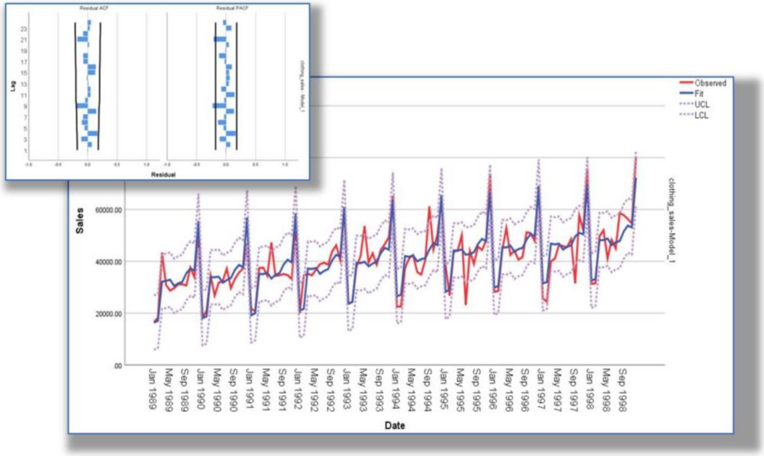


Training
Webinar

Introduction to Time Series Forecasting with IBM SPSS Statistics

Jarlath Quinn

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

Materials

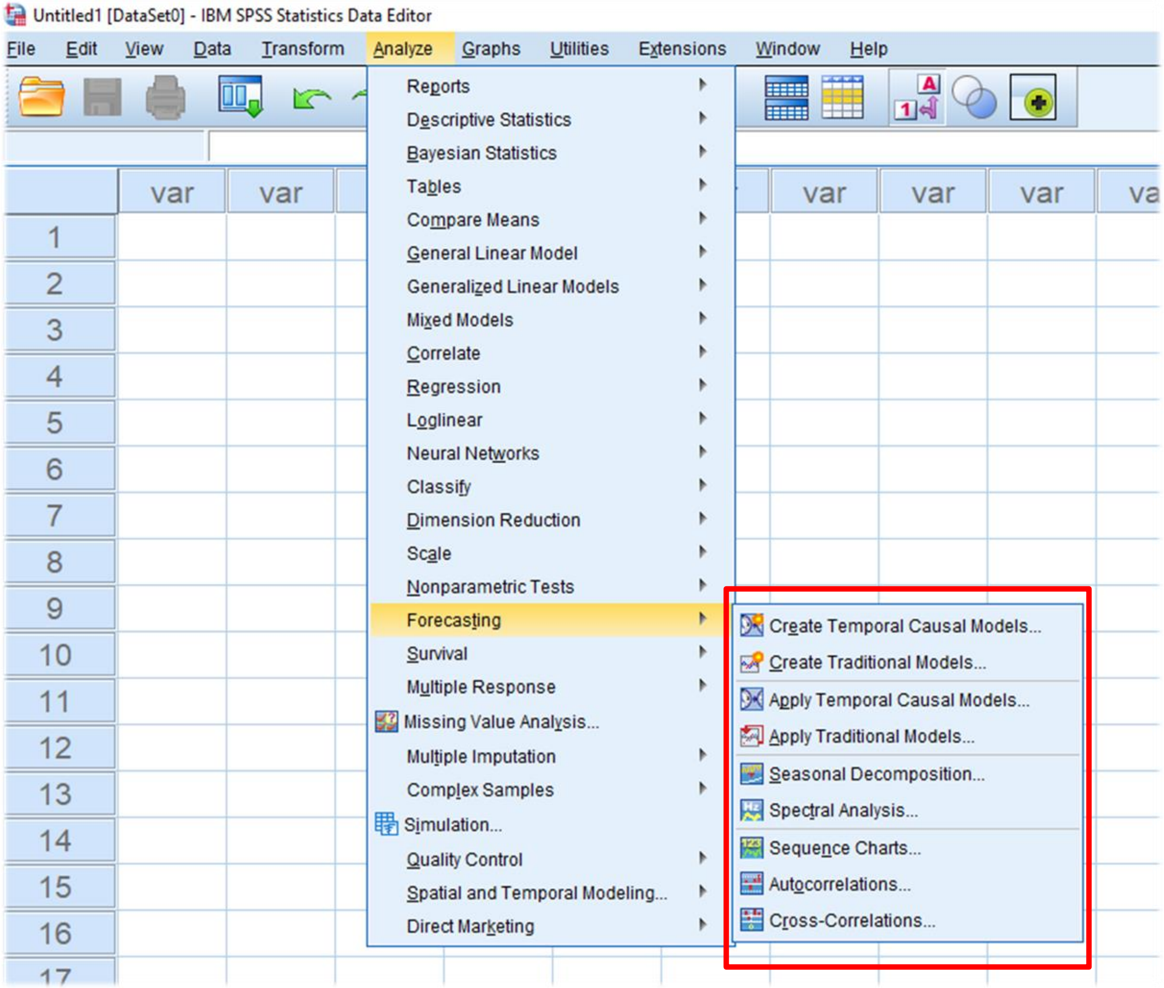
- All demonstrations are documented in these slides with accompanying SPSS data files
- If you have any follow-up questions, please email us at: **info@sv-europe.com**

IBM SPSS Statistics Base & Associated Modules

* You will need the SPSS Statistics Forecasting Module

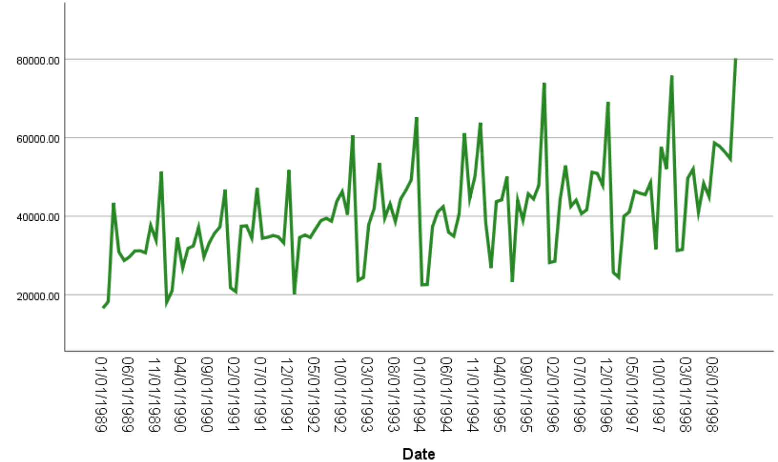
- Missing Values
- Exact Tests
- Categories
- Tables
- Neural Networks
- Advanced Statistics
- SPSS Base
- Regression
- Complex Samples
- Decision Trees
- Data Preparation
- Conjoint
- Direct Marketing
- Forecasting *
- Bootstrapping

- **If you can see sub-menu, then it is installed on your copy of SPSS Statistics**



Agenda

- The Principles of Time Series Forecasting
- Visualising Time Series
- Smoothing Techniques
- Exponential Smoothing Methods
- Interpreting Output and Model Fit
- Using Predictor Fields with ARIMA Modelling
- Generating Forecasts



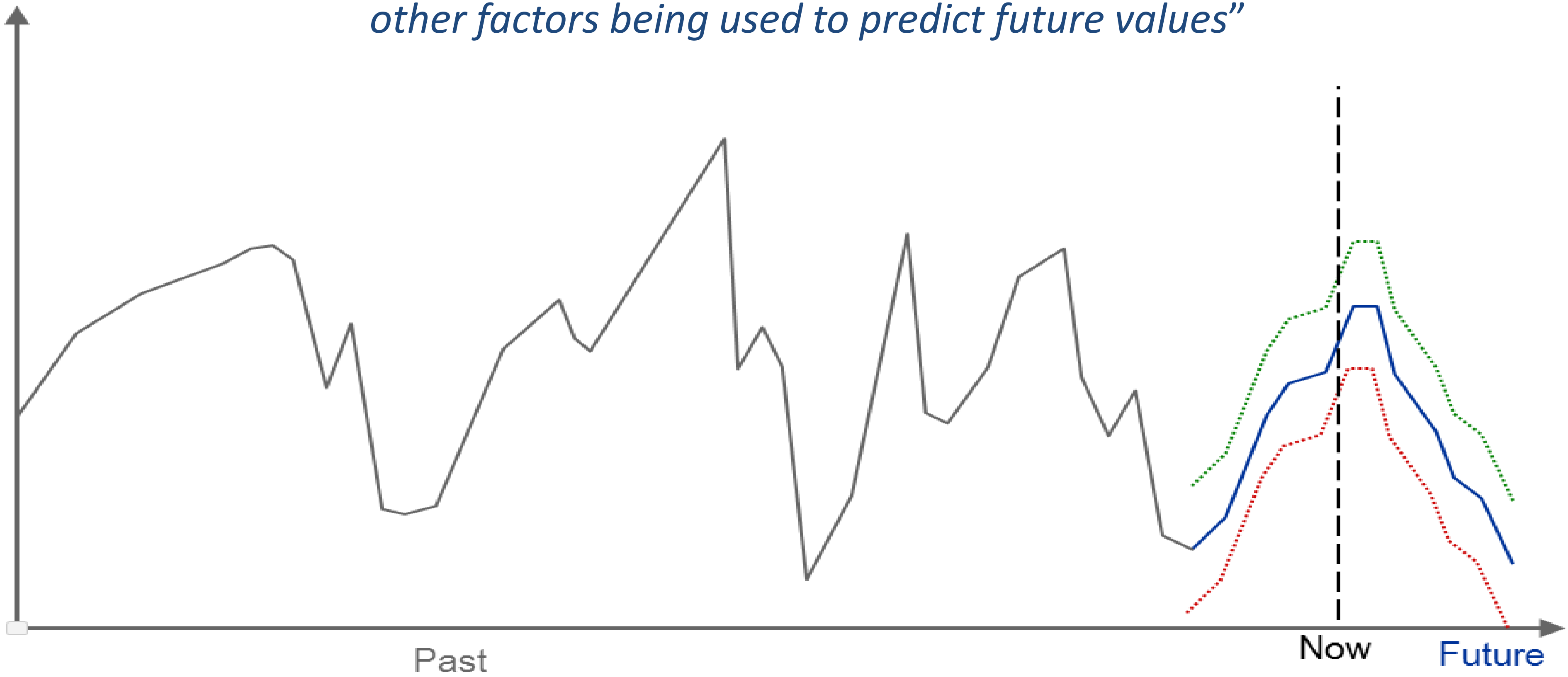
The Principles of Time Series Forecasting

What is Time Series?

- A ‘Time Series’ is simply a series of values of a quantity collected over a specific time period, often with equal intervals between them
- Examples of time series include:
  - Airline passenger numbers for a particular country over the last 40 years
  - Daily website hits during a three-month period
  - Hourly traffic volumes over the course of a week

Time Series Forecasting

“ ’Things’ that are observed repeatedly over time, with past values and other factors being used to predict future values”

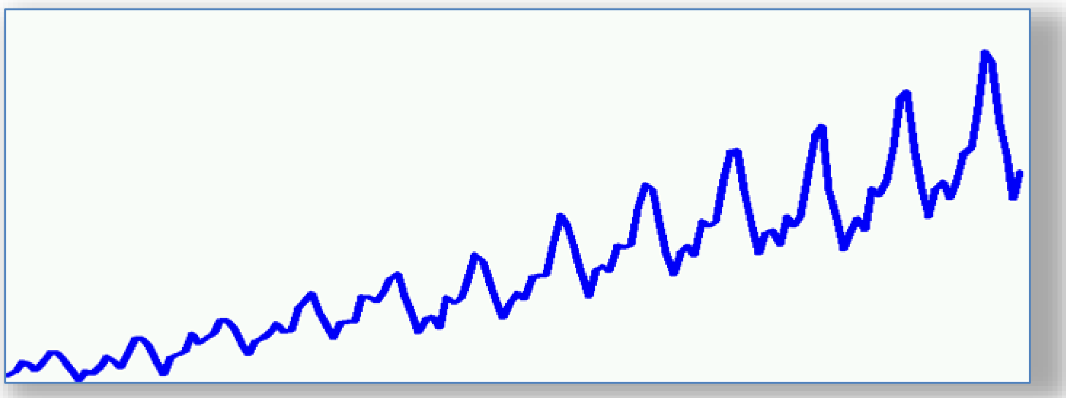


What is Time Series?

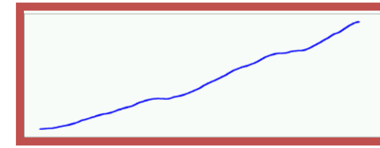
- Time series analysis is based on the principle that the past provides a model for the future
- Time series forecasting models often don’t require predictor/independent variables
- The goal of time series analysis is separate the random variability (‘noise’) from the variability that can be explained
- A single time series may have several elements that enable effective forecasting

What's *in* a 'Time Series'?

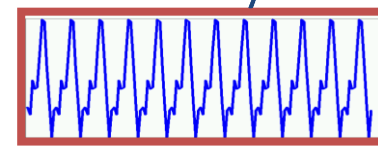
Time Series



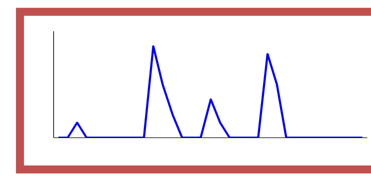
Trend



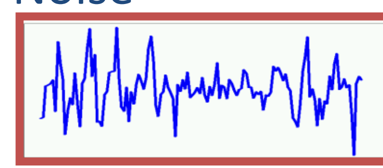
Seasonality



Events



Noise

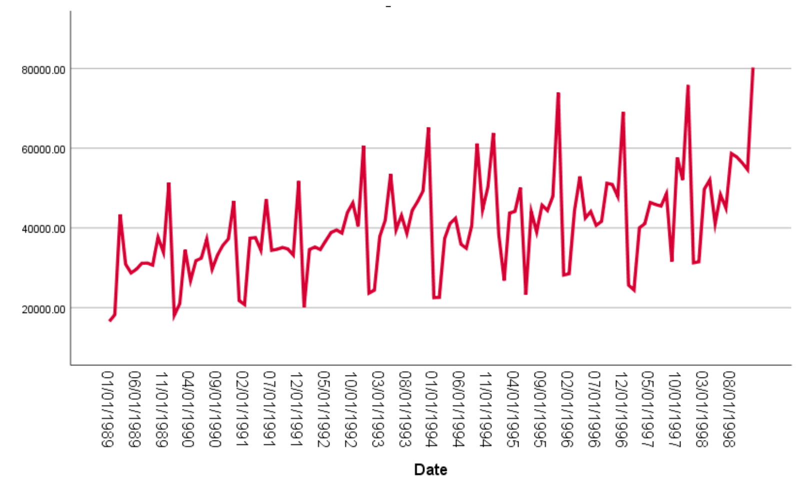


SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Considerations

- Are the data points regularly spaced?
- How far can I forecast into the future?
- What is the periodicity?
- Should I exclude data?
- Do I have other predictor fields?
- Are there special events that need to be marked?
- Do I need to forecast more than one series?

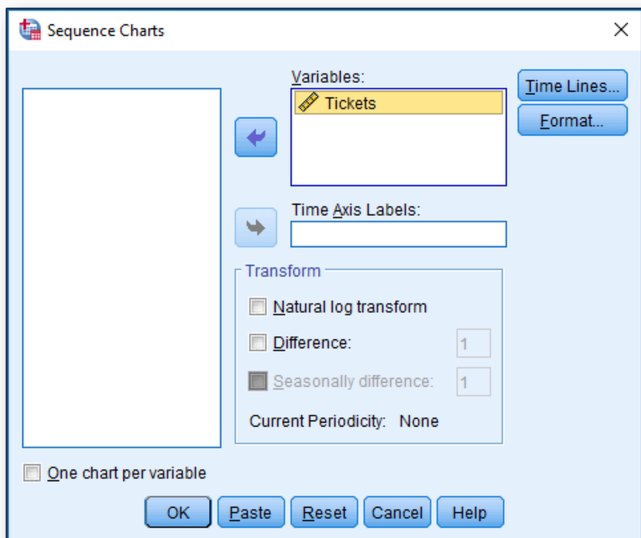


Visualising Time Series

Demo 1: Visualising Time Series

- Using the dataset **Ticket Sales Missing.sav**
- Visualise the series by clicking:
  - **Analyze**
    - **Forecasting**
      - **Sequence Charts**

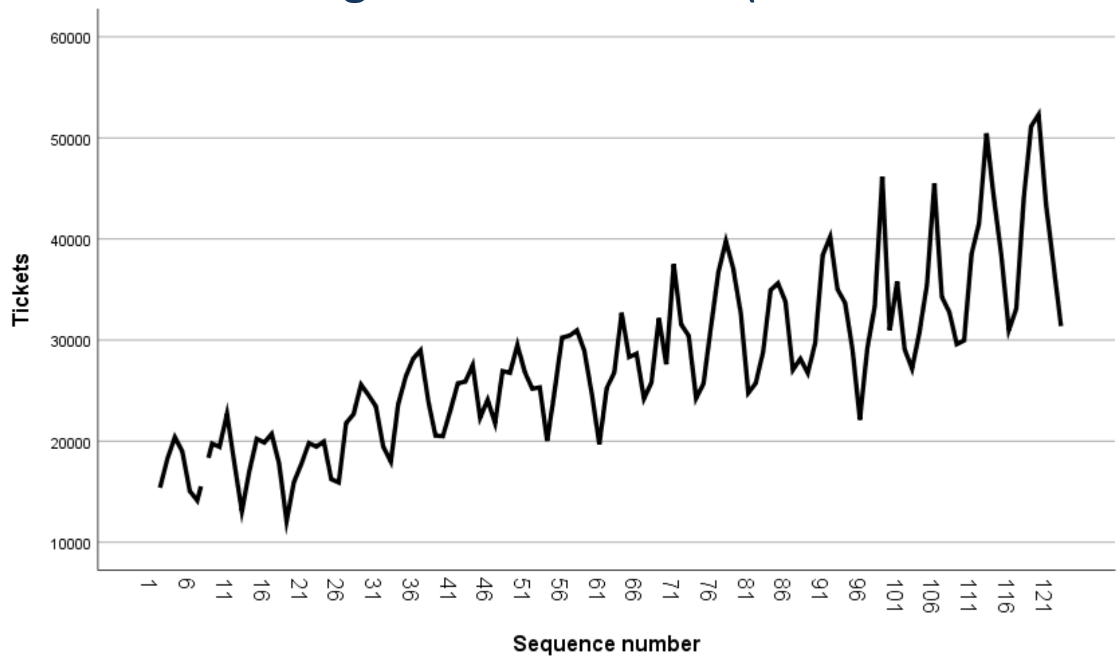
	Tickets	var
1	.	
2	15397	
3	18270	
4	20384	
5	18993	
6	15059	
7	14125	
8	.	
9	19760	
10	19435	
11	22703	
12	17771	
13	13091	
14	17039	
15	20224	
16	19862	
17	20723	
18	17754	
19	12072	
20	15986	
21	17767	
22	19823	



A SELECT INTERNATIONAL COMPANY

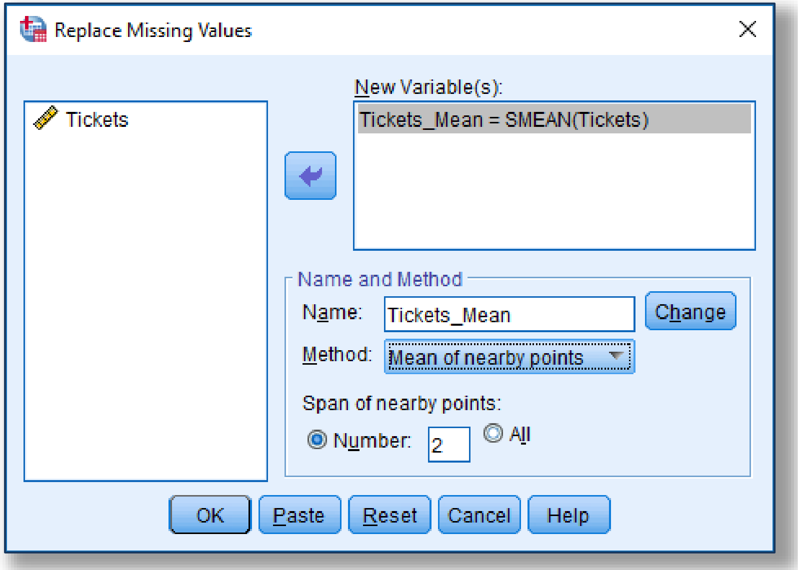
Demo 1: Visualising Time Series

- A few things to note:
  - You don't *need* a variable for the time axis
  - There are gaps in the series i.e. missing data
  - The peaks and troughs increase in magnitude over time (more on this later)



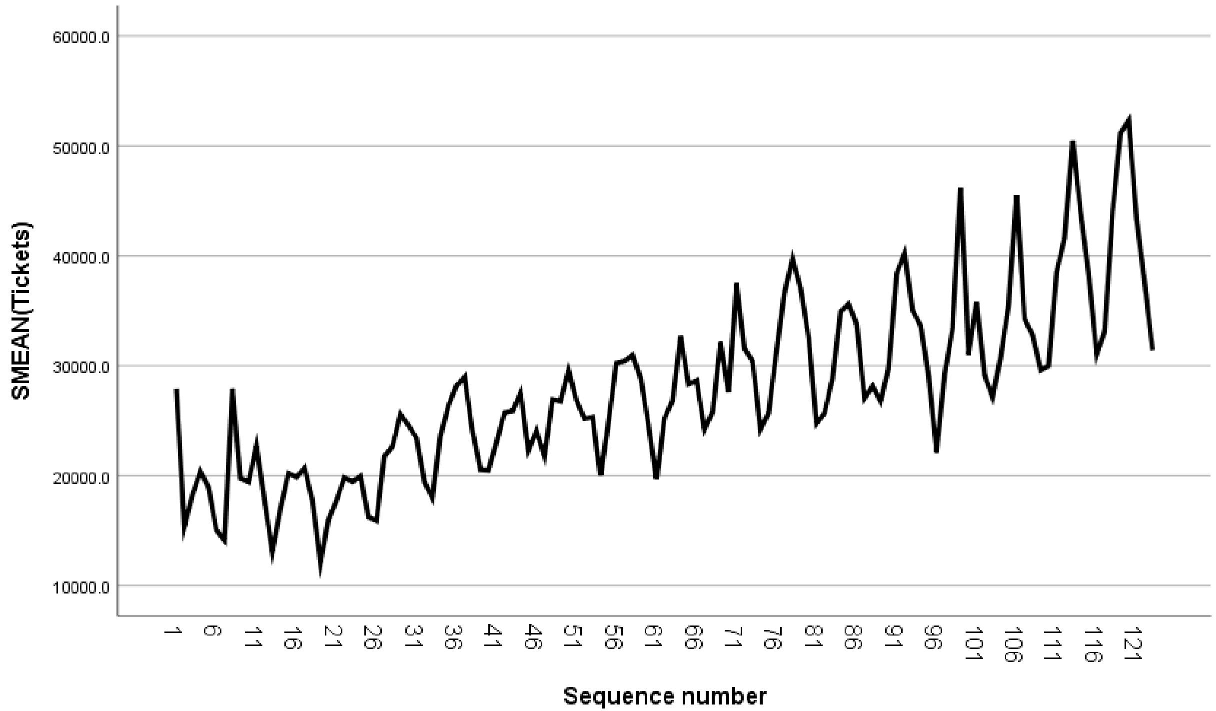
Demo 1: Visualising Time Series

- Let's deal with the missing data first. Click:
  - **Transform**
    - **Replace Missing Values**
- Create a variable called **Tickets_Mean**
- Fill in the missing values using the method '**Mean of nearby points**'
- Remember to click **'Change'**
- Click: **OK**



Demo 1: Visualising Time Series

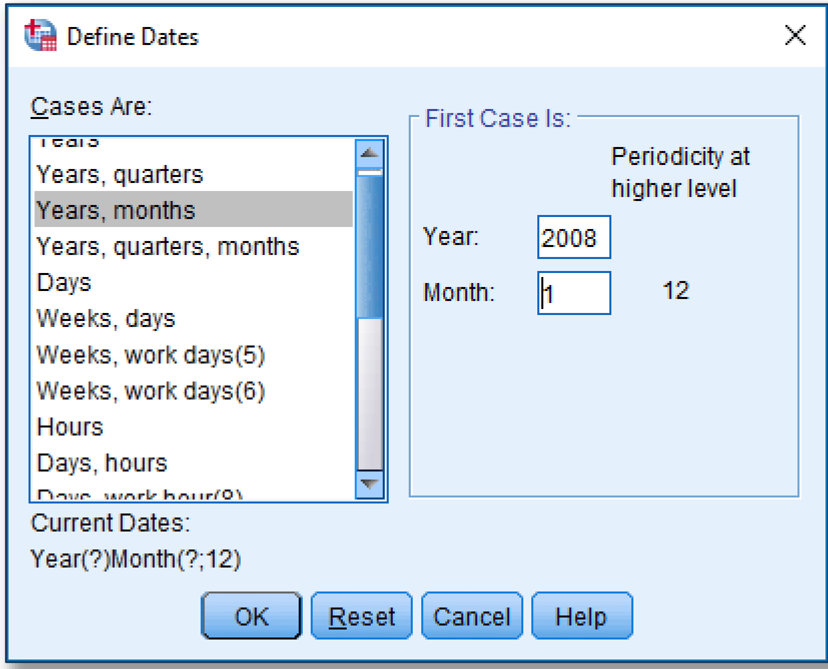
- Re-run the Sequence Chart but this time on the new variable **Tickets_Mean**
- This is one simple way in which missing data can be dealt with



SMARTVISION Europe

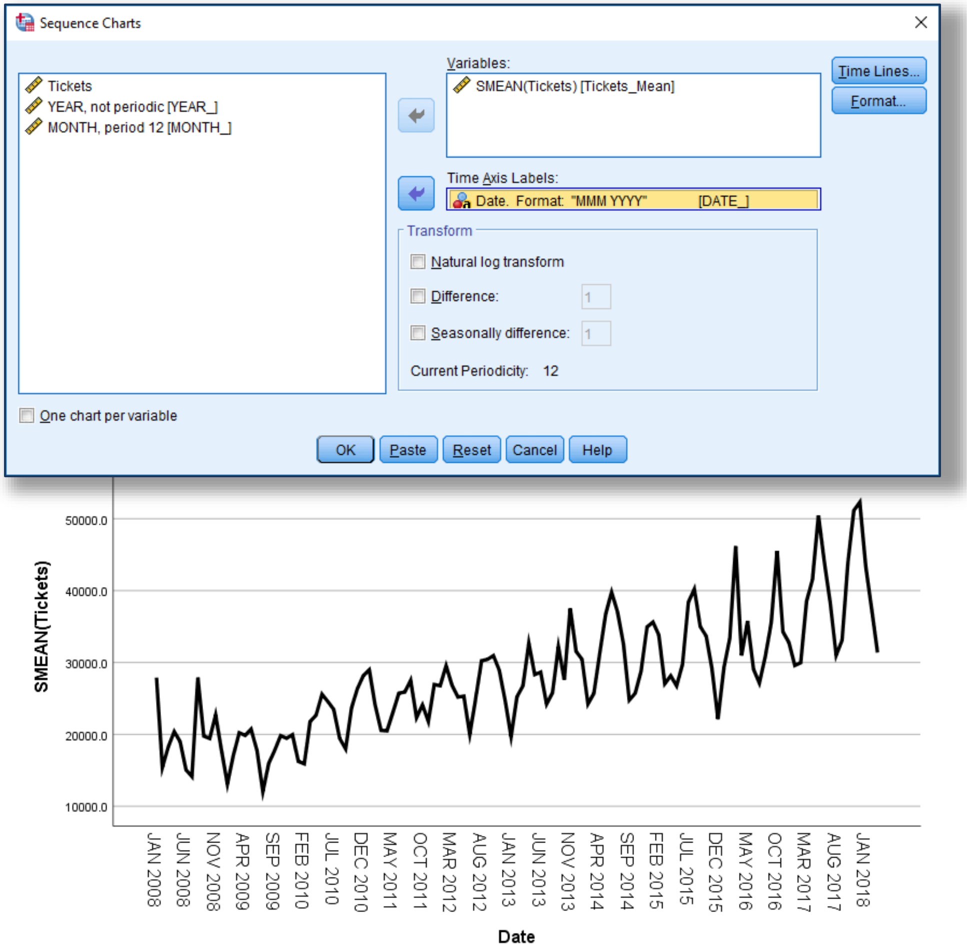
Demo 1: Visualising Time Series

- Now lets assign some date values to the sequence
- This is called **assigning periodicity**
- Click:
  - **Data**
    - **Define Date and Time**
- Define the periodicity as **'Years, Months'**
- Start at the **year 2008** and **month 1**
- Click: **OK**

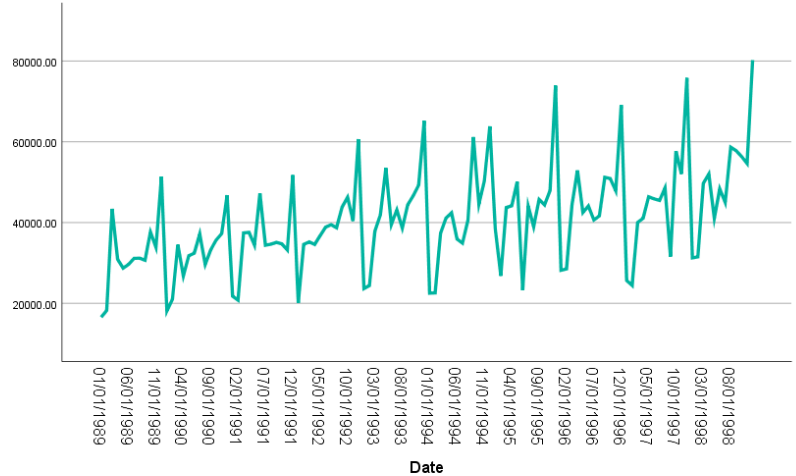


Demo 1: Visualising Time Series

- SPSS creates a series of fields that it understands as periodic date values
- It can use these fields to detect the inherent periodicity in the data
- Re-run the Sequence Chart using the new date field containing the Year and the Month



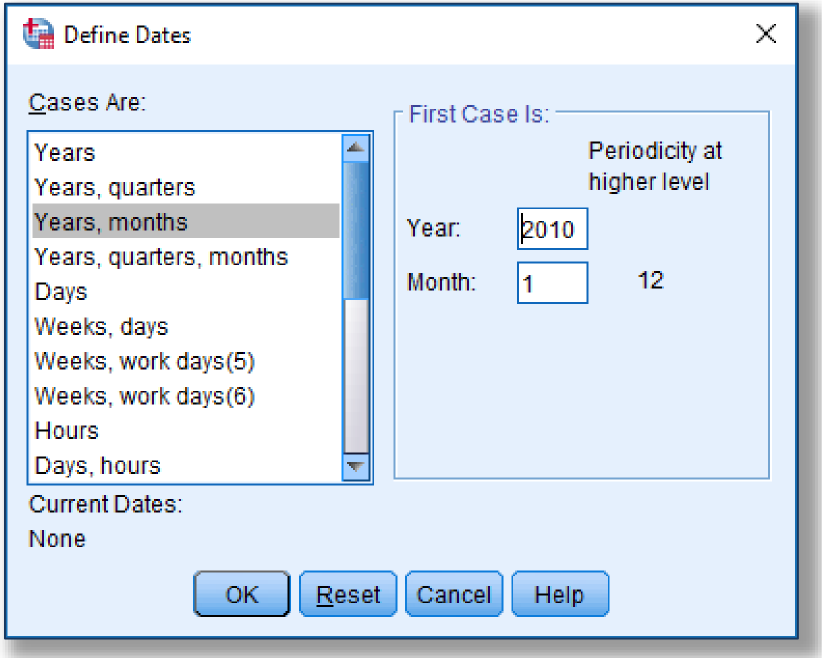
SMARTVISION Europe



Smoothing Series

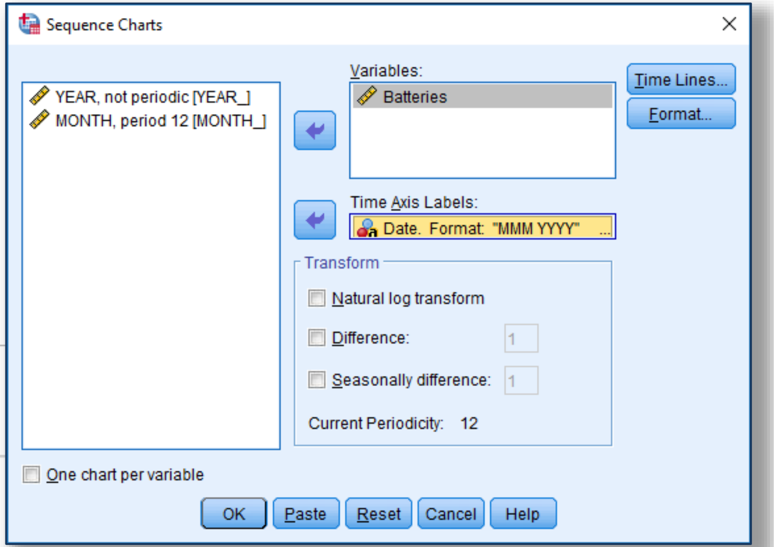
Demo 2: Smoothing Series

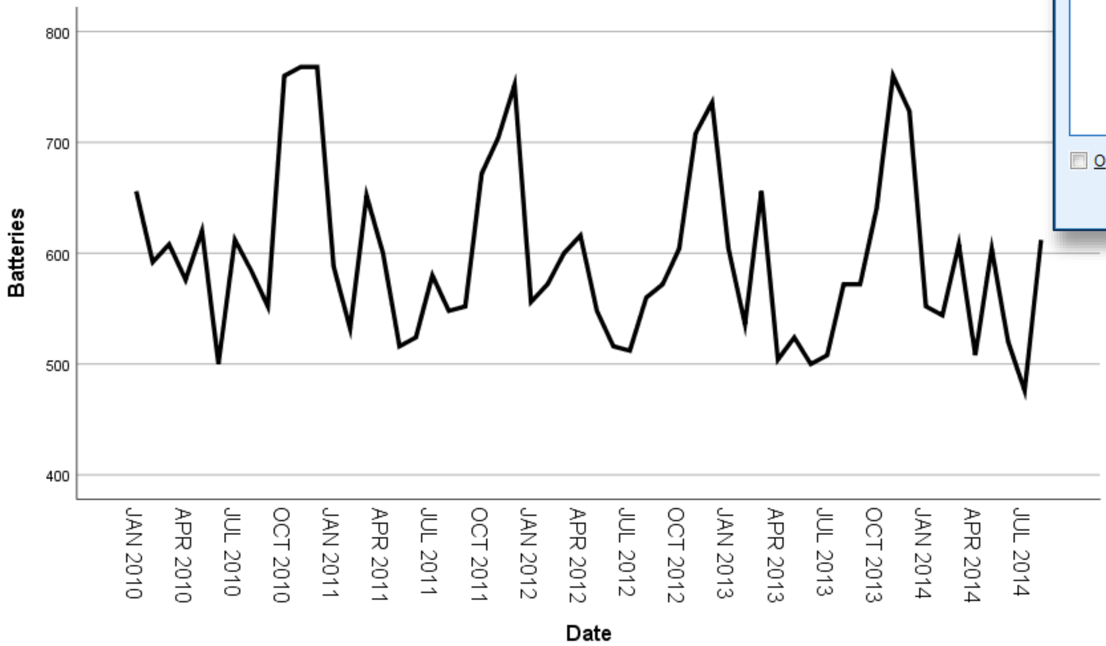
- Using the dataset **Battery Sales.sav**
- Click:
  - **Data**
    - **Define date and time**
- Define the periodicity as '**Years, Months**'
- Start at the **year 2010** and **month 1**
- Click: **OK**



Demo 2: Smoothing Series

- Create a sequence plot of the Batteries sales incorporating the newly created date variable





Demo 2: Smoothing Series

- We can use SPSS to create a **smoothed** version of this series
- Smoothing is usually done to reveal a clearer picture of the series by simplifying it and removing some of the randomness.
- Smoothing a series is similar to the methods that **pure** Time Series techniques employ to create a model for forecasting.
- The are *many different* ways that smoothing techniques can be applied.
- In the previous example we used **the mean of nearby points** to replace missing values. What if we used a similar technique to replace the entire series?

Moving Average Smoothing Example

- One of the most common forms of smoothing is using a moving average
- Calculating a Moving Average – with a span of 3 cases

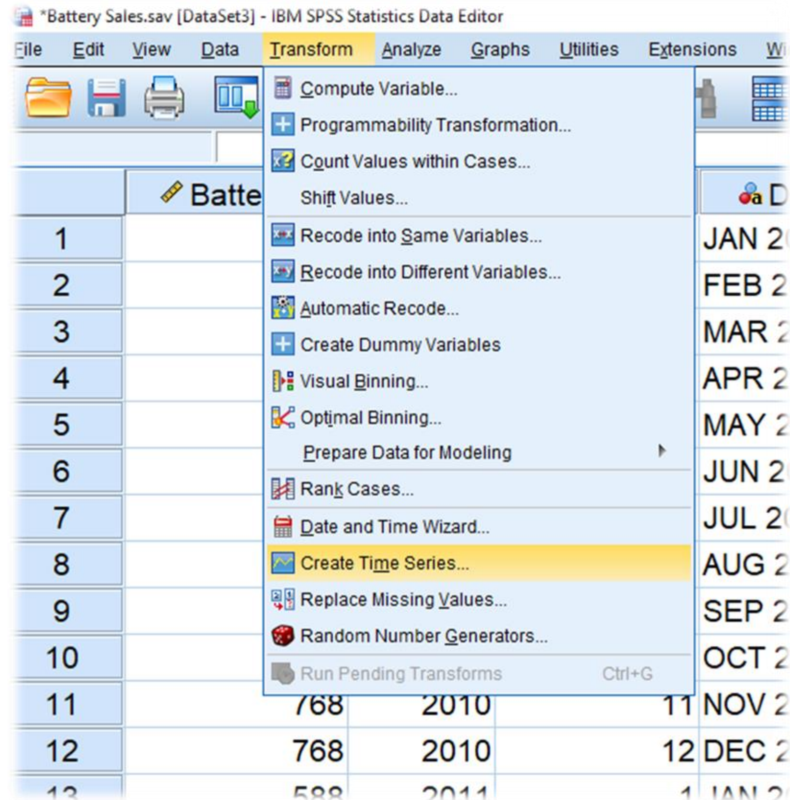
Case_ID	Variable	MA_3
1	1	
2	1	1.3
3	2	2
4	3	3.3
5	5	5.3
6	8	8.7
7	13	14
8	21	

Case_ID	Variable	MA_3
1	1	
2	1	1.3
3	2	2
4	3	3.3
5	5	5.3
6	8	8.7
7	13	14
8	21	

Case_ID	Variable	MA_2
1	1	
2	1	1.3
3	2	2
4	3	3.3
5	5	5.3
6	8	8.7
7	13	14
8	21	

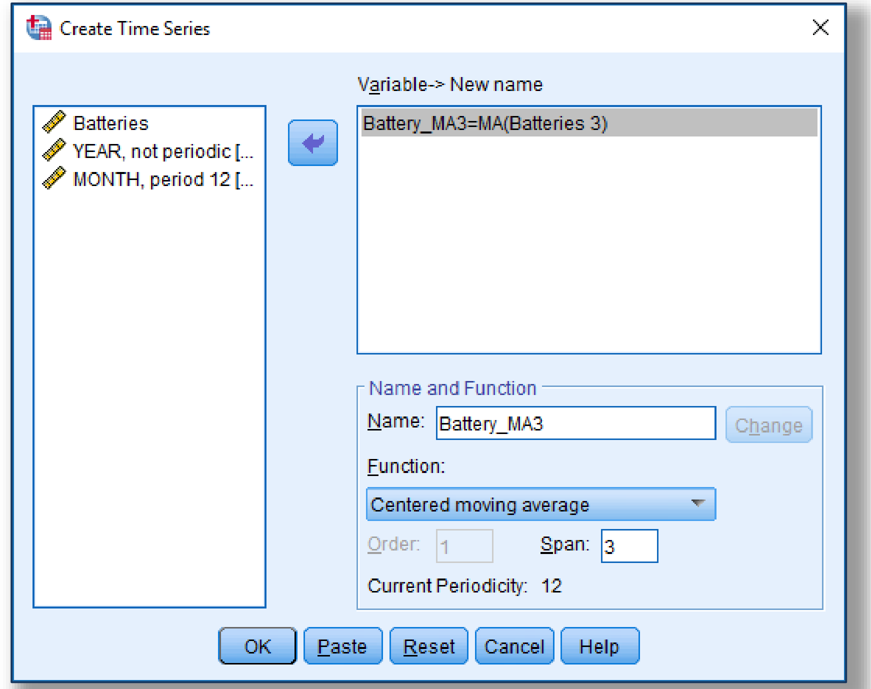
Demo 2: Smoothing Series

- To create a moving average of the Battery sales sequence using a span of 3 cases, click:
  - **Transform**
    - **Create Time Series**
  - This will generate a dialog allowing you to choose from a range of smoothed series options



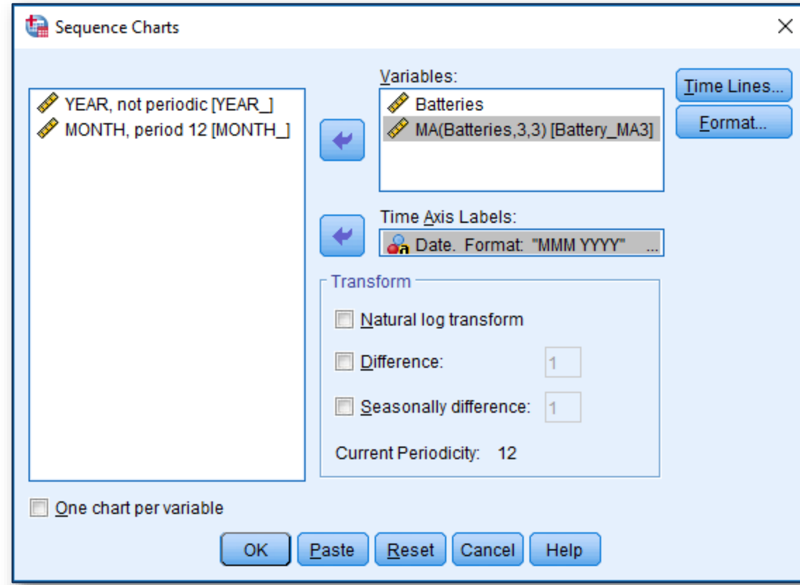
Demo 2: Smoothing Series

- In the dialog:
  - Choose **the Batteries variable**
  - Request **Centred moving average**
  - Choose a **span of 3**
  - Give the new series the name **'Battery_MA3'**
  - Remember to click **Change**
  - Click **OK** to create the new series



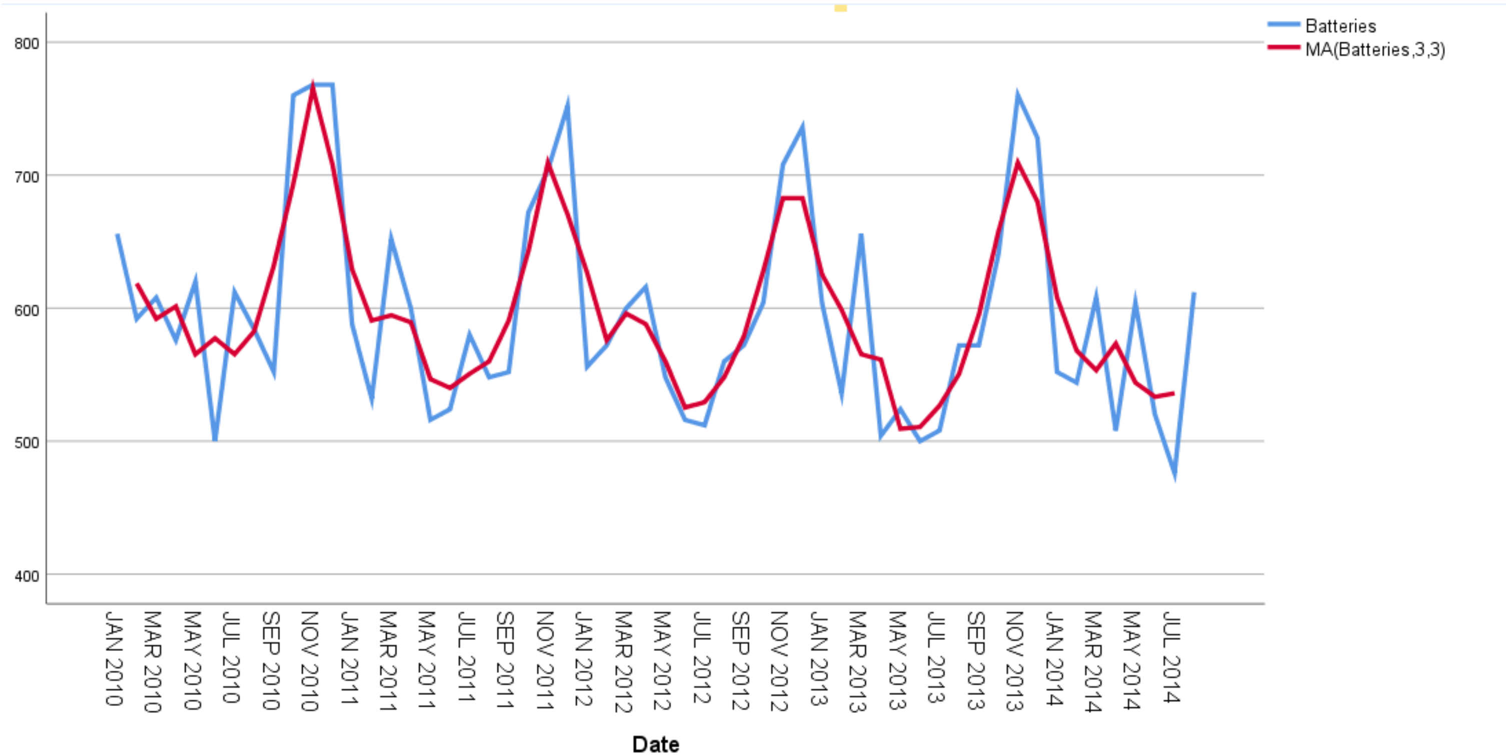
Demo 2: Smoothing Series

- Now let’s request a new sequence chart showing both the **Batteries** and the new **Battery_MA3** variables



Demo 2: Smoothing Series

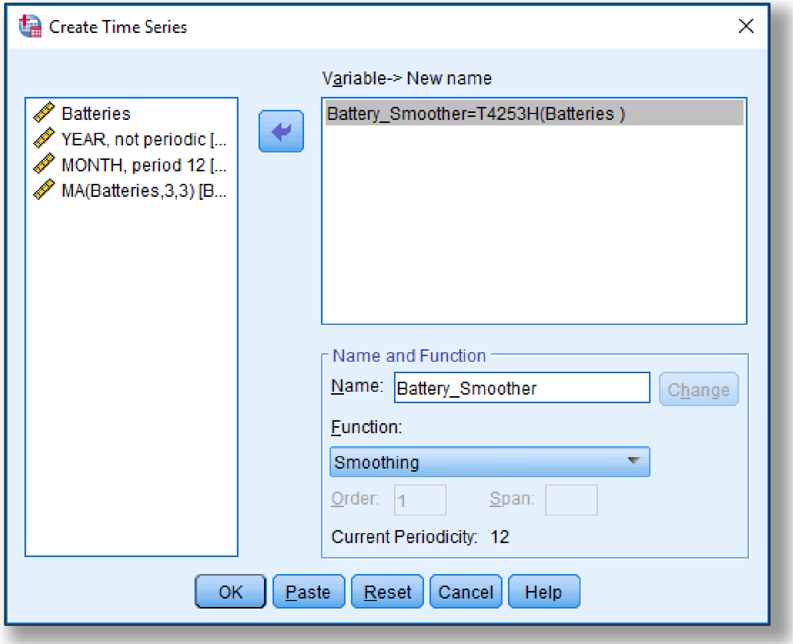
- The sequence chart clearly shows that the smoothed series moving average is a less noisy, albeit simplified version of the Batteries sales variable



SMARTVISION Europe

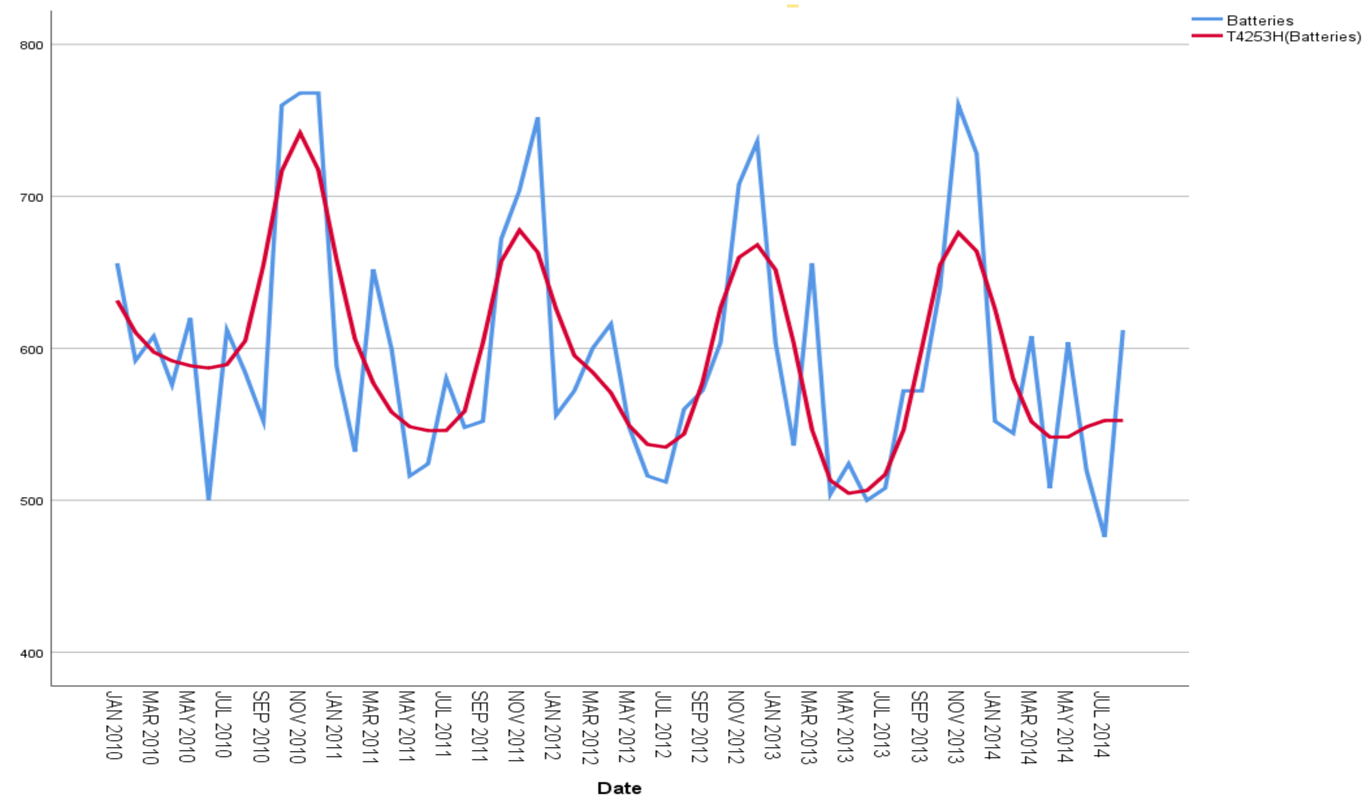
Demo 2: Smoothing Series

- Some analysts create complex smoothers where one smoothed sequence is smoothed again (and again) using different measures (such as means and medians) and various span widths
- In fact, among the smoothing options is a ‘pre-built’ smoother based on median values
- To illustrate, create a new smoothed sequence using the function **Smoothing**
- Notice it labels the new series as **T4253H** – this represents a smoother based on running medians that have been smoothed using a span of 4, then again based on a span of 2, followed by spans of 5 and 3 respectively.

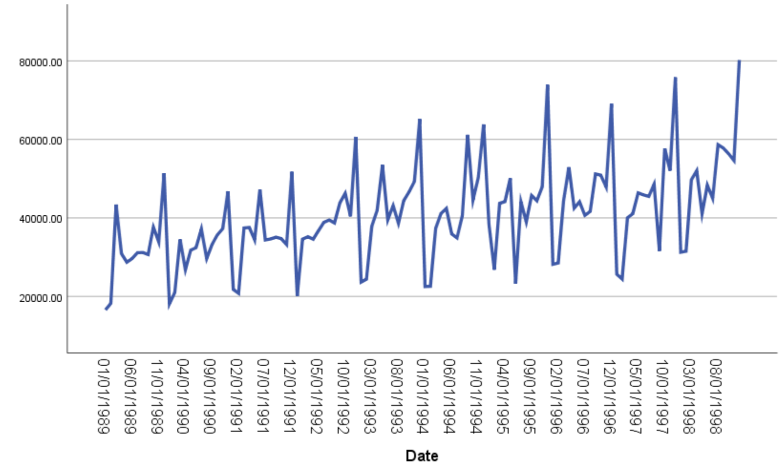


Demo 2: Smoothing Series

- The resultant sequence chart shows that the T4253H Smoother further simplifies the series showing clear seasonal peaks and troughs



SMARTVISION Europe



Exponential Smoothing Methods

Exponential Smoothing

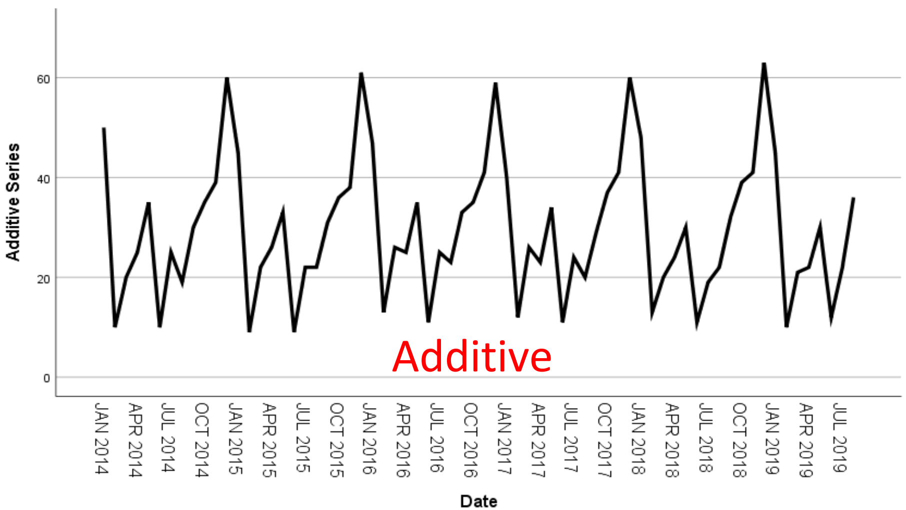
- Using the simple moving average smoothing techniques that we saw earlier, each time point in the calculation has equal weight. For example, by using a span of say 5, the cases that are two time points away are treated as equally as those that are one time point away when the moving average is calculated.
- In ***Exponential* smoothing**, values that are *closer* (in time) are *given greater weight* than those that are further away.
- This approach can be employed in forecasting where the values that are *more recent* have a greater influence on estimating the future than those that are less recent.
- Simple exponential models use a parameter value called **Alpha (α)**. This is a weight ranging between 0 and 1. The closer alpha is to 1, the more the exponential smoothing weights the most *recent* observation. So **Alpha (α) = 0.85** relies more on the most recent observations than **Alpha (α) = 0.7**

Additive vs Multiplicative Series

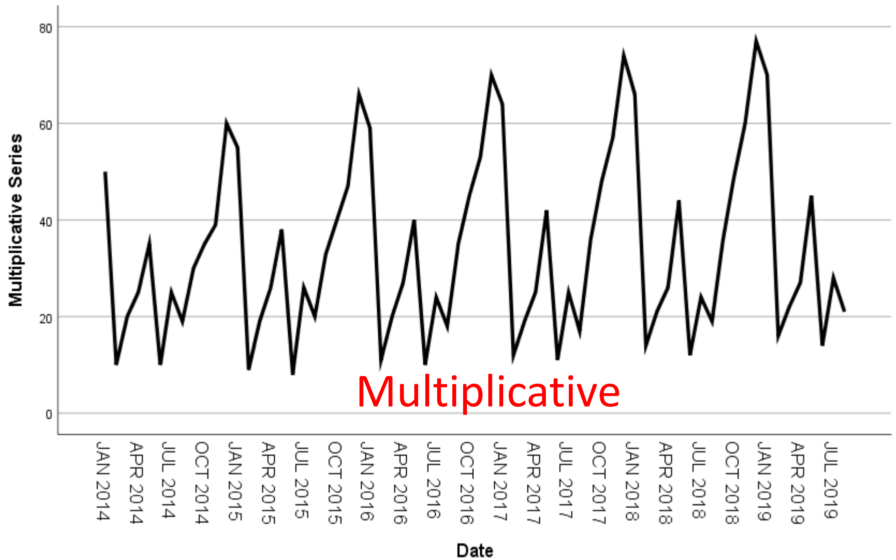
- Before we go any further, it's useful to understand something about two fundamentally different kinds of series: **Additive** and **Multiplicative**
- **Additive** series show a more or less constant average range of variation across time. In other words the difference between the values in June and December are, on average, the same from one year to the next. This can be true for series where sample measurements are taken (such as pollen counts) over time.
- In a **Multiplicative** series however, the differences appear to get greater over time. This often occurs where with series dealing with quantities of goods or demand for services. Although the *proportional difference* between June and December can be the same, the **absolute** difference can be increasing.
- As we shall see, the SPSS Forecasting module can use the **Expert Modeler** mode to automatically detect whether a series is **Additive** or **Multiplicative**

Additive vs Multiplicative Series

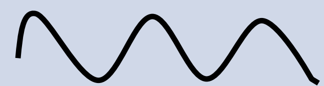
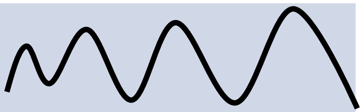
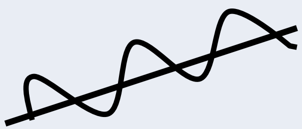
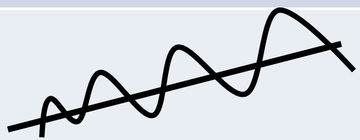
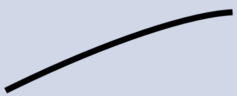
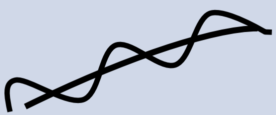
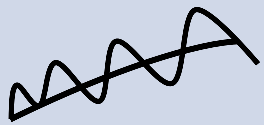
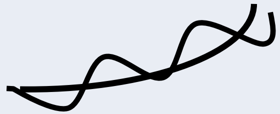
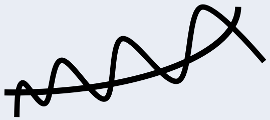
Seasonal Variation Constant



Seasonal Variation Changing

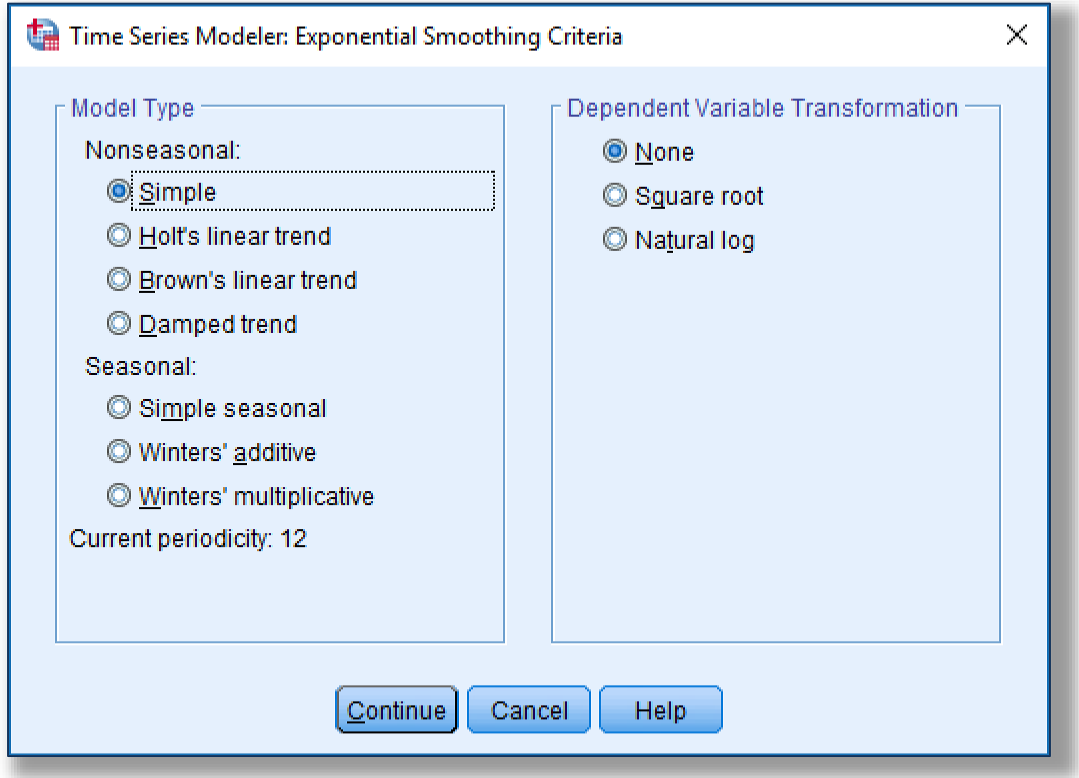


Basic Exponential Smoothing Methods

	Non Seasonal	Additive Seasonal	Multiplicative Seasonal
Constant Level			
Linear Trend			
Damped Trend			
Exponential Trend			

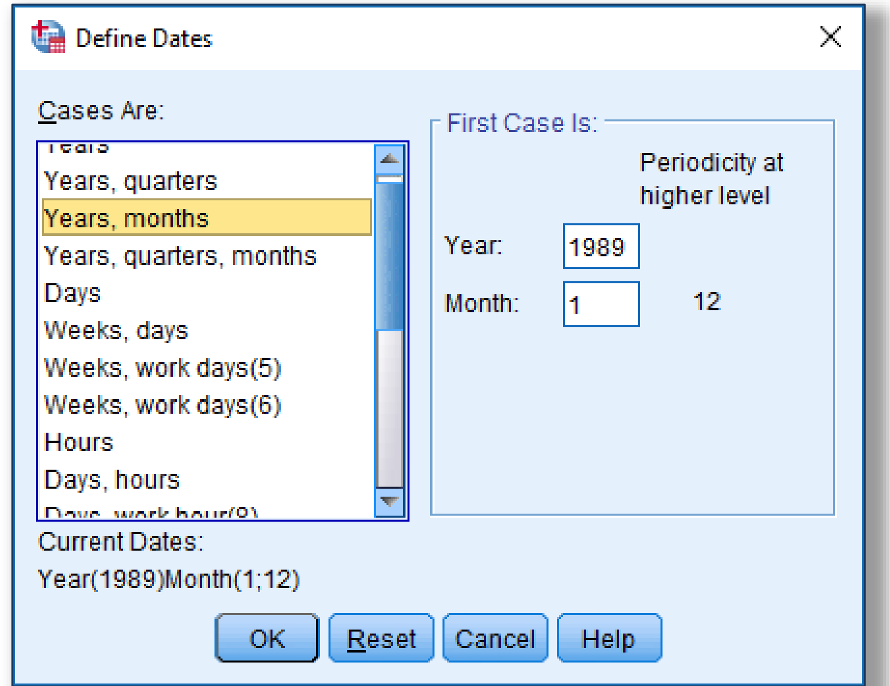
SPSS Statistics Exponential Smoothing Models

- SPSS Statistics has **7** standard Exponential smoothing models
- **3** of the standard models are for *seasonal* data.
- Including:
  - a model for an **additive** series
  - a model for a **multiplicative** series
- SPSS Forecasting **Expert Modeler** *automatically* chooses a model type for the series



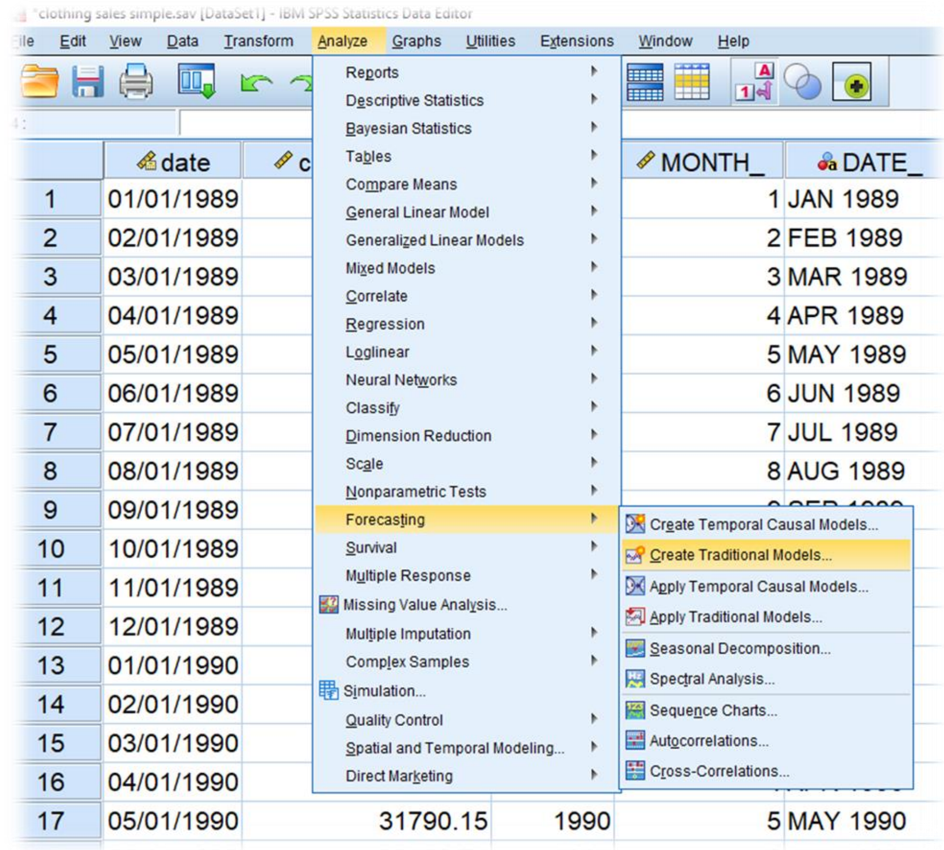
Demo 3: Forecasting with SPSS Expert Modeler

- Use the file **Clothing sales simple.sav**
- From the main menu click:
  - **Data**
    - **Define date and time**
- Define the periodicity as **Years, months** beginning in January 1989



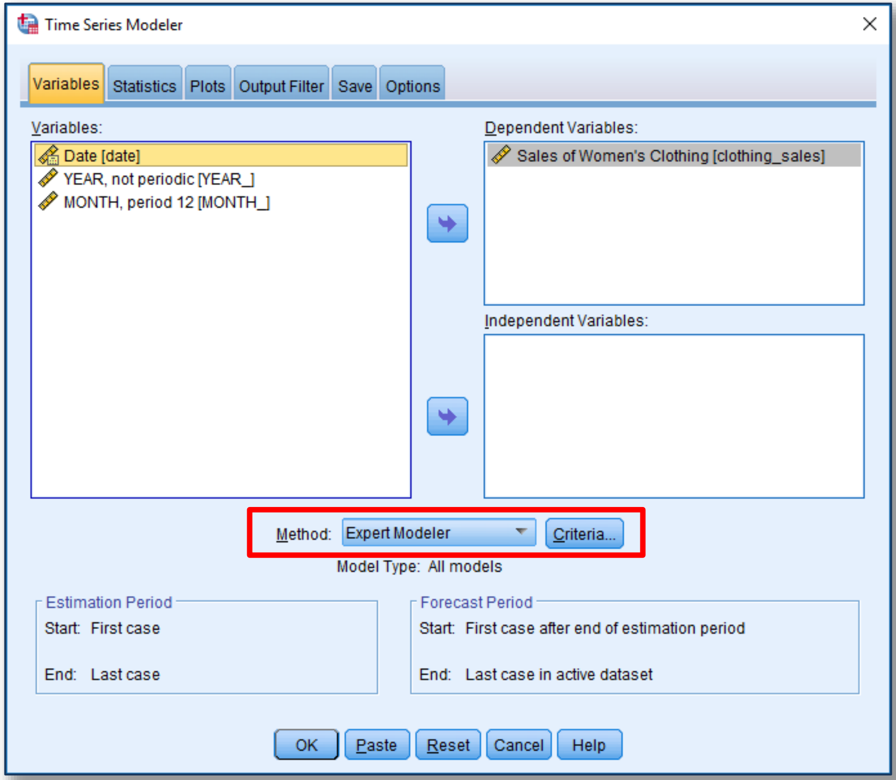
Demo 3: Forecasting with SPSS Expert Modeler

- To create a forecast using Expert Modeler, from the main menu click:
  - **Analyze**
    - **Forecasting**
      - **Create Traditional Models**



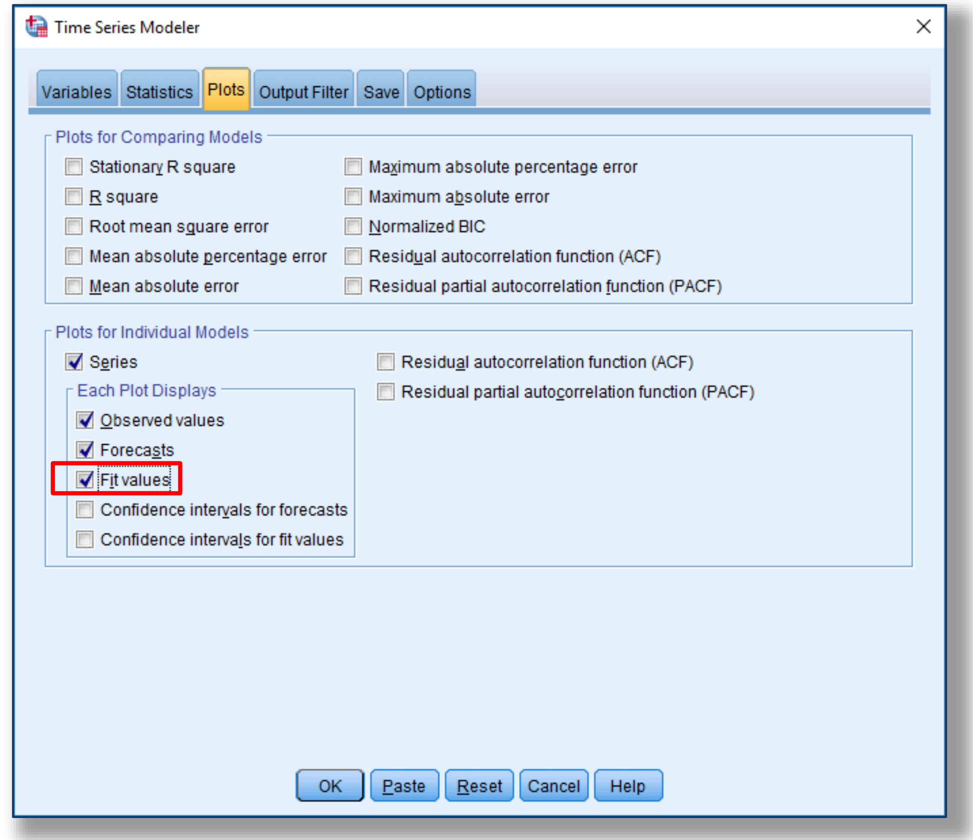
Demo 3: Forecasting with SPSS Expert Modeler

- Specify **Sales of Women's' Clothing** as the dependent variable
- Note that the default **Method** for model fitting is **Expert Modeler**
- Based on a measure of model fit, this method will try a number of model types and select the one with the best fit over the series
- This method is especially useful when fitting models against many series at the same time
- Nevertheless, you can override this functionality and specify your own model(s)



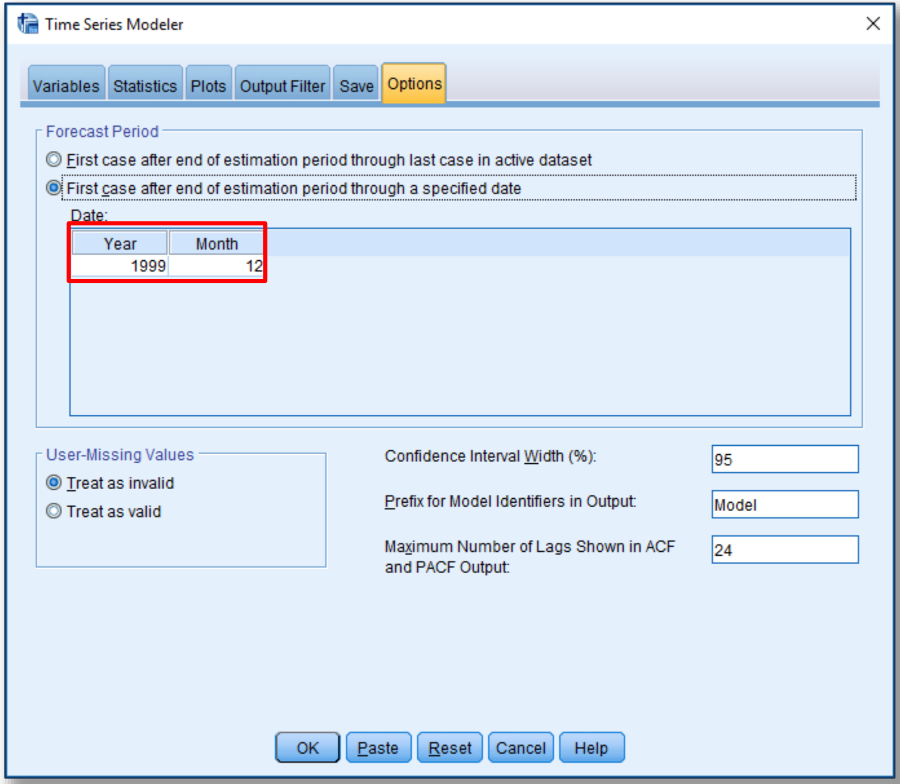
Demo 3: Forecasting with SPSS Expert Modeler

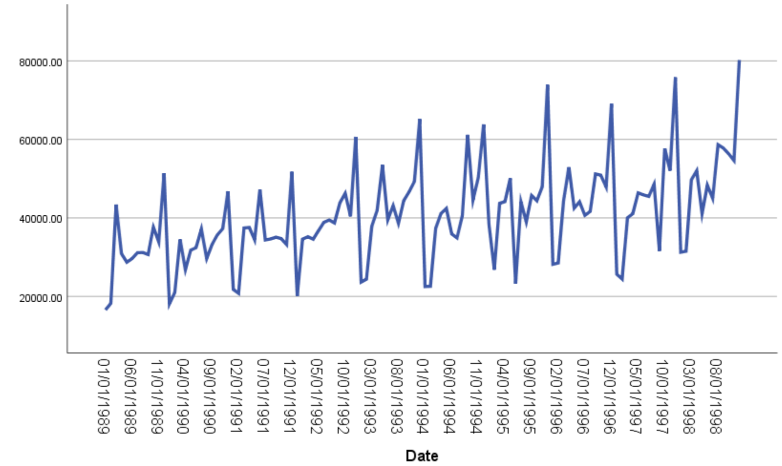
- Click the tab:
  - **Plots**
- Check the box:
  - **Fit values**
- This will enable us to see how the model as been fitted to actual series in the resultant time series plot



Demo 3: Forecasting with SPSS Expert Modeler

- Click the tab:
  - **Options**
- Click the radio button:
  - **First case after end of estimation through a specified date**
- Specify the Year and Month as:
  - **1999 and 12**
- Click:
  - **OK**





Interpreting Output and Model Fit

Demo 3: Forecasting with SPSS Expert Modeler

- Note that expert modeler chooses **Winters' Additive** as the best model type – this model is used with series showing a trend and seasonal effects

Model Description

			Model Type
Model ID	Sales of Women's Clothing	Model_1	Winters' Additive

Model Summary

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile 5	10	25	50	75	90	95
Stationary R-squared	.735	.	.735	.735	.735	.735	.735	.735	.735	.735	.735
R-squared	.815	.	.815	.815	.815	.815	.815	.815	.815	.815	.815
RMSE	5296.069	.	5296.069	5296.069	5296.069	5296.069	5296.069	5296.069	5296.069	5296.069	5296.069
MAPE	9.412	.	9.412	9.412	9.412	9.412	9.412	9.412	9.412	9.412	9.412
MaxAPE	83.036	.	83.036	83.036	83.036	83.036	83.036	83.036	83.036	83.036	83.036
MAE	3624.699	.	3624.699	3624.699	3624.699	3624.699	3624.699	3624.699	3624.699	3624.699	3624.699
MaxAE	19302.778	.	19302.778	19302.778	19302.778	19302.778	19302.778	19302.778	19302.778	19302.778	19302.778
Normalized BIC	17.269	.	17.269	17.269	17.269	17.269	17.269	17.269	17.269	17.269	17.269

We can ignore these statistical values as they are used to compare the performance of *multiple models* so aren't relevant here

Model Fit Statistics

- **Stationary R square –** A *stationary* model is effectively one with the trend removed so that the values have the same variance and mean over time. This measure is therefore appropriate when the series has a trend or obvious seasonality. As such, it attempts to estimate how well the model fits if we ignore the effects these elements. Stationary R-square values can be negative but cannot exceed +1. Positive values mean that the model under consideration is better than the baseline model.
- **R square -** An estimate of the proportion of the total variation in the series that is explained by the model. This measure is most useful when the series is already stationary (no trend). Again, R-Square values can be negative but cannot exceed +1. Negative values mean that the model under consideration is worse than a baseline model. This measure is rarely used to compare model performance in Time Series.
- **Root mean square error (RSME) -** Squared errors are based on the square of the differences between *the fitted values and the observed values*. It's similar to a standard deviation value.
- **Mean absolute percentage error (MAPE) –** The average error values in percentage terms. Values such as 0.15 equate to an average of 15% error.
- **Mean absolute error -** Mean of the absolute values of the forecast errors. MAE is in the same units as the dependent series. MAE is appropriate when the cost of the forecast errors is proportional to the absolute size of the forecast error.
- **Maximum absolute percentage error -** The largest forecast error, expressed as a percentages. This measure gives a worst case scenario indication of model performance. It works best if there are no extremes to the data.
- **Maximum absolute error -**The largest forecast error. Expressed in the same units as the dependent series. This measure gives a worst case scenario indication of model performance.
- **Normalized BIC -** The Normalized *Bayesian Information Criterion* (BIC) fit measure that enables you to compare different models for the same series. Normalized BIC "rewards" simpler models that fit better, while it "penalizes" models that use more parameters. It is based on a mean squared error so smaller values indicate a better fit. **This is the fit measure that Expert Modeler uses when comparing candidate models.**

Demo 3: Forecasting with SPSS Expert Modeler

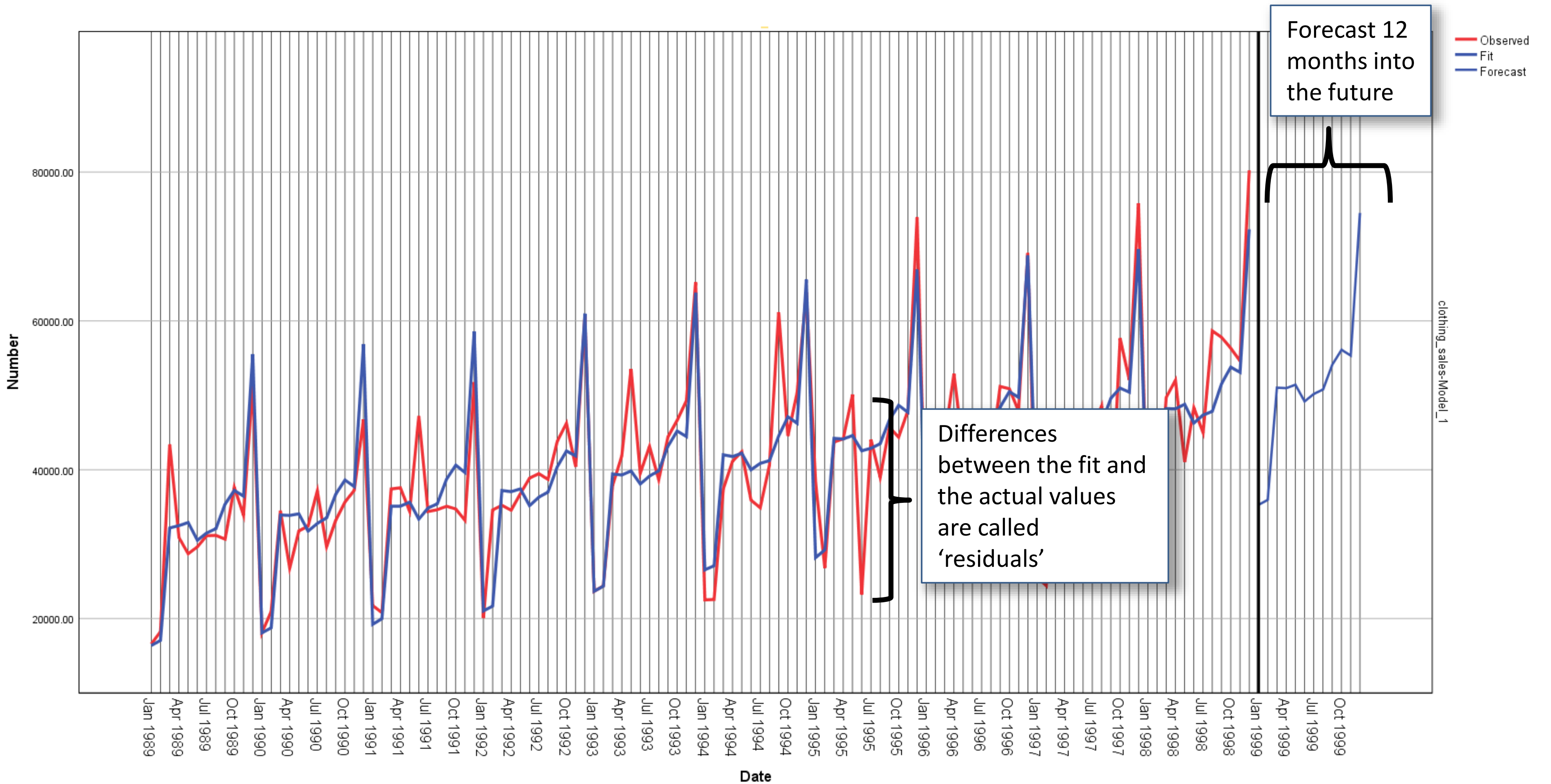
- **Ljung-Box Q: a lack of fit test to check that the model is correctly specified**

Model Statistics

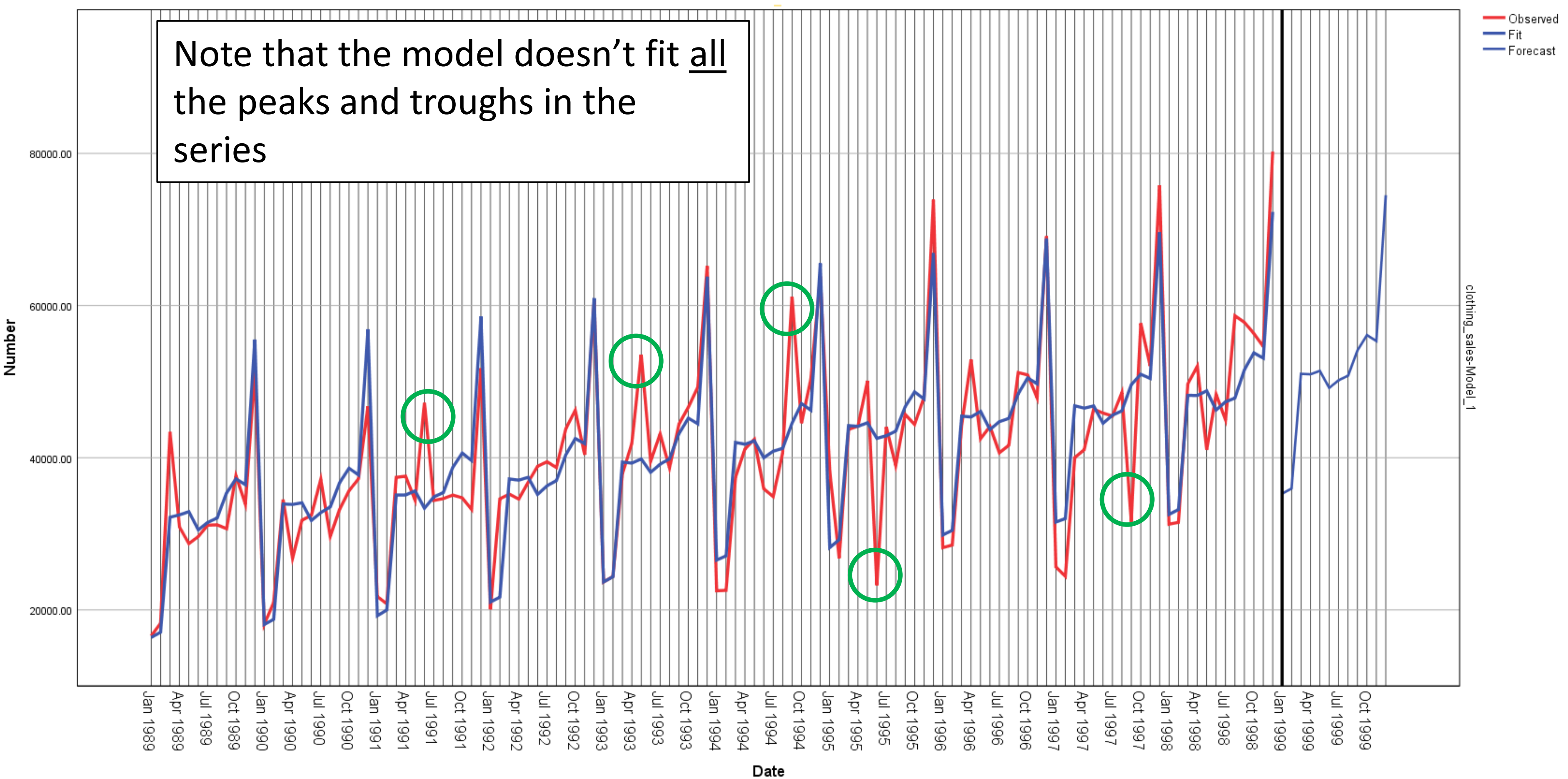
Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Sales of Women's Clothing-Model_1	0	.735	23.635	15	.072	0

- More specifically it tests for autocorrelation among the residuals (errors) of the model. Autocorrelation refers to a situation where knowing the previous value in a sequence might help you estimate the next value. After the model has been fitted, the correlation between the residuals in the sequence should be random (white noise).
- Values in the Sig. column above 0.05 indicate that the model doesn't leave any significant correlations after the model has been specified. So we *might* assume that it's doing a reasonable job (overall) of fitting the series.

Demo 3: Forecasting with SPSS Expert Modeler



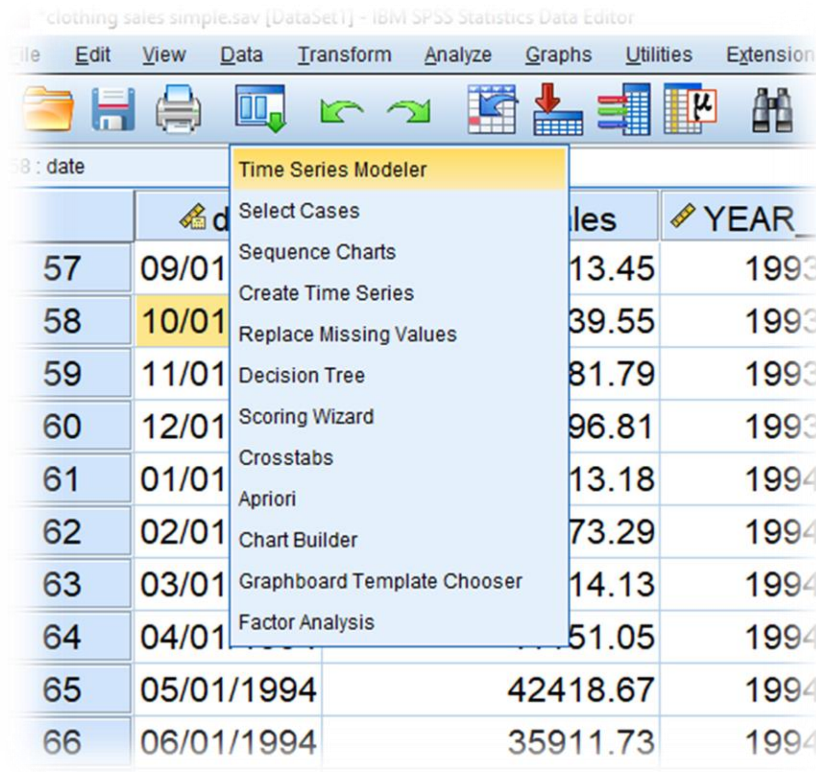
Demo 3: Forecasting with SPSS Expert Modeler



Demo 3: Forecasting with SPSS Expert Modeler

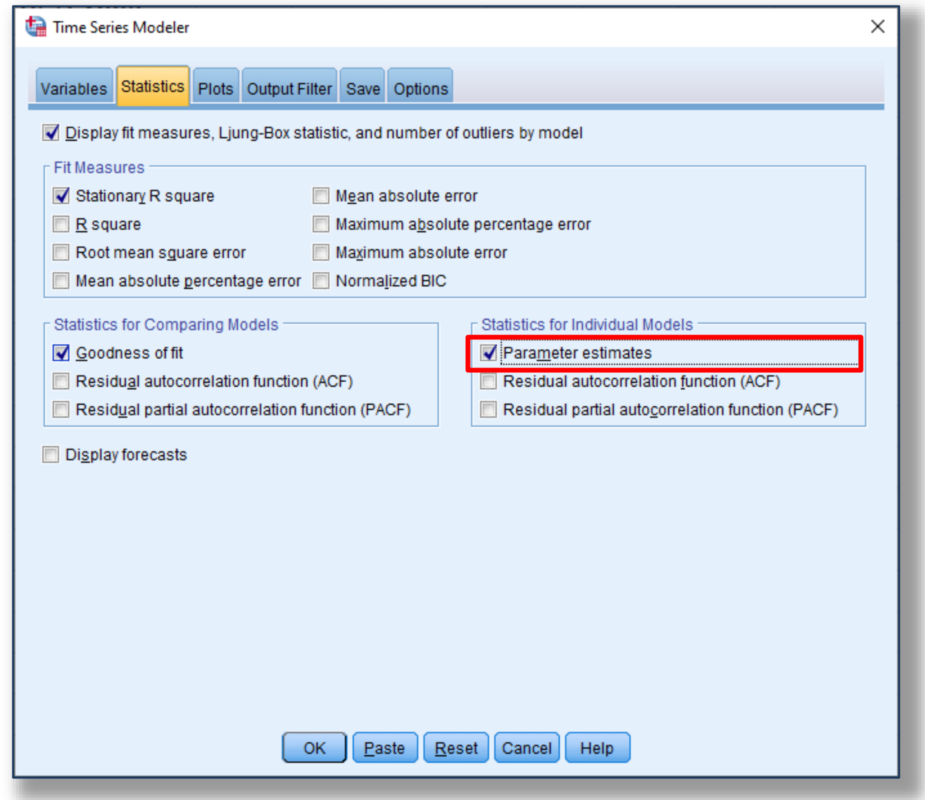
- Use the **Recall Dialog** button on the Data Editor Toolbar to request the last Time Series Modeler dialog box.
- Click:

  - **Time Series Modeler**



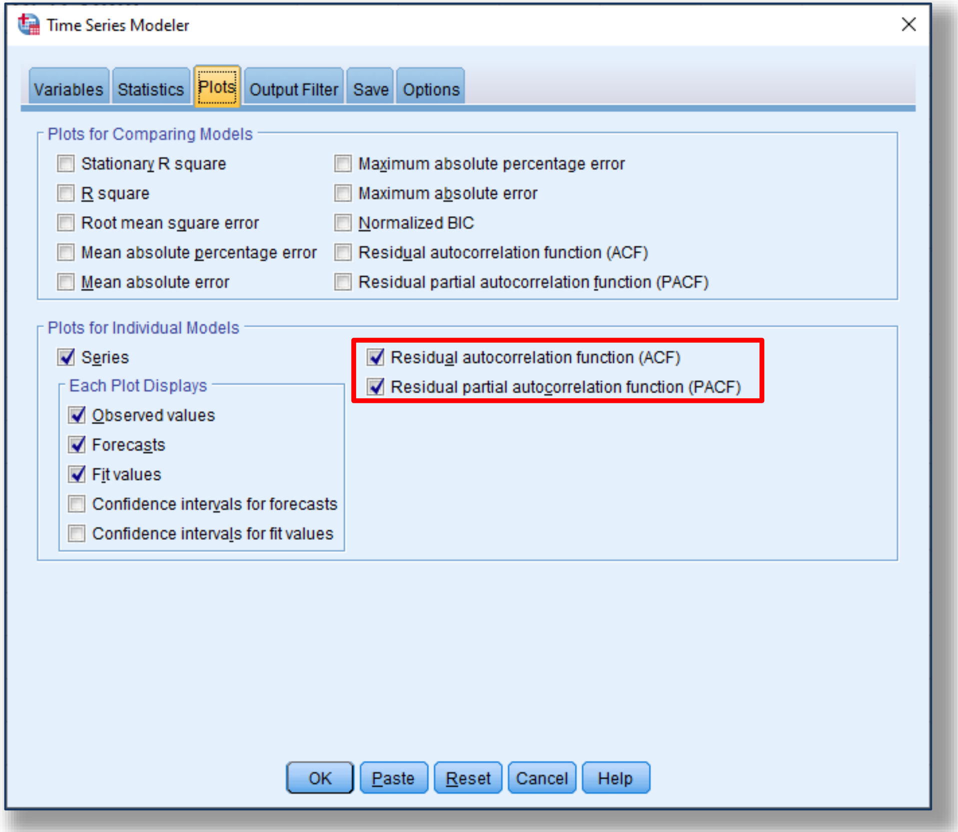
Demo 3: Forecasting with SPSS Expert Modeler

- Within the **Statistics** tab, check the box marked **Parameter estimates**
- This will produce output showing coefficient values that indicate how the model has been specified by Expert Modeler



Demo 3: Forecasting with SPSS Expert Modeler

- Within the **Plots** tab, check the boxes marked:
  - **Residual autocorrelation function (ACF)**
  - **Residual partial autocorrelation function (PACF)**
- This will produce *diagnostic* charts showing how strongly previous *residual* values correlate with more recent values
- These charts represent the correlations between time lags.
- In this case a lag of 1 refers to the previous month's residual. A lag of 2 refers to the residuals 2 months ago and so on.



Demo 3: Forecasting with SPSS Expert Modeler

- **The Model Parameters table shows 3 components: Alpha (Level), Gamma (Trend) and Delta (Season)**

Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	No Transformation	Alpha (Level)	.034	.019	1.814	.072
		Gamma (Trend)	2.406E-7	.001	.000	1.000
		Delta (Season)	.001	.033	.030	.976

- The estimates for the three component coefficients are all close to zero. The standard errors are also small so the significance tests indicate that they are could be in fact be zero (the values in the Sig. column are all above 0.05).
- In reality, we don't interpret this result as meaning that the components are not adding to the model. As we would, for example, in Linear Regression.

Demo 3: Forecasting with SPSS Expert Modeler

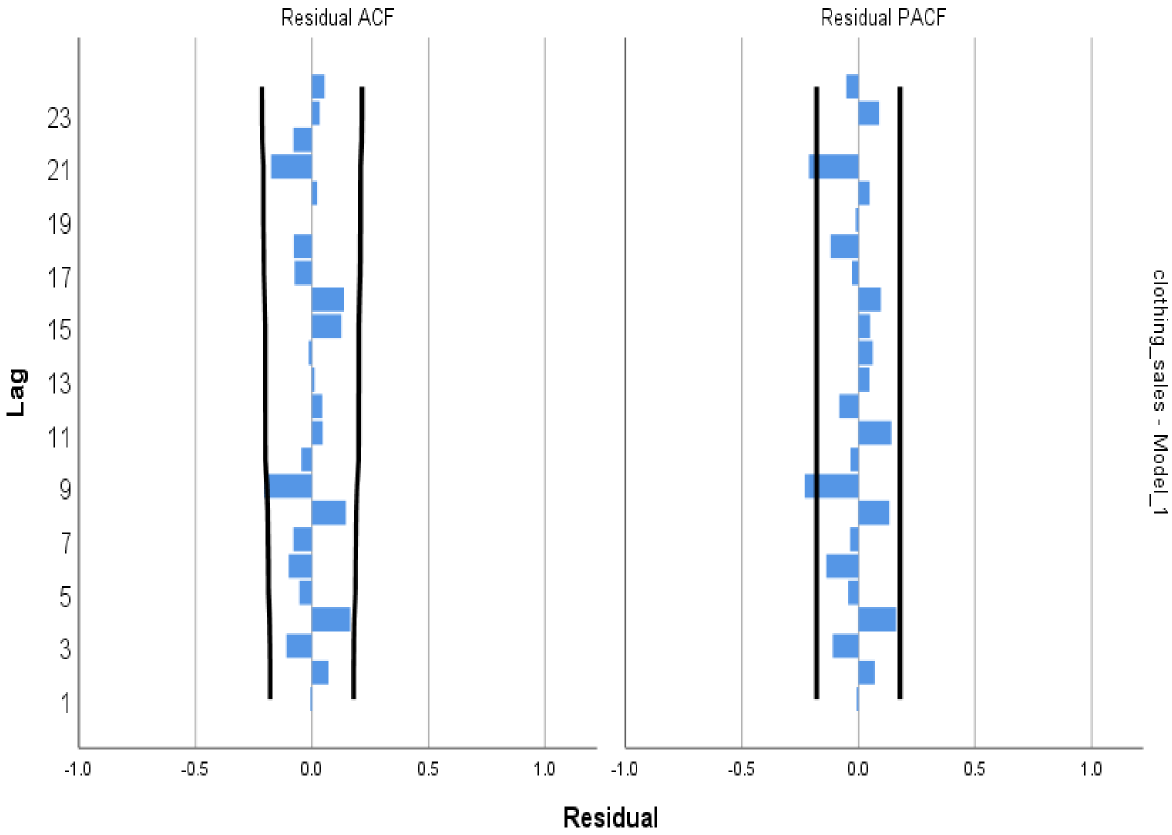
Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	No Transformation	Alpha (Level)	.034	.019	1.814	.072
		Gamma (Trend)	2.406E-7	.001	.000	1.000
		Delta (Season)	.001	.033	.030	.976

- The **Alpha** value is the smoothing parameter that indicates the degree to which more recent observations are used in the model fit. When Alpha is 1, only the single the most recent observation is used. When it is equal to 0 then all the previous values have equal weight.
- The **Gamma** value captures the trend component of the series. However, there can still be a trend in the series even if the Gamma coefficient estimate is 0. A Gamma value of 0 simply indicates that the trend is constant over time. A value closer to 1 would indicate that the trend is changing and so more recent observations would be used to account for this component.
- The **Delta** value looks at seasonality. A value closer to 1 would indicate that greater weight is given to values from the previous season whereas a value closer to 0 would indicate that all the previous seasons are involved in the model formulation.

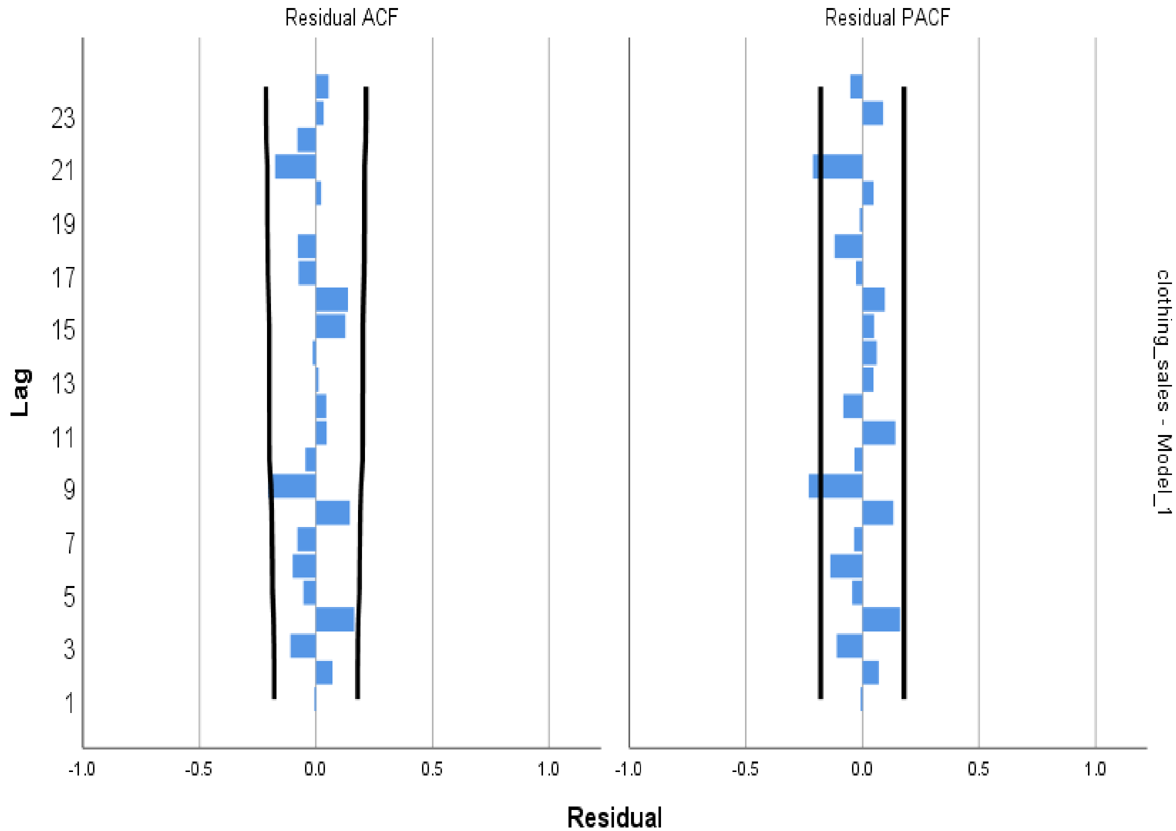
Demo 3: Forecasting with SPSS Expert Modeler

- The bars shown in the autocorrelation plots help us to see if there is a signal in the residuals.
- We can see here that the strongest correlation is at lag 9. Indicating that the residual values at any given point show a correlation the values 9 months previously.
- Bars that exceed the black 95% confidence lines, indicate that the autocorrelation at that lag point may be regarded as statistically significant.



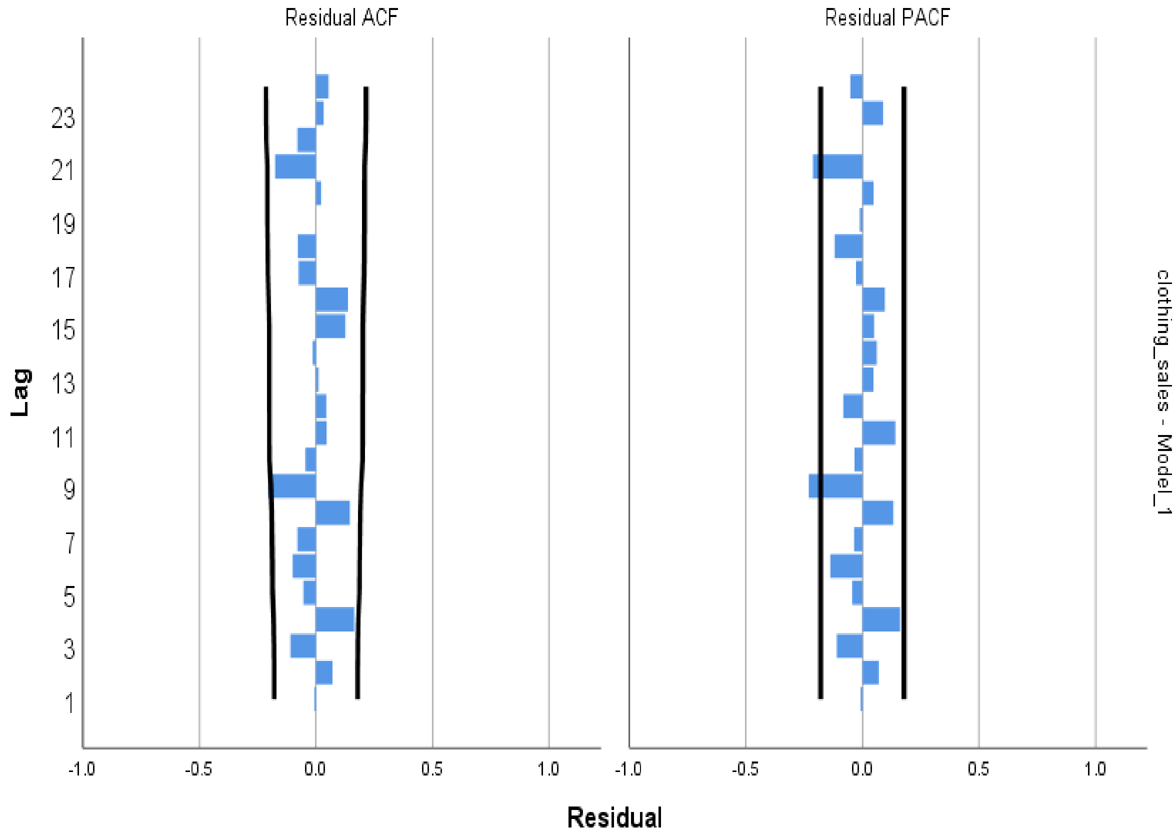
Demo 3: Forecasting with SPSS Expert Modeler

- Autocorrelation plots are a handy way to check if the model *systematically* overestimates or underestimates a value rather than *randomly.*
- The PACF is more sensitive form of the ACF plot. The PACF plot shows *Partial* autocorrelation. Partial correlations are those that *control for* the presence of other variables.
- Partial Autocorrelations show the correlation at a given lag point *controlling for* residuals values *up to that time point.*



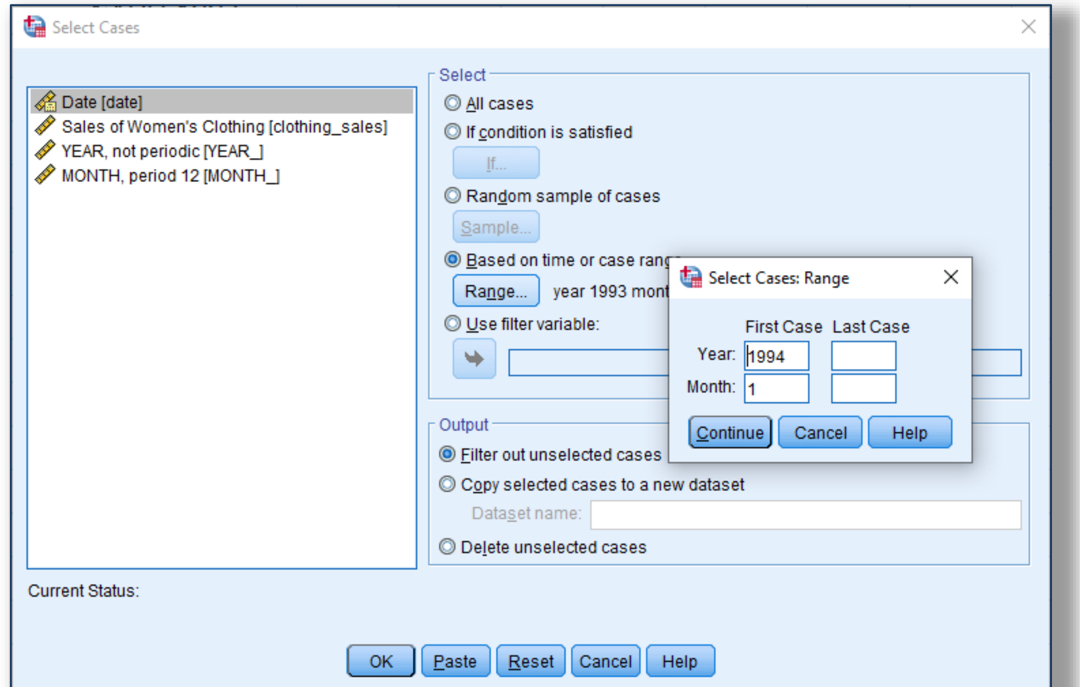
Demo 3: Forecasting with SPSS Expert Modeler

- Although the ACF indicates borderline strong signal at lag point 9, it's unlikely to cause problems.
- Generally speaking the higher order autocorrelations (those at larger lag points) are less troublesome than those that are indicate autocorrelation with more recent values (i.e. lower order lags).
- ACF and PACF charts are especially useful for analysts trying to fit ARIMA models (of which more later).



Demo 3: Forecasting with SPSS Expert Modeler

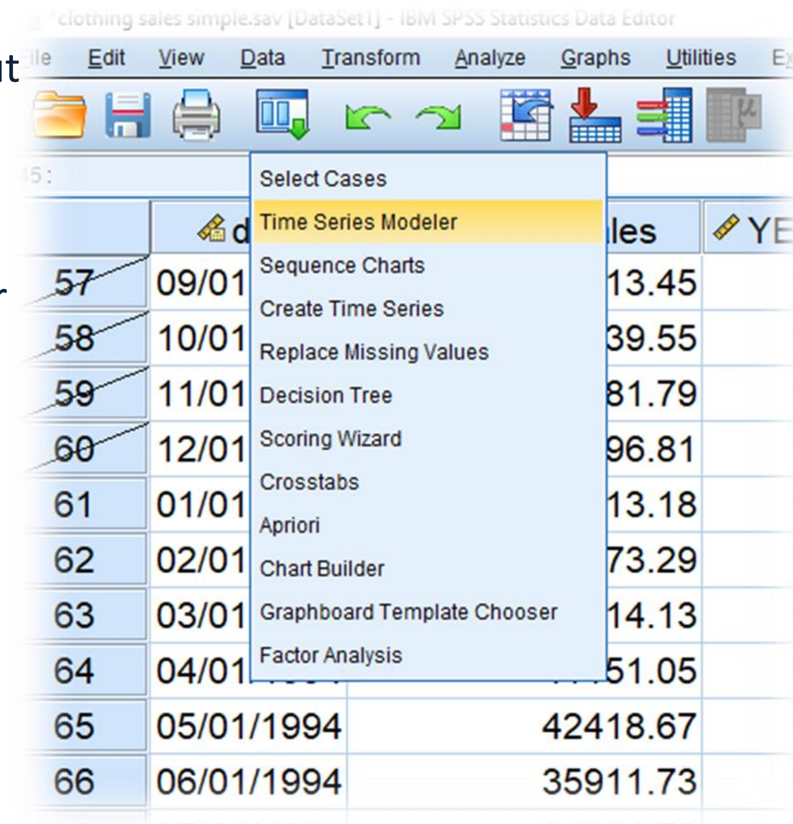
- Finally, lets look at the effect of changing the *range of time* that we submit to the Expert Modeller algorithm.
- From the main menu, click:
  - **Data**
    - **Select Cases**
- Click the radio button '**Based on time or case range'** and click:
  - **Range**
- Specify the first case as:
  - Year: **1994**
  - Month: **1**
- Click:
  - **Continue**
    - **OK**



Demo 3: Forecasting with SPSS Expert Modeler

- The **Select Cases** procedure has now filtered out the first 4 years of data.
- To re-run the Time Series Modeler simply use the **Recall Dialog** button on the Data Editor Toolbar to request the last Time Series Modeler dialog box.
- Click:

  - **Time Series Modeler**
    - **OK**



Demo 3: Forecasting with SPSS Expert Modeler

- Although the chosen Model Type is still Winter's Additive. The fit statistics are slightly different when using only the last 5 years of data

Fit Statistic	Mean
Stationary R-squared	.735
R-squared	.815
RMSE	5296.069
MAPE	9.412
MaxAPE	83.036
MAE	3624.699
MaxAE	19302.778
Normalized BIC	17.269

Model 1: 1989 - 1998

Fit Statistic	Mean
Stationary R-squared	.822
R-squared	.815
RMSE	5552.480
MAPE	9.634
MaxAPE	68.892
MAE	3897.216
MaxAE	19307.360
Normalized BIC	17.449

Model 2: 1994 - 1998

Demo 3: Forecasting with SPSS Expert Modeler

Model 1: 1989 - 1998

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Sales of Women's Clothing-Model_1	0	.735	23.635	15	.072	0

Model 2: 1994 - 1998

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Sales of Women's Clothing-Model_1	0	.822	21.646	15	.117	0

Demo 3: Forecasting with SPSS Expert Modeler

Model 1: 1989 - 1998

Exponential Smoothing Model Parameters

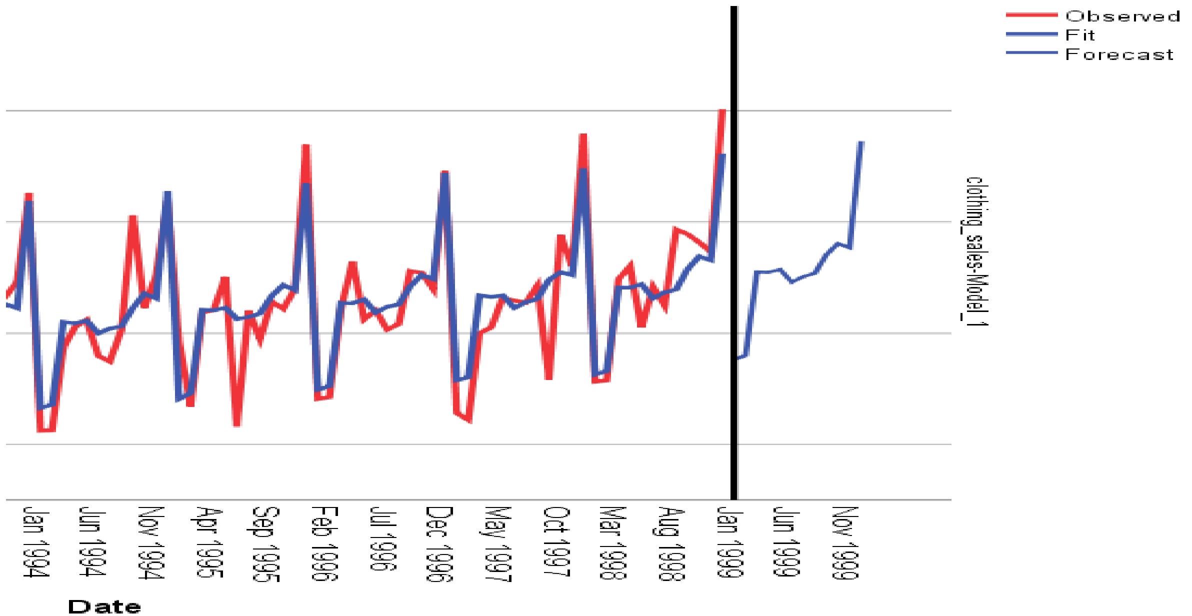
Model			Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	No Transformation	Alpha (Level)	.034	.019	1.814	.072
		Gamma (Trend)	2.406E-7	.001	.000	1.000
		Delta (Season)	.001	.033	.030	.976

Model 2: 1994 - 1998

Exponential Smoothing Model Parameters

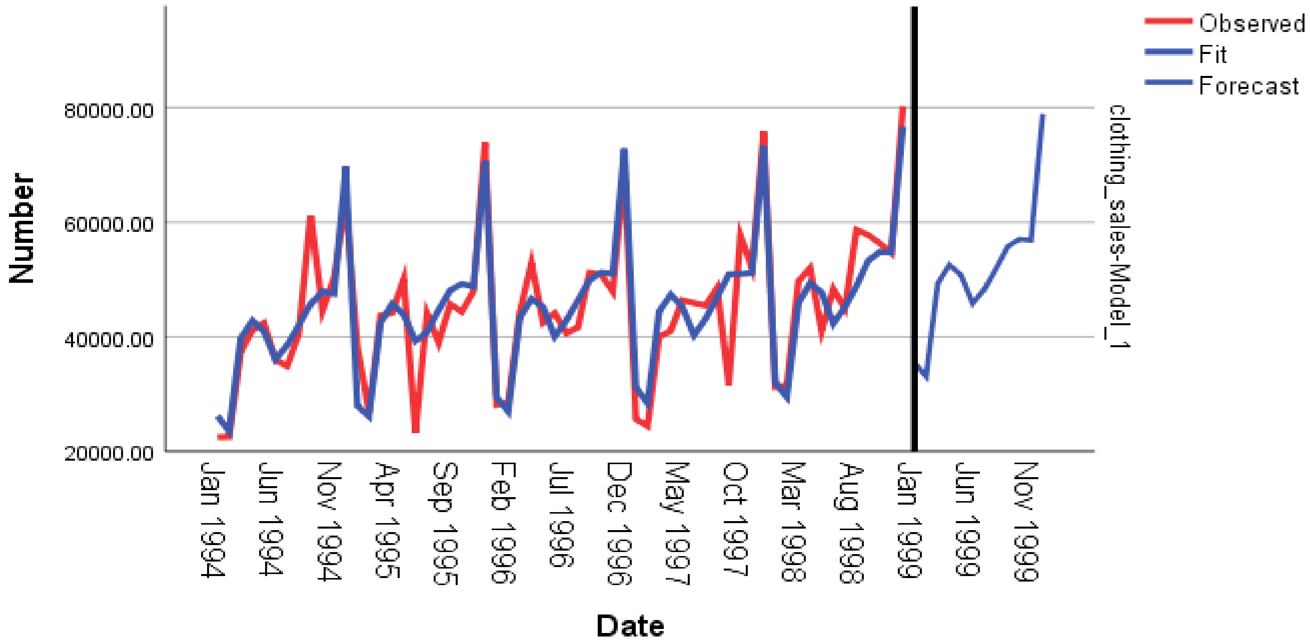
Model			Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	No Transformation	Alpha (Level)	.061	.036	1.686	.097
		Gamma (Trend)	2.516E-5	.002	.011	.991
		Delta (Season)	.001	.071	.014	.989

Demo 3: Forecasting with SPSS Expert Modeler



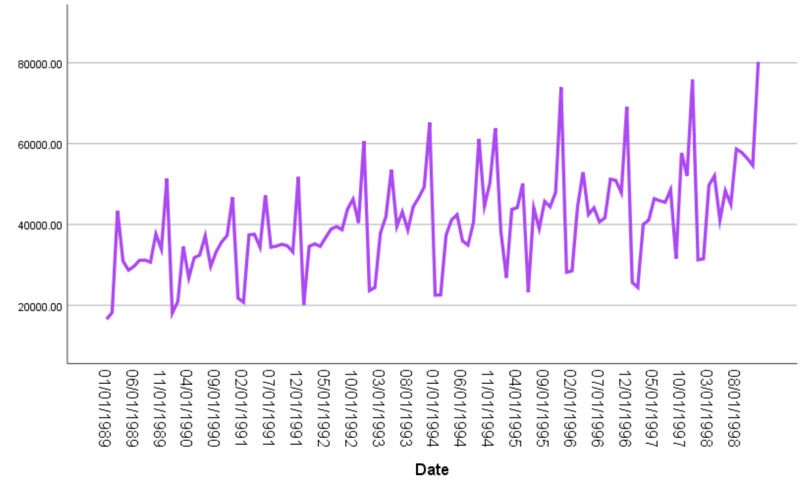
Model 1: 1989- 1998 (showing only last 4 years)

Demo 3: Forecasting with SPSS Expert Modeler



Model 2: 1994 - 1998

A SELECT INTERNATIONAL COMPANY



Using Predictor Fields with ARIMA

Using predictor fields with ARIMA

- As we have seen, the Expert Modeler in SPSS Forecasting attempts to automatically find the best-fitting model for each dependent series. So far we have only been working with Exponential Smoothing models However, by default the Expert Modeler considers two types of Time Series model: Exponential Smoothing and **ARIMA**.
- If we choose to work with any **independent** (predictor) variables, then the Expert Modeler will select an ARIMA model if any of these independent variables have a statistically significant relationship with the dependent series.
- The acronym ARIMA refers to the **three components** of this modelling approach.
- It stands for Autoregressive (**AR**) Integrated (**I**) Moving Average (**MA**)
- This is the **structure** of an ARIMA model. But not all ARIMA models use **all** these elements. In fact, most don't.

Using predictor fields with ARIMA

- ARIMA (sometimes call "Box-Jenkins" models) is a powerful and functionally rich approach to forecasting.

- The three components within the ARIMA model are usually shortened to ***p,d,q***

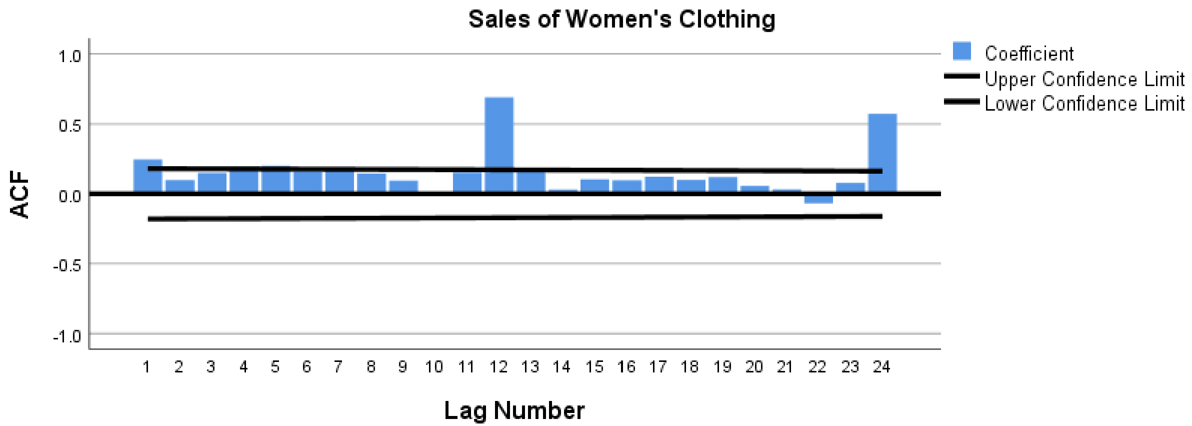
  - **p** refers to the **autoregression** component

  - **d** refers to the integration or ***differencing***

  - **q** refers to the **moving average** component

Using predictor fields with ARIMA

- **Autoregression (p)** refers to the correlation between a value in a series and the previous value(s). If I want to know the temperature today, is it useful to know the temperature yesterday? If so, then the autoregression ***p*** component is equal to **1** as this represents a lag of 1 day (a *first order* autorepression). If however the *day before yesterday* is a better component, then a value of **2** should be used to signify this (a *second order* autoregression).
- Autocorrelations are displayed in a ACF plot like the one below.



Using predictor fields with ARIMA

- Integration or ***differencing (d)***. Many analysts prefer the term ‘differencing’ rather than ‘integration’ for the **d** component of ARIMA.
- Differencing is a technique to used to make a time series **stationary**. A stationary series is one where the mean, the variance and autocorrelation values are constant over time. In other words the series has no trend and the peaks and troughs are the same proportion apart throughout.
- You may recall that Time Series analysis attempts to break a series up into different components by isolating elements like the trend component. ARIMA models require a series to be stationary in order to estimate correctly and the differencing component deals with this.
- Differencing **(d)** simply refers to the difference between one value in a sequence and another. If we were looking at daily temperatures, A first order differencing would result is a sequence of values where showing the difference between one day and the previous day. These values could of course be positive or negative.

Using predictor fields with ARIMA

- **Moving Average (q)**. We shouldn't confuse the term moving average here with the kind of moving average *smoothing* exercise that we undertook earlier. In fact the **q** component is another kind of autoregression. Except that this time it focusses on the *errors (*sometimes called 'shocks'*)* in the forecasting model.
- A first order **q** component (denoted as '1') in an ARIMA model indicates that the model's error in the immediate previous period is related to what the dependent variable will be now. For example, knowing that the model overestimated by 20% in the previous period helps to predict the current value.
- The **q** component is used to extract autocorrelated patterns (*systematic* errors) in the series that could otherwise be included in the model error.

Using predictor fields with ARIMA

- The **p,d,q** components in the ARIMA model could refer to values the *current season* **and/or** the values in *previous seasons*. So for seasonal models they are often shown as:

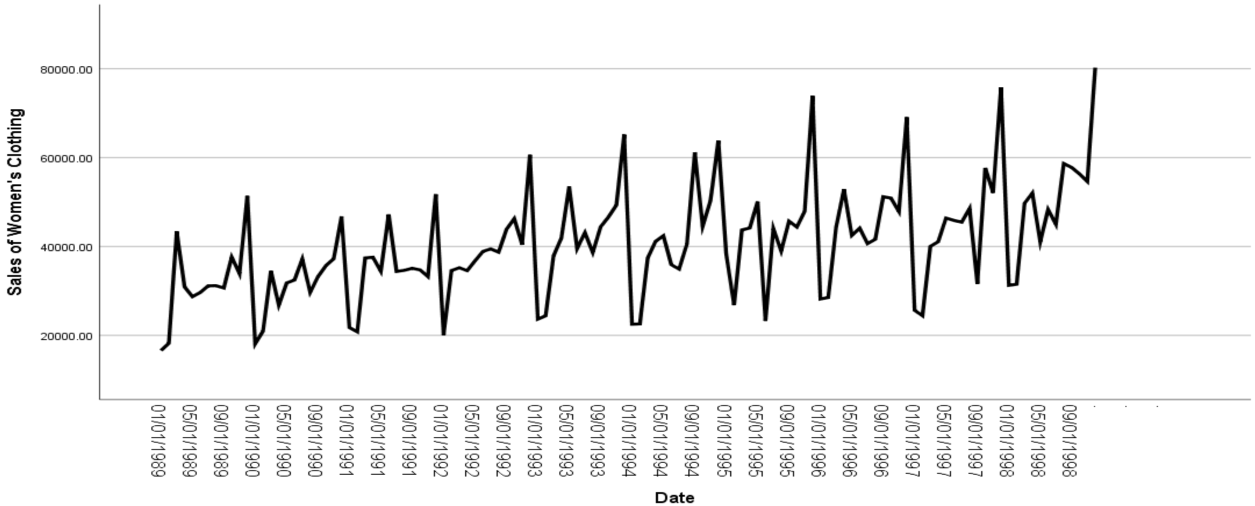
Non-Seasonal Seasonal

(p,d,q) (P,D,Q)

- An ARIMA model specified as **(1,0,0) (0,0,0)** would be one that only uses the correlation with the *most recent value* (first order autoregression) to estimate the current value.
- Whereas an ARIMA model specified as **(0,0,0) (1,0,0)** only uses the correlation with the value at the *same timepoint last season* to (first order seasonal autoregression) to estimate the current value.

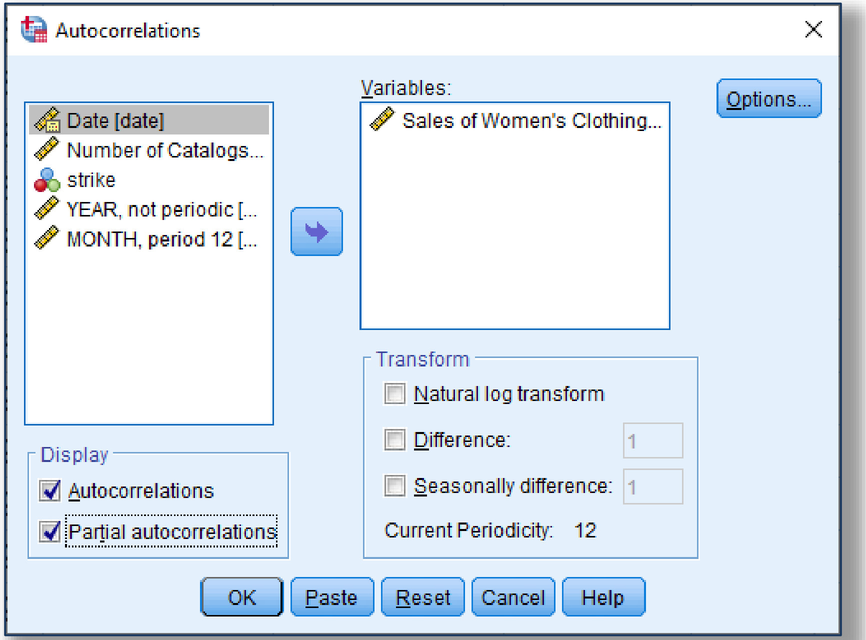
Demo 4: Using predictor fields with ARIMA

- Open the data file **clothing sales with independent variables.sav**
- This version of the file has the periodicity date variables already added
- It also has two predictor variables: **mail** and **strike**
- The variable **strike** has only two values: 1 and 0. This is called an **event** variable.
- Let’s remind ourselves what the series looks like:



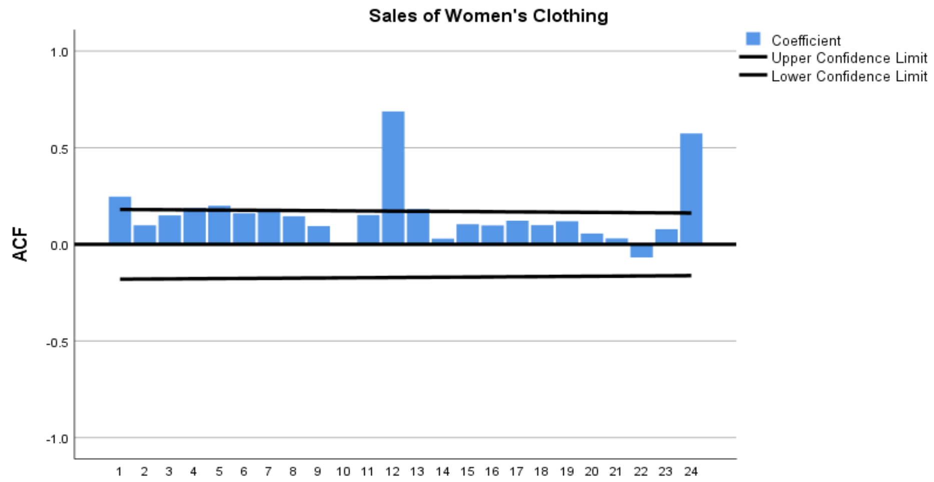
Demo 4: Using predictor fields with ARIMA

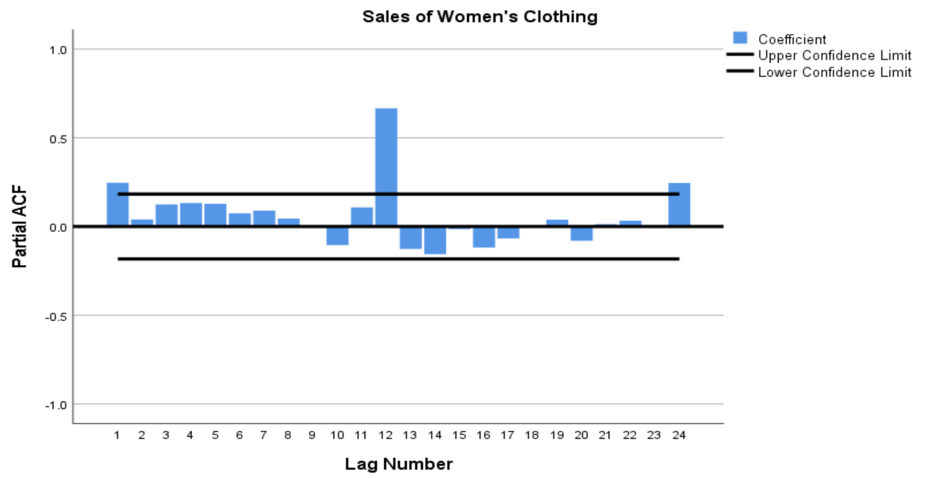
- Before building an ARIMA model, we can investigate the autocorrelation in the series. Click:
  - **Analyze**
    - **Forecasting**
      - **Autocorrelations**
- Request autocorrelations and partial autocorrelations for the Sales variable. Click:
  - **OK**



Demo 4: Using predictor fields with ARIMA

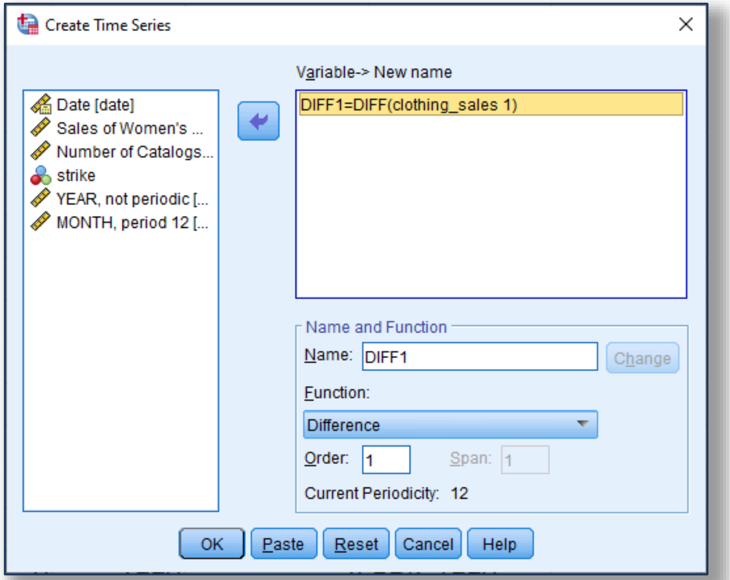
- Both the ACF and PACF plots show strong seasonal correlations at the 12 and 24 month lag points.
- A weaker correlation is also shown at the non-seasonal 1st order (i.e. lag 1)
- Based on this, the ARIMA model should probably have at least this structure: (0,0,0) (1,0,0)





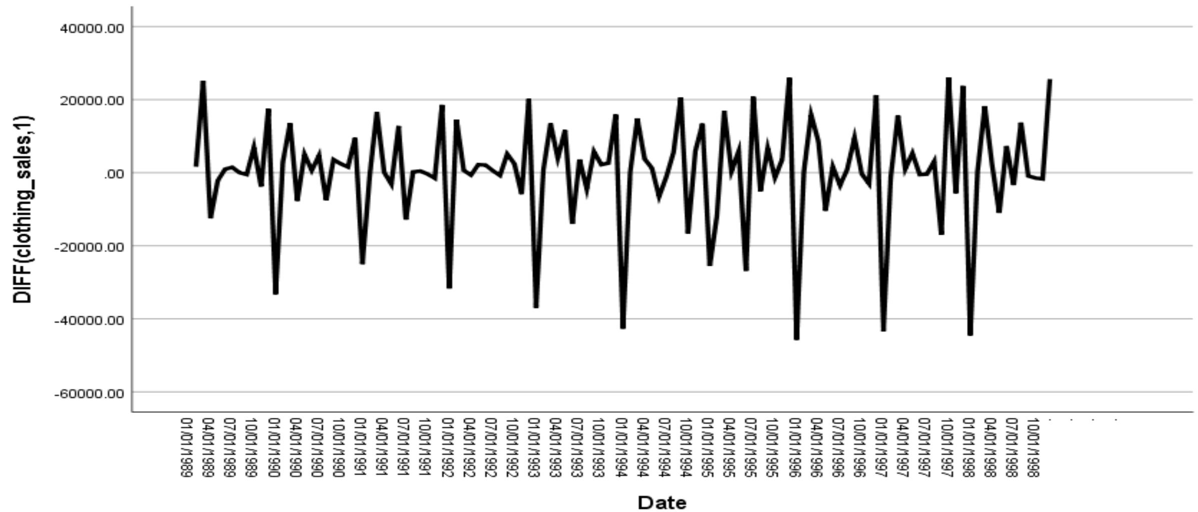
Demo 4: Using predictor fields with ARIMA

- However the earlier sequence chart of the sales variable looked like a trend was evident (and maybe slightly multiplicative). So the series mean and variance at least, are unlikely to be constant over the entire time window.
- In this case, differencing may be required. To illustrate the effect of differencing a series, click:
  - **Transform**
    - **Create Time Series**
- Using the Difference function create a field called **DIFF1** based on the Sales variable.
- The resulting variable will simply show the difference between the current value and the previous one.



Demo 4: Using predictor fields with ARIMA

- The sequence chart below shows the differenced series as represented by the new variable Diff1.
- You can see that the trend has been removed and the variation is more or less constant.
- Nevertheless the series is not random. We can see peaks and troughs at regular seasonal intervals.
- This is **d** component in the **p,d,q** structure of an ARIMA model.

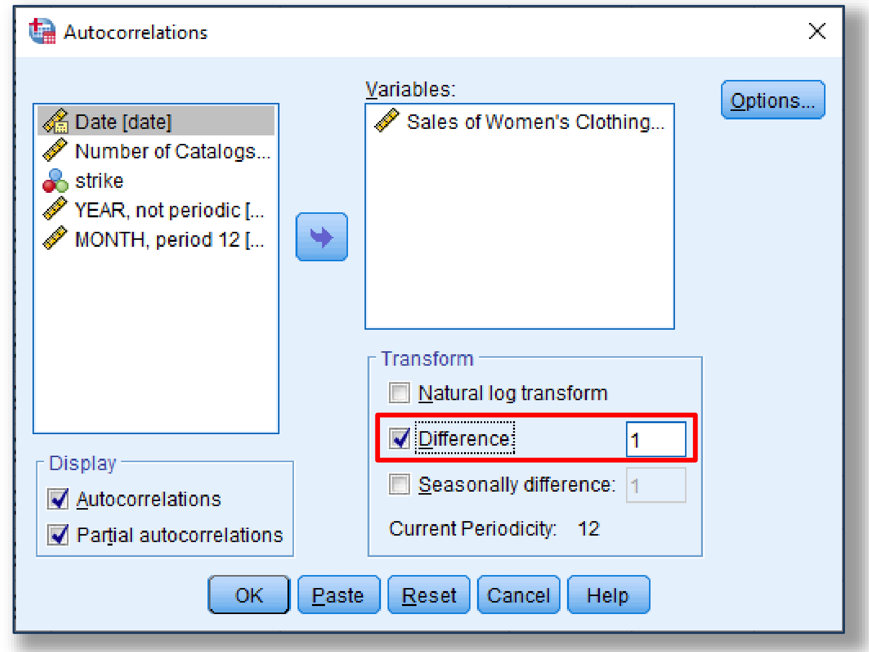


SMARTVISION Europe

Demo 4: Using predictor fields with ARIMA

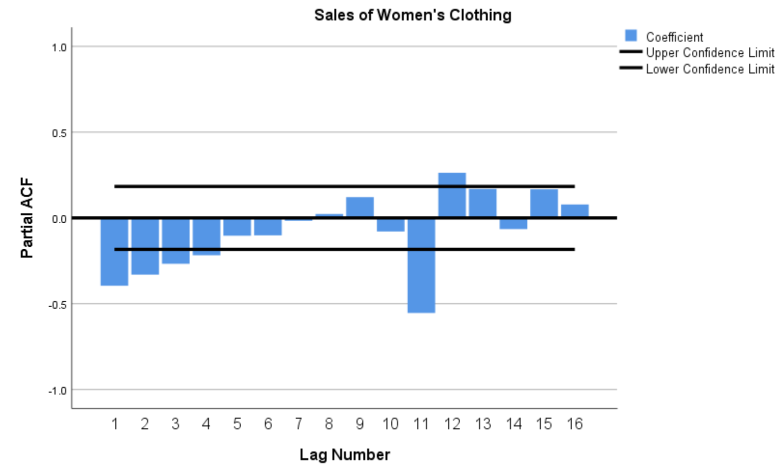
- In fact we can request Autocorrelations of the *differenced* (i.e. stationary) series.
- We don't even have to create a differenced variable to do this. Click the dialog recall button:

  - **Autocorrelations**
- Check the box marked **Difference**.
- Click:
  - **OK**



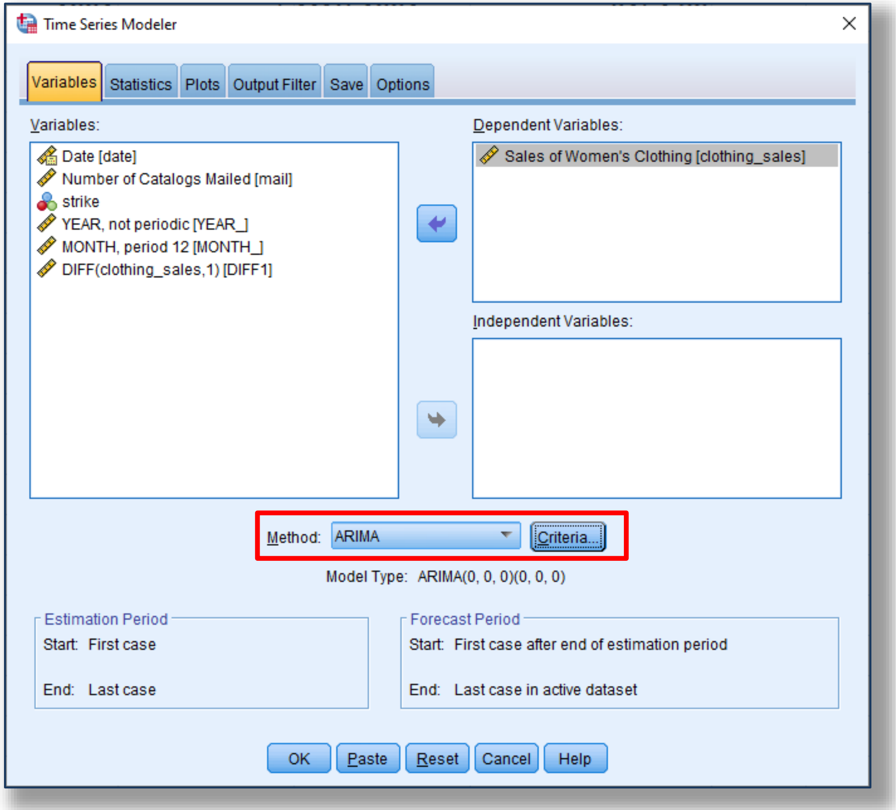
Demo 4: Using predictor fields with ARIMA

- The now stationary series shows strong seasonal correlations at the 12 month lag point. There are strong *negative* correlations at the 1st lag point.
- It should be noted that January tends to exhibit weaker sales revenues and so has negative values in the differenced series.



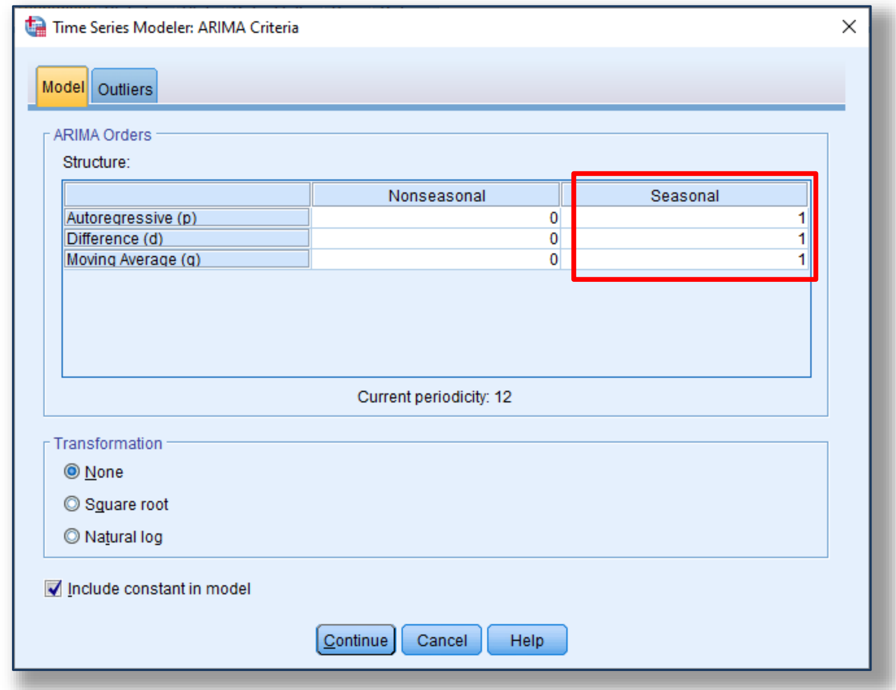
Demo 4: Using predictor fields with ARIMA

- As an experiment, let's return to the Time Series Modeler dialog an **manually** specify and ARIMA model. Click:
  - **Analyze**
    - **Forecasting**
      - **Create Traditional Models**
- Click the drop-down button marked **Method** and change it to **ARIMA**
- We can now override **Expert Modeler** and specify our own **ARIMA** model. To do so, click the button marked:
  - **Criteria**



Demo 4: Using predictor fields with ARIMA

- We know that the autocorrelation plots are showing strong signals of first order correlation at the seasonal level **(P).** We also know that the series has a trend and is not constant, so we should include differencing (**D**). It's possible that the model errors might autocorrelated with each other (**Q**).
- With this in mind let's specify the P,D,Q values in the seasonal column as (1,1,1).
- This means our ARIMA model is (0,0,0), (1,1,1)
- To see how the model performs click:
    - **Continue**
        - **OK**



Demo 4: Using predictor fields with ARIMA

Winter's Additive Exponential Smoothing

Fit Statistic	Mean
Stationary R-squared	.735
R-squared	.815
RMSE	5296.069
MAPE	9.412
MaxAPE	83.036
MAE	3624.699
MaxAE	19302.778
Normalized BIC	17.269

ARIMA (0,0,0) (1,1,1)

Fit Statistic	Mean
Stationary R-squared	.306
R-squared	.753
RMSE	6084.494
MAPE	11.374
MaxAPE	84.309
MAE	4409.612
MaxAE	21318.207
Normalized BIC	17.557

- The first thing we notice is that the Stationary R-Squared value is much smaller. This is the effect of adding the **D** (differencing component to the model). But apart from that there's not a **huge** difference between the fit values.

Demo 4: Using predictor fields with ARIMA

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	Ljung-Box Q(18): Statistics	Ljung-Box Q(18): DF	Ljung-Box Q(18): Sig.	Number of Outliers
Sales of Women's Clothing-Model_1	0	.306	14.109	16	.591	0

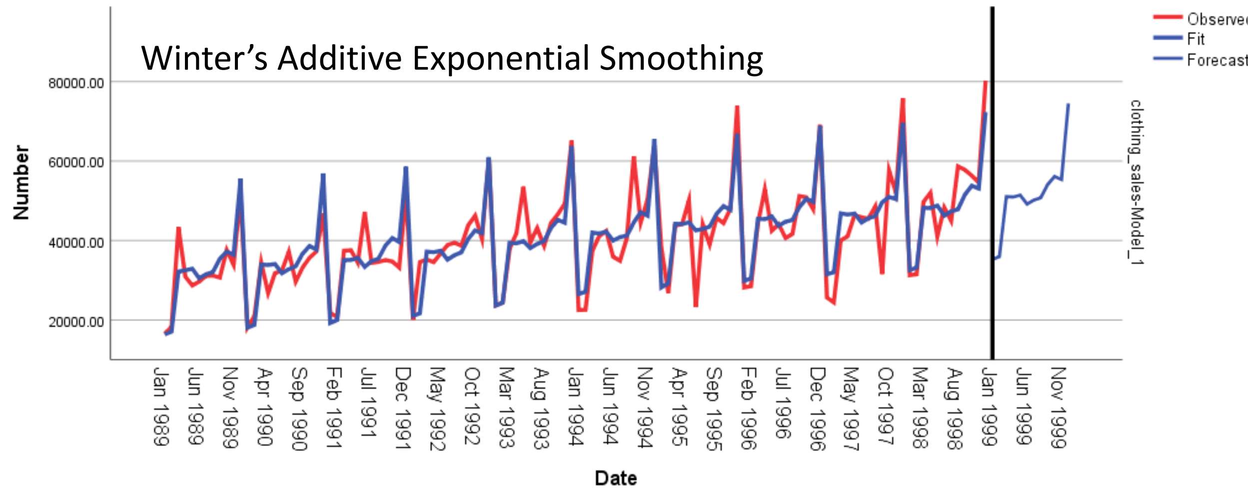
ARIMA Model Parameters

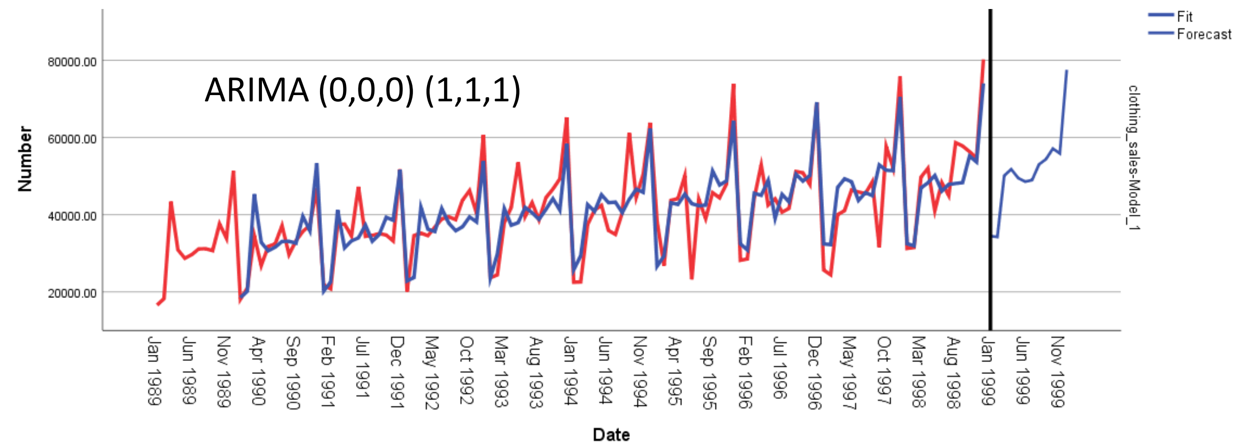
					Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	Sales of Women's Clothing	No Transformation	Constant		1936.846	225.145	8.603	.000
			AR, Seasonal	Lag 1	.084	.165	.510	.611
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	.784	.171	4.586	.000

- The Ljung-Box Q test indicates the model is correctly specified.
- The model paraments now include a constant value of $1,936 (think of it as a minimum sales)
- The coefficient for the Seasonal Autoregression term (P) is shown to be quite small (0.084)
- The coefficient for the Seasonal Moving Average term (Q) is much larger (0.784)

Demo 4: Using predictor fields with ARIMA

- The sequence charts for the Exponential Smoothing model vs our manually specified ARIMA model show that they look very similar.
- Note that the ARIMA model doesn't include a model fit line for the first 12 months of the data (as it would need the previous 12 months to do so)

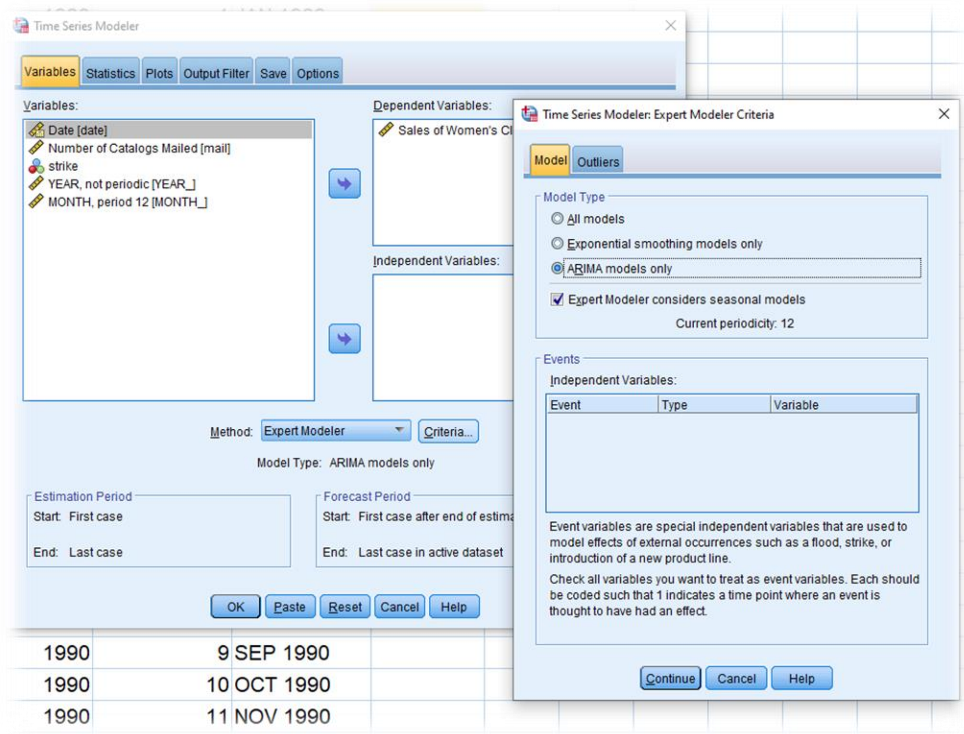




SMARTVISION Europe

Demo 4: Using predictor fields with ARIMA

- Now let's return to the **Time Series Modeler** dialog and use the **Expert Modeler** function to specify a model for us.
- Obviously if we just run this procedure as we did at the start, it will select a Winters Additive Exponential Smoothing model again.
- To force it to choose and ARIMA model, change the Method back to:
  - **Expert Modeler**
- In the sub-dialog choose:
  - **ARIMA models only**
- Click:
  - **Continue**
    - **OK**



Demo 4: Using predictor fields with ARIMA

Model Description

			Model Type
Model ID	Sales of Women's Clothing	Model_1	ARIMA(0,0,0)(0,1,1)

- The results show that Expert Modeler has dropped the seasonal autogestion term (P). You may recall that this has a very small value in the coefficients table.
- The model Ljung-Box fit test and the parameter table is shown below.

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Sales of Women's Clothing-Model_1	0	.304	15.143	17	.585	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	Sales of Women's Clothing	No Transformation	Constant		1936.590	236.015	8.205	.000
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	.706	.099	7.093	.000

Demo 4: Using predictor fields with ARIMA

Winter's Additive Exponential Smoothing

Fit Statistic	Mean
Stationary R-squared	.735
R-squared	.815
RMSE	5296.069
MAPE	9.412
MaxAPE	83.036
MAE	3624.699
MaxAE	19302.778
Normalized BIC	17.269

ARIMA (0,0,0) (1,1,1)

Fit Statistic	Mean
Stationary R-squared	.306
R-squared	.753
RMSE	6084.494
MAPE	11.374
MaxAPE	84.309
MAE	4409.612
MaxAE	21318.207
Normalized BIC	17.557

ARIMA (0,0,0) (0,1,1)

Fit Statistic	Mean
Stationary R-squared	.304
R-squared	.752
RMSE	6062.496
MAPE	11.364
MaxAPE	85.067
MAE	4400.095
MaxAE	21802.869
Normalized BIC	17.506

- Remember that the Expert Modeler attempts to find models with the smallest Normalised BIC value. That's why it chose Winter's Additive in our first example and the ARIMA (0,0,0) (0,1,1) when we forced it to only consider ARIMA models. Nevertheless the improvement in fit over our manually specified model is very small.

Demo 4: Using predictor fields with ARIMA

- The dataset also contains two predictor fields which can be specified as independent variables. In this context Exponential Smoothing algorithms don’t make use of predictor fields, *but ARIMA can*.

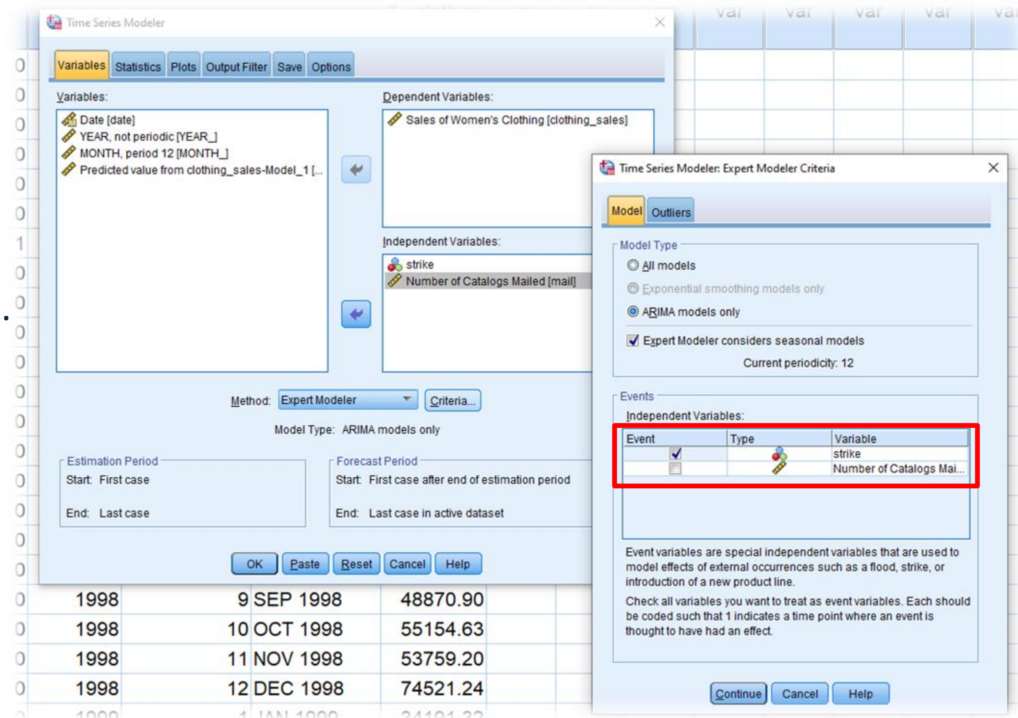
- The variable **mail** refers to the number of catalogues that were posted to customers. Note that if we wish to use this to forecast into the future we would need to specify the future values for each monthly mailing volume. You can see these anticipated future values at the bottom of the data file.

- The variable **strike** indicates whether or not a postal strike *event* had occurred that month.

mail	strike	YEA
11288	0	19
11096	0	19
11224	0	19
11483	0	19
11643	0	19
10893	0	19
11147	1	19
12260	0	19
11168	0	19
14370	0	19
11890	0	19
11722	0	19
11589	0	19
11633	0	19
11951	0	19
11706	0	19

Demo 4: Using predictor fields with ARIMA

- In the Time Series Modeler dialog we can specify both of these fields as independent variables (predictors).
- By clicking the **Criteria** button we can also indicate that **strike** is an **event** variable. These kinds of variables have only two values 1 and 0. the value 1 indicates that the event took place at that timepoint.
- To do so, simply check the adjacent box under the **Event column**.
- Event variables are useful for explaining unusual spikes or troughs in the series caused by events such as warehouse fires, unusual weather events or accidents.



Demo 4: Using predictor fields with ARIMA

Winter's Additive Exponential Smoothing

Fit Statistic	Mean
Stationary R-squared	.735
R-squared	.815
RMSE	5296.069
MAPE	9.412
MaxAPE	83.036
MAE	3624.699
MaxAE	19302.778
Normalized BIC	17.269

ARIMA (0,0,0) (0,1,1) without independent variables

Fit Statistic	Mean
Stationary R-squared	.304
R-squared	.752
RMSE	6062.496
MAPE	11.364
MaxAPE	85.067
MAE	4400.095
MaxAE	21802.869
Normalized BIC	17.506

ARIMA (0,0,0) (0,1,1) with independent variables

Fit Statistic	Mean
Stationary R-squared	.492
R-squared	.803
RMSE	5349.490
MAPE	9.810
MaxAPE	46.514
MAE	4045.744
MaxAE	16525.282
Normalized BIC	17.309

- The model fit seems to have improved slightly on the previous ARIMA model that did not include independent variables. Many of the statistics in the Exponential Smoothing model still seem to indicate a slightly better overall fit. But the biggest change is in the MaxAPE and MaxAE where the maximum absolute percentage error and maximum absolute error values are both quite smaller. This is almost certainly due to the inclusion of the **strike** variable.

Demo 4: Using predictor fields with ARIMA

Model Description

			Model Type
Model ID	Sales of Women's Clothing	Model_1	ARIMA(0,0,0)(0,1,1)

- Expert Modeler has not altered the structure of the ARIMA model.
- The parameter table now shows that on average a strike event costs $20,101 and that the company makes around $4.58 for each catalogue mailed.

Model Statistics

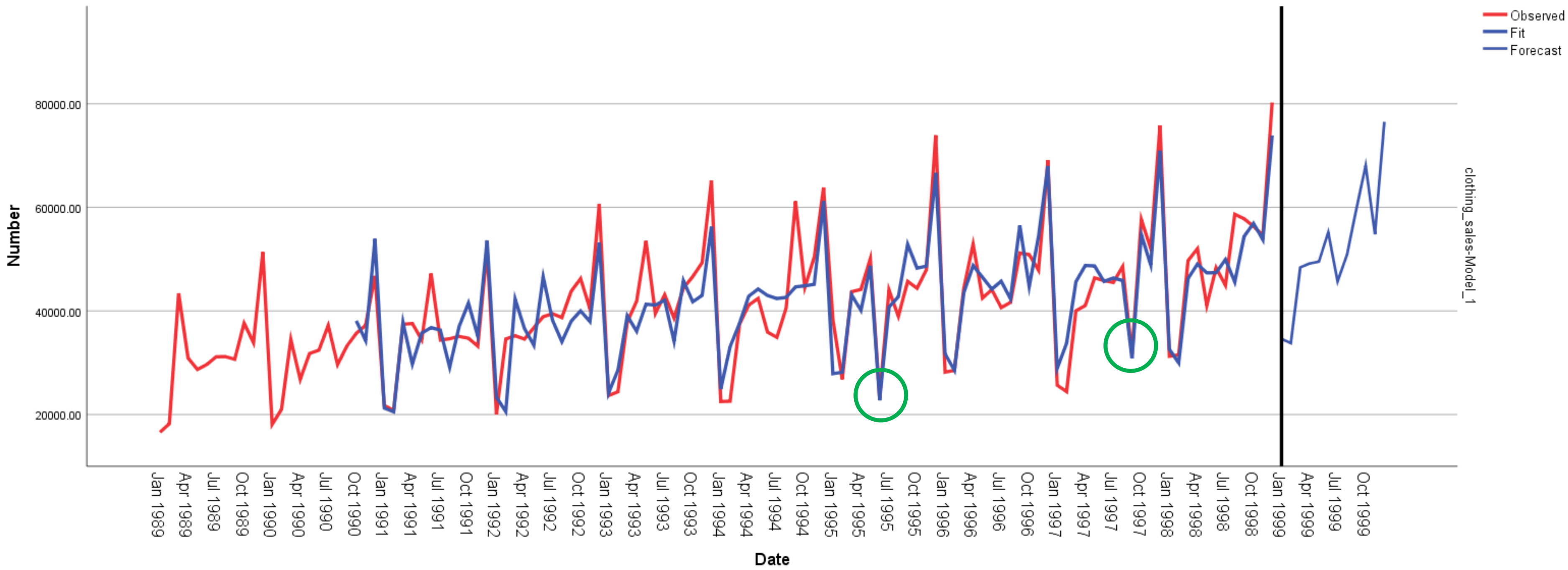
Model	Number of Predictors	Model Fit statistics: Stationary R-squared	Ljung-Box Q(18): Statistics	Ljung-Box Q(18): DF	Ljung-Box Q(18): Sig.	Number of Outliers
Sales of Women's Clothing-Model_1	2	.492	12.964	17	.739	0

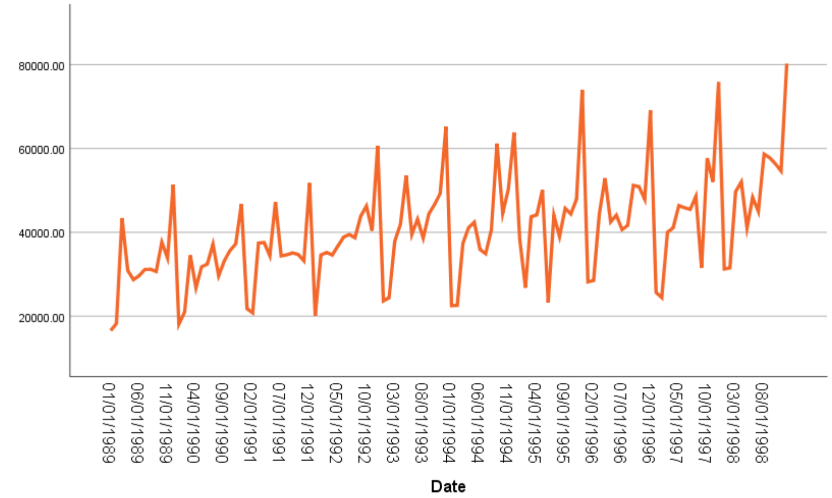
ARIMA Model Parameters

					Estimate	SE	t	Sig.
Sales of Women's Clothing-Model_1	Sales of Women's Clothing	No Transformation	Seasonal Difference		1			
			MA, Seasonal	Lag 1	.680	.118	5.773	.000
	strike	No Transformation	Numerator	Lag 0	-20101.492	3459.333	-5.811	.000
			Seasonal Difference		1			
	Number of Catalogs Mailed	No Transformation	Delay		9			
			Numerator	Lag 0	4.576	.528	8.670	.000
			Seasonal Difference		1			

Demo 4: Using predictor fields with ARIMA

- The immediate effect of adding an event variable can be seen in the sequence chart for the model. Postal strikes occurred in June 1996 and September 1997. As such, the event variable contains the value 1 at each of these time points. The model immediately picks this up so the fit line correctly reflects the downturn in revenue on these two occasions. Thus an event variable can be used to help explain anomalous values caused by external effects with the aim of providing more accurate estimates.





Generating Forecasts

Demo 5: Generating Forecasts

- One of the simplest ways to create new data *showing forecasted values* is to simply add the forecasts to the end of the data file itself.
- In our current example file, it's necessary to add values for the two independent variables **mail** and **strike** because they are part of the model and will therefore be needed to generate these new values.
- Of course, often analysts won't **know** the values of these independent variables in advance and may have to make educated guesses as to what they will be (or create a separate forecast).
- However, this same problem offers the opportunity for the analyst *to experiment with 'what if' analysis* by, for example, increasing or decreasing the anticipated marketing budget and noting it's effect on the of forecasts. Or estimating the impact of a particular event occurring again in the future.

*clothing sales with independent variables.sav [DataSet1] - IBM SPSS Statistics Data Editor

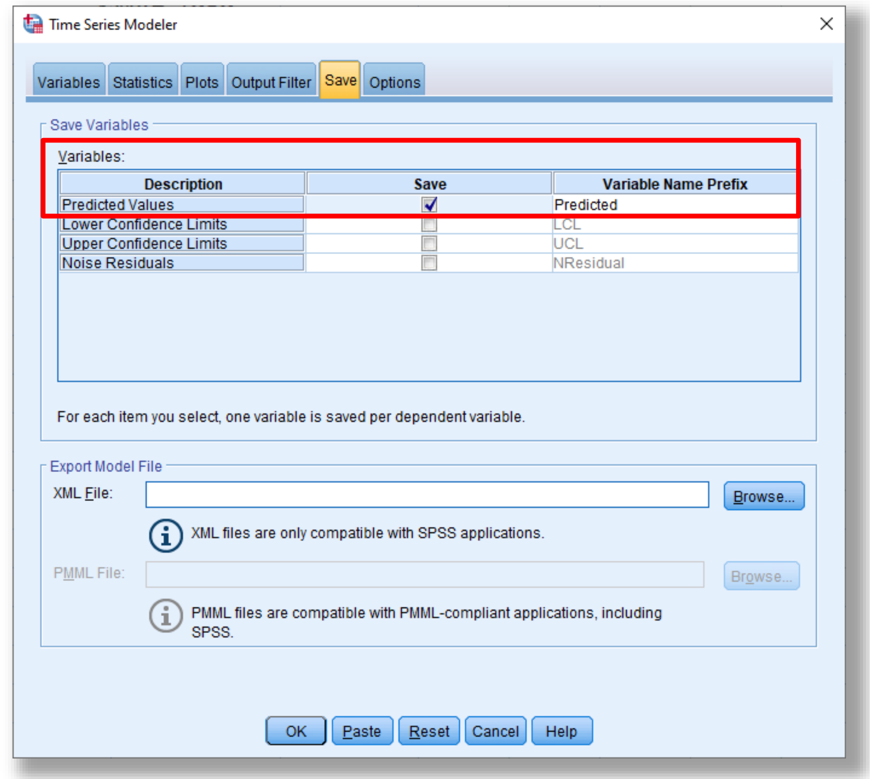
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

132 : strike 0

	date	clothing_sales	mail	strike	YEAR_	MONTH_	DATE_
113	05/01/1998	41038.75	11951	0	1998	5	MAY 1998
114	06/01/1998	48329.36	11706	0	1998	6	JUN 1998
115	07/01/1998	44961.98	11460	0	1998	7	JUL 1998
116	08/01/1998	58660.76	11808	0	1998	8	AUG 1998
117	09/01/1998	57791.14	12781	0	1998	9	SEP 1998
118	10/01/1998	56329.40	11690	0	1998	10	OCT 1998
119	11/01/1998	54617.35	11393	0	1998	11	NOV 1998
120	12/01/1998	80245.97	15263	0	1998	12	DEC 1998
121	.	.	14370	0	1999	1	JAN 1999
122	.	.	11890	0	1999	2	FEB 1999
123	.	.	11722	0	1999	3	MAR 1999
124	.	.	11589	0	1999	4	APR 1999
125	.	.	11633	0	1999	5	MAY 1999
126	.	.	11951	0	1999	6	JUN 1999
127	.	.	11706	0	1999	7	JUL 1999
128	.	.	11460	0	1999	8	AUG 1999
129	.	.	11808	0	1999	9	SEP 1999
130	.	.	12781	0	1999	10	OCT 1999
131	.	.	11690	0	1999	11	NOV 1999
132	.	.	11393	0	1999	12	DEC 1999

Demo 5: Generating Forecasts

- By simply returning to the Time Series dialog and clicking the tab marked:
  - **Save**
- Then checking the Predicted Values box also marked:
  - **Save**
- Before finally re-running the procedure by clicking:
  - **OK**
- This will create the forecasted values in the dataset.



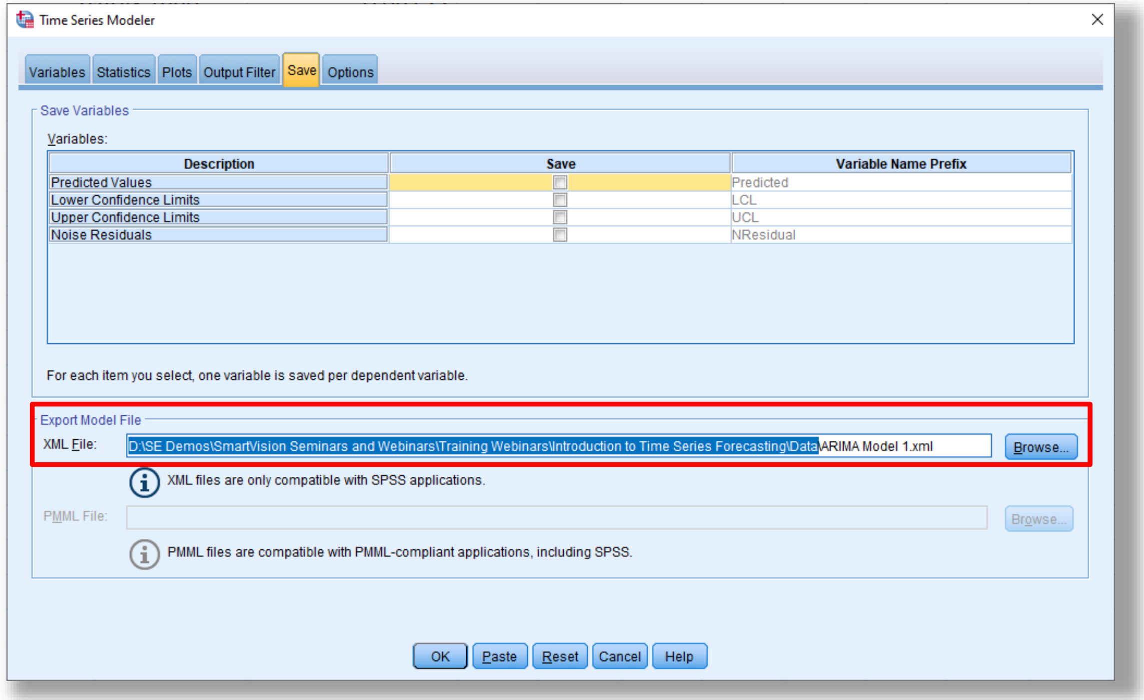
Demo 5: Generating Forecasts

- The forecasted clothing sales for the year 1999 are now shown in the dataset.

date	clothing_sales	mail	strike	YEAR_	MONTH_	DATE_	Predicted_clothing_sales_Model_1
05/01/1998	41038.75	11951	0	1998	5	MAY 1998	47359.96
06/01/1998	48329.36	11706	0	1998	6	JUN 1998	47391.22
07/01/1998	44961.98	11460	0	1998	7	JUL 1998	49934.48
08/01/1998	58660.76	11808	0	1998	8	AUG 1998	45590.14
09/01/1998	57791.14	12781	0	1998	9	SEP 1998	54405.31
10/01/1998	56329.40	11690	0	1998	10	OCT 1998	56927.51
11/01/1998	54617.35	11393	0	1998	11	NOV 1998	53767.85
12/01/1998	80245.97	15263	0	1998	12	DEC 1998	73884.71
.	.	14370	0	1999	1	JAN 1999	34621.09
.	.	11890	0	1999	2	FEB 1999	33784.23
.	.	11722	0	1999	3	MAR 1999	48400.32
.	.	11589	0	1999	4	APR 1999	49167.51
.	.	11633	0	1999	5	MAY 1999	49522.23
.	.	11951	0	1999	6	JUN 1999	55169.71
.	.	11706	0	1999	7	JUL 1999	45732.82
.	.	11460	0	1999	8	AUG 1999	50807.40
.	.	11808	0	1999	9	SEP 1999	59576.83
.	.	12781	0	1999	10	OCT 1999	68085.74
.	.	11690	0	1999	11	NOV 1999	54808.51
.	.	11393	0	1999	12	DEC 1999	76528.76

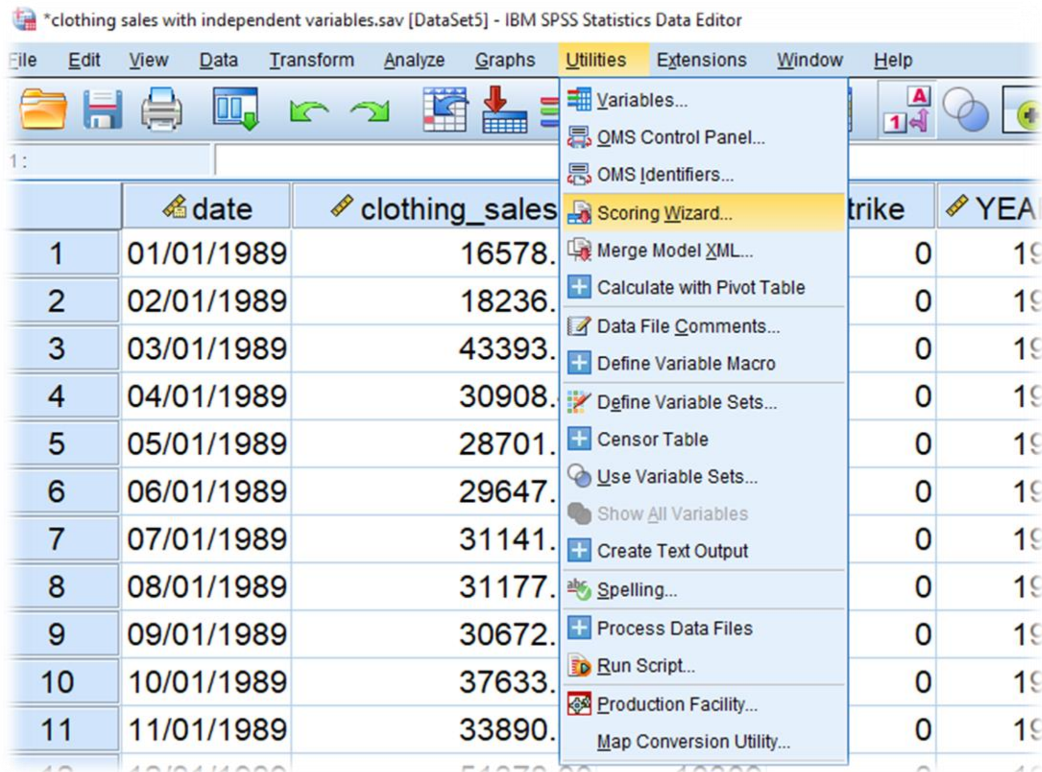
Demo 5: Generating Forecasts

- An alternative approach is to save the model as a file.
- Again, simply return to the Time Series dialog and click the tab marked:
    - **Save**
- Specify a file name for the model in Export File Section box marked:
    - **XML File**
- Again re-run the procedure by clicking:
    - **OK**
- This will write out the model as an xml file for re-use in the future.



Demo 5: Generating Forecasts

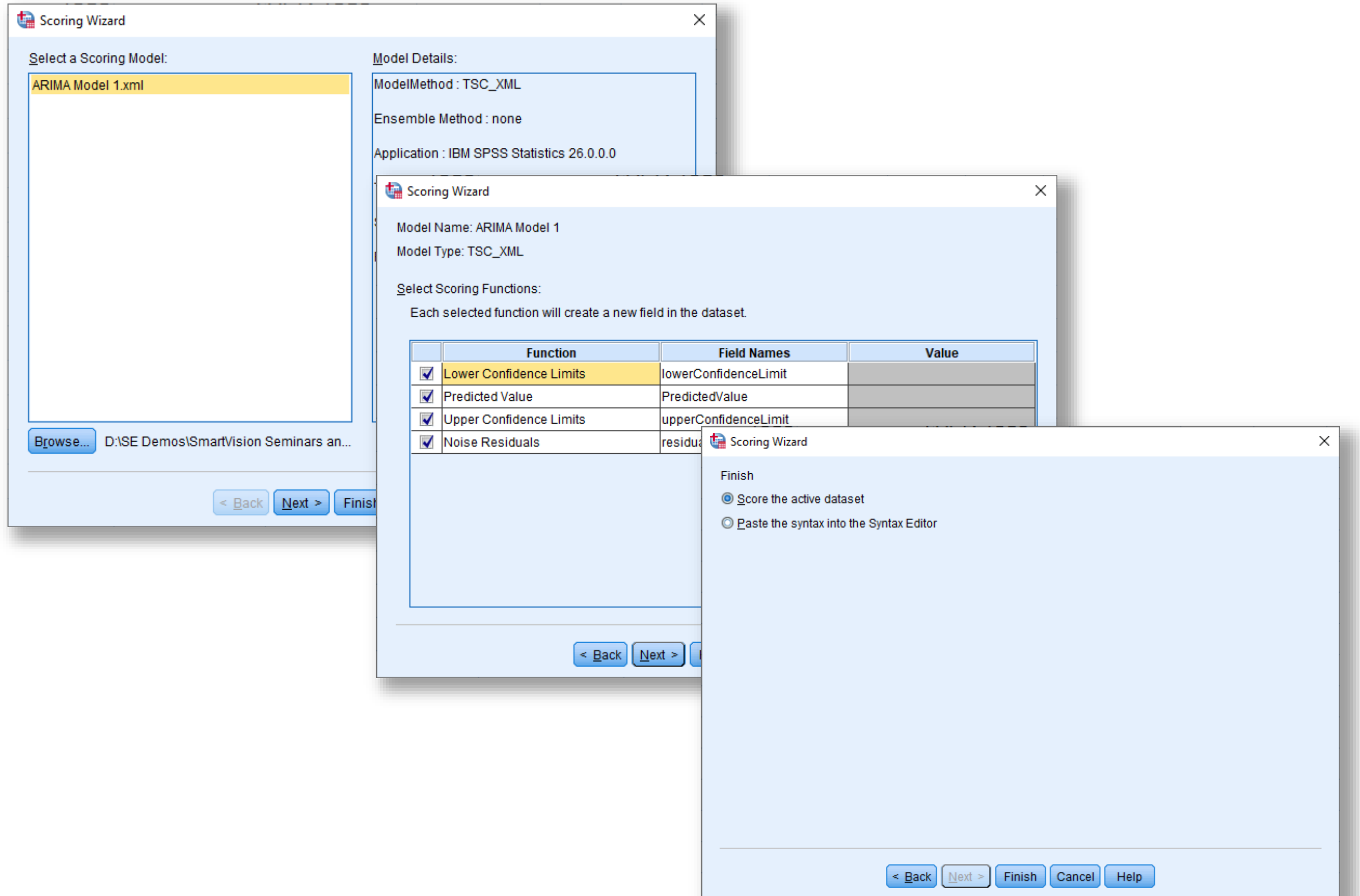
- We can apply the xml model file to a new dataset by returning to the Data Editor window and clicking:
  - **Utilities**
    - **Scoring Wizard**



SMARTVISION Europe

Demo 5: Generating Forecasts

- Browsing for the xml file in the Scoring Wizard dialog allows us to step through the wizard by clicking
  - **Next**
- Before finally clicking:
  - **Finish**



Demo 5: Generating Forecasts

- The XML model is applied to the dataset and the forecasted clothing sales for the year 1999 (including confidence intervals) are now shown in the data editor window

*clothing sales with independent variables.sav [DataSet5] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

1 : PredictedValue_clot...

	date	clothing_sales	mail	strike	YEAR_	MONTH_	DATE_	Predicted Value_clothing_s..	lowerConfidenceLimit_clothi	upperConfidenceLimit_clothi
113	05/01/1998	41038.75	11951	0	1998	5	MAY 1998	.	.	.
114	06/01/1998	48329.36	11706	0	1998	6	JUN 1998	.	.	.
115	07/01/1998	44961.98	11460	0	1998	7	JUL 1998	.	.	.
116	08/01/1998	58660.76	11808	0	1998	8	AUG 1998	.	.	.
117	09/01/1998	57791.14	12781	0	1998	9	SEP 1998	.	.	.
118	10/01/1998	56329.40	11690	0	1998	10	OCT 1998	.	.	.
119	11/01/1998	54617.35	11393	0	1998	11	NOV 1998	.	.	.
120	12/01/1998	80245.97	15263	0	1998	12	DEC 1998	.	.	.
121	.	.	14370	0	1999	1	JAN 1999	34621.09	24359.82	44882.37
122	.	.	11890	0	1999	2	FEB 1999	33784.23	23522.96	44045.51
123	.	.	11722	0	1999	3	MAR 1999	48400.32	38139.04	58661.59
124	.	.	11589	0	1999	4	APR 1999	49167.51	38906.23	59428.78
125	.	.	11633	0	1999	5	MAY 1999	49522.23	39260.95	59783.50
126	.	.	11951	0	1999	6	JUN 1999	55169.71	44908.43	65430.98
127	.	.	11706	0	1999	7	JUL 1999	45732.82	35471.54	55994.09
128	.	.	11460	0	1999	8	AUG 1999	50807.40	40546.12	61068.67
129	.	.	11808	0	1999	9	SEP 1999	59576.83	49315.56	69838.11
130	.	.	12781	0	1999	10	OCT 1999	68085.74	57825.91	78345.57
131	.	.	11690	0	1999	11	NOV 1999	54808.51	44548.68	65068.34
132	.	.	11393	0	1999	12	DEC 1999	76528.76	66268.93	86788.59

Working with Smart Vision Europe Ltd.

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - http://www.sv-europe.com/buy-spss-online/
- **Training and Consulting Services**
  - Guided consulting & training to develop in house skills
  - Delivery of classroom training courses / side by side training support
  - Identification & recruitment of analytical skills into your organisation
- **Advice and Support**
  - offer ‘no strings attached’ technical and business advice relating to analytical activities
  - Technical support services

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In
Sign up for our Newsletter

Thank you