ACTA ASTRONOMICA
Vol. 39 (1989) pp. 85-100

Search for Outlying Data Points in Multivariate Solar Activity Data Sets

by

A. Bartkowiak
Institute of Computer Science, University of Wrocław, Poland

and

M. Jakimiec
Astronomical Institute, University of Wrocław, Poland

ABSTRACT

The aim of this paper is the investigation of outlying data points in the solar activity data sets. Two statistical methods for identifying of multivariate outliers are presented: the chi2-plot method based on the analysis of Mahalanobis distances and the method based on principal component analysis, i.e. on scatterdiagrams constructed from the first two or last two eigenvectors. We demonstrate the usefullness of these methods applying them to same data of solar activity. The methods allow to reveal quite precisely the data vectors containing some errors and also some untypical vectors, i.e. vectors with unusually large values or with values revealing untypical relations as compared with the common relations between the appropriate variables.

1. Introduction

When analysing astronomical observations we are very often confronted with the problem that the data arising from various sources are of different quality. It is a common practice that the data lying far from the main data bulk are used with a reduced weight. One can suspect that such remote data vectors can comprise errors introduced by the data processing or the observation deficiency (the problem of uncertain data is mentioned e.g. by Pfleiderer, 1983). It can happen, however, that such remote data points are not faulty observations and are very substantial for a considered problem. With such situation met Giannuzzi (1981) who analysed the relation between total masses of close binary systems (and also total orbital angular momenta) and the mass ratios for semidetached systems, and estimated the values of the regression coefficients on the basis of data comprising also several remote points relevant for the problem. Thus, the outlying values may contain erroneous observations as well as individual points with an important information about the analysed problem. Pfleiderer and Krommidas (1982) suggest to make first a preliminary analysis based on examination of the probability distribution of the considered variable to identify gross errors, if the exist, in the data set. For the multivariate data, however, the analysis based on univariate probability distributions only may be too simple and unsatisfactory. Moreover, even in the univariate case, the probabilistic considerations are difficult to carry out for non-normal distributions which occur very often in astronomical practice. For instance, Bartkowiak and Jakimiec (1986) analysing multivariate data comprising daily characteristics of sunspot groups reported that the distributions of these characteristics were very skew. So, we infer that some outlaying values of one or several of these characteristics may be observed on a given day, and yet they should not be treated as erroneous but as proper and relevant for the interrelation structure of the analysed variables.

In every statistical analysis a check for outliers should be carried out at the beginning of the study. Many methods for detection of univariate and multivariate outliers are known (see e.g. Gnanadesikan and Kettenring, 1972; Belsley *et al.* 1980; Jolliffe, 1986; Atkinson, 1987). In the present paper we want to demonstrate the efficiency of some of these methods – namely chi2-plots and principal components – using solar activity data sets. Furthermore, we will demonstrate that the applied methods allow us to discover observational vectors with values which disagree with the internal covariance structure of the considered variables. We will also show how strongly one erroneous observation (due to a mistake in the components of a data vector) may influence the results of the statistical investigations of the internal covariance structure of the data.

2. The data

A comprehensive description of the data used in the present analysis is given by Jakimiec and Wanke-Jakubowska (1989). Now we analyse a complex of fourteen variables characterizing sunspot groups of D, E, F Zurich classes for which the daily gradient of the area is negative, i.e. sunspot groups being in the decay phase. The variables $x1$ – $x7$ describe the sunspot group daily characteristics: McIntosh class (McI), sunspot group area (A), calcium plage area (CaA) and intensity (CaI), magnetic class (Mag), magnetic field strength (H) and magnetic field index (MFI). Further seven variables $x8$ – $x14$ describe flare activity of the sunspot group for a given day: maximum value of X-ray flare flux (maxX), number of faint flares (NFF) and number of stronger flares (NSF), the daily sum of the X-ray flare fluxes in the wavelength interval 1–8Å (Fs), hardness index (HI), the daily sum of the X-ray flare fluxes in the wavelength interval 0.5–4Å (Fh) and the daily maximum value of the six-hour hardness index (maxh).

The distributions of the variables $x1$, $x2$, $x7$, $x8$ and $x11$ were very skew. Therefore the logarithmic transformation $X' = \log X$ was applied to these variables. For the same reason we transformed also the variables $x3$ and $x13$, however, we used here the formula $X' = \log X - 2$ and $X' = \log X + 2$, respectively (X stands for the values before the transformation, and X' for the values after the transformation).

The values of the mentioned fourteen variables $x1$ — $x14$ compiled for a sunspot group on a given day will be called a data vector or an item. The considered data were put together as a $n \times p$ matrix X, with $n = 149$ denoting the number of items, and $p = 14$ denoting the number of the variables.

3. Detection of outliers by chi2-plots

The method of constructing chi2-plots is described e.g. by Gnanadesikan and Kettenring (1972). Let us suppose that our data comprise observations of p variables for n data vectors (items). Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ denote the vector of observations for the i-th item (with $i = 1, \ldots, n$). This vector can be viewed as a point in the variable space R^p. Let $\bar{\mathbf{x}}_. = (\bar{x}_{.1}, \ldots, \bar{x}_{.p})$ be the mean vector of the values x_{ij}. From a geometrical point of view $\bar{\mathbf{x}}_.$ can be considered as the center of gravity of all data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. For each item i we calculate D_i^2, the Mahalanobis distance of the point $\mathbf{x}_i$ from the center of gravity $\bar{\mathbf{x}}_.$:

$$D_i^2 = (\mathbf{x}_i - \mathbf{x}_{.i}) S^{-1} (\mathbf{x}_i - \mathbf{x}_{.i})', \tag{1}$$

where S is the covariance matrix of the considered p variables. We order the values of D_i^2 in a nondecreasing sequence:

$$D_{(1)}^2 \leqslant D_{(2)}^2 \leqslant D_{(3)}^2 \leqslant \ldots \leqslant D_{(n)}^2$$

and associate with each ordered value a corresponding quantile of the χ_p^2 distribution. Putting these quantiles againts the ordered Mahalanobis

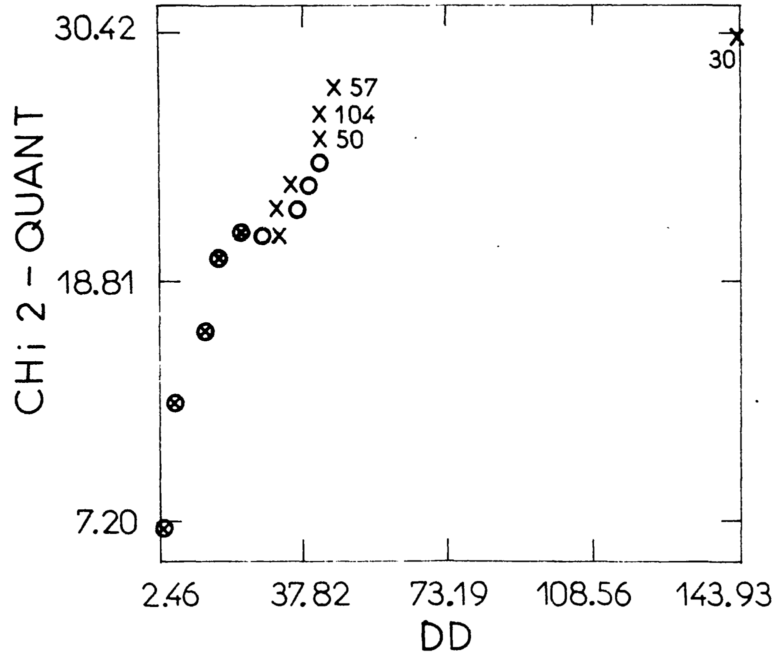

Fig. 1. Chi2-plot for $n = 149$ items with one big outlier (erroneous data). Notations: crosses — single points, open circles — multiple points, crossed circles — class representatives. A unit on the x-axis is equal to 3.537 and on the y-axis — 1.161.

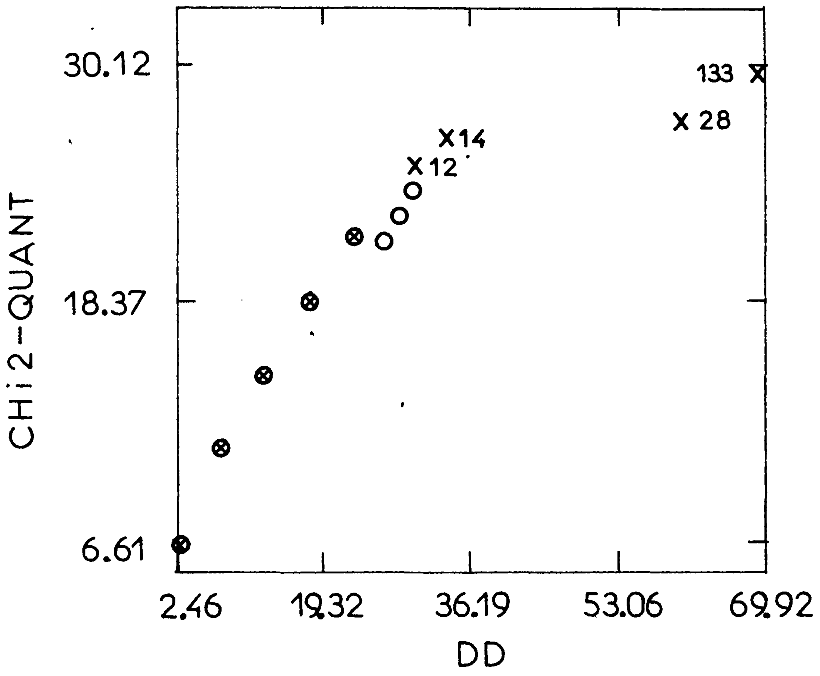

Fig. 2. Chi2-plot for erroneous data after trimming off the $\alpha=0.1$ part of items with the largest values of D^2. Notations as in Fig. 1. Units: x-axis — 1.687, y-axis — 1.175.

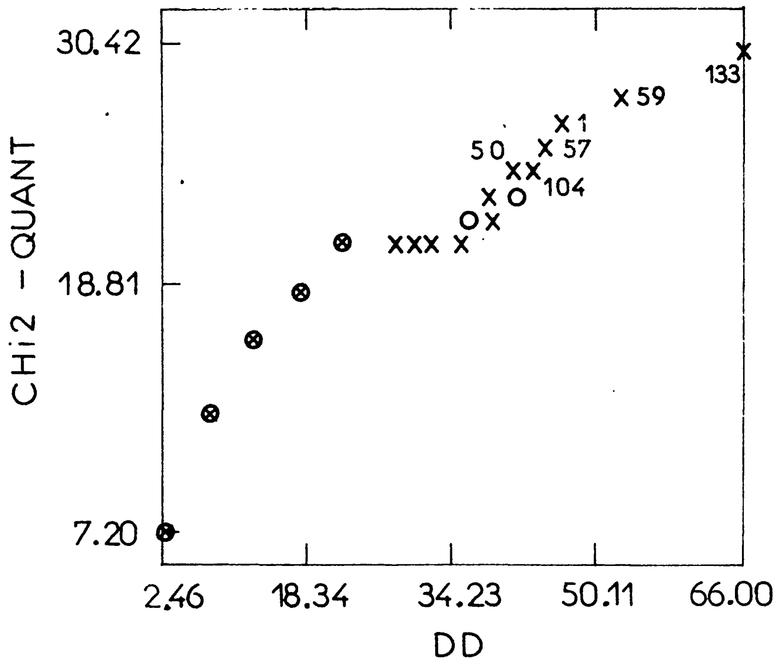

Fig. 3. Chi2-plot for $n=149$ items after correcting the item no. 30 (corrected data). Notations as in Fig. 1. Units: x-axis — 1.589, y-axis — 1.161.

distances $D^2_{(i)}$ we obtain a chi2-plot. Such chi2-plots constructed for our data are shown in Figures 1 – 4. The outliers appear in these plots as isolated points with large values of the Mahalanobis distance. To speed up the calculations we retained exactly only D^2 values surpassing some threshold value T.

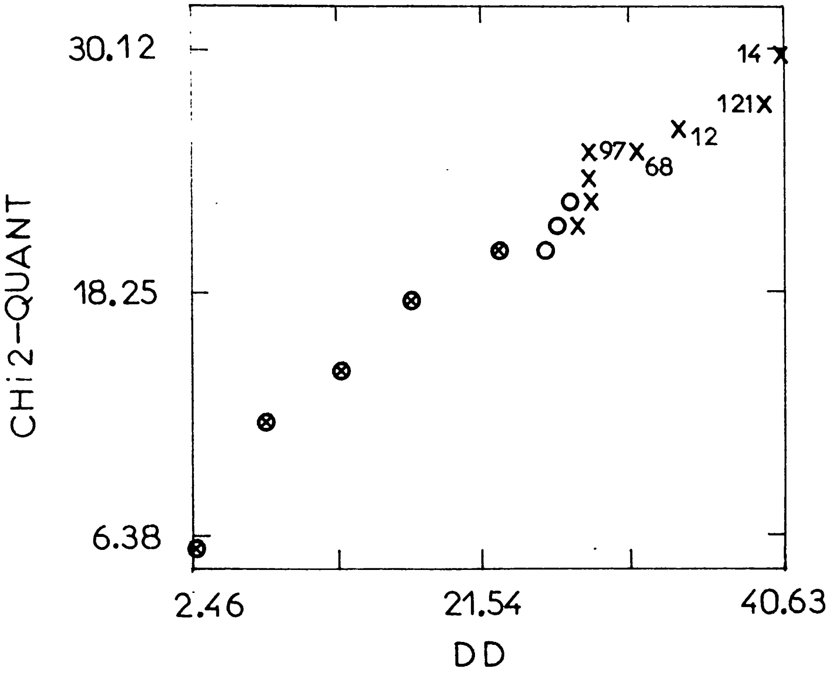

Fig. 4. Chi2-plot for the corrected data after trimming off the $\alpha = 0.1$ part of items with the largest values of D^2. Notations as in Fig. 1. Units: x-axis – 0.954, y-axis – 1.187.

The interval $(0, T)$ was subdivided into five classes established at the beginning of the calculations. The D^2 values smaller than T were assigned to the appropriate classes. For the circled-crossed points in Figures 1 to 4 abscissa gives the centres of these classes, and ordinate gives the appropriate ranks of these centres obtained from the cumulative frequencies. The distances D^2 surpassing the treshold T are presented in Figures 1 – 4 with the exact values.

In Figure 1 we see one really outlying point which corresponds to the item no. 30. The other points seem to be close together. Fourteen items with largest values of the Mahalanobis distance are shown in Table 1 (first row). They constitute an $\alpha = 0.1$ part of the all considered data vectors. We removed these items from our data and repeated the construction of the chi2-plot. It is shown in Figure 2 as the chi2-plot for the trimmed data. We see there two clearly isolated points: the items no. 133 and no. 28. From this it follows that after removing the 14 items with the largest Mahalanobis distances we do not obtain a homogeneous set of data.

Checking the correctness of our data for the revealed outlier (the item no. 30) we found that there was a serious (grave) error in this data vector. The erroneous and also the correct data for this item are shown in Table 2. The

Table 1
Items with largest values of the Mahalanobis distance

Rank Kind of data	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Erroneous data with item no.30	30	57	103	50	22	1	59	58	102	19	35	15	8	68
Erroneous data after cut-off	133	28	14	12	105									
Corrected data	133	59	1	57	104	50	22	58	35	15	19	102	28	8
Corrected data after cut-off	14	121	12	68	97									

Table 2
Erroneous and correct values of the variables $x1$ — $x14$ for the item no. 30 and mean and standard deviation values calculated for all data, i.e. for $n = 149$ items

Variable	Data for item no.30 Erroneous	Correct	Mean	Standard deviation
x1 - McI	2.06	2.06	1.62	0.28
x2 - A	2.89	2.89	2.32	0.42
x3 - CaA	1.79	1.79	1.49	0.28
x4 - CaI	0.54	0.54	0.50	0.05
x5 - Mag	2.00	2.00	2.65	1.13
x6 - H	5.00	5.00	4.05	0.80
x7 - MFI	1.60	1.60	1.11	0.46
x8 - maxX	0.48	0.48	0.45	0.53
x9 - NFF	8.00	8.00	3.36	2.85
x10 - NSF	1.04	1.00	1.30	2.34
x11 - Fs	0.02	1.04	0.71	0.63
x12 - HI	1.41	0.02	0.04	0.04
x13 - Fh	0.04	1.41	1.13	0.96
x14 - maxh	0.90	0.04	0.05	0.06

Table 3
Covariances between the $x10$ — $x14$ variables

	Erroneous data					Corrected data				
Variable	10	11	12	13	14	10	11	12	13	14
	(big error in the item no.30)					(the item no.30 corrected)				
10	5.455					5.456				
11	1.072	0.396				1.070	0.394			
12	0.063	0.014	0.014			0.066	0.021	0.002		
13	1.641	0.591	0.025	0.935		1.638	0.587	0.035	0.927	
14	0.084	0.023	0.010	0.039	0.008	0.086	0.027	0.002	0.045	0.003
	(after cut-off an 0.10 part)					(after cut-off an 0.10 part)				
10	2.508					2.237				
11	0.718	0.326				0.642	0.306			
12	0.040	0.015	0.001			0.031	0.013	0.001		
13	1.121	0.489	0.026	0.777		0.998	0.456	0.022	0.724	
14	0.052	0.021	0.002	0.035	0.002	0.041	0.018	0.001	0.030	0.002

mean values and the values of standard deviations calculated for each variable are given also in Table 2. The errors were found for values of the variables $x10 - x14$ (a missing value of the variable $x10$ and a shift of the values for the variables $x11 - x14$). The erroneous and corrected (for the item no. 30) covariances are shown in Table 3. For the corrected data (i.e. for the data with the corrected item no. 30) we constructed anew the chi2-plots both for the whole set of data and for the data after trimming off an $\alpha = 0.10$ part of the items with the largest values of the Mahalanobis distances. The plots are shown in Figure 3 (the whole set of data) and in Figure 4 (after trimmnig of an 0.10 part of the data). The extreme points visible in the figures are listed in Table 1. Except the points no. 30 and no. 68 they are the same as found previously from Figure 1, but now they appear in different order. It is curious that the most outstanding point (the item no. 133) was not found previously among the fourteen most outstanding points, and it was revealed only when these fourteen items were trimmed off from the considered data.

4. Detection of outliers by principal components

The method of principal components is described e.g. by Morrison (1967). We calculate the eigenvalues $l_1, l_2, \ldots, l_p$ and the appropriate eigenvectors $\mathbf{a}_1$, $\mathbf{a}_2, \ldots, \mathbf{a}_p$ of the covariance matrix S. Then we make a projection of points from R^p onto planes spanned by the pairs $(\mathbf{a}_h, \mathbf{a}_k)$ centered at the mean of the cloud.

Let us consider the influence of the erroneous data vector no. 30 on the structure of the eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ and on the principal components. The differences between the covariances for the variables $x10 - x14$ may be seen in Table 3. The structure of the eigenvectors calculated for the erroneous and corrected data coincide for the first three vectors. In vector $\mathbf{a}_4$ we note a change in the sign of the components. The structure is different in the eigenvectors $\mathbf{a}_9$, $\mathbf{a}_{11}$, $\mathbf{a}_{13}$ and $\mathbf{a}_{14}$ – especially in the 11 to 14-th components. Table 4 shows the eigenvectors $\mathbf{a}_9 - \mathbf{a}_{14}$ calculated from the erroneous and corrected data (i.e. with the proper values for the item no. 30). The underlined numbers indicate the values which differ strongly for the erroneous and corrected data.

It is generally believed (see e.g. Gnanadesikan and Kettenring 1972) that the plots constructed from the first few or last few eigenvectors reveal the outliers. We made the scatterdiagrams of the projections on the planes $(\mathbf{a}_1, \mathbf{a}_2)$ and $(\mathbf{a}_{13}, \mathbf{a}_{14})$. The components of the eigenvectors $\mathbf{a}_1$ and $\mathbf{a}_2$ calculated from the erroneous and corected data are very similar, and so look also the scatterdiagrams visualising the projections in planes obtained from these vectors. In Figure 5 we show the scatterdiagram obtained from the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ evaluated for the erroneous data. The item no. 30 is not seen as outlier here. Instead, we see here at outstanding positions the points (items) nos. 1, 50, 57 and 58 found already when considering the Mahalanobis distances.

Table 4

Eigenvectors $\mathbf{a}_9$ — $\mathbf{a}_{14}$ from the covariances evaluated for the erroneous and corrected data

Component	Eigenvectors evaluated for the erroneous data					
	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}
1	0.048	0.861	0.006	−0.008	0.026	0.003
2	−0.165	−0.492	0.066	0.014	0.033	0.002
3	0.194	−0.013	0.052	0.037	0.053	0.018
4	0.009	−0.014	0.040	−0.998	−0.017	−0.034
5	0.020	−0.008	0.001	−0.002	0.000	0.000
6	0.027	0.015	−0.030	−0.005	−0.008	−0.002
7	−0.061	0.028	−0.036	0.002	0.000	−0.004
8	0.743	−0.110	−0.291	−0.007	0.341	−0.015
9	0.023	−0.012	−0.004	0.001	0.034	−0.002
10	0.002	−0.004	0.001	0.000	0.010	−0.001
11	−0.107	−0.001	−0.493	0.000	−0.733	0.039
12	−0.398	−0.030	0.562	0.051	−0.471	−0.539
13	0.376	0.051	0.434	0.022	0.227	−0.034
14	0.255	−0.021	0.397	−0.008	−0.255	0.840

Component	Eigenvectors evaluated for the corrected data					
	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}
1	−0.108	−0.855	−0.034	−0.009	0.010	−0.001
2	0.179	0.483	−0.046	0.015	−0.003	0.004
3	−0.105	0.016	−0.100	0.033	0.046	0.002
4	0.014	0.012	−0.022	−0.994	−0.092	−0.018
5	−0.012	0.007	−0.002	−0.003	0.000	0.000
6	−0.035	−0.013	0.019	−0.005	−0.006	0.000
7	0.010	−0.026	0.020	0.004	−0.012	−0.001
8	−0.809	0.158	−0.282	0.008	−0.087	0.009
9	−0.029	0.014	−0.035	0.002	−0.004	−0.001
10	−0.002	0.004	−0.010	0.001	−0.002	0.000
11	−0.078	0.017	0.882	−0.023	0.134	0.005
12	−0.012	0.013	−0.103	−0.040	0.607	−0.785
13	−0.531	−0.089	−0.313	0.033	−0.088	−0.006
14	0.002	0.009	−0.137	−0.080	0.767	0.619

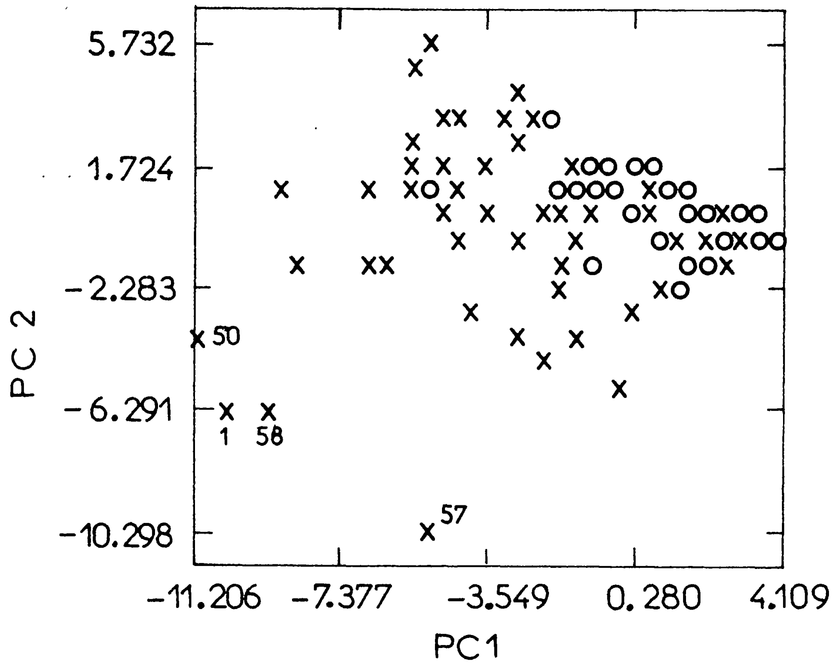

Fig. 5. Projections of points-items onto the plane ($\mathbf{a}_1$, $\mathbf{a}_2$). Eigenvectors calculated for the erroneous data. The abscissa and ordinate are the principal components PC1 and PC2, respectively. Notations: crosses — single points, open circles — multiple points.

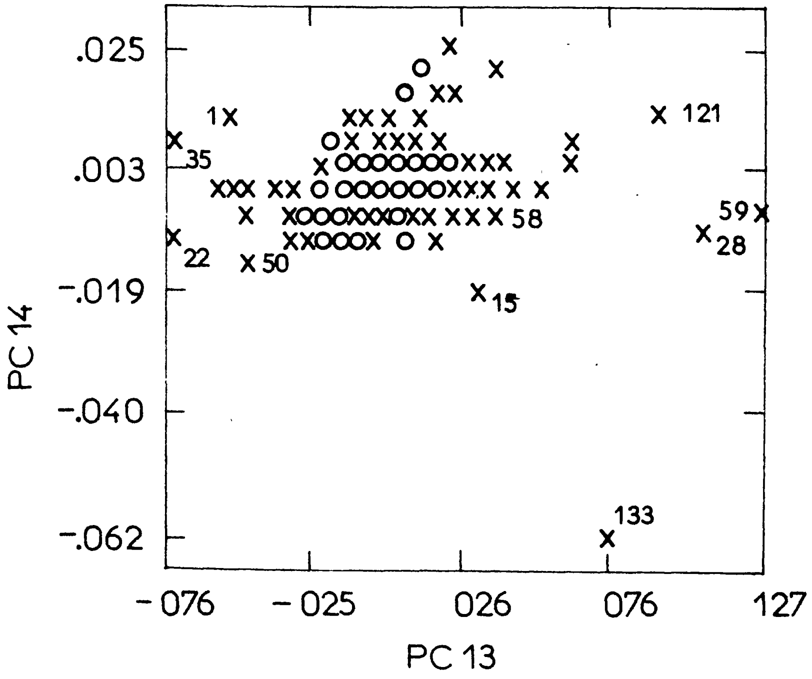

Fig. 6. Projection of points-items onto the plane (a_{13}, a_{14}). Eigenvectors calculated for the corrected data. The abscissa and ordinate are the principal components PC13 and PC14, respectively. Notations: crosses – single points, open circles – multiple points.

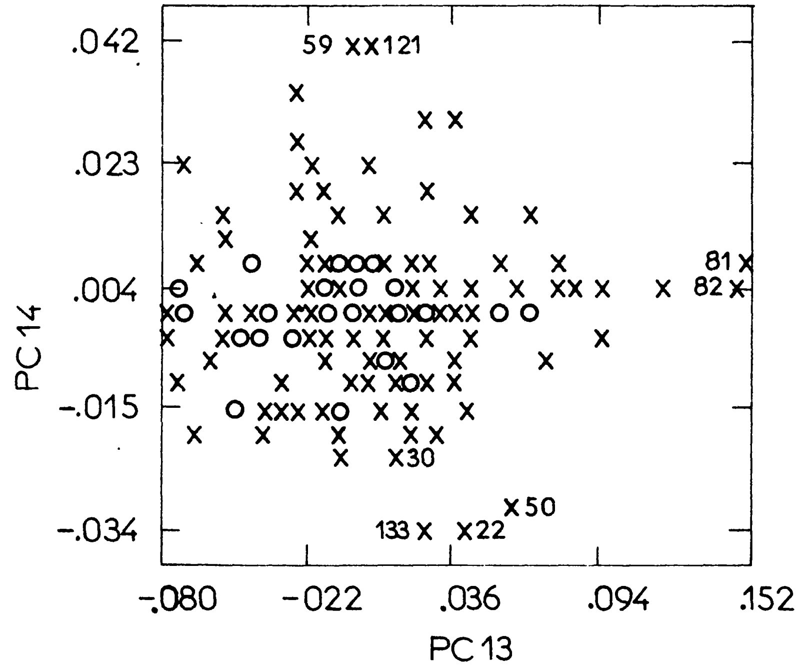

Fig. 7. Projection of points-items onto the plane (a_{13}, a_{14}). Eigenvectors calculated for the erroneous data. The abscissa and ordinate are the principal components PC13 and PC14, respectively. Notations: crosses – single points, open circles – multiple points.

We constructed also the scatterdiagrams based on the last two eigenvectors, i.e. $\mathbf{a}_{13}$, $\mathbf{a}_{14}$. In Figure 6 we show the scatterdiagram constructed for the corrected for the item no. 30 data – one can see here clearly the outstanding position of the point-item no. 133. We find also at some extreme positions the points nos. 15, 28, 59, 21, 1, 35 and 22. All these points have large Mahalanobis distances (see Table 1). An analogous scatterdiagram (Figure 7) constructed from the erroneous data yields the similar informations with the difference that the outstanding position of the point no. 133 is not so clearly pronounced. The scatterdiagrams constructed from the last two eigenvectors also did not reveal the erroneous item no. 30.

5. A closer look on the revealed outliers

We analysed in more detail the data vectors for the items having large Mahalanobis distances and appearing at extreme positions in the scatterdiagrams constructed from the first two or last two principal components. In the following we will call these items "suspicious outliers".

The values x_{ij} (where i is the no. of the item, and j is the no. of the variable) for twelve such items are shown in Table 5 (with normal characters). To get an idea whether the values x_{ij} are big or small as compared with other observed values of the j-th variable we calculated the standardized values t_{ij} using the formula:

$$t_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_j}, \tag{2}$$

where $\bar{x}_{.j}$ and s_j are the sample mean and standard deviation, respectively. In our case the means $\bar{x}_{.j}$ and the standard deviations s_j (for $j = 1, \ldots, 14$) were calculated from the data after correcting them for the erroneous item no. 30. The values t_{ij} calculated for the twelve analysed items are given in Table 5•(with bold characters).

Let us remind that for data points coming from a normal population $N(\mu_j, \sigma_j)$ with expected values $\mu_j = \bar{x}_{.j}$ and variances $\sigma_j^2 = s_j^2$ about ninety five percent of analysed observations should have values t_{ij} belonging to the interval $(-1.96, +1.96)$.Individuals having the absolute values of t_{ij} larger than 2.0 can be suspected to be nonhomogeneous with the remaining part of the data – in other words, they can be suspected as sampled from a population with other probability distribution as the assumed $N(\mu_j, \sigma_j)$ distribution.

We do not show in Table 5 the item no. 30. After correcting its mistakenly shifted values (see Table 2) it became homogeneous with the remaining individuals and it is revealed neither in the chi2-plot nor in the scatterdiagrams constructed from the first two or last two principal components.

Looking at the t-values shown in Table 5 one can see that eight of them (the items nos. 1, 15, 28, 50, 57, 58, 59 and 104) have outstanding values with

Table 5

The values x_{ij} (normal characters) and their standardized t_{ij} values (bold characters) for some suspicious outliers

Item			V	a	r	i	a	b	l	e	s			
no.	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
1	1.91	2.43	1.74	0.48	4	4	1.51	2.00	8	12	2.32	0.08	3.21	0.12
	1.0	**0.3**	**0.9**	**-0.4**	**1.2**	**-0.1**	**0.9**	**2.9**	**1.6**	**4.6**	**2.6**	**1.0**	**2.2**	**1.3**
15	1.95	2.80	~~1.84~~	0.54	5	4	1.87	2.34	3	2	2.38	0.18	3.62	0.18
	1.2	**1.2**	**1.1**	**0.8**	**2.1**	**-0.1**	**1.6**	**3.6**	**-0.1**	**0.3**	**2.6**	**3.2**	**2.6**	**2.4**
22	1.65	2.32	1.83	0.48	1	5	1.00	0.18	2	0	0.51	0.02	1.81	0.02
	0.1	**0.0**	**1.2**	**-0.4**	**-0.5**	**1.2**	**-0.2**	**-0.5**	**-0.5**	**-0.6**	**-0.3**	**-0.4**	**0.7**	**-0.5**
28	2.06	2.84	1.65	0.54	3	5	1.60	2.11	9	7	2.30	0.24	3.68	0.29
	1.5	**1.2**	**0.6**	**0.8**	**0.3**	**1.2**	**1.1**	**3.1**	**2.0**	**2.4**	**2.5**	**4.6**	**2.6**	**4.3**
35	1.95	3.00	1.68	0.54	2	4	1.20	1.18	4	0	0.72	0.02	0.95	0.02
	1.2	**1.6**	**0.7**	**0.8**	**-0.6**	**-0.1**	**0.2**	**1.4**	**0.2**	**-0.6**	**0.0**	**-0.4**	**-0.2**	**-0.5**
50	2.06	2.85	1.74	0.48	5	5	1.67	0.85	10	11	1.67	0.05	2.34	0.05
	1.5	**1.3**	**0.9**	**-0.4**	**2.1**	**1.2**	**1.2**	**0.8**	**2.3**	**4.2**	**1.5**	**0.3**	**1.3**	**0.0**
57	1.95	2.95	0.95	0.60	5	4	1.87	1.40	1	12	1.83	0.12	2.92	0.16
	1.2	**1.5**	**-1.9**	**2.0**	**2.1**	**-0.1**	**1.6**	**1.8**	**-0.8**	**4.6**	**1.8**	**1.9**	**1.9**	**2.0**
58	2.06	2.87	1.00	0.60	3	4	2.18	1.85	7	12	2.06	0.20	3.36	0.26
	1.5	**1.3**	**-1.7**	**2.0**	**0.3**	**-0.1**	**2.3**	**2.6**	**1.3**	**4.6**	**2.1**	**3.7**	**2.3**	**3.8**
59	1.95	2.87	1.00	0.54	4	5	1.97	1.70	11	5	1.95	0.26	3.37	0.33
	1.2	**1.3**	**-1.7**	**0.8**	**1.2**	**1.2**	**1.9**	**2.4**	**2.7**	**1.6**	**2.0**	**5.1**	**2.3**	**5.1**
104	1.30	2.20	0.18	0.48	2	4	0.00	0.78	2	1	0.90	0.10	1.91	0.13
	-1.1	**-0.3**	**-4.6**	**-0.4**	**-0.6**	**-0.1**	**-2.4**	**0.6**	**-0.5**	**-0.1**	**0.3**	**1.4**	**0.8**	**1.5**
121	1.78	2.56	1.43	0.48	2	5	1.26	0.48	3	1	0.60	0.07	0.48	0.10
	0.6	**0.6**	**-0.2**	**-0.4**	**-0.6**	**1.2**	**0.3**	**0.1**	**-0.1**	**-0.1**	**-0.2**	**0.7**	**-0.7**	**0.8**
133	1.38	2.43	1.75	0.48	2	4	0.70	0.30	2	0	0.54	0.10	0.60	0.02
	-0.8	**0.3**	**0.9**	**-0.4**	**-0.6**	**-0.1**	**-0.9**	**-0.3**	**-0.5**	**-0.6**	**-0.3**	**1.4**	**-0.5**	**-0.5**

$|t| > 2.0$ at least in one analysed variable. This means that these items could be detected as outstanding when one would consider only univariate distributions. For the remaining items the multivariate methods used in our study yielded substantial informations needed for revealing their singularity.

The aim of our further analysis of these twelve data vectors is to explain the reasons of the statement that they are "suspicious outliers". We want to distinguish erroneous data vectors and individuals vectors having untypical values of the variables.

5.1. Identification of erroneous values

The "outliers" were checked once more for correctness of the recorded values. It happened that for eight items (nos. 22, 35, 57, 59, 104, 121 and 133) some errors in the data values were found (in Table 5 the erroneous values are underlined).

Table 6

Erroneous and correct values x_{ij} and t_{ij} for the selected items

item	variable	erroneous values		correct values	
no. i	no. j	x_{ij}	t_{ij}	x_{ij}	t_{ij}
22	13	1.81	0.7	0.81	-0.3
35	8	1.18	1.4	0.18	-0.5
57	3	0.95	-1.9	1.95	1.6
58	3	1.00	-1.7	2.00	1.8
59	3	1.00	-1.7	2.00	1.8
104	3	0.18	-4.6	1.18	-1.1
121	13	0.48	-0.7	1.48	0.4
133	12	0.10	1.4	0.01	-0.8

The erroneous and correct values x_{ij} are listed in Table 6 together with the corresponding values t_{ij}. It is quite astonishing that only one of these erroneous values (in the item no. 104) could be detected considering solely univariate statistics. The meaning of the discovered errors will be discussed below.

5.2. Untypical data vectors

One can see from Table 5 that for seven items (nos. 1, 15, 28, 50, 57, 58 and 59) the t-values are higher than 2.0 at least for two variables characterizing the X-ray flare activity. In Table 7 we put together the analysed twelve data vectors (the items with a detected erroneous value are marked by an "E"). The method which revealed the analysed vector to be an outlier is marked by "+" and the method which did not identify the vector as an outlier is marked by "−".

The first four items (nos. 1, 50, 57, 58) are characterized by great number of strong flares described by the variable $x10$ (all these four items have $x10 > 10$ while the mean value $\bar{x}_{.10} = 1.30$). All these items have large values of the

Table 7
The used methods which revealed data vectors suspected to be outliers

item no.	chi2-plot	$\langle a_1, a_2 \rangle$ plane	$\langle a_{13}, a_{14} \rangle$ plane
1	+	+	-
50	+	+	-
57	+	+	+
58	+	+	+
15	+	-	+
28	+	-	+
59	+	-	+
22	+	-	+
35	+	-	+
121	+	-	+
133	+	-	+
104	+	-	-

Mahalanobis distances (D^2). Moreover, these four items and only they were found at the outstanding positions in the (a_1, a_2) plane. The items nos. 15, 28, 57, 58 and 59 correspond to sunspot groups with very strong X-ray flares for the given day with extremely high values of the hardness index ($x12>0.12$ while $\bar{x}_{.12}=0.04$). These items have also large D^2 values, however, they were found in the ($\mathbf{a}_{13}$, $\mathbf{a}_{14}$) plane. Then, they were identified as outliers because high values of the $x12$ cause untypical relations between the variables characterizing flare activity. So, their identification as outliers is fully justified.

A detailed analysis of the past of the sunspot groups described by the seven vectors (the items nos. 1, 15, 28, 50, 57, 58 and 59) demonstrated that all these sunspot groups indicated conspicuously strong flare activity also during several successive days preceding the day for which the data vectors were established. So, these items are truly singular and outstanding because they correspond to the sunspot groups which were flaring untypically strongly.

One should remember, however, that the items nos. 57, 58, 59 had also some errors in the values of the variable $x3$. Nevertheless, these errors are not connected with the variables describing flare activity, and therefore, our conclusions on the untypicallity of these data vectors due to the large sunspot group flare activity are still valid.

The next four items in Table 7 (nos. 22, 35, 121 and 133) have the absolute values of t_{ij} less than 2.0 for all considered variables; for the variables characterizing flare activity the t-values are negative in most cases, which means that these data vectors describe sunspot groups with low flare activity. It follows from the detailed analysis of the past that for these sunspot groups the flare activity was also very low during several preceding days. A low flare activity is typical for the analysed data vectors and therefore these four items should not be identified as outliers. Nevertheless, just these vectors contained some erroneous values (see Table 6) which caused untypical relations between the variables describing flare activity. For instance, for the item no. 133 the error induced large perturbation in the relations between the variable $x12$ (high value – $t=1.4$) and the remaining variables characterizing flare activity (low values – $t<0$). The erroneous value of the variable $x12$ is not very far from the mean value. The chi2-plot method did not reveal at once this error because it was marked by the much greater and serious errors in the item no. 30. However, the principal component method reveals the item no. 133 as an outlier for the erroneous data (Figure 7). After correction of the data for the item no. 30 the four mentioned items can be seen at outstanding positions in the (a_{13}, a_{14}) plane (Figure 6).

The item no. 104 (see Table 5) exhibits different bahaviour. It describes a sunspot group with the values of all variables very near to the mean values – except the value of the variable $x3$ which is faulty. So, the high value of the Mahalanobis distance may be due solely to the mistakenly low value of the variable $x3$. This item, like the faulty item no. 30 was not discovered by the principal component method, i.e. it was not revealed at outstanding positions in the scatterdiagrams constructed from the first two or last two eigenvectors.

6. The impact of one erroneous data vector on the covariance structure

The data considered so far were based on the observations made during the decay phase of sunspot group evolution. We had also analogous data for the increasing phase. Our question was, whether the internal covariance structure for the 14 variables calculated for the increase and decay phase is the same.

Let Σ_1 and Σ_2 denote the covariance matrices for the populations of sunspot groups being in the increase and decay phase, respectively. First we tested the hypothesis H_0: $\Sigma_1=\Sigma_2$. We carried out the calculations using the program UNI1 from the SABA package (Bartkowiak, 1984). This hypothesis was rejected because we obtained the very great value of the statistic *CHI*: $CHI=802.52$ with the number of degrees of freedom $DF=105$ and so, the probability $P(\chi^2>CHI/H_0)<0.00005$. However, there are the mentioned previously errors in the item no. 30 for the decay phase data. After correcting the data for the item no. 30 we obtained:

$CHI=200.44$ with $DF=105$ and $P(\chi^2>CHI/H_o)<0.00005$, what means that as before the hypothesis H_0 should be rejected.

Next step was the testing of the hypothesis that the two matrices Σ_1 and Σ_2 have a common principal component structure with possible differences in the spread along the principal component axes. This hypothesis can be formulated as follows: we seek for an orthogonal $p\times p$ matrix B that diagonalises simultaneously the Σ_1 and Σ_2 matrices.

$$H_c\colon\ B^T\Sigma_1B=\Lambda_1,\ B^T\Sigma_2B=\Lambda_2,$$

where Λ_1 and Λ_2 are diagonal matrices. To test the hypothesis H_c we used

Table 8

Diagonalized covariance matrices Λ_1 and Λ_2 converted to correlation matrices

va ria ble	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
x1	100	20 **20**	4 **-4**	9 **9**	9 **-8**	-15 **15**	9 **-9**	10 **9**	-3 **4**	3 **3**	-0 **3**	11 **-7**	7 **-9**	-10 **-10**
x2	-12 **-12**	100	9 **-9**	5 **5**	15 **-17**	-4 **5**	-13 **13**	7 **7**	5 **-8**	6 **5**	-2 **-7**	-13 **-5**	4 **18**	-3 **-10**
x3	-3 **3**	-6 **6**	100	-5 **5**	9 **9**	-4 **-5**	-2 **-2**	0 **-0**	9 **8**	-4 **4**	-8 **9**	2 **14**	14 **1**	-7 **7**
x4	-6 **-6**	-3 **-3**	3 **-3**	100	-2 **-0**	-1 **1**	-4 **4**	-14 **-13**	-2 **-0**	-2 **-2**	5 **1**	-14 **-10**	9 **11**	2 **-4**
x5	-4 **4**	-7 **7**	-3 **-3**	0 **0**	100	-3 **-3**	2 **2**	4 **4**	-8 **-11**	-4 **3**	-7 **2**	18 **-4**	-3 **9**	2 **-7**
x6	11 **-11**	3 **-4**	4 **4**	1 **1**	6 **6**	100	6 **6**	-8 **8**	-9 **-6**	-2 **1**	-2 **5**	-10 **7**	6 **3**	-11 **8**
x7	-5 **6**	8 **-8**	10 **1**	2 **-2**	-2 **-2**	-1 **1**	100	6 **-6**	-0 **-0**	-1 **-1**	9 **-10**	2 **-9**	-9 **-7**	12 **-7**
x8	-5 **-5**	-4 **-4**	-0 **0**	8 **8**	-3 **3**	3 **-3**	-3 **3**	100	11 **-11**	19 **-18**	1 **-0**	9 **-6**	6 **-11**	-15 **-13**
x9	1 **-2**	-2 **-2**	-4 **-4**	1 **0**	5 **8**	3 **3**	0 **0**	-5 **6**	100	-0 **0**	-3 **-3**	31 **-3**	-1 **2**	10 **2**
x10	-2 **-0**	-3 **-3**	2 **-3**	1 **-1**	3 **-2**	1 **-1**	0 **-0**	-12 **-12**	-0 **0**	100	6 **5**	-2 **-1**	1 **1**	-3 **-5**
x11	0 **-2**	1 **4**	4 **-6**	-3 **-1**	5 **-1**	1 **-2**	-6 **7**	-0 **0**	3 **4**	-4 **-4**	100	-42 **-2**	-2 **4**	-7 **6**
x12	-1 **4**	1 **3**	-1 **-9**	2 **6**	-3 **3**	1 **-4**	-0 **7**	-1 **4**	-2 **2**	0 **0**	-2 **1**	100	10 **-1**	24 **16**
x13	-4 **5**	-3 **-6**	-9 **-1**	-6 **-6**	3 **-6**	-3 **-2**	6 **4**	-4 **6**	1 **1**	-0 **-1**	1 **-2**	11 **1**	100	-5 **2**
x14	4 **4**	1 **4**	3 **-3**	-1 **2**	-1 **4**	4 **-3**	-5 **4**	7 **6**	-5 **-1**	1 **2**	3 **-3**	-69 **-7**	2 **-1**	100

a method proposed by Flury (1984). After carrying out Flury's test we obtained:

$CHI = 360.68$ with $DF = 91$ and $P(\chi^2 > CHI/\mathrm{H_c}) < 0.000005$, what means that the hypothesis $\mathrm{H_c}$ should also be rejected. After removing the error in the item no. 30 from the decay phase data we obtained:

$CHI = 128.84$ with $DF = 91$ and $P(\chi^2 > CHI/\mathrm{H_c}) \cong 0.006$.

One can see the drastic change (decrease) of the values *CHI* as compared with the values $CHI = 360.68$ due solely to the errors in the item no. 30. In Table 8 we show the correlation matrices obtained from the diagonalized matrices Λ_1 and Λ_2 (to illustrate them clearly each element of these matrices was multiplied by 100 and we show only the integer obtained after this multiplication). Transformed correlations for the increase phase of sunspot group evolution are in the lower left part, and for the decay phase they are in the upper right part of the table. The values obtained for the corrected data are given by the bold characters. Nonetheless, the diagonalisation is not complete – particularly remained some peculiarities in the variables $x8$ – $x14$. This might be due, in some part, to the erroneous values of the variables characterizing flare activity of sunspot groups. After removing all errors from the data we obtained:

$CHI = 135.38$ with $DF = 91$ and $P(\chi^2 > CHI/\mathrm{H_c}) = 0.001$.

So, we should reject the hypothesis that the internal covariance structure is the same for the increase and decay phase of sunspot group evolution. The problem needs further examination.

7. Conclusions

The methods employed in this paper allowed us to reveal quite precisely the data vectors displaying various kinds of peculiarity: vectors concealing some errors introduced during the process of the data formation; vectors corresponding to untypical items with unusually big values of the variables; and also vectors with values revealing unusual relations compared with the common relations between the considered variables.

The multivariate methods, which were used here are relatively easy and simple in use, and, as it may be inferred from the present study, they are very efficient in application. It is important that they are more efficient than the methods based on an one-dimensional analysis.

The importance of singular, nonhomogeneous data vectors in the construction of the predicting algorithms needs further investigation. Among others, the question arises, how untypical observational vectors can influence the estimated values of the parameters which appear in the predicting function. For the solution of this problem, the methods applied in the present study do not seem to be sufficient, and therefore, additional methods should be applied.

REFERENCES

Atkinson, A. T., 1987, *Plots, transformations and regression*, Cambridge University Press (second edition).

Bartkowiak, A., 1984, *SABA — An Algol package for statistical data analysis on the ODRA 1305 computer*, University of Wrocław Press.

Bartkowiak, A. and Jakimiec, M., 1986, in *Solar Terrestrial Prediction Proceedings*, P. A. Simon, G. Heckman and M. A. Shea (eds.) p. 285.

Belsley, D. A., Kuh, E. and Welcsch, R. A., 1980, *Regression diagnostics: Identifying influential data and sources of collinearity*, Wiley, New York.

Flury, B. N. 1984, *Common principal components in k groups*, JASA **79**, 892.

Giannuzzi, M. A., 1981, *Astr. Astrophys.*, **103**, 111.

Gnanadesikan, R., Kettenring, J. R., 1972, *Robust estimates, residuals and outlier detection with multiresponse data*, Biometrics **28**, 81.

Jakimiec, M. and Wanke-Jakubowska, M., 1989, *Acta Astr.* **38**, no. 4, (in press).

Jolliffe, I. T., 1986, *Principal component analysis*, Springer, New York, Berlin.

Morrison, D. F., 1967, *Multivariate statistical methods*, Mc Graw Hill, New York.

Pfleiderer, J., 1983, in *Statistical Methods in Astronomy*, ESA, Paris, France.