

## Demographic Map Visualizer (Regions)



Revised: 10/9/2017



Summary .....	1
Data Input.....	3
Statlet .....	8
Analysis Options.....	11
Boundary Files .....	17

### Summary

The **Demographic Map Visualizer (Regions)** Statlet is designed to illustrate changes in regional statistics over time. Given data for each of  $k$  regions during  $p$  time periods, the program generates a dynamic display that illustrates how the data have changed in each region. Typical applications include plotting:

1. Population and other demographic measurements.
2. Unemployment indices, housing starts, and other economic indices.
3. Quarterly sales of products such as automobiles.

Each region is drawn using a color that illustrates the level of the selected variable. As time changes, the analyst can follow changes in the data within each region. Various options are offered for smoothing the data and for dealing with missing values.

**Sample StatFolio:** *visualizeus.sgp*

## Sample Data

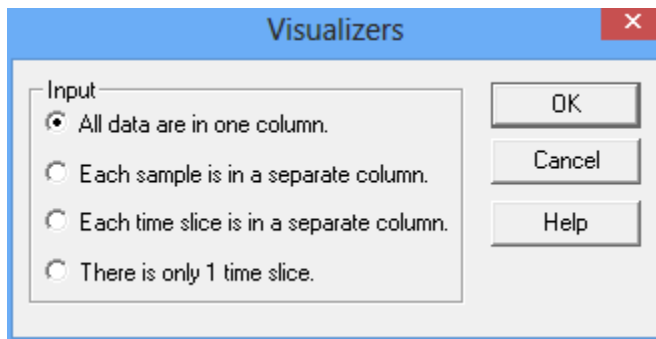
The file *crimerates.sgd* contains data for each state in the United States and the District of Columbia over  $p = 46$  years (1965-2010). The data were obtained from the FBI's Uniform Crime Reporting (UCR) program. The first several rows and columns of the file are shown below:

State	Year	Population	Total Crime Rate per 100,000 population	Violent Crime Rate total	Murder and Manslaughter
Alaska	1965	253000	2603.6	149.0	6.3
Alaska	1966	272000	2785.7	150.4	12.9
Alaska	1967	272000	2884.6	160.7	9.6
Alaska	1968	277000	3320.9	175.5	10.5
Alaska	1969	282000	3880.1	221.3	10.6
Alaska	1970	302173	3801.8	278.0	12.2
Alaska	1971	313000	4164.5	355.3	13.4
Alaska	1972	325000	4478.5	370.5	9.5
Alaska	1973	330000	4943.3	384.5	10.0
Alaska	1974	337000	5239.8	453.1	13.6

## Data Input

### Data Input Dialog Box

Data to be displayed by the *Demographic Map Visualizer* may be arranged in any of 4 ways. When selected from the main menu, the procedure first displays a dialog box which specifies how the data are structured:



1. **All data are in one column:** All data are placed in a single data column, as with the crime data file displayed above. Additional columns are then constructed to identify the regions (“samples”) and years (“slices”) associated with each row of the file. This format allows for more than one variable to be stored in the same data file.
2. **Each sample is in a separate column:** The data for each region or sample are placed in a separate data column, as in the data file shown below:

	Year	Alaska	Alabama	Arkansas	Arizona	California	Colorado	Connecticut	Delaware
1	1965	2603.6	1592.5	1274.2	3547.8	4319.4	2704.5	1834.4	2408.7
2	1966	2785.7	1758.3	1382.9	4135.8	4549.4	3009.6	1982.3	2619.7
3	1967	2884.6	1851.0	1628.5	4837.9	5055.1	3309.1	2281.2	2891.6
4	1968	3320.9	1999.0	1958.5	4874.4	5721.1	3862.6	2890.4	3165.5
5	1969	3880.1	2126.6	2188.5	5224.6	6099.7	4498.2	3225.5	3501.7
6	1970	3801.8	2479.5	2421.2	5914.2	6339.1	5318.2	3489.4	4263.1
7	1971	4164.5	2498.4	2328.2	5941.5	6690.1	5517.0	3646.2	5015.1
8	1972	4478.5	2394.5	2352.4	5933.3	6413.1	5593.6	3403.1	4523.7
9	1973	4943.3	2582.3	2756.5	6703.9	6304.9	5495.8	3664.4	4582.6
10	1974	5239.8	3000.1	3300.7	8221.7	6846.8	6165.8	4407.0	5949.6
11	1975	6196.6	3472.5	3540.1	8341.5	7204.6	6675.5	4957.0	6668.2
12	1976	6220.7	3808.3	3406.7	7886.4	7234.0	6782.4	5004.6	6264.4

3. **Each time slice is in a separate column:** The data for each time period or slice are placed in a separate data column, as in the data file shown below:

	State	1965	1966	1967	1968	1969	1970	1971	1972
1	Alaska	2603.6	2785.7	2884.6	3320.9	3880.1	3801.8	4164.5	4478.5
2	Alabama	1592.5	1758.3	1851.0	1999.0	2126.6	2479.5	2498.4	2394.5
3	Arkansas	1274.2	1382.9	1628.5	1958.5	2188.5	2421.2	2328.2	2352.4
4	Arizona	3547.8	4135.8	4837.9	4874.4	5224.6	5914.2	5941.5	5933.3
5	California	4319.4	4549.4	5055.1	5721.1	6099.7	6339.1	6690.1	6413.1
6	Colorado	2704.5	3009.6	3309.1	3862.6	4498.2	5318.2	5517.0	5593.6
7	Connecticut	1834.4	1982.3	2281.2	2890.4	3225.5	3489.4	3646.2	3403.1
8	Delaware	2408.7	2619.7	2891.6	3165.5	3501.7	4263.1	5015.1	4523.7
9	Florida	3320.2	3716.3	4103.6	4498.5	4742.5	5317.2	5673.0	5376.9
10	Georgia	1764.3	1879.8	1954.2	2156.1	2399.5	2881.6	3047.9	3051.8
11	Hawaii	3252.3	3503.2	3719.4	4438.3	4532.9	5265.1	5458.8	4612.5

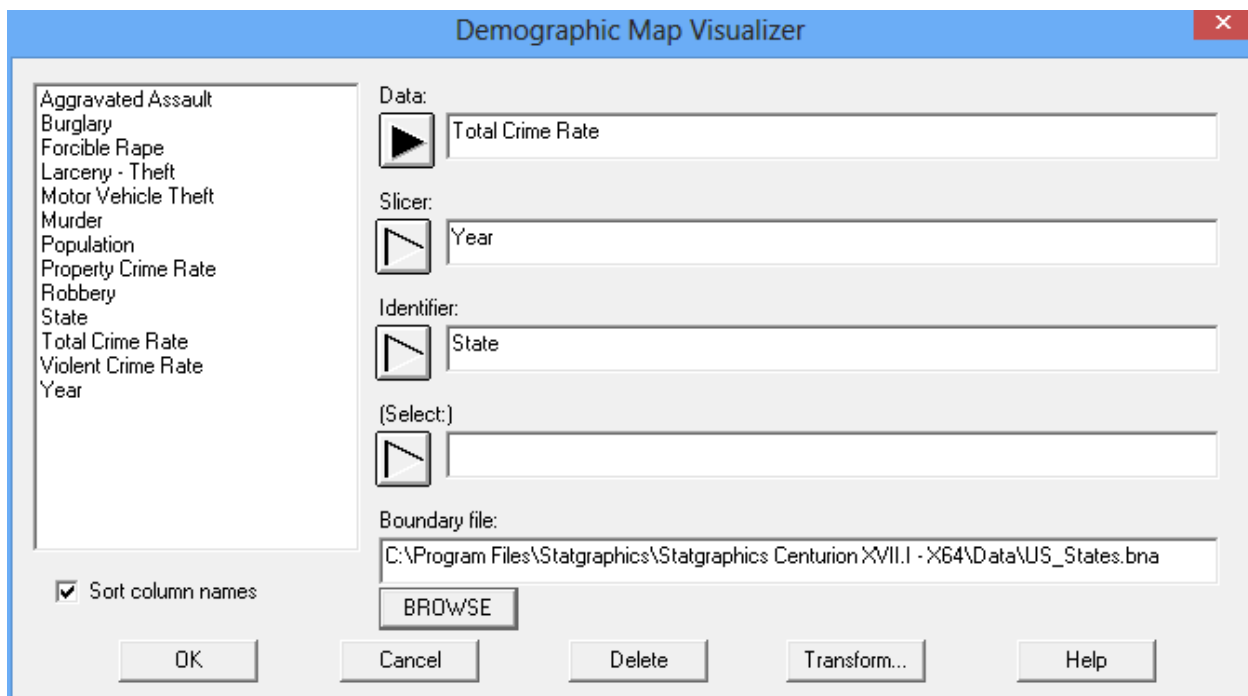
4. **There is only 1 time slice:** The data consist of 1 column with the measurements and a second column identifying the locations:

	State	Population	Total Crime Rate	Violent Crime Rate	Murder	Forcible Rape	Robbery	Aggrava Assault
			per 100,000 population	total				
1	Alaska	710231	3491.3	638.8	4.4	75.0	83.6	475.8
2	Alabama	4779736	3894.6	377.8	5.7	28.2	99.6	244.2
3	Arkansas	2915918	4064.2	505.3	4.7	45.0	81.3	374.3
4	Arizona	6392017	3942.1	408.1	6.4	33.9	108.5	259.3
5	California	37253956	3076.4	440.6	4.9	22.4	156.0	257.4
6	Colorado	5029196	3005.0	320.8	2.4	43.7	62.3	212.4
7	Connecticut	3574097	2474.6	281.4	3.6	16.3	99.4	162.0
8	Delaware	897934	4069.1	620.9	5.3	34.7	203.7	377.1
9	Florida	18801310	4100.8	542.4	5.2	28.6	138.7	369.8
10	Georgia	9687653	4043.8	403.3	5.8	21.6	127.7	248.2
11	Hawaii	1360301	3576.9	262.7	1.8	26.8	77.5	156.7
12	Iowa	3046355	2516.0	273.5	1.3	27.4	33.2	211.6
13	Idaho	1567582	2216.8	221.0	1.3	33.5	13.7	172.6

Structures #2 and #3 can only be used with a single data variable.

All data in one column

After specifying the structure of the data, a second data input dialog box requests the names of the columns containing the data values to be analyzed. For structure #1, the dialog box requests the name of a single data column:

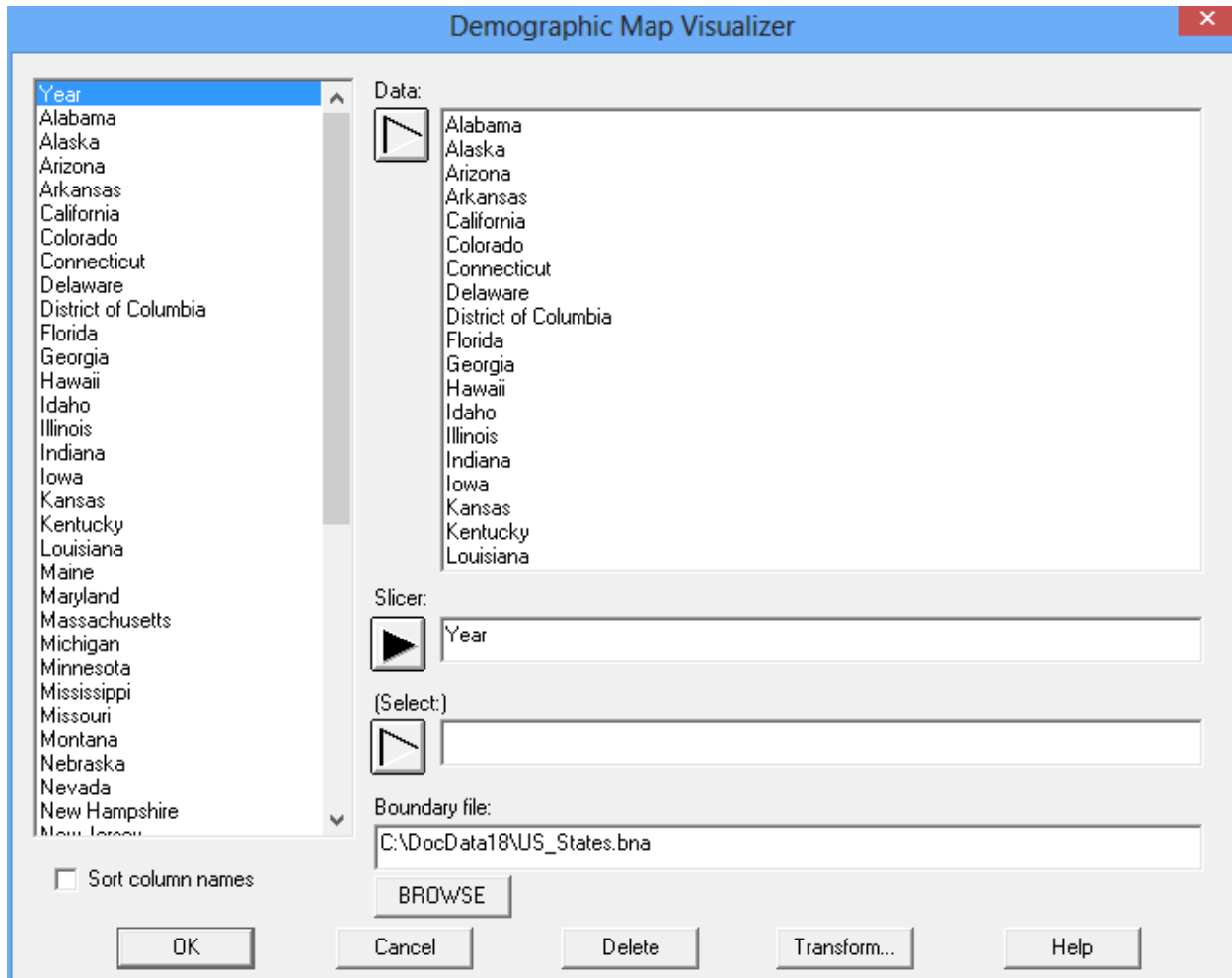


- **Data:** name of the numeric column containing the observations to be analyzed. There should be a total of  $k$  times  $p$  observations.
- **Slicer:** name of the numeric column used to define subsets of the data. This variable, often a measure of time, is changed dynamically to illustrate changes in the other variables. There should be  $p$  unique values of this variable.
- **Identifier:** column containing the region names or other identifier. The entries may be in any order but must match exactly one of the identifier columns contained in the selected boundary file.
- **Select:** optional subset selection.
- **Boundary file:** name of a BNA or SHP file containing the definition of polygon boundaries for the map to be drawn. Two types of boundary files may be used:
  - BNA or *Atlas Boundary Files* are a common type of file for defining map boundaries. Some sample BNA files are supplied with Statgraphics, including *US\_States.bna* which contains definitions of the boundaries for all 50 states and the District of Columbia.
  - SHP *shape files* are widely used for defining map boundaries and other features. Statgraphics supplies some sample SHP files such as *worldmap.shp* which contains the boundaries of countries around the world, and *us\_states.shp* which contains state boundaries.

As an example, the data to be analyzed here is the total crime rate in each state, measured as crimes per 100,000 population.

Each sample or time slice in a separate column

For structures #2 and #3, multiple data columns must be specified:

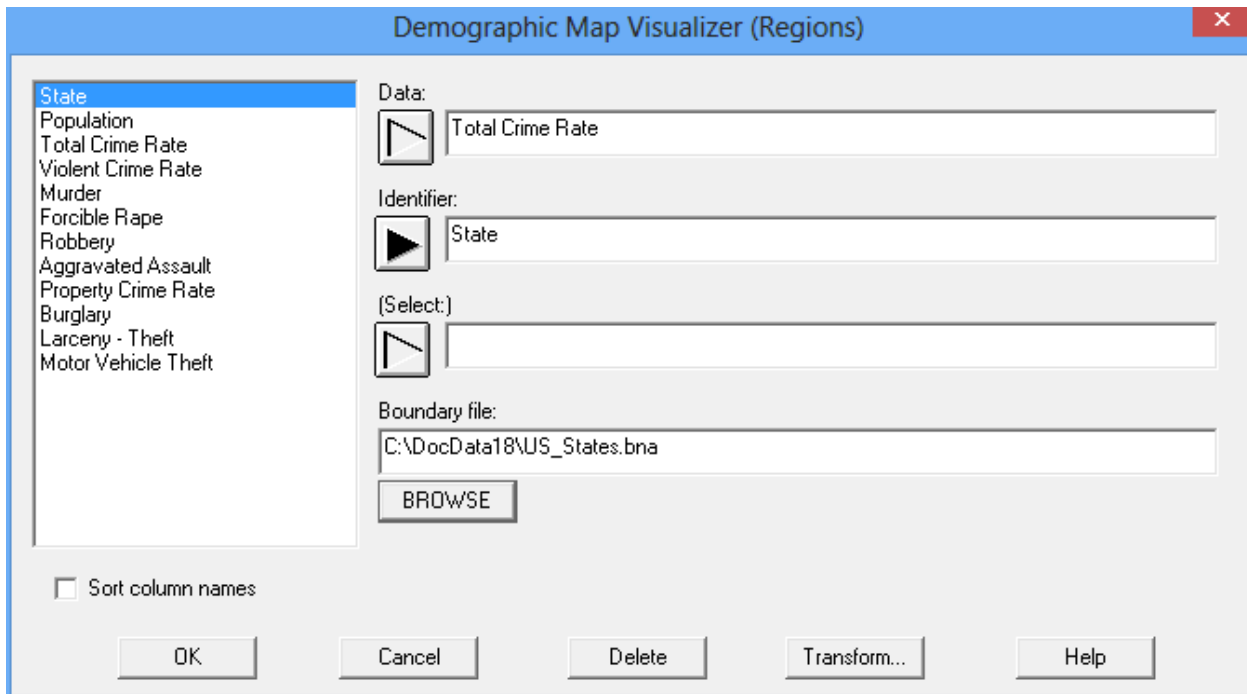


- **Data:** names of 2 or more numeric columns containing the observations to be analyzed.
- **Slicer** (structure #2 only): name of the numeric column used to define subsets of the data. This variable, often a measure of time, is changed dynamically to illustrate changes in the other variables.
- **Identifier** (structure #3 only): column containing the region names or other identifier. The entries may be in any order but must match exactly one of the identifier columns contained in the selected boundary file.
- **Select:** optional subset selection.

**Boundary file:** name of a BNA or SHP file containing the definition of polygon boundaries for the map to be drawn.

## One time slice only

For structure #4, the data input dialog box has the following form:

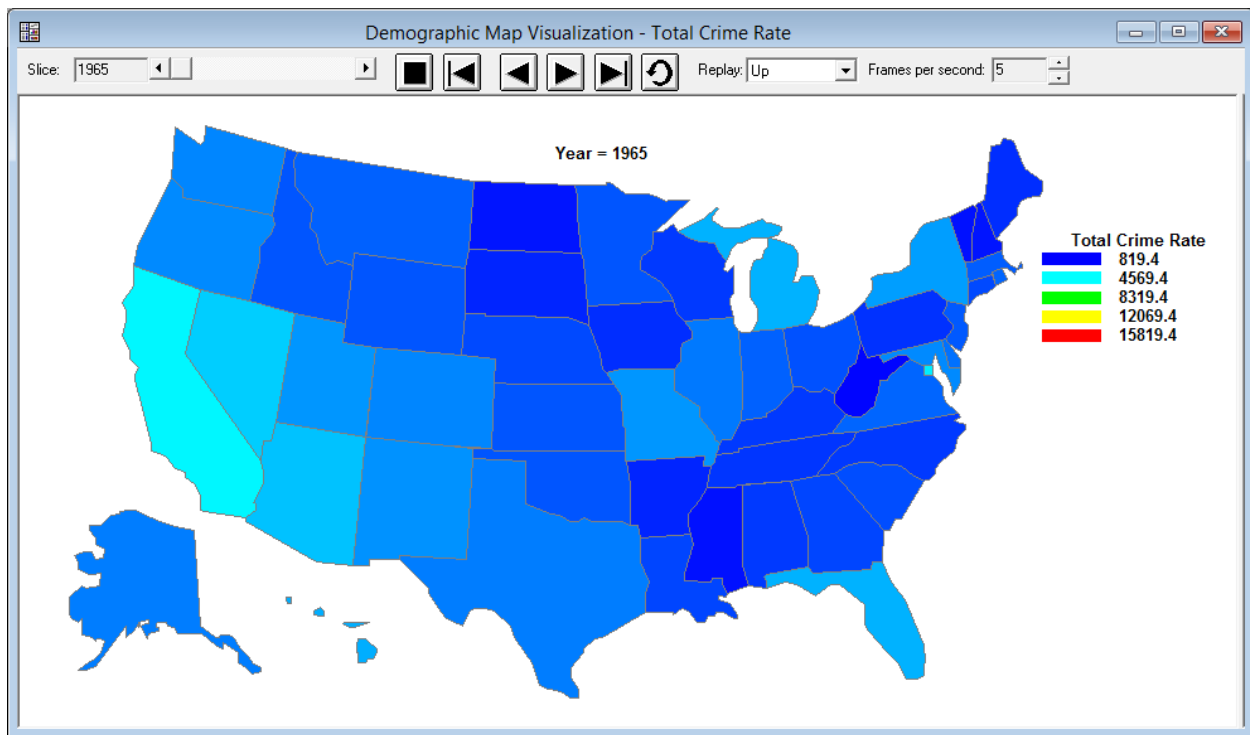


- **Data:** name of a single numeric column containing the observations to be analyzed.
- **Identifier:** column containing the location names or other identifier.
- **Select:** optional subset selection.
- **Boundary file:** name of a BNA or SHP file containing the definition of polygon boundaries for the map to be drawn.

## 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:





Each state is colored on a sliding scale from blue to red, depending on the value in that state. 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 forward button:** advances to the last time period.

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

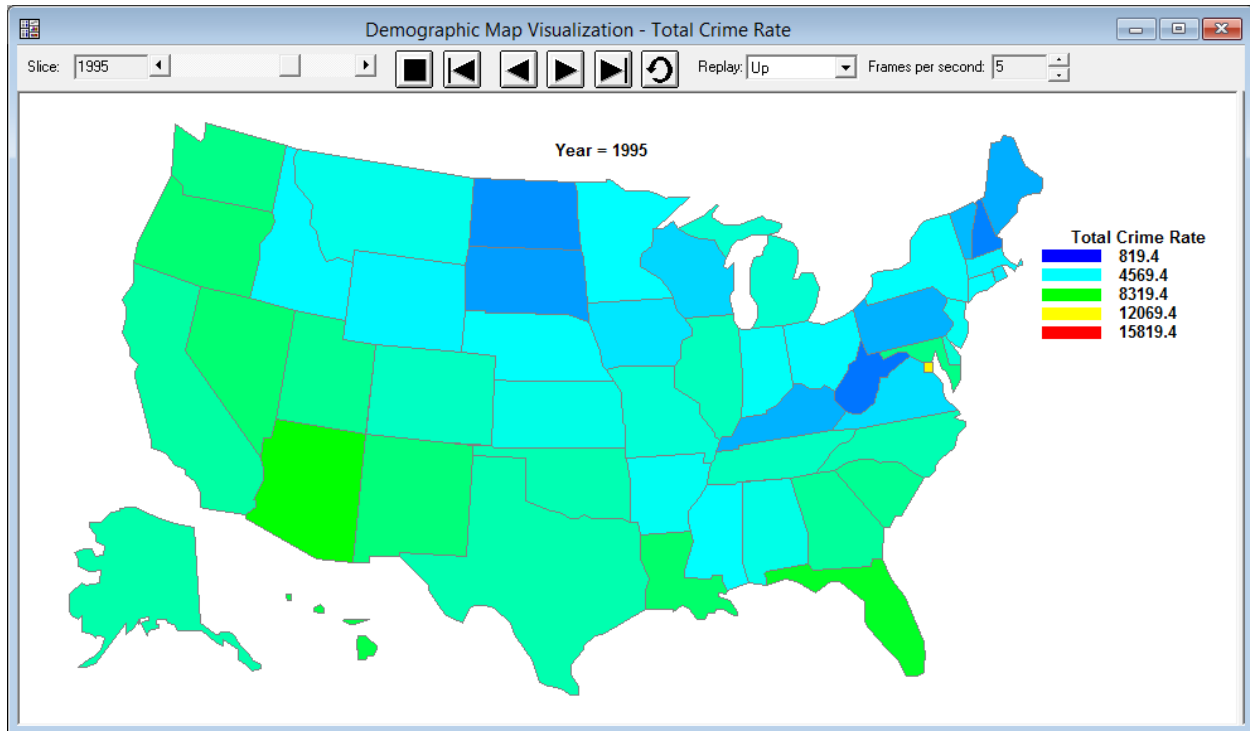
**Replay button:** causes the sequence of time periods to be replayed over and over.

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

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

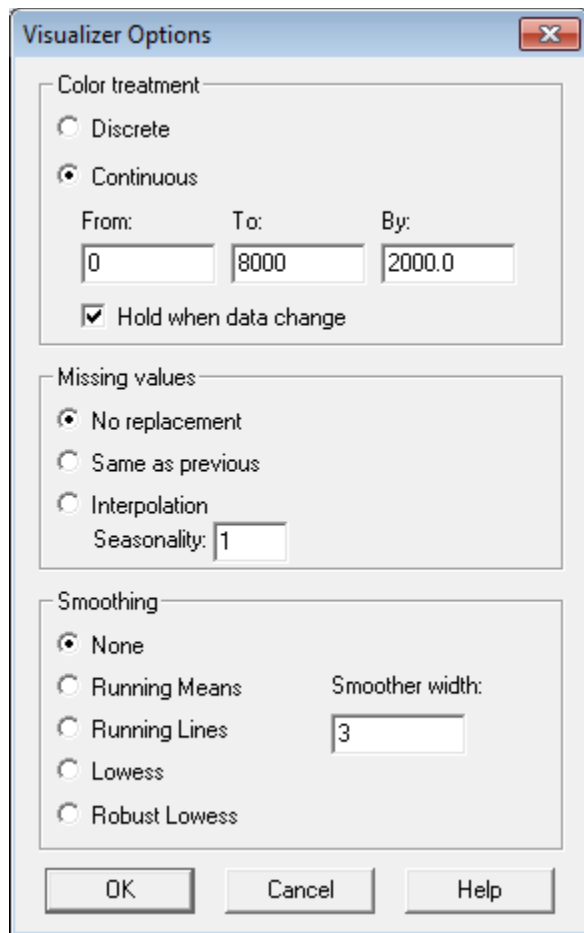
Frames per second:  **Frames per second spinner:** specifies the rate at which the time period is changed.

Changing the time period to 1995 shows a dramatic increase in the crime rate in most of the states.



## Analysis Options

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

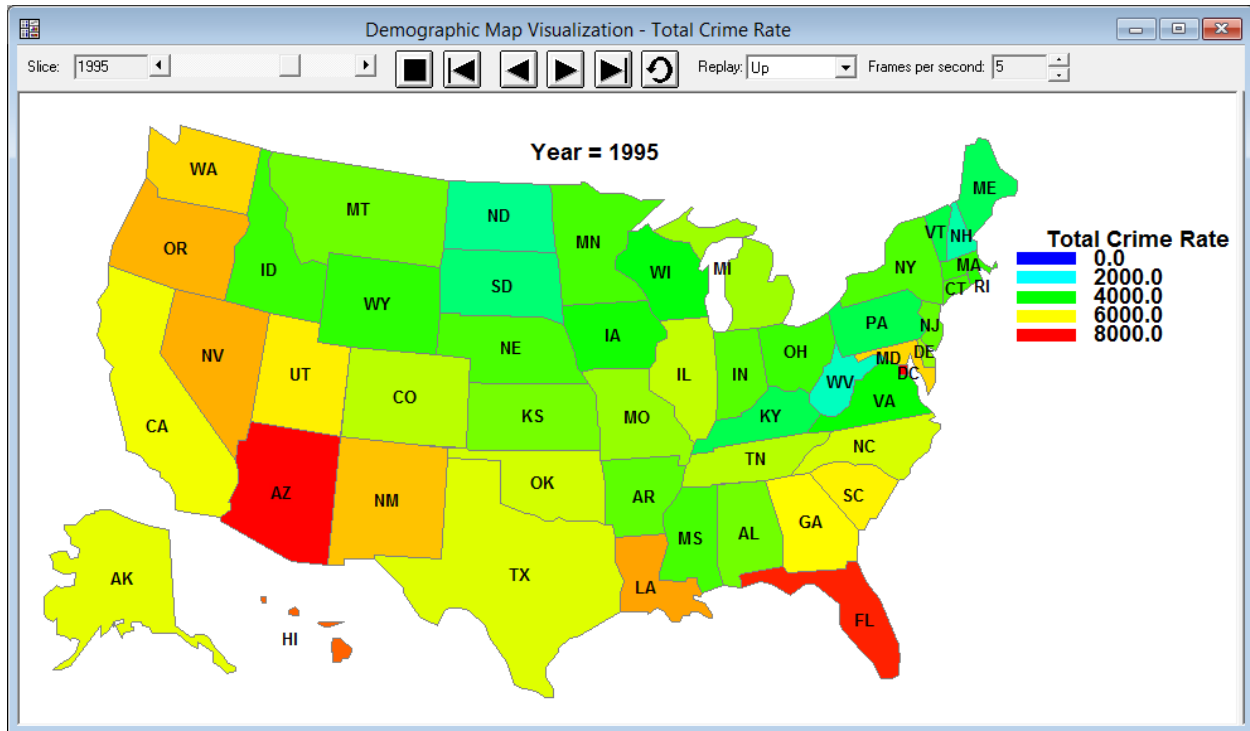


The following options are available:

- 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.
- Missing values:** specifies how missing values should be treated. By default, missing values are not plotted (the states are not filled). Selecting *Same as previous* will cause missing values to be replaced with the closest previous value which is not missing. 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,  $s = 1$ .

- 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 states to change color more smoothly as time is changed.

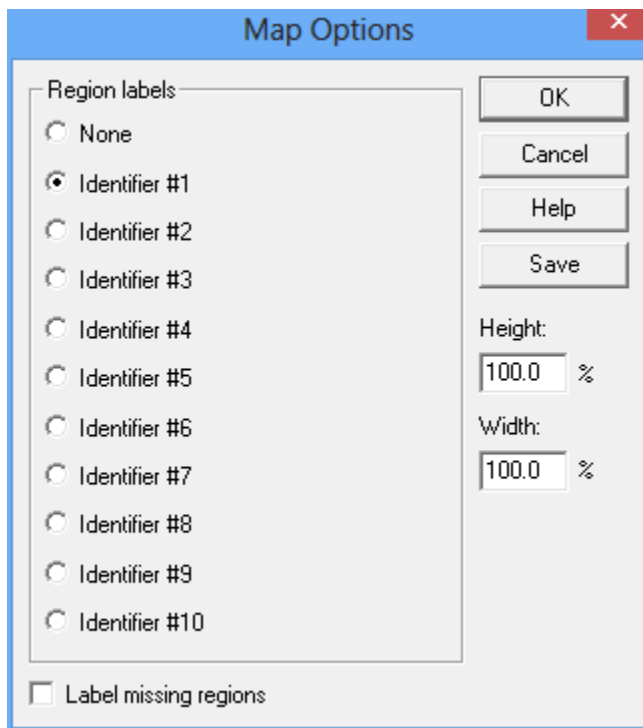
The plot below shows the output for the 1995, changing the limits on which the colors are based.



In 1995, Florida, Arizona and the District of Columbia had the highest crime rates, while the crime rate was lowest in West Virginia and New Hampshire.

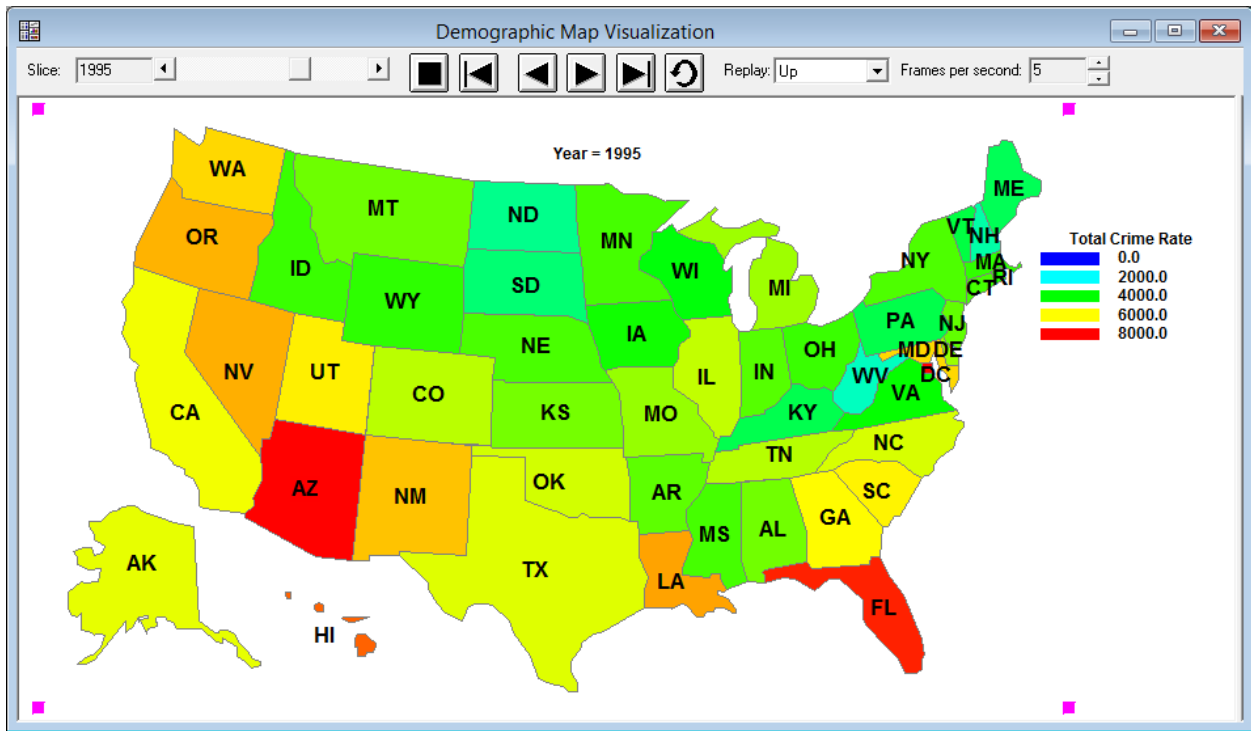
### Pane Options

Labels may be added to each region using the *Pane Options* dialog box:



- **Region labels:** identifier to be used to label each region on the graph. Identifiers 1 through 10 correspond to the labels in the header record of the BNA file for each region.
- **Label missing regions:** whether to include labels for regions that have no data.
- **Height:** height of map relative to the available plotting area. This value may be changed to give a proper aspect ratio for the area being plotted.
- **Width:** width of map relative to the available plotting area. This value may be changed to give a proper aspect ratio for the area being plotted.
- **Save:** saves the current label positions and makes them the default positions for future maps. The positions are saved in a file with the same name as the boundary file, except that the file extension is ".bnp" instead of ".bna" or "shq" instead of "shp". You must have write access to the directory containing the boundary file in order to save the positions.

The graph below shows 2-letter identifiers for each state, after repositioning some of the labels:



## 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:

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

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

$$y_t = \frac{-y_{t-2s} + 3y_{t-s} + y_{t+s}}{3} \quad (2)$$

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

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

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

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

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

$$y_t = \frac{y_{t-s} + 3y_{t+s} - y_{t+2s}}{3} \quad (5)$$

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:

$$y_t = -y_{t-2s} + 2y_{t-s} \quad (6)$$

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:

$$y_t = \frac{y_{t-2s} + 2y_{t+s}}{3} \quad (7)$$

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:

$$y_t = \frac{y_{t-2s} + y_{t+2s}}{2} \quad (8)$$

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:

$$y_t = \frac{y_{t-s} + y_{t+s}}{2} \quad (9)$$

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:

$$y_t = \frac{2y_{t-s} + y_{t+2s}}{3} \quad (10)$$

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:

$$y_t = 2y_{t+s} - y_{t+2s} \quad (11)$$

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.



## Boundary Files

### BNA files

A definition of the BNA file format may be found at:

[http://www.softwright.com/faq/support/boundary\\_file\\_bna\\_format.html](http://www.softwright.com/faq/support/boundary_file_bna_format.html)

The file named *US\_States.bna* contains boundary definitions for the 50 states in the United States plus the District of Columbia. When creating a data file, you may use either of the 2 identifiers shown in the table below.

#### *State Names and Abbreviations*

Alabama	AL
Alaska	AK
Arizona	AZ
Arkansas	AR
California	CA
Colorado	CO
Connecticut	CT
Delaware	DE
District of Columbia	DC
Florida	FL
Georgia	GA
Hawaii	HI
Idaho	ID
Illinois	IL
Indiana	IN
Iowa	IA
Kansas	KS
Kentucky	KY
Louisiana	LA
Maine	ME
Maryland	MD
Massachusetts	MA
Michigan	MI
Minnesota	MN
Mississippi	MS
Missouri	MO
Montana	MT
Nebraska	NE
Nevada	NV
New Hampshire	NH
New Jersey	NJ
New Mexico	NM

New York	NY
North Carolina	NC
North Dakota	ND
Ohio	OH
Oklahoma	OK
Oregon	OR
Pennsylvania	PA
Rhode Island	RI
South Carolina	SC
South Dakota	SD
Tennessee	TN
Texas	TX
Utah	UT
Vermont	VT
Virginia	VA
Washington	WA
West Virginia	WV
Wisconsin	WI
Wyoming	WY

The file named *Departements\_Metropole.bna* contains boundary definitions for the 95 departments in France. When creating a data file, you may use either of the 2 identifiers shown in the table below. (Note: the file is © 2013 by IGN - GEOFLA ([www.ign.fr](http://www.ign.fr)) and is distributed with their permission.)

01	AIN
02	AISNE
03	ALLIER
04	ALPES-DE-HAUTE-PROVENCE
05	HAUTES-ALPES
06	ALPES-MARITIMES
07	ARDECHE
08	ARDENNES
09	ARIEGE
10	AUBE
11	AUDE
12	AVEYRON
13	BOUCHES-DU-RHONE
14	CALVADOS
15	CANTAL
16	CHARENTE
17	CHARENTE-MARITIME
18	CHER
19	CORREZE
21	COTE-D'OR
22	COTES-D'ARMOR

23	CREUSE
24	DORDOGNE
25	DOUBS
26	DROME
27	EURE
28	EURE-ET-LOIR
29	FINISTERE
2A	CORSE-DU-SUD
2B	HAUTE-CORSE
30	GARD
31	HAUTE-GARONNE
32	GERS
33	GIRONDE
34	HERAULT
35	ILLE-ET-VILAINE
36	INDRE
37	INDRE-ET-LOIRE
38	ISERE
39	JURA
40	LANDES
41	LOIR-ET-CHER
42	LOIRE
43	HAUTE-LOIRE
44	LOIRE-ATLANTIQUE
45	LOIRET
46	LOT
47	LOT-ET-GARONNE
48	LOZERE
49	MAINE-ET-LOIRE
50	MANCHE
51	MARNE
52	HAUTE-MARNE
53	MAYENNE
54	MEURTHE-ET-MOSELLE
55	MEUSE
56	MORBIHAN
57	MOSELLE
58	NIEVRE
59	NORD
60	OISE
61	ORNE
62	PAS-DE-CALAIS
63	PUY-DE-DOME
64	PYRENEES-ATLANTIQUES
65	HAUTES-PYRENEES
66	PYRENEES-ORIENTALES

67	BAS-RHIN
68	HAUT-RHIN
69	RHONE
70	HAUTE-SAONE
71	SAONE-ET-LOIRE
72	SARTHE
73	SAVOIE
74	HAUTE-SAVOIE
75	PARIS
76	SEINE-MARITIME
77	SEINE-ET-MARNE
78	YVELINES
79	DEUX-SEVRES
80	SOMME
81	TARN
82	TARN-ET-GARONNE
83	VAR
84	VAUCLUSE
85	VENDEE
86	VIENNE
87	HAUTE-VIENNE
88	VOSGES
89	YONNE
90	TERRITOIRE DE BELFORT
91	ESSONNE
92	HAUTS-DE-SEINE
93	SEINE-SAINT-DENIS
94	VAL-DE-MARNE
95	VAL-D'OISE

Also included with the program is a data file named *France\_Dept\_Data.sgd* which contains population, employment, and unemployment statistics for each department in France. The file is ©2012 by INSEE ([www.insee.fr](http://www.insee.fr)) and is distributed with their permission.

### *U.S. County Boundaries*

BNA files with county boundaries for each state in the United States may be found at:

Center for International Earth Science Information Network - CIESIN. 1996. Archive of Census Related Products (ACRP): 1992 Boundary Files. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <http://sedac.ciesin.columbia.edu/data/set/acrp-boundary-1992>.

**Note:** You may use the *Demographic Map* procedure on the *Plot* menu to create an empty map which displays all of the regions in a specific BNA file.



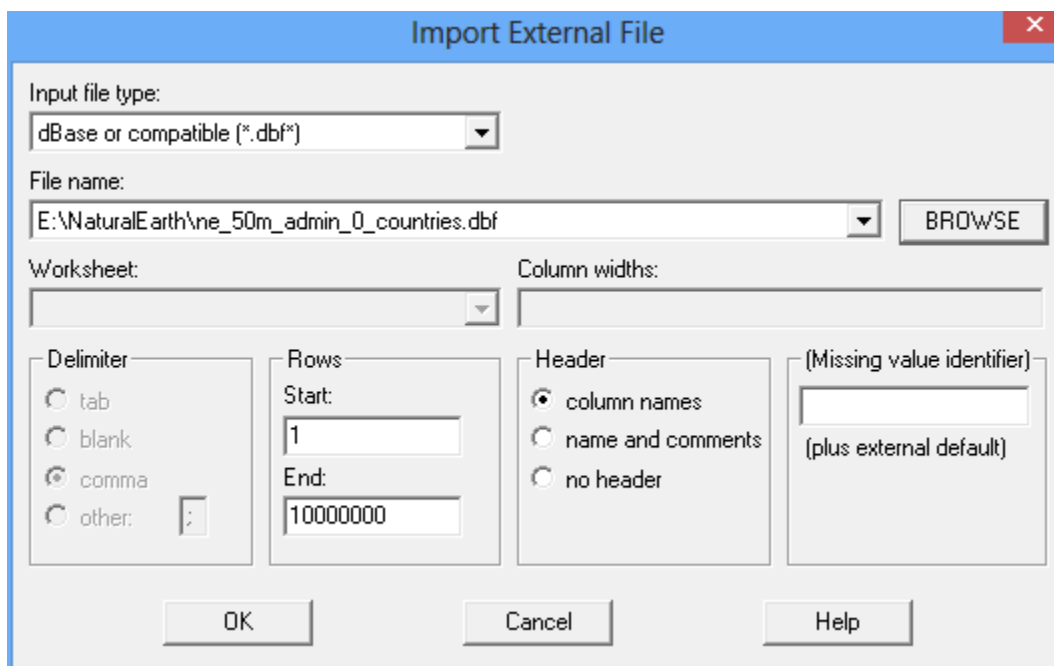
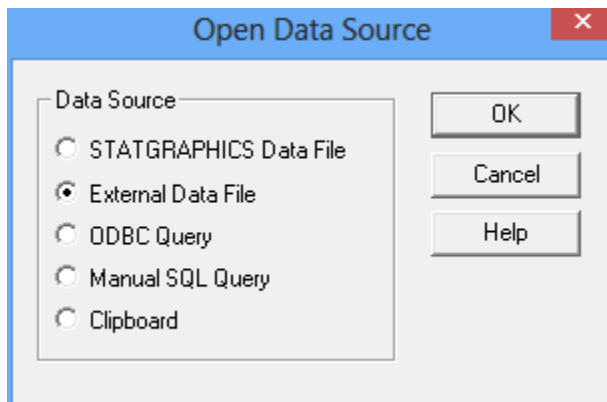
## SHP files

The shapefile format is a popular geospatial vector data format developed and regulated by ESRI. To create a map, you must have at least 2 files with the same name but different extensions:

1. A file with the extension *.shp* which contains the region boundaries.
2. A file with the extension *.dbf* which identifies the regions.

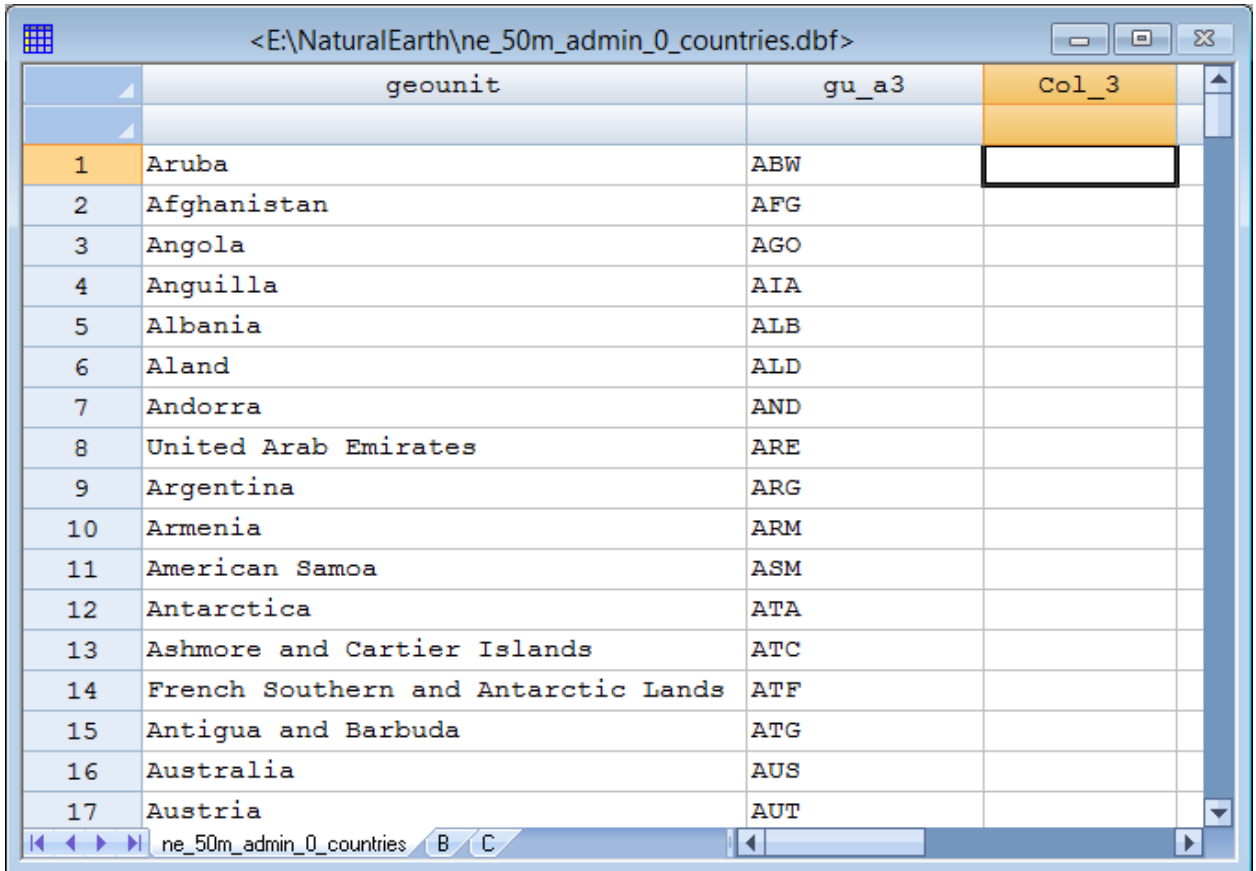
For example, the *worldmap* file supplied with Statgraphics was obtained from *naturalearthdata.com* by downloading the admin 0 countries file in medium format. The 2 relevant files are *ne\_50m\_admin\_0\_countries.shp* and *ne\_50m\_admin\_0\_countries.dbf*. The SHP file may be used without any changes. The DBF file, which contains the region identifiers, needs to be converted to a standard Statgraphics data file as follows:

*Step 1:* Open the DBF by selecting *File – Open – Open Data Source* from the main Statgraphics menu. Fill in the dialog boxes as shown below:



This loads the file into a Statgraphics datasheet.

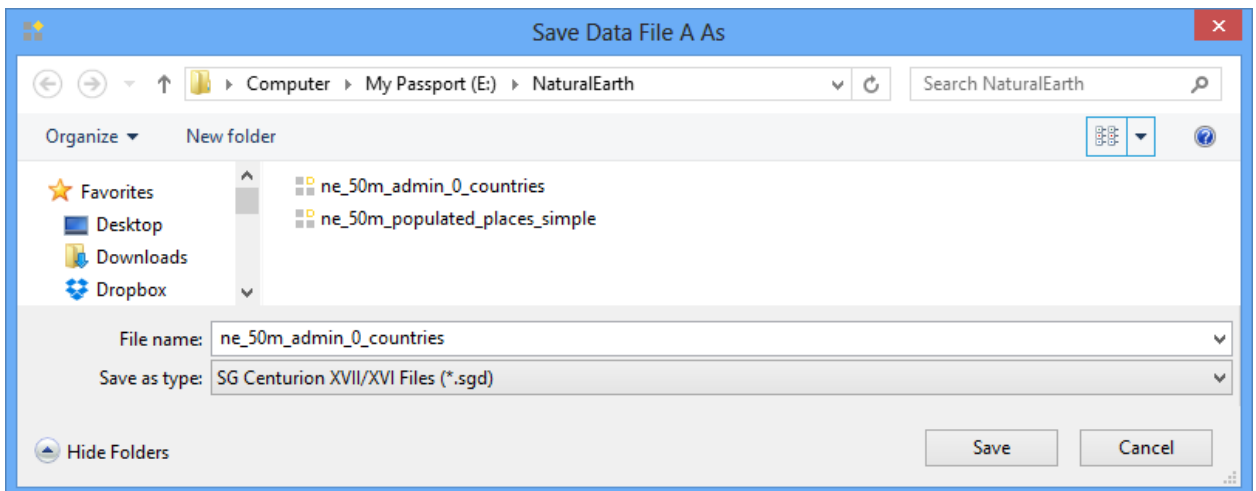
*Step 2:* Delete all columns except those needed to identify the regions:



	geounit	gu_a3	Col_3
1	Aruba	ABW	
2	Afghanistan	AFG	
3	Angola	AGO	
4	Anguilla	AIA	
5	Albania	ALB	
6	Aland	ALD	
7	Andorra	AND	
8	United Arab Emirates	ARE	
9	Argentina	ARG	
10	Armenia	ARM	
11	American Samoa	ASM	
12	Antarctica	ATA	
13	Ashmore and Cartier Islands	ATC	
14	French Southern and Antarctic Lands	ATF	
15	Antigua and Barbuda	ATG	
16	Australia	AUS	
17	Austria	AUT	

You may use up to 10 columns as alternative identifiers.

*Step 3:* Save the data file with the same name as the SHP file but with an SGD extension:



When constructing a data file to be displayed on the map, use any of the columns in the saved SGD file as your identifier column. That will match your data with the proper regions.