

***Heat Map***



Revised: 10/9/2017



Summary .....	1
Sample Data .....	2
Data Input.....	3
Analysis Summary .....	6
Heat Map.....	7
Analysis Options .....	8
Calculations.....	10

**Summary**

The **Heat Map** procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent values of the quantitative variable.

**Sample StatFolio:** *heatmap.sgp*

## Sample Data

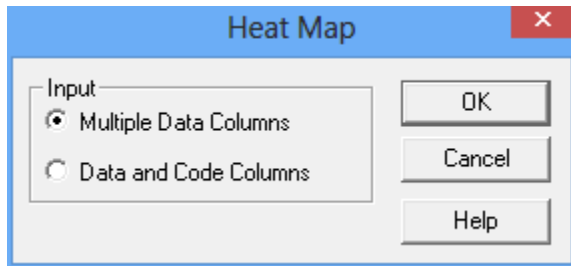
The file *crimerates.sgd* contains data for each state in the United States and the District of Columbia over 46 years (1965-2010). The data were obtained from the FBI's Uniform Crime Reporting (UCR) program. Several rows and columns of the file are shown below:

State	Year	Population	Total Crime Rate per 100,000 population	Burglary per 100,000 population	Larceny - Theft per 100,000 population	Motor Vehicle Theft per 100,000 population
Alaska	1965	253000	2603.6	554.5	1492.9	407.1
Alaska	1966	272000	2785.7	593.0	1600.7	441.5
Alaska	1967	272000	2884.6	688.6	1630.9	404.4
Alaska	1968	277000	3320.9	747.3	1915.9	482.3
Alaska	1969	282000	3880.1	870.6	2196.8	591.5
Alaska	1970	302173	3801.8	789.9	2182.5	551.3
Alaska	1971	313000	4164.5	848.6	2438.0	522.7
Alaska	1972	325000	4478.5	970.8	2639.1	498.2
...	...	...	...	...	...	...

## Data Input

### Data Input Dialog Box

Data to be displayed by the *Heat Map* may be arranged in either of 2 ways. When selected from the main menu, the procedure first displays a dialog box which specifies how the data are structured:



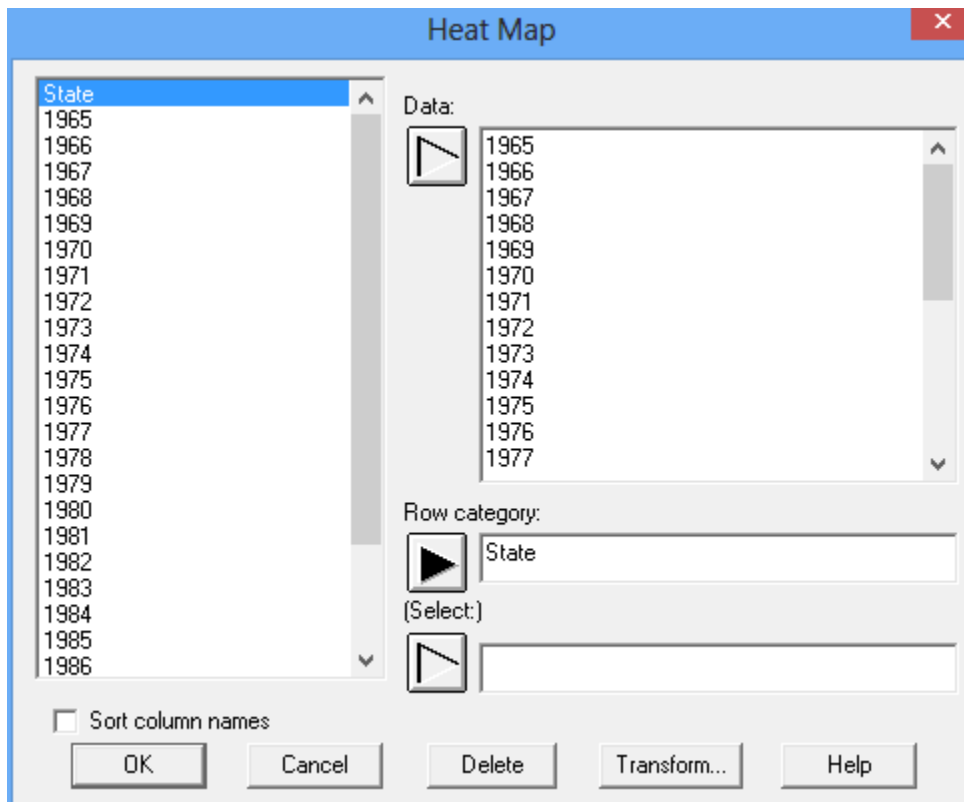
1. **Multiple Data Columns:** The data for each time period are placed in a separate column, as in the file below:

total crime by year.sgd									
	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

2. **Data and Code Columns:** All data are placed in a single data column, as with the crime data file displayed earlier. Additional columns are then constructed to identify the 2 categorical factors. This format allows for more than one variable to be stored in the same data file.

### Multiple Data Columns

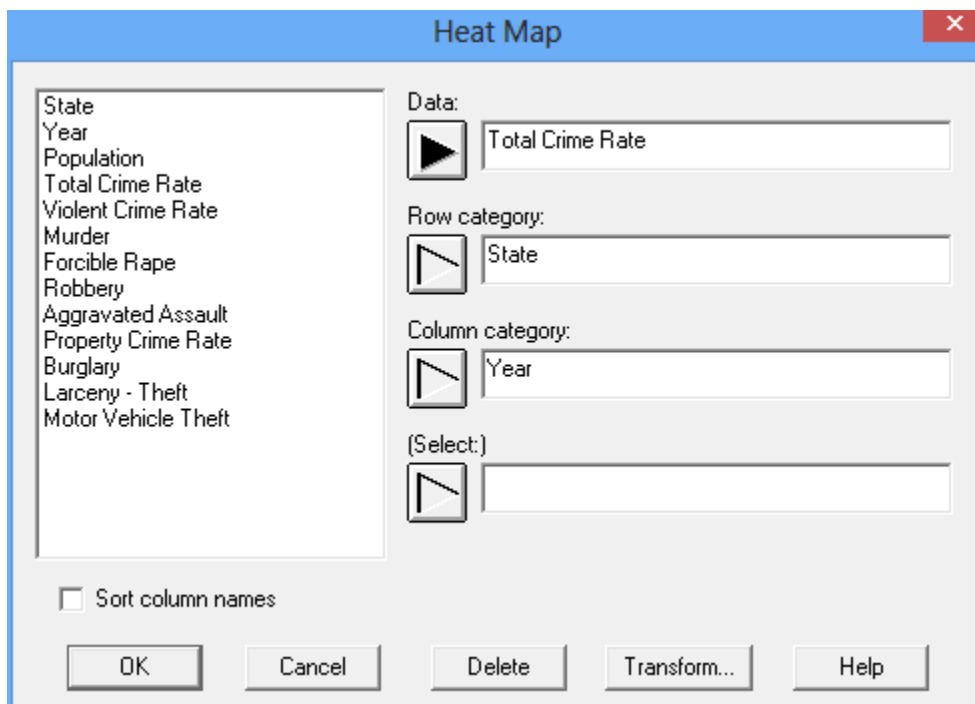
After specifying the structure of the data, a second data input dialog box requests the names of the columns containing the data values to be analyzed. For structure #1, the dialog box requests the names of multiple data column:



- **Data:** names of 2 or more numeric columns containing the observations to be analyzed.
- **Row category:** name of the column used to define the rows of the heat map. The variable may be either numeric or non-numeric.
- **Select:** optional subset selection.

### Data and Code Columns

For structure #2, the data input dialog box has the following form:



- **Data:** name of a single numeric column containing the observations to be analyzed.
- **Row category:** name of the data column used to define the rows of the heat map. The variable may be either numeric or non-numeric.
- **Column category:** name of the data column used to define the columns of the heat map. The variable may be either numeric or non-numeric.
- **Select:** optional subset selection.

Each row of the data file represents one combination of the row and column categories.

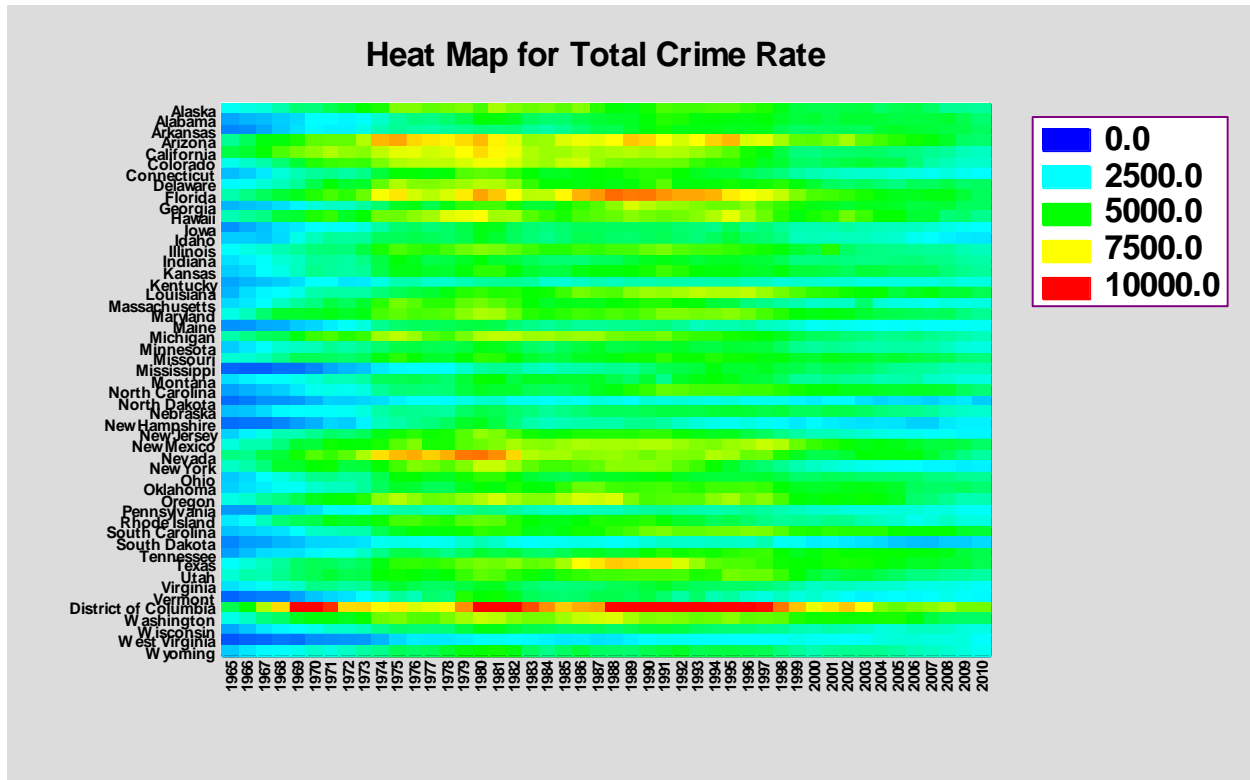
## Analysis Summary

The Analysis Summary displays a table containing the data to be plotted. A portion of the table is shown below:

<b>Heat Map - Total Crime Rate</b>											
Data variable: Total Crime Rate (per 100,000 population)											
Row variable: State											
Column variable: Year											
Total Crime Rate											
State	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
Alaska	2603.6	2785.7	2884.6	3320.9	3880.1	3801.8	4164.5	4478.5	4943.3	5239.8	6196.6
Alabama	1592.5	1758.3	1851	1999	2126.6	2479.5	2498.4	2394.5	2582.3	3000.1	3472.5
Arkansas	1274.2	1382.9	1628.5	1958.5	2188.5	2421.2	2328.2	2352.4	2756.5	3300.7	3540.1
Arizona	3547.8	4135.8	4837.9	4874.4	5224.6	5914.2	5941.5	5933.3	6703.9	8221.7	8341.5
California	4319.4	4549.4	5055.1	5721.1	6099.7	6339.1	6690.1	6413.1	6304.9	6846.8	7204.6
Colorado	2704.5	3009.6	3309.1	3862.6	4498.2	5318.2	5517	5593.6	5495.8	6165.8	6675.5
Connecticut	1834.4	1982.3	2281.2	2890.4	3225.5	3489.4	3646.2	3403.1	3664.4	4407	4957
Delaware	2408.7	2619.7	2891.6	3165.5	3501.7	4263.1	5015.1	4523.7	4582.6	5949.6	6668.2
Florida	3320.2	3716.3	4103.6	4498.5	4742.5	5317.2	5673	5376.9	5960.3	7387.3	7721.2
Georgia	1764.3	1879.8	1954.2	2156.1	2399.5	2881.6	3047.9	3051.8	3430.3	3912.4	4625.9
Hawaii	3252.3	3503.2	3719.4	4438.3	4532.9	5265.1	5458.8	4612.5	4958.8	6071.7	6026.6
Iowa	1397.6	1643.8	1881	2143.7	2252.5	2505.1	2625.6	2531.7	2831.6	3413.7	3908.7
Idaho	1982.5	2011.4	1905.7	2121.4	2728.3	3154.8	3507.5	3420.4	3457.8	4082.6	4141.1
Illinois	2545.9	2759.3	2910.5	3237.9	3468.5	3779.6	3885.3	3791.1	4324.9	5184.3	5703.5
Indiana	2170.5	2336.1	2592.5	2920.2	3089.4	3484.9	3522.7	3409.7	3726.7	4336.9	4911.4
Kansas	1990.2	2102	2474.5	2723.7	3138	3643.5	3495.2	3404.8	3513.8	4300.4	4747
Kentucky	1589.7	1738.4	1856.3	1991	2164.9	2484.1	2481.7	2233.6	2265.3	2759.7	3264.4
Louisiana	1771.3	2158.1	2388.8	2507.3	2752.1	3315.9	3456.2	3382.5	3402.9	3816.4	4123.4
Massachusetts	2121.4	2275.3	2482.1	3067.9	3478.6	3746.3	4347.6	4107.1	4521	5382.9	6077.8
Maryland	2759.4	3163.6	3910.4	4579.7	4573.3	4641	4799.2	4628.7	4791.4	5650.1	5907.5
Maine	1416.1	1427.2	1622.9	1682.1	1941.2	2060	2316.3	2590.6	2842.4	3600.2	3959.6
Michigan	3318.9	3708.3	4062	4170.6	4804.2	5507.4	5715.2	5363.6	5489.4	6519.6	6800.3
Minnesota	2011.4	2234.1	2593.2	2963.3	3076.6	3200.9	3536.9	3354.1	3535.6	3931	4298.7
Missouri	2912.3	2920.3	3123.7	3584.2	4179.2	4154.7	4045.6	3933.2	4141.4	4788.1	5397.8
Mississippi	977.2	897.2	900.2	1087.2	1094.6	1302.4	1600.5	1805.3	1926.3	2249.2	2410.7
Montana	2150.3	2356	2457.1	2542	2689.6	2882.9	3183.9	3205.3	3395.3	4083.8	4188.9
North Carolina	1568.4	1704.2	1892.3	2006.9	2251.7	2642.7	2729.2	2659.1	2811.9	3511.2	3816.7
North Dakota	1066.3	1207.1	1399.4	1405.1	1485.9	1706.5	2050.4	1987.8	2078.4	2160.1	2337.2
Nebraska	1831.8	1925.6	2092.8	2408	2451.3	2552.9	2538.9	2628.5	2811.2	3344.3	3614
...											

## Heat Map

The heat map uses a color gradient to represent the value of the data variable:

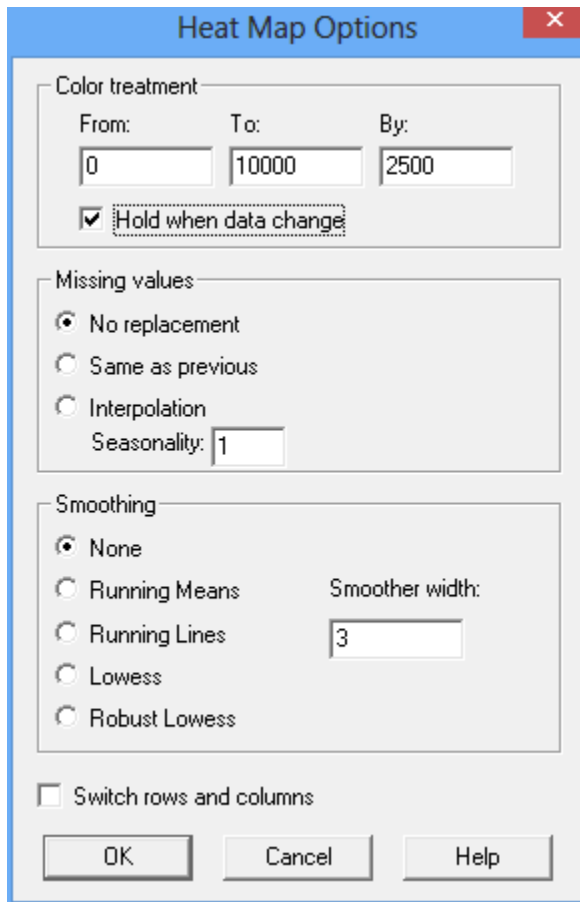


By default, the colors range from dark blue for very low values to dark red for very high values. The color palette may be changed by clicking the right mouse button on the plot and selecting *Graphics Options*.

In the plot above, notice the predominance of blue at the left of the plot when crime was low throughout the country. More red and yellow appears during the 80's and 90's, but disappears soon after the turn of the century. Note also that certain states contain most of the reds and yellows.

## Analysis Options

The *Analysis Options* dialog box lets users control various aspects of the output:



The following options are available:

- **Color treatment:** specifies the values used to define the color scale.
- **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:** smooths 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.



- **Switch rows and columns:** if checked, rows and columns are reversed.

## 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.