QÜESTIIÓ, Vol 18, 1 pp. 7-38, 1994

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 $X_1, \ldots, X_p$ 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 Wasiucionek, 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 $X1, ..., X8$) 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 $X = logX$ 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.

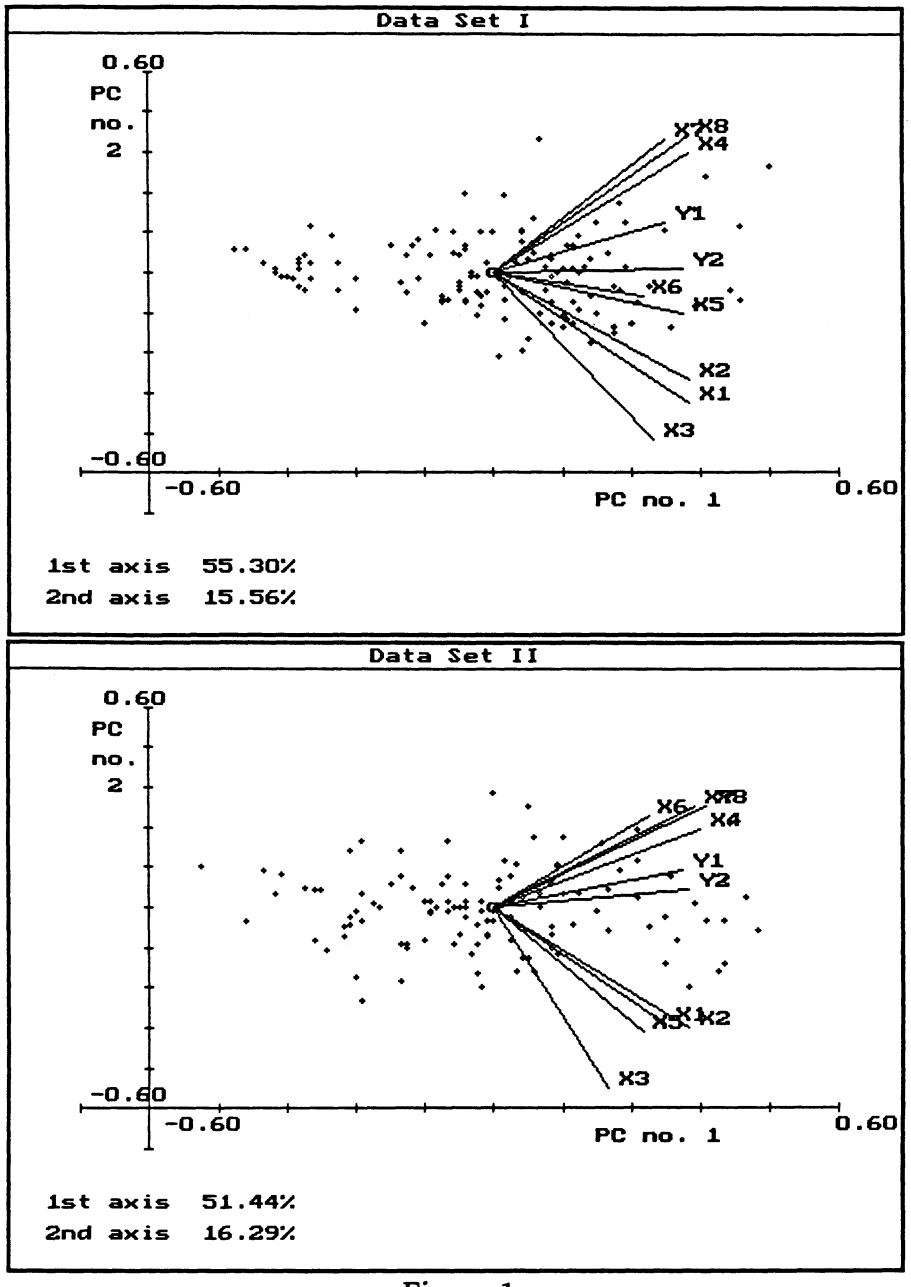

Figure 1.

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

$$\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{A}^T, \tag{1}$$

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

$$\mathbf{X} = \mathbf{G}\mathbf{H}^T, \tag{2}$$

with

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

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:

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

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 photoshere— 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 photosheric 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 $X5$ 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 continous 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, REGDIST, 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 $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$. 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):

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

L1-norm Distance (L1):

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

Gower's Distance (GOWER):

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

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

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: $\mathbf{D} = \left(d_A^2(i,j)\right)$, 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) $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^k$ 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:

$$\mathbf{A} = (a_{ij}), \quad \text{with} \quad a_{ij} = -(d_{ij}^2)/2,$$

$$\mathbf{B} = \mathbf{HAH}, \quad \text{with} \quad \mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^T. \tag{7}$$

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:

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

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

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 $h = 1, ..., n-1$.

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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq \epsilon$. 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 $\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_m$; the `tilde ~` should remind us that these are rescaled eigenvectors:

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

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

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

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

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 $\tilde{\mathbf{\Gamma}}$, i.e. the vector $\underline{\tilde{\gamma}}_l$ can

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

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

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 $\mathbf{y} = (y_1, \ldots, y_n)^T$ its sample values.

We compute the Pearsonian correlation of each of the established principal coordinates $\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_m$ 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 $\tilde{\Gamma}_{i_1}, \ldots, \tilde{\Gamma}_{i_k}$ mostly correlated with the vector $\mathbf{y}$. For simplicity the retained PC's will be denoted from now by $\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_k$. These principal coordinates can be put together as column vectors into the matrix $\tilde{\mathbf{\Gamma}}_{(k)} = (\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_k)$. Obviously:

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

with $\mathbf{\Lambda}_{(k)}$ 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:

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

This equation corresponds to an ordinary regression model. The unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$ can be computed by known methods. The mutual orthogonality of the involved PC's, i.e. of the column vectors $\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_k$ 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 $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)$ 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 $\tilde{\mathbf{\Gamma}}_{(k)}$ 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 identic 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) $$\hat{y}_i = \beta_0 + \beta_1 \tilde{\gamma}_{i1} + ... + \beta_k \tilde{\gamma}_{ik}$$

with $\beta_0 = \overline{y}$ and $\boldsymbol{\beta} = \boldsymbol{\Lambda}_{(k)}^{-1} \tilde{\boldsymbol{\Gamma}}_{(k)}^T \mathbf{y}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$.

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 $\widehat{y^*}$ 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 $\breve{\mathbf{b}}$ and $\mathbf{d}^*$, where

$$\breve{\mathbf{b}} = (\breve{b}_1, \ldots, \breve{b}_n) = (b_{11}, ..., b_{1n})$$

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

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

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 R^m (of all evaluated PC's) and R^k (the retained PC's mostly correlated with $\mathbf{y}$) are respectively:

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

and

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

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 $\widehat{y^*}$ can be evaluated by the formula :

(10) $$\widehat{y^*} = \overline{y} + \mathbf{c}_{(k)} \boldsymbol{\Lambda}_{(k)}^{-1} \tilde{\boldsymbol{\Gamma}}_{(\mathbf{k})}^T \mathbf{y}.$$

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 **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.

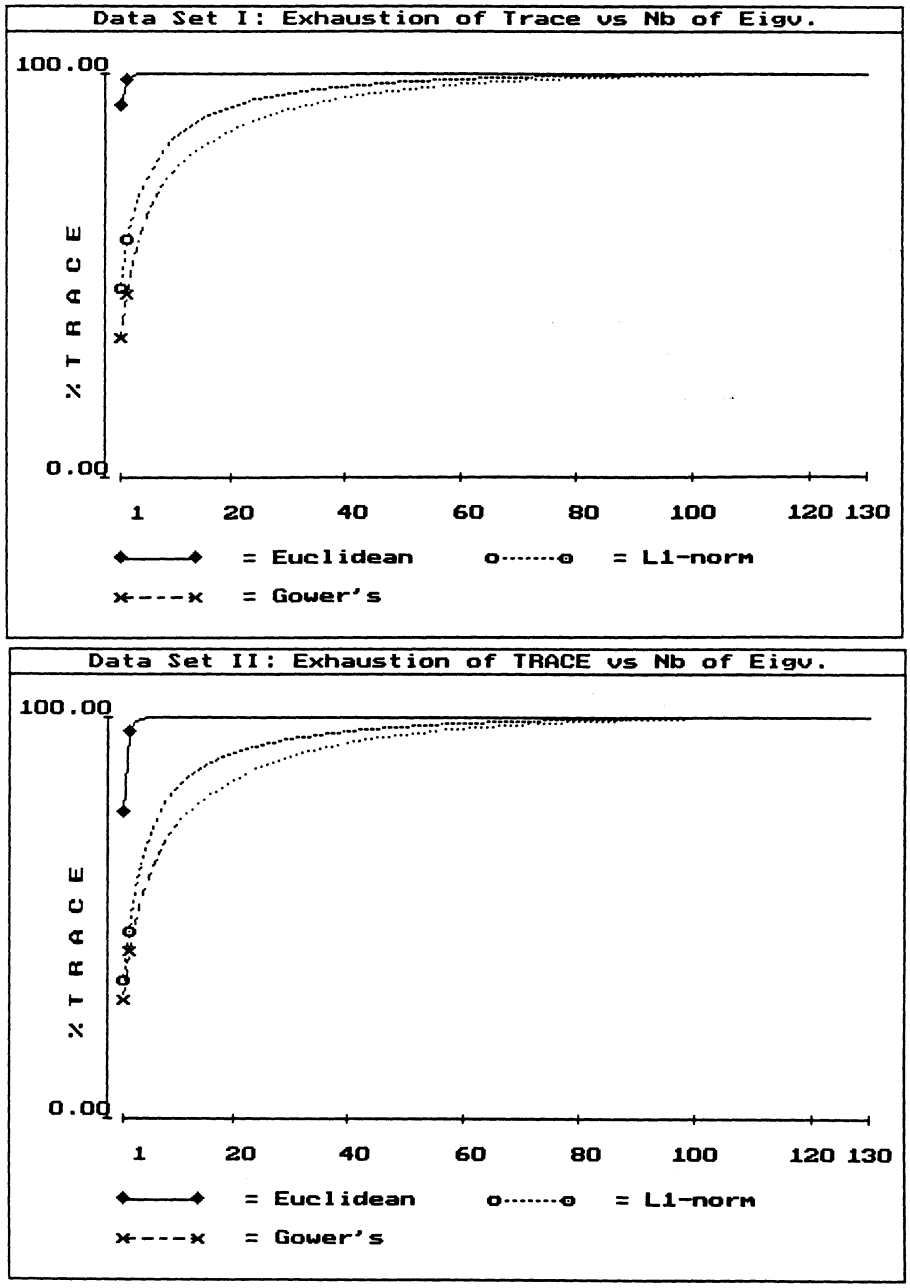

Figure 2.

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 $i = 1, ..., 8$ (all positive eigenvalues) for the Euclidean distance, and $i = 1, ..., n - 1$ (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	$Y1$ = MvXY				$Y2$ = Fs			
	Data Set I		Data Set II		Data Set I		Data Set II	
Distance	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	$Y1$ = MvXY				$Y2$ = Fs			
	Data Set I		Data Set II		Data Set I		Data Set II	
Distance	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

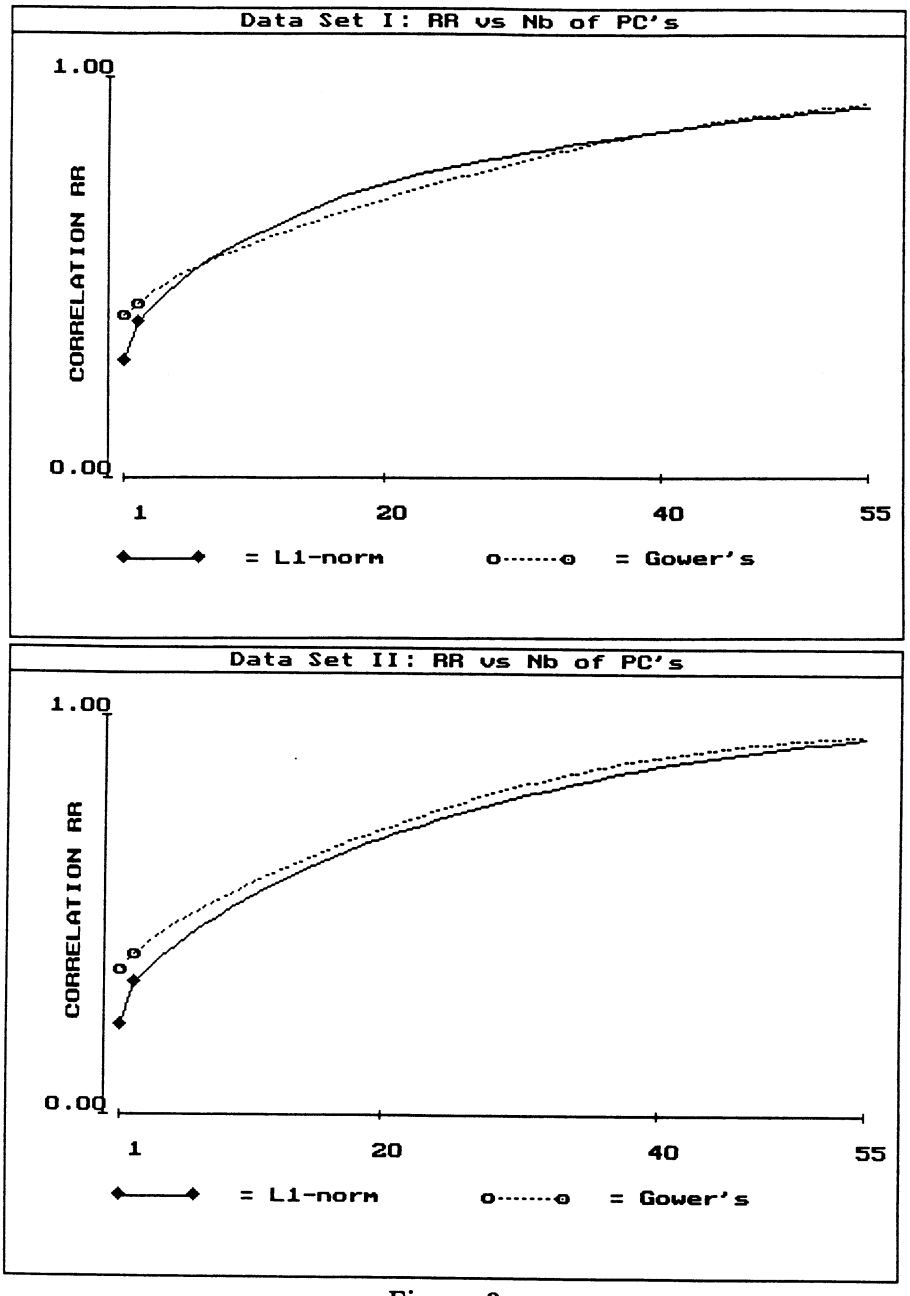

Figure 3.

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 Gowers'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	$Y1$	I	1	5	4	2	8	6	7	3	0.3800
Distance		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	$Y1$	I	1	46	2	82	35	23	24	6	0.4993
Distance		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	$Y1$	I	1	2	45	23	38	80	6	12	0.4773
Distance		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	$Y1$	I	1	4	3	2	7	5	6	—	0.3734
Distance		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	$Y1$	I	1	40	2	5	111	96	72	—	0.4937
Distance		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	$Y1$	I	1	2	40	74	5	23	39	—	0.5002
Distance		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

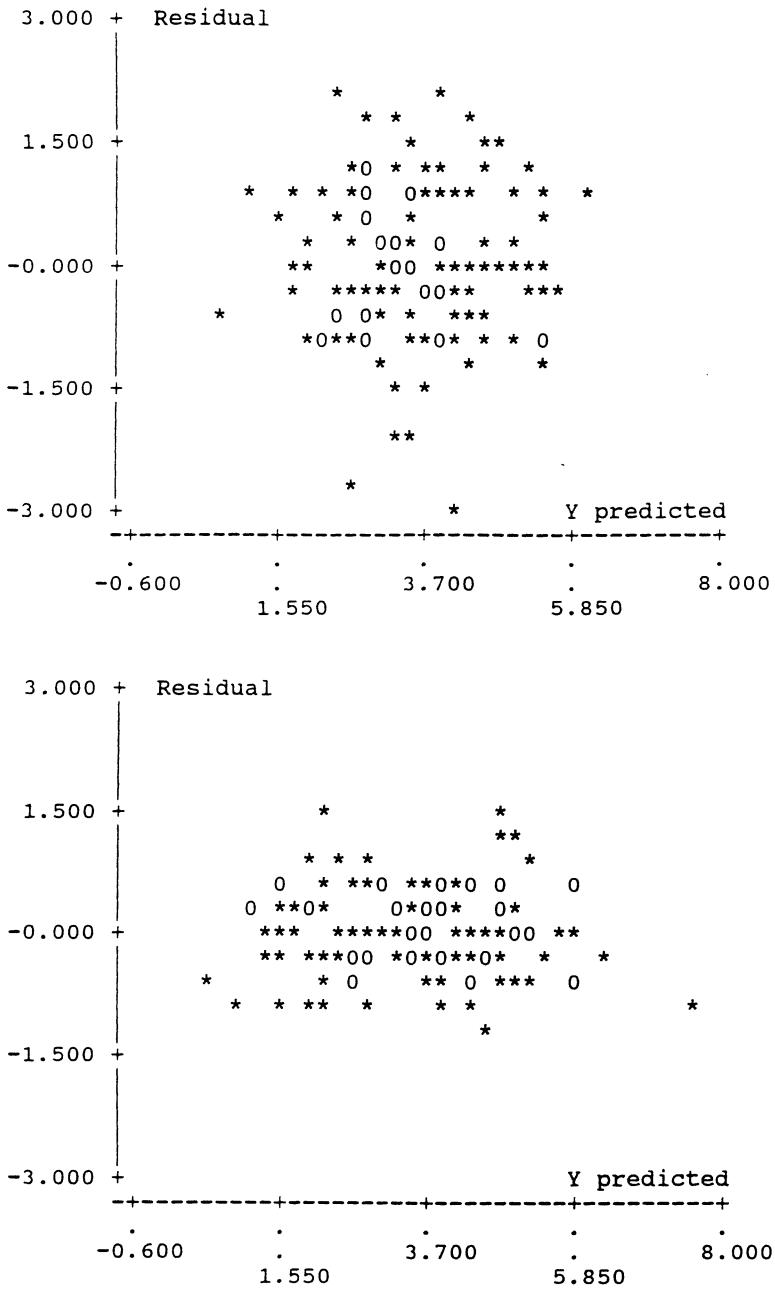

Figure 4.

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

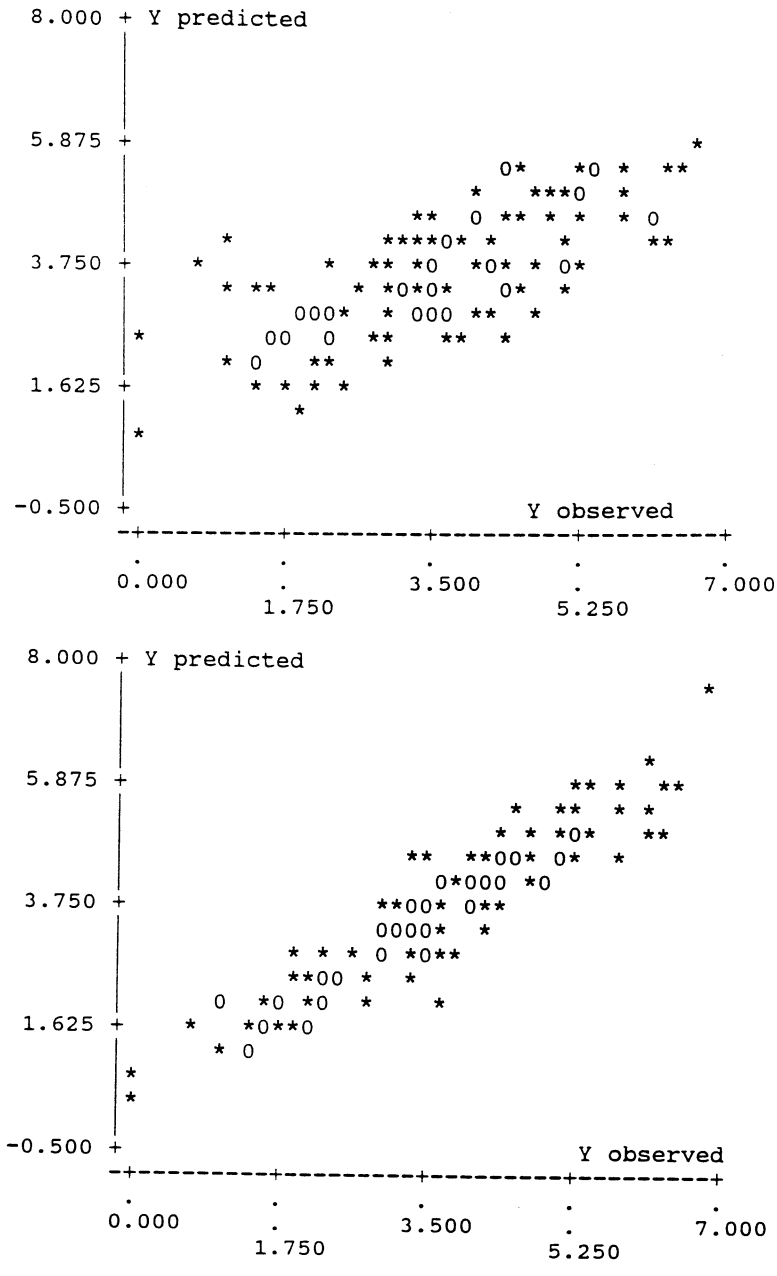

Figure 5.

Predicted values of Y2=Fs 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 $\hat{y}$ 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 $i = 1, ..., 130$, i.e. for all the items belonging to this data set. The respective residuals were evaluated as:

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

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:

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

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 $\mathbf{\Lambda}_{(k)}$ 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 $i = 1, ..., 117$ 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		Estimation from foreign data set		
					SD_r	q_r	AV_v	SD_v	q_v
Mv XY	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
Fs	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 significancy 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'Estadistica.

[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.* 2ª 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 Estadisticas". *Estadistica Española*, **30**, n° **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 Estadisticos 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.** & **Wasiucionek, 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.