SESUG Proceedings (c) SESUG, Inc (http://www.sesug.org) The papers contained in the SESUG proceedings are the property of their authors, unless otherwise stated. Do not reprint without permission.
SEGUG papers are distributed freely as a courtesy of the Institute for Advanced Analytics (http://analytics.ncsu.edu).

Paper DM05

A METHODOLOGICAL APPROACH TO PERFORMING CLUSTER ANALYSIS WITH SAS®

William F. McCarthy

Maryland Medical Research Institute, Baltimore, MD

ABSTRACT

The purpose of this paper is to present an outline of steps that help to ensure that a cluster analysis is performed in a methodological manner. These steps allow for proper data exploration, verification and iteration. The approach presented is in the spirit of data mining/exploratory analysis. Each step in the outline is linked to SAS code and SAS output. This presentation will use a well-known data set from Rouncefield M. (1995) The Statistics of Poverty and Inequality, Journal of Statistics Education, Vol. 3, No. 2. The data considered are live birth rate per 1,000 of population, death rate per 1,000 of population, infant deaths per 1,000 of population, and country (97 countries considered).

METHODOLOGY

The following steps help to ensure that a cluster analysis is performed in a methodological manner that allows for proper data exploration, verification and iteration. The approach presented is in the spirit of data mining/exploratory analysis.

1. ***Look at the raw data graphically*** (in the spirit of John W. Tukey, Exploratory Data Analysis; Addison-Wesley Publishing Company, 1977) that will be used to perform the cluster analysis. ***Make scatter plots of each variable in a pair-wise manner***. The appearance of the scatter plots will inform you if your data may need transformation prior to cluster analysis. Data with poorly separated or elongated patterns need to be transformed. Also, variables with different units of measurement or with different size variances will need to be transformed as well. ***SAS provides various PROCS for such transformations prior to cluster analysis (PROC STANDARD, PROC ACECLUS)*** **[Refer to lines 67 thru 92 in the attached SAS code], [Refer to Figures 1,2,3].**

RESULTS: Figures 1,2 and 3 show patterns suggestive of elongated elliptical clusters. Thus, one needs to perform a linear transformation on the raw data before the cluster analysis.

2. ***Transform the data if required*** **[Refer to lines 94 thru 106 in the attached SAS code].**

RESULTS: PROC ACECLUS was used to preprocess the raw data subsequent to cluster analysis. PROC ACECLUS is used to obtain approximate estimates of the pooled within-cluster covariance matrix and to compute canonical variables for subsequent cluster analysis. The proportion of pairs used for estimating the within-cluster covariance was p= .03.

3. ***Perform the cluster analysis*** (on the transformed data if required, on the raw data if permitted) **[Refer to lines 107 thru 111 in the attached SAS code].**

RESULTS: PROC CLUSTER was used to perform an initial cluster analysis to estimate the minimal number of clusters that best accounts for the variability within the transformed data set.

4. ***Determine the number of clusters to use in the cluster analysis***. One can use the following approaches:

Plot a horizontal tree diagram with respect to R-Squared (Proportion of the Variance Explained) **[Refer to lines 119 thru 124 in the attached SAS code], [Refer to Figure 4]** *Use Occam's Razor*, the simplest explanation is the best – go with the minimal number of clusters that best accounts for the variability within the data set (unless subject matter suggests otherwise or interpretation is more meaningful with more);

Plot the results of the Cubic Clustering Criterion (CCC), Pseudo F and the T Squared **[Refer to lines 194 thru 204 in the attached SAS code], [Refer to Figures 5,6]** Take the first local maxima as the number of putative clusters to use for CCC; take the first local maxima as the number of putative clusters to use for Pseudo F; and take the first local maxima plus one as the number of putative clusters to use for T Squared;

Plot each of the variables used in the cluster analysis, in a pair-wise manner, with respect to the putative clusters to be used **[Refer to lines 130 thru 171 in the attached SAS code], [Refer to Figures 7,8,9]** Look for adequate discrimination between the putative clusters; which pair(s) of variables best discriminate the clusters? ; and

Plot the putative clusters in terms of the first two canonical variables **[Refer to lines 179 thru 187 in the attached SAS code], [Refer to Figure 10].** This allows you to see how much of the discrimination between the putative clusters (if any) is done by the first canonical variable; and how much is done (if any) by the second canonical variable.

RESULTS: Figures 4, 5, 6, 7, 8, 9 and 10 all showed that the 3 was the likely estimate of the minimal number of clusters that best accounts for the variability within the transformed data set.

5. ***Compare putative clusters using PROC UNIVARIATE: descriptive statistics and box plots*** **[Refer to lines 222 thru 239 in the attached SAS code], [Refer to attached SAS Output].** In a descriptive sense, do you have adequate discrimination of the putative clusters, do you see any outliers that might be confounding things, etc.

RESULTS: The PROC UNIVARIATE output shows, both with descriptive statistics and box plots, that there is adequate discrimination (separation) of the 3 clusters.

6. ***Look at who is contained in each putative cluster*** **[Refer to lines 211 thru 215 in the attached SAS code], [Refer to attached SAS Output].** From a subject matter perspective, do the elements making up each putative cluster make sense?

RESULTS: From a subject matter perspective, the countries contained in each of the 3 clusters make sense.

7. ***Based on diagnostics (in terms of statistics) and verification process (in terms of subject matter), loop back and make adjustments (if required) in terms of data trimming of outliers, variables to be used, transformations, number of putative clusters to be used, etc***.

RESULTS: No further adjustments to the cluster analysis were required.

The example presented used the agglomerative hierarchical clustering procedure with clusters determined by Ward's minimum-variance.

SAS CODE FOR METHODOLOGY

The data being analyzed is already in the form of a SAS dataset. This presentation will use a well-known data set from Rouncefield M. (1995) The Statistics of Poverty and Inequality. Journal of Statistics Education, Vol. 3, No. 2. The data considered are live birth rate per 1,000 of population, death rate per 1,000 of population, infant deaths per 1,000 of population, and country (97 countries considered).

```
 1 data Poverty;
 2      input Birth Death InfantDeath Country $20. @@;
 3      datalines;
 4   24.7  5.7  30.8 Albania              12.5 11.9  14.4 Bulgaria
 5   13.4 11.7  11.3 Czechoslovakia       12   12.4   7.6 Former_E._Germany
 .                                        .
 .                                        .
 .                                        .
52   41.7 10.3    66 Zimbabwe
53 ;
54 run;
55 ************************************************************************;
56 *                                                                      *;
57 * It is often useful when beginning a cluster analysis to look at the data     *;
58 * graphically. The following statements use the GPLOT procedure to make a      *;
59 * scatter plot of the variables Birth and Death.                       *;
60 *                                                                      *;
61 * Plots of the other variable pairs should be done as well.            *;
62 * The clusters that comprise these data may be poorly separated and elongated. *;
63 * Data with poorly separated or elongated clusters must be transformed.        *;
64 *                                                                      *;
65 ************************************************************************;
66
67   axis1 label=(angle=90 rotate=0) minor=none;
68   axis2 minor=none;
69 * Birth*Death;
70   proc gplot data=poverty;
71      plot Birth*Death/
72         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
73   run;
74   quit;
```

```
 75
 76   axis1 label=(angle=90 rotate=0) minor=none;
 77   axis2 minor=none;
 78 * Birth*InfantDeath;
 79   proc gplot data=poverty;
 80      plot Birth*InfantDeath/
 81         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
 82   run;
 83   quit;
 84
 85   axis1 label=(angle=90 rotate=0) minor=none;
 86   axis2 minor=none;
 87 * Death*InfantDeath;
 88   proc gplot data=poverty;
 89      plot Death*InfantDeath/
 90         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
 91   run;
 92   quit;
 93
 94 ********************************************************************************;
 95 *                                                                              *;
 96 * If you know the within-cluster covariances, you can transform the data to make   *;
 97 * the clusters spherical. However, since you do not know what the clusters are,    *;
 98 * you cannot calculate exactly the within-cluster covariance matrix. The ACECLUS   *;
 99 * procedure estimates the within-cluster covariance matrix to transform the data,  *;
100 * even when you have no knowledge of cluster membership or the number of clusters. *;
101 *                                                                              *;
102 ********************************************************************************;
103
104 proc aceclus data=Poverty out=Ace p=.03;
105      var Birth Death InfantDeath;
106   run;
107 proc cluster data=Ace outtree=Tree method=ward ccc pseudo;
108      var can1 can2 can3 ;
109       id Country;
110      copy Birth--Country;
111   run;

112 goptions vsize=8in htext=1pct htitle=2.5pct;
113   axis1 order=(0 to 1 by 0.2);
114 ********************************************************************************;
115 *                                                                              *;
116 * PLOTTING HORIZONTAL TREE DIAGRAM WITH RESPECT TO R_SQUARED                   *;
117 *                                                                              *;
118 ********************************************************************************;
119 proc tree data=Tree out=New nclusters=3
120             graphics haxis=axis1 horizontal;
121      height _rsq_;
122      copy can1 can2 ;
123      id country;
124   run;
125 ********************************************************************************;
126 *                                                                              *;
127 * PROC TREE FOR PLOTS BELOW                                                    *;
128 *                                                                              *;
129 ********************************************************************************;
130 proc tree data=tree out=New nclusters=3 noprint;
131      copy Birth Death InfantDeath can1 can2 ;
132      id Country;
133   run;
134
135 ********************************************************************************;
136 * The following statements invoke the GPLOT procedure, using the SAS data set      *;
137 * created by PROC TREE.                                                        *;
138 * The first set of plot statements requests a scatter plot of the two variables    *;
139 * Birth and Death, etc using the variable CLUSTER as the identification variable.  *;
140 *                                                                              *;
141 * The second PLOT statement requests a plot of the two canonical variables,        *;
142 * using the value of the variable CLUSTER as the identification variable.          *;
143 *                                                                              *;
144 ********************************************************************************;
145 *Birth*Death=cluster;
146   legend1 frame cframe=ligr cborder=black position=center
```

```
147      value=(justify=center);
148   axis1 label=(angle=90 rotate=0) minor=none;
149   axis2 minor=none;
150   proc gplot data=New;
151      plot Birth*Death=cluster/
152         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
153   run;
154 * Birth*InfantDeath=cluster;
155      legend1 frame cframe=ligr cborder=black position=center
156      value=(justify=center);
157   axis1 label=(angle=90 rotate=0) minor=none;
158   axis2 minor=none;
159   proc gplot data=New;
160      plot Birth*InfantDeath=cluster/
161         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
162   run;
163 * Death*InfantDeath=cluster;
164      legend1 frame cframe=ligr cborder=black position=center
165      value=(justify=center);
166   axis1 label=(angle=90 rotate=0) minor=none;
167   axis2 minor=none;
168   proc gplot data=New;
169      plot Death*InfantDeath=cluster/
170         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
171   run;
172
173 ********************************************************************************;
174 *                                                                              *;
175 * PLOTTING CLUSTERS IN TERMS OF CAN1 AND CAN2                                  *;
176 *                                                                              *;
177 ********************************************************************************;
178
179 legend1 frame cframe=ligr cborder=black
180           position=center value=(justify=center);
181
182   axis1 label=(angle=90 rotate=0) minor=none order=(-10 to 20 by 5);
183   axis2 minor=none order=(-10 to 20 by 5);
184   proc gplot data=New ;
185      plot can2*can1=cluster/frame cframe=ligr
186                      legend=legend1 vaxis=axis1 haxis=axis2;
187   run;
188 ************************************************************************;
189 *                                                                      *;
190 *  CCC, PSEUDO F AND T_SQUARED PLOTS                                   *;
191 *                                                                      *;
192 ************************************************************************;
193
194 legend1 frame cframe=ligr cborder=black
195           position=center value=(justify=center);
196   axis1 label=(angle=90 rotate=0) minor=none order=(0 to 600 by 100);
197   axis2 minor=none order=(1 to 30 by 1);
198   axis3 label=(angle=90 rotate=0) minor=none order=(0 to 7 by 1);
199   proc gplot data=tree;
200      plot _ccc_*_ncl_  /
201         frame cframe=ligr legend=legend1 vaxis=axis3 haxis=axis2;
202      plot _psf_*_ncl_  _pst2_*_ncl_  /overlay
203         frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
204    run;
205 ************************************************************************;
206 *                                                                      *;
207 *  LOOK AT WHICH COUNTRIES ARE IN EACH CLUSTER                         *;
208 *                                                                      *;
209 ************************************************************************;
210
211 proc sort data=new;
212 by cluster;
213 run;
214 proc print data=new;
215 run;
216
```

```
217 ********************************************************************************;
218 *                                                                              *;
219 * COMPARING CLUSTERS WITH PROC UNIVARIATE                                      *;
220 *                                                                              *;
221 ********************************************************************************;
222 proc sort data=poverty;
223 by country;
224 run;
225 proc sort data=new;
226 by country;
227 run;
228 data compare;
229 merge poverty new;
230 by country;
231 run;
232 proc sort data=compare;
233 by cluster;
234 run;
235 proc univariate data=compare plots;
236 by cluster;
237 var Birth Death InfantDeath;
238 where cluster in (1,2,3);
239 run;
```

SAS OUTPUT (SELECTIVE EXAMPLES)

Figure 1.

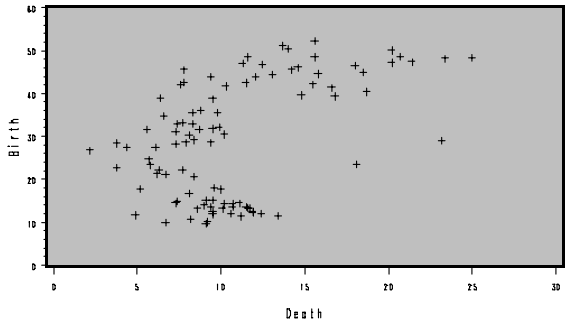


Figure 2.

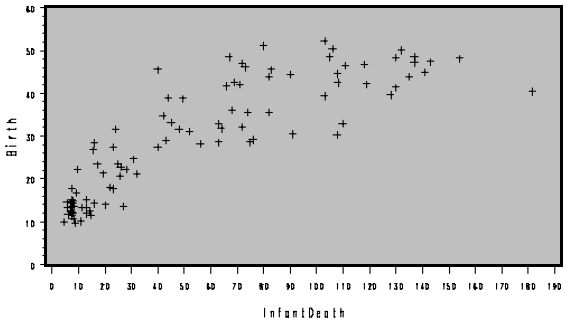
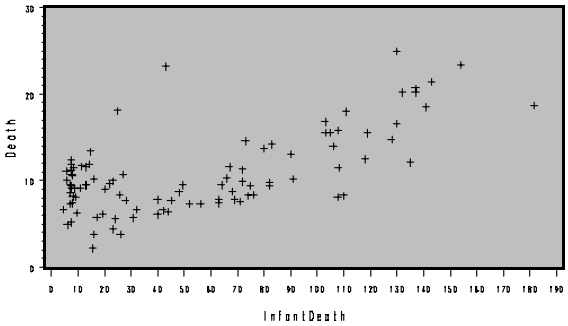


Figure 3.



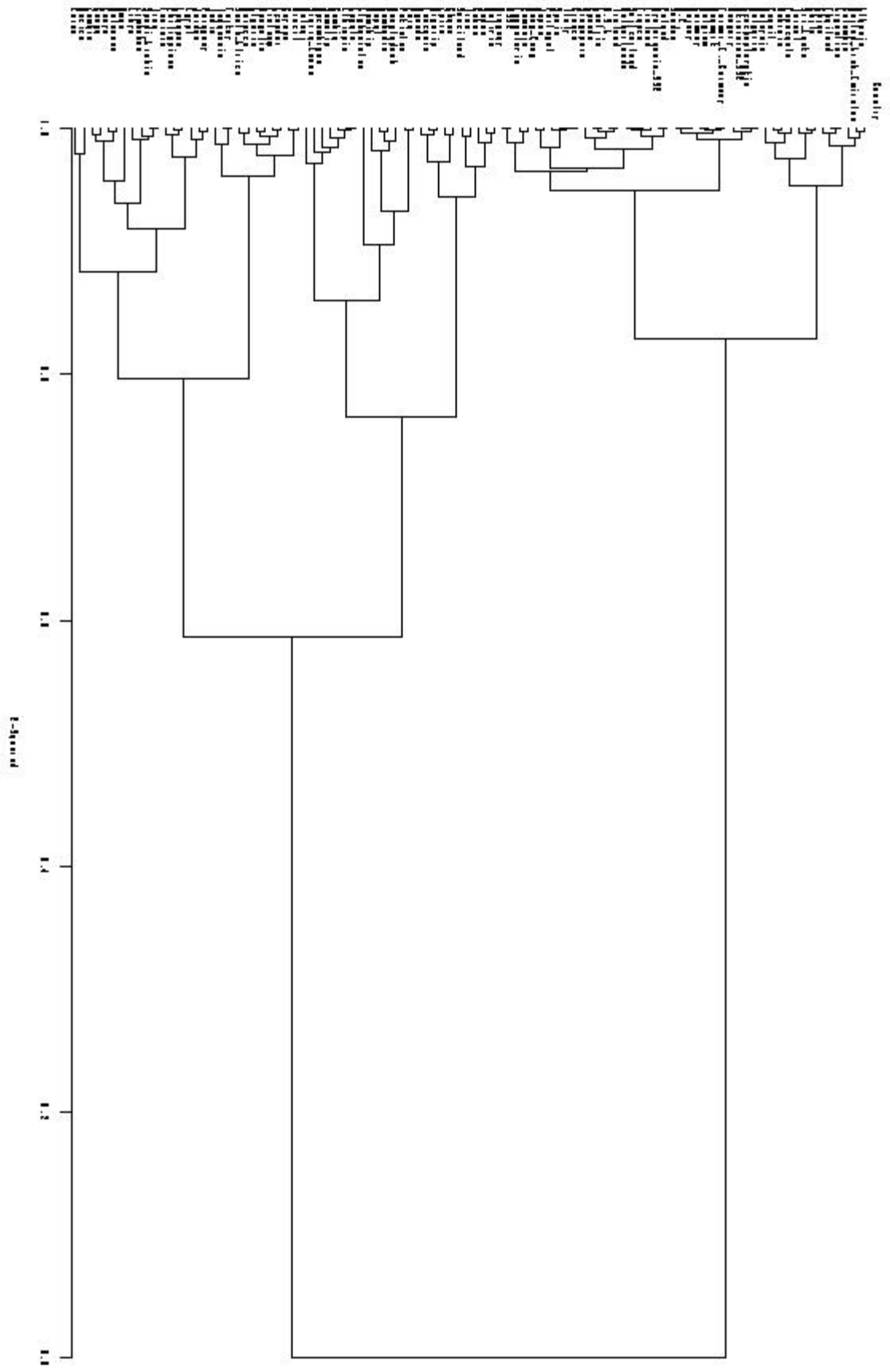


Figure 5.

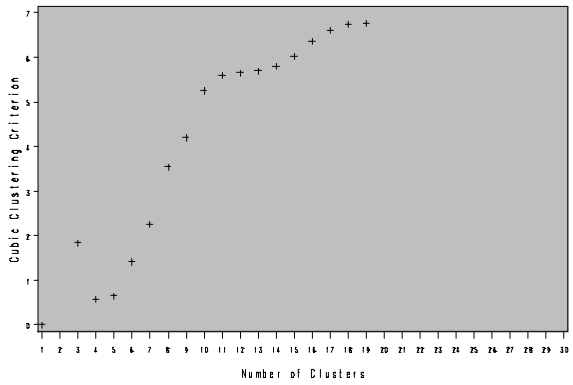


Figure 6.

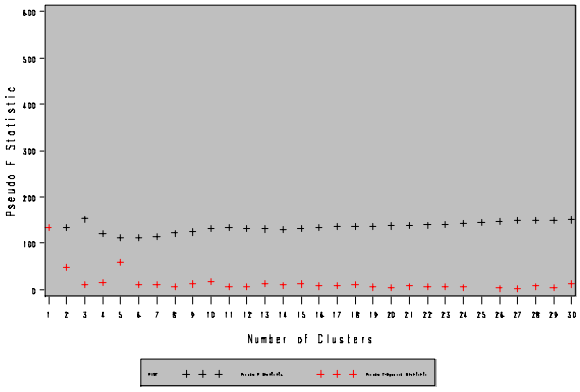


Figure 7.

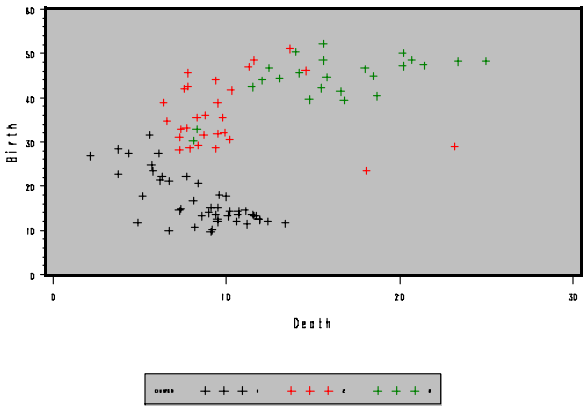


Figure 8.

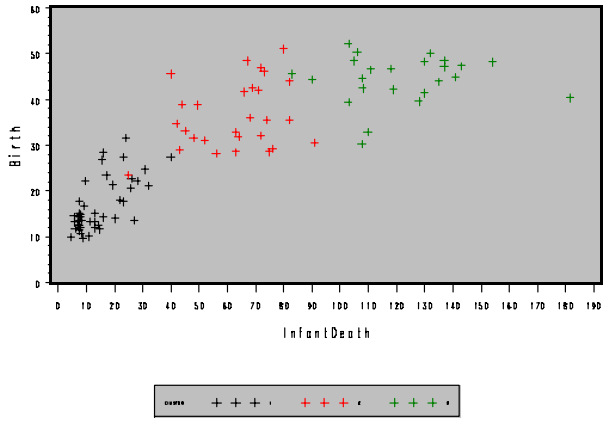


Figure 9.

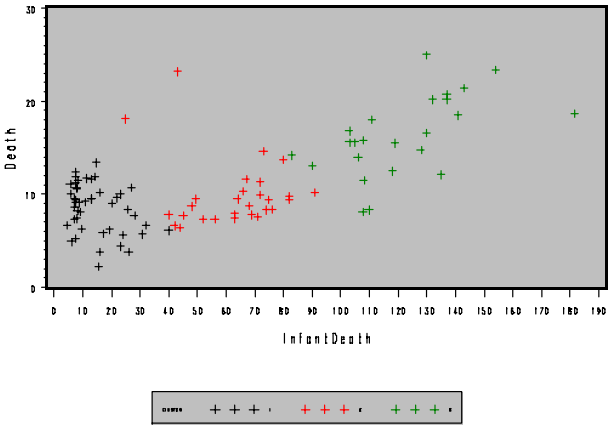
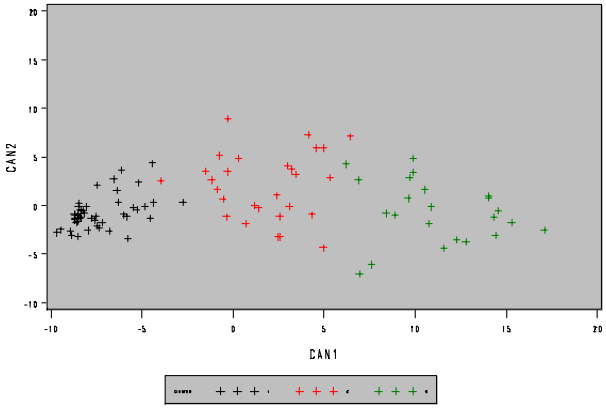


Figure 10.



The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	64.5500051	54.7313223	0.8091	0.8091
2	9.8186828	4.4038309	0.1231	0.9321
3	5.4148519		0.0679	1.0000

Root-Mean-Square Total-Sample Standard Deviation = 5.156987
Root-Mean-Square Distance Between Observations = 12.63199

Obs	Country	Birth	Death	Infant Death	Can1	Can2	CLUSTER	CLUSNAME
1	Austria	14.9	7.4	8.0	-8.49023	-0.84518	1	CL5
2	Canada	14.5	7.3	7.2	-8.65908	-0.93748	1	CL5
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
44	Columbia	27.4	6.1	40.0	-2.75049	0.37902	1	CL5
45	Malaysia	31.6	5.6	24.0	-4.43240	4.35460	1	CL5
46	Iraq	42.6	7.8	69.0	3.24949	3.72944	2	CL4
47	Saudi_Arabia	42.1	7.6	71.0	3.43441	3.20093	2	CL4
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
72	Korea	23.5	18.1	25.0	-3.9601	2.5664	2	CL4
73	Oman	45.6	7.8	40.0	-0.2801	8.9755	2	CL4
74	Angola	47.2	20.2	137.0	14.3314	-1.2233	3	CL3
75	Ethiopia	48.6	20.7	137.0	14.5614	-0.5016	3	CL3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
96	Malawi	48.3	25.0	130.0	14.0470	1.0009	3	CL3
97	Afghanistan	40.4	18.7	181.6	19.3225	-10.5363	3	CL3

---------- CLUSTER=1 ----------

The UNIVARIATE Procedure
Variable: Birth

Moments

N	45	Sum Weights	45
Mean	16.6622222	Sum Observations	749.8
Std Deviation	5.73792364	Variance	32.9237677
Skewness	0.98132575	Kurtosis	-0.0872774
Uncorrected SS	13941.98	Corrected SS	1448.64578
Coeff Variation	34.436725	Std Error Mean	0.85535915

Basic Statistical Measures

Location		Variability	
Mean	16.66222	Std Deviation	5.73792
Median	14.30000	Variance	32.92377
Mode	13.60000	Range	21.90000
		Interquartile Range	8.70000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	19.4798	Pr > |t|	<.0001
Sign	M	22.5	Pr >= |M|	<.0001
Signed Rank	S	517.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	31.6
99%	31.6
95%	27.5
90%	26.8
75% Q3	21.2
50% Median	14.3
25% Q1	12.5
10%	11.4
5%	10.1
1%	9.7
0% Min	9.7

---------- CLUSTER=1 ----------

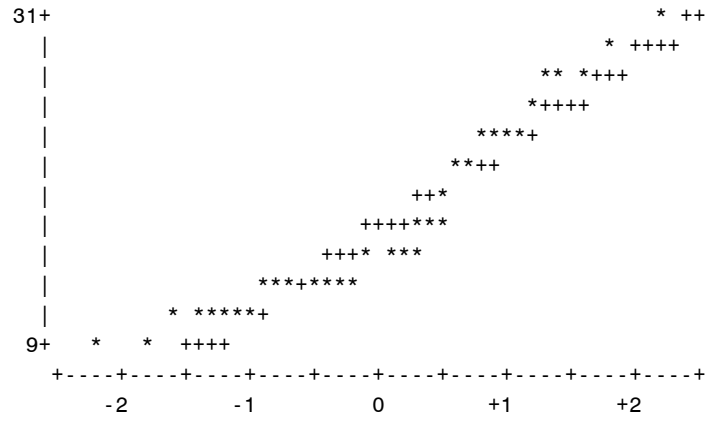
The UNIVARIATE Procedure
Variable: Birth

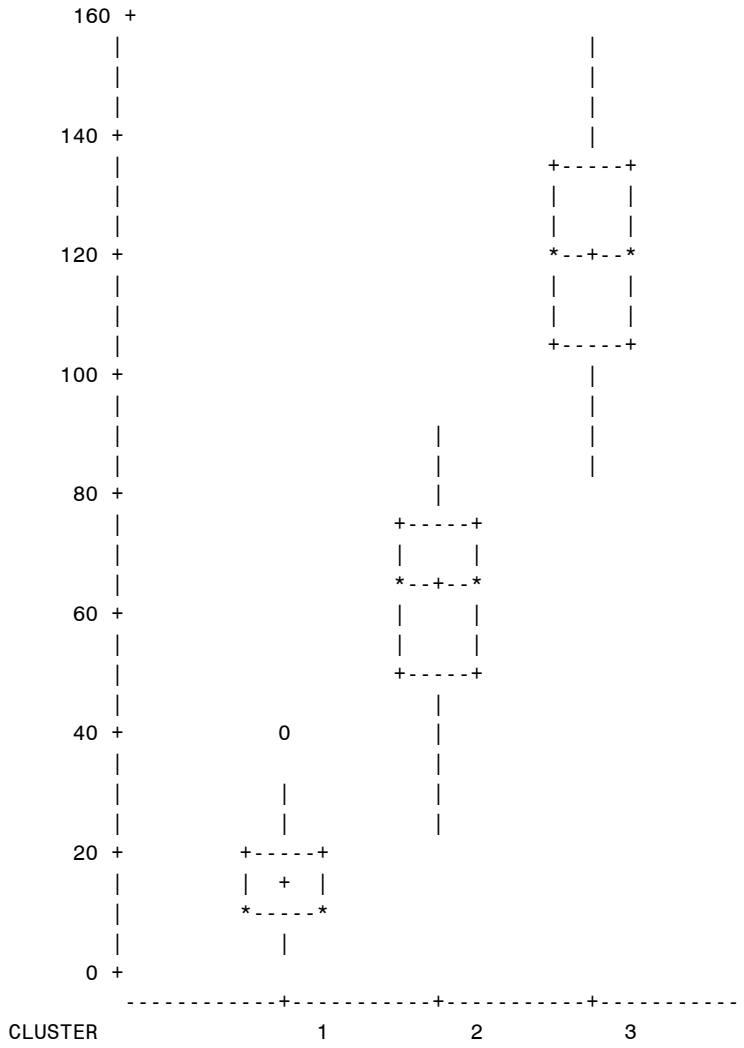
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
9.7	23	26.8	25
9.9	24	27.4	11
10.1	18	27.5	44
10.7	33	28.4	4
11.4	17	31.6	26

```
Stem Leaf                     #             Boxplot
  30 6                        1                |
  28 4                        1                |
  26 845                      3                |
  24 7                        1                |
  22 3384                     4                |
  20 723                      3             +-----+
  18 0                        1             |     |
  16 778                      3             |  +  |
  14 03355912                 8             *-----*
  12 004552244666            12             +-----+
  10 174679                   6                |
   8 79                       2                |
     ----+----+----+----+
```

Normal Probability Plot





CONTACT INFORMATION

William F. McCarthy
Director, Clinical Trials, Clinical Statistics and SAS Programming
Principal Statistician
Maryland Medical Research Institute
600 Wyndhurst Avenue
Baltimore, Maryland 21210-2425
(410) 435-4200
wmccarthy@mmri.org

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.