ACTA ASTRONOMICA
Vol. **38** (1988) pp. 431–442

Problem of Short Term Prediction of Solar Flare Activity. IV. Overestimations in Solar Flare Activity Predictions

by

M. Jakimiec and M. Wanke-Jakubowska

Astronomical Institute of Wrocław University, Wrocław

Received March 1, 1988

ABSTRACT

The bias of a training data set owing to the sample selection effect is analysed. The investigation is based on the predictions performed by use of the MVRA for four flare characteristics. The sample burdening with the selection effect is due to the fact that, usually, daily values of solar flare characteristics are used as the predictors and that the flare activity changes in time shorter than 24 hours. The asymmetry effect consisting in the overestimation of the forecasted values for low flare activity predictions seems to be a result of this selection. This effect is found to be statistically significant.

1. Introduction

Generally, one assumes *a priori* without evidencing the faultlessness of the data, although the problem of quality of the data base employed in prediction procedure is quite frequently not omitted. Indeed, when analysing various kinds of the data errors one concludes that most of these errors do not burden the data base. This fact is optimistically emphasized by Sawyer *et al.* (1986). They said: "The problem of errors in the data base, although far from trivial, is presumably correctable". However, one should be sure that the samples used in prediction procedure are random and unbiased and this problem was not investigated with sufficient care as yet. The problem was discussed in Paper I of this series (Jakimiec 1987) and in the present paper will be studied further on.

First, we will consider the following three facts:

(1) Flare activity characteristics are, without doubt, the best predictors of flare activity occurring the next day (*e.g.* Hirman *et al.* 1980; Jakimiec and Wasiucionek 1980; Neidig *et al.* 1986; Bartkowiak and Jakimiec 1986).
(2) Prediction methodology is seriously limited by the fact that we use the daily values of solar activity characteristics (*i.e.* available observations are performed every 24 hours only).
(3) Flare activity is changing strongly in time range shorter than 24 hours. These three facts together cause the sample (*i.e.* the training data set used for estimation of algorithm parameters) to be burdened with the effect of selection, *i.e.* the sample is not random and unbiased (*e.g.* Pfleiderer (1983) has discussed comprehensively the problem of sample selection).

Generally, one employs in the prediction procedure the daily characteristics of the solar flare activity adjusting them to the other characteristics of an active region. In Paper III of the series we found that the time interval between the appearing of the strong flare activity and the appearing of the precursor – the harder X-ray flux enhancement – is shorter than 24 hours. This fact becomes very important when the flare activity increases rapidly after an interval of low activity in a given active region. At that time it may occur that both the precursor and the strong flares appear in the same time interval of 24 hours, and so the information announcing the activity increase is lost. Consequently, the data used in prediction (training data set) will be burdened with the effect of the sample selection. The sample bias results in a bad prediction quality. Of course, the bad quality of prediction can also be due, besides other reasons, to the lack of more appropriate predicting variables.

In summary, we conclude that the sample selectiveness and/or the lack of more appropriate variable result in the algorithm deficiency in rendering the variability range of the observed values of the predicted variable, *Y*. So, it seems likely that it may cause the regression function in multidimensional space of variables to change somewhat the inclination in comparison with the inclination expected for an unselected data sample. This inclination change may be reflected in the fact that the predicting algorithm shows a tendency to give a smaller dispersion of the forecasted values around the mean value than the actual dispersion of the observed values do. Thus the variance $s_{\hat{y}}^2$, calculated for the forecasted values ($\hat{y}$), is lower than the variance s_y^2, calculated for the observed values (y). In the confrontation of the forecasted and actually observed values of the predicted variable in form of a test table, the inclination of the predicting function may appear as an effect of asymmetry.

In the solar flare activity prediction the effect of asymmetry is twofold: (1) Firstly, there is the effect of underestimation of strong flare activity (U-SFA), as discussed in Paper III. (2) Secondly, there is the effect of the overestimation of the low flare activity (O-LFA). This second effect occurs, rather commonly in almost all predictions. The first time this effect was noticed by Jakimiec (1983) in an analysis of a number of various predictions. Moreover, the same (O-LFA)

effect can be also seen *e.g.* in the test tables reported by Beirong *et al.* (1986), Neidig *et al.* (1986) or Bartkowiak and Jakimiec (1986).

In this paper, using the actual data set for solar flare activity prediction, we will examine whether the O-LFA effect of asymmetry is associated with the selection effect of the employed sample. With this intent we form two data sets: a training data set (TDS) and a choosen one (TDS-bis) taken out from the former set. We expect that the changed training data set will give as a result the decrease of the selection effect and the improvement of the prediction quality. In Section 2.1. we present the way of the TDS-bis formation.

2. Analysis of the data

2.1. Data sets.

The observational data cover the time range from January 1979 to June 1980, and have been collected from the Solar Geophysical Data – SGD (1979, 1980). We analyse the complex of 18 daily characteristics for the D, E, F Zurich class sunspot groups. The first seven variables X, describe sunspot group characteristics (the same were used by Jakimiec and Bartkowiak, 1986). They are as follows: *x1* – McI, McIntosh sunspot class; *x2* – A, sunspot group area; *x3* – CaA, calcium plage area; *x4* – CaI, calcium plage intensity; *x5* – Mag, magnetic class; *x6* – H, magnetic field strength; *x7* – MFI, magnetic field index.

The further seven variables describe solar flare activity of the sunspot group on a given day. Now, it is useful to remind of the notations used in Paper III: *fs* and *fh* are maximum values of solar flare X-ray flux in the wavelength intervals 1-8 Å and 0.5-4 Å, respectively; *Fs* (Total Flare Flux) is the sum of *fs* values for the sunspot group on a given day, and correspondingly *Fh* is the sum of *fh* values; *h* is the quotient of the sum [*fh*] and the sum [*fs*] calculated consecutively every three hours for six-hour time intervals. So, the variables are as follows: *x8* – maxX, the maximum value of *fs* for the sunspot group on a given day; *x9* – NFF, the number of faint flares (for which $fh < 0.08 \cdot 10^{-2}$ erg cm^{-2} s^{-1}) per day; *x10* – NSF, the number of stronger, flares (for which $fh > 0.08 \cdot 10^{-2}$ erg cm^{-2} s^{-1}) per day; *x11* – *Fs*, Total Flare Flux (1-8 Å); *x12* – Fh, Total Flare Flux (0.5-4 Å); *x13* – HI, Hardness Index, *i.e.* the quotient of *Fh* and *Fs* values; *x14* – maxh, maximum value of eight *h* values for a given day.

Four predicted variables, Y, concerning the flare activity on the next day are: *y1* – maxX′, the maximum value of *fs* for the sunspot group on the next day; *y2* – NSF′, the number of stronger flares (for which $fh > 0.08 \cdot 10^{-2}$ erg cm^{-2} s^{-1}) per day; *y3* – Fs′, Total Flare Flux (1-8 Å); *y4* – Fh′, Total Flare Flux (0.5-4 Å).

The frequency distribution of the most of the 18 analysed variables reveal very strong skewness. It concerns the following variables: *x1*, *x2*, *x3*, *x7*, *x8*, *x11*, *x12*, *y1*, *y3*, *y4* and therefore for further analysis we use the variable values obtained after the logarithmic transformation: $X \Rightarrow \log X$.

In order to analyse the consequence of the sample selectivity for the prediction quality, three data sets were formed. Two of them, TDS and TDS-bis, covering 1979 year are the base for the estimation of predicting algorithm parameters. The third, control data set (CDS), cover the first half of 1980 year, and will be used for the prediction quality evaluation. The choosen data set (TDS-bis) was set up from TDS as follows: we remove from TDS the cases, for which flare activity for a given day was very low (*i.e.* $x13 < 0.04$), and for which strong flare activity or the harder X-ray flux enhancement occurred the next day. We hope that the way we perform the transformation of TDS cause the predictions made for TDS-bis to be less burdened by the fact we use the daily data.

We have included into our data sets (TDS, TDS-bis and CDS) only those sunspot group observations for which the full set of 18 characteristics could be picked up. The sample sizes are: $N_{TDS} = 383$, $N_{TDS\text{-}bis} = 291$, $N_{CDS} = 234$, respectively.

2.2. *Correlation matrices analysis.*

For both data sets, TDS and TDS-bis, two variants of the correlation matrices were calculated: (*a*) for the relation between predicted (Y) and predicting (X) variables, and (*b*) for the relation between Y and X^2 variables. We have analysed three maximum values of correlation coefficient, r, belonging to the matrices, for each of the predicted variables and for two, (*a*) or (*b*) relations. Without any loss of information only the (*a*) relation between Y and X variables can be taken into account for further analysis because the differences between the r values calculated for (*a*) and (*b*) relations are not significant. We found also that the set of three variables with maximum r values in the covariance matrices for TDS and TDS-bis, and also for (*a*) and (*b*) relations is very similar. It is very important that among these variables only flare characteristics such as Fs, Fh, NSF and maxX are present.

Table 1 shows the values of correlation coefficient between the variables *x8*, *x10*, *x11*, *x12*, *y1*, *y2*, *y3* and *y4* characterizing flare activity. Comparing the r values for the variables X (upper-left part) or Y (lower-right part of the matrix) we see that the variables reveal very strong interrelation, and also, the r values calculated for the TDS and TDS-bis sets are not significantly different. Instead, the X variables are correlated less strongly with the predicted variables Y taken for the next day (lower-left part of the matrix for TDS, and upper-right part of the matrix for TDS-bis). This means that the 24 hours time interval makes the random factors more prevailing for the relation of Y and X. So,

Table 1

Values of the correlation coefficient between the variables characterizing flare activity. TDS – lower part, and TDS-bis – upper part of the table.

TDS-bis / TDS	x8 maxX	x10 NSF	x11 Fs	x12 Fh	y1 maxX'	y2 NFS'	y4 Fs'	y5 Fh'
x8-maxX	-	0.75	0.93	0.95	0.63	0.58	0.65	0.65
x10-NSF	0.75	-	0.76	0.76	0.63	0.60	0.66	0.69
x11-Fs	0.92	0.74	-	0.98	0.65	0.56	0.67	0.69
x12-Fh	0.93	0.75	0.97	-	0.64	0.57	0.68	0.68
y1-maxX'	0.52	0.52	0.55	0.54	-	0.75	0.93	0.92
y2- NSF'	0.52	0.53	0.51	0.52	0.72	-	0.73	0.71
y3- Fs'	0.54	0.53	0.60	0.58	0.92	0.72	-	0.96
y4- Fh'	0.54	0.55	0.58	0.57	0.92	0.69	0.96	-

using the daily values in predictions we lose an essential amount of information. The information loss is different for TDS and TDS-bis, namely, on the average, the random variance for TDS-bis is less than for TDS by about 0.11. This fact may be considered as the first indication that the sample bias effect may be important for the prediction. So, in further analysis we will examine whether the consequence of data set selection by use of the flare characteristics may affect the prediction quality, as it follows from above analysis.

2.3. *Analysis of predicting functions.*

We use for the predicting function evaluation the multi-variable regression analysis (MVRA) methods. We employ the linear regression of the Y predicted variables on X variable vector:

$$yk = a_0 + a_1\, x1 + a_2\, x2 + \ldots + a_{14}\, x14 + ek, \tag{1}$$

where yk (k = 1, 2, 3, 4) are the predicted variables, $x = (x1, x2, \ldots, x14)$ is the vector of explanatory variables, a_0 is the free term, $a = (a_1, \ldots, a_{14})$ is the vector of the regression coefficients, and ek is the error.

The values of the regression function (RF) coefficients were estimated for both data sets, TDS and TDS-bis. In order to obtain a proper subset of the predicting variables (X) we applied the stepwise search procedure MAXSTEPREGR (Bartkowiak 1978). We work with two significance levels: $\alpha = 0.10$ and $\alpha = 0.04$, and we adopt the number of predicting variables not greater than five. It means we assume that to draw out the major part information included into the set of explanatory variables it is sufficient to take five predicting variables only. So, we cannot get a new important information for the prediction including to RF any additional variable. The multiple correlation coefficient, RR, was used as a measure of the closeness of the Y on X relation.

Table 2 presents the set of predicting variables choosen for the *RF* (with $\alpha = 0.10$), and the values of *RR* (in parentheses). We see that besides the variables (such as *x10*, *x11* or *x14*) characterizing flare activity, the predicting variable set contains also the variables characterizing sunspot groups (such as *x2* or *x7*). The underlined variables were choosen in the procedure of the *RF* estimation with the significance level $\alpha = 0.04$. Then we note that for both data sets the obtained variable subset is the same.

Table 2
Set of the predicting variables and the determination coefficient values.

	TDS		TDS-bis	
	X	RR	X	RR
y1	x2, x4, x10, x11	(0.36)	x2, x6, x7 x10, x11	(0.49)
y2	x2, x5, x10, x14	(0.37)	x2, x5, x6, x10, x14	(0.45)
y3	x1, x4, x7, x10, x11	(0.42)	x2, x6, x7, x10, x11	(0.55)
y4	x4, x5, x7, x10, x11	(0.41)	x5, x7, x9, x10, x11	(0.55)

It can also be seen from Table 2 that *RR* values calculated for TDS-bis set are systematically greater than those calculated for TDS. This result confirms the importance of differences between the values of correlation coefficient shown in Table 1. This may be the evidence that the random variance for TDS-bis is lower than for TDS either because of the rejection of the outlying points (when we form the TDS-bis) or because the *RF* calculated for TDS-bis (TDS-bis-*RF*) is tilted to the *RF* obtained for TDS (TDS-*RF*) (or these two lines are shifted one from another).

For each of the predicted variables we receive two predicting functions, for TDS and TDS-bis; *i.e.* we obtain two regression coefficient vectors. As a measure of the mentioned inclination we use the ϕ angle between these two vectors (estimated with $\alpha = 0.04$). Moreover, as a measure of TDS-bis-*RF* shift we take the absolute value of the subtraction result of the corresponding a_0 values ($|\Delta a_0|$). It should be noted that to calculate the ϕ and $|\Delta a_0|$ values we reduced the units of the *RF* coefficients to one standard deviation of the predicted variable. The ϕ and $|\Delta a_0|$ values are given in the 3-th and 4-th columns of Table 5, respectively.

The calculations concerning the MVRA methods have been performed on an Odra 1305 computer with programmes from the SABA package by Bartkowiak (1981).

2.4. Analysis of prediction quality.

Next we will examine the differences in the quality of predictions which are performed by use of the algorithms calculated for TDS and TDS-bis sets (now

we use the RL estimation with $\alpha = 0.10$). The comparison of the forecasted ($\hat{y}$) and observed (y) values of the predicted variables was made for the control data set (CDS). In test tables each of four predicted variables are divided into four categories (see Table 3). In Table 4 are put together the test tables for the TDS prediction evaluation (left part) and for the TDS-bis prediction evaluation (right part). One can easily see that the O-LFA effect of asymmetry is quite prominent in the TDS tables and is far less prominent for the TDS-bis tables.

Now we will examine whether the O-LFA effect is eliminated in the TDS-bis test tables. Let $[n_{ij}]$ be the matrix of the event numbers, and $[p_{ij}]$ –

Table 3
The division of variable Y values into categories.

	1	2	3	4
y1	≤ 0.3	0.31 – 0.8	0.81 – 1.3	≥ 1.31
y2	0	1, 2	3, 4	≥ 5
y3	≤ 0.3	0.31 – 0.8	0.81 – 1.3	≥ 1.31
y4	≤ 0.1	0.11 – 1.2	1.21 – 2.2	≥ 2.21

Table 4
Matrices of forecasts vs observed events.

observed, y		forecasted, ŷ TDS 1	2	3	4	TDS-bis 1	2	3	4	Total
y1	1	73	39	1	0	93	19	1	0	113
	2	18	33	8	0	27	24	8	0	59
	3	3	31	11	2	11	28	7	1	47
	4	3	7	4	1	4	6	4	1	15
Total		97	110	24	3	135	77	20	2	234
y2	1	64	44	7	0	81	28	6	0	115
	2	18	38	12	1	26	33	9	1	69
	3	2	16	11	2	4	15	10	2	31
	4	1	7	6	5	2	7	5	5	19
Total		85	105	36	8	113	83	30	8	234
y3	1	26	60	11	0	69	24	4	0	97
	2	9	16	12	1	21	10	6	1	38
	3	2	14	24	5	7	21	13	4	45
	4	0	14	27	13	4	21	19	10	54
Total		37	104	74	19	101	76	42	15	234
y4	1	2	72	6	0	18	60	2	0	80
	2	0	30	20	1	7	34	29	1	51
	3	0	27	33	3	2	34	24	3	63
	4	0	11	21	8	0	21	15	4	40
Total		2	140	80	12	27	149	50	8	234

the matrix of the probabilities in a given test table. Then the O-LFA effect is manifested by the fact that n_{12} exceeds n_{21}. In order to investigate whether the O-LFA effect in the test table is statistically significant we put the null hypothesis: $H(p_{12} = 0.5)$. Probability p_{12} can be estimated by the fraction:

$$f_{12} = \frac{n_{12}}{n_{12}+n_{21}}. \tag{2}$$

If the null hypothesis is right, the statistics:

$$u = \frac{f_{12}-p_{12}}{\sqrt{p_{12}p_{21}}}\sqrt{n_{12}+n_{21}} = \frac{f_{12}-0.5}{0.5}\sqrt{n_{12}+n_{21}} \tag{3}$$

has asymptotically normal distribution, *i.e.* $u \in N(0, 1)$. In the 5-th and 6-th columns of Table 5 the f_{12} and u values are given for each of predicted variables and for TDS and TDS-bis predictions, $u_\alpha = 2.58$ is the critical value for the

Table 5

Values of parameters used for the comparison of the quality of predictions performed by use of the TDS and TDS-bis data sets.

variable	data set	ϕ	$\lvert\Delta a_0\rvert$	f_{12}	u	$\bar{v}$	WS	AS
y1	TDS	7°.8	0.710	0.684	2.77	-0.13	0.38	0.15
	TDS-bis			0.413	1.18	-0.32	0.39	0.37
y2	TDS	7°.8	0.054	0.710	3.31	0.06	0.32	0.06
	TDS-bis			0.519	0.28	-0.09	0.24	0.09
y3	TDS	9°.4	0.705	0.870	6.15	0.08	0.49	0.07
	TDS-bis			0.533	0.44	-0.36	0.43	0.33
y4	TDS	9°.0	0.742	1.00	8.46	0.17	0.49	0.14
	TDS-bis			0.896	6.48	-0.10	0.61	0.10

significance level $\alpha = 0.01$. Jakimiec (1986) formed three indices of the prediction quality: e – a measure of the dispersion of v deviations ($v = \hat{y} - y$), WS ("wings" index) and AS (asymmetry index) – measure of the large deviations. The last three columns of Table 5 give the values of mean deviation ($\bar{v}$), and WS and AS values. The critical values for WS and AS indices are $WS_c = 0.35$ and $AS_c = 0.25$, respectively. The values higher than the corresponding critical values are underlined. The calculated values of index e are not presented in Table 5 because they are less than the critical values for all test tables. It means that the v values are concentrated around the diagonal ($v = 0$) quite good for all predictions.

From Table 5 we see that the O-LFA effect is statistically significant ($u > u_\alpha$) for the predictions obtained by use of the TDS data. This effect is absent ($u < u_\alpha$) for the *y1*, *y2*, *y3*, variables and is somewhat reduced for the variable *y4* when the predicting algorithm coefficients are estimated for the

TDS-bis data set. The elimination of the O-LFA effect seems to be associated with the fact that the TDS-bis-*RFs* are tilted to the TDS-*RFs* (at $\phi \sim 8° - 9°$). However, we found that the TDS-bis-*RFs* are shifted from TDS-*RFs* ($|\Delta a_0| \sim 0.70$) for the variables *y1*, *y3* and *y4*. So, we have the worsening of the prediction quality ($\bar{v} \ll 0$, $WS > WS_c$ and/or $AS > AS_c$) which consists in the increase of the U-SFA effect.

We found in Paper III that the enhancement of the hardness index (see variable *x14*) occurs frequently before strong *X*-ray flares in time interval shorter than 24 hours. We want to know whether the underestimated events in Table 4 (for which the deviation values $v = \hat{y} - y$ are negative) were forerun by the hardness index enhancement, *i.e.* by the appearance of small, faint flares for which the harder (0.5-4 Å) *X*-ray flux was greater than $0.08 \cdot 10^{-2}$ erg

Table 6

The number of underestimated events, n_{UE}, and the corresponding number of strong events for corrected forecast, n_{SE}.

		y1		y2		y3		y4	
	v	n_{UE}	n_{SE}	n_{UE}	n_{SE}	n_{UE}	n_{SE}	n_{UE}	n_{SE}
TDS	−1	53	23	40	21	50	20	48	26
	−2	10	6	9	6	16	10	11	7
	−3	3	1	1	1	0	0	0	0
TDS-bis	−1	59	24	46	23	61	21	56	23
	−2	17	10	11	7	28	17	23	14
	−3	4	1	2	2	4	3	0	0

cm^{-2} s^{-1}. Table 6 shows the number of underestimated events n_{UE} (left column) and the corresponding number of events n_{SE} (right column) for which strong flares ($\geqslant$ C5, *i.e.* $fs \geqslant 0.07$) were found occurring during the second half of the day, and also for which during the first half of the day the enhancement of the harder X-ray flux was observed. Very low flare activity occurring the preceeding day may render the proper forecasting of these strong flares impossible. We note that we would gain the corrected forecasts for about 48% of such events if we were able to employ the flare data from the first half of a given day. Considering that some of the events with $v = -1$ may be included into the central part of the v normal distribution (Jakimiec 1986), the mean percentage of the corrected forecasts is about 80%.

In Section 1 the question of the predicting algorithm deficiency was discussed. The deficiency arises from the fact that the algorithm shows the tendency to increase the concentration of the forecasted values around the mean value of the predicted variable (this effect is most prominent for the prediction of the variable *y4* — see Table 4). As a consequence, the variance $s_{\hat{y}}^2$ of the forecasted ($\hat{y}$) values is significantly less than the s_y^2 variance of the observed (y) values. This effect can be expressed by the values of the Variance

Quotient:

$$VQ = \frac{s_y^2}{s_{\hat{y}}^2}. \tag{4}$$

The VQ values calculated for the TDS and TDS-bis predictions are given in Table 7. The VQ statistics, for normally distributed variable Y, has F-Snedecor distribution with $(N-1, N-1)$ degrees of freedom. The critical value is $F_\alpha = 1.39$ for the significance level $\alpha = 0.01$. Table 7 shows that only for the

Table 7
Values of VQ quotient of the variances of the forecasted and observed values.

	y1	y2	y3	y4
TDS	1.890	1.438	2.124	3.352
TDS-bis	1.949	1.371	1.729	2.802

TDS-bis prediction of the $y2$ variable, the VQ value is less than the critical F_α value. It means that only for $y2$ variable we obtain the predicting algorithm which gives the correct variability range of the predicted variable. We can also see from Table 7 that for other variables the VQ values are greater than the critical value. This conclusion is in good agreement with the conclusion drawn above from the analysis of the prediction quality (Table 5). So, we think that also the VQ statistics may be used as a certain indicator of the prediction quality.

3. Conclusions

In paper I of this series the different effects associated with the prediction quality were discussed. The choice of the adequate set of the explanatory variables may be of great importance for this problem. The question of the usefulness of these variables is also very significant for the predicton quality. The quality of prediction may be affected by the facts that the observational data are not enough faultless and that the sample selected for the predicting function estimation may not be random and unbiased. Further effect associated with the prediction quality that should be mentioned is the unhomogeneity of a population from which we select both, TDS and CDS data sets. Problem of the homogeneity is of major importance in the proper prediction procedure where the predicting function is extrapolated forward.

It is very important to recognize the consequences of these effects for the prediction quality. Bad prediction quality denotes the appearance of high number of large deviations which appear in the form of the overestimated or

underestimated events. In present work we suppose the overestimations (O-LFA effect) to be the result of the fact that TDS is selected not randomly. If the calculation result does not depend on the way that we select the sample, we may regard this sample as randomly selected. However, really, the estimated values of the predicting function coefficients depend on the variables characterizing solar flare activity. For the reason that we use the daily values of solar activity characteristics, and that flare activity changes strongly in time interval shorter than 24 hours, these variables are essential to the procedure of the TDS formation. It means we may think that the result (the calculated values of the predicting function coefficients) is burdened with the selection effect. From Tables 4 and 5 we can see that for the predictions based on the TDS the O-LFA effect is present and is statistically significant. To examine whether the selection effect may be removed we have taken purposefully another training data set (TDS-bis). We reject from TDS such observational vectors that contain very low flare activity for the given day and strong flare activity for the next day. We found that the predicting functions calculated for the TDS-bis are inclined to those calculated for the TDS at the angle of about 8 to 9 degrees (see Table 5). Comparing the prediction qualities performed for both data sets we see that the O-LFA effect is highly reduced for TDS-bis predictions. This fact means that the employing of the daily values in the short-term predictions of flare activity may be the reason of the O-LFA effect which have been stated in many predictions previously performed.

It was mentioned above that the bad quality of predictions may be due, besides the sample bias, to the lack of more appropriate predicting variables. Generally, this lack is believed to be the most important reason of the bad prediction quality. It can be seen from Tables 4 to 7 that the prediction quality is not the same for various variables. This fact seems to argue that we should to search for another sunspot group characteristics in the predictions of flare X-ray flux (particularly for the *y4* variable). However, if we use the sample with any new sunspot group characteristic (*e.g.* magnetic shear index) we do not remove the 24-hour selection effect.

In most predictions performed as yet for which the O-LFA effect is stated the training and control data sets were constructed for the same time interval. Therefore there was no problem of the predicting function extrapolation and this problem have not been discussed as the reason of bad quality of predictions. In present paper the TDS and CDS are constructed for two different time intervals (1979 and 1980). So, analysing the quality of predictions, the problem of the predicting function extrapolation can not be omitted. Particularly, the bad quality of the *y4* variable prediction may be the result of the extrapolation. For the first half of 1980 year for which the CDS was formed, we can notice strongly increased solar activity level. This solar activity rise can change strongly the structure of the interrelations of various sunspot group characteristics, particularly, the harder X-ray flux of solar flares. In the further

work we will study in more detail the difference in the structure of interrelations of sunspot group characteristics comparing 1979 and 1980 years.

In this paper the multivariable regression analysis (MVRA) alone is used for the predicting function estimation. We believe that the result would look similarly also for the predictions performed by use of the other methods, *e.g.* for multivariable discriminant analysis (MVDA) methods which are based also on the correlation matrix.

Now, we will mention that the analyzed O-LFA effect may be dependent on the solar numerical force of sunspot groups, *i.e.* it might be connected with 11-year cycle (being greater for the high solar activity and smaller for the decaying phase of solar cycle). This fact, besides other reasons, may results in the worsening of the quality of prediction performed for the maximum phase of solar activity cycle. Hirman *et al.* (1980) who used data for 1977 year obtained better prediction quality than the prediction quality obtained by Neidig *et al.* (1986) who used data for time interval from January 1977 to January 1979. Similarly, Jakimiec and Wasiucionek (1980) showed that the prediction quality is better for the period of lower solar activity (1973-75) than for the period of high solar activity (1971-72). The problem needs further investigation for the entire 11-year cycle of solar activity.

REFERENCES

Bairong, Z., Baorong, L., Zhihuang, L. 1986, in *Solar-Terrestrial Prediction Proceedings*, P. A. Simon, G. Heckman and M. A. Shea, eds. 324.

Bartkowiak, A., 1978, *Applicationes Mathematicae* **XVI**, 2, 293 1981, SABA, the technical description of statistical programs in the ALGOL 1500 language (in Polish), Wrocław University.

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

Hirman, J. W., Neidig, D. F., Seagraves, P. H., Flowers, W. E. and Wiborg, P. H.: 1980, in *Sol.-Terrest. Pred. Proc.*, Vol. 3, R. F. Donnelly, ed. C-64.

Jakimiec, M., 1983, in *Statistical Methods in Astronomy* (ESA, Paris, France), 109.

– 1986, in *Solar-Terrestrial Prediction Proceedings* P. A. Simon, G. Heckman and M. A. Shea, eds. 311

– 1987, *Acta Astr.* **37**, 270.

Jakimiec, M., and Bartkowiak, A., 1986, in *Solar-Terrestrial Prediction Proceedings*, P. A. Simon, G. Heckman and M. A. Shea eds. 294.

Jakimiec, M., and Wanke-Jakubowska, M., 1987, *Acta Astr.* 37, 299.

Jakimiec, M., and Wasiucionek, J., 1980, in *Sol-Terrest. Pred. Proc.*, Vol. 3., R. F. Donnelly, ed. C-54.

Neidig, D. F., Wiborg, P. H., Seagraves, P. H., Hirman, J. W., and Flowers, W. E., 1986, in *Solar-Terrestrial Prediction Proceedings*, P. A. Simon, G. Heckman and M. A. Shea, eds., 300.

Pfleiderer, H., 1983, in *Statistical Methods in Astronomy* (ESA, Paris, France).

Sawyer, C., Warwick, J. W., and Dennett, J. T., 1986, *Solar Flare Prediction* (Colorado Associated University Press, Boulder, USA).

SGD 1979, 1980 Solar Geophysical Data – Comprehensive Reports.