ACTA ASTRONOMICA
Vol. **40** (1990) pp. 159–168

Atypical Data Vectors and their Influence on Interrelations among Sunspot Group Characteristics

by

M. Jakimiec

Astronomical Institute of Wrocław University, Wrocław, Poland

A. Bartkowiak

Institute of Computer Science, University of Wrocław, Poland

Received September 16, 1989

ABSTRACT

The influence of atypical data vectors on the interrelation structure of fourteen daily characteristics of solar active regions is considered. Seven characteristics describe sunspot groups features and seven X-ray flare activity. We found atypical data vectors in the investigated data set. Some correspond to very strongly flaring sunspot groups, and some reveal non-typical interrelations of the characteristics. The factor analysis carried out for the considered data showed that the interrelation structure evaluated for a reduced data set, (*i.e.* for the data set obtained after removal of some atypical data vectors) remains practically the same.

1. Introduction

Usually, a set of predicting variables describing solar active region or sunspot group features for the given day is the basis for flare activity predictions (see *e.g.* Sawyer *et al.* 1986). The knowledge of the relationships among these variables has in the first place a cognitive meaning. A good knowledge of these relationships permits us to construct statistical models, which, in turn, could be used in the procedure of short-term predictions of solar flare activity. The problem of a statistical description of the relationships of 21 active region characteristics was considered by Jakimiec and Bartkowiak (1986).

In this paper we consider another set of data comprising $p = 14$ daily characteristics of active regions for a given day. The variables are the same

as described by Bartkowiak and Jakimiec (1989). However, now the data set is augmented, *i.e.* we consider sunspot groups both in the increase and in the decay phase of evolution. Fourteen considered variables were employed for the construction of the predicting function by Jakimiec and Bartkowiak (1989). It was found that the data set comprises several unquestionably atypical data vectors (outliers). The non-typicalness is related to very strongly flaring sunspot groups, *i.e.* to unusually large values of the flare characteristics, or to values revealing atypical relations as compared with the general relations between the appropriate variables.

There is a question: what is the impact of such atypical data vectors on the interrelations of the considered characteristics. Are the interrelations of the predicted variables stable, or are they strongly influenced by some non-typical data vectors contained in the data. We will examine whether, after removing some (*e.g.* atypical) data vectors from our data set, we would obtain quite a changed interrelation structure.

To establish the interrelation structure we apply factor analysis. First, we describe how atypical data vectors can be identified. Next, we carry out factor analysis twice: for entire data set (together with atypical data vectors), and for a reduced data set (without the atypical data vectors).

2. The data

The observational data are gathered for the year 1979 from Solar Geophysical Data. We analyze a complex of $p = 14$ variables characterizing sunspot groups of D,E,F Zurich classes. The analysis is carried out separately for sunspot groups in the increase phase (INC) of the evolution (with the positive daily gradient of the area or with no area change) and for sunspot groups in the decay phase (DEC) of the evolution, *i.e.* with negative gradient of the area. The appropriate sample sizes are $N_{INC} = 283$ and $N_{DEC} = 149$

Seven 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). Another seven variables (x8 - x14) describe flare activity of the sunspot group for the given day: maximum value of X-ray 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 1–8 A (Fs), hardness index (HI=Fh/Fs) with (Fh) the daily sum of the X-ray flare fluxes in the wavelength interval 0.5-4 A, and the daily maximum value of the six-hour hardness index (maxh). Because of very skew distributions of the variables, for the variables $x1, x2, x7, x8$, and $x11$ the logarithmic transformation $X' = \log X$ was performed, and for the variables $x3$ and $x13$ the transformation $X' = \log X + 2$ was applied.

3. Identification of atypical data vectors

We use here methods similar to employed by Bartkowiak and Jakimiec (1989), *i.e.* χ^2-plots constructed from Mahalanobis distances D^2, and also scatter-diagrams constructed from the first two and last two principal components. These methods are described *e.g.* by Gnanadesikan and Kettenring (1972). Bartkowiak and Jakimiec (1989) presented in more detail the process of search, identification and explanation of atypical data vectors in the data set refering to sunspot groups in the decay phase of the evolution (DEC). At present, we employ similar methods for a data set describing sunspot groups in the increase phase (INC). We construct two χ^2-plots: (a) for all $N_{INC} = 234$ data vectors (items), and (b) for a reduced data set obtained from the original data set after removing $k = 18$ items with largest Mahalanobis distances D^2. These two χ^2-plots are presented in Figs. 1a and 1b respectively. In Fig. 1a we see clearly one conspicuous outlier and a dozen or so further items with large Mahalanobis distances. After removing these data vectors from the data we construct another plot (Fig. 1b), with a linear pattern. It means, that the reduced data set is homogeneous and does not contain atypical data vectors. Then we construct scatter-diagrams: (a) from the first two (PC1,PC2) and (b) from the last two (PC13,PC14) principal components calculated from the covariance matrix of the 14 considered characteristics. The scatter-diagrams are shown in Figs. 2a and 2b. It is believed that the first few principal components are able to reveal outliers, which are atypical in range of the values of some variables, while the last few principal components reveal outliers which are atypical in interrelations of some considered variables. From the components of eigenvectors being the base for constructing principal components (not shown in this paper) we can infer that the first two principal components (*i.e.* PC1 and PC2) reveal, first of all, the non-typicalness appearing in variables $x9, x10$ and $x13$, which describe the number of flares and the total hard X-ray flux. The last two principal components (*i.e.* PC13 and PC14) reveal mainly the nontypicalness in the variables $x12$ and $x14$, which describe the hardness indices. Every point located in the scatterdiagrams (a) and (b) far from the bulk of points was previously revealed in the χ^2-plot (Fig. 1a) as an item with a great Mahalanobis distance. However, not all data vectors with great Mahalanobis distances are revealed as outliers in the scatter-diagrams (PC1,PC2) or (PC13,PC14). This is due to the fact that these items may be outliers in other variables than those describing flare activity. Most of the data vectors, revealed as atypical, similarly to the result found by Bartkowiak and Jakimiec (1989), describe strongly flaring sunspot groups (more than ten flares), and also describe sunspot groups with non-typical relationship between variables.

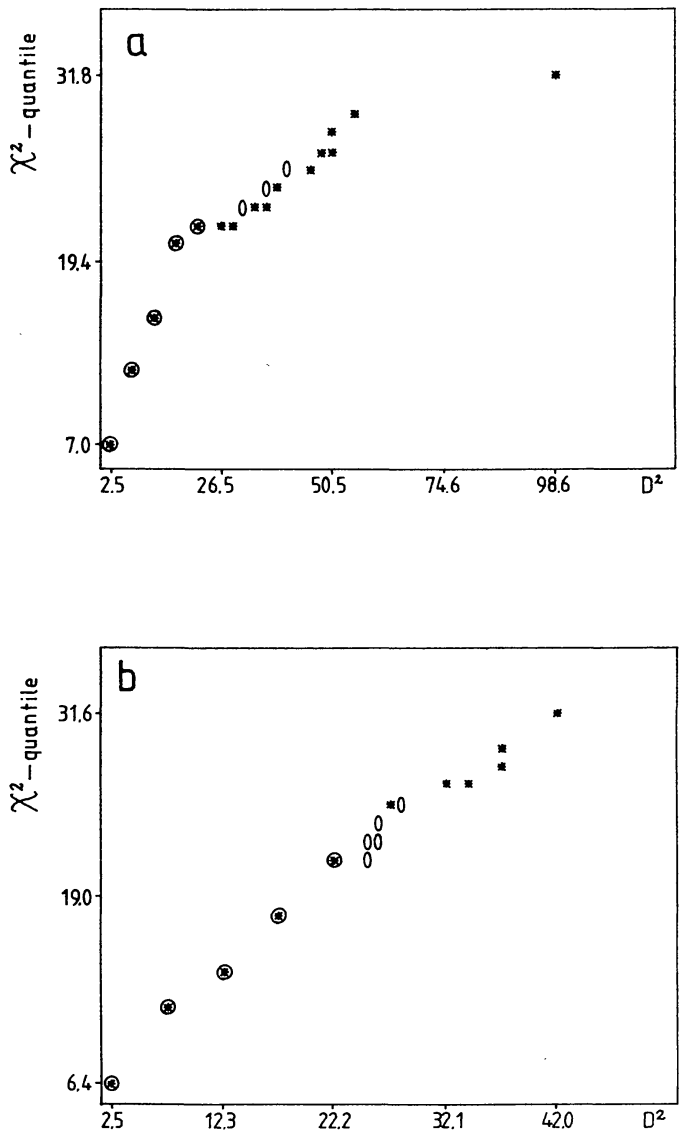

Fig. 1. χ^2-plots of Mahalanobis distances D^2 obtained (a) for entire data set and (b) after removing 18 items with the largest D^2 values. Overlaping points are marked by a "zero" sign. Circled points show five χ^2 quantiles for the five D^2 classes containing the main part of the data vectors.

4. Factor analysis

We employ the factor analysis method in order to discover some common factors explaining the interrelation structure of the considered p characteristics. We carried out factor analysis according to principles presented *e.g.* by Morrison (1967). There are two main points: how many factors are needed to explain the correlation structure, and what is the meaning of the obtained factors.

The calculations are performed for four data sets, two of which are the original data describing the sunspot groups in the increase (INC) and decay

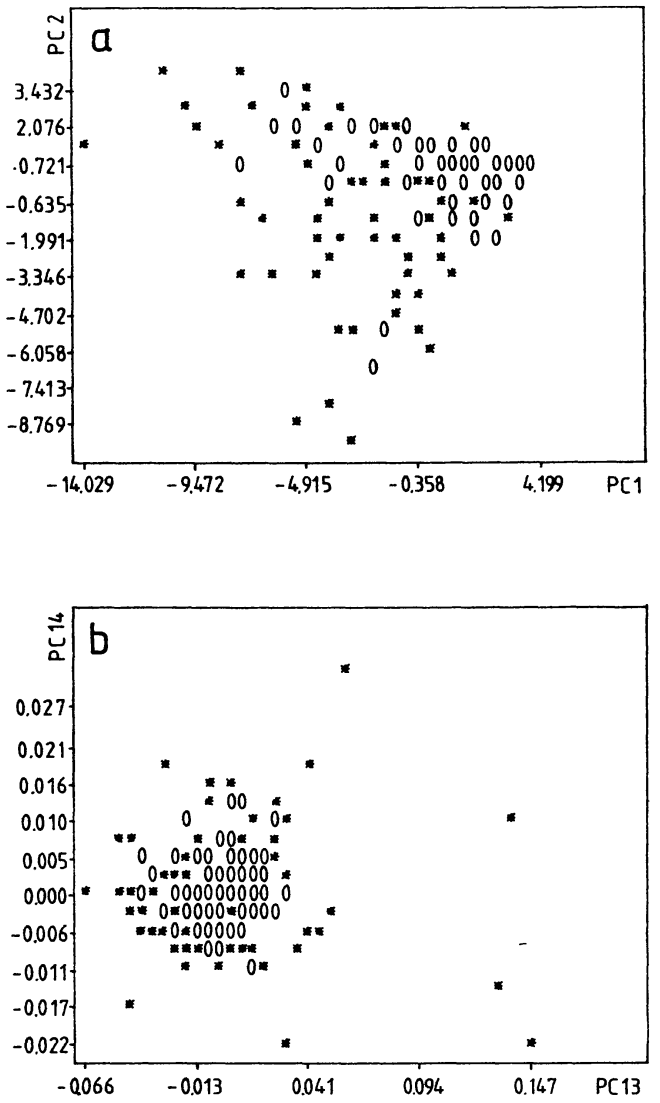

Fig. 2. Scatterdiagrams constructed (a) from the first two principal components (PC1,PC2), and (b) from the last two principal components (PC13,PC14).

(DEC) phase of the evolution. In the INC data set eighteen (see Section 3), and in the DEC data set fourteen (Bartkowiak and Jakimiec 1989) atypical items were identified. We remove from the INC fourteen and from the DEC seven very conspicuous non-typical data vectors. So, we obtain two reduced data sets, INC-R and DEC-R, respectively. For all the four data sets the factor analysis was performed, considering their correlation matrices.

To answer the question how many factors should be considered, we investigate the exhaustion of the correlation matrices by progressively augmented number of factors. It is known that the extracted factors should reproduce all correlations between the considered variables, *i.e.* should reproduce the correlation matrix R of these variables. Suppose now, we extracted m $(m < p)$ factors only. These factors reproduce only a part of the matrix R, say $R^{(m)}$, and the other part remain unexplained.

$$R = R^{(m)} + V^{(m)} \tag{1}$$

The goodness of approximation of R by $R^{(m)}$ can be measured by a mathematical index called "trace of a matrix". For the decomposition given by Eq. (1) we have:

$$\text{trace}(R) = \text{trace}(R^{(m)}) + \text{trace}(V^{(m)}) \tag{2}$$

When R is a correlation matrix, we have trace(R)=p, where p is the # number of considered variables. It can be shown that (*e.g.* Morrison, 1967):

$$\text{trace}(R^{(m)}) = \sum_{i=1}^{p}\sum_{j=1}^{m} l_{ij}^2 = \sum_{i=1}^{p} com(i,m) \tag{3}$$

where l_{ij} is the loading of the $j-$th factor in the $i-$th variable The sum $\sum_{j=1}^{m} l_{ij}^2$ (see Eq.3) is called "communality" of the $i-$th variable with m factors. The element $v_{ii}^{(m)}$ of the matrix V^m is called "specificity" of the $i-$th variable.

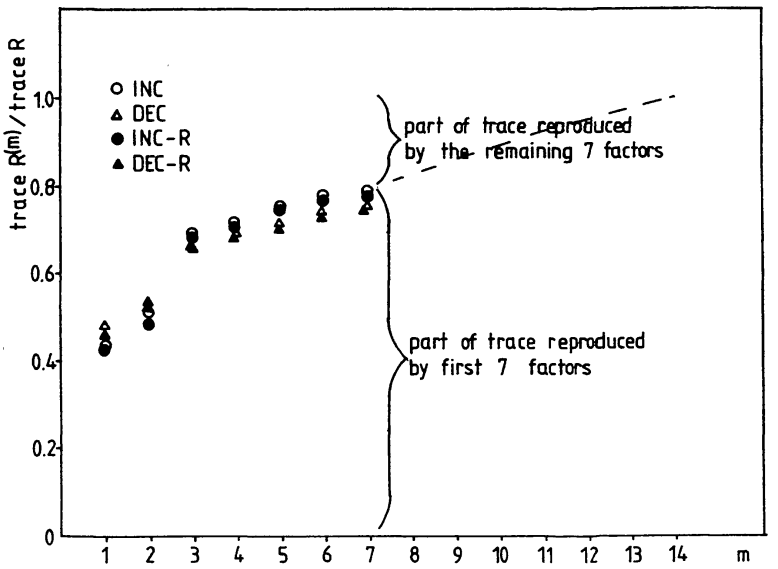

Fig. 3. Exhaustion of trace(R) by progressively augmented number m of factors.

Putting $y =$ trace($R^{(m)}/p$) against m we obtain a plot visualizing the speed of exhaustion of trace(R) by progressively augmented number m of factors. For $m = p$ we should obtain $y = 1$ meaning a total exhaustion of the matrix R,*i.e.* a total reproduction of this matrix by the extracted factors. The process of exhaustion of trace(R) by the sums of the communalities $\sum_{i=1}^{p} com(i,m)$ with increasing m is shown in Fig. 3, for $m = 1,2,...,7$. One can see that the first two factors account for nearly 50% of the whole

trace(R). The contribution of subsequent factors (for $m = 3,4,5,6,7$) is much smaller. We decided to stop calculations with $m = 7$ factors, which reproduce the traces of the the appropriate matrices R evaluated for the data sets INC, INC-R, DEC, DEC-R in 78.54%, 77.42%, 76.02%, 74.91%, respectively. One can see that the pattern of these plots is very similar for all the mentioned data sets. The factor loadings for the extracted 7 factors were subjected to a rotation varimax. The absolute values of the rotated loadings $|\widetilde{l_{ij}}|$ (where $i = 1,2,...,14$ is a number of the variables, and $j = I, II, ..., VII$ is a number of the factors) are shown in Fig. 4a (for the INC and INC-R data sets), and in Fig. 4b (for the data sets DEC and DEC-R). In each figure the loadings obtained from the whole data set are put together (parallel bars) with those obtained for the reduced data sets. The specificities $v_{ii}^{(7)}$, are plotted on the top of Figures 4a and 4b. In Table 1 the percentages of the total variance, explained by each of the seven common factors, are given for the considered data sets. The sum shows the percentage of the total variance explained by seven factors. One can see from the Table 1 and from Figs. 4a and 4b that the pattern of the factor loadings, obtained for the entire data sets and for the reduced sets is very similar.

Jakimiec and Bartkowiak (1986) ascribe some meaning to the obtained factors. A similar meaning may be ascribed to the factors given in Figs. 4a and 4b:

(a) the first factor (I) explaining about 30-35% of the total variance (TVar) contains mainly the characteristics of strong flare activity ($x8, x10-, x11, x12, x13, x14$), which are poorly correlated with sunspot group characteristics;

(b) the second factor (II) - about 8% of the TVar - explains the interrelations between the characteristics of faint flares, *i.e.* this factor explains the contribution of faint flares ($x9$) to the total flare fluxes ($x11, x13$);

(c) the interrelations of the variables describing the sunspot group features reveal in three factors: the factor III contains the variables from $x1$ to $x7$ (this factor explains about 20% of the TVar); the factor V (3-4% of the TVar) explains the correlation between the McIntosh class ($x1$) and the sunspot group area ($x2$); the interrelations of the variables $x1, x2, x3$ and $x4$ are explained also by the factor VI (5-7% of the TVar). The factor VI shows that the calcium plage characteristics ($x3, x4$) are correlated more strongly for sunspot groups in the decay (Fig. 4a) than in the increase (Fig. 4b) phase of the evolution. Similarly, factor VII shows for the decay phase stronger correlations between the magnetic class ($x5$) and the characteristics of flare activity ($x10, x11, x13$). For the decay phase the factor VII explains only 1% of the TVar.

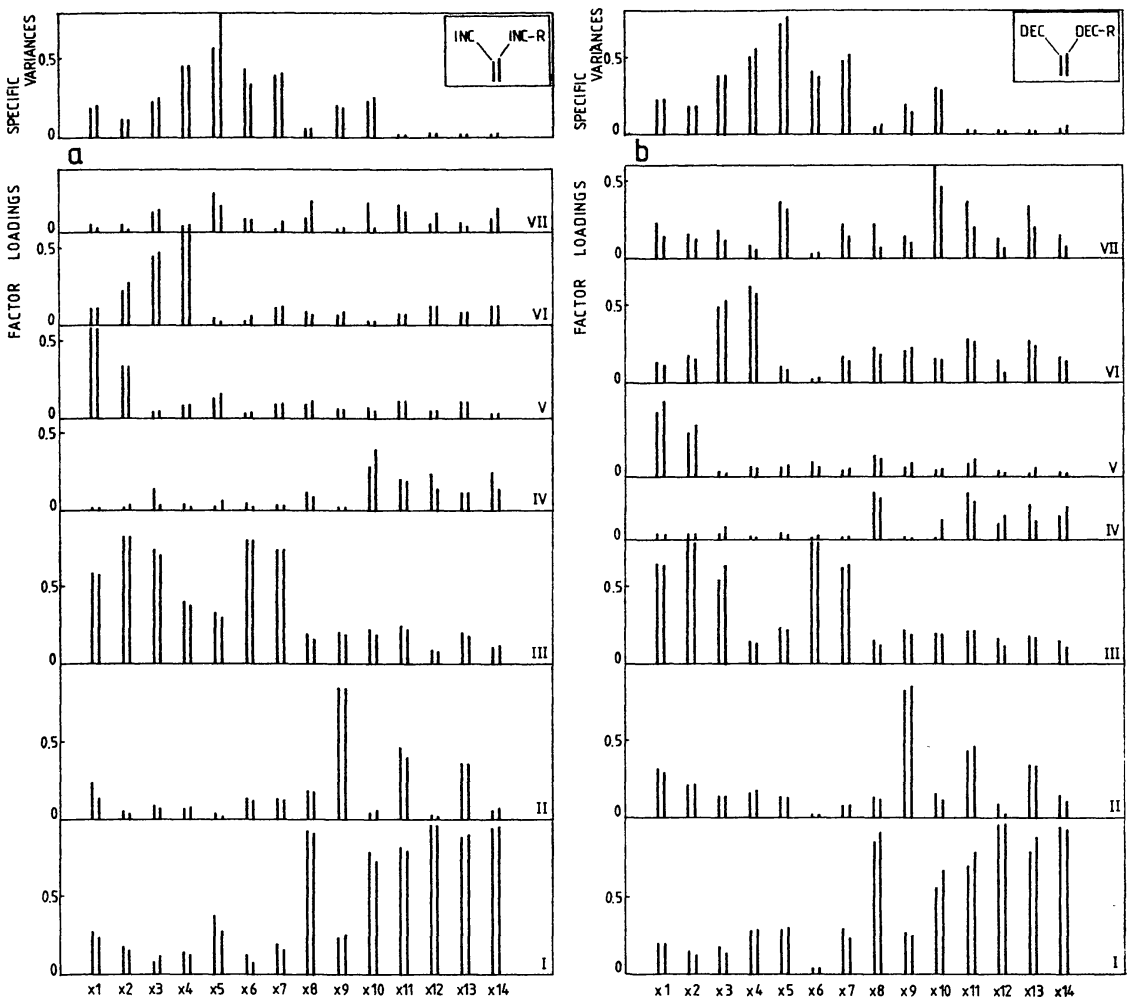

Fig. 4. Factor loadings $|l_{ij}|$ and specific variances $v_{ii}^{(m)}$ obtained for $m = 7$ factors (a) for the data sets INC and INC-R, and (b) for the data sets DEC and DEC-R. The loadings obtained for the whole data set are put together with those obtained for the reduced data set (parallel bars).

5. Conclusions

There are two questions of special interest. First, how many atypical data vectors are contained in the data sets describing sunspot groups in the

Table 1

The percentages of the total variance explained by the common factors.

factor j	data sets INC	INC-R	DEC	DEC-R
I	30.6	33.8	36.2	34.7
II	8.9	9.0	8.2	8.4
III	18.7	18.1	23.0	21.9
IV	1.9	1.6	2.0	2.0
V	2.1	2.8	3.7	4.0
VI	7.1	6.4	4.6	5.4
The sum	76.1	74.9	78.6	77.5

increase (INC) and in the decay (DEC) phase of the evolution. Using two methods: χ^2-plots, constructed from the Mahalanobis distances, and scatterdiagrams constructed from the principal components, we found fourteen atypical items in the INC data set and eighteen atypical items in the DEC data set. Some of the atypical data vectors describe very strongly flaring sunspot groups, and some describe sunspot groups with non-typical interrelations of the considered variables.

The second question we were interested in is how strongly the revealed atypical items may influence the interrelation structure of the considered variables. To establish this structure we used the factor analysis method. From Figs. 3, 4a and 4b and from Table 1 we can conclude that about 75-80% of the total variance is explained by seven common factors. One can see also that there are no pronounced differences betweeen the factor loadings calculated for the entire data sets (INC or DEC) and for the reduced data sets (INC-R or DEC-R). The differences between the factor loadings obtained for sunspot groups in the increase (INC) and in the decay (DEC) phases of the evolution seem to be more pronounced. These differences will be discussed more comprehensively in another work. Fact, that the interrelation structure is, in principle, defined by the bulk of the data vectors means that the intrinsinc structure is sufficiently stable inspite of the atypical data vectors contained in the data sets.

REFERENCES

Bartkowiak, A., and Jakimiec, M. 1989, *Acta Astr.*, **39**, 85.
Gnanadesican, R., and Kettenring, J.R. 1972, *Biometrics*, **28**, 81.
Jakimiec, M., and Bartkowiak, A. 1986, *Solar-Terrestrial Prediction* , P.A.Simon, G.Heckman and M.Shea eds.(Boulder, USA), 294 .

Jakimiec, M., and Bartkowiak, A. 1989, *Acta Astr.*, **39**, 257.
Morrison, D.F. 1967, *Multivariate Statistical Methods*, (McGraw-Hill, New York).
Sawyer, C., Warwick, J.W., and Dennett, J.T. 1986, *Solar Flare Prediction*, (Colorado Associated University Press Boulder, USA).