UNIVERSITA' DEGLI STUDI DI MILANO

Facoltà di Medicina e Chirurgia

Dipartimento di Scienze Cliniche e di Comunità

Scuola di Dottorato in **Scienze Biomediche cliniche e sperimentali**

Dottorato in **Statistica Biomedica – XXV ciclo**

DIETARY PATTERNS AND ESOPHAGEAL CANCER: A POSTERIORI DIETARY PATTERNS IDENTIFIED THROUGH FACTOR ANALYSIS AND CLUSTER ANALYSIS

Dott. **Francesca Bravi**

Matricola R08573

Tutor: Ch.mo Prof. **Adriano Decarli**

Coordinatore del Dottorato: Ch.mo Prof. **Adriano Decarli**

anno accademico 2011-2012

SUMMARY

ABSTRACT

Background: Because of the complexity of diet and the potential interactions between dietary components, the use of dietary patterns has been proposed, to describe variations in overall dietary intakes in a specific population and to analyze the relationship between diet and cancer risk. In the present work, factor analysis and cluster analysis were used in combination to identify groups of subjects with similar dietary patterns.

Patients and methods: We analyzed data from an Italian case–control study, including 304 cases with squamous cell carcinoma of the esophagus and 743 hospital controls. Dietary habits were evaluated using a food frequency questionnaire. A posteriori dietary patterns were identified through principal component factor analysis performed on 28 selected nutrients. A varimax rotation was applied to achieve a simpler loading structure. Nutrients with absolute rotated factor loading greater or equal to 0.63 on a given pattern were used to name the patterns. For each pattern, participants were grouped into categories according to quartile of factor scores among the control population, and the odds ratios (OR) and corresponding 95% confidence intervals (CI) were estimated using unconditional multiple logistic regression models accounting for potential confounding variables.

Then, cluster analysis was performed on factor scores obtained from factor analysis. The main analysis was carried out using the *k*-means method with Euclidean distance. The initial seeds were obtained performing preliminarily a hierarchical method (Ward's) and cutting the resulting dendrogram at the level corresponding to 6 clusters. Results from the main analysis were compared with those from other clustering solutions identified using the *k*-means method with Manhattan, Lagrange and Correlation coefficient similarity measure distances and the Partitioning around Medoids method, with both Euclidean and Manhattan distances.

The identified clusters were characterized by examining the distribution of several sociodemographic and lifestyle variables, and the average consumption of selected nutrients and food groups, within cluster. The ORs were estimated for each of the identified clusters, and corresponding 95% CIs were obtained referring to the floating absolute risks method.

Results: PCFA allowed to identify five major dietary patterns, which explained about 80% of the total variance in the original nutrients. The *Animal products and related components* pattern (with high factor loadings on calcium, phosphorus, riboflavin, animal protein, saturated fatty acids, cholesterol, and zinc) was positively related to esophageal cancer risk (OR=1.64, 95% CI: 1.06-2.55). The *Vitamins and fiber* (with high loadings on vitamin C, total fiber, beta-carotene equivalents, soluble carbohydrates, and total folate) and the *Other polyunsaturated fatty acids and vitamin D* (with high loadings on other polyunsaturated fatty acids, vitamin D, and niacin) were inversely related to esophageal cancer (OR=0.50, 95% CI: 0.32-0.78, and OR=0.48, 95% CI: 0.31-0.74, respectively), while no relationship with this cancer was observed for the *Starch-rich* (starch, vegetable protein, and sodium) characterized by high loadings on (OR=0.80, 95% CI: 0.50-1.28) and the *Other fats* (with high loadings on linoleic acid, linolenic acid, and vitamin E) patterns (OR=1.04, 95% CI: 0.67-1.63). The naming of the factors, based on high factor scores characterizing each pattern, was confirmed by the distributions of selected nutrients and food groups.

The subsequent cluster analysis, based on differences in the dietary patterns, yielded 6 clusters, one of which (C3) was characterized by the lowest intakes of all nutrients and food groups considered, while the remaining clusters were determined by an extreme value of the dietary patterns, one-by-one. Subjects in the C1 cluster were characterized by the highest values of the *Vitamins and fiber* pattern, subjects in the C2 cluster had the highest values of the *Other polyunsaturated fatty acids* pattern, the C4 cluster was characterized by the highest

scores of the *Animal products and related components*, subjects in the C5 cluster had the highest values of the *Other fats* pattern, the C6 cluster was characterized by the highest scores of the *Starch-rich* pattern and had the highest intakes of bread, and pasta and rice. Significant inverse relations were observed between the C1, C5 and C6 clusters (OR=0.59, 95% CI:0.40-0.88, OR=0.42, 95% CI:0.20-0.86, and OR=0.60, 95% CI: 0.42-0.86, respectively) – which were characterized by high values of the *Vitamins and fiber*, *Other fats*, and *Starch-rich* patterns, respectively – as compared to the C3 cluster. No significant risk was observed for the C2, and C4 clusters (OR=0.76, 95% CI: 0.51-1.13, and OR=1.29, 95% CI: 0.80-2.07).

Conclusion: The combined application of factor and cluster analyses, allows to identify key dietary aspects in a specific population, and to obtain mutually exclusive groups of subjects who are similar for these characteristics. The two techniques have limitations that arise from the subjective decisions involved in the analyses. In this application, various alternative options were tried, to check robustness and solution stability. Among these complementary analyses, results from PCFA were compared with those from another principal axis factoring, and those from PCFA analyses performed separately in strata of center and gender, and in randomly generated split samples. Moreover, the internal consistency of the identified patterns was evaluated using the Cronbach's coefficient alphas. All these checks supported the decisions adopted in the main analyses. As concern cluster analysis, to limit the influence of the starting point, the initial seeds used in the *k*-means method were obtained performing a hierarchical clustering (Ward's method) and cutting the corresponding dendrogram at the level k=6. Moreover, some alternative solutions were identified through different methods and distances, yielding comparable clustering solutions. Another limitation of cluster analysis is its sensitivity to the presence of outliers; however, the exclusion of 8 potential outliers did not materially change the results.

INTRODUCTION

Because of the complexity of diet and the potential interactions between dietary components, such as foods and nutrients, approaches that focus on single components may be not completely adequate to describe the complex relationships between diet and cancer etiology. To overcome this limitation, the use of dietary patterns has been proposed, given their ability to describe variations in overall dietary intakes in a specific population, thus allowing to better understand and describe the association between diet as a whole and cancer (Hu, 2002; Moeller *et al*, 2007; Newby *et al*, 2004). Moreover, allowing the identification of beneficial or detrimental dietary profiles in a specific population, they facilitate the dissemination of dietary recommendations aiming at the specific context.

Dietary patterns can be defined as combinations of dietary components aimed at summarizing diet as a whole, or key aspects of the diet for the population under consideration. Three main approaches have been proposed to identify dietary patterns: an exploratory or *a posteriori* approach, which empirically derives dietary patterns directly from the data; a hypothesis-oriented or *a priori* approach, which is based on the available evidence on the association between specific food components and the disease; methods combining characteristics of the exploratory and the hypothesis-oriented approaches, including *reduced rank regression* and *partial least square* analyses (DiBello *et al*, 2008; Hu, 2002; Meyer *et al*, 2011; Moeller *et al*, 2007; Newby *et al*, 2004).

Among exploratory methods, principal component and factor analyses are used to reduce the data dimension by transforming an original larger dataset of correlated dietary components into a smaller set of uncorrelated variables, called principal components or factors respectively, which explains the largest possible amount of variance of the original dietary

components. These analyses produce a continuous summary score for each subject and for each factor, which indicates the degree to which a subject's diet conforms to each identified dietary pattern. Cluster analysis is another exploratory method, which is applied to identify mutually exclusive groups of subjects with specific dietary habits, who are highly similar within group and highly dissimilar between different groups. Principal component or factor analysis and cluster analysis may be usefully applied in combination in an overall statistical strategy for data reduction and clustering. This may allow the joint identification of some key nutritional features in the population under examination, and of a partition of the subjects based on similarity in these characteristics.

In the *a priori* approach, dietary patterns are indexes or scores built upon scientific evidence or theories for the specific disease analyzed, or based on current nutritional guidelines, recommendations, or specific dietary compositions.

PRINCIPLES OF FACTOR ANALYSIS

Definition of the factor model

The essential aim of factor analysis is to describe the relationships among many variables in terms of a few underlying, but unobservable, random quantities called *factors*. The basic idea is that variables can be grouped according to their correlations. If variables within a particular group are highly correlated among themselves and have relatively small correlations with variables in a different group, then it is plausible that each group of variables represents a single underlying (and unobservable) factor (Johnson & Wichern, 2002).

The observable random vector $\mathbf{X}$, with p components, has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model postulates that $\mathbf{X}$ is linearly dependent upon a few unobservable random

variables F_1, F_2, …, F_m, called *common factors*, and p additional sources of variations ε_1, ε_2, …, ε_m, called *errors*. In particular, the factor analysis model is as follows:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

or, in matrix notation:

$$\underset{(p\text{x}1)}{\mathbf{X} - \boldsymbol{\mu}} \; \underset{=(p\text{x}m)(m\text{x}1)}{= \mathbf{L}\,\mathbf{F}} + \underset{(p\text{x}1)}{\boldsymbol{\varepsilon}}$$

The coefficient l_{ij} is called *factor loading* of the ith variable on the jth factor, so the matrix **L** is the *matrix of factor loadings*. The ith specific factor ε_i is associated only with variable X_i. The p deviations $X_1 - \mu_1$, $X_2 - \mu_2$, ..., $X_p - \mu_p$ are expressed in terms of $p + m$ random variables F_1, F_2, ..., F_m, ε_1, ε_2, ..., ε_p which are unobservable.

Moreover it is assumed that:

$$\underset{(m\text{x}1)}{E(\mathbf{F}) = \mathbf{0}}, \quad \underset{(m\text{x}m)}{\text{Cov}(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \mathbf{I}}$$

$$\underset{(p\text{x}1)}{E(\boldsymbol{\varepsilon}) = \mathbf{0}}, \; \underset{(p\text{x}p)}{\text{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Psi}} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$$

and that **F** and **ε** are independent, so:

$$\underset{(p\text{x}m)}{\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{0}}$$

With these assumptions constitute the *orthogonal factor model*. The *orthogonal factor model* implies a covariance structure for **X**, as follows:

$$(\mathbf{X} - \boldsymbol{\mu})\,(\mathbf{X} - \boldsymbol{\mu})' = (\mathbf{LF} + \boldsymbol{\varepsilon})(\mathbf{LF} + \boldsymbol{\varepsilon})'$$

$$= (\mathbf{LF} + \boldsymbol{\varepsilon})((\mathbf{LF})' + \boldsymbol{\varepsilon}')$$

$$= \mathbf{LF}(\mathbf{LF})' + \boldsymbol{\varepsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$$

so that

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}) = E(\mathbf{X} - \boldsymbol{\mu})\,(\mathbf{X} - \boldsymbol{\mu})'$$

$$= \mathbf{L}E(\mathbf{FF}')\mathbf{L}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{L}' + \mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$$

$$= \mathbf{LL}' + \boldsymbol{\Psi}$$

according to the previous assumptions on Cov(**F**) and Cov($\boldsymbol{\varepsilon}$).

Also assuming independence, Cov($\boldsymbol{\varepsilon}$,**F**) = $E(\boldsymbol{\varepsilon},\mathbf{F}')$= **0**.

Moreover, $(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = (\mathbf{LF} + \boldsymbol{\varepsilon})\mathbf{F}' = \mathbf{LFF}' + \boldsymbol{\varepsilon}\mathbf{F}'$, so Cov(**X**, **F**) = $E(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = \mathbf{L}E(\mathbf{FF}') + E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{L}$.

The model $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\,\mathbf{F} + \boldsymbol{\varepsilon}$ is linear in the common factors.

The proportion of variance of the *i*th variable contributed by the *m* common factors is called *i*th *communality*. The portion of Var(X_i) = σ_{ii} due to the specific factor is called the *uniqueness*, or *specific variance*. Indicating the communality with h_i^2 the variance σ_{ii} can also be written as:

$$\sigma_{ii} = \underbrace{l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2}_{\text{communality}} + \psi_i$$

Var (X_i) = communality + specific variance

or, substituting $h_i^2 = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2$ it is obtained:

$$\sigma_{ii} = h_i^2 + \psi_i \quad i = 1, 2, \ldots, p$$

where the *i*th communality is the sum of squares of the loadings of the *i*th variable on the *m* common factors.

When $m > 1$, there is always some ambiguity associated with the factor model. Let **T** be any m x m orthogonal matrix so that $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$, then the factor model can be written as:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} = \mathbf{LTT}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$$

where

$$\mathbf{L}^* = \mathbf{LT} \text{ and } \mathbf{F}^* = \mathbf{T}'\mathbf{F}$$

Since

$$E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$$

and

$$\text{Cov}(\mathbf{F}^*) = \mathbf{T}'\text{Cov}(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$$
(m x m)

it is impossible on the basis of observation on **X**, to distinguish the loadings **L** from the loadings $\mathbf{L}^*$. That is, the factors **F** and $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ have the same statistical proprieties, and even if the loadings $\mathbf{L}^*$ are different from the loadings **L**, they both generate the same covariance matrix **Σ**. This ambiguity provides the rationale for the *factor rotation*, since orthogonal matrices correspond to rotations (and reflections) of the coordinate system for **X**. Thus, the analysis proceeds by imposing conditions allowing to uniquely estimate **L** and **Ψ**. The loading

matrix is then rotated (i.e. multiplied by an orthogonal matrix), the rotation being determined by some criterion that facilitate the interpretation.

Estimation methods

The sample covariance matrix **S** is an estimator of the unknown population covariance matrix **Σ**. If the off-diagonal elements of **S** are small or those of the sample correlation matrix **R** are essentially zero, the variables are not related and thus a factor analysis should not be performed. In such a situation, the *uniqueness* play a dominant role, while the major purpose of factor analysis is to determine a few important *common factors*, that adequately describe the phenomenon under consideration.

If **Σ** appears to deviate significantly from a diagonal matrix, then a factor analysis can be applied, and the initial problem is to estimate the factor loadings l_{ij} and the specific variances ψ_i.

Among the methods of parameter estimation, the most popular are the *principal component* (and the related *principal factor*) *method* and the *maximum likelihood method*. The solutions of these methods can be rotated in order to simplify the interpretation of factors.

Principal component solution of the factor model

The principal component factor analysis of the sample covariance matrix **S** is specified in terms of its eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1)$, $(\hat{\lambda}_2, \hat{\mathbf{e}}_2)$, …, $(\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$. Let $m < p$ be the number of common factors, the matrix of the estimated factor loadings is given by:

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 \mid \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 \mid \ldots \mid \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m \right]$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}'$, so

$$\tilde{\Psi} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_1 \end{bmatrix} \quad \text{with } \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{l}_{ij}^2$$

The principal component factor analysis of the sample correlation matrix is obtained starting with **R** in place of **S**.

Maximum likelihood solution of the factor model

If the common factors **F** and the specific factors **ε** can be assumed to be normally distributed, then the maximum likelihood estimates of the factor loadings and specific variances may be obtained. When $\mathbf{F}_j$ and $\boldsymbol{\varepsilon}_j$ are jointly normal, the observations $\mathbf{X}_j - \boldsymbol{\mu} = \mathbf{L}\mathbf{F}_j + \boldsymbol{\varepsilon}_j$ are then normal and the likelihood can be written depending on **L** and **Ψ** through $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$. Since **L** is not uniquely defined, a computationally convenient *uniqueness condition* is imposed:

$$L'\Psi^{-1}L = \Delta, \text{ a diagonal matrix}$$

Number of factors to retain

A crucial aspect of factor analysis is the choice of the number of factors to retain. Three main criteria are generally used. The first one is to retain those factors with eigenvalues greater than 1, so that the unity represents the smallest eigenvalue for which a pattern is retained. Since the eigenvalue represent the amount of variance in all of the nutrients that can be explained by a given pattern, the retained patterns would account for more than their share of the total variance of the original nutrients. However, this should be considered as a minimum standard, and other thresholds above 1 may be chosen in some circumstances. The second criterion is to

add successive factors until the cumulative percentage of variance explained by the retained factors is satisfactory. To terminate the factor extraction process, a cumulative explained variance of 75-80% is considered adequate. The third one, is to plot the extracted factors against their eigenvalues in descending order of magnitude to identify distinct breaks in the slope of the plot, called *scree plot*. To determine where the break occurs, a straight line should be drawn with a ruler through the lower values of the plotted eigenvalues. That point where the factors curve above the straight line drawn through the smaller eigenvalues identifies the optimal number of factors to retain.

Finally, a researcher should also consider factor interpretability, in determining the number of factors to retain.

Factor rotation

All factor loadings obtained from the initial loadings by an orthogonal transformation have the same ability to reproduce the covariance (or correlation) matrix. If $\hat{\mathbf{L}}$ is the *p* x *m* matrix of estimated factor loadings obtained by any method, then

$$\hat{L}^{*} = \hat{L}T \quad \text{where } \mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$$

is a *p* x *m* matrix of "rotated" loadings. The estimated covariance (or correlation) matrix remains unchanged, since:

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}\mathbf{T}\mathbf{T}'\hat{\mathbf{L}} + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}^{*}\hat{\mathbf{L}}^{*\prime} + \hat{\mathbf{\Psi}}$$

This equation indicates that also the residual matrix remains unchanged:

$$\mathbf{S}_n - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\mathbf{\Psi}} = \mathbf{S}_n - \hat{\mathbf{L}}^{*}\hat{\mathbf{L}}^{*\prime} - \hat{\mathbf{\Psi}}$$

Moreover the specific variances $\hat{\psi}_i$, and so the communalities $\hat{h}_i^2$, are unchanged as well. Thus, from a mathematical point of view, it doesn't make any difference whether $\hat{\mathbf{L}}$ or $\hat{\mathbf{L}}^*$ is obtained. Since the original loadings may not be easily interpretable, it is usual practice to rotate them until a "simple structure" is obtained. Ideally, it is desirable to have a pattern of loadings such that each variable loads highly on a single factor and has small-to-moderate loadings on the remaining factors.

The simplest case of rotation is an *orthogonal rotation* in which the angle between the reference axes of factors are maintained at 90 degrees; this implies that the rotated factors remain uncorrelated. Other forms of rotation, indicated as *oblique rotations*, allow the angle between the reference axes to vary, i.e., factors are allowed to be correlated with each other. Orthogonal rotation procedures are more commonly used than oblique rotation procedures, and should be performed when the common factors are assumed to be independent. In other situations where the correlations between the underlying constructs are not assumed to be zero, oblique rotations may yield simpler and more interpretable factor solutions.

Peculiar issues of factor analysis in nutritional epidemiology

In nutritional epidemiology, the factors identified through factor analysis are commonly called *dietary patterns*.

In addition to the subjective decisions necessary throughout the factor analysis process, which may have an impact on the number and type of patterns identified, a relevant issue in this field is the choice of the type of dietary data to work on. Nutrients are continuous variables, while food groups are discrete variables with peaky distributions where a few categories have very high frequencies and most categories have frequencies near zero. Thus factor analysis would be more properly applied on nutrients. Moreover, compared to food groups, nutrients offer the

advantage of being directly involved in the biological processes, although they depend on the food groups they are derived from.

Another relevant decision concerns the list of variables to be included in the analysis, which should be comprehensive enough to represent the overall diet in the population under investigation.

Moreover, the researcher may choose some data transformation: data standardization would be useful to avoid problems related to the use of different measurement units and amount of intake; logarithmic transformation may be applied to obtain normal distributed variables, although interpretation in this case would become more difficult; moreover, energy-adjustment may be performed as well.

PRINCIPLES OF CLUSTER ANALYSIS

Clustering may be defined as the partitioning of a set of observations into groups so that observations within a group are "similar" and observations in different groups are "dissimilar". Thus, the essential aim of cluster analysis is the identification of "natural" structures in a dataset. Clustering can provide an informal means for assessing dimensionality, identifying outliers, and suggesting interesting hypothesis concerning relationships. Clustering is different from classification, since the latter imply the allocation of the observations to known groups, and the objective is to assign new observations to one of these predefined groups. On the contrary, in cluster analysis no *a priori* assumptions are made on the number and characteristics of the groups, and grouping is based on similarities or distances between the observations at hand. Thus, the inputs required are similarity measures.

Similarity measures

Most efforts to obtain a rather simple group structure from a complex dataset require the choice of a measure of "closeness" or "similarity". A great subjectivity is involved in this choice, and it may considerably influence the final solution. When observations are clustered, proximity is generally indicated by some kind of distance, while variables are usually grouped on the basis of correlation coefficients or measures of associations.

The statistical distance between two observations is of the form:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \mathbf{A} (\mathbf{x} - \mathbf{y})}$$

Usually, $\mathbf{A} = \mathbf{S^{-1}}$, where $\mathbf{S}$ contains the sample variances and covariances. However, without previous knowledge of the distinct groups, these sample quantities cannot be computed. For this reason, Euclidean distance is often preferred for clustering.

A general distance measure is the Minkowski metric, defined as follows:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{i=1}^{p} |\boldsymbol{x}_i - \boldsymbol{y}_i|^m\right]^{1/m}$$

For $m = 1$, $d(\boldsymbol{x}, \boldsymbol{y})$ measures the Manhattan or "city-block" distance between two points in p dimensions. For $m = 2$, $d(\boldsymbol{x}, \boldsymbol{y})$ becomes the Euclidean distance. For $m \to \infty$ $d(\boldsymbol{x}, \boldsymbol{y})$ becomes the Lagrange distance. In general, varying *m* changes the weight given to larger and smaller differences.

Other definitions may be adopted depending on the context and aim of the analysis.

Whenever possible, the measure adopted should satisfy the following proprieties:

given two points *P* and *Q*, and an intermediate point *R*,

$d(P, Q) = d(Q, P)$

$d(P, Q) > 0 \; if \; P \neq Q$

$d(P, Q) = 0 \; if \; P = Q$

$d(P, Q) \leq d(P, R) + d(R, Q)$ (triangle inequality)

Hierarchical Clustering Methods

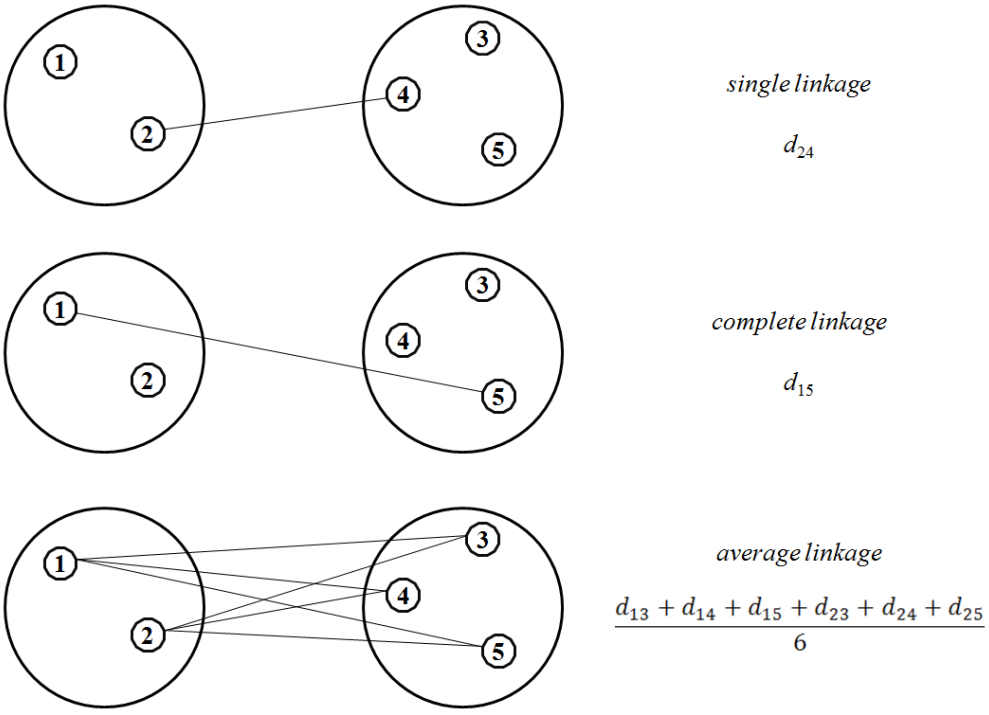
Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. *Agglomerative hierarchical methods* start with individual objects. Thus, there are initially as many clusters as objects. The most similar objects are first

grouped, and these initial groups are then merged according to their similarities. *Divisive hierarchical methods* work in the opposite direction. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" those in the other group. These subgroups are further divided into dissimilar subgroups.

The results of both agglomerative and divisive methods may be displayed in the form of two-dimensional diagram, known as dendrogram.

Among agglomerative hierarchical methods, *single linkage*, *complete distance*, and *average distance* methods are largely used. In the *single linkage* method, groups are merged according to the distance between the nearest members. In the *complete linkage* method groups are merged according to the distance between the farthest members. In the *average linkage* method groups are merged according to the average distance between pairs of members.

Figure 1 – examples for single, complete and average linkage



Ward's method

Another hierarchical method is the *Ward's method*, based on the minimization of the "loss of information" from joining two groups. The method is usually implemented with loss of information represented by an increase in an error sum of squares criterion, ESS. Let ESS_k be, for a given cluster *k*, the sum of squared deviations of every observation in the cluster from the cluster mean. If there are *K* clusters, ESS is defined as follows:

$$ESS = \sum_{j=1}^{K} ESS_k$$

At each step of the analysis, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS (minimum loss of information) are joined. Initially, each cluster consists of a single observation and, if there are *N* observations, the value of ESS is given by

$$ESS = \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$$

where $\mathbf{x}_j$ is the multivariate measurement associated with the *j*th observation and $\bar{\mathbf{x}}$ is the mean of all the observations.

Nonhierarchical Clustering Methods

Nonhierarchical clustering techniques are designed to group observations into a collection of *K* clusters. The number of clusters *K* may be either fixed in advance or determined as part of the clustering procedure. Since it is not required to calculate the matrix of distances, nonhierarchical methods can be applied to much larger datasets than can hierarchical techniques.

These methods starts either from (a) an initial partition of observations into groups or (b) an initial set of seed points, which will form the cores of clusters. One way to start is to randomly select seed points from among the observations or to random partition the observations into initial groups.

K-means Method

The *K-means* method assigns each observation to the cluster having the nearest center (mean). The process is composed of the following steps:

(a) Partition the observations into *K* initial clusters.

(b) Proceed through the list of observations, assigning an observation to the cluster whose center (mean) is nearest. Recalculate the center for the cluster receiving the new observation and for the cluster losing the observation.

(c) Repeat the procedure until no more reassignments occur.

Rather than starting with a partition of all the observations into *K* preliminary groups, it is possible to specify *K* initial centers (seed points) and then proceed as before. The final assignment of the observations to clusters will be, to some extent, dependent upon the initial partition or the initial selection of the seed points.

Choice of the number of clusters

Some arguments weigh for not fixing the number of clusters in advance, including the following:

(a) If two or more seed points inadvertently lie within a single cluster, their resulting clusters will be poorly differentiated.

(b) The existence of an outlier might produce at least one group with very disperse observations.

(c) Even if the population is known to consist of *K* groups, the sampling method may be such that data from the rarest group do not appear in the sample. Forcing the data into *K* groups would lead to nonsensical clusters.

Thus, it is usually preferable to repeat the clustering procedure and to compare the different solutions obtained varying the number of clusters *K*.

A partition can be said of good quality when:

(a) Observations within a cluster are homogeneous (small within-cluster variability).

(b) The observations in one cluster differ from those in other ones (high between-clusters variability).

Peculiar issues of cluster analysis in nutritional epidemiology

The choice of the distance measure has consequences on the identified clusters, which are relevant also from a nutritional standpoint.

In the subsequent application, some different distance measures will be used, including:

(a) Euclidean distance:

$$\left(\sum_{i=1}^{p}|x_i - y_i|^m\right)^{1/m}, \qquad m = 2 \rightarrow \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

(b) Manhattan distance:

$$\left(\sum_{i=1}^{p}|x_i - y_i|^m\right)^{1/m} \quad , \qquad m = 1 \rightarrow \sum_{i=1}^{p}|x_i - y_i|$$

(c) Lagrange distance:

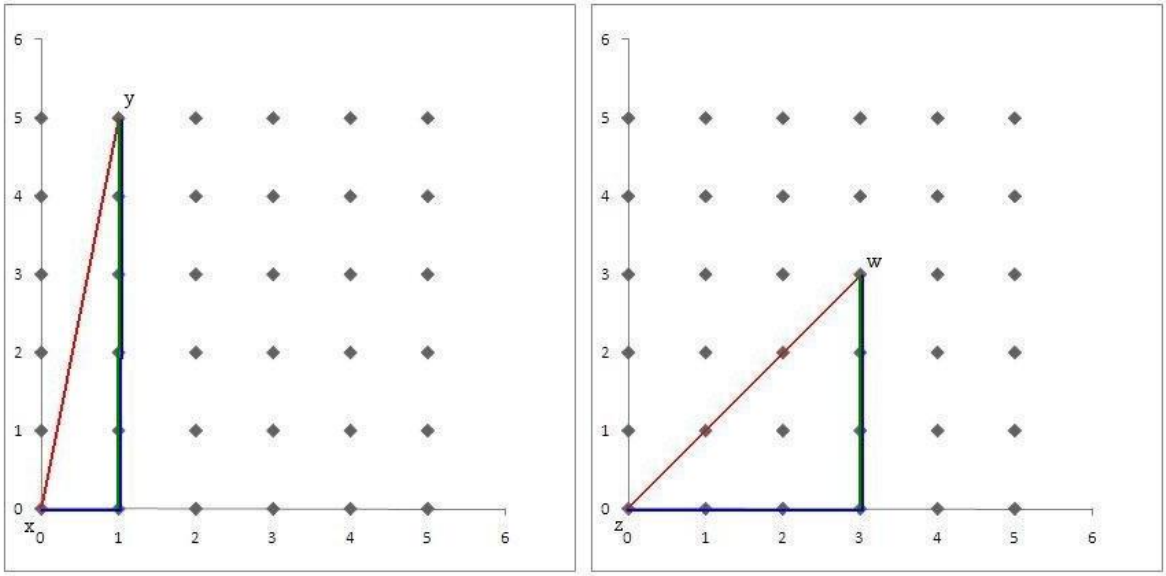
$$\lim_{m\to\infty}\left(\sum_{i=1}^{p}|x_i - y_i|^m\right)^{1/m} = \max_i|x_i - y_i|$$

(d) Correlation coefficient similarity measure:

$$\frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

Figure 2 and **Figure 3** show 2-dimensional examples on these different measures; these examples are provided to discuss the proprieties of these measures, and their use in nutritional epidemiology.

Figure 2 – examples for Manhattan, Euclidean and Lagrange distances



$x = (0,0)\ y = (1,5)$

$d_{Man}\ (x, y) = |1 + 5| = 6$

$d_{Eucl}(x, y) = \sqrt{1^2 + 5^2} = \sqrt{26} = 5.1$

$d_{Lagr}\ (x, y) = \max(1,5) = 5$

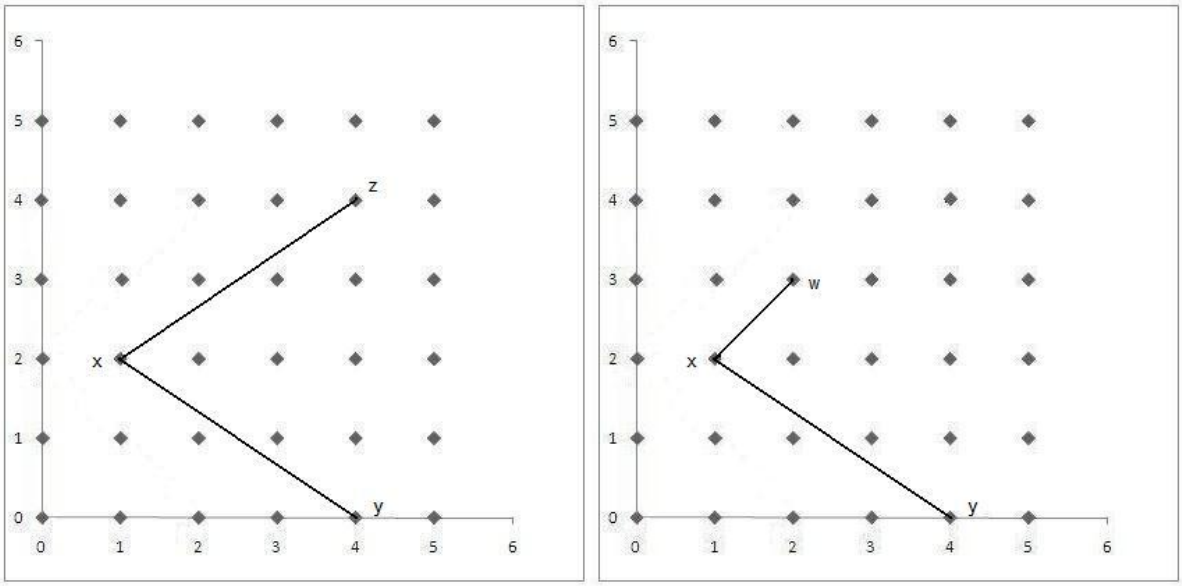
$z = (0,0)\ w = (3,3)$

$d_{Man}\ (z, w) = |3 + 3| = 6$

$d_{Eucl}(z, w) = \sqrt{3^2 + 3^2} = \sqrt{18} = 4.2$

$d_{Lagr}\ (z, w) = \max\ (3,3) = 3$

Figure 3 – examples for the Correlation coefficient similarity measure



$x = (1,2)\ \ y = (4,0)\ \ z = (4,4)\ \ w = (2,3)$

$$d_{Correl}(x,y) = \frac{(1-2.5)(2-1)+(4-2.5)(0-1)}{\sqrt{[(1-2.5)^2+(4-2.5)^2][(2-1)^2+(0-1)^2]}}$$

$$= \frac{-1.5(1)+1.5(-1)}{\sqrt{[(-1.5)^2+1.5^2][1^2+(-1)^2]}} = \frac{-1.5-1.5}{\sqrt{[2.25+2.25][1+1]}} = \frac{-3}{\sqrt{4.5*2}}$$

$$= \frac{-3}{\sqrt{9}} = \frac{-3}{3} = -1$$

$$d_{Correl}(x,z) = \frac{(1-2.5)(2-3)+(4-2.5)(4-3)}{\sqrt{[(1-2.5)^2+(4-2.5)^2][(2-3)^2+(4-3)^2]}}$$

$$= \frac{-1.5(-1)+1.5(1)}{\sqrt{[(-1.5)^2+1.5^2][(-1)^2+1^2]}} = \frac{1.5+1.5}{\sqrt{[2.25+2.25][1+1]}} = \frac{3}{\sqrt{4.5*2}}$$

$$= \frac{3}{\sqrt{9}} = \frac{3}{3} = 1$$

$$d_{Correl}(x,w) = \frac{(1-1.5)(2-2.5)+(2-1.5)(3-2.5)}{\sqrt{[(1-1.5)^2+(2-1.5)^2][(2-2.5)^2+(3-2.5)^2]}}$$

$$= \frac{-0.5(-0.5)+0.5(0.5)}{\sqrt{[(-0.5)^2+0.5^2][(-0.5)^2+0.5^2]}} = \frac{0.25+0.25}{\sqrt{[0.25+0.25][0.25+0.25]}}$$

$$= \frac{0.5}{\sqrt{0.5*0.5}} = 1$$

These examples shows that:

(a) $d_{Manhattan}(P,Q) \geq d_{Euclidean}(P,Q) \geq d_{Lagrange}(P,Q)$

(b) the Euclidean and Lagrange distances provide an higher measure when the difference between *P* and *Q* is unbalanced on its dimensions, while the Manhattan distance provide the same measure when the sum of the distances of each dimension is the same, independently of the presence of unbalance.

(c) the Correlation coefficient similarity measure does not satisfy the condition that $d(P,Q) > 0\ if\ P \neq Q$.

(d) the Correlation coefficient similarity measure depends more on the angular coefficient between the segment joining *x* and *y* and the horizontal axis, than on the position in the space of the two points. In the examples in Figure 3 $d_{Correl}(x,y) \neq d_{Correl}(x,z)$, although *y* and *z* have the same distance in the space from *x*; further, $d_{Correl}(x,y) < d_{Correl}(x,w)$, although *w* is closer to *x* than *y*; moreover, $d_{Correl}(x,z) = d_{Correl}(x,w)$, although *z* and *w* have different distances in the space from *x*.

This, from a nutritional point of view, implies that the Euclidean and Lagrange distances are more appropriate to identify subjects with unbalanced nutritional components, while the Manhattan distance gives greater emphasis to overall larger dietary intakes, without taking into account the presence of unbalance and which nutritional component most contribute to increase the intake.

The Correlation coefficient similarity measure does not take into account unbalances nor overall differences in nutritional intake between subjects, but allows to identify trends in differences in nutritional intakes.

APPLICATION ON DATA FROM A CASE-CONTROL STUDY

This section describes an application of factor and cluster analyses on data from an Italian case-control study on esophageal cancer (Bravi *et al*).

Design and participants

A case-control study, carried out between 1992 and 1997 in the provinces of Milan, Pordenone and Padua, included 304 subjects (275 men and 29 women) with incident, histologically confirmed diagnosis of squamous cell carcinoma of the esophagus under 77 years (median age, 60 years), and 743 hospital controls (593 men and 150 women) under 77 years (median age, 60 years). Controls were subjects admitted to the same network of hospitals as cases for a wide spectrum of acute, non neoplastic conditions, not related to smoking, alcohol consumption, or long-term modifications of diet. Controls were frequency matched with cases by 5-year age groups, sex, period of interview and study center, with a control-to-case ratio of about 2 for men and about 5 for women. Response rate was greater than 95% for both cases and controls.

For both cases and controls, data were collected during their hospital stay by centrally trained interviewers. The questionnaire included information on socio-demographic characteristics, anthropometric measures, selected lifestyle habits, such as tobacco smoking and alcohol drinking, a personal medical history and a family history of cancer. A satisfactorily reproducible and valid (Decarli *et al*, 1996; Franceschi *et al*, 1993) food frequency questionnaire (FFQ) was used to assess the patients' usual diet in the two years before diagnosis (for cases) or hospital admission (for controls). The FFQ included questions on 78 foods and beverages, including a range of the most common recipes in Italian diet. Subjects

were asked to indicate the average weekly frequency of consumption and corresponding portion size (small, medium, large) for each dietary item. To estimate micro- and macro-nutrients, an Italian food composition database was used, integrated with other sources, when needed (Gnagnarella *et al*, 2004; Salvini *et al*, 1998). Losses due to cooking were subtracted from the computation of the content of vitamins, when appropriate.

Statistical analysis: factor analysis

Variable selection

As discussed before, a crucial role in factor analysis is played by the selection of the variables to be included in the analysis. According to the previous considerations, in this application nutrients were preferred to food groups. The analyses were carried out on 28 macro- and micro-nutrients. Existing relationships among nutrients were evaluated to avoid overrepresentation of single nutrients and subsequent artificially higher correlations.

Factorability of the original matrix

A preliminary examination of the correlation matrix of the original nutrients was carried out to assess its factorability. In particular, a visual inspection was done to identify variables that were:

- too highly correlated ($r \geq 0.80$) with one another; this situation reflects problems of multicollinearity, so that one or more of these variables would be dropped from the analysis;
- not sufficiently correlated ($r < 0.30$) with one another; this means that these variables will not share much of the common variance, thus potentially leading to solutions with patterns characterized by a single nutrient.

Moreover, measures of sampling adequacy that compare the simple and partial correlation coefficients may be defined either overall or for single variables. The overall measure, called Kaiser-Meyer-Olkin statistic (KMO), is defined as follows (Pett *et al*, 2003):

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2}$$

where ΣΣ are the sum over all variables in the matrix when variable $i \neq$ variable j, r_{ij} is the Pearson correlation coefficient between i and j, and a_{ij} the partial correlation coefficient between i and j. Individual measures of sampling adequacy are computed using only the simple and partial correlation coefficients involving the specific variable under consideration. The overall and individual measures range between 0 and 1. Small values indicate that the squared Pearson correlation coefficient is small relative to the squared Pearson correlation coefficient and therefore a factor analysis may be imprudent. If the sum of the squared partial correlation coefficients is small compared with the sum of the squared correlation coefficients, the measures approach 1.

Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is an identity matrix. It is a chi-square test (Pett *et al*, 2003), whose statistic is defined as follows:

$$\chi^2 = -\left[(N-1) - \left(\frac{2k+5}{6}\right)\right] \log|R|$$

where the χ^2 is the calculated chi-square value for Bartlett's test, N is sample size, k is the number of variables in the matrix and $|\boldsymbol{R}|$ the determinant of the correlation matrix. The degrees of freedom for this χ^2 statistic are $k(k-1)/2$. Larger values of the test suggest that the null hypothesis should be rejected.

Since Bartlett's test statistic depends explicitly on the sample size, *N*, for larger samples this test tends to indicate that the correlation matrix is not an identity matrix. For this reason, it should be used only as a minimum standard for assessing the quality of the correlation matrix.

Identification of dietary patterns through factor analysis

Exploratory factor analysis was performed according to the principal component method (PCFA hereafter). This approach assumes that the variables included in the analysis can be calculated by the extracted components or factors. Because each standardized variable has a mean of 0 and a variance of 1, the initial estimate of communality for each variable is 1. This is what will be placed initially on the diagonal of the correlation matrix. The first pattern is a linear combination of the original nutrients, such that it explains the maximum amount of the variance among the original nutrients. After this extraction, a residual correlation matrix is created, which contains the variances not explained by the first pattern on the diagonal, and the partial correlations of the nutrients with each other after the extraction of the first pattern on the off-diagonal. The second pattern is then extracted from this residual matrix, thus it is uncorrelated to the first one. This procedure of extraction is repeated on subsequent residual matrices, until the elements in the residual variance-covariance matrix are reduced to random errors.

Estimation of factor scores

Factor scores were estimated for each subject and pattern. They indicate the degree to which each subject's diet conforms to one of the identified dietary profiles. In the main analysis they were calculated through the weighted least square method, where variables that have lower loadings on the pattern are given less weight than those with higher loadings, in the calculation of factor scores.

Choice of the number of dietary patterns to retain

As discussed before, the choice of the number of patterns to retain is a crucial decision in performing factor analysis. In this application this choice was based on the combination the following three criteria: examination of the scree plot; factor eigenvalue greater than 1; and pattern interpretability.

Rotation of the identified dietary patterns

To improve the interpretation of the identified patterns, a rotation was performed. An orthogonal rotation was chosen, since it was assumed that the dietary patterns are independent. Specifically, a varimax rotation was performed, that consists in rotating the axes to orientations that maximize the variances of the loadings within the patterns, while maximizing differences between the high and low loadings on a particular pattern.

Naming of the identified dietary patterns

The interpretation and naming of the identified dietary patterns, was based on those nutrients having factor loadings greater or equal to 0.63 in absolute value on a given pattern. Since the contribution that a pattern gives to a nutrient's sample variance is equal to the square of its loading on that pattern, a threshold of 0.63 warrants a minimum contribution of the pattern on the nutrient variance of approximately 0.40.

Evaluation of the identified solution

To determine the internal consistency of the identified patterns the Cronbach's coefficient alphas (α) were examined, considering those nutrients having rotated factor loading greater or equal to 0.40 in absolute value on any pattern (Cronbach, 1951; Pett *et al*, 2003). The standardized *Cronbach's coefficient alpha when item deleted* were also calculated for each

pattern and for each variable. This measure of reliability represents the proportion of total variance in a given pattern that can be attributed to a common source.

The general formula for α is given as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sum \sigma_i^2 + 2\sum \sigma_{ij}}\right)$$

where *k* is number of variables, $\sum \sigma_i^2$ is the sum of the variances of the variables and $\sum \sigma_{ij}$ is the sum of the covariances of all possible pairs of variables.

When the variances of the items are all equal, the formula for standardized coefficient α is given as follows:

$$\alpha = \frac{k\bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}}$$

Values for α should range between 0 and 1. If there is little correlation among the variables, α will be equal to 0. The higher the correlation among the variables, the higher will be the value of α. Its value is influenced not only by the size of the correlation among the variables but also by the number of variables in the set. Indeed, increasing the number of variables will increase the size of α, even when the correlations among the variables are small.

To evaluate the robustness of the dietary patterns identified with PCFA, a series of checks of the identified solution were carried out. First, another estimation method was applied, specifically principal axis factoring. Briefly, it consists in adopting the squared multiple correlation coefficients of each variable with all the other ones as an estimate of the initial communality. Then, the analysis is undertaken in the same way as that outlined for PCFA. This approach gave essentially the same results as PCFA.

Second, PCFA analyses were carried out separately within subgroups of gender.

Third, factor scores were also calculated applying the multiple regression method, as follows:

$$\widehat{P}_{ij} = \sum_{k=1}^{p} W_{jk} z_{jk}$$

$\widehat{P}_{ij} =$ estimated standardized score for respondent *i* on pattern *j*

$W_{jk} =$ factor score coefficient for variable *k* on pattern *j*

and standardizing the results (Pett *et al*, 2003). The correlations between scores referring to the same factor calculated with different methods were 0.99 for all the comparisons.

Fourth, to confirm internal reproducibility of the identified patterns, subjects were randomly placed into one of two equally sized groups, or split-samples, and factor analysis was performed separately in both split-samples using the same approach of the main analysis. Each split sample contained cases selected by chance together with the corresponding matched controls.

Interpretation of the identified solution

To further facilitate the interpretation of the identified dietary patterns the Spearmen rank correlation coefficients were calculated between the continuous factor scores derived in the main analysis and the weekly number of portions of 29 selected food groups defined in the same dataset.

Risk estimates

For each dietary pattern, subjects were grouped into four categories according to the quartile distribution of factor scores among controls. Odds ratios (OR) and the corresponding 95%

confidence intervals (CI) were estimated for each category using unconditional multiple logistic regression models (Breslow & Day, 1980). Separate models for each dietary pattern were fitted; a composite model which included all the dietary patterns simultaneously was also fitted. A set of potential confounding variables and risk factors was also included in the models, in both situations.

Statistical analysis: cluster analysis

Selection of input variables

Cluster analysis was performed on factor scores of the five dietary patterns obtained from the main analysis based on PCFA.

Examination of potential outliers

Since cluster analysis is sensitive to the presence of outliers, an examination of the role of potential outliers was performed through two strategies. First, the matrix of Euclidean distances between the subjects was calculated, and the minimum distance was extracted for each subject. These distances were plotted, to identify subjects who were potentially apart from the population. Second, since, clustering is useful in itself to identify potential outliers, a series of cluster analysis was performed (with *K*-means method and Euclidean distance), with predefined number of clusters equal to 30 to identify small groups, changing the starting point at random.

Choice of the number of clusters

To identify the most reasonable number of clusters in a population, several indexes and statistical tests have been proposed, which are generally based on the sum of squares within

and between the clusters (Milligan & Cooper, 1985). However, there are no completely satisfactory methods for determining the number of population clusters for any type of cluster analysis (Everitt, 1979).

To identify the appropriate number of clusters, a series of cluster analyses was performed using *K*-means method with Euclidean distance and varying the predefined number of clusters from 3 to 15. Results were compared on the basis of 3 measures:

(a) the R^2 pooled for overall variables;
(b) the Cubic Clustering Criterion (CCC);
(c) the Pseudo-*F* statistic.

The CCC was developed by Sarle (Sarle, 1983), and is based on the null hypothesis that the data has been sampled on an hyperbox, against the alternative hypothesis that the data has been sampled from a mixture of spherical multivariate normal distributions, with equal variances and sampling probabilities.

The Pseudo-F statistic, developed by Calinski and Harabasz (Calinski & Harabasz, 1974), measures the separation among the clusters at the current level in the hierarchy; large values indicate that the mean vectors of all the clusters are different.

Method and distance measure

The main analysis was carried out with the *K*-means clustering method, using Euclidean distance. Since this method is highly sensitive to the initial selection of cluster centers, a preliminary cluster analysis was performed through a hierarchical method (specifically Ward's method); the dendrogram obtained through this method was cut at the *k* level, and the

resulting centers were used as initial seeds to carry out the *K*-means clustering (Venables & Ripley, 2002).

Comparison with clustering obtained through other methods and distance measures

The results of the main analysis were compared with clustering obtained through other methods and distance measures. The *K*-means clustering method was carried out using also the Manhattan, Lagrange, and Correlation coefficient similarity measure distance. These clustering solutions were compared to that of the main analysis through a concordance table of the frequency distributions. The proportion of observed agreement and the *k* stastistic were computed (Landis & Koch, 1977). Then, the four clustering solutions were compared in terms of distribution of the factor scores of the five identified dietary patterns.

Other clustering solutions were obtained through the Partitioning Around Medoids methods, both using the Euclidean and Manhattan distances (Kaufman & Rousseeuw, 1990).

Interpretation of the clustering solution

The main clustering solution, based on the *K*-means method and Euclidean distance, was characterized by examining the distributions of the five identified dietary patterns, both in the tabular form and through a scatterplot. Moreover, the distributions of several sociodemographic and lifestyle variable, and selected nutrients and food groups, within each cluster, were examined.

Risk estimates

The ORs were estimated for each of the identified clusters. Corresponding 95% CIs were estimated referring to floating absolute risks method (Easton *et al*, 1991). This method assigns a floating standard error (SE) to each cluster, that is independent of the choice of the reference

category. A CI for the OR between two groups can then be calculated from the floating SEs and is indicated as floating confidence interval (FCI). Floating SE estimates have been derived from a covariance structure model applied to the covariance matrix of the log relative risk estimates (Plummer, 2004).

Results

The correlation matrix resulted adequate to perform factor analysis. All the nutrients showed at least 10 correlation coefficients greater or equal to 0.30 in absolute value, thus allowing to perform the analyses on the whole set of selected nutrients. Bartlett's test of sphericity allowed to reject the null hypothesis that the correlation matrix is an identity matrix (p-value < 0.0001). The Keiser-Meyer-Olkin statistic was equal to 0.83, suggesting that the sample size was adequate to the number of nutrients. The individual measures of sampling adequacy were generally very high, with 26 nutrients having measures greater or equal to 0.70. Overall, the correlations among the nutrients were strong enough to indicate that the correlation matrix was factorable.

Table 1 gives the factor loading matrix for the five retained dietary patterns. These patterns explained more than 79% of the total variance in the original nutrients, accounting for about 25%, 15%, 15%, 12%, and 11% respectively. The first pattern, named *Animal products and related components* had the greatest loadings on calcium, phosphorus, riboflavin, animal protein, saturated fatty acids, cholesterol, and zinc. The second pattern, named *Vitamins and fiber*, was based on vitamin C, total fiber, beta-carotene equivalents, soluble carbohydrates, and total folate. The third pattern, named *Starch-rich*, was characterized by starch, vegetable protein, and sodium. The fourth pattern, named *Other polyunsaturated fatty acids and vitamin*

D, had high loadings on other polyunsaturated fatty acids, vitamin D, and niacin. The fifth pattern, named *Other fats* was based on linoleic acid, linolenic acid, and vitamin E.

Table 1 - Factor loading matrix[1] and explained variance (VAR) for the five major dietary patterns identified.

Nutrient	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
Animal protein	**0.76**	0.11	0.23	0.46	0.18
Vegetable protein	0.29	0.29	**0.85**	0.11	0.19
Cholesterol	**0.69**	0.11	0.25	0.40	0.21
Saturated fatty acids	**0.76**	0.18	0.27	0.22	0.29
Monounsaturated fatty acids	0.33	0.30	0.27	0.40	0.41
Linoleic acid	0.19	0.12	0.17	0.16	**0.88**
Linolenic acid	0.26	0.14	0.15	0.11	**0.87**
Other PUFAs	0.24	0.16	0.15	**0.86**	0.17
Soluble carbohydrates	0.46	**0.63**	0.18	-	-
Starch	0.31	0.12	**0.88**	-	0.14
Sodium	0.59	0.11	**0.66**	-	0.12
Calcium	**0.87**	0.20	0.10	-	-
Potassium	0.54	0.53	0.39	0.31	0.22
Phosphorus	**0.82**	0.19	0.38	0.24	0.19
Iron	0.48	0.15	0.40	0.37	0.26
Zinc	**0.67**	0.20	0.50	0.36	0.20
Thiamin (vitamin B1)	0.57	0.45	0.46	0.28	0.19
Riboflavin (vitamin B2)	**0.82**	0.36	0.16	0.15	0.13
Vitamin B6	0.51	0.49	0.38	0.45	0.23
Total folate	0.49	**0.63**	0.37	0.21	0.18
Niacin	0.36	0.34	0.42	**0.64**	0.24
Vitamin C	0.14	**0.85**	-	0.18	-
Retinol	0.45	-	-	0.30	-
Beta-carotene equivalents	-	**0.70**	-	0.12	0.27
Lycopene	-	0.17	0.55	0.32	0.14
Vitamin D	0.25	0.18	-	**0.82**	0.13
Vitamin E	0.18	0.46	0.22	0.33	**0.70**
Total fiber	0.20	**0.78**	0.39	0.16	0.14
Proportion of explained VAR (%)	25.07	15.49	15.09	12.83	10.72
Cumulative explained VAR (%)	25.07	40.55	55.64	68.46	79.18

[1]Estimated from principal component factor analysis performed on 28 nutrients. [2] Loadings greater or equal to 0.63 in absolute value were shown in bold typeface; loadings smaller than 0.10 in absolute value were not shown. PUFA: polyunsaturated fatty acids.

Table 2 shows the values for the standardized Cronbach's coefficient alpha for each dietary pattern and for standardized Cronbach's *coefficient alpha when item deleted* for each pattern and nutrient. Coefficient alphas for each pattern were equal to 0.966, 0.941, 0.931, 0.935 and 0.891, respectively. Almost all of the standardized *coefficient alphas, when item deleted*, were

lower than the corresponding overall standardized coefficient alpha for the same pattern, but the differences were small. These results indicate that all of the nutrients are contributing to the pattern, and none of the nutrients would materially modify the value of coefficient alpha if removed from the pattern. There might be some gain in removing a few nutrients; however, the benefit would be limited, and the interpretation of the patterns would become less convincing. Moreover, the examination of the *coefficient alphas, when item deleted* for each nutrient and pattern allowed to confirm the selection of the dominant nutrients selected according to the 0.63 cut-off.

Table 2 - Standardized Cronbach's coefficient alpha for each pattern and standardized Cronbach's *coefficient alpha when item deleted* for each pattern and nutrient.

	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
Standardized alpha	0.966	0.941	0.931	0.935	0.891
Standardized alpha *when item deleted*					
Deleted nutrient					
Animal protein	0.963			0.920	
Vegetable protein			0.913		
Cholesterol	0.964			0.928	
Saturated fatty acids	0.963				
Monounsaturated fatty acids				0.936	0.914
Linoleic acid					0.857
Linolenic acid					0.858
Other PUFAs				0.924	
Soluble carbohydrates	0.967	0.938			
Starch			0.919		
Sodium	0.965		0.921		
Calcium	0.965				
Potassium	0.962	0.928			
Phosphorus	0.961				
Iron	0.965		0.927		
Zinc	0.961		0.913		
Thiamin (vitamin B1)	0.962	0.930	0.917		
Riboflavin (vitamin B2)	0.962				
Vitamin B6	0.962	0.927		0.918	
Total folate	0.963	0.927			
Niacin			0.919	0.915	
Vitamin C		0.938			
Retinol	0.971				
Beta-carotene equivalents		0.944			
Lycopene			0.944		
Vitamin D				0.929	
Vitamin E		0.937			0.803
Total fiber		0.929			

Table 3 shows the values of the Spearman rank correlation coefficients between continuous factor scores derived from the main factor analysis and weekly portion for 29 selected food groups defined on the same data. The *Animal products and related components* pattern had the highest values for cheese, milk, eggs, liver, red meat, sugar and candies, and butter and margarine. For the *Vitamins and fiber* pattern, the highest values were found for citrus fruit,

other fruit, fruiting vegetables, leafy vegetables, cruciferous vegetables, other vegetables, cruciferous vegetables, and olive oil. For the *Starch-rich* pattern the highest values were found for bread, pasta and rice, and red meat. The *Other polyunsaturated fatty acids and vitamin D* pattern had the highest values for fish, white meat, red meat, and olive oil. The *Other fats* pattern had an high value for unspecified seed oil.

Table 3 – Spearman rank correlation coefficients between continuous factor scores derived from factor analysis and weekly number of portions for 29 selected food groups defined on the same data.

Food groups	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
Milk[1]	**0.49**	**0.27**	-0.15	-0.11	-
Coffee	-	-	-	-	-
Tea and decaffeinated coffee	-	0.12	-	-	-
Bread	0.15	-	**0.74**	-	-
Pasta and rice	-	-	**0.37**	0.17	-
Soups	0.16	-	-	-	-
Eggs	**0.33**	-	-	0.19	0.17
White meat	-	0.13	-	**0.41**	-
Red meat	**0.27**	-	**0.31**	**0.36**	0.19
Liver	**0.31**	-	-	0.23	-
Processed meat	0.19	-	0.12	0.22	-
Fish	-	0.13	-0.11	**0.61**	-
Cheese	**0.63**	-	-	-	-
Potatoes	0.18	0.10	0.14	0.14	0.21
Pulses	0.11	0.19	0.12	0.11	-
Leafy vegetables	-	**0.35**	-	-	0.20
Fruiting vegetables	-	**0.44**	-	0.17	0.18
Root vegetables	-0.12	0.20	-	-	-
Cruciferous vegetables	-	**0.30**	-	0.16	-
Other vegetables	0.12	**0.35**	-	0.17	0.16
Citrus fruit	-	**0.56**	-	-	-0.13
Other fruit	-	**0.70**	-	-	-
Soft drinks and fruit juice	0.16	0.15	-	-	-
Desserts	0.21	0.21	0.11	-	-
Sugar and candies	**0.25**	-	0.11	-	-
Butter and margarine	**0.25**	-	-	0.10	-
Specified seed oil	-	-	-	-	0.17
Unspecified seed oil	0.10	-0.12	-	-	**0.50**
Olive oil	-	**0.30**	0.15	**0.31**	-

[1]Correlation coefficients greater or equal to 0.25 in absolute value were shown in bold typeface, those smaller than 0.1 were not shown. PUFAs: polyunsaturated fatty acids.

Table 4 gives the ORs and corresponding CIs for esophageal cancer according to quartiles of factor scores for the five retained dietary patterns. Results refer to the composite model including all the five patterns simultaneously, and major confounding and risk variables as well. An increased risk of esophageal cancer was observed for the *Animal products and related components* pattern (OR=1.64, 95% CI: 1.06-2.55, for the highest versus the lowest quartile category of factor score, p for trend = 0.006). An inverse relationship was found for

the *Vitamins and fiber* (OR=0.50, 95% CI: 0.32-0.78, p for trend < 0.001) and the *Other polyunsasturated fatty acids* (OR=0.48, 95% CI: 0.31-0.74, p for trend < 0.001) patterns. No significant association was found for the *Starch-rich* (OR=0.80, 95% CI: 0.50-1.28, p for trend = 0.21) and the *Other fats* patterns (OR=1.04, 95% CI: 0.67-1.63, p for trend = 0.84). Consistent results were obtained from the five models including each pattern separately.

Table 4 – Odds ratios (ORs)[1] and corresponding 95% confidence intervals (CIs) for quartiles of factor scores among esophageal cancer cases and controls.

Dietary patterns	Quartile category, OR (95% CI)				
	I[2]	II	III	IV	*p*[3]
Animal products and related components	1[c]	0.82 (0.51-1.32)	1.01 (0.64-1.60)	1.64 (1.06-2.55)	0.006
Vitamins and fiber	1[c]	0.52 (0.35-0.78)	0.37 (0.23-0.57)	0.50 (0.32-0.78)	<0.001
Starch-rich	1[c]	1.17 (0.74-1.84)	1.03 (0.65-1.63)	0.80 (0.50-1.28)	0.21
Other PUFAs and vitamin D	1[c]	0.62 (0.41-0.95)	0.56 (0.36-0.86)	0.48 (0.31-0.74)	<0.001
Other fats	1[c]	1.01 (0.64-1.60)	1.01 (0.65-1.58)	1.04 (0.67-1.63)	0.84

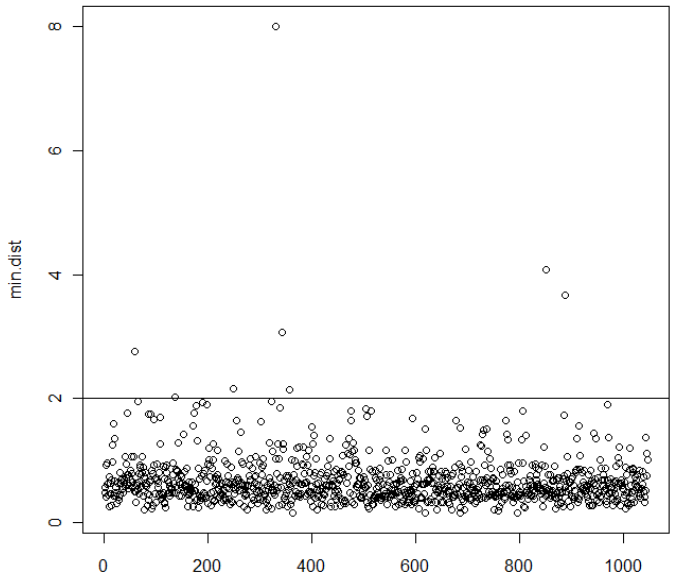
[1]Estimated from a multiple logistic regression model adjusted for age, sex, study center, education, alcohol drinking, tobacco smoking, and body mass index. Results refer to the composite model including all the five patterns simultaneously. [2]Reference category. [3]P-value for linear trend. PUFAs: polyunsaturated fatty acids.

Preliminary examination of potential outliers

Before performing the main cluster analysis, based on dietary patterns obtained through factor analysis, potential outliers were searched, through the strategies described before.

Figure 4 shows the plot of the minimum distance for each subjects. The examination of the plot allowed to identify 8 potential outliers.

Figure 4 – Minimum distances between each subject and all the remaining ones.



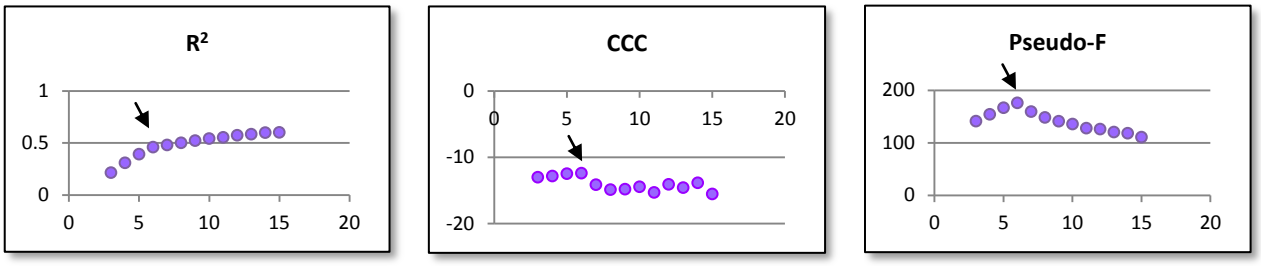
Results from the clustering performed with predefined number of clusters equal to 30 confirmed the role of potential outliers of these 8 observations.

Thus, all the analyses were carried out both including and excluding the 8 potential outliers (hereafter referred as "full dataset" and "reduced dataset", respectively).

Choice of the number of clusters

After running a series of cluster analyses with a predefined number of clusters varying from 3 to 15, results were compared on the basis of the R^2, the CCC, and the Pseudo-*F* statistic distributions. These measures were plotted against the corresponding number of clusters, to identify where the curve levels off (for R^2) or has a peak (for CCC and Pseudo-*F*) (**Figure 5**).

Figure 5 – Distributions of R^2, Cubic Clustering Criterion, and Pseudo-F, according to number of clusters



The three measures were consistent and indicated as a plausible solution that having a number of clusters equal to 6.

In the following, the 6 clusters obtained in the main analysis will be referred to as: C1, C2, C3, C4, C5, C6.

Comparison of results from the datasets including and excluding potential outliers

Table 5 shows the distribution of subjects, according to the 6 clusters identified from the full dataset and the reduced dataset (excluding 8 potential outliers). Clusters obtained from the reduced dataset were named R1, R2, R3, R4, R5 and R6.

Most subjects were allocated to the same cluster in the two situations, indicating a good agreement of the two clustering solutions. Out of the 8 potential outliers, 1 was allocated to the C3 cluster, 4 to the C4 cluster, 3 to the C5 cluster.

Table 5 - Distribution of subjects, according to the 6 clusters identified from the full dataset and the reduced dataset

		Full dataset						
	Cluster	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**Total**
Reduced dataset	**-**	0	0	1	4	3	0	8
	R1	221	0	2	1	0	0	224
	R2	0	165	0	0	0	2	167
	R3	1	0	303	0	0	10	314
	R4	0	1	7	88	0	3	99
	R5	0	1	1	0	54	2	58
	R6	0	0	0	1	0	176	177
	Total	222	167	314	94	57	193	1047

Table 6 provides an insight in the characteristics of the 8 potential outliers. Among these subjects 3 were esophageal cancer cases and 8 controls. As concerns nutritional characteristics, 1 subject had a low total energy intake (1488 kcal/day), while the remaining had very high energy intakes (more than 3000 kcal/day). All subjects had an extreme value on at least one dietary pattern, although the composition of the dietary pattern was quite different among the subjects.

Table 6 – Characteristics of the 8 potential outliers

Case/control status	Total energy	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
case	4750.63	5.10	0.24	-0.29	2.51	-0.48
case	5617.31	-0.07	5.16	2.04	3.58	0.62
case	7825.71	5.73	2.66	4.87	-3.07	-1.65
control	1488.22	-1.54	-1.24	-1.12	3.50	-1.23
control	3274.02	3.45	1.18	-2.97	0.29	-0.06
control	3942.34	-3.13	7.76	-1.06	-1.75	8.02
control	4770.24	2.01	-0.24	-0.14	-1.10	5.58
control	6200.98	1.48	1.91	1.60	-1.32	6.75

Since the exclusion of these potential outliers did not materially modify the clustering solution, in the following, results are provided referring to the full dataset. Results concerning the reduced dataset are shown in Appendix.

Clustering solution from K-means method with Euclidean distance

Figure 6 shows the dendrogram obtained from hierarchical clustering (Ward's method), which was preliminarily performed. The dendrogram was cut at level 6 and the centers of the 6 identified clusters were used as initial seeds to carry out cluster analysis with the *K*-means method and Euclidean distance.

Figure 6 – Dendrogram obtained from Ward's method

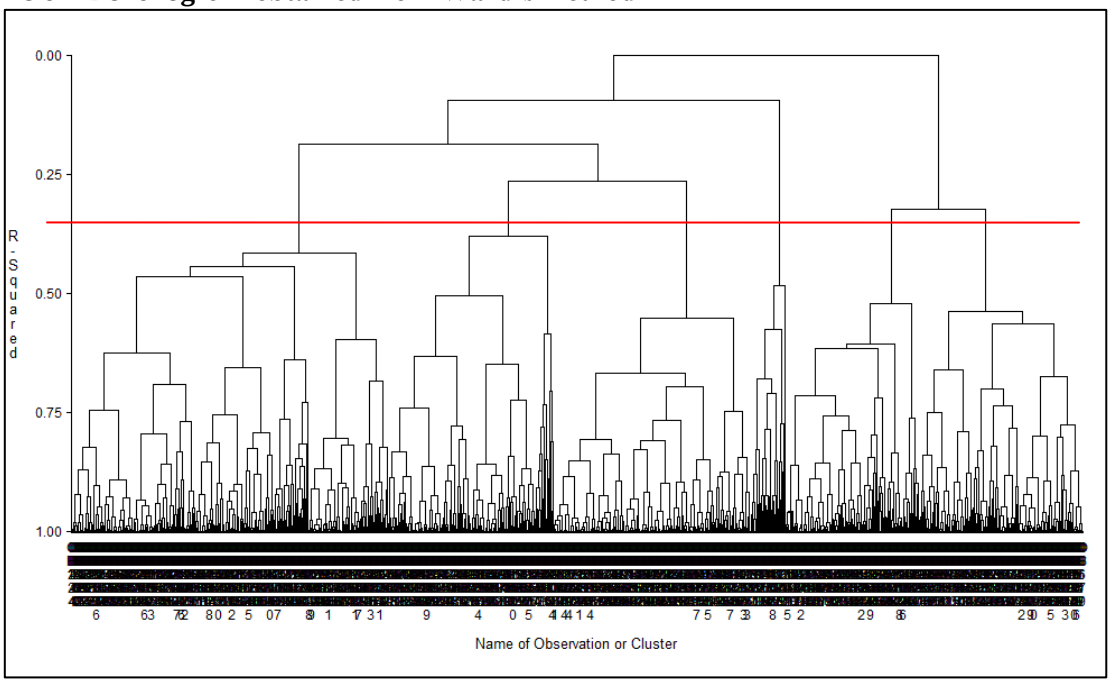


Table 7 shows a description of the identified clusters in terms of esophageal cancer cases and controls, according to the original dietary patterns.

The C3 cluster showed the highest number of subjects. The C5 cluster was small, including about 5% of the subjects. C3 and C4 had an higher percentage of cases as compared to that of controls, C1 had an higher percentage of controls, while the other groups had a similar composition, in terms of cases and controls. The C3 cluster had negative means on all the dietary patterns. All the remaining clusters were characterized by the highest mean on one of the dietary patterns in turn. The C1 cluster was characterized by an high contribution of the *Vitamins and fiber* pattern. The C2 cluster was high in the *Other polyunsaturated fatty acids and vitamin D* pattern. The C4 cluster was characterized by an high contribution of the *Animal products and related components*. The C5 cluster had an high mean of the *Other fats* pattern. The C6 cluster had an high mean of the *Starch-rich* pattern.

Table 7* – Description of the identified clusters in terms of cases and controls, and according to the original dietary patterns.

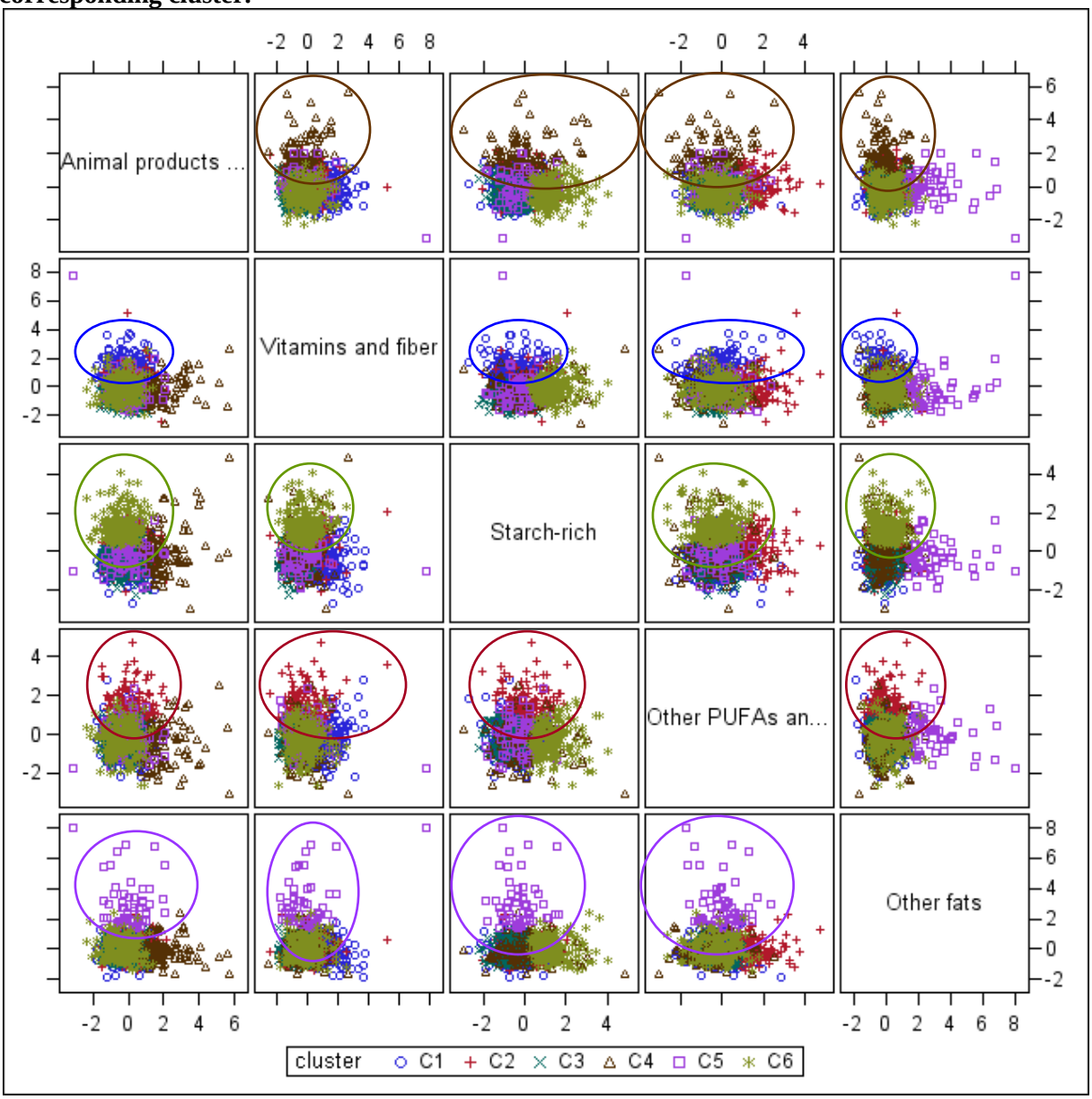
Cluster	Cases	Controls	All subjects	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
	N (%)	N (%)	N (%)	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)
C1	40 (13.16)	182 (24.50)	222 (21.20)	-0.16 (0.70)	1.13 (0.72)	-0.49 (0.68)	-0.26 (0.76)	-0.30 (0.61)
C2	42 (13.82)	125 (16.82)	167 (15.95)	0.17 (0.73)	-0.02 (0.88)	0.07 (0.73)	1.47 (0.76)	-0.04 (0.66)
C3	109 (35.86)	205 (27.59)	314 (29.99)	-0.37 (0.62)	-0.66 (0.51)	-0.49 (0.55)	-0.29 (0.61)	-0.23 (0.49)
C4	40 (13.16)	54 (7.27)	94 (8.98)	2.07 (1.03)	-0.19 (0.88)	0.02 (1.21)	-0.50 (1.04)	-0.13 (0.75)
C5	13 (4.28)	44 (5.92)	57 (5.44)	0.03 (0.95)	-0.09 (1.39)	-0.25 (0.79)	0.01 (0.91)	2.95 (1.52)
C6	60 (19.74)	133 (17.90)	193 (18.43)	-0.38 (0.70)	-0.09 (0.78)	1.36 (0.72)	-0.27 (0.79)	-0.05 (0.71)

*Results obtained from the full dataset.

Figure 7 shows the scatterplots of the dietary patterns, plotted with different symbols according to the 6 identified cluster.

Each dietary pattern contributed to the separation of one group from the others. The *Animal products and related components* pattern allowed to identify the C4 cluster. The *Vitamin and fiber* pattern was able to separate the C1 cluster. The *Starch-rich* pattern contributed to the identification of the C6 cluster. The *Other polyunsaturated fatty acids and vitamin D* pattern allowed to separate the C2 cluster. The *Other fats* pattern was able to identify the C5 cluster.

Figure 7* - Scatterplots of the dietary patterns, plotted with different symbols according to the corresponding cluster.



*Results obtained from the full dataset.

Table 8 shows the distribution of selected sociodemographic and lifestyle variables, for each cluster. The C1 cluster had the highest proportion of women as compared to other groups, most subjects were less educated, non drinkers or moderate drinkers, never or ex smokers, of normal weight, with a total energy intake lower than 2590 kcal. In the C2 cluster subjects were more likely to be males, with a low level of education, heavy drinkers, never or ex smokers, overweight or obese, with a total energy intake higher than 2590 kcal. In the C3

cluster most subjects were older than 60 years, men, less educated, heavy drinkers, never or ex smokers, of normal weight, with a total energy intake lower than 2590 kcal. Subjects in the C4 cluster were more likely to be men, with a low level of education, heavy drinkers, never or ex smokers, normal or overweight, with a total energy intake higher than 3200 kcal. In the C5 cluster most subjects were men, less educated, heavy drinkers, never or ex smokers, overweight or obese, with a total energy intake higher than 3200 kcal. In the C6 cluster subjects were more likely to be men, less educated, heavy drinkers, never or ex smokers, of normal weight, with a total energy intake between 2080 and 3200 kcal.

Table 8* – Distribution of sociodemographic and lifestyle variables, for each cluster

	C1 (N=222)	C2 (N=167)	C3 (N=314)	C4 (N=94)	C5 (N=57)	C6 (N=193)
Age						
<55	66 (29.73)	49 (29.34)	44 (14.01)	25 (26.60)	16 (28.07)	68 (35.23)
55-59	53 (23.87)	37 (22.16)	60 (19.11)	20 (21.28)	13 (22.81)	34 (17.62)
60-64	43 (19.37)	27 (16.17)	65 (20.70)	23 (24.47)	11 (19.30)	39 (20.21)
65-69	39 (17.57)	34 (20.36)	72 (22.93)	9 (9.57)	11 (19.30)	34 (17.62)
≥70	21 (9.46)	20 (11.98)	73 (23.25)	17 (18.09)	6 (10.53)	18 (9.33)
Sex						
Male	144 (64.86)	149 (89.22)	253 (80.57)	84 (89.36)	52 (91.23)	186 (96.37)
Female	78 (35.14)	18 (10.78)	61 (19.43)	10 (10.64)	5 (8.77)	7 (3.63)
Education						
<7	128 (57.66)	95 (56.89)	204 (64.97)	72 (76.60)	47 (82.46)	124 (64.25)
7-11	63 (28.38)	43 (25.75)	73 (23.25)	16 (17.02)	8 (14.04)	51 (26.42)
≥12	31 (13.96)	29 (17.37)	37 (11.78)	6 (6.38)	2 (3.51)	18 (9.33)
Alcohol drinking						
Non drinker	69 (31.08)	20 (11.98)	55 (17.52)	11 (11.70)	5 (8.77)	18 (9.33)
Drinker <4 drinks/day	84 (37.84)	50 (29.94)	97 (30.89)	19 (20.21)	16 (28.07)	46 (23.83)
Drinker ≥4 drinks/day	69 (31.08)	97 (58.08)	162 (51.59)	64 (68.09)	36 (63.16)	129 (66.84)
Tobacco smoking						
Never smoker	85 (38.29)	44 (26.35)	81 (25.80)	21 (22.34)	16 (28.07)	31 (16.06)
Ex smoker	77 (34.68)	56 (33.53)	133 (42.36)	29 (30.85)	18 (31.58)	83 (43.01)
Current smoker <15 cigarettes/day	26 (11.71)	24 (14.37)	34 (10.83)	19 (20.21)	6 (10.53)	16 (8.29)
Current smoker 15-24 cigarettes/day	27 (12.16)	30 (17.96)	48 (15.29)	17 (18.09)	11 (19.30)	43 (22.28)
Current smoker ≥25 cigarettes/day	7 (3.15)	13 (7.78)	18 (5.73)	8 (8.51)	6 (10.53)	20 (10.36)
Body mass index						
≤18.5	3 (1.35)	1 (0.60)	2 (0.64)	5 (5.32)	1 (1.75)	1 (0.52)
18.6-24.9	122 (54.95)	72 (43.11)	159 (50.64)	40 (42.55)	18 (31.58)	114 (59.07)
25-29.9	68 (30.63)	70 (41.92)	111 (35.35)	41 (43.62)	25 (43.86)	58 (30.05)
≥30	29 (13.06)	24 (14.37)	42 (13.38)	8 (8.51)	13 (22.81)	20 (10.36)
Total energy intake						
<2080	63 (28.38)	9 (5.39)	177 (56.37)	3 (3.19)	1 (1.75)	9 (4.66)
2080-2590	77 (34.68)	29 (17.37)	101 (32.17)	13 (13.83)	8 (14.04)	34 (17.62)
2591-3230	57 (25.68)	62 (37.13)	31 (9.87)	17 (18.09)	16 (28.07)	78 (40.41)
≥3231	25 (11.26)	67 (40.12)	5 (1.59)	61 (64.89)	32 (56.14)	72 (37.31)

*Results obtained from the full dataset.

Table 9 shows the mean daily intake of selected standardized nutrients and the mean daily total energy and non-alcoholic energy intakes, for each cluster.

There were no dominant nutrients for the C3 cluster, which had the lowest mean intakes for each nutrient. The C1 cluster had the highest mean intakes of vitamin C and total fibre. The C2 cluster had the highest mean intakes of other polyunsaturated fatty acids, vitamin D, niacin, and vitamin B6. The C4 cluster had the highest mean intakes of animal protein, cholesterol, saturated fatty acids, soluble carbohydrates, sodium, calcium, potassium, phosphorus, zinc, thiamin, riboflavin, total folate, and retinol. The C5 cluster had the highest mean intakes of monounsaturated fatty acids, linoleic acid, linolenic acid, iron, beta-carotene equivalents, and vitamin E. The C6 cluster had the highest mean intakes of vegetable protein, starch.

Table 9[*] – Description of the identified clusters: mean daily intake of selected standardized nutrients and the mean daily total energy and non-alcoholic energy intakes, for each cluster.

Nutrient	C1 N=222	C2 N=167	C3 N=314	C4 N=94	C5 N=57	C6 N=193
	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)
Animal protein	-0.26 (0.71)	0.82 (0.72)	-0.66 (0.60)	1.32 (1.15)	0.53 (0.91)	-0.14 (0.78)
Vegetable protein	-0.23 (0.71)	0.26 (0.79)	-0.79 (0.51)	0.48 (1.21)	0.32 (1.05)	0.99 (0.79)
Cholesterol	-0.25 (0.73)	0.71 (0.91)	-0.58 (0.56)	1.14 (1.34)	0.52 (1.04)	-0.10 (0.81)
Saturated fatty acids	-0.15 (0.77)	0.47 (0.84)	-0.66 (0.55)	1.49 (1.20)	0.63 (0.99)	-0.07 (0.73)
Monounsaturated fatty acids	-0.07 (0.83)	0.74 (1.06)	-0.66 (0.55)	0.46 (1.03)	0.84 (1.31)	0.04 (0.82)
Linoleic acid	-0.27 (0.62)	0.23 (0.74)	-0.49 (0.52)	0.14 (0.78)	2.61 (1.55)	0.07 (0.74)
Linolenic acid	-0.26 (0.52)	0.12 (0.60)	-0.50 (0.48)	0.33 (0.76)	2.74 (1.69)	0.04 (0.75)
Other polyunsaturated fatty acids	-0.22 (0.76)	1.30 (0.84)	-0.55 (0.61)	0.07 (1.13)	0.56 (0.91)	-0.17 (0.75)
Soluble carbohydrates	0.51 (0.83)	0.11 (0.88)	-0.69 (0.53)	0.79 (1.32)	0.28 (1.11)	-0.02 (0.95)
Starch	-0.42 (0.66)	0.18 (0.75)	-0.64 (0.51)	0.61 (1.39)	0.16 (0.94)	1.03 (0.84)
Sodium	-0.36 (0.70)	0.23 (0.75)	-0.65 (0.51)	1.16 (1.39)	0.21 (0.96)	0.64 (0.89)
Calcium	0.01 (0.71)	0.09 (0.77)	-0.54 (0.64)	1.88 (1.17)	0.23 (0.96)	-0.20 (0.69)
Potassium	0.18 (0.74)	0.53 (0.84)	-0.89 (0.56)	0.81 (1.17)	0.59 (1.03)	0.22 (0.82)
Phosphorus	-0.22 (0.69)	0.50 (0.72)	-0.75 (0.58)	1.54 (1.18)	0.53 (0.89)	0.13 (0.75)
Iron	-0.25 (0.67)	0.62 (0.82)	-0.66 (0.76)	0.67 (1.24)	0.69 (1.02)	0.29 (0.89)
Zinc	-0.28 (0.70)	0.66 (0.80)	-0.77 (0.57)	1.13 (1.15)	0.49 (0.97)	0.31 (0.82)
Thiamin (vitamin B1)	0.07 (0.69)	0.53 (0.87)	-0.86 (0.49)	0.96 (1.25)	0.42 (1.02)	0.27 (0.86)
Riboflavin (vitamin B2)	0.08 (0.74)	0.34 (0.76)	-0.68 (0.58)	1.53 (1.39)	0.33 (0.88)	-0.13 (0.74)
Vitamin B6	0.11 (0.73)	0.75 (0.84)	-0.89 (0.54)	0.71 (1.12)	0.57 (1.05)	0.16 (0.82)
Total folate	0.36 (0.82)	0.38 (0.89)	-0.88 (0.56)	0.72 (1.09)	0.42 (1.08)	0.22 (0.81)
Niacin	-0.09 (0.80)	1.04 (0.76)	-0.83 (0.54)	0.30 (1.05)	0.66 (0.94)	0.21 (0.80)
Vitamin C	0.94 (1.02)	0.23 (0.91)	-0.71 (0.49)	0.02 (0.83)	0.05 (0.86)	-0.14 (0.82)
Retinol	-0.11 (0.83)	0.47 (1.16)	-0.20 (0.87)	0.64 (1.39)	0.16 (0.94)	-0.31 (0.70)
Beta-carotene equivalents	0.51 (0.89)	0.23 (0.72)	-0.54 (0.44)	-0.05 (0.76)	0.56 (2.70)	-0.04 (0.67)
Lycopene	-0.22 (0.80)	0.43 (0.95)	-0.48 (0.70)	-0.24 (0.90)	0.25 (0.91)	0.71 (1.16)
Vitamin D	-0.15 (0.84)	1.24 (1.03)	-0.52 (0.63)	0.09 (0.86)	0.41 (0.95)	-0.22 (0.76)
Vitamin E	0.09 (0.74)	0.52 (0.93)	-0.73 (0.51)	0.04 (0.93)	1.86 (1.33)	0.06 (0.75)
Total fibre	0.63 (0.81)	0.23 (0.92)	-0.85 (0.56)	0.15 (1.04)	0.27 (1.22)	0.31 (0.82)

[*]Results obtained from the full dataset.

Table 10 gives the median weekly intake of selected food groups by clusters. The C3 cluster was characterized by the lowest median intakes of almost all the food groups. The C1 cluster had the highest intakes of white meat, leafy vegetables, cruciferous vegetables, other vegetables, citrus fruit, other fruit. The C2 cluster had the highest intakes of coffee, egg, white meat, red meat, processed meat, fish, potatoes, cruciferous vegetables, soft drinks and fruit juice, and olive oil. The C4 cluster had the highest intakes of milk, egg, processed meat, cheese, soft drinks and fruit juice, desserts, and butter and margarine. The C5 cluster had the highest intakes of soup, egg, white meat, processed meat, potatoes, leafy vegetables, fruiting vegetables, soft drinks and fruit juices, sugar and candies, and seed oils. The C6 cluster had the highest intakes of bread, and pasta and rice. The table also shows the mean daily total energy, alcohol and nonalcoholic energy intakes. The C3 cluster was characterized by the lowest daily energy intake, either total and nonalcoholic. On the other hand, the C5 cluster had the highest total and nonalcoholic energy intakes.

Table 10[*] – Description of the identified clusters: median weekly intake of selected food groups by cluster.

Food groups	C1 N=222	C2 N=167	C3 N=314	C4 N=94	C5 N=57	C6 N=193
	median	median	median	median	median	median
Milk	7.00	5.00	3.00	14.00	7.00	1.50
Coffee	14.00	18.50	14.00	14.00	14.50	14.00
Bread	15.25	21.50	15.00	24.50	21.25	35.00
Pasta and rice	4.75	5.50	4.25	4.75	4.75	6.00
Soup	2.00	2.00	1.50	2.13	2.50	2.25
Egg	1.00	2.00	1.00	2.00	2.00	1.00
White meat	2.00	2.00	1.00	1.00	2.00	1.00
Red meat	3.50	5.75	3.75	4.88	5.00	4.75
Processed meat	2.00	3.00	2.00	3.00	3.00	2.50
Fish	1.50	2.00	1.50	1.00	1.50	1.00
Cheese	4.18	4.38	3.67	8.77	5.12	4.05
Potatoes	1.00	2.00	1.00	1.50	2.00	1.00
Pulses	1.50	1.50	1.00	1.50	1.50	1.50
Leafy vegetables	7.00	5.50	3.50	4.75	7.00	5.25
Fruiting vegetables	3.50	3.00	1.50	2.08	3.75	2.50
Cruciferous vegetables	0.50	0.50	0.25	0.25	0.42	0.25
Other vegetables	2.17	1.83	0.67	0.83	1.17	1.00
Citrus fruit	5.00	3.50	1.50	3.50	2.50	2.00
Other fruit	18.67	11.50	8.40	11.35	11.67	10.83
Soft drinks and fruit juices	0.00	0.50	0.00	0.50	0.50	0.00
Desserts	3.00	2.92	1.50	5.00	3.50	2.25
Sugar and candies	28.00	35.50	25.00	43.50	45.50	40.00
Butter and margarine	1.88	2.27	1.46	5.68	2.88	1.75
Seed oils	0.66	2.11	2.07	2.22	53.26	4.03
Olive oil	25.96	39.37	14.45	19.71	1.89	25.13
	mean[+]	mean[+]	mean[+]	mean[+]	mean[+]	mean[+]
Total energy	2436.50	3113.47	2015.48	3633.10	3485.51	3113.71
Alcohol (g)	28.55	54.35	46.35	73.45	70.07	58.84
Non alcoholic energy	2236.66	2733.05	1691.05	3118.97	2995.03	2701.83

[*]Results obtained from the full dataset. [+] Daily intake.

Risk estimates

Table 11 gives the ORs and corresponding FCI for the six identified clusters. The C3 cluster was chosen as reference category, since it had the highest number of subjects and it was not characterized by high values on any dietary patterns. After accounting for major confounding variables, significant decreased esophageal cancer risk were observed for the C1 cluster (characterized by the highest factor scores of the *Vitamins and fiber* pattern, OR=0.59, 0.95% FCI: 0.40-0.88), the C5 cluster (characterized by the highest values of the *Other fats* pattern, OR=0.42, 95% FCI: 0.20-0.86), and the C6 cluster (characterized by the highest factor scores

of the *Starch-rich* pattern, OR=0.60, 95% FCI=0.42-0.86). Non significant estimates were observed for the C2 and C4 clusters (characterized by high factor scores of the *Other polyunsaturated fatty acids and vitamin D* and *Animal products and related components* patterns, OR=0.76, 95% FCI: 0.51-1.13, and OR=1.29, 95% FCI: 0.80-2.07, respectively).

Table 11* - Odds ratios (OR) of esophageal cancer and corresponding 95% floating confidence intervals (FCI), by cluster

Cluster	Cases	Controls	Total	crude OR (95% FCI)	adjusted OR (95% FCI)[1]
C1	40	182	222	0.41 (0.29-0.58)	0.59 (0.40-0.88)
C2	42	125	167	0.63 (0.45-0.90)	0.76 (0.51-1.13)
C3	109	205	314	1.00 (0.79-1.26)	1.00 (0.75-1.33)
C4	40	54	94	1.39 (0.93-2.10)	1.29 (0.80-2.07)
C5	13	44	57	0.56 (0.30-1.03)	0.42 (0.20-0.86)
C6	60	133	193	0.85 (0.63-1.15)	0.60 (0.42-0.86)

[1]Adjusted for age, sex, center, education, alcohol drinking, tobacco smoking, body mass index.
*Results obtained from the full dataset.

Comparison with other clustering solutions

The following tables (**Table 12 – Table 17**) show the comparisons of pairs of clustering solutions based on *k*-means method and different distances (Euclidean, Manhattan, Lagrange and Correlation coefficient similarity measures). The identified clusters were named M1-M6 (Manhattan distance), L1-L6 (Lagrange distance), CC1-CC6 (Correlation coefficient similarity measure).

A good agreement was observed between the solutions based on Euclidean, Manhattan, Lagrange distances, the *k* statistics ranging from 0.65 to 0.83. The solution based on the Correlation coefficient similarity measure was only partially in agreement with the previous ones, with *k* statistics of 0.40 to 0.46.

Thus, further analyses were carried out on the subset of 730 subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distances. Results are shown in Appendix.

Table 12* - Agreement between clustering solution obtained through *k*-means method and Euclidean and Manhattan distances.

Euclidean distance	Manhattan distance M1	M2	M3	M4	M5	M6	Total
C1	190	1	2	9	7	13	222
	(85.59)	(0.52)	(0.88)	(6.72)	(7.87)	(7.22)	21.20
C2	9	142	2	1	7	6	167
	(4.05)	(73.20)	(0.88)	(0.75)	(7.87)	(3.33)	15.95
C3	17	23	217	38	10	9	314
	(7.66)	(11.86)	(95.18)	(28.36)	(11.24)	(5.00)	29.99
C4	4	8	0	78	4	0	94
	(1.80)	(4.12)	(0.00)	(58.21)	(4.49)	(0.00)	8.98
C5	0	0	0	1	56	0	57
	(0.00)	(0.00)	(0.00)	(0.75)	(62.92)	(0.00)	5.44
C6	2	20	7	7	5	152	193
	(0.90)	(10.31)	(3.07)	(5.22)	(5.62)	(84.44)	18.43
Total	222	194	228	134	89	180	1047
	21.20	18.53	21.78	12.80	8.50	17.19	100.00

*Results obtained from the full dataset. *k*=0.75

Table 13* - Agreement between clustering solution obtained through *k*-means method and Euclidean and Lagrange distances.

Euclidean distance	Lagrange distance L1	L2	L3	L4	L5	L6	Total
C1	169	12	24	8	0	9	222
	(93.89)	(6.56)	(7.52)	(8.25)	(0.00)	(4.33)	21.20
C2	6	143	13	1	2	2	167
	(3.33)	(78.14)	(4.08)	(1.03)	(3.33)	(0.96)	15.95
C3	0	10	278	1	1	24	314
	(0.00)	(5.46)	(87.15)	(1.03)	(1.67)	(11.54)	29.99
C4	1	2	2	86	0	3	94
	(0.56)	(1.09)	(0.63)	(88.66)	(0.00)	(1.44)	8.98
C5	0	1	0	0	56	0	57
	(0.00)	(0.55)	(0.00)	(0.00)	(93.33)	(0.00)	5.44
C6	4	15	2	1	1	170	193
	(2.22)	(8.20)	(0.63)	(1.03)	(1.67)	(81.73)	18.43
Total	180	183	319	97	60	208	1047
	17.19	17.48	30.47	9.26	5.73	19.87	100.00

*Results obtained from the full dataset. *k*=0.83

Table 14* - Agreement between clustering solution obtained through *k*-means method and Euclidean and Correlation coefficient similarity measure distances.

	Correlation coefficient similarity measure						
Euclidean distance	CC1	CC2	CC3	CC4	CC5	CC6	Total
C1	170	0	35	9	8	0	222
	(79.07)	(0.00)	(22.44)	(5.06)	(5.37)	(0.00)	21.20
C2	15	113	0	7	4	28	167
	(6.98)	(60.11)	(0.00)	(3.93)	(2.68)	(17.39)	15.95
C3	27	75	33	68	73	38	314
	(12.56)	(39.89)	(21.15)	(38.20)	(48.99)	(23.60)	29.99
C4	2	0	3	88	0	1	94
	(0.93)	(0.00)	(1.92)	(49.44)	(0.00)	(0.62)	8.98
C5	1	0	0	0	56	0	57
	(0.47)	(0.00)	(0.00)	(0.00)	(37.58)	(0.00)	5.44
C6	0	0	85	6	8	94	193
	(0.00)	(0.00)	(54.49)	(3.37)	(5.37)	(58.39)	18.43
Total	215	188	156	178	149	161	1047
	20.53	17.96	14.90	17.00	14.23	15.38	100.00

*Results obtained from the full dataset. k=0.43

Table 15* - Agreement between clustering solution obtained through *k*-means method and Manhattan and Lagrange distances.

	Lagrange distance						
Manhattan distance	L1	L2	L3	L4	L5	L6	Total
M1	152	17	38	8	0	7	222
	(84.44)	(9.29)	(11.91)	(8.25)	(0.00)	(3.37)	21.20
M2	5	129	31	9	1	19	194
	(2.78)	(70.49)	(9.72)	(9.28)	(1.67)	(9.13)	18.53
M3	2	6	197	0	0	23	228
	(1.11)	(3.28)	(61.76)	(0.00)	(0.00)	(11.06)	21.78
M4	6	3	37	76	1	11	134
	(3.33)	(1.64)	(11.60)	(78.35)	(1.67)	(5.29)	12.80
M5	1	12	8	4	57	7	89
	(0.56)	(6.56)	(2.51)	(4.12)	(95.00)	(3.37)	8.50
M6	14	16	8	0	1	141	180
	(7.78)	(8.74)	(2.51)	(0.00)	(1.67)	(67.79)	17.19
Total	180	183	319	97	60	208	1047
	17.19	17.48	30.47	9.26	5.73	19.87	100.00

*Results obtained from the full dataset. k=0.65

Table 16* - Agreement between clustering solution obtained through *k*-means method and Manhattan and Correlation coefficient similarity measure distances.

	Correlation coefficient similarity measure						
Manhattan distance	CC1	CC1	CC1	CC1	CC1	CC1	Total
M1	175	9	23	11	4	0	222
	(81.40)	(4.79)	(14.74)	(6.18)	(2.68)	(0.00)	21.20
M2	10	117	0	16	2	49	194
	(4.65)	(62.23)	(0.00)	(8.99)	(1.34)	(30.43)	18.53
M3	19	57	25	32	61	34	228
	(8.84)	(30.32)	(16.03)	(17.98)	(40.94)	(21.12)	21.78
M4	4	1	11	114	2	2	134
	(1.86)	(0.53)	(7.05)	(64.04)	(1.34)	(1.24)	12.80
M5	3	3	1	4	76	2	89
	(1.40)	(1.60)	(0.64)	(2.25)	(51.01)	(1.24)	8.50
M6	4	1	96	1	4	74	180
	(1.86)	(0.53)	(61.54)	(0.56)	(2.68)	(45.96)	17.19
Total	215	188	156	178	149	161	1047
	20.53	17.96	14.90	17.00	14.23	15.38	100.00

*Results obtained from the full dataset. *k*=0.46

Table 17* - Agreement between clustering solution obtained through *k*-means method and Lagrange and Correlation coefficient similarity measure distances.

	Correlation coefficient similarity measure						
Lagrange distance	CC1	CC1	CC1	CC1	CC1	CC1	Total
L1	148	1	27	2	2	0	180
	(68.84)	(0.53)	(17.31)	(1.12)	(1.34)	(0.00)	17.19
L2	19	110	5	7	9	33	183
	(8.84)	(58.51)	(3.21)	(3.93)	(6.04)	(20.50)	17.48
L3	42	76	24	75	70	32	319
	(19.53)	(40.43)	(15.38)	(42.13)	(46.98)	(19.88)	30.47
L4	5	0	4	88	0	0	97
	(2.33)	(0.00)	(2.56)	(49.44)	(0.00)	(0.00)	9.26
L5	1	1	1	0	56	1	60
	(0.47)	(0.53)	(0.64)	(0.00)	(37.58)	(0.62)	5.73
L6	0	0	95	6	12	95	208
	(0.00)	(0.00)	(60.90)	(3.37)	(8.05)	(59.01)	19.87
Total	215	188	156	178	149	161	1047
	20.53	17.96	14.90	17.00	14.23	15.38	100.00

*Results obtained from the full dataset. *k*=0.40

Table 18 shows the distribution of the 5 dietary patterns identified through factor analysis, within clusters, according to the three solutions based on Euclidean, Manhattan and Lagrange distances. Clusters based on Manhattan distance, tended to have lower median and means on the factor that characterized each cluster, as compared to corresponding clusters based on the Euclidean and Lagrange distances.

Table 18* – Distribution of the factors scores within clusters, according to the solutions based on Euclidean, Manhattan and Lagrange distances.

		Euclidean			Manhattan			Lagrange		
		Range	median	mean (std)	range	median	mean (std)	range	median	mean (std)
C1	Animal products and related components	-1.81 - 1.52	-0.16	-0.16 (0.70)	-1.81 - 3.45	-0.13	-0.12 (0.74)	-1.81 - 1.52	-0.20	-0.22 (0.67)
	Vitamins and fiber	0.17 - 3.77	0.92	1.13 (0.72)	-0.18 - 3.77	0.84	1.00 (0.72)	0.16 - 5.16	1.02	1.27 (0.80)
	Starch-rich	-2.74 - 1.58	-0.55	-0.49 (0.68)	-2.97 - 0.96	-0.66	-0.64 (0.62)	-2.14 - 2.04	-0.49	-0.39 (0.74)
	Other PUFAs and vitamin D	-2.22 - 2.81	-0.26	-0.26 (0.76)	-2.22 - 3.29	-0.12	-0.11 (0.84)	-2.22 - 3.58	-0.22	-0.20 (0.75)
	Other fats	-1.93 - 1.71	-0.39	-0.30 (0.61)	-1.93 - 1.04	-0.44	-0.37 (0.53)	-1.93 - 1.19	-0.38	-0.33 (0.58)
C2	Animal products and related components	-1.54 - 2.20	0.17	0.17 (0.73)	-1.83 - 5.10	0.22	0.30 (0.89)	-1.54 - 2.68	0.05	0.12 (0.75)
	Vitamins and fiber	-2.50 - 5.16	-0.05	-0.02 (0.88)	-2.50 - 5.16	-0.29	-0.19 (0.87)	-1.67 - 2.59	0.02	0.05 (0.82)
	Starch-rich	-2.07 - 2.04	0.07	0.07 (0.73)	-1.77 - 3.50	0.15	0.20 (0.81)	-2.07 - 2.58	0.03	0.10 (0.80)
	Other PUFAs and vitamin D	0.21 - 4.75	1.35	1.47 (0.76)	0.08 - 4.75	1.15	1.25 (0.72)	-0.59 - 4.75	1.26	1.35 (0.80)
	Other fats	-1.38 - 2.26	-0.18	-0.04 (0.66)	-1.62 - 1.53	-0.16	-0.09 (0.60)	-1.38 - 2.26	-0.20	-0.05 (0.68)
C3	Animal products and related components	-1.68 - 1.12	-0.42	-0.37 (0.62)	-1.68 - 0.67	-0.66	-0.62 (0.47)	-1.74 - 1.91	-0.38	-0.32 (0.65)
	Vitamins and fiber	-1.92 - 0.35	-0.63	-0.66 (0.51)	-1.92 - 0.42	-0.69	-0.72 (0.47)	-2.50 - 1.66	-0.54	-0.56 (0.60)
	Starch-rich	-2.22 - 0.69	-0.50	-0.49 (0.55)	-1.88 - 0.79	-0.51	-0.46 (0.53)	-2.74 - 0.86	-0.55	-0.56 (0.57)
	Other PUFAs and vitamin D	-1.77 - 1.06	-0.32	-0.29 (0.61)	-1.65 - 3.5	-0.38	-0.33 (0.64)	-1.69 - 2.11	-0.31	-0.25 (0.63)
	Other fats	-1.16 - 1.37	-0.34	-0.23 (0.49)	-1.24 - 1.22	-0.36	-0.27 (0.46)	-1.16 - 1.35	-0.32	-0.23 (0.48)
C4	Animal products and related components	0.71 - 5.73	1.82	2.07 (1.03)	0.12 - 5.73	1.22	1.43 (1.02)	0.65 - 5.73	1.69	2.00 (1.04)
	Vitamins and fiber	-2.57 - 2.66	-0.26	-0.19 (0.88)	-2.57 - 2.66	-0.41	-0.34 (0.91)	-1.91 - 2.66	-0.13	-0.12 (0.89)
	Starch-rich	-2.97 - 4.87	-0.16	0.02 (1.21)	-2.22 - 4.87	-0.15	-0.07 (1.11)	-2.97 - 4.87	-0.25	-0.12 (1.13)
	Other PUFAs and vitamin D	-3.07 - 2.51	-0.36	-0.50 (1.04)	-3.07 - 0.76	-0.71	-0.75 (0.77)	-3.07 - 2.51	-0.40	-0.53 (1.08)
	Other fats	-1.65 - 2.39	-0.22	-0.13 (0.75)	-1.65 - 2.9	-0.21	-0.13 (0.67)	-1.65 - 2.39	-0.23	-0.14 (0.76)
C5	Animal products and related components	-3.13 - 2.04	-0.01	0.03 (0.95)	-3.13 - 3.11	-0.06	0.05 (1.00)	-3.13 - 2.04	-0.03	0.01 (0.96)
	Vitamins and fiber	-1.77 - 7.76	-0.24	-0.09 (1.39)	-1.77 - 7.76	-0.10	-0.03 (1.20)	-1.77 - 7.76	-0.25	-0.09 (1.36)
	Starch-rich	-1.91 - 1.63	-0.22	-0.25 (0.79)	-1.91 - 3.59	-0.21	-0.18 (0.84)	-1.91 - 1.63	-0.22	-0.19 (0.79)
	Other PUFAs and vitamin D	-1.75 - 2.38	0.07	0.01 (0.91)	-1.75 - 3.18	0.07	0.12 (0.91)	-1.75 - 2.38	0.07	0.04 (0.93)
	Other fats	1.34 - 8.02	2.38	2.95 (1.52)	0.5 - 8.02	1.89	2.32 (1.50)	1.37 - 8.02	2.36	2.88 (1.51)
C6	Animal products and related components	-2.35 - 1.39	-0.39	-0.38 (0.70)	-2.35 - 1.39	-0.50	-0.48 (0.63)	-2.35 - 2.09	-0.37	-0.37 (0.72)
	Vitamins and fiber	-1.91 - 2.50	-0.08	-0.09 (0.78)	-1.77 - 2.56	0.05	0.15 (0.86)	-2.57 - 1.46	-0.18	-0.19 (0.72)
	Starch-rich	0.29 - 4.05	1.21	1.36 (0.72)	0.04 - 4.05	1.19	1.30 (0.70)	-0.24 - 4.05	1.03	1.22 (0.80)
	Other PUFAs and vitamin D	-2.67 - 1.80	-0.26	-0.27 (0.79)	-2.67 - 2.33	-0.30	-0.30 (0.81)	-2.67 - 1.80	-0.36	-0.40 (0.77)
	Other fats	-1.37 - 2.40	-0.18	-0.05 (0.71)	-1.73 - 2.4	-0.25	-0.15 (0.66)	-1.37 - 2.40	-0.24	-0.09 (0.70)

*Results obtained from the full dataset.

Table 19 gives the distribution of the 5 dietary patterns identified through factor analysis, within clusters, according to the solution based on the Correlation coefficient similarity measure. In this solution, none of the clusters is characterized by the lowest means on all the dietary patterns, as was the C3 cluster in the main solution. Moreover, two clusters are characterized by high means on two dietary patterns. The CC1 cluster was characterized by the highest mean of the *Vitamins and fiber* pattern. The CC2 cluster had the highest mean of the *Other PUFAs and vitamin D* pattern. The CC3 cluster was characterized by the highest mean of the *Starch-rich* pattern and an high mean of the *Vitamins and fiber* one. The CC4 cluster had the highest mean of the *Animal products and related components*. The CC5 cluster had the highest mean on the *Other fats* pattern. The CC6 cluster was characterized by the highest mean of the *Starch-rich* pattern and an high mean of the *Other PUFAs and vitamin D* one.

Table 19* – Distribution of the factors scores within clusters, according to the solutions based on the Correlation Coefficient Similarity Measure.

		Correlation Coefficient Similarity Measure		
		Range	median	mean (std)
CC1	Animal products and related components	-1.81 - 1.99	-0.15	-0.14 (0.71)
	Vitamins and fiber	-0.54 - 5.16	0.86	1.05 (0.86)
(35 cases	Starch-rich	-2.74 - 2.04	-0.67	-0.65 (0.63)
180 controls)	Other PUFAs and vitamin D	-2.22 - 3.58	-0.09	-0.03 (0.83)
	Other fats	-1.93 - 1.82	-0.39	-0.34 (0.56)
CC2	Animal products and related components	-1.68 - 2.2	-0.13	-0.06 (0.73)
	Vitamins and fiber	-2.5 - 1.08	-0.47	-0.47 (0.66)
(55 cases	Starch-rich	-2.07 - 1.17	-0.39	-0.39 (0.60)
133 controls)	Other PUFAs and vitamin D	-0.54 - 4.75	1.01	1.12 (0.87)
	Other fats	-1.38 - 2.26	-0.32	-0.19 (0.60)
CC3	Animal products and related components	-1.86 - 5.73	-0.32	-0.27 (0.84)
	Vitamins and fiber	-1.29 - 2.66	0.31	0.42 (0.80)
(46 cases	Starch-rich	-0.70 - 4.87	0.78	0.94 (0.95)
110 controls)	Other PUFAs and vitamin D	-3.07 - 0.39	-0.81	-0.85 (0.62)
	Other fats	-1.73 - 2.40	-0.37	-0.20 (0.69)
CC4	Animal products and related components	-0.38 - 5.56	1.00	1.24 (1.09)
	Vitamins and fiber	-1.91 - 1.56	-0.54	-0.46 (0.75)
(85 cases	Starch-rich	-2.97 - 3.07	-0.33	-0.27 (0.88)
93 controls)	Other PUFAs and vitamin D	-2.65 - 2.51	-0.51	-0.49 (0.83)
	Other fats	-1.62 - 2.39	-0.28	-0.17 (0.59)
CC5	Animal products and related components	-3.13 - 2.04	-0.51	-0.40 (0.78)
	Vitamins and fiber	-1.77 - 7.76	-0.43	-0.31 (0.99)
(44 cases	Starch-rich	-1.91 - 1.63	-0.40	-0.36 (0.65)
105 controls)	Other PUFAs and vitamin D	-1.75 - 2.38	-0.34	-0.27 (0.72)
	Other fats	-0.63 - 8.02	0.99	1.40 (1.59)
CC6	Animal products and related components	-2.35 - 2.09	-0.57	-0.48 (0.68)
	Vitamins and fiber	-2.57 - 1.35	-0.52	-0.48 (0.65)
(39 cases	Starch-rich	-0.44 - 3.59	0.98	1.05 (0.74)
122 controls)	Other PUFAs and vitamin D	-1.26 - 2.08	0.26	0.35 (0.69)
	Other fats	-1.37 - 2.08	-0.33	-0.25 (0.58)

*Results obtained from the full dataset.

A good agreement was also observed with the clustering solutions obtained through the Partitioning Around Medoids method.

Discussion

The present work is based on the subsequent application of factor and cluster analyses to data from a case-control study on esophageal cancer.

PCFA allowed to identify five major dietary patterns, which explained about 80% of the total variance in the original nutrients. The *Animal products and related components* pattern was positively related to esophageal cancer risk. The *Vitamins and fiber* and the *Other polyunsaturated fatty acids and vitamin D* were inversely related to esophageal cancer, while no relationship with this cancer was observed for the *Starch-rich* and the *Other fats* patterns. The naming of the factors, based on high factor scores characterizing each pattern, was confirmed by the distributions of selected nutrients and food groups.

The subsequent cluster analysis, based on differences in the dietary patterns, yielded 6 clusters, one of which was characterized by the lowest intakes of all nutrients and food groups considered, while the remaining clusters were determined by an extreme value of the dietary patterns, one-by-one. Subjects in the C1 cluster were characterized by the highest values of the *Vitamins and fiber* pattern and had the highest intakes of white meat, vegetables, and fruit. Subjects in the C2 cluster had the highest values of the *Other polyunsaturated fatty acids* pattern and were more likely to consume high intakes of coffee, egg, meat, fish, potatoes, cruciferous vegetables, soft drinks and fruit juice, and olive oil. The C4 cluster was characterized by the highest scores of the *Animal products and related* components, and subjects in this cluster had high intakes of milk, egg, processed meat, cheese, soft drinks and fruit juice, desserts, and butter and margarine. Subjects in the C5 cluster had the highest values of the *Other fats* pattern and were more likely to have high intakes of soup, egg, white meat, processed meat, potatoes, leafy vegetables, fruiting vegetables, soft drinks and fruit

juices, sugar and candies, and seed oils. The C6 cluster was characterized by the highest scores of the *Starch-rich* pattern and had the highest intakes of bread, and pasta and rice. The C3 cluster included the highest number of subjects, was characterized by moderate intakes on each pattern and subjects in this cluster had the lowest daily energy intake. Significant inverse relations were observed between the C1, C5 and C6 clusters – which were characterized by high values of the *Vitamins and fiber*, *Other fats*, and *Starch-rich* patterns, respectively – as compared to the C3 cluster. No significant risk was observed for the C2, and C4 clusters.

Factor analysis is a method that allows to estimate cancer risk more comprehensively than others based on single foods or nutrients, as it accounts for the complex forms of interaction existing among dietary components. Limitations of factor analysis arise from the subjective decisions involved in the definition of dietary patterns, including the set of variables included in the analysis, the number of retained patterns, the type of rotation, and the naming of the patterns. In this application, various alternative options were tried to check robustness and solution stability. Among these complementary analyses, results from PCFA were compared with those from another factor analysis method (i.e., principal axis factoring), and those from PCFA analyses performed separately in strata of center and gender, and in randomly generated split samples. Moreover, the internal consistency of the identified patterns was evaluated using the Cronbach's coefficient alphas. All these checks supported the decisions adopted in the main analyses.

The main characteristic of cluster analysis is its ability to identify mutually exclusive subgroups within a population, with similarities on given variables. Its use in nutritional epidemiology, allows to identify groups of patients with specific dietary behaviors. The application of cluster analysis on dietary patterns derived through factor analysis represent an interesting – although rarely used – statistical strategy for data reduction an clustering. Insight

into dietary behaviors of different clusters within a population can help to tailor dietary recommendations and health promotion interventions.

Cluster analysis has, however, some limitations, that arise from the subjective decisions required at various steps of the analysis, including the choice of the initial variables, method and distance measure, identification of the optimal number of clusters. In this application, to limit the influence of the starting point, the initial seeds used in the *k*-means method were obtained performing a hierarchical clustering (Ward's method) and cutting the corresponding dendrogram at the level k=6. Moreover, some alternative solutions were identified through different methods and distances, yielding comparable clustering solutions. Another limitation of cluster analysis is its sensitivity to the presence of outliers; however, the exclusion of 8 potential outliers did not materially change the results.

REFERENCES

Bravi F, Edefonti V, Randi G, Garavello W, La Vecchia C, Ferraroni M, Talamini R, Franceschi S, Decarli A Dietary patterns and the risk of esophageal cancer. *Ann Oncol*.

Breslow NE, Day NE (1980) *Statistical methods in cancer research. Vol. I. The analysis of case-control studies. IARC Sci Publ No. 32*. Lyon, France: IARC.

Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. *Communications in statistics* **3:** 1-27.

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* **16:** 297-334.

Decarli A, Franceschi S, Ferraroni M, Gnagnarella P, Parpinel MT, La Vecchia C, Negri E, Salvini S, Falcini F, Giacosa A (1996) Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients. *Ann Epidemiol* **6:** 110-8.

DiBello JR, Kraft P, McGarvey ST, Goldberg R, Campos H, Baylin A (2008) Comparison of 3 methods for identifying dietary patterns associated with risk of disease. *Am J Epidemiol* **168:** 1433-43.

Easton DF, Peto J, Babiker AG (1991) Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Stat Med* **10:** 1025-35.

Everitt B (1979) Unresolved problems in cluster analysis. *Biometrics* **35:** 169-181.

Franceschi S, Negri E, Salvini S, Decarli A, Ferraroni M, Filiberti R, Giacosa A, Talamini R, Nanni O, Panarello G, et al. (1993) Reproducibility of an Italian food frequency questionnaire for cancer studies: results for specific food items. *Eur J Cancer* **29A:** 2298-305.

Gnagnarella P, Parpinel M, Salvini S, Franceschi S, Palli D, Boyle P (2004) The update of the Italian food composition database. *J Food Comp Analysis* **17:** 509-522.

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

Johnson R, Wichern D (2002) *Applied multivariate statistical analysis*, 5th edn. Upper Saddler River, NJ: Prentice Hall.

Kaufman L, Rousseeuw PJ (1990) *Findings groups in data: an introduction to cluster analysis*: John Wiley & Sons.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* **33:** 159-74.

Meyer J, Doring A, Herder C, Roden M, Koenig W, Thorand B (2011) Dietary patterns, subclinical inflammation, incident coronary heart disease and mortality in middle-aged men from the MONICA/KORA Augsburg cohort study. *Eur J Clin Nutr* **65:** 800-7.

Milligan G, Cooper M (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50:** 159-179.

Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, Krebs-Smith SM, Guenther PM (2007) Dietary patterns: challenges and opportunities in dietary patterns research an Experimental Biology workshop, April 1, 2006. *J Am Diet Assoc* **107:** 1233-9.

Newby PK, Muller D, Tucker KL (2004) Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *Am J Clin Nutr* **80:** 759-67.

Pett MA, Lackey NR, Sullivan JJ (2003) *Making sense of factor analysis: the use of factor analysis for instrument development in health care research*: CA: Sage.

Plummer M (2004) Improved estimates of floating absolute risk. *Stat Med* **23:** 93-104.

Salvini S, Parpinel M, Gnagnarella P, Maisonneuve P, Turrini A (1998) *Banca dati di composizione degli alimenti per studi epidemiologici in Italia*. Milano, Italia: Istituto Europeo di Oncologia.

Sarle W (1983) Cubic Clustering Criterion. *SAS Technical Report A-108, Cary, NC: SAS Institute Inc.*

Venables W, Ripley B (2002) *Modern applied statistics with S*, 4th edn. New York, NY: Springer Verlag.

APPENDIX

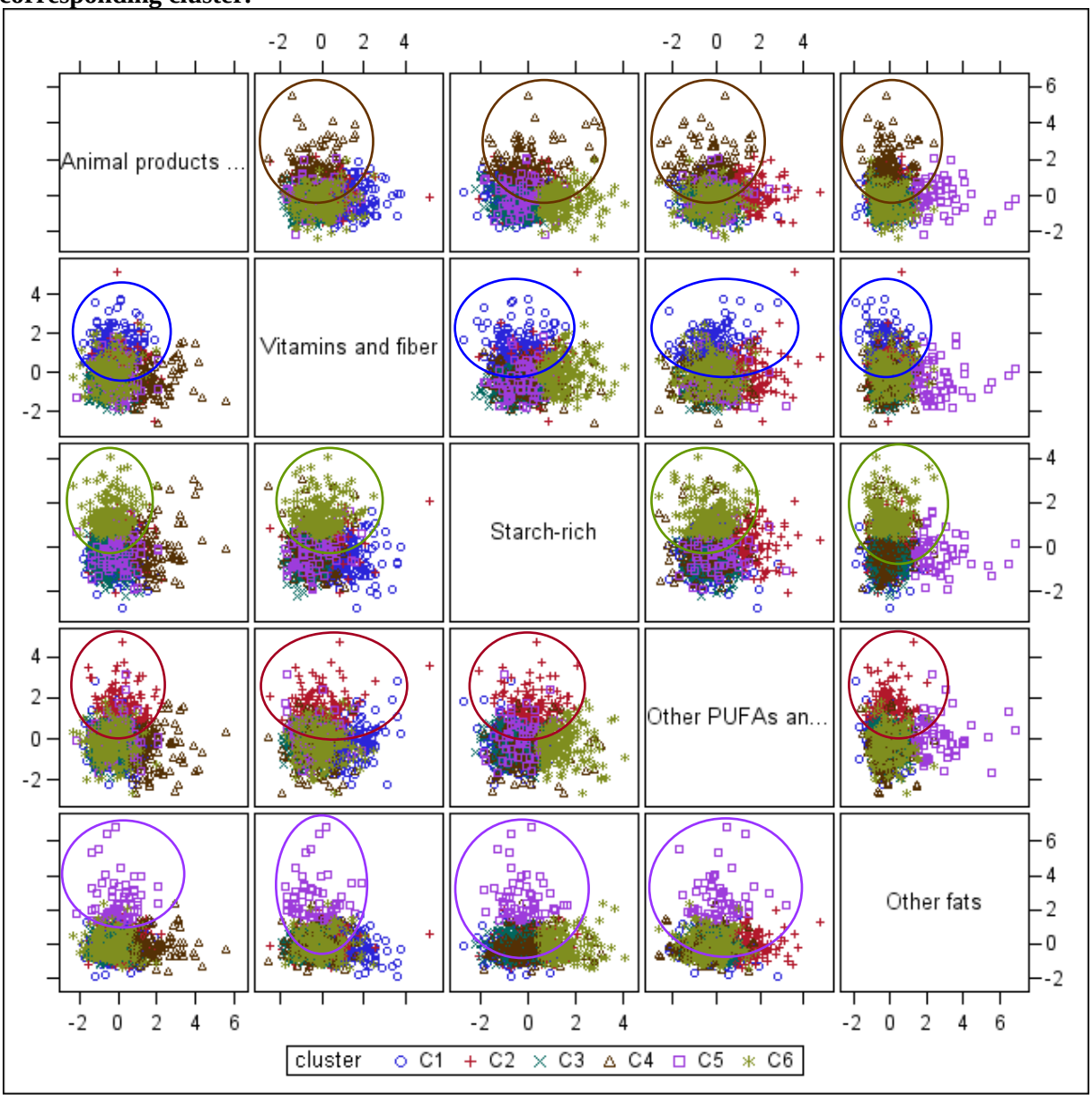
RESULTS FROM THE REDUCED DATASET (EXCLUDING 8 POTENTIAL OUTLIERS)

Table 7R* – Description of the identified clusters in terms of cases and controls, and according to the original dietary patterns.

Cluster	Cases	Controls	All subjects	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
	N (%)	N (%)	N (%)	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)
C1	40 (13.29)	184 (24.93)	224 (21.56)	-0.14 (0.71)	1.12 (0.73)	-0.49 (0.68)	-0.26 (0.75)	-0.29 (0.61)
C2	41 (13.62)	126 (17.07)	167 (16.07)	0.15 (0.74)	-0.02 (0.88)	0.09 (0.73)	1.46 (0.75)	-0.05 (0.64)
C3	107 (35.55)	207 (28.05)	314 (30.22)	-0.41 (0.59)	-0.64 (0.50)	-0.46 (0.57)	-0.29 (0.61)	-0.24 (0.49)
C4	46 (15.28)	53 (7.18)	99 (9.53)	1.87 (0.94)	-0.36 (0.81)	-0.03 (1.03)	-0.54 (0.96)	-0.10 (0.73)
C5	14 (4.65)	44 (5.96)	58 (5.58)	-0.02 (0.84)	-0.33 (0.89)	-0.25 (0.76)	0.12 (0.94)	2.67 (1.22)
C6	53 (17.61)	124 (16.80)	177 (17.04)	-0.35 (0.71)	-0.02 (0.77)	1.44 (0.71)	-0.25 (0.78)	-0.08 (0.71)

*Results obtained from the reduced dataset (excluding 8 potential outliers).

Figure 7R* - Scatterplots of the dietary patterns, plotted with different symbols according to the corresponding cluster.



*Results obtained from the reduced dataset (excluding 8 potential outliers).

Table 8R* – Distribution of sociodemographic and lifestyle variables, for each cluster

	C1 (N=224)	C2 (N=167)	C3 (N=314)	C4 (N=99)	C5 (N=58)	C6 (N=177)
Age						
<55	67 (29.91)	50 (29.94)	44 (14.01)	25 (25.25)	15 (25.86)	63 (35.59)
55-59	53 (23.66)	37 (22.16)	59 (18.79)	22 (22.22)	13 (22.41)	32 (18.08)
60-64	44 (19.64)	27 (16.17)	63 (20.06)	24 (24.24)	12 (20.69)	38 (21.47)
65-69	39 (17.41)	34 (20.36)	73 (23.25)	10 (10.10)	12 (20.69)	30 (16.95)
≥70	21 (9.38)	19 (11.38)	75 (23.89)	18 (18.18)	6 (10.34)	14 (7.91)
Sex						
Male	144 (64.29)	149 (89.22)	255 (81.21)	90 (90.91)	53 (91.38)	170 (96.05)
Female	80 (35.71)	18 (10.78)	59 (18.79)	9 (9.09)	5 (8.62)	7 (3.95)
Education						
<7	129 (57.59)	93 (55.69)	207 (65.92)	74 (74.75)	49 (84.48)	113 (63.84)
7-11	64 (28.57)	45 (26.95)	72 (22.93)	18 (18.18)	7 (12.07)	46 (25.99)
≥12	31 (13.84)	29 (17.37)	35 (11.15)	7 (7.07)	2 (3.45)	18 (10.17)
Alcohol drinking						
Non drinker	70 (31.25)	20 (11.98)	56 (17.83)	11 (11.11)	4 (6.90)	16 (9.04)
Drinker <4 drinks/day	84 (37.50)	49 (29.34)	98 (31.21)	20 (20.20)	16 (27.59)	44 (24.86)
Drinker ≥4 drinks/day	70 (31.25)	98 (58.68)	160 (50.96)	68 (68.69)	38 (65.52)	117 (66.10)
Tobacco smoking						
Never smoker	86 (38.39)	44 (26.35)	80 (25.48)	22 (22.22)	15 (25.86)	30 (16.95)
Ex smoker	77 (34.38)	56 (33.53)	136 (43.31)	31 (31.31)	21 (36.21)	75 (42.37)
Current smoker <15 cigarettes/day	27 (12.05)	23 (13.77)	31 (9.87)	19 (19.19)	6 (10.34)	15 (8.47)
Current smoker 15-24 cigarettes/day	27 (12.05)	30 (17.96)	50 (15.92)	19 (19.19)	10 (17.24)	38 (21.47)
Current smoker ≥25 cigarettes/day	7 (3.13)	14 (8.38)	17 (5.41)	8 (8.08)	6 (10.34)	19 (10.73)
Body mass index						
≤18.5	3 (1.34)	1 (0.60)	1 (0.32)	5 (5.05)	1 (1.72)	1 (0.56)
18.6-24.9	122 (54.46)	72 (43.11)	159 (50.64)	43 (43.43)	18 (31.03)	106 (59.89)
25-29.9	70 (31.25)	71 (42.51)	111 (35.35)	42 (42.42)	25 (43.10)	52 (29.38)
≥30	29 (12.95)	23 (13.77)	43 (13.69)	9 (9.09)	14 (24.14)	18 (10.17)
Total energy intake						
<2080	63 (28.13)	9 (5.39)	179 (57.01)	3 (3.03)	1 (1.72)	6 (3.39)
2080-2590	78 (34.82)	29 (17.37)	100 (31.85)	15 (15.15)	10 (17.24)	30 (16.95)
2591-3230	57 (25.45)	63 (37.72)	30 (9.55)	23 (23.23)	16 (27.59)	72 (40.68)
≥3231	26 (11.61)	66 (39.52)	5 (1.59)	58 (58.59)	31 (53.45)	69 (38.98)

*Results obtained from the reduced dataset (excluding 8 potential outliers).

Table 9R* – Description of the identified clusters: mean daily intake of selected standardized nutrients and the mean daily total energy and non-alcoholic energy intakes, for each cluster.

Nutrient	C1 N=224	C2 N=167	C3 N=314	C4 N=99	C5 N=58	C6 N=177
	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)
Animal protein	-0.25 (0.71)	0.80 (0.72)	-0.68 (0.60)	1.14 (1.14)	0.46 (0.94)	-0.09 (0.79)
Vegetable protein	-0.22 (0.70)	0.27 (0.78)	-0.77 (0.52)	0.31 (1.00)	0.18 (0.89)	1.09 (0.79)
Cholesterol	-0.23 (0.73)	0.68 (0.89)	-0.61 (0.54)	0.95 (1.27)	0.44 (0.96)	-0.05 (0.82)
Saturated fatty acids	-0.14 (0.77)	0.45 (0.85)	-0.68 (0.54)	1.30 (1.20)	0.52 (0.95)	-0.03 (0.75)
Monounsaturated fatty acids	-0.06 (0.82)	0.73 (1.06)	-0.66 (0.56)	0.33 (1.05)	0.72 (1.31)	0.08 (0.82)
Linoleic acid	-0.26 (0.62)	0.21 (0.71)	-0.50 (0.51)	0.08 (0.77)	2.32 (1.12)	0.09 (0.74)
Linolenic acid	-0.25 (0.53)	0.11 (0.59)	-0.51 (0.46)	0.26 (0.76)	2.48 (1.50)	0.04 (0.74)
Other polyunsaturated fatty acids	-0.22 (0.76)	1.29 (0.83)	-0.56 (0.62)	-0.06 (1.05)	0.52 (0.96)	-0.12 (0.74)
Soluble carbohydrates	0.52 (0.86)	0.12 (0.88)	-0.70 (0.52)	0.55 (1.12)	0.12 (1.10)	0.06 (0.95)
Starch	-0.41 (0.66)	0.19 (0.75)	-0.63 (0.51)	0.45 (1.16)	0.06 (0.83)	1.12 (0.84)
Sodium	-0.35 (0.71)	0.22 (0.76)	-0.66 (0.50)	1.03 (1.17)	0.06 (0.81)	0.72 (0.87)
Calcium	0.03 (0.71)	0.08 (0.78)	-0.57 (0.61)	1.70 (1.15)	0.10 (0.96)	-0.17 (0.70)
Potassium	0.18 (0.74)	0.52 (0.84)	-0.90 (0.56)	0.60 (1.05)	0.43 (0.93)	0.30 (0.83)
Phosphorus	-0.21 (0.68)	0.49 (0.72)	-0.77 (0.56)	1.34 (1.08)	0.41 (0.86)	0.18 (0.76)
Iron	-0.25 (0.67)	0.61 (0.82)	-0.66 (0.76)	0.57 (1.20)	0.61 (1.02)	0.33 (0.90)
Zinc	-0.27 (0.70)	0.65 (0.80)	-0.78 (0.56)	0.95 (1.11)	0.37 (0.92)	0.38 (0.83)
Thiamin (vitamin b1)	0.08 (0.69)	0.52 (0.88)	-0.86 (0.50)	0.69 (1.07)	0.26 (0.88)	0.36 (0.89)
Riboflavin (vitamin b2)	0.08 (0.75)	0.33 (0.76)	-0.71 (0.57)	1.24 (1.15)	0.19 (0.77)	-0.06 (0.77)
Vitamin b6	0.11 (0.73)	0.73 (0.84)	-0.89 (0.54)	0.48 (1.08)	0.43 (0.96)	0.24 (0.82)
Total folate	0.36 (0.82)	0.38 (0.89)	-0.89 (0.56)	0.49 (0.96)	0.22 (0.92)	0.30 (0.81)
Niacin	-0.09 (0.80)	1.03 (0.76)	-0.83 (0.54)	0.14 (1.02)	0.57 (0.94)	0.29 (0.78)
Vitamin c	0.95 (1.01)	0.22 (0.92)	-0.71 (0.48)	-0.13 (0.81)	-0.06 (0.78)	-0.09 (0.84)
Retinol	-0.10 (0.85)	0.47 (1.16)	-0.23 (0.86)	0.45 (1.08)	0.12 (0.89)	-0.30 (0.71)
Beta-carotene equivalents	0.51 (0.89)	0.22 (0.73)	-0.53 (0.44)	-0.17 (0.72)	0.10 (0.83)	0.00 (0.67)
Vitamin d	-0.15 (0.84)	1.22 (1.03)	-0.53 (0.63)	-0.01 (0.80)	0.37 (0.99)	-0.18 (0.75)
Vitamin e	0.10 (0.73)	0.51 (0.93)	-0.73 (0.52)	-0.07 (0.93)	1.61 (1.14)	0.10 (0.74)
Total fibre	0.63 (0.82)	0.22 (0.92)	-0.84 (0.56)	-0.04 (0.96)	0.07 (0.99)	0.41 (0.80)

*Results obtained from the reduced dataset (excluding 8 potential outliers).

Table 10R* – Description of the identified clusters: median weekly intake of selected food groups by cluster.

Food groups	C1 N=224	C2 N=167	C3 N=314	C4 N=99	C5 N=58	C6 N=177
	median	median	median	median	median	median
Milk	7.0	4.8	2.8	12.0	5.1	2.0
Coffee	14.0	18.5	14.0	14.0	14.3	14.0
Bread	15.3	21.5	15.0	24.3	20.5	35.0
Pasta and rice	4.8	5.8	4.3	4.5	4.8	6.0
Soup	2.0	2.0	1.5	2.3	2.1	2.3
Egg	1.0	2.0	1.0	2.0	2.0	1.0
White meat	2.0	2.0	1.0	1.0	2.0	1.0
Red meat	3.5	6.0	3.8	4.8	5.0	4.8
Processed meat	2.0	3.0	2.0	3.0	2.5	2.5
Fish	1.5	2.0	1.5	1.0	1.5	1.0
Cheese	4.3	4.4	3.6	8.7	4.6	4.1
Potatoes	1.0	2.0	1.0	1.5	1.5	1.0
Pulses	1.5	1.5	1.0	1.5	1.5	1.5
Leafy vegetables	7.0	5.5	3.5	4.2	7.0	5.3
Fruiting vegetables	3.5	3.0	1.7	1.8	3.5	2.5
Cruciferous vegetables	0.5	0.5	0.3	0.3	0.5	0.3
Other vegetables	2.2	1.8	0.7	0.8	1.1	1.0
Citrus fruit	5.0	3.5	1.5	2.5	2.2	2.0
Other fruit	18.7	11.5	8.4	10.6	10.9	11.7
Soft drinks and fruit juices	0.0	0.5	0.0	0.5	0.3	0.0
Desserts	3.0	3.0	1.5	4.5	2.9	2.3
Sugar and candies	28.0	36.0	25.0	43.0	44.3	42.0
Butter and margarine	1.9	2.3	1.5	5.2	3.0	1.6
Seed oils	0.7	2.1	1.9	3.1	49.9	3.9
Olive oil	26.0	39.4	14.6	18.4	2.0	25.8
	mean+	**mean+**	**mean+**	**mean+**	**mean+**	**mean+**
Total energy	2442.71	3108.76	2011.37	3492.20	3358.91	3166.79
Alcohol (g)	28.30	54.29	46.16	77.38	71.14	57.21
Non alcoholic energy	2244.61	2728.74	1688.27	2950.57	2860.93	2766.29

*Results obtained from the reduced dataset (excluding 8 potential outliers). + Daily intake.

Table 11R* – Odds ratios (OR) of esophageal cancer and corresponding 95% floating confidence intervals (FCI), by cluster

Cluster	Cases	Controls	Total	crude OR (95% FCI)	adjusted OR (95% FCI)[1]
C1	40	184	224	0.42 (0.30-0.59)	0.61 (0.41-0.90)
C2	41	126	167	0.63 (0.44-0.90)	0.73 (0.49-1.09)
C3	107	207	314	1.00 (0.79-1.26)	1.00 (0.75-1.33)
C4	46	53	94	1.68 (1.13-2.49)	1.59 (1.00-2.52)
C5	53	124	177	0.83 (0.60-1.14)	0.59 (0.41-0.86)
C6	14	44	58	0.62 (0.34-1.12)	0.43 (0.21-0.86)

[1]Adjusted for age, sex, center, education, alcohol drinking, tobacco smoking, body mass index.
[*]Results obtained from the full dataset.

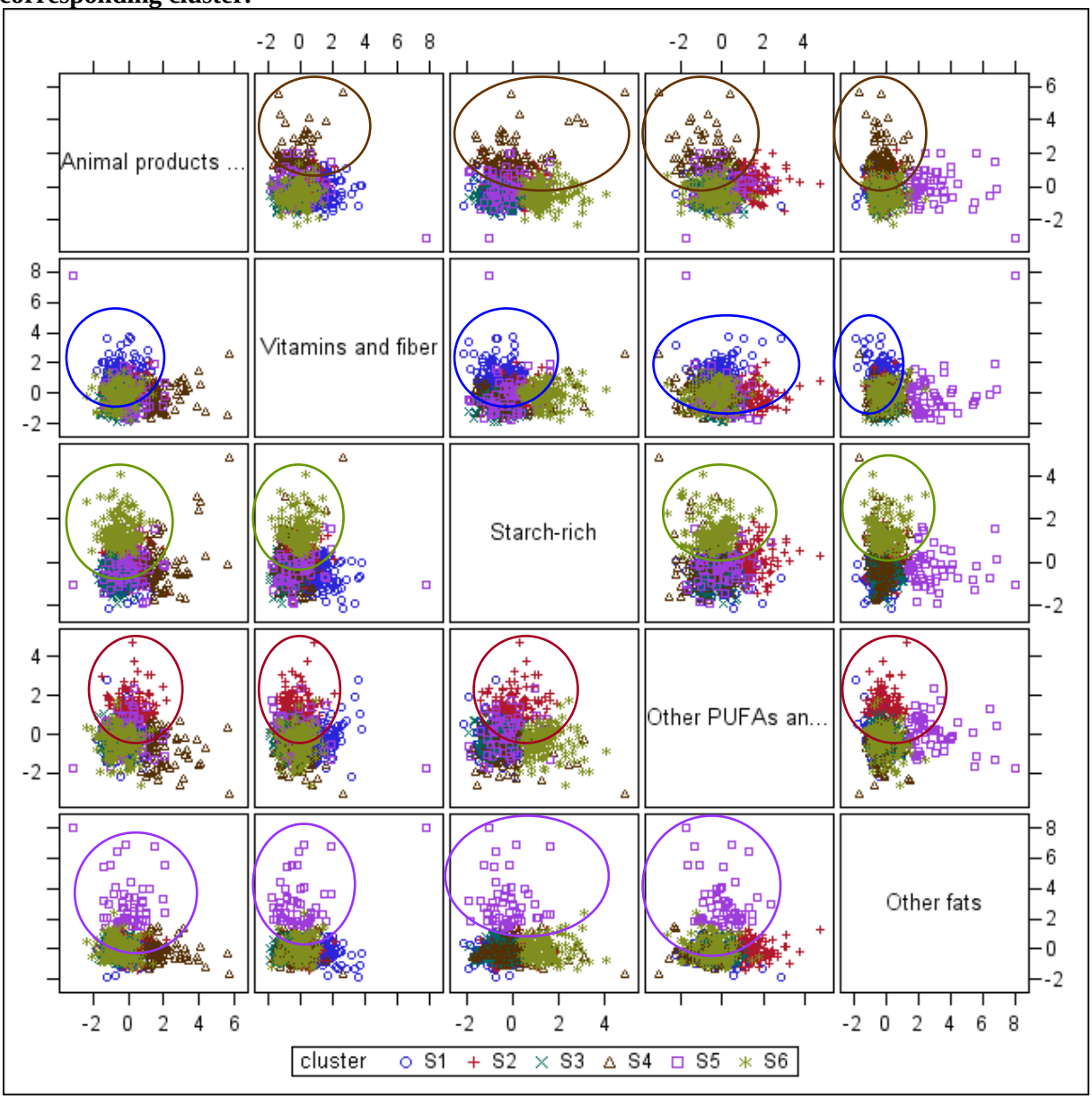
RESULTS FROM THE SUBSET OF SUBJECTS WHO WERE CLASSIFIED IN THE SAME WAY IN THE THREE SOLUTIONS BASED ON EUCLIDEAN, MANHATTAN AND LAGRANGE DISTANCES

Table 7S[*] – Description of the identified clusters in terms of cases and controls, and according to the original dietary patterns.

Cluster	Cases	Controls	All subjects	Animal products and related components	Vitamins and fiber	Starch-rich	Other PUFAs and vitamin D	Other fats
	N (%)	N (%)	N (%)	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)
S1	24 (11.37)	126 (24.28)	150 (20.55)	-0.25 (0.63)	1.19 (0.72)	-0.57 (0.55)	-0.20 (0.68)	-0.35 (0.55)
S2	30 (14.22)	91 (17.53)	121 (16.58)	0.22 (0.72)	-0.10 (0.68)	0.09 (0.66)	1.53 (0.68)	-0.10 (0.58)
S3	70 (33.18)	127 (24.47)	197 (26.99)	-0.59 (0.48)	-0.71 (0.48)	-0.57 (0.46)	-0.33 (0.56)	-0.26 (0.44)
S4	31 (14.69)	40 (7.71)	71 (9.73)	2.01 (1.03)	-0.25 (0.81)	-0.04 (1.15)	-0.75 (0.89)	-0.17 (0.65)
S5	13 (6.16)	42 (8.09)	55 (7.53)	0.00 (0.95)	-0.08 (1.42)	-0.24 (0.79)	0.04 (0.90)	2.98 (1.53)
S6	43 (20.38)	93 (17.92)	136 (18.63)	-0.46 (0.67)	-0.06 (0.67)	1.42 (0.70)	-0.41 (0.72)	-0.10 (0.69)

[*]Results obtained from the subset of subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distances.

Figure 7S[*] - Scatterplots of the dietary patterns, plotted with different symbols according to the corresponding cluster.



[*]Results obtained from the subset of subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distances.

Table 8S* – Distribution of sociodemographic and lifestyle variables, for each cluster

	S1 (N=150)	S2 (N=121)	S3 (N=197)	S4 (N=71)	S5 (N=55)	S6 (N=136)
Age						
<55	45 (30.00)	40 (33.06)	27 (13.71)	17 (23.94)	15 (27.27)	48 (35.29)
55-59	36 (24.00)	27 (22.31)	42 (21.32)	15 (21.13)	13 (23.64)	24 (17.65)
60-64	33 (22.00)	20 (16.53)	45 (22.84)	17 (23.94)	11 (20.00)	29 (21.32)
65-69	23 (15.33)	19 (15.70)	46 (23.35)	7 (9.86)	11 (20.00)	25 (18.38)
≥70	13 (8.67)	15 (12.40)	37 (18.78)	15 (21.13)	5 (9.09)	10 (7.35)
Sex						
Male	90 (60.00)	109 (90.08)	158 (80.20)	63 (88.73)	51 (92.73)	130 (95.59)
Female	60 (40.00)	12 (9.92)	39 (19.80)	8 (11.27)	4 (7.27)	6 (4.41)
Education						
<7	86 (57.33)	73 (60.33)	126 (63.96)	59 (83.10)	45 (81.82)	86 (63.24)
7-11	41 (27.33)	28 (23.14)	44 (22.34)	9 (12.68)	8 (14.55)	37 (27.21)
≥12	23 (15.33)	20 (16.53)	27 (13.71)	3 (4.23)	2 (3.64)	13 (9.56)
Alcohol drinking						
Non drinker	50 (33.33)	15 (12.40)	39 (19.80)	10 (14.08)	5 (9.09)	15 (11.03)
Drinker <4 drinks/day	53 (35.33)	33 (27.27)	59 (29.95)	14 (19.72)	14 (25.45)	33 (24.26)
Drinker ≥4 drinks/day	47 (31.33)	73 (60.33)	99 (50.25)	47 (66.20)	36 (65.45)	88 (64.71)
Tobacco smoking						
Never smoker	63 (42.00)	32 (26.45)	45 (22.84)	18 (25.35)	15 (27.27)	21 (15.44)
Ex smoker	46 (30.67)	42 (34.71)	92 (46.70)	22 (30.99)	17 (30.91)	58 (42.65)
Current smoker <15 cigarettes/day	19 (12.67)	16 (13.22)	18 (9.14)	12 (16.90)	6 (10.91)	10 (7.35)
Current smoker 15-24 cigarettes/day	17 (11.33)	23 (19.01)	27 (13.71)	11 (15.49)	11 (20.00)	32 (23.53)
Current smoker ≥25 cigarettes/day	5 (3.33)	8 (6.61)	15 (7.61)	8 (11.27)	6 (10.91)	15 (11.03)
Body mass index						
≤18.5	1 (0.67)	1 (0.83)	1 (0.51)	4 (5.63)	1 (1.82)	0 (0.00)
18.6-24.9	88 (58.67)	53 (43.80)	99 (50.25)	31 (43.66)	18 (32.73)	82 (60.29)
25-29.9	41 (27.33)	47 (38.84)	67 (34.01)	32 (45.07)	24 (43.64)	39 (28.68)
≥30	20 (13.33)	20 (16.53)	30 (15.23)	4 (5.63)	12 (21.82)	15 (11.03)
Total energy intake						
<2080	46 (30.67)	6 (4.96)	142 (72.08)	3 (4.23)	1 (1.82)	6 (4.41)
2080-2590	59 (39.33)	20 (16.53)	46 (23.35)	12 (16.90)	8 (14.55)	27 (19.85)
2591-3230	36 (24.00)	47 (38.84)	8 (4.06)	13 (18.31)	15 (27.27)	58 (42.65)
≥3231	9 (6.00)	48 (39.67)	1 (0.51)	43 (60.56)	31 (56.36)	45 (33.09)

*Results obtained from the subset of subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distances.

Table 9S[*] – Description of the identified clusters: mean daily intake of selected standardized nutrients and the mean daily total energy and non-alcoholic energy intakes, for each cluster.

Nutrient	S1 (N=150)	S2 (N=121)	S3 (N=197)	S4 (N=71)	S5 (N=55)	S6 (N=136)
	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)	Mean (std)
Animal protein	-0.32 (0.68)	0.87 (0.70)	-0.88 (0.50)	1.12 (0.97)	0.53 (0.92)	-0.26 (0.71)
Vegetable protein	-0.32 (0.56)	0.28 (0.75)	-0.94 (0.47)	0.34 (1.18)	0.33 (1.06)	1.02 (0.77)
Cholesterol	-0.31 (0.66)	0.80 (0.93)	-0.76 (0.44)	1.00 (1.30)	0.52 (1.06)	-0.20 (0.75)
Saturated fatty acids	-0.23 (0.66)	0.48 (0.83)	-0.87 (0.47)	1.39 (1.11)	0.61 (0.98)	-0.18 (0.63)
Monounsaturated fatty acids	-0.09 (0.78)	0.74 (0.94)	-0.77 (0.51)	0.27 (0.77)	0.85 (1.33)	-0.04 (0.76)
Linoleic acid	-0.34 (0.49)	0.18 (0.70)	-0.60 (0.44)	0.03 (0.73)	2.64 (1.57)	0.03 (0.71)
Linolenic acid	-0.31 (0.48)	0.08 (0.52)	-0.60 (0.43)	0.22 (0.66)	2.78 (1.70)	-0.04 (0.73)
Other polyunsaturated fatty acids	-0.23 (0.68)	1.33 (0.74)	-0.66 (0.55)	-0.22 (0.86)	0.58 (0.91)	-0.29 (0.69)
Soluble carbohydrates	0.48 (0.85)	0.12 (0.84)	-0.83 (0.48)	0.72 (1.36)	0.29 (1.12)	0.02 (0.98)
Starch	-0.52 (0.54)	0.23 (0.72)	-0.77 (0.47)	0.52 (1.43)	0.16 (0.95)	1.07 (0.80)
Sodium	-0.47 (0.60)	0.27 (0.74)	-0.83 (0.44)	1.00 (1.26)	0.20 (0.96)	0.62 (0.89)
Calcium	-0.06 (0.66)	0.06 (0.74)	-0.78 (0.51)	1.90 (1.17)	0.20 (0.94)	-0.27 (0.65)
Potassium	0.15 (0.70)	0.48 (0.74)	-1.08 (0.50)	0.64 (1.18)	0.62 (1.03)	0.16 (0.78)
Phosphorus	-0.30 (0.61)	0.50 (0.69)	-0.99 (0.48)	1.42 (1.17)	0.53 (0.91)	0.03 (0.68)
Iron	-0.29 (0.62)	0.61 (0.76)	-0.82 (0.71)	0.44 (1.15)	0.72 (1.02)	0.20 (0.89)
Zinc	-0.35 (0.64)	0.71 (0.81)	-1.00 (0.50)	0.94 (1.05)	0.50 (0.98)	0.23 (0.76)
Thiamin (vitamin b1)	0.02 (0.65)	0.57 (0.88)	-1.04 (0.42)	0.74 (1.12)	0.43 (1.04)	0.24 (0.85)
Riboflavin (vitamin b2)	-0.03 (0.66)	0.36 (0.73)	-0.91 (0.47)	1.40 (1.36)	0.34 (0.89)	-0.20 (0.69)
Vitamin b6	0.09 (0.73)	0.76 (0.78)	-1.09 (0.45)	0.48 (1.02)	0.60 (1.05)	0.08 (0.77)
Total folate	0.29 (0.70)	0.38 (0.79)	-1.09 (0.50)	0.56 (1.01)	0.42 (1.10)	0.17 (0.76)
Niacin	-0.11 (0.76)	1.07 (0.74)	-1.00 (0.49)	0.01 (0.90)	0.69 (0.95)	0.15 (0.74)
Vitamin c	1.05 (1.03)	0.13 (0.76)	-0.81 (0.43)	-0.08 (0.73)	0.06 (0.87)	-0.19 (0.67)
Retinol	-0.18 (0.79)	0.58 (1.17)	-0.39 (0.62)	0.44 (1.07)	0.16 (0.95)	-0.36 (0.67)
Beta-carotene equivalents	0.53 (0.92)	0.22 (0.70)	-0.59 (0.47)	-0.18 (0.57)	0.55 (2.75)	-0.04 (0.63)
Vitamin d	-0.15 (0.76)	1.28 (0.95)	-0.59 (0.61)	-0.06 (0.69)	0.42 (0.95)	-0.34 (0.67)
Vitamin e	0.06 (0.66)	0.47 (0.79)	-0.84 (0.46)	-0.10 (0.76)	1.89 (1.35)	0.00 (0.72)
Total fibre	0.57 (0.77)	0.21 (0.85)	-1.00 (0.49)	0.02 (1.00)	0.29 (1.23)	0.32 (0.76)

[*]Results obtained from the subset of subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distances.

Table 10S* – Description of the identified clusters: median weekly intake of selected food groups by cluster.

Food groups	S1 (N=150)	S2 (N=121)	S3 (N=197)	S4 (N=71)	S5 (N=55)	S6 (N=136)
	median	median	median	median	median	median
Milk	7.00	4.00	2.00	14.00	7.00	2.00
Coffee	14.00	14.50	14.00	14.00	14.50	14.00
Bread	15.00	22.00	14.75	23.75	21.25	35.00
Pasta and rice	4.75	5.50	4.00	4.75	5.00	6.00
Soup	2.00	2.00	1.50	2.25	2.50	2.25
Egg	1.00	2.00	1.00	2.00	2.00	1.00
White meat	2.00	2.00	1.00	1.00	2.00	1.00
Red meat	3.50	6.00	3.50	4.75	5.25	4.38
Processed meat	2.00	3.00	1.50	3.00	3.00	2.50
Fish	1.50	2.00	1.50	1.00	1.50	1.00
Cheese	4.08	4.40	3.27	8.83	5.07	3.92
Potatoes	1.00	2.00	1.00	1.00	2.00	1.00
Pulses	1.25	1.50	1.00	1.50	1.50	1.50
Leafy vegetables	7.00	5.50	3.25	4.50	7.00	5.00
Fruiting vegetables	3.50	3.00	1.50	1.92	3.75	2.50
Cruciferous vegetables	0.50	0.50	0.25	0.25	0.33	0.25
Other vegetables	2.17	1.83	0.83	0.83	1.17	0.92
Citrus fruit	5.67	3.33	1.00	3.50	3.00	2.00
Other fruit	19.50	11.50	7.83	11.17	11.83	11.60
Soft drinks and fruit juices	0.25	0.50	0.00	0.50	0.00	0.00
Desserts	2.75	3.50	1.42	4.50	3.50	2.19
Sugar and candies	28.00	35.00	21.00	50.00	45.50	42.00
Butter and margarine	1.85	2.27	1.40	6.18	2.88	1.57
Seed oils	0.54	2.22	2.24	1.83	56.24	3.47
Olive oil	26.73	39.67	14.12	18.39	1.58	24.97
	mean+	**mean+**	**mean+**	**mean+**	**mean+**	**mean+**
Total energy	2368.11	3114.71	1853.44	3486.98	3499.94	3064.72
Alcohol (g)	28.32	51.62	44.68	69.44	71.55	55.14
Non alcoholic energy	2169.89	2753.4	1540.64	3000.91	2999.09	2678.71

*Results obtained from the subset of subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distances. + Daily intake.

Table 11S* – Odds ratios (OR) of esophageal cancer and corresponding 95% floating confidence intervals (FCI), by cluster

Cluster	Cases	Controls	Total	crude OR (95% FCI)	adjusted OR (95% FCI)[1]
S1	24	126	150	0.41 (0.29-0.58)	0.39 (0.28-0.55)
S2	30	91	121	0.63 (0.45-0.90)	0.64 (0.45-0.91)
S3	70	127	197	1.00 (0.79-1.26)	1.00 (0.78-1.27)
S4	31	40	71	1.39 (0.93-2.10)	1.35 (0.89-2.05)
S5	13	42	55	0.56 (0.30-1.03)	0.54 (0.29-1.02)
S6	43	93	136	0.85 (0.63-1.15)	0.82 (0.60-1.12)

[1]Adjusted for age, sex, center, education, alcohol drinking, tobacco smoking, body mass index.
[*]Results obtained from the subset of subjects who were classified in the same way in the three solutions based on Euclidean, Manhattan and Lagrange distance.