J. Space Weather Space Clim. 3 (2013) A09
DOI: 10.1051/swsc/2013032

SWSC

RESEARCH ARTICLE OPEN ACCESS

Statistics of sunspot group clusters

Ryszarda Getko

Astronomical Institute, University of Wrocław, Kopernika 11, 51–622 Wrocław, Poland
Corresponding author: e-mail: r-get@wp.pl

Received 27 February 2012 / Accepted 16 February 2013

ABSTRACT

The Zubrzycki method is utilized to find all sunspot groups which are close to each other during each Carrington rotation. The sunspot group areas and their positions for the years 1874–2008 are used. The descending, the ascending and the maximum phases of solar cycles for each solar hemisphere are considered separately. To establish the size of the region D where the clusters are searched, the correlation function dependent on the distance between two groups is applied. The method estimates the weighted area of each cluster. The weights dependent on the correlation function of distances between sunspot groups created each cluster. For each cluster the weighted position is also evaluated. The weights dependent on the areas of sunspot groups created a given cluster. The number distribution of the sunspot groups created each cluster and the cluster statistics within different phases of the 11-year cycle and within all considered solar cycles are also presented.

Key words. Sun – sunspot – statistics and probability – variability

1. Introduction

In the last decades, many studies investigated the statistical properties of active regions. Harvey & Zwaan (1993) used the daily full-disc magnetograms taken by the National Solar Observatory/Kitt Peak to examine the size distribution of bipolar active regions. They considered the new active regions and the existing active regions. An analysis indicates that their size distribution function decreases with increasing size. The shape of the size distribution function of new active regions is independent of the phase of a solar cycle, only the total number of region rises and falls with rise and decline of the solar cycle. This similarity in the shape of the size distribution functions was also found by Tang et al. (1984). They studied bipolar active regions in the Mount Wilson Observatory magnetograms. Although all the above authors used different selection criteria, Harvey & Zwaan (1993) showed that the cumulative size distributions of the emerging bipoles from their study, the sample of existing sunspot regions and the Tang et al. (1984) sample of magnetic bipolar regions had the same shape and had different slopes only. These results are also consistent with the Schrijver (1988) size distribution function of Ca II plage regions. Parnell et al. (2009) studied photospheric magnetic field data from different instruments in order to investigate the distribution function of magnetic features over a wide range of scale. They and a number of authors (Harvey 1993; Harvey & Zwaan 1993; Schrijver & Harvey 1994) showed that the bipolar magnetic regions, from the largest regions with major sunspot to ephemeral active regions without sunspots, follow the power law. Moreover, the shape of the size distribution function is an invariant solar feature.

The tendency for active regions to emerge clustered both in position and in time is a property of solar activity (Gaizauskas et al. 1983; Castenmiller et al. 1986; Brouwer & Zwaan 1990; van Driel-Gesztelyi et al. 1992). Gaizauskas et al. (1983) define a complex of activity as a cluster or a sequence of many active regions which are related in proximity and by continuity in their emergences. These active regions emerge in the same belt of activity and within well-defined zones of longitude. The complex of activity is maintained for typically 3–6 Carrington rotations by repeated injections of new magnetic flux which is concentrated in active regions within that complex. Getko (2004) identified very strong activity complexes creating a strong Wolf number fluctuation during different phases of the 11-year cycle and presented their development curves with time. This study extends same of these researches. The focus of the investigation reported in this paper is to determine the statistical properties of activity complexes as a step in understanding the emergence of magnetic flux on the Sun on all size scales. For this study, the sunspot groups are used as tracers for activity complexes. In particular, *sunspot area* as representatives of the freshly emerged surface magnetic field is chosen. To localize sunspot groups which create a complex of activity during one Carrington rotation, the Zubrzycki method (Zubrzycki 1957) based on the theory of stochastic processes is applied. In this analysis each sunspot group appearance is represented by a data point with two coordinates: the mean latitude (x) and the mean Carrington longitude (y). For each sunspot group the sum of areas during the days the group has been visible during one Carrington rotation is considered. This sum as a function of each sunspot group location is a parameter which is treated as a stochastic process. The Zubrzycki method is used to find the clusters. The cluster includes the objects (in this paper, sunspot groups) with low distances among the cluster members. By applying statistical clustering criteria given by Zubrzycki, a fraction of the groups existed during one Carrington rotation may be grouped into clusters. Because the members of the cluster could change from one Carrington rotation to another the search of clusters during successive rotations is independent. These clusters could form long-lived structures

like complexes of activity. Such an approach seems to be consistent with the Gaizauskas definition (Gaizauskas et al. 1983). To describe the properties of clusters the statistics of the cluster data sets and the histograms of numbers of sunspot groups creating one cluster evaluated during different phases of the solar cycle and during each of 12 considered solar cycles in each of solar hemisperes are presented.

2. Data and methods

The data used for this study are the daily sunspot groups positions and the areas (in units of millionths of a hemisphere) for the northern hemisphere and the southern hemisphere for the years 1878–2010 (solar cycles 12–23) available at the National Geophysical Data Center (NGDC) (http://solarscience.msfc.nasa.gov/greenwch/). For each sunspot group the mean position and the sum of its areas during each Carrington rotation for each solar hemisphere separately are evaluated. To describe the properties of the clusters the data from each solar cycle are divided into three groups: the ascending, the maximum and the descending phases of the solar cycle. For each solar cycle all i for which the values of monthly smoothed Wolf numbers R_i are less than $0.75 \times \max(R_i)$ define the ascending and the descending phases. The remaining months define the maximum phase. All 12 data sets from the ascending (the maximum, the descending) phases are linked together. Finally, six data sets (for the ascending, the maximum and the descending phases and for each solar hemisphere) are considered.

The procedure used here to find the sunspot group clusters is described in Zubrzycki (1957). This method was introduced to the probabilistic description of the geological deposits. Usually it dealt with many characteristics of the deposit, such as the content of the ore, the thickness or the additives. These parameters are estimated on the basis of drilling at selected points, and then their values are evaluated in a predetermined area. However, the estimators such as the arithmetic average, the weighted average or the geometric average do not take into account the distribution of the samples. Thus, a method based on the theory of stochastic processes is proposed. The deposit structure can be described by constants and the correlation function between the samples. The correlation function determines the region (D) where the cluster members are localized. The parameter values at the points $p_1 = p(x_1,y_1)$, $p_2 = p(x_2,y_2)$, …, $p_k = p(x_k,y_k)$ belonging to the region D are represented by k random variables $Y_1 = Y(p_1)$, $Y_2 = Y(p_2)$,...,$Y_k = Y(p_k)$. The method needs a few assumptions. First of all, all the random variables have the same expected value

$$E(Y_i) = m, \tag{1}$$

and the same variance

$$D^2(Y_i) = s^2, \tag{2}$$

for $i = 1, 2, \ldots, k$. In addition, the correlation coefficient $R(Y_i, Y_j)$ between two random variables Y_i and Y_j depends only on the distance $d = \mathrm{d}(p_i, p_j)$ between the points p_i and p_j:

$$R(Y_i, Y_j) = f(\mathrm{d}(p_i, p_j)), \tag{3}$$

where $\mathrm{d}(p_i,p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. This means that a geological structure of the deposit is isotropic and does not depend on the direction. So, it can be treated as an isotropic stochastic process. Moreover, the conditions (1)–(3) indicate that it is stationary. The mean value of the parameter Y in the region D is defined by the following integral:

$$Y_D = \frac{1}{|D|} \iint_D Y(p)\mathrm{d}p, \tag{4}$$

where $|D|$ is the area of the region D and $Y(p)$ is the value of the parameter Y at the point $p = p(x, y) \in D$. It can be estimated on the basis of the parameter value Y at the points $p_1, \ldots, p_k$ belonging to the region D. The true values $Y_1, ..., Y_k$ at the points $p_1, \ldots, p_k$ can be measured at these points. The results are denoted by $Y_i^*, \ldots, Y_k^*$. Their expected values are also $E(Y_i^*) = E(Y_i + \epsilon_i) = m$ because $E(\epsilon_i) = 0$ (ϵ_i is the measurement error of Y_i). Their variances are $D^2(Y_i^*) = s^{*2}$. To find the region D the sample correlation function is given by:

$$R(d) = \frac{(2\sum Y_i^* Y_j^*)/2n - ((\sum Y_i^* + \sum Y_j^*)/2n)^2}{\left(\sum Y_i^{*2} + \sum Y_j^{*2}\right)/2n - ((\sum Y_i^* + \sum Y_j^*)/2n)^2}, \tag{5}$$

where d is the distance between two points p_i and p_j. The function $R(d)$ is cut for $d \geq d^*$, such that $R(d^*)$ is close to zero. The region D is defined as the square $2(d^* - 1) \times 2(d^* - 1)$. The function $R(d)$ is fitted by the following correlation functions:

Model I:

$$f^*(d,a,c) = \begin{cases} c \times \frac{2}{\pi}\left\{\arccos\frac{d}{a} - \frac{d}{a}\sqrt{1 - \left(\frac{d}{a}\right)^2}\right\} & \text{for } d < a, \\ 0 & \text{for } d \geq a, \end{cases}$$

Model II:

$$f^*(d,a,c) = c \times \exp\left\{-\frac{d^2}{a^2}\right\},$$

Model III:

$$f^*(d,a,c) = c \times \exp\left\{-\frac{d}{a}\right\}.$$

The parameters a and c minimalize the sum:

$$B^2 = \frac{1}{N}\sum_{i=1}^{k} n_i (r_i - f^*(d_i,a,c))^2, \tag{6}$$

where $r_1, \ldots, r_k$ denote the correlation coefficients calculated for the distances $d_1, \ldots d_k$ and $n_1, \ldots n_k$ are the number of pairs for which the correlation coefficients $r_{1, \ldots,} r_k$ are calculated ($N = n_1 + \ldots + n_k$). Because the minimalization of the expression (6) leads to the complicated equations, the trial method is applied. In order to determine the parameters a and c, the function $f^*(d, a, c)$ is represented as:

$$f^*(d,a,c) = cf^*(d/a,1,1), \tag{7}$$

and the expression (6) is written as:

$$B^2 = \frac{1}{N}\sum_{i=1}^{k} n_i (r_i - cf^*(d_i/a,1,1))^2. \tag{8}$$

To find the best solution the parameter a is fixed and the value B^2 is minimalized towards c. It follows that:

$$c = \frac{\sum_{i=1}^{k} n_i r_i f^*(d_i/a, 1, 1)}{\sum_{i=1}^{k} n_i f^{*2}(d_i/a, 1, 1)}. \tag{9}$$

For each pair c and a the expression (6) is calculated. In result, the pair for which the value B^2 is the smallest is chosen. This best fitting points out the function $f^*(d, a, c)$ to estimate the weighted area of a given cluster. This function $f^*(d, a, c) = f^*(d)$ is also used to determine the variance of $Y_1, \ldots, Y_k$:

$$s^2 = s^{*2} f^*(0+), \tag{10}$$

where $f^*(0+) = \lim_{d\to 0+} f^*(d)$. Thus, the following estimators are proposed:

$$E(Y_1^*) = E(Y_2^*) = \ldots = E(Y_k^*) = m, \tag{11}$$

$$E(Y_D) = m|D|, \tag{12}$$

$$D^2(Y_1^*) = D^2(Y_2^*) = \ldots = D^2(Y_k^*) = s^{*2}, \tag{13}$$

$$\omega\left(Y_i^*, Y_j^*\right) = s^2 f^*\left(\mathrm{d}(p_i, p_j)\right), \qquad i \neq j, \tag{14}$$

$$\omega(Y_j^*, Y_D) = s^2 \iint_D f^*(\mathrm{d}(p_j, q))\mathrm{d}q; \tag{15}$$

$$\omega(Y_D, Y_D) = s^2 \iint_D \left[\iint_D f^*(\mathrm{d}(p, q))\mathrm{d}q\right]\mathrm{d}p. \tag{16}$$

where $\omega(Y_i^*, Y_j^*)$ denotes the covariance between two random variables Y_i^* and Y_j^*.

The following estimators of the integral Y_D are taken into consideration.

I. The first is defined by the sum $c_0 + c_1 Y_1^* + \ldots + c_k Y_k^*$. The coefficients c_0, c_1, …, c_k minimalize the expression $E(c_0 + c_1 Y_1^* + \ldots + c_k Y_k^* - Y_D)^2$. Thus, the error estimation is $s_I(Y_D, Y_1^*, \ldots, Y_k^*) = \min E(c_0 + c_1 Y_1^* + \ldots + c_k Y_k^* - Y_D)^2$. The Least Square Method is applied. The constants c_1, …, c_k are given by the following normal equations:

$$\begin{aligned} &\omega\left(Y_1^*, Y_1^*\right)c_1 + \cdots + \omega\left(Y_1^*, Y_k^*\right)c_k = \omega(Y_1^*, Y_D) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\omega\left(Y_k^*, Y_1^*\right)c_1 + \ldots + \omega\left(Y_k^*, Y_k^*\right)c_k = \omega(Y_k^*, Y_D) \end{aligned}$$

The constant c_0 is determined from the equation:

$$c_0 + c_1 E(Y_1^*) + \ldots + c_k E(Y_k^*) = E(Y_D).$$

The error s_I is of the form:

$$s_I = \left|\begin{array}{c} \omega\left(Y_1^*, Y_1^*\right) \ldots \omega(Y_1^*, Y_D) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \omega(Y_1^*, Y_D) \cdots \omega(Y_D, Y_D) \end{array}\right| : \left|\begin{array}{c} \omega\left(Y_1^*, Y_1^*\right) \cdots \omega(Y_1^*, Y_k^*) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \omega\left(Y_k^*, Y_1^*\right) \cdots \omega(Y_k^*, Y_k^*) \end{array}\right|.$$

II. The second is expressed as the sum $c_1 Y_1^* + \ldots + c_k Y_k^*$. The coefficients c_1, …, c_k minimalize the expression $E(c_1 Y_1^* + \ldots + c_k Y_k^* - Y_D)^2$. Thus, the error estimation is $s_{II}(Y_D, Y_1^*, \ldots, Y_k^*) = \min E(c_1 Y_1^* + \ldots + c_k Y_k^* - Y_D)^2$. The constants $c_1, \ldots, c_k$ are given by the following normal equations:

$$\begin{aligned} &E\left(Y_1^* Y_1^*\right)c_1 + \ldots + E\left(Y_1^* Y_k^*\right)c_k = E(Y_1^* Y_D) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ &E\left(Y_k^* Y_1^*\right)c_1 + \ldots + E\left(Y_k^* Y_k^*\right)c_k = E(Y_k^* Y_D) \end{aligned}$$

where $E(Y_i^* Y_j^*) = \omega(Y_i^*, Y_j^*) + E(Y_i^*)E(Y_j^*)$.

The error s_{II} is calculated as:

$$s_{II} = \left|\begin{array}{cc} E\left(Y_1^* Y_1^*\right) \ldots E\left(Y_1^* Y_k^*\right) & E(Y_1^* Y_D) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ E\left(Y_k^* Y_1^*\right) \ldots E\left(Y_k^* Y_k^*\right) & E(Y_k^* Y_D) \\ E\left(Y_D Y_1^*\right) \ldots E\left(Y_D Y_k^*\right) & E(Y_D Y_D) \end{array}\right| : \left|\begin{array}{c} E\left(Y_1^* Y_1^*\right) \ldots E\left(Y_1^* Y_k^*\right) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ E\left(Y_k^* Y_1^*\right) \ldots E\left(Y_k^* Y_k^*\right) \end{array}\right|.$$

III. The third is written as the sum $c_1 Y_1^* + \ldots + c_k Y_k^*$. The coefficients c_1, …, c_k minimalize the expression $E(c_1 Y_1^* + \ldots + c_k Y_k^* - Y_D)^2$ and the constraint $E(c_1 Y_1^* + \ldots + c_k Y_k^*) = E(Y_D)$ should be satisfied. The Method of Lagrange Multipliers is applied. The error estimation is $s_{III}(Y_D, Y_1^*, \ldots, Y_k^*) = \min_{E(c_1 Y_1^* + \ldots + c_k Y_k^*) = E(Y_D)} E(c_1 Y_1^* + \ldots + c_k Y_k^* - Y_D)^2$. Hence, the constants $c_1, \ldots, c_k$ are given by the following equations:

$$\begin{aligned} &\omega\left(Y_1^*, Y_1^*\right)c_1 \ldots \omega\left(Y_1^*, Y_k^*\right)c_k + \lambda = \omega(Y_1^*, Y_D) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\omega\left(Y_k^*, Y_1^*\right)c_1 + \ldots + \omega\left(Y_k^*, Y_k^*\right)c_k + \lambda = \omega(Y_k^*, Y_D) \end{aligned}$$

$$c_1 + \ldots + c_k = |D|.$$

The error s_{III} has the form:

$$\begin{aligned} s_{III} = {} & c_1\left\{c_1\omega\left(Y_1^*, Y_1^*\right) + \cdots + c_k\omega\left(Y_1^*, Y_k^*\right) - \omega\left(Y_1^*, Y_D\right)\right\} \\ & + \cdots + c_k\left\{c_1\omega\left(Y_k^*, Y_1^*\right) + \cdots + c_k\omega\left(Y_k^*, Y_k^*\right)\right. \\ & \left.-\omega\left(Y_k^*, Y_D\right)\right\} - \left\{\omega\left(Y_D, Y_1^*\right) + \cdots + \omega\left(Y_D, Y_k^*\right)\right. \\ & \left.-\omega(Y_D, Y_D)\right\}. \end{aligned}$$

IV. The fourth is given as the sum $cY_1^* + \ldots + cY_k^*$. The coefficient c minimalizes the expression $E(cY_1^* + \ldots + cY_k^* - Y_D)^2$ and the constraint $E(cY_1^* + \ldots + cY_k^*) = E(Y_D)$ should be satisfied. The Method of Lagrange Multipliers yields the solution. The error estimation is $s_{IV}(Y_D, Y_1^*, \ldots, Y_k^*) = \min_{E(cY_1^* + \ldots + cY_k^*) = E(Y_D)} E(cY_1^* + \ldots + cY_k^* - Y_D)^2$.

Taking into consideration the constraint $E(cY_1^* + \ldots + cY_k^*) = E(Y_D)$ and equations (11) and (12), the constant c is described by:

$$c = \frac{|D|}{k}.$$

The error s_{IV} is:

$$s_{IV} = c^2 \sum_{i,j} \omega\left(Y_i^*, Y_j^*\right) - c \sum_i \omega\left(Y_i^*, Y_D\right) - \sum_i \omega\left(Y_i^*, Y_D\right) + \omega(Y_D, Y_D).$$

V. The fifth is defined by the constant C, such that $E(Y_D) = C$. This constant minimalizes the expression $E(Y_D - C)^2$. The error estimation is $s_V(Y_D) = \min E(Y_D - C)^2$. Hence, $C = E(Y_D)$ and $s_V = \omega(Y_D, Y_D)$.

The smallest error estimation indicates the best estimator of the integral Y_D. Thus, by applying statistical criteria the k members of a cluster in the region D and the best estimator of the integral Y_D are chosen.

This method can be widely used in almost all researches where the clustering is useful. For the purpose of this study an iterative approach for automated clustering as addition to the Zubrzycki method is proposed. The algorithm is divided into two stages. In the first stage, the size of the region D should be estimated. To find it the sample correlation function $R(d)$ is evaluated using the positions of sunspot groups. For each value $R(d_i)$ ($i = 1, 2, \ldots$) the statistical significance is tested. The null hypothesis H_0: $R(d_i) = 0$ (the alternative hypothesis H_1: $R(d_i) \neq 0$ is verified). In the null case the statistic

$$U = \frac{R(d_i)}{1 - [R(d_i)]^2} \sqrt{n},$$

is approximately normally distributed. This means that the hypothesis H_0 is rejected when $|U| > u(1 - \frac{\alpha}{2})$, where $u(1 - \frac{\alpha}{2})$ is the quantile of the standard normal distribution. The smallest value d_i for which the null hypothesis is approved gives the value d^*. Therefore, for the values $R(d_i)$ such that $d_i < d^*$ the best fitting functions (Models I–III) are computed. Each of these functions is nonlinear. Moreover, the measurement error for each fitted value r_i is different (because each r_i is calculated for different number of pairs n_i). In such a case the value of the χ^2 goodness-of-fit decides about the quality of the fit. The smallest χ^2 value determines the best model. To test that the model is assumed to be qualitatively correct, the Pearson χ^2 test is also used. The second stage is a cluster search stage, where for a given Carrington rotation the sunspot groups with low distance to each other are found. Suppose that during this Carrington rotation the m of sunspot groups were observed. The position of the lth ($l = 1, \ldots, m$) group defines the centre of the region D. Suppose, moreover, that the number of groups in the region D is k. For these groups five estimators (I–V) of the integral Y_D and their errors are computed. The smallest error gives the optimal set of constants c_j, ($j = 0, 1, \ldots, k$) and the best estimator of the integral Y_D. Thus, for each of m groups, which its position defines the region D, one can find a set of groups located in the region D, the best estimator of the integral Y_D and the corresponding error estimation. Among all the estimators ($l = 1, \ldots, m$) the set of groups for which the error estimation is smallest is chosen. This set forms the first cluster. Sunspot groups, which form this cluster, are not taken into account when the next clusters are searched. Thus, each group belongs to one cluster only. For the other groups, which do not create the first cluster, the procedure is repeated from the second stage, until the groups remain single, not forming clusters or until the set of groups under consideration is empty. For the next Carrington rotation the clusters are searched from the second stage.

3. Results

To describe the statistical properties of clusters the sunspot data from the both hemispheres and from cycles 12 to 23 together are studied. The division into the ascending, the maximum and the descending phases of the solar cycles in each solar hemisphere are also considered, but the best fitting is obtained for the entire sunspot data set. Moreover, the values d^* calculated for different phases of solar cycles are similar, although for the ascending and the descending phases of solar cycles statistics of pairs created values r_i for $i < 7$ rotation are poor.

The sample correlation function with the distance d between sunspot groups from the entire data set is shown in Figure 1a. It can be seen that the correlation coefficient $R(d_i)$ for $d_i = 7$ rotations is close to zero. This is confirmed by the test for the presence of correlation. For $d_i = 7$ rotations the H_0 hypothesis is approved (p-value is 0.6). Thus, the value d^* is 7° and the region D is defined as the square 12° × 12°. Figures 1b–1d present the sample correlation functions with the distance $d < d^*$ (dashed lines) and the best fittings (solid lines) for Models I–III respectively. Table 1 shows the estimated parameters a, c, and the value of the χ^2 goodness-of-fit for Models I–III. The χ^2 statistic is designed to test the null hypothesis, H_0, that the distribution of outcomes is consistent with each of models. This statistic is asymptotically distributed as the χ^2 distribution with $k - p - 1$ degrees of freedom, where k is the number of bins and p is the number of parameters of each model. For each model $k - p - 1 = 7 - 2 - 1 = 4$ degrees and the p-value is close to 1. This means that all models give very good fit, but Model II is the best.

In order to find all sunspot group clusters during each Carrington rotation, the region D for each sunspot group position (u, v) is determined, such that $D = \{(z,w)$: $z \in [u - d^* + 1, u + d^* - 1]$, $w \in [v - d^* + 1, v + d^* - 1]\}$. Therefore, for each sunspot group and for all sunspot groups from its region D the five estimators I–V are computed and the smallest error estimation is chosen. Note, that during a given rotation each sunspot group could be close to one another. Thus, for this group and for all groups which belong to its region D the best integral Y_D and the best estimation error could be calculated. From these errors the smallest is chosen. Such a sunspot group set is treated as a cluster. This procedure is repeated for the remaining sunspot groups (which do not form the cluster) until all clusters are evaluated.

The method presented above is applied to sunspot groups from three different phases (the ascending, the maximum and the descending) of 12 solar cycles for each solar hemisphere. For the ascending phases about 26% of all sunspot groups create the clusters (1156 groups from the northern hemisphere form 513 clusters and 950 groups from the southern hemisphere form 426 clusters). For the solar maxima this percentage is about 35% of groups create the clusters (4242 groups from the northern hemisphere form 1832 clusters and 3941 groups from the southern hemisphere form 1740 clusters). For the descending phases about 30% of all sunspot groups create the clusters (2170 groups from the northern hemisphere form 947 clusters and 2179 groups from the southern hemisphere form 976 clusters). This study has revealed, that for the fixed size of the region D and for both hemispheres the percentage of groups creating clusters is almost the same regardless of the solar cycle phase.

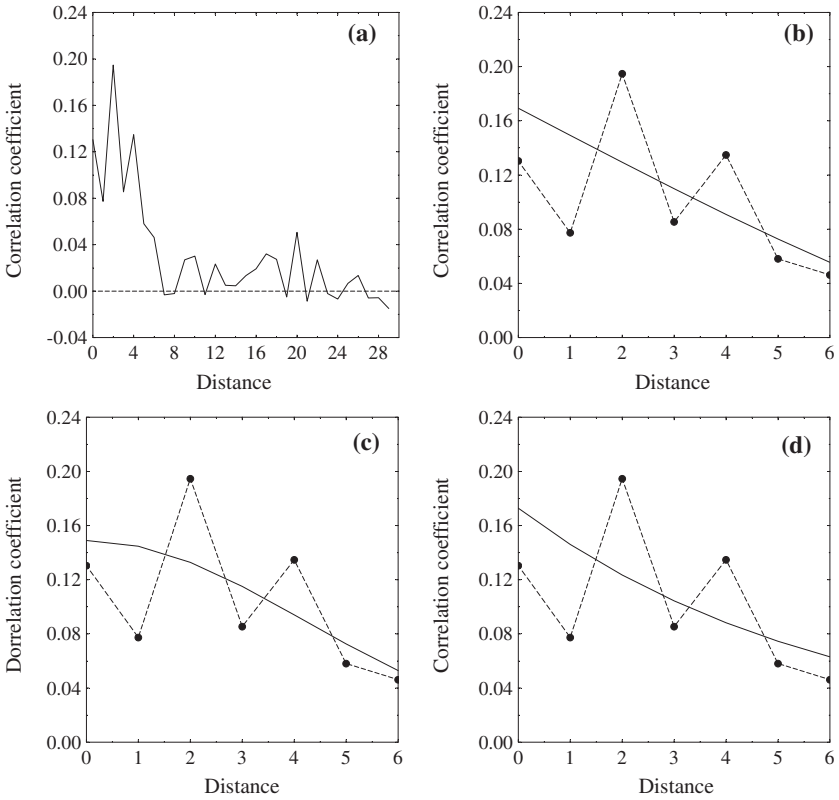

Fig. 1. (a) The sample correlation function $R(d)$ for $d = 0, \ldots, 29$ between sunspot group areas from the entire data set. (b) The sample correlation function $R(d)$ for $d = 0, \ldots, d^*$ (dashed line) and the best fitting for the Model I (solid line). (c) The same as for (b), but for the Model II. (d) The same as for (b), but for the Model III.

To analyse the evaluated clusters the number distribution functions of the sunspot groups creating one cluster for each phase of 12 solar cycles in each solar hemisphere are determined. Figure 2a demonstrates the histograms of number of sunspot groups creating one cluster for the phases of ascending in the northern hemispheres (filled triangles) and in the southern hemisphere (open circles). The same, but for the maximum and the descending phases, is plotted in Figures 2d and 2g. The shapes of these histograms are essentially the same. Moreover, they are roughly linear in the log-linear scaled plot (Figs. 2b, 2e, 2h and Figs. 2c, 2f, 2i for the northern and the southern hemisphere respectively). Almost all points are inside the 95% confidence bands (dashed lines). The correlation coefficient and the regression equation are given in the left lower corner of each figure. As is easily seen the correlation coefficients are almost -1 for the phases of ascending. For the maximum and descending phases they are less than -0.97. Moreover, after removing a few clusters which contain more than 7 sunspot groups for the phases of maximum the correlation is almost -1 and all histogram points are inside the confidence bands. The same is for the phases of descending, but after removing a few clusters which contain more than five sunspot groups. To compare the histograms for both hemispheres the regression lines in the log-log scaled plot are presented. Figures 3a–3c show them for the phases of ascending, maximum and descending respectively. The correlation coefficients and the regression equations are presented in the right lower corners. For all three cases the dependence is very strong (the correlation coefficients are almost 1) and the slopes of regression lines are close to 1 which means that these histograms are almost the same. Similarity in the shape of size distribution functions at different phases of the solar cycle also is found for bipolar active regions (Harvey & Zwaan 1993) for sunspot umbrae (Bogdan et al. 1988) and for Ca II H and K plages, except for the smaller (Schrijver 1988).

Table 1. Values of the parameters a, c and the value of the χ^2 goodness-of-fit for Models I–III.

Model	a	c	χ^2
I	10.7	0.17	0.0013
II	5.9	0.15	0.0011
III	5.9	0.17	0.0014

The number distribution functions of the sunspot groups creating one cluster calculated for each of 12 solar cycles (cycles 12–23) in each solar hemisphere are also investigated. As in Figure 2 the histograms of the number of groups creating one cluster for cycles 12–23 in each solar hemisphere shown in Figure 4 indicate that 70–90% of all clusters in each solar cycle are formed by two sunspot groups, 9−21% of all clusters are formed by three sunspot groups. The remaining clusters are created by four and more (up to eight) sunspot groups. The shape of all distribution functions is also invariant solar feature. The histograms of the number of groups creating one cluster for cycles 12–23 calculated for each solar hemisphere separately (Figs. 5 and 6) are roughly linear in the log-linear scaled plot (the correlation coefficients are close to -1). The correlation between the slope of the regression line and the strength of each cycle is not statistically significant. For the northern hemisphere it is 0.34 (p-value is 0.28) and the southern hemisphere it is 0.1 (p-value is 0.81). The dependence between the histogram points from the northern and the southern hemispheres for each solar

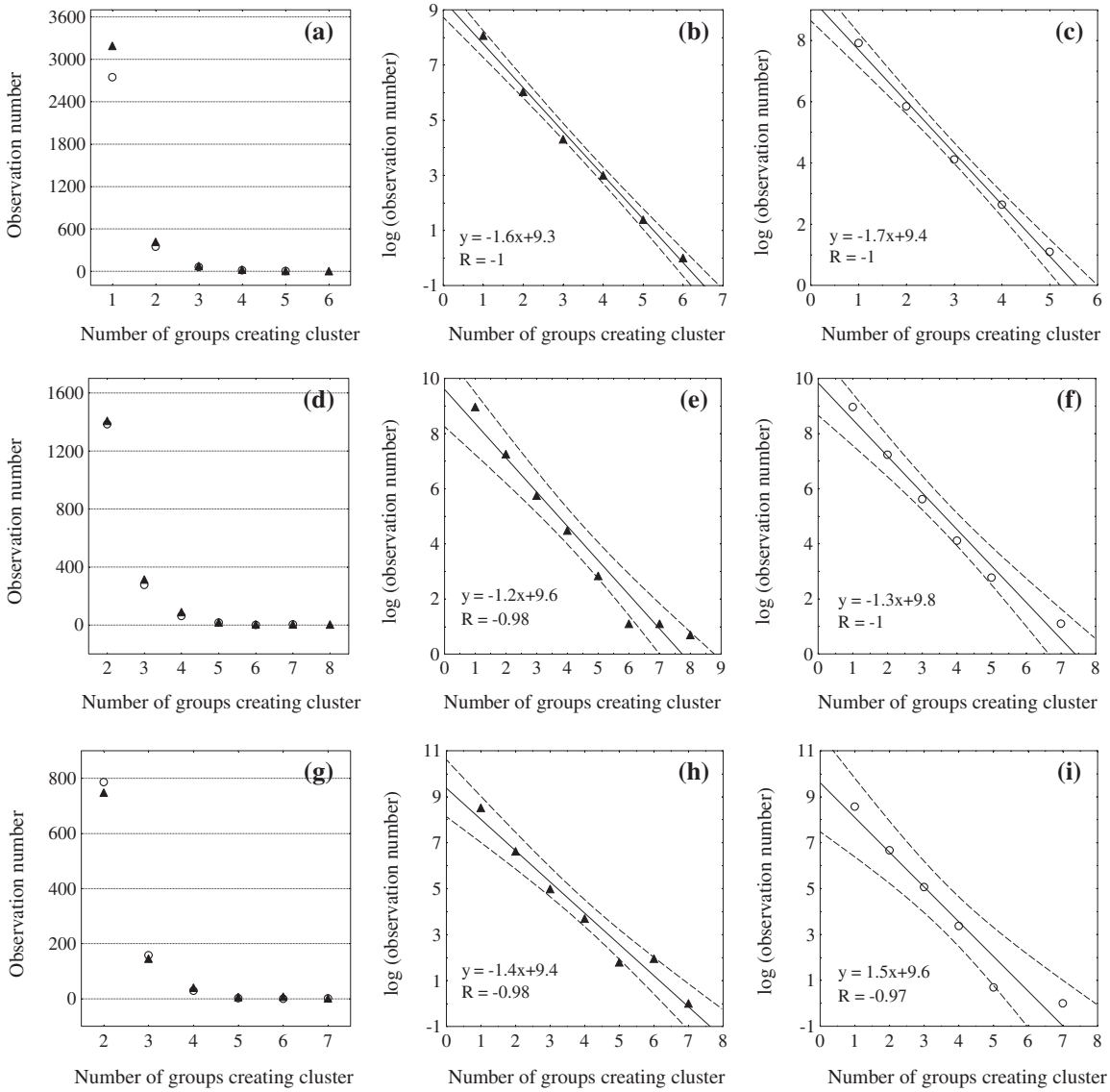

Fig. 2. *The upper row*: (a) the histograms of number of sunspot groups creating one cluster for the ascending phases of solar cycles in the northern (filled triangles) and the southern hemispheres (open circles). (b) The fit of a regression line (solid line) with the 95% confidence bands (dashed lines) to the histogram points for the ascending phases of solar cycles in the northern hemisphere (filled triangles), but in the log-linear scaled plot. The correlation coefficient and the regression equation are given in the left lower corner. (c) The same as in (b), but for the histogram points from the southern hemisphere (open circles). *The middle row*: the same as in the upper row, but for the maximum phases of solar cycles. *The lower row*: the same as in the upper row, but for the descending phases of solar cycles.

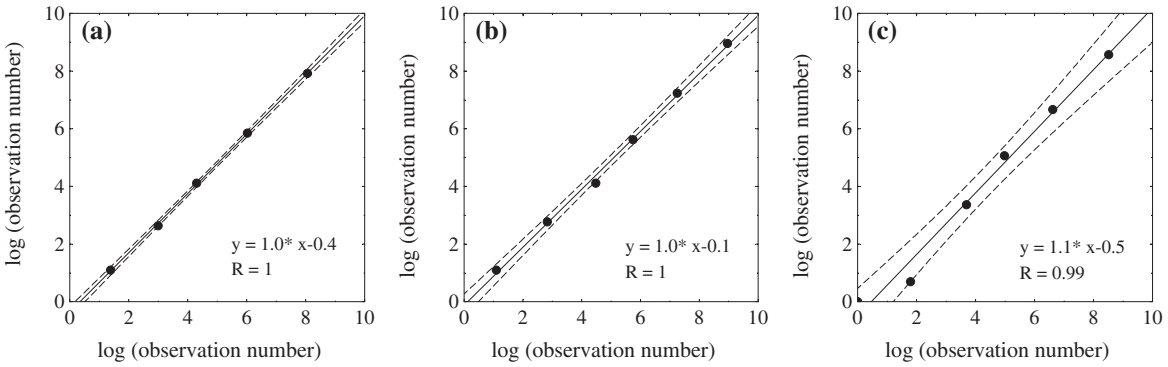

Fig. 3. (a) The fit of a regression line (solid line) with the 95% confidence bands (dashed lines) to the histogram points of number of sunspot groups creating one cluster from the northern and the southern hemispheres for the ascending phases of solar cycles in the log-log scale plot. The correlation coefficient and the regression equation are presented in the right lower corner. (b) The same, but for the maximum phases of solar cycles. (c) The same, but for the descending phases of solar cycles.

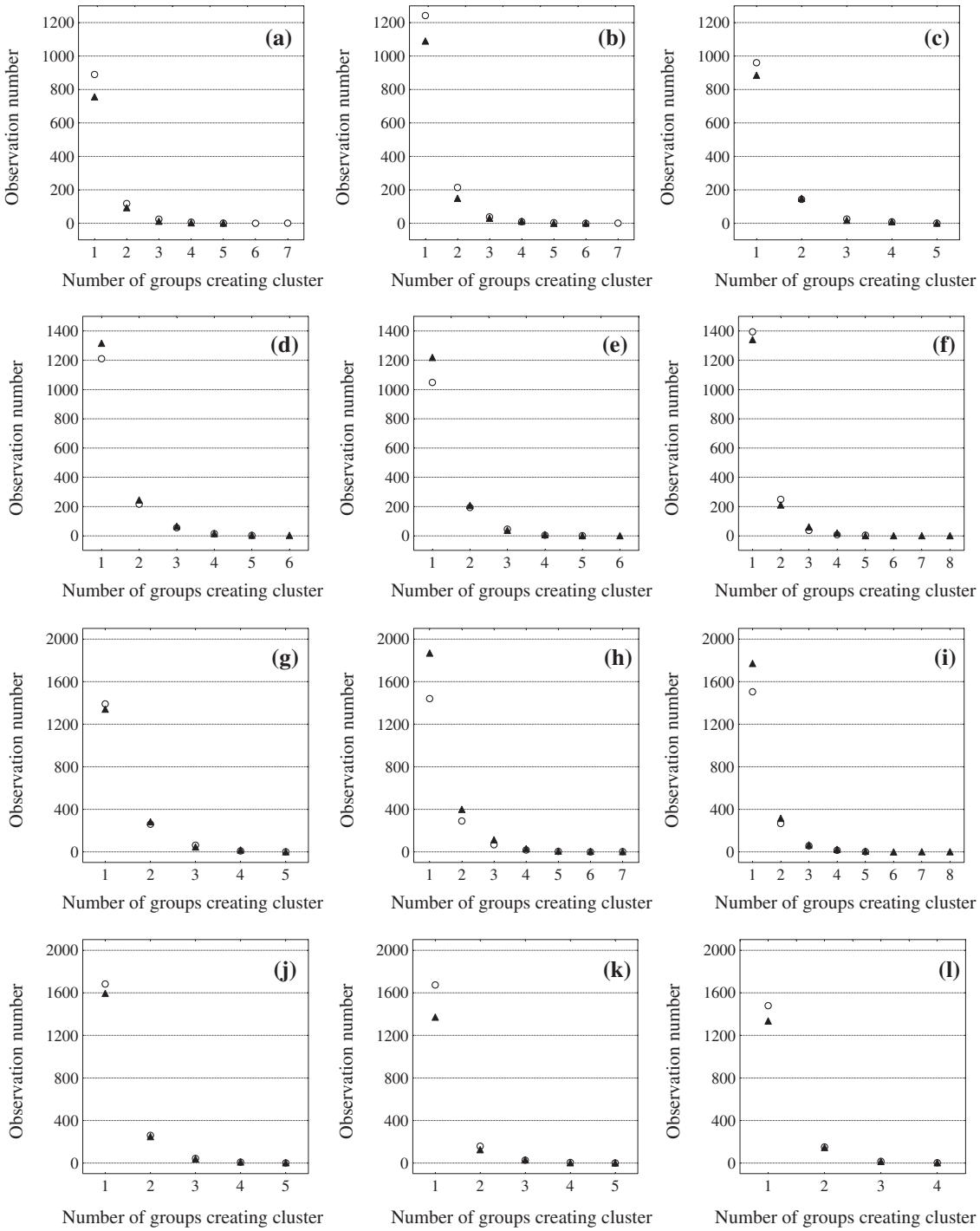

Fig. 4. The histograms of number of sunspot groups creating one cluster in the northern (filled triangles) and the southern hemispheres (open circles) for (a) cycle 12, (b) cycle 13, (c) cycle 14. (d–f) The same as in the upper row, but for cycles 15–17. (g–i) The same as in the upper row, but for cycles 18–20. (j–l) The same as in the upper row, but for cycles 21–23.

cycle in the log-log scaled plot is demonstrated in Figures 7 and 8. The correlations between them are between 0.94 (for cycle 17) and 1 (±0.001) for cycles 14, 15, 16, 19 and 23. Most clusters were found for cycle 19 (556 in the northern hemisphere and 380 in the southern hemisphere), the smallest number of clusters was found for cycle 12 (108 in the northern hemisphere and 150 in the southern hemisphere). The presence of the N-S asymmetry between cluster numbers in each solar cycle needs more investigations in future.

4. Discussion

The Zubrzycki method proves to be successful in localizing sunspot clusters. In many clustering methods the personal judgement enters in the adoption of the numerical criteria. These clustering criteria influence the scale and the properties of the resulting clusters. Zubrzycki proposes to estimate the size of the region where the clusters are searched. The sample correlation function between the positions of sunspot groups

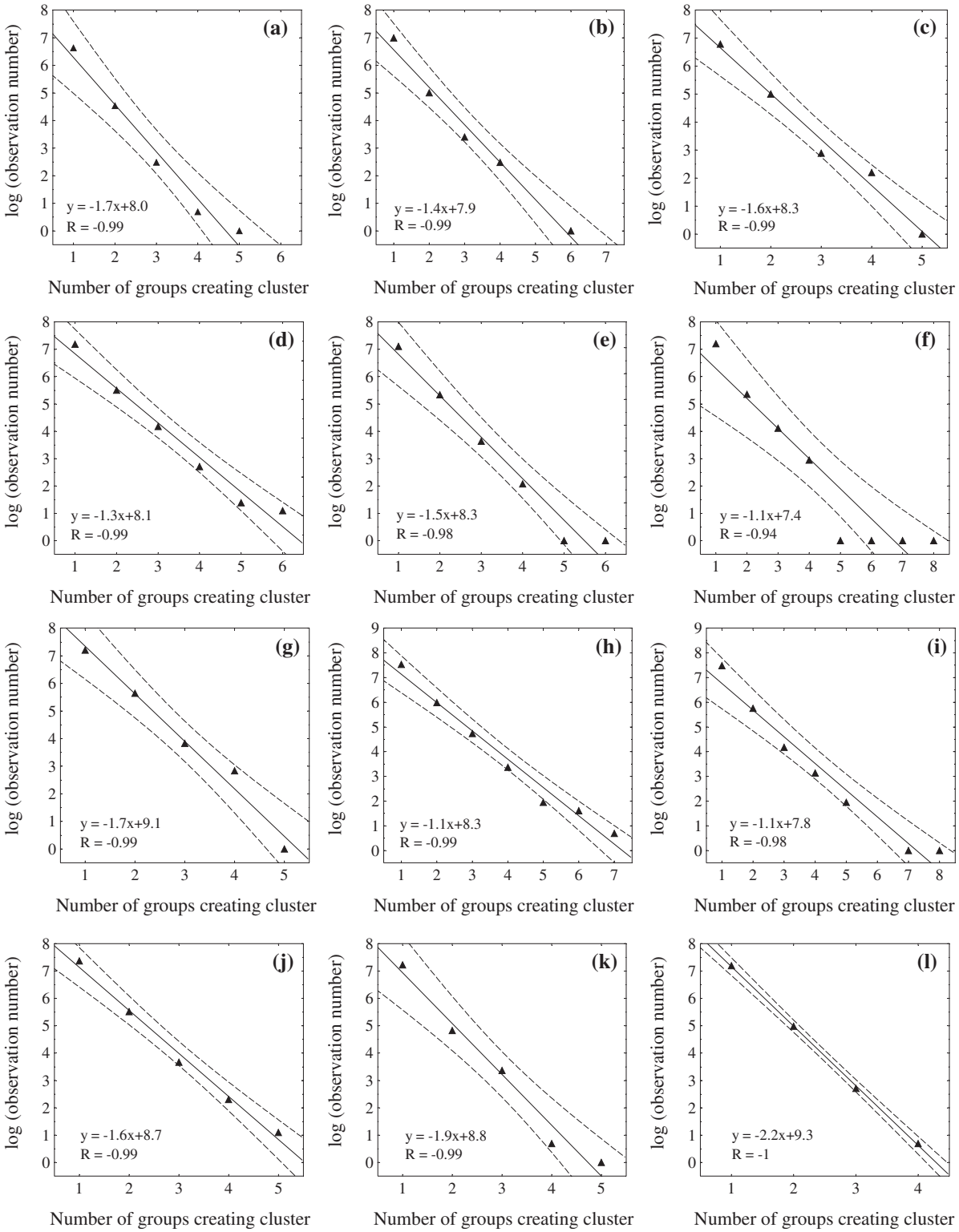

Fig. 5. (a) The fit of a regression line (solid line) with the 95% confidence bands (dashed lines) to the histogram points of number of sunspot groups creating one cluster from the northern hemisphere in the log-linear scale plot for (a) cycle 12, (b) cycle 13, (c) cycle 14. (d–f) The same as in the upper row, but for cycles 15–17. (g–i) The same as in the upper row, but for cycles 18–20. (j–l) The same as in the upper row, but for cycles 21–23. The correlation coefficients and the regression equations are presented in the left lower corners.

forming the clusters determines the sizes of the region D for the entire sunspot data set (cycles 12–23 for both solar hemispheres). Thus, most clusters have the extension in longitude up to 12°. A clustering tendency was also investigated by several authors. Castenmiller et al. (1986) consider the sunspot groups from the descending phase of cycle 19. They search the sunspot nest which is defined as a relatively small space on the solar surface where a number of sunspot groups appear and disappear, one after the other. These possible nests are localized by eye. Their extension in longitude is typically between 5° and 25° and in latitude between 3° and 13°. After correction their nests occupy areas smaller than 120° square

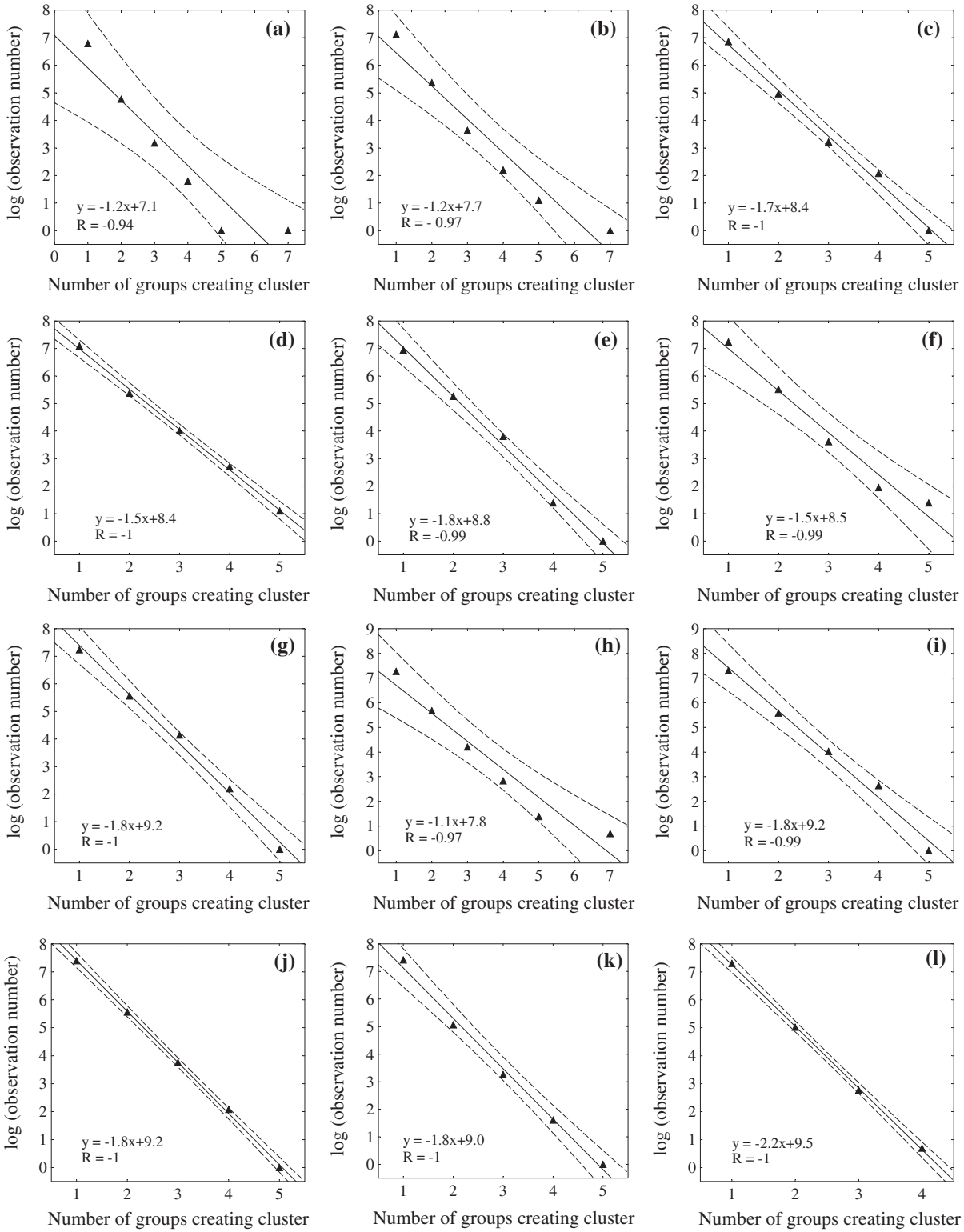

Fig. 6. The same as in Figure 5, but for the southern hemisphere.

deg. The significant difference between the complex defined by Gaizauskas and the nest is that the nest does not show a tendency to expand or to contract during its lifetime and there are time gaps (up to 2.5 rotation) in the appearance of new active regions. Despite these differences both are of the same nature. Brouwer & Zwaan (1990) used the same data as Castenmiller et al. (1986), but the clustering method was a standard mathematical technique of cluster analysis. They defined different clustering criteria and considered more compact nests. The majority of their nests have areas smaller than about 30° square deg. The authors emphasize that there is a clear tendency for small-scale clusters to cluster once again in larger clusters. They found the double nests for which the distance between components in longitude is less than 25°. The double nests were also found by Castenmiller et al. (1986). Both the authors compare their components to the branches of some of their

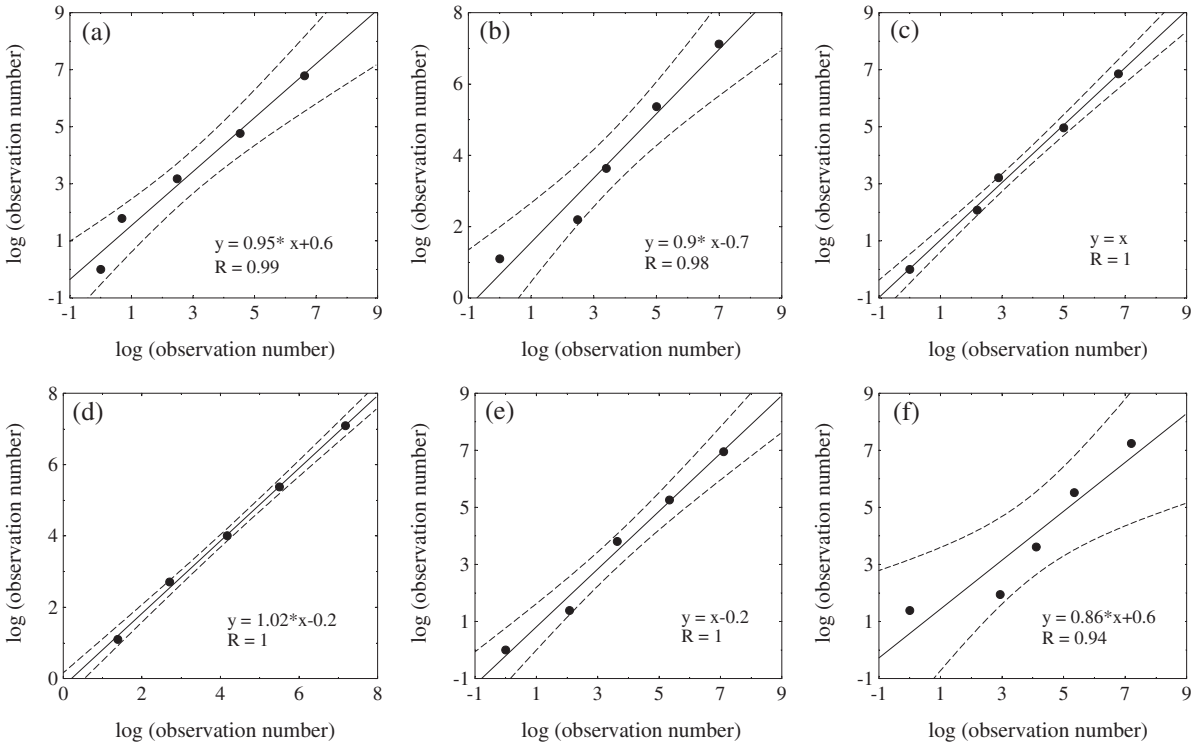

Fig. 7. The same as in Figure 3, but for cycles 12–17. The upper row shows the regression lines for cycles 12–14, the lower row presents them for cycles 15–17.

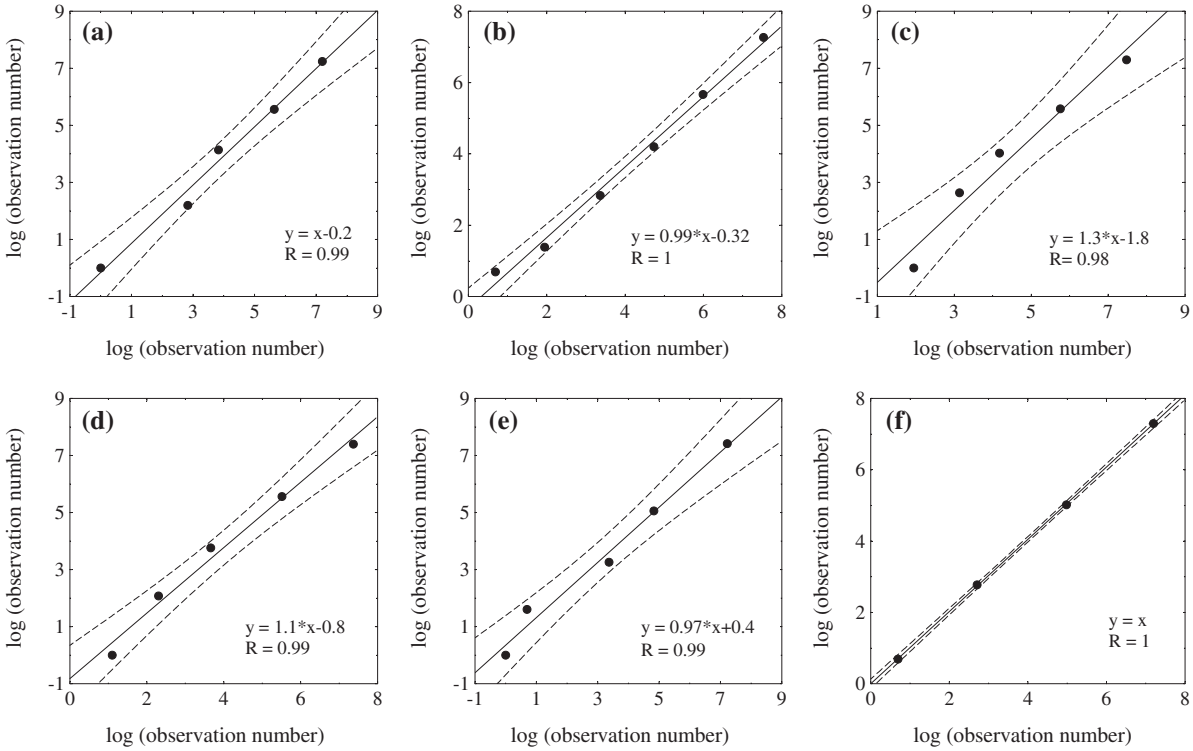

Fig. 8. The same as in Figure 3, but for cycles 18–23. The upper row shows the regression lines for cycles 18–20, the lower row presents them for cycles 21–23.

"complexes of activity", which were found by Gaizauskas et al. (1983). Moreover, Brouwer & Zwaan (1990) convince the reader of a large-scale in the clustering tendency of active regions as a next step in the complex large scale patterns in the appearance of magnetic flux in solar atmosphere. This could be confirmed by the sample correlation function (Fig. 1a), which contains two positive peaks for $\tau \in [10°, 12°]$. Such peaks were also found in the correlation functions calculated

for the ascending (at $\tau = 10°, 13°$), the maximum (at $\tau = 10°, 12°$) and the descending (at $\tau = 9°$) phases of solar cycles in each solar hemisphere.

To estimate the area of a cluster the weights are evaluated by the best fitting of the sample correlation function. The presented models (I–III) assume that the correlation between the cluster members decreases with the distance between them. Such an assumption gives very good fitting. Moreover, a detailed analysis indicates that in majority of cluster members the best estimator of the integral Y_D is created on the basis of the same weights. This means that the integral Y_D can be estimated by the arithmetic average of the member areas. Note, that such estimated values of the integral Y_D are not used in this study and can be applied in future investigations.

An application of the Zubrzycki method to sunspot data also carries negative consequences due to the assumption of stationarity. This is particularly evident when during the solar maxima very strong and significantly weaker sunspot groups appear. Then, the assumptions of the constant expected values and the constant variances in every points of the region D are false. This fact has a significant impact on the estimated value of the integral Y_D. For such cases, the estimated integral may be less than the area of the largest group. However, a statistical analysis of all cases (all possible clusters for all Carrington rotations) shows that for almost all clusters (99% of all clusters) the weights $c_1, \ldots, c_k$ are equal and the arithmetic average of the member areas should be adopted to estimate the integral Y_D.

It is also worth noting that the method determines clusters in the two-dimensional space (longitude-latitude) for each rotation separately. It may therefore be the first step to determine the long-lived structures such as complexes of activity and active longitudes and consequently be the starting point for lots of studies in various field of solar physics: evolution of them, interaction between them, fragmentation of flux tube, periodicities in solar activity, asymmetry of solar hemispheres and modelling of the magnetic flux emergences through the photosphere.

5. Conclusions

The following results have been obtained:

1. For the ascending and the descending phases of solar cycles about 30% of all sunspot groups create clusters. For the solar maxima they give about 35% of all groups. The number distribution function of the sunspot groups creating one cluster decreases with increasing number.
2. The number distribution function of the sunspot groups creating one cluster is roughly linear in the log-linear scale independent of the phase of a solar cycle.
3. The shape of number distribution function is also preserved for each solar cycle in each solar hemisphere.
4. The hemispheric imbalance in cluster numbers during each solar cycle is found.

Acknowledgements. I would like to thank Prof. B. Kopocinski for the helpful dicussions and for the suggestion to apply the Zubrzycki method to the sunspot group data. I also thank the referees for careful consideration of the paper, valuable criticism, comments and suggestions.

References

Bogdan, T.J., P.A. Gilman, I. Lerche, and R. Howard, Distribution of sunspot umbral areas – 1917–1982, *Astrophys. J.*, **327**, 451–456, 1988.

Brouwer, M. P., and C. Zwaan, Sunspot nests as traced by a cluster analysis, *Sol. Phys.*, **129**, 221–246, 1990.

Castenmiller, M.J.M., C. Zwaan, and E.B.J. van der Zalm, Sunspot nests – Manifestations of sequences in magnetic activity, *Sol. Phys.*, **105**, 237–255, 1986.

Gaizauskas, V., K.L. Harvey, J.W. Harvey, and C. Zwaan, Large-scale patterns formed by solar active regions during the ascending phase of cycle 21, *Astrophys. J.*, **265**, 1056–1065, 1983.

Getko, R., Activity complexes as a cause of strong positive monthly Wolf number fluctuations, *Sol. Phys.*, **224**, 291–301, 2004.

Harvey, K.L., *Magnetic bipoles on the Sun*, PhD thesis. Utrecht University, 1993.

Harvey, K.L., and C. Zwaan, Properties and emergence of bipolar active regions, *Sol. Phys.*, **148**, 85–117, 1993.

Parnell, C.E., C.E. DeForest, H.J. Hagenaar, B.A. Johnston, D.A. Lamb, et al., A Power-law distribution of solar magnetic fields over more than five decades in flux, *Astrophys. J.*, **69**, 75–82, 2009.

Schrijver, C.J., Radiative fluxes from the outer atmosphere of a star like the Sun – A construction kit, *A&A*, **189**, 163–172, 1988.

Schrijver, C.J., and K.L. Harvey, The photospheric magnetic flux budget, *Sol. Phys.*, **150**, 1–18, 1994.

Tang, F., R. Howard, and J.M. Adkins, A statistical study of active regions 1967–1981, *Sol. Phys.*, **91**, 75–86, 1984.

van Driel-Gesztelyi, L., E.B.J. van der Zalm, and C. Zwaan, Active nests on the Sun, *Proc. of the National Solar Observatory/ Sacramento Peak 12th Summer Workshop*, ASP Conference Series, ASP, San Francisco, **27**, 89, 1992.

Zubrzycki, S., About an estimation of the parameters of geological deposits, *Wiadomosci matem.*, **III**, 105–153, 1957.

Cite this article as: Getko R: Statistics of sunspot group clusters. J. Space Weather Space Clim., 2013, **3**, A09.