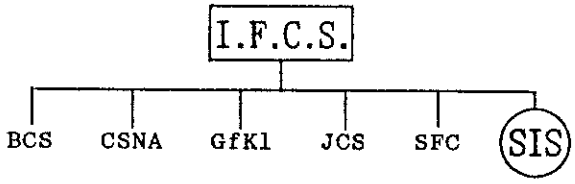
SOCIETA' ITALIANA DI STATISTICA
GRUPPO ITALIANO ADERENTE ALL'I.F.C.S.
(International Federation of Classification Societies)

# ATTI DELLE GIORNATE DI STUDIO DEL GRUPPO ITALIANO ADERENTE ALL'I.F.C.S.

ERICE - TRAPANI
24 - 25 Ottobre 1988



# EXPLORATORY ANALYSIS OF DIETARY DATA DERIVING FROM A CASE CONTROL STUDY

ADRIANO DECARLI - ETTORE MARUBINI
Istituto di Statistica Medica e Biometria
Istituto Nazionale Tumori
Milano

MONICA FERRARONI
Istituto di Statistica Medica e Biometria
Milano

## 1. Introduction

The relevance of results obtained from epidemiological research on lifestyle factors (i.e alcohol, tobacco, drug use, etc) together with the increasing availability of computing hardware and software, has recently stimulated the collection and analysis of large quantities of data also on exposures which are difficult to quantify, with the consequent possibility of formulating and testing newer aetiological hypotheses.

One of the major fields in which the afore mentioned problems (large amount of data, difficulty of quantifying the level of exposure and the degree of accuracy of its measure) are present simultaneously is the study of the relationship between dietary habits and health (Wahrendorf, 1986; Rasanen and Pietinen, 1986).

Data collection in epidemiological studies on diet is usually aimed at the evaluation of a well defined hypothesis (for example the association between vitamin A intake and breast cancer). Nevertheless, it is necessary to collect large quantities of data not directly connected with the hypothesis under study. This is done in order to achieve a better understanding of dietary habits, as for example the evaluation of total calorie intake (Willet and Stampfer, 1986). In any case, diet varies considerably in different populations and time periods and is characterized by a large number of independent variables with complex interrelations. Therefore the researcher, after testing the main hypothesis, is obviously stimulated to use the large amount of available information to define new hypotheses.

In case-control studies this phase is usually approached by estimating the existing degree of association between the intake level of each food item and the possibile onset of the pathological condition under examination. The fit of a logistic model including, along with confounding variables, all food items

which, singly, have proved to be significantly associated with the disease, enables to select a group of variables which convey an independent contribution to the disease risk. These variables are then usually put together in order to estabilish a risk score (Trichopoulos et al., 1985).

In the exploratory analysis of data, a promising approach to whereby the relationships between food items and the disease under investigation can be understood, is the definition of groups according to multivariate methods, such as cluster analysis. In these circumstances clusters are characterized by different dietary patterns and allow one to assess the association between dietary patterns and disease occurence.

When the diet as a whole is taken into consideration, we have to cope with a large number of highly correlated variables; as a consequence, the classification methods can lead to largely unstable and unreproducible results (MacGee, 1984). Aim of this paper is to suggest a multistep strategy of analysis the core of which is the use of standard multivariate techniques for discrete variables, multiple correspondence analysis (MCA) (Benzecri, 1973), as well as for continuous variables, principal component analysis (PCA) (Jolliffe, 1986), in order to identify dietary clusters.

These multivariate classification methods allow the user to define orthogonal axes which can be easily interpreted since they represent a linear combination of the original variables. This improves the interpretation of clusters in terms of dietary habits.

We will discuss the basic findings from an exploratory analysis carried out both with traditional methods, as well as with PCA or MCA on 214 cases and 215 controls of a study on breast cancer recently carried out at the Istituto Nazionale dei Tumori (I.N.T.) of Milan (Marubini et al., 1988; Gerber et al., 1988).

A comparison between results obtained with the different procedures is illustrated; this enables to point out the potentialities of the suggested procedure. Such procedure needs in any case further investigation, especially in relation to some methodological aspects.

## 2. Subjects

The data utilized refers to a case-control study on diet and breast cancer. The main features of the study have been illustrated elsewhere (Marubini et al., 1988).

Briefly, the sample is made up of 214 cases of primary carcinoma of the breast selected from consecutive admissions to the I.N.T. of Milan from May 1982 to June 1985. 215 controls were selected from female patients consecutively admitted to one of the major University hospitals of Milan (San Raffaele) during the same period, with the exception of patients admitted for malignant tumors and for hepatic, vascular and metabolic diseases.

Both cases and controls had the following characteristics: age between 30 and 65 years; residence in Milan or its province;

negative medical history of breast cancer or other malignancies. With reference to Milan and its province the catchment area of the two hospitals are comparable. A blood sample was taken from each subject and tested for retinol, B-carotene, Vitamin E, Vitamin C and Vitamin B, total cholesterol, triglycerides, high density lipoproteins (HDL), low density lipoproteins (LDL) copper and zinc.

All subjects were interviewed in the same standard way, during hospitalization, by previously trained interviewers. No subject refused to co-operate.

## 2.1 Selection of cases

Cases were chosen by means of a periodic check from the patients admitted to the I.N.T.

Subjects with an admission diagnosis of primary breast cancer and with the required characteristics were randomly selected. 258 women were interviewed: 42 subjects were excluded because of negative histological diagnosis (benign breast diseases) and 2 because they had previous malignancies. The remaining 214 cases were aged 31-65 (median age=48).

## 2.2 Selection of controls

Controls were chosen by means of a periodic examination from patients admitted to the San Raffaele Hospital.

Subjects were chosen in such a way as to have an equal number of cases and controls during the same period of time. The controls were hospitalized for orthopedic illnesses (46%), acute surgical condition (22%) and other including peripheral venous diseases (32%).

A total number of 222 controls was interviewed: 7 subjects were excluded because the discharge diagnosis did not agree with the protocol. The remaining 215 controls were aged 30-64 (median age=47).

## 3. Questionnaire

Dietary information was collected by two dieticians who interviewed an approximately equal number of cases and controls. Socio-economic information, medical history, and dietary information based on a dietary history questionnaire (Burke, 1947; Lyon et al., 1983; Block, 1982; Byers et al., 1987) was collected. Subjects were asked to try to remember their usual weekly consumption of 69 foodstuffs. For eight other Vitamin A rich foods the questionnaire gave an estimate referring to annual consumption. Particular attention was taken in collecting information about seasonal food stuffs such as fruit and vegetables.

From these 77 foodstuffs or groups of foodstuffs the consumption was transformed into the daily intake of nutrients by means of tables derived from different sources (Paul and Southgate, 1979; Fidanza et al., 1974).

## 4. Statistical Methods

The majority of variables was categorized in terms of the quintiles estimated in the control group. For some variables a dichotomic coding (no intake/intake) or a three level code (no intake/low and medium intake/high intake) was used.

The exploratory step of the analysis was carried out on the 64 food items assumed by more than 5% of subjects by following two different approaches.

In the first one the analysis was performed by the approach currently adopted in epidemiology (Breslow and Day, 1980; Trichopoulos et al., 1985). After all, food variables associated in a significant way with the risk of breast cancer were identified. This was done by fitting a logistic model for each re-coded food variable. Each model was adjusted for potential confounding factors: age, age of menarche, menopausal status, age at first birth, education, Quetelet index. All food items, which in this phase showed a $X^2$ value for trend higher than the preselected threshold of 3.84 ($\alpha$=0.05) were picked up. For each of them the odds ratios (OR) of breast cancer were computed together with their 95% approximate confidence intervals (CI) according to various consumption levels.

All these variables (plus confounding factors) were simultaneously inserted into a multiple logistic regression model the constants of which were estimated by maximum likelihood.

Finally, all items resulting significantly associated with breast cancer risk, also after this step, were utilized for calculating a risk score. The OR and its approximate confidence interval was finally estimated for each score value. All analyses were performed using package GLIM (Payne, 1985).

The second approach resorts to the use of structural analysis tecniques in order to define different dietary habit patterns (Stellman, 1986). To this end all dietary information on the 64 food items previously selected, was considered.

The steps of the analysis may be summarized as follows: a) reduction of the dimension of variables space; b) definition of dietary groups; c) evaluation of association between dietary groups and disease.

Among various available techniques, the attention was focused on the application of: i) multiple correspondence analysis (MCA); ii) principal component analysis (PCA).

The first method enables to investigate the association among variables taking into account the original space of variables. It provides concise information through orthogonal factorial axes identified by linear combinations of different modalities of the original variables. Also PCA permits the definition of sets of orthogonal axes which are linear combinations of the original variables.

In both cases, the axes have an interpretation in terms of dietary habits. However, it is worth noting that MCA is more appropriate for the identification of non linear associations between variables (Lauro and Decarli, 1982).

Usually, both the first PCA factorial axis and the first MCA factorial axis provide an estimate of the dimension of phenomena under investigation, enabling to discriminate high intake

from low intake subjects.

The definition of dietary groups was performed in two different ways: i) by using the first 12 factorial axes defined through MCA; ii) by utilizing continuous variables in PCA. As before the first 12 principal components were employed for definition of clusters.

The number of axes was chosen in such a way as to obtain comparable values of the variance explained by the same number of axes. An aggregation algorithm "nuages dinamiques" to identify stable groupings of subjects was utilized, both for discrete and for continuous data. This algorithm makes all possible comparisons of basic clusters, obtained through aggregation of subjects around moving centres determined at random (Diday, 1979, p.11-27). An arbitrary number of 5 clusters was adopted in both analyses; finally the clusters was interpreted in terms of dietary habits. The whole procedure was accomplished by using SPAD statistical package (Lebart and Morineau, 1982).

The evaluation of the association between dietary pattern and disease was performed by assigning to each subject, in addition, to her own socio-economic variables, a dummy variable in order to define the corresponding cluster. This variable can be considered a global risk score.

Owing to the fact that in this context any choice of reference category suitable for estimating the OR is arbitrary, the totality of data (cases+controls) was here adopted as reference set.

Also in this case the results were adjusted for age, education, age at first birth, age at menarche, menopausal status and Quetelet index. This analysis was carried out by means of GLIM Package.

## 5. Results

### 5.1 Analysis using Multiple Logistic Regression

For all food items considered in the analysis, the frequency distributions of cases and controls in the categories defined according to the quintiles determined on controls are presented in Table 1. Estimates of the odds ratios associated with each consumption level are also shown. The table gives the values of $X^2$ for trend computed by considering the qualitative variable defining the quintile, as a continuous variable. Only food items relative to sausages, giblets, potatoes, apples, butter, margarine, tea and turnips have a statistically signifiant $X^2$ value. The eight food items showing significant associations with breast cancer were inserted in a multiple logistic regression model including major non-dietary risk factors (age, age of menarche, age at first birth, menopausal status, education and Quetelet index). Table 2 gives estimates of OR and the corresponding confidence intervals estimated from the model for the above mentioned food items. It follows from these analyses that the effects of these food items on breast cancer risk are quite independent.

Tab. 1:Relation of breast cancer risk with frequencies of use of selected food items and beverage evaluated on 215 cases of breast cancer and 214 hospital controls collected in Milan, 1985-1986.

| Food or beverage | Frequencies of consumption (quintile) (N°cases:N°controls) | | | | | Odds Ratios estimates (£) | | | | | $X^2$(trend) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 |
| Beef | 42:43 | 48:43 | 43:43 | 34:43 | 47:43 | 1* | 1.23 | 1.11 | 0.88 | 1.36 | 0.17 |
| Pork | 47:43 | 31:42 | 31:46 | 48:43 | 57:41 | 1* | 0.70 | 0.65 | 1.09 | 1.54 | 3.48 |
| Veal | 55:67 | 35:32 | 28:30 | 34:42 | 62:44 | 1* | 1.19 | 1.29 | 0.81 | 1.57 | 1.05 |
| Chicken | 54:44 | 49:52 | 40:41 | 25:38 | 46:40 | 1* | 0.79 | 0.77 | 0.54 | 0.97 | 0.36 |
| Sausages | 37:46 | 30:41 | 38:43 | 47:42 | 62:43 | 1* | 0.97 | 1.23 | 1.44 | 2.30 | 8.48 |
| Giblets | 147:122 | 66:93 | | | | 1* | 0.58 | | | | 6.29 |
| Fresh fish | 78:73 | 55:62 | 81:80 | | | 1* | 0.90 | 0.95 | | | 0.03 |
| Preser.fish | 143:135 | 71:80 | | | | 1* | 0.83 | | | | 0.67 |
| Eggs | 41:45 | 39:41 | 39:43 | 38:39 | 57:47 | 1* | 1.02 | 1.00 | 1.11 | 1.24 | 0.57 |
| Milk | 50:42 | 41:44 | 35:43 | 43:44 | 45:42 | 1* | 0.81 | 0.68 | 0.84 | 0.84 | 0.20 |
| Skimmed milk | 149:152 | 65:63 | | | | 1* | 0.98 | | | | 0.00 |
| Half fat cheese | 53:46 | 33:40 | 40:43 | 45:43 | 43:43 | 1* | 0.84 | 0.91 | 0.89 | 0.92 | 0.03 |
| Full fat cheese | 44:43 | 35:44 | 31:42 | 41:43 | 63:43 | 1* | 0.66 | 0.69 | 0.89 | 1.33 | 1.81 |
| Bread | 34:43 | 43:43 | 40:43 | 51:43 | 46:43 | 1* | 1.34 | 1.30 | 1.65 | 1.46 | 1.68 |
| Brown bread | 203:206 | 11:9 | | | | 1* | 1.16 | | | | 0.08 |
| Crackers | 134:143 | 80:72 | | | | 1* | 1.11 | | | | 0.24 |
| Brown crackers | 201:195 | 13:20 | | | | 1* | 0.56 | | | | 2.23 |
| Biscuits | 98:101 | 75:61 | 41:53 | | | 1* | 1.21 | 0.73 | | | 0.78 |
| Pastry | 68:68 | 41:48 | 43:61 | 62:38 | | 1* | 0.81 | 0.67 | 1.48 | | 0.72 |
| Pasta | 49:43 | 48:43 | 45:43 | 32:43 | 40:43 | 1* | 0.88 | 0.85 | 0.57 | 0.81 | 1.30 |
| Rice | 40:49 | 32:40 | 49:43 | 48:42 | 45:41 | 1* | 0:90 | 1.28 | 1.14 | 1.08 | 0.27 |
| Potatoes | 36:44 | 37:43 | 34:42 | 49:43 | 58:43 | 1* | 0.90 | 0.91 | 1.34 | 1.66 | 4.09 |
| Polenta | 189:179 | 25:36 | | | | 1* | 0.65 | | | | 2.05 |
| Dried pulses | 188:182 | 26:33 | | | | 1* | 0.79 | | | | 0.56 |
| Fresh pulses | 58:54 | 35:45 | 27:31 | 38:46 | 56:39 | 1* | 0.83 | 0.85 | 0.87 | 1.63 | 2.30 |
| Green veg. | 41:43 | 48:44 | 36:42 | 31:43 | 58:43 | 1* | 1.47 | 1.09 | 0.81 | 1.68 | 0.55 |
| Root veget. | 25:45 | 42:44 | 52:40 | 61:43 | 34:43 | 1* | 1.75 | 2.55 | 2.45 | 1.40 | 1.57 |
| Fresh veget. | 31:45 | 45:41 | 42:43 | 45:43 | 51:45 | 1* | 1.55 | 1.55 | 1.57 | 1.88 | 2.80 |
| Tomatoes | 64:43 | 38:43 | 36:43 | 37:44 | 39:42 | 1* | 0.69 | 0.68 | 0.58 | 0.68 | 1.85 |
| Citrus Fruit | 37:37 | 56:49 | 28:39 | 54:47 | 39:43 | 1* | 1.15 | 0.72 | 1.25 | 1.00 | 0.00 |
| Apples | 74:43 | 45:46 | 30:45 | 28:43 | 37:38 | 1* | 0.59 | 0.34 | 0.30 | 0.45 | 9.85 |
| Peach | 51:43 | 25:35 | 30:30 | 65:65 | 43:42 | 1* | 0.71 | 0.81 | 0.87 | 0.88 | 0.02 |
| Bananas | 152:152 | 62:63 | | | | 1* | 1.03 | | | | 0.02 |
| Figs | 101:79 | 71:81 | 42:55 | | | 1* | 0.71 | 0.61 | | | 3.79 |
| Fresh fruit | 60:52 | 32:34 | 55:50 | 27:37 | 40:42 | 1* | 0.80 | 0.94 | 0.62 | 0.83 | 0.67 |
| Dried fruit | 181:187 | 33:28 | | | | 1* | 0.98 | | | | 0.00 |
| Nurs | 151:142 | 63:73 | | | | 1* | 0.88 | | | | 0.30 |
| Sugar | 52:43 | 49:45 | 35:40 | 33:44 | 45:43 | 1* | 1.01 | 0.60 | 0.54 | 0.89 | 1.16 |
| Chocolate | 156:152 | 58:63 | | | | 1* | 0.81 | | | | 0.86 |
| Candy | 176:181 | 38:34 | | | | 1* | 1.09 | | | | 0.10 |
| Marmalade | 111:116 | 51:46 | 52:53 | | | 1* | 1.06 | 0.94 | | | 0.02 |
| Puddings | 139:153 | 75:62 | | | | 1* | 1.33 | | | | 1.76 |
| Ice-cream | 131:127 | 52:57 | 31:31 | | | 1* | 0.82 | 1.02 | | | 0.04 |
| Dairy cream | 138:187 | 31:28 | | | | 1* | 1.32 | | | | 0.91 |
| Butter | 33:43 | 42:44 | 25:42 | 61:43 | 53:43 | 1* | 1.32 | 0.72 | 1.92 | 1.65 | 3.87 |
| Margarine | 164:180 | 50:35 | | | | 1* | 1.66 | | | | 3.97 |
| Olive oil | 49:43 | 48:43 | 42:44 | 44:42 | 31:43 | 1* | 0.94 | 0.83 | 0.94 | 0.70 | 0.81 |
| Sunflower oil | 141:156 | 19:27 | 57:32 | | | 1* | 0.69 | 1.63 | | | 2.30 |
| Seed oils | 84:82 | 54:51 | 41:39 | 35:43 | | 1* | 0.97 | 1.03 | 0.90 | | 0.05 |
| Juice | 173:172 | 41:43 | | | | 1* | 0.90 | | | | 0.16 |
| Soft drinks | 179:175 | 35:40 | | | | 1* | 0.76 | | | | 1.04 |
| Beer | 194:184 | 20:30 | | | | 1* | 0.66 | | | | 1.53 |
| Wine | 51:57 | 30:35 | 34:42 | 50:41 | 49:40 | 1* | 1.08 | 0.97 | 1.55 | 1.47 | 2.62 |
| Alcohol | 163:160 | 51:55 | | | | 1* | 0.92 | | | | 0.11 |
| Coffee | 54:43 | 52:47 | 26:39 | 39:46 | | 1* | 0.82 | 0.50 | 0.60 | 0.87 | 0.69 |
| Tea | 150:132 | 28:40 | 36:43 | | | 1* | 0.64 | 0.61 | | | 4.19 |
| Liver | 68:58 | 43:50 | 68:55 | 35:52 | | 1* | 0.65 | 0.94 | 0.49 | | 3.04 |
| Carrots | 51:44 | 25:42 | 45:43 | 60:43 | 33:43 | 1* | 0.44 | 0.91 | 1.17 | 0.60 | 0.01 |
| Cabeage | 92:87 | 42:47 | 40:38 | 40:43 | | 1* | 0.83 | 1.15 | 0.90 | | 0.00 |
| Broccoli | 150:147 | 44:37 | 20:31 | | | 1* | 1.35 | 0.78 | | | 0.03 |
| Turnip | 157:131 | 33:42 | 24:42 | | | 1* | 0.76 | 0.53 | | | 4.55 |
| Spinach | 47:57 | 63:57 | 51:53 | 53:48 | | 1* | 1.40 | 1.38 | 1.44 | | 1.26 |
| Melons | 54:44 | 46:42 | 43:43 | 30:43 | 41:43 | 1* | 0.86 | 0.91 | 0.59 | 0.82 | 1.00 |
| Apricot | 48:43 | 39:43 | 39:43 | 37:36 | 51:50 | 1* | 0.73 | 0.78 | 0.83 | 0.82 | 0.16 |

* Reference category ; (£) Adjusted for age, age at menarche, age at first birth, menopausal status, education, Quetelet index.

Giblets, potatoes, apples, tea, butter and margarine shown in table 2 were included in a single risk score, computed by summing up the levels of the positively associated foods (potatoes, butter, margarine) and substracting those of the thre major inversely associated ones (tea, apple, giblets) (Trichopoulos, 1985; La Vecchia, 1987).

For the sake of simplicity in computing the score, intake was considered as divided into three categories (1 = low intake, 2 = average intake, 3 = high intake) with exclusion of margarine and giglets (1 = no intake, 3 = intake).

Tab. 2: Food items significantly related with breast cancer risk. Case-control study of breast cancer. Milan, 1985-1986.

| Food items | Odds Ratios estimate (tertile) (*)(+) 1 | 2 | 3 |
|---|---|---|---|
| Sausages | 1£ | 1.09 (0.62-1.92) | 1.53 (0.88-2.62) |
| Giblets | 1£ | 0.52 (0.33-0.81) | |
| Potatoes | 1£ | 0.85 (0.49-1.48) | 1.55 (1.02-2.51) |
| Apples | 1 £ | 0.47 (0.27-0.80) | 0.46 (0.27-0.79) |
| Butter | 1£ | 1.10 (0.62-1.94) | 1.64 (1.02-2.66) |
| Margarine | 1£ | 1.92 (1.12-3.33) | |
| Tea | 1 £ | 0.59 (0.32-1.10) | 0.48 (0.27-0.85) |
| Turnip | 1£ | 0.84 (0.46-1.53) | 0.54 (0.28-1.02) |

£ Reference category
(*) Adjusted for age, age at menarche, age at first birth, menopausal status, education and Quetelet index by multiple logistic regression.
(+) Giblets and margarine with dichotomic code (1= no intake, 3=intake).

The obtained score, whose values range between -6 (subjects with level 3 for tea, apples and giblets and level 1 for potatoes, butter and margarine), and +6 (subjects with level 1 for tea, apples and giblets and level 3 for potatoes, butter and margarine) was subdivided into 3 levels.

The corresponding multivariate odds ratios, reported in Table 3, indicate that subjects with low consumption of apples, tea and giblets and high intake of potatoes, butter and margarine have a risk of breast cancer 4.06 times that of people reporting high intake of apples, tea and giblets and low intake of potatoes, butter and margarine (95% CI=2.2-7.4).

Tab. 3: Relation of breast cancer risk to a combined score of apples, tea, margarine, giblets, potatoes and butter. case-control study of breast cancer, Milan,1985-1986.

| score | N°cases:N°controls | Odds Ratios estimate (*) |
|---|---|---|
| - 2 | 28:57 | 1 ($) |
| (-1,0,1) | 97:108 | 2.10 (1.20-3.69) |
| 2 | 89:50 | 4.06 (2.21-7.45) |
| $X^2_1$ (trend) | | 21.89 P < 0.001 |

($) Reference category.
(*) Adjusted for age, age at first birth, age at menarche, menopausal status, education, Quetelet index by multiple logistic regression.

## 5.2 Exploratory analysis with use of MCA and PCA

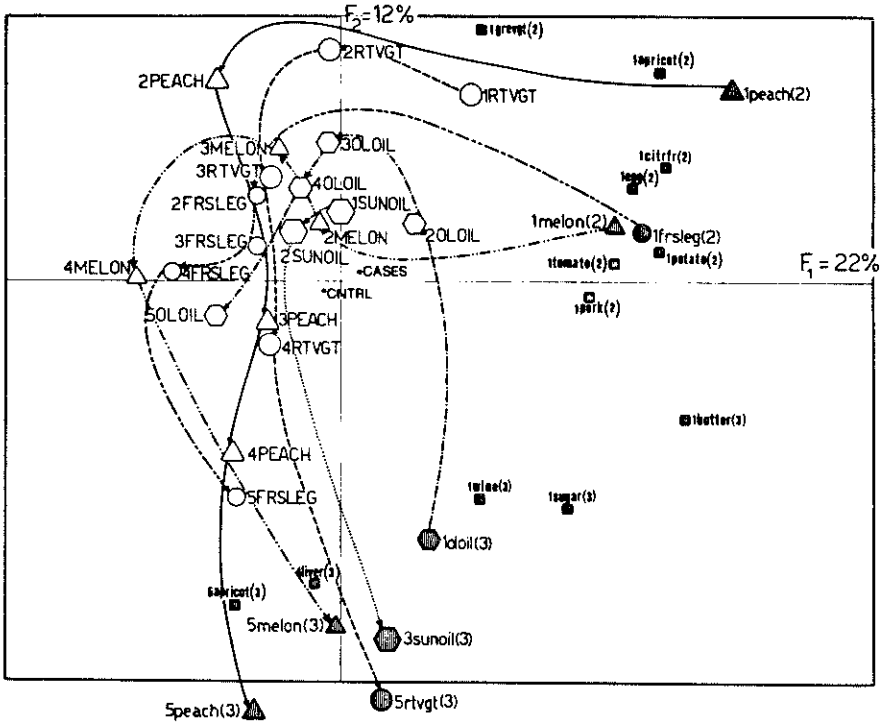
- Multiple correspondence analysis -

On the set of 64 variables the MCA analysis led to the identification of 12 factorial axes (variability explained 63%), which were utilized in successive cluster analyses.

The first factorial axis (horizontal), is characterized, as it was expected, by the counterposition of low intakes (on the positive part of axis) and high intakes (on the negative part). The second one is characterized by different consumption of many types of vegetables and vegetable fats.

As an example in Fig.1, the first factorial plane and the food items which contribute more to explain the first two axes, (root vegetables, fresh legumes, peaches, melon, olive oil, sunflower oil) are reported.

**Fig.1: Multiple correspondence analysis on breast cancer data (Milan, Italy, 1985-1986). Graphical display, on the first factorial plane, of the categories of some food items, including those which mostly characterize the second and third clusters.**




**Legend of abbreviated names given to the food items represented in the figure.**

| Food Items abbrev.(*) | name | symbol | absolute contribution to F1 | F2 |
|---|---|---|---|---|
| peach | peach | △ | 4.5 | 6.9 |
| melon | melon | △ | 2.8 | 3.3 |
| sunoil | sun flower oil | ⬡ | 0.1 | 3.2 |
| oloil | olive oil | ⬡ | 0.7 | 2.7 |
| rtvgt | root vegetable | ○ | 0.8 | 7.9 |
| frsleg | fresh legume | ○ | 3.7 | 1.5 |
| citrfr | citrus fruit | □ | 4.0 | 1.6 |
| grevgt | green vegetable | □ | 0.8 | 3.6 |

(*) The first number before each abbreviation indicates the category of the variable. Variables categories written in small letters are those characterising clusters. whose number is in brackets. indicates that the relative category charaterizes a cluster.

The number of subjects in the identified clusters, the estimated relative risks and the main features of each of them are reported in Tab 4. Table 4 also presents food items and their intake level which mostly characterize the 5 identified clusters.

Tab. 4: Multiple correspondence analysis on breast cancer data, Milan, 1985-1986. Description in terms of dietary habit of 5 identified clusters.

| Clusters | N° Cases | N° Controls | OR(*) | Dietary characteristics |
|---|---|---|---|---|
| 1 | 57 | 79 | 0.67 | Medium calorie intake. Medium consumption of butter, figs,citrus fruits. High consumption of tomatoes, eggs and beef. |
| 2 | 46 | 51 | 0.83 | Very low calorie intake. No consumption of pork, potatoes,peach, eggs and low consumption of green vegetables. |
| 3 | 34 | 30 | 0.87 | High consumption of fruit and vegetables (peach,melons,apple carrots and green vegetables). Low wine, butter, nuts and sugar consumption. |
| 4 | 55 | 40 | 1.35 | High calorie intake. Very high consumption of pasta, sugar, potatoes, butter, bread, and full fat cheese. No consumption of brown crackers, giblets and carrots. |
| 5 | 22 | 15 | 1.37 | Low calorie intake. No consumption of green and fresh vegetables, giblets, tomatoes and fresh fruits. Very low consumption of fruit and vegetables.High salami and other sausages consumption. |

(*) Adjusted for age, age at first birth, age at menarche, menopausal status education and Quetelet index by multiple logistic regression.

Some food items characterising the second and the third cluster can be seen on the first factorial plane in Fig 1. They are displayed with sketched symbols. On the upper right, the foods marked with (2) characterize 97 subjects (46 cases and 51 controls) with low calorie intake. The estimated odds ratios in

this group was 0.83. In the lower part of the figure the foods marked with (3) concern 64 women (34 cases and 30 controls) with elevated vegetable and low wine and sugar consumption. The corresponding relative risk was 0.87.

- Principal Component Analysis -

Also in this case the 64 selected variables are used. The first 12 principal axes were identified (variability explained 58%).

In Table 5 the list of variables which provide a major contribution to the explanation of the axes is given. As an example only the first six factorial axes are reported.

Tab. 5: Principal component analysis on breast cancer data,Milan, 1985-1986. Distribution of 215 cases of breast cancer and 214 hospital controls according to quintile distribution of values along the first 6 components. Food items with higher positive and negative correlation are reported.

| Component Number | | Quintile 1 | 2 | 3 | 4 | 5 | $X^2$(*) | Sign of correlation coefficient + | - |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ca | 52 | 34 | 41 | 47 | 40 | 13.91 | brown crack. | fresh fru. |
| | co | 34 | 51 | 45 | 39 | 46 | | tea | apricot |
| | | | | | | | | wine | puddings |
| 2 | ca | 50 | 42 | 40 | 46 | 36 | 7.64 | fresh fruit | butter |
| | co | 35 | 44 | 46 | 41 | 49 | | peach | bread |
| | | | | | | | | apple | sugar |
| 3 | ca | 47 | 46 | 44 | 44 | 33 | 3.39 | melon | carrot |
| | co | 37 | 41 | 42 | 42 | 53 | | tomato | root veget. |
| | | | | | | | | broccoli | rice |
| 4 | ca | 35 | 45 | 42 | 44 | 48 | 3.33 | pastry | green veg. |
| | co | 51 | 40 | 44 | 43 | 37 | | milk | olive oil |
| | | | | | | | | juice | citrus fr. |
| 5 | ca | 41 | 46 | 38 | 43 | 46 | 1.35 | liver | apple |
| | co | 44 | 40 | 48 | 44 | 39 | | giblets | fig |
| | | | | | | | | juice | half fat cheese |
| 6 | ca | 40 | 37 | 47 | 35 | 55 | 13.64 | chicken | juice |
| | co | 45 | 49 | 39 | 51 | 31 | | margarine | olive oil |
| | | | | | | | | candy | marmalade |

(*) $X^2$ for linear trend adjusted for age, age at first birth, age at the menarche, menopausal status, education, Quetelet index.

In this analysis the first axis can be interpreted as an axis of the size of the phenomenon under investigation. The

value of each subject on the coordinate of the first axis is, in fact, proportional to a total intake of the subject itself.

The other axes are more strongly characterized by the counterposition of some foods. The sign of each principal component is arbitrary, and hence the sign of the correlation coefficients of foods listed in table 5 is arbitrary, too. For instance, the second axis is strongly characterized by the counterposition of fresh fruit, peach and apples versus bread, butter and sugar.

Similar interpretations can be given for other axes.

A further suggestion for utilizing PCA in the present context is given by the fact that principal axes can be thought as complex orthogonal dietary variables.

Table 5 shows a partition of cases and controls into five classes defined according to the quintile of the new variables identified by the first 6 principal axes. The axes which most contribute to discriminate cases from controls are the 1st, 2nd and 6th which, in terms of food items are characterized by fresh fruits, tea, brown crackers, apricot, peach, butter, bread, chicken, margarine, juice and olive oil.

The cluster procedure led to the identify groups including 138, 85, 81, 85, and 42 subjects respectively. Table 6 gives the coordinates of the centres of gravity along the principal axes and the estimated odds ratios for each identified cluster.

Tab. 6: Principal component analysis on breast cancer data,Milan, 1985-1986. Coordinates of the centres of gravity along th e principal axes and the estimated relative risk for each identified clusters.

| Clusters | N° Cases | N° Cont. | OR(*) | Coordinates of the centres of gravity along the first 6 axes 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 65 | 1.08 | 0.219 | 0.068 | -0.024 | 0.035 | -0.023 | -0.012 |
| 2 | 50 | 35 | 1.39 | 0.025 | -0.181 | -0.061 | 0.058 | -0.005 | 0.095 |
| 3 | 36 | 45 | 0.85 | -0.067 | -0.131 | 0.054 | -0.170 | 0.007 | -0.028 |
| 4 | 35 | 50 | 0.69 | -0.158 | 0.203 | -0.047 | -0.045 | 0.067 | -0.056 |
| 5 | 22 | 20 | 1.14 | -0.310 | -0.013 | 0.193 | 0.190 | -0.064 | 0.015 |

(*) Adjusted for age, age at first birth, age at menarche, menopausal status, education and Quetelet index by multiple logistic regression. Reference category is the whole set of cases and controls.

It should be noted that the definition of clusters and their subsequent evaluation in terms of risk provides an independent information as compared with the information contained in Table 5. In this case, the principal axes 1,2 and 4 are those explaining most of the cluster inertia. When some axes provide the most efficient partition between cases and controls, together with the clearest distinction between dietary habits, stronger support can be given to the association between dietary habits and disease. In the example, the observed agreement is particularly satisfactory.

The last two tables help one to try to characterize various clusters in terms of dietary habits. For instance cluster 2 (OR=1.39) is chiefly characterized by the second principal axis, with reference to low intake of fresh fruit, peach and apples and high consumption of butter, bread and sugar. The cluster 4 (OR=0.69) is characterized by a low calorie intake and high consumption of fresh fruit.

It is worthnoting that PCA, providing components build up by the counterposition of food items, may suggest hypotheses concerning interactions between various foods.

## 6. Conclusions

In the exploratory analysis of dietary data epidemiologists have to cope with a great number of highly correlated variables. Defining strategies suitable for simplifing the analysis of dietary data and for leading to meaningful results is thus relevant.

The joint application of multivariate methods, such as those suggested here, appears to be suitable for pursuing the aims. In fact MCA and PCA allow one to identify orthogonal axes which can be interpreted in terms of dietary habits. Furthermore the interpretation of clusters subsequently picked up is made easier by this approach (Stellman, 1986).

The exploratory nature of these methods and the structure of the axes characterized by the counterposition of the consumption of specific food items permit, sometimes, the formulation of very complex hypotheses on the relationship between diet and disease.

Findings emerging from this example, i.e. the discrimination of dietary patterns with different risks, even if not statistically significant, represents a challenge for researches in this field to improve the suggested strategy.

It must be emphasised that, though the effect of diet on breast cancer is still debatable, some of the findings revealed during this tentative analysis agree with some recently suggested hypotheses (Rafian and Boin, 1987).

In particular the results of MCA (Tab. 4) seem to suggest that the risk of breast cancer is positively associated with a low consumption of fruit and vegetable and/or a high calorie intake. Using PCA (Tab. 5) an estimate of OR=1.39 is obtained for the second cluster which is characterized along the second axis by a strong negative value of the coordinates of the gravity center. This, as shown in Tab. 5 means that the cluster consists

of subjects with a high consumption of butter, bread and sugar and a low consumption of fresh fruit.

Procedures of analyis currently used nowadays, as those presented in tables 1-3, resorting to a high number of univariate analysed often lead to pick up effects due to single items even when they are negligible.

In the example discussed in previous section, the results obtained with PCA and MCA seem support the hypothesis that the 6 food items (giblets, apples, potatoes, tea, butter, margarine) selected by multiple logistic regression can be interpreted more in terms of indicators of a particular diet than of actual risk/protection factors for breast cancer.

The use of the structural analysis techniques appears to help in understanding the possible links between diet and specific pathologies when, as in this example, some of the principal components are highly associated with the risk of developing the disease under study. If cases and controls are differently arranged along some of the axes, the suggestion of an association between diet and cancer appears fortified. Furthermore the principal component can explain the variability existing between subjects in terms of dietary habits.

Since the aim of this paper is to discuss an approach for the analysis of dietary data and to verify its main peculiarities we did not tackle some methodological problems which require further research. Just to mention a few of them, discussion is still open about the definition of the most suitable metrics and aggregation criteria for cluster analysis in order to study the dietary data.

The need of comparing the results obtained with different multivariate methods, will lead to a deeper comprehension of their prons and cons.

In conclusion, we belive that the study of the relationship between diet and health is certainly interesting for the epidemiologist, but even more for the statistician given the stimulating methodological problems involved.

## ACKNOLEDGEMENTS

This work was conducted within the framework of the CNR (Italian National Research Council) Applied Projects "Oncology" Grant N. 84.00666.44 (Director Prof. E. Marubini).

The contribution of the Italian Association for Research on Cancer, is gratefully acknowledged.

REFERENCES


Benzecri J.P. (1973), L'analyse des donnes. Tome 2: L'analyse des correspondences. Dunod, Paris.

Block G. (1982), "Rewiew of Validation of Dietary Assessment Methods", Am.J.Epidemiol., 115, pp. 492-505.

Breslow N., Day N.E. (1980), "Statistical Methods in Cancer Research". IARC, Sci. Publ. 32.

Burke B.S. (1947), "The Dietary History as a Tool in Research".J. Am. Diet. Ass., 23, pp. 1041-1046.

Byers T.,Marshall J.,Anthony E., Fiedler R., Zielezny M. (1987), "The Reliability of Dietary History from the Distant Past", Am. J. Epidemiol., 125, pp. 999-1011.

Diday E. (1979), Optimisation en Classification Automatique. Tome 1, pag. 11-27, INRIA, Le Chesnay.

Fidanza F., Liquori G., Mancini F. (1974), Lineamenti di Nutrizione Umana. Napoli: Idelsom.

Gerber M., Cavallo F., Marubini E., Richardson S., Barberi A., Capitelli E., Costa A., Crasres De Paulet A., Crastes De Paulet P., Decarli A., Pastorino U., Pujol H. (1988), "Liposoluble vitamins and lipid parameters in breast cancer. A joint study in northern Italy and souhern France. International Journal of cancer., 42, pag. 489-494.

Jolliffe I.T. (1986) Principal Component Analysis. Series in Statistics. Springer Verlag. New York.

Lauro N.C., Decarli A. (1982)," Correspondence Analysis and Log Linear Models in Multiway Contingency Table Study: Some Remarks on Experimental data". Metron, XL, pp. 213-234.

Lebart L., Morineau A. (1982), SPAD. Systeme Portable pour l'Analyse des Donnes. Vol I. CESIA, Paris.

La Vecchia C., Negri E., Decarli A., D'Avanzo B., Franceschi S. (1987), "A Case Control Study of Diet and Gastric Cancer in Northern Italy". Int. J. Cancer, 40, pp. 484-489.

Lyon J.L., Gardner J.W., West D.W., Mahoney A.M. (1983), "Methodological Issues in Epidemiological Studies of Diet and Cancer". Cancer Research, 43 (suppl.), pp. 2392-2396.

MacGee D., Reed D., Yano K. (1984), "The Results of Logistic Analyses when the variables are Highly Correlated: an Empirical Example using Diet and CHD Incidence".J.Chron.Dis., 37, pp. 712-719.

Marubini E., Decarli A., Costa A., Mazzoleni C., Andreoli C., Barbieri A., Capitelli E., Carlucci M., Cavallo F., Monferroni N., Pastorino U., Salvini S. (1988)," The relationship of dietary intake and serum levels of retinol and beta-carotene with breast cancer: results from a case-control study",Cancer, 1, pp. 173-179.

Paul A.A.,Southgate D.A.T. (1979), MacCance and Widdowson's the composition of foods. London: Her Mayesty's Stationery Office.

Payne C.D., (ed.). (1985), The GLIM System. Relase 3.77. Generalised Linear Interactive Modeling Manual. NAG, Oxford.

Rafian T.E., Boin C.J. (1987), "Diet in the Etiology of Breast Cancer". Epidemiologic Reviews, 9, pp. 120-145.

Stellman, S.D. (1986), Chairman's Remarks. Workshop on the Selection. Follow-up and Analysis in Prospective Studies. National Cancer Institute, Monograph Nx67, pp. 145-147.

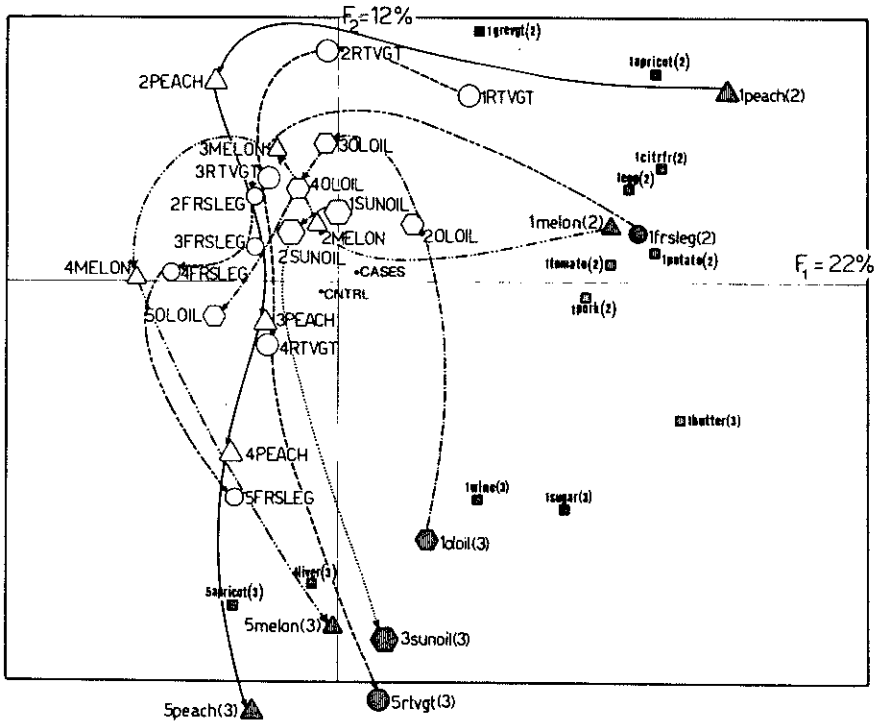
Rasanen L., Pietinen P. (1986), "Keynote Address: The Roleof Diet and Nutrition in Cancer". Cancer, 58 (suppl.), pp. 1791- 1975.

Trichopoulos D., Ouranos G., Day N.E., Tzonou A., Manouson O., Papadimitriou C.H., Trichopoulous A. (1985), "Diet and Cancer of the Stomach: a case-control study in Greece". Int.J.Cancer, 36, pp. 291-297.

Wahrendorf J. (1986), "The Changing Face of Cancer Epidemiology". Statistics in Medicine, 5, pp. 547-553.

Willett W., Stampfer M.J. (1986), "Total Energy Intake: Implications for Epidemiologic Analyses". Am.J.Epidemiol., 124, pp. 17-27.

**Fig.1: Multiple correspondence analysis on breast cancer data (Milan, Italy, 1985-1986). Graphical display, on the first factorial plane, of the categories of some food items, including those which mostly characterize the second and third clusters.**




Legend of abbreviated names given to the food items represented in the figure.

| Food Items abbrev.(*) | name | symbol | absolute contribution to F1 | F2 |
|---|---|---|---|---|
| peach | peach | △ | 4.5 | 6.9 |
| melon | melon | △ | 2.8 | 3.3 |
| sunoil | sun flower oil | ⬡ | 0.1 | 3.2 |
| oloil | olive oil | ⬡ | 0.7 | 2.7 |
| rtvgt | root vegetable | ○ | 0.8 | 7.9 |
| frsleg | fresh legume | ○ | 3.7 | 1.5 |
| citrfr | citrus fruit | □ | 4.0 | 1.6 |
| grevgt | green vegetable | □ | 0.8 | 3.6 |

(*) The first number before each abbreviation indicates the category of the variable. Variables categories written in small letters are those characterising clusters, whose number is in brackets. indicates that the relative category charaterizes a cluster.

# EXPLORATORY ANALYSIS OF DIETARY DATA DERIVING FROM A CASE CONTROL STUDY

Ferraroni M. Istituto di Statistica Medica e Biometria
Via Venezian,1 - 20133 Milano -Italy -Tel 02/238908

Decarli A. Istituto di Statistica Medica e Biometria and
Istituto Nazionale Tumori, Milano - Italy

Marubini E. Istituto di Statistica Medica e Biometria and
Istituto Nazionale Tumori, Milano - Italy

SUMMARY

The utilization of multivariate techniques (principal component and multiple correspondence analysis) is considered in order to determine, in exploratory analyses of case-control data, the dietary patterns associated with risk of, or protection against, a defined disease. The application of the techniques on data from a case-control study of breast cancer (214 cases and 215 controls) recently conducted at the Istituto Nazionale dei Tumori of Milan is compared to results obtainable by using methods such as multiple logistic regression. It is thereby emphasized how this new approach is capable of pointing out more complex associations among variables.

The need for a refinement of some methodological aspects is stressed, such as a study of aggregational criteria and metrics appropriate for the analysis of these kinds of data.

KEY WORDS : Dietary Data, Cluster Analysis, Correspondence Analysis, Breast Cancer.