Multivariate Data Analysis Using Statgraphics Centurion: Part 2

Dr. Neil W. Polhemus
Statpoint Technologies, Inc.

Multivariate Statistical Methods

The simultaneous observation and analysis of more than one response variable.

*Primary Uses

1. Data reduction or structural simplification
2. Sorting and grouping
3. Investigation of the dependence among variables
4. Prediction
5. Hypothesis construction and testing

*Johnson and Wichern, Applied Multivariate Statistical Analysis

Data set: U.S. Demographics

us_state_data.sgd

	State	Education	Income per Capita	Urban Percentage	Population Density	Non-English Language	Median Age	Unemployment rate
		% completing college	2011 dollars	2000 census	per square mile	percentage	years	2012
1	Alabama	22.0	34879	55.4	87.6	0.7	35.8	7.3
2	Alaska	27.2	45665	65.7	1.1	2.4	32.4	7.0
3	Arizona	26.4	35061	88.2	45.2	5.6	34.2	8.3
4	Arkansas	19.6	33739	52.4	51.3	1.1	36	7.3
5	California	30.2	43646	94.5	217.1	9.6	33.3	10.5
6	Colorado	36.3	44053	84.5	41.5	3.4	34.3	8.0
7	Connecticut	35.7	57901	87.7	702.9	4.4	37.4	8.4
8	Delaware	28.0	41448	80.0	401.1	1.8	36	7.1
9	Florida	26.0	39635	89.3	296.4	5.9	38.7	8.6
10	Georgia	27.5	35979	71.7	141.4	2.4	33.4	9.0
11	Hawaii	29.5	42925	91.6	188.6	6.9	36.2	5.8
12	Idaho	24.6	32881	66.4	15.6	2.0	33.2	7.1
13	Illinois	30.7	43721	87.8	223.4	4.7	34.7	8.9
14	Indiana	22.7	35689	70.8	169.5	1.3	35.2	8.4
15	Iowa	24.9	41155	61.1	52.4	1.2	36.6	5.2
16	Kansas	29.7	40883	71.4	32.9	2.1	35.2	5.7
17	Kentucky	20.6	33988	55.7	101.7	0.7	35.9	8.2
18	Louisiana	21.1	38549	72.7	102.6	1.7	34	6.4
19	Maine	27.1	38299	40.2	41.3	1.4	38.6	7.3
20	Maryland	36.1	50655	86.1	541.9	2.4	36	6.8
21	Massachusetts	38.7	53471	91.4	809.8	4.7	36.5	6.7
22	Michigan	25.3	36264	74.7	175.0	1.7	35.5	9.1

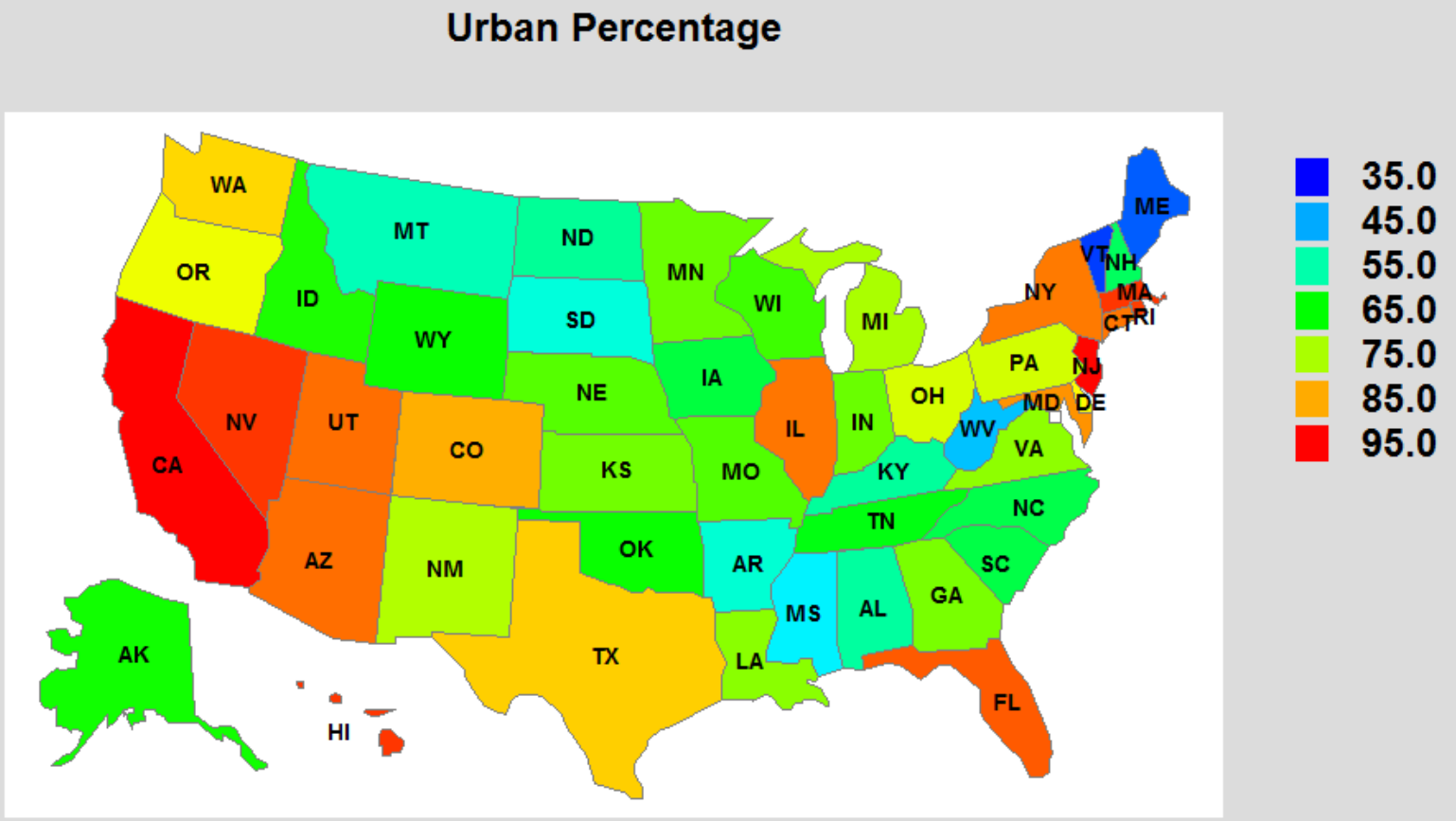
us_state_data B C

Variables

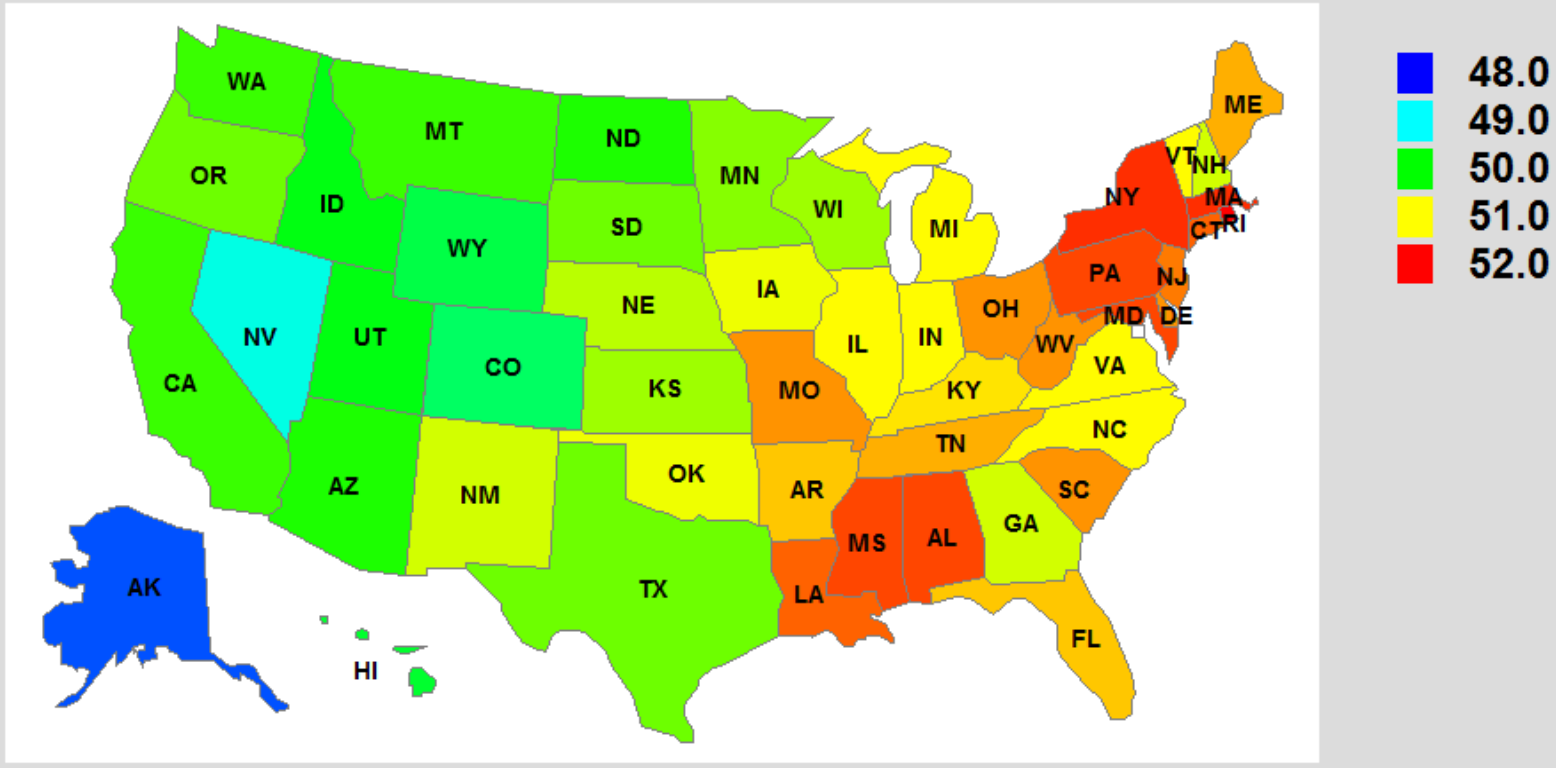
- *Education*: percent completing college
- *Income per Capita*: in 2011 dollars
- *Urban Percentage*: percent living in urban areas
- *Population Density*: people per square mile
- *Non-English Language*: percent that don't speak English
- *Median Age*: in years
- *Unemployment Rate*: for 2012
- *Percent Female*: of general population
- *Crime Rate*: violent crimes per 100,000
- *Manufacturing*: % of GDP
- *Infant Mortality*: 2006
- *Poverty Rate*: percentage below poverty level

- *Election 2012*: results of 2012 presidential election (Obama or Romney)

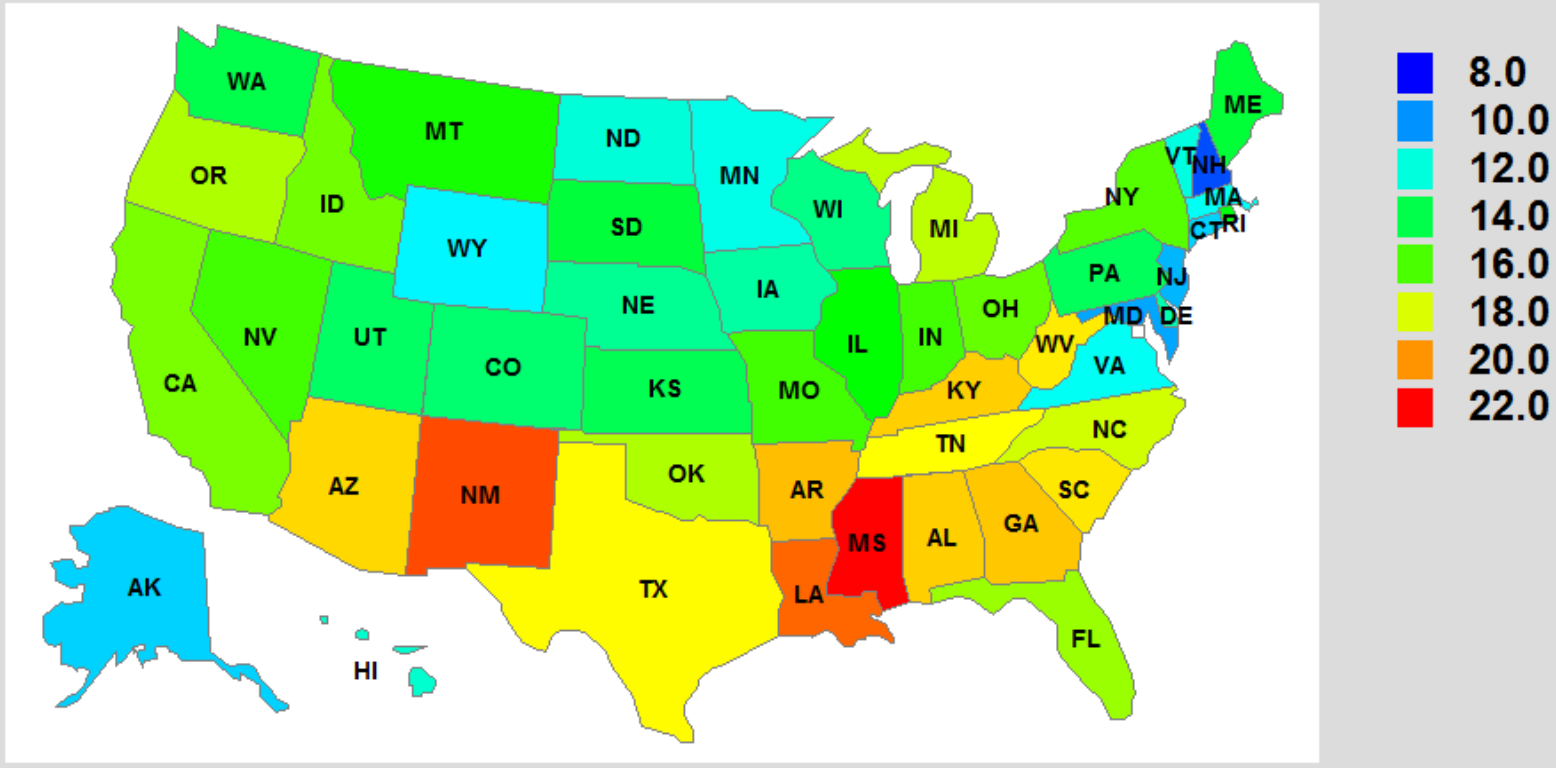
Urban Percentage



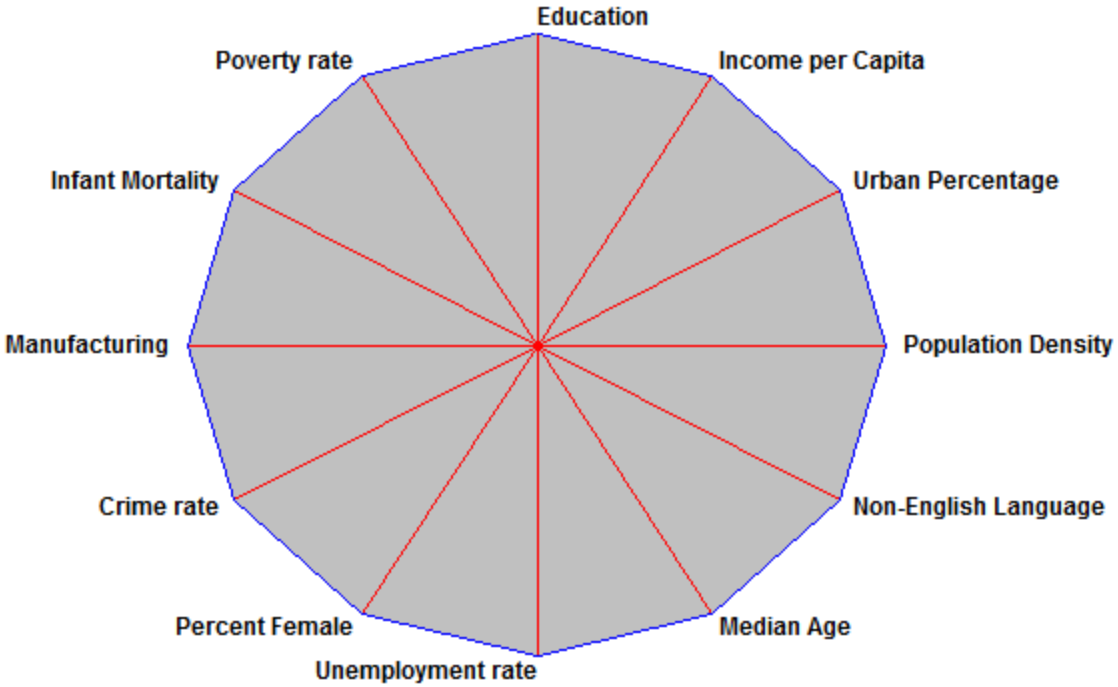
Percent Female



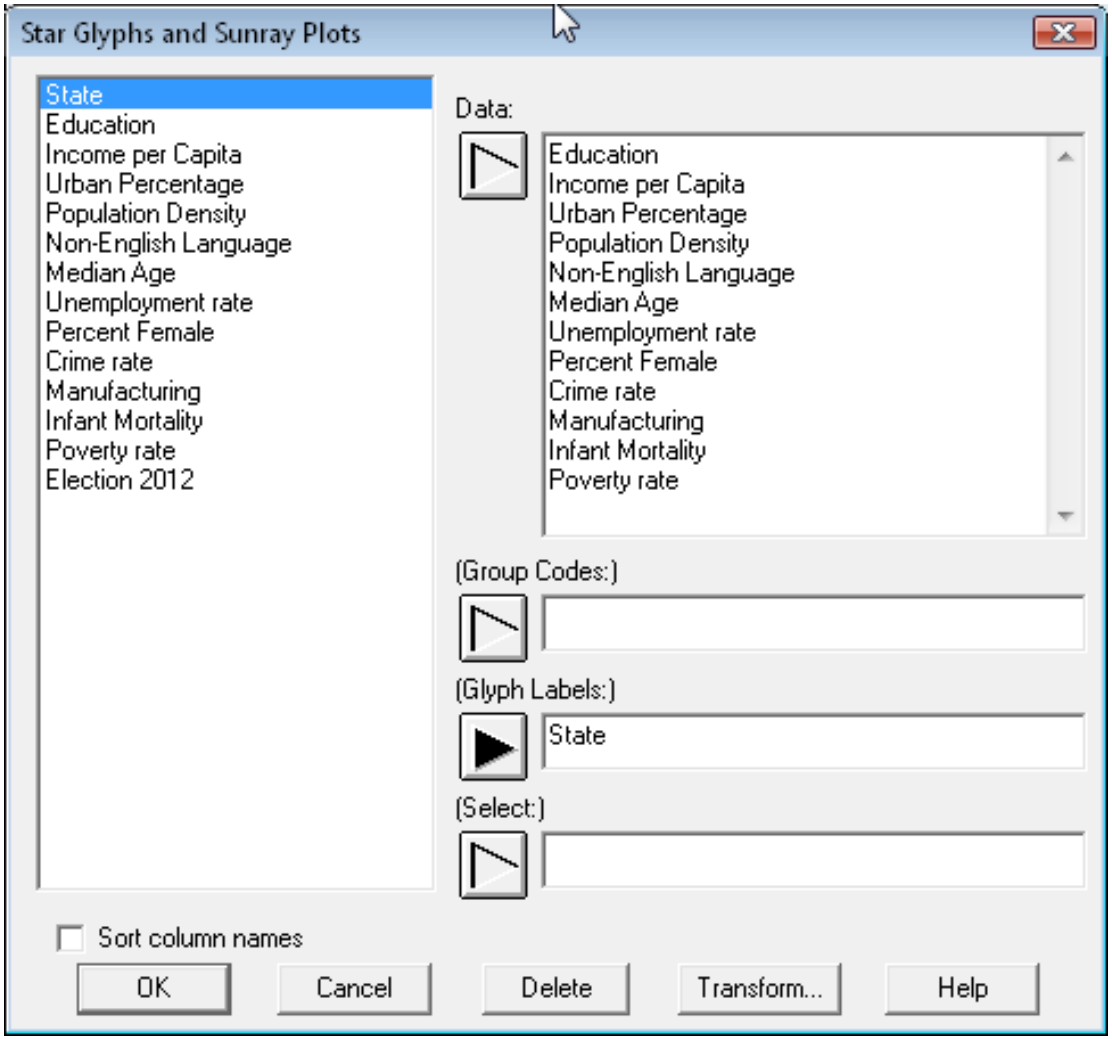
Poverty Rate

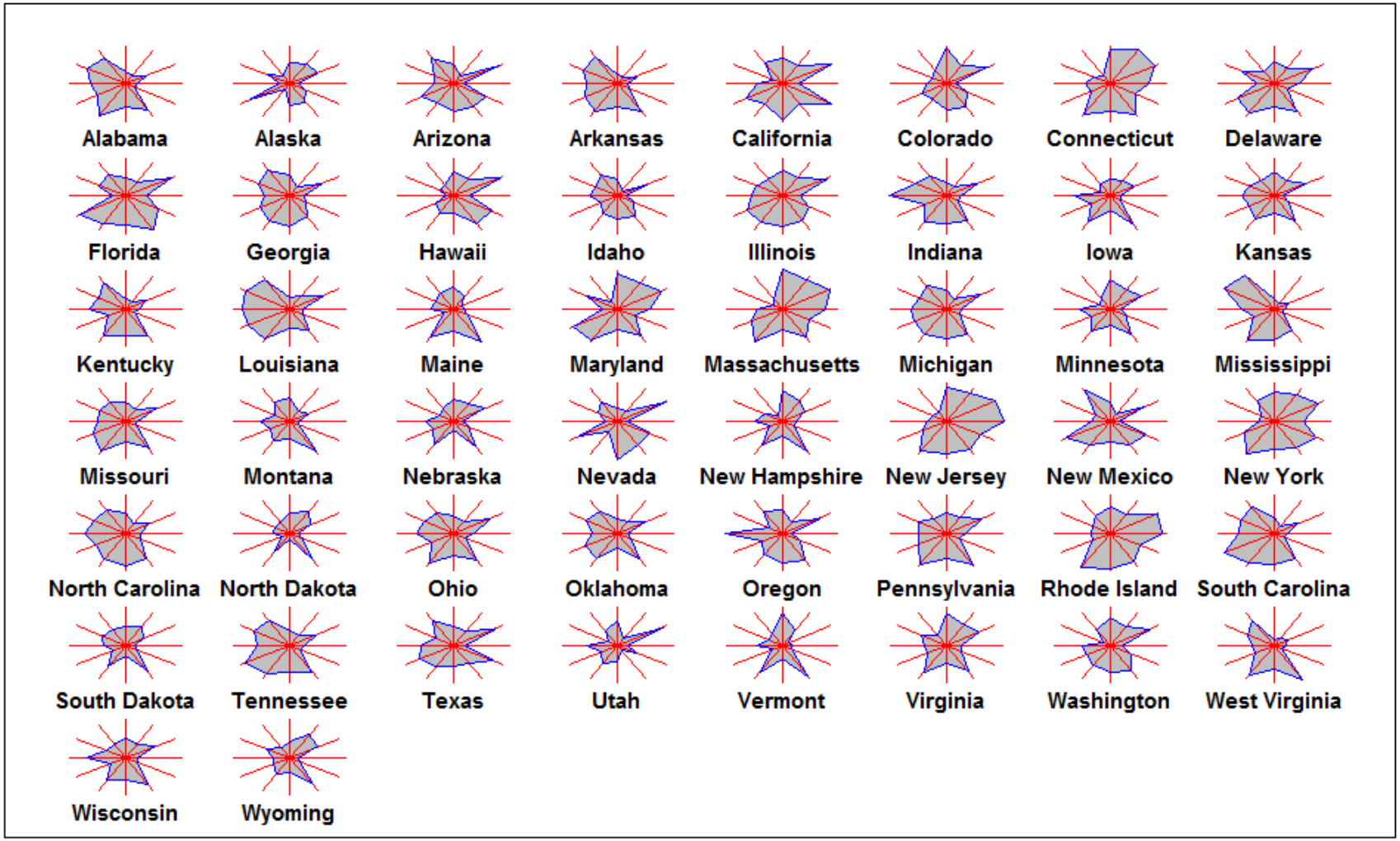


Star Glyphs


Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate

Data Input


Star Glyphs and Sunray Plots
State
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
Election 2012
Data:
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
(Group Codes:)
(Glyph Labels:)
State
(Select:)
Sort column names
OK
Cancel
Delete
Transform...
Help


Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

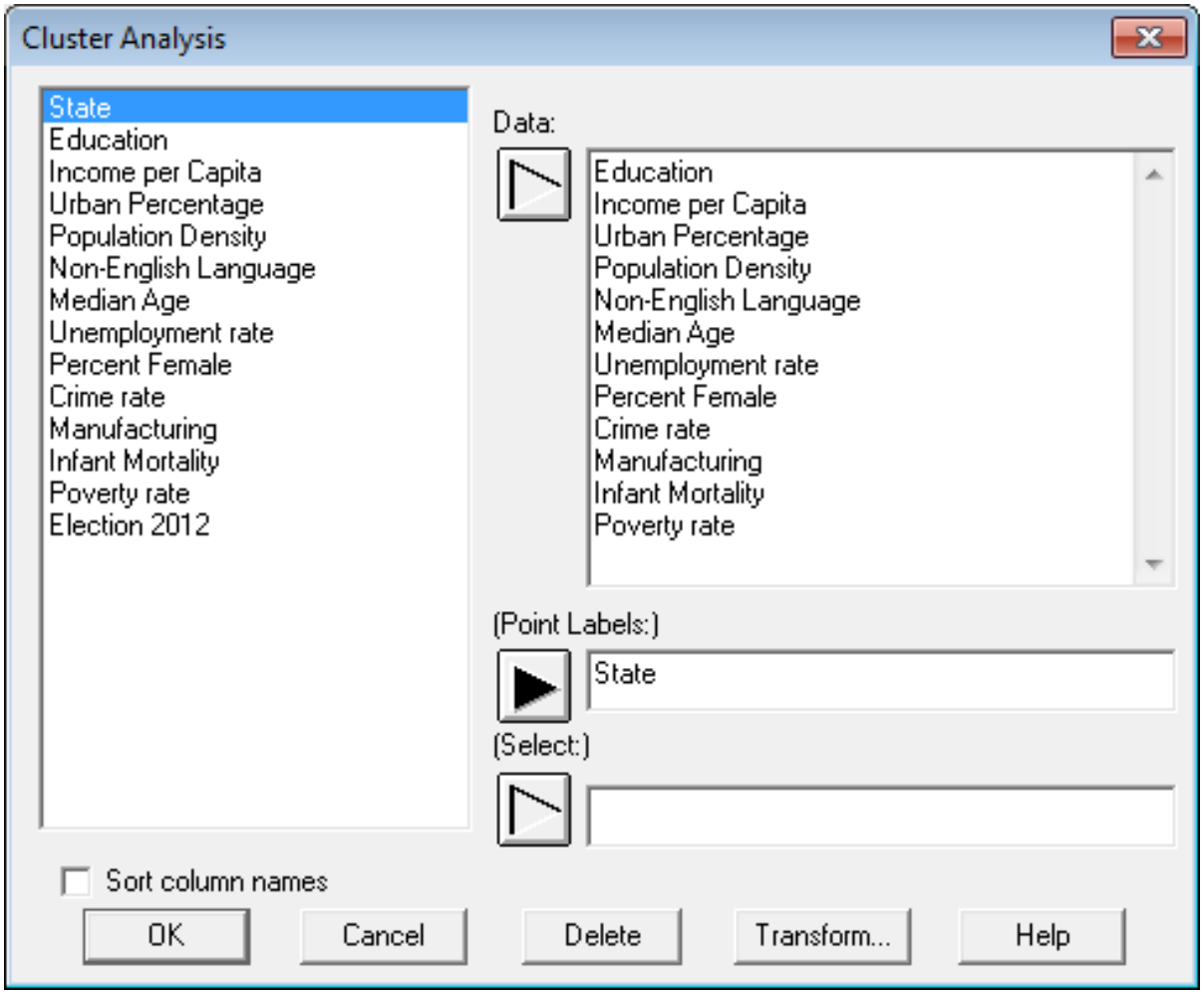
Cluster Analysis

Method for assigning objects to groups so that objects in the same group are more similar to each other than they are to objects in other groups.

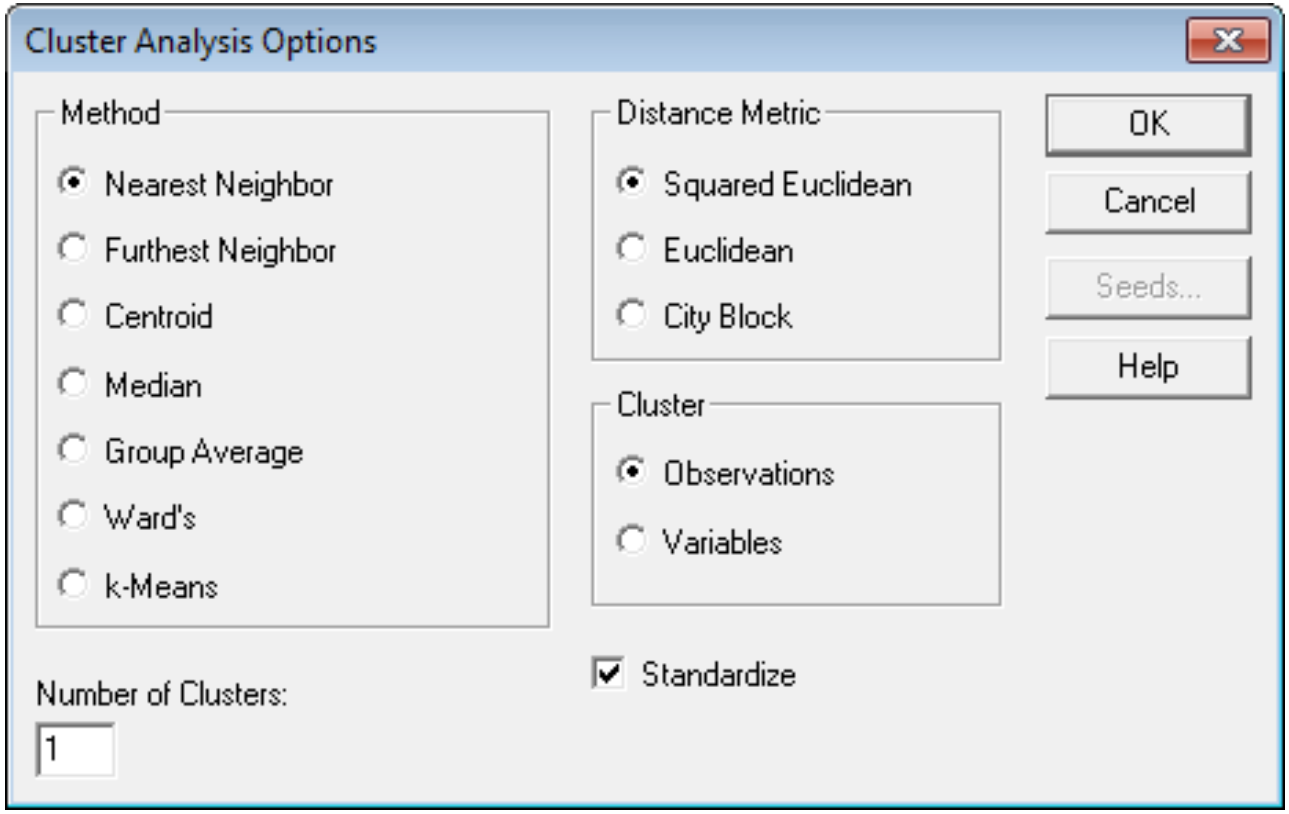
Some questions:

1. How many natural groups are there?
2. Which objects belong to each group?

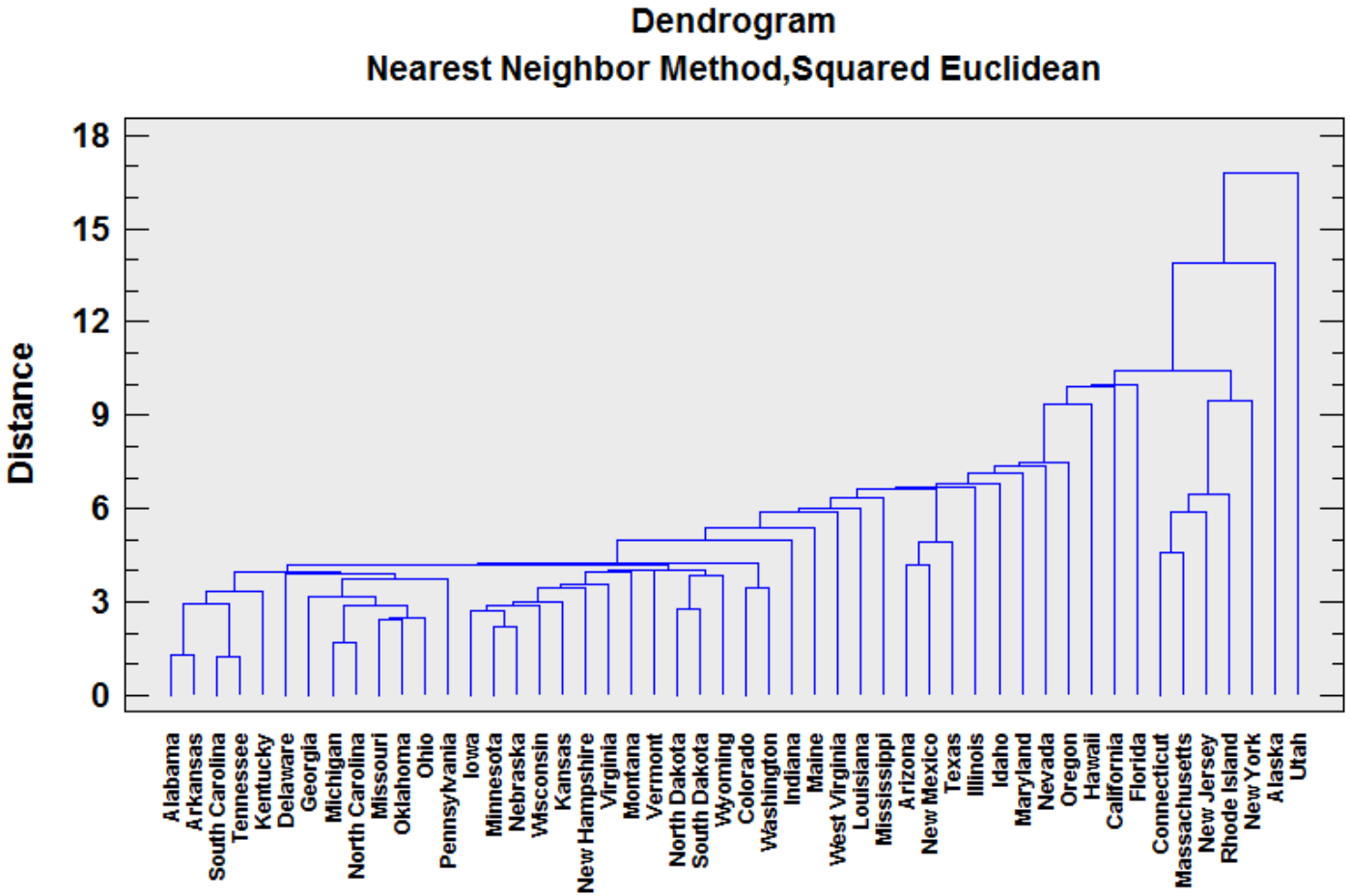
Data Input


Cluster Analysis
State
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
Election 2012
Data:
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
(Point Labels:)
State
(Select:)
Sort column names
OK
Cancel
Delete
Transform...
Help

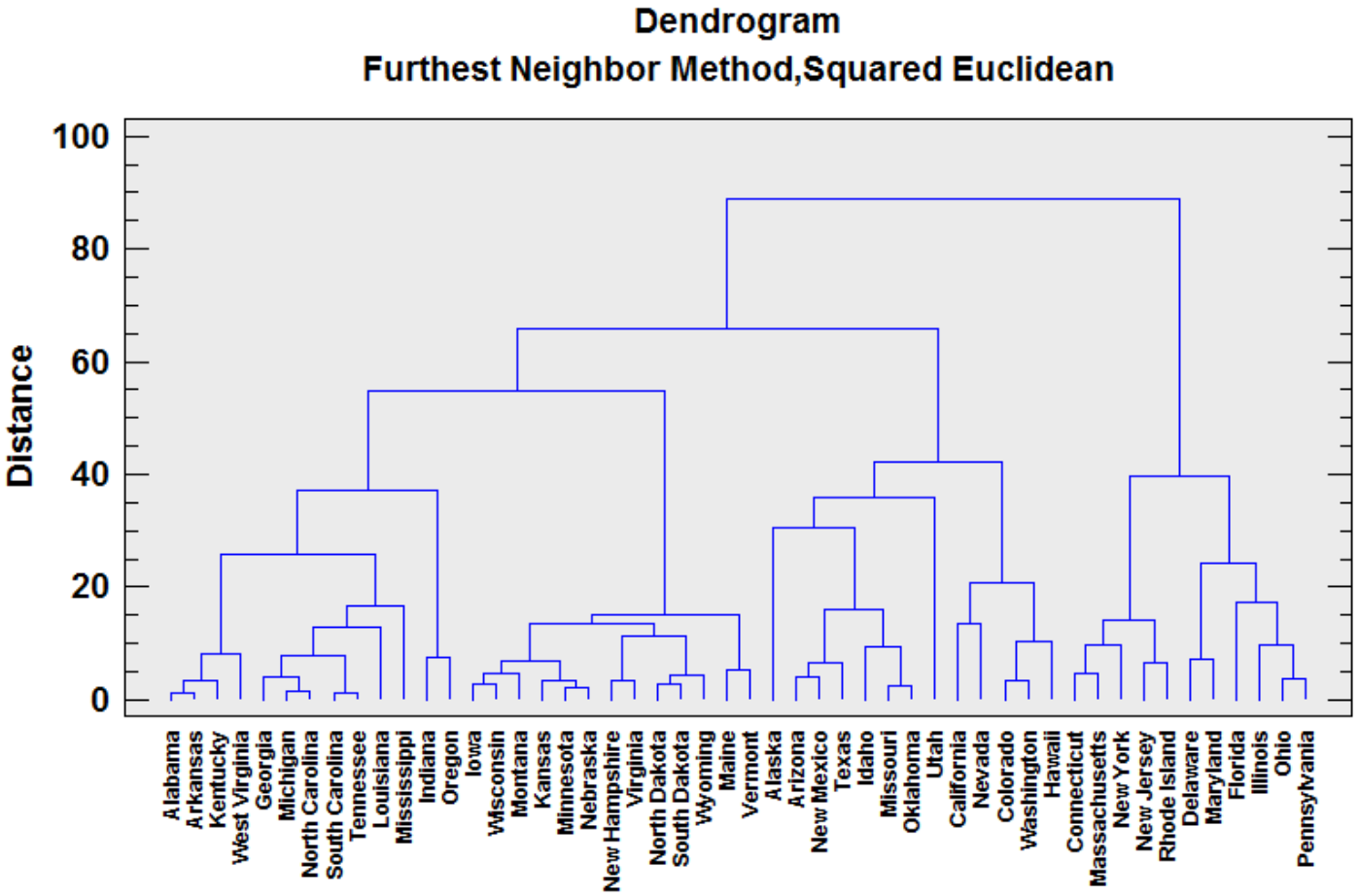
Analysis Options



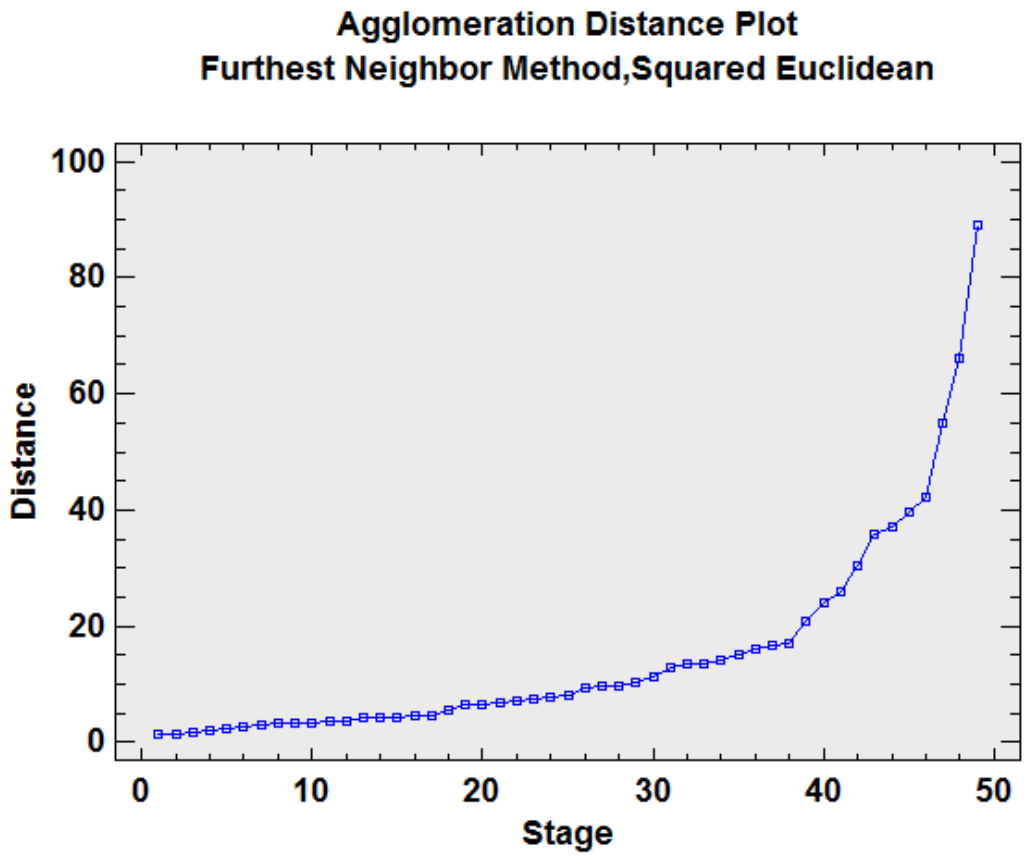
Dendrogram – Nearest Neighbor



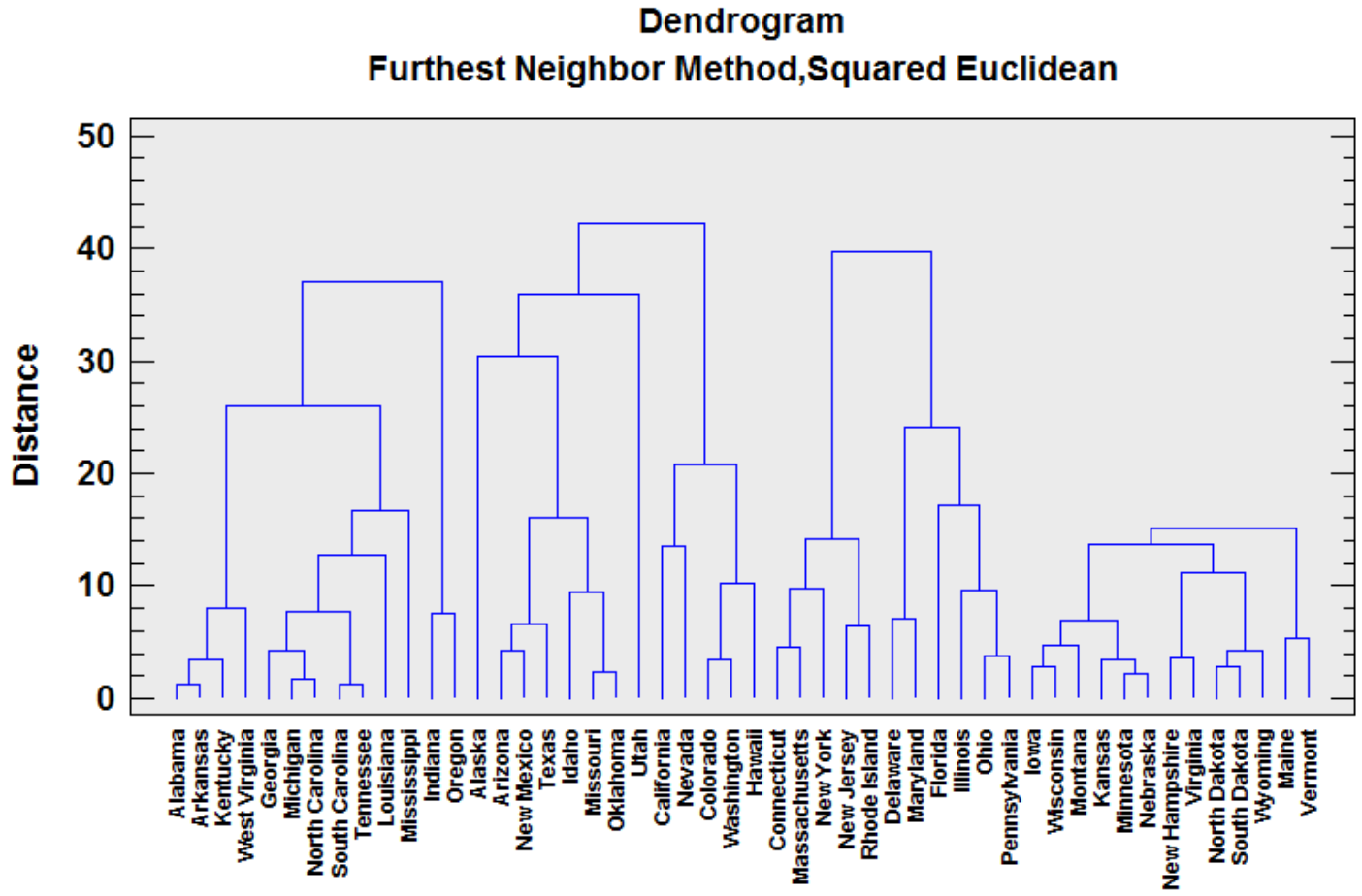
Dendrogram – Furthest Neighbor



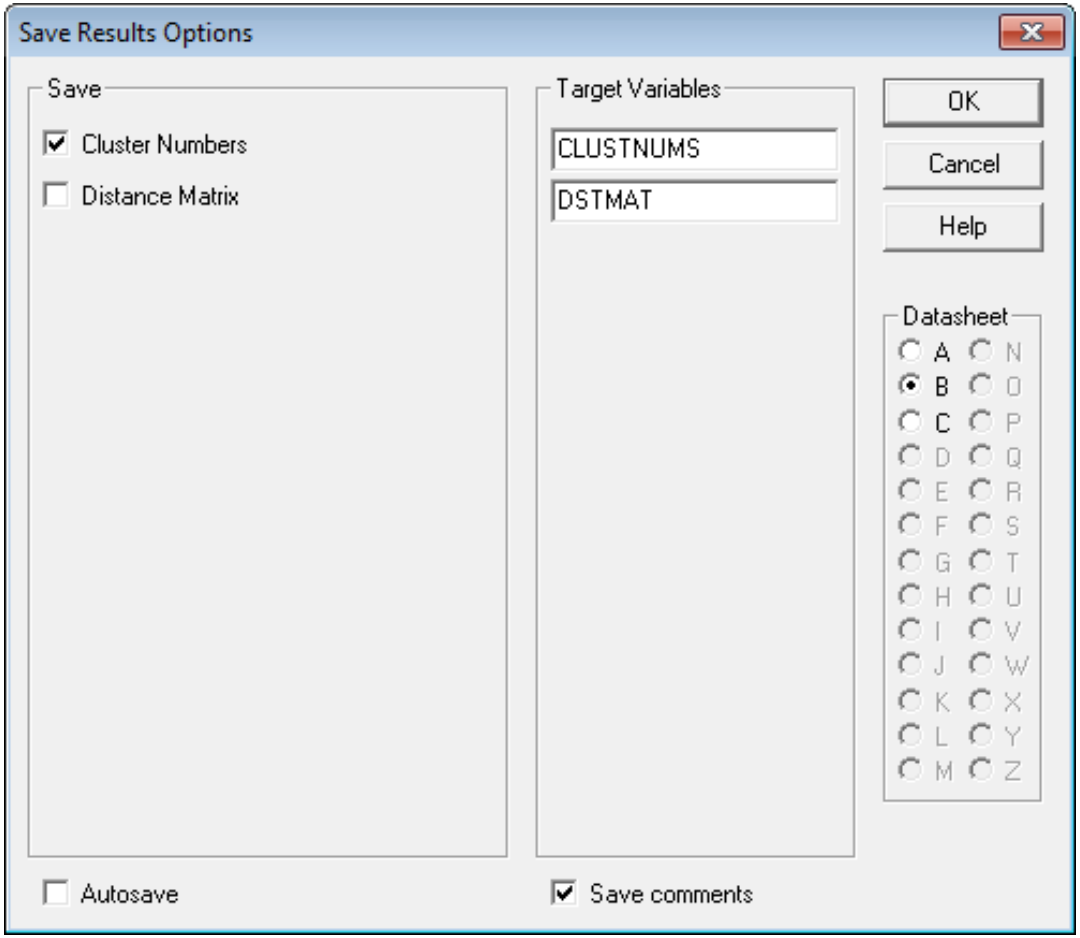
Agglomeration Distance Plot



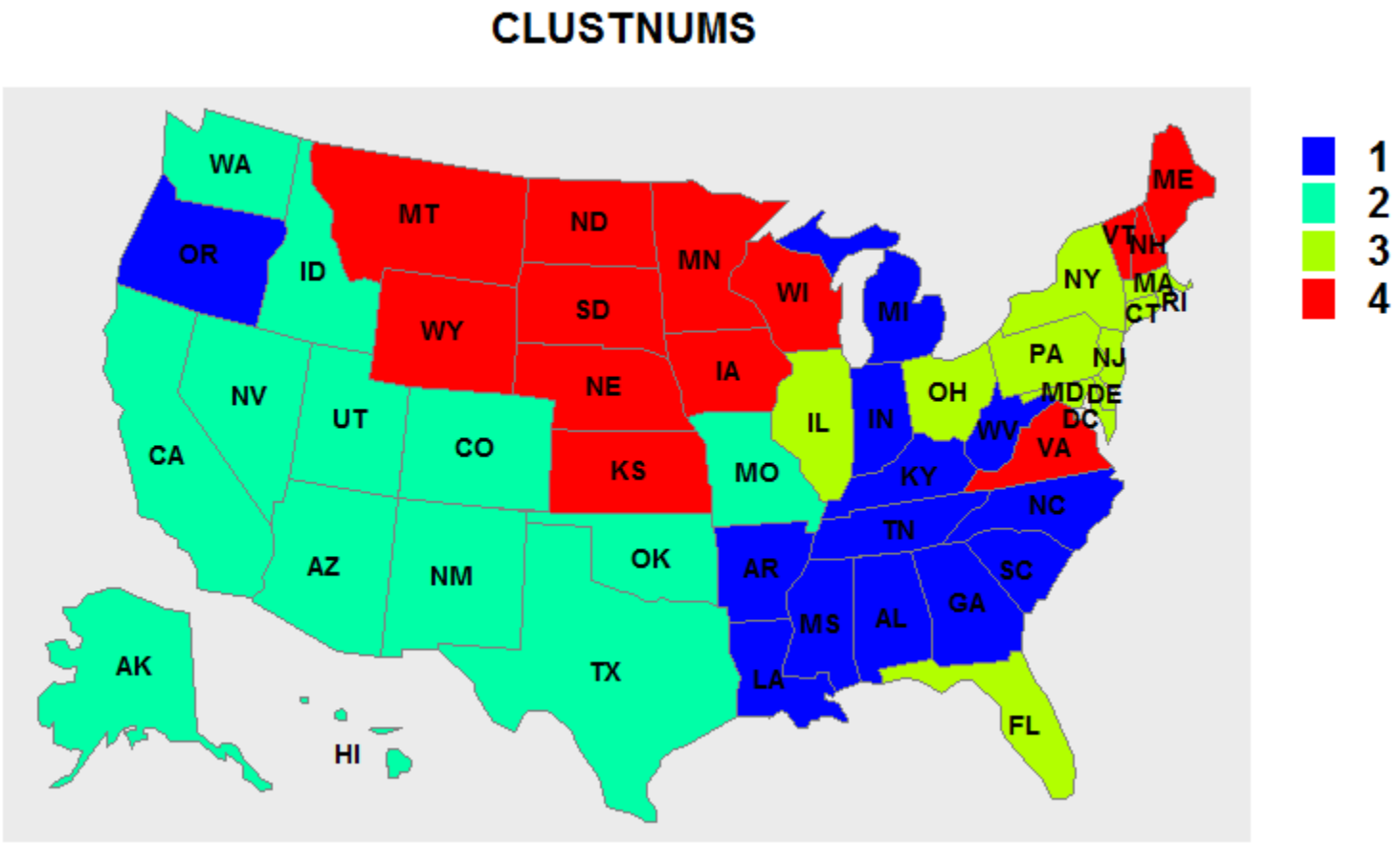
Dendrogram - 4 Clusters



Save Cluster Numbers



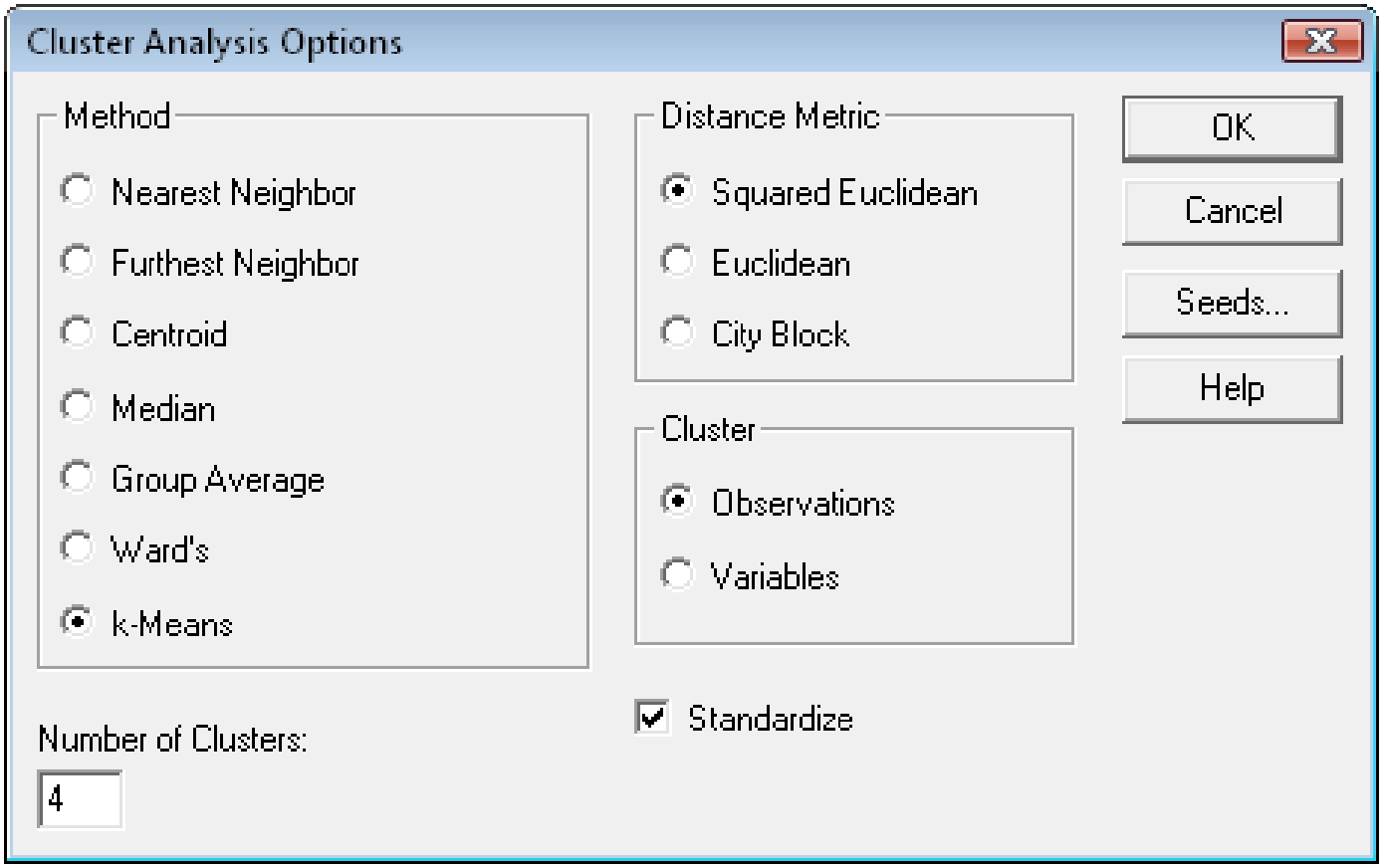
Map of Cluster Numbers



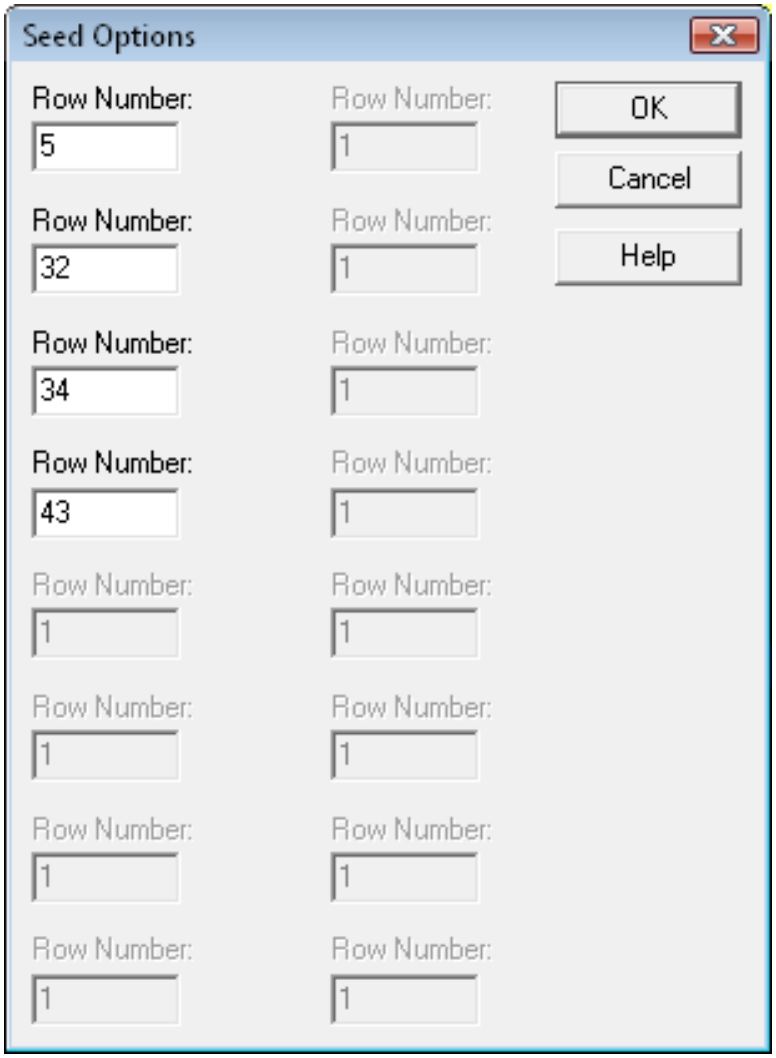
Method of k-Means

1. Starts with a set of typical objects (seeds) for each cluster.
2. Each object is assigned to group with nearest seed.
3. Centroids of groups are computed.
4. Objects are assigned to different groups if nearer to a different centroid.
5. Steps 3 and 4 are repeated until no change.

Method of k-Means



Seeds



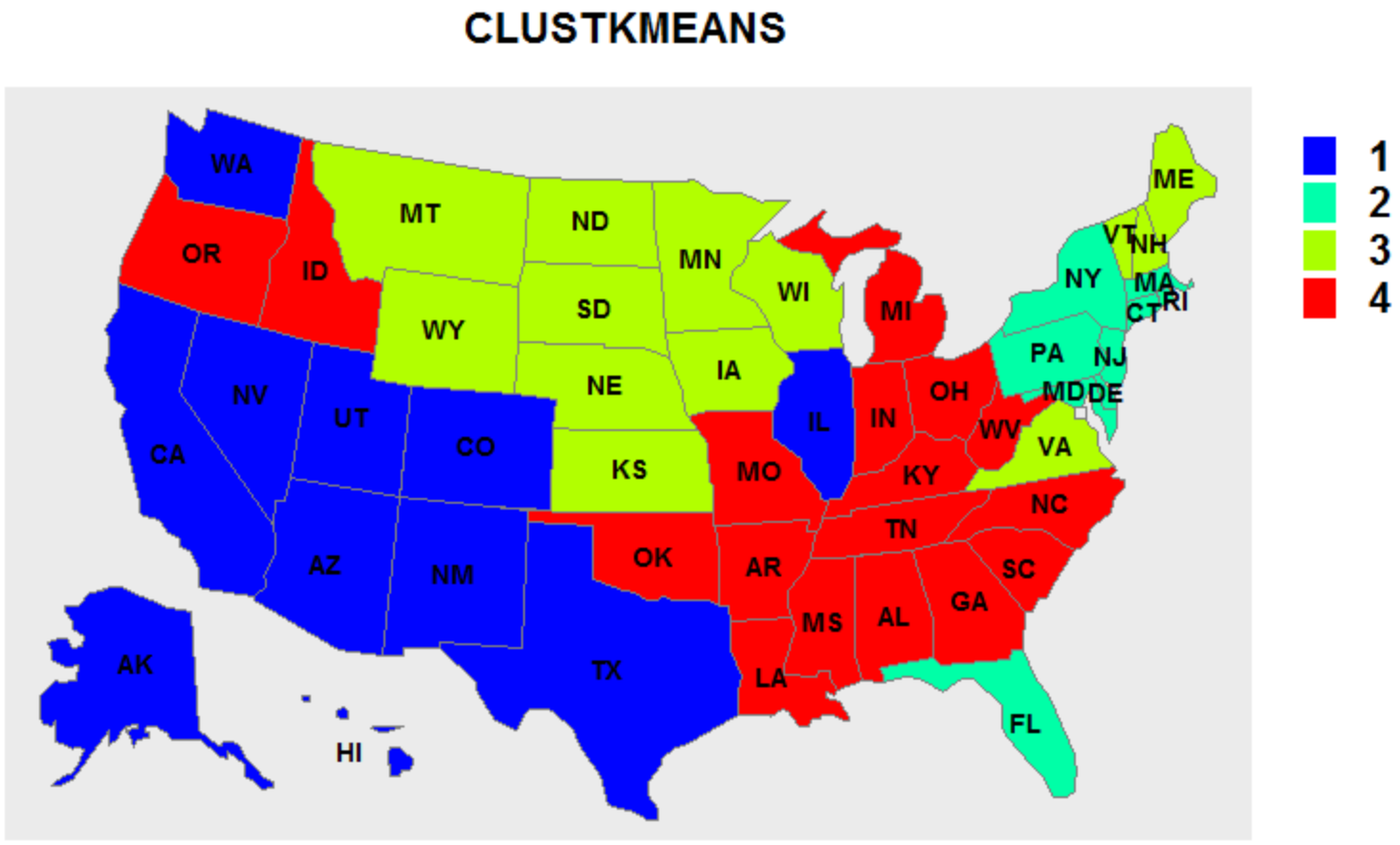
Row 5 = California

Row 32 = New York

Row 34 = North Dakota

Row 43 = Texas

Map for k-Means



Classification

The classification problem deals with the *assignment* of objects to *known* groups. Goals include:

1. Understanding what differentiates the groups.
2. Predicting group membership for unclassified objects.

Methods

- **Discriminant Analysis** (on the Centurion *Relate* menu) – based on Fisher's linear discriminant functions.

- **Neural Network Classifier** (on the Centurion *Relate* menu) – based on a Bayesian classifier with priors and costs.

- **Uniwin** (companion program to Statgraphics) – includes methods for dealing with qualitative factors.

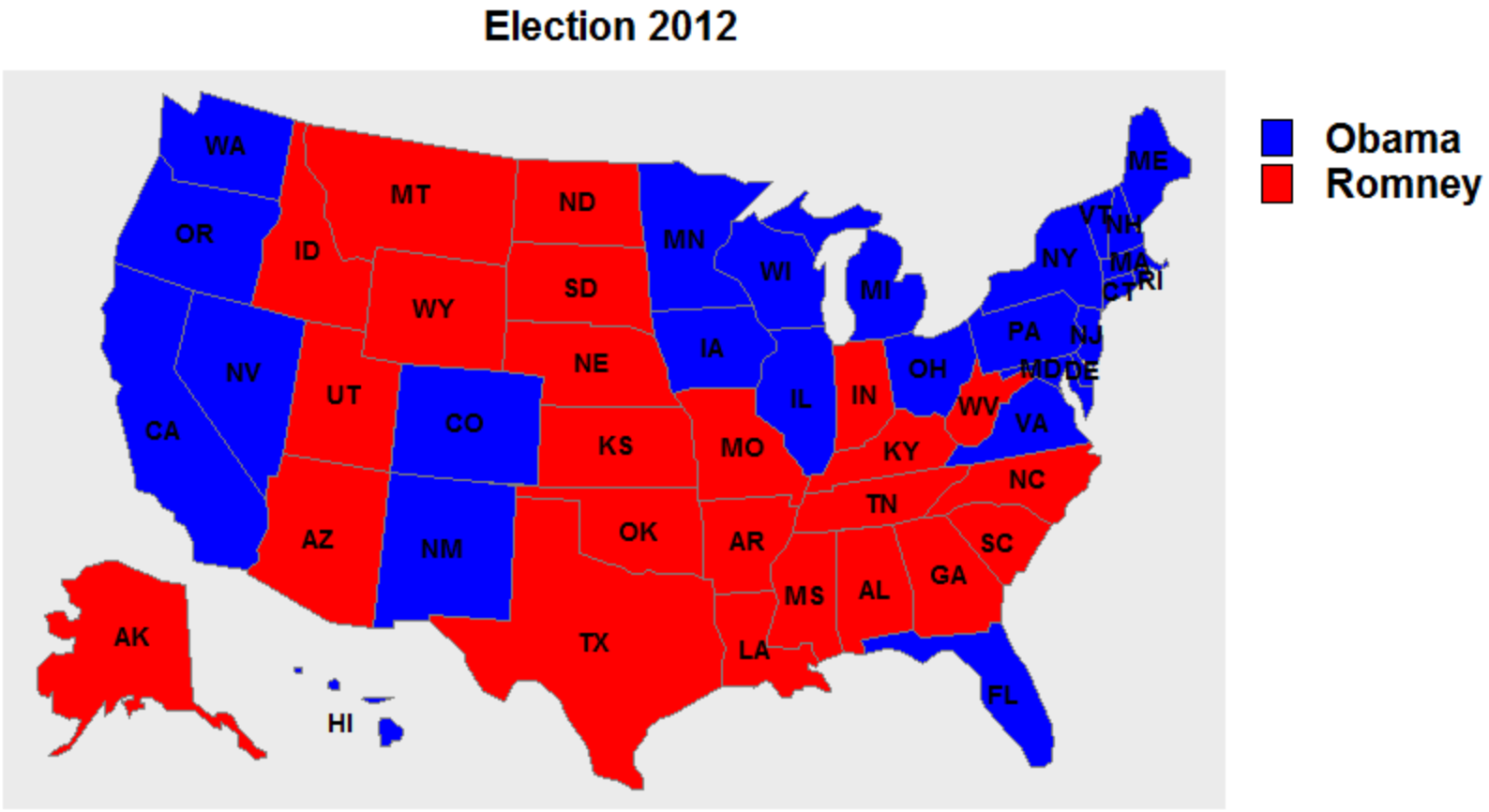
(Linear) Discriminant Analysis

Constructs linear functions of the standardized classification variables:

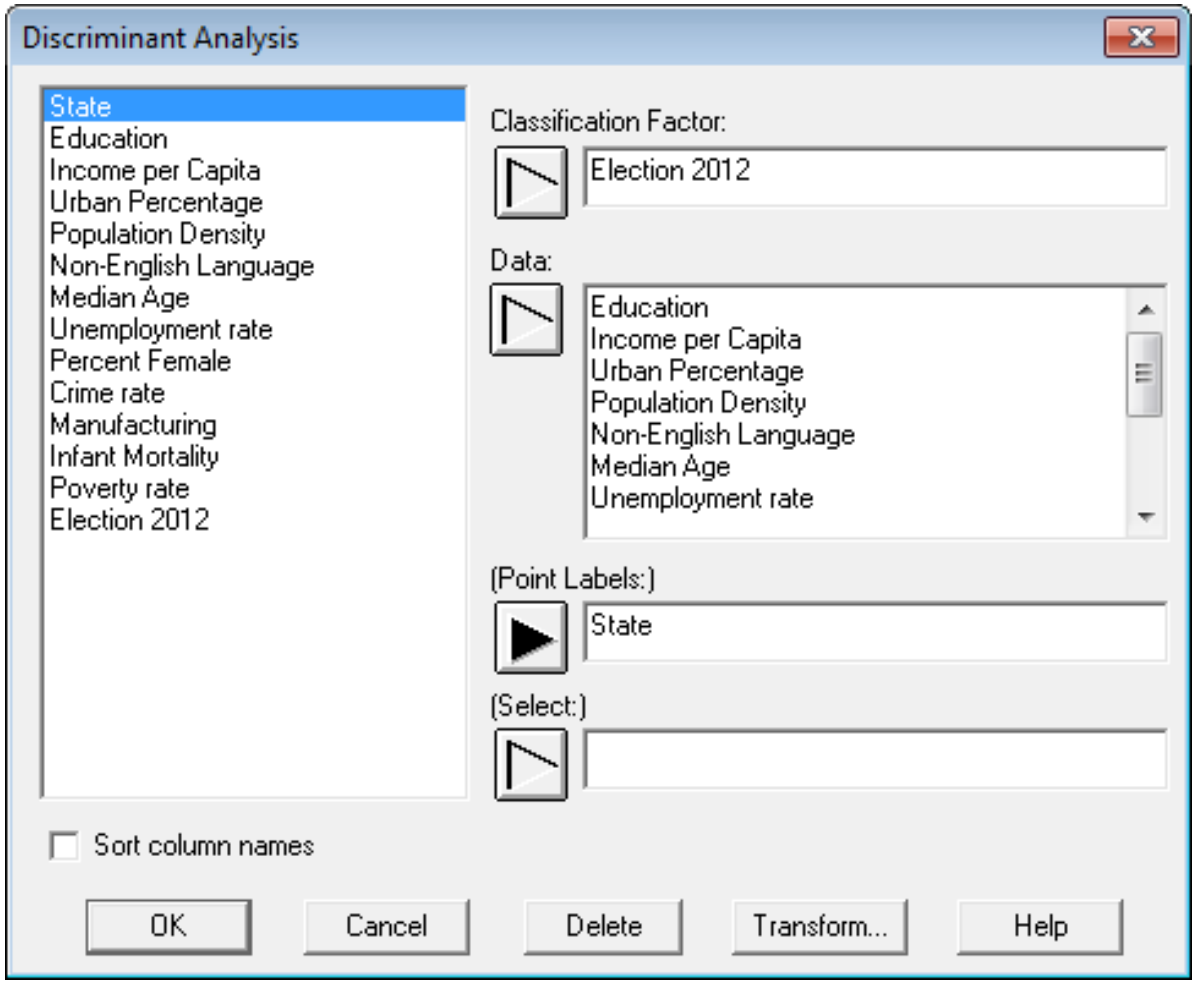
$$D_j = d_{j1}Z_1 + d_{j2}Z_2 + \ldots + d_{jp}Z_p$$

These functions are derived so as to maximize the separation of the groups. If there are *g* groups, there are *g-1* discriminant functions.

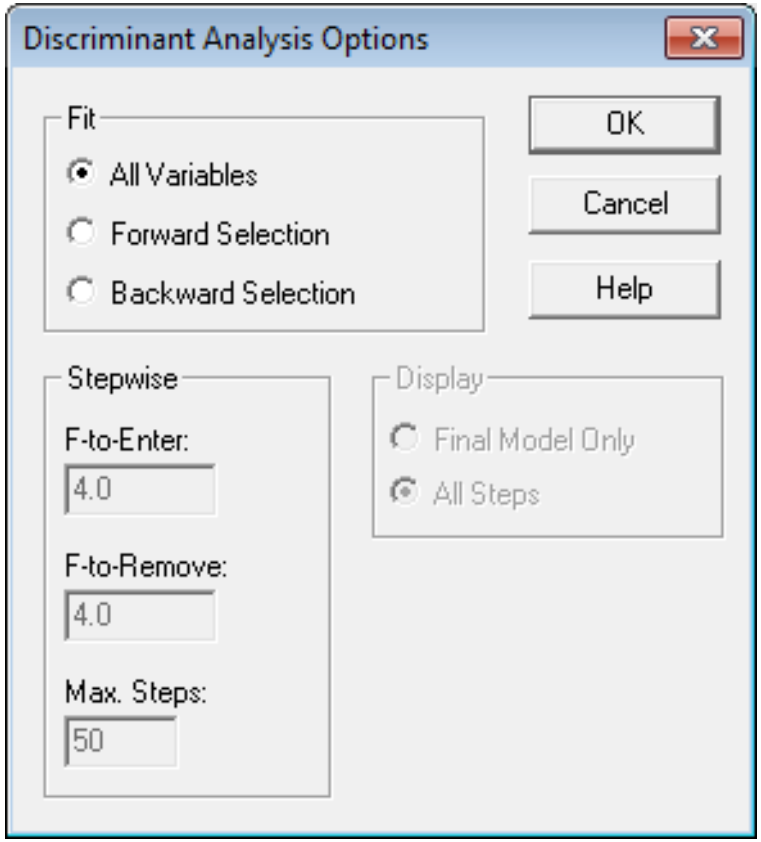
Example


Election 2012
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME
AK
HI
Obama
Romney

Data Input


Discriminant Analysis
State
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
Election 2012
Classification Factor:
Election 2012
Data:
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
(Point Labels:)
State
(Select:)
Sort column names
OK
Cancel
Delete
Transform...
Help

Analysis Options

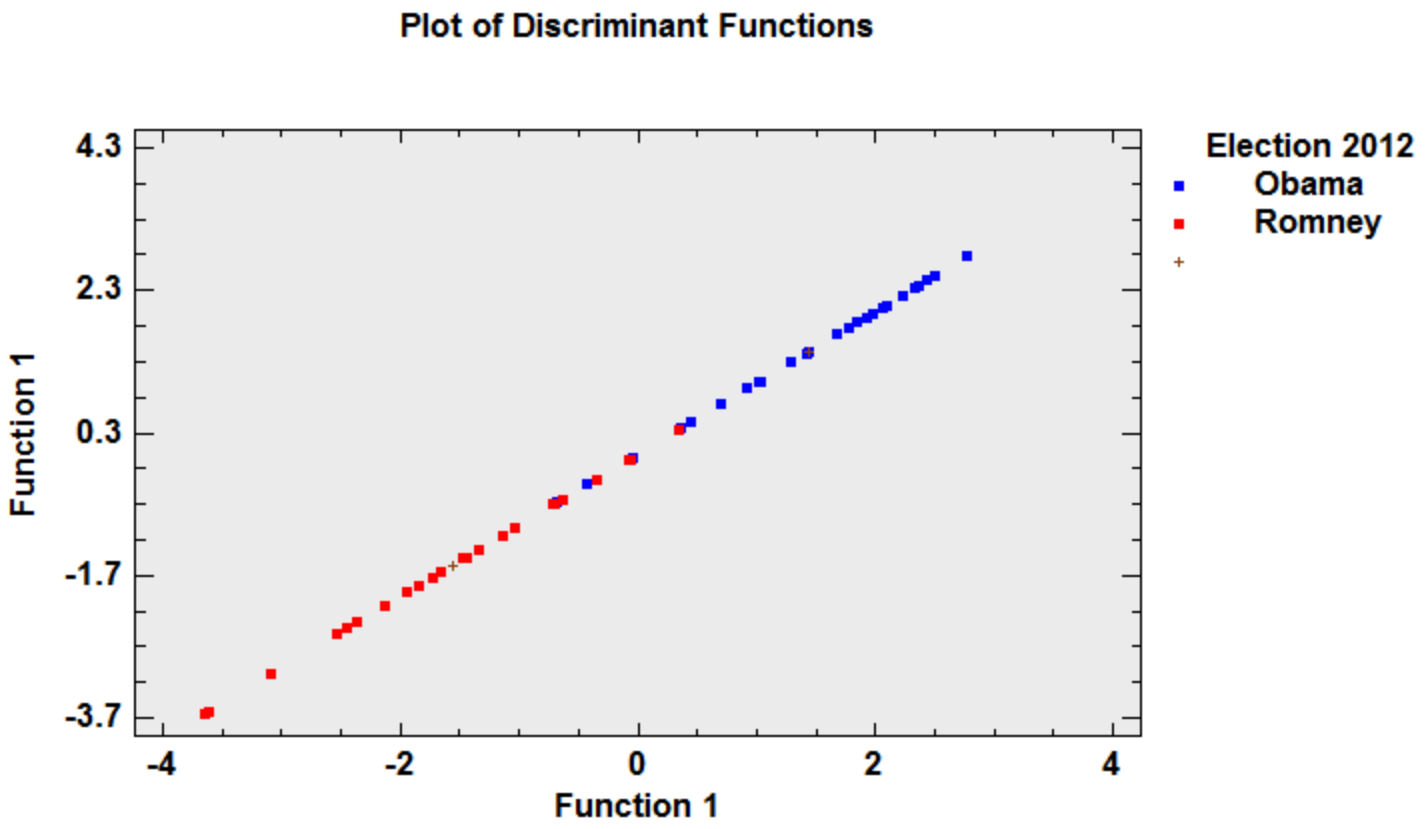


All Variables

Discriminant Function Coefficients for Election 2012

	1
Education	0.395536
Income per Capita	-0.360961
Urban Percentage	0.63466
Population Density	-1.15086
Non-English Language	0.475647
Median Age	0.623294
Unemployment rate	0.925253
Percent Female	1.17913
Crime rate	-0.13596
Manufacturing	-0.13972
Infant Mortality	-0.338133
Poverty rate	-1.23874

Discriminant Function Plot



Backward selection

Discriminant Function Coefficients for Election 2012

	1
Urban Percentage	0.819124
Population Density	-1.0466
Median Age	0.547071
Unemployment rate	0.988681
Percent Female	1.13951
Infant Mortality	-0.703059
Poverty rate	-1.10642

Classification Functions

These functions create a score for each group:

$$C_j = c_{j1}X_1 + c_{j2}X_2 + \ldots + c_{jp}X_p + c_{j0}$$

Objects are assigned to whatever group has the largest value of (C_j * $prior_j$) where $prior_j$ is the prior probability of belonging to group j.

Classification Functions

Classification Function Coefficients for Election 2012

	Obama	*Romney*
Urban Percentage	7.52144	7.35091
Population Density	-1.04647	-1.03366
Median Age	4.02402	3.1769
Unemployment rate	70.5856	68.958
Percent Female	364.974	360.8
Infant Mortality	-64.7134	-63.0789
Poverty rate	-52.0814	-51.0366
CONSTANT	-9218.23	-8981.45

Classification Summary

Classification Table

Actual	*Group*	*Predicted*	*Election 2012*
Election 2012	*Size*	*Obama*	*Romney*
Obama	26	25	1
		(96.15%)	(3.85%)
Romney	24	2	22
		(8.33%)	(91.67%)

Percent of cases correctly classified: 94.00%

Classification Table

		Actual	Highest	Highest	Squared		2nd Highest	2nd Highest	Squared	
Row	Label	Group	Group	Value	Distance	Prob.	Group	Value	Distance	Prob.
1	Alabama	Romney	Romney	9063.2	0.190741	0.9940	Obama	9058.09	10.41	0.0060
2	Alaska	Romney	Romney	8526.19	1.86488	0.9995	Obama	8518.49	17.2667	0.0005
3	Arizona	Romney	Romney	9003.84	1.05538	0.7360	Obama	9002.82	3.10603	0.2640
4	Arkansas	Romney	Romney	8920.24	0.85143	0.9984	Obama	8913.77	13.7822	0.0016
5	California	Obama	Obama	9274.07	0.506866	0.9972	Romney	9268.19	12.2616	0.0028
6	Colorado	Obama	Obama	9112.55	0.127363	0.9476	Romney	9109.66	5.91866	0.0524
7	Connecticut	Obama	Obama	9318.2	2.71709	0.9998	Romney	9309.71	19.6965	0.0002
8	Delaware	Obama	Obama	9176.07	0.677643	0.8313	Romney	9174.48	3.8672	0.1687
9	Florida	Obama	Obama	9234.89	0.313588	0.9957	Romney	9229.44	11.2205	0.0043
10	Georgia	Romney	Romney	8963.96	0.079138	0.9908	Obama	8959.28	9.43119	0.0092
11	Hawaii	Obama	Obama	9011.55	0.212008	0.9313	Romney	9008.94	5.42548	0.0687
12	Idaho	Romney	Romney	8818.46	0.593405	0.9976	Obama	8812.42	12.6739	0.0024
13	Illinois	Obama	Obama	9341.46	0.0631484	0.9900	Romney	9336.87	9.2477	0.0100
14	Indiana	Romney	Romney	9144.64	1.85605	0.5226	Obama	9144.54	2.0373	0.4774
15	Iowa	Obama	Obama	9286.51	0.0456147	0.9643	Romney	9283.22	6.63646	0.0357
16	Kansas	Romney	Romney	9118.58	1.27521	0.6772	Obama	9117.84	2.75712	0.3228
17	Kentucky	Romney	Romney	8991.35	0.0028434	0.9769	Obama	8987.61	7.48781	0.0231
18	Louisiana	Romney	Romney	8942.78	2.96204	0.9998	Obama	8934.09	20.347	0.0002
19	Maine	Obama	Obama	9287.42	0.0751031	0.9580	Romney	9284.29	6.32854	0.0420
20	Maryland	Obama	Obama	9307.34	0.0631806	0.9900	Romney	9302.74	9.24809	0.0100
21	Massachusetts	Obama	Obama	9232.47	0.880308	0.9985	Romney	9225.96	13.8977	0.0015
22	Michigan	Obama	Obama	9169.03	0.999779	0.7506	Romney	9167.93	3.20345	0.2494
23	Minnesota	Obama	Obama	9268.2	0.0118297	0.9731	Romney	9264.62	7.18746	0.0269
24	Mississippi	Romney	Romney	8877.5	6.44242	1.0000	Obama	8866.53	28.3865	0.0000
25	Missouri	Romney	*Obama	9308.98	0.576118	0.8549	Romney	9307.21	4.12368	0.1451
26	Montana	Romney	Romney	8912.5	0.490746	0.8740	Obama	8910.56	4.36466	0.1260
27	Nebraska	Romney	*Obama	9169.94	1.87659	0.5174	Romney	9169.87	2.0159	0.4826
28	Nevada	Obama	Obama	9059.19	0.0219447	0.9867	Romney	9054.88	8.63094	0.0133
29	New Hampshire	Obama	Obama	9297.53	0.0650424	0.9901	Romney	9292.93	9.27046	0.0099
30	New Jersey	Obama	Obama	9020.75	0.826247	0.9984	Romney	9014.32	13.6803	0.0016
31	New Mexico	Obama	*Romney	9036.19	0.123342	0.9484	Obama	9033.28	5.94632	0.0516
32	New York	Obama	Obama	9363.85	0.685872	0.9980	Romney	9357.65	13.0891	0.0020

Validation Set

May leave some of the objects out of the "training" set and use them to validate the results.

		Actual	*Highest*	*Highest*	*Squared*		*2nd Highest*	*2nd Highest*	*Squared*	
Row	*Label*	*Group*	*Group*	*Value*	*Distance*	*Prob.*	*Group*	*Value*	*Distance*	*Prob.*
41	South Dakota		Romney	7717.54	1.10608	0.9979	Obama	7711.39	13.3976	0.0021
42	Tennessee		Romney	7901.73	0.00882873	0.9746	Obama	7898.09	7.30368	0.0254
43	Texas		Romney	7878.05	0.510392	0.8233	Obama	7876.51	3.58782	0.1767
44	Utah		Obama	8023.29	0.0768165	0.9358	Romney	8020.61	5.43549	0.0642
45	Vermont		Romney	7973.29	0.757301	0.7563	Obama	7972.15	3.02183	0.2437
46	Virginia		Obama	8100.11	0.239598	0.8934	Romney	8097.98	4.49051	0.1066
47	Washington		Obama	8187.44	1.17996	0.9980	Romney	8181.2	13.6518	0.0020
48	West Virginia		Romney	7883.79	0.162729	0.9885	Obama	7879.33	9.07195	0.0115
49	Wisconsin		Obama	8051.58	0.414302	0.8486	Romney	8049.86	3.86086	0.1514
50	Wyoming		Romney	7776.7	0.0147521	0.9563	Obama	7773.61	6.18573	0.0437

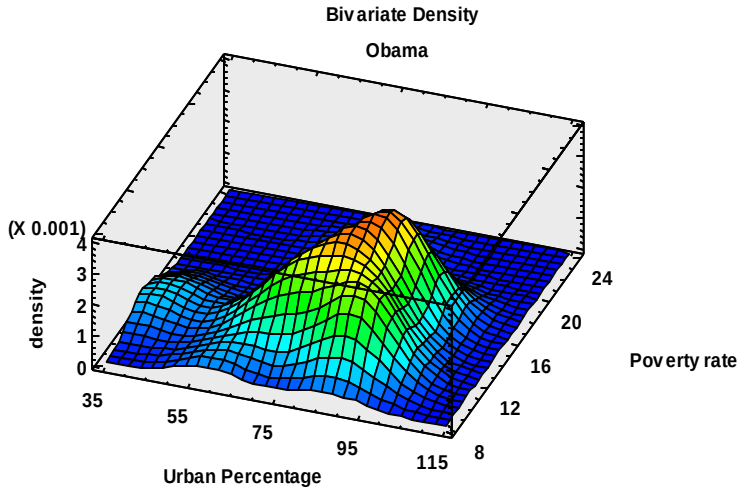
Correctly classified 8 out of 10 omitted states.

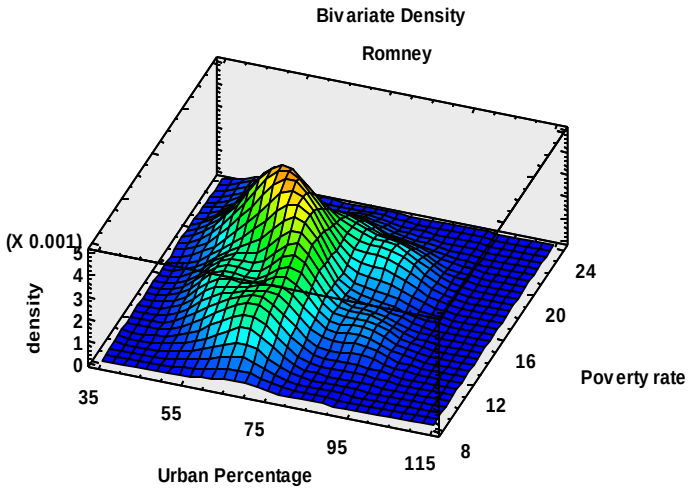
Neural Network Classifier

Implements a nonparametric method for classifying objects based on the product of 3 quantities:

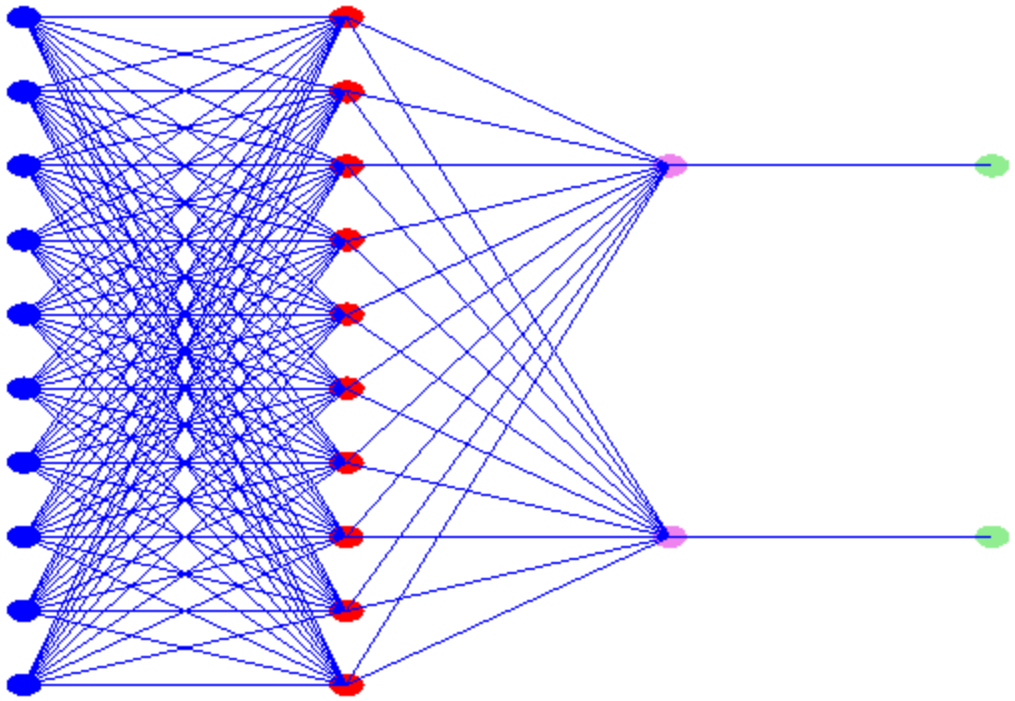
1. The estimated density function in the neighborhood of the object (given a specified value of σ).
2. The prior probabilities of belonging to each group.
3. The costs of misclassifying cases that belong to a given group.

Bivariate Density (σ = 0.5)

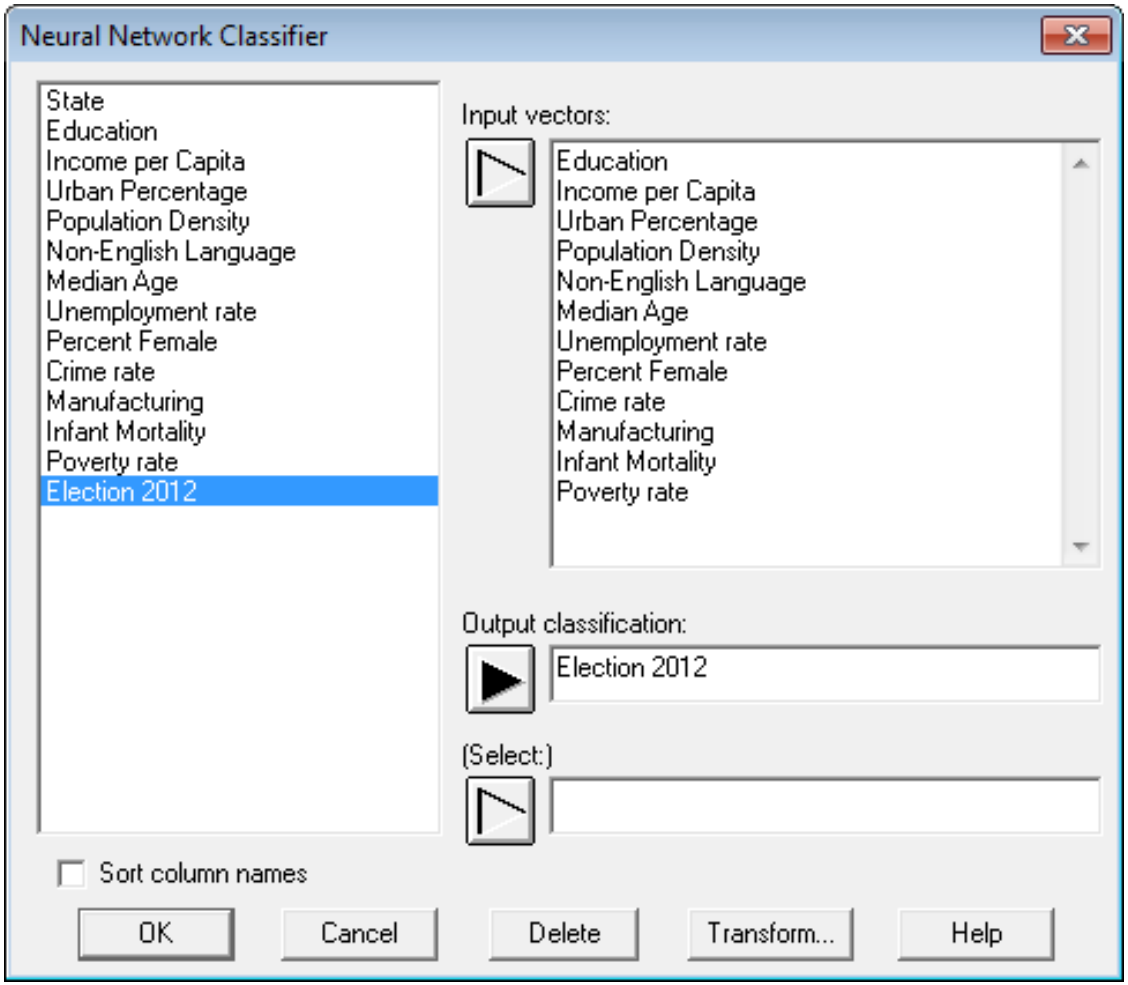




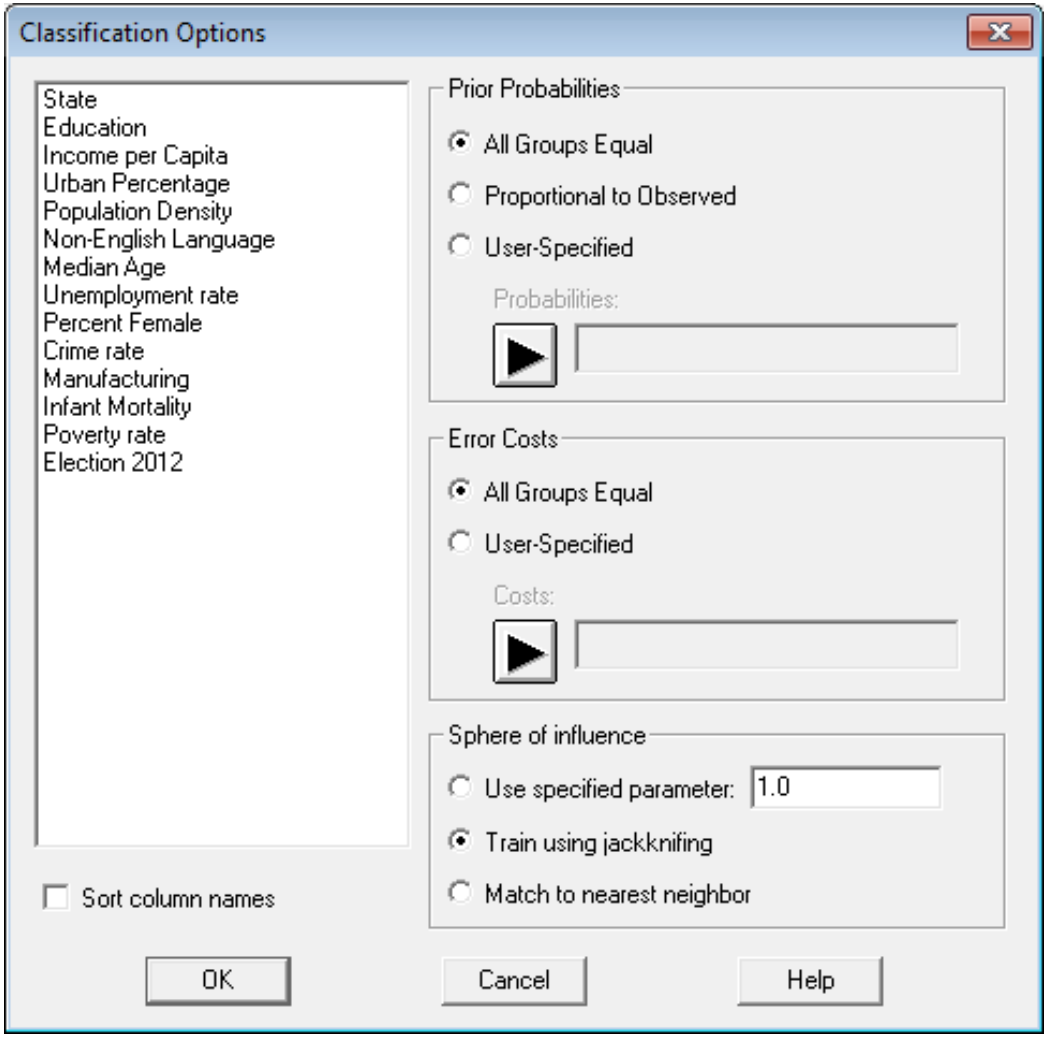
Schematic Diagram



Data Input


Neural Network Classifier
State
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
Election 2012
Input vectors:
Education
Income per Capita
Urban Percentage
Population Density
Non-English Language
Median Age
Unemployment rate
Percent Female
Crime rate
Manufacturing
Infant Mortality
Poverty rate
Output classification:
Election 2012
(Select:)
Sort column names
OK
Cancel
Delete
Transform...
Help

Analysis Options



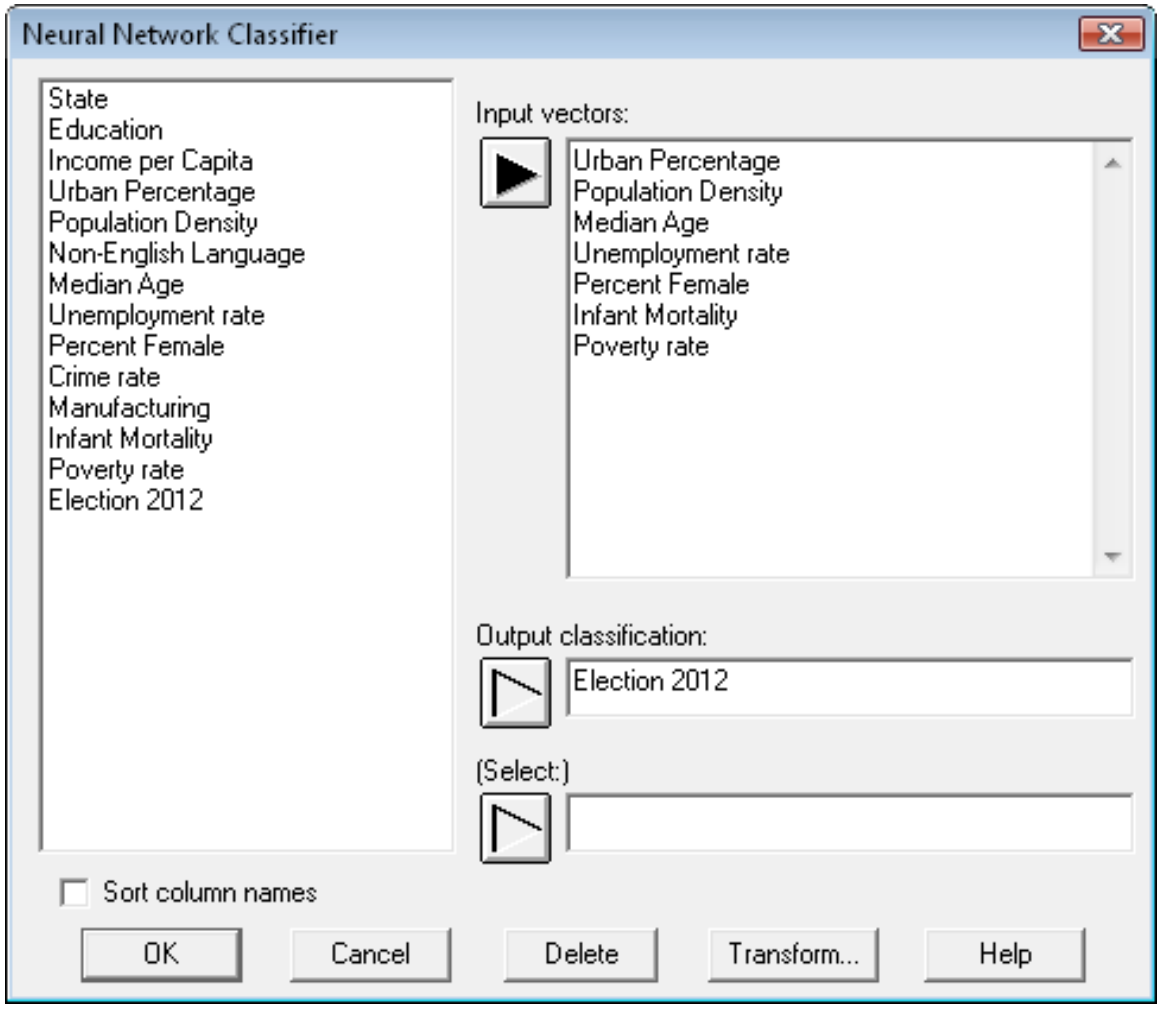
Classification Summary

Classification Table

Actual	*Group*	*Predicted*	
Election 2012	*Size*	*Obama*	*Romney*
Obama	26	18	8
		(69.23%)	(30.77%)
Romney	24	3	21
		(12.50%)	(87.50%)

Percent of training cases correctly classified: 78.00%

Reduced Set of Variables



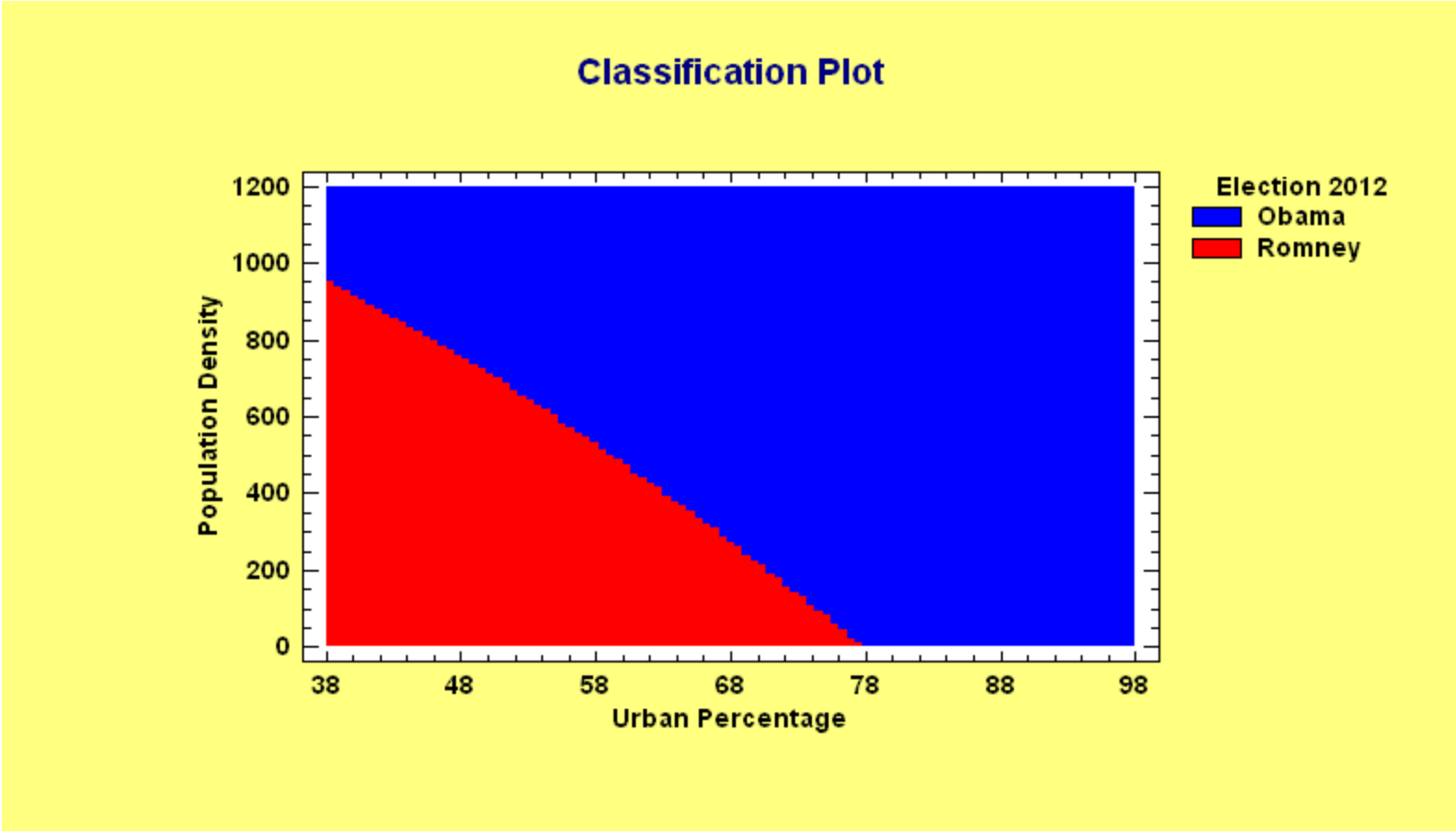
Some Improvement

Classification Table

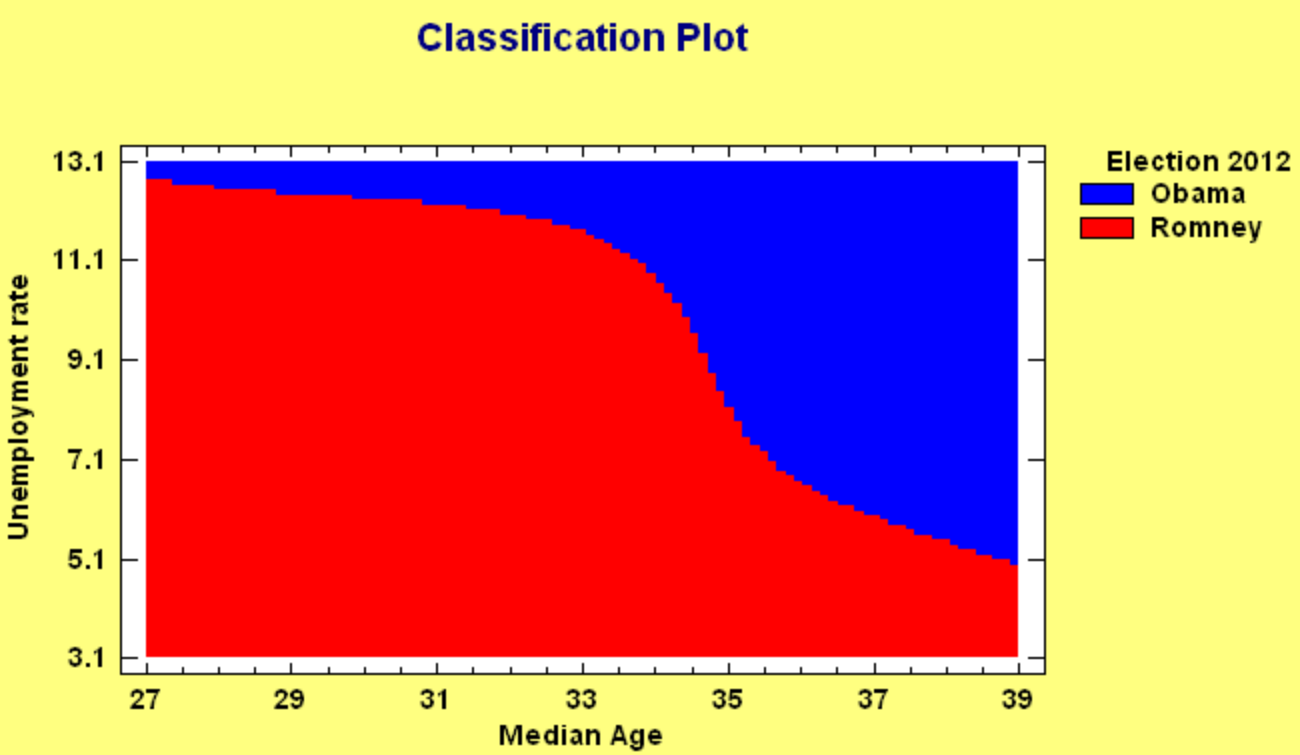
Actual	*Group*	*Predicted*	
Election 2012	*Size*	*Obama*	*Romney*
Obama	26	22	4
		(84.62%)	(15.38%)
Romney	24	3	21
		(12.50%)	(87.50%)

Percent of training cases correctly classified: 86.00%

Classification Plot



Classification Plot



UNIWIN Plus from Sigma Plus

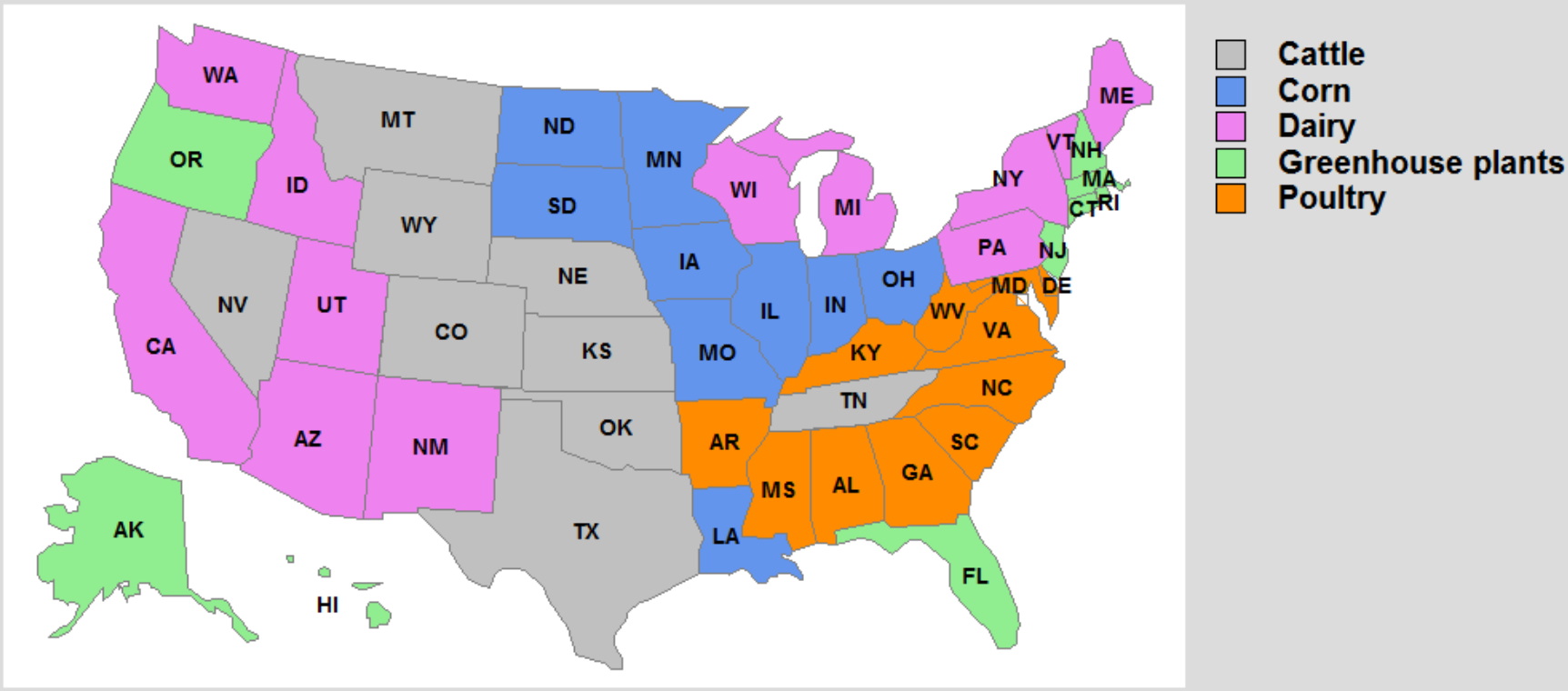
Software package written by our longtime colleague Christian Charles at Sigma Plus in France.

Contains additional features for multivariate analysis.

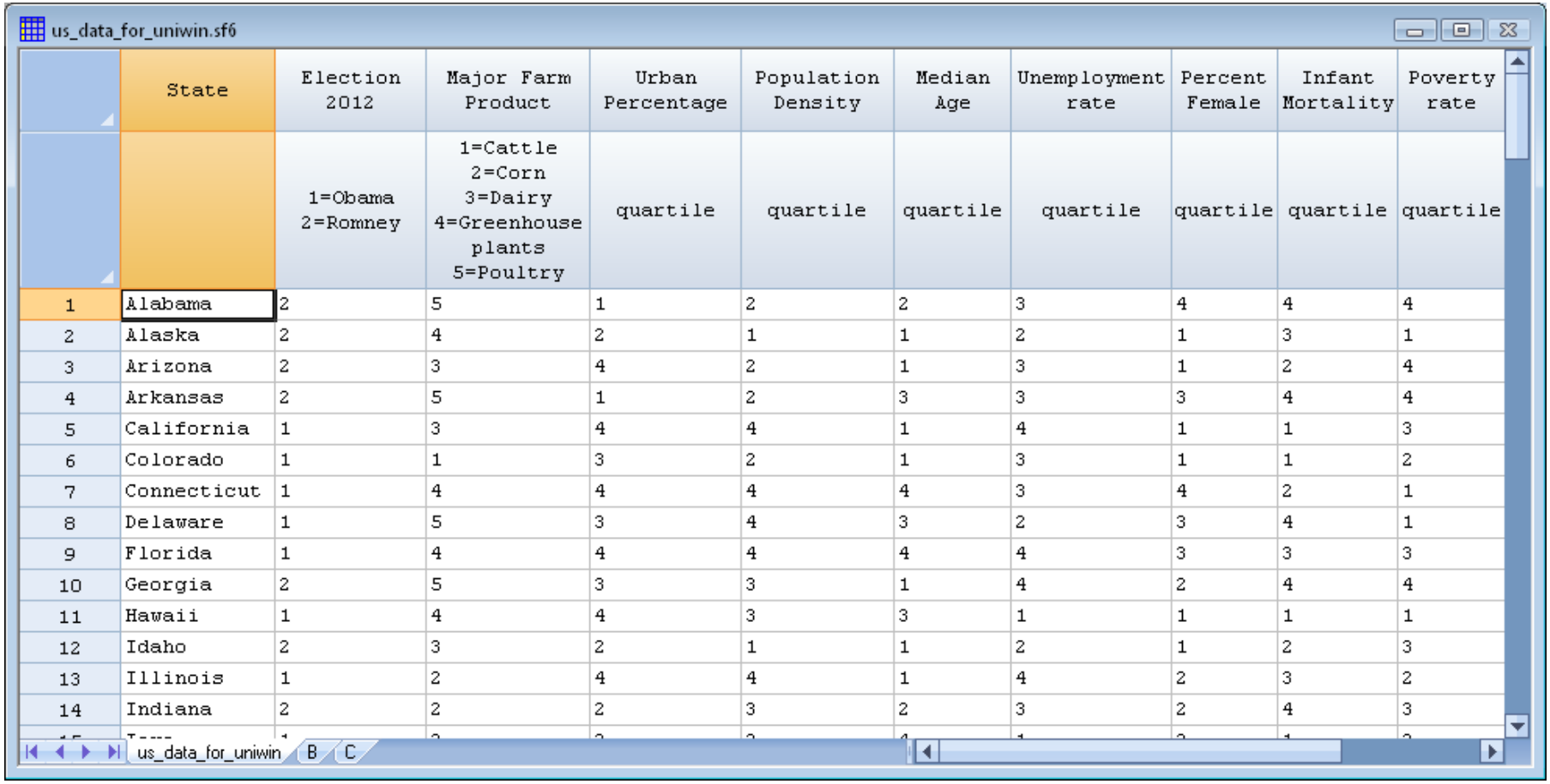
Reads Statgraphics data files.

Uniwin Cluster and Classification

Handle qualitative variables.



New Variables

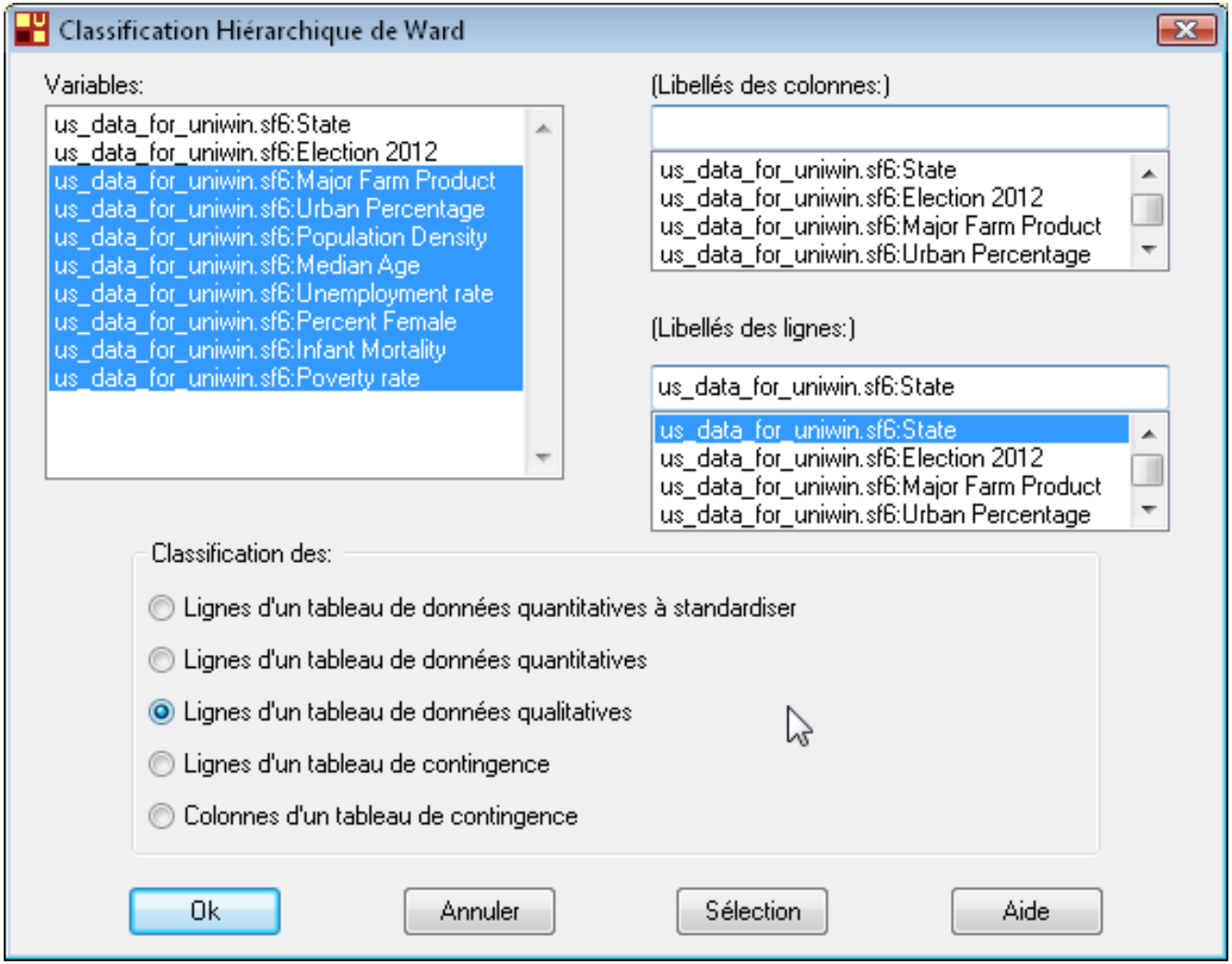


us_data_for_uniwin.sf6

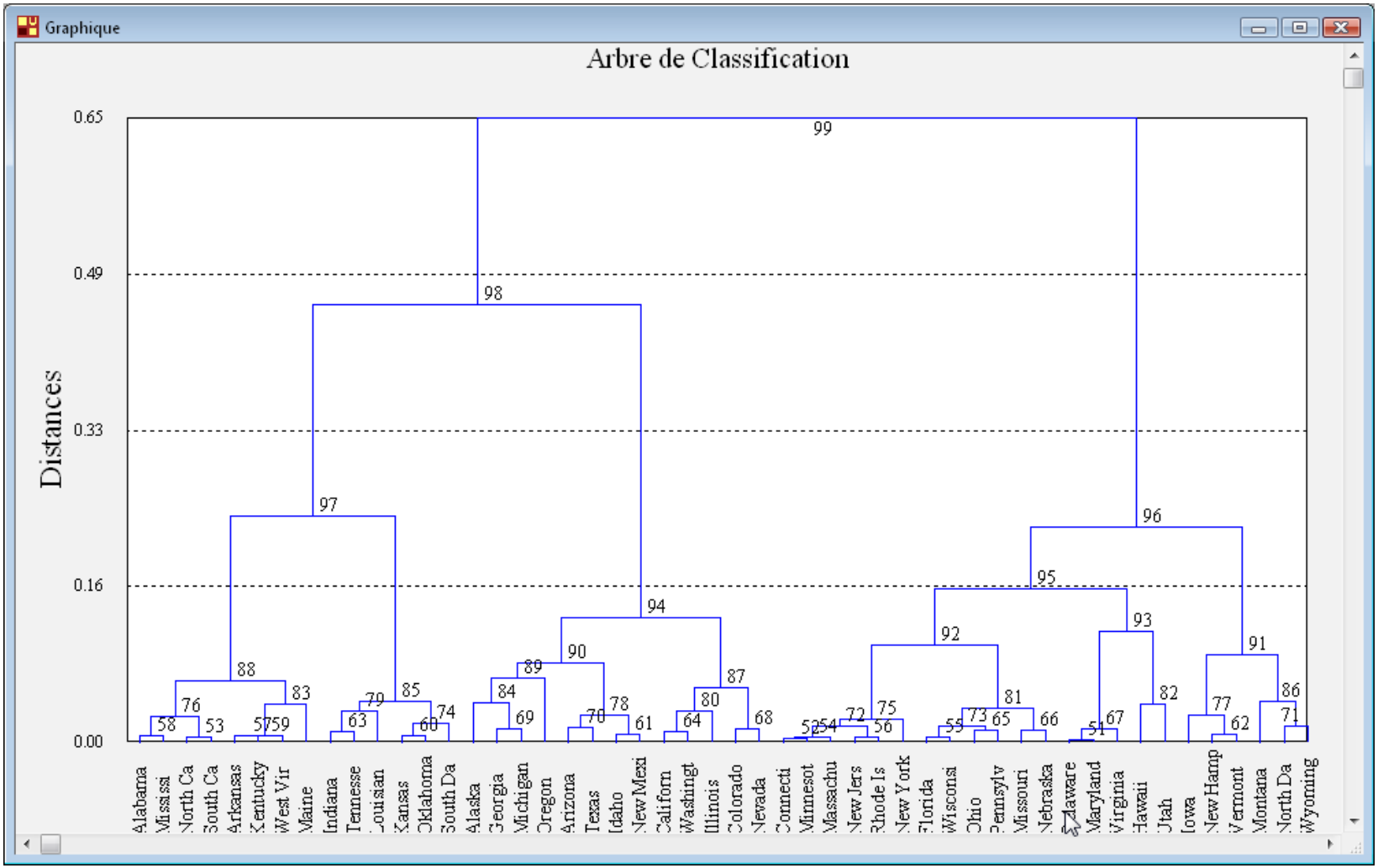
	State	Election 2012	Major Farm Product	Urban Percentage	Population Density	Median Age	Unemployment rate	Percent Female	Infant Mortality	Poverty rate
		1=Obama 2=Romney	1=Cattle 2=Corn 3=Dairy 4=Greenhouse plants 5=Poultry	quartile	quartile	quartile	quartile	quartile	quartile	quartile
1	Alabama	2	5	1	2	2	3	4	4	4
2	Alaska	2	4	2	1	1	2	1	3	1
3	Arizona	2	3	4	2	1	3	1	2	4
4	Arkansas	2	5	1	2	3	3	3	4	4
5	California	1	3	4	4	1	4	1	1	3
6	Colorado	1	1	3	2	1	3	1	1	2
7	Connecticut	1	4	4	4	4	3	4	2	1
8	Delaware	1	5	3	4	3	2	3	4	1
9	Florida	1	4	4	4	4	4	3	3	3
10	Georgia	2	5	3	3	1	4	2	4	4
11	Hawaii	1	4	4	3	3	1	1	1	1
12	Idaho	2	3	2	1	1	2	1	2	3
13	Illinois	1	2	4	4	1	4	2	3	2
14	Indiana	2	2	2	3	2	3	2	4	3

us_data_for_uniwin B C

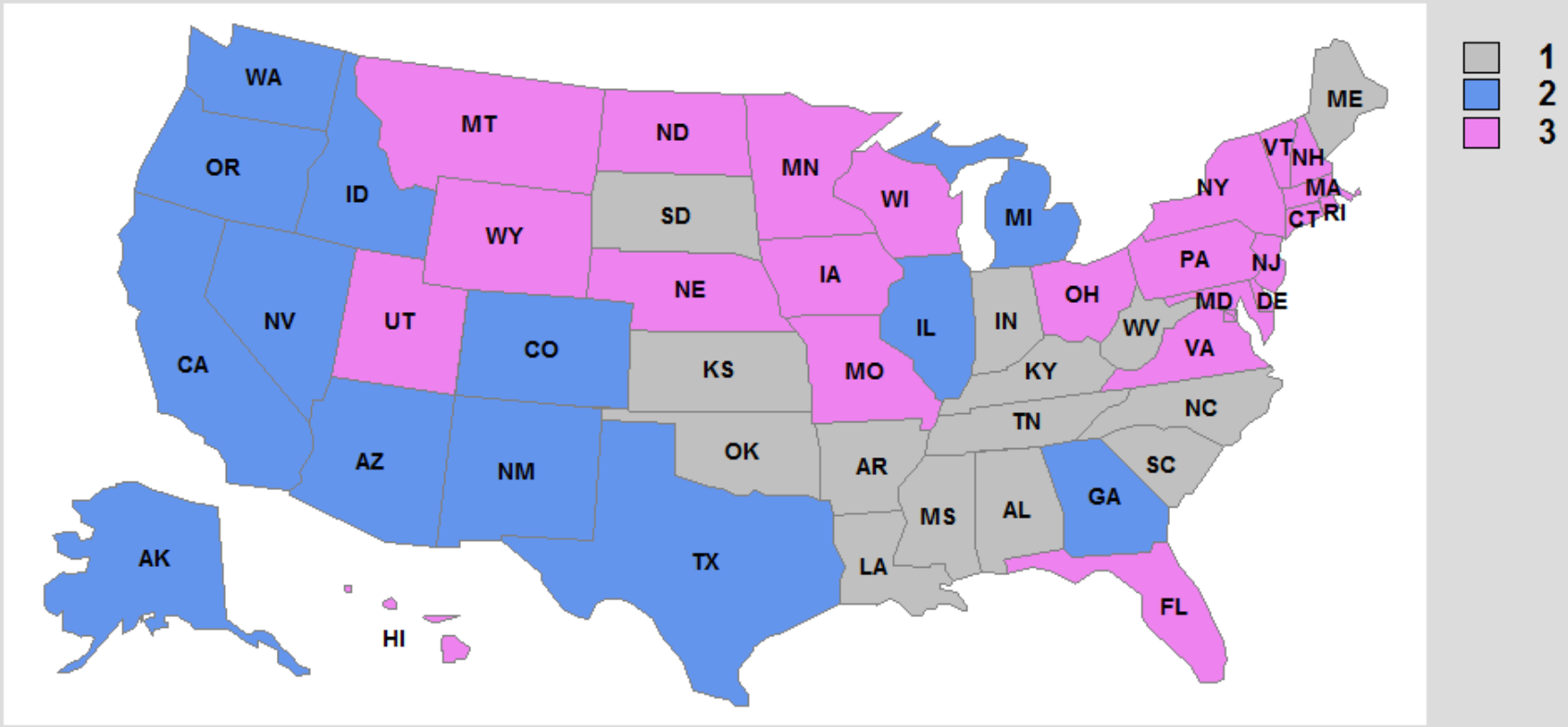
Uniwin Cluster Analysis



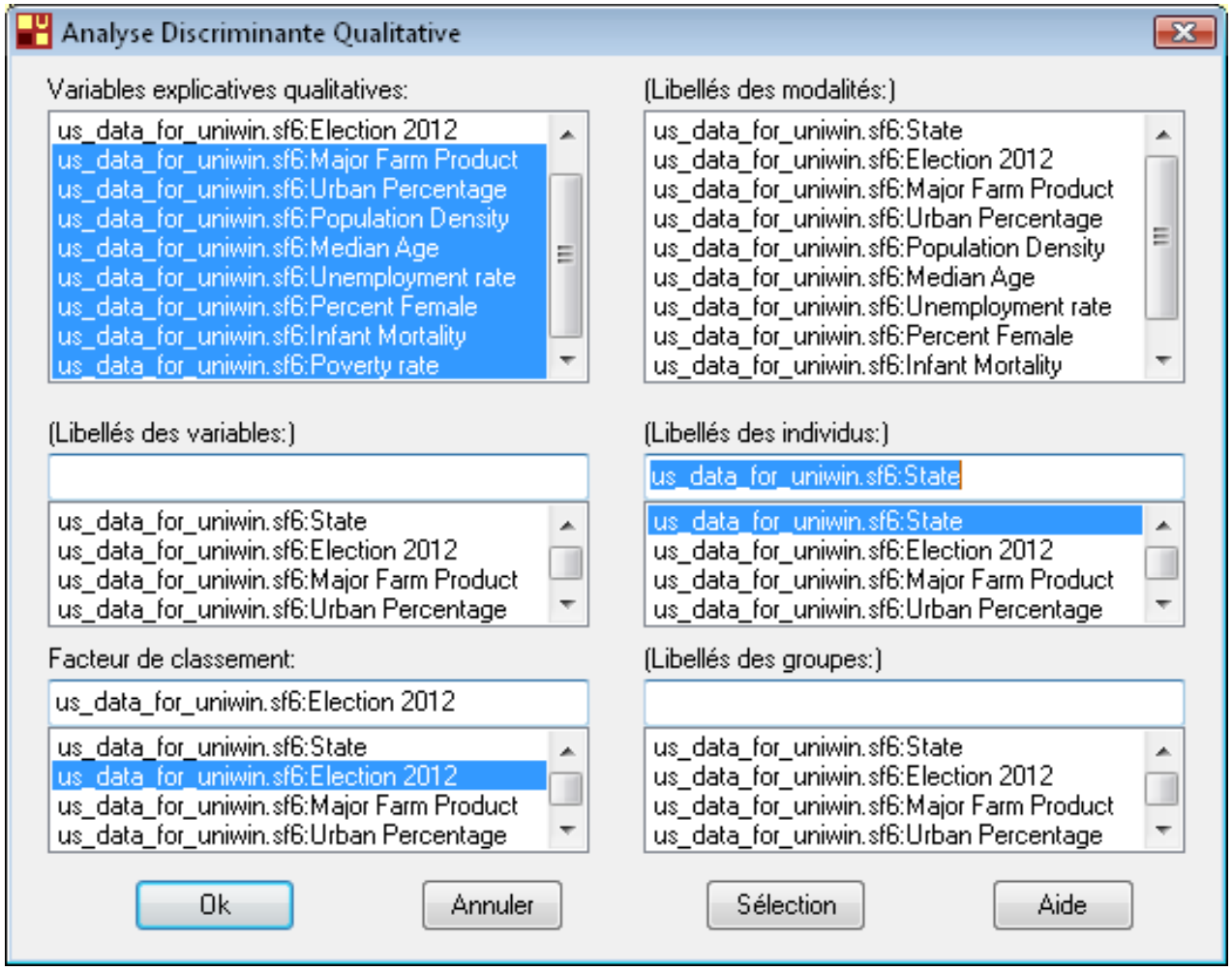
Dendrogram


Graphique
Arbre de Classification
Distances
0.65
0.49
0.33
0.16
0.00
99
98
97
96
95
94
93
92
91
90
89
88
87
86
85
84
83
82
81
80
79
78
77
76
75
74
73
72
71
70
69
68
67
66
65
64
63
62
61
60
59
58
57
56
55
54
53
52
51
Alabama
Mississi
North Ca
South Ca
Arkansas
Kentucky
West Vir
Maine
Indiana
Tennesse
Louisian
Kansas
Oklahoma
South Da
Alaska
Georgia
Michigan
Oregon
Arizona
Texas
Idaho
New Mexi
Californ
Washingt
Illinois
Colorado
Nevada
Connecti
Minnesot
Massachu
New Jers
Rhode Is
New York
Florida
Wisconsi
Ohio
Pennsylv
Missouri
Nebraska
Delaware
Maryland
Virginia
Hawaii
Utah
Iowa
New Hamp
Vermont
Montana
North Da
Wyoming

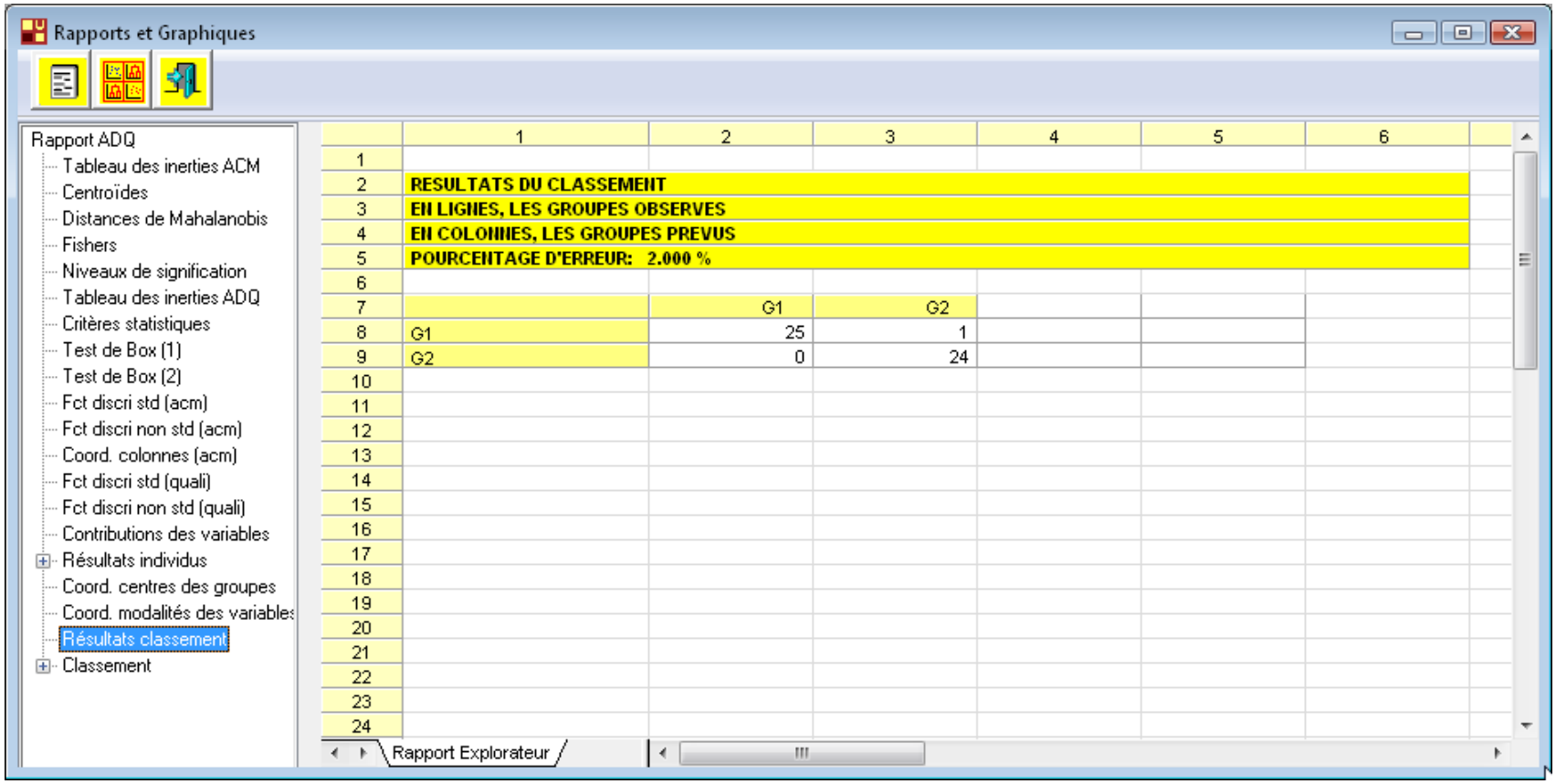
Uniwin Clusters



Uniwin Classification



Classification Results



Wrongly classified only Nevada (only cattle state to vote for Obama).

More Information

Statgraphics Centurion: www.statgraphics.com

Uniwin Plus: www.statgraphics.fr or www.sigmaplus.fr

Or send e-mail to info@statgraphics.com

Join the Statgraphics Community on:

Follow us on