Time Series – Smoothing

Summary

The **Smoothing** procedure is designed to help illustrate any trends and cycles present in a time series. A *time series* consists of a set of sequential numeric data taken at equally spaced intervals, usually over a period of time or space. The procedure applies one or more selected smoothers to the data and plots the results. The smoothers available include both classic moving averages and nonlinear smoothers that are resistant to outliers.

Sample StatFolio: *tssmooth.sgp*

Sample Data:

The file *baseball.sgd* contains the leading batting average in U. S. Major League Baseball for each year between 1901 and 2004. Batting averages represent the proportion of times that a player gets a hit out of all at-bats that result in either a hit or an out. The table below shows a partial list of the data from that file. The batting averages are expressed as the number of points out of 1000, such that a player batting 333 would have gotten a hit one-third of the time.

Year	Leading average
1901	422
1902	376
1903	355
1904	381
1905	377
1906	358
1907	350
1908	354
1909	377
1910	385
…	…
2004	372

Data Input

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

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

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

- **Sampling Interval:** If time indices are not provided, this defines the interval between successive observations.

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

- **Trading Days Adjustment**: a numeric variable with n observations used to normalize the original observations, such as the number of working days in a month.

The observations in the *Data* column will be divided by these values before being plotted or analyzed.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* displays the number of observations in the time series, the length of seasonality (if any), any adjustments made to the data, and the selected smoothers.

Smoothing - Leading average
Data variable: Leading average

Number of observations = 104
Start index = 1901
Sampling interval = 1.0 year(s)
First smoother: 3RSSH resistant smoother
Second smoother: simple moving average of 5 terms

Note: a limited amount of missing data is permitted, providing there are not too many missing values close together. Missing values are replaced by interpolated values according to the method outlined in the *Calculations* section of the *Time Series – Descriptive Methods* documentation.

Analysis Options

Analysis Options selects the smoothers to be applied to the data:

- **Smoother 1**: first smoother to apply to the data.

- **Smoother 2**: smoother to apply to the smoothed values resulting from the first smoother, if any.

- **Length of Moving Average**: the span k of the moving average when using the *Simple Moving Average* or *Henderson's Moving Average*. k must be greater than or equal to 2 for the simple MA and greater than or equal to 3 and odd for Henderson's MA. The higher the value of k, the smoother the result.

- **EWMA Smoothing Constant**: the smoothing parameter λ when using the EWMA. The smoothing constant must satisfy $0 < \lambda < 1$. The smaller the value of λ, the smoother the result.

The available smoothers are:

1. *Simple Moving Average*: The smoothed value at time t is the average of the k data values centered at time t. If k is odd, then the smoothed value is

$$S_t = \frac{\sum_{j=-(k-1)/2}^{(k-1)/2} y_{t+j}}{k} \qquad (1)$$

If k is even, a centered moving average is used, defined by

$$S_t = \frac{0.5 y_{t-k/2} + \sum_{j=-k/2+1}^{k/2-1} y_{t+j} + 0.5 y_{t+k/2}}{k} \qquad (2)$$

2. *Spencer's 15-term and 21-term moving averages*: weighted moving averages developed for actuarial work. The smoothed value at time t is given by

$$S_t = \frac{\sum_{j=-m}^{m} w_j y_{t+j}}{\sum_{j=-m}^{m} w_j y_{t+j}} \qquad (3)$$

where $m = 7$ for the 15-term moving average and $m = 10$ for the 21-term moving average. The weights are shown below:

j	w_j (15-term MA)	w_j (15-term MA)
-10		-5/1750
-9		-15/1750
-8		-25/1750
-7	-3/320	-25/1750
-6	-6/320	-10/1750
-5	-5/320	30/1750
-4	3/320	90/1750
-3	21/320	165/1750
-2	46/320	235/1750
-1	67/320	285/1750
0	74/320	300/1750
1	67/320	285/1750
2	46/320	235/1750
3	21/320	165/1750
4	3/320	90/1750
5	-5/320	30/1750
6	-6/320	-10/1750
7	-3/320	-25/1750
8		-25/1750
9		-15/1750
10		-5/1750

3. *Henderson's Weighted Moving Average* – a weighted moving average, often used to smooth census data. The span k must be odd and at least 3, but otherwise it may be specified by the user. Common values for k are 5, 9, 13, and 23. The weights are given by

$$w_j = \frac{315\left[(m-1)^2 - j^2\right]\left[m^2 - j^2\right]\left[(m+1)^2 - j^2\right]\left[(3m^2 - 16) - 11j^2\right]}{8m(m^2 - 1)(4m^2 - 1)(4m^2 - 9)(4m^2 - 25)} \quad (4)$$

where $m = (k+3) / 2$.

4. *EWMA* – exponentially weighted moving average. This smoother includes all observations from up to and including y_t, with weights that decay exponentially with the distance from t. Given a smoothing constant λ, the smoothed value if given by

$$S_t = \frac{\sum_{j=0}^{t-1} \lambda(1-\lambda)^j y_{t-j}}{\sum_{j=0}^{t-1} \lambda(1-\lambda)^j} \quad (5)$$

5. 3RSS, 3RSSH, 5RSS, 5RSSH, and 3RSR: nonlinear smoothers developed by John Tukey (1977). These smoothers are very good at ignoring outliers, and are often applied as a first step to reduce the influence of potential outliers before a moving average is applied. Each symbol in the name of the smoother indicates an operation that is applied to the data:

"3": A moving window of size 3 is passed through the data, and the median of the 3 values in the window is kept as the smooth.

"5": A moving window of size 5 is passed through the data, and the median of the 5 values in the window is kept as the smooth.

"R": The smoothed data is *resmoothed* by applying the previous operation to the smoothed result. This is repeated until no further change occurs.

"S": Since running medians tend to leave flat mesas and dales, these flat spots are *split* to make the result smoother than it would otherwise be. The main smoothing operation is then performed again.

"H": The data is smoothed using a weighted average of the form

$$S_t = 0.25 y_{t-1} + 0.5 y_t + 0.25 y_{t+1} \quad (6)$$

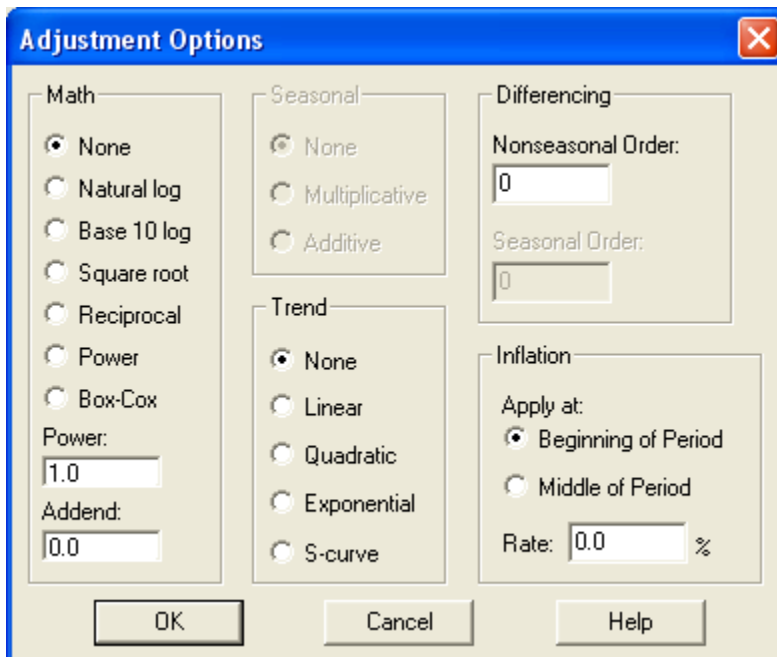which is referred to as *hanning* after its proponent Julius von Hann.

For a detailed discussion of nonlinear smoothing, see Velleman and Hoaglin (1981).

The baseball data was smoothed using two smoothers in sequence: 3RSSH, followed by a simple moving average of 5 terms. This means that:

(i)   Running medians of 3 were taken and repeated until there was no further change.
(ii)   Mesas and dales were split.
(iii)   Running medians of 3 were taken and repeated until there was no further change.
(iv)   Mesas and dales were split.
(v)   Running medians of 3 were taken and repeated until there was no further change.
(vi)   Hanning was performed.
(vii)   A simple moving average of 5 terms was taken.

The result is quite satisfactory.

- **Adjustments**: this button displays a second dialog box that defines adjustments to be applied to the data before it is smoothed:

o   **Math**: transforms the data by performing the indicated mathematical operation.

o   **Seasonal**: seasonally adjusts the data using the indicated method.

o   **Trend**: removes a trend by fitting and subtracting the indicated type of trend.

o   **Differencing**: transforms the data by taking nonseasonal differences of order d and/or seasonal differences of order D.

o   **Inflation**: adjusts the data for inflation using the specified inflation rate

If more than one transformation is requested, they are applied in the following order:
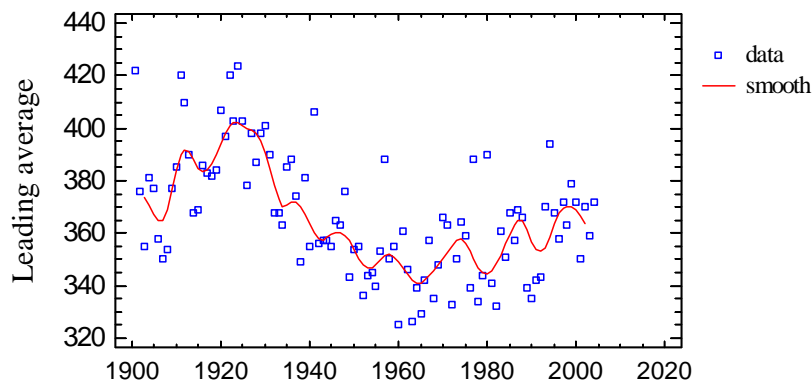
1.   trading day adjustment
2.   inflation adjustment
3.   math adjustment
4.   seasonal adjustment
5.   trend adjustment
6.   differencing

For a detailed discussion of the transformation options, refer to the *Time Series – Descriptive Methods* documentation.

Time Sequence Plot

The *Time Sequence Plot* displays the time series data in sequential order, together with the result of applying one or more smoothers to it:

Smoothed Time Series Plot for Leading average

The baseball data show some interesting features:

1.   A general upward trend through about 1923, then a decreasing trend through 1966, and then another upward trend thereafter.
2.   Oscillations around the trend line with a period that varies between about 10 years and 15 years.

Data Table

The data table displays the original observations and the smoothed result:

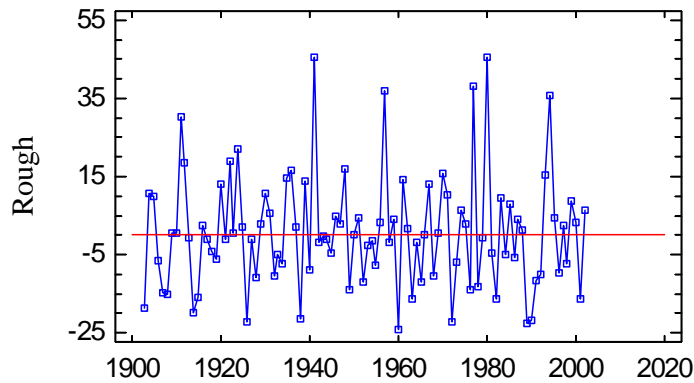| Period	Data	Smooth	Rough
Data Table for Leading average First smoother: 3RSSH resistant smoother Second smoother: simple moving average of 5 terms			
1901	422.0		
1902	376.0		
1903	355.0	373.6	-18.6
1904	381.0	370.4	10.6
1905	377.0	366.95	10.05
1906	358.0	364.6	-6.6
1907	350.0	364.9	-14.9
1908	354.0	369.0	-15.0
1909	377.0	376.4	0.6
1910	385.0	384.3	0.7
…	…	…	…

The table includes:

- **Data**: the original observations y_t.

- **Smooth**: the final smoothed values S_t.

- **Rough**: the residuals that were left behind, defined as $R_t = y_t - S_t$.

Residual Plot

The *Residual Plot* displays the rough:

Time Series Plot of Rough for Leading average

It can be used to identify outliers or other unusual observations, such as the outstanding achievements in 1941 by Ted Williams when he hit .406 and in 1980 by George Brett when he hit .390.

Save Results

The following results may be saved to the datasheet:

1. *Data* – the original data, including any interpolated values used to replace observations that were missing.

2. *Adjusted data* – the data after applying any requested adjustments.

3. *Smooth* – the smoothed values S_t.

4. *Rough* – the residual values R_t.