A robust multivariate analysis to identify dietary patterns

F. HABI-RACHEDI[1], P. RONDEAU[1], S. MARQUE [1], B. A. HOLMES [2]

[1] *Danone Research, Biometrics Department - Clinical Studies Platform, Palaiseau, France*

[2] *Danone Research, Global Nutrition Department, Palaiseau, France*

Fatiha.HABI-RACHEDI @danone.com , Pascale.RONDEAU@danone.com

CONTEXT & OBJECTIVES

The objective of this analysis was to identify dietary patterns in a sample of 308 women aged between 18 and 60 years old recruited from one area in the North of France.

The purpose of most pattern detection methods is to represent the variation in a data set in a manageable form by recognising classes or groups. There are basically two approaches that have been carried out in many types of studies with large data sets; Principal Component Analysis (PCA) and Clustering Analysis (CA).

MATERIAL & METHODS

Dietary data were collected using three non-consecutive multiple pass 24-hour dietary recalls carried out over the telephone by trained dieticians. The data was entered directly by the dietician into a web-based tool developed by Medical Expert Systems (MXS, Paris, France). The 24-hour recall method involves the subject recalling all foods and drinks consumed the preceding day using a structured interview with specific neutral probes (Thompson and Byers, 1994).

Data were collected on two weekdays and one weekend day (Sunday). Each item of food and drink consumed was linked to a food composition database and grouped into one of 27 pre-defined categories.

The results presented here are based on mean intakes over the three days.

The results of the PCA on the dietary data were not easy to interpret and more than three components were needed to have a good representation of the data.

Two types of CA were selected, the k-means (non hierarchical method) and Ward's Agglomerative hierarchical clustering (AHC based on Ward's Minimum-Variance). The grouping of subjects is done on the basis of similarities (dissimilarities) in eating behaviours, measured by distances (Euclidian) between the subject intakes.

All food categories were standardized to a mean of 0 and a standard deviation of 1.

Data were analysed using SAS® 9.2 and SAS® Enterprise Guide® 4.2 (SAS Institute Inc., Cary, NC, USA).

In the initial k-means algorithm, the k cluster centers are generated randomly, making it sensitive to starting conditions. An alternative way which greatly enhanced robust cluster recovery was developed by Milligan (Milligan, 1980). The FASTCLUS procedure in SAS is based on this modified algorithm.

DIETARY PATTERN ANALYSIS

There are three statistical parameters that indicate the measure of fit of the k-means or the Ward's Minimum-Variance methods: Pseudo-F statistic (PFS), Cubic Clustering Criterion (CCC) and all approximate expected R-squared (R2). In general, the goal is to maximize each parameter.

The values of CCC for different numbers of clusters for Ward's AHC were all negative, indicating that the data distribution by clusters were close to uniform (no clusters) distribution.

In the k-means method, the number of clusters must be established a priori and therefore several solutions were compared with a varying number (N) of clusters (N from two to seven). The number of clusters was chosen based on the three statistical parameters described above considering also a good balance of subjects in each cluster.

Table 1 : Statistical parameters of goodness of fit for k-means

N	PFS	R2	CCC	% min	% max
2	16,88	0,05	-4,11	11%	89%
3	18,93	0,09	1,35	4%	60%
4	18,7	0,12	4,62	7%	44%
5	14,93	0,14	-0,9	3%	72%
6	15	0,16	2,8	1%	40%
7	15,59	0,18	8	1%	45%

The small clusters were characterized by the extreme values of the food categories with a low percentage of consumers such as nuts and appetizers and shellfish.

❑ ***Winsorized k-means***

In order to avoid the effect of extreme values on the k-means clustering, the largest values for each food category were capped at a given value using the winsorized approach which avoids the need to delete observations from the analysis (Tukey, 1962, Mingxin, 2006). There are two techniques to determine the cap values which we compared: percentiles and box plot with fences. The box plot with fences approach identifies extreme values in the tails of the distribution using quantities based on the inter quartile range. Values were capped at the upper inner fence: Q3 + 1.5*IQR where IQR= inter quartile range and Q3 the third quartile.

The statistical parameters were better than those for k-means without correction and the clusters were more evenly sized according to the two techniques percentiles and box plot. Between the two correction techniques the Pseudo-F and the R2 didn't differ greatly, however the CCC parameter is improved using the box plot technique. In addition, with the box plot technique the correction is standardised across the food groups while the correction using percentiles can be done by the 90, 95 or the 99 percentiles according to the food category.

Table 2 : Statistical parameters of goodness of fit for winsorized k-means

	PFS		R^2		CCC		min%		max%	
N/Method	WPerct	WBP	WPerct	WBP	WPerct	WBP	WPerct	WBP	WPerct	WBP
2	18,18	15,60	0,04	0,04	7,38	9,22	47	37	53	63
3	16,30	14,82	0,07	0,06	9,74	13,25	29	28	42	39
4	13,27	13,65	0,10	0,09	5,95	14,32	16	18	31	32
5	11,96	11,90	0,12	0,11	4,82	11,58	13	13	27	32
6	11,24	11,08	0,14	0,13	4,78	11,17	11	9	23	27
7	0,90	10,35	0,16	0,14	5,97	10,36	7	7	22	24

Wperct : correction using the 90, 95 and 99 percentiles
WBP: correction using the box plot

The optimum number of clusters (k) was selected as four (k=4) with a reasonable balance of subjects per cluster: Cluster 1, 58 subjects, Cluster 2, 94 subjects, Cluster 3, 100 subjects and Cluster 4, 56 subjects.

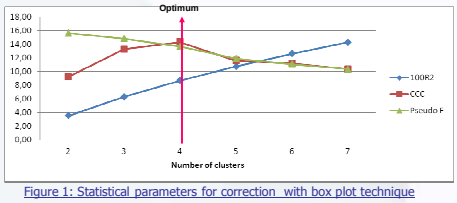

Figure 1: Statistical parameters for correction with box plot technique

CHARACTERISATION OF THE CLUSTERS

❑ **Canonical Discriminate Analysis** Using canonical discriminate analysis (CDA), the food categories were transformed into three canonical variables which enables the visualization of the clusters and the associated food categories.

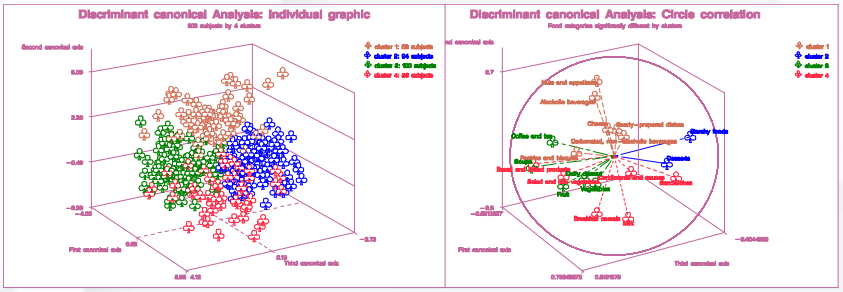

Figure 2 : CDA displaying the division of the subjects (n=308) by the four clusters and the associate circle correlation

❑ **Tests:** The dietary data were not normally distributed and therefore to test for differences in level of intake between the clusters the Kruskal-Wallis (KW) test was used. The Mann-Whitney test was used to test for differences between each pair of clusters. The Bonferroni correction was applied for the multiple comparisons. A significant difference in mean intakes across the four clusters was observed for 19 of the 27 food categories which appear on the circle correlation of CDA.

A significant difference was observed in the mean age of subjects (KW-test) across the four clusters. Despite this difference, adjusting for age with the Van-Elteren test (stratified KW-test) revealed only a few small differences in the food categories that characterized the clusters.

DISCUSSION AND CONCLUSION

❖ We selected the k-means clustering for this sample recognizing a limitation of this method which is sensitivity to extreme values. The winsorized k-means is a solution to overcome this issue.

❖ For each cluster there was a negligible difference between the winsorized and the non-winsorized means for all food categories.

❖ The winsorized k-means is a robust method which is able to identify dietary patterns even in small sample sizes and without effect of age on almost all the clusters.

❖ k-means appears to be a relevant method for identifying dietary patterns in this study population. Given that this separation is linear, future testing using kernel k-means for nonlinear separation will be undertaken.

REFERENCES

1. Hu F (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol, 13, 3-9.

2. Kim S (1992) The metrically trimmed mean as a robust estimator of location. Ann. Statist. 20.

3. Milligan GW & Cooper MC (1985) An Examination of Procedures for Determining the Number of Clusters in a Data Set. Psychometrika, 50, 159-179.

4. Mingxin Wu (2006) Trimmed and winsorized estimators, A dissertation for the degree of Doctor of Philosophy, Michigan State University Probability and Statistics Department.

5. Sarle WS (1983) Cubic Clustering Criterion. SAS Technical Report A-108, Cary, NC. SAS Institute Inc.

6. Thompson F and Byers T (1994) Dietary assessment resource manual. J. Nutr. 124: 2245S-2317S.

7. Tukey J. W. (1962) The Future of Data Analysis, The Annals of Mathematical Statistics, 33, p. 18.

ACKNOWLEDGMENTS: The Danone Research Clinical Studies Platform and all the people involved in the study, Damien PAINEAU (Danone Research, Global Nutrition Department) for his useful comments on this work.

BIOMETR χ
Gives meaning to your data