac{\omega(B)}{\delta(B)} \right\} \Upsilon_t + \left\{ \frac{\theta\ \alpha(b)(B)}{\Phi\beta(b)\ (B)} \right\} A_t$$

which gives us:

$$\varsigma_t = \left\{ \frac{\omega(B)}{\delta(B)} \Upsilon_{tb} \right\} + N_t$$



where $\Upsilon_{tb}$ is N.I.I.D. Thus we get a clear picture of the true relationship between X and Y:

$$\frac{\omega(B)}{\delta(B)}$$

This explains the need for filtering and exposes the relationship of the prewhitening filter in identifying the form of the inter-relationship between observed series. Note: After prewhitening the Seriesy/Seriesx data, the correlation between the two surrogate series was null thus concluding that the x series did not substantially improve the prediction of the y series, given that the past of y is included in the model.

SHUFFLE ALONG

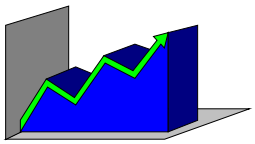
This section discusses the fatal flaw in applying traditional cross-sectional regression methods to time series data.

"A particular use of regression is to forecast time series variables and/or to relate time series variables to each other." Unfortunately, many time series regression analyses are invalid because they violate a number of the basic regression assumptions. They violate assumptions because time series data have a special characteristic: they always come in a predetermined sequence. This mandatory sequencing of data is not an issue in most non-time series analysis.

Most time series data tend to flow or drift in somewhat cohesive patterns over time. This flowing behavior wreaks havoc on a number of the regression assumptions, such as the requirement for uncorrelated residuals. In the time series context, the common term for the correlation is autocorrelation. Autocorrelation can be particularly strong in time series.

Using time as the independent variable is almost always a bad practice, even though it is still common. The resulting projections have a much higher probability of being wrong and will almost certainly provide extremely limited insight into the behavior of the dependent variable. The resulting R-squared is usually very high, but virtually meaningless.

True time series regression (transfer function) techniques are quite involved, but they operate on the same principles discussed here. An important part of the technique is to render the variables "stationary" over time - to make them look like a good residuals graph - before beginning the



regression. This is done with differencing of various types and mathematical transformations as well as with more esoteric techniques. A good rule is to examine the residuals and modify your differencing and transformation schemes until the residuals conform to the rules."

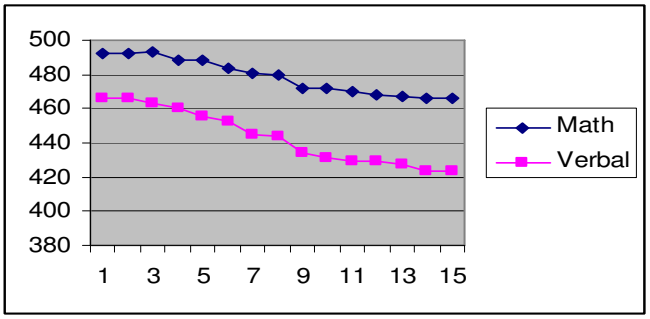
Early researchers would either difference or detrend their data prior to estimating a model. These were ad hoc attempts to pre-whiten or filter the data for purposes of creating surrogate series which could be modeled using standard cross-sectional regression procedures. An ARIMA model is a superset of these ad hoc filters.

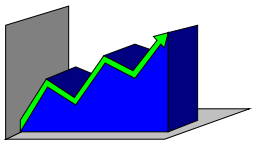
Galton developed regression for independent samples, each sample had potentially correlated measurements. Suppose, for example, we had observed 15 independent samples of math SAT scores. Suppose we had sampled 15 different countries (cross-sectional analysis), it would then be reasonable to equally weight all 15 samples.

Sometimes statisticians build weighted regression models where they apply different weights or believability to each observation. This feature (weighted ARIMA/TF models) is related but a different issue than the issue that concerns us here. We will now illustrate the graphical relationship and then compute the parameters of the model,

$$Y_t = W0 + W1\ X_t + A_t$$

and then illustrate the effect of "shuffling the data"



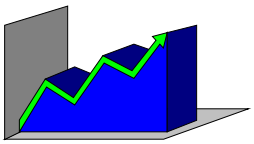


## CROSS-SECTIONAL DATA

| | | CHARACTERISTICS (MEASUREMENTS) A | B | ... Z |
|---|---|---|---|---|
| U | INDEPENDENT SAMPLE 1 | X1A | X1B | X1Z |
| N | | | | |
| C | INDEPENDENT SAMPLE 2 | X2A | X2B | X2Z |
| O | | | | |
| R | . | . | . | . |
| R | . | . | . | . |
| E | . | . | . | . |
| L | . | . | . | . |
| A | . | . | . | . |
| T | . | . | . | . |
| E | | | | |
| D | INDEPENDENT SAMPLE N | XNA | XNB | XNZ |

## TIME SERIES DATA

| | | CHARACTERISTICS (MEASUREMENTS) A | B | ... Z |
|---|---|---|---|---|
| | CORRELATED READING 1 | X1A | X1B | X1Z |
| C | CORRELATED READING 2 | X2A | X2B | X2Z |
| O | | | | |
| R | . | . | . | . |
| R | . | . | . | . |
| E | . | . | . | . |
| L | . | . | . | . |
| A | . | . | . | . |
| T | . | . | . | . |
| E | | | | |
| D | CORRELATED READING T | XNA | XNB | XNZ |



LINEAR REGRESSION RESULTS

| VARIABLE NAME | LAG | COEFFICIENT | STANDARD ERROR | T-RATIO |
|---|---|---|---|---|
| CONSTANT | | 193.488 | 9.23785 | 20.9452 |
| VERBAL.DAT | 0 | .643141 | .208246E-01 | 30.8838 |

THE RESIDUAL STATISTICS

| | | | |
|---|---|---|---|
| SUM OF SQUARES | : 20.235 | DEGREES OF FREEDOM | : 13 |
| MEAN SQUARE | : 1.5565 | NUMBER OF RESIDUALS | : 15 |
| R SQUARED | : .98656 | | |

MATH
(Y) 492 492 493 488 488 484 481 480 472 472 470 468 467 466 466
VERBAL
(X) 466 466 463 460 455 453 445 444 434 431 429 429 427 424 424

Now suppose we reorder the pairs of data, insuring that the relationship between an x and a y remains unchanged. For example, if we interchange the first and the last data point we would have,

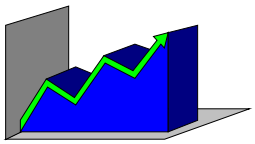
MATH
(Y) 466 492 493 488 488 484 481 480 472 472 470 468 467 466 492
VERBAL
(X) 424 466 463 460 455 453 445 444 434 431 429 429 427 424 466

How do you think the regression between x and y is effected? (Or do you think the estimates of the coefficients will change in the model?)

$$Y = W0_t + W1\ X_t$$

Most people think that the answers are different, but they are not. **They are identical**.

$$\Sigma Y = N A + \Sigma X B$$
$$\Sigma XY = \Sigma X A + \Sigma X^2 B$$



Ordinary least squares gives equal weight to all pairs of readings thus we minimize the vertical sum of squares.

$$\text{MIN} \quad \Sigma \quad (Y - Y)^2$$

A transfer function is essentially a superset, where the assumption of equality in the pairs of readings is empirically tested. Should the data evidence the need for structure to account for serial autocorrelation then one should incorporate same. This structure on

the noise $\frac{\theta(b)}{\phi(b)}$ can be inverted and is called the phi-weights.

$$\left(\frac{\theta(b)}{\phi(b)}\right)^{-1} = 1 - \text{PHIB-PHI2B}^2 \text{ -PHI3B}^3 \ldots\ldots = \text{PI(B)}$$

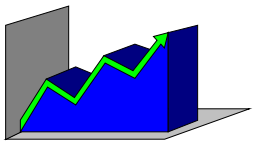
Thus we minimize $\Sigma\left(\text{PI(B)}\ (Y - \hat{Y})\right)^2$.

THEORY & PRACTICE

Statistical forecasting is part art and part science. AFS has developed "smart" or "expert" forecasting software that replicates the art or statistical judgement that was needed in older software. These products are then an example of artificial intelligence, insofar as a machine has been "taught" to think like a good statistician. Some have likened them to automatic cameras, where you just point and shoot. All of our products offer both this feature and the feature of allowing the user to focus or model themselves.

The best forecast is often a combination of two different inputs:

- The best human, street-smart, judgmental opinion
- The best statistical objective projection.



AFS products deal with the latter in an optimal fashion, offering just one input into the "final forecast". The strength and the weakness of our forecasting is its objectivity. Different applications can benefit from good statistical projections:

- Accountants can use them in the analytical review process to identify irregular activity or to identify changes.

- Planners and financial analysts can use them as input to models. Rather than assume that sales and expenses grow at specific rates, Box-Jenkins methods provide bona fide, objective baseline projections.

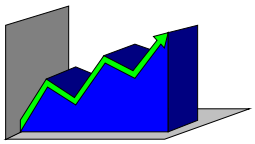
- Inventory control analysts can use them to determine stocking strategies and optimal re-order points.

- Quality control engineers can improve the accuracy of their charts by incorporating sequential relationships between readings. Shewart control charts assume independence of sequential readings; this is not realistic for time series data.

Statistical forecasting or modeling sometimes requires a candidate list of "cause" or "input" variables which are then evaluated as to their importance or statistical significance. As a point of fact, no time series is truly dependent on its past. It is dependent on other factors which may be known or unknown and whose data may or may not be available. If these input series continue to affect the output series as they did in the past, then the history of the series may be sufficient.

Autobox performs transfer function modeling which is a superset of regression methods. This uses Box-Jenkins ARIMA modeling, which is an extrapolitive approach to projecting a single time series from its own past. This apparently "naive" procedure of extrapolating a series based upon its past is in reality a very powerful way of incorporating omitted variables.

In short, Box-Jenkins models for a single series are simply the "best" weighted moving average of the past. These weights are termed the pi weights. ARIMA models are a parsimonious representation of the pi weights. Rather than assuming the appropriate weighting structure, the Box-Jenkins modeling process allows the data to "speak" for itself.



REGRESSION MODELS VERSUS TRANSFER FUNCTION MODELS

A very natural question arises in the selection and utilization of models. One asks,"Why not use simple models that provide uncomplicated solutions?" The answer is very straightforward, "Use enough complexity to deal with the problem and not an ounce more". Restated, let the data speak and validate all assumptions underlying the model. Don't assume a simple model will adequately describe the data. Use straight-forward identification/validation schemes to identify the structure.

Since regression is a particular subset of a transfer function, it is possible to constrain the transfer function to deliver regression results. The following is a development of a transfer function model from the bottom up. In other words, we will start with a simple model and allow complications to enter one-by-one. In this manner, we will learn what the real requirements are for a simple regression model.

Simple model:

$$Y_t = B0 + B1\ X_t + A_t$$

restating in terms of a more general class, we get:

$$Y_t - k = + w0\ X_t + A_t$$

introducing the notion of a polynomial, we get:

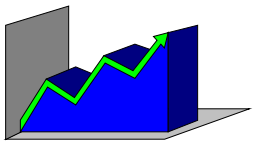
$$Y_t - k = + w(B)\ X_t + A_t$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

incorporating the identity element (1) doesn't change anything.

$$Y_t - k = + \frac{w(B)}{1} X_t + A_t$$

$$Y_t - k = + \frac{w(B)}{1} X_t + \frac{1}{1} A_t$$



recognizing that there might be the need for an autoregressive process to operate on the error term we get the following:

$$Y_t - k = + \frac{w(B)}{1} X_t + \frac{1}{\phi(B)} A_t$$

where $\phi(B) = 1$.

recognizing that there might be the need for an moving average process to operate on the error term we get the following:

$$Y_t - k = + \frac{w(B)}{1} X_t + \frac{\theta(B)}{\phi(B)} A_t$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$ where $\phi(B) = 1$ and $\theta(B) = 1$

recognizing that there might be the need to incorporate a period of delay (b) before Y responds to X we get:

$$Y_t - k = + \frac{w(B)}{1} X_{t-b} + \frac{\theta(B)}{\phi(B)} A_t$$

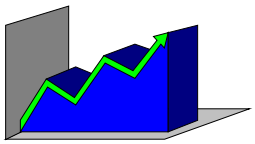
$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$

incorporating the identity element (1) doesn't change anything.

$$Y_t - k = + \frac{w(B)}{1} X_{t-b} 1 + \frac{\theta(B)}{\phi(B)} A_t$$



$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$

recognizing the possibility that the level of Y might depend on changes in X, rather than the level of X suggests:

$$Y_t - k \quad 1 \quad = \quad + \quad \left[\frac{w(B)}{1}\right] X_{t-b} \quad \{[1-B^o]]^d \quad + \quad \frac{\theta(B)}{\phi(B)} A_t$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag" where $\phi(B) = 1$ and $\theta(B) = 1$ ; $\{[1-B^o]]^d$

or more generally

$$Y_t - k \quad 1 \quad = \quad + \quad \left[\frac{w(B)}{1}\right] X_{t-b} \quad \{[1-B^{ox}]]^{dx} \quad + \quad \frac{\theta(B)}{\phi(B)} A_t$$

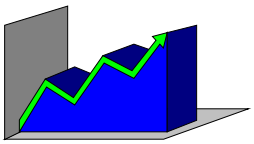
$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$ ; $\{[1-B^{ox}]]^{dx} = 1$ for ox=0 and dx=0

recognizing the possibility that changes in Y might depend on changes in X, leads to:

$$A_t\left[ Y_t - k \right] \quad \{[1-B^{oy}]]^{dy} = \quad + \quad \left[ \frac{w(B)}{1} \right] X_{t-b} \quad \{[1-B^{ox}]]^{dx} \quad + \quad \frac{\theta(B)}{\phi(B)}$$



$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$ ; $\{[1-B^{ox}]]^{dx} = 1$ for ox=0 and dx=0; $\{[1-B^{oy}]]^{dy} = 1$ for oy=0 and dy=0

recognizing the possibility that changes when X changes it effects a sequence of Y's leads to a dynamic model of:

$$A_t \left(Y_t - k\right) \{[1-B^{oy}]]^{dy} = + \left[\frac{w(B)}{\delta(B)}\right] X_{t-b} \{[1-B^{ox}]]^{dx} + \frac{\theta(B)}{\phi(B)}$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$ ; $\{[1-B^{ox}]]^{dx} = 1$ for ox=0 and dx=0; $\{[1-B^{oy}]]^{dy} = 1$ for oy=0 and dy=0

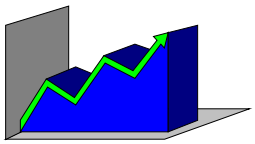
incorporating the identity element (1) for an appropriate power transformation to do-couple linear dependence between the first and second moment delivers:

$$A_t^{1} \left(Y_t - k\right) \{[1-B^{oy}]]^{dy} = + \left[\frac{w(B)}{\delta(B)}\right] X_{t-b}^{1} \{[1-B^{ox}]]^{dx} + \frac{\theta(B)}{\phi(B)}$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$ ; $\{[1-B^{ox}]]^{dx} = 1$ for ox=0 and dx=0; $\{[1-B^{oy}]]^{dy} = 1$ for oy=0 and dy=0



Allowing these power transforms to be other than unity leads to:

$$A_t \left( Y_t - k \right)^{a} \{[1-B^{oy}]]^{dy} = + \left( \frac{w(B)}{1} \right) X_{t-b}^{a} \{[1-B^{ox}]]^{dx} + \frac{\theta(B)}{\phi(B)}$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$ ; $\{[1-B^{ox}]]^{dx} = 1$ for ox=0 and dx=0; $\{[1-B^{oy}]]^{dy} = 1$ for oy=0 and dy=0 ; a=1

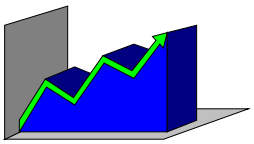
expanding to more than one input(X) leads to:

$$\mathbf{A_t} \left( \mathbf{Y_t - k} \right)^{a} \{[1-B^{op}]]^{dp} = + \sum^{N} \left( \left( \frac{w(B)}{1} \right) \mathbf{X_{t-b}}^{a} \{[1-B^{on}]]^{dn} \right) + \frac{\theta(B)}{\phi(B)}$$

$w(B) = w0$ where $B^1X_t = X_{t-1}$ and in general $B^kX_t = X_{t-k}$

b = delay or "dead time" or "lag"

where $\phi(B) = 1$ and $\theta(B) = 1$ ; **BOLD IS A MATRICES** $\{[1-B^{on}]]^{dn} = 1$ for on=0 and dn=0; $\{[1-B^{op}]]^{dp} = 1$ for op=0 and dp=0 ; a=1



That is why the ordinary regression model is seen to be a particular subset of a more powerful class of model where the following set of assumptions are in place:

- power transforms are known or assumed for all series =1.
- the delay is known =0
- the required level of differences is known for all series =0
- the distributed lag structure w(B) is known for all series =w0
- the dynamic lag structure ë(B) is known for all series =w0
- the AR lag structure í(B) is known =1.
- the MA lag structure é(B) is known =1.

These are the restrictions that shackle the transfer function you assume the ordinary regression model. Thus the model is:

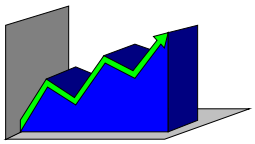
$$Y_t = B0 + B1\ X_t + A_t$$

Sometimes a numerical example can clear up algebraic confusion.

$$\left(Y_t\right) = + .03\ X_t \quad A_t$$

Generalizing

$$\left(Y_t\right) = + \left(.03\right) X_t \quad \left(\frac{1}{(1 - .7B)}\right) A_t$$



Generalizing

$$\left(Y_t\right)\ \{\ [1-B\ ]^{12}\}^1 \quad = \quad + \quad \left(\ .3\right)\ X_{t-3} \quad \left(\frac{1}{(1\ -\ .7B)}\right)\ A_t$$

Generalizing

$$\left(Y_t\right)\ \{\ [1-B\ ]^{12}\}^1 \quad = \quad + \quad \left(\frac{.3}{(1\ -\ .6B)}\right)\ X_{t-3} \quad \left(\frac{1}{(1\ -\ .7B)}\right)\ A_t$$

Generalizing

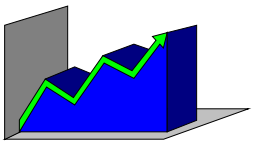
$$\left(Y_t\right)\ \{\ [1-B\ ]^{12}\}^1 \quad = \quad + \quad \left(\frac{.3}{(1\ -\ .6B)}\right)\ X_{t-3} \quad \left(\frac{1}{(1\ -\ .7B)}\right)\ A_t$$

Generalizing

$$\left(Y_t\right)\ \{\ [1-B\ ]^{12}\}^1 \quad = \quad + \quad \left(\frac{.3}{(1\ -\ .6B)}\right)\ X_{t-3} \quad \left(\frac{1}{(1\ -\ .7B)}\right)\ A_t$$

and in a rearranged form:

$$(1\text{-}.7B)\ [1-B\ ]^{12}\ Y_t = \quad + \quad \frac{(1\text{-}.7B)(.03)}{(1\ -\ .6B)}\ X_{t-3} \quad A_t$$



or approximately

$(1-.7B-B^{12}+.7B^{13})\,Y_t \approx + .03\ X_{t-3} + A_t$

$Y_t\ 7Y_{t-1}- Y_{t-12} - Y_{t-13} \approx + .03\ X_{t-3} + A_t$

$Y_t \approx 7Y_{t-1}+ Y_{t-12} + Y_{t-13} + .03\ X_{t-3} + A_t$

GLOSSARY OF TERMS AND PROGRAM OPTIONS

This glossary is not a complete glossary of all terms, but rather it is a list of the most commonly questioned items. The topics include definitions of a few statistical phrases, a number of computational algorithms and some discussions of the more complicated Autobox options.

AUTOCORRELATION

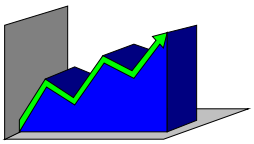
Autocorrelation is a measure of the unconditional dependence that exists between observations in a time series that are separated by a particular time interval, called lag. The value of the autocorrelation lies between +1 and -1. The closer the autocorrelation is to +1 and -1, the more highly correlated are the observations separated by the particular lag being considered. Similarly, the closer the autocorrelation is to 0, the less correlated are the observations separated by a particular time interval. If the autocorrelation is closer to +1, a positive correlation is indicated; if it is closer to -1, a negative correlation exists. In summary, the autocorrelation measures the unconditional relationship between lags.

Conceptually, the ACF for lag 1 is equivalent to the regression coefficient in the model:

$Y_t = B + B\ Y_{t-1} + A_t$

and the ACF for lag 2 is equivalent to the regression coefficient in the model:

$Y_t = B + B\ Y_{t-2} + A_t$ and so on. Thus the ACF measures the unconditional or simple dependency between values k periods apart.



More compactly where Z is the simple mean for the entire range.

$$\text{ACF at lag k} = r_k = \frac{\sum_{t=1}^{n-k} (Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^{n} (Z_t - Z)^2}$$

The regression approach does not give the same answer as the compact formula due to the fact that the regression approach recomputes the denominator each and every time. Otherwise the answers will be very similar.

PARTIAL AUTOCORRELATION

Partial autocorrelation is a measure of the conditional dependence that exists between observations in a time series that are separated by a particular time interval, the autocorrelations between observations at all smaller time intervals being already known. The value of the partial autocorrelation lies between +1 and -1. The closer the partial autocorrelation is to +1 and -1, the more highly correlated are the observations separated by the particular lag being considered.

Similarly, the closer the autocorrelation is to 0, the less correlated are the observations separated by a particular time interval. If the partial is closer to +1, a positive correlation is indicated; if it is closer to -1, a negative correlation exists. In summary, the partial autocorrelation measures the conditional correlation between lags.

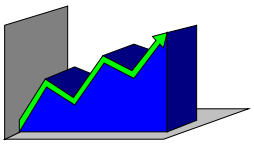
Conceptually, the PACF for lag1 is equivalent to the regression coefficient in the model:

$$Y_t = B_o + B\ Y_{t-1} + A_t \quad \text{and is identical to the ACF at lag1}$$

And the PACF for lag2 is equivalent to the lag2 regression coefficient in the model:

$$Y_t = B_o + B\ Y_{t-2} + B\ Y_{t-2} + A_t \quad \text{and is identical to the ACF at lag1}$$

measures the conditional or compound dependency between values k periods apart, given intermediate lags. It is the partial regression coefficient in the framework of a regression model.



PACF at lag k = [see algorithm on page 65 of the Box and Jenkins (1976) text.

CROSS AUTOCORRELATION

Cross correlation is a measure of the dependence that exists between observations in two time series that are separated by a particular time interval, called lag. The value of the cross correlation lies between +1 and -1. The closer the cross correlation is to +1 and -1, the more highly correlated are the observations separated by the particular lag being considered. Similarly, the closer the cross correlation is to 0, the less correlated are the observations separated by a particular time interval. If the correlation is closer to +1, a positive correlation is indicated; if it is closer to -1, a negative correlation exists. In summary, the cross correlation measures the strength of the relationship between the lags of two time series.

CCF at lag k = [See algorithm on page 374 of the Box and Jenkins (1976) text.

IMPULSE RESPONSE WEIGHT

The impact of time series X on output series Y can be represented by a distributed lag model:

$$Y_t = V_oX_t + V1X_{t-1} + V1X_{t-2} + A_t$$

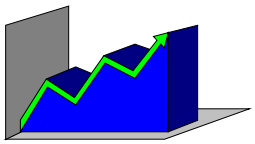
The weights Vo, V1, V2 ... are called the impulse response weights. In the context of B-J modeling, the transfer function portion of a multiple input model is actually a parsimonious representation of the impulse response weights. An estimate of the impulse response weights, for the purpose of model identification, can be obtained from the cross-correlations of the prewhitened X series and the prewhitened Y series.

LAMBDA VALUE (TRANSFORMATION PARAMETER)

The lambda value in a Box-Jenkins model is the value (L) specified as $Z_t^{(L)}$ . It represents the power transformation (L) that is applied to series Z in order to induce variance stationarity. The lambda value may take on any value in the range -1 through +1. A lambda value of 1 indicates the original (untransformed) series; a value of .5 indicates the square roots of the series; a value of 0 indicates the natural logarithms of the series; a -1 indicates the inverse of the series; etc.

RESIDUALS AND THE RESIDUAL AND MODEL STATISTICS

The residuals from a model are generated by subtracting the fit values from the actual values. The fit values are obtained by "fitting" the equation to the original time series. Defining $a_t^2$ as the residuals, N as the number of residuals, and M as the number of parameters in the model, the following statistics are computed:



$$\text{Residual Sum of Squares} = \sum_{t=1}^{N} a_t^2 \qquad 1 \text{ to } t$$

$$\text{Residual Mean Square} = \frac{\text{Residual Sum of Squares}}{\text{degrees of freedom}}$$

Residual Standard Error = ✔Residual Mean Square

Akaike Criteria (AIC) = N*ln(Residual Mean Square) + 2M

These statistics provide a measure of model adequacy. The AIC is only defined for models where the data is in the original metric.

SEASONALITY (PERIODICITY) OF THE DATA

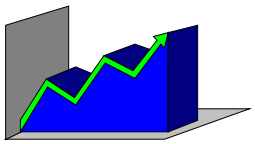
The automatic algorithms require the user to specify the periodicity of the data. Periodicity is an integer value which represents the seasonal structure of the data. For example, periodicity of 1 indicates annual data, periodicity of 4 indicates quarterly data and periodicity of 12 indicates monthly data. The reason for the necessity of entering this information is that the automatic algorithm(s) need to "know" the seasonality in order to appropriately identify any seasonal structure. If the periodicity is unknown, then a safe response is a seasonality of one.

DISPLAY THE MODEL DEVELOPMENT VIA THE PRINT OPTIONS

Each automatic modeling algorithm is an iterative process of tentative model formation, estimation, diagnostic checking and model improvement. The result is a model which can be used for forecasting. You can choose to see the intermediate solutions of the iterative modeling process by requesting the program to display the model development. A word of advice, this output can be quite lengthy.

START AUTOMATIC PROCESS WITH A MODEL

Each automatic algorithm starts processing by examining the key statistics and choosing the first tentative model. You can bypass this stage by personally specifying the form of the first model. The advantage of this option is that you are able to utilize the automatic features to "fine tune" a



model of your choice. A possible disadvantage is that, if you enter a particularly poor model, the program may not be able to recover a good model for the series.

The key is that, if you ask Autobox to automatically identify an initial model, this will be groundwork on which the entire structure of necessity and sufficiency testing starts. If you elect not to have Autobox automatically identify an initial model, then, if needed, the program will use the existing model stored in the Database. If a model doesn't exist but is needed, then a null model will be used. A null model is simply a 'mean model with value 0 '.

WITHHOLDING OBSERVATIONS FOR FORECAST UPDATING

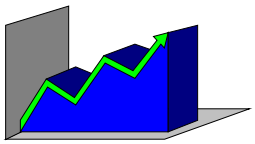
One feature of Autobox is the ability to compute updated forecasts. Naturally, this is an option within the forecasting phase of the program. However, you must be aware of this option and how you are to use it before you invoke Autobox so that you can understand the sequence of its decisions.

If you withhold 10 observations and ask for a 20 period forecast, the program will check to see if an additional 20 actuals exist in the Database to be used. If they do not, the program then checks to see if 20 forecasts exist for F_seriesname starting at the correct time period. If they do exist, then the program will treat them as actuals and this will affect the width of the resultant confidence intervals for the dependent series. If both of the above fail, the program will then generate the 20 future values using the existing ARIMA model. If the series is deterministic, then the last value will be used to populate the 20 element projection vector (input series). However, if you only request a 10 period forecast then only 10 additional or new values would be required. Note that these scenarios were based on the assumption that the user requested updated forecasts to be made at each new origin.

In the case where no updating is requested, the program uses the forecast length to determine if there are enough values known either through the existence of actuals or the existence of stored forecasts (F_) at the required forecast launch point.

AGGREGATE FORECASTS

The forecasting phase of the program will compute the aggregate forecast from any origin for any length, within the limits imposed by the number of observations. The aggregate forecast is the sum of the forecast values.



The program also computes the aggregate uncertainty, which is a function of the forecast variance and the correlation between the forecast errors. A description of the formula used for this calculation can be found in Appendix A5.1.2 of the Box and Jenkins text (1976, pp. 159-160). For the cases where an actual value is used in the computation of the aggregate forecast, the aggregate uncertainty is only computed for the time periods that were forecast. The aggregate interval forecasts are then offset by the actual values at that time period.

## THE PURPOSE OF PREWHITENING MODELS

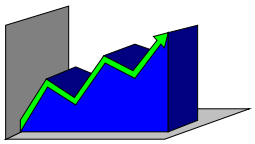
The program can store prewhitening models based on ARIMA estimation or the user can key in models via the Edit Model option. If no model exists, then the program simply uses a mean model with a value of 0. We would like to explain the role that these models play. Please note that the focus is on the program's need for the information in these files. One can consider a prewhitening model to be a filter which converts time series data to cross-sectional data allowing for system or model identification to proceed. In this way, Transfer Function model are a superset of regression models.

## THE PURPOSE OF PREWHITENING MODELS IN THE IDENTIFICATION PHASE

The process of identifying a transfer function model, as outlined by Box and Jenkins, is in itself a multi-step procedure. The very first step is to develop the optimal ARIMA model for each time series that is to be included in the transfer function equation. These ARIMA models play a critical role in the analysis of the relationship between each input series and the output series. AFS suggests that the ARIMA model used in prewhitening be kept as simple as possible, but not too simple. We suggest that interventions, both mean and variance, should not be included as it is possible to mask interrelationships by fitting overly complicated intrarelationships.

Autobox offers the option to use differencing factors in the "Identification phase". AFS has found that over-differencing can lead to poor identification. Box and Jenkins , faced with "expensive computing costs" took the shortcut and used the ARIMA model differences, possibly different, in the Transfer Function identification phase. We suggest that the user try it both ways, that is with and without the differencing factors.

Box and Jenkins developed the concept of empirical identification in the case of one stochastic input series. Researchers are also concerned with the case of multiple inputs, where the input



series and/or their respective error processes may or may not be interrelated. It is quite natural to ask about the applicability of the bivariate approach of Box and Jenkins when faced with data arising under different scenarios.

The bivariate approach, two series at a time, to identify a multi-input model may be valid if the noise processes underlying each of the inputs is uncorrelated. If these noise processes are cross-correlated, then a more general approach to model identification is required. This approach was suggested by Liu and Hannsens.

In selecting this option, the common-filter option uses a slightly different approach. Some researchers incorrectly conclude that for proper identification and forecasting of transfer function models it is necessary that the input series be independent of each other. This is not true. The requirement is that for the Box and Jenkins approach to transfer function identification to be valid for multiple inputs the prewhitened input series have to be independent of each other. The common filter approach allows the prewhitened series to be cross-correlated as identfication of the form of all inputs is simultaneous, thus avoiding the requirement placed by the Box-Jenkins approach.

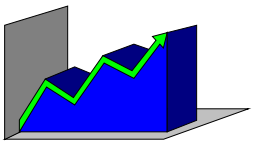
1) Transfer function modeling using the Box-Jenkins
cross correlation (CC) method. (the Default)
"Use the common filter method (YES/NO)" = NO

During the transfer function identification stage of modeling, the program prewhitens each series according to the rules laid out by Box and Jenkins (1976). Each stochastic time series is made stationary by filtering it with the stationarity factors (lambda and differencing) obtained from its own prewhitening model. Each stochastic input series is further filtered by the autoregressive/moving average/trend factors from its own prewhitening model. The stationary output series is also filtered by the series prewhitening model, one model at a time.

2) Transfer function modeling using the
Liu-Hanssens common filter (CF) method.

"Use the common filter method (YES/NO)" = YES

During the transfer function identification stage of automatic modeling, the program prewhitens each series according to the rules laid out by Liu and Hanssens (1982). Each stochastic time series is made stationary by filtering it with the stationarity factors (lambda and differencing)



obtained from its own prewhitening model. Each time series is further filtered by the autoregressive/moving average/trend factors from the common filter. Autobox determines the form of the common filter by examining the autoregressive/moving average/ trend factors of each prewhitening model, hence, although these factors are not directly used to prewhiten, they do play an important role in that the average filter or common filter is used across all model series. Note that if you do not store prewhitening models prior to multivariate analysis the program assumes the classical, and broadly incorrect specification, that the prewhitening models are identically null and of the form Y(t)= μ with μ=0.0 .

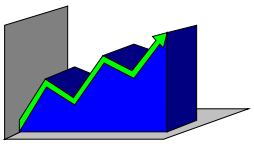
Thus transfer function identification uses the classical and very ordinary cross-correlations. This is the very assumption which has led to incorrect interrelationships given the presence of intrarelationships.

THE PURPOSE OF PREWHITENING MODELS IN THE ESTIMATION PHASE

Estimating a transfer function model with Autobox is rather straightforward. The program needs to know the data, the type of the input series (be it stochastic or intervention) and the form of the model whose coefficient values are to be estimated. The only information in the prewhitening model that is of importance to ESTIMATING the model is the lambda value. Each stochastic input series is transformed to the power lambda before the model coefficients are estimated.

The output series may also be transformed, but since the transformation value to be used is obtained from the transfer function model form, the prewhitening model for the output series is not used. However, the word ESTIMATING was capitalized for a reason - estimating the model is only part of what Autobox does in the estimation phase. The other part is diagnostic checking, where the program computes a number of residual statistics that are needed to evaluate

the necessity and sufficiency of the model. In transfer function modeling, checking the cross correlations of the residual series is a key test for model sufficiency. For each of the stochastic input series, Autobox needs the entire prewhitening model in order to filter the series for this cross-correlation test.

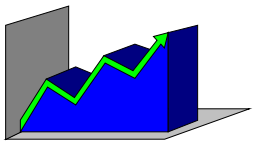


THE PURPOSE OF PREWHITENING MODELS IN THE FORECASTING PHASE

Forecasting with a transfer function model implies that the forecast values generated for the output series are a function of both the output series itself and the various input series. As the computation of the forecast values proceeds out into the future, the forecasts that are generated may in fact come off forecast values that have already been computed. To this end, the program needs to compute the forecast values for each input series in the equation. Hence, one need for the prewhitening models in the forecasting phase is to compute the forecast values for each of the input series. (If the user specifies the forecast values for an input series, then this need obviously disappears). Another reason that the prewhitening model is needed is that this program obtains the lambda value (the transformation parameter) that is to be applied to each time series from its own prewhitening model.

THE PURPOSE OF STARTING MODELS IN AUTOMATIC MODELING

Starting the modeling process with a user specified model is an option to both automatic ARIMA and automatic transfer function modeling. If this option is enabled, then a starting model form must be entered. The program begins the automatic process with an estimation of the coefficients of the starting model, thereby totally bypassing the up-front identification portion of the automatic algorithm. After estimating the starting model, the program continues with the normal diagnostic checking and model updating process.



*5*

# References


Anderson, O.D. (1975). Time Series Analysis and Forecasting. London: Butterworth & Co.

Bell, W. (1983). "A Computer Program for Detecting Outliers in Time Series," in American Statistical Association 1983 Proceedings of the Business Economic Statistics Section, Toronto, pp. 624-639.

Box, G.E.P., and Cox, D.R. (1964). "An Analysis of Transformations (with discussion)", Journal of the Royal Statistical Society, B26, pp. 211-252.

Box, G.E.P., and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, 2nd ed. San Francisco: Holden Day.

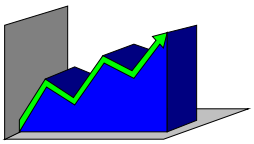
Box, G.E.P., and Tiao, G. (1975). "Intervention Analysis with Applications to Economic and Environmental Problems," Journal of the American Statistical Association, Vol 70,70-79

Chang, I., and Tiao, G.C. (1983). "Estimation of Time Series Parameters in the Presence of Outliers," Technical Report #8, Statistics Research Center, Graduate School of Business, University of Chicago, Chicago.

Downing, D.J., and McLaughlin, S.B. (1986). "Detecting Shifts in Radical Growth Rates by the Use of Intervention Detection," Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge.

Franses, P., (1991) "Seasonality, non-stationarity and forecasting of monthly time series", International Journal of Forecasting", (7), pp 192-208.

Liu, L.M., and Hanssens, D.M. (1982). "Identification of Multiple-Input Transfer Function Models," Communications in Statistics - Theory and Methods, 11(3), pp. 297-314.



Mabert, V.A. (1975). "An Introduction to Short Term Forecasting Using the Box-Jenkins Methodology", Publication #2 AII.E. Monograph Series, Production Planning and Control Division, Norcross, Georgia.
Makridakis, S., Wheelright, S. and McGee, V. (1983). Forecasting: Methods and Applications. New York: Wiley.

McCleary, R., and Hay, R. (1980). Applied Time Series Analysis for the Social Sciences. Los Angeles: Sage.

Nelson, C.R. (1973). Applied Time Series Analysis for Managerial Forecasting. San Francisco: Holden Day.

Pack, D.J. (1978). "Forecasting Time Series Affected by Identified Isolated Events," College of Administrative Science, The Ohio State University, Ohio.

Pack, D.J. (1977). "Revealing Time Series Interrelationships," Decision Sciences, 8, pp. 377-402.

Quenouille, M.H. (1957). The Analysis of Multiple Time Series. London: Charles Griffin & Company.

Reilly, D.P. (1980). "Experiences with an Automatic Box-Jenkins Modeling Algorithm," in Time Series Analysis, ed. O.D. Anderson. (Amsterdam: North-Holland), pp. 493-508.

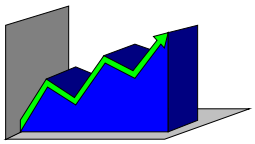
Reilly, D.P. (1987). "Experiences with an Automatic Transfer Function Algorithm," in Computer Science and Statistics Proceedings of the 19th Symposium on the Interface, ed. R.M. Heiberger, (Alexandria, VI: American Statistical Association), pp. 128-135.

Shumway, R.H. (1988). Applied Statistical Time Series Analysis. Englewood Cliffs: Prentice Ha Hall.

Tiao, G.C., and Box, G.E.P. (1981). "Modeling Multiple Time Series with Applications," Journal of the American Statistical Association, Vol. 76, pp. 802-816.

Tsay, R.S. (1986). "Time Series Model Specification in the Presence of Outliers," Journal of the American Statistical Society, Vol. 81, pp. 132-141.

Tsay, R.S., and Tiao, G.C. (1983). "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models,"



Technical Report #2, Statistics Research Center, Graduate School of Business, University of Chicago, Chicago.

Vandaele, W. (1983). Applied Time Series and Box-Jenkins Models. Orlando: Academic Press.

# 6

# Reviews

Tim C. Ireland and Ramesh Sharda, A Test of Box-Jenkins Forecasting Expert Systems for Microcomputers, Department of Management, Oklahoma State University, College of Business Administration, Stillwater, OK 74078, Phone: (405) 744-8642, (405) 744-8638.

John C. Picket (1991), Autobox 3.0, International Journal of Forecasting, 7:395-398.

John C. Picket (1991), Autobox 3.0, OR/MS Today, pp. 32-35, April 1991.

David P. Reilly (1980) "Experiences with an Automatic Box-Jenkins Modeling Algorithm", in Time Series Analysis, ed. O.D. Anderson. (Amsterdam: North-Holland). Pp. 493-508

David P. Reilly (1987) "Experiences with an Automatic Transfer Function Algorithm", in Computer Science and Statistics Proceedings of the 19th Symposium on the interface, ed. R.M. Heiberger, ( Alexandria, Va.: American Statistical Association), pp. 128-135.

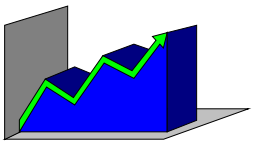
J. Keith Ord (1986), Autobox - Software Review, International Journal of Forecasting, 2:511-513.

R.H. Shumway (1986), Autobox (Version 1.02), The American Statistician, 40(4):299-300.

Leonard J. Tashman and Michael L. Leach (1991), Automatic forecast software: A survey and evaluation, International Journal of Forecasting, 7:209-230.

Pamela A. Texter and J. Keith Ord, (1989), Forecasting using automatic identification procedures: A comparative analysis, International Journal of Forecasting, 5:209-215.

Pamela Texter-Geriner and J. Keith Ord, (1991), Automatic forecasting using explanatory variables: A comparative study, International Journal of Forecasting, 7:127-140.



# Troubleshooting

If you get an error or are very dissatisfied with the forecast, model, etc. We would like to receive an email with a zip file with the following:

Data file ("*.ASC")
*.AFS files
Autobox.exe, freefore.dll, freel.dll
ENGINE.*,
Create a file named 'snoop.afs' and 'presnoop.afs' and run Autobox and save all files that were created in the Autobox directory during the process of running the ASC file (You can use Windows Explorer and sort the directory to see these newly created files). Open a Word document and take print screens of each of your selections up so we can see exactly what you did. Upon receipt of your email, we will review and respond.

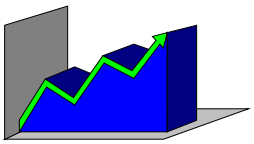
IF AUTOBOX STOPS WORKING (YOU CAN TELL THIS BY VIEWING DETAILS.HTM ON YOUR FREEFORE DIRECTORY THEN AUTOBOX IS NOT FUNCTIONING PROPERLY) AND YOU NEED TO STOP IT, DON'T TURN OFF YOUR COMPUTER AND DO FOLLOW THESE DIRECTIONS

HIT CTRL-ALT-DEL AT THE SAME TIME
CHOOSE "TASK MANAGER"
CHOOSE "PROCESSES"
SORT ON IMAGE NAME BY CLICKING ON IMAGE NAME
SELECT AFSENGINE.EXE AND CLICK ON "END PROCESS"
SELECT AFSLITE.EXE AND CLICK ON "END PROCESS"
CHOOSE "APPLICATIONS"
SELECT "AUTOBOX" AND CHOOSE "END TASK"

Contact us for any questions:

Afs Inc.

P.o. Box 563



Hatboro PA 19040

sales@autobox.com

Phone 215-675-0652  Fax 215-672-2534