ACTA ASTRONOMICA
Vol. **41** (1991) pp. 61–71

A Multivariate Comparison of Delta and of Other Magnetic Class Sunspot Groups

by

M. Jakimiec and J. Paciorek

Astronomical Institute, University of Wrocław, Poland

A. Bartkowiak

Institute of Computer Science, University of Wrocław, Poland

Received August 10, 1990

ABSTRACT

Multivariate statistical analysis of sunspot group characteristicsis is applied to visualize the differences between sunspot groups of δ and of other magnetic configurations. We try to visualize graphically these differences through the projection of points (sunspot groups) from multivariate space of variables onto a two-dimensional plane. We found that flare activity characteristics can discriminate distinctly the δ sunspot groups from other ones. Investigating multivariate regression functions estimated separately for sunspot groups of δ and of other magnetic classes we do not get significant differences – the hypotheses that the regressions estimated for sunspot groups of δ and of other magnetic classes are parallel or equal can not be rejected. It means that for predictions of flare activity the same prediction function may be used.

1. Introduction

It is generally known that in the sunspot groups with δ magnetic configuration (δ-sunspot groups) very strong flare activity appears. But it is also known that the δ configuration alone is not sufficient for the flare appearance. Priest (1984) emphasizes that the occurrence of large degree magnetic complexity is the main condition for strong flare activity. Zirin and Liggett (1987) accentuate that great flares occure only in sunspot groups of the δ configuration. However they also found that some of δ-sunspot groups, those formed by collision between two separate but growing bipolar groups, are not strongly flaring.

In short term predictions of solar flare activity one of the variables characterizing magnetic field configuration, the magnetic class, is usually introduced to

the analysis in the form of discrete variable. Our intent in this study is to visualize graphically the differences between the sunspot groups of δ and of other (α, β, $\beta\gamma$, γ) magnetic classes employing multivariate statistical analysis of the variables characterizing flare activity. Further, we will investigate if differentiation with respect to magnetic classes is connected also with differences in the prediction functions.

2. The data

The full set of data was previously described by Jakimiec and Bartkowiak (1990a). We analyze here the complex of 15 daily characteristics of sunpot groups of D, E, F Zurich classes. The first seven variables X describe characteristics as follows: $x1 - McI$, McIntosh class; $x2 - A$, area of sunspot group, $x3 - CaA$; calcium plage area; $x4 - CaI$, calcium plage intensity; $x5 - M$, magnetic class; $x6 - H$, magnetic field strength; $x7 - MFI$, magnetic field index. The further six variables describe sunspot group flare activity. They are as follows: $x8 - maxX$, the maximum value on a given day of flare X-ray flux for the sunpot group; $x9 - NFF$, the number of faint flares per day; $x10 - NSF$, the number of stronger flares per day; $x11 - Fs$, the total flare flux (1–8 Å); $x12 - HI$, hardness index, *i.e.* the quotient of Fh and Fs values; $x13 - Fh$, the total flare flux (0.5–4 Å). As predicted variables Y we will consider the flare activity on the next day: $y1 - Fs'$ and $y2 - Fh'$.

We have included into our data set only those sunspot group observations for which the full set of 15 characteristics can be completed. The sample size is $n = 383$. The frequency distributions of most of the analyzed variables reveal very high skewness. Therefore, for further analysis we use the values obtained after the logarithmic transformation: $X \Rightarrow \log X$ for the variables $x1$, $x2$, $x3$, $x7$, $x8$, $x11$, $y1$, and $X = \log X + 2$ for the variables $x13$ and $y2$.

The whole data set was divided into two parts: set I – the sample containing sunpot groups of α, β, $\beta\gamma$ or γ magnetic classes, (*i.e.* the items in which $x5 < 5$); set II – the sample containing the δ sunspot groups, (the items with $x5 = 5$). The sample sizes are: $n_I = 332$, $n_{II} = 51$. In the following we are going to analyze the differences between these two data sets, I and II, taking into account the introduced 15 variables.

3. Statistical comparison of sunspot groups of δ and other magnetic configuration

We perform the statistical analysis in four steps: first, we compare the data sets I and II analysing separately the individual characteristics of sunspot groups; second, we investigate the differences between the data sets I and II, considering all flare characteristics together by applying multivariate methods; third, a graphical

visualization of the differences is presented; lastly, we compare the prediction functions estimated for the two data sets (I and II).

3.1. *Univariate analysis of sunspot group characteristics*

The mean values ($\bar{x}$) and variances (s^2) obtained for the two data sets I and II are shown in Table 1. One can see that all the means in the data set II are greater than in the data set I. We test the hypotheses that both the mean values and the variances in two data sets are equal. We verify these hypotheses using the t-Student and the F-Snedecor statistics. We carry out the calculations using the program UNI1 from the SABA package (Bartkowiak, 1984). The critical values at the significance level $\alpha = 0.01$ are $t_{0.01} = 2.58$ and $F_{0.01} = 1.74$ (when $s_I^2 > s_{II}^2$) or 1.59 (when $s_I^2 < s_{II}^2$). Bold numbers in Table 1 mean that the values of statistics F or t are greater than the appropriate critical values, *i.e.* the variance or the mean value calculated for the data set I is significantly different from those calculated for the data set II.

Table 1

Means ($\bar{x}$) and standard deviations (s) evaluated for data sets I and II. F and t are statistics used for testing equality of variances and equality of means in both subsets. Bold fonts indicate means and standard deviations for which $t > t_{0.01}$ and $F > F_{0.01}$, respectively.

Variable	Subset I, $n_I = 332$ $\bar{x}_I$	s_I	Subset II, $n_{II} = 51$ $\bar{x}_{II}$	s_{II}	F*	t**
x1 - McI	1.56	0.27	**1.82**	0.27	1.02	6.28
x2 - A	2.27	0.38	**2.63**	0.40	1.11	17.20
x3 - CaA	1.41	**0.31**	**1.62**	0.22	1.86	6.06
x4 - CaI	0.49	0.06	**0.52**	0.05	1.51	4.70
x5 - M	2.19	0.62	5.00	0.00	-	-
x6 - H	3.92	**0.82**	**4.45**	0.58	2.04	5.75
x7 - MFI	0.98	0.47	**1.48**	0.37	1.61	7.25
x8 - maxX	0.35	0.42	**0.87**	**0.68**	2.63	5.26
x9 - NFF	3.18	2.93	**4.76**	3.40	1.34	3.16
x10 - NSF	0.93	1.78	**3.06**	**3.02**	2.89	4.70
x11 - Fs	0.61	0.56	**1.23**	0.66	1.41	6.37
x12 - HI	0.03	0.03	**0.07**	**0.07**	3.69	4.40
x13 - Fh	0.96	0.86	**1.94**	0.99	1.33	6.70
y1 - Fs'	0.56	0.59	**1.08**	0.63	1.14	5.54
y2 - Fh'	0.88	0.90	**1.67**	0.99	1.20	5.38

* critical values: $F_{0.01} = 1.74$ (when $s_I > s_{II}$) or 1.59 (when $s_I < s_{II}$),

** critical value $t_{0.01} = 2.58$.

One can see from Table 1 that the variances of some variables ($x3 - CaA$, $x6 - H$) are markedly smaller for the δ sunspot groups. However, the variances

of some variables characterizing flare activity ($x8 - maxX$, $x10 - NSF$ and $x12 - HI$) are visibly greater for the δ sunspot groups than for sunspot groups of other magnetic classes. At the same time, all mean values calculated for the δ sunspot groups are significantly greater than those calculated for the I data set. The result confirms the commonly known fact that the characteristics of δ sunspot groups have in general higher values and higher flare activity than the others. The differences between the I and II data sets are statistically significant.

3.2. *Multivariate analysis of the differences between the two data sets*

Now, investigating the differences between the I and II data sets of sunspot groups we take into account only the characteristics of flare activity. Three following variants of the variable subsets are considered:
A: all variables characterizing flare activity on the given day $\{x8, x9, x10, x11, x12, x13, x14\}$;
B: variables characterizing only high flare activity, *i.e.* as in variant A, but without the variable $x9$;
C: variables characterizing the lower flare activity $(x9, x11, x13)$.

In first place we take into account global differences between the δ sunspot groups and sunspot groups of other magnetic classes calculating the Mahalanobis distances (D^2) between two data sets, I and II. This calculation is done separately for the sets of variables denoted as variants A, B and C. The obtained values D^2 are shown in Table 2. One can see that the values D^2 evaluated for the three subsets of variables (variants A, B, and C) are similar, especially those obtained for the variants A and B. Generally, the obtained distances are not big. Next we perform a step-wise search of subsets of variables yielding the largest values of D^2 (the "best subset"). The found subsets of variables are shown also in Table 2, together with the respective values D^2 evaluated for them. One can see that the Mahalanobis distances evaluated for the full sets (A), (B) or (C) and also for the "best subset" of these variables are very near. Assuming that the data sets I and II both are sampled from a p dimensional normal population we test the null hypothesis that the sunspot groups of δ and of other magnetic classes have the same expected values. If this hypothesis is true, then the statistics

$$F = \frac{n_I \cdot n_{II}(n_I + n_{II} - p - 1)}{(n_I + n_{II})(n_I + n_{II} - 2)} D^2$$

has the F-Snedecor distribution with p and $(n_I + n_{II} - p - 1)$ degrees of freedom. In the formula above n_I and n_{II} are the sample sizes and we take for p in turn 7, 6, 3, the number of variables considered in the variants A, B, C, respectively. Using the statistic F we check whether the variables $x8 - x13$ can discriminate with statistical significance sunspot groups of δ from other magnetic classes. In Table 2 we show F_c (the calculated values of the F statistics) and $P(F > F_c/H_0)$ (the probabilities that the random variable F exceeds the calculated values F_c, when

the null hypothesis is true). One can see that the P-values are very small. That means that the null hypothesis (stating the same expected values of the variables in the two data sets, I and II) should be rejected. We can say that the considered variables have a statistically significant discriminative power. One can see from Table 2 that the best discriminating variables are: $x10$ (NSF, the number of strong flares) and $x13$ (Fh, the Total Flare Flux, 0.5–4 Å). That means that, among all, these two variables contain the much important information discriminating the δ sunspot groups from sunspot groups of the other magnetic classes.

Table 2

Discriminative power of the variables describing flare activity. D – Mahalanobis distance between the subsets I and II, F_c – statistic testing the significance of the discrimination, H_0 – null hypothesis stating that the expected values of the considered variables are the same in the data sets I and II.

Variant	Mahalanobis distance D^2	Statistic F_c	$P(F>F_c/H_0)$
A: {x8,x9,x10, x11,x12,x13,x14}	1.472	9.15	<0.0005
the best subset {x10,x13}	1.399	30.85	<0.0005
B: {x8,x10,x11, x12,x13,x14}	1.459	10.61	<0.0005
the best subset {x10,x13}	1.399	30.85	<0.0005
C: {x9,x11,x13}	1.260	18.54	<0.0005
the best subset {x13}	1.250	55.39	<0.0005

3.3. *Graphical visualization of the differences between the data sets I and II*

We have found in previous section that the differences in variables $x8 - x14$ for the δ and the other magnetic classes are statistically significant. Now we will illustrate these differences by a graph in a plane. We employ here a method proposed by Bartkowiak *et al.* (1987). We seek for a projection of item-points (located in the multivariate space of variables) onto such a plane that the point projections belonging to different data subsets are separated as much as possible. The first projection vector is the solution of the matrix equation

$$(B - \lambda W)\mathbf{a} = \mathbf{0}$$

where B and W are the matrices of covariances between data subsets and within data subsets, respectively (Rao, 1973); λ and $\mathbf{a}$ are to be found. The vector $\mathbf{a}$

is the required projection vector, yielding the first discriminant variate ($z1$). The second variate ($z2$) is constructed as orthogonal to the first one and also as having the largest variance. So, we obtain for each item-point (sunspot group) from the multivariate space of variables two values ($z1$, $z2$) which can be considered and visualized as coordinates of a point located in a plane. We made separately the projections for the three considered sets of variables described above as the variants A, B and C. The three projections look much the same, and therefore we show in Fig. 1 only the scatterdiagram of the projection points ($z1$, $z2$) obtained for the variant A, *i.e.* the projections from the 7-dimensional space R^7 of the variables $\{x8, x9, x10, x11, x12, x13, x14\}$. The projections of point-items belonging to the data set I are marked by bars, multiple points from this data set (*i.e.* when more than one point from R^7 gave the same projection) are marked with a "+" sign. Similarly, projections of points-items belonging to the data set II are marked by circles, and multiple points by dark points. From Fig.1 one can see that both the bars and circles are intermixed, although, the projection of points from the II data set are more clustered at the left side of the scatterdiagram, while the points from the I data set (and also the "plus"es) are more dense in the central and right part of the scatterdiagram.

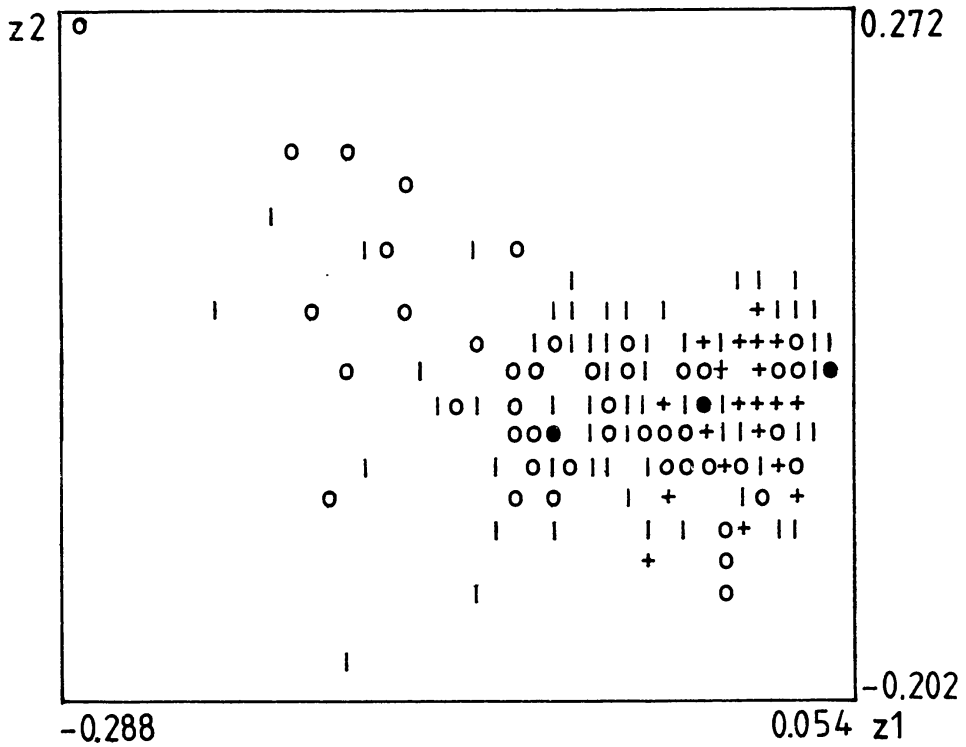

Fig. 1. The scatterdiagram of the projection points ($z1$, $z2$) obtained for the variant A. Bars and circles mean items belonging to the data sets I and II, respectively. The plusses and dark points mean the multiple points.

To exhibit more clearly the differences between the I and II data sets we subdivided the $z1$ axis into seven parts called in the following "septiles" (the

subdivision was performed on the basis of the histogram constructed from the "$z1$" values, for the two data sets, I and II). Next, the points belonging to subsequent septiles were counted separately for the I and II data sets. In such a way we obtained two frequency distributions of points in the fixed septiles. They are shown in Fig. 2. One can see that the frequency distribution obtained for the set I is nearly uniform, except the first septile, which has a visibly lower frequency. On the other hand, the frequency of counts made for the set II reveals a high pick in the first septile and rather low values in the last four septiles.

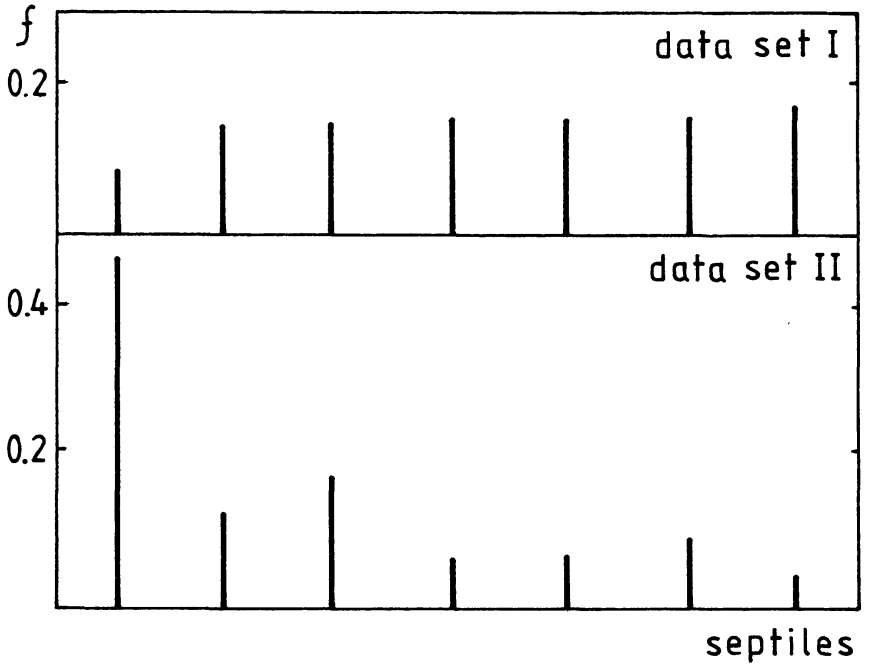

Fig. 2. Two frequency distributions (f) of points in the fixed septiles (x-axis)

To find out what have in common (with respect to the characteristics of flare activity) the points (sunspot groups) located at extreme position on left side in Fig. 1 we looked back at the observed values of the variables $x8 - x14$ for these items. They are shown in Table 3. Each value was standardized (*i.e.* from each value the appropriate mean value calculated for the given data set was subtracted and the obtained difference was divided by the appropriate standard deviation). The values, that after standardization are in absolute value greater than 2.0, are underlined. One can see that each of the items listed in Table 3 has at least one value underlined. From that we infer that points located at the left side of Fig. 1 (negative values of $z1$) correspond to strongly flaring sunspot groups. The items identified as atypical by Bartkowiak and Jakimiec (1989) and by Jakimiec and Bartkowiak (1989) are marked by *. As much as 19.6% (10/51) of the considered data vectors describing the δ sunspot groups were revealed as atypical ones while only 3.3% (11/332) of data vectors from the I data set were in the cited papers revealed as atypical. One can also see from Table 3 that just as the II data set also

the I data set contains a number of strongly flaring sunspot groups. However, their relative proportion to the respective data set size is much larger for the second data set. The magnetic classes of these strongly flaring sunspot groups in the I data set (the upper part of Table 3) are various; three of them are the γ class, five are the $\beta\gamma$ class and two are the β class.

Table 3

Values of the variables $x8 - x13$ for the items located in Figure 1 at extreme left position. Underlined are the values which after standardization are greater in absolute values than 2.0, * means the item identified in former analysis as atypical.

Subset	Item	x8	x9	x10	x11	x12	x13
I	5*	1.30	3	8	1.68	0.12	2.78
	116	1.30	5	7	1.82	0.10	2.82
n_I=332	189*	1.30	3	6	1.61	0.21	2.93
	196	1.30	2	6	1.73	0.09	2.66
	205*	2.00	8	12	2.32	0.08	3.21
	220	2.30	7	2	2.33	0.09	3.30
	229*	2.11	9	7	2.30	0.24	3.68
	247	1.30	2	6	1.76	0.10	2.77
	251*	1.85	7	12	2.06	0.20	3.36
	252*	1.70	11	5	1.95	0.26	3.37
II	11	2.18	4	6	2.25	0.20	3.49
	13*	2.51	5	10	2.58	0.33	4.11
n_{II}=51	23	1.48	2	4	1.60	0.14	2.74
	24*	2.00	6	4	2.08	0.26	3.49
	25*	2.08	4	9	2.20	0.14	3.34
	26*	0.90	3	8	1.52	0.05	2.25
	30*	2.00	9	6	2.23	0.10	3.18
	31	2.11	4	5	2.18	0.14	3.32
	32*	2.34	3	2	2.38	0.18	3.62
	38*	0.85	10	11	1.67	0.05	2.34
	41*	1.40	1	12	1.83	0.12	2.92

3.4. Comparison of prediction functions for the I and II data subsets

In the former section we found that flare activity characteristics can in some degree distinguish the δ sunspot groups from other magnetic class sunspot groups. Now we are going to investigate whether these differences between the data sets I and II will be reflected also in the prediction functions. We consider the linear regression:

$$Y = b_0 + \sum_{i=0}^{p} b_i x_i + e$$

for two predicted variables taken in turn: $Y = y1\ (Fs')$ and $Y = y2\ (Fh')$. The regression coefficients b_i $(i = 0, 1, ..., p)$ are estimated separately for each of two predicted variables and separately for the I and II data sets. The data sets I and II are very unbalanced (their sizes are $n_I = 332$, $n_{II} = 51$), therefore, the estimated regression would be established in high degree by items from the first data set. To avoid that we choose randomly from the data set I three subsets (III, IV, V), each of them of the size equal to 51. For these three additional data subsets the appropriate regression coefficients b_i are estimated too. In Table 4 the estimated for all data sets I – V the regression coefficients $b(i)$ (for the predicted variable $y1 = Fs'$) are given together with their standardized values $t(i)$. In the bottom line of the Table 4 the values of the multiple correlation coefficients RR are given. At first sight the differences between the corresponding values of the regression coefficients seem to be high.

Table 4

The regression coefficients $b(i)$ and their standardized values $t(i)$ obtained for the predicted variable $y1$ (Fs) from the data sets I – V.

	Data set I		Data set II		Subset III		Subset IV		Subset V	
i	b(i)	|t(i)|	b(i)	|t(i)|	b(i)	|t(i)|	b(i)	|t(i)|	b(i)	|t(i)|
0	-0.697	-	-0.022	-	-1.313	-	-0.731	-	-0.804	-
1	0.151	1.00	-0.246	0.50	-0.306	0.67	0.287	0.80	0.365	0.56
2	0.107	0.82	0.668	1.91	0.741	2.19	0.276	0.75	0.836	1.47
3	0.165	1.30	-0.485	1.16	0.742	1.94	0.319	0.82	-0.703	1.40
4	1.049	1.88	1.221	0.64	-0.456	0.35	1.172	0.74	-0.656	0.24
6	-0.100	2.26	-0.207	1.30	-0.100	1.00	-0.355	2.63	-0.072	0.33
7	0.170	2.12	0.051	0.23	-0.016	0.09	0.459	2.07	-0.161	0.44
8	-0.154	0.75	0.526	1.07	-0.023	0.05	-0.054	0.07	-0.873	0.98
9	0.025	1.64	-0.005	0.17	0.001	0.04	0.009	0.15	-0.034	0.58
10	0.082	3.45	0.038	1.05	0.135	1.89	0.033	0.26	0.049	0.32
11	0.138	0.50	-1.048	1.27	-0.080	0.09	0.490	0.61	2.030	1.57
12	-0.198	0.11	-7.102	2.13	-8.666	0.81	6.848	1.24	24.800	1.57
13	0.076	0.44	1.050	1.92	0.386	0.55	-0.442	0.86	-0.970	0.80
RR	0.383		0.566		0.667		0.468		0.322	
n	332		51		51		51		51	
category	non-δ		δ		1 subsample from I		2 subsample from I		3 subsample from I	

Then we compare the regression obtained for the data set II with the regressions obtained for the data set I and for the data subsets III, IV and V. We consider two null hypotheses:

(1) the hypothesis that the regressions are parallel, *i.e.*

$$H_0^{(1)}: \quad b_1^{(II)} = b_1^{(U)}, \quad b_2^{(II)} = b_2^{(U)}, \quad \ldots, \quad b_{13}^{(II)} = b_{13}^{(U)}$$

where for U we take in turn the I, III, IV and V data subsets, and
(2) the hypothesis that the regressions are equal, *i.e.*

$$H_0^{(2)}: \; b_0^{(II)} = b_0^{(U)}, \; b_1^{(II)} = b_1^{(U)}, \; \ldots, \; b_{13}^{(II)} = b_{13}^{(U)}$$

We verify these hypotheses by use of the F-Snedecor statistics. The probabilities $P = P(F > F_{calc}/H_0)$ are given in Table 5. One can see from Table 5 that the hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ can not be rejected at the $\alpha = 0.01$ significance level either for the predicted variable Fs' or for Fh'. It means that although there are great differences in the mean values of the predicted variables y (see Table 1), the prediction functions can be assumed to be the same for the compared data sets.

Table 5

Values of probability $P(F > F_c/H_0)$, where H_0 is the hypothesis stating that the regressions in the data set II (δ sunspot groups) and in the remaining data sets I, III, IV, V (sunspot groups of other magnetic classes) are parallel or equal

Compared data sets	Predicted variable	$P(F > F_c/H_0)$	
		Parallelity of the regressions	Equality of the regressions
I and II	y1 = Fs'	0.3508	0.4870
	y2 = Fh'	0.1015	0.4202
II and III	y1 = Fs'	0.2614	0.9097
	y2 = Fh'	0.2828	0.8127
II and IV	y1 = Fs'	0.4664	0.7435
	y2 = Fh'	0.2349	0.8879
II and V	y1 = Fs'	0.6762	0.0352
	y2 = Fh'	0.3930	0.2002

4. Conclusions

Analysing the variables describing sunspot groups and their flare activity by univariate and multivariate statistical methods we found that there are great differences in characteristics between sunspot groups of the δ and of the other magnetic classes (Table 1). The proportion of strongly flaring sunspot groups in the data set comprising items of the δ magnetic class is larger than in the data set with items of other magnetic classes. This is visualized in Figs. 1 and 2. We found that $x10$ (number of strong flares) and $x13$ (total X-ray flare flux in the wavelength range 0.5–4 Å) are the variables most strongly discriminating the δ sunspot groups from groups of other magnetic classes (Table 2). However, the analysis of the regression functions (Table 5) did not reveal significant differences between the regressions

evaluated from the two considered data sets. It means that for the predictions of flare activity the same prediction functions may be used for sunspot groups of δ and other magnetic classes.

REFERENCES

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

Bartkowiak, A., Łukasik, S., Chwistecki, L., Mrukowicz, M., and Morgenstern, W. 1987, *Proceedings MEDINFO'87*, **III**, eds. A.Serio, R.O'Moore, A.Tardini, F.H.Roger (Rome), p.1262.

Bartkowiak, A., and Jakimiec, M. 1989, *Acta Astr.*, **39**, 85.

Bartkowiak, A., and Jakimiec, M. 1990, *Robust Regression in Short-Term Prediction of Solar Flare Activity, in Solar-Terrestrial Predictions, Proceedings of Workshop, Leura, Australia, 1989.*

Jakimiec, M., and Bartkowiak, A. 1990, *Influence of Atypical Data Vectors on the Relationships among Characteristics of Solar Active Regions, in Solar-Terrestrial Predictions, Proceedings of Workshop, Leura, Australia, 1989.*

Jakimiec, M., and Bartkowiak, A. 1991, *Acta Astr.*, **41**, 1.

Priest, E.R. 1984, *Adv, Sp. Rev.*, **4**, 7,37.

Rao, C.R. 1973, *Linear statistical inference and its applications*, Willey, New York.

Zirin, H., and Liggett, M. 1987, *Sol. Phys.*, **113**, 267.