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?
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).

```sas
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 axis1 label=(angle=90 rotate=0) minor=none;
67 axis2 minor=none;
68 * Birth*Death;
69 proc gplot data=poverty;
70   plot Birth*Death/                 frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
71 run;
72 quit;
```
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
* Birth*InfantDeath;
proc gplot data=poverty;
   plot Birth*InfantDeath/
      frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;
quit;

axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
* Death*InfantDeath;
proc gplot data=poverty;
   plot Death*InfantDeath/
      frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;
quit;

************************************************************************************;
*                                                                                  *
* If you know the within-cluster covariances, you can transform the data to make   *
* the clusters spherical. However, since you do not know what the clusters are,   *
* you cannot calculate exactly the within-cluster covariance matrix. The ACECLUS   *
* procedure estimates the within-cluster covariance matrix to transform the data, *
* even when you have no knowledge of cluster membership or the number of clusters. *
*                                                                                  *
* **********************************************************************************;

proc aceclus data=Poverty out=Ace p=.03;
   var Birth Death InfantDeath;
run;

proc cluster data=Ace outtree=Tree method=ward ccc pseudo;
   var can1 can2 can3;
   id Country;
   copy Birth--Country;
run;

goptions vsize=8in htext=1pct htitle=2.5pct;
axis1 order=(0 to 1 by 0.2);

**********************************************************************************;
*                                                                                *
* PLOTTING HORIZONTAL TREE DIAGRAM WITH RESPECT TO R_SQUARED                     *
*                                                                                *
**** **** *
**********************************************************************************;

proc tree data=Tree out=New nclusters=3 graphics haxis=axis1 horizontal;
   height _rsq_;
   copy can1 can2;
   id Country;
run;

***********************************************************************************;
*                                                                                 *
* PROC TREE FOR PLOTS BELOW                                                       *
*                                                                                 *
***********************************************************************************;

proc tree data=tree out=New nclusters=3 noprint;
   copy Birth Death InfantDeath can1 can2;
   id Country;
run;

***********************************************************************************;
* The following statements invoke the GLOT procedure, using the SAS data set      *
* created by PROC TREE.                                                          *
* The first set of plot statements requests a scatter plot of the two variables   *
* Birth and Death, etc using the variable CLUSTER as the identification variable. *
* The second PLOT statement requests a plot of the two canonical variables,       *
* using the value of the variable CLUSTER as the identification variable.         *
*                                                                                *
*Birth*Death=cluster;
legend1 frame cframe=ligr cborder=black position=center
axis1 label=(angle=90 rotate=0) minor=none;
proc gplot data=New;
   plot Birth*Death=cluster/
   frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;

* Birth*InfantDeath=cluster;
legend frame cframe=ligr cborder=black position=center
value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
proc gplot data=New;
   plot Birth*InfantDeath=cluster/
   frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;

* Death*InfantDeath=cluster;
legend frame cframe=ligr cborder=black position=center
value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
proc gplot data=New;
   plot Death*InfantDeath=cluster/
   frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;

*************************************************************************************;
*                                                                                   *
* PLOTTING CLUSTERS IN TERMS OF CAN1 AND CAN2                                       *
*                                                                                   *
*************************************************************************************;

legend1 frame cframe=ligr cborder=black
   position=center value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none order=(-10 to 20 by 5);
axis2 minor=none order=(-10 to 20 by 5);
proc gplot data=New;
   plot can2*can1=cluster/frame cframe=ligr
   legend=legend1 vaxis=axis1 haxis=axis2;
run;

*******************************************************************************;
*                                                                        *
*  CCC, PSEUDO F AND T_SQUARED PLOTS                                     *
*                                                                        *
*******************************************************************************;

legend1 frame cframe=ligr cborder=black
   position=center value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none order=(0 to 600 by 100);
axis2 minor=none order=(1 to 30 by 1);
axis3 label=(angle=90 rotate=0) minor=none order=(0 to 7 by 1);
proc gplot data=tree;
   plot _ccc_*_ncl_/ frame cframe=ligr legend=legend1 vaxis=axis3 haxis=axis2;
   plot _psf_*_ncl_ _pst2_*_ncl_ /overlay
   frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;

*******************************************************************************;
*                                                                        *
*  LOOK AT WHICH COUNTRIES ARE IN EACH CLUSTER                           *
*                                                                        *
*******************************************************************************;

proc sort data=new;
by cluster;
run;
proc print data=new;
run;
proc sort data=poverty;
  by country;
run;
proc sort data=new;
  by country;
run;
data compare;
merge poverty new;
  by country;
run;
proc sort data=compare;
  by cluster;
run;
proc univariate data=compare plots;
  by cluster;
var Birth Death InfantDeath;
where cluster in (1,2,3);
run;
SAS OUTPUT (SELECTIVE EXAMPLES)

Figure 1.

Figure 2.

Figure 3.
Figure 5.

Figure 6.

Figure 7.
The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64.5500051</td>
<td>0.8091</td>
<td>0.8091</td>
</tr>
<tr>
<td>2</td>
<td>9.8186828</td>
<td>0.1231</td>
<td>0.9321</td>
</tr>
<tr>
<td>3</td>
<td>5.4148519</td>
<td>0.0679</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

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

Infant

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

Moments

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>45</td>
<td>Sum Weights</td>
<td>45</td>
</tr>
<tr>
<td>Mean</td>
<td>16.6622222</td>
<td>Sum Observations</td>
<td>749.8</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>5.73792364</td>
<td>Variance</td>
<td>32.9237677</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.98132575</td>
<td>Kurtosis</td>
<td>-0.0872774</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>13941.98</td>
<td>Corrected SS</td>
<td>1448.64578</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>34.436725</td>
<td>Std Error Mean</td>
<td>0.85535915</td>
</tr>
</tbody>
</table>

Basic Statistical Measures

Location | Variability
---------|---------
Mean     | 16.6622222 |
Median   | 14.30000   |
Mode     | 13.60000   |

Tests for Location: Mu0=0

<table>
<thead>
<tr>
<th>Test</th>
<th>-Statistic-</th>
<th>-----p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student's t</td>
<td>t</td>
<td>19.4798</td>
</tr>
<tr>
<td>Sign</td>
<td>M</td>
<td>22.5</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S</td>
<td>517.5</td>
</tr>
</tbody>
</table>

Quantiles (Definition 5)

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>31.6</td>
</tr>
<tr>
<td>99%</td>
<td>31.6</td>
</tr>
<tr>
<td>95%</td>
<td>27.5</td>
</tr>
<tr>
<td>90%</td>
<td>26.8</td>
</tr>
<tr>
<td>75% Q3</td>
<td>21.2</td>
</tr>
<tr>
<td>50% Median</td>
<td>14.3</td>
</tr>
<tr>
<td>25% Q1</td>
<td>12.5</td>
</tr>
<tr>
<td>10%</td>
<td>11.4</td>
</tr>
<tr>
<td>5%</td>
<td>10.1</td>
</tr>
<tr>
<td>1%</td>
<td>9.7</td>
</tr>
<tr>
<td>0% Min</td>
<td>9.7</td>
</tr>
</tbody>
</table>
The UNIVARIATE Procedure
Variable: Birth

Extreme Observations

<table>
<thead>
<tr>
<th>Value</th>
<th>Obs</th>
<th>Value</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.7</td>
<td>23</td>
<td>26.8</td>
<td>25</td>
</tr>
<tr>
<td>9.9</td>
<td>24</td>
<td>27.4</td>
<td>11</td>
</tr>
<tr>
<td>10.1</td>
<td>18</td>
<td>27.5</td>
<td>44</td>
</tr>
<tr>
<td>10.7</td>
<td>33</td>
<td>28.4</td>
<td>4</td>
</tr>
<tr>
<td>11.4</td>
<td>17</td>
<td>31.6</td>
<td>26</td>
</tr>
</tbody>
</table>

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

---+---+---+---+
-2  -1  0   +1  +2
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.