

Visual Representation of Longitudinal Change in Cluster Analysis with SAS®

Robert Bauserman, Maryland Medical Research Institute, Baltimore, MD

Ann Albertson, Bell Institute for Research In Nutrition, Minneapolis, MN

Nort Holschuh, Bell Institute for Research In Nutrition, General Mills, Minneapolis, MN

William McCarthy, Maryland Medical Research Institute, Baltimore, MD

ABSTRACT

When cluster analysis is done with data from a single point in time, it is simple to identify and describe mutually exclusive clusters. However, many designs call for multiple measurements over time for each subject. One way to cluster analyze such data sets is to separately cluster data from each time point. This approach may lead to complex results. The optimal number of clusters (indicated by diagnostic measures such as the cubic clustering criterion) may change over time. The more clusters at each time point, the greater the number of possible pathways that participants may follow over time. This presentation uses PROC CLUSTER results from a longitudinal study of dietary patterns in childhood and adolescence to illustrate how a single graphic in Word® can show migration of subjects between dietary clusters over time. The same approach can be used with longitudinal sets of cluster analyses based on any characteristics of interest.

INTRODUCTION

Cluster analysis is typically done with data from a single point in time for each subject. However, many research designs call for multiple measurements at specified points in time for each subject. One way to perform cluster analysis with such data sets is to separately analyze the data from each time point. When data is collected from multiple time points, and one is interested in describing both (1) the cluster structure of the data at each time point and (2) how study participants shift between clusters over time, then it is necessary to decide how best to efficiently present such data. Is it better to use tables or figures, and how does one structure the presentation format to best show the flow of participants between clusters over time?

Cluster analyzing each time point separately avoids the problem of correlated observations due to the inclusion of the same subjects at multiple points in time. It also acknowledges that descriptive characteristics of the clusters may not be consistent over time, because it does not try to force measurements from different time points into the same cluster. However, this approach may lead to complex results. The optimal number of clusters, as indicated by diagnostic measures such as the cubic clustering criterion, the pseudo-F statistics, and the pseudo-t statistic, may shift from one time point to another. Also, at any given time point, specific clusters may closely match clusters from earlier or later time points in terms of descriptive variables of interest; however, there is no guarantee that they will do so.

A likely goal of such longitudinal analysis is to identify and describe how study participants change over time, by moving into specific clusters at each time point. The challenge addressed in this paper is how best to summarize and represent such change. Tabular representations of all possible patterns may be complex and difficult for readers to follow or visualize. Consequently, graphical presentations may be the best solution. In addition, the ODS system can create formatted output files that can be linked directly to the graph with embedded hyperlinks. Below we present cluster analyses of multi-year dietary data from adolescent girls and boys, along with a graphical approach to present the cluster structure and link to descriptive output of the clusters.

GRAPHIC REPRESENTATION OF CLUSTER ANALYSIS AT MULTIPLE TIME POINTS

The dietary data used here is from 418 participants (215 males and 203 females) with 5 full years of food diary records in the DISC study (Dietary Intervention Study in Children). This was a clinical intervention designed to lower cholesterol in boys and girls with high ($\geq 85^{\text{th}}$ percentile) cholesterol levels for their age and sex. The intervention focused on nutritional education to emphasize healthier food choices (lower in saturated fats and cholesterol) from multiple food groups.

The data set for the DISC study included detailed dietary records from five different years for all participants included in the analyses. For each year, 3 days of detailed breakfast food records were analyzed and categorized, resulting in a single observation for each subject for that year. This observation included the average daily number of servings of

“good” and “bad” foods (called “Go” and “Whoa” foods in the intervention) from each of seven categories: grains/breads, dairy, meat/fish/poultry, fats/oils, fruits, vegetables, and snacks).

The statistical procedure used to analyze the data from each year is a SAS® macro developed in-house at MMRI that invokes the following PROCs: ACECLUS, CLUSTER, TREE, and GPLOT. Details of the coding are described elsewhere (McCarthy, 2007), as are details of the cluster analysis output for identifying the number of dietary clusters in the data at different points in time (Bauserman, Albertson, Holschuh, and McCarthy, 2007). Briefly, data from each year is run through the macro separately. ACECLUS generates orthogonal, canonical variables which are then entered into CLUSTER to generate the actual clusters. Diagnostic measures (such as the cubic clustering criterion, or CCC) are then graphed with GPLOT so that the best number of clusters can be identified.

Finally, the TREE procedure can be set to output cluster assignments for each participant (observation) as a separate variable, which can later be used to compare clusters on characteristics of interest. When the optimal number of clusters for a specific time point have been identified from the cluster diagnostics, then

- (1) rerun the macro with the TREE procedure set to output cluster assignments for that specific number;
- (2) label the output variables to identify the visit that generated that set of cluster assignments;
- (3) merge them back into your data sets for each visits; and
- (4) use any appropriate statistical techniques to identify and describe differences between clusters at that time point.

The techniques used to summarize or compare cluster characteristics depend on the requirements of the particular analysis you need. In this case, the presentation of the cluster structure may require nothing more than a description of the dietary characteristics of each cluster. Here's how the different baseline clusters compare with each other:

Mean servings from each food group at baseline, by cluster

Food Group	Cluster 1 (n=289) Moderate	Cluster 2 (n=29) Pizza	Cluster 3 (n=30) Bad veggies	Cluster 4 (n=70) Meat and bread
Bread/grain Go	3.50	3.37	3.51	3.75
Bread/grain Whoa	0.63 ^b	0.62 ^b	0.79 ^b	1.78 ^a
Dairy Go	1.46 ^{ab}	1.18 ^b	1.39 ^{ab}	1.74 ^a
Dairy Whoa	0.63	0.70	0.60	0.55
Fats/oils Go	0.66 ^b	0.60 ^b	0.97 ^b	1.73 ^a
Fats/oils Whoa	0.30 ^b	0.22 ^b	0.31 ^{ab}	0.60 ^a
Meat/fish/poultry Go	1.45 ^{bc}	1.24 ^c	1.93 ^b	2.66 ^a
Meat/fish/poultry Whoa	1.93	2.28	1.57	2.07
Snacks Go	2.39	2.84	2.42	2.53
Snacks Whoa	1.53	1.45	1.41	1.82
Vegetables Go	0.69 ^{ab}	0.51 ^b	0.86 ^a	0.97 ^a
Vegetables Whoa	0.22 ^c	0.50 ^b	1.33 ^a	0.42 ^b
Dessert, All	0.97	1.30	0.92	1.05
Fruit, All	1.35	1.11	1.35	1.23
Pizza, All	0.35 ^b	2.13 ^a	0.27 ^b	0.15 ^b

Means with different superscripts are significantly different.

As it turns out, visit 3 also results in a 4-cluster solution. But now the dietary patterns of the four clusters are not the same as those found at baseline – which is good, because hopefully a multi-year dietary intervention had some effect. And none of the clusters at visit 3 are identical to those from baseline in terms of the participants who make

up those clusters. It is difficult to imagine a single table that could somehow summarize the cluster structure at each visit, and also show how participants shifted over time. Here's where the capability of the ODS system to output into different file formats, and the capability of Word to create graphics and embed hyperlinks, allows us to present cluster structure and cluster characteristics, at different time points, in a single image.

First, the ODS system allows us to change our output from the LISTING format to other options such as HTML or even PDF. So all we need to do is ensure that our output is sent to a properly formatted external file that can then be linked to Word. The bare minimum of coding to do this requires turning off the LISTING output destination and substituting the format you want.

```
ODS LISTING CLOSE;  
ODS PDF;
```

To send the output to a specific location outside your SAS directories for later retrieval, you can specify the FILE="path\filename" option.

```
ODS PDF FILE='my-directory\sub-directory\filename';
```

Then, just insert the code for your analysis (REPORT, TABULATE, MEANS, whatever).

If you used the FILE option in the ODS PDF statement, then every additional procedure you execute will try to output to that file, so it's a good idea to turn it off.

```
ODS PDF CLOSE;  
ODS LISTING;
```

In this example, we really need only the mean servings from each food group to characterize the dietary clusters.

```
ods pdf file="C:\Documents and Settings\rbauserman\My  
Documents\diets\dietclustersbase1.pdf";  
  
proc means data=disc_cluster_base1;  
  
  var bgg_w0 bgw_w0 desg_w0 desw_w0 dyg_w0 dyw_w0 fog_w0 fow_w0 frg_w0  
mfp_g_w0 mfpw_w0 pizzag_w0 pizzaw_w0 sng_w0 snw_w0 vg_w0 vw_w0 desgw_w0  
fogw_w0 frgw_w0 pizzagw_w0;  
  
  title "Food group servings, Baseline visit";  
  
  title2 "Moderate cluster";  
  
run;
```

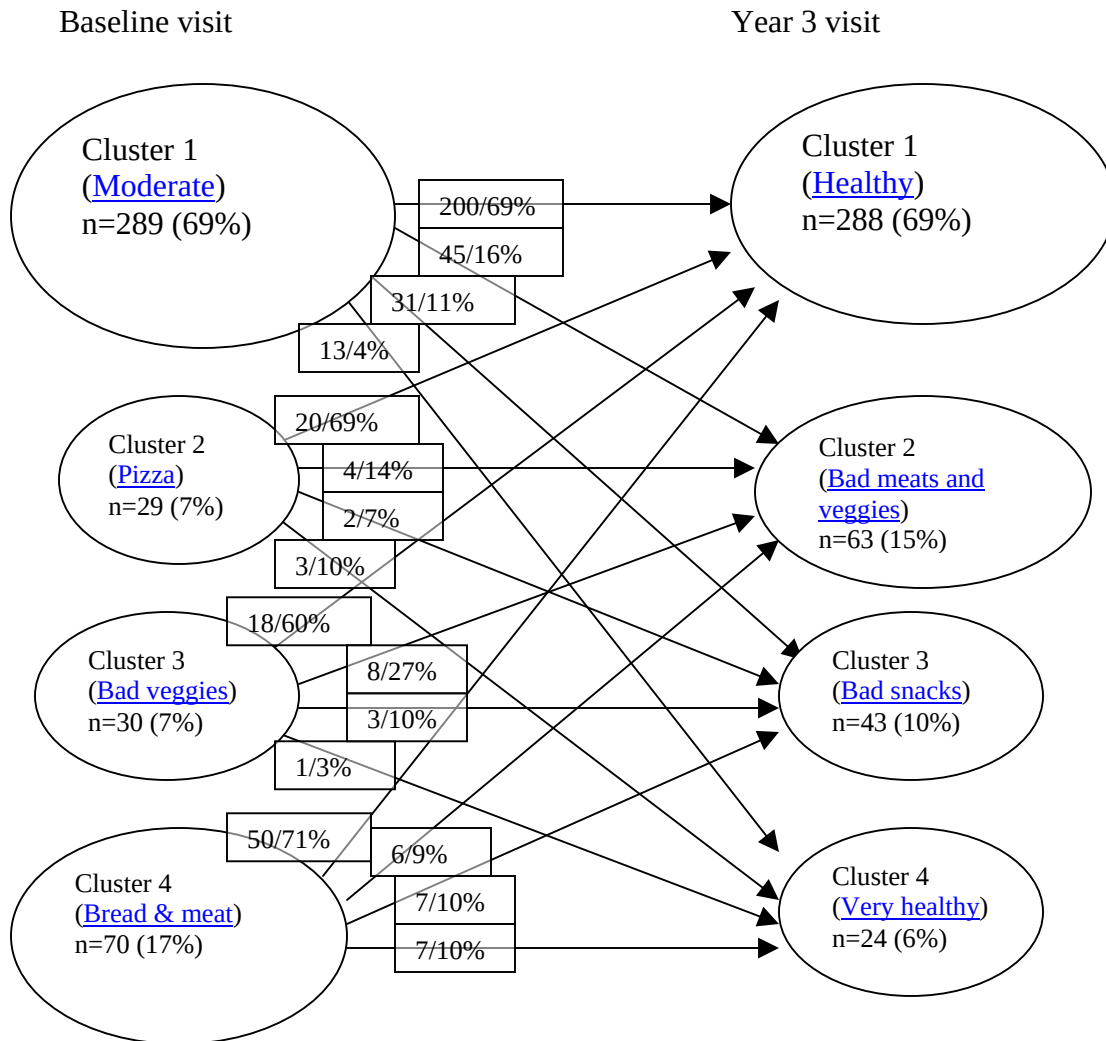
Now, it's time to create your graph. The Word program in the Microsoft office suite has a drawing toolbar with a wide variety of options. These include shapes, arrows, and so on. Shapes can be resized and moved even after you have placed them on the page, so you can adjust your figure as much as necessary as you go along. Here's an example. Note some of the features:

- (1) Figures are sized differently (although not to scale) to indicate which clusters are larger and smaller.
- (2) The figure uses ellipses, but circles, squares, or rectangles can be created just as easily. Both the "Basic shapes" and "Flowchart" options allow selection of an even wider variety of flowchart shapes.
- (3) Arrow styles can be adjusted in a variety of ways (thickness, presence of arrowheads, solid vs. "block" arrows, etc.).
- (4) Text inside text boxes can be formatted (font, style) just as ordinary text can.

The data user can see, in a single figure, (a) how many clusters there are at each time point; (b) the number of study participants in each cluster (and a visual cue of their relative size); and (c) how many participants, or what percentage, or both, move into any given cluster at the next point in time.

Figure 1.

Shifts in dietary cluster membership, Baseline to Year 3 (all subjects)



You create hyperlinks to external files from within the figure from the “Insert” menu on the Word toolbar. The steps are:

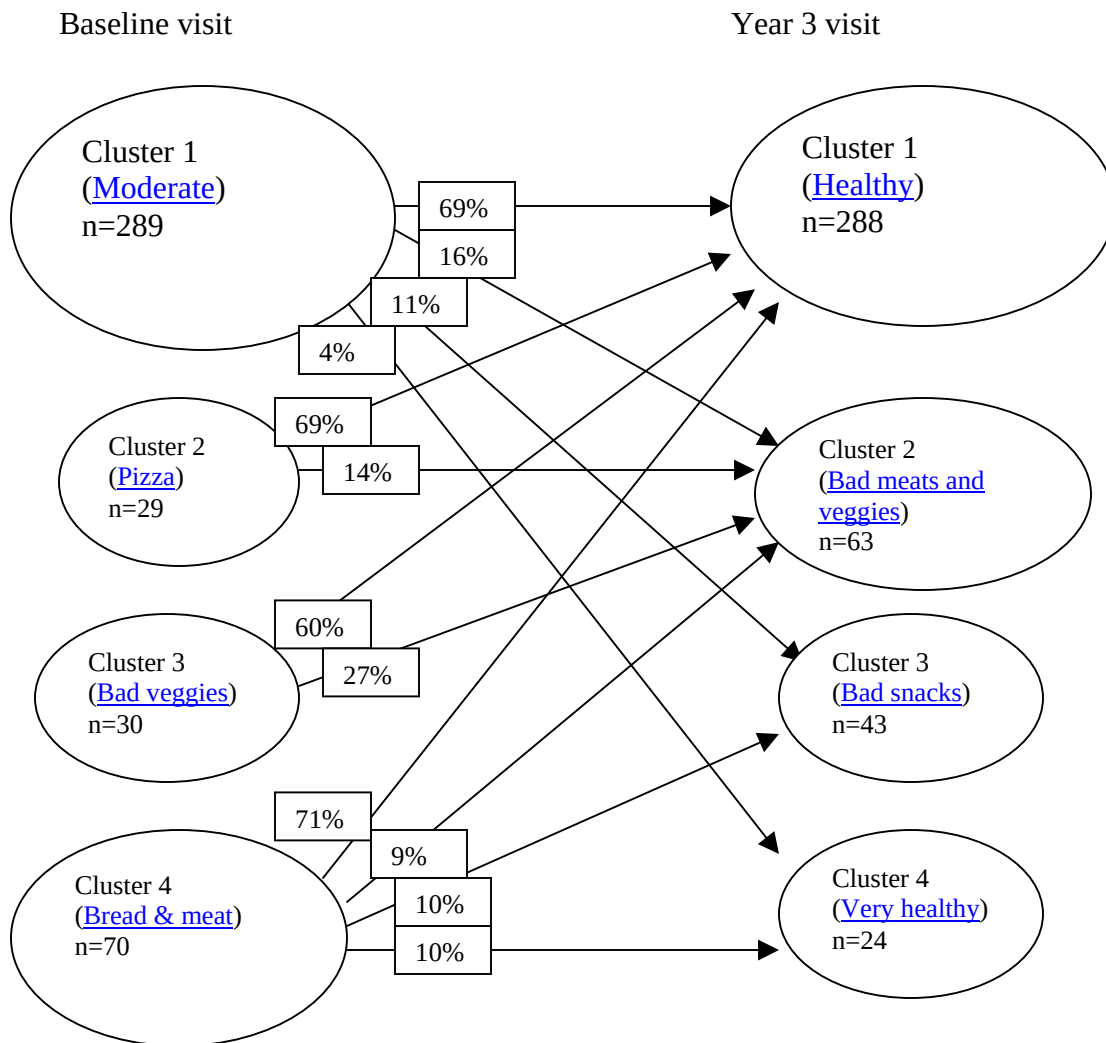
- 1) Highlight the text that you want to turn into a hyperlink;
- 2) Click on “Insert”;
- 3) Click on the “Hyperlink” button;

If we click on the “Moderate” hyperlink for the first cluster under “Baseline” visit, we see the output reproduced on the final page as Appendix A.

But this figure might still be a little cluttered. So, we can make some adjustments to remove some of the clutter that might be less informative.

Figure 2.

Shifts in dietary cluster membership, Baseline to Year 3 (all subjects) - revised



This second graph simply drops every connection and text box from the first that represents only 1, 2, or 3 participants. Dropping connections with very small n's, or that form very small percentages of a cluster, sacrifices little information from the graph but can substantially increase its clarity. This version also simplifies the text boxes by displaying only the percentage, rather than the raw n, that moves from each cluster into different clusters at the following visit – which may be the most meaningful information for data consumers anyway - and makes sure the text boxes are “filled” to keep the connector arrows from showing through. These are only some of the possible adjustments that could be made to the formatting to enhance clarity of the data presentation for end users of the data.

Note that you may still need tables to summarize cluster characteristics at specific time points – such as the table shown earlier that summarized the dietary patterns of each cluster at baseline. However, a graphic representation may be much more effective for conveying the “big picture” of how the number and size of clusters changes from visit to visit, and how participants shift between clusters over time.

Conclusion

Cluster analysis is a useful tool for identifying data patterns that may not be apparent from univariate or bivariate analyses, and SAS provides multiple cluster analysis tools. When data on subjects includes measurements at multiple points in time, cluster analysis presents challenges. Separate cluster analyses of the measurements from

each point in time avoid such problems as correlated observations over time, but confront users with potentially complex results.

Tabular data can coherently summarize cluster analysis from a single point in time, but graphical representation may be easier to understand when the goal is to show how cluster solutions shift over time and how subjects migrate between clusters. Combining the Output Delivery System (ODS) from SAS with Microsoft Word graphic tools (from the Drawing toolbar) and the Insert-Hyperlink feature, you can construct a single graph that can show at a glance

- (1) the number and relative size of clusters at each point in time; and
- (2) the flow of participants, in absolute numbers or percentages or both, across clusters over time; and
- (3) embedded hyperlinks that directly access formatted SAS output that describes cluster characteristics.

Such graphs can themselves become awkward and cluttered, but several rules of thumb may help simplify the presentation while preserving the bulk of the data.

- (1) Small clusters (small n's or small percentages of the total sample) may be dropped from the graph while still allowing the large majority of subjects to be included.
- (2) Arrows and text boxes used to show subject flow between clusters may be dropped if they represent a very small subset of the participants from a particular cluster, reducing clutter and making the graph simpler to read.
- (3) Naming clusters on the basis of their predominant characteristics may help ensure that data consumers do not assume that strong links between clusters from different time points imply identical cluster characteristics.

REFERENCES

Bauserman, Robert, Albertson, Ann, Holschuh, Nort, & McCarthy, William (2007). Longitudinal cluster analysis with dietary data over time. SESUG 2007 Conference, Hilton Head, South Carolina.

McCarthy, William (2007). A Methodological approach to performing cluster analysis with SAS. SESUG 2007 Conference, Hilton Head, South Carolina.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the first author at:

Robert Bauserman
Maryland Medical Research Institute
600 Wyndhurst Ave.
Baltimore, MD 21210
Work Phone: 410-532-0249
Fax: 410-433-1117
E-mail: rbauserman@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.

Appendix A
PDF output for Food group servings, Baseline visit
Moderate cluster

15:00 Monday, July 16, 2007 1

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
bgg_w0	Servings of go bread and grains per day	289	3.5046713	1.3483789	0.3100000	7.1100000
bgw_w0	Servings of whoa bread and grains per day	289	0.6325606	0.6789067	0	3.4400000
desg_w0	Servings of go desserts per day	289	0.4061938	0.5408366	0	2.6700000
desw_w0	Servings of whoa desserts per day	289	0.5638408	0.7604984	0	4.3200000
dyg_w0	Servings of go dairy per day	289	1.4559516	1.0445121	0	4.3200000
dyw_w0	Servings of whoa dairy per day	289	0.6322837	0.6567298	0	2.9100000
fog_w0	Servings of go fats and oils per day	289	0.6604844	0.6617305	0	3.2800000
fow_w0	Servings of whoa fats and oils per day	289	0.2991003	0.4485260	0	2.8300000
frg_w0	Servings of go fruit per day	289	1.3502768	1.2015698	0	5.1700000
mfp_g_w0	Servings of go meat, fish, poultry per day	289	1.4447059	0.9739032	0	5.8000000
mfpw_w0	Servings of whoa meat, fish, poultry per day	289	1.9331142	1.3960022	0	6.7100000
pizzag_w0	Servings of go bread and grains per day	289	0.1193080	0.2994478	0	1.9200000
pizzaw_w0	Servings of go bread and grains per day	289	0.2286851	0.3968421	0	2.0200000
sng_w0	Servings of go snacks per day	289	2.3866090	1.6220986	0	7.9800000
snw_w0	Servings of whoa snacks per day	289	1.5347059	1.7218716	0	8.1600000
vg_w0	Servings of go vegetables per day	289	0.6857439	0.6131031	0	2.6500000
vw_w0	Servings of whoa vegetables per day	289	0.2195156	0.2648934	0	1.6700000
desgw_w0	All desserts	289	0.9668166	0.9288607	0	4.5600000
fogw_w0	All fats/oils	289	0.9619031	0.7605030	0	3.5000000
frgw_w0	All fruits	289	1.3506920	1.2015824	0	5.1700000
pizzagw_w0	All pizza	289	0.3479931	0.4654659	0	2.0200000