A European perception of food using two methods of correspondence analysis

Eric J. Beh[a1], Rosaria Lombardo[b] & Biagio Simonetti[c]

[a]*School of Mathematical & Physical Sciences, University of Newcastle, Callaghan, NSW, 2308, Australia.*
[b]*Department of Strategy and Quantitative Methods, Second University of Naples, Gran Priorato di Malta, 81043 Capua (CE), Italy*
[c]*Department of Analysis of the Economic System and Society, University of Sannio, 82100, Benevento, Italy.*

Abstract
In a recent issue of this journal, Guerrero *et al.* (2010) studied an interesting data set involving the analysis of consumer-driven associations to the word ‘‘Traditional”, from a food perspective, in six European countries. As part of their analysis, they demonstrated the sources of association between the words studied and the country of origin of those interviewed using correspondence analysis. In this paper we focus on this association by assuming that the country of origin is a predictor of the words associated with “Traditional”. This analysis is performed using another member of the correspondence analysis family - non-symmetric correspondence analysis. This paper will also explore the use of both these correspondence analysis techniques on their data and consider the dendrogram and the semantic differential plot as alternative approaches to visually summarising the association.

Keywords: Correspondence analysis, Non-symmetrical correspondence analysis, association, cultural perception of food.

1. Introduction

Correspondence analysis (CA) is a popular graphical tool that is used to explore the symmetric association structure between categorical variables. It has gained a high level of exposure and use in a diversity of areas including those involved in the many aspects concerned with the study of food. In particular, in the third volume of this journal McEwan & Schlich (1991/92) described in some detail the mechanics of CA while investigating the association between sensory assessors and attributes of jam samples. Since then, there have been a large number of articles published in *Food Quality & Perception* that consider various aspects of CA – for example, see Torres & van de Velden (2007), van Herk, & van de Velden (2007), Bécue-Bertaut, Álvarez-Esteban & Pagès (2008), Marshell & Bell (2003) and Guerreo *et al.* (2010). These articles have all helped with the growing success of the development and application of CA. However, one must keep in mind that by the very nature of the mathematical properties of CA, the categorical variables considered in the study are assumed to be “globally” associated, or symmetrically related. That is, they are both considered to be such that neither of them is regarded a response variable of another variable. The reader is directed to, for example, Greenacre (1984) and Beh, (2004) for a mathematical description of CA, and Clausen (1988) and Greenacre & Blasius (1994) for a more applied focus. Certainly, the relative simplicity of performing CA when analyzing multiple categorical variables by treating them in a symmetrical manner makes it a highly informative and flexible graphical statistical technique. Using the traditional approach to CA, it is also very difficult to ensure that the dependent/independent variable structure (which is referred to here as “predictive” association) between multiple categorical variables is preserved. For example, there are many studies where one categorical variable is dependent on a second, or multiple, categorical variables. One may consider the recent study undertaken by Guerrero *et al.* (2010). Their data considered how the word “Traditional” (from a food perspective) was perceived across six regions in six European countries; Flanders in Belgium, Burgundy (Dijon) in France, Lazio in Italy, Akershus and Østfold

[1] Corresponding author. Tel + 61 2 4921 5113; fax. + 61 2 4921 6898.
Email address: eric.beh@newcastle.edu.au

in Norway, Mazovia (Warsaw) in Poland and Catalonia in Spain. The two variables of interest in their analysis were the country of origin from where a participant of the study came and 28 different words that they associated with "Traditional" food; these variables have been labeled here as *Country* and *Word Association*, respectively, for convenience. The data appear here in Table 1 and is based on the six histograms in their Fig. 1. Refer to Guerrero *et al.* (2010) for an excellent and comprehensive description of the methods used to collect the data and the CA that was performed using this histogram data. It must be noted that the authors were careful to ensure that their conclusions reflected a "global association" of the variables studied rather than a "predictive" association which is considered here.

Since their analysis involves only two categorical variables (*Country* and *Word Association*), we shall refer to the CA approach conducted by Guerrero *et al.* (2010) as *simple correspondence analysis* (or just simple CA). This helps to distinguish it from *multiple correspondence analysis* which involves the analysis of three or more categorical variables – this paper will not consider such a situation, but the interested reader may refer to Greenacre and Blasius (2006) for more on this topic.

- - - - - - - - - - - - - - - - - - -
Table 1 about here
- - - - - - - - - - - - - - - - - - -

By performing a simple CA on Table 1, a researcher can graphically identify the association structure between the *Country* and *Word Association* variables. For the analysis of Table 1, this is certainly an appropriate analysis to perform and provides a means of identifying how a country is associated with a word, or multiple words – thus, it reflects the "global" association structure between *Country* and *Word Association*. This is one of the primary aims of the study undertaken by Guerrero *et al.* (2010). However, if the researcher wishes to understand the nature of the association by taking into account that a person's country of origin may impact upon how they perceive "Traditional" food, then an alternative approach to simple CA may be considered to graphically depict this "predictive" association structure. This adaptation is referred to as non-symmetrical correspondence analysis (NSCA) and treats one categorical variable as a predictor, or independent, variable and treats the second variable as a response variable. Hence NSCA can be used to graphically reflect the "predictive" association structure between *Country* and *Word Association*. For Table 1, we can treat *Country* as the predictor variable, and determine how it impacts upon the outcome of *Word Association*. An account of the mathematical and practical issues of this cousin of simple CA was proposed by D'Ambra & Lauro (1989) and discussed further by Kroonenberg & Lombardo (1999), Lombardo, Kroonenberg & D'Ambra (2000) and Lombardo, Beh & D'Ambra (2007). Therefore the reader is invited to consider any of these for more detail on NSCA.

Therefore the aim of the remainder of this paper is to explore the practical distinctions between CA and NSCA by analysing Table 1. Section 2 provides a description of the application of both techniques, while the *Discussion* (Section 3) provides some final remarks when interpreting the results obtained from performing NSCA.

2. Perception of "Traditional" in terms of food in Europe

Consider the data described and analysed by Guerrero *et al.* (2010) which is summarised by Table 1. To determine the extent to which *Country* and *Word Association* are statistically and significantly associated, the Pearson chi-squared statistic can be calculated and compared with the critical value obtained from the chi-squared distribution with $(28 - 1)(6 - 1) = 135$ degrees of freedom. For Table 1, the Pearson chi-squared statistic is 663.8 and, with a p-value < 0.0000, there is ample evidence to indicate that there is a significant association between *Country* and *Word Association*. To graphically depict this "global" association, Guerrero *et al.* (2010) perform a simple CA of the data where the two-dimensional correspondence plot is given here by Fig. 1a). Note that the configuration of points of this plot is identical to Fig. 2. of Guerrero *et al.* (2010). These authors provide some insightful remarks concerning the "global" association structure between the 28 words studied and the six countries of Table 1.

Suppose we know from which of the six countries a participant of the study hails. Given this information, we can determine the nature of the "predictive" association between *Country* and *Word Association* by performing NSCA where *Country* is treated as a predictor variable and *Word Association* is the response variable. Rather than measuring the extent to which the variables are asymmetrically related using the Pearson chi-squared statistic, a more legitimate measure is to consider the C-statistic which, for Table 1, is also chi-squared distributed with 135 degrees of freedom. This C-statistic (Light & Margolin, 1971) is based on the Goodman-Kruskal tau index (Goodman & Kruskal, 1954). The C-statistic of Table 1 is 698.3, and with p-value < 0.0000 there is ample evidence to suggest that the country of origin of a person (from those six countries sampled) is a factor in determining how they perceive "Traditional" food. Making such a conclusion concerning the "predictive" nature of the association between the two variables using the chi-squared statistic (as we did using CA) is not possible since the statistic focuses purely on the "global" association between categorical variables. To graphically determine more precisely how *Country* helps to "predict" *Word Association*, Fig. 1.b) is the two-dimensional correspondence plot obtained from performing NSCA of Table 1. By observing the proximity of the points in this plot one can see the link between the 28 different words and determine which countries are more likely to lead to such associations. For example, suppose that we know an individual is from Norway. We can infer from Fig 1.b) that they will generally perceive "Traditional" food with the words *Christmas*, *Country* and *Rural*. Similarly, knowing that an individual is from Spain, we can infer that they identify "Traditional" food with *Ancient*, *Grandmother* and *Old-fashioned*, while for Italians the word is associated with *Home-made*, *Holidays* and *Tasty*.

There are some key discrepancies in the interpretation of the configuration of points for the two plots of Fig. 1. This is because of the different ways in which the association between *Country* and *Word Association* is considered. Guerrero *et al.* (2010), and Fig. 1a) indicate that the perception of "Traditional" in a food context in France is closely associated with *Cooking* and *Holiday* as well as *Restaurant*, *Meal*, *Recipe*, and *Dish*. This does not imply that knowing that a randomly selected person is from France will mean that they generally associate "Traditional" food with these words. In contrast, it is valid to make the statement that, from Fig. 1b), knowing that an individual is from France they will generally associate "Traditional" food with *Regional*, *Healthy*, *Simple*, *Natural*, *Family* and *Cooking*. While the histograms of Fig. 1 in Guerrero *et al.* (2010) indicate that *Healthy*, *Simple* and *Natural* have relatively low frequencies for those surveyed from France, they also have a relatively very low frequency across the remaining five countries included in the study.

- - - - - - - - - - - - - - - - - - -

Fig. 1 about here

- - - - - - - - - - - - - - - - - - -

Fig. 1a) suggests that there is an association between Spain and the words *Good*, *Home* and *Ancient*. This is not surprising since the cell frequency of these three words is relatively high for Spain when compared with their frequency of other countries. Note that, Fig. 1a) does not imply that knowing that an individual is from Spain will lead them to generally associate "Traditional" food with *Good*, *Home* and *Ancient*. Such a one-way association can only be made using NSCA. Doing so, Fig. 1b) suggests that knowing that a randomly chosen individual is from Spain, they generally perceive "Traditional" food with *Ancient* and *Grandmother*. However, the two-dimensional correspondence plot of Fig. 1b) also highlights a major problem often overlooked when interpreting the association between categorical variables. Given that we are considering an individual who is from Spain, Fig. 1b) indicates that they are likely to associate "Traditional" food with *Old-Fashioned*, something which is clearly not true since no Spaniard made such a word association; note the zero frequency for this cell in Table 1. This discrepancy is because Fig. 1b) depicts just over 56% of the asymmetric association between the *Country* and *Word Association* variables; this value can be seen from the labels of this plot. Therefore, when performing NSCA, about 44% of the information in Table 1 is being ignored by considering the graphical summary of Fig. 1b). Similarly, for simple CA, Fig. 1a) graphically summarises only 49.5% of the association, thereby ignoring about half of the association structure that exists in the table; while Guerrero *et al.* (2010) did not explicitly state this, it is apparent from the percentages on their

Fig. 2 that this is the case. The misinterpretations can be accounted for by not taking into consideration the information contained in the third and higher dimensions. To graphically summarise all of the association between *Country* and *Word Association* we can perform hierarchical clustering. Here, we consider the coordinates that would normally be plotted in the non-symmetric correspondence plot using the optimum number of dimensions needed to graphically represent 100% of the association. For Table 1, the optimum number of dimensions needed is min(28, 6) – 1 = 5. Due to the difficulty in viewing a five dimensional space, a dendrogram of the coordinates can be constructed, using the complete-linkage method of the Euclidean distances to identify the clusters. Fig. 2 is a dendrogram from performing such an analysis of Table 1. See, for example Hair, Babin, Anderson & Tatham (2006, Chapter 8), or Lattin, Carrol & Green (2003, Chapter 8) for the description and application of clustering methods. Other plotting procedures, including Andrew's plots (Andrews, 1972; Rovan, 1994), can also be used for viewing points in multi-dimensional space.

- - - - - - - - - - - - - - - - - - -
Fig. 2 about here
- - - - - - - - - - - - - - - - - - -

The dendrogram of Fig. 2 identifies five distinct clusters that help explain how *Country* impacts upon *Word Association*:

- Cluster 1: Suppose we consider a Belgian. Their perception of "Traditional" to describe food is best summarised by the words *Feast*, *Old-Fashioned* and *Restaurant*. Note that if we explore the association between the two variables by ignoring how *Country* can help predict *Word Association*, there is an association between Belgium, *Old-Fashioned* and *Restaurant*; see also Guerrero *et al.* (2010, pg 229) who reached such a conclusion. That is, there is not just an association between these categories, but NSCA reveals that being Belgium does generally lead to "Traditional" food being perceived by these words.
- Cluster 2: Given that someone is from Poland or France, their perception of "Traditional" food is best summarised by *Family*, *Meal*, *Dish*, *Recipe* and *Cooking*. Indeed, ignoring the predictability of *Word Association* given *Country*, Guerrero *et al.* (2010, pg 229), and Fig. 1a) also concludes that there is an association between Poland and *Family*, and also between France and *Cooking*, *Meal*, *Recipe* and *Dish*. NSCA reveals that being from France and Poland leads to these word associations, not just that an association exists.
- Cluster 3: Given that someone is from Norway, their perception of "Traditional" to describe food is best summarised by the words *Christmas*, *Country*, *Rural*. The CA performed by Guerrero *et al.* (2010) also identified that an association existed, but did not consider the impact of *Country* influencing on *Word Association*.
- Cluster 4: Given that someone is from Italy, their perception of "Traditional" to describe food is best summarised by the words *Regional*, *Good*, *Simple*, *Home-made* and *Natural*. Note that ignoring the "predictive" association structure of the two variables, Fig 1a) reveals that there is a relationship between Italy and these last two words. Fig 1a) also shows that Italy and *Old* are associated. However, considering the association between *Country* and *Word Association* using NSCA shows that the word *Old* is more closely linked with Spain than Italy; see Fig. 2.
- Cluster 5: Given a Spaniard, their perception of "Traditional" food is best summarised by the words *Ancient*, *Home*, and *Old*. To a lesser extent, a Spaniards perception can also be described by the words *Culture*, *Habit*, *Grandmother* and *Quality*.

There is also a sixth cluster where five words are not clearly predictable given an individual's country of origin. These are *Dinner*, *Holidays*, *Kitchen*, *Healthy* and *Tasty*. However, Fig. 2 shows that there is a fairly weak predictive link with those from Norway.

A limitation of the dendrogram is that, unlike the correspondence plot, it is not possible to identify those countries that dominate the "predictive" association of *Word Association* given *Country*. However an alternative way to visualize this aspect of the association is to consider the semantic differential plot (Snider & Osgood, 1969) of Fig. 3. This plot highlights, in a more descriptive manner than

NSCA assesses the association structure, those words that can be "predicted" as important word associations with "Traditional" food. It shows that the Norwegians who participated in the study associate "Traditional" food with *Christmas* (dark blue line), *Rural* (light blue line) and *Country* (dark red line); a finding that is consistent with the observing Fig. 2. Similarly Spaniards associate "Traditional" food with *Old* and *Home*, a conclusion also reached by forming NSCA. It is interesting to note that many of the words that were not associated with a country in the dendrogram (that lie in the sixth cluster described) are not deemed to be important words that people attribute to "Traditional" food. This is especially the case for *Dinner*, *Holidays* and *Kitchen*. If one were to view the "predictive" association in a five-dimensional plot, these words would lie close to the origin.

- - - - - - - - - - - - - - - - - - -
Fig. 3 about here
- - - - - - - - - - - - - - - - - - -

Fig. 3. Also highlights that Norway, Spain and France are all important countries that help define the "predictive" association between the two variables of Table 1, while Italy does not appear as dominant. In the five-dimensional plot obtained by performing NSCA, Italy would be situated closer to the origin than any of the other five countries; a finding that is not apparent when inspecting the distance of Italy from the origin in Fig 1b). In fact, NSCA reveals, by inspecting Table 2, that Norway contributes very heavily to the "predictive" association measured by the C-statistic, followed by Belgium and France. As Fig. 3 suggests, and Table 3 reveals, Italy is the least dominant contributor of the association of all six countries that participated in the study.

- - - - - - - - - - - - - - - - - - -
Table 2 about here
- - - - - - - - - - - - - - - - - - -

3. Discussion

The data analysis performed by Guerrero *et al.* (2010) provides a valuable insight into how specific words are associated with the word "Traditional", considered from a food perspective, and how the country of origin of a participant in the study is associated with these words. However, additional insight into the "predictive" structure of the variables can be uncovered by considering how the country of origin can be used as a predictor of the words listed in Table 1. NSCA allows for such a structure to be revealed. Certainly, this short paper has been concerned with understanding the association structure between *Country* and *Word Association* by taking into consideration that *Country* can be used as variable that elicits a response where the responses consist of the words included in the *Word Association* variable. Regardless of which CA approach is used, it is important to ensure that the interpretations of a two-dimensional display, such as those of Fig. 1, graphically reflects a large proportion of the association that exists between two variables. In situations where this does not occur (such as in Fig. 1), alternative graphical techniques, such as the dendrogram or the semantic differential plot, may be used by incorporating all of the information contained in the data, not just the information reflected in the first two dimensions. Other graphical summaries that are linked with CA, such as biplots, can also be considered. However, rather than elaborating further on this issue the reader is invited to consider Kroonenberg & Lombardo (1999) and Lombardo, Kroonenberg & D'Ambra (2000) for a description of using biplots with CA.

From an equally practical perspective, when performing CA, a researcher must be careful not confuse "global association" with "predictive association". When studying the many aspects concerned with food, interest on how variables are associated is important, but understanding how one categorical variable helps to predict the outcome of another (or multiple) categorical variable is different matter. By performing the traditional approach to CA (as is often done in many studies that do use CA as the method of analysis) the researcher is actually investigating the "global association" between variables, while NSCA allows for one to investigate the properties surrounding the "predictive association" of the variables. NSCA has been presented in the statistical literature as a variant of CA where the main

computational difference concerns the fact that the weighted metric (in our case, the marginal frequencies) is only used for the *Country* and not for the *Word Association* variable. One effect of this is that those words with small proportions (for example, *Simple*, *Quality*, *Recipe* and *Kitchen*) will have little influence on NSCA in contrast with CA. Such distinctions between these two CA techniques can help researchers shed more light on the behaviour of their data than they may otherwise be able to obtain.

Bibliography

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, 125 – 136.

Bécue-Bertaut, M., Álvarez-Esteban, R. & Pagès, J. (2008). Rating of products through scores and free-text assertions: Comparing and combining both. *Food Quality and Preference*, 19, 122 – 134.

Beh, E. J. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review*, 72, 257-284.

Clausen, S.-E. (1988). *Applied Correspondence Analysis: An Introduction*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-121. Thousand Oaks, CA: Sage.

D'Ambra, L. & Lauro, N. C. (1989). Non-symmetrical correspondence analysis for three-way contingency table. In *Multiway Data Analysis* (Coppi, R. and Bolasco, S. eds.) Elsevier: Amsterdam, pp. 301–315.

Goodman, L. A. & Kruskal,W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732–764.

Greenacre, M. (1984). *Theory and Application of Correspondence Analysis*. Academic Press: London.

Greenacre, M. & Blasius, J. (eds.) (1994). *Correspondence Analysis in the Social Sciences*. Academic Press: London.

Greenacre, M. & Blasius, J. (eds.) (2006). *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC: London.

Guerrero, L., Claret, A., Verbeke, W., Enderli, G., Zakowska-Biemans, S., Vanhonacker, F., Issanchou, S., Sajdakowska, M., Granli, B. S., Scalvedi, L., Contel, M. & Hersleth, M. (2010). Perception of traditional food products in six European regions using free word association. *Food Quality and Preference*, 21, 225-233.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. (2006). *Multivariate Data Analysis (6th Edition)*. Pearson.

Kroonenberg, P. & Lombardo, R. (1999). Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioural Research*, 34, 367–397.

Lattin, J., Carroll, J. D. & Green, P. E. (2003). *Analyzing Multivariate Data*. Brooks/Cole – Thomson Learning: Victoria (Australia).

Light, R.J. and Margolin, B.H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66, 534-544.

Lombardo R., Kroonenberg P., D'Ambra L. (2000). Nonsymmetric Correspondence Analysis: a simple tool in market share distribution. *Journal of the Italian Statistical Society*, 3, 107-126.

Lombardo, R., Beh, E. J. & D'Ambra, L. (2007). Non-symmetric correspondence analysis with ordinal variables. *Computational Statistics and Data Analysis*, 52, 566 – 577.

Marshall, D. & Bell, R. (2003). Meal construction: exploring the relationship between eating occasion and location. *Food Quality and Preference*, 14, 53 – 64.

McEwan, J. A. & Schlich, P. (1991/92). Correspondence analysis in sensory evaluation. *Food Quality & Preference*, 3, 23 – 36.

Rovan, J. (1994). Visualizing solutions in more than two dimensions. In *Correspondence Analysis in the Social Sciences* (eds Greenacre, M. & Blasius, J.), Academic Press: London, pg 210 – 229.

Snider J. G. & Osgood C.E. (1969). *Semantic Differential Technique: A Sourcebook.* Aldine Press: Chicago.

Torres, A. & van de Velden, M. (2007). Perceptual mapping of multiple variable batteries by plotting supplementary variables in correspondence analysis of rating data. *Food Quality & Preference*, 18, 121-129.

van Herk, H. & van de Velden (2007). Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. *Food Quality & Preference*, 18, 1096-1105.

Table 1
Cross-classification of individuals according to nationality and the 28 most common free-word associations of "Traditional" when describing food.

		Country					
Code	*Word Association*	*Belgium*	*France*	*Italy*	*Norway*	*Poland*	*Spain*
1	Ancient	0	0	2	0	0	15
2	Christmas	0	1	0	22	11	9
3	Cooking	2	17	0	0	5	9
4	Country	0	4	0	17	12	6
5	Culture	4	3	0	0	1	7
6	Dinner	0	0	0	3	14	0
7	Dish	0	10	0	0	11	0
8	Family	3	16	5	10	16	13
9	Feast	8	9	0	2	0	0
10	Good	1	2	6	5	2	12
11	Grandmother	10	4	1	5	2	9
12	Habit	14	3	0	3	0	25
13	Healthy	4	1	3	2	11	6
14	Holidays	8	0	0	0	12	1
15	Home	1	1	0	0	5	12
16	Home-made	0	0	8	2	12	9
17	Kitchen	4	0	2	1	4	2
18	Meal	1	6	1	0	11	2
19	Natural	2	2	5	2	5	5
20	Old	2	8	0	6	11	22
21	Old-fashioned	10	0	0	5	0	0
22	Quality	3	4	0	1	0	4
23	Recipe	0	7	0	0	6	0
24	Regional	1	7	4	2	1	1
25	Restaurant	5	5	1	2	1	1
26	Rural	0	12	1	19	5	11
27	Simple	2	1	3	1	3	1
28	Tasty	8	3	4	0	17	8

Table 2
Proportional contribution of each country to the C-statistic

Belgium	*France*	*Italy*	*Norway*	*Poland*	*Spain*
0.646	0.519	0.410	0.810	0.478	0.468

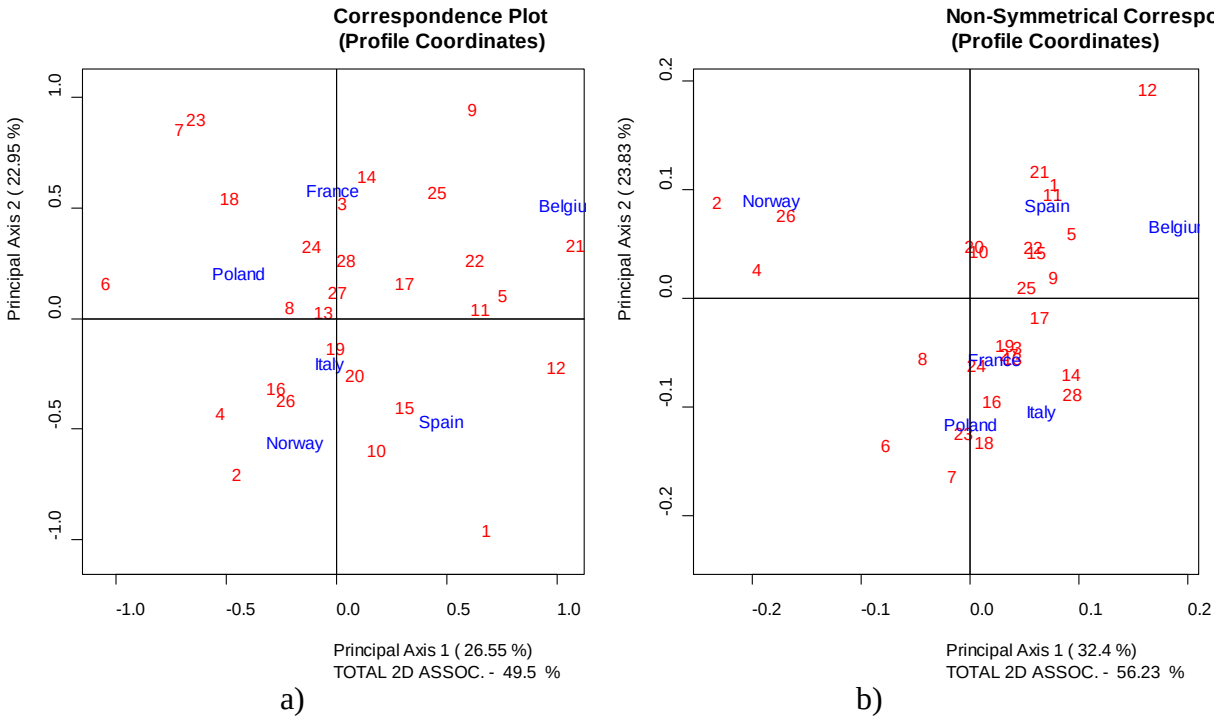

Fig. 1. a) Correspondence plot and b) non-symmetric correspondence plot of Table 1

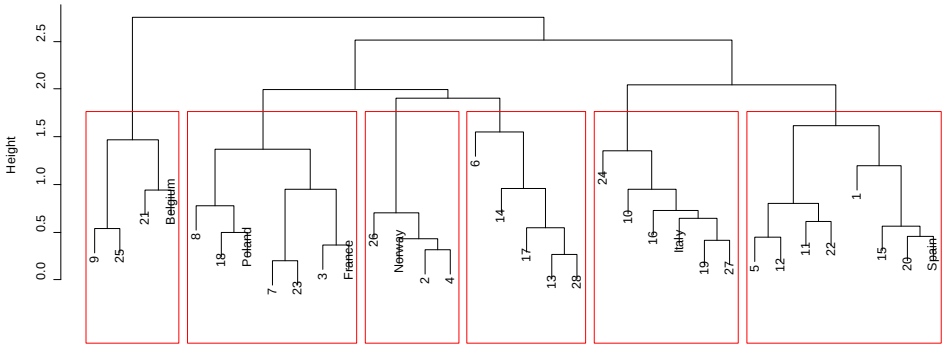

Fig. 2. Dendrogram from a hierarchical clustering of the NSCA coordinates in the five-dimensional correspondence plot.

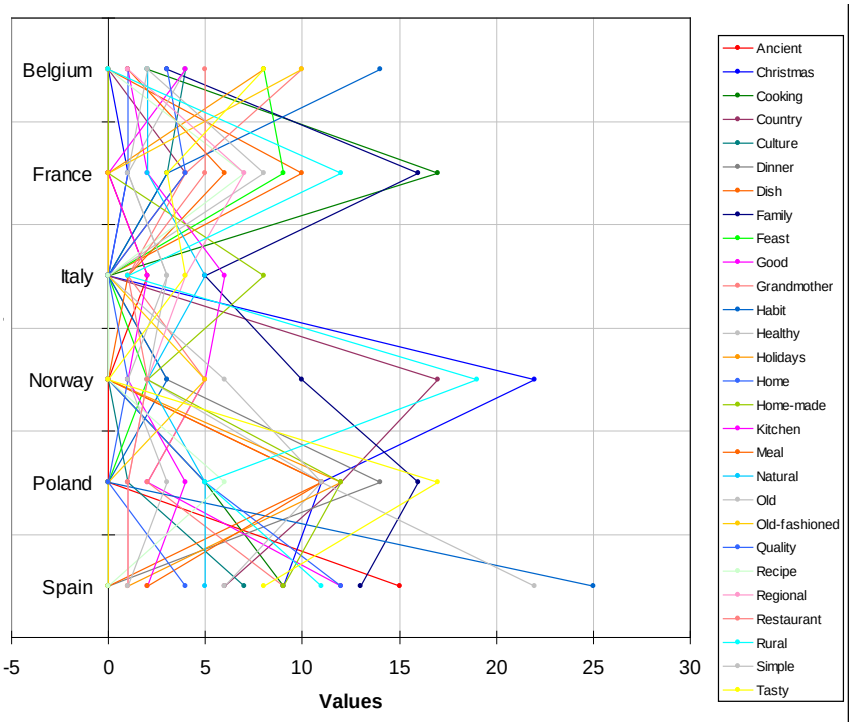

Fig. 3. Semantic differential plot of Table 1

Table 1
Cross-classification of individuals according to nationality and the 28 most common free-word associations of "Traditional" when describing food.

Table 2
Proportional contribution of each country to the C-statistic

Fig. 1. a) Correspondence plot and b) non-symmetric correspondence plot of Table 1

Fig. 2. Dendrogram from a hierarchical clustering of the NSCA coordinates in the five-dimensional correspondence plot.

Fig. 3. Semantic differential plot of Table 1