Three-Dimensional Visualizer

Summary

The Three-Dimensional Visualizer Statlet is designed to plot multiple time series in a manner that helps users visualize the changes in multiple variables over time. Given *n* time series observed over *p* time periods, the program generates a dynamic display that illustrates how each of the variables has changed over time. Typical applications include plotting:

1. Yearly demographic variables for different countries.
2. Quarterly sales figures for multiple divisions within a company.
3. Monthly economic indices on a state-by-state basis.
4. Daily closing stock prices for multiple equities within a portfolio.

The basic plot shows bubbles plotted on an X-Y-Z display. The positions along the X, Y and Z axes represent the values of three primary data variables. The size and color of the bubbles may be used to illustrate other variables. As time increases, the analyst can follow changes in all of the variables simultaneously. Various options are offered for smoothing the data and for dealing with missing values.

Sample StatFolio: *visualize3d.sgp*

Sample Data

The file *worldbank.sgd* contains data for *n* = 188 countries over *p* = 50 years (1961-2010). It was obtained from the World Bank (worldbank.org). Variables in the file include:

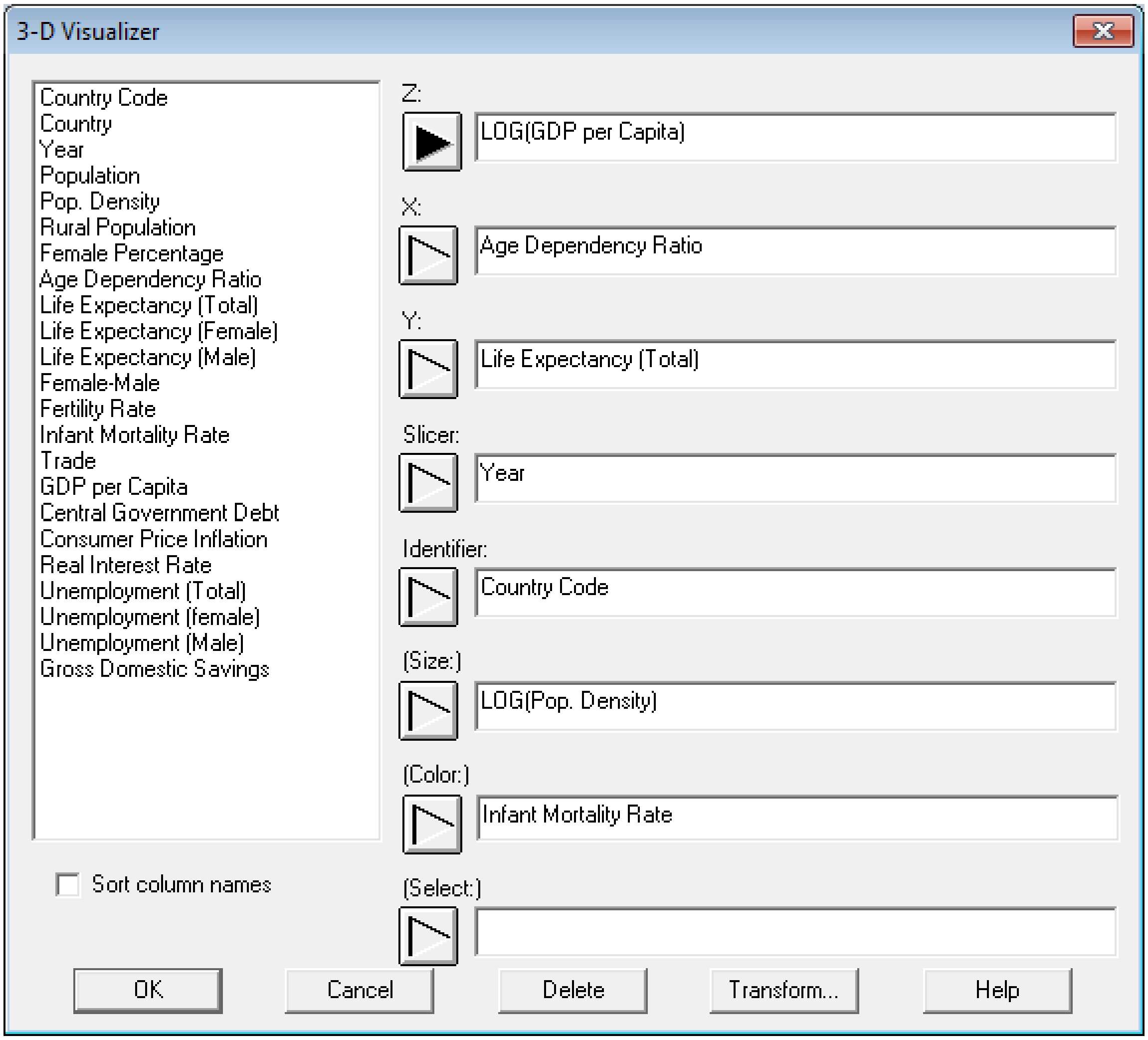
- Population, population density, female population, rural population
- Age dependency ratio
- Life expectancy, infant mortality, fertility rate
- Central government debt
- Consumer price inflation
- Interest rates
- Unemployment rates
- GDP
- Savings rates
- Trade rates

The first several rows and columns of the file are shown below:

Country Code	Country	Year	Population	Population Density	Rural Population	Female Percentage	Age Dependency Ratio	Life Expectancy (Total)
			total	people per sq. km of land area	% of total population	% of total population	% of working-age population	years
ABW	Aruba	1961	55436	307.98	49.22	50.97	85.56	65.99
ABW	Aruba	1962	56227	312.37	49.24	50.97	84.44	66.37
ABW	Aruba	1963	56698	314.99	49.26	50.99	83.02	66.71
ABW	Aruba	1964	57031	316.84	49.28	51.01	81.42	67.04
ABW	Aruba	1965	57362	318.68	49.3	51.03	79.76	67.37
ABW	Aruba	1966	57714	320.63	49.32	51.04	78.03	67.7
ABW	Aruba	1967	58052	322.51	49.34	51.04	76.26	68.03
ABW	Aruba	1968	58388	324.38	49.36	51.05	74.4	68.38
ABW	Aruba	1969	58725	326.25	49.38	51.07	72.47	68.73
ABW	Aruba	1970	59066	328.14	49.4	51.1	70.44	69.09

Data Input

The data input dialog box requests the names of the columns containing the data values to be analyzed:



Country Code
Country
Year
Population
Pop. Density
Rural Population
Female Percentage
Age Dependency Ratio
Life Expectancy (Total)
Life Expectancy (Female)
Life Expectancy (Male)
Female-Male
Fertility Rate
Infant Mortality Rate
Trade
GDP per Capita
Central Government Debt
Consumer Price Inflation
Real Interest Rate
Unemployment (Total)
Unemployment (female)
Unemployment (Male)
Gross Domestic Savings

Z: LOG(GDP per Capita)

X: Age Dependency Ratio

Y: Life Expectancy (Total)

Slicer: Year

Identifier: Country Code

(Size:) LOG(Pop. Density)

(Color:) Infant Mortality Rate

(Select:)

- **Z:** name of the numeric column containing the observations to be plotted on the vertical axis. There should be a total of *n* times *p* observations.

- **X:** name of the numeric column containing the observations to be plotted on the horizontal axis. There should be a total of *n* times *p* observations.

- **Y:** name of the numeric column containing the observations to be plotted on the axis that extends into the screen. There should be a total of *n* times *p* observations.

- **Slicer:** name of the numeric column used to define subsets of the data. This variable, often a measure of time, is changed dynamically to illustrate changes in the other variables. There should be *p* unique values of this variable.

- **Identifier:** name of the numeric or non-numeric column used to define each sample. There should be *n* unique values of this variable. A separate bubble is created for each value.

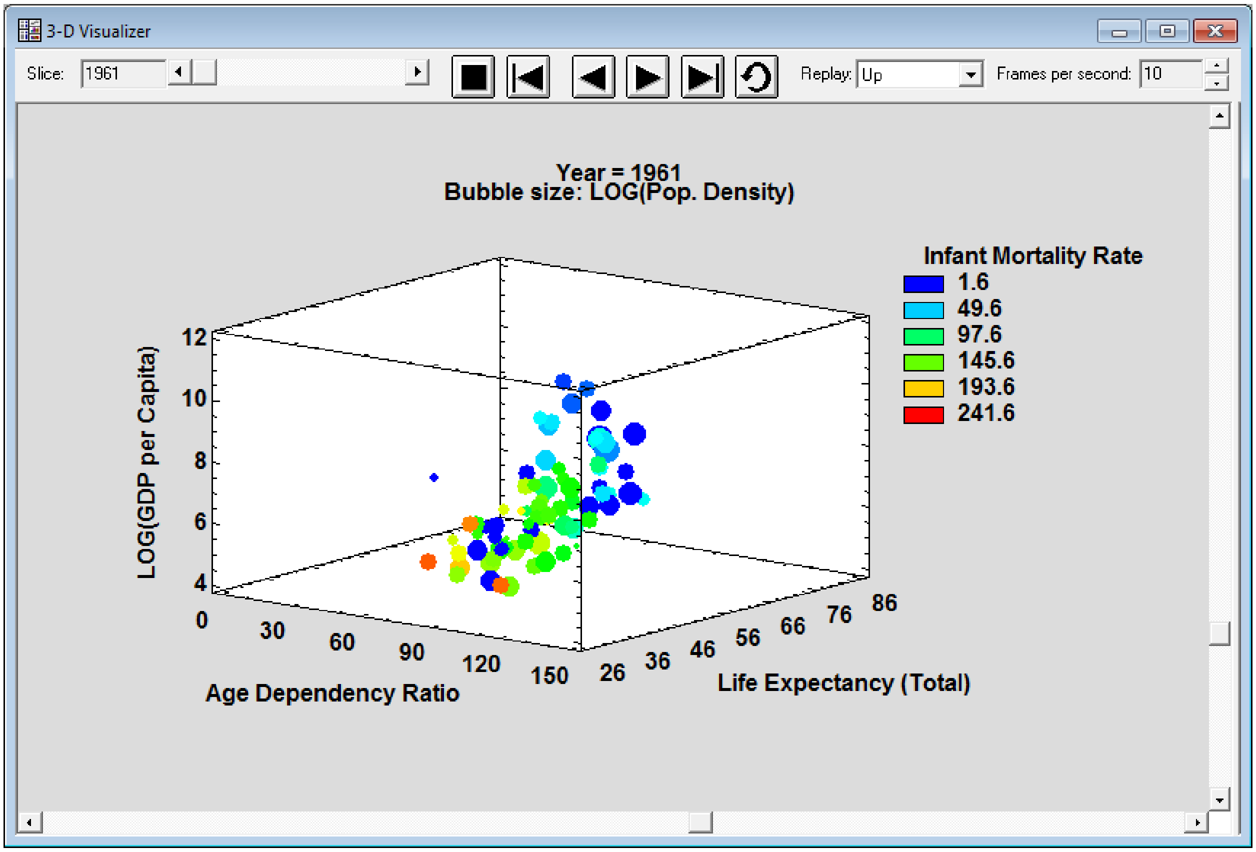
- **Size:** name of the numeric column used to scale the size of the bubbles.

- **Color:** name of the numeric or non-numeric column used to determine the color of each bubble.

- **Select:** optional subset selection.

Statlet

The output of this procedure is displayed in a dynamic Statlet window. When first created, the window displays data for the first time period (or first value of the *Slicer*) as shown below:



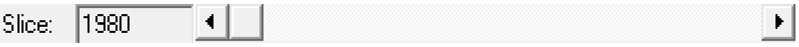
3-D Visualizer

Slice: 1961

Replay: Up

Frames per second: 10

There is a single bubble for each country in the file. The Statlet toolbar contains the following controls:



Slice: 1980

Slice scrollbar: used to change the time period at which the data are displayed.

▶ **Forward button**: used to start a timer which plots the data for each time period in increasing order.

◀ **Backward button**: used to start a timer which plots the data for each time period in decreasing order.

▶| **Fast foward button**: advances to the last time period.

|◀ **Rewind button**: rewinds to the first time period.

⟲ **Replay button**: causes the sequence of time periods to be replayed over and over.

■ **Stop button**: stops the timer or replay.

Replay: Up

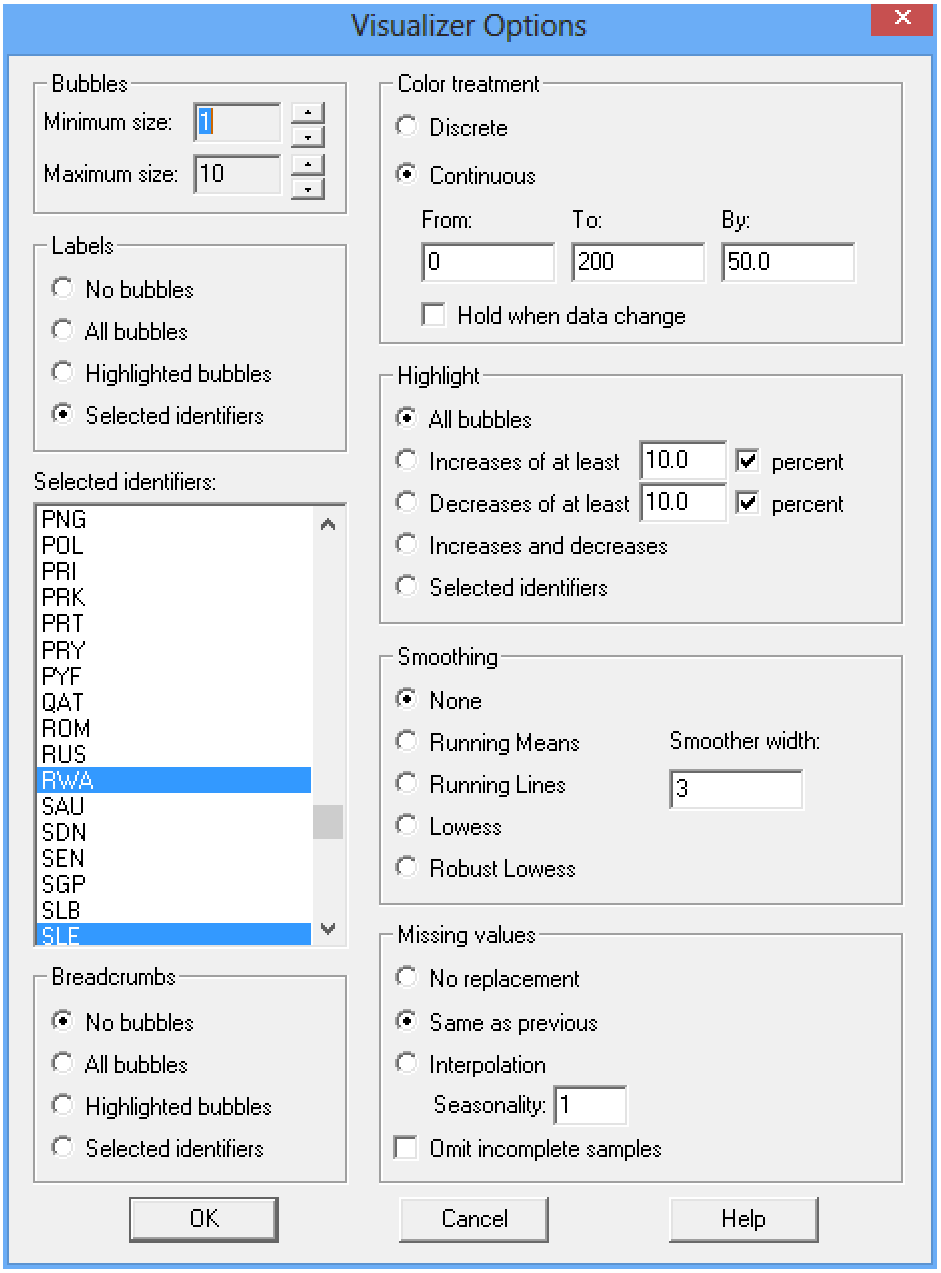
Replay pulldown list: specifies the direction for the time sequence when the replay button is pushed.

Frames per second: 1

Frames per second spinner: specifies the rate at which the time period is changed.

Analysis Options

The *Analysis Options* dialog box allows for various special effects to be created:



Bubbles
Minimum size: 1
Maximum size: 10

Labels
○ No bubbles
○ All bubbles
○ Highlighted bubbles
● Selected identifiers

Selected identifiers:
PNG
POL
PRI
PRK
PRT
PRY
PYF
QAT
ROM
RUS
RWA
SAU
SDN
SEN
SGP
SLB
SLE

Breadcrumbs
● No bubbles
○ All bubbles
○ Highlighted bubbles
○ Selected identifiers

Color treatment
○ Discrete
● Continuous
From: 0 To: 200 By: 50.0
☐ Hold when data change

Highlight
● All bubbles
○ Increases of at least 10.0 ☑ percent
○ Decreases of at least 10.0 ☑ percent
○ Increases and decreases
○ Selected identifiers

Smoothing
● None
○ Running Means
○ Running Lines
○ Lowess
○ Robust Lowess
Smoother width: 3

Missing values
○ No replacement
● Same as previous
○ Interpolation
Seasonality: 1
☐ Omit incomplete samples

The following options are available:

- **Bubbles**: sets the minimum and maximum size of the bubbles. The units are in pixels.
- **Labels**: specifies which bubbles are to be labeled with their sample identifiers.
- **Color treatment**: specifies the manner in which the color variable should be handled. *Discrete* will display a different color for each unique value of that variable. *Continuous* will color each variable based on a continuous palette, which by default ranges from blue

at the lowest level to red at the highest level. *Graphics Options* may be used to change the palette.

- **Highlight**: controls which bubbles are plotted at full intensity. You can elect to highlight a subset of the bubbles in order to demonstrate special features of the data. *Increase* and *decrease* may be used to highlight data which have changed significantly from the previous time period.

- **Selected identifiers**: When labeling or highlighting only selected bubbles, click on each sample identifier that you wish to select.

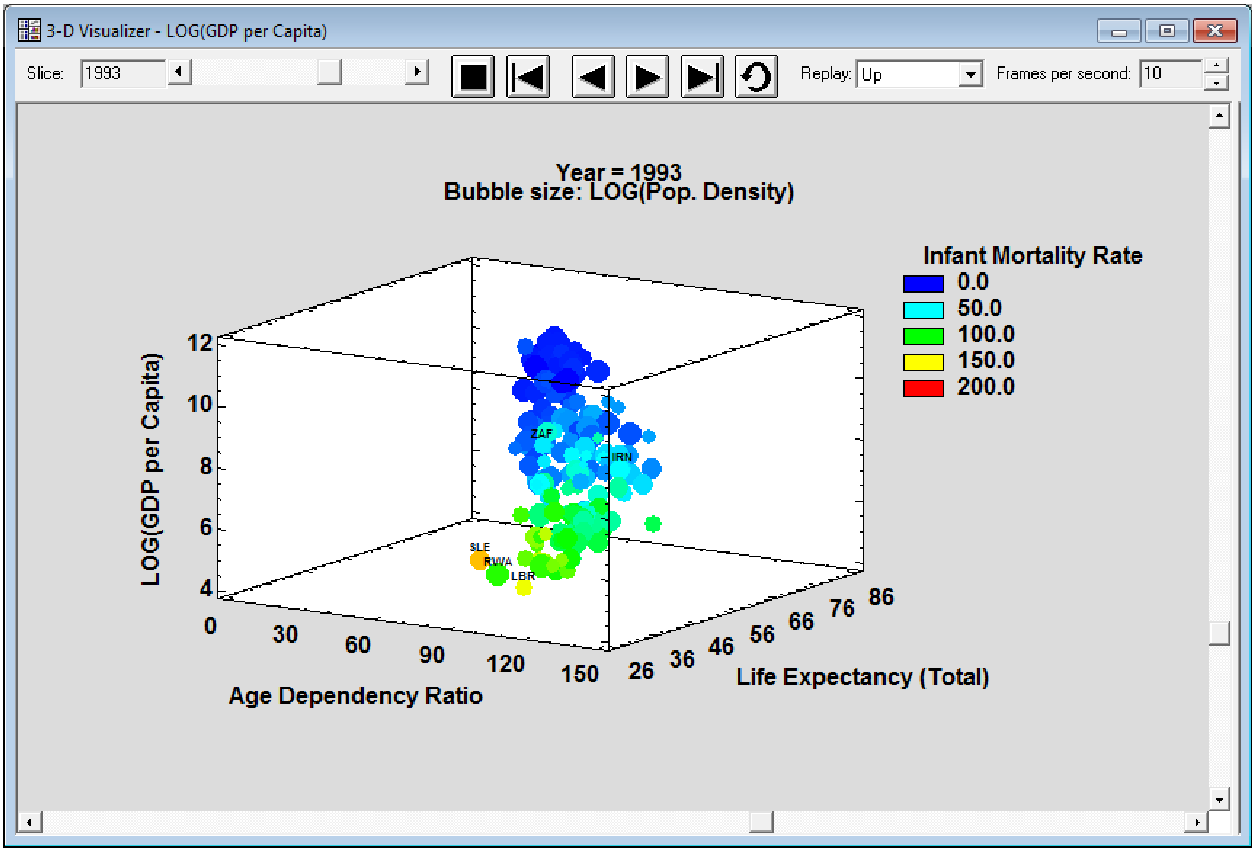
- **Smoothing**: smoothes each time series using one of four methods. These are the same methods used to smooth X-Y scatterplots as described in the PDF document titled *Graphics Options*. If the data contain a large amount of sampling error, smoothing the time series will cause the points to move more smoothly as time is changed.

- **Breadcrumbs**: leaves a transparent image of the data at earlier time periods for the selected bubbles.

- **Missing values**: specifies how missing values should be treated. By default, missing values are not plotted, so that bubbles may appear and disappear as time changes. Selecting *Same as previous* will cause missing values to be replaced with the closest previous value which is not missing, which will cause bubbles to pause in one place but not disappear. Interpolation fills in missing values using an interpolation of 4 adjacent values, as described in the *Calculations* section of this document. If the data are seasonal, indicate the length of seasonality s to be used in the interpolation (for seasonal monthly data, $s = 12$). For nonseasonal data, $s = 1$.

- **Omit incomplete traces**: plots only time series with no missing data (after the missing value substitution is performed).

The plot below shows the output for 1993, replacing missing values with the previous values, labeling the bubbles for several countries, and rescaling the color treatment.



3-D Visualizer - LOG(GDP per Capita)

Slice: 1993

Replay: Up

Frames per second: 10

Dramatic changes in unemployment, life expectancy, and other variables occurred between 1961 and 1993.

Calculations

The *interpolation* method may be used to replace a limited number of missing values in each time series, provided there are not too many missing values close together. Before the data is analyzed, missing values are replaced by interpolated values, determined using the following rule:

1. If y_t, the observation at time t, is missing, find the two observations in the same season that precede time t (y_{t-s} and y_{t-2s}) and the two observations in the same season that come after time t (y_{t+s} and y_{t+2s}).

2. If none of the four observations are missing, then the replacement value for y_t is:

$$y_t = \frac{-3y_{t-2s} + 12y_{t-s} + 12y_{t+s} - 3y_{t+2s}}{18} \tag{1}$$

3. If y_{t+2s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{-y_{t-2s} + 3y_{t-s} + y_{t+s}}{3} \tag{2}$$

4. If y_{t+s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{-3y_{t-2s} + 8y_{t-s} + y_{t+s}}{6} \tag{3}$$

5. If y_{t-s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + 8y_{t+s} - 3y_{t+2s}}{6} \tag{4}$$

6. If y_{t-2s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-s} + 3y_{t+s} - y_{t+s}}{3} \tag{5}$$

7. If y_{t+s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = -y_{t-2s} + 2y_{t-s} \tag{6}$$

8. If y_{t-s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + 2y_{t+s}}{3} \tag{7}$$

9. If y_{t-s} and y_{t+s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + y_{t+2s}}{2} \tag{8}$$

10. If y_{t-2s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-s} + y_{t+s}}{2} \tag{9}$$

11. If y_{t-2s} and y_{t+s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{2y_{t-s} + y_{t+2s}}{3} \tag{10}$$

12. If y_{t-2s} and y_{t-s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = 2y_{t+s} - y_{t+2s} \tag{11}$$

If more than 2 of the four observations are missing, the missing value will not be replaced.

The interpolated values are designed to perfectly reproduce a quadratic trend (if only one observation is missing) or a linear trend (if two observations are missing), provided no noise is present.