One-Dimensional Visualizer

Summary

The One-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 display. The position along the Y axis represents the value of the primary data variable. 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. Various options are offered for smoothing the data and for dealing with missing values.

Sample StatFolio: *visualize1d.sgp*

Sample Data

The file *oecd.sgd* contains data for *n* = 22 countries over *p* = 31 years (1980-2010). It was obtained from the Organisation for Economic Co-Operation and Development (oecd.org). Variables in the file include:

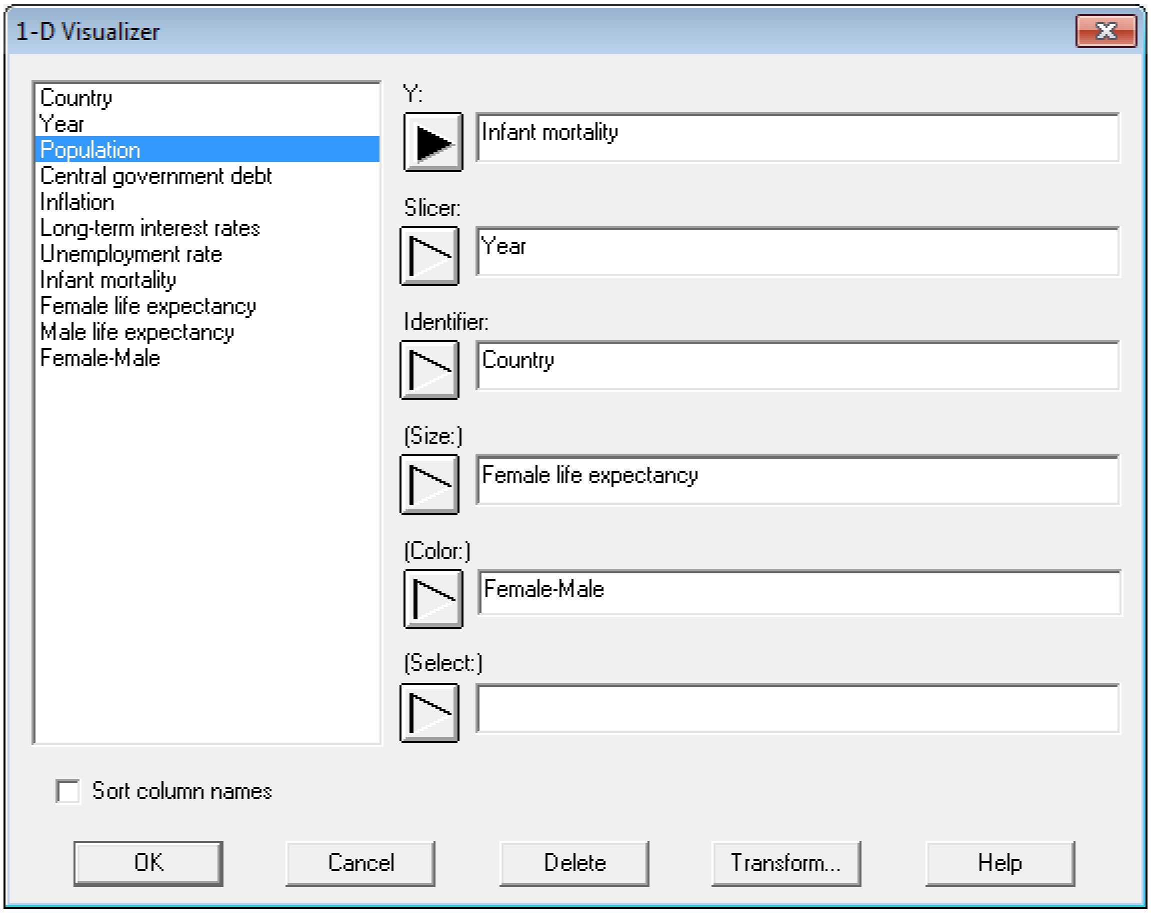
- Population
- Central government debt
- Inflation
- Long-term interest rates
- Unemployment rate
- Infant mortality
- Female life expectancy
- Male life expectancy

Several rows and columns of that file are shown below:

Country	Year	Population	Infant Mortality	Female Life Expectancy	Male Life Expectancy	Female - Male
			deaths per 1,000 live births	at birth	at birth	
Australia	1980	14695360	10.7	78.1	71	7.1
Australia	1981	14923260	10	78.4	71.4	7
Australia	1982	15184250	10.3	78.2	71.2	7
Australia	1983	15393470	9.6	78.8	72.1	6.7
Australia	1984	15579390	9.2	79	72.5	6.5
Australia	1985	15788310	9.9	78.8	72.4	6.4
Australia	1986	16018350	8.8	79.2	72.9	6.3
Australia	1987	16263870	8.7	79.5	73.1	6.4
Australia	1988	16532160	8.7	79.5	73.1	6.4
Australia	1989	16814420	8	79.6	73.3	6.3

Data Input

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



Y: Infant mortality

Slicer: Year

Identifier: Country

(Size:) Female life expectancy

(Color:) Female-Male

(Select:)

□ Sort column names

- **Y:** 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.

- **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.

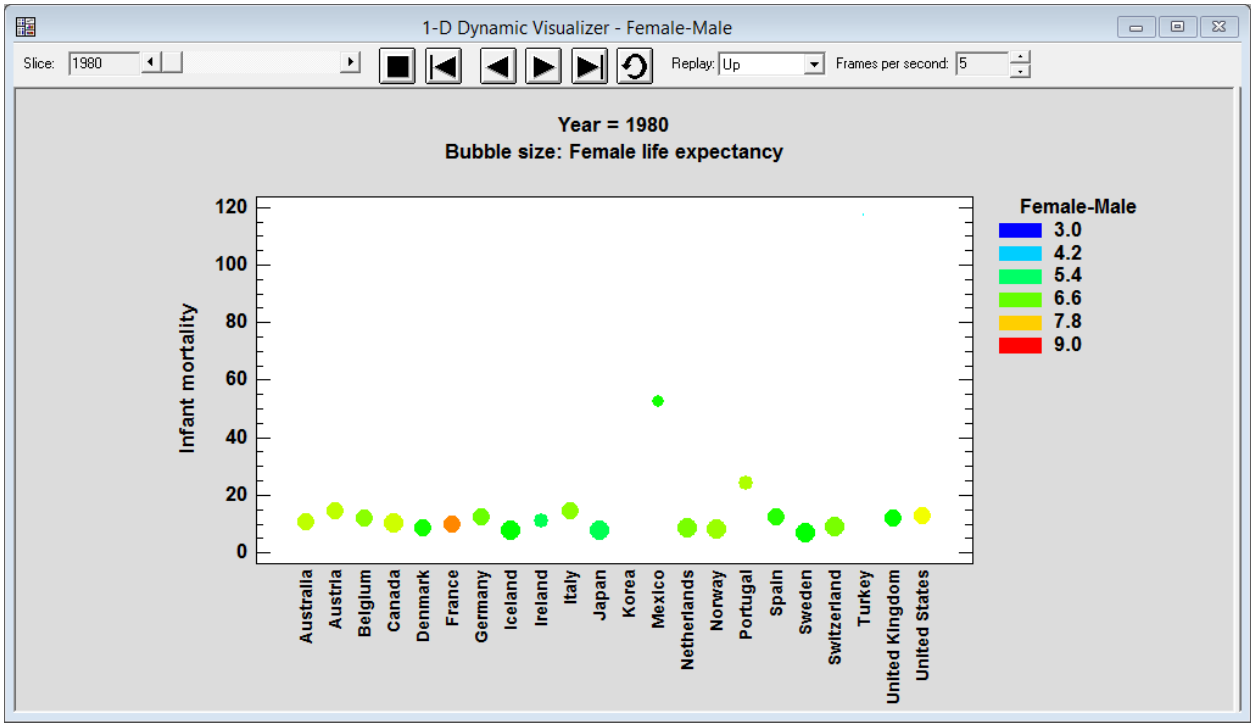
- **Size:** name of the numeric column used to scale the size of the bubbles.

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

- **Select:** optional subset selection.

Statlet

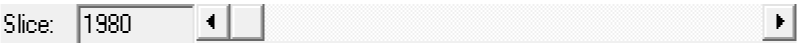
The output of this procedure is displayed in a dynamic Statlet window. When first created, the window displays data for the first time period (or first value of the *Slicer*) as shown below:


Slice: 1980

Replay: Up

Frames per second: 5

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

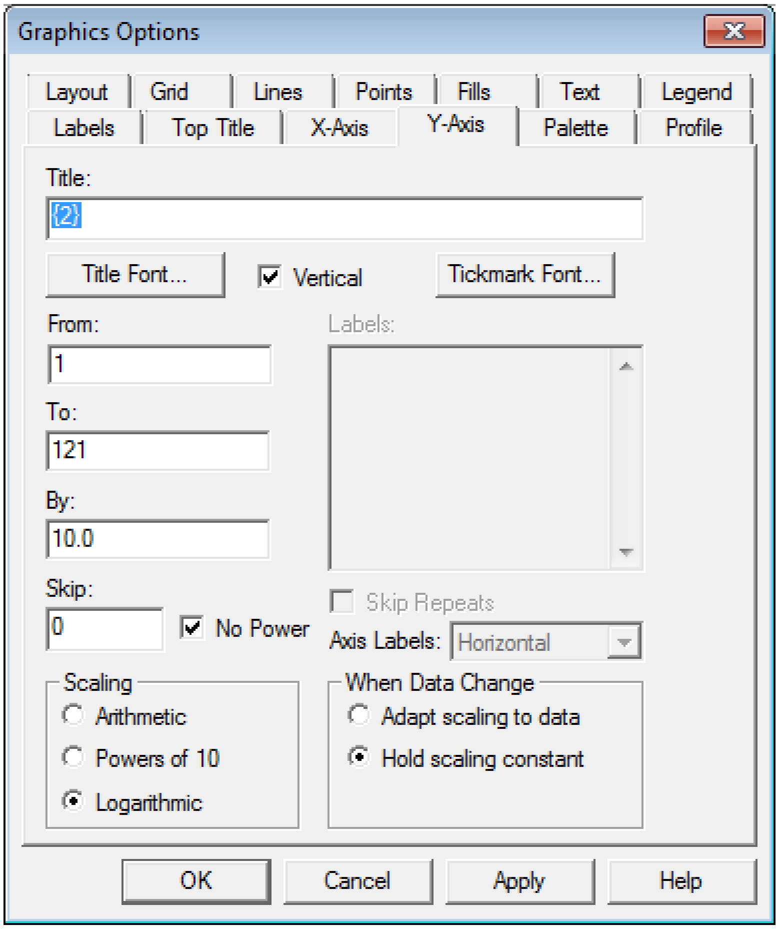
Replay pulldown list: specifies the direction for the time sequence when the replay button is pushed.

Frames per second: 1

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

Graphics Options

The standard *Graphics Options* tabbed dialog box may be used to change various aspects of the plot. For example, if the variable used to scale the Y axis varies over an order of magnitude, it may be useful to use a logarithmic scale for that axis. Click anywhere on the graph with the right mouse button and select *Graphics Options* from the popup menu. Set the *Y-Axis* tab as shown below:



Graphics Options

Title: {2}

☑ Vertical

From: 1

To: 121

By: 10.0

Skip: 0

☑ No Power

Labels:

□ Skip Repeats

Axis Labels: Horizontal

Scaling

○ Arithmetic

○ Powers of 10

● Logarithmic

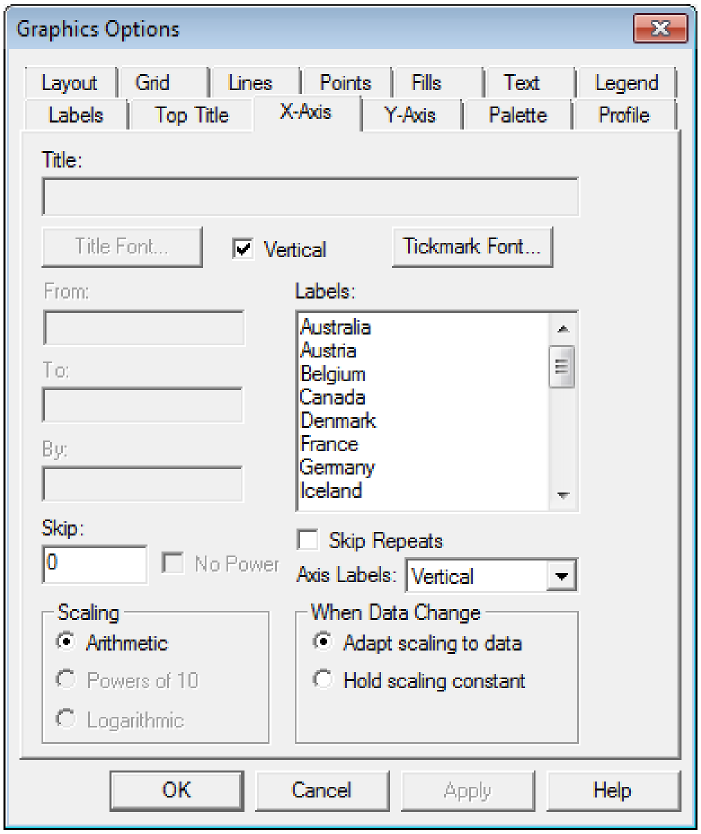
When Data Change

○ Adapt scaling to data

● Hold scaling constant

Applying a logarithmic transformation to the Y axis for the current example shows more detail about countries with low infant mortality (be sure to check the *Hold scaling constant* checkbox so that the scaling does not revert back to the default settings when the data change).

It is also helpful to rotate the X-axis ticmarks so that they are vertical and reduce the ticmark font size. To do so, click on the X-axis tab. Change the *Axis Labels* setting to *Vertical*. Press the *Tickmark Font* button and change the font size to 9.



Title:

☑ Vertical

Labels:
Australia
Austria
Belgium
Canada
Denmark
France
Germany
Iceland

Skip: 0

□ No Power

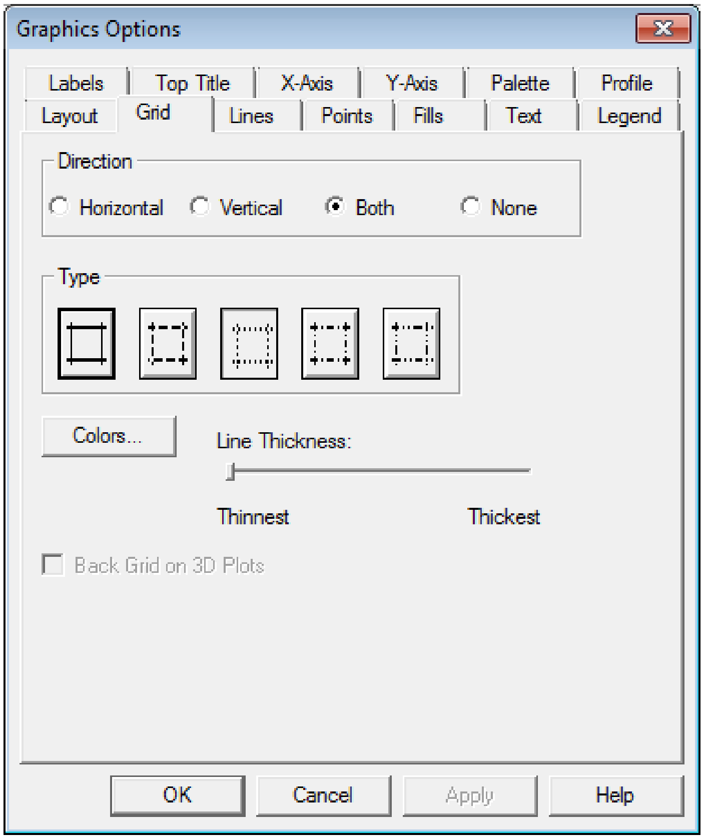
□ Skip Repeats

Axis Labels: Vertical

Scaling
● Arithmetic
○ Powers of 10
○ Logarithmic

When Data Change
● Adapt scaling to data
○ Hold scaling constant

Finally, adding a grid may also be helpful:

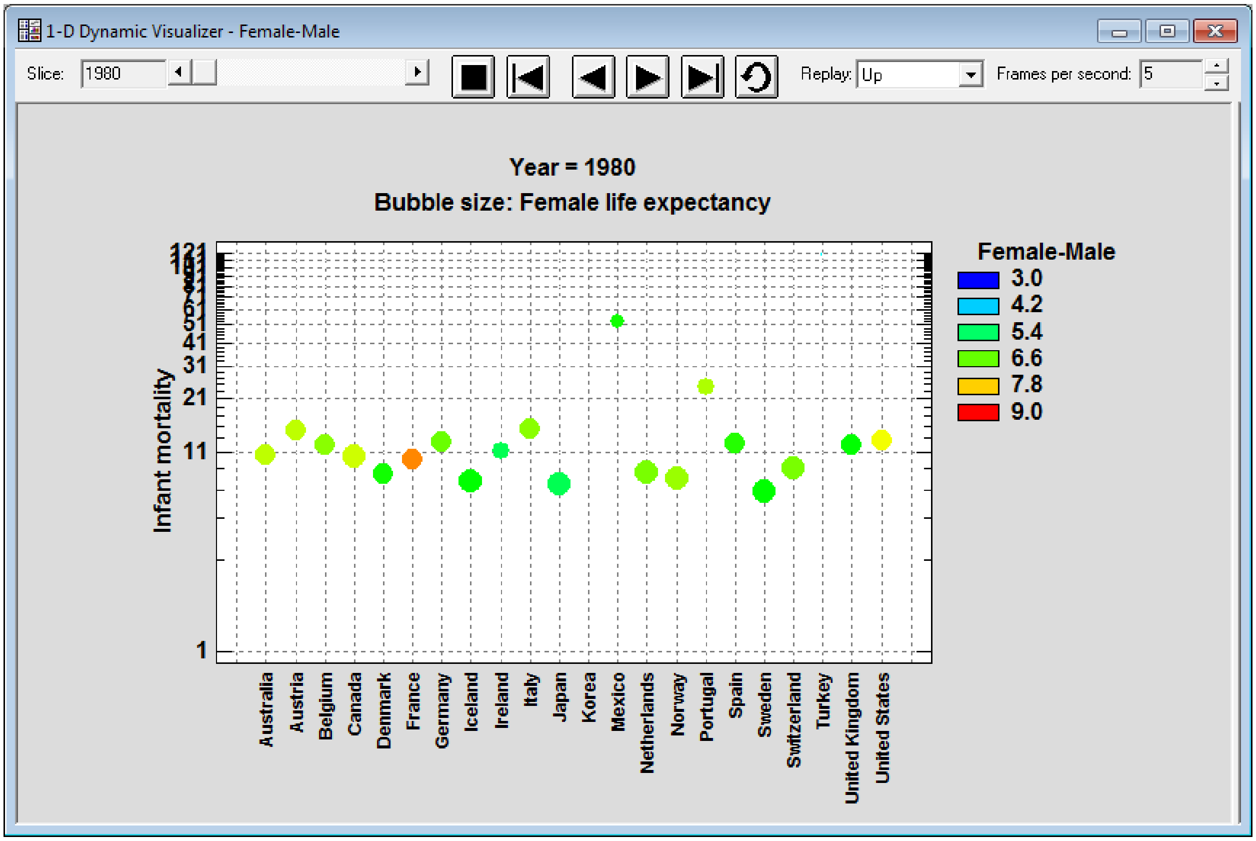


Direction: ○ Horizontal ○ Vertical ● Both ○ None

Line Thickness: Thinnest

□ Back Grid on 3D Plots

The resulting display is shown below:



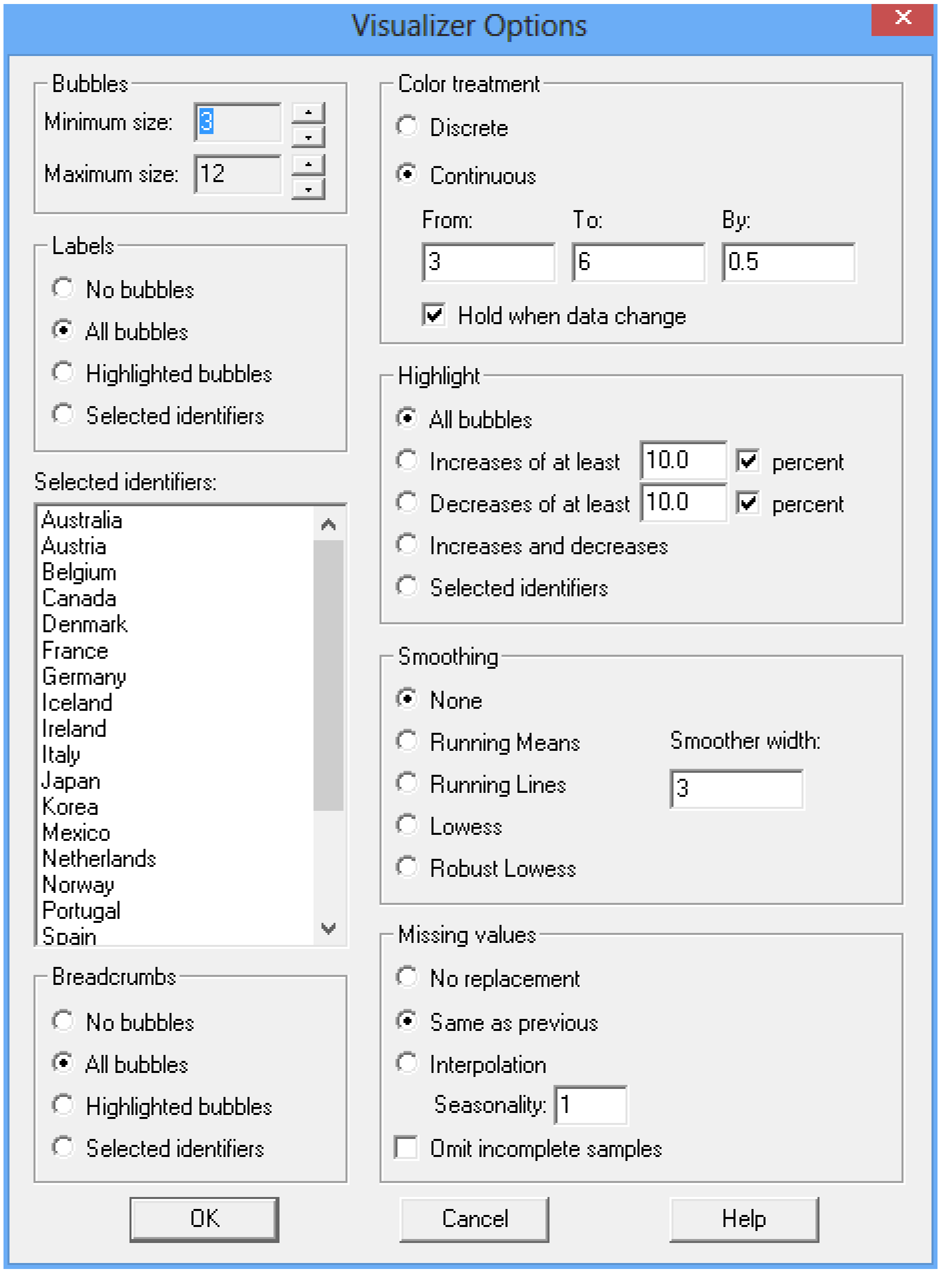
Slice: 1980

Replay: Up

Frames per second: 5

Analysis Options

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



Visualizer Options

Bubbles
Minimum size: 3
Maximum size: 12

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

Selected identifiers:
Australia
Austria
Belgium
Canada
Denmark
France
Germany
Iceland
Ireland
Italy
Japan
Korea
Mexico
Netherlands
Norway
Portugal
Spain

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

Color treatment
○ Discrete
● Continuous
From: 3 To: 6 By: 0.5
☑ 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**: specifies 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.

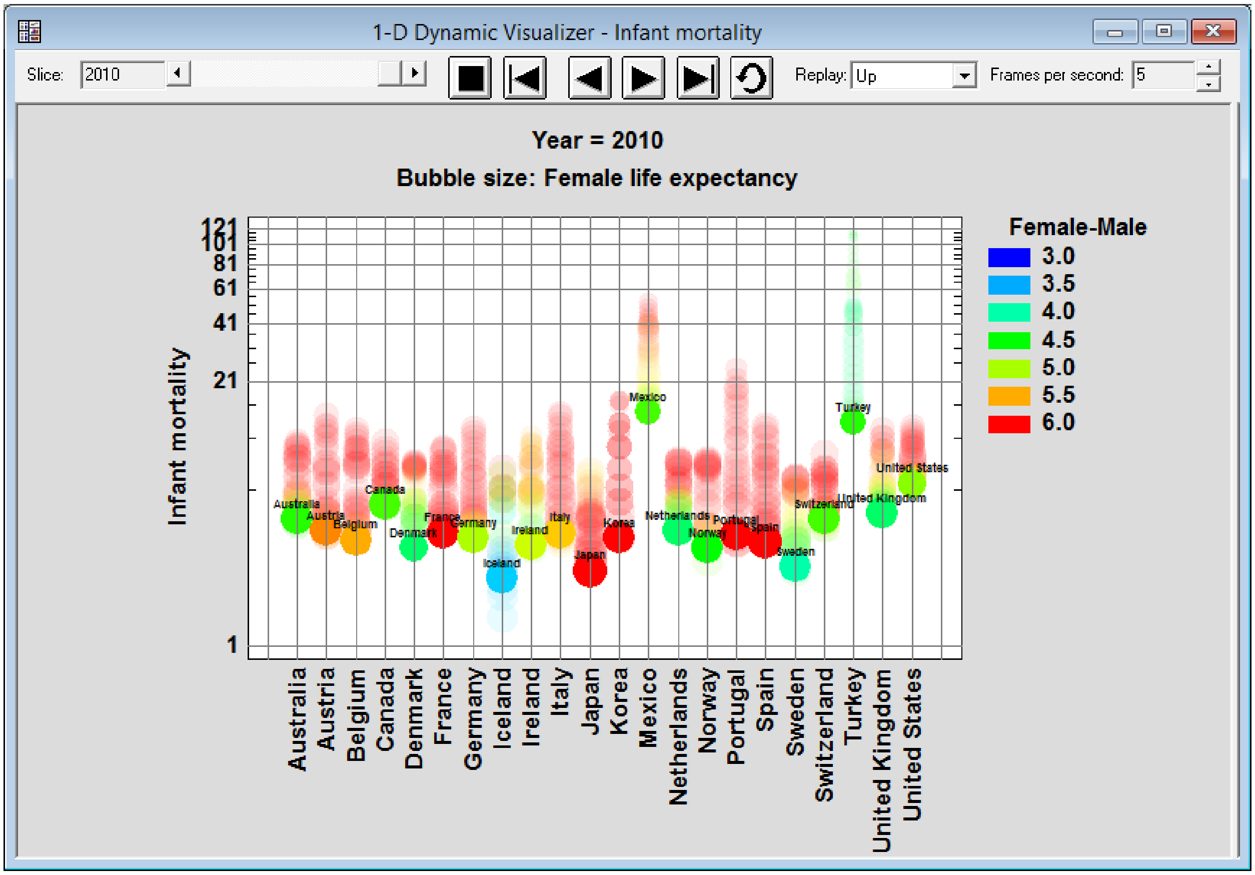
- **Smoothing**: smoothes each time series using one of four methods. These are the same methods used to smooth X-Y scatterplots as described in the PDF document titled *Graphics Options*. If the data contain a large amount of sampling error, smoothing the time series will cause the points to move more 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 the final time period, replacing missing values with the previous values and plotting breadcrumbs.



1-D Dynamic Visualizer - Infant mortality

Slice: 2010

Replay: Up

Frames per second: 5

In all countries, life expectancy has increased significantly between 1980 and 2010. Notice that infant mortality in Turkey and Mexico was high in the 1980’s, but decreased rapidly. In 2010, the difference in life expectancy between males and females is much higher in countries such as Japan and Korea than in Iceland or Sweden.

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.