# 13 Illustrative Benchmark Analyses

## 13.1 Introduction

**Raymond Bisdorff**
*CRP-GL, Luxembourg*

The development of the SODAS software based on symbolic data analysis was extensively described in the previous chapters of this book. It was accompanied by a series of benchmark activities involving some official statistical institutes throughout Europe. Partners in these benchmark activities were the National Statistical Institute (INE) of Portugal, the Instituto Vasco de Estadistica Euskal (EUSTAT) from Spain, the Office For National Statistics (ONS) from the United Kingdom, the Inspection Générale de la Sécurité Sociale (IGSS) from Luxembourg and marginally the University of Athens[1].

The principal goal of these benchmark activities was to demonstrate the usefulness of symbolic data analysis for practical statistical exploitation and analysis of official statistical data.

This chapter aims to report briefly on these activities by presenting some significant insights into practical results obtained by the benchmark partners in using the SODAS software package as described in chapter 14 below.

Our editorial criteria for compiling the illustrative examples of this chapter are as follows: – First, we aim at readability of the examples for general statisticians not necessarily working in a national official statistical institute. – A second framing aspect was externally imposed upon us by the SODAS software package available at that time in the sense that all statistical results discussed were obtained by a practical application of version 1.031 of the SODAS workbench which had been released at that time. The chapter is divided into three parts:

- First, we present the results obtained in the context of social security statistics where professional careers of retired working persons from Luxembourg are analysed with the help of the SODAS software.
- Second, we illustrate some results obtained from a common exploitation and analysis of the labour force survey provided by EUSTAT and INE. This example shows the usefulness of symbolic data analysis for combining

[1]The essential contribution of the Greek partner, under the scientific direction of Prof. H. Papageorgiou, is confined to an enhancement of the development of the DB2SO method. The aim was to generate and handle complex mother-daughter variables (see Chapter 5 for a more detailed description of this scientific contribution).

statistical results from different statistical institutes as is common in the European Union at the level of EUROSTAT.

- Finally, a last example concerns processing of ONS Census data.

## 13.2 Professional Careers of Retired Working Persons

**Raymond Bisdorff**
*CRP-GL, Luxembourg*

### 13.2.1 Basic Statistical Data Matrix

The statistical data set we propose describes exhaustively all completed (40 years) professional careers of persons having worked in Luxembourg and retiring within the year 1991. In the terminology of Chapter 2, our set $\Omega^1$ of individuals contains 1223 professional careers extracted from the administrative records of the Social Security Office in Luxembourg. Each individual career is described by the following 85 classic statistical variables: – birth year (1926 - 1936); – gender (F(0) - M(1)); – pension fund (workers (0), employees (1), liberal professions (2), farmers (3)); – monthly pension; – activity sector (17 main NACE rev. 1 sectors[2]); – yearly salaries from 1991 to 1952 (40 observations); – and finally, yearly health care allowances from 1991 to 1952 (40 observations). The description of all variables is summarized in *Table* 13.1.

| variable $Y_j$ | Range $\mathcal{Y}_j$ |
|---|---|
| GENDER | {0 (female), 1 (male)} |
| BIRTHYEAR | $\{1926, \ldots, 1936\}$ |
| PENSION FUND | {0 (workers), 1 (employees), 2 (liberal professions), 3 (farmers)} |
| NACE SECTOR | $\{A, C, D, E, F, G, H, I, J, K, L, N, ND, O, P\}$ |
| MONTHLY PENSION | $[0, 23184] \subset \mathbb{R}_+$ |
| SALARY$t$, $t = 91, \ldots, 52$ | $[0, 450] \subset \mathbb{R}_+$ |
| ALLOW$t$, $t = 91, \ldots, 52$ | $[0, 250] \subset \mathbb{R}_+$ |

**Table 13.1:** *Statistical description of the professional careers*

Gender (GENDER), pension fund (FUND) and NACE rev. 1 (NACE) are qualitative variables of *nominal* type (see Section 2.3). The great majority of our

[2]NACE revision 1 from 1990 is the official nomenclature for economic activities in use in the European Union from 1990 on. The 17 sector aggregation we use here is the usual one used in official publications about the Luxembourg economy.

individuals (92.2%) are male persons. The distribution among the four possible pension funds is the following: workers (50.8%), employees (33.4%), liberal professions (6.9%) and farmers (8.8%). The distribution among the 17 economic sectors presented in NACE rev. 1 nomenclature is shown in *Table* 13.2.

Half of our persons worked in the manufacturing industry (51%), which was in fact the main labour force reservoir for working persons in Luxembourg within the period under consideration. The second and third most important sectors are "Trade, recover and repair services" (6.7%) and "General government services" (5.0%), respectivley. Notice the high rate of missing values. The social security administration commonly allocates a default activity sector (ND) to private (home) employers so that we may conclude that the residual class (ND) contains mainly domestic services.

| *Sector* | *Activity* | *Size* | *%* |
|---|---|---|---|
| A | Agriculture, forestry and fishery | 3 | 0.2 |
| C | Extracting industry | 1 | 0.1 |
| D | Manufacturing industry | 624 | 51.0 |
| E | Power and water industry | 5 | 0.4 |
| F | Building and Construction | 26 | 2.1 |
| G | Trade, recover and repair services | 82 | 6.7 |
| H | Lodging and catering services | 3 | 0.2 |
| I | Transport and communication services | 16 | 1.3 |
| J | Credit and insurance services | 33 | 2.7 |
| K | Estate agencies, renting and other services | 9 | 0.7 |
| L | General government services | 61 | 5.0 |
| N | Health and social care services | 13 | 1.1 |
| O | Other non-market services | 7 | 0.6 |
| P | Domestic services | 5 | 0.4 |
| ND | missing value (not defined) | 335 | 27.4 |

**Table 13.2:** *Range of NACE variable with distribution of individuals*

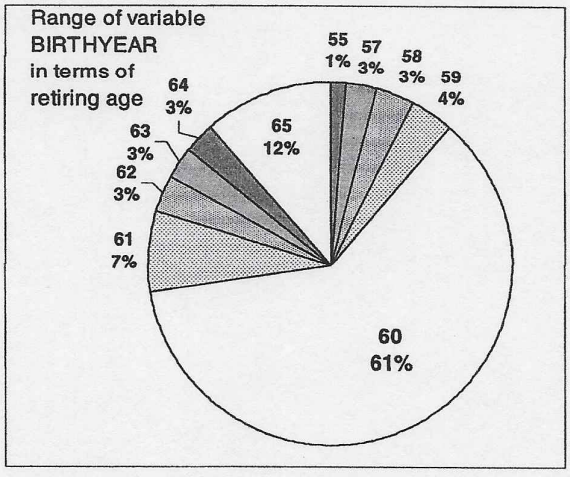


**Figure 13.2.1:** *Distribution of individuals in terms of retiring age*

The birth year is a *qualitative* variable BIRTHYEAR of *ordinal* type (see Section 2.3) with range {1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1936}. The ordinal character of this variable implicitly carries the age of the persons in question. Indeed, the youngest persons (birth year 1936) just reached the age of 55 when retiring in 1991, the oldest persons (birth year 1926) reached in 1991 the age of 65 in 1991, whereas those born in 1931 retired around the age of 60. One may notice in Figure 13.2.1 that the distribution with respect to the age of the retired persons in 1991 is rather unbalanced in favour of the years 1931 and 1926. A majority of persons (61%) request their retirement precisely at the age of 60, the first possible legal age of retirement with full pension. However, a significant proportion (12%) of the population stays in workforce activity until the age of 65, the uppermost age limit for official retirement of working persons. We will see below that these ages configure different retirement strategies with respect to the health situations of the persons.

Monthly pension (PENSION), yearly salaries (SALARY$t$, $t = 91, \ldots, 52$) as well as yearly health care allowances from 1952 to 1991 (ALLOW$t$, $t = 91, \ldots, 52$) are all *quantitative* variables of a *continuous* type (see Section 2.3). The monthly pension is represented in constant Luxembourgish francs expressed in base 100 in 1948 and is written as $\mathrm{LF}_{1948}$. All salaries and health care allowances are expressed in unit $10^3\mathrm{LF}_{1948}$, i.e., thousands of constant $\mathrm{LF}_{1948}$ and are confined to a common positive real range $[0, 450] \subset \mathbb{R}_+$.

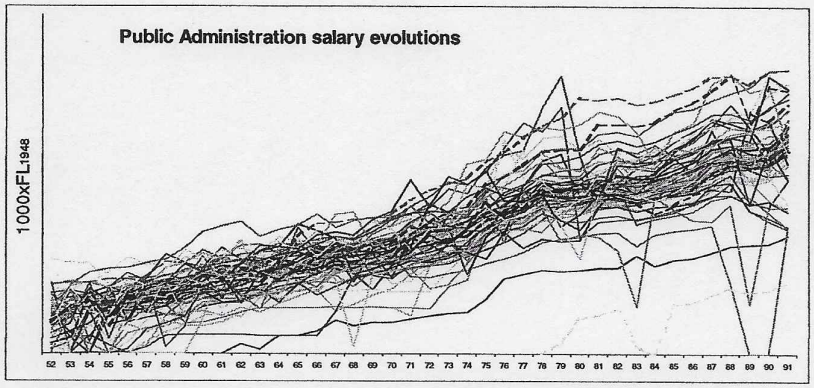


**Figure 13.2.2:** *Sample of salary evolutions from 1952 to 1991*

Following the terminology of Chapter 2, we can now compile all 85 variables in an 85-dimensional column vector variable $X$ = (GENDER, BIRTHYEAR, FUND, NACE, PENSION, SALARY$t$, ALLOW$t$)$'$ with $t = 91, \ldots, 52$, which takes values in the Cartesian product $\mathfrak{X} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{85}$ of the corresponding domains (see Table 13.1). For each individual career $k \in \Omega^1$ we obtain an 85-dimensional column vector $x_k = (x_1, \ldots, x_{85})'$ with $k = 1, \ldots, 1223$. By putting

all 1223 observations together, we obtain an example of a classical data matrix $\underset{\sim}{X}^1 = (x_{kj})_{1223\times 85}$ underlying all the following illustrative analyses. For our practical work, this data matrix was entered as a Microsoft Access table of corresponding design and content.

A quick glance (see for instance Figure 13.2.2) at these $1223 \times 2$ time series reveals that not all information is relevant. Indeed, the general economic and institutional contexts of the professional careers induce strong correlations among successive yearly salaries. Furthermore, health care allowances only occur in case of illnesses so that positive amounts of yearly health care allowances remain as rare events for most persons at the beginning of their working career. For these reasons we want to choose an appropriate small subset of years capturing a sufficient amount of the total variance over all these 40 years. To detect these most relevant variables, we will use the divisive clustering method DIV (see Section 11.2) provided in the SODAS software package.

## 13.2.2 Divisive Clustering of Professional Careers

### 13.2.2.1 Determination of important years in the professional careers

For this first analysis, we transform our basic classic data matrix $\underset{\sim}{X}^1$ into a first-order symbolic data array $\underline{X}^1$ (see Section 3.5) with the help of the SODAS-DB2SO method (see Chapter 5). The original set $\Omega^1$ of $n = 1223$ observations becomes a set $E^1$ of $N = 1223$ first-order *symbolic* data units and the original classic variables are canonically transformed into corresponding *symbolic* variables with identical range. Thus each *symbolic* data unit $u \in E^1$ is described by a symbolic data vector $\xi_u = x_u$ and all these data vectors may again be gathered in a symbolic data matrix $\underline{X}^1 = (\xi_{uj})^1_{1223\times 85}$.

To obtain an idea about the most relevant yearly salaries and health care allowances for discriminating between professional careers, we decided to use a decision or segmentation tree approach. To do so, we submit our previous first-order symbolic data matrix $\underline{X}^1$ to the SODAS-DIV method (see Section 11.2) providing a divisive clustering on symbolic data. This method basically consists in clustering the global set $E^1$ of symbolic data units by successively bipartitioning one of the clusters obtained at the previous step, starting with the universe $E^1$ as an initial unique cluster. The choice of the cluster to be split at a given step is based on a minimum-within-classes variance criterion for the partition of the set $E^1$ obtained in this way. The user has to provide the maximal number of clusters he wants to obtain eventually. This tree segmenting method works either on qualitative or on quantitative symbolic data.

```
THE CLUSTERING TREE :
---------------------
- the number noted at each node indicates
  the order of the divisions
- Ng <-> yes and Nd <-> no

                +----  Cluster 1 (Ng=120)
                !
            !----4- [SALARY73 <= 38.571449]
            !   !
            !   +----  Cluster 5 (Nd=52)
            !
       !----2- [ALLOW91 <= 74.242449]
       !    !
       !    +----  Cluster 3 (Nd=125)
       !
  !----1- [SALARY85 <= 116.415001]
       !
       !         +----  Cluster 2 (Ng=174)
       !         !
       !    !----6- [SALARY71 <= 117.911999]
       !    !    !
       !    !    !    +----  Cluster 7 (Ng=297)
       !    !    !    !
       !    !    !---7- [SALARY89 <= 181.783501]
       !    !         !
       !    !         +----  Cluster 8 (Nd=91)
       !    !
       !----3- [SALARY82 <= 208.027504]
            !
            !         +----  Cluster 4 (Ng=29)
            !         !
            !    !---8- [SALARY71 <= 103.574997]
            !    !   !
            !    !   +----  Cluster 9 (Nd=115)
            !    !
            !----5- [SALARY75 <= 196.376999]
                 !
                 !    +----  Cluster 6 (Ng=214)
                 !    !
                 !---9- [SALARY58 <= 147.545998]
                      !
                      +----  Cluster 10 (Nd=6)
```

**Figure 13.2.3:** *The clustering of professional careers resulting from the program* SODAS DIV

For our purpose, we have chosen all 40 yearly salaries, i.e., SALARY$t$ for $t = 91, \ldots, 52$ as well as all health care allowances ALLOW$t$ for $t = 91, \ldots, 52$. All these symbolic variables appear as classical continuous real-valued variables and the method applies a classical Euclidean distance function to assess distances between symbolic units. In order to get a reasonably rich subset of splitting variables, we asked for up to 10 clusters.

Figure 13.2.3 above shows the text output[3] provided by the SODAS-DIV method.

[3]Given the rather limited resources available in the development project of the SODAS software, it was not possible to implement a graphical representation of the resulting tree. Such a graphical tool will surely represent an interesting further enhancement of the SODAS software package.

The main splitting variable appears to be SALARY85, then ALLOW91, i.e., health care allowances from the last working year. In a third stage we obtain SALARY82, then SALARY73, SALARY75 and SALARY71. In the 7th position we obtain SALARY89, i.e., very recent salaries, and finally SALARY58, i.e., a very early salary in the professional career. We may conclude, following the advice of the segmentation tree, that we should restrict our attention to the salaries from the following seven years: 1989, 1985, 1982, 1975, 1973, 1971 and 1958. Furthermore, only the health care allowances from the last working year 1991 seem to be of importance in discriminating the professional careers under review.

Non-trivial as it is, this result shows:

1. that salary for the early years, except 1958, are of low interest for discrimination;
2. that the year 1971, when the official weekly working time was lowered in Luxembourg from 48 to 44 hours and moreover all hourly salaries were increased by 9.1%, has a clear discriminating effect on the salary declarations;
3. that the year 1973, when the social minimum salary was equalized for workers and employees in Luxembourg, has also a discriminating effect on careers;
4. that the salaries in 1975, after the major economic (industrial) crisis of the period under review, discriminate the population;
5. that for the last part of a career, apparently only the salaries from the years 1982, 1985 and 1989 seem to be of importance.

The result is of great practical usefulness since a classical approach for selecting a subset of variables would have suggested a uniform distribution of the years to consider, such as, for instance, every fifth year.

In order to evaluate the semantic outcome of the result we now turn our attention to the actual clusters which are proposed by the DIV segmentation tree. Indeed, the resulting 10 clusters may be considered as symbolic objects for which we may develop a corresponding second-level symbolic data analysis. This will be the subject of the next section.

#### 13.2.2.2 Emerging clusters of professional careers

Each of the ten clusters can be associated with a specific decision rule involving a subset of the relevant yearly salaries and/or the health care allowances from 1991. Let us look, for instance, at the decision rules for cluster '1' and cluster '3':

$$C_1 = \text{cluster '1'}: \quad ([\text{SALARY85} \leq 116.415001] \wedge [\text{ALLOW91} \leq 74.242449] \wedge [\text{SALARY73} \leq 38.571449]).$$

$$C_3 = \text{cluster '3'}: \quad ([\text{SALARY85} \leq 116.415001] \wedge [\text{ALLOW91} > 74.242449]).$$

On the one hand, cluster $C_1$ gathers careers where the 1985 salary is below $116,415\,\mathrm{FL}_{1948}$, the 1991 health care allowance is below $74,250\,\mathrm{FL}_{1948}$ and the 1973 salary is below $38,571\,\mathrm{FL}_{1948}$. On the other hand, cluster $C_3$ gathers careers where the 1985 salary is again below $116,415\,\mathrm{FL}_{1948}$, but this time the health care allowance in 1991 is above $74,242\,\mathrm{FL}_{1948}$. Thus, the 10 clusters appear as *symbolic objects of assertion type* (see Section 4.3) with the subset of individual professional careers obtained in each cluster as the associated extension. We now construct these second-level symbolic data units describing the extension of the 10 clusters within the 1223 individual professional careers from 1991, i.e. our original basic data matrix $\underset{\sim}{X}^1$.

The assertion, or decision rule, describing each cluster allows us to enlarge our basic classic data matrix $\underset{\sim}{X}^1$ by a new single-valued nominal variable CLUSTER with domain $\{1,\ldots,10\}$ such that for each of the 1223 careers $k$, CLUSTER($k$) specifies its corresponding cluster label. In this way, we restrict our relevant data matrix to the yearly salaries from 1989, 1985, 1982, 1975, 1973, 1971 and 1958, as well as to the health care allowances solely from the last working year, i.e., 1991. We thus obtain a new reduced classical data matrix denoted below as $\underset{\sim}{X}^2$ with 1223 rows and only 8 columns.

Importing matrix $\underset{\sim}{X}^2$ again via the SODAS-DB2SO importation method (see Chapter 5) allows us to construct a second set $E^2 = \{C_1,\ldots,C_{10}\}$ of 10 second-order symbolic data units called assertion objects (or symbolic objects of assertion type, see Section 4.3) describing each of the 10 clusters with $\Omega^2 = \{C_1,\ldots,C_{10}\}$.

The resulting symbolic variables together with their symbolic ranges are shown below:

```
variable GENDER, nominal {"0", "1"}, multiple,mode=probabilist;
variable FUND, nominal {"0", "1", "2", "3"}, multiple, mode=probabilist;
variable NACE,
  nominal {"A", "C", "D", "E", "F", "G", "H", "I", "J", "K", "L", "N", "ND",
"O", "P"},
  multiple, mode=probabilist;
variable BIRTHDYEAR,
  nominal {"1926", "1927", "1928", "1929", "1930", "1931", "1932", "1933",
"1934", "1936"},
  multiple, mode=probabilist;
variable PENSION,  real [0:23184]   interval;
variable SALARY89, real [0:428.651] interval;
variable SALARY85, real [0:409.821] interval;
variable SALARY82, real [0:404.979] interval;
variable SALARY75, real [0:281.57]  interval;
variable SALARY73, real [0:269.089] interval;
variable SALARY71, real [0:262.606] interval;
variable SALARY58, real [0:298.839] interval;
variable ALLOW91,  real [0:235.514] interval;
```

Due to the cluster aggregation, our original nominal variables GENDER, FUND, NACE and BIRTHYEAR become *symbolic* nominal variables of a *multiple-valued probabilistic* type (see Section 3.4) where for each cluster, the 'value' of the variable is defined by an empirical frequency distribution on its finite nominal range. The 8 original quantitative variables are transformed into (real-valued) interval variables (see Section 3.4) with corresponding range limits. For each cluster

$u \in \Omega^2$ out of the ten clusters described in the decision tree, we obtain a new (column) data vector $x_u = (\xi_1, \ldots, \xi_{13})'$ with $u = 1, \ldots, 10$. Combining these symbolic data vectors $x_u$, we obtain a second-order symbolic data matrix denoted in the sequel as $\underline{X}^2 = (\xi_{uj})^2_{10\times 13}$.

To illustrate the constructed second-order symbolic data units, let us have a look at the description file for $C_1$ as provided by the output of the DB2SO method:

```
os "1"(120) =
   [GENDER    = {"0"(0.55), "1"(0.45)}]
  ^[BIRTHYEAR = {"1932"(0.00833333), "1927"(0.0666667), "1926"(0.525),
                 "1929"(0.075), "1930"(0.108333), "1928"(0.075),
                 "1931"(0.133333),"1933"(0.00833333)}]
  ^[FUND = {"1"(0.0333333), "3"(0.866667), "0"(0.0666667), "2"(0.0333333)}]
  ^[NACE = {"A"(0.00833333), "D"(0.00833333), "N"(0.00833333),
            "ND"(0.916667), "O"(0.0166667), "L"(0.00833333),
            "P"(0.025), "G"(0.00833333)}]
  ^[PENSION  = [1139:6994]]
  ^[SALARY89 = [0:203.028]]
  ^[SALARY85 = [0:87.4357]]
  ^[SALARY82 = [0:138.361]]
  ^[SALARY75 = [0:54.6471]]
  ^[SALARY73 = [0:37.9166]]
  ^[SALARY71 = [0:57.3486]]
  ^[SALARY58 = [0:56.194]]
  ^[ALLOW91  = [0:6.03215]]
```

The symbolic data unit os "1" (120) represents the description of Cluster $C_1$ with respect to the classical data matrix $\underline{X}^2$. It has a size of 120 individual careers and is characterized by the assertion above, where we may notice:

1. the exceptionally high proportion of female persons (55%) compared to a global proportion of 8%;
2. the large proportion of persons of age 65 at the time of their retirement;
3. the fact that the great majority of persons are supported by the agricultural pension fund (FUND = "3").

All this information can be simultaneously visualized by the SODAS - Symbolic Object Editor (see Chapter 7) under the form of *Zoom Stars* (see Figure 13.2.4).

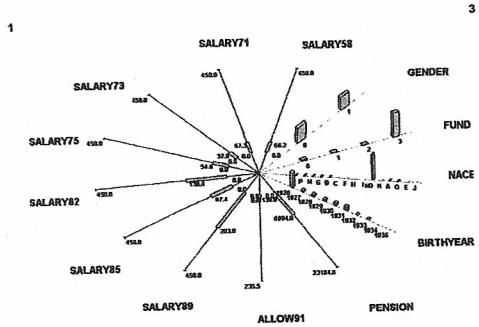


**Figure 13.2.4:** *Cluster $C_1$*

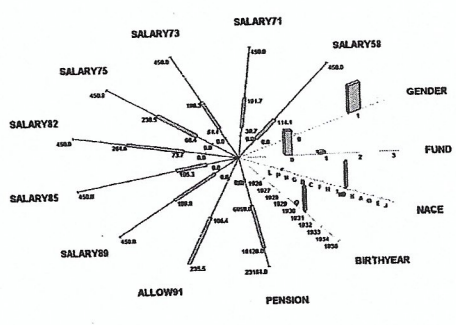


**Figure 13.2.5:** *Cluster $C_3$*

Such a visualization is useful, e.g., when comparing the description of cluster $C_1$ to the description of cluster $C_3$ given by:

```
os "3"(125) =
   [GENDER    = {"1"(1)}]
  ^[BIRTHYEAR = {"1930"(0.08), "1931"(0.92)}]
  ^[FUND      = {"1"(0.088), "0"(0.912)}]
  ^[NACE      = {"ND"(0.992), "P"(0.008)}]
  ^[PENSION   = [6959:18128]]
  ^[SALARY89  = [0:189.845]]
  ^[SALARY85  = [0:105.296]]
  ^[SALARY82  = [73.6666:264.56]]
  ^[SALARY75  = [66.3658:236.501]]
  ^[SALARY73  = [51.1063:198.345]]
  ^[SALARY71  = [39.6677:191.742]]
  ^[SALARY58  = [0:114.126]]
  ^[ALLOW91   = [106.45:235.514]]
```

Here, we see that $C_3$ contains 125 persons, all male, of which 92.1% are supported by the workers pension fund and who have retired almost all (91%) around the age of 60. The corresponding Zoom Star is shown in Figure 13.2.5. Comparing both stars, we notice that the 1991 health care allowances for cluster $C_3$ are very high compared to those of $C_1$. Moreover, salaries tend to be significantly lower for cluster $C_3$.

Let us conclude this illustration with a third cluster, namely cluster $C_6$:

```
os "6"(214) =
   [GENDER = {"1"(0.981308), "0"(0.0186916)}]
  ^[BIRTHYEAR = {"1927"(0.0233645), "1929"(0.0607477), "1926"(0.088785),
                 "1931"(0.649533), "1932"(0.0186916), "1930"(0.0794393),
                 "1936"(0.00934579), "1933"(0.0186916), "1928"(0.0373832),
                 "1934"(0.0140187)}]
  ^[FUND = {"0"(0.0420561), "2"(0.0046729), "1"(0.953271)}]
  ^[NACE = {"I"(0.00934579), "E"(0.0140187), "F"(0.0233645), "G"(0.0607477),
            "H"(0.0046729), "M"(0.00934579), "D"(0.719626), "J"(0.0981308),
            "O"(0.0140187), "ND"(0.0327103), "K"(0.0046729), "L"(0.00934579)}]
  ^[PENSION  = [14130:23184]]
  ^[SALARY89 = [167.629:428.651]]
  ^[SALARY85 = [152.094:409.821]]
  ^[SALARY82 = [209.484:404.979]]
  ^[SALARY75 = [196.465:281.57]]
  ^[SALARY73 = [121.126:269.089]]
  ^[SALARY71 = [119.405:253.677]]
  ^[SALARY58 = [0:133.303]]
  ^[ALLOW91  = [0:15.8113]]
```

Here, we have 214 employees, mostly (98.1%) male, mainly from the manufacturing industry 'D' (72%) as is evidenced by the corresponding Zoom Star (see Figure 13.2.6). Compared to both previous clusters, we observe here the highest salaries and very low health allowances in the last working year, a result which confirms the employee status of persons clustered in $C_3$.

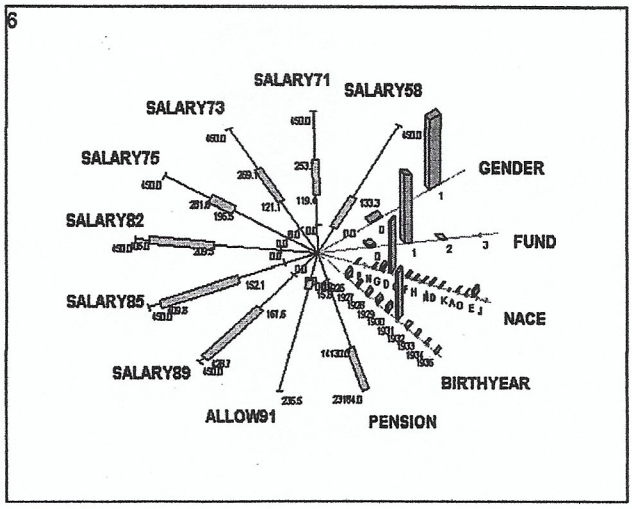


**Figure 13.2.6:** *Cluster 6 represented as a 3D zoom-star*

In order to assess the overall quality of our divisive clustering results, we will now control the classification results with the help of a classical principal component analysis.

### 13.2.2.3 Visualizing symbolic data units by a classical principal component analysis

We have submitted our second (restricted) classical data matrix $\underset{\sim}{X}^2$ to a principal component analysis (PCA)[4]. As active continuous variables to be considered in the PCA, we have chosen all seven yearly salaries from 1989, 1985, 1982, 1975, 1973, 1971 and 1958, as well as the 1991 health care allowances. First we calculate the corresponding correlation matrix $S$ (Table 13.3).

| | *variable* | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
|---|---|---|---|---|---|---|---|---|---|
| 1. | SALARY85 | 1.00 | | | | | | | |
| 2. | ALLOW91 | -0.62 | 1.00 | | | | | | |
| 3. | SALARY82 | 0.77 | -0.04 | 1.00 | | | | | |
| 4. | SALARY73 | 0.60 | 0.05 | 0.82 | 1.00 | | | | |
| 5. | SALARY75 | 0.58 | 0.07 | 0.83 | 0.93 | 1.00 | | | |
| 6. | SALARY71 | 0.53 | 0.09 | 0.77 | 0.94 | 0.90 | 1.00 | | |
| 7. | SALARY89 | 0.96 | -0.62 | 0.74 | 0.55 | 0.53 | 0.48 | 1.00 | |
| 8. | SALARY58 | 0.30 | 0.15 | 0.53 | 0.65 | 0.66 | 0.67 | 0.28 | 1.00 |

**Table 13.3:** *Correlation matrix $S$ for basic data matrix $X^2$*

[4]The computation was realized with the help of the SPAD version 3.5 software package.

We notice that the health care allowances for 1991 are negatively correlated to the three salaries considered in the eighties, i.e., the salaries for 1989, 1985 and 1982. Furthermore, the three salaries from the first half of the seventies, i.e., 1975, 1973 and 1971, appear as highly correlated to each other ($\geq 0.90$).

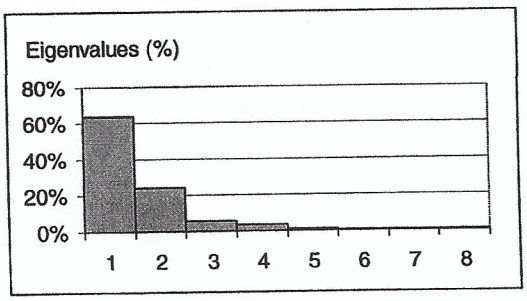


**Figure 13.2.7:** *Histogram of the CPA eigenvalues*

As we can see from Figure 13.2.7, the two first eigenvalues account for 87.5% of the total variance, and the quality of the representation of our individuals in the factorial plot by the two first principal axes may be observed in Figure 13.2.8.

The divisive clustering obtained in the previous section can be easily recognized along the first axis where we find the three clusters presented above, i.e. cluster $C_1$ completely on the left, cluster $C_3$ almost in the middle and cluster $C_6$ to the right. In fact, the left-to-right ordering of the 10 clusters matches the vertical order of the leaves of the SODAS DIV-segmentation tree in Fig. 13.2.4.

Furthermore, we can plot the complete salary evolutions of the careers collected, for instance, in clusters $C_3$ and $C_6$ (see Figures 13.2.9 and 13.2.10). First, we notice that each of these clusters indeed discriminates a certain subset of more or less homogeneous working careers. For instance, Cluster $C_3$ gathers mainly persons who had no yearly salaries after 1985 until their retirement in 1991. Cluster $C_6$ comprises mainly persons with a regular salary evolution at a comparatively high level.

Even if our curiosity for further exploration of these clusters certainly does not fade away at this point, we would like to stop this illustrative analysis here and turn our attention towards a different symbolic data analysis problem, namely discrimination of specific retiring strategies on the basis of the age of the persons under review.

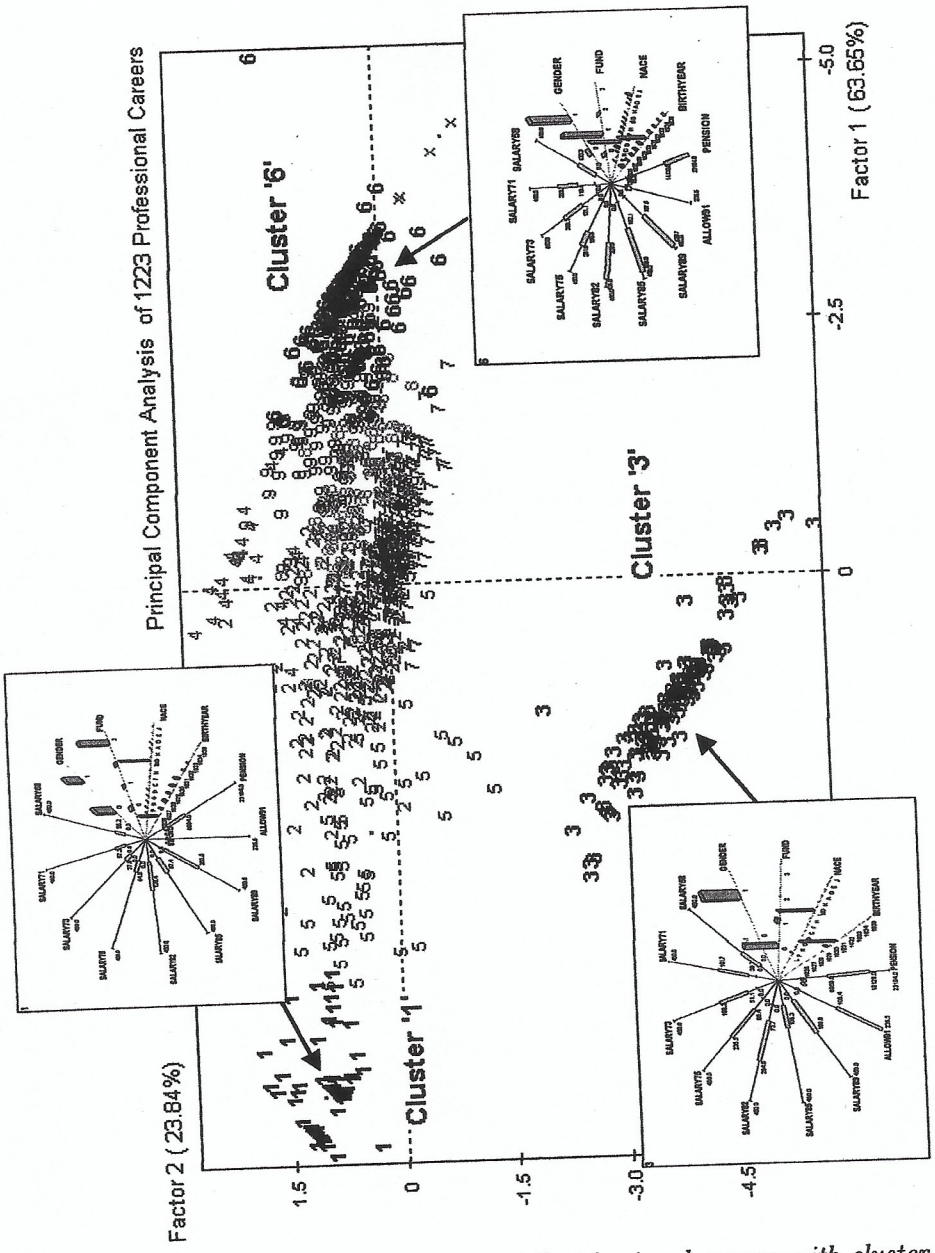


**Figure 13.2.8:** *Factorial plot of all 1223 professional careers with cluster illustration*

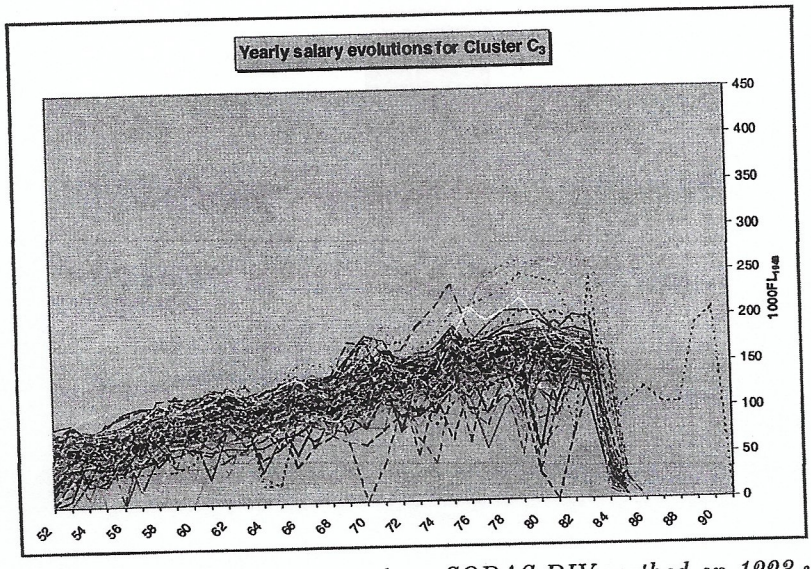


**Figure 13.2.9:** *Cluster $C_3$ resulting from SODAS-DIV method on 1223 working careers of persons retiring in 1991*

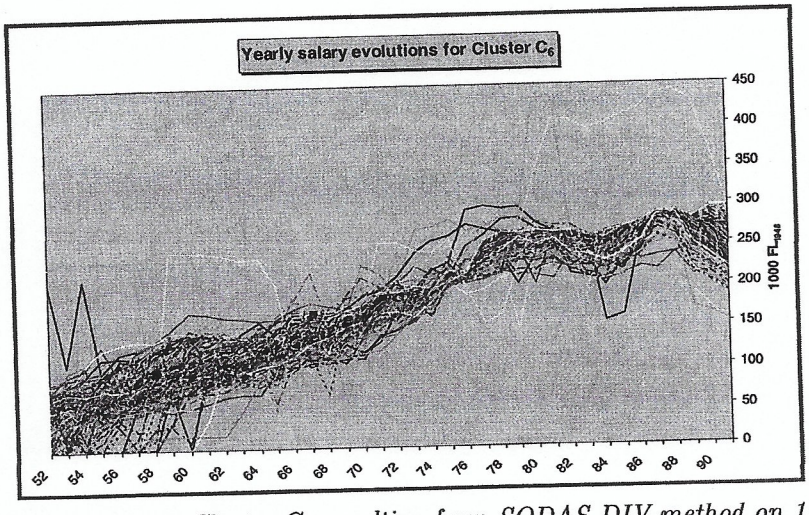


**Figure 13.2.10:** *Cluster $C_6$ resulting from SODAS-DIV method on 1223 working careers of persons retiring in 1991*

### 13.2.3 About the Discrimination of the Retiring Age from the Professional Careers

#### 13.2.3.1 Distinguishing three retiring scenarios

Turning our attention now to the actual age of the persons when they retire, we notice roughly speaking three possible retiring scenarios:

1. some persons (11%), in those cases where they can claim a full working career of 40 years, retire before the normal legal retiring age of 60;
2. most persons (68%) retire as soon as they reach the age of 60 or in the following year;
3. some others (21%) go on working possibly until the final age of 65.

We can now assign each individual working career to one of the three scenarios using the following three-partition of the range of the BIRTHYEAR variable: $S_1 = \{1936, 1934, 1933, 1932\}$, $S_2 = \{1931, 1930\}$ and $S_3 = \{1929, 1928, 1927, 1926\}$, where $S_1$ corresponds to the youngest persons retiring before the age of 60, $S_2$ corresponds to the normal retiring scenario at the age of 60 or 61 and $S_3$ corresponds to the eldest persons who continue to work despite a retirement opportunity. In this way, we enlarge our original classical data matrix $\underset{\sim}{X}^1$ by adding an *ordinal* variable SCENARIO with range $\{1, 2, 3\}$ corresponding to the three scenarios above such that $\forall u \in \Omega^1$:

$$\begin{aligned} \text{SCENARIO}(u) = 1 &\Leftrightarrow \text{BIRTHYEAR}(u) \in S_1, \\ \text{SCENARIO}(u) = 2 &\Leftrightarrow \text{BIRTHYEAR}(u) \in S_2, \\ \text{SCENARIO}(u) = 3 &\Leftrightarrow \text{BIRTHYEAR}(u) \in S_3. \end{aligned}$$

Projecting this new illustrative variable on to the factorial plot resulting from the previous PCA (see Figure 13.2.8) which takes into account only the seven most relevant years out of the complete career, unfortunately does not suggest any really significant clustering. We will therefore introduce a different subset of variables, apparently more relevant for discriminating our retiring scenarios namelythe very last years in the working career. Indeed, as can be seen in *Figure 13.2.9* showing the salary evolutions for cluster $C_3$, loss of salaries associated with high health care allowances occurring at the end of the working career point to health problems that should a priori give one of the most plausible reasons why persons were retiring as soon as possible.

#### 13.2.3.2 Assigning each working career to one of the three retiring scenarios

In order to illustrate these phenomena statistically, we will apply a *symbolic discriminant analysis* (see Chapter 10) to our data. The SODAS TREE method

(see Section 10.3 and Chapter 14), which we are going to use in this analysis, is a tree-growing method which renders a description of the three retiring scenarios in the form of a binary decision tree concerning the scenarios' descriptive variables, in this case salaries and health care allowances mainly from the end of the working career. Instead of working as before on first order symbolic data we consider this time basic symbolic data aggregates built from the Cartesian product SCENARIO × GENDER × FUND × NACE, i.e., formed with the help of our nominal illustrative variables.

Let us consider a new basic data matrix $\underset{\sim}{X}^3$ involving the following variables: SCENARIO × GENDER × FUND × NACE, SALARY$t$ and ALLOW$t$ for $t \in \{91, 90, 89, 88, 87, 86, 85, 82, 75\}$ where the first variable SCENARIO × GENDER × FUND × NACE is a compound nominal variable constructed from the original four nominal variables in such a way that for all $k \in \Omega^1$, SCENARIO × GENDER × FUND × NACE$(k) =$ "$s\_g\_f\_n$" where $s =$ SCENARIO$(k)$, $g =$ GENDER$(k)$, $f =$ FUND$(k)$ and $n =$ NACE$(k)$. For each individual career, $k \in \Omega^1$ we thus obtain a 19-dimensional column vector $x_k = (x_1, \ldots, x_{19})'$ with $k = 1, \ldots, 1223$. Gathering all 1223 observations together, we obtain a third classical data matrix $\underset{\sim}{X}^3 = (x_{kj})_{1223 \times 19}$.

Importing this matrix $\underset{\sim}{X}^3$ again via the SODAS-DB2SO importation method (see Chapter 4) into the SODAS software allows us to construct a third set $E^3$ of second level symbolic data units of Cartesian type (see Section 4.3). Theoretically we should have from three scenarios, two genders, four funds and 15 sectors, up to 360 symbolic data units, but only 86 of these render a non-null volume in terms of individual working careers gathered. All resulting 18 symbolic variables, SALARY$t$ and ALLOW$t$ for $t \in \{91, 90, 89, 88, 87, 86, 85, 82, 75\}$, are again of continuous interval type. We add, for each $u =$ "$s\_g\_f\_n$" $\in E^3$ of the 86 symbolic data units, a nominal single-valued symbolic variable SCENARIO such that SCENARIO$(u) = s$. Thus we obtain for each $u \in E^3$ a 19-dimensional column symbolic data vector $\xi_u$ and together we obtain as usual the symbolic data matrix $\underline{X}^3_{86 \times 19}$.

As the SODAS-TREE relies on the assumption that all symbolic descriptor variables of interval type in fact support a uniform distribution inside each interval range, we must, if possible, avoid over-generalization which occurs by considering all individual careers, even atypical ones. The SODAS-DB2SO method therefore provides an assertion reduction method (see Chapter 5) which specializes each of the previous 86 symbolic data units by rejecting atypical individual working careers according to a volume criterion measuring the generality index of each union of individuals. The DB2SO assertion reducing procedure tries to maximize this criterion under a minimum covering threshold constraint. Here we stay with the default minimum covering threshold of 80%. The resulting specialization of the assertions for each symbolic data unit can be graphically inspected (see Figure 13.2.11 for instance).

The reduced symbolic data matrix $\underline{X}^3_{86\times 19}$ now becomes input data to the SODAS-TREE method.

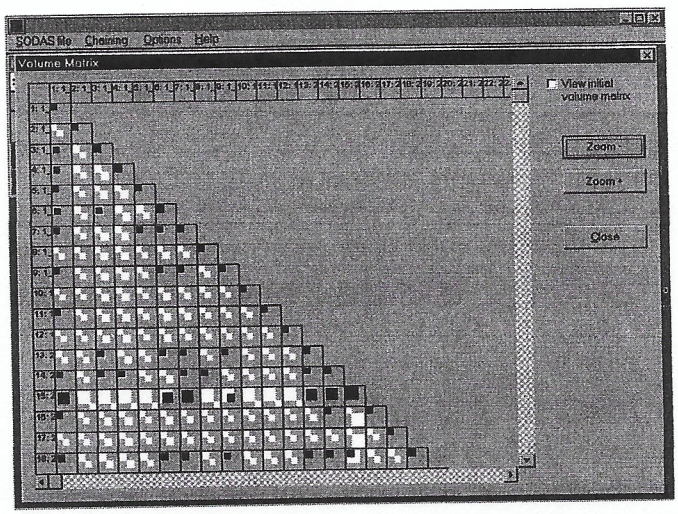


**Figure 13.2.11:** *SODAS-DB2SO specialization of symbolic data units illustrated by a volume matrix*

#### 13.2.3.3 Discriminating the retiring scenarios

The nominal variable SCENARIO with range $\{1, 2, 3\}$ is considered as prediction *class identifier* whereas all SALARY$t$ and ALLOW$t$ for $t \in \{$ 91, 90, 89, 88, 87, 86, 85, 82, 75$\}$ act as *predictor* variables. The quality of the discrimination of the decision tree obtained from the SODAS TREE method is described with the help of a confusion matrix and corresponding misclassification rates as shown in 13.4. The retiring scenarios $S_2$ and $S_3$, even if somewhat confounded with each other (up to 20%), are well discriminated against scenario $S_1$. Inversely, scenario $S_1$ is rather confused with scenario $S_2$ (25%) but not with scenario $S_3$.

| *scenarios* | $S_1$ | $S_2$ | $S_3$ | *Total* | *Error/Size* | *Frequency* |
|---|---|---|---|---|---|---|
| $S_1$ | 15 | 6 | 3 | 24 | 9/24 | 37.50 |
| $S_2$ | 0 | 24 | 6 | 30 | 6/30 | 20.00 |
| $S_3$ | 1 | 5 | 26 | 32 | 6/32 | 18.75 |
| *Total* | 16 | 35 | 35 | 86 | 21/86 | 24.42 |

**Table 13.4:** *Confusion matrix for retiring scenario prediction*

Through the computed decision tree, each one of the three possible retiring scenarios may now be associated with a disjunction of specific decision rules involving

a subset of the considered yearly salaries and/or the health care allowances. Let us outline for instance some of these decision rules describing scenario $S_1$:

$$\begin{aligned} S_1 = (\ &[\text{ALLOW91} > 6.30279] \wedge [\text{ALLOW89} \leq 5.50446]\ ) \\ \vee (\ &[\text{ALLOW91} > 6.30279] \wedge [\text{ALLOW89} > 5.50446] \\ &\wedge [\text{SALARY82} \leq 118.282007]\ ). \end{aligned}$$

We observe here that people retired before the age of 60 in the case they had a certain amount of health care allowances ($\text{ALLOW91} > 6,302\text{FL}_{1948}$) during their last working year and they had a rather low yearly salary in 1882 ($SALARY82 \leq 118,282\text{FL}_{1948}$). Perhaps some developing health problems convinced these persons to retire early, even before reaching the age of 60.

We can compare this result with two decision rules concerning retiring scenario $S_3$:

$$\begin{aligned} S_3 = (\ &[\text{ALLOW91} \leq 6.30279] \wedge [\text{ALLOW90} \leq 7.68887] \wedge \\ &[\text{ALLOW75} \leq 1.24844] \wedge [\text{ALLOW89} \leq 13.0329] \wedge \\ &[\text{SALARY86} \leq 73.162903] \wedge [\text{SALARY75} \leq 22.0166]\ ) \\ &\vee \\ (\ &[\text{ALLOW91} \leq 6.30279] \wedge [\text{ALLOW90} \leq 7.68887] \wedge \\ &[\text{ALLOW75} \leq 1.24844] \wedge [\text{ALLOW89} \leq 13.0329] \wedge \\ &[\text{SALARY86} > 73.162903] \wedge [\text{SALARY88} > 177.410995] \wedge \\ &[\text{SALARY91} > 266.53299] \wedge [\text{SALARY86} > 251.070999]\ ) \end{aligned}$$

This time we notice in both rules above that health care allowances are generally rather limited ($\text{ALLOW91} \leq 6,300\text{FL}_{1948}$, $\text{ALLOW90} \leq 7,688\text{FL}_{1948}$ and $\text{ALLOW89} \leq 13,032\text{FL}_{1948}$). Thus, no major health problems prevent these persons from staying active at work after the age of 60. However, two very different social situations:

- one with low salaries, and
- another with comparatively high salaries may motivate this choice.

In order to conclude our investigation, let us now visualize, as in the previous section, the three discriminated retiring scenarios with the help of a classical PCA of our basic complete data matrix $\underset{\sim}{X}^1$.

#### 13.2.3.4 Visualizing the retiring scenarios with a classical principal component analysis

We have again submitted our first classical data matrix $\underset{\sim}{X}^1$ to a principal component analysis (PCA)[5].

[5]The computation was realized with the help of the SPAD version 3.5 software package.

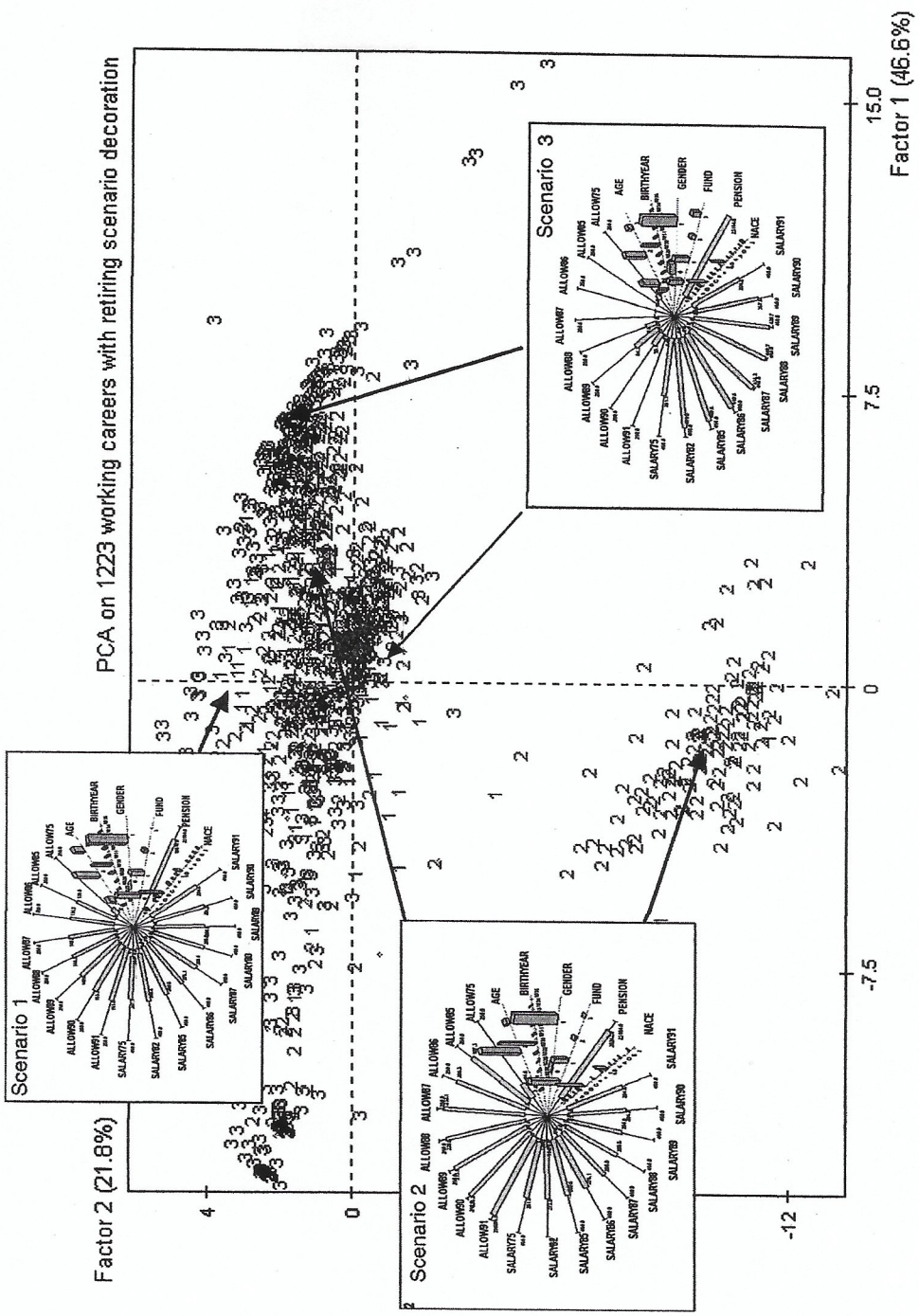


**Figure 13.2.12:** *Factorial plot of all 1223 professional careers with retiring scenario decoration*

As active continuous variables to be considered we have chosen the complete set of yearly salaries and health care allowances from 1952 to 1991. The corresponding factorial plot using the two first principal axes is shown in Figure 13.2.12. These two principal factors, with a weight of 46.6% and 21.8%, confirm the result of the previous PCA in the sense that again the $X$ axis discriminates low versus high salaries from left to right whereas the $Y$ axis discriminates between low and high health care allowances from top to bottom.

The retiring scenarios illustration clearly shows that persons with rather high health care allowances all retired as soon as possible (scenario $S_2$), whereas persons continuing to work after the age of 61, i.e., retiring scenario $S_3$, frequently appear at both ends of the salary spectrum. Finally, as expected, retiring scenario $S_1$ appears rather confused with scenario $S_2$, which confirms the confusion matrix and the comparatively higher misclassification rate for this scenario (see Table 13.4).

## 13.3 Comparing European Labour Force Survey Results from the Basque Country and Portugal

**Anjeles Iztueta, Patricia Calvo**

*EUSTAT - Inst. Vasco de Estadistica Euskal, Vitoria-Gasteiz, Spain*

### 13.3.1 The European Labour Force Survey Data

The benchmark data consists of statistical results from the quarterly Portuguese (INE) and Basque (EUSTAT) Labour Force Survey (LFS), a total of 56,049 records describing individual persons. Due to space limitations, only a subset of 17 important variables have been selected for illustration. These variables have been adapted by both institutes in order to harmonize them and to obtain the same structure before joining them in a common basic data set. The selected variables are given in Table 13.5.

The first sixteen variables are of qualitative nominal type (see Section 2.4) whereas the sampling weight variable ELEVA is of a quantitative continuous type (see Section 2.3).

Due to the complex structure of the LFS questionnaire, some of our variables are logically dependent on another (or others), like for instance variable BUSQ2