

Qüestió

Quaderns d'Estadística
i Investigació Operativa

Any 1994, volum 18, núm. 1
Segona època

Entitats patrocinadores:

Universitat de Barcelona
Universitat Politècnica de Catalunya
Institut d'Estadística de Catalunya



Generalitat de Catalunya
**Institut d'Estadística
de Catalunya**

Sumari

Editorial	3
-----------------	---

Articles originals

Estadística General i Investigació Operativa

Distance-based regression in prediction of solar flare activity	7
Anna Bartkowiak and Maria Jakimiec	

Notes on the bias of dissimilarity indices for incomplete data sets: the case of archaeological classification	39
Angela Montanari and Stefania Mignani	

Building a knowledge base for correspondence analysis	51
M^a Carmen Bravo Llatas	

The use of third-order moments in structural models	75
Erik Meijer and Ab Mooijaart	

Estadística Oficial

Estadística oficial i coordinació del sistema estadístic català	87
Josep A. Casco Robledo	

El Pla estadístic de Catalunya: la Llei 30/1991 i la planificació de l'activitat estadística pública	127
J. Oliveras i Prats	



EDITORIAL

Aquest és el segon número dedicat a publicar les comunicacions presentades a la Seventh International Conference on Multivariate Analysis (Barcelona, 21 al 24 de setembre de 1992), que van ésser sotmeses a **Qüestió** i que posteriorment han estat acceptades. En conseqüència i per raons d'espai, la Secció Docent (article docent i problemes) es posposa per al Volum 18, número 2.

Aproximadament la meitat de les comunicacions avaluades han estat acceptades, i són les que ara es publiquen plegades en aquests dos números. El Comitè Organitzador de la Sevent International Conference on Multivariate Analysis està molt agraït a T. Aluja, C. Arenas, J.M. Corcuera, C.M. Cuadras, H.L. Kiers, M. Nikulin, F. Oliva, J.M. Oller, V. Quesada, M.C. Ruiz de Villa, A. Satorra i a altres avaluadors anònims, per la seva tasca de revisió i crítica dels originals, que ens ha permès seleccionar els que són més adequats per a la publicació.

Pel que fa a les conferències invitades, aquestes han estat publicades independentment en forma de llibre: *Multivariate analysis: future directions 2* (C.M. Cuadras and C.R. Rao, editors), North-Holland, Amsterdam, 1993.

La part d'Estadística Oficial del present número de **Qüestió** enfoca en dos articles els conceptes bàsics de l'estadística oficial en general i les característiques específiques del sistema estadístic de Catalunya en particular. Aquest tema té un doble interès per a totes les persones que treballen en el camp de l'Estadística, des de la recerca teòrica fins a les aplicacions immediates. D'una banda constitueix una àrea de l'estadística que tots els ciutadans haurien de conèixer. D'altra banda comporta unes línies de recerca, desenvolupament i difusió de l'administració pública que tenen cada cop més pes en el progrés general de l'estadística.

En l'article "Estadística oficial i coordinació del sistema estadístic català", el senyor Josep A. Casco, subdirector general d'Assistència Tècnica Estadística, remarca algunes distincions bàsiques entre estadística pública i oficial, entre estadística oficial i resultats oficials i entre activitats estadístiques i administratives. Després presenta algunes característiques del sistema estadístic de Catalunya i els seus instruments de coordinació.

En l'article "El Pla estadístic de Catalunya: la Llei 30/1991 i la planificació de l'activitat estadística pública", el senyor Jordi Oliveras, director general de l'Institut d'Estadística de Catalunya descriu el procés d'elaboració del Pla i la filosofia de la Llei del pla, basada en la planificació estratègica i els criteris de decisió de preferència. El Pla és un instrument per donar forma a una voluntat política i establir unes pautes per dur-la a terme.

Estadística General i Investigació Operativa

DISTANCE-BASED REGRESSION IN PREDICTION OF SOLAR FLARE ACTIVITY

ANNA BARTKOWIAK* and MARIA JAKIMIEC†

Short-term prediction of solar flare activity using multiple regression methods was considered. The variables describing active regions the given day were used to predict the flare activity on the next day. Two groups of observational data covering the years 1988 and 1989 were dealt with. Some variants of the distance-based regression as proposed by Cuadras and Arenas (1990) appeared to be superior to the ordinary least squares method —by describing more accurately the data sets under consideration.

Key words: Distance based regression, Euclidean distance, L1-norm distance, Gower distance, Predictions

1. INTRODUCTION

Usually, a set of predictor variables $X_1, \dots, X_p$ which describe the features of solar active region on a given day is the basis in the procedure of short-term, i.e. one day ahead, predictions of solar flare activity. Both the predictor and the predicted variables can be of various types: continuous or discrete, i.e. categorical.

*Anna Bartkowiak. Institute of Computer Science, University of Wrocław. Przesmyckiego 20, 51-151 Wrocław, Poland.

†Maria Jakimiec. Astronomical Institute, University of Wrocław. Kopernika 11, 51-622 Wrocław, Poland.

—Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.

—Acceptat l'octubre de 1993.

Usually, the prediction algorithms are constructed by the use of multivariate regression functions (see e.g. Jakimiec and Wasiucioneck, 1980, or Bartkowiak and Jakimiec, 1986), discriminant functions (see e.g. Hirman *et al.*, 1980) or logistic regression functions (see Vecchia *et al.*, 1980). However, the results obtained by these methods are not satisfactory, what can be seen, a.o. in the papers by Jakimiec and Bartkowiak (1989) and Bartkowiak and Jakimiec (1990). First, the linear regression model does not fit ideally the investigated solar data. In the prediction of solar flare activity we encounter the effect of asymmetry consisting in overestimation of low flare activity and in underestimation of strong flare activity. Second, the quality of the predictions, as measured by the multiple correlation coefficient, is not very high; e.g. the mentioned authors, using the classical LSE regression, got a correlation coefficient ≈ 0.50 . Therefore, it seems to be necessary to look for more sophisticated models and for more sophisticated statistical methods.

Distance-based regression, as proposed by Cuadras and Arenas (1990), (see also Fortiana, 1992), is a substantially new method. The approach is similar as in Cuadras (1989) and utilizes some concepts studied formerly by Cuadras (e.g. Cuadras, 1988). The analysis starts from a distance matrix between the considered data vectors characterizing the considered objects. In distance-based regression the distance matrix can be evaluated on the base of a mixed set of variables (e.g. some of them can be continuous, the other categorical) which is a great advancement as compared to the classical methods. A convenient method of evaluating distances for mixed type variables was proposed by Gower (1971).

The aim of this paper is the comparison of results obtained by the ordinary least squares (OLS) regression and by the distance-based regression (DBR) proposed by Cuadras and Arenas. We consider three distances : Euclidean, L1-norm and Gower's. The formulæ for evaluation of these distances are given in section 3.1 of the paper. To compare the results yielded by these methods we perform our analysis using continuous variables only. Cuadras and Arenas have proved that DBR evaluated from Euclidean distances based on p variables is equivalent to the ordinary least squares regression evaluated for the same p variables. Hence our analysis will compare, as a matter of fact, the OLS results with those yielded by the DBR using L1-norm and Gower's distances.

2. THE DATA

Our data comprise daily characteristics of sunspot groups as published in SGD (Solar Geophysical Data 1988, 1989). We have chosen for analysis those

sunspot groups which, according to Zirin *et al.* (1991), were BEARALERT regions in 1988 and 1989. So called BEARALERT's are issued by the Big Bear Observatory, on the base of solar maps. The alerts are announced only for those sunspot groups for which a strong flare activity is expected in a near future. However, it can happen that in such indicated regions no strong flaring would occur soon. For many sunspot groups, for which it is judged that the probability of the flare occurrence is much less than 0.01, the alert is not issued at all.

In this paper we analyse 10 characteristics of sunspot groups. Eight of them (denoted in the following as $X1, \dots, X8$) are taken as predictor variables and two (denoted as $Y1$ and $Y2$) as predicted variables. We make the predictions by constructing regression equations allowing to express the expected values of the predictor variables as linear functions of the predictors. In our problem the predictor variables (explanatory variables in the considered regression) describe the daily characteristics of a sunspot group observed the given day. The predicted variables (explained variables in the considered regression) characterize the flare activity in the given sunspot group the next day. So, the statement of the problem is a very classical one. Before starting the proper calculations we have noticed that the frequency distributions of some variables were very skew. So, to diminish their skewness, we have performed the logarithmic transformation $X = \log X$ for the variables $X2, X4, Y1$ and $Y2$ which have exhibited the highest coefficient of skewness.

The meaning of the variables $X1 - X8$ and $Y1 - Y2$ taken for our further consideration is as follows:

- $X1$ - McI - McIntosh class determined (see Hirman *et al*, 1980) as a product of three McIntosh's parameters of sunspot group.
- $X2$ - Area - Sunspot group area ($\log$).
- $X3$ - Cnt - The number of spots in the sunspot group.
- $X4$ - MvXX - The maximum value of solar flare X-ray flux ($\log$).
- $X5$ - NF - Total number of H_α flares.
- $X6$ - AvHF - The average hardness index of the faint X-ray flares.
- $X7$ - AvHA - The average hardness index obtained for all flares.
- $X8$ - THI - The total hardness index.
- $Y1$ - MvXY - The maximum value of solar flare X-ray flux on the next day ($\log$).
- $Y2$ - Fs - The total sum of the maximum values of solar flare X-ray fluxes observed the next day ($\log$).

The gathered data were subdivided into two sets: I — the data for the year 1988, and II — the data for the year 1989. The respective sample sizes are $n_I = 130$ and $n_{II} = 117$. These data sets comprise individual data vectors, each composed from $p = 8$ values of predictor variables and from 2 values of the predicted variables. The data are complete, i.e. there were no missing values.

In the following the individual data vectors will be also sometimes referred to as “data items” or simply “items” .

Table 1.

Averages and Standard Deviations for the variables $X1 - X8$ and $Y1 - Y2$ considered in data sets I and II

Variable	Data Set I n=130		Data Set II n=117	
	Average	Standard Deviation	Average	Standard Deviation
$X1 - McI$	75.14	40.38	97.27	31.76
$X2 - Area$	5.99	1.14	6.72	0.76
$X3 - Cnt$	29.38	21.89	45.84	21.00
$X4 - MvXX$	1.68	1.40	2.79	1.65
$X5 - NF$	7.36	5.32	8.13	4.55
$X6 - AvHF$	2.51	1.51	3.46	2.31
$X7 - AvHA$	4.09	2.85	5.89	3.48
$X8 - THI$	6.37	4.59	9.12	5.76
$Y1 - MvXY$	1.62	1.41	2.79	1.58
$Y2 - Fs$	2.35	1.56	3.56	1.45

The averages and standard deviations for the considered variables are shown in Table 1. One can see that the averages obtained for the data set II are higher than those for the data set I. For all variables, except the $X5=NF$, the differences are statistically significant. The photospheric variables ($X1, X2, X3$) and the number of flares ($X5$) reveal higher variances in 1988 than in 1989 year, while the variables characterizing flare activity ($X4, X6 - X8, Y1, Y2$) reveal higher variances in the 1989 than in the 1988 year. One can see that, although both data sets contain the BEARALERT sunspot groups only, in the year 1989 the flare activity was much stronger than in the year 1988. This reflects the non-stationarity of solar activity in the solar eleven-year cycle, which is evident even in the adjacent years 1988 and 1989. We can add that the peak of the 11-year cycle was observed in the year 1989.

To find out something about the similarity or dissimilarity on the covariance structure of the variables recorded in the two groups of data we have constructed biplots. They are shown in Fig. 1.

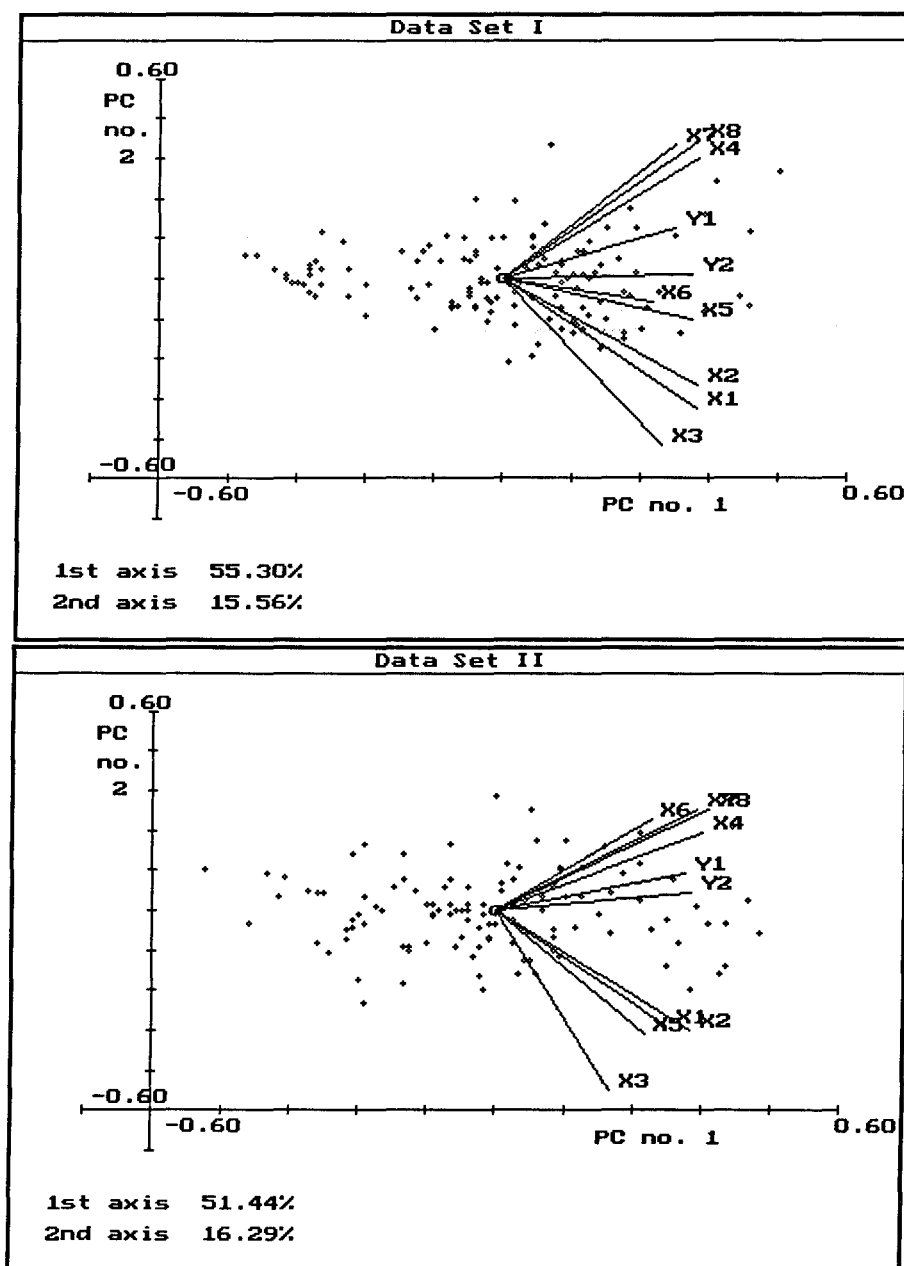


Figure 1.

Biplots established from data sets I (upper figure) and II (lower figure). Both biplots constructed from the variables X1-X8 and Y1-Y2.

The biplots were constructed according to the algorithm described by Jolliffe (1986), see also Krzanowski (1988). Speaking generally, we make first a singular value decomposition (s.v.d.) of the data matrix $\mathbf{X}$ of size $n \times p$ and of rank k by factorizing it into the form

$$(1) \quad \mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{A}^T,$$

where $\mathbf{U}$ of size $n \times k$ and $\mathbf{A}$ of size $p \times k$ are rank k matrices columnwise orthonormal, and $\mathbf{L}$ of size $k \times k$ is diagonal with nonnegative elements.

The factorization can be rewritten as

$$(2) \quad \mathbf{X} = \mathbf{G}\mathbf{H}^T,$$

with

$$(3) \quad \mathbf{G} = \mathbf{U}\mathbf{L}^\alpha \quad \text{and} \quad \mathbf{H}^T = \mathbf{L}^{1-\alpha}\mathbf{A}^T$$

In our construction we took as $\mathbf{X}$ in formula (1) the matrix $\tilde{\mathbf{X}} = (\tilde{x}_{ij})$ obtained from the original data matrix $\mathbf{X}$ by the following standardization:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_j \sqrt{n-1}}, \quad i = 1, \dots, n, \quad j = 1, \dots, p.$$

After such a standardization the crossproduct $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ appeared as the correlation matrix $\mathbf{R}$ of the variables under consideration. The s.v.d. of $\tilde{\mathbf{X}}$ was then obtained from properly rescaled eigenvectors and eigenvalues of the matrix $\mathbf{R}$. The final $\mathbf{G}\mathbf{H}^T$ decomposition was obtained by taking the constant α appearing in (3) as $\alpha = 1.0$. The biplot was then constructed by using the first two columns of $\mathbf{G}$ and the first two rows of $\mathbf{H}^T$ for graphing the points-items (corresponding to the rows of $\tilde{\mathbf{X}}$) and the points-variables (corresponding to the columns of $\tilde{\mathbf{X}}$), respectively. The points-items appear in our graphs as dots, and the points-variables as vectors emanating from the (0,0) point.

In the following we will be concerned only with the representation of points-variables.

The first two components taken for constructing the biplot do account for 70.9% and 67.7% of total inertia in set I and set II, respectively. These percentages are reasonably high —thus the representation of the interdependency structure hidden in the data should be reasonably good. Among others, the angles between the vectors representing the variables should reflect the correlations between the respective variables: vectors for variables whose correlation coefficient r is equal $r = 1$ should overlap, and vectors for variables with $r = 0$ should be perpendicular. Let us underline that this happens only when the interdependency structure is totally reproduced in the graph, i.e. the first two components reproduce in 100% the total inertia. In our graphs the first two

axes account only for $\approx 70\%$ of total inertia — therefore the relations among the variables are represented only in an approximate way.

Looking at the biplots shown in Fig.1 one can state that, in principle, the vectors representing the variables under consideration exhibit a much similar pattern.

All the variables are correlated positively. It is quite amazing that the vectors $X1, X2, X3$ —describing the photosphere— are located on one side of the bunch of the vectors-variables, and the vectors $X4, X7, X8$ —describing the flaring activity— on the other side of this bunch. This can be seen clearly in both plots.

The vectors $Y1$ and $Y2$ representing the predicted variables are located somehow in the middle of the bunch of the variables.

From this configuration one is tempted to infer that both groups of predictors, i.e. both the photospheric and the flare variables are necessary to make a good prediction.

The described structure is somehow blurred by the variables $X5$ and $X6$ which are located in the two constructed biplots in different positions with respect to the mentioned groups of variables. This happens mainly with the variable $X6$ whose position is quite different in the two biplots representing the data from the two years.

After stating the blurring role of the variable $X6$ we have constructed on the basis of the other nine variables new biplots (not shown in this paper). The above described patterns of location of the variables in the biplots remained the same.

Now let us return to the main topic of the paper, namely to the predictions of the variables $Y1$ and $Y2$ from the variables $X1 - X8$. Clearly both the photospheric ($X1, X2, X3$) and the flare ($X4, X7, X8$) variables should be included into the predicting equation; despite the fact that in the year 1989 (data set II) the variables $X1, X2, X3$ —as compared with the flare variables— will have less impact on the predictions (this may be inferred from the fact that in the year 1989 both $Y1$ and $Y2$ are closer to the “flare” variables and more distant to the photospheric variables).

A final conclusion after inspecting the two biplots in Fig. 1. —especially when dropping the variable $X6$ — could be, that despite of some minor dissimilarities in the location of the vectors representing the variables, their intercorrelation structure in both data sets appears to be the same, with minor departures only. The same seems to be true for the predicted variables $Y1$ and $Y2$. Howe-

ver, since the position of these variables is a little shifted in the two plots, and additionally the variable X_5 has changed its position in these plots, it would be overoptimistic to expect that the predicting equations in the two data sets (representing two adjacent years) will be the same. Nonetheless, predictions made in one set from the equation established in the other set should not be hopeless. In Section 4 we will try to find out how such predictions do work. An analysis of the interdependence structure when taking two additional explanatory variables is presented in a paper by Jakimiec and Bartkowiak (1994).

3. EVALUATING THE DISTANCE-BASED REGRESSION

We consider the distance based regression as proposed by Cuadras and Arenas (1990) and implemented in MULTICUA (Arenas *et al.*, 1991). In dependence of the chosen distance matrix their method allows to consider predictor variables also of mixed type. However our analysis will be based on continuous variables only because we want to compare the results obtained from the distance-based method to those obtained for the OLS method.

The distance-based analysis proceeds in the following steps:

- Step 1.** Taking into consideration the explanatory variables evaluate the distance matrix between the individuals taken for the analysis. Let this be the matrix $\mathbf{D}$ of size $n \times n$ (the matrix $\mathbf{D}$ is symmetric, so we need in fact only, say, the lower triangle of this matrix).
- Step 2.** Using the methods of Multidimensional Scaling (see, e.g., Mardia, Kent & Bibby, 1979) get an equivalent representation of the data items in an Euclidean space by evaluating the principal coordinates (PC's) with respect to the inner product matrix $\mathbf{B}$ derived from the distance matrix $\mathbf{D}$. The obtained PC's are mutually orthogonal.
- Step 3.** Evaluate the Pearsonian correlation coefficient of each PC with the predicted variable Y . Retain k principal coordinates yielding the largest squared correlation with Y . Since the principal coordinates are mutually orthogonal, the multiple determination coefficient of Y with the retained principal coordinates is simply the sum of the squared correlation coefficients evaluated between Y and the individual PC's.
- Step 4.** Perform an OLS regression of Y from the retained principal coordinates. If desired then evaluate the predicted values of Y for given data vectors $\mathbf{x}$ and also the residuals, i.e. the differences between the observed and the predicted values of Y .

The calculations for our analysis were done using a special program, REG-DIST, developed in the Institute of Computer Science, University of Wrocław.

Now we give some details on implementation of these steps.

3.1. Implementation of step 1. Choosing the distance matrix

Here the choice of the formula for the evaluation of the distance between two individuals is crucial. Say, we consider n individuals and p variables, the i -th individual described by the vector of features $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$. Say, we want to evaluate the squared distance $d^2(i, j)$ between the individuals *no. i* and *no. j*. We will consider in our further analysis the following (squared) distances:

Euclidean Distance (EUCL):

$$(4) \quad d_E^2(i, j) = \sum_{h=1}^p (x_{ih} - x_{jh})^2$$

L1-norm Distance (L1):

$$(5) \quad d_L^2(i, j) = \sum_{h=1}^p |x_{ih} - x_{jh}|$$

Gower's Distance (GOWER):

$$(6) \quad d_G^2(i, j) = \left(\sum_{h=1}^p |x_{ih} - x_{jh}| / G_h \right) / p$$

$$\text{where } G_h = \max_i(x_{ih}) - \min_i(x_{ih})$$

The EUCL and the L1 distances depend strongly on the units in which the subsequent variables are expressed. This would impose a dominance of some variables depending on their units. To avoid it we can normalize the observed values by dividing them by the respective standard deviations or ranges.

The Gower distance is invariant under linear transformations of units in which the considered variables are expressed and therefore it does not need any normalization.

Let $\mathbf{D}$ denote the matrix of squared distances: $\mathbf{D} = (d_A^2(i, j))$, where for A we may substitute E (Euclidean), L (L1-norm) or G (Gower's). It is known (see, e.g. Mardia *et al.*, 1979, or Cuadras, 1991)) that all the three distances introduced above yield distance matrices $\mathbf{D}$ which are Euclidean from the Multidimensional Scaling perspective, which means that there exists a configuration of points in some Euclidean space whose interpoint distances are given by $\mathbf{D}$; that is, for some k there exist points (row vectors) $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathbb{R}^k$ such that

$$d_{rs}^2 = (\mathbf{z}_r - \mathbf{z}_s)(\mathbf{z}_r - \mathbf{z}_s)^T.$$

3.2. Implementation of step 2. Evaluation of eigenvectors of the transformed distance matrix

Using Multidimensional Scaling methodology we first transform the considered (square) distance matrix $\mathbf{D}$ according to the formulæ:

$$\begin{aligned} \mathbf{A} &= (a_{ij}), \quad \text{with} \quad a_{ij} = -(d_{ij}^2)/2, \\ (7) \quad \mathbf{B} &= \mathbf{H}\mathbf{A}\mathbf{H}, \quad \text{with} \quad \mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^T. \end{aligned}$$

The matrix $\mathbf{B}$ will be called in the following the *centered inner product* matrix or simply the *inner product* matrix. Let us notice that the matrix $\mathbf{B}$ has dimensions $n \times n$ and is symmetric. It should be stressed that the matrix $\mathbf{B}$ itself is not a distance matrix.

It is known that the matrix $\mathbf{B}$ derived from the introduced in Section 3.1 distances is nonnegative definite. To get from this matrix a representation of the data items in an Euclidean space we perform the spectral decomposition of $\mathbf{B}$ by computing its eigenvalues and eigenvectors:

$$\mathbf{B} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T,$$

with $\mathbf{\Gamma}$ denoting the matrix of eigenvectors $\Gamma_1, \dots, \Gamma_{n-1}$ contained in $\mathbf{\Gamma}$ columnwise: $\mathbf{\Gamma} = (\Gamma_1, \dots, \Gamma_{n-1})$, and $\mathbf{\Lambda}$ being the diagonal matrix of eigenvalues: $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_{n-1})$. The (column) vectors $\Gamma_1, \dots, \Gamma_{n-1}$ are orthonormal, i.e. $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}$. Obviously each vector Γ_h , $h = 1, \dots, n$, has dimension n : $\Gamma_h^T = (\gamma_{1h}, \dots, \gamma_{nh})^T$.

When the size n is large, then the computing of the eigenvalues and of the eigenvectors of this matrix can cause some computational problems. E.g. for $n = 130$ we should compute possibly 129 eigenvectors, each of dimension 130, and we might have some problems with memory allocations, especially when using smaller computers.

Since the inner product matrix $\mathbf{B}$ —when using the three distances introduced in Section 3.1— is nonnegative definite, all its eigenvalues should be nonnegative and their sum should be equal to the sum of diagonal elements of $\mathbf{B}$. Each eigenvalue can be regarded as the variance of the principal coordinate derived from the eigenvector corresponding to this eigenvalue (see next subsection for a definition of a principal coordinate). Of course, principal coordinates with zero variance have no meaning for us.

Taking these facts into account we decided to proceed in the following manner:

First of all we fix $\epsilon > 0$, a small real number (in our program we accepted $\epsilon = 0.00001$). Next we continue stepwise, computing subsequent eigenvalues in a loop for $h = 1, \dots, n - 1$.

Say, we have already computed m eigenvalues ($0 \leq m \leq n - 1$). We compute the next, $(m + 1)$ -th eigenvalue, denoted by λ_{m+1} , and check the inequality: $\lambda_{m+1} \geq \epsilon$. If this inequality holds, we retain the $(m + 1)$ -th eigenvalue, otherwise we retain only m eigenvalues and stop the procedure.

After finishing these evaluations we check whether the sum of the extracted eigenvalues reproduces the trace of the matrix $\mathbf{B}$.

3.3. Implementation of step 3. Choosing the relevant principal coordinates

Say we got from the spectral decomposition of the matrix $\mathbf{B}$ exactly m eigenvalues satisfying $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq \epsilon$. For each of these eigenvalues we evaluate the corresponding principal coordinate by rescaling of the eigenvector by its eigenvalue. We will denote the derived principal coordinates (PC's) by $\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_m$; the tilde ~ should remind us that these are rescaled eigenvectors:

$$\tilde{\Gamma}_h = \Gamma_h \lambda_h^{1/2}, \quad h = 1, \dots, m.$$

Obviously $\tilde{\Gamma}_h^T \tilde{\Gamma}_h = \lambda_h$ and $\tilde{\Gamma}_g^T \tilde{\Gamma}_h = 0$ for $g \neq h$, $g, h = 1, \dots, k$.

Analogously as the eigenvectors $\Gamma_1, \dots, \Gamma_m$, also the derived principal coordinates can be put together into a matrix of size $n \times m$, $m \leq (n - 1)$ denoted now as $\tilde{\Gamma}$:

$$\tilde{\Gamma} = (\tilde{\gamma}_{ij}) = (\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_m).$$

In such a way we obtain a representation of the n data items in an Euclidean space $\mathbb{R}^m$. In this situation the l -th row of the matrix $\tilde{\Gamma}$, i.e. the vector $\tilde{\gamma}_l$ can

be viewed as the vector of coordinates of a point $P_l \in \mathbb{R}^m$. Obviously

$$\tilde{\gamma}_l = (\tilde{\gamma}_{lm}, \dots, \tilde{\gamma}_{lm}).$$

So far only the predictor variables were dealt with. Now we should relate them with the predicted variables $Y1$ and $Y2$. We shall do it separately for each of them.

Let Y denote the considered predicted (explained) variable and $\mathbf{y} = (y_1, \dots, y_n)^T$ its sample values.

We compute the Pearsonian correlation of each of the established principal coordinates $\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_m$ with the vector $\mathbf{y}$. For further considerations we take into account instead of the direct Pearsonian correlation r its square, called also the coefficient of determination. We denote the squared correlation coefficient by RR . The magnitude of RR gives information about the part of the total variability of the variable Y that can be explained by the impact (regression) of the considered principal coordinate. The obtained RR 's are then ordered by decreasing values and their cumulative sums are computed. Our computer program presents in the screen the ordered RR 's together with their cumulative values and asks for k , the number of principal coordinates to be taken for further analysis.

Suppose we have retained k principal coordinates $\tilde{\Gamma}_{i_1}, \dots, \tilde{\Gamma}_{i_k}$ mostly correlated with the vector $\mathbf{y}$. For simplicity the retained PC's will be denoted from now by $\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_k$. These principal coordinates can be put together as column vectors into the matrix $\tilde{\Gamma}_{(k)} = (\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_k)$. Obviously:

$$\tilde{\Gamma}_{(k)}^T \tilde{\Gamma}_{(k)} = \Lambda_{(k)},$$

with $\Lambda_{(k)}$ being the diagonal matrix of the eigenvalues corresponding to the chosen PC's.

The multiple squared correlation coefficient between $\mathbf{y}$ and the k retained principal coordinates will be denoted by $RR(k)$.

3.4. Implementation of step 4. Evaluation of the regression and of the residuals

Due to the fact that all the principal coordinates are mutually orthogonal the computations of the ordinary regression are straightforward. We use here the formulæ given by Cuadras and Arenas. Our interest is focused on the predicted values $\hat{y}_i$ obtained from this regression. We are also concerned with the residuals $r_i = y_i - \hat{y}_i$.

The regression equation establishing the linear relationship between the predicted variable Y and the retained PC's taken as predictors is:

$$(8) \quad Y = \beta_0 + \beta_1 \tilde{\Gamma}_1 + \dots + \beta_k \tilde{\Gamma}_k + e_{(k)}.$$

This equation corresponds to an ordinary regression model. The unknown parameters $\beta_0, \beta_1, \dots, \beta_k$ can be computed by known methods. The mutual orthogonality of the involved PC's, i.e. of the column vectors $\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_k$ makes the computations especially easy.

Before considering expected (predicted) values of the variable Y to be evaluated from the established regression equation given by (8) let us point out two possible situations:

- The data vector comprising the values of the predictor variables recorded for the item for which we want to make the prediction — is one of the data vectors on the base of which the above regression was established.
- The data vector is a completely new vector.

Corresponding to these two situations let us introduce the following notations: We will call the data set from which the PC's were evaluated — the *base data set*. A vector $\mathbf{x}^*$ belonging to this set will be called *own data vector*.

A set of vectors not belonging to the base data set will be referred to as *foreign data set* and the vectors belonging to this set will be called *foreign data vectors*.

Now suppose we have a data vector $\mathbf{x}^* = (x_1^*, \dots, x_p^*)$ with known values and we want to evaluate for this vector the predicted value of the variable Y . We have to distinguish between two cases:

- (i) the vector $\mathbf{x}^*$ comes from the same base data set, from which the principal coordinates contained in the matrix $\tilde{\Gamma}_{(k)}$ were evaluated, i.e. it is an own data vector;
- (ii) the vector $\mathbf{x}^*$ does not belong to this set, i.e. it is a foreign data vector, which means that it belongs to a foreign data set.

Suppose the vector $\mathbf{x}^*$ belongs to the base data set. In this case we can find such a subscript i that the vector $\mathbf{x}^*$ is identical with the data vector $\mathbf{x}_i$ from the base data set. To evaluate the predicted value $\hat{y}^* = \hat{y}_i$ for the respective item we use the formula (see Cuadras and Arenas, 1990, formula (14)):

$$(9) \quad \hat{y}_i = \beta_0 + \beta_1 \tilde{\gamma}_{i1} + \dots + \beta_k \tilde{\gamma}_{ik}$$

with $\beta_0 = \bar{y}$ and $\beta = \Lambda_{(k)}^{-1} \tilde{\Gamma}_{(k)}^T \mathbf{y}$, where $\beta = (\beta_1, \dots, \beta_k)$.

Now suppose the vector $\mathbf{x}^*$ corresponds to a new data item not comprised in the base data set, i.e., it is a foreign data vector. The evaluation of $\hat{y}^*$ is in this case more difficult. The problem was solved by Cuadras and Arenas. First we position the point $\mathbf{x}^*$ in the spaces $\mathbb{R}^m$ and $\mathbb{R}^k$. To do it we have to evaluate the vectors $\check{\mathbf{b}}$ and $\mathbf{d}^*$, where

$$\check{\mathbf{b}} = (\check{b}_1, \dots, \check{b}_n) = (b_{11}, \dots, b_{1n})$$

is a row vector comprising the diagonal elements of the inner product matrix $\mathbf{B}$ (see eq.(7)) evaluated from the base data set, and

$$\mathbf{d}^* = (d_1^2, \dots, d_n^2)$$

is the row vector comprising the squared distances of the new data item $\mathbf{x}^*$ from all the n data items contained in the base data set.

The coordinates $\mathbf{c}$ and $\mathbf{c}_{(k)}$ of this new data vector $\mathbf{x}^*$ in the formerly established spaces $\mathbb{R}^m$ (of all evaluated PC's) and $\mathbb{R}^k$ (the retained PC's mostly correlated with $\mathbf{y}$) are respectively:

$$\mathbf{c} = \frac{1}{2}(\check{\mathbf{b}} - \mathbf{d}^*)\tilde{\Gamma}\Lambda^{-1}, \quad \mathbf{c} \in \mathbb{R}^m,$$

and

$$\mathbf{c}_{(k)} = \frac{1}{2}(\check{\mathbf{b}} - \mathbf{d}^*)\tilde{\Gamma}_{(k)}\Lambda_{(k)}^{-1}, \quad \mathbf{c}_{(k)} \in \mathbb{R}^k.$$

Obviously the components of the vector $\mathbf{c}_{(k)}$ are identic with the first k components of the vector $\mathbf{c}$.

Now the predicted value $\hat{y}^*$ can be evaluated by the formula :

$$(10) \quad \hat{y}^* = \bar{y} + \mathbf{c}_{(k)}\Lambda_{(k)}^{-1}\tilde{\Gamma}_{(k)}^T \mathbf{y}.$$

After evaluation of the predicted value we calculate the residuals as the differences between the observed and predicted values. To make clear whether the residuals were obtained using predictions by formula (9) or by formula (10) we shall use the following denotations: r_i is the residual obtained when evaluating the predicted value on the base of PC's from the own data set, i.e. by (9); and v_i is the residual obtained when evaluating the predicted values on the base of PC's from a foreign data set, i.e. by (10).

4. RESULTS

We have analyzed the data presented in Section 2 of the paper . Let us remind that these were two sets of data, each comprising two predicted variables $Y1 = MvXY$ and $Y2 = Fs$, both predicted from the same set of $p = 8$ predictor variables.

The evaluations of the EUCL and of the L1 distance were done with normalized variables, i.e. at the begin of the calculations the data values were divided by the respective standard deviations.

The covariance structure in both data sets was similar except of the variables $X5$ and $X6$ (see Fig.1). Since the predictor variable $X6$ could blurr the covariance structure in both data sets, the calculations were repeated omitting this variable.

The evaluations were carried out along the steps described in Section 3. We have considered the distances evaluated by three methods: Euclidean (EUCL), L1-norm (L1) and Gower's (GOWER).

Our interest was focused on the following points:

1. To what degree of accuracy can the inner product matrix $\mathbf{B}$ be reproduced by h ($h \ll n - 1$) eigenvectors?
2. Suppose that we retain k PC's yielding the largest determination coefficients RR with the predicted variable Y . Let $RR(k)$ denote the squared multiple correlation coefficient (determination coefficient) of the considered variable Y with the k retained PC's. We ask: Is there any difference between the values of $RR(k)$ obtained when considering different distances?
3. How much differ the residuals obtained when using different numbers of PC's and different distances?

The results are as follows:

4.1. Exhaustion of the trace of the inner product matrix B by subsequent eigenvalues

The percentages of cumulative sum for subsequent eigenvalues —when considering $p=8$ predictors— are visualized in Figure 2.

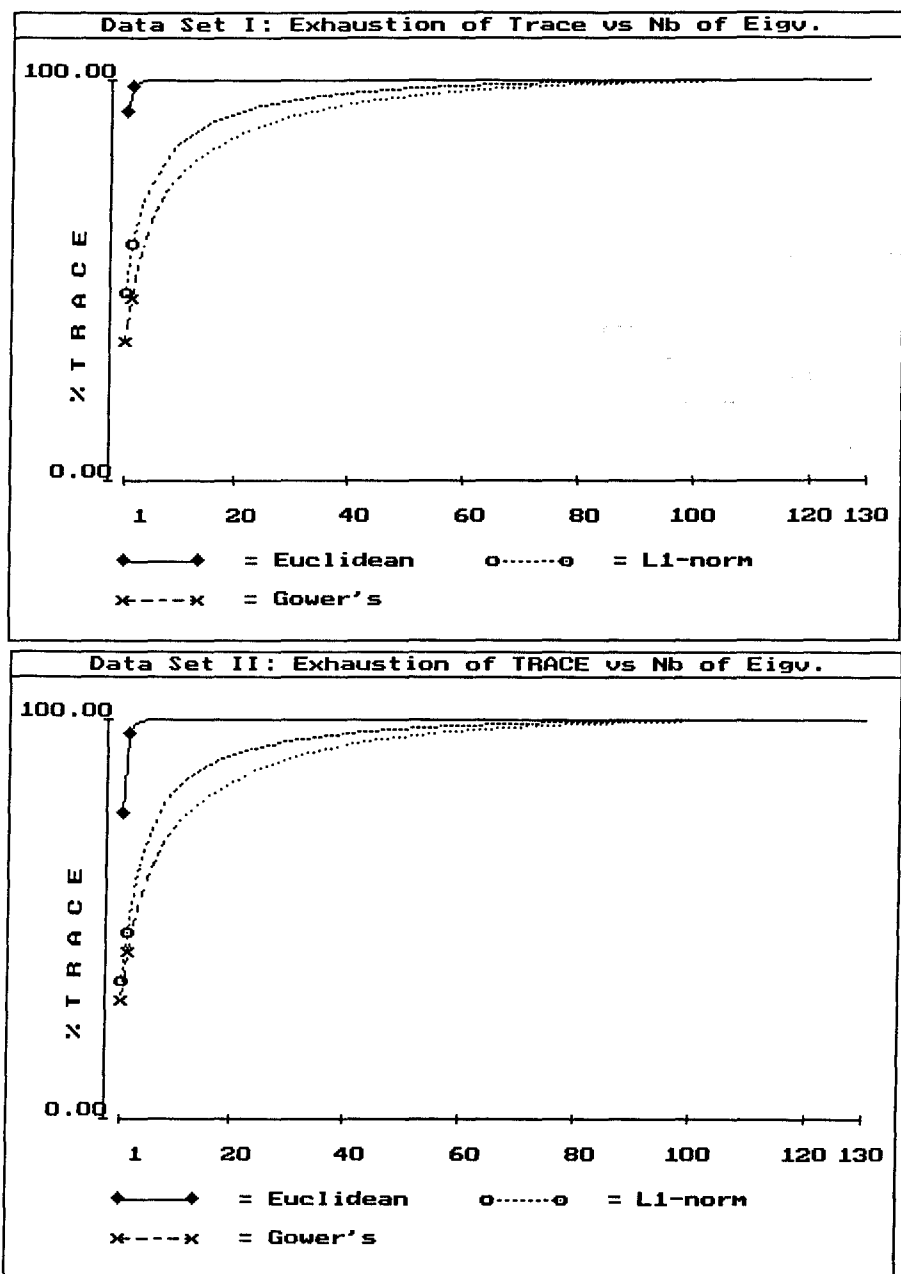


Figure 2.

Percentage of exhaustion of the trace of the inner product matrix B vs the Number of subsequently extracted eigenvalues — stated in data set I (upper figure) and II (lower figure).

One can see the substantial difference between the curve obtained for the Euclidean distance and the two curves obtained for the L1-norm and Gower's distances. For Euclidean distances the complete, i.e. 100 percentage exhaustion is achieved with $p = 8$ PC's, while for the other two distances that happens only with $n - 1$ PC's.

Considering the Euclidean distance the trace is practically exhausted by the first two eigenvalues. Thus the mutual inter-point configuration of the predictor variables can be visualized by a cluster of points in a plane obtained from the first two principal coordinates.

Quite different situation is met when considering matrices obtained for the other two methods. Here the cumulative sum increases steadily and it is difficult to decide how many PC's should be taken to obtain a satisfactory representation of the interdependency structure of the predictors.

4.2. Evaluation of the correlation coefficients $R(PC_i, Y)$

We have computed the squared Pearsonian correlation coefficients $RR(PC_i, Y)$ taking $i = 1, \dots, 8$ (all positive eigenvalues) for the Euclidean distance, and $i = 1, \dots, n - 1$ (all PC's) for the two other distances. The obtained RR's were ordered according to their decreasing values. This was done for both predicted variables considered in the data sets I and II.

Obviously a cumulative sum, evaluated for, say, k PC's, is in fact the coefficient of determination, or the square of the multiple correlation coefficient between the considered variable Y and the k principal coordinates yielding the largest squares of the univariate correlation coefficients.

In upper part of Table 2 we show the multiple determination coefficients $RR(k)$ obtained for the three analysed distances when considering various numbers (k) of principal coordinates taken according to the ordered RR's. For the Euclidean distance with $p = 8$ predictors at most $k = 8$ PC's could be evaluated. Comparing the multiple determination coefficients obtained for the L1-norm and Gower's distances one can see that these are higher by about 0.10 as compared with those obtained from the Euclidean distances. The $RR(k)$ values increase, up to about 0.95, when more and more principal coordinates, up to $k = 60$, are taken into consideration. The limiting value of the RR's when using the L1-norm and the Gower distances is $RR(n - 1) = 1.0$. The gains in the RR's obtained from $k = 8$ PC's, when using the L1-norm or Gower's distances instead of the Euclidean distance, range from 0.0672 to 0.1640.

Table 2.

Determination coefficients $RR(k)$ for three analysed distances and for varying numbers k of PC's. Results when considering 8 and 7 predictors.

8 predictors Distance	Y1 = MvXY				Y2 = Fs			
	Data Set I		Data Set II		Data Set I		Data Set II	
	k	$RR(k)$	k	$RR(k)$	k	$RR(k)$	k	$RR(k)$
EUCL	7	0.3800	7	0.3768	7	0.4885	7	0.4043
	8	0.3800	8	0.3768	8	0.4886	8	0.4044
L1-NORM	7	0.4771	7	0.5211	7	0.5536	7	0.5056
	8	0.4993	8	0.5408	8	0.5717	8	0.5253
	14	0.6100	14	0.6427	14	0.6606	14	0.6239
	16	0.6368	16	0.6711	16	0.6843	16	0.6523
	24	0.7299	24	0.7667	24	0.7666	24	0.7446
	60	0.9505	60	>.9503	60	0.9513	60	0.9508
GOWER	7	0.4554	7	0.5103	7	0.5405	7	0.5202
	8	0.4773	8	0.5313	8	0.5558	8	0.5378
	14	0.5893	14	0.6359	14	0.6345	14	0.6352
	16	0.6204	16	0.6657	16	0.6581	16	0.6648
	24	0.7241	24	0.7616	24	0.7396	24	0.7658
	60	0.9509	60	>.9518	60	0.9504	60	>.9520

7 predictors Distance	Y1 = MvXY				Y2 = Fs			
	Data Set I		Data Set II		Data Set I		Data Set II	
	k	$RR(k)$	k	$RR(k)$	k	$RR(k)$	k	$RR(k)$
EUCL	7	0.3734	7	0.3606	7	0.4760	7	0.3795
L1-NORM	7	0.4937	7	0.4984	7	0.5945	7	0.5214
	8	0.5116	8	0.5179	8	0.6098	8	0.5418
	14	0.6047	14	0.6186	14	0.6810	14	0.6429
	16	0.6317	16	0.6446	16	0.7003	16	0.6710
	24	0.7240	24	0.7361	24	0.7728	24	0.7539
	60	0.9348	60	0.9512	60	0.9486	60	0.9458
GOWER	7	0.5002	7	0.4922	7	0.5680	7	0.5044
	8	0.5285	8	0.5112	8	0.5878	8	0.5218
	14	0.6424	14	0.6015	14	0.6817	14	0.6108
	16	0.6679	16	0.6273	16	0.7064	16	0.6351
	24	0.7578	24	0.7195	24	0.7856	24	0.7234
	60	0.9491	60	0.9474	60	0.9533	69	0.9463

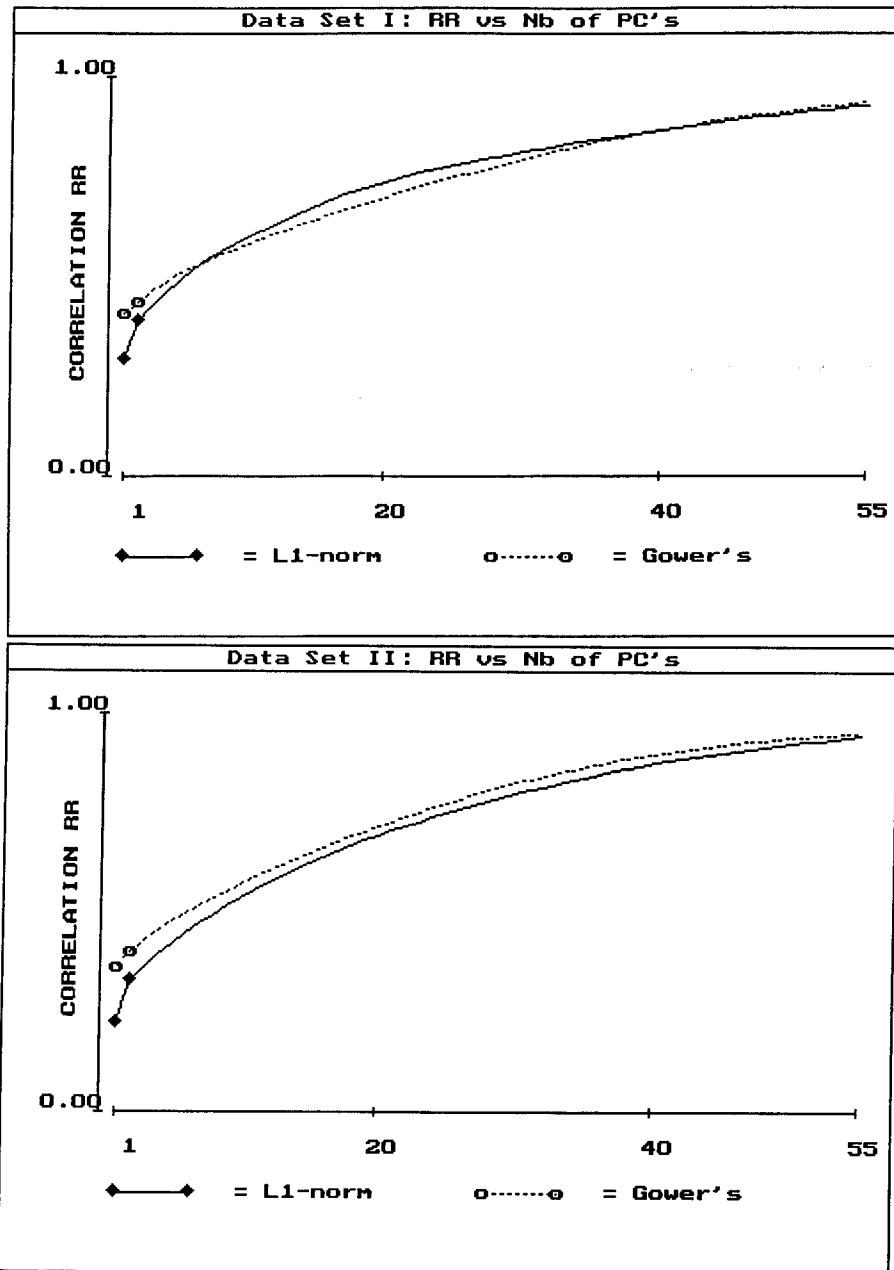


Figure 3.

Squared correlation (RR) vs the Number of selected Principal Coordinates (PC's) considered as predictors for $Y2=Fs$ — in data set I (upper figure) and II (lower figure).

For comparison with the $p = 8$ predictor set used when evaluating the determination coefficients shown in upper part of Table 2, we show in lower part of Table 2 analogous determination coefficients obtained when using only $p = 7$ predictor variables. Working with Euclidean distance we could obtain only 7 PC's. Therefore, when displaying the RR's obtained by use of the two other distances, we show the respective values $RR(k)$ for comparative purposes for $k = 7, 8, 14, 16, 24, 60$. The gains from the use of the L1-norm or Gower's distances, instead of the Euclidean distance, range from 0.0920 to 0.1419.

The cumulative sum $RR(k)$ of the ordered RR's versus k , the number of PC's yielding the subsequent sums, is visualized in Figure 3. This is shown only for the L1-norm and Gower's distances and for 55 PC's yielding the greatest RR's. One can see the steadiness of the increase of the respective cumulative sums. The upper graph in this Figure corresponds to data set I and the lower graph to data set II.

In Table 3 we show the *no.*'s (i.e. the *id* numbers) of the PC's that gave the largest RR's with the predicted variables. One can see that, for both predicted variables $Y1$ and $Y2$, the 1-st principal coordinate gave always the largest values of RR, i.e. in all cases it was the first PC that was identified as the most important one.

The *no.*'s of the second most important PC vary for various cases. There are no similarities between the *id* numbers obtained for $p = 8$ and $p = 7$ predictors. When the L1-norm or Gower's distances are used, quite often a PC with a high *no.*, was identified as playing the second most important role. We find here the *no.* 2 as the smallest *id* number and the *no.* 74 as the highest one.

The same can be said when looking at the *no.* of the third most important principal coordinate —obtained when using the L1 or Gower's distances. Among the *id* numbers of PC's contained in the first triplet of most important PC's we have such high *no.*'s as 23, 45 and 46 for $p = 8$, and 61, 74 and 94 for $p = 76$.

One can also see in Table 3 that among the eight PC's yielding the largest cumulative sum of the RR's there are PC's with as high *no.*'s as *no.* 121 and 123. This obviously can be stated again only for the L1 and Gower's distances.

From this we might conclude that using the L1 or Gower's distances we should, in fact, compute the all possibly available PC's and next check all of them for their correlations with the dependent variable.

Table 3.

No.'s of eight Principal Coordinates yielding the largest values of the determination coefficient RR in data sets I and II. Analysis is performed first with all 8 predictor variables and next with 7 predictor variables only.

Distance	Y	Set	No.s of eight Principal Coordinates								$RR(8)$
Euclidean Distance	Y1	I	1	5	4	2	8	6	7	3	0.3800
		II	1	6	2	8	7	3	4	5	0.3768
	Y2	I	1	4	5	8	2	6	3	7	0.4886
		II	1	6	8	7	3	4	2	5	0.4044
L1-norm Distance	Y1	I	1	46	2	82	35	23	24	6	0.4993
		II	1	13	17	67	38	73	102	28	0.5408
	Y2	I	1	23	6	65	80	12	79	46	0.5717
		II	1	17	13	69	81	67	38	102	0.5253
Gower's Distance	Y1	I	1	2	45	23	38	80	6	12	0.4773
		II	1	13	16	102	22	38	17	29	0.5313
	Y2	I	1	6	23	12	80	121	79	107	0.5558
		II	1	22	13	102	81	17	79	86	0.5378

Distance	Y	Set	No.'s of seven Principal Coordinates								$RR(7)$
Euclidean Distance	Y1	I	1	4	3	2	7	5	6	—	0.3734
		II	1	2	5	7	6	4	3	—	0.3606
	Y2	I	1	3	4	7	5	2	6	—	0.4760
		II	1	5	3	2	7	4	6	—	0.3795
L1-norm Distance	Y1	I	1	40	2	5	111	96	72	—	0.4937
		II	1	61	19	12	56	103	97	—	0.4984
	Y2	I	1	40	5	72	10	74	20	—	0.5945
		II	1	19	94	73	104	103	61	—	0.5214
Gower's Distance	Y1	I	1	2	40	74	5	23	39	—	0.5002
		II	1	12	26	61	11	74	19	—	0.4922
	Y2	I	1	74	5	39	14	40	123	—	0.5680
		II	1	12	19	26	79	104	11	—	0.5044

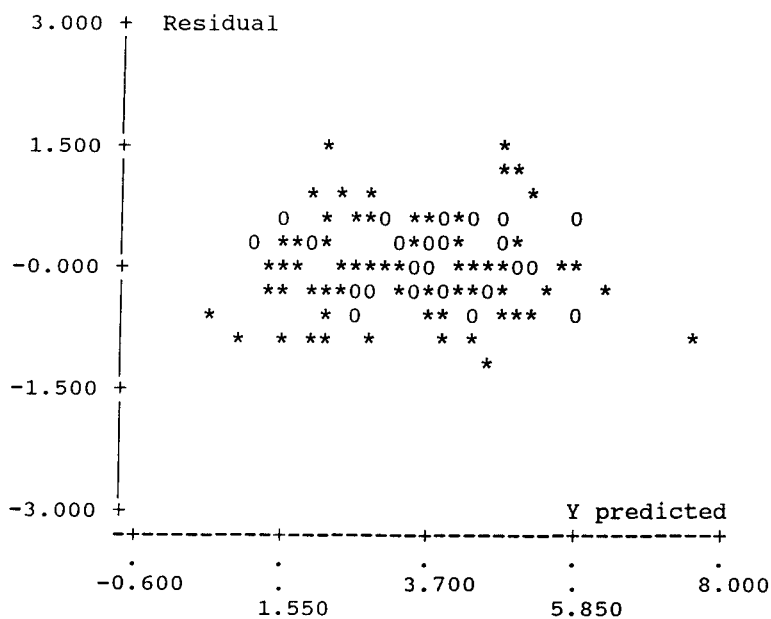
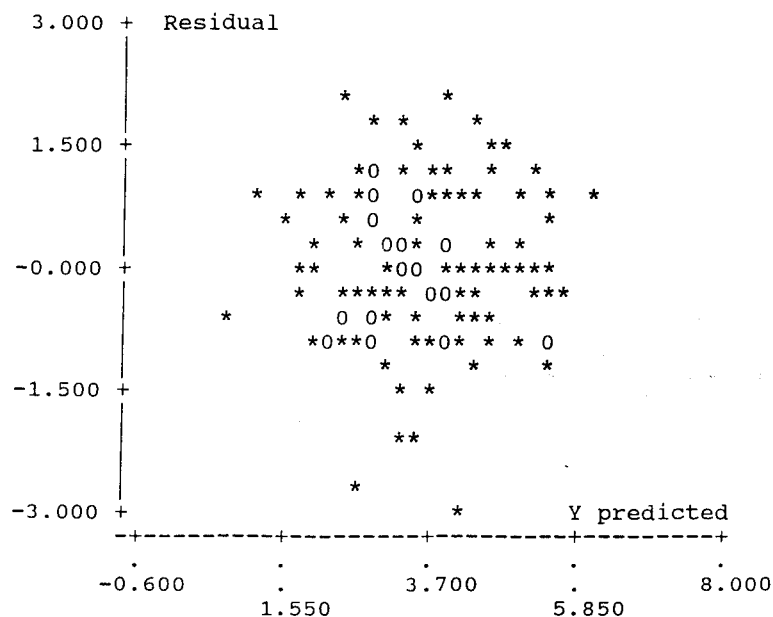


Figure 4.

Residuals against Predicted values of the variable $Y = F_s$ when considering $k=8$ PC's (upper figure) and $k=35$ PC's (lower figure) * denotes single point 0 denotes multiple (overlapping) points.

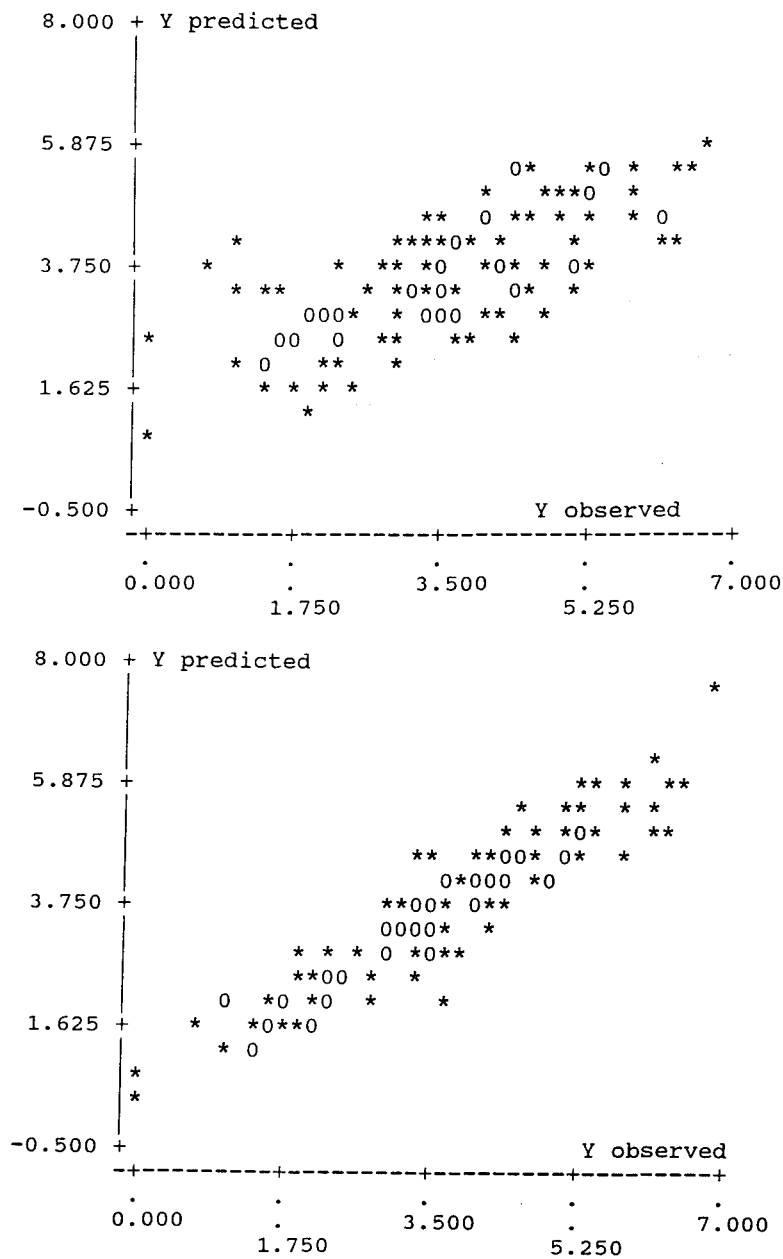


Figure 5.

Predicted values of $Y_2 = F_s$ against observed values of this variable when considering $k=8$ PC's (upper figure) and $k=35$ PC's (lower figure) * denotes single point 0 denotes multiple (overlapping) points.

4.3. Evaluation of predicted values and analysis of residuals

The main goal of our work was to compare the usefulness of the distance based regression in flare activity predictions. The quality of a regression method can be gauged in a very appealing way by looking at the residuals obtained from the respective regressions.

To illustrate directly the differences between the predicted values $\hat{y}_i$ calculated on the basis of increasing number of PC's we have constructed several scatterdiagrams visualizing the relations between the residuals r_i and the values $\hat{y}_i$, predicted on the basis of various numbers of retained PC's. In Figure 4 we show exemplary scatterdiagrams obtained for the variable Y2 when making predictions in set II on the basis of PC's evaluated in the same set. The predicted values $\hat{y}_i$ were evaluated by use of the Gower distance method both for the number $k = 8$ of PC's (upper part of the figure) and for $k = 35$ of PC's (lower part of the figure). Multiple (overlapping) points are marked by 0's. One can see the substantial difference between these two clusters of points: the values of r_i obtained for $k = 35$ PC's are distributed more closely to zero line than those obtained for $k = 8$ PC's.

Similarly, for the same case (i.e. for the data set II and for the variable Y2), we show in Figure 5 the predicted values $\hat{y}_i$ versus the observed values $y_i^{(obs)}$. The predicted values were also evaluated by use of the Gower distance method with $k = 8$ PC's (upper part of figure) and with $k = 35$ PC's (lower part of figure). One can see that the points obtained for $k = 35$ of PC's are distributed more closely to the diagonal line than those obtained for $k = 8$ of PC's.

Looking at these figures one can see that when retaining a higher number of PC's we can describe more accurately the interdependency structure in the data set under consideration; hence the residuals become smaller and smaller with increasing number of PC's taken for prediction.

In our analysis with two sets of data we wanted to evaluate the quality of predictions not only for the own data sets but also make some inference for data vectors belonging to foreign data sets. We proceeded as follows: First we considered the data set I as the base data set. We evaluated for this set the PC's and next —using formula (9)— the predicted values $\hat{y}_i$ for the data items $i = 1, \dots, 130$, i.e. for all the items belonging to this data set. The respective residuals were evaluated as:

$$r_i = y_i^{(obs)} - \hat{y}_i \quad \text{for } i = 1, \dots, 130.$$

Next, with the same base data set we took in turn the data vectors from the second data set, considering them as new observations. We evaluated for them

the predicted values by means of formula (10). The respective residuals were now evaluated as:

$$v_i = y_i^{(obs)} - \hat{y}_i \quad \text{for } i = 1, \dots, 117,$$

where $y_i^{(obs)}$ denotes now, for the i -th data item, the value of the predicted variable Y considered in the second data set, and $\hat{y}_i$ denotes the respective value evaluated by formula given by eq.(10) using the matrices $\tilde{\Gamma}_{(k)}$ and $\Lambda_{(k)}$ evaluated from set I.

Next we interchanged the meaning of the two sets. Set II was taken as the base data set and set I was considered as a foreign data set. We evaluated for the data vectors belonging to set II the predicted values $\hat{y}_i$ for $i = 1, \dots, 117$ and the residuals r_i from PC's obtained from the own data set. In turn we considered the data vectors from the first data set as new data vectors and evaluated for them the predicted values y_i and the residuals v_i using the PC's from the set II.

The evaluations were carried out for each data set in six setups established by combinations of the two predicted variables $Y1 = MvXY$ and $Y2 = Fs$ crossed with the three methods of computing distances. For each setup by taking into account the two groups of data and the two kinds of residuals (i.e. r_i and v_i) we got four series of residuals.

Summary of the analysis is presented in Table 4.

We present in Table 4 only the results obtained when considering the full set of predictors, i.e. the variables $X1 - X4$. The results obtained after dropping the variable $X6$ from the set of predictors (see the discussion in Section 2) look very similar and are not shown here.

The series of residuals were characterized by their means (AV_r, AV_v), standard deviations (SD_r, SD_v), and the quotients $q_r = SD_r/SD_y$ and $q_v = SD_v/SD_y$ obtained as the ratios of the standard errors of the residuals r_i and v_i to the standard deviation of the considered variable Y , respectively. The means AV_r are all equal to zero, and therefore they are not shown in Table IV. The evaluations of the predicted values were done taking into account $k = 3$ and $k = 8$ PC's. The squared multiple correlation coefficients $RR(k = 3)$ and $RR(k = 8)$ are shown in Table IV too.

Statements made when analysing values shown in Table IV:

1. Estimation from the own data set

- (a) All the means AV_r of residuals obtained when making the prediction from own data set appeared to be equal to zero and therefore they are not shown in Table 4.

Table 4.

Averages (AV_v), standard deviations (SD_r, SD_v) and quotients (q_r, q_v) for the predicted variables $Y1=MvXY$ and $Y2=Fs$.

Y	Data Set	Distance	k	$RR(k)$	Estimation from				
					own data set		foreign data set		
					SD_r	q_r	AV_v	SD_v	q_v
M v X Y	I	EUCL	3	0.3566	1.14	0.81	-0.47	1.20	0.85
			8	0.3800	1.11	0.79	-0.34	1.22	0.87
		L1- NORM	3	0.3686	1.12	0.80	-0.54	1.30	0.92
			8	0.4993	1.01	0.72	-0.44	1.46	1.04
		GO- WER	3	0.3559	1.13	0.80	-0.56	1.21	0.91
			8	0.4773	1.02	0.72	-0.64	1.50	1.06
	II	EUCL	3	0.3667	1.26	0.80	0.78	1.33	0.84
			8	0.3768	1.25	0.79	0.91	1.35	0.85
		L1- NORM	3	0.4162	1.21	0.77	0.67	1.34	0.85
			8	0.5408	1.07	0.68	0.49	1.37	0.87
		GO- WER	3	0.4133	1.24	0.78	0.70	1.31	0.83
			8	0.5313	1.08	0.68	0.74	1.37	0.87
F s	I	EUCL	3	0.4786	1.13	0.72	-0.42	1.23	0.79
			8	0.4886	1.12	0.72	-0.40	1.21	0.78
		L1- NORM	3	0.4685	1.14	0.73	-0.61	1.26	0.81
			8	0.5717	1.02	0.65	-0.42	1.45	0.93
		GO- WER	3	0.4649	1.14	0.73	-0.74	1.32	0.85
			8	0.5558	1.04	0.67	-0.91	1.59	1.02
	II	EUCL	3	0.3916	1.14	0.79	0.67	1.19	0.82
			8	0.4044	1.12	0.77	0.79	1.22	0.84
		ABS	3	0.4231	1.10	0.76	0.50	1.19	0.82
			8	0.5253	1.00	0.69	0.57	1.26	0.87
		GO- WER	3	0.4302	1.10	0.76	0.51	1.21	0.83
			8	0.5378	0.99	0.68	0.61	1.29	0.89

- (b) Obviously the values $RR(k = 8)$ are consequently higher than those of $RR(k = 3)$ evaluated for the same data set, the same predicted variable and the same method. This has to be so. Introduction of more explanatory variables (in our case more PC's) into the regression cannot decrease the explained variance, equivalently can not decrease the squared multiple correlation coefficient.

Comparing the squared correlation coefficients RR , evaluated with $k = 8$ PC's, obtained for the three used distances we state that for all setups the L1-norm and Gower distances yield systematically higher values of RR than the Euclidean distance do, and consequently lower values of SD_r and q_r .

- (c) Analogously, comparing the standard deviations SD_r 's and the quotients q_r 's evaluated for the three distances we state that the appropriate values obtained in the case of estimation from $k = 3$ PC's are systematically **higher** than those for $k = 8$ PC's. This could be expected for the same reasons as explained above: with increase of the determination coefficient RR the corresponding standard deviation of the residuals should decrease (unless there are not too few degrees of freedom left).

2. Estimation from the foreign data set

- (a) The means AV_v of residuals obtained when making the predictions from a foreign data set are systematically either **negative** (when we predict the variables $Y1$ and $Y2$ in set I using the regression equation established from set II) or **positive** (when predicting the same variables in set II from the regression established in set I). That means that for our data all the predictions based on a foreign regression (i.e. on a foreign set of PC's) are systematically biased and we make regularly either overestimation or underestimation of the predicted values. This could eventually be corrected by adding a suitable constant to the predicted values.
- (b) The standard deviations (SD_v) and the respective quotients (q_v) characterizing the residuals obtained when using for prediction $k = 3$ PC's established in a foreign data set —are **systematically higher** than the analogous SD_r 's and q_v 's obtained when making the respective predictions from the own data set. The same is true when considering predictions made on the basis of $k = 8$ retained PC's. That means that predictions from a foreign data set are generally worse than analogous predictions made from the own data set— and this was stated in our data using both $k = 3$ and $k = 8$ PC's for predictions. The differences in the quotients obtained when predicting

from the own and the foreign data set are within the range (0.04 - 0.12) and (0.06 - 0.35) for $k = 3$ and $k = 8$, respectively.

- (c) Comparing the standard deviations SD_v 's of residuals and the quotients q_v 's —obtained when making the prediction from $k = 3$ as opposed to $k = 8$ PC's we are surprised to find that in 11 of the 12 performed analyses the values of SD_v and q_v obtained when predicting from $k = 8$ PC's are **worse** than the analogous values obtained from $k = 3$ PC's only.

This is a warning against taking too many PC's for prediction!

- (d) Comparing the standard deviations of the residuals and the resulting quotients within the methods EUCL, L1 and GOWER we state that:
 - i. using $k = 3$ PC's for prediction we obtain in majority of the setups nearly the same values of the residuals and of the quotients for all three considered distances;
 - ii. using $k = 8$ PC's the respective values of SD_v and q_v obtained with L1-norm and Gower's distances are decidedly higher than those yielded by the Euclidean distances. It happened even, that in two of four presented in Table 4 analyses the Gower distance yielded quotients $q_v=1.06$ and $q_v=1.02$, which means that the variance of the residuals is greater than the variance of the predicted variables.

This again is a strong warning against using too many PC's that are known to describe in a satisfactory manner the interdependency structure in the own data set, however could be totally unsuitable for the new data set under consideration.

5. DISCUSSION

Generally we found that the regression obtained when using the L1-norm or the Gower's distance is superior to that based on Euclidean distances —when describing the given data set. This can be seen when comparing the coefficients of determination (RR) obtained with the the same (i.e. $k = 8$) number of principal coordinates.

Since it is known (the proof is presented by Cuadras and Arenas, 1990) that the regression based on Euclidean distances with p PC's is equivalent (in results, e.g. in the coefficient of determination and in the predicted values of Y) to the ordinary least squares method with p variables, we may infer that the

distance-based regression using the L1-norm or the Gower's distance describes the considered data set more accurately than this is done by means of the classical OLS method. Moreover, retaining more and more principal coordinates (which is possible, when using L1-norm and Gower's distances, but what is not possible when using the Euclidean distances) we can make the residuals smaller and smaller —up to the final pace, when with $n - 1$ PC's we obtain a total explanation of the variable Y by the considered PC's.

However, these fantastic results are achieved by accounting for the particular configuration of the observed data vectors. It is sure that a certain part of this configuration arose from random effects. Therefore, it should be put clearly that a high number of PC's is certainly useful to describe the given data set (a random sample from a population under consideration), however it may be quite useless for describing the same features in another data set representing another sample for the same population, especially if this data differs in some particular aspects from the given data set.

In our paper we have considered two sets of data: one describing sunspot group activity during the 1988 year, the other during the 1989 year. We have made predictions for the year 1989 using the distance-based regression evaluated for the 1988, and vice-versa. The interdependency structure in the considered sets of data was discussed in Section 2. To some extent this structure was found to be similar, although not identic. This made us feel permitted to try to make predictions in one of these sets of data from regression (PC's) established in another set. We stated (see Table 4) that taking $k = 3$ PC's from the foreign data set one can get reasonable predictions —although systematically biased (with some additional knowledge on the solar activity process the bias could be removed in a reasonable way).

Taking for predictions $k = 8$ PC's from the foreign data set we got often much worse or even very bad predictions than considering only $k = 3$ PC's.

From this we might infer that describing the data set by $k = 8$ PC's we include into the description a part of random noise which has added to the given data set some particular features which are not likely to appear in another sample.

For our data the model is much more complicated because the solar activity is a nonstationary process changing according to an eleven-year cycle. Both our data sets I and II refer to the time interval very near to the maximum of the solar eleven-year cycle (end of 1989 year). However, the structure of the interrelations among characteristics of solar active regions has changed to some extent in this time interval (see Jakimiec and Bartkowiak (1994) for more details). Therefore, our predictions performed for the PC's of the foreign data set could be bad. We

might hope that making predictions for intervals being more close in time to the base data set we would obtain quite satisfactory estimates of flare activity. This problem needs further investigation.

Cuadras and Arenas give one statistical test, permitting to judge the statistical significance of the retained PC's. The test works under assumption of normality. Our data are not normal and, therefore, we did not apply the mentioned test. Perhaps an evaluation of the statistical significance of the retained PC's by a kind of resampling methods would be more suitable in these circumstances.

From computational point of view the distance-based regression is much more difficult to carry out. The algorithm we have used works with the lower triangle of a square matrix of size $n \times n$ simulated in an one-dimensional array and to compute the eigenvalues and eigenvectors the whole structure has to be kept in memory. Our program can deal mostly with $n = 130$ data-items (individuals). An analogous program from MULTICUA can deal with perhaps $n = 160$ data-items. With larger number of data vectors special computational methods are needed. This is a disadvantage of the distance based methods.

Recently we have learned that Cuadras *et al.* (1993) established another algorithm for identifying the relevant principal coordinates. In principle, the new algorithm does not require the evaluation of all eigenvectors of the inner product matrix **B** and therefore it is more economic.

In any case the numerical obstacles can surely be surmounted. The most important is the solution which seems, in the case of the distance-based regression, to be a very general one describing the intrinsic features of the data considered.

ACKNOWLEDGEMENT

The work was supported by the KBN (Polish Research Council) grant n° 650/2/91.

REFERENCES

- [1] Arenas, C.; Cuadras, C.M. & Fortiana, J.F. (1991). *MULTICUA, Paquete no standard de Análisis Multivariante*. Universitat de Barcelona, Departament d'Estadística.
- [2] Bartkowiak, A. & Jakimiec, M. (1986). "Short-Term Flare Activity Predictions by Means of Regression Functions Calculated for Various Zurich Class Groups Observed Over the Period 1977-1979". In: *Proc. of Solar-Terrest. Pred.* P.A. Simon, G. Heckman and M. Shea, eds., Boulder, USA, 285-293.
- [3] Bartkowiak, A. & Jakimiec, M. (1990). "Short-Term Predictions of Flare Activity Using α — Trimmed Regression Method". *Acta Astr.*, **40**, 169-181.
- [4] Cuadras, C.M. (1991). *Metodos de Analisis Multivariante*. 2^a edition, PPU Barcelona.
- [5] Cuadras, C.M. (1989). "Distance analysis in discrimination and classification using both continuous and categorical variables". In: Y. Dodge, (Ed): *Statistical Data Analysis and Inference*. Elsevier, North Holland, pp. 459-473.
- [6] Cuadras, C.M. (1988). "Distancias Estadísticas". *Estadística Española*, **30**, n^o **119**, 295-378.
- [7] Cuadras, C.M., Arenas, C. and Fortiana, J. (1993). "Further aspects of a distance based model for prediction including non linear regression". *LINSTAT'93*, Poznań (Poland), 1-4 June 1993. Report.
- [8] Cuadras, C.M. & Arenas, C. (1990). "A distance based regression model for prediction with mixed data". *Commun. Statist. — Theory Meth.*, **19**, 2261-2279.
- [9] Fortiana, J. (1992). *Enfoque basado en Distancias de algunos Metodos Estadísticos Multivariantes*. Ph. D. dissertation, Universitat de Barcelona, 155 pp.
- [10] Gower, J.C. (1971). "A general coefficient of similarity and some of its properties". *Biometrics*, **27**, 857-874.
- [11] Hirman, J.W.; Neidig, D.F.; Seagraves, P.H.; Flowers, W.E. & Wiborg, P.H. (1980). "The Application of Multivariate Discriminant Analysis to Solar Flare Forecasting". In: *Proc. of Solar-Terrest. Pred.* R.F. Donnelly, ed. 3, C-64.
- [12] Jakimiec, M. & Bartkowiak, A. (1989). "Investigation of the Influential Points in a Regression Problem Occuring in Short-Term Prediction of Solar Flare Activity". *Acta Astr.*, **39**, 257-273.
- [13] Jakimiec, M. & Bartkowiak, A. (1994). "Short-term Solar Flare Predictions by Distance-Based Regression. I. Bearalert Regions in 1988 and 1989 — Continuous predictors". *Acta Astr.*, **44**, 115-140.

- [14] **Jakimiec, M. & Wasiucioneck, J.** (1980). "Short-Term Flare Predictions and Their Stationarity During the 11-Year Cycle". In: *Proc. of Solar-Terrest. Pred.* R.F. Donnelly, ed. 3, C-54.
- [15] **Jolliffe I.T.** (1986). *Principal Component Analysis*. Springer New York.
- [16] **Krzanowski W.** (1988). *Principles of Multivariate Analysis. A User's Perspective*. Oxford. Clarendon Press.
- [17] **Mardia, K.V.; Kent J.T. & Bibby J.M.** (1979). *Multivariate Analysis*. Academic Press, London, New York.
- [18] **SGD.** *Solar Geophysical Data, Comprehensive Results. 1988, 1989.* NOAA Boulder, USA.
- [19] **Vecchia, D.F.; Caldwell, G.A.; Tryon, P.V. & Jones, R.H.** (1980). "Logistic Regression for Solar Flare Probability Forecasting". In: *Proc. of Solar-Terrest. Pred.* R.F. Donnelly, ed. 3, C-76.
- [20] **Zirin, H. & Marquette, W.** (1990). "BEARALERTS: A Successful Flare Prediction System". *Sol. Phys.*, **131**, 1, 149–164.

NOTES ON THE BIAS OF DISSIMILARITY INDICES FOR INCOMPLETE DATA SETS: THE CASE OF ARCHAEOLOGICAL CLASSIFICATION

ANGELA MONTANARI and STEFANIA MIGNANI

Department of Statistics

University of Bologna, Italy

The problem of missing values is particularly present in archaeological research where, because of the fragmentariness of the finds, only a part of the characteristics of the whole object may be observed. The performance of various dissimilarity indices differently weighting missing values is studied on archaeological data via a simulation. An alternative solution consisting in randomly substituting missing values with character states is also examined. Gower's dissimilarity coefficient seems to be the least biased one either with 25% missing values and 49%; it has not however a constant behaviour as to the sign of the bias. The simulation experiment has also shown that when average linkage cluster analysis is performed on an incomplete data set either using Gower's index or randomly substituting missing values gives satisfactory results while the modified indices fail to detect the cluster structure.

Key words: Dissimilarity indices, missing values, classification, cluster analysis, archaeological research.

—Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.

—Acceptat el febrer de 1993.

1. INTRODUCTION

The problem of missing values is particularly present in archaeological research where time and possible geological, climatic and historical changes do not always allow findings of complete remains but only fragments on which just a portion of the characteristics of the whole object may be observed, often different from one object to another. The gain obtained observing, for each unit, the largest number of variables is in fact partially dimmed by the possible increase in missing values, with an ensuing decrease of reliability in comparisons. When one's aim is to detect clusters of observations, each measure of similarity between two units is in fact the more reliable the more it is capable to control for the effect of missing data and to use all the available information.

This paper deals with the definition of dissimilarity indices between archaeological reports of raw pottery, whose age lies between IV century B.C. and XIII century A.D., found out in an excavation nearby Castel Raimondo (Udine, Italy)¹, in order to detect groups homogeneous for characteristics that could allow us to date the finds. The fragmentariness of the manufacts is probably due to the presence in the clayey cob of sometimes big mineral inclusions as well as to the rudimentary production and baking; the work of baking, in fact, took place at very low temperatures (500 – 600°C) and for too short a time to allow a complete clay transformation (Guermandi, 1990).

When one has to deal with incomplete data sets, statistical literature sometimes suggests to restrict the analysis only to the units without missing values. This didn't seem to be a viable solution in our context as we had no complete object on which all the 28 qualitative variables considered could be observed (from kind of pottery to decoration, from brim descriptive parameters to ceramic cob). In order to reduce the proportion of missing values in the data set to be used for statistical analysis, we selected at first only the 436 units which had no missing values in the four variables that describe brim characteristics. This aspect seems to be the one most closely connected to time variation. (In what follows this data set will be called *data1*). In this new data set the overall percentage of missings was about 49%; therefore two seemed the possible solutions to the problem of comparison between two units: using a dissimilarity index apt to suitably weight missing values thus sizing the differences between units with a good reliability or substituting missing values with possible states of

¹The research has been realized by the "Istituto Beni Culturali" of "Regione Emilia Romagna" and by the "Istituto di Archeologia" of the University of Bologna under the direction of Sara Santoro Bianchi. The results have been published in two volumes: *Castel Raimondo. Scavi 1988-90. I Lo scavo* (S. Santoro Bianchi ed.) L'Erma di Bretschneider, Roma, 1992; *Castel Raimondo. Scavi 1988-90. II I materiali*, L'Erma di Bretschneider, Roma, in press.

the character, in analogy to what is commonly done with quantitative variables (Brothwell and Krzanowski, 1974; Beale and Little, 1975).

An often suggested approach in the statistical literature (Gordon, 1981) is to handle missing observations on qualitative variables by assigning the missing value to the state which occurs most frequently in the set of objects most closely resembling the incompletely recorded unit. In our context however such a solution is not appropriate for two reasons: the high number of missing makes it difficult to determine suitable frequencies for the various states of the characters and, on the other hand, substituting the most frequent state for missing values would lead to underestimate dissimilarities. Infrequent characteristics in observed data may in fact have been very frequent in those reports in which they are now missing because of the manufacts fragmentariness (for example objects with thin brim may poorly be represented in the observed set of objects as such a characteristic favours object brittleness). We have therefore decided to substitute incompletely recorded data with a character state randomly selected among possible ones with probability equal to the reciprocal of the number of states each character can assume.

2. SIMILARITY AND DISSIMILARITY INDICES: A COMPARISON BETWEEN DIFFERENT SOLUTIONS

When, in multivariate statistical analysis, one has to deal with qualitative variables, a similarity index between two generic units i and j may be determined as the ratio between the number of attributes on which the two units have the same category, divided by the total number of attributes considered. Such an index may be easily modified, allowing for missing values, by dividing the number of matches for the number of variables for which comparison is possible and then multiplying by the ratio between the total number of variables and the number of possible comparisons (Seber, 1984; Krzanowski, 1988). If no comparison is possible the similarity between two units is set equal to 0. Yet this index is not normalized and so it is not suitable to make comparisons between different empirical situations: that's why it will not be examined in what follows.

An alternative and more general solution which allows for comparisons between units on which both qualitative and quantitative characters have been examined, has been put forward by Gower in 1971; one version of that coefficient is obtained, for two units i and j , by assigning a score $0 \leq s_{ijk} \leq 1$ and a weight w_{ijk} for character k . The coefficient is described as:

$$S_G = \frac{\sum_{k=1}^n s_{ijk} s_{ij'k}}{\sum_{k=1}^n w_{ijk}}$$

and the weight w_{ijk} is set to 1 when a comparison is considered valid for character k and to 0 when the value of the state for character k is unknown for one or both observational units; in such a case also s_{ijk} is conventionally set equal to 0. For quantitative characters one has:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{x_{k \max} - x_{k \min}}$$

where x_{ik} and x_{jk} are the values of unit i and unit j on variable k , while for qualitative characters s_{ijk} is 1 for matches between states for that character and 0 for a mismatch. If no comparison is possible the similarity between two units is set equal to 0. This coefficient, denoted as D_1 in what follows, has recently found fairly extensive archaeological use (Doran and Hodson, 1975; Shennan, 1988).

However, as Sneath and Sokal (1973) underline, the number of states in a multistate character is thus not taken into consideration and Gower's coefficient resembles, in this aspect, a simple matching coefficient applied to a data matrix involving multistate characters. Besides this coefficient does not behave uniformly in all situations: the similarity coefficient between two units with only one match and incomparable for all the other characters is in fact set equal to one, that is maximum similarity, thus showing, in that case, a tendency to overestimate similarity between units; on the contrary, when two units differ in one character state and are incomparable for the remaining ones the coefficient considers them as completely different, tending, in such a case, to underestimate similarity.

An alternative solution Gower's index allows for, is to substitute the weight $w_{ijk} = 0.5$ for $w_{ijk} = 0$ when comparisons are impossible, assuming that, for missing observations, similarity or dissimilarity between two units are equally likely for variable k . (This index will be indicated as D_2 in what follows).

This approach neglects again the different number of states each character can assume and so does not take into account that the probability for a match to occur on character k decreases as the number of such values increases; a further modification may thus be obtained by setting the weight w_{ijk} equal to

the reciprocal of the number of states. (From now on this index will be indicated as D_3).

In what follows reference will be made to dissimilarity coefficients, obtained as the complement to 1 of the corresponding similarity index, as the statistical software available for further analysis requires dissimilarity or distance matrices as input.

3. THE SIMULATION

In order to evaluate the performance of the coefficients described in the previous section, a subset of data1 was drawn. The aim was to obtain a set of units with complete records, of suitable size for statistical analysis; so we dropped one variable at a time² and counted the number of units with no missing value; we stopped when we got at least 100 units. A set of 122 complete units on 9 variables was thus obtained and the matrix containing the dissimilarities between such units calculated. As no missing values appeared in the data matrix, all the considered indices gave the same dissimilarity values. Under the hypothesis of random distribution of missings in the data set —as random is the finding of one fragment instead of another— a simulation experiment was then run as follows. Some data observations were randomly selected and substituted with missing values (first according to the proportion of missings in the data set of the 436 initial manufactures, *i.e.* 49%, and then in about half that proportion) and the various dissimilarity coefficients were calculated. This process was repeated 400 times, and the bias of each index was calculated as the difference between the mean of the dissimilarity matrices over the 400 replications and the dissimilarity matrix of the 122 real objects. All calculations were implemented in Gauss on a PC.

At last, to determine whether missing data might be better handled by modifying dissimilarity coefficients or by imputing possible values, each missing, generated as previously described, was substituted by a character state³, chosen at random among all the possible ones, with probability equal to the reciprocal of the number of states and once again, in 400 replicates, dissimilarity matrices and corresponding bias were computed. (From now on this situation will be

²One must remember that no entire object was found or assembled and so in data1 the observational units are only fragments.

³The same matrix has been used either to replace missings with imputed values or to compute the different dissimilarity coefficients.

denoted as D_4). The results thus obtained have then been compared with those derived using the various dissimilarity indices.

The capability the considered methods have in handling missing values has been further checked studying their performance in cluster analysis (Wishart, 1978). From the set of 9 variables previously used in the simulation experiment, 4 have been chosen, namely “kind of clayey cob” and “brim characteristics” (three), which could well separate between “open” and “closed” shapes⁴ of pottery and which had no missings in the considered data set. Thus two populations have been defined and samples of 50 objects have been randomly selected from each population. The couples of samples have then been put together, thus forming a sample of 100 units on which average linkage cluster analysis has been performed (Sneath and Sokal, 1973; Gordon, 1981; Rizzi, 1985). This kind of clustering method was chosen because of its good performances in previous analyses of data1 (see Capitanio, Mignani, Montanari, 1993). Besides, as the dissimilarity indices used are weighted averages, the recourse to average linkage coherently extends the properties of arithmetic mean to the clustering criterion.

Missing values have then been randomly introduced in the sample in the way and in the proportions previously described. Average linkage cluster analysis has been performed again, for each of the different ways of handling missing values, *i.e.* for each of the methods denoted as D_1, D_2, D_3, D_4 . The process has been replicated many times and on different samples. In the end the results obtained for each method have been compared with the one derived from the original sample in which no missings were present with the aim of detecting which, among D_1, D_2, D_3, D_4 , could best recover the original cluster structure.

4. RESULTS AND CONCLUSIONS

For each coefficient and for the case in which values have been randomly assigned to replace missing data, table 1 and 3 report the overall mean bias, obtained on the values below the principal diagonal in the matrix containing the mean dissimilarity bias for each pair of units (in the 400 replicates), and its range. In symbols, denoted by $\bar{b}_{ij}$ the generic mean dissimilarity bias for units i and j , the overall mean bias is given by:

⁴“Open” shapes are “baking pans”, “large bowls” and “bowls”; “closed” shapes are “glasses”, “jars” and “little jars”.

$$\bar{b} = \frac{\sum_{i,j} \bar{b}_{ij}}{n(n-1)/2}$$

for $i = 1, \dots, n-1$, $j > i$ and $n = 122$.

Table 2 and 4 report the percentage of cases each index overestimates or underestimates the distance between two units for both 49% and 25% missing values.

As clearly appears from the tables, D_1 has the least mean bias for 25% missing values but not for 49% missings; in that case it is slightly more biased than D_4 ; its bias has however the minimum range in both situations and so it seems to be the most reliable index among those considered. It half the times overestimates and half times underestimates the “true” distance consistently with what aforesaid in section 2. D_2 and D_3 more often give positively biased estimates but D_2 is more biased than D_3 .

Table 1.

Results of the simulation with 49% missing values.

Dissimilarity index	Mean bias	Minimum value of bias	Maximum value of bias
D_1	0.0044	-0.0857	0.1093
D_2	0.2040	0	0.5149
D_3	0.1387	0	0.3613
D_4	0.0034	-0.30	0.4022

Table 2.

Proportion of positive and negative values of bias for each index with 49% missing values.

	D_1	D_2	D_3	D_4
Positive	58	100	100	51
Negative	42	0	0	49

Table 3.

Results of the simulation with 25% missing values.

Dissimilarity index	Mean bias	Minimum value of bias	Maximum value of bias
D_1	-0.0008	-0.0859	0.0943
D_2	0.1108	-0.0109	0.2816
D_3	0.0649	0.0147	0.1716
D_4	0.0371	-0.15	0.2922

Table 4.

Proportion of positive and negative values of bias for each index with 25% missing values.

	D_1	D_2	D_3	D_4
Positive	49	100	99	63
Negative	51	0	1	37

The presence of 100% positive values in table 2 for the two indices can be explained considering that the quoted values are expected values over 400 replicates and that positive biases are larger than negative ones. D_4 gives a small mean bias but has a large bias range; so replacing missings by randomly selected character states does not seem to be a correct way to handle missing values.

In conclusion Gower's dissimilarity coefficient in its original version is the one giving the least biased results even if it has not a constant behaviour as to the sign of the bias.

Cluster analysis⁵ for D_1, D_2, D_3, D_4 seems to confirm the results previously stated; Gower's index succeeds in recovering the original cluster structure better than all the other indices although also D_4 works fairly well particularly with 49% missing values. A previous study on the relationships between the variables showed that the characters considered are significantly connected; in this context the good performance of D_4 strengthens the assumption of random distribution of missing values. Further checks are however going on.

⁵ Cluster analysis has been performed using SAS/STAT CLUSTER PROCEDURE.

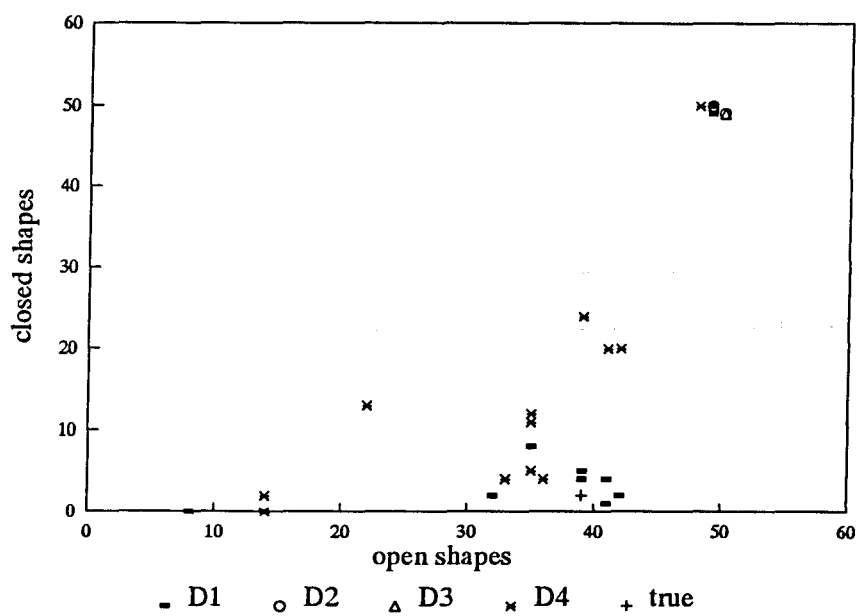


Figure 1.

Cluster results for each index with 49% missing values.

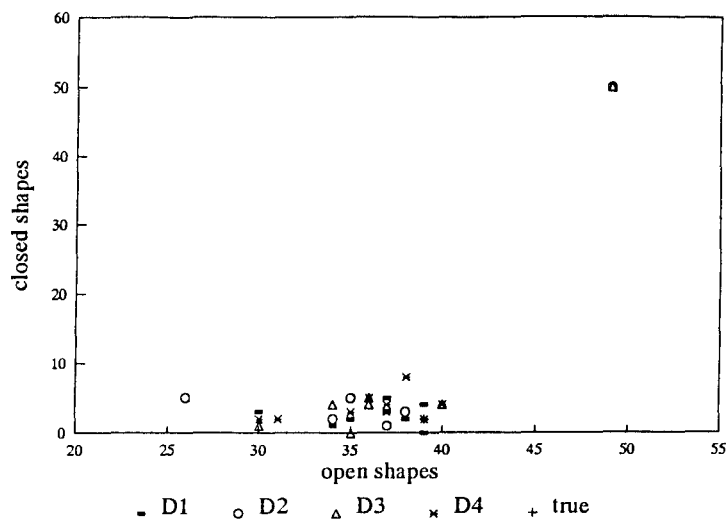


Figure 2.

Cluster results for each index with 25% missing values.

D_2 and D_3 completely fail in detecting the existing clusters, always giving rise to a single big cluster containing all the units. When the proportion of missing values decreases D_1, D_2, D_3, D_4 give similar results. All samples studied gave results similar to the ones shown for one of them in pictures 1 and 2 (only the count of the units of one cluster is shown, the one for the other cluster may be determined as difference); to avoid congestion on the pictures only a small subset of the simulation results has been represented.

REFERENCES

- [1] Beale, E.M. and Little, R.J.A. (1975). "Missing values in multivariate analysis". *Journal of the Royal Statistical Society*, **B 37**, 129–145.
- [2] Brothwell, D.R. and Krzanowski, W.J. (1974). "Evidence of biological differences between early British populations from Neolithic to Medieval times as revealed by eleven commonly available cranial vault measurements". *Journal of Archaeological Sciences*, **1**, 249–260.
- [3] Capitano, A., Mignani, S. and Montanari, A. (1992). "Le elaborazioni statistiche". In: S. Santoro Bianchi (ed.) *Castel Raimondo. Scavi 1988–90. II I materiali*. L'Erma di Bretschneider, Roma (in press).
- [4] Doran, J. and Hodson, F. (1975). *Mathematics and computers in archaeology*. Edinburgh University Press. Edinburgh.
- [5] Gordon, A.D. (1981). *Classification*. Chapman and Hall. London.
- [6] Gower, J.C. (1971a). "A general coefficient of similarity and some of its properties". *Biometrics*, **27**, 857–872.
- [7] Gower, J.C. (1971b). "Statistical methods of comparing different multivariate analyses of the same data". In: *Mathematics in the archaeological and historical sciences*. (F.R. Hodson, D.G. Kendall, P. Tatu eds.). University Press. Edinburgh.
- [8] Gower, J.C. and Legendre, P. (1986). "Metric and euclidean properties of dissimilarity coefficients". *Journal of Classification*, **3**, 5–48.
- [9] Guermandi, M.P. (1990). "La ceramica grezza. Analisi computerizzata e classificazione: problemi di metodo". *Antichità Altoadriatiche*, **XXXVI**.
- [10] Krzanowski, W.J. (1988). *Principles of multivariate analysis*. Clarendon Press. Oxford.
- [11] Rizzi, A. (1985). *Analisi dei dati*. Nis. Roma.
- [12] Seber, G.A.F. *Multivariate observations*. John Wiley and Sons. New York.

- [13] **Shennan, S.** (1988). *Quantifying archaeology*. Edinburgh University Press. Bristol.
- [14] **Sneath, P.H.A.** and **Sokal, R.R.** (1973). *Numerical taxonomy*. Freeman and Co. San Francisco.
- [15] **Wishart, D.** (1978). "Treatment of missing values in cluster analysis". In: *Compstat 1978* (C.A. Corsten, J. Hermans eds.). Physica-Verlag. Wien.

BUILDING A KNOWLEDGE BASE FOR CORRESPONDENCE ANALYSIS¹

M^a CARMEN BRAVO LLATAS[†]

Universidad Complutense de Madrid

This paper introduces a statistical strategy for Correspondence Analysis. A formal description of the choices, actions and decisions taken during data analysis is built. Rules and heuristics have been obtained from the application of this technique to real case studies.

The strategy proposed checks suitability of certain types of data matrices for this analysis and also considers a guidance and interpretation of the application of this technique. Some algorithmic-like rules are presented and specific criteria are given for application.

This strategy has been implemented in a statistical knowledge-based system prototype using hypertext and rules.

Key words: Statistical Strategy, Correspondence Analysis, Knowledge-Based System, Knowledge Enhancement System, Hypertext.

1. This paper has been partially supported by Spanish CICYT Project INF91-74.

[†] M^a Carmen Bravo Llatas. Centro de Proceso de Datos. Universidad Complutense. Avda. Paraninfo s/n. 28040 Madrid. Spain.

—Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.

—Acceptat el gener de 1994.

1. INTRODUCTION

The increasing number of user friendly statistical packages together with the daily use of Statistics by non-experts may lead to its misuse. This is the main reason why statistical knowledge-based systems or statistical knowledge enhancement systems (Hand (1987, 1990)) are necessary. A first aim of these systems should be to help researchers in data analysis tasks providing them with guidance to choose a technique for analysis, to perform a successful analysis and to interpret the obtained results. Another goal of these systems should be the knowledge improvement in a statistical technique both by non-experts and experts in Statistics.

When dealing with the creation of a statistical knowledge base for a system, attention could be paid to two types of knowledge: the statistical theoretical background about a technique or a set of techniques and, the heuristics and rules that data analysts adopt in their work. In this paper, we concentrate on the latter and propose a statistical strategy for simple Correspondence Analysis. It also can be applied when more than two variables are present. The multiple Correspondence Analysis can be reduced to the two-way analysis partitioning the variable set into two groups, each of them composed of independent variables. In this case, a submatrix of the Burt matrix is analysed, with different row and column categories, Lebart *et al.* (1984), Benzécri and Maïti (1990), Benzécri (1990).

A statistical knowledge enhancement system prototype has been developed combining these two types of knowledge. Rules and hypertext have been the tools employed to represent statistical knowledge. A deeper view of this system may be seen in Ferrán (1991), who organizes knowledge about Correspondence Analysis in order to implement it in a hypertext, and Ribes (1991), who describes the prototype. In Bravo (1991) a methodology for building a statistical strategy is proposed, and a part of the statistical strategy presented here is outlined.

For further information on these systems where hypertext is one tool employed to represent knowledge, see Hand (1987, 1990). Concerning strategies for Correspondence Analysis, Jambu (1991) gives some rules of action for application.

The rest of the paper is organized as follows: In Section 2, the statistical strategy concept is defined and a general view of Correspondence Analysis strategy is outlined. In Section 3, determining suitability of matrix substrategy is described, with emphasis on variables and categories. Two trees representing the main points of the strategy are also presented. Sections 4 to 6 show strategies for selection of the number of axes to be retained, interpretation based on one

space and simultaneous interpretation. Some algorithms with specific criteria are presented and some ideas on how to implement them in the prototype are outlined. Section 7 presents some information about the implementation of the prototype together with a short example.

2. STATISTICAL STRATEGY

Hand (1986) defines a statistical strategy as a formal description of the choices, actions and decisions taken during data analysis. It is not merely a sequence of actions but a recipe to choose an action as a result of an external event.

There are two well-known types of statistical strategies depending on their main purpose. One is related to the selection of some statistical techniques suitable for analysing a specific data set and the other provides a guidance to the analysis and interpretation of results, once a statistical technique has been chosen. For the development of our strategy, Correspondence Analysis is assumed to be the technique chosen to analyse data.

As many authors agree (Lubinsky and Pregibon (1988), Thisted (1986), Olford and Peters (1986), Pregibon (1986)), a statistical strategy may be represented by a set of trees where every node is a different representation of data. Furthermore, data analysis consists in navigating through this set where any movement from one node to another is done by means of any transformation, such that: a test, a complicated procedure, a researcher's preference or even data themselves.

Hand (1986), Huber (1986) and Pregibon (1986) consider that data analysis does not merely consist in data cleaning followed by analysing but in these four stages:

- (a) **Objective formulation.** In the application context researcher's objectives are formulated and suitable techniques are considered. Dependent and independent variables are determined as well. In our case, there is no need to select a technique because Correspondence Analysis has been chosen for application. Even though, we will determine if our main aim is basically descriptive or scores on axes will be used in a subsequent analysis; select active and illustrative variables; combine categories to form illustrative ones; or make a decision for missing value categories.
- (b) **Formalization.** In this stage, the translation of objectives to statistical terms is carried out.

- (c) **Numerical processing.** It consists in modeling test, data cleaning and selection of data transformations. Undoubtedly this is the most valuable part a knowledge enhancement system should contain. At this stage, we would give criteria or algorithms to select axes, interpret them or determine well represented categories. Some data cleaning is also included such as, for example, if only a few categories explain the maximum inertia axis we should reconsider the matrix to be analysed.
- (d) **Interpretation of results.** In subject-matter context, we will interpret retained axes and proximities between points projected onto axes.

Although these stages are followed in a sequential way it is not a rigid structure. Iterative or recursive cycles may occur in numerical processing. Besides, from numerical processing and from the interpretation it is possible to go back to the first stage to look for another objective. Similarly, if disparity between objectives and results is encountered a reformulation of an old objective may be done.

Note also that interaction between the system and the user may take place in the four stages. Whenever it is possible, the user's preferences should be taken into account to choose the way analysis should proceed.

As in most data analysis tasks, I believe that Correspondence Analysis strategy has hierarchical structure that can be represented by a tree. It is subdivided in two main interrelated parts connected to the two types of statistical strategy: determining suitability of the matrix to analyse and, guiding and interpreting the analysis. The interrelation between these two hierarchical parts can be explained by displacements from a node of one branch to the other branch. The first strategy is composed of two substrategies: one forms the data matrix to be analysed and the other determines when the analysis is appropriate for this data set. The second strategy is composed of four substrategies: selection of the number of axes to be retained, interpretation based on one space, simultaneous interpretation and interpretation of interesting and illustrative categories. These substrategies are also broken down into some substrategies and so on.

3. DETERMINING SUITABILITY OF MATRIX

3.1. Determining data matrix

Since Correspondence Analysis is an exploratory data analysis technique revealing interrelations between two spaces of categories, we shall usually use it

to analyse large data matrices of positive numbers. This is motivated by the difficulty of drawing significant information just from viewing these matrices.

This substrategy, presented in figures 1 and 2, deals with the configuration of data matrix for analysis, with regard to variables and categories. It also implies some data cleaning and selects illustrative variables or categories to be projected onto the axes.

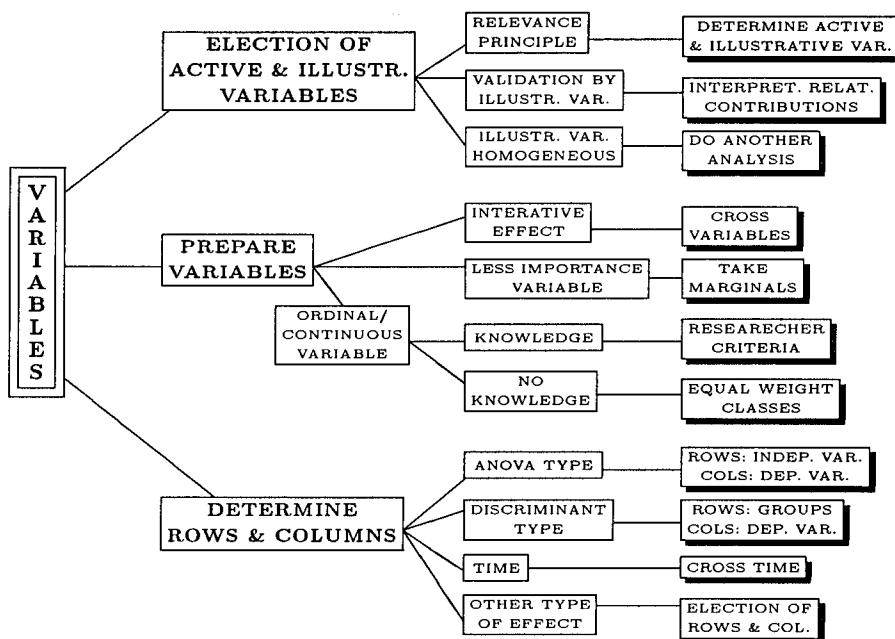


Figure 1.
Statistical strategy for variables.

3.1.1. VARIABLES

Globally, the substrategy related to variables deals with the selection of active and illustrative variables, as well as the preparation of these variables and the determination of row and column spaces of the matrix to be analysed.

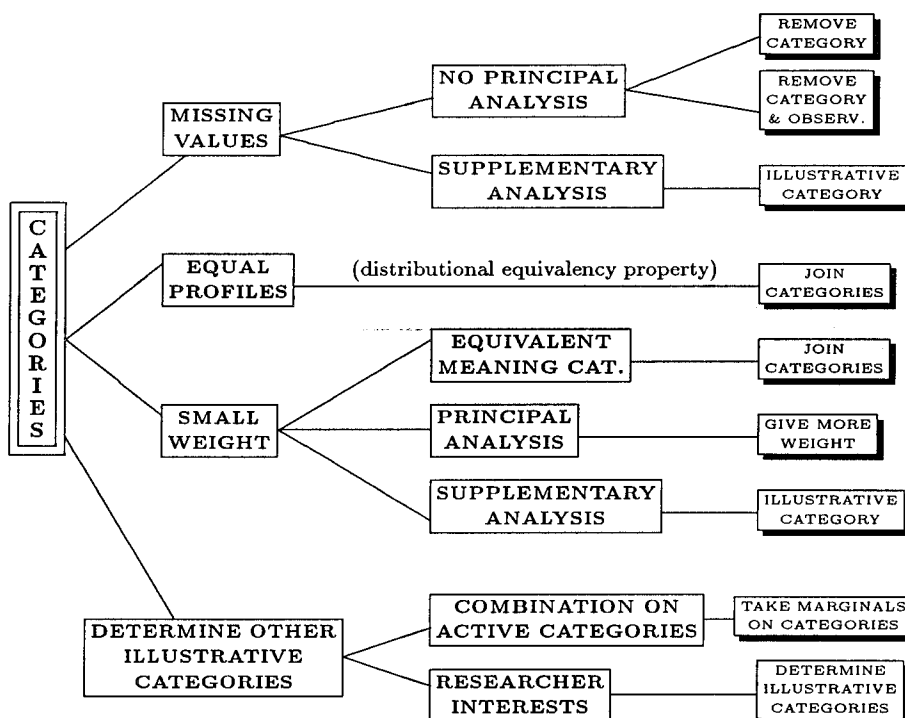


Figure 2.
Statistical strategy for categories.

Taking into account the application context, the steps below should be followed to specify that matrix:

- (a) Select active and illustrative variables following the Relevance Principle to preserve homogeneity. In Correspondence Analysis, we require not only measure homogeneity of variables, but also their semantic homogeneity. The Relevance Principle should be applied to keep for principal analysis those variables interrelated in a specific point of view. Not being illustrative variables in the analysis, they could be used as a validation test for principal analysis, Lebart *et al.* (1984) and Benzécri *et al.* (1990).
- (b) If there are continuous variables, categorize them following the researcher's knowledge or any other criteria as, for example, subdividing in similar weight classes, Tekaiia *et al.* (1990).

- (c) If there are ordinal variables, consider their categorization if the number of possible values is greater than a fixed number, i.e., four or five. It is convenient to follow the researcher's interests and not to unbalance the weight of classes. Even, extreme value categories could be joined when it is judged reasonable.
- (d) If more than two active variables are present:
- ▶ determine sets of variables with interactive effects on the others. Afterwards, variables in each of these sets should be crossed to create a new active one, which replaces those. Commonly, a temporal variable will be crossed with those of interest in evolution, Benzécri *et al.* (1990), Alawieh (1990), Maravalle (1990).
 - ▶ determine variables of less important to take marginals on them.
 - ▶ determine row and column spaces of the matrix to be analysed, which is built extracting the corresponding submatrices of the Burt matrix.
- (e) Even when Correspondence Analysis is an exploratory technique, sometimes it is useful to apply it to data matrices traditionally analysed by other confirmatory techniques, such as, for example, analysis of variance, non-parametric multiple comparison tests or discriminant analysis. Although these techniques did not reveal significant effects, Correspondence Analysis may show important interrelations among variables, which are easy to check with the data matrix, Benzécri *et al.* (1990), Maïti (1989), Tekaïa *et al.* (1990).

3.1.2. CATEGORIES

This substrategy considers the treatment given to special types of categories such as, missing value, similar profile and small weight and determines illustrative categories. It is subdivided in these parts:

- (a) For missing value categories, the strategies that could be followed are: define them as illustrative categories, keep them in principal analysis or remove them from principal analysis. In case of elimination of categories from principal analysis we also could consider the elimination of experimental units with missing values on them.
- (b) In each space, similar profile categories will be joined, as a consequence of distributional equivalence property of χ -squared distance, Lebart *et al.* (1982).

- (c) If there are small weight categories, reconsider their presence in principal analysis. For each of them, three strategies may be followed: join the category to another of similar meaning, define the category as an illustrative one or give more weight to it.

A category could be considered of small weight when its weight is lower than $1/(\alpha \times p)$, p being the number of categories corresponding to the variable and α a real number greater than 1, e.g., 3.

- (d) Define other illustrative categories taking marginals on combinations of active categories or considering the researcher's interests.

3.2. Determining suitability of analysis

The aim of Correspondence Analysis is the distribution of total inertia in a few axes. Although it could be applied in the independent case, it is more valuable when rows and columns of the matrix are not independent. In this case, the analysis tries to explain interrelations among row and column points. Note, however, that if the distribution of total inertia on axes were homogeneous, Correspondence Analysis would not be the best suitable descriptive technique for these data matrices, Lebart *et al.* (1984).

4. SELECTION ON THE NUMBER OF AXES TO BE RETAINED

The selection of the number of axes to be retained depends on our main purpose, if it is merely descriptive or if in subsequent analysis the original variables are going to be replaced with scores on the axes.

The proposed strategy, which is not based on a statistical test, tries to get a compromise among well-known criteria. These criteria are: a) to retain only a few axes, b) to keep as much of total inertia as possible, c) to avoid that the inertia of any retained axis being similar to the inertia of any other non-retained axis, d) to retain those axes whose proportion of inertia is greater than the 'inertia mean'.

Algorithm below combines specific versions of those aspects:

Algorithm 1.

Step 0. Initialization:

 $\Lambda(=0)$: inertia explained by retained axes.

$k(=0)$: number of retained axes.

$\epsilon (= 0.25|0.20)$: coefficient to determine a big step between two eigenvalues.

m, s : degrees of freedom of rows and columns.

k' : maximum number of axes to be retained.

λ_k : k^{th} eigenvalue or inertia explained by k^{th} axis.

$\Sigma\lambda_l$: total inertia minus one.

Step 1. $k = k + 1$

$$\Lambda = \Lambda + \lambda_k$$

Step 2. If $\left(\left(\frac{\Lambda}{\Sigma \lambda_l} \geq 0.80 \text{ and } |\lambda_{k+1} - \lambda_k| > \epsilon \times \lambda_k \right) \text{ OR } (k \geq \min(m, s)) \right)$

then go to **Step 3**.

else go to *Step 1*.

Step 3. If $k > k'$ then

Reinitialization: $k = 0$

$$\Lambda = 0$$

3.1. $k = k + 1$

If $\lambda_k > \frac{\sum \lambda_l}{\max(m, s)}$ then $\Lambda = \Lambda + \lambda_k$
Go to **3.1**.

else $k = k - 1$

Go to Step 4.

else go to **Step 4**.

Step 4. Results:

k : number of axes to be retained.

$$\frac{\text{Lambda}}{\Sigma \lambda_l}$$
 : proportion of inertia explained by retained axes.

End.

In this algorithm, all computations are supposed to start at the second eigenvalue because the first one is always one and it doesn't make sense.

5. ONE SPACE INTERPRETATION

This strategy deals with the fact of interpretation based on one space. Usually, for each retained axis the row or column spaces will be chosen to interpret it. Afterwards, the projections on the axis of the other space points will be interpreted, taking into account that badly represented categories over the retained axes set should lead us to reconsider their presence in the analysis.

5.1. Interpretation of an axis

The algorithm introduced for the interpretation of an axis provides criteria to select the categories that explain mostly the axis according to the category absolute contributions. It also determines the relative importance of these categories in this explanation and locates the category projections on both sides of the axis. Note that we should also include those categories whose absolute contributions are similar to one of those of the explicative categories. Lebart *et al.* (1982), define the absolute contribution of a point to an axis as the proportion of total inertia of the axis explained by the point.

The algorithm also considers the case when very few categories explain the maximum inertia axis. In this case, the matrix to be analysed should be reconsidered, according to the following substrategies for those categories: join them to another category of similar meaning, remove them from principal analysis or give them less weight.

Once chosen one of the two spaces, the algorithm proposed to interpret an axis is as follows:

Algorithm 2.

Step 0. Initialization:

LIMIT ($= 1/n$) : limit to retain an explicative category.

n : number of categories in the space.

$k(= 1)$: number of categories explaining the axis.

$\theta(= 0.5)$: coefficient to detect similarities between absolute contributions.

$\alpha(= 0.20)$: coefficient to obtain N .

N : maximum number of categories that explain the first axis prior to reconsidering the matrix to be analysed.

$N = \text{INT}[n \times \alpha]$ if $n < 20$; $N = 4$ if $n \geq 20$.

$CABS_i (i = 1, \dots, n)$: absolute contribution of i^{th} point to the axis.
 $PROJ_i (i = 1, \dots, n)$: projection of i^{th} point onto the axis.

Step 1. Order categories decreasingly according to their absolute contributions:

$CABS_1 > \dots > CABS_n$
 Verify that $CABS_1 > LIMIT$
 $SIGN = SIGN(PROJ_1)$

Step 2. $k = k + 1$

If $CABS_k < LIMIT$ then *go to Step 5.*

Step 3. $I = 1$

3.1. If $(CABS_1 - CABS_k) < \theta \times LIMIT \times I$ then the relative importance of the two points in the construction of the axis is I -similar.
Go to Step 4.

else $I = I + 1$
Go to 3.1.

Step 4. If $SIGN(PROJ_k) = SIGN$ then the k^{th} point is projected on the same semiaxis of the point of greatest importance.

else k^{th} and the point of greatest importance are in opposite semiaxes.
Go to Step 2.

Step 5. $k = k - 1$

If $(CABS_{k+1} - CABS_k) < (\theta/2) \times LIMIT$
 then readjust the number of categories, k , due to similarity between absolute contributions.

Step 6. If it is the first axis and $k < N$ then reconsider the configuration of the matrix following the strategy given above the algorithm.

Go to End.

else *Results:*
 the first k categories explain the axis.
 their relative importance is given in **Step 3.**
 their relative position on the axis is given in **Step 4.**
End.

The criteria specified in the algorithm may be substituted as, for example, the limit for retaining categories could be $LIMIT = \mu/(n - 1)$, being $\mu = 0.75$ or any other constant smaller than one.

Practically speaking, in the prototype built, the set of categories that explain an axis is divided only into three groups of importance. The interval between the limit to accept explicative categories and the greatest absolute contribution among categories is divided into three equal size subintervals.

5.2. Representation of categories on an axis

The algorithm proposed for interpreting categories on an axis provides criteria to select well represented categories on the axis according to the category relative contributions. It also positions the projections of well represented categories on both sides of the axis and considers the selection of categories whose relative contributions are similar to one of those of the well represented categories. Beforehand, the presence of badly represented categories in the retained axes set should lead us to reconsider their presence in the analysis. According to Lebart *et al.* (1982) relative contribution of an axis to a point is the squared cosine of the angle formed by the point with the axis.

Note that retaining a category to interpret an axis (based on absolute contribution), position it on a semiaxis and not to represent this category onto it because of a low relative contribution, may be confusing. It may be due to the fact that this category would participate in the explanation of some other axes, Lebart *et al.* (1984).

Once interpreted an axis by the categories of one space, the algorithm for interpretation of the other space projections onto it is given by:

Algorithm 3.

Step 0. Initialization:

LIMIT(= 0.40): minimum value for relative contributions to select a well represented category.

CAT: the other space category of greatest absolute contribution.

SIGN: sign of the projection of CAT.

n : number of categories in the space.

$k(= 1)$: number of well represented categories on the axis.

$\text{CREL}_i(i = 1, \dots, n)$: relative contribution of axis to i^{th} point.

$\text{PROJ}_i(i = 1, \dots, n)$: projection of i^{th} point onto the axis.

$\alpha(= 0.05)$: coefficient to detect similarities between relative contributions.

Step 1. Order categories decreasingly according to their relative contributions:

$$\text{CREL}_1 > \dots > \text{CREL}_n.$$

Verify that $\text{CREL}_1 > \text{LIMIT}$.

Step 2. $k = k + 1$

If $\text{CREL}_k < \text{LIMIT}$ then go to **Step 5**.

Step 3. If $\text{SIGN}(\text{PROJ}_k) = \text{SIGN}$ then the k^{th} point and CAT are on the same semiaxis.

else the k^{th} point and CAT are in opposite semiaxes.

Step 4. $I = 1$

4.1. If $(1 - \text{CREL}_k) < \alpha \times I$ then the level of good representation of k^{th} point is I .

Go to **Step 2**.

else $I = I + 1$

Go to 4.1.

Step 5. $k = k - 1$

If $(\text{CREL}_{k+1} - \text{CREL}_k) < \alpha/2$ then readjust the number of categories, k , diminishing or augmenting it.

Step 6. Results:

The first k categories can be explained by the axis.

The way they are correlated with the axis and other space categories is given by relative contributions and by **Step 3**.

The levels of good representation are given in **Step 4**.

End.

All the specified criteria in this algorithm may be modified. For example, the limit in order to select well represented categories, could be related to the global quality of representation of categories in the retained axes set, provided that these qualities are acceptable. This limit could be $\text{LIMIT} = Q/\text{NRA}$ or $\text{LIMIT} = 1/\text{NRA}$, where Q is the global quality of representation of the category and NRA is the number of retained axes.

In our prototype, the set of well represented categories is divided into three groups whose relative contributions lie in $[0.40, 0.60)$, $[0.60, 0.80)$ and $[0.80, 1]$, respectively. Considering that the limits of these groups depend on the dimensionality of the data, they could be changed depending on the number of inertia axes in each problem.

6. SIMULTANEOUS REPRESENTATION AND INTERESTING AND ILLUSTRATIVE CATEGORIES INTERPRETATION

Simultaneous interpretation is done by explaining proximities between projected points of both spaces onto axes. It should be taken into account their quality of representation and weights, as well as Euclidean distances between points and from the origin, which is simultaneously the center of gravity of row and column points.

For well represented categories on an axis, the proximity of a row (column) point to a column (row) point is encountered when the component of this column (row) in the row (column) profile has a higher value than the same component in the average row (column) profile. The row (column) profile components with lower values than the corresponding components in the average row (column) profile are the column (row) categories which are projected on different semiaxis than the row (column) category. Similarly, two row (column) points are close to each other when their profiles are similar and are located on different semiaxes when they have very different profiles. In fact, the proximity between two row (column) points is explained by higher values in their profiles than the average row (column) profile of the components corresponding to the nearest column (row) points, and lower values than the average row (column) profile of components corresponding to the other semiaxis column (row) points.

The above interpretations will be more accurate when points are farther from the origin, better represented, and have more weight, Lebart *et al.* (1984). It would be interesting to investigate ways of combining these criteria in order to obtain a measure of the relative importance of these interpretations.

For categories that are illustrative or of interest for the researcher, the axes in which they are well represented have to be selected. Proximities to other well represented categories on those axes may be interpreted as the above paragraphs suggest. Section 5 also may be applied to interpret projections onto axes.

In the prototype built, simultaneous representation of well represented points on axes is shown, and some guidelines are given to the user for interpretation.

7. IMPLEMENTATION OF THE PROTOTYPE. AN EXAMPLE

A prototype for PC MS-DOS has been built. Both the statistical strategy and the theoretical background of the technique have been included. There is

a continuous interaction between the user and the prototype, allowing them to participate in matrix building and analysis and to learn how the analysis is performed and interpreted by the prototype. The user's knowledge about the technique also can be improved using the hypertext and explaining commands.

Tools employed in the prototype are: a) KES^R II² (Knowledge Engineering System II, Release 2.2), an expert system tool, b) HIPER1, the hypertext, written in C (Mosich *et al.*, 1988), c) ANCO, a FORTRAN program that performs the analysis. KES^R II is composed of a rule-based knowledge base and a backward chaining inference engine.

KES^R II helps us to build a menu-driven interface between the user and the prototype. It displays two main menu screens: One to determine the matrix for analysis and another to perform the analysis; both of them include several submenu screens and questions. It is possible to ask for explanation of technical terms about those questions or reasons leading to them by the EXPLAIN and WHY commands.

Both menu screens allow for linking to HIPER1, which contains the theoretical background of the technique in textual cards. HIPER1 allows the navigation through its knowledge base in a non-sequential way, letting the user to establish interrelations among statistical terms according to their needs. Graphical aids would make more understandable some concepts and strategies, and could be included in later versions.

During matrix construction, the prototype tries to obtain some metainformation available of data. It suggests the application of Relevance Principle to select active and illustrative variables; inquires into dependencies among variables (anova, discriminant, time effect, others) and interactive effect of some variables on the others; and asks for: the user's interests in categorization of ordinal and continuous variables, how to deal with non-complete cases, missing value and small weight categories, and the determination of some illustrative categories as for example marginals and others of user's interest.

Once built the data matrix, the analysis is performed by ANCO, which creates some communication files with KES^R II, which applies the strategy.

Having proved the success of the conjunction of hypertext, rules and programming techniques in the construction of the prototype, we suggest the development of a more powerful system. We would suggest also the use of a hypertext

²KES^R II (c) (1985) Software Architecture & Engineering, Inc., distributed and supported by Sperry Corp.

as the interface between the user and the system and the incorporation of tools that would allow the system to get as metainformation of data as possible.

The following short example illustrates the use of the prototype. It concerns a study of the influence of body position in some respiratory measures, taking into account sex, age, body mass index (bmi), smoking and exercise. The initial menu screen is shown in figure 3.

What do you want to do?

1. Identification of principal and illustrative variables. (Relevance Principle).
2. Identification of rows and columns.
3. Preparation of variables. (Type of variables, interactive effects, less important variables).
4. Determination of categories. (Missing value, equal profile, small weight).
5. Determination of illustrative categories.
6. Determination of meaningful categories.
7. Enter into help menu. (Hypertext).
8. Perform analysis.
9. Exit.

=?

Figure 3.
Menu for determining the matrix for analysis.

The matrix for analysis will be determined by clicking points displayed in this menu. Through point 1, the prototype asks for selection of active variables among the observed ones, according to the Relevance Principle: *sex*, *age*, *bmi*, *position*, *smoking*, *exercise* and *some respiratory parameters*. *Smoking and exercise* are considered unimportant. Point 2, see figure 4, finds a discriminant variable, *position*; variables characterizing the groups, *respiratory parameters*; and other variables influencing on these, *sex*, *age* and *bmi*. Point 3 helps to detect an interactive effect of *sex* and *age* on *respiratory parameters*; determines the categorization of continuous variables in three classes: *age* according to the user's interests, *bmi* and *respiratory parameters* in equivalent weight categories, defining different *bmi* limits for each *sex*; and variables of less importance, *smoking*, *exercise* and *some respiratory parameters*. Point 4 considers the presence of missing value, small weight and equal profile categories.

Type of problem:

1. Discriminant.
2. Anova.
3. Time effect.
4. Other interrelations.

=? EXPLAIN 1

A problem is of discriminant type if there are some groups that have to be characterized by some parameters.

=? 1

Which variable identifies your groups?

- | | |
|--------------|-----------|
| 1. Sex. | 7. Fvc. |
| 2. Age. | 8. Pef. |
| 3. Bmi. | 9. Mef50. |
| 4. Position. | 10. Pif. |
| 5. Smoking. | 11. Tpef. |
| 6. Exercise. | 12. Ttot. |

=? 4

Which variables characterize your groups?

- | | |
|--------------|-----------|
| 1. Sex. | 7. Fvc. |
| 2. Age. | 8. Pef. |
| 3. Bmi. | 9. Mef50. |
| 4. Position. | 10. Pif. |
| 5. Smoking. | 11. Tpef. |
| 6. Exercise. | 12. Ttot. |

=? 7&8&9&10&12

Is there any other variable that incides in your characteristic variables?

- | | |
|--------------|-----------|
| 1. Sex. | 7. Fvc. |
| 2. Age. | 8. Pef. |
| 3. Bmi. | 9. Mef50. |
| 4. Position. | 10. Pif. |
| 5. Smoking. | 11. Tpef. |
| 6. Exercise. | 12. Ttot. |

=? 1&2&3

Figure 4.
Identification of rows and columns of the matrix.

Point 5 defines some illustrative categories, such as marginals on *sex and age and other categories*. Finally, point 6 determines meaningful categories to be interpreted. This menu helps to determine the matrix for analysis. Here, it is the submatrix extracted from the Burt matrix with rows the categories of *position, bmi* and the combination of *sex and age*, and columns, the categories of the *respiratory parameters* of interest in principal analysis.

Figure 5 displays the hypertext, which may be called from the two main menu screens. Once built the matrix for analysis, point 8 of the initial menu leads to the second main menu screen shown in figure 6. Point 1 of this second menu shows the histogram of proportion of inertia explained by the axes and gives some explanation about the number of axes that are retained in this step. Other points are self explanatory enough. Figures 7 and 8 display part of the output of the session, but not the dialogue between the user and the prototype. Output in figure 7 is related to points 2 and 3 of this second menu, while figure 8 is related to point 4.

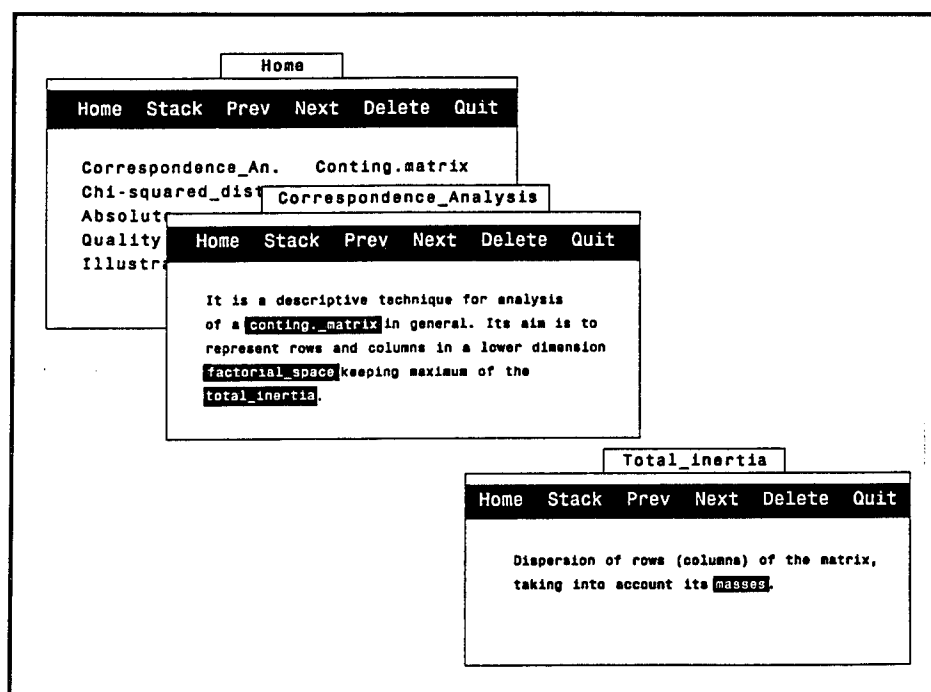


Figure 5.
The hypertext.

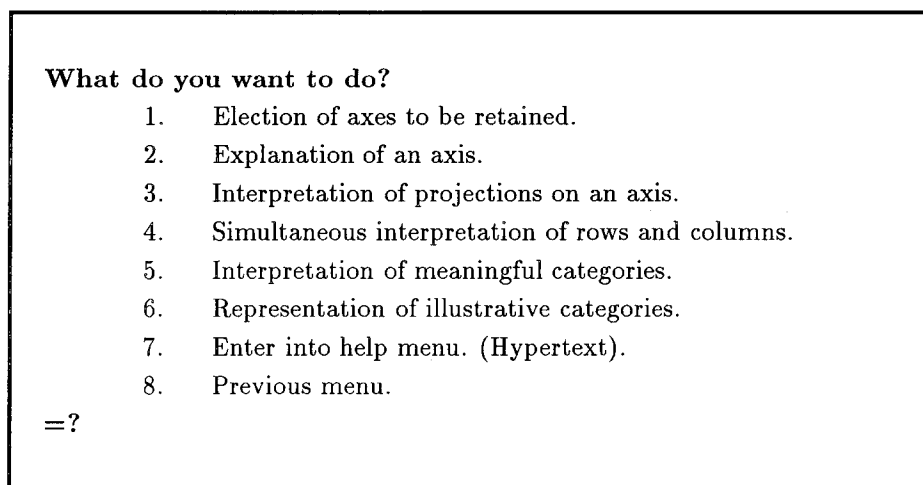


Figure 6.
Menu for performing the analysis.

8. CONCLUDING REMARKS

The main goal of developing statistical strategies is to build statistical knowledge enhancement systems that not only lead to a better use of data analysis techniques but also to the improvement of user's knowledge about them.

The developed strategy and the organized knowledge about Correspondence Analysis have been implemented in a knowledge enhancement system prototype. Next step should be the transformation of this prototype into a modular system to which new modules related to other techniques could be added subsequently. This modular system would represent a valuable tool to analyse data correctly and to learn or improve users' knowledge in data analysis techniques.

According to our experience, these systems not only help researchers in other fields to perform analysis but also benefit Statistics and data analysts. Statistics is protected against misapplication of techniques to data and data analysts may improve their own knowledge about certain techniques.

(Point 2. Determined an axis and row points to explain it)

First axis is explained by the categories:

mc^- , mc^+ , $pos1$, $pos2$, and $pos3$

Below, groups of categories by their importance in determining the axis (absolute contributions):

Maximum importance categories: mc^+ , $pos1$, and $pos3$

Medium importance categories: mc^- , and $pos2$

Moreover, the categories: mc^+ , $pos1$, and $pos2$

are opposed in this axis to the categories: mc^- , and $pos3$

Ready for command:

(Point 3.)

Being explained this axis by row categories, column categories will be explained.

The categories well represented onto first axis are:

fv^- , $fv^=$, fv^+ , pef^- , $pef^=$, pef^+ , $m50^-$, $m50^=$, $m50^+$, pif , pif^+ , t^- ,
and t^+

Below, groups of categories by levels of good representation (relative contributions):

Rather well represented categories: fv^- , pef^- , $m50^-$, $m50^+$, and t^-

Quite well represented categories: $fv^=$, fv^+ , pef^+ , pif^+ , and t^+

Fairly well represented categories: $pef^=$, $m50^=$, and pif^-

Moreover, the categories: fv^- , pef^- , and t^+

are projected onto positive semiaxis being correlated with the categories: mc^+ , $pos1$, and $pos2$

and the categories: $fv^=$, fv^+ , $pef^=$, pef^+ , $m50^-$, $m50^=$, $m50^+$, pif^- , pif^+ , and t^-

are projected onto negative semiaxis being correlated with the categories: mc^- , and $pos3$

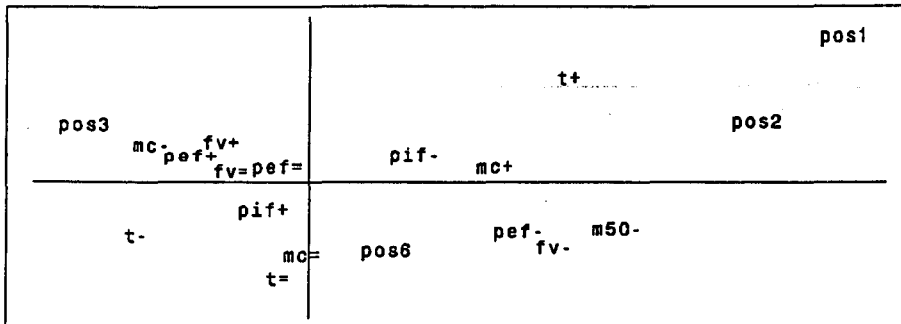
Ready for command:

Figure 7.

Parts of the output displayed by the prototype during analysis.

(Point 4.)

Categories well represented onto first (horizontal) and second (vertical) axes:



For simultaneous interpretation of well projected categories in each axis take into account:

- One row (column) point is near one column (row) point because the corresponding column (row) component in the row (column) profile is higher than this component in the average row (column) profile; the column (row) points on the other semiaxis have lower values in their row (column) profile components than the corresponding components in the average row (column) profile.
- The proximity between two row (column) points is derived from higher values than the average row (column) profile of the components corresponding to the nearest column (row) points, and lower values than the average row (column) profile of the components corresponding to the other semiaxis column (row) points.
- The more weighty and farther the points are from the origin the more accurate these interpretations are.
- It should be taken into account quality of representation of points.

Figure 8.

Part of the output displayed by the prototype in point 4 of second menu.

ACKNOWLEDGEMENTS

The author wishes to thank Professor José Miguel García-Santesmases, Professor Magdalena Ferrán and technician Belén Ribes without whose help and skills this work would not be done.

REFERENCES

- [1] **Alawieh, A.** (1990). "Répartition en France par Régions de la Collecte pour L'Ensemble des Céréales". *Les Cahiers de L'Analyse des Données*, **XV**, 3, 367–370.
- [2] **Benzécri, J.P.** (1990). "Analyse des Données Biologiques et Pathologie Clinique". *Les Cahiers de L'Analyse des Données*, **XV**, 3, 285–304.
- [3] **Benzécri, J.P.** and **Maïti, G.D.** (1990). "Etude des Réactions de 2000 Sujets à une Thérapeutique". *Les Cahiers de L'Analyse des Données*, **XV**, 4, 463–474.
- [4] **Benzécri, J.P., Maïti, G.D.** and **Kremer, C.** (1990). "Sevrage d'un Traitement Hypnotique ou Anxiolytique et Thérapeutique Substitutive". *Les Cahiers de L'Analyse des Données*, **XV**, 4, 447–462.
- [5] **Bravo, M.C.** (1991). "Estrategia Estadística en el Sistema de Mejora del Conocimiento MACABE". *Trabajo de Investigación. Departamento de Estadística e Investigación Operativa*. Universidad Complutense de Madrid.
- [6] **Ferrán, M.** (1991). "Generación de Bases de Conocimientos para Sistemas de Mejora del Conocimiento. Aplicación al Sistema MACABE". *Trabajo de Investigación. Departamento de Estadística e Investigación Operativa*. Universidad Complutense de Madrid.
- [7] **Hand, D.J.** (1986). "Patterns in Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 355–387.
- [8] **Hand, D.J.** (1987). "A Statistical Knowledge Enhancement System". *The Journal of the Royal Statistical Society, Series A (General)*, Vol. **150**, Part 4, 334–345.
- [9] **Hand, D.J.** (1990). "Emergent Themes in Statistical Expert Systems". *Knowledge, Data and Computer-Assisted Decisions, Nato Asi Series*, Vol. **F 61**, Springer-Verlag, 279–288.
- [10] **Huber, P.J.** (1986). "Environments for Supporting Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 285–294.

- [11] **Jambu, M.** (1991). *Exploratory and Multivariate Data Analysis*. Academic Press.
- [12] **Lebart, L., Morineau, A. and Warwick, K.M.** (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley & Sons.
- [13] **Lebart, L., Morineau, A. and Fénelon, J.P.** (1982). *Traitement des Données Statistiques*. Dunod.
- [14] **Lubinsky, D. and Pregibon, D.** (1988). "Data Analysis as Search". *Journal of Econometrics*, **38**, 247–268.
- [15] **Maïti, G.D.** (1989). "Etude Comparative de L'Efficacité de Deux Médicaments Antalgiques et de Leur Association". *Les Cahiers de L'Analyse des Données*, **XIV**, 2, 157–162.
- [16] **Maravalle, M.** (1990). "Géographie Politique de L'Italie D'Après les Votes à Quatre Scrutins Nationaux de 1983 à 1989". *Les Cahiers de L'Analyse des Données*, **XV**, 2, 191–208.
- [17] **Mosich, D., Shammass, N. and Flamig, B.** (1988). *Advanced TURBO C^R Programmer's Guide*. Willey.
- [18] **Oldford, R.W. and Peters, S.C.** (1986). "Implementation and Study of Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 335–353.
- [19] **Pregibon, D.** (1986). "A DIY Guide to Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 389–399.
- [20] **Ribes, B.** (1991). "Un sistema de Representación del Conocimiento Estadístico. Sistema MACABE". *Trabajo de Investigación. Departamento de Estadística e Investigación Operativa*. Universidad Complutense de Madrid.
- [21] **Tekaïa, F., Sansonetti, Ph. and Claverie, J.M.** (1990). "Estimation du Estade de L'Infection par le VIH Chez les Sujets Séro-positifs". *Les Cahiers de L'Analyse des Données*, **XV**, 3, 261–278.
- [22] **Thisted, R.A.** (1986). "Representing Statistical Knowledge for Expert Data Analysis Systems". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 267–283.

THE USE OF THIRD-ORDER MOMENTS IN STRUCTURAL MODELS

ERIK MEIJER and AB MOOIJAART

Leiden University

Structural models are usually estimated using only second order moments (covariances or correlations). When variables are not multivariate normally distributed, however, methods that also fit higher order moments, such as skewnesses, are theoretically asymptotically preferable. This article reports results from a Monte Carlo simulation study in which estimators that fit both second-order moments and third-order moments are compared with estimators that fit only second-order moments.

Key words: Structural models, robustness, elliptical distribution, asymptotically distribution-free, generalized least squares.

1. INTRODUCTION

Structural models, as most other statistical and data-analytical models and techniques, are usually estimated by fitting only covariances or correlations (e.g. Jöreskog, 1967; Jöreskog & Goldberger, 1972; Bentler, 1983; Browne, 1982,

*Erik Meijer & Ab Mooijaart. Department of Psychometrics and Research Methodology. Leiden University. P.O. Box 9555. NL-2300 RB Leiden. The Netherlands.

—Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.

—Acceptat el maig de 1993.

1984). The reason for this stems from the classical normal theory: if the variables are multivariate normally distributed then all information from the sample—assuming that the means are of no interest—is contained in its covariance matrix.

In practice, however, variables are often not normally distributed and more information can be extracted from the higher-order moments. This implies, for example, that otherwise unidentified models may be identified by fitting higher order moments (see, e.g., Van Montfort, Mooijaart, & De Leeuw, 1987; Mooijaart & Bentler, 1986), and it implies that estimators that use higher order moments are generally asymptotically more efficient (see Kano, Bentler, & Mooijaart, 1993). In particular, Mooijaart (1985) showed that the exploratory factor analysis model is identified if the factors are independent, whereas the usual factor analysis estimators all have the well-known rotation problem.

On the other hand, it is well-known that higher order moments are estimated with considerably less precision than covariances, and it may be questionable, particularly in small samples, whether the use of higher order moments leads to better estimates because of the extra information that is used, or to worse estimates because sample statistics are used that may not be good estimates of their population counterparts.

Besides identification and precision, correctness of inference is a third aspect. If one tests whether a model must be rejected by using a test statistic that is based on strong assumptions regarding the distribution of the variables then the test statistic may reject the (structural) model too often or not often enough, depending on the specific (wrong) assumptions that are made and the specific (correct) distribution of the variables. This applies also to confidence intervals that are used to test whether a parameter may be equal to a certain value (typically 0 or 1). This problem of *robustness* of statistical inference is discussed by several authors; see, for example, Browne & Shapiro (1988), Mooijaart & Bentler (1991), and the references therein. Their results are quite positive in this respect: test statistics of normal theory based estimators are asymptotically correct within a broad range of models, and confidence intervals concerning some classes of parameters in structural models are asymptotically correct, but other confidence intervals are not.

In this article, the focus of attention is on the precision of estimators and correctness of statistical inference, rather than on the identification problem. Estimators that fit third-order moments as well as second-order moments are compared with estimators that fit only second-order moments in a Monte-Carlo simulation study.

2. METHOD

To compare the properties of estimators that use higher-order moments with the properties of estimators that use only second-order moments, a Monte Carlo simulation study was conducted. Samples were generated from a 4-variable, 1-factor, factor analysis model, with loadings 0.8, 0.8, 0.6, and 0.6, respectively. Sample sizes varied from 300 to 10,000. The nonnormalities in the data were generated by drawing the factor from a lognormal distribution and then correcting it to have mean zero, variance one, and prespecified skewness, which was varied between 0.1 and 4.0. The errors were generated from a normal distribution with mean zero and variances 0.36, 0.36, 0.64, and 0.64, respectively. Consequently, the variables all had expectations zero and variances one, and the skewnesses varied —between variables and samples— from 0.02 to 2.0. With each combination of sample size and factor skewness, 300 independent samples were drawn.

The parameters were estimated by *Generalized Least Squares* (GLS) estimators, and by *Linearized Generalized Least Squares* (LGLS) estimators. LGLS estimators are asymptotically equivalent to GLS estimators, but are computationally more efficient. See Bentler & Dijkstra (1985) for a discussion of GLS and LGLS. Maximum Likelihood estimators were not used, because these estimators apply only when normality is assumed, and in this study the focus is on methods that fit higher-order moments, which implies that nonnormality is assumed.

The estimators that fit higher-order moments that were used in this study were introduced by Mooijaart (1985). These estimators fit second-order and third-order moments, which implies that they are based on the assumption that the variables are nonsymmetrically (skewly) distributed, which is not a very restrictive assumption. The discrepancy function that is minimized is based on the *Asymptotically Distribution-free* (ADF) theory of Browne (1984). These estimators will therefore be called ADF3. The ADF3 estimators were compared with estimators that fit only second-order moments, namely Normal theory (NORM), Elliptical theory (ELL), and Asymptotically Distribution-free (ADF2) estimators. The NORM estimators were described by Jöreskog & Goldberger (1972), and these estimators are asymptotically equivalent to normal theory maximum likelihood estimators. The ELL estimators were described by Bentler (1983), and use the assumption that the variables are elliptically distributed, which is a relaxation of the normality assumption. The ADF2 estimators were described by Browne (1982, 1984), and use no specific assumptions about the distributions of the variables. Their asymptotic properties are independent of the distribution of the variables. The NORM, ELL, and ADF2 estimators are implemented in the program EQS (Bentler, 1989).

The discrepancy function (loss function) that had to be minimized has the following form:

$$F(\theta) = (s - \sigma(\theta))' W (s - \sigma(\theta)),$$

where s contains sample statistics, $\sigma(\theta)$ is the (asymptotic) expectation of s according to the model, θ is the parameter vector that has to be estimated, and W is the *weight matrix*. For the second-order estimators, s consists of the nonduplicated elements of the sample covariance matrix, and for the ADF3 estimators, s consists of all nonduplicated second and third-order moments. The vector $\sigma(\theta)$ reflects the structural model that has to be estimated. It is well-known (e.g., Bentler & Dijkstra, 1985) that W is optimal if $\text{plim}(W) = \Gamma^{-1}$, where Γ is the asymptotic covariance matrix of $\sqrt{N}(s - \sigma)$, and N is the sample size. Therefore, W was chosen in such a way that W is a consistent estimator of Γ^{-1} . But the elements of Γ , and consequently the elements of W , depend on the distributions of the variables. Therefore, W reflects the distributional assumptions. In LGLS, $\sigma(\theta)$ is linearized around an initial consistent estimate θ_1 of θ .

The structural model $\sigma(\theta)$ can be tested by the chi-square test statistic $NF(\hat{\theta})$, where $F(\hat{\theta})$ is the minimum of $F(\theta)$, and $\hat{\theta}$ is the argument for which the minimum is attained. This test statistic follows an asymptotic chi-square distribution if σ and W are correctly specified.

The asymptotic covariance matrix of $\sqrt{N}(\hat{\theta} - \theta)$ can be consistently estimated by $\left(\Delta(\hat{\theta})' W \Delta(\hat{\theta})\right)^{-1}$, where

$$\Delta(\theta) = \frac{\partial \sigma}{\partial \theta'}(\theta),$$

and confidence intervals for the elements of θ can be easily obtained from the (diagonal) elements hereof and the (consistent) estimates $\hat{\theta}$ of θ .

The model that had to be estimated was equal to the model with which the data were generated, that is, a one-factor model with a skewly distributed factor and symmetrically distributed errors. The variance of the factor was fixed to one and the factor loadings were free parameters.

Based on the results of robustness studies (e.g., Browne, 1987), it was expected that the NORM, ADF2, and ADF3 test statistics would perform well, but that the ELL test statistic would not perform well. Furthermore, it was expected that the confidence intervals of the ADF3 and ADF2 methods would be approximately correct and that the confidence intervals of the error standard deviations of the NORM estimators would be approximately correct, but that the confidence

intervals of the loadings of the NORM estimators would be incorrect, and that the confidence intervals of the ELL estimators would be incorrect. If the model had been specified with one factor loading fixed to one and unspecified factor variance, which would have rendered a model that is statistically indistinguishable from the one used, the confidence intervals of the loadings would have been consistently estimated by the NORM estimators. On the other hand, this would have implied that the factor variance is a free parameter that had to be estimated, and the confidence interval of the factor variance would be inconsistently estimated, so the problem of inconsistency is only shifted from one set of parameters to another. We have chosen the first parametrization, because we think that this is the most exploratory one. In confirmatory factor analysis, the second parametrization is often used.

From the results of Kano, Bentler, & Mooijjaart (1993), it was expected that the ADF3 estimators would be more efficient than the second-order estimators, and from the results of robustness studies (e.g., Mooijjaart & Bentler, 1991), it was expected that the ADF2, NORM, and ELL estimators would be equally efficient.

3. RESULTS

Test statistic

Some results on the properties of the test statistics are given in Tables 1a and 1b. The ADF3 test statistic performed well with sample sizes 1000 and larger, and its performance was independent of factor skewness. The ADF2 and NORM test statistics accepted the model in all cases about 95% of the replications, independent of sample size or skewness. The ELL test statistic nearly always accepted the model if the factor skewness was 2.0 or larger, where the skewness of the variables was 0.4 or larger. The GLS and LGLS test statistics generally performed equally well, except in the ADF3 case: at relatively small sample sizes the GLS test statistic performed better.

Parameter estimates

From Table 2 it can be seen that the ADF3 estimators were slightly biased downward. The GLS estimators were more biased than the LGLS estimators. This bias diminished at larger sample sizes, as expected because of the consistency of the estimators. The second order estimators were practically unbiased.

Table 1.Percentage of times the model was accepted ($\alpha = .05$).

(a) Factor skewness = 2.0.

sample size	ADF3		ADF2		NORM		ELL	
	GLS	LGLS	GLS	LGLS	GLS	LGLS	GLS	LGLS
300	84	76	91	91	93	93	100	100
500	89	87	95	95	95	95	99	99
700	89	86	93	93	94	94	98	98
1000	94	93	95	95	95	95	99	99
2000	95	94	95	95	95	95	100	100
10000	90	90	95	95	95	95	100	100

(b) Sample size = 10000.

factor skewness	ADF3		ADF2		NORM		ELL	
	GLS	LGLS	GLS	LGLS	GLS	LGLS	GLS	LGLS
0.1	95	95	97	97	97	97	97	97
0.25	97	97	95	95	95	95	95	95
0.5	95	95	96	96	96	96	96	96
1.0	97	97	95	95	95	95	96	96
2.0	90	90	95	95	95	95	100	100
3.0	96	96	97	97	97	97	100	100
4.0	96	94	92	92	92	92	100	100

Table 2.

Average bias of the loadings estimates. (Factor skewness = 2.0.)

sample size	ADF3		ADF2		NORM		ELL	
	GLS	LGLS	GLS	LGLS	GLS	LGLS	GLS	LGLS
300	-0.04	-0.02	-0.01					
500	-0.04	-0.03	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
700	-0.02	-0.01						
1000	-0.02	-0.01						
2000	-0.01	-0.01						
10000								

Note. Blank means absolute value of bias was less than 0.005.

From Table 3 it can be seen that the ADF3 estimators had indeed slightly smaller standard deviations (were more efficient) than the second order estimators. The differences, however, were very small. At smaller sample sizes LGLS estimators were better than GLS estimators for the ADF methods.

Table 3.

Average standard deviation of the loadings estimates. (Factor skewness = 2.0.)

sample size	ADF3		ADF2		NORM		ELL	
	GLS	LGLS	GLS	LGLS	GLS	LGLS	GLS	LGLS
300	0.085	0.078	0.127	0.082	0.081	0.081	0.081	0.081
500	0.058	0.058	0.069	0.063	0.063	0.063	0.063	0.063
700	0.050	0.050	0.065	0.052	0.052	0.052	0.052	0.052
1000	0.042	0.041	0.065	0.044	0.043	0.043	0.043	0.043
2000	0.030	0.030	0.031	0.031	0.031	0.031	0.031	0.031
10000	0.013	0.013	0.014	0.014	0.014	0.014	0.014	0.014

Note. Figures were calculated as follows: In each of 300 replications with the same skewness and sample size, factor loadings were estimated by each of the eight methods. For each of the four loadings, the standard deviation was calculated per method. These four standard deviations were averaged.

Table 4.

Percentage of times the confidence intervals for the loadings covered the true value. (Factor skewness = 2.0; $\alpha = .05$)

sample size	ADF3		ADF2		NORM		ELL	
	GLS	LGLS	GLS	LGLS	GLS	LGLS	GLS	LGLS
300	74	82	91	93	81	81	94	94
500	81	85	91	92	83	83	93	93
700	85	88	94	94	85	85	95	95
1000	88	90	93	94	84	84	94	94
2000	88	90	94	94	82	82	95	95
10000	94	94	94	94	83	83	95	95

Confidence intervals

The ADF3 confidence intervals were correct only in very large samples (see Table 4). This result was independent of factor skewness (Table 5), but strongly related to bias (cf. Table 2). The ADF2 confidence intervals were generally satisfactory. At larger skewnesses (larger departure from symmetry), the NORM

confidence intervals for the loadings and the ELL confidence intervals for the error standard deviations were unreliable (see Tables 4 and 5), a result that was independent of sample size, but due to the inconsistency of the estimated asymptotic covariance matrix (cf. Browne, 1987, for the normal case). For the ADF methods the LGLS confidence intervals were better than the GLS ones, most probably because of the smaller bias of the LGLS estimators.

Table 5.

Percentage of times the confidence intervals for the error standard deviations covered the true value. (Sample size = 1000; $\alpha = .05$.)

factor skewness	ADF3		ADF2		NORM		ELL	
	GLS	LGLS	GLS	LGLS	GLS	LGLS	GLS	LGLS
0.1	90	90	96	96	97	97	97	97
0.25	88	89	95	95	95	96	96	96
0.5	89	90	96	96	96	96	96	96
1.0	88	89	94	94	94	94	97	97
2.0	90	89	95	95	96	96	99	99
3.0	88	88	94	94	95	95	100	100
4.0	89	88	96	96	96	96	100	100

4. DISCUSSION

The ADF3 method generally performed a little worse than the ADF2 and NORM methods, although confidence intervals of the NORM estimators should be corrected, for example by using the “robust covariance matrix of the estimator” in EQS (Bentler, 1989). However, the differences in performance were small in larger samples, and the advantages of the ADF3 method, such as identification and information, may outweigh its somewhat worse performance. This may particularly be the case if the ADF3 estimates are corrected for bias, for example by using bootstrap or jackknife estimators (see, e.g., Efron, 1982).

The LGLS estimators generally did not perform worse than their GLS counterparts, and they may be preferred because of their computational efficiency.

Although not the main focus of attention, the ELL estimators were also studied, and because the test statistic and the confidence intervals performed so badly, they cannot be recommended.

The results obtained here apply only to the relatively simple 4-variable, 1-factor, factor analysis model with normal errors and a skewly distributed factor. It is hoped that the conclusions generalize to more complicated models with more variables and factors, but this remains to be investigated.

REFERENCES

- [1] **Bentler, P.M.** (1983). "Some contributions to efficient statistics in structural models: specification and estimation of moment structures". *Psychometrika*, **48**, 493–517.
- [2] **Bentler, P.M.** (1989). *EQS structural equations program manual*. Los Angeles: BMDP Statistical Software.
- [3] **Bentler, P.M. & T. Dijkstra** (1985). "Efficient estimation via linearization in structural models". In: P.R. Krishnaiah (ed.), *Multivariate Analysis VI* (pp. 9–42). Amsterdam: Elsevier Science.
- [4] **Browne, M.W.** (1982). "Covariance structures". In: D.M. Hawkins (ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.
- [5] **Browne, M.W.** (1984). "Asymptotically distribution-free methods for the analysis of covariance structures". *British Journal of Mathematical and Statistical Psychology*, **37**, 62–83.
- [6] **Browne, M.W.** (1987). "Robustness of statistical inference in factor analysis and related models". *Biometrika*, **74**, 375–384.
- [7] **Browne, M.W. & A. Shapiro** (1988). "Robustness of normal theory methods in the analysis of linear latent variate models". *British Journal of Mathematical and Statistical Psychology*, **41**, 193–208.
- [8] **Efron, B.** (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.
- [9] **Jöreskog, K.G.** (1967). "Some contributions to maximum likelihood factor analysis". *Psychometrika*, **32**, 443–482.
- [10] **Jöreskog, K.G. & A.S. Goldberger** (1972). "Factor analysis by generalized least squares". *Psychometrika*, **37**, 243–260.
- [11] **Kano, Y., P.M. Bentler & A. Mooijaart** (1993). "Additional information and precision of estimators in multivariate structural models". In: K. Matusita, M.L. Puri, T. Hayakawa (eds.), *Statistical sciences and data analysis, Proceedings of the Third Pacific Area Statistical Conference* (pp. 187–196). Utrecht, The Netherlands: VSP.
- [12] **Mooijaart, A.** (1985). "Factor analysis for non-normal variables". *Psychometrika*, **50**, 323–342.

- [13] **Mooijaart, A. & P.M. Bentler** (1986). "Random polynomial factor analysis". In: E. Diday, Y. Escoufier, L. Lebart, J.P. Pagès, Y. Schektmann, & R. Tomassone (eds.), *Data analysis and informatics, IV* (pp. 241–250). Amsterdam: North Holland.
- [14] **Mooijaart, A. & P.M. Bentler** (1991). "Robustness of normal theory statistics in structural equation models". *Statistica Neerlandica*, **45**, 159–171.
- [15] **Van Monfort, K., A. Mooijaart & J. De Leeuw** (1987). "Regression with errors in variables: estimators based on third order moments". *Statistica Neerlandica*, **41**, 223–239.

Estadística Oficial

ESTADÍSTICA OFICIAL I COORDINACIÓ DEL SISTEMA ESTADÍSTIC CATALÀ

JOSEP A. CASCO ROBLEDO

Institut d'Estadística de Catalunya

S'hi expliquen els continguts conceptuals que corresponen a tres expressions terminològiques d'ús comú: estadística oficial, sistema estadístic i coordinació estadística. El model estadístic de Catalunya serveix com a exemple i fil conductor d'aquesta reflexió.

Official statistics and coordination of the Catalan statistical system.

Key words: Estadística oficial, sistema estadístic, coordinació estadística, estadística pública, organització estadística.

INTRODUCCIÓ

Aquest article pretén mostrar alguns aspectes bàsics de les institucions estadístiques denominades oficials.

Un element rellevant en l'organització estadística el constitueix la col·laboració participativa dels ciutadans. Són molt rares les estadístiques oficials que es puguin fer per observació directa dels fenòmens. El qüestionari o l'entrevista són els instruments imprescindibles d'observació. En aquest aspecte juga un paper cabdal la garantia jurídica del respecte als drets fonamentals de les persones i, especialment, la preservació de la intimitat.

—Article rebut el desembre de 1993.

—Acceptat el febrer de 1994.

La informació estadística és una part de la informació necessària per nodrir el coneixement sobre determinats fenòmens col·lectius que afecten el funcionament de la societat. Per això resideix en els organismes governamentals. Per tant, un altre tema important és el de la dependència/independència funcional de l'estadística oficial. En els països democràtics l'han de garantir les lleis, és a dir, el poder legislatiu i el poder judicial. D'aquí la importància que adquireix l'organització institucional, les seves normes i relacions que anomenem sistema estadístic.

És la pròpia legislació estadística la que ha de garantir una organització estadística que condueixi, de forma objectiva a l'obtenció de resultats estadístics per procediments científics al servei del Govern i dels agents socials amb independència de preferències ideològiques o partidistes.

El tema de la coordinació posa de manifest la filosofia subjacent de les estructures organitzatives que són els instituts d'estadística i posseeix relació directa amb la coherència i la comparabilitat de les estadístiques.

Com a mostra de tot això es contempla el sistema estadístic català que és modern i incorpora experiències que han resultat positives en altres sistemes. No obstant això, és necessari advertir que es tracta d'una interpretació personal de la legislació estadística catalana.

1. ESTADÍSTICA OFICIAL

La definició i la delimitació dels termes de l'enunciat de l'article no és una fixació escolàstica i acadèmica sinó una necessitat intrínseca del tema. No hi ha una definició de l'estadística oficial sobre la qual existeixi un consens universal. Aquesta és una raó més per aventurar una aproximació conceptual.

Sense massa rigor, però amb força encert des d'una perspectiva sociològica, seria vàlid definir l'estadística oficial com aquella activitat estadística duta a terme per organismes governamentals contraposada a l'estadística feta per institucions acadèmiques i privades. Probablement, aquesta és la idea referencial més universal i vista des de fora. Però no és una definició precisa.

Situats dins de l'estadística oficial, és possible que tampoc no hi hagi massa consens en la concreció terminològica. En canvi, es pot dir que existeix unanimitat respecte al concepte encara que aquest no s'hagi explicitat amb claredat. No obstant això, s'ha de dir que l'expressió "estadística oficial" no és afortunada per les seves connotacions equívokes. Però l'expressió "estadística pública" és conceptualment errònia encara que terminològicament sona millor.

1.1. Estadística pública i estadística oficial

L'estadística oficial s'acostuma a identificar amb l'estadística pública com si fossin dues expressions intercanviables. Conceptualment són diferents. Tota estadística oficial és pública però no tota l'estadística pública és oficial.

L'activitat estadística es qualifica de pública quan és duta a terme per les institucions que són o que depenen de l'Administració pública, directament o indirecta. Aquesta és una condició necessària a tota estadística oficial però no suficient i es podria dir que és una concepció premoderna, és a dir, anterior a la creació dels instituts d'estadística i de les lleis d'estadística que corresponen a un procés històric que neix els segles XVIII i XIX i segons el país desemboca en l'actual. En realitat es pot afirmar que l'estadística oficial neix quan s'acaba l'estadística pública, entesa com la activitat censal històrica dels estats.

El respecte a la llibertat i a la intimitat personal com un dret de les persones enfront de l'Estat, dona pas al naixement del secret estadístic com a contrapartida a la col·laboració obligatòria en el subministrament d'informació. Es crea, llavors, un òrgan especialitzat a l'Administració pública que es caracteritza per recollir informació dels ciutadans per procediments i canals diferents dels administratius habituals i que és l'institut d'estadística. S'acostuma a assenyalar els censos de població de Suècia i d'Estats Units de 1749 i 1790, respectivament, com el moment en què neix l'estadística moderna¹. A Espanya es crea el 1856 la Comisión General de Estadística del Reino i l'any següent es fa el primer cens de població de l'estadística moderna espanyola.

Per tant, es pot afirmar que l'activitat estadística és oficial quan és elaborada per institucions públiques legitimades per fer-la. La legitimitat inclou el reconeixement legal de competència o potestat per fer activitats estadístiques i, també, la submissió a les normes que regulen aquestes activitats. Aquesta és una convenció generalitzada en els països occidentals. Pràcticament, cadascun dels estats s'ha dotat de la legislació específica en matèria estadística necessària segons la seva pròpia organització administrativa i territorial.

Les lleis d'estadística, en els països occidentals, desenvolupen els principis i les garanties jurídiques en virtut de les quals les administracions públiques estan facultades per fer investigacions estadístiques sobre les persones, les empreses

¹Casco Robledo, José A. "Los Censos de Población ante la estrategia de la información territorial". "Ciudad y Territorio" n.º 1/71. (Instituto de Estudios de Administración Local. Madrid 1971), pàg. 63, on s'explica la vinculació del secret estadístic al naixement dels instituts d'estadística i el canvi a la modernitat relacionat amb els primers censos de caràcter estadístic i finalitat exclusivament estadística. Anglaterra, França i Bèlgica precedeixen Espanya, a la qual segueixen Itàlia, Portugal, Alemanya, Canadà, la Índia, Rússia i Japó. Aquest darrer el 1898.

i les institucions, amb dret a exigir respostes veraces, és a dir, gaudeixen del privilegi de la col·laboració obligatòria en el subministrament de la informació. Aquesta legitimació ens permet identificar les estadístiques oficials com a diferents de les estadístiques públiques. Aquestes últimes pel fet de ser dutes a terme per una entitat pública no gaudeixen del privilegi de resposta obligatòria.

La regulació jurídica de l'activitat estadística amb rang de Llei es justifica, doncs, per tres tipus de necessitats. La protecció del dret a la intimitat dels ciutadans que va lligat a l'obligatorietat de subministrar informació per part d'aquests i que té un origen històric a cada país. En segon lloc, la necessitat d'establir i de fixar el règim jurídic de l'activitat estadística pública respecte a qui decideix fer estadístiques, com i quan, com s'han de tractar i conservar les dades i com s'han de difondre els resultats. En tercer lloc la necessitat de crear els instruments institucionals i organitzatius que puguin satisfer les necessitats d'informació estadística per a la presa de decisions dels governs i dels agents econòmics i socials i el control parlamentari i públic de les polítiques governamentals².

L'estadística oficial, a més d'estar regulada per lleis, sol estar planificada. La planificació aplicada a l'activitat estadística és relativament recent i ha sorgit, no només d'una expansió en altres camps i contextos, sinó de la necessitat de coordinació i eficàcia de l'activitat estadística pública. No és habitual que la planificació de l'estadística es faci mitjançant un instrument jurídic amb rang de llei, però a Catalunya ha estat així.

Quant a Catalunya, l'article 9.33 de l'Estatut d'Autonomia atorga a la Generalitat competència exclusiva sobre l'estadística del seu interès. Posteriorment, el Parlament de Catalunya va crear el marc legal que permet el desenvolupament d'aquesta competència estatutària mitjançant la Llei 14/1987, de 9 de juliol, d'Estadística. De la mateixa manera, la Constitució espanyola, en el seu article 149.1.31 atribueix competència exclusiva a l'Estat per a les seves pròpies finalitats i la Llei 12/1989, de 9 de maig, de la funció estadística pública estableix les regles per les quals s'ha de regir l'estadística pública de l'Estat.

La Llei d'estadística de Catalunya de 1987 esmentada, estableix que el Pla estadístic s'ha d'aprovar per Llei del Parlament de Catalunya amb una vigència de quatre anys i s'acompanya així l'èmfasi conceptual del paper de la planificació expressat en aquesta Llei amb l'elevació del rang jurídic d'aquesta activitat. En

²Servicios de Estadística del Principado de Asturias. "El Sistema Estadístico". Asturias. 1994. *Documento de trabajo*. Pàg. 14. Estableix quatre necessitats per justificar la Llei de la funció estadística pública, però poden resumir-se en aquestes tres si es fa una afirmació més general. Aquest estudi és de gran interès per tal de conèixer la legislació estadística espanyola comparada.

compliment d'això, el Parlament de Catalunya va aprovar la Llei 30/1991, de 13 de desembre, del Pla estadístic de Catalunya 1992-1995. D'aquesta manera, Catalunya compta amb dues lleis en aquesta matèria: la de regulació bàsica o constitucional de l'activitat estadística pública de Catalunya i la de la seva planificació amb vigència temporal de quatre anys. Per completar aquesta informació manca afegir el Decret 341/1989, d'11 de desembre, de creació de l'Institut d'Estadística de Catalunya.

La pròpia Llei d'estadística de Catalunya obliga el Govern a elaborar un programa anual d'actuació estadística que desenvolupi en forma executiva la Llei del pla estadístic. Aquesta previsió legal s'ha complert en temps i en forma pel Govern de Catalunya i s'han aprovat per Decret els programes corresponents a 1992, 1993 i 1994 i tot això fa pensar que quan arribi el seu moment, s'aprovarà també el programa d'actuació estadística de 1995, últim del Pla estadístic vigent.

1.2. Estadística oficial i resultats estadístics oficials

El llenguatge col·loquial denomina estadístiques tant a les activitats estadístiques com als seus resultats, la qual cosa, per extensió, ens condueix a confondre les “activitats estadístiques oficials” amb els “resultats estadístics oficials”. Confusió en la qual no incorre la legislació catalana, entre altres³.

La legislació estadística catalana introdueix el concepte que es refereix als “resultats estadístics oficials”. Es tracta d'una prerrogativa legal que té el Govern català en virtut de la qual pot atribuir “valor oficial” als resultats estadístics a l'efecte de les relacions jurídiques sobre la regulació de les quals tingui competència la Generalitat. Aquesta declaració d'oficialitat només pot recaure sobre resultats obtinguts d'estadístiques elaborades per l'òrgan estadístic de la Generalitat —Institut d'Estadística de Catalunya— directament o en col·laboració amb altres entitats⁴.

En realitat, es tracta d'una precisió que estableix la distinció entre activitat estadística i resultats estadístics. Per tant, podem establir que existeixen tres conceptes a diferenciar:

- L'“estadística pública” que correspon a l'activitat estadística de les administracions públiques, la qual pot estar o no legitimada.

³També es confonen dades i resultats, és a dir les observacions amb les xifres derivades del seu tractament.

⁴Llei d'estadística de Catalunya esmentada. Art. 38.

- L'“estadística oficial” que correspon a l'activitat estadística pública regulada per les lleis específiques. A Catalunya, a més a més, planificada.
- Els “resultats estadístics oficials” que són les expressions numèriques finals derivades d'algunes activitats estadístiques oficials sobre les quals ha recaigut una declaració governamental d'oficialitat. En general, es persegueix una finalitat jurídica.

Les declaracions d'oficialitat dels resultats són comunes a molts països occidentals i solen recaure sobre resultats derivats dels censos i no de mostres. La finalitat és l'establiment d'una sola referència numèrica, l'oficial, a l'efecte de determinar drets o obligacions entre administracions públiques o entre aquestes i els ciutadans⁵. De la mateixa manera, la Unió Europea requereix resultats d'estadístiques oficials per tal de concretar polítiques comunitàries. Aquesta declaració és diferent de la contractual, en la qual persones o institucions s'imposen una estadística de referència per determinar la quantitat exacta que les parts hauran de respectar. És molt freqüent en els arrendaments urbans i en operacions de crèdit.

Malgrat aquestes precisions, quan es parla de l'estadística oficial s'acostuma a incloure l'activitat estadística i els seus resultats, en un sentit ampli, però s'és restrictiu quant a l'autoria, que ha de ser l'Administració pública legitimada per fer-la.

1.3. Activitat estadística estricta i activitat estadística instrumental

Per acabar de completar conceptualment la definició de l'estadística oficial és convenient al·ludir a l'activitat estadística estricta i a la instrumental. Es tracta d'una aportació original de la legislació catalana.

Tradicionalment, l'activitat estadística oficial s'ha definit a partir de la descripció de les fases que es requereixen per a la disponibilitat de resultats útils per al coneixement i l'anàlisi quantitatius “de la realitat geogràfica, econòmica, demogràfica i social” d'un país. Així, per exemple, la llei catalana d'estadística defineix l'activitat estadística com “la recopilació, l'elaboració, l'ordenació sistemàtica de dades i la publicació i la difusió de resultats”⁶.

⁵A Espanya, el Consell de Ministres aprova els resultats dels censos de població per municipis i passen a ser “xifres oficials”, que serveixen de base per al càlcul de la contribució estatal a les finances dels ajuntaments respectius i d'altres assumptes importants. A Catalunya s'aproven les xifres comarcals de població i també les del coneixement del català.

⁶Llei d'estadística de Catalunya esmentada. Art. 2.

En canvi, la Llei del pla estadístic de Catalunya, quatre anys posterior, reitera la definició d'activitat estadística tradicional i afegeix el concepte d'activitat estadística instrumental, que inclou "totes les activitats prèvies o complementàries de les anteriors —les compreses en la definició tradicional— i legalment exigibles o tècnicament necessàries per tal de poder complir els requisits que estableix la legislació sobre estadística ..." ⁷.

Per tant, es dona unitat a l'activitat estadística des de l'òptica de l'ofici estadístic, del qual destaquen les tasques de recerca i desenvolupament tècnic, metodològic i normatiu en el camp estadístic. Aquestes activitats s'integren també en l'apartat d'activitats estadístiques oficials.

La consideració d'activitat estadística oficial atribuïda a la producció de resultats gaudeix d'una major acceptació per raó del seu efecte immediat a la societat pel qual es justifica la necessitat de lleis d'estadística i la creació dels instituts. Això explica que aquest concepte es correspon amb les regulacions clàssiques de l'estadística pública i que la inclusió del concepte instrumental sigui més moderna i propera a l'ofici estadístic i a la planificació. A l'ofici estadístic ja que són molts els professionals que només fan tasques instrumentals en els processos de producció estadística que quedarien exclosos de l'estricta definició legal clàssica. Però també es deu a les tasques modernes de planificació ja que la necessitat de recursos s'ha de justificar per les activitats a realitzar i, per tant, s'han d'incloure les activitats instrumentals per tal que puguin ser valorades i dotades.

Gran part de les activitats instrumentals, sobretot les que es refereixen a la utilització d'eines estadístiques i informàtiques no requereixen una legitimació específica i només es poden denominar activitats estadístiques oficials per extensió o per la seva inclusió en el procés estadístic oficial. En canvi, altres activitats instrumentals del tipus de les classificacions i els codis i les directes de coordinació i normalització estadística gaudeixen d'una acceptació més generalitzada com a activitats estadístiques oficials per raó de ser pròpies de l'Administració pública però segueixen sent instrumentals i no directament referides a les definicions clàssiques de la legislació estadística.

Aquesta reflexió, no obstant això, no treu mèrits a l'aportació de la legislació catalana —Llei del pla estadístic— que en aquest aspecte ha estat considerada molt positivament i ràpidament importada a altres sistemes estadístics propers.

⁷Llei del pla estadístic de Catalunya esmentada. Art. 7.

1.4. Activitat estadística com a activitat administrativa

Com a conseqüència del discurs anterior s'ha de concloure que l'activitat estadística oficial és una activitat administrativa especialitzada.

Entre els principis explícits que ha d'observar l'activitat estadística oficial hi ha, en primer lloc, el d'interès públic. La declaració d'interès públic l'efectua el Parlament —cas de Catalunya— i el Govern i es refereix a cada activitat per separat. És una característica exclusiva de l'activitat administrativa. El segueixen els principis d'objectivitat i de correcció tècnica que són comuns a qualsevol activitat estadística segons l'ètica professional amb la diferència de ser, a més, un precepte legal. El principi de l'obligatorietat de la col·laboració ciutadana només correspon a l'estadística oficial quant al fet que és administrativa. El respecte a la intimitat és també un principi ètic comú a tota activitat estadística però no així el de l'observança del secret estadístic, que és una institució exclusiva del professional estadístic de l'Administració pública. Per últim, el principi de la publicitat de resultats estadístics és també d'aplicació obligatòria només en l'activitat administrativa⁸.

L'activitat estadística oficial no és ni superior ni inferior a altres activitats estadístiques realitzades en institucions no governamentals des d'un punt de vista de la seva correcció tècnica. En canvi, s'ha de subratllar la responsabilitat addicional dels professionals de l'estadística oficial quant a la implicació legal i administrativa dels seus actes.

1.5. Conclusió

D'acord amb la interpretació de la legislació catalana, l'activitat estadística oficial és l'activitat estadística estricta o instrumental que porten a terme directament o indirecta, les institucions públiques catalanes sotmeses a la regulació i a la planificació establertes per les lleis.

Els principis i garanties jurídiques que, de forma explícita o implícita, afecten i regulen l'estadística oficial són els següents:

- *D'interès públic*: es faran només les estadístiques necessàries i útils per a la presa de decisions de les institucions públiques i agents socials però l'interès públic ha de ser declarat pel Parlament per mitjà de la Llei del

⁸Llei d'estadística de Catalunya esmentada. Art. 19.1. Aquest article presenta un resum dels principis tècnics i garanties jurídiques de l'activitat estadística.

pla estadístic i pel Govern per mitjà dels programes o per decret justificat per raons d'urgència.

- *D'obligatorietat*: les estadístiques oficials gaudeixen del privilegi de col·laboració o resposta obligatòria.
- *De legalitat*: les estadístiques s'ajustaran a les normes reguladores generals i les específiques de cadascuna d'elles.
- *D'objectivitat*: neutralitat subjectiva en el plantejament i respecte als resultats obtinguts.
- *De respecte a la intimitat personal*: els qüestionaris i la forma de recollida de la informació s'han de sotmetre a normes especials de control de l'anonimat quan continguin preguntes relacionades amb l'honor personal o amb les conviccions religioses o polítiques de l'informant. A més, estan protegits pel secret estadístic.
- *De transparència informativa*: els interessats han de conèixer les seves obligacions i drets, la finalitat que persegueix l'estadística i altres característiques de cada activitat.
- *De coherència*: les estadístiques han de formar un conjunt coherent respecte a les necessitats d'informació i les estadístiques que ja estan en curs.
- *D'homogeneïtat i normalització*: les estadístiques han de ser comparables mitjançant el compliment de normes tècniques comunes obligatòries.
- *D'especialitat*: la finalitat ha de ser exclusivament estadística.
- *De proporcionalitat*: les estadístiques no han d'exigir més informació que la necessària i s'han de minimitzar les molèsties als ciutadans i el cost.
- *De cooperació*: les institucions productores d'estadístiques no han d'incórrer en duplictat i estan obligades a col·laborar.
- *De publicitat*: els resultats estadístics han de tenir publicitat, han d'estar disponibles i s'han de difondre.
- *De secret estadístic*: en virtut d'aquesta obligació les dades individuals de comunicació obligatòria no poden ser fetes públiques ni comunicades a cap persona o entitat, ni tan sols a les administracions públiques, llevat d'aquelles institucions o entitats que també estan vinculades per l'obligació del secret estadístic i exclusivament amb la finalitat d'ésser usades per a operacions estadístiques.

2. EL SISTEMA ESTADÍSTIC CATALÀ

L'activitat estadística oficial està burocratitzada i subjecta a unes regles que poden tenir el rang de llei, com ja hem vist. Aquesta és la situació generalitzada en els països occidentals. Aquestes regles s'apliquen a un espai concret geogràfic, poblacional i institucional com a element imprescindible de tota norma. Això constitueix l'àmbit d'aplicació, com ja és sabut.

Però, a més, existeix una consideració més àmplia que l'esmentat àmbit jurídic la qual està determinada per l'especificitat del tema, la manera d'afectar les persones i les institucions, les relacions entre unes i altres i d'elles entre si, les pautes i els criteris sobre la utilització dels mitjans i dels objectius que es persegueixen, la censura i el règim sancionador dels comportaments desviats, etc., la qual cosa constitueix l'àmbit o l'espai sociocultural diferenciat i propi.

D'acord amb les teories sistèmiques, quan aquest espai sociocultural s'eleva al rang de model organitzatiu i se'n destaquen l'estructura, les funcions i els processos a complir pels agents que hi intervenen i la seva interrelació, i els plans, programes i activitats ordenats en un conjunt harmònic, es pot qualificar de sistema. Per aquesta raó, hi ha autors que defensen la identitat entre sistema i organització. Des de fa quaranta anys són nombrosos els treballs realitzats sobre la teoria de sistemes i les seves aplicacions (la revista "General Systemes" va aparèixer el 1956). El professor Azorín va desenvolupar el tema amb aplicació a l'organització estadística⁹.

De forma resumida i esquemàtica, el sistema estadístic és el conjunt ordenat i harmònic de mètodes, procediments i resultats estadístics dels diferents agents institucionals, en un espai jurídic concret.

2.1. Característiques del sistema estadístic català

El sistema estadístic català és distint del sistema estadístic espanyol amb el qual s'autocoordina i es relaciona, i no constitueix un subsistema d'aquell. El model organitzatiu català és diferent però no molt d'altres existents. El que és més significatiu del model català és l'atenció que ha prestat a algunes funcions, latents en altres contextos i que a Catalunya s'han explicitat i potenciat.

Les idees força que han determinat el model de l'estadística catalana són: el pes que té el Parlament —poder legislatiu— en el procés estadístic de Catalunya,

⁹Vegeu Azorín Poch, Francisco. "Estadística, información, organización y sistema". 1986. Conferència publicada pel Centro de Información Estadística de Galicia.

la rellevància del sistema estadístic per sobre de les institucions individualitzades i el paper particular de l'Institut d'Estadística. El més important en el model estadístic català és que funcioni el sistema estadístic en el seu conjunt. L'Institut d'Estadística de Catalunya gaudeix d'una doble consideració legal en el sistema estadístic català. D'una part, és l'òrgan especialitzat de la Generalitat i de l'altra és un agent més del sistema estadístic. Quant a agent del sistema estadístic, l'Institut ha de produir les estadístiques que li assigna el Pla estadístic i el Govern en els programes anuals. Tanmateix, la funció bàsica de l'Institut d'Estadística com a òrgan especialitzat en estadística de la Generalitat de Catalunya consisteix a gestionar el sistema estadístic català¹⁰.

Com la majoria dels models organitzatius de l'estadística oficial occidental el model català és descentralitzat quant a la producció estadística. Però presenta una integració particular dels departaments governamentals i d'altres administracions públiques. Resulta molt radical respecte a la planificació estadística que es converteix en condició indispensable per a la participació en el propi sistema i com a conseqüència és el criteri de legitimació de les institucions i les seves activitats estadístiques. El sistema estadístic català està integrat, en primer lloc, per la Generalitat de Catalunya i els seus organismes dependents que formen el sistema estadístic de la Generalitat de Catalunya. En segon lloc, per la mateixa Generalitat i les entitats de l'Administració local que participen en el Pla estadístic i que en conjunt formen el sistema estadístic integral de Catalunya¹¹.

La integració de les diferents institucions en el sistema estadístic de Catalunya es fa per mitjà de la participació en el Pla estadístic de Catalunya. Aquesta participació és obligatòria per als departaments i organismes de la Generalitat però voluntària per a les entitats de l'Administració local. Però queda ben entès que l'activitat estadística oficial només es pot dur a terme dins del Pla estadístic aprovat per llei del Parlament. Per tant, les activitats estadístiques de fora del Pla estadístic no són estadístiques oficials i no gaudeixen de la protecció de la Llei d'estadística. O, el que és el mateix, fora del Pla estadístic i, per tant, fora del sistema estadístic no hi ha estadístiques oficials¹².

El Pla estadístic de Catalunya explicita els objectius operatius que seran objecte de les activitats d'estadística i assigna les activitats estadístiques instrumentals només a l'Institut d'Estadística de Catalunya mentre que les activitats de producció estadística les assigna a tots els departaments de la Generalitat inclòs, també, l'Institut d'Estadística de Catalunya. En una primera assignació nominal el Pla estadístic sanciona les activitats estadístiques en curs seguint el criteri realista de relacionar-les amb les matèries incloses en les competències de

¹⁰Llei del pla estadístic de Catalunya esmentada. Art. 6.

¹¹Llei del pla estadístic esmentada. Art. 1 i 2.

¹²Llei d'estadística esmentada. Art. 8 i 14.2.

cada Departament. Posteriorment, els programes anuals d'actuació estadística incideixen en el mateix criteri d'assignació respecte de les activitats estadístiques noves¹³.

El model estadístic català, doncs, descentralitza la producció estadística en funció de criteris sectorials entre tots els departaments però centralitza les activitats instrumentals en l'Institut d'Estadística de Catalunya quant a òrgan especialitzat de la Generalitat. Les activitats de producció assignades a l'Institut són, en general, les que no encaixarien en les competències sectorials de cap departament en concret o que serien comunes o polivalents, les activitats de síntesi i les grans operacions censals.

Aquestes activitats que corresponen a l'Institut quant a agent productor del sistema, podrien, en teoria, ser assignades a una altra institució del sistema però no les instrumentals de coordinació, de normativa i de desenvolupament metodològic, les quals li corresponen com a òrgan especialitzat.

2.2. Estructura organitzativa del sistema estadístic català

L'organització estadística catalana en l'àmbit de la qual es desenvolupa l'estadística oficial per a finalitats de la pròpia comunitat està integrada per:

- L'Institut d'Estadística de Catalunya que és l'òrgan especialitzat de la Generalitat.
- El Consell Català d'Estadística, òrgan consultiu de l'Institut.
- (Els serveis estadístics) dels departaments.
- (Els serveis estadístics) d'altres entitats públiques de caràcter territorial que hi participen.

Des del punt de vista estructural bàsic l'organització estadística catalana és comuna a la majoria d'altres països i de comunitats autònomes espanyoles. L'única diferència es pot establir per defecte per l'escassa definició i concreció de les unitats estadístiques dels departaments i altres entitats públiques. De fet, les normes fan referència a l'activitat estadística dels departaments i d'altres entitats públiques de caràcter territorial sense fer esment dels serveis o unitats específiques d'estadística com a elements concrets de l'estructura organitzativa estadística catalana. Una característica constant del sistema estadístic català és la seva flexibilitat organitzativa que, en principi, té una aparença de debilitat o

¹³Llei del pla estadístic esmentada. Art. 15, 18 i 20.

d'inacabament. No només passa en els textos legals sinó també en decrets tan específics com el de la pròpia organització que comentem¹⁴.

De tota manera, encara que l'estructura derivada de les lleis no té en compte aquest aspecte organitzatiu de l'estadística oficial dels departaments i altres entitats públiques, aquestes unitats estadístiques estan formalitzades en l'estructura interna dels departaments i en alguns d'aquests són subdireccions específiques. Però això no salva ni substitueix una organització estadística formal que comporta, normalment, la regulació jurídica del secret estadístic i la seva neutralitat funcional.

Per tant, l'estructura organitzativa del sistema estadístic català emfasitza el paper de l'Institut d'Estadística de Catalunya i això respon al paper central que ocupa el Pla estadístic del qual és responsable últim l'Institut. En canvi, el model de producció de la informació és descentralitzat i destaca la prevalença del sistema estadístic sobre les institucions. Aquesta dualitat conceptual té, doncs, aquest reflex en l'estructura organitzativa del sistema estadístic.

2.2.1. L'Institut d'Estadística de Catalunya

Com és preceptiu la Llei d'estadística autoritza el Govern a crear l'Institut d'Estadística de Catalunya. Així la creació efectiva i la seva estructura organitzativa es va fer per Decret del Govern com ja hem esmentat. Això introdueix una major flexibilitat que si hagués estat per llei en cas de modificacions.

L'Institut d'Estadística de Catalunya és un organisme autònom de caràcter administratiu. Aquesta forma jurídica és també comuna a altres instituts d'estadística.

L'adscripció dels instituts d'estadística té importància en relació amb el paper que se'ls assigna. Així ha estat històricament. L'Institut d'Estadística de Catalunya està adscrit al Departament d'Economia i Finances. Aquesta adscripció resulta molt freqüent en temps recents, encara que en temps passats era més comuna l'adscripció dels instituts d'estadística a la presidència del Govern. Aquest era el cas de Catalunya durant la Generalitat de la II República i la Generalitat provisional de la nova democràcia espanyola. També l'Instituto Nacional de Estadística (INE) va estar adscrit a la presidència del Govern fins al 1973. És d'interès notar que aquests canvis d'adscripció estan relacionats amb canvis en el pes de la temàtica de les estadístiques oficials. És paradigmàtic a aquest respecte el fet que les estadístiques estelars de l'INE siguin l'índex de preus

¹⁴Decret de creació de l'Institut d'Estadística de Catalunya esmentat.

de consum (IPC) i l'enquesta de població activa (EPA) o que les estadístiques agràries constitueixin el centre de les estadístiques de la Unió Europea. El progressiu intervencionisme dels governs en l'economia dels països ha exigít una informació estadística econòmica més precisa i freqüent.

L'estructura organitzativa de l'Institut d'Estadística de Catalunya és molt senzilla. Hi ha dos òrgans: la Junta de Govern i el director de l'Institut.

Existeix un òrgan col·legiat de direcció de l'Institut que es denomina Junta de Govern. Aquest òrgan està presidit pel conseller d'Economia i Finances i integrat pels representats dels departaments de la Generalitat que tenen interès directe en la producció i la difusió de les seves pròpies estadístiques i com a conjunt o òrgan de direcció de l'Institut en la gestió i el control del sistema estadístic. Com a mínim, el nivell d'aquests representants és de director general però la majoria són secretaris generals ja que aquests posseeixen, per raó del seu estatus administratiu, una visió més global del seu departament que el director general. Aquests són nomenats pel conseller respectiu. A més, són membres de la Junta de Govern de l'Institut, el director de l'Institut d'Estadística, el director general de Serveis d'Informàtica i el director de l'Institut Cartogràfic per tal de fer la coordinació necessària per raó de la matèria.

No hi ha representants de l'Administració local a la Junta de Govern de l'Institut d'Estadística de Catalunya malgrat que el sistema estadístic integral de Catalunya inclou aquelles entitats públiques de caràcter territorial que participen en el Pla estadístic. Tampoc no és gaire freqüent aquesta participació en altres comunitats autònomes, ja que només en dos dels sis instituts autonòmics d'estadística es dona aquest supòsit. A Catalunya, la forma de participar les entitats locals en el sistema estadístic és per mitjà del Pla estadístic i de forma voluntària. Això representa una congruència lògica, encara que s'hagués pogut arbitrar una fórmula de presència condicionada a la seva integració en el Pla estadístic i en els programes anuals.

Les competències de la Junta de Govern són molt àmplies i coincideixen pràcticament amb les funcions que la llei atribueix a l'Institut d'Estadística del qual és òrgan de direcció. Però aquestes seran tractades més endavant per tal d'integrar-les a l'activitat de coordinació. Com es podrà observar, aquestes funcions estan orientades principalment a la gestió i a la coordinació del sistema.

A més de la Junta de Govern, de la qual en forma part, el director és també òrgan de govern i administració de l'Institut. La seva funció bàsica és executiva i exerceix l'autoritat màxima sobre l'equip de persones que realitzen els treballs tècnics i administratius a la institució. És nomenat pel Govern de la Generalitat a proposta del conseller d'Economia i Finances.

La major part de les funcions que el Decret de creació de l'Institut d'Estadística de Catalunya atribueix al director es refereixen a la gestió immediata de l'organisme autònom i a la coordinació de les seves activitats. Des del punt de vista de la coordinació del sistema estadístic s'esmenten només dues funcions directes. D'una part, ha d'establir els contactes amb altres organismes i persones per a la consecució dels convenis que constitueixen la via de relació o participació amb el Pla estadístic i els programes anuals d'actuació estadística d'altres institucions de dins i de fora del sistema estadístic català o per a l'acompliment d'accions conjuntes. D'altra part, el director de l'Institut ha d'emetre les certificacions oficials pertinents sobre les estadístiques incloses en el Pla estadístic amb independència de quina sigui la institució responsable de la seva execució. Només l'Institut d'Estadística de Catalunya està facultat per emetre certificacions oficials sobre les estadístiques incloses en el Pla estadístic.

Ara bé, la gran tasca de coordinació del sistema des d'un punt de vista efectiu la fa el director de l'Institut basant-se en la seva obligació d'executar els acords que aprova la Junta de Govern que, com s'ha dit, està orientada, prioritàriament, a la gestió del sistema estadístic català.

L'estructura organitzativa interna de l'Institut quant a organisme executor no consta en el Decret de creació de l'Institut d'Estadística de Catalunya. Això incideix de nou en la idea ja expressada de flexibilitat organitzativa. Aquesta estructura interna es farà a partir del reglament de règim interior.

Provisionalment funciona una estructura que té el seu interès per raó de les seves característiques. L'Institut d'Estadística compta amb cinc subdireccions. Dues d'elles són verticals, la d'Estadístiques Econòmiques i la de Demogràfiques i Socials. Cadascuna d'elles compta amb tots els recursos personals tècnics necessaris, personal de camp i personal auxiliar i el maquinari i programari informàtics. Aquesta és una opció molt controvertida respecte a les diferents estructures organitzatives dels instituts d'estadística. L'altra opció és la centralització, en sentit horitzontal de les fases de camp i tractament informàtic. Cap d'elles no està exempta d'inconvenients o d'avantatges¹⁵.

Hi ha tres subdireccions de caràcter horitzontal. La Subdirecció General d'Assistència Tècnica Estadística és la més característica de l'Institut d'Estadística de Catalunya. No existeix una altra d'igual en els instituts d'estadística del nostre entorn. Per mitjà d'aquesta Subdirecció, el director de l'Institut executa els acords de la Junta de Govern relacionats amb la gestió i la coordinació del sistema. El personal d'aquesta Subdirecció manté les relacions amb les institucions o agents del sistema per mitjà dels interlocutors tècnics i les comissions

¹⁵Instrucció n.º 3 del director de l'Institut d'Estadística de Catalunya de 18 de gener de 1994. Document intern. Aquesta instrucció en modifica una d'anterior.

de seguiment dels convenis. Realitzen, per tant, les tasques preparatòries del Pla estadístic en col·laboració amb les institucions implicades. També fan el seguiment de les activitats estadístiques del Pla i l'esborrany de l'informe sobre la seva execució i compliment de les normes reguladores. Aquesta mateixa Subdirecció ha de presentar i impulsar les propostes sobre normatives de caràcter tècnic, de recerca i desenvolupament tècnic i de formació del personal.

S'ha de notar que, a més de les tasques descrites, l'assistència tècnica estadística té com a missió específica garantir la correcció tècnica i jurídica de les activitats estadístiques de tots els agents del sistema estadístic català. Però això no s'ha d'aconseguir de forma burocràtica i fiscalitzadora, sinó a partir de l'oferta de suport tècnic i instrumental als agents del sistema i l'organització de cursos de formació.

Hi ha dues altres subdireccions horitzontals més que són comunes a altres instituts i que corresponen a les tasques de difusió de resultats i a les de gerència o serveis interns. Aquesta integra el Centre de Procés de Dades i els Serveis d'Administració i Manteniment. La Subdirecció de Difusió s'encarrega de la publicitat dels resultats estadístics utilitzant tots els mitjans tradicionals i moderns de comunicació social i de la informació estadística. És característic de l'Institut d'Estadística de Catalunya el Servei de Documentació Estadística a partir de bases de dades accessibles des de l'exterior i des de la sala de consulta de la Biblioteca especialitzada que compta amb la més completa col·lecció de fons documentals sobre estadística espanyola i catalana des de 1960 de tot l'Estat.

Només resta ressenyar un altre aspecte organitzatiu de l'Institut d'Estadística de Catalunya, característic per omissió, que es refereix a l'absència d'una estructura territorial. No existeixen delegacions pròpies de l'Institut al territori de Catalunya. Això reforça la idea de flexibilitat, ja que algunes operacions estadístiques han requerit punts de recolzament de caràcter territorial i s'han resolt de diferent manera utilitzant sempre l'estructura territorial d'alguns departaments afins al tema d'investigació. A més dels ajuntaments existeix una xarxa potencial de suport de l'activitat estadística als consells comarcals els quals constitueixen una divisió territorial pròpia de Catalunya que encara no s'ha fet servir però que proporcionaria una bona base territorial en el futur.

En resum, l'Institut d'Estadística de Catalunya compta amb només dos òrgans, un de col·legiat que és la Junta de Govern que integra una exhaustiva representació dels departaments de la Generalitat i un altre d'unipersonal que és el director. L'estructura interna de l'Institut és flexible i comuna a altres instituts però amb la singularitat d'una àrea dedicada exclusivament a la funció bàsica de l'òrgan estadístic de la Generalitat: la gestió del sistema estadístic.

2.2.2. El Consell Català d'Estadística

El Consell Català d'Estadística és l'òrgan de caràcter consultiu i d'assessorament de l'Institut d'Estadística de Catalunya.

La idea bàsica de fons en la constitució d'aquests consells consultius en la majoria de sistemes estadístics és trobar representants de la societat civil a l'objecte de recollir les seves opinions sobre les necessitats d'informació estadística i la verificació de l'ajustament de l'oferta estadística a la demanda. Presidits per una persona relacionada amb l'estadística i nomenada pel Govern de la Generalitat a proposta del conseller d'Economia i Finances, els membres del Consell són vint-i-vuit. Entre ells hi ha una representació de la Junta de Govern a més del director de l'Institut d'Estadística de Catalunya que n'és el secretari. L'Administració local és present per una representació dels consells comarcals, una representació de les dues associacions d'ajuntaments, un representant de l'Ajuntament de Barcelona i un altre per a cada una de les quatre diputacions. Les administracions públiques compten amb divuit membres, cinc representants de la Generalitat i tretze de l'Administració local. L'estament universitari compta amb un representant per a cada una de les universitats catalanes. Les cambres i les associacions empresarials estan representades per un total de quatre membres. Hi ha dos representants dels sindicats i un de l'Associació de Consumidors de Catalunya.

El Consell Català d'Estadística té nombroses i variades competències que no apareixen en altres contextos propers. Especialment sobresurten aquelles atribucions del Consell en relació amb l'assessorament tècnic. En aquest sentit, no existeix una relació directa entre les especialitats dels representants socials i institucionals i aquesta missió tècnica. De tota manera, els representants de les universitats cobririen aquest aspecte però hauria d'interpretar-se pel que fa a interès pel seu caràcter coordinador i homogeneïtzador. Això no impedeix que entre els representants d'altres institucions hi hagi especialistes en temes estadístics de caràcter tècnic. En aquest capítol, el Consell Català d'Estadística ha de ser consultat sobre els conceptes, definicions i classificacions utilitzades en les estadístiques de totes les institucions que actuen en l'àmbit territorial de Catalunya.

Com correspon al pes del Pla estadístic en el sistema estadístic català la major part de les competències del Consell es refereixen a la seva participació en l'avantprojecte de la Llei del pla estadístic i el seguiment dels programes anuals d'actuació estadística, sempre pel que fa a assessorament i consulta. Una característica comuna a altres sistemes estadístics de les comunitats autònomes és l'especial atenció a les entitats públiques catalanes de caràcter territorial i les seves estadístiques.

És interessant, per la seva singularitat, la competència referida a l'estudi sobre l'equilibri relatiu de l'oferta estadística catalana respecte a les necessitats d'informació, però des de l'òptica de la seva relació amb l'estadística de la resta de l'Estat i de la Unió Europea.

2.2.3. Els departaments de la Generalitat

Com ja hem assenyalat no existeix una norma precisa sobre l'organització estadística dels departaments de la Generalitat. Això és congruent amb la idea de flexibilitat organitzativa tant en el sistema estadístic pel que fa a òrgans superiors, com en el cas de la pròpia estructura de l'Institut d'Estadística.

En canvi, els departaments de la Generalitat són els més importants productors d'estadístiques en el sistema estadístic català. No hi ha cap departament que no tingui assignades estadístiques en la seva participació obligatòria en el Pla estadístic. Estadístiques tan importants com les agràries, energètiques, sanitàries, laborals, de l'habitatge, obres públiques i d'ensenyament són fetes pels departaments responsables d'aquests sectors a la Generalitat. Es tracta d'estadístiques que requereixen la utilització d'instruments estadístics d'observació, d'anàlisi i tractament en sentit estricte. Moltes altres estadístiques són d'origen administratiu i són dutes a terme per tots els departaments, que produeixen, a més, transparència informativa sobre la gestió pública de la Generalitat.

Per tant, no és una frase retòrica l'afirmació que el model català és descentralitzat, però ho és en relació amb la producció i difusió estadístiques. És centralitzat en relació amb la gestió i la coordinació del sistema que es fa des de l'Institut d'Estadística. De tota manera, una observació atenta sobre l'òrgan superior de direcció de l'Institut d'Estadística condueix a la conclusió que els departaments de la Generalitat són a la gestió i coordinació del sistema per mitjà de la seva representació política a la Junta de Govern. El que succeeix en realitat és que el personal estadístic dels departaments desenvolupa activitats en la funció de producció d'estadístiques que té assignades el departament corresponent. El personal estadístic de l'Institut, en canvi, du a terme aquesta funció de producció d'estadístiques com a agent del sistema i a més realitza, en exclusiva, les activitats instrumentals que preparen la normativa tècnica de coordinació. L'exclusivitat, s'entén, es refereix a la missió específica de l'òrgan estadístic però no ho és respecte a les tasques d'estudi i definició que són consultades i de vegades compartides, amb el personal estadístic dels departaments.

Com ja hem assenyalat anteriorment cada departament té la seva pròpia estructura organitzativa del personal estadístic i de la seva activitat. Mancaria, no obstant això, la concreció funcional i jurídica de l'estatut del personal estadístic.

Pel que fa a la relació amb el personal estadístic de l'Institut s'ha arribat a una fórmula pràctica. S'ha posat en marxa una figura no existent en la normativa jurídica però que ha estat consensuada i acordada per la Junta de Govern de l'Institut d'Estadística en la qual estan representats tots els departaments. Aquesta figura és la de l'interlocutor tècnic del departament designat oficialment per a això i que ostenta la representació tècnica de la unitat o del servei d'estadística del departament en les seves relacions amb l'Institut d'Estadística. Aquestes relacions, com és obligat en el sistema estadístic català, es produeixen per mitjà del Pla estadístic i els programes anuals d'actuació estadística, tant en el moment de la seva proposta com en el seu seguiment o com a assistència per al seu compliment.

2.2.4. Les entitats públiques catalanes de caràcter territorial

Com ja s'ha dit anteriorment, l'instrument marc de la col·laboració institucional entre la Generalitat de Catalunya i els seus organismes i empreses d'una banda i les entitats públiques catalanes de caràcter territorial d'una altra, és la participació en el Pla estadístic. Però, com s'explica a continuació la via legal i la fórmula formal d'aquesta relació és el conveni específic per mitjà de l'Institut d'Estadística de Catalunya. S'ha de posar de relleu que es tracta de la porta d'entrada al sistema estadístic integral de Catalunya¹⁶.

La legislació catalana reconeix la competència de l'Administració local, per a l'elaboració d'estadístiques del seu interès i en compliment de les seves funcions de govern i d'administració. Aquestes poden ser fetes dins o fora del Pla estadístic. Si estan dins del Pla, passen a ser estadístiques oficials, si estan fora del Pla estadístic no són estadístiques oficials, és a dir, no gaudeixen del privilegi de col·laboració obligatòria, només serien estadístiques públiques però no estadístiques oficials. Però hi ha una excepció que es reserva per a aquella estadística de l'Administració local a la qual el Parlament atribueixi l'esmentat privilegi. Però ha de ser a petició de l'entitat territorial al Parlament i mitjançant el Govern i amb un informe favorable de l'Institut d'Estadística. En qualsevol cas, totes les estadístiques s'han de sotmetre a les obligacions de la Llei d'estadística¹⁷.

La via ordinària per a l'exercici de la competència estadística de les entitats territorials és la participació en el Pla estadístic de Catalunya mitjançant conveni amb l'Institut d'Estadística de Catalunya per a un programa concret.

¹⁶Llei del pla estadístic de Catalunya esmentada. Art. 2.

¹⁷Llei d'estadística de Catalunya esmentada. Art. 13, 14 i 15.

No es tracta d'una petita porta d'entrada al sistema estadístic integral de Catalunya sinó l'afirmació de la voluntarietat del procés quant al fet que el Parlament respecta la llibertat de col·laboració d'aquestes entitats. Per altra part, la capacitat tècnica diversa i la diferent capacitat econòmica de les entitats territorials no suportaria la participació obligatòria en la producció estadística.

Existeixen dos tipus de participació de les entitats territorials en el Pla estadístic. La participació activa consisteix a proposar l'elaboració d'estadístiques coincidents amb els objectius previstos en el Pla, com a organismes responsables de l'execució en l'àmbit territorial propi. Existeix un altre tipus de participació que es pot qualificar de passiva que consisteix a requerir, també mitjançant conveni, la realització d'explotacions específiques a un nivell de desagregació territorial superior al previst o la inclusió d'aspectes temàtics específics en la investigació estadística¹⁸.

Els convenis entre aquestes entitats territorials i l'Institut d'Estadística són declaracions protocolàries, ja que impliquen veritables obligacions amb efectes contrastables. Per això, es crea una comissió de seguiment integrada per les parts i que està integrada per tècnics amb presència política. En aquest cas, la tasca d'assistència tècnica i l'activitat de seguiment del compliment del programa es pot superposar amb la feina de la comissió del conveni.

A la pràctica, donat el nombre elevat d'entitats territorials a Catalunya, la fórmula del conveni anual resulta difícil pel que fa a la seva formalització administrativa per mitjà dels òrgans de representació de les pròpies entitats i de l'Institut i la pròpia gestació i perfeccionament dels documents incideixen en un consum excessiu de temps de l'Institut, només enfront de nombroses entitats. De tota manera, la línia de solució vindrà, d'una part, per la rellevància de la modalitat de la participació activa en els programes anuals d'actuació estadística en els quals s'adquireix un compromís i una responsabilitat per part de les entitats territorials. De fet, el contingut de la participació passiva podria coincidir amb alguns dels serveis que presta l'Institut i que també tenen un fonament legal. Els convenis, en alguns casos, haurien de ser de col·laboració genèrica.

Malgrat la dificultat esmentada s'ha de destacar que la inclusió de l'activitat estadística de les entitats territorials ofereix una imatge integradora i també resulta complementària des de l'òptica de la informació estadística referida al territori que té una gran demanda. Al mateix temps havien de ser un instrument de coordinació, ja que només l'Institut d'Estadística de Catalunya pot fer convenis en nom propi o dels departaments amb les entitats territorials¹⁹.

¹⁸Llei del pla estadístic de Catalunya esmentada. Art. 23 al 27.

¹⁹Llei del pla estadístic esmentada. Art. 27.

Una situació especial en aquest context del sistema estadístic català és la dels consells comarcals. Catalunya posseeix una divisió territorial administrativa pròpia que comprèn quaranta-una comarques, que compten amb els seus respectius òrgans de govern: els consells comarcals esmentats. Aquests consells tenen atribuïdes, legalment, competències estadístiques i estan obligats a realitzar el Pla d'actuació comarcal que requereix molta informació estadística. Existeixen dos problemes que s'han de solucionar progressivament: el risc que l'estadística sigui esporàdica i la manca de recursos tècnics.

Els consells comarcals són òrgans molt propers a les necessitats específiques i diferencials de territoris relativament poc extensos i, en general, amb un nombre reduït d'habitants. Les excepcions comprenen zones de màxima concentració de l'activitat econòmica industrial que, a Catalunya, és molt acusada. Això representa avantatges i inconvenients respecte del funcionament del sistema estadístic català. Els avantatges de la progressiva integració dels consells comarcals en el sistema estadístic de Catalunya se situen en la potencialitat instrumental de ser plataformes territorials intermèdies de recollida d'informació dels agents governamentals del sistema coordinats per l'Institut i unitats de producció estadística, per elles mateixes, en operacions de caràcter censal que no tindrien lloc si no existissin aquestes entitats. Això podria suplir la manca de xarxa territorial de l'Institut.

Des d'un punt de vista organitzatiu, no se citen les unitats d'estadística de les entitats de caràcter territorial. Aquestes, no obstant això, existeixen en l'organització interna dels ajuntaments grans i en alguna diputació però pateixen de manca de regulació del secret estadístic i d'estatut del personal estadístic.

3. LA COORDINACIÓ DEL SISTEMA ESTADÍSTIC CATALÀ

Quan es parla de la coordinació del sistema estadístic s'és conscient que es tracta d'una expressió tan àmplia que desborda les possibilitats d'acotació o delimitació. En aquest mateix escrit s'ha intentat una aproximació a una definició general del sistema estadístic i s'han assenyalat les principals característiques del sistema estadístic català. Hem descrit els òrgans superiors, les seves estructures organitzatives i alguns aspectes funcionals i s'ha fet esment dels agents que la integren. Han quedat apuntats els instruments legals que regulen i que planifiquen l'activitat estadística de Catalunya, destacant-ne només alguns aspectes. Però tot això ofereix una visió parcial del conjunt d'elements a coordinar, però al mateix temps pot suggerir la imatge de desordre i complexitat.

Situats en la mateixa idea de sistema, la coordinació és una funció que exerceix algú sobre altres que han de ser coordinats. Però no es tracta de coordinar directament els agents del sistema sinó de coordinar la seva activitat estadística. Per això, el sistema s'ha de dotar dels instruments necessaris de coordinació. Això es pot aconseguir de maneres diferents.

3.1. El Pla estadístic i els programes anuals d'actuació estadística

Existeix un primer element objectiu de coordinació a partir del Pla estadístic i dels programes anuals d'actuació estadística. El Pla estadístic és una llei que assenyalava quins objectius operatius, en el pla del coneixement, ha d'aconseguir en el període de quatre anys. Els programes anuals d'actuació estadística són decrets del Govern que concreten les activitats estadístiques que s'han d'executar per aconseguir els objectius del Pla estadístic a la pràctica.

Quan diem que el Pla i els programes són elements objectius de coordinació, s'entén en el sentit que ambdós són instruments que, en principi, en la seva preparació ja enclouen una base de coherència, tant en el conjunt d'objectius entre si com en les activitats, que el seu seguiment i la seva execució ha de resultar coordinat sense necessitat d'una altra intervenció de qualsevol subjecte o agent. La primera activitat de coordinació ja s'ha fet a l'hora de seleccionar les propostes d'objectius i activitats sobre els mateixos llistats. La base de la coordinació de les activitats estadístiques radica, també, que les normes són precises i concretes de manera que s'elimina qualsevol risc de duplicitat o de dubte sobre quines són les activitats estadístiques que s'han de realitzar dins del sistema estadístic català i quines no s'han de fer perquè no són necessàries o perquè un agent de fora del sistema, per exemple l'INE, les du a terme.

Com a il·lustració per a aquells lectors que no han pogut observar de prop el Pla estadístic català, es reproduïx un objectiu dels 188 que aquest preveu:

“3.5. Estadístiques de salut.

Conèixer l'estat de salut de la població catalana fent el disseny i l'execució gradual d'una enquesta a les famílies. Operació complexa, de cost alt i disponibilitat a mitjà termini²⁰.”

Tots els objectius operatius mantenen el mateix esquema de redacció: un títol, un infinitiu en el pla del coneixement a obtenir, un gerundi en el pla pràctic d'actuar i una valoració orientativa que estableix límits quant a la complexitat

²⁰Llei del pla estadístic de Catalunya esmentada. Annex III, 3.5.

tècnica i cost econòmic de l'operació que s'espera. Per a un professional de l'estadística oficial la redacció anterior oferiria pocs dubtes sobre l'activitat a dur a terme. Però la Llei obliga el Govern a realitzar un programa amb la descripció precisa de cada activitat estadística per a coneixement de qualsevol ciutadà obligat a col·laborar en una determinada activitat estadística.

Una il·lustració del Programa anual d'actuació estadística de 1994 que es correspon amb l'objectiu ressenyat seria excessivament extensa perquè l'explicació de cada activitat ocupa una pàgina del *Diari oficial de la Generalitat de Catalunya* (DOGC). Però en síntesi diu:

“El Departament de Sanitat i Seguretat Social ha de fer amb la col·laboració de l'Institut d'Estadística de Catalunya una enquesta per mostreig dirigida a la població en general. Es farà per entrevista personal; el qüestionari inclourà, a més dels ítems referits a la salut de la persona enquestada, variables sociodemogràfiques o estat d'opinió sobre la utilització dels serveis públics sanitaris. Els resultats hauran d'estar disponibles passats dos anys. Es farà una publicació impresa amb els resultats desagregats per regions sanitàries. El mateix Departament és també el responsable de la publicació²¹.

Un altre element de coordinació derivat de les pròpies normes és l'assignació de la responsabilitat d'execució d'una activitat a un departament o a una entitat concrets. En relació amb les activitats estadístiques que estaven en curs en el moment de la redacció del Pla estadístic, la pròpia a assignar els objectius als departaments que havien de continuar-les. En canvi, els objectius per a estadístiques noves no van tenir assignació concreta en el Pla estadístic però sí en el Programa anual d'actuació estadística.

Per tant, la Llei del pla estadístic i els decrets dels programes anuals d'actuació estadística ofereixen un conjunt coherent d'objectius i activitats que permet saber quines són les activitats estadístiques necessàries per dur a terme i quins agents del sistema tenen la responsabilitat de portar-les a terme. Tot això representa un element objectiu de coordinació que no es podria garantir si l'activitat estadística no estigués planificada.

És per això que la pròpia Llei d'estadística considera el Pla estadístic com a l'instrument d'ordenació de l'estadística de la Generalitat i de les entitats públiques catalanes de caràcter territorial que hi participen i dels organismes i empreses que en depenen. També, per aquesta raó s'explica el radicalisme respecte a la participació en el Pla estadístic per tal que l'estadística pública sigui considerada estadística oficial de Catalunya.

²¹Decret 331/1993, de 28 de desembre, pel qual s'estableix el Programa anual d'actuació estadística per a 1994. DOGC. N.º 1.857. 8.2.1994 pàg. 865.

3.2. Nomenclatures, classificacions, i codis territorials oficials

L'estadística oficial ha de ser comparable per exigències no només tècniques, sinó també per eficàcia administrativa. Però en el sistema estadístic català, com en altres, és també una exigència legal expressa. Es tracta d'un tema de coordinació d'alta complexitat i eficàcia.

Existeixen elements o condicions necessàries per a la comparabilitat, tant horitzontal com longitudinal en el temps, que es deriven de la univocitat conceptual i de la uniformitat dels mètodes de recollida i tractament estadístic de les observacions, que són pràcticament inassolibles. És un problema l'estudi del qual és d'actualitat candent i que tracta d'obtenir millores amb ajut dels "diccionaris comuns de dades" i amb les "bases de dades de meta-informació". De moment és un aspecte de la coordinació feta per mitjà d'agents del sistema i que no és un tema a tractar per normativa objectivable. Per això no el considerarem aquí, encara que podria ser objecte d'un escrit en un altre moment.

En canvi, ja fa bastant temps que els instituts d'estadística proposen classificacions concretes i nomenclatures referides a les activitats econòmiques o a les ocupacions professionals, com ara la CCAE (Classificació catalana d'activitats econòmiques), la CCO (Classificació catalana d'ocupacions), PRODCOM (Producció comunitària), la CCBS (Classificació catalana de béns i serveis), a les quals s'han d'afegir les nombroses classificacions que s'utilitzen en les estadístiques del comerç exterior. Les denominacions i les sigles poden canviar segons l'idioma o el país, però els conceptes i les definicions han de coincidir.

Cada sector de l'activitat estadística té la necessitat de denominar amb termes precisos les característiques diferencials de cada element observat en la investigació. En aquest cas, es requereix el terme precís seleccionat entre altres termes sinònims, que millor i més universalment caracteritzi l'element observat. Si cadascun dels agents fa aquest procés lliurement, les estadístiques no seran comparables. Quan aquests termes passen a ser obligatoris per a tots els agents del sistema estadístic som davant d'una "nomenclatura oficial". D'aquesta manera, es poden comparar les diferents estadístiques. Per això, cal, a més, que els termes estiguin ben aplicats d'acord amb les notes explicatives que ajuden a la identificació terminològica. L'efecte pràctic visible es produeix en la fase de difusió i publicació de resultats.

El pas següent consisteix a classificar els elements i els seus termes. La classificació és una nomenclatura estructurada. Aquesta estructura es fa d'acord amb criteris de categorització o reunió d'elements relacionats entre ells en funció de la matèria. S'exigeix un coneixement elevat del tema en estudi per obtenir aquesta ordenació. Dins de cada categoria els diversos elements s'ordenen de

forma inclusiva des del més general fins a l'element menys divisible, o en el seu cas, indivisible i únic. És freqüent trobar classificacions particulars utilitzades per cada investigador. Aquestes poden tenir molta qualitat tècnica però no permeten la comparabilitat. Per tot això, és necessari que tots obligatòriament n'utilitzin una. D'aquí la necessitat de les classificacions oficials.

Les diferents classificacions posseeixen un tercer requisit que constitueix el sistema de codificació. Cada categoria i els seus elements s'identifica amb un número que, de forma sistemàtica, assigna un dígit per cada nivell d'inclusió segons el sistema decimal. Aquest codi podria ser també alfabètic i alfanumèric i seguir una ordenació lliure segons l'autor de l'estadística. De vegades passa així. En canvi, en els sistemes estadístics és obligatòria una codificació oficial. No obstant això, el problema de comparabilitat no seria tan greu si existeix una taula de correspondència biunívoca.

Els mateixos processos se segueixen en els sistemes de codificació geogràfica. La diferència rau en la seva més gran facilitat perquè les denominacions o els topònims ja estan fixats encara que, de vegades, es requereix una tasca prèvia de normalització sobretot en la seva ortografia. De la mateixa manera, les categories estan establertes per les divisions administratives a diferents nivells. En aquest cas, el marge de llibertat de l'investigador es redueix al sistema de codis. Llibertat que tampoc no existeix en el sistema estadístic, que imposa una codificació obligatòria i comuna per a tots.

Com queda patent, el valor coordinador de les nomenclatures, classificacions i sistemes de codificacions és importantíssim. Aquestes constitueixen, juntament amb el Pla estadístic i els programes anuals que garanteixen sobretot la coherència de les estadístiques, l'instrument de coordinació que permet la comparabilitat de les estadístiques. Ja s'ha dit també que l'homogeneïtat tècnica requereix la uniformitat dels conceptes i les definicions operatives. Aquest darrer aspecte és el més complicat d'assolir i per això no s'ha produït, de forma generalitzada, la internacionalització que ha tingut lloc en el cas de nomenclatures i classificacions. De la mateixa manera que en la creació de les classificacions internacionals inicials, cada vegada que se'n modifica o se n'actualitza alguna, s'actua de forma coordinada i consensuada entre els diferents sistemes estadístics. Encara que els sistemes estadístics autonòmics espanyols —que gaudeixen de competències exclusives— són independents entre ells i en relació amb el sistema estadístic de l'Administració central que també és independent respecte a la resta de sistemes d'altres països, es pot afirmar que existeix una gran permeabilitat en relació amb el tema de les classificacions. En la Unió Europea s'ha produït un canvi favorable a partir del suport legal EUROSTAT i per tant obligatori per als països membres, que s'ha desenvolupat en aquests últims deu anys. En l'àmbit mundial els organismes internacionals de major projecció en

l'activitat estadística com l'OIT, l'OMS o la UNESCO han coordinat de forma indirecta els sistemes estadístics de cada país.

Per últim, s'ha d'assenyalar que segons la legislació catalana, correspon a l'Institut d'Estadística de Catalunya proposar la normativa tècnica en aquests temes, encara que amb caràcter estadístic. Una de les prioritats del Pla estadístic vigent és el desenvolupament d'aquesta activitat estadística instrumental.

Un altre aspecte important de la coordinació a partir de les nomenclatures i classificacions és la seva influència en la gestió administrativa que condueix a la creació d'arxius i registres normalitzats de forma indirecta. Aquest és un tema que, en algunes normes catalanes sobre codis geogràfics de caràcter estadístic, s'ha tingut en compte de forma explícita i s'ha estès el seu ús obligatori en la gestió administrativa²².

3.3. El paper coordinador de la Junta de Govern de l'Institut d'Estadística de Catalunya

La Junta de Govern de l'Institut d'Estadística de Catalunya té atribuïts dos tipus de funcions: aquelles que estan adreçades al propi òrgan estadístic i aquelles que es refereixen a la gestió del sistema estadístic.

Es destaquen les funcions de la Junta de Govern que interessen més en relació amb el sistema estadístic. Aquestes es poden resumir en les que segueixen:

- Sobre la planificació i la programació de l'activitat estadística.

La Junta de Govern de l'Institut d'Estadística de Catalunya efectua les propostes de pla estadístic a partir d'un inventari de les necessitats d'informació estadística. El Govern presenta un projecte de llei de pla estadístic basat en aquesta proposta. Aquesta funció és importantíssima si es té en consideració que el Pla estadístic és l'eix del sistema estadístic català. De la mateixa manera, la Junta de Govern de l'Institut fa la proposta de cada programa anual d'actuació estadística que serà aprovat pel Govern de Catalunya. El programa és molt concret i inclou les normes reguladores de cada estadística. Aquestes consisteixen a determinar quina activitat s'ha de fer, per part de quin organisme, com i quan estarà disponible i quin serà el mitjà de difusió dels resultats, fonamentalment.

²²Ordre de 25 de febrer de 1993, per la qual s'estableix un nou sistema de codificació de les entitats locals, ens de gestió i consorcis amb participació pública local, amb finalitats estadístiques.

- *Sobre la correcció tècnica de les estadístiques oficials de Catalunya.*

Per tal d'assolir la millor qualitat i la comparabilitat de les estadístiques, la Junta de Govern de l'Institut prepara i proposa les normes tècniques que per Decret del Govern o per ordres del conseller seran d'obligat compliment per a l'elaboració de les estadístiques de la Generalitat i de les entitats territorials. Aquest tipus de normativa pretén també l'homogeneïtat metodològica i és un instrument addicional adequat de coordinació del sistema estadístic català que adquireix una transcendència en les normatives sobre classificacions, nomenclatures i codis. (En la pràctica traspasa els límits estadístics i afecta processos administratius de forma indirecta).

- *Sobre el compliment de la legalitat estadística i, especialment, la defensa dels drets dels ciutadans per mitjà de l'observança del secret estadístic.*

La Junta de Govern prepara i promou normes, notes i informes jurídics sobre el secret estadístic i mitjançant el personal destinat a aquesta finalitat, fiscalitza els processos de recollida i tractament de la informació estadística dels organismes responsables dins del sistema estadístic. Per mitjà de l'assistència tècnica de l'Institut col·labora amb els serveis jurídics d'altres departaments de la Generalitat i de les entitats territorials sobre els aspectes de dubtosa aplicació de la normativa a les situacions concretes.

- *Sobre la utilització d'arxius i registres administratius per a l'obtenció de resultats estadístics i sobre els requisits i normes tècniques relatives a les tasques de formació, conservació i actualització d'aquests arxius i registres.*

Aquesta és una funció que en el sistema estadístic català és operativa a partir que constitueix un dels criteris establerts per la Llei del pla estadístic. Aquest consisteix a esgotar aquesta via administrativa com a prioritària i alternativa a les estadístiques que requereixen treball de camp amb la finalitat de no molestar innecessàriament el ciutadà i estalviar recursos econòmics. L'explotació estadística dels fitxers administratius que són necessaris per a l'exercici d'altres competències no estadístiques ha tingut un reflex directe en cent vuit activitats previstes en el Programa anual d'actuació estadística de 1994 que utilitzen aquesta font d'informació. Això afavoreix, a més, la transparència informativa de l'Administració pública mitjançant la difusió.

No obstant això, hem de dir que no s'han abordat justament arxius administratius que, per la seva dificultat especial, requeririen un estudi complex del seu funcionament i les propostes tècniques de modificació conseqüents. Aquest tema s'ha fet més complex amb les noves exigències derivades de la Llei orgànica del tractament automatitzat de dades personals (LORTAD) que, entre altres qüestions, planteja la necessitat de regulació dels propis arxius. Malgrat tot, existeix una activitat inicial per a la formació d'un

nou arxiu de caràcter administratiu en l'àmbit de la Generalitat que podria ser marc per a l'extracció de mostres estadístiques.

- *Sobre l'aprovació de convenis de col·laboració amb altres institucions públiques i privades per a l'actuació conjunta en l'àmbit estadístic.*

La figura del conveni és la via que preveu la Llei del pla estadístic per a la participació d'altres entitats públiques en el propi Pla. Aquesta via s'ha utilitzat amb profusió a l'Institut en relació amb les entitats públiques catalanes de caràcter territorial i amb l'INE. En aquest sentit, els convenis amb les entitats de caràcter territorial ha significat l'eixamplament i la constitució progressiva del sistema estadístic integral de Catalunya. Per tant, és una funció directament inclosa en el sistema estadístic. La inclusió en el Pla de les estadístiques públiques d'alguna entitat territorial amb un pes relatiu molt notable a Catalunya ha significat un element coordinador molt important.

- *Sobre la realització de l'informe preceptiu per a l'aprovació dels resultats de les activitats estadístiques oficials.*

L'aprovació de resultats estadístics és un tràmit administratiu quasi exclusiu de Catalunya. En altres sistemes estadístics es dona per suposat que l'ordre de publicació per part del responsable de l'organisme és un tràmit d'aprovació suficient. Entre altres raons perquè no existeix un concepte unívoc de l'expressió d'aprovació de resultats. A Catalunya es tracta d'un concepte jurídic, no tècnic, almenys en primera instància.

És comú a les institucions estadístiques un procés de revisió i validació tècnica dels resultats estadístics i, és clar, també a Catalunya. Però l'expressió d'"aprovació de resultats" significa, a Catalunya, una declaració favorable sobre el compliment de les normes reguladores generals i particulars que afecten cada estadística.

De forma molt resumida, les normes reguladores generals de les activitats estadístiques són l'expressió dels principis i les garanties jurídiques que imposa la Llei d'estadística i l'ajustament de l'activitat a algun objectiu del Pla estadístic. En canvi, les normes reguladores específiques són pròpies de cada activitat i coincideixen amb l'esquema metodològic que correspon a cadascuna. Per tant, en segona instància, significa que s'han aplicat les millores tècniques possibles en la relació cost/benefici. S'espera, en conseqüència, que els resultats siguin estadísticament correctes.

Per tant, l'aprovació de resultats a Catalunya és una declaració de cada Conseller sobre les estadístiques del seu Departament on es diu que la seva realització s'ha ajustat al que estava previst per la Llei i pel Programa d'actuació estadística. És a dir, com que és molt difícil afirmar que els resultats són correctes o fiables des de fora de l'autor o autors de l'activitat,

es comprova que s'ha complert la Llei i que el mètode és correcte. Per això, abans d'aquesta declaració l'Institut d'Estadística emet un informe preceptiu sobre el compliment de les normes reguladores de cada estadística.

Hi ha altres obligacions addicionals a les anteriors i que han estat afegides per la Llei del pla estadístic però que es recollien també a la llista de funcions de l'Institut d'Estadística que assenyalava la Llei d'estadística²³.

- *Sobre la formulació i el manteniment de l'inventari de la informació i la documentació estadístiques a Catalunya.*

L'objectiu de l'activitat estadística és aconseguir la disponibilitat de la informació estadística. El principal suport de la informació estadística continua essent el paper, és a dir, el llibre. D'aquí que s'uneixin en una mateixa funció. Com ja s'ha dit, part de la informació estadística sobre Catalunya es produeix fora del sistema estadístic català. Però, encara que no fos així en aquest moment, les sèries històriques van ser produïdes per l'INE i les ministeris. El mínim que s'exigeix és l'inventari que inclou la relació i la descripció de les estadístiques existents. Quant al suport de la informació s'ha d'entendre també com a mínim. A l'actualitat, adquireix gran importància l'inventari referit a les estadístiques del propi sistema estadístic català. Aquesta és doncs una obligació de la Junta de Govern de l'Institut d'Estadística de Catalunya.

- *Sobre el seguiment de les estadístiques incloses en el Pla estadístic.*

S'assigna a l'Institut d'Estadística la vigilància del compliment de les activitats estadístiques que han de fer els agents del sistema estadístic català. La pròpia Llei del pla estadístic estableix que el Govern pot atribuir a l'Institut d'Estadística de Catalunya "la facultat inspectora de les activitats estadístiques sotmeses al Pla estadístic". Una de les activitats instrumentals atribuïdes a l'Institut d'Estadística de Catalunya en els programes anuals d'actuació estadística consisteix en el seguiment de l'activitat de tots els agents. Com a resultat de l'estudi detallat de cada activitat estadística, la Junta de Govern aprova cada any un informe en el qual consta l'estat d'execució i la informació disponible a cadascuna de les activitats. L'esmentat informe serveix de base a les resolucions sobre aprovació de resultats ja esmentades. D'aquestes resolucions, el Govern en dóna compte, anualment, al Parlament de Catalunya.

²³ L'article 6 d'aquesta Llei assenyalava, de forma clara i contundent, les obligacions de l'Institut d'Estadística respecte al sistema estadístic. La llista de funcions de l'Institut d'Estadística estan recollides en l'article 45 de la Llei d'estadística i en l'art. 2 del Decret de creació de l'Institut.

- *Sobre la prestació d'assistència tècnica a les administracions públiques catalanes que elaboren les estadístiques del Pla estadístic.*

Aquesta és una funció singular del sistema estadístic català. Està orientada directament a fer que el sistema funcioni. Per tractar-se d'una activitat oberta de coordinació que pot incloure molts aspectes i matisos la tractarem en l'apartat següent.

3.4. Assistència tècnica de l'Institut d'Estadística de Catalunya

Es tracta d'una activitat condicionada al requeriment de les institucions. No representa cap tipus de relació de superioritat i inferioritat entre les institucions però sí una divisió d'especialitzacions. Se suposa que les institucions que realitzen les estadístiques del Pla estadístic estan especialitzades en el tema sectorial en què actuen i que l'Institut s'especialitza en els instruments metodològics, en les normes sobre classificacions i codis en la coordinació del conjunt. És normal que les institucions productores d'estadístiques comptin amb personal estadístic, per la qual cosa cobren una importància més gran en l'assistència tècnica els aspectes normatius i de coordinació que els instrumentals estadístics. Entre aquests últims, només es poden assenyalar els referits al mostreig probabilístic i a l'organització de grans operacions de camp com a aspectes més específics de l'Institut.

No obstant això, s'ha de tenir en consideració l'existència de diversitat de nivells tècnics del tipus d'activitat i la novetat dels processos estadístics d'algunes entitats que participen en el Pla estadístic. Alguns processos constitueixen una novetat i un camp encara experimental en el curriculum de qualsevol estadístic professional per la qual cosa l'assistència tècnica consisteix en l'estudi compartit entre tècnics. Per exemple algunes estadístiques de medi ambient.

Ara bé, com que la funció d'assistència tècnica és una oferta de l'Institut, aquesta s'ha de definir institucionalment. Així doncs, l'activitat d'assistència tècnica de l'Institut d'Estadística de Catalunya consisteix a donar suport tècnic a les institucions que participen en l'execució del Pla estadístic i els programes anuals d'actuació estadística per tal d'assegurar la correcció tècnica de les operacions, la comparabilitat dels resultats i la legalitat estadística.

El concepte oposat a assistència tècnica és el de substitució parcial o total, per part de l'Institut, de les activitats estadístiques dels agents productors del sistema estadístic català. Aquesta alternativa és contrària als principis que regeixen el model organitzatiu d'aquest sistema. No hi ha subsidiaritat.

La responsabilitat jurídica de cadascun dels agents del sistema estadístic català s'estén al compliment d'aquelles exigències tècniques que garanteixin la

qualitat i la comparabilitat dels resultats estadístics, en un règim de completa autosuficiència de recursos i autonomia tècnica.

L'assistència tècnica estadística de l'Institut als agents del sistema estadístic comprèn aquelles tasques d'ajut directe al personal que treballa en les diferents fases de definició dels projectes, en els seus aspectes conceptuals, en les tècniques i instruments d'observació, com ara qüestionaris i formularis, en el disseny i selecció de mostres estadístiques, en els elements organitzatius de recollida de dades, en els sistemes de seguiment, depuració i enregistrament de dades, plans de tractament i tabulació de la informació i en els diferents mitjans de difusió de resultats. En suma, en totes aquelles tasques necessàries per tal de garantir la correcció tècnica de la metodologia que han d'emprar en les activitats estadístiques pròpies. Indirectament, gran part dels cursos de formació que organitza l'Institut s'han d'inscriure en aquesta línia.

És també una activitat d'assistència tècnica la divulgació general i específica de les normes de coordinació com les de codis, nomenclatures, classificacions i d'altres elements que assegurin la comparabilitat de les dades i dels resultats estadístics.

De la mateixa manera, la vigilància de la legalitat estadística conté elements d'assistència tècnica en els assessoraments puntuals sobre la interpretació de les lleis i les normes sobre els procediments, les tècniques i els mètodes que faciliten la preservació del secret estadístic i la defensa dels drets dels ciutadans i la publicitat de la informació estadística.

Finalment, l'Institut presta assistència tècnica als agents del sistema estadístic en els treballs preparatoris del Pla estadístic i dels programes anuals d'actuació estadística, en la formulació d'objectius i en la descripció de les activitats estadístiques. En compliment de la Llei del pla estadístic de 1992-1995, l'assistència tècnica de l'Institut s'orienta també en ajuts puntuals per a l'observança dels criteris sobre prioritització d'activitats i, molt especialment tal i com preveu l'objectiu central del Pla estadístic, en l'aprofitament dels arxius i registres administratius com a font d'informació estadística.

3.5. Relacions amb altres sistemes estadístics i l'autocoordinació

El sistema estadístic català no està tancat en ell mateix encara que no constitueix un subsistema del sistema estadístic central espanyol format per l'INE i els ministeris. No obstant això, la via d'obertura fonamental és la del

conveni interinstitucional i com és característic a partir de la participació en el Pla estadístic i els programes anuals d'actuació estadística²⁴.

A diferència dels convenis de col·laboració entre l'Institut d'Estadística de Catalunya i les entitats catalanes de caràcter territorial, els convenis amb l'INE o amb els ministeris no produeixen l'entrada en el sistema estadístic català sinó en el Pla estadístic i els programes com a col·laboradors d'altres sistemes. Per tant, l'obertura del sistema a altres sistemes no persegueix la formació d'un sistema de tercer grau sobre el de la Generalitat i l'integral de Catalunya. En canvi, com ja hem dit, les entitats territorials catalanes participen i estan en el sistema estadístic integral de Catalunya.

Aquestes relacions institucionals amb l'INE i els ministeris persegueixen tres objectius: la coherència, l'homogeneïtat i la comparabilitat del sistema estadístic de Catalunya amb les del seu entorn. Es tracta de l'autocoordinació²⁵.

La raó de considerar la coherència del sistema estadístic català, a més d'interna, fora del sistema té un sentit lògic. Per raons històriques, la producció estadística de l'INE i dels ministeris era l'única existent i abastava el territori estatal i les seves divisions administratives. Encara en l'actualitat, la major part de les estadístiques importants sobre Catalunya són produïdes per organismes del sistema estadístic del Govern central. El bon sentit de l'eficàcia administrativa condueix a evitar la duplicitat en la producció i per tant a la col·laboració interinstitucional. D'aquí que es persegueixi la coherència entre les estadístiques a dur a terme, qui ha d'assumir la responsabilitat d'executar-les i com s'han de difondre. Tot això s'ha de concretar en els convenis.

És d'interès políticoadministratiu notar que la coherència entre els sistemes estadístics autonòmics i del Govern central es podria haver produït mitjançant diferents fórmules de relació. A títol d'exemple il·lustratiu que estalviarà raonaments extensos es pot citar l'organització estadística descentralitzada alemanya entre l'activitat estadística dels Ländes i l'Institut Central d'Estadística. Són aquells els que nodreixen d'estadístiques el propi govern i el govern federal per bé que la coordinació és, de vegades, complexa. Podria també haver-se adoptat en termes teòrics el model francès que, com ja és sabut, és centralista tant pel que fa a l'INSEE com dels ministeris. En aquest cas l'Institut central té les seves delegacions en les regions sense cap problema de coordinació i coherència de la producció estadística però l'autonomia i l'adaptació a les necessitats regionals són escasses. És també notable el fet que a l'Estat espanyol de les autonomies la majoria dels serveis estadístics dels ministeris centrals han seguit un model de relació amb els seus homònims de les autonomies més semblant al sistema

²⁴Llei del pla estadístic de Catalunya esmentada. Art. 3.

²⁵Llei del pla estadístic de Catalunya esmentada. Art. 3.

alemany. En general, juntament amb la transferència de competències sectorials es va produir també el traspàs de les activitats estadístiques. En aquests casos, les estadístiques són fetes per les respectives conselleries de les comunitats autònomes, les quals envien els resultats als ministeris per tal que aquests realitzin les agregacions pertinents i obtenir, així, les estadístiques d'àmbit estatal. Atesa l'heterogeneïtat competencial de les comunitats autònomes i el seu desenvolupament estadístic diferent, les situacions són també diferents. S'han posat de manifest problemes de coordinació que, en alguns casos, han derivat en la interrupció d'algunes sèries estadístiques importants²⁶.

Tota aquesta problemàtica organitzativa de la producció estadística espanyola és conseqüència de la indefinició normativa en les relacions entre sistemes estadístics, la gran diferència de recursos tècnics i econòmics entre les comunitats autònomes entre elles mateixes i amb el Govern central i la concurrència de competències exclusives en matèria estadística. En principi, l'única via, tant des de l'òptica de l'INE com des del sistema estadístic català per tal d'aconseguir la coherència de l'activitat estadística és la del conveni de col·laboració.

El segon objectiu que persegueixen els convenis de col·laboració institucional entre la Generalitat de Catalunya i les entitats públiques estatals és l'homogeneïtat metodològica.

Existeixen dos aspectes a distingir en relació amb l'homogeneïtat metodològica. Un aspecte instrumental tècnic i un altre de caràcter normatiu. El primer està obert a la llibertat de l'exercici professional estadístic i s'orienta a la consecució de la correcció tècnica en aplicació dels coneixements i de les experiències més avançades que serveixin millor a l'objectiu d'investigació. L'aspecte de caràcter normatiu es refereix a classificacions, nomenclatures i codis que s'han d'observar en el sistema estadístic.

La col·laboració institucional per aconseguir l'homogeneïtat metodològica no és un problema si es decideix que aquests aspectes han de ser definits pel responsable de l'activitat estadística i aquesta responsabilitat ja està assignada. En aquesta hipòtesi la institució responsable de l'activitat decideix el disseny metodològic i, per tant, la institució col·laboradora accepta la metodologia prevista com una qüestió de fet. Per tant, el tema de l'homogeneïtat metodològica es remet al tema de la coherència i l'evitació de la duplicitat que hem descrit. Ara bé, quan es tracta d'activitats estadístiques diferents en cada sistema estadístic,

²⁶Casco Robledo, José A. "La coparticipación como condición para la eficacia de la estadística pública, estatal, autonómica y local", 1988. Ponència encarregada per a les JIAL, publicada en el llibre d'actes. Barcelona 1988. Existeix separata distribuïda pel CIDC, Barcelona 1988. L'autor defineix un model operatiu no estanc, amb distribució variable de funcions entre les administracions públiques sense duplicitat d'estadístiques.

la col·laboració per a l'homogeneïtat metodològica o bé es refereix als aspectes conceptuals —definicions operatives— i passa a ser un tema de comparabilitat, o bé es refereix a classificacions i codis que corresponen a l'aspecte normatiu.

El tema de fons, quant a la col·laboració per a l'homogeneïtat tècnica no està resolt en les relacions inter-sistemes estadístics. Això no s'ha d'interpretar en el sentit que no existeix homogeneïtat en les estadístiques —cosa que passa amb escassa freqüència— sinó que el que no existeix és la col·laboració interinstitucional formal per al disseny homogeni de les estadístiques que ho requerissin. No hi haurà una efectiva col·laboració fins que no s'anivellin les institucions respecte dels recursos tècnics.

En relació amb l'homogeneïtat de les normes tècniques referides a codis territorials, nomenclatures i classificacions del sistema estadístic català i el sistema estadístic espanyol de l'INE i ministeris, no existeixen problemes pel que fa a la teoria. La pròpia Llei d'estadística de Catalunya estableix que les activitats estadístiques s'han d'atendre a les normes tècniques estatals de compliment obligat i a les normes tècniques internacionals vigents. Per tant, teòricament no es requereix la via dels convenis perquè existeix imperatiu legal. Però aquest imperatiu legal afectaria només la part obligatòria imprescindible per aconseguir l'homogeneïtat i la comparabilitat²⁷.

En l'aspecte pràctic, en canvi, existeixen dos ordres de problemes que poden requerir la via del conveni inter-sistemes. D'una part, s'ha de considerar l'aspecte jurídic ja que és la Generalitat, a proposta de l'òrgan estadístic, la que ha d'establir l'obligatorietat de determinades normes tècniques per als agents del sistema estadístic català. Això vol dir, pràcticament, que s'han de repetir els instruments jurídics. D'altra part, hi ha el tema de la llengua catalana com a llengua pròpia de Catalunya, per la qual cosa es requereix la traducció normalitzada de la terminologia que no és un tema menor en el cas de les nomenclatures i les classificacions²⁸.

Per últim, els convenis institucionals entre el sistema estadístic català i l'estatal han de perseguir la comparabilitat dels resultats estadístics. És evident que es tracta d'una conseqüència dels dos objectius anteriors però, a la vegada, té el valor d'un principi. Les estadístiques han de ser comparables.

No obstant això, sense ànim de complicar el tema, hem de recordar que en estadística, dues o més metodologies homogènies no ofereixen resultats iguals en

²⁷Llei d'estadística de Catalunya. Art. 19.2.

²⁸La supervisió lingüística correspon a organismes específics de la pròpia Generalitat. De vegades no existeix una correspondència biunívoca entre els termes i s'ha de buscar el sentit exacte dels ítems per aconseguir una plena correspondència.

ser aplicades en moments o en àmbits diferents. Això és així perquè l'estadística no és una ciència exacta i els seus resultats poden variar. Però la variació ha d'estar dins d'uns límits controlats, si no s'hauria de procedir a un procés de validació recíproc. Això sí que ha de ser objecte de conveni. Seríem al capítol de la difusió de resultats.

De tota manera, l'aspecte més important en la comparabilitat de les estadístiques consisteix en la utilització de conceptes i definicions operatives uniformes. Si no fos així, seria preferible dissenyar dues operacions estadístiques separades ja que no es tractaria de duplicitat de l'estadística perquè serien diferents.

Un element a destacar en aquest apartat es refereix a l'Institut d'Estadística de Catalunya com a interlocutor únic o part signant en els convenis amb l'INE i els ministeris. És cert que el plantejament inicial de la Llei del pla estadístic de Catalunya fa referència genèrica a la Generalitat i les entitats dependents, d'una part, i a altres entitats estatals, de l'altra. Però la Llei és taxativa quan assenyalava l'Institut d'Estadística en el procediment de relació i conveni amb altres sistemes estadístics. Així doncs, els convenis entre sistemes no es fan entre ministeris i conselleries, per exemple, sinó entre l'Institut d'Estadística i cada ministeri. Això no invalida els convenis signats amb els ministeris anteriors a la Llei del pla estadístic²⁹.

Un altre element important referit a les relacions entre el sistema estadístic català i el sistema estadístic espanyol el constitueix la creació per la Llei de la funció estadística pública, del Comitè Interterritorial d'Estadística. És un organisme del sistema estadístic estatal presidit pel president de l'INE i en el qual són presents els representants dels serveis estadístics dels ministeris, els representants dels òrgans estadístics de les comunitats autònomes i altres representants de l'INE. Aquest Comitè té per objecte vetllar "per la coordinació, la cooperació i l'homogeneïtzació en matèria estadística entre l'Estat i les comunitats autònomes".

El Comitè Interterritorial d'Estadística no té homònim en el sistema estadístic català. Entre altres raons, la Llei d'estadística de Catalunya es va aprovar dos anys abans que l'estatal, per la qual cosa no es coneixien encara les orientacions del sistema estadístic estatal respecte a les relacions amb altres sistemes estadístics.

A diferència de la via dels convenis que és individualitzada i bilateral, les relacions per mitjà del Comitè són col·lectives. Es tracta de donar informació i de rebre opinions. El tema bàsic és el Pla estadístic nacional i els programes

²⁹Llei del pla estadístic de Catalunya esmentada. Art. 28.2

anuals. De tota manera, encara és aviat per jutjar la seva eficàcia. Sembla que podria resultar amb potencialitat de coordinació en el futur.

En relació amb organismes internacionals o amb sistemes estadístics d'altres països de l'entorn, la doctrina derivada de la legislació catalana no és diferent respecte al que hem explicat. La via de col·laboració és el conveni que ha de signar l'Institut d'Estadística en representació del sistema estadístic català. L'aspecte més important i viu és el relatiu a les classificacions internacionals. En aquest cas, no és necessari el conveni, perquè la pròpia Llei d'estadística de Catalunya preceptua la seva utilització si aquestes classificacions són de compliment obligat per a l'INE, el que succeeix en el marc de la Unió Europea, com a mínim per a algunes i fins a un nombre determinat de dígits. La tendència des dels anys setanta es troba en la línia de la universalització de les classificacions estadístiques a escala mundial. No és estrany, doncs que la legislació catalana obligui en aquesta línia. Un altre tipus de relacions amb països del nostre entorn té a Catalunya, per raons de proximitat, una vigència experimental respecte a regions del sud de França. Una de les col·laboracions s'ha materialitzat en una publicació conjunta que ha servit també per posar de manifest la necessitat d'homogeneïtzar els mètodes per tal de fer possible la comparació de les estadístiques³⁰.

REFERÈNCIES

Llei 14/1987, de 9 de juliol, d'estadística de Catalunya (DOGC núm. 869, de 27 de juliol de 1987).

Llei 30/1991, de 13 de desembre, del pla estadístic de Catalunya 1992-1995 (DOGC n.º 539 de 10 de gener de 1992).

Ley 12/1989, de 9 de mayo, de la Función Estadística Pública (BOE n.º 112, de 11 de mayo de 1989).

Decret 341/1989, d'11 de desembre, de creació de l'Institut d'Estadística de Catalunya (DOGC n.º 1.241, de 12 de gener de 1990).

Decret 123/1992, d'11 de maig, pel qual s'estableix el Programa anual d'actuació estadística per a 1992 en desplegament de la Llei 30/1991, de 13 de desembre, del Pla estadístic de Catalunya 1992-1995 (DOGC n.º 1.605, d'11 de juny de 1992).

³⁰ Aquesta publicació *Memento Statistique. Memòria estadística. Euroregió. Catalunya, Languedoc-Roussillon, Midi-Pyrénées*, serà distribuïda per l'Institut d'Estadística de Catalunya en un futur immediat.

Decret 320/1992, de 14 de desembre, pel qual s'estableix el Programa anual d'actuació estadística per a 1993 en desplegament de la Llei 30/1991, de 13 de desembre, del Pla estadístic de Catalunya 1992-1995 (DOGC 1.694, de 14 de gener de 1993).

Decret 331/1993, de 28 de desembre, pel qual s'estableix el Programa anual d'actuació estadística per a 1994 en desplegament de la Llei 30/1991, de 13 de desembre, del Pla estadístic de Catalunya 1992-1995 (DOGC 1.857, de 8 de febrer de 1994).

ENGLISH SUMMARY:

OFFICIAL STATISTICS AND COORDINATION OF THE CATALAN STATISTICAL SYSTEM

Josep A. Casco Robledo

The first part of this text is devoted to define the "official statistics" concept. According to the author, the statistical activities of the modern States began on the 18th and 19th centuries. At that time, statistics-especialized governmental bodies were created, which realized population censuses, different from those previously carried out by other administrative agencies for military or fiscal purposes. The new censuses were intended for statistical information, with special rules to grant respect for the citizens rights.

When a specific legislation governs statistical activities, these are then legitimized as "official statistics". As a consequence, all official statistical activities are public, although this does not mean that all public statistical activities are official. Only the official ones are subject to specific principles and rules. More recently, from the second half of the current century, rules oblige to plan statistical activities. This is the case of Catalonia, among other countries.

Accordingly, the "official statistics" concept is applied only to the statistical activities carried out, either directly or indirectly, by the public institutions subjected to the regulations and plans established by law. The mandatory cooperation expected from citizens is justified because of parliament's and government's

statements about the general interest that some statistical activities have, as well as the granted statistical secrecy of the professionals involved in them, in front of any one else, the State included. To these basic principles there are other to be added regarding publicity of the results, technical accuracy, proportionality and others taken into account in the more recent legislation, being one among them the requirement that the statistical activities be described in a specific plan and programme.

The second part deals with the description of the Western countries statistical systems and, particularly, the Catalan statistical one. It has been only in this century that statistics have been organized systematically and referred to as "statistical systems". The most recent statistical regulations already includes this expression.

The Catalan statistical system agents are the various Generalitat's Departments and the public institutions depending on them, together with other bodies of the Local Administration. All them participate in the statistical system through the statistical Plan, its ruling tool.

The basic legislation of the Catalan statistical system is composed of: "Llei d'Estadística de Catalunya" (Catalan Statistical Law) of 1987, and "Llei del Pla Estadístic de 1992-1995" (Statistical Plan Law 1992-1995). Each year, the Catalan Government pass a bill with the statistical activities to be carried out during the year, in order to accomplish the 4-year Plan. There exist statistical activities out of the Plan, but these are not under the Statistical Law and therefore are not official statistics.

An important feature of the Catalan statistical system, is the influence of the Parliament in its operation, and the role assigned to "Institut d'Estadística de Catalunya" (Catalan Statistical Institute). Indeed, the Catalan Parliament declares the public interest of the statistics, knows the yearly statistical operational programmes, receives submitted reports about their implementation and receives the results. The Institute has a two fold mission: as a system's agent, it produces the statistics assigned by the Plan and its programmes; and as a specialized body, it manages the entire statistical system.

Besides the "Institut d'Estadística de Catalunya", the Catalan statistical system has a consultative body: the "Consell Català d'Estadística" (Catalan Statistical Council). This is composed by an ample representation of the country's academic, social and economic bodies, as well as of the local public administrations and the Institute's Steering Board itself.

The third and last part of this article relates to the official statistical activities coordination, carried out by the Catalan Statistical Institute. To this

purpose, the main tool used is the Statistical Plan, both in its preparation and proposals design phases, and in its execution monitoring and control. Another coordinating tool are the Institute's technical standardization tasks, specially nomenclatures, classifications and coding systems. These standards are translations of the EUROSTAT ones into Catalan language.

As a coordinating agent, the Catalan Statistical Institute acts directly on the Generalitat's Departments, through its Steering Board. This Board is composed of an exhaustive representation of all Catalan Government's Departments, and it is chaired by the "Conseller d'Economia i Finances" (Economy and Finance Councillor), the Department on which the Institute is depending. The Institute has also a single-person body, the Director, who cares of the execution of the agreements passed by the Board, and manages the Institute.

The most characteristic activity of the Catalan Statistics Institute is the technical advice provided to the system's agents. It consists on technical support given to the participating institutions in the Statistical Plan, in order to assure the technical quality of the operations, comparability of the results and statistical legality. So, it is an important coordination component of the whole system.

Finally, it is observed that the Catalan system is no closed on itself. Although it is not an INE's nor State government ministeries subsystem, it is linked to them by mean of cooperation agreements.

EL PLA ESTADÍSTIC DE CATALUNYA: LA LLEI 30/1991 I LA PLANIFICACIÓ DE L'ACTIVITAT ESTADÍSTICA PÚBLICA

J. OLIVERAS i PRATS

Director de l'Institut d'Estadística de Catalunya

A partir del marc legal preestablert i de l'aplicació dels actuals mètodes de planificació pública per objectius, la Llei del pla estadístic regula el desenvolupament de l'activitat estadística oficial al nostre país per tal d'afavorir al màxim el coneixement de la seva realitat econòmica, demogràfica i social. Una vegada fixades les relacions de col·laboració institucionals, que han de permetre assolir un sistema estadístic català homogeni i comparable amb els dels països europeus més avançats, s'analitza l'objectiu fonamental d'aquesta Llei, la seva estructura i jerarquització d'objectius i, d'una manera sintètica, els objectius temàtics i sectorials més significatius que incorpora.

The Statistics Plan Act of Catalonia: the Law 30/1991 and the planification of public statistical activity.

Key words: Planificació estratègica per objectius, sistema estadístic, criteris de decisió estrictes i de preferència, activitat estadística instrumental.

Nota de la Redacció: aquest treball forma part d'uns articles del director de l'Institut d'Estadística de Catalunya que difonen les idees fonamentals de la Llei del pla estadístic de Catalunya 1992-1995 en diversos àmbits de la societat. D'entre aquests escrits, s'ha decidit editar una versió revisada de l'article publicat al 1992 a la revista *Nota d'Economia* (maig 1992, n.º. 43), que edita el Dept. d'Economia i Finances de la Generalitat de Catalunya.

1. ANTECEDENTS

La Llei 30/1991, de 13 de desembre, del Pla estadístic de Catalunya 1992-1995, representa la culminació del procés de desenvolupament legislatiu de les competències exclusives de la Generalitat en les estadístiques d'interès per a Catalunya. L'esmentat procés s'obre amb la Llei 14/1987 d'estadística, que defineix els principis tècnics i garanties jurídiques que han de regir en tota l'activitat estadística pública i que institueix l'obligatorietat del Pla estadístic com a instrument del Parlament per a l'ordenació i planificació d'aquestes activitats estadístiques a Catalunya .

Ens trobem, doncs, amb una Llei que aprova un pla, que preveu quines estadístiques farà la Generalitat durant el proper quadrienni, que preveu la col·laboració interinstitucional en el camp estadístic, que estableix objectius en matèria estadística adaptables a les possibles noves necessitats de dades i a les disponibilitats econòmiques i organitzatives, que garanteix el compliment de les normes legals sobre drets i deures dels ciutadans i que assegura la qualitat tècnica i la fiabilitat de la producció estadística.

2. EL PROCÉS D'ELABORACIÓ DE LA LLEI DEL PLA

Per tal de copsar la naturalesa del procés d'elaboració del Projecte de llei, cal referir-se breument al marc i els agents institucionals implicats. En aquest sentit, la Llei dedica el primer capítol a la definició del sistema estadístic i a l'assignació de funcions i de papers dels seus diferents agents. De fet, distingeix tres sistemes: el sistema estadístic de la Generalitat, el sistema estadístic integral de Catalunya i la resta de sistemes estadístics.

Quant al sistema estadístic de la Generalitat, el Pla estadístic és l'instrument d'ordenació i de planificació de l'estadística de la Generalitat de Catalunya i de tots els organismes i entitats que en depenen. El sistema estadístic integral de Catalunya està format per l'Administració de la Generalitat de Catalunya i l'Administració de les entitats territorials de Catalunya. La resta dels sistemes estadístics es defineixen per exclusió i també en sentit genèric referit a entitats de dins i de fora de l'Estat espanyol. Aquesta Llei, doncs, defineix el sistema estadístic de Catalunya no com un sistema tancat, sinó com un sistema que s'integra a sistemes estadístics d'abast estatal i europeu, que busca la màxima homogeneïtat i comparabilitat de les dades i que procura economies per mitjà de la complementarietat de les actuacions.

En conseqüència, el procés d'elaboració del Projecte de llei del pla estadístic ha estat alimentat per un llarg procés de consultes, d'assessorament, i d'assistència tècnica, que ha dut a terme el mateix Institut d'Estadística de Catalunya. En aquest procés s'ha interaccionat amb diversos interlocutors dels diferents departaments de la Generalitat i de les entitats locals de Catalunya per tal de detectar i analitzar críticament, de forma conjunta, les activitats estadístiques en curs, com també les necessitats plantejades en aquesta matèria. Al mateix temps, s'han mantingut relacions amb l'INE, l'organisme estatal responsable d'estadística, per cercar la màxima complementarietat dels treballs estadístics de l'Estat i els de les institucions catalanes, i amb la mateixa finalitat s'han estudiat les normes europees sobre la matèria. El resultat és un pla que, a més de satisfer les previsions esmentades a la Llei d'estadística, asseguren la coherència de les activitats en els contextos, europeu, estatal, català i de l'Administració local.

3. L'OBJECTIU CENTRAL DEL PLA

La Llei defineix l'objectiu central del Pla estadístic de Catalunya 1992-1995 com el "d'aconseguir un conjunt coherent, fiable i actualitzat de dades estadístiques al mínim cost possible, aprofitant al màxim les fonts existents, que permeti el coneixement de la realitat econòmica, demogràfica i social de Catalunya i que sigui útil a la presa de decisions de les institucions públiques i els agents socials, minimitzant les molèsties als ciutadans i garantint el secret estadístic".

Aquest objectiu central és el criteri bàsic i essencial per a la presa de decisions de tots els òrgans implicats en l'execució del Pla estadístic i ha de servir de criteri interpretatiu principal per a l'aplicació d'aquesta Llei i de les disposicions per al seu compliment i desenvolupament.

El Pla estadístic pretén, doncs, aconseguir les estadístiques que facilitin el coneixement de la realitat catalana que servirà per a la presa de decisions en el sector públic i privat. Aquesta utilitat equival a la declaració d'interès públic de les activitats estadístiques incorporades en els successius programes anuals d'actuació estadística que desenvolupen la Llei del pla.

Aquestes estadístiques, però, han de tenir coherència entre elles, amb les prioritats que determinin les necessitats i amb l'existència d'altres estadístiques ja en curs, tot intentant l'equilibri entre l'àmbit temàtic i territorial.

La fiabilitat és una altra característica molt present a la Llei d'estadística, que es deriva de l'exigència de correcció tècnica. Això comporta la necessitat de dissenyar acuradament els projectes tècnics i les metodologies i tècniques aplicades a cada operació estadística. L'actualització i la periodificació de les estadístiques responen a una necessitat cada dia més imperiosa a causa de l'enorme mobilitat i del canvi que caracteritzen tots els nivells de la vida social.

Així doncs, la coherència, la fiabilitat i l'actualització són característiques del conjunt d'estadístiques a aconseguir i responen als principis tècnics exigits per la Llei d'estadística. Però també tenen la seva traducció pràctica en les normes i els criteris exigits en el procés de selecció de les estadístiques que seran incorporades en els successius programes anuals d'actuació estadística.

L'objectiu central també fa una al·lusió clara a la forma de treballar en la realització de les activitats estadístiques. Per una part, s'han de seleccionar mètodes alternatius amb un cost més baix, i per això s'introdueixen objectius d'activitats instrumentals amb aquesta finalitat. D'alguna manera, aquesta exigència va lligada a l'aprofitament de fonts existents com a alternativa als feixucs treballs de camp però, sobretot, marca un model d'actuació eficaç que evita les duplicitats amb altres operacions estadístiques i afavoreix la utilització d'informació d'origen administratiu. Tot això té una expressió clara en els criteris de prioritització en la selecció d'objectius operatius i d'activitats que la pròpia Llei imposa.

Finalment, l'objectiu central fa referència a les garanties jurídiques per als ciutadans que hagin de col·laborar en el subministrament de la informació. D'una banda, minimitzant les molèsties als ciutadans d'acord amb el respecte del principi de proporcionalitat, no exigint més col·laboració que l'estrictament necessària, tant en nombre d'intervencions com en extensió i complexitat dels qüestionaris. Això enllaça amb l'aprofitament d'altres fonts d'informació. D'altra banda, fent ús d'aquelles tècniques que incrementen la seguretat i la garantia de la salvaguarda del secret estadístic. També són nombrosos els objectius operatius sobre el secret estadístic que s'hauran de desenvolupar en el present Pla estadístic i que consten en el text legal.

4. ESTRUCTURA DE LA LLEI DEL PLA

L'objectiu central de la llei es desplega en objectius més precisos en diferents camps. En el camp de les activitats estadístiques en curs de realització, es defineix cadascuna de les activitats censades en forma d'objectiu i s'estableixen uns criteris de decisió estrictes que hauran de complir per tal d'assegurar que s'ajusten als requisits del Pla.

D'aquesta manera, la seva incorporació progressiva als programes anuals d'actuació estadística permetrà la validació de la seva correcció tècnica dels mètodes emprats i de la seva adequació a l'objectiu central.

En el camp de les activitats estadístiques noves que han de dur-se a terme a partir de la vigència del Pla, s'identifica cadascuna d'elles mitjançant un objectiu que estableix les prestacions que se li demanen, com també estimacions de la complexitat i del cost. Igualment, es defineixen els tipus de criteris de decisió que han de condicionar les prioritats del Govern per a la posada en marxa d'aquestes activitats.

D'altra banda, aquesta Llei defineix les activitats estadístiques instrumentals com aquelles que són prèvies o complementàries de les activitats estadístiques de producció i legalment exigibles, o tècnicament necessàries, per poder complir els requisits que estableix la legislació sobre estadística, com les de formació, i les de recerca i desenvolupament tècnic, metodològic i normatiu en el camp estadístic. Aquesta definició permet sistematitzar i planificar les activitats de recerca, de desenvolupament tècnic i metodològic i de formació en matèria estadística. Aquests tipus d'activitats s'atribueixen a l'Institut d'Estadística de Catalunya que d'aquesta manera, i en virtut d'aquesta Llei, es configura no solament com un organisme especialitzat en la producció estadística, sinó també com un agent prestador de serveis de suport tècnic i metodològic al conjunt descentralitzat de productors d'estadístiques inserits en el nostre sistema.

Finalment, la Llei del pla estadístic estableix les normes de procediment relatives a la realització del Programa anual d'actuació estadística. L'Institut d'Estadística de Catalunya té l'obligació de presentar el projecte de Programa anual d'actuació estadística. Aquest Programa serà aprovat per Decret del Govern de la Generalitat. És important la norma sobre l'aprovació de resultats estadístics i la seva publicitat. El capítol X de la Llei està dedicat a aquests temes.

5. LES ESTADÍSTIQUES ECONÒMIQUES EN LA LLEI DEL PLA

La classificació temàtica és útil per avaluar el grau d'abast dels objectius en estadística econòmica recollits a la Llei del pla. El sistema econòmic es representa sintèticament per unes activitats fetes per uns agents o sectors institucionals de l'economia. Aquesta representació forma part de l'acord internacional sobre l'estructura de la comptabilitat integrada d'una economia nacional o regional.

Naturalment, no tots els sectors institucionals presenten un nivell d'informació equivalent. El sector de les empreses no financeres, és a dir, les empreses de producció agrària, industrial, de la construcció i dels serveis (tant els privats com els que ofereix l'Administració), és el més estudiat per l'estadística econòmica. L'estadística d'aquests sectors és, en general, d'establiments, d'ocupació i de comptes d'explotació dels establiments (facturació, costos de personal, beneficis, etc...) i també d'inversió. També són importants les estadístiques d'unitats físiques, específicament lligades a activitats particulars (de produccions agràries, de llicències de construcció, de moviment de turistes, etc.)

A més del sector institucional de les empreses no financeres, tenim el sector financer, el sector públic, el de les famílies i el sector exterior. L'aproximació institucional queda completada amb l'atenció al mercat de treball i a l'estadística de preus.

Complementàriament a aquestes estadístiques fonamentals, existeixen òptiques complementàries de gran importància: la síntesi estructural global, la que posa l'èmfasi en el seguiment conjuntural de l'economia i, finalment, l'estadística de base i territorial, formada per activitats de caràcter censal, per la confecció de directoris i per estadístiques de síntesi per a delimitacions territorials que, en el nostre cas, són de tipus comarcal.

La Llei del pla abasta tots aquests àmbits. En el Quadre 1, es visualitzen els objectius de la Llei segons les àrees temàtiques anunciades. És interessant destacar que els objectius en referència amb l'annex II de la Llei corresponen a estadístiques en curs, mentre que els designats en l'annex III corresponen a noves activitats que s'hauran de dur a terme en els successius programes anuals d'actuació estadística.

5.1. Estadística dels sectors productius

Aquest és l'àmbit més important de l'estadística econòmica. El seu objectiu és proporcionar una idea de l'estructura i l'evolució anual dels diferents sectors productius. L'aproximació a les activitats productives es pot fer de diferents formes. Es pot diferenciar entre l'estadística dels establiments, dels ocupats, dels resultats del compte d'explotació dels establiments amb conceptes de tipus macroeconòmic, el que es coneix com a comptabilitat nacional, i una estadística de tipus microeconòmic més lligada amb els conceptes de comptabilitat empresarial. També resulta molt important una estadística de demanda lligada als sectors productius, com és la de la inversió o la de la compra de béns de producció. Finalment, cal destacar la importància de l'estadística d'unitats físiques diferents en funció del sector (agrícola, pesquer, energètic, de la construcció, del turisme).

La Llei del pla abasta d'una forma important aquesta estadística. En l'àrea d'estadística i comptes, actualment només es disposa del que correspon al sector agrari, a les indústries agroalimentàries i al comerç interior. Els objectius de la Llei abasten la totalitat dels sectors industrials, de la construcció i, d'una forma probablement plurianual, els diferents sectors dels serveis, tenint una especial atenció en els sectors d'oferta turística i de serveis a les empreses i a les famílies. Pel que fa a l'estadística de caràcter microeconòmic, ja es disposa de la xarxa comptable agrària i de les ràtios econòmico-financeres del conjunt de l'economia.

Pel que fa a l'estadística física, el balanç és millor especialment per als sectors agrari, energètic i de la construcció. En el cas del sector turístic es pot millorar. Una de les carències en aquest tipus d'estadística es refereix a l'estadística física de la indústria, que per la seva gran diversitat és molt difícil d'abordar. De tota manera, amb la renovació de l'enquesta industrial es pot esperar una millora d'aquestes estadístiques.

5.2. Estadística econòmica d'altres àmbits

Els sectors financer, de les famílies, de l'Administració pública i del sector exterior estan considerats en la Llei. En aquest cas la majoria dels objectius corresponen a noves estadístiques. Només l'estadística dels crèdits al sector de la construcció i la de pressupostos i liquidacions estan en curs. Cal destacar que hi ha dificultats tècniques notables, en particular les derivades de la problemàtica de regionalització de l'activitat financera, de l'Administració central i també del sector exterior, especialment en l'àmbit dels fluxos interregionals.

L'estadística del mercat de treball a Catalunya es millora de forma significativa, tant a partir d'una tabulació pròpia de l'enquesta de població activa, que és una font d'informació potent i parcialment infrautilitzada des d'un punt de vista regional, com de l'enquesta de salaris. Pel que fa als preus, ja es disposa de preus del sector agrari i la millora establerta per la Llei implica tenir també preus per als productes industrials. Aquesta millora és molt important, ja que molts indicadors de competitivitat i també l'avaluació rigorosa del creixement real de la indústria catalana requereixen conèixer l'efecte monetari que afecta l'evolució del valor afegit expressat a preus corrents.

La culminació de l'estadística estructural d'una economia queda reflectida en la síntesi comptable. Aquesta síntesi recull tant l'aproximació a les branques productives, com l'aproximació institucional. A més, de forma quinquennal, queda recollida l'anàlisi de les relacions intersectorials que es coneixen com a taules input-output.

El seguiment de la conjuntura té tres operacions de caràcter específic: l'avanç del producte interior brut (PIB), l'índex de producció industrial (IPI) i l'enquesta d'expectatives empresarials. També s'identifica un objectiu de síntesi que faci possible seguir, de forma global, el moment de l'economia catalana.

L'estadística econòmica de base és un element fonamental per a qualsevol sistema estadístic. A la Llei del pla queden recollits, a més dels dos censos econòmics (el de locals i l'agrari), els directoris d'empreses i d'establiments de caràcter genèric, i els directoris específicament lligats a registres administratius particulars. Finalment, es contempla una aproximació sintètica de caràcter territorial que pretén donar una visió articulada de l'evolució de les economies comarcals.

6. LES ESTADÍSTIQUES DEMOGRÀFIQUES I SOCIALS EN LA LLEI DEL PLA

6.1. Estadístiques demogràfiques

La línia bàsica del Pla, pel que fa a les estadístiques demogràfiques, és definir un sistema integrat de les diferents àrees d'informació demogràfica, completant el marc estadístic preexistent, amb unes noves activitats que tanquin i donin coherència al conjunt del sistema d'informació demogràfica. Així doncs, les estadístiques en curs integrades per les estadístiques d'estructura de la població, derivades dels censos i dels padrons municipals, es completen amb les estadístiques de fluxos demogràfics a partir del moviment natural de la població (naixements, defuncions i matrimonis) i de les migracions. La Llei dona prioritat a les noves estadístiques de contingut econòmic i de previsions demogràfiques. Així doncs, s'especifica en aquest sentit i com a objectius operatius (vegeu el Quadre 2): avaluar l'evolució de la distribució de la població comarcal per períodes intercensals, estimar la distribució futura de la població de les comarques de Catalunya per sexe i edat a fi de planificar equipaments i serveis i, finalment, conèixer les dades d'esperança de vida i probabilitat de supervivència de la població catalana per edat i sexe per tal d'obtenir les taules de vida de Catalunya.

6.2. Estadístiques socials

En aquesta àrea, la Llei recopila i integra activitats estadístiques dutes a terme pels diversos departaments de la Generalitat, atès el model descentralitzat del sistema estadístic de Catalunya.

Així, d'una banda, s'assumeixen estadístiques altament consolidades i, d'una altra, es potencien estadístiques en una fase més incipient però que la Llei, en recopilar-les, els dona ple suport. Una visió exhaustiva d'ambdós vessants es recull en el Quadre 3.

Pel que fa a fixar objectius per a les noves activitats estadístiques, la Llei especifica que, per iniciar aquestes activitats, s'han de satisfer, entre altres, els criteris de decisió de diversificar les estadístiques socials disponibles, donant prioritat a noves estadístiques socials sobre condicions de vida de la població.

En aquest sentit, la Llei fixa la necessitat de disposar d'una informació estructurada sobre els hàbits i les condicions de vida de la població catalana, fent el disseny i l'execució d'una enquesta polivalent a les famílies. Aquest objectiu es complementa amb tres objectius addicionals que s'assoleixen per explotacions específiques de l'esmentada enquesta, i que són els de conèixer els comportaments culturals, sanitaris, esportius i associatius de la població de Catalunya, conèixer les principals característiques i el grau d'utilització i satisfacció dels equipaments i els serveis i les condicions urbanístiques de l'hàbitat immediat de la població i, finalment, conèixer els principals aspectes de la mobilitat social i de la marginació a Catalunya per raons econòmiques, culturals i socials.

Aquesta àrea prioritària es complementa amb els objectius de conèixer l'estructura social de la població i disposar d'informació sobre les condicions de vida de la població fent un sistema d'indicadors socials. És a dir, s'han considerat els objectius més rellevants des d'una òptica general de les estadístiques socials. A més, hi ha objectius sectorialment molt destacables en les àrees de salut pública, d'ensenyament, de benestar social, de cultura, de medi ambient, i d'infraestructures, transports i comunicacions. D'altra banda, cal remarcar les estadístiques de caràcter sectorial i/o multidisciplinar que es basen, fonamentalment, en l'aprofitament d'informació d'origen administratiu, d'acord amb els preceptes que l'objectiu central de la Llei del pla estableixen en relació amb l'optimització dels recursos en l'activitat estadística pública.

7. LES ACTIVITATS ESTADÍSTIQUES INSTRUMENTALS EN LA LLEI DEL PLA

L'activitat estadística és molt àmplia i comprèn múltiples actuacions centrades, en primera instància, en la producció i difusió d'informació estadística. Igualment, però, incorpora tasques envers la regulació jurídica i la coordinació tècnica d'aquestes actuacions, i també els desenvolupaments metodològics i ins-

trumentals. Finalment, inclou activitats orientades a la professionalització progressiva i continuada del personal estadístic.

Per tant, l'activitat estadística consumeix un volum elevat de recursos; alhora, és una activitat tècnicament i organitzativa complexa que requereix ineludiblement instruments de regulació i planificació adequats. Al mateix temps, exigeix una àmplia reflexió sobre les necessitats d'informació per a la presa de decisions.

En aquest sentit, la gestió eficient dels recursos destinats a les activitats de producció i difusió estadística es preveu explícitament a la Llei del pla mitjançant la incorporació d'activitats de caràcter instrumental, enteses com les actuacions orientades a augmentar els nivells de qualitat en el tractament de dades i, al mateix temps, procurar economies d'escala en despesa i temps, a partir de la millora de la seva coordinació i correcció tècnica.

En el nostre model de producció estadística descentralitzada, l'Institut d'Estadística de Catalunya ha de gestionar el sistema estadístic català i assumir, entre altres, la responsabilitat de proporcionar els elements metodològics i tècnics que el regulen i l'enforteixen. Així, les activitats estadístiques instrumentals, juntament amb les tasques de reciclatge i perfeccionament professional, s'orienten en tres direccions diferenciades (tal com se sintetitza en el Quadre 4).

- a) Primerament, garantir l'homogeneïtat i correcció dels resultats estadístics mitjançant l'establiment de normatives i directrius per a la coordinació, correcció tècnica, comparabilitat i sistematització del treball estadístic. En aquest sentit, els objectius establerts pretenen cobrir les necessitats de definició i articulació d'un sistema harmonitzat i integral de nomenclatures de les unitats estadístiques bàsiques i, simultàniament, facilitar l'acompliment dels preceptes legals (preservació del secret estadístic, entre altres) i tècnics inherents a la tasca de producció, difusió i accés a les dades estadístiques.
- b) En segon lloc, augmentar els nivells d'eficiència en la generació, tractament i ús d'informació estadística mitjançant desenvolupaments tècnics que siguin estalviadors de recursos i temps. El caràcter altament automatitzable dels processos regulars en l'obtenció i tractament de dades originals permet plantejar-se legitimament avanços significatius en diversos àmbits dels sistemes d'informació; en aquest sentit, poden destacar-se activitats tendents a disposar de sistemes automàtics per a la codificació i imputació de dades provinents d'enquesta, procediments àgils pel seu processament i tractament mecanitzat, i la implementació de sistemes d'emmagatzemament i de difusió d'informació estadística, convenientment estandarditzada i estructurada de cara al seu màxim aprofitament.

- c) Finalment, elevar la qualitat de les estadístiques a partir de l'aplicació dels perfeccionaments metodològics més adients, destinats bàsicament a la producció d'informació estadística. En estadística regional, aquesta orientació es concreta en objectius com l'establiment de metodologies d'estimació i valoració de la fiabilitat estadística (sobretot, en àrees geogràfiques reduïdes), la quantificació de fluxos interregionals o dels seus efectes encreuats (simulant-los amb modelitzacions econòmiques interregionals), a partir de la promoció i difusió de la recerca estadística aplicada en aquest camp.

Els mitjans previstos per a la difusió i aprenentatge dels avenços que es produeixin en els desenvolupaments esmentats són l'assistència tècnica als agents del sistema estadístic, i el perfeccionament professional o reciclatge del personal estadístic especialitzat. Així, en la mesura que la realització de totes aquestes activitats s'atribueix a l'Institut d'Estadística de Catalunya, aquest organisme es configura no solament com un òrgan especialitzat en la producció estadística, sinó com un agent prestador de serveis de suport tècnic i metodològic al conjunt descentralitzat d'agents productors d'estadístiques inserits en el sistema institucional.

L'execució del Pla estadístic de Catalunya, mitjançant els seus Programes anuals d'actuació estadística, constitueix el mecanisme a partir del qual es preveu, doncs, encarar la complexitat organitzativa i tècnica de l'activitat estadística. Els avenços que es produeixin en aquest camp han de permetre afinar, a la vegada, els mitjans de planificació i programació, sobretot tenint en compte que aquest plantejament té el ferm propòsit de reforçar l'orientació de l'estadística catalana cap als estàndards metodològics emprats en l'activitat estadística dels països més avançats.

Quadre 1. L'estadística econòmica en la Llei del pla estadístic de Catalunya 1992-1995		
Àrees temàtiques	Objectius	Articulat a la Llei
Estadística dels sectors productius: establiments, ocupació i comptes econòmics	Estadística i Comptes	
	Comptes econòmics del sector agrari	Annex II objectiu 8.7
	Comptes econòmics del sector industrial	Annex III objectiu 1.6
	Comptes econòmics de les indústries agroalimentàries	Annex II objectiu 8.9
	Comptes econòmics del sector de la construcció	Annex III objectiu 1.13
	Comptes econòmics del sector comerç	Annex II objectiu 3.5
	Comptes econòmics del sector turístic	Annex III objectiu 1.15
	Serveis a empreses i famílies	Annex III objectiu 1.18
	Estadística microeconòmica	
	Xarxa comptable agrària	Annex II objectiu 8.8
	Ràtios econòmiques i financeres de les empreses	Annex II objectiu 3.1
	Quaderns d'estadística sectorial	Annex III objectiu 1.20
Estadística dels sectors productius: inversió	Inversió industrial	Annex III objectiu 1.7
	Indicadors de la indústria amb inversió registrada	Annex III objectiu 1.8
	inversió estrangera autoritzada	Annex III objectiu 1.9
Estadística dels sectors productius: unitats físiques	Sector agrari	
	Estructures agràries	Annex III objectiu 1.5
	Superfície i producció agrícoles	Annex II objectiu 8.1
	Ramaderia	Annex II objectiu 8.2
	Sector forestal	Annex II objectiu 8.3
	Pesca	Annex II objectiu 8.4
	Sector energia	
	Energia elèctrica	Annex II objectiu 11.1
	Consums energètics comarcals	Annex II objectiu 11.2
	Balanç energètic	Annex II objectiu 11.3
	Fluxos de combustibles sòlids	Annex III objectiu 1.29
	Sector de la construcció	
	Habitatges nous i rehabilitats	Annex II objectiu 7.1
	Préstecs protegits per a la compra i la rehabilitació d'habitatges	Annex II objectiu 7.2
	Llicències d'edificació	Annex II objectiu 7.3
	Oferta i demanda d'habitatges	Annex II objectiu 7.4
	Sector turístic	
	Oferta extrahotelera dels municipis turístics	Annex III objectiu 1.16
	Fluxos turístics per carretera i aeroports	Annex III objectiu 1.17
	Turisme entrat per la Jonquera	Annex II objectiu 12.3
Estadística del sector financer	Sistema financer	Annex III objectiu 1.19
	Crèdits sectors de la construcció	Annex II objectiu 1.14
Estadística de les famílies	Pressupostos i despesa de les famílies	Annex III objectiu 1.22
	Hàbits de consum	Annex III objectiu 1.23
Estadística del sector exterior	Comerç exterior	Annex III objectiu 1.26
	Transport per carretera	Annex III objectiu 1.27
	Transporta ferroviari, aeri i marítim	Annex III objectiu 1.28
Estadística de l'Administració pública	Pressupostos i liquidacions de les administracions	Annex III objectiu 2.4
	Comptes econòmics de les administracions públiques	Annex III objectiu 1.24
Estadística del mercat de treball	Mercat laboral	Annex III objectiu 1.24
	Evolució dels salaris	Annex III objectiu 1.25
Estadística de preus	Preus agraris en llotges i mercats	Annex II objectiu 8.6
	Preus rebuts i pagats pels pagesos	Annex II objectiu 8.5
	Índex de preus industrials	Annex III objectiu 1.11
Síntesi comptable	Comptes econòmics de Catalunya	Annex III objectiu 1.1
	Taules Input-Output	Annex III objectiu 1.2
Estadística de la conjuntura	Evolució conjuntural	Annex III objectiu 1.4
	Avanç de la taxa de variació del PIB	Annex II objectiu 3.2
	Índex de producció industrial	Annex III objectiu 1.10
	Expectatives empresarials	Annex III objectiu 1.12

Quadre 1. L'estadística econòmica en la Llei del pla estadístic de Catalunya 1992-1995 (continuació)		
Àrees temàtiques	Objectius	Articulat a la Llei
Estadística de base i territorial	Cens agrari	Annex II objectiu 3.3
	Cens de locals	Annex III objectiu 1.30
	Directorí d'empreses i d'establiments	Annex III objectiu 1.31
	Directorí de les explotacions agràries	Annex II objectiu 3.4
	Directorí del sector de la construcció	Annex III objectiu 1.33
	Directorí de les entitats financeres i asseguradores	Annex III objectiu 1.34
	Directorí d'establiments turístics	Annex II objectiu 12.2
	Directorí de cafeteries i restaurants amb llicència de joc	Annex II objectiu 2.3
	Directorí de les empreses de transport	Annex II objectiu 7.5
	Directorí de les societats cooperatives	Annex II objectiu 9.7
	Directorí de centres de treball	Annex II objectiu 9.8
	Actualització del registre industrial	Annex II objectiu 11.5
	Base de dades d'establiments turístics	Annex II objectiu 12.1
	Economies comarcals	Annex III objectiu 1.3

Quadre 2. L'estadística demogràfica en la Llei del pla estadístic de Catalunya 1992-1995		
Àrees temàtiques	Objectius	Articulat a la Llei
Estadístiques de la síntesi demogràfica	Projeccions de població	Annex III objectiu 2.2
	Taules de vida	Annex III objectiu 2.3
Estadístiques d'estructura i fluxos demogràfics	Cens de Població	Annex II objectiu 3.6
	Mobilitat espacial obligada	Annex II objectiu 3.9
	Naixements i matrimonis	Annex II objectiu 3.10
	Moviments migratoris	Annex II objectiu 3.11
	Evolució intercensal de la població comarcal	Annex III objectiu 2.1

Quadre 3. L'estadística social en la Llei del pla estadístic de Catalunya 1992-1995		
Àrees temàtiques	Objectius	Articulat a la Llei
Estadístiques de condicions de vida de les famílies	Condicions de vida de la població	Annex III objectiu 3.1
	Comportament de la població	Annex III objectiu 3.2
	Hàbitat de la població	Annex III objectiu 3.3
	Mobilitat social i marginació	Annex III objectiu 3.4
	Estructura social	Annex III objectiu 3.7
	Indicadors socials	Annex III objectiu 3.8
Estadístiques d'edificis i habitatge	Cens d'habitatges	Annex II objectiu 3.7
	Habitatges nous i rehabilitats	Annex II objectiu 7.1
	Prèstecs protegits per a la compra i rehabilitació d'habitatges	Annex II objectiu 7.2
	Llicències d'edificació	Annex II objectiu 7.3
	Oferta i demanda d'habitatges	Annex II objectiu 7.4
	Cens d'edificis	Annex III objectiu 3.32
Estadístiques d'ensenyament i recerca	Directori de centres d'ensenyament	Annex II objectiu 4.1
	Ensenyament infantil i primari	Annex II objectiu 4.2
	Ensenyament secundari	Annex II objectiu 4.3
	Ensenyament musical, artístic i altres ensenyaments especialitzats reglats	Annex II objectiu 4.4
	Educació especial	Annex II objectiu 4.5
	Despesa de l'ensenyament públic	Annex III objectiu 3.9
	Despeses i ingressos de l'ensenyament privat	Annex III objectiu 3.10
	Ensenyament universitari	Annex III objectiu 3.11
	Recerca i desenvolupament	Annex III objectiu 3.12
Estadístiques de cultura i esports	Cens d'equipaments esportius	Annex II objectiu 1.1
	Directori de clubs i associacions esportives	Annex II objectiu 1.2
	Pràctica esportiva	Annex II objectiu 1.3
	Cens lingüístic	Annex II objectiu 3.8
	Directori d'empreses editorials, periodístiques, cinematogràfiques i discogràfiques	Annex II objectiu 5.1
	Directori de cinemes i teatres	Annex II objectiu 5.2
	Biblioteques	Annex III objectiu 3.17
	Bases de dades culturals	Annex III objectiu 3.18
	Cens de clubs i associacions esportives	Annex III objectiu 3.19
Estadístiques de salut	Directori d'establiments sanitaris	Annex II objectiu 6.1
	Centres hospitalaris	Annex II objectiu 6.2
	Causas de mort	Annex II objectiu 6.3
	Morbiditat hospitalària	Annex II objectiu 6.4
	Malalties de declaració obligatòria	Annex II objectiu 6.5
	Control sanitari d'aigües	Annex II objectiu 6.6
	Programa de vacunació	Annex II objectiu 6.7
	Directori de les entitats d'assegurança d'assistència mèdico-farmacèutica	Annex II objectiu 6.8
	Drogodependències	Annex II objectiu 6.9
	Hàbits de consum de tabac	Annex II objectiu 6.10
	Despesa sanitària	Annex II objectiu 6.11
	Estadístiques de salut	Annex III objectiu 3.5
	Taules de morbiditat	Annex III objectiu 3.6
Estadístiques de benestar social	Directori d'entitats, serveis i establiments socials	Annex II objectiu 13.1
	Educació permanent d'adults	Annex II objectiu 13.2
	Despeses dels serveis socials	Annex III objectiu 3.13
	Disminucions, discapacitats i minusvalies	Annex III objectiu 3.14
	Beneficiaris dels serveis socials	Annex III objectiu 3.15
	Serveis socials	Annex III objectiu 3.16
Estadístiques d'infraestructura, transports i comunicacions	Emissores de radiodifusió i televisió local	Annex II objectiu 1.4
	Ocupació de l'espectre radioelèctric	Annex II objectiu 1.5
	Atenció al ciutadà en matèria de recepció de senyals	Annex II objectiu 1.6
	Actuacions Pla únic d'obres i serveis	Annex II objectiu 2.6
	Directori de les empreses de transport	Annex II objectiu 7.5
	Xarxa ferroviària de la Generalitat de Catalunya	Annex II objectiu 7.6
	Trànsit marítim	Annex II objectiu 7.7
	Actuacions en obres públiques	Annex II objectiu 7.8
	Vehicles registrats	Annex II objectiu 7.9
	Xarxa d'infraestructures viàries i de comunicacions	Annex III objectiu 3.29
	Xarxes de transport ferroviari	Annex III objectiu 3.30

Quadre 3. L'estadística social en la Llei del pla estadístic de Catalunya 1992-1995 (continuació)		
Àrees temàtiques	Objectius	Articulat a la Llei
Estadístiques del medi ambient	Hidrografia	Annex II objectiu 7.10
	Residus sòlids urbans	Annex III objectiu 3.20
	Aigües residuals urbanes	Annex III objectiu 3.22
	Aigües residuals industrials	Annex III objectiu 3.23
	Autoritzacions d'abocadors	Annex III objectiu 3.24
	Residus industrials	Annex III objectiu 3.25
	Indicadors de medi ambient	Annex III objectiu 3.26
	Entitats subministradores d'aigua	Annex III objectiu 3.27
	Climatologia	Annex III objectiu 3.28
Estadístiques de condicions laborals	Vagues i tancaments patronals	Annex II objectiu 9.1
	Regulació d'ocupació	Annex II objectiu 9.2
	Conciliacions individuals	Annex II objectiu 9.3
	Sancions, conciliacions col·lectives i altres	Annex II objectiu 9.4
	actuacions públiques en matèria laboral	
	Sinistralitat laboral	Annex II objectiu 9.5
	Eleccions sindicals	Annex II objectiu 9.6
	Directori de les societats cooperatives, societats anònimes laborals, associacions professionals i mutualitats laborals	Annex II objectiu 9.7
	Convenis col·lectius	Annex III objectiu 3.33
	Formació ocupacional	Annex III objectiu 3.36
	Plans d'ocupació	Annex III objectiu 3.37
L'estadística administrativa en la Llei del pla estadístic de Catalunya 1992-1995		
Àrees temàtiques	Objectius	Articulat a la Llei
Estadístiques de justícia	Institucions penitenciàries	Annex II objectiu 10.1
	Reinserció juvenil	Annex II objectiu 10.2
	Directori de fundacions i d'associacions	Annex II objectiu 10.3
	Directori d'edificis i locals judicials	Annex II objectiu 10.4
Estadístiques de seguretat	Protecció i seguretat	Annex II objectiu 2.1
	Accidents de trànsit	Annex II objectiu 2.2
	Accidents industrials	Annex II objectiu 11.6
	Estat mecànic del parc de vehicles	Annex II objectiu 11.7
	Indicadors de la seguretat de les instal·lacions industrials	Annex III objectiu 3.21
	Indicadors de l'estat mecànic del parc de vehicles	Annex III objectiu 3.31
Estadístiques de control administratiu del joc	Directori de cafeteries i restaurants amb llicència de joc	Annex III objectiu 2.3
	Joc	Annex III objectiu 2.7
Estadístiques d'eleccions	Electoral	Annex III objectiu 2.5
Estadístiques multidisciplinàries	Base de dades d'estadístiques municipals	Annex II objectiu 3.12
	Anuari estadístic	Annex II objectiu 3.13
	Base de dades ESPAN	Annex II objectiu 3.14
	Biblioteca de l'Institut d'Estadística de Catalunya	Annex II objectiu 3.15
	Anuari estadístic comarcal	Annex III objectiu 3.35

Quadre 4. Les activitats estadístiques instrumentals en la Llei del pla estadístic de Catalunya 1992-95		
Àrees temàtiques	Objectius	Articulat a la Llei
Homogeneïtzació estadística	Codis oficials d'entitats territorials i administratives	Annex IV objectiu 1.1
	Sistemes de nomenclatures i classificacions	Annex IV objectiu 1.2
	Documentació tècnica d'operacions estadístiques	Annex IV objectiu 1.3
	Aprofitament estadístic d'informació d'origen administratiu	Annex IV objectiu 1.4
Correcció tècnica	Mètodes per a la preservació del secret estadístic	Annex IV objectiu 2.1
	Elaboració de mostres estadístiques	Annex IV objectiu 2.2
	Disseny de qüestionaris	Annex IV objectiu 2.3
	Correcció tècnica d'estadístiques incloses en el Pla estadístic	Annex IV objectiu 2.4
	Seguiment i suport tècnic en l'execució del Pla i Programes anuals	Annex IV objectiu 2.5
	Anàlisi de necessitats estadístiques	Annex IV objectiu 2.6
Formació en procediments estadístics	Reciclatge i perfeccionament del personal estadístic	Annex IV objectiu 3.1
	Formació del personal de l'administració i d'usuaris d'estadística	Annex V objectiu 3.1
Desenvolupament tècnic	Sistema de codificació automàtica	Annex V objectiu 1.1
	Sistema automàtic d'imputació de dades	Annex V objectiu 1.2
	Seguiment d'operacions de camp i processament de dades	Annex V objectiu 1.3
	Procediments d'acarament automàtic de fitxers	Annex V objectiu 1.4
	Diccionari comú de dades	Annex V objectiu 1.5
	Experimentació i seguiment de sistemes d'informació	Annex V objectiu 1.6
	Adaptació de sistemes de referenciació geogràfica	Annex V objectiu 1.7
	Determinació de nivells de desagregació	Annex V objectiu 1.8
Perfeccionament metodològic	Sistematització de monografies metodològiques	Annex V objectiu 2.1
	Modelització economètrica regional	Annex V objectiu 2.2
	Col·laboració estadística amb centres de recerca	Annex V objectiu 2.3
	Territorialització de dades per a petites àrees geogràfiques	Annex V objectiu 2.4
	Metodologia d'estimació de fluxos comercials interterritorials	Annex V objectiu 2.5

ENGLISH SUMMARY:

THE STATISTICS PLAN ACT OF CATALONIA: THE LAW 30/1991 AND THE PLANIFICATION OF PUBLIC STATISTICAL ACTIVITY

J. Oliveras i Prats

THE INSTITUCIONALIZATION OF A REGIONAL STATISTICS SYSTEM

The Catalan Parliament voted the law on statistics (1987) similar in contents to statistic laws in other countries that established the technical and juridical guarantees in statistics: citizen collaboration, secrecy (of statistics) and public interest. It also stipulated that statistics had to be planned and that an official Board of Statistics had to be created, in order to apply the law. So the Catalan Institute of Statistics must produce the statistics which the Parliament and the government entrusts it with, through the Plans and Programmes for statistical activity, but it also has the responsibility of managing the statistical system. If the Catalan Statistics Institute does not take on this role of giving support, help and technical assistance to the statistical agents, it runs the risk of being one more producer of statistics, quite a qualified producer of statistics and a bureaucratic controller of statistical activity.

The fact that the Statistical Plan has to be approved by Parliament demonstrates Catalonia's desire to create a statistical system capable of satisfying our country's needs and ordering and harmonizing the statistical activity of Catalan institutions. We consider all the Catalan statistical activity as a system: an ordered and harmonic whole of methods, procedures and statistical results from the different institutional agents. In the same way, we define the Catalan statistical system not as a limited system, but one which integrates statistical systems of state or European scope, striving for a maximum homogeneity and comparability of data.

The process of elaboration of the Statistical Plan Act has undergone a long process of consultation, advice and technical assistance, carried out by the Catalan Institute of Statistics. In this process there have been wide and varied interactions among the different departments of the Catalan Government and among local administrations entities in Catalonia, with the aim of detecting

and critically analyzing together current statistical activity, as well as the needs suggested.

STRATEGIC PLANNING BY OBJECTIVES IN STATISTICS

Regional and state statistics Plans normally consist of a rigid list of statistics which governments are bound to execute over a specific period of time. The drawbacks are: the necessity of defining all the information needs for 4 years at a time, without the possibility of amending or adapting them to future needs. There is a tendency to include a large number of insufficiently-designed statistical projects to this list so as not to miss the opportunity offered, just once every 4 years.

The worst thing is that this list of statistics inhibits the politician, since there are no decisions to take because the statistical plan incorporates all of the statistical black hole. To try to solve this problem, we have adopted an alternative method, which is applying strategic planning by objectives within the legislation of statistical matters.

- * We have described the statistical operations in terms of objectives to reach, with a group of decision-making rules about priorities. This definition of statistical needs in the form of objectives allow policy makers to identify statistical activity with their own information needs.
- * In the Plan Act, a series of rules and criteria of decisions have been established in order to facilitate the priorities of statistical objectives: strict criteria and priority criteria. Strict criteria of decisions "sine qua non" conditions refer to guarantees of technical quality of projects, the legality of the methodology and the comparability of the statistical results, the periodicity of their publication and the non duplication with other statistical activities. The priority criteria refer to subjects and to the methodology to be used.
- * As to the methodology used, priority is given to those that use administrative sources or available statistical sources, over those derived from field work. As it is well known, sometimes the same information need can be solved with different statistical activities in cost, complexity and variable availabilities. This is why in the Statistical Plan Act there are the description of the information need to be covered and the way to achieve it. If the objective which describes an information need cannot be executed within the parameters of cost, complexity and availability contemplated by law,

the government is freed from the obligation of carrying out this statistic unless Parliament provides the necessary resources.

- * These *statistical objectives* described and subject to the decision-making rules about priorities of the Statistical Plan Act become statistical activities through the Annual Statistical Programme. In the Annual Statistical Programme the executive board decides which statistical objectives of the Plan Act will become statistical activities for the following year.

STATISTICAL ACTIVITIES BY SUBJECTS AND TECHNICAL REQUIREMENTS

The subjects of the statistical objectives related to the Statistical Plan Act tend to balance the available information giving priority to economic statistics followed by social ones. In the case of economic statistics the most important objective is to know institutional dynamics. Additionally, it is important to investigate the industrial economy, public administration, the financial activities, foreign trade, households, labour market and prices. Finally, other subjects considered are Input-Output analysis, National Accounting, small areas economies (county economies) and economics census.

Social statistics define its priorities in the population habits and welfare, and a welfare indicators system. Demography is considered to be sufficiently covered in Catalonia at the moment; however it is programmed to continue working in the correspondence between population census data and administrative information about natality, mortality and immigration statistics. It is also aimed (above all) to provide regular forecasts on future age and sex population distribution.

With the Planning Act we also faced another problem: in order to put the legal guarantees and technical correction requirements into operation, the Catalan Statistics Institute has to carry out a series of activities of methodological, technical and organizational character which are very expensive. So the activities and resources for these functions had to be planned.

We call instrumental statistical activities those that are legally required, or technically necessary to be able to comply with the requisites which the statistics legislation establishes. For instance, training, research and technical methodological and normative development in the field of statistics. Instrumental statistical activities consume important resources in official statistical bodies, and it is therefore appropriate to be able to identify them when planning statistical

activity and to allocate resources to them, allowing for the planning and systematizing of the activities of research, development and training in statistical methodology.

Through identifying instrumental statistics activity in the Statistics Plan the policy makers become aware of the importance and complexity of statistics works.

Butlleta de subscripció a la revista **Qüestió**

Nom i cognoms _____	

Empresa/Institució _____	

Adreça _____	

Codi postal _____	ciutat _____
Tel. _____	Fax _____
Data d'expedició _____	
Signatura:	DNI o NIF _____

Desitjo subscriure'm a **Qüestió** per a l'any 1994.

El preu de la subscripció és de 2.700 PTA.

Forma de pagament

- ☐ Transferència al compte de la Caixa de Catalunya número 6985/77,
Agència 100, Comte d'Urgell 162, 08036 Barcelona
- ☐ Domiciliació bancària
- ☐ Taló nominatiu a l'Institut d'Estadística de Catalunya
- ☐ Gir postal
- ☐ En efectiu

Retornar aquesta butlleta (o una fotocòpia) a:

Qüestió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista **Qüestió**

El sotasignat _____
autoritza el Banc/Caixa _____
Agència núm. _____ adreça _____
Codi postal _____ ciutat _____
a abonar a la revista **Qüestió** amb càrrec al meu compte número _____
_____, les subscripcions a **Qüestió**.
_____, a _____ d _____ de 19 ____

(Signatura)

El sotasignat _____
autoritza la revista **Qüestió** a carregar al meu compte número _____
_____, al Banc/Caixa _____

Agència núm. _____ adreça _____
Codi postal _____ ciutat _____
l'import de les tarifes vigents de les subscripcions a la revista **Qüestió**.
_____, a _____ d _____ de 19 ____

(Signatura)