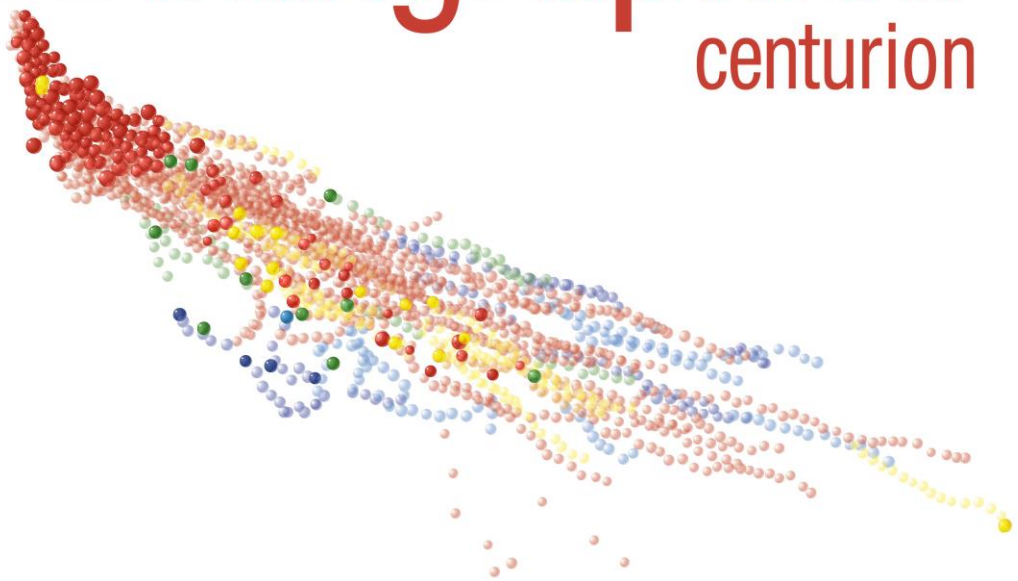
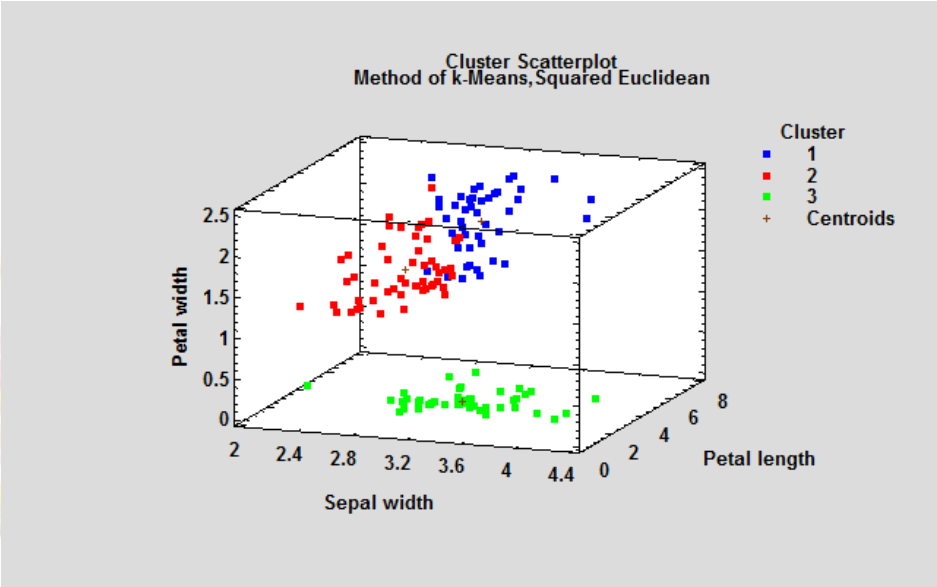
statgraphics®
centurion



Cluster Analysis using Statgraphics

Presented by
Dr. Neil W. Polhemus

Cluster Analysis



"A statistical classification technique in which cases, data, or objects (events, people, things, etc.) are sub-divided into groups (clusters) such that the items in a cluster are very similar (but not identical) to one another and very different from the items in other clusters. It is a discovery tool that reveals associations, patterns, relationships, and structures in masses of data."

from *businessdictionary.com*

statgraphics®
statgraphics.com centurion

Important Applications

- **Market research** – partitioning consumers into market segments.
- **Medical imaging** – identifying different types of tissue.
- **Bioinformatics** – separating sequences into gene families.
- **Social networking** – dividing people into communities.
- **Educational data mining** – identifying groups of students with different needs.
- **Climatology** – identifying weather patterns.

Example

THE WORLD BANK

C:\Data\Statistical practice\worldbank2009.sgd

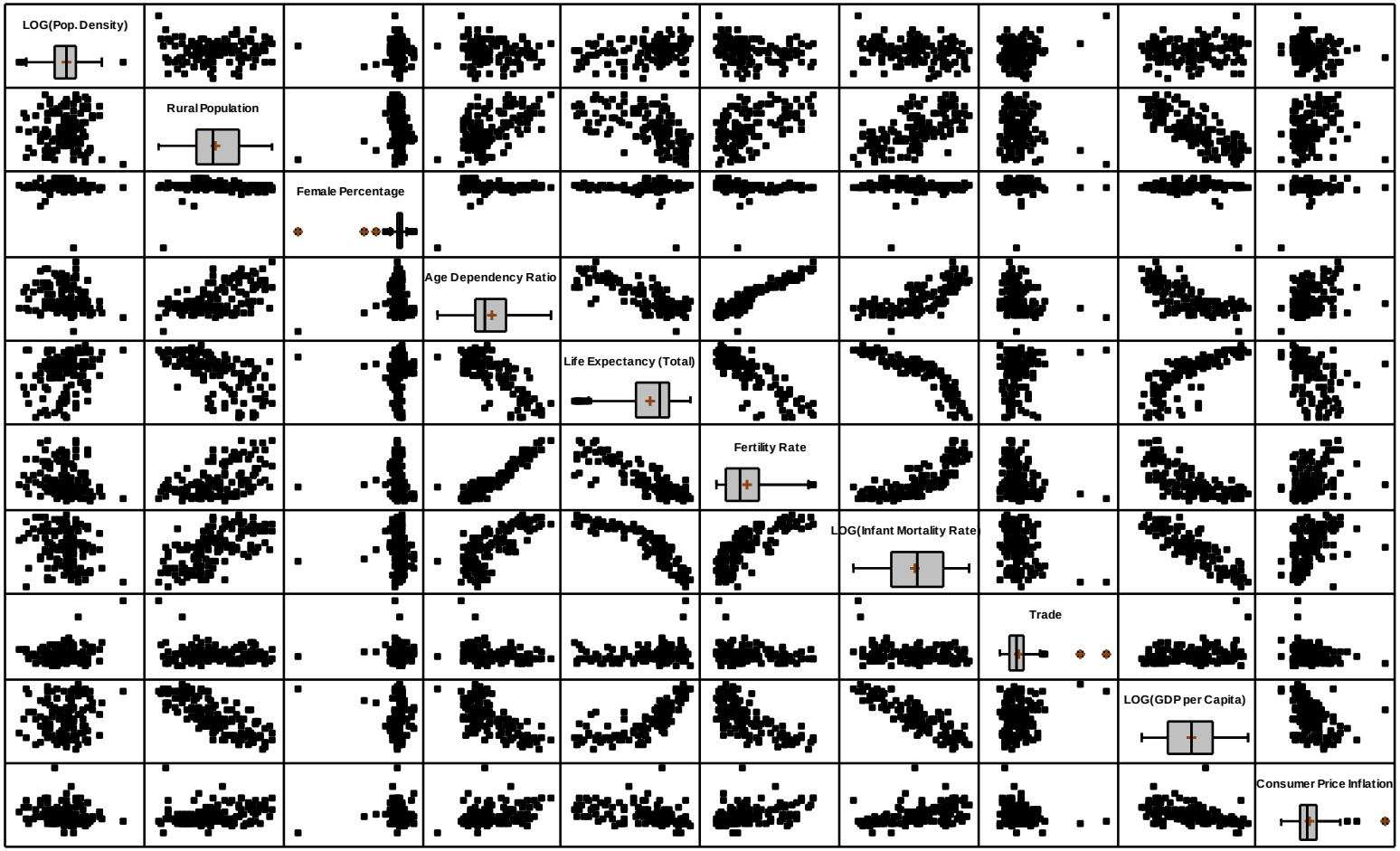
	Country	Population	Pop. Density	Rural Population	Female Percentage	Age Dependency Ratio	Life Expectancy (Total)	Life Expectancy (Female)	Life Expectancy (Male)	Fertility Rate	Infant Mortality Rate
		total	people per sq. km of land area	% of total population	% of total population	% of working-age population	years at birth	years at birth	years at birth	births per woman	per 1,000 live births
1	Aruba	106612	592.29	53.16	52.59	40.68	74.83	77.29	72.49	1.71	
2	Afghanistan	33438329	51.27	75.58	48.26	96.09	47.9	48.05	47.75	6.42	103.2
3	Angola	18555115	14.88	42.4	50.49	97.38	50.25	51.72	48.85	5.58	99.9
4	Albania	3192723	116.52	52.64	49.93	48.78	76.76	79.97	73.71	1.56	17.2
5	United Arab Emi	6938815	83	22.06	30.49	21.51	76.4	77.34	75.5	1.79	6.4
6	Argentina	40062470	14.64	7.8	51.08	55.32	75.46	79.32	71.79	2.23	12.7
7	Armenia	3084979	108.32	36.22	53.45	46.27	73.65	76.96	70.49	1.74	18.4
8	Australia	21951700	2.86	11.08	50.22	47.79	81.54	83.9	79.3	1.9	4.2
9	Austria	8365275	101.46	32.62	51.24	47.64	80.08	82.9	77.4	1.39	3.6
10	Azerbaijan	8947150	108.29	47.94	50.6	38.47	70.32	73.36	67.42	2.3	41.1
11	Burundi	8170853	318.18	89.3	50.97	70.23	49.42	50.8	48.11	4.47	88.5
12	Belgium	10796493	356.56	2.62	51	51.93	79.74	82.4	77.2	1.84	3.6
13	Benin	8601771	77.76	58.4	50.74	88.33	55.17	57.09	53.35	5.37	74.7
14	Burkina Faso	15984479	58.42	80.02	50.41	91.01	54.47	55.47	53.51	5.89	93.3
15	Bangladesh	147030145	1129.52	72.38	49.28	57.37	68.33	68.9	67.77	2.3	40
16	Bulgaria	7585131	69.84	28.6	51.63	44.9	73.41	77.1	69.9	1.57	11.1
17	Bahrain	1169578	1538.92	11.44	38.03	29.5	74.89	75.57	74.24	2.57	8.9
18	Bahamas	338358	33.8	16.1	51.12	42.28	74.98	78.16	71.95	1.9	14
19	Bosnia and Herz	3767683	73.59	51.98	51.93	41.36	75.25	77.92	72.71	1.16	7.5
20	Belarus	9507000	46.86	26.12	53.47	40.24	70.41	76.4	64.7		4.5
21	Belize	333200	14.61	47.8	50.68	65.81	75.62	77.09	74.22	2.85	14.9
22	Bolivia	9773441	9.02	33.96	50.15	69.88	65.96	68.22	63.82	3.41	43.4
23	Brazil	193246610	22.84	13.96	50.77	48.65	72.76	76.4	69.29	1.86	18.4
24	Barbados	272750	634.3	59.68	50.37	41.12	76.41	79.73	73.25	1.54	17.1
25	Brunei Darussal	391837	74.35	24.74	49.44	42.88	77.8	80.19	75.52	2.07	5.9

worldbank2009 | B | C

statgraphics® centurion

statgraphics.com

Scatterplot matrix



statgraphics®
statgraphics.com centurion

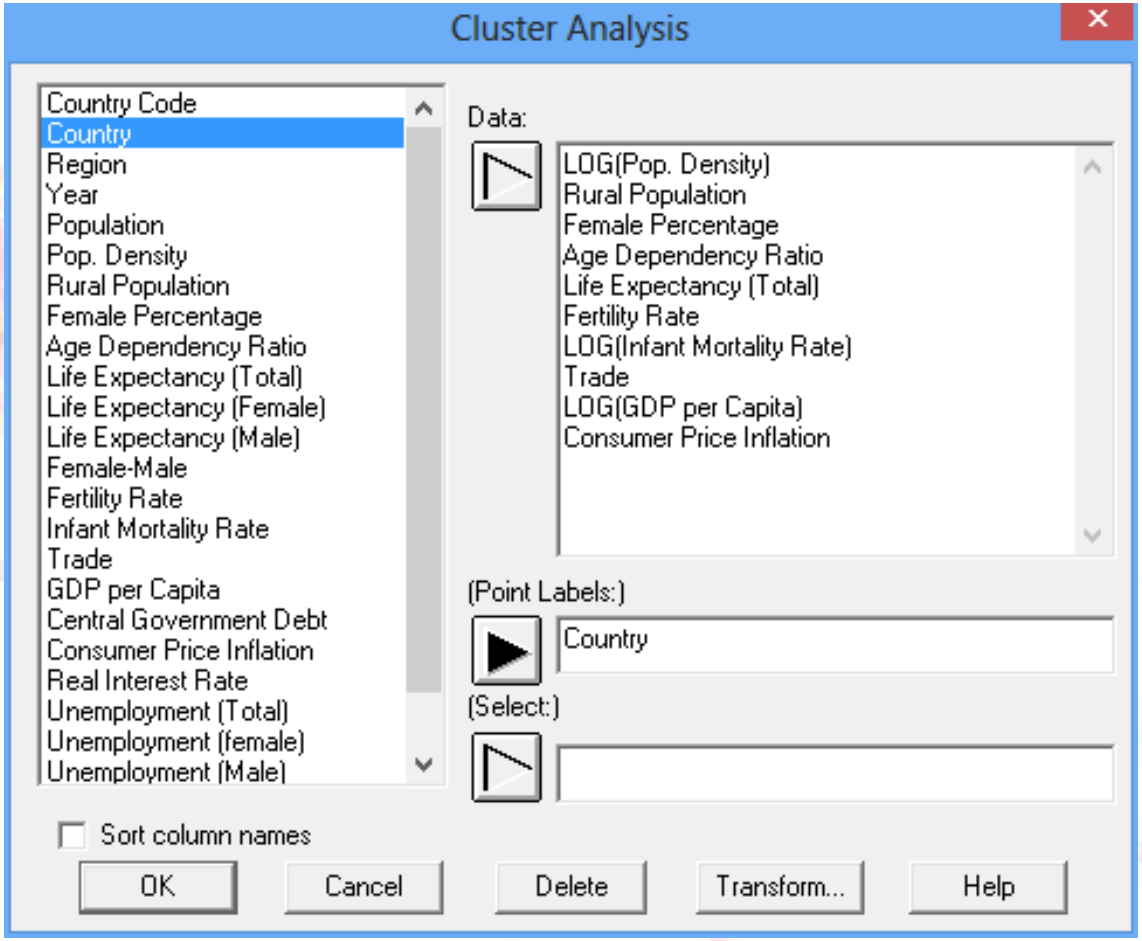
Notation

n multivariate observations

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & \vdots & & \\ & \vdots & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

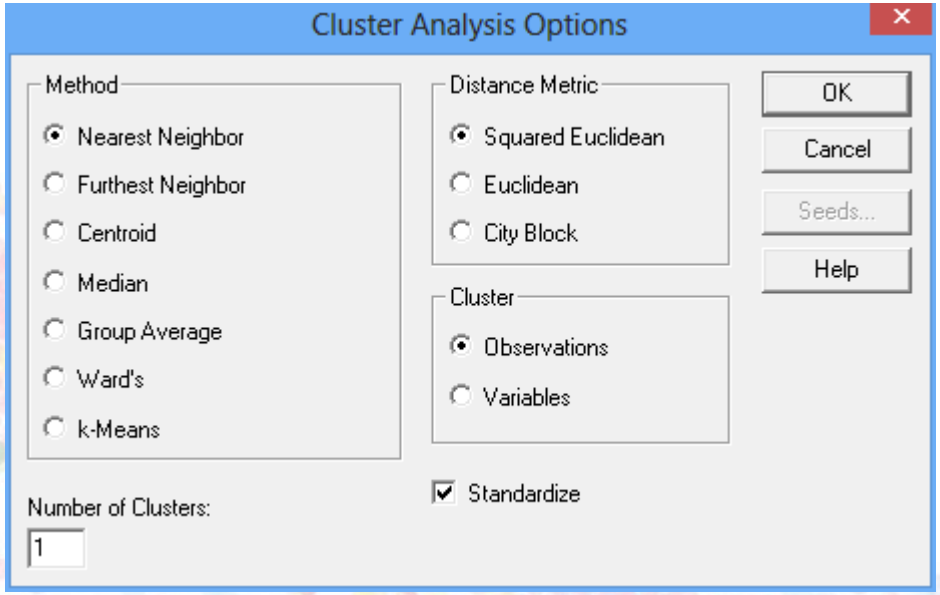
x_{ij} = value of j^{th} variable for the i^{th} item

Data Input Dialog



statgraphics®
statgraphics.com centurion

Analysis Options



Cluster – may cluster either observations or variables.
Method – procedure used to create clusters.
Distance metric – how distance between points or clusters are measured.
Standardize – whether to standardize the variables before calculating distance.
Number of clusters – target number of clusters to be created.

Distance Metrics

- Squared Euclidean

$$d(x,y) = \sum_{j=1}^{p} \left(x_j - y_j\right)^2$$

- Euclidean

$$d(x,y) = \sqrt{\sum_{j=1}^{p} \left(x_j - y_j\right)^2}$$

- City Block

$$d(x,y) = \sum_{j=1}^{p} \left|x_j - y_j\right|$$

Clustering Methods in Statgraphics Centurion

- **Agglomerative hierarchical methods**
  - Begin with *n* clusters.
  - Combine 2 clusters together based on how close they are to each other.
  - Continue until only *k* clusters remain.

- **Method of k-means**
  - Begin by creating *k* clusters using selected observations as seeds.
  - Assign other observations to the closest cluster.
  - Move points from one cluster to another until no more changes are indicated.

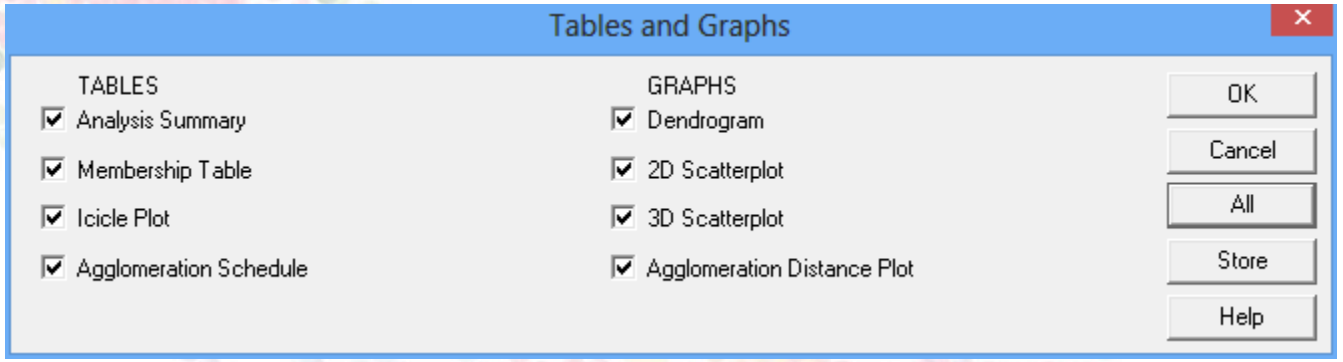
Single Linkage (Nearest Neighbor)

1. Begin with *n* clusters, one for each observation.
2. Calculate the minimum distance between all pairs of points that are located in different clusters.
3. For the pair that are closest to each other, join those 2 clusters together.
4. Repeat steps 2 to 3 until the number of clusters has been reduced to that desired.

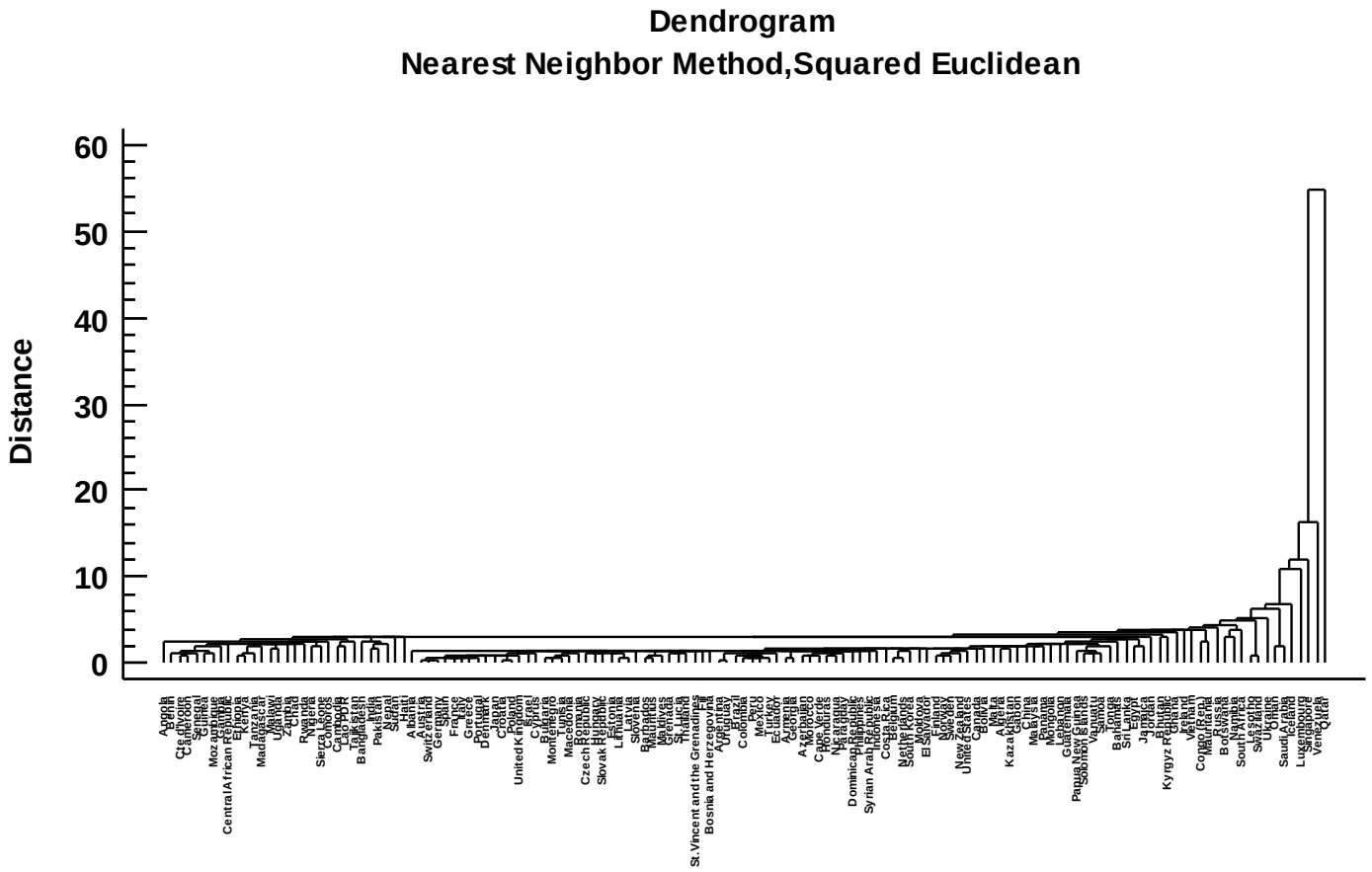
Other Hierarchical Methods

- *Complete linkage (farthest neighbor)* – distance between clusters is maximum distance between members of the 2 clusters.
- *Average linkage* – distance between clusters is average distance between all pairs of points.
- *Centroid* – distance between clusters is distance between their centroids.
- *Median* – distance between clusters is distance between their multivariate medians.
- *Ward's method* – distance between clusters is defined by increase in sums of squared deviations around the means if clusters were joined.

Tables and Graphs

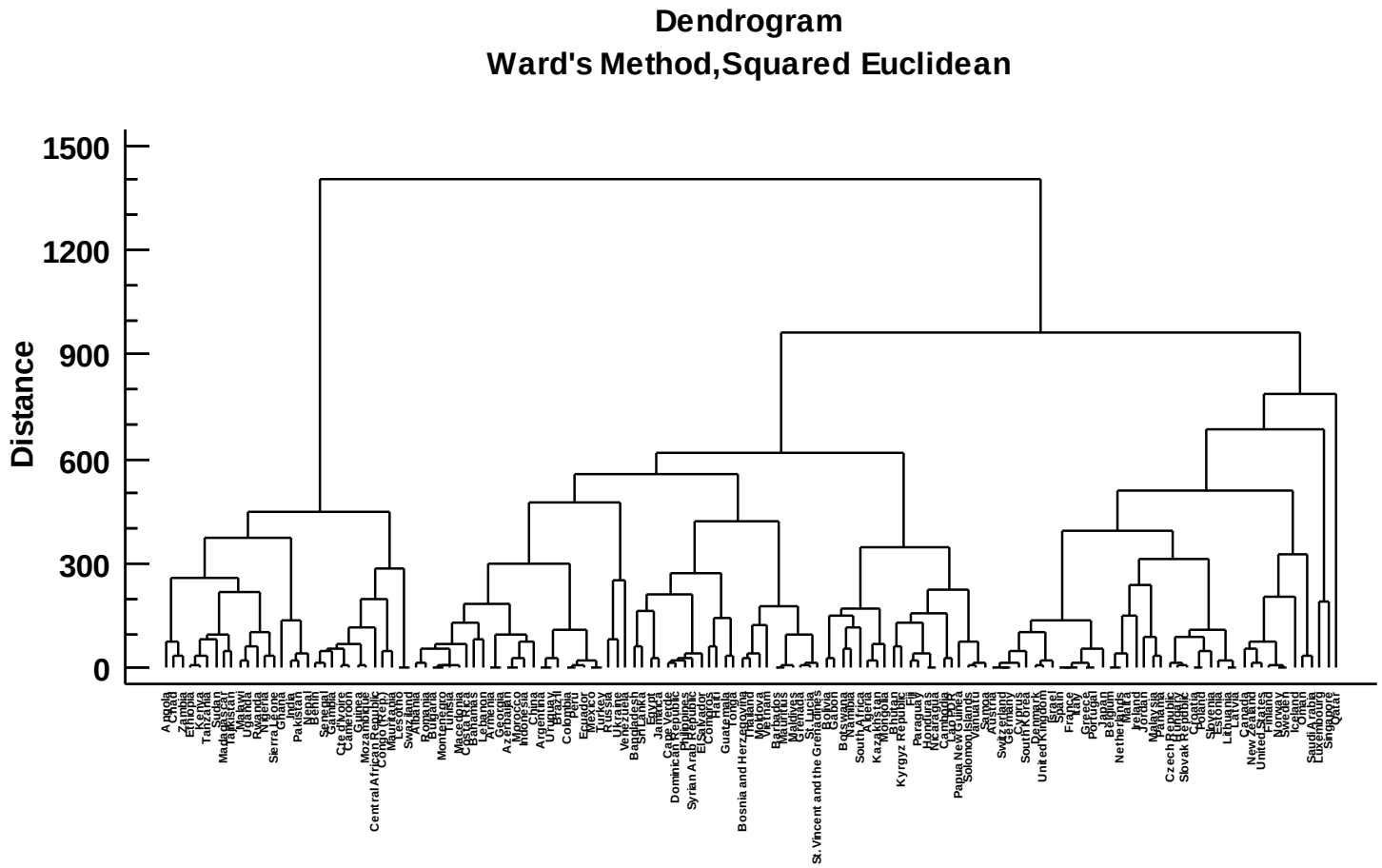


Dendrogram



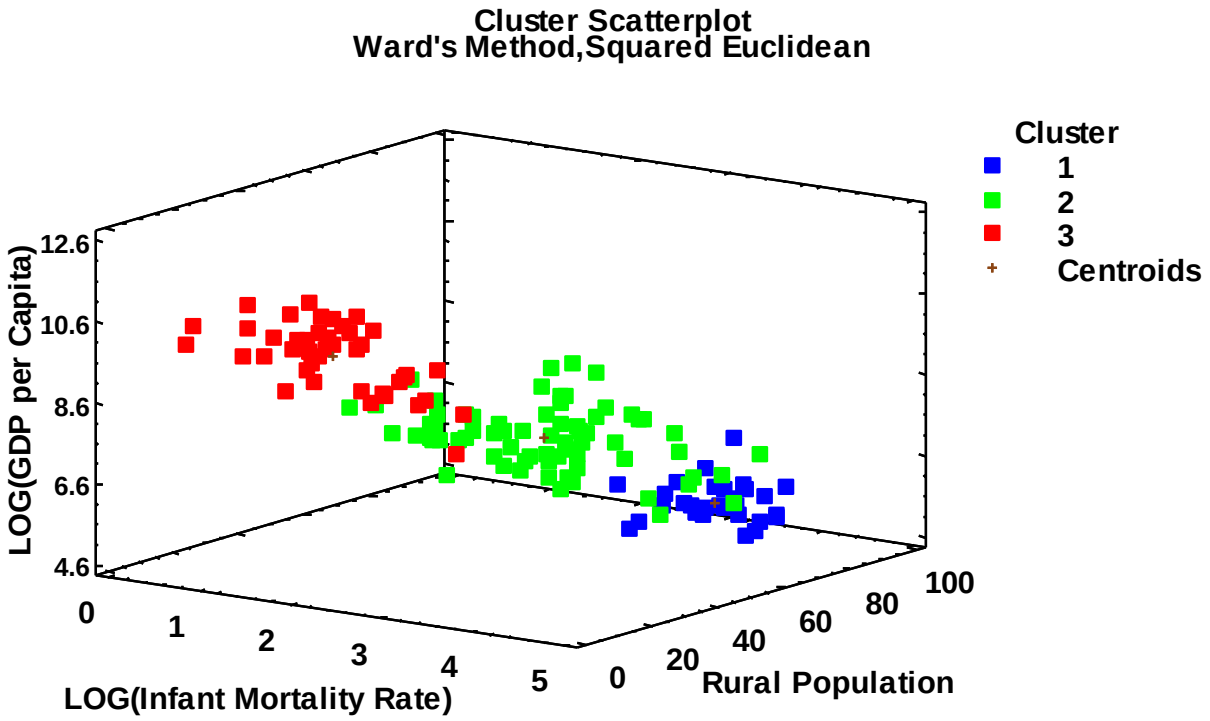
statgraphics®
statgraphics.com centurion

Ward's Method



statgraphics®
statgraphics.com centurion

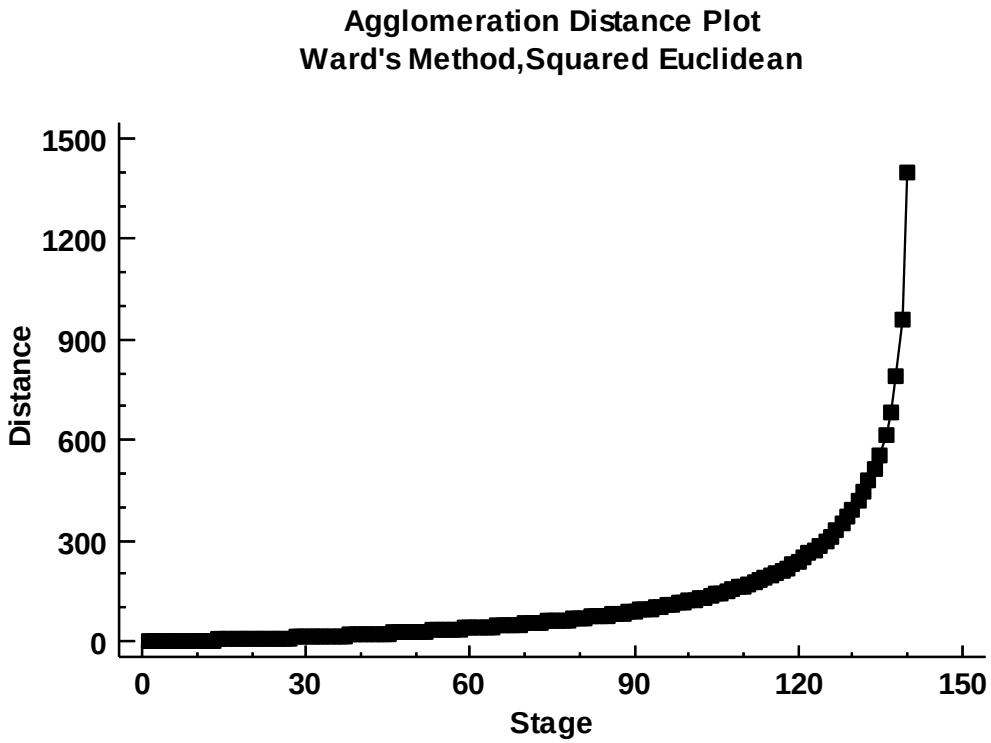
Cluster Scatterplot



Icicle Plot

statgraphics®
statgraphics.com centurion

Agglomeration Distance Plot

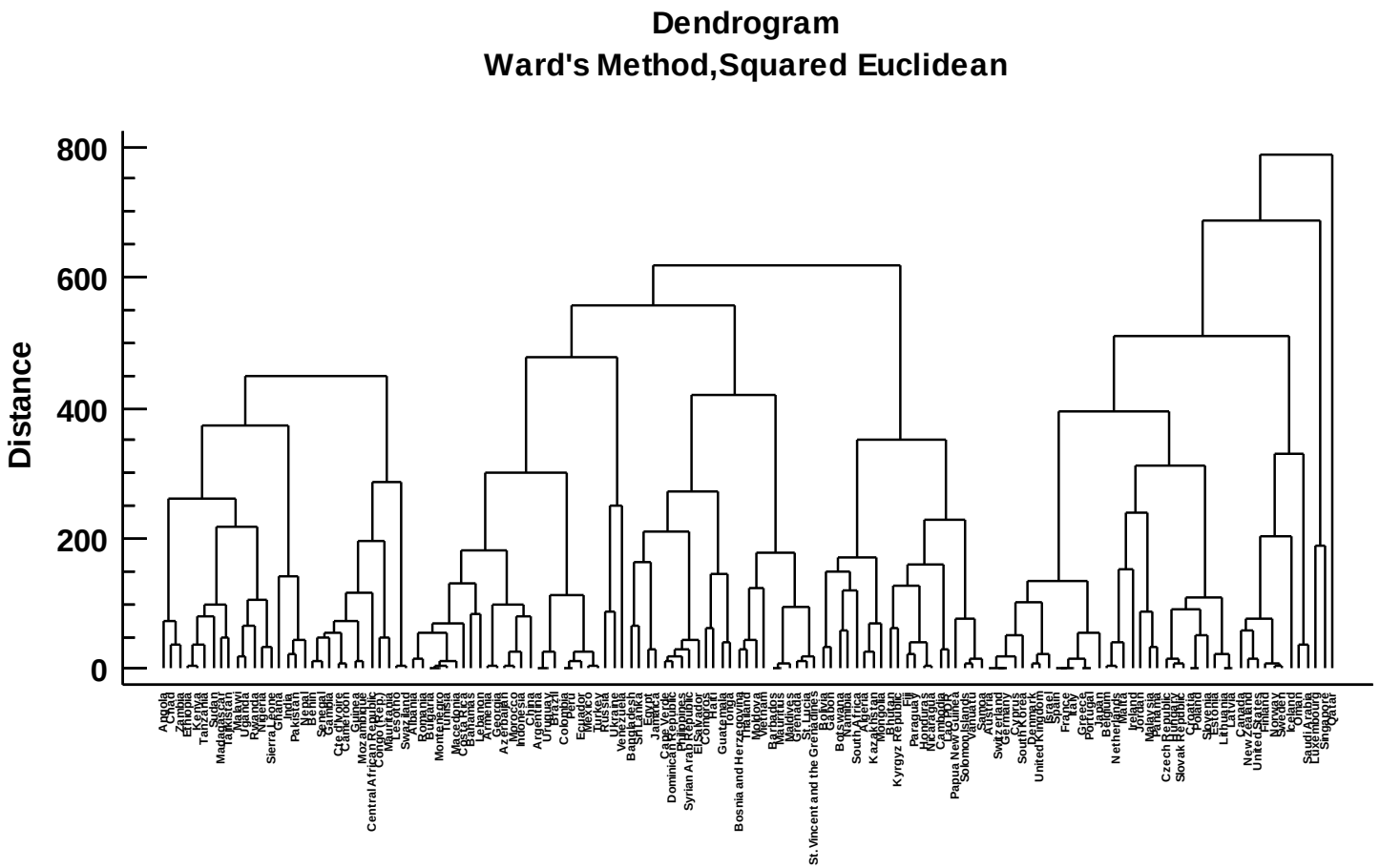


Agglomeration Schedule

Agglomeration Schedule
Clustering Method: Ward's
Distance Metric: Squared Euclidean

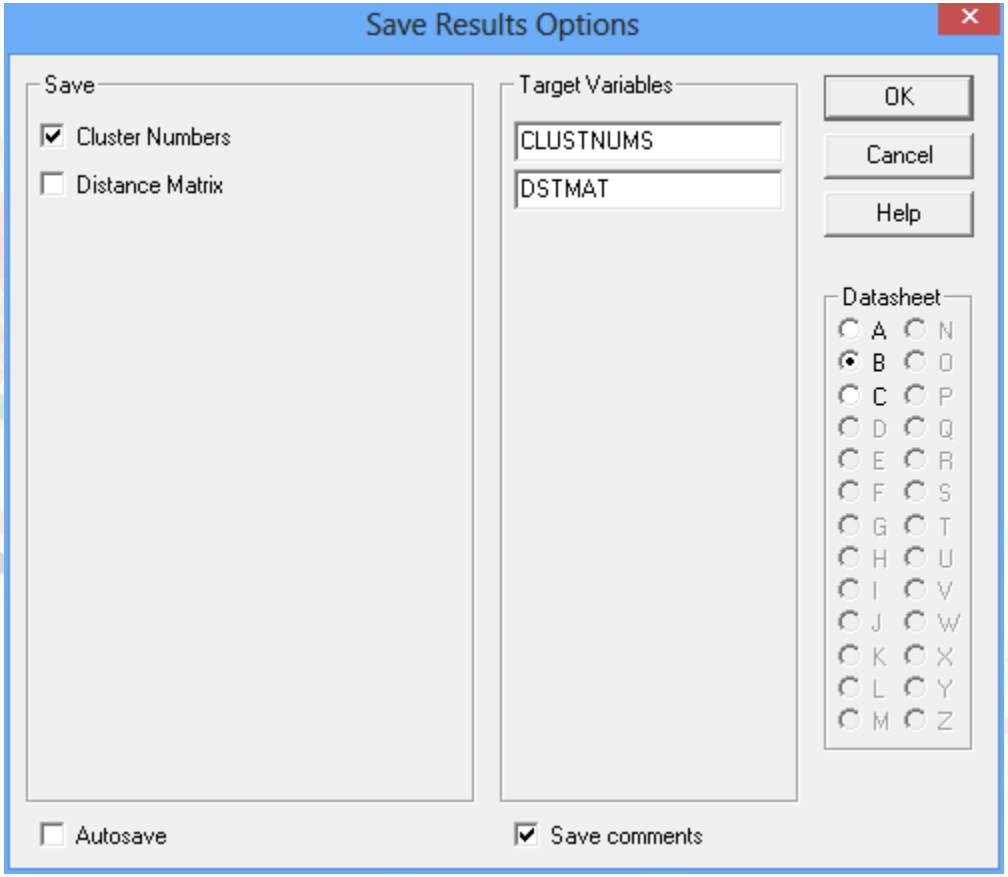
	Combined	Combined		Previous Stage	Previous Stage	Next
Stage	Cluster 1	Cluster 2	Distance	Cluster 1	Cluster 2	Stage
1	6	175	0.185525	0	0	48
2	30	44	0.372855	0	0	9
3	74	137	0.618574	0	0	69
4	37	134	0.89071	0	0	29
5	67	140	1.169	0	0	35
6	101	103	1.45245	0	0	42
7	52	57	1.75582	0	0	11
8	16	114	2.07097	0	0	20
9	9	30	2.39446	0	2	38
10	24	118	2.72627	0	0	22
11	52	84	3.08718	7	0	35
12	7	61	3.45595	0	0	94
13	109	171	3.82543	0	0	47
14	12	127	4.20585	0	0	62
15	73	126	4.60172	0	0	61
16	128	159	5.00619	0	0	25
17	100	160	5.4328	0	0	124
18	10	105	5.86132	0	0	49
19	54	89	6.29619	0	0	58
20	16	170	6.7368	8	0	30
21	34	35	7.2095	0	0	73
22	24	108	7.71322	10	0	92
23	46	60	8.23682	0	0	45
24	76	157	8.80092	0	0	36
25	55	128	9.37869	0	16	84
26	151	182	9.96316	0	0	33
27	63	116	10.5558	0	0	82

Creating 3 Clusters

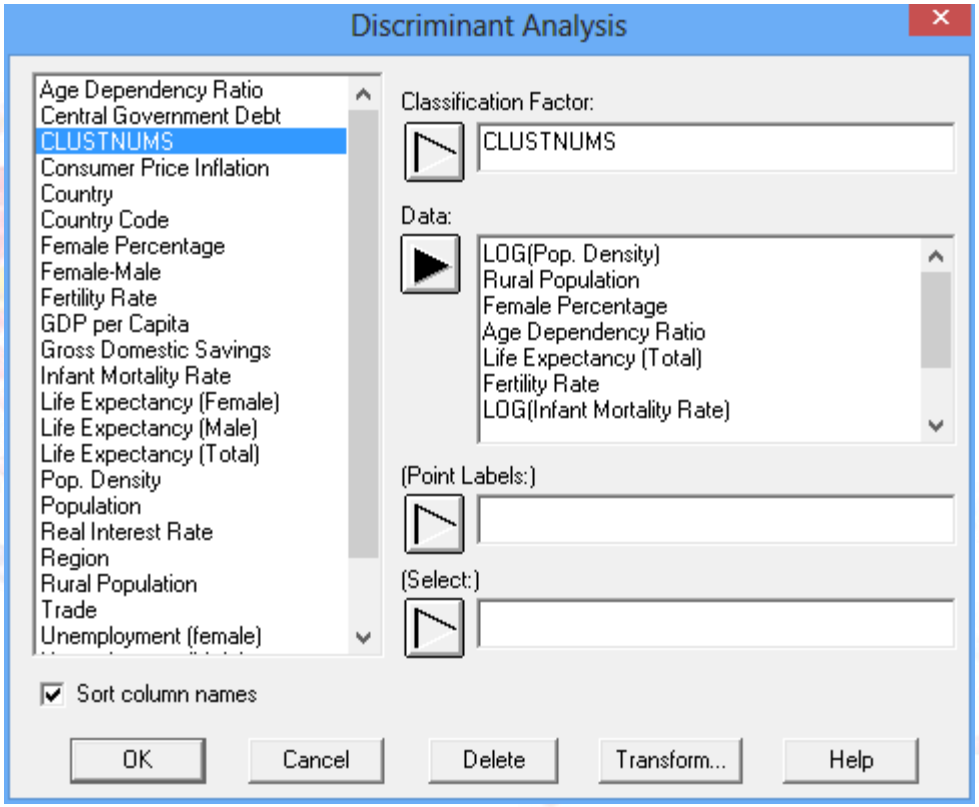


statgraphics®
statgraphics.com centurion

Save Cluster Numbers

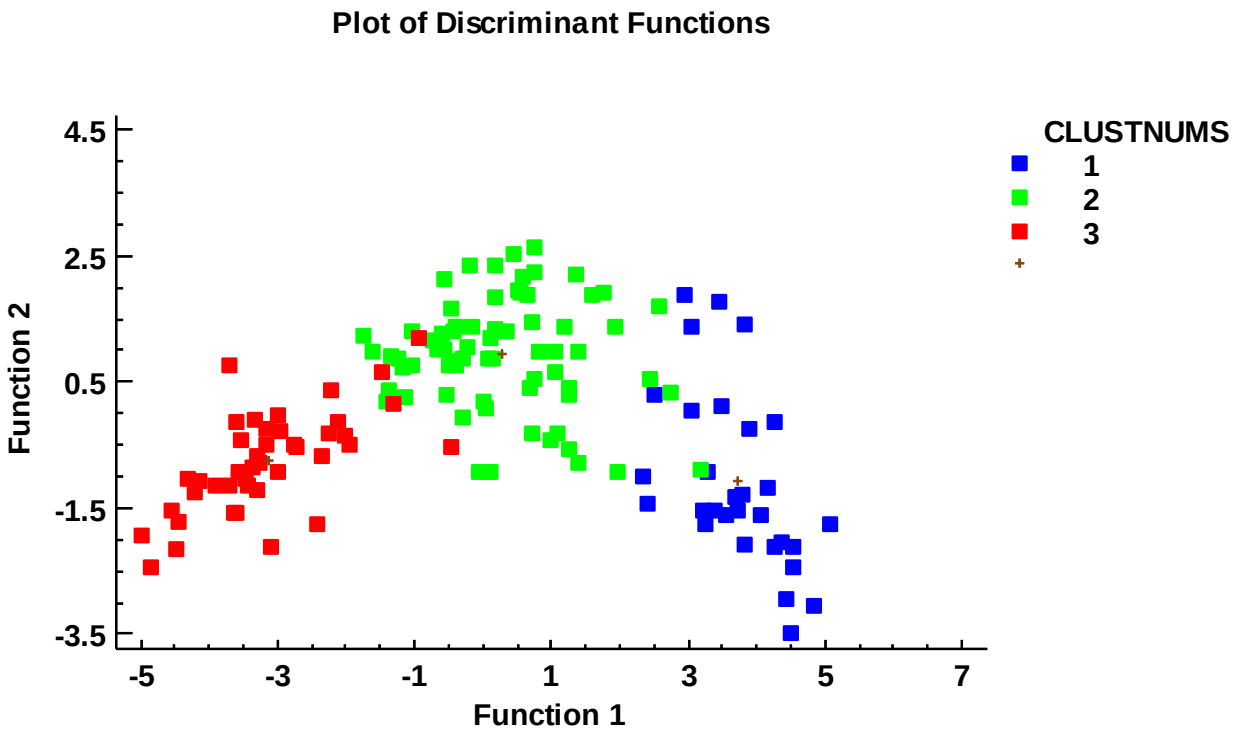


Discriminant Analysis



statgraphics®
centurion
statgraphics.com

Discriminant Function Plot

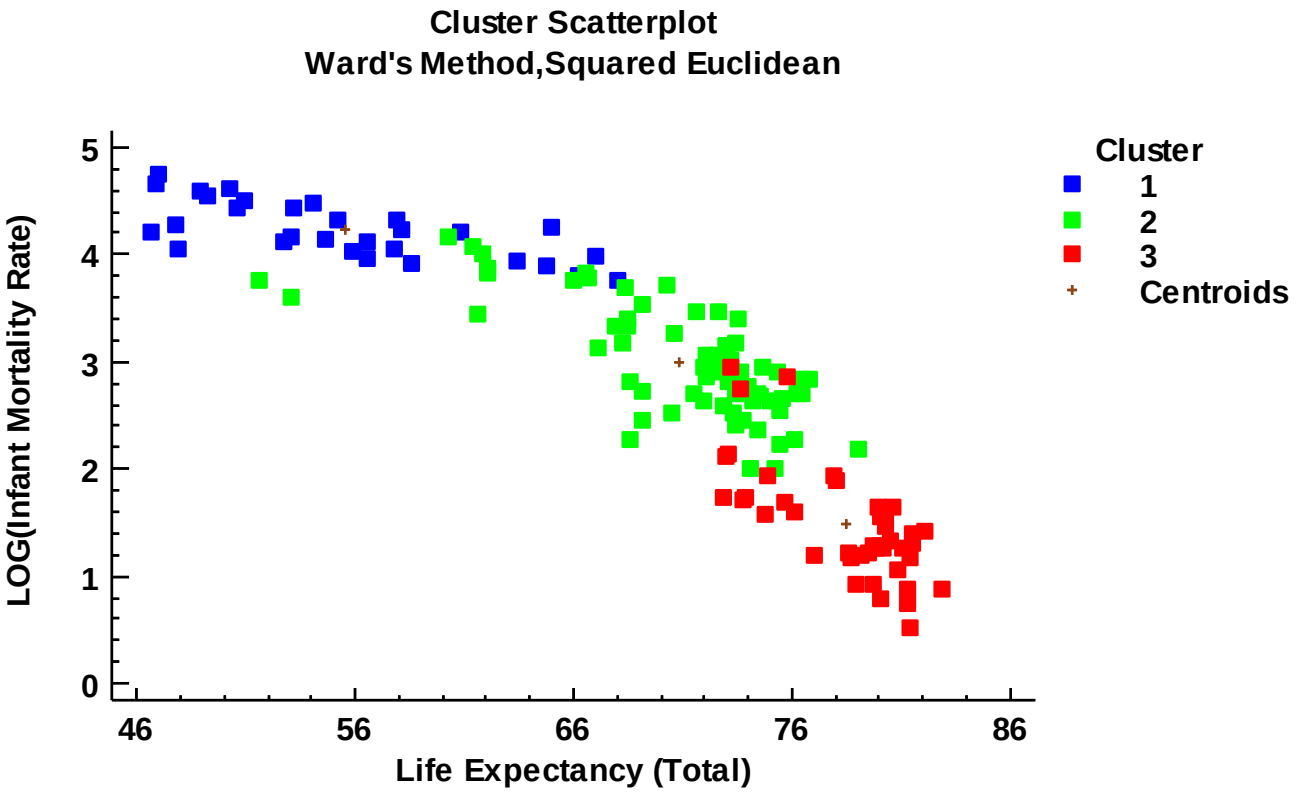


Discriminant Function Coefficients

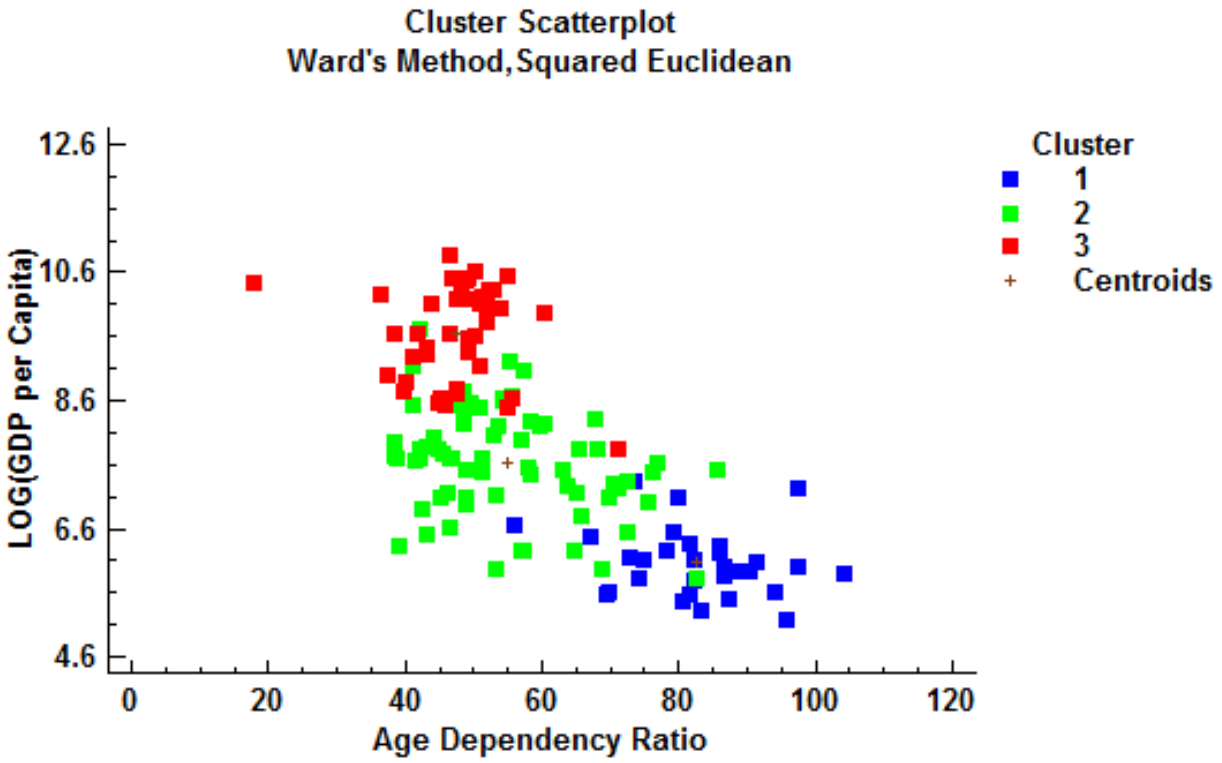
Discriminant Function Coefficients for CLUSTNUMS

	1	2
LOG(Pop. Density)	0.206202	-0.292762
Rural Population	0.118498	0.228847
Female Percentage	0.0495294	0.164767
Age Dependency Ratio	0.205199	-0.171682
Life Expectancy (Total)	-0.196611	0.884014
Fertility Rate	-0.0844127	-0.578296
LOG(Infant Mortality Rate)	0.61384	1.01239
Trade	-0.110333	0.0102956
LOG(GDP per Capita)	-0.253587	-0.0161386
Consumer Price Inflation	0.336886	-0.0141494

Cluster Scatterplot



Cluster Scatterplot



Method of k-Means

1. k observations are selected to be the initial seeds.
2. All remaining observations are assigned to cluster with nearest seed.
3. The centroids of each cluster are calculated.
4. Each observation is checked to see if it is closer to the centroid of another cluster. If so, it is switched to that cluster.
5. Step 4 is repeated until there are no further changes.

Seeds

- Select USA, China and India as initial seeds.

statgraphics® centurion

statgraphics.com

Cluster Summary

Cluster Summary

Cluster	*Members*	*Percent*
1	60	42.55
2	43	30.50
3	38	26.95

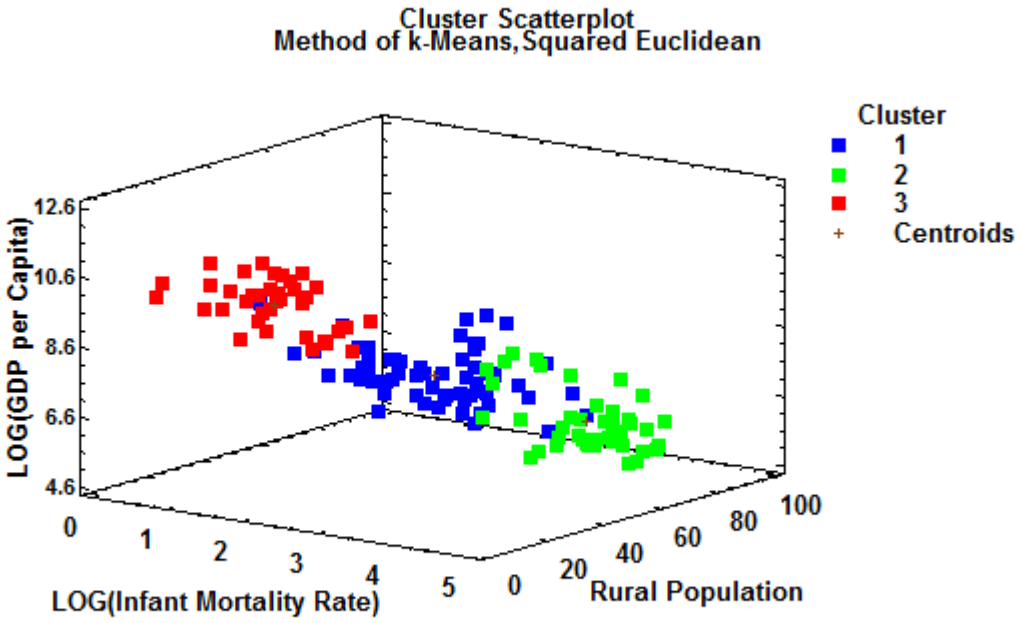
Centroids

Cluster	*LOG(Pop. Density)*	*Rural Population*	*Female Percentage*	*Age Dependency Ratio*	*Life Expectancy (Total)*
1	4.28565	40.7383	49.8538	50.986	72.4028
2	3.87485	64.6284	50.0347	79.2144	58.3921
3	4.59232	24.8284	50.9718	47.4158	78.875

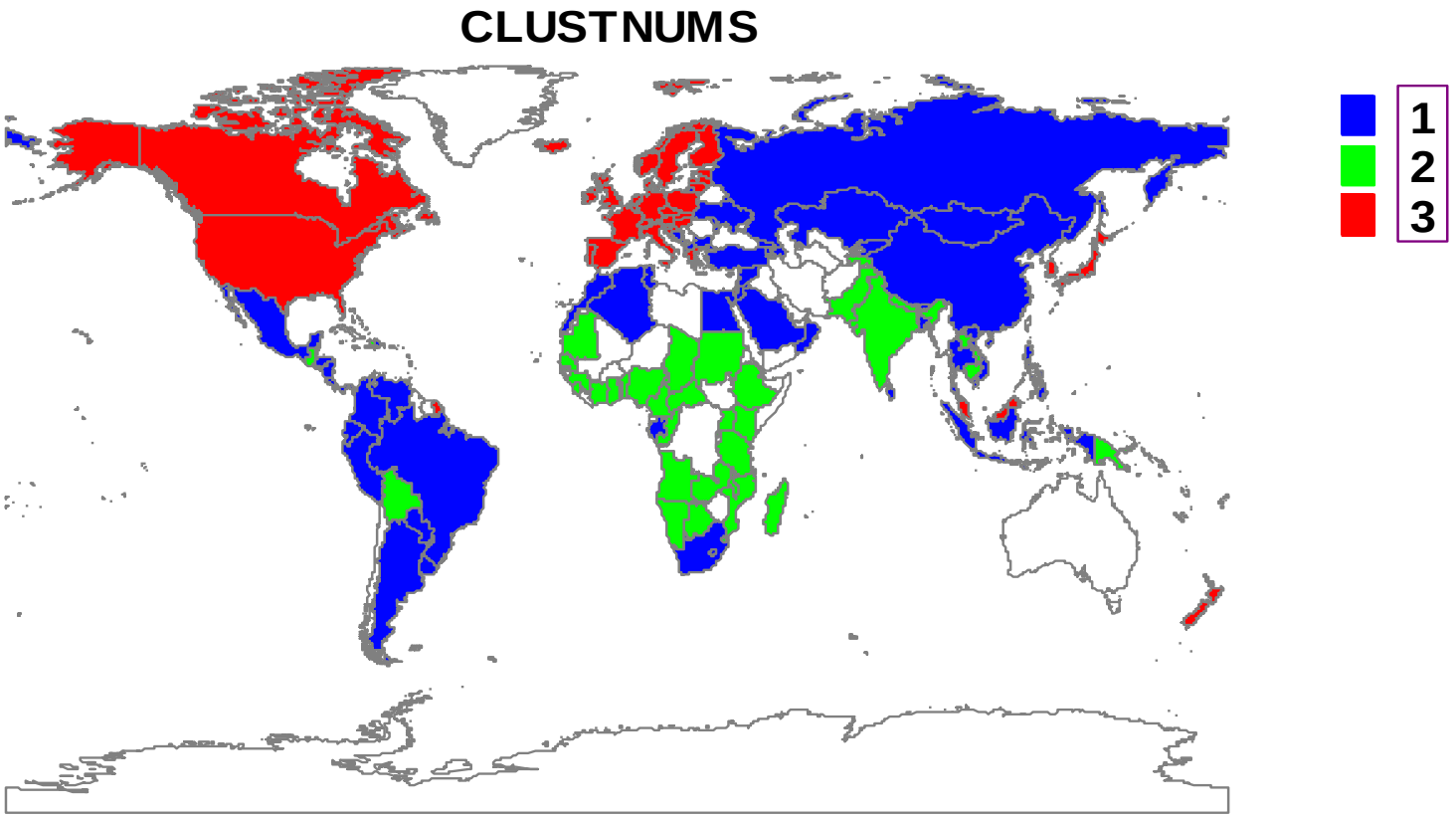
Cluster	*Fertility Rate*	*LOG(Infant Mortality Rate)*	*Trade*	*LOG(GDP per Capita)*	*Consumer Price Inflation*
1	2.221	2.86738	80.5958	7.86765	4.23267
2	4.40512	3.99465	74.7758	6.3524	6.78767
3	1.69026	1.37328	104.216	9.73032	1.30342

Cluster Scatterplot

USA=#3, China=#1, India=#2



World Map (coming soon)



statgraphics®
statgraphics.com centurion

References

- Johnson and Wichern (2012) Applied Multivariate Analysis, sixth edition.

- Copy of slides and presentation at:

  www.statgraphics.com/webinars