Cluster analysis for pattern recognition in solar butterfly diagrams

E. Illarionov[1], **D. Sokoloff**[2], **R. Arlt**[3⋆], and **A. Khlystova**[4]

[1] Department of Mechanics and Mathematics, Moscow State University, Moscow 119991, Russia
[2] Department of Physics, Moscow State University, Moscow 119992, Russia
[3] Leibniz-Institut für Astrophysik Potsdam, An der Sternwarte 16, D-14482 Potsdam, Germany
[4] Institute of Solar-Terrestrial Physics, Siberian Branch, Russian Academy of Sciences, Irkutsk 664033, Russia

The dates of receipt and acceptance should be inserted later

Key words magnetic fields – sunspots – solar cycles – Sun

We investigate to what extent the wings of solar butterfly diagrams can be separated without an explicit usage of Hale's polarity law as well as the location of the solar equator. We apply two algorithms of cluster analysis for this purpose, namely DBSCAN and C-means, and demonstrate their ability to separate the wings of contemporary butterfly diagrams based on the sunspot group density in the diagram only. Then we apply the method to historical data concerning the solar activity in the 18th century (Staudacher data). The method separates the two wings for Cycle 2, but fails to separate them for Cycle 1. In our opinion, this finding supports the interpretation of the Staudacher data as an indication of the unusual nature of the solar cycle in the 18th century.

1 Introduction

The physical nature of the solar activity cycle is associated with an activity wave propagation from solar mid-latitudes to the solar equator. The waves can be easily seen in the well-known butterfly diagram as inclined wide strips of increased sunspot density. More specifically, a cycle contains one wave in each solar hemisphere. Sunspots, as places where magnetic lines of an initially toroidal magnetic field rise above the solar surface or return back into the solar interior, can be associated with each other in sunspot groups. A polarity of a given group can be introduced according to the magnetic field direction in the leading spot of the group. The polarity law by Hale says that each of the activity waves (wings of the butterfly diagram) in a given cycle has a preferred polarity of sunspot groups, while the northern and southern activity waves of a given cycle have opposite polarities. Activity waves in a given hemisphere in two consecutive cycles have opposite polarities as well.

In previous studies, two approaches were employed to separate cycle wings. The first approach relies on the magnetic field sign in the leading polarity of sunspot groups or Hale's polarity law (Hale & Nicholson 1925; Richardson 1948; Mouradian & Soru-Escaut 1993). The second approach uses the latitudinal migration of the emergence locations of sunspot groups, also known as Spörer's law. It implies that sunspots appear at mid-latitudes at the beginning of a sunspot cycle; the sunspot formation zone then gradually migrates to the equator, reaching latitudes below 10° at the end of the cycle. The mean latitudes of sunspot groups are usually determined by the center of mass for the northern and southern hemispheres, but boundaries of the cycle branches are determined visually. Li et al. (2001) suggested to identify the wings of butterfly diagram using the migration rate of sunspot groups in the course of a solar cycle.

The ability of the visual method is limited, however. Several percent of the sunspot groups do not follow Hale's polarity law (Khlystova & Sokoloff 2009; Sokoloff & Khlystova 2010 and references therein). A consistent determination of the exact number of the groups which violate the law requires, however, a well-defined algorithm for the activity wave identification. Sometimes it is not obvious to which cycle and activity wave a particular sunspot group belongs. This makes an explicit formulation of an algorithm desirable which separates the groups between cycles and activity waves. The desired algorithm is expected to avoid an explicit usage of the group polarities to be used in particular to estimate the accuracy of Hale's polarity law. Of course, the activity waves isolated by the desired algorithm should agree in general with the visual impression, but a deviation from the traditional identification may be used to extract some physical information concerning the underlying dynamo action.

The desired algorithm being of some interest for contemporary activity cycles becomes especially important for the historical butterfly diagrams obtained from archive data (e.g. Ribes & Nesme-Ribes 1993; Arlt 2009) because the polarity of the groups can be determined for the last century only.

The aim of this paper is to present a method for the above described aim and discuss how it works with contemporary as well as historical data. The method suggested is based on the cluster analysis technique which has already been exploited to study sunspot nests. Brouwer & Zwaan (1990) examined the properties of sunspot nests using a

⋆ Corresponding author: rarlt@aip.de

single-linkage clustering technique. Petrovay & Abuzeid (1991) found the existence of several levels of sunspot clusters by means of a Bayesian iteration procedure associated with the cluster analysis. Note however that the identification of wings in butterfly diagrams was out of the scope of these papers.

2 Data

The sunspot data of the Royal Greenwich Observatory and the USAF/NOAA[1] for the Cycles 14–22 (1874–2010) and the historical data from the Staudacher sunspot record (Arlt 2009) for the Cycles 1–4 (1749–1797) were used. While we tested the method for the Cycles 14–22, we will show the results for the Cycles 21 and 22 exemplarily below and add Cycles 1 and 2 from the historical data. For the first data set we used the time of the first observation of a sunspot group and corresponding latitude. Historical data are however much less detailed and sunspots are not separated in sunspot groups. For the second data set we used the latitude and time of observation of each sunspot. For the sake of consistence, we applied the same approach for the first data set to conclude that the shape of the clusters identified are stable after such modification of the data set (see below, Sect. 4).

3 Clusters of sunspot groups

According to the aim of this study, we have to base the desired method on the expectation that the sunspot group density in a wing of the butterfly diagram is higher than the one in the surrounding area. We do not prescribe the expected shape of the wing, so we refer to the wings just as clusters.

In order to determine the sunspot group density in the butterfly diagram, we have to define how to calculate the areas in the diagram. The point is that the time t and the latitude β have different units. We introduce dimensionless quantities by measuring time in units of the nominal cycle length (11 years) and latitude in units of the typical maximum latitude at which sunspot groups occur (40°).

We start the process of cluster identification from the most straightforward algorithm known as DBSCAN (Sander et al. 1998; the GDBSCAN algorithm mentioned in the title of their paper is a generalization of DBSCAN which is unimportant for our aims).

3.1 The DBSCAN algorithm

The underlying idea of the algorithm can be described as follows: two points in the butterfly diagram belong to the same cluster if they can be connected by a sequence of circles of a given radius r, where each of them contains at least m cluster points. In the Greenwich/USAF record, these cluster points are sunspot groups, while in the historical record, the cluster points are individual sunspots.

Speaking more specifically, the algorithm works as follows. Let $B_r(p)$ be a circle with the center p and the radius r. We refer to a point q which belongs to $B_r(p)$ as directly attainable from p if $B_r(p)$ contains more then m points. A point q is referred to as density-reachable from p if there is a sequence $p_1, ..., p_n$, $p_1 = p, p_n = q$ so that the following point is directly attainable from the previous one. We refer to two points p and q as density-connected if both of them are density-reachable from a point w. The algorithm under discussion identifies a set of density-connected points in the diagram as a cluster. If a spot located in a point p does not belong to a cluster and $B_r(p)$ contains less then m points it is considered as noise. The subtle distinction between density-reachable and density-connected points is important if a cluster contains two parts connected by a short and narrow bridge. This can happen in principle, however, is not very realistic for the butterfly shapes.

Note that the above defined clusters do not depend on the starting point chosen to build a cluster nor on the order of testing the points on attainment with the cluster points. An attractive property of the method is also that the form of clusters is not prescribed from the beginning.

The choice of the parameters r and m affects the result as follows. By reducing r (for a given m) we reproduce more details of the shape of a cluster. However, if r is too small, a cluster which represents a physical entity can be artificially separated into several clusters. By enlarging r we reduce the number of points which are not included in clusters (reduce the noise), but we take the risk to join several real clusters to a single, artificial one. In particular, we take the risk of joining artificially the northern and southern wings of a butterfly.

By enlarging m (for a given r) we enlarge the density of points in clusters, reduce the spatial size of the clusters, enlarge the noise and take the risk of separating a physical cluster into several artificial ones. By playing with m and r, one can match the shape of the cluster obtained with the visual impression.

The algorithm was initially developed for a general case where no *a priori* information concerning the expected shape of the clusters is available. The choice of r and m is then based on the shape of the cumulative distribution function of the distances between the neighboring points in the diagram. Applying the recommendations given to Cycle 21 yields $r = 0.03$ and $m = 4$. We tested the algorithm with these values of r and m to learn that it separates the following cycle from the preceding one reasonably well, but fails to separate the northern and southern wings of the butterfly.

Then we chose r and m deviating from the expected values for the shape of a typical wing. We presumed that the diameter ($2r$) should be about $1/10$ of the expected size of the wing (i.e. about 1.1 yr in time and about 4° in latitude according to our above definition) and m should be about 1% of the total amount of the spots in the wing (about

[1] http://solarscience.msfc.nasa.gov/greenwch.shtml

40 groups) and varied the parameters in the vicinity of these raw estimates to get a separation between cycles and wings, i.e. to obtain a result which resembles the visual impression. Note that the raw estimates of r and m ($r = \bar{r} = 0.05$ and $m = \bar{m} = 40$) mean that the critical point density is close to the mean density of sunspot groups in the diagram in the course of the cycle. This looks like a natural choice.

As a result we find that we have to take m and r somewhat larger than $\bar{m}$ and $\bar{r}$, i.e. $r = 0.07$ and $m = 80$ (Fig. 1). Note that the density of points in the circles remained the same as for $\bar{m}, \bar{r}$, since both r and m were increased. Smaller values of r and m result in a breakup of the wing into several smaller clusters.

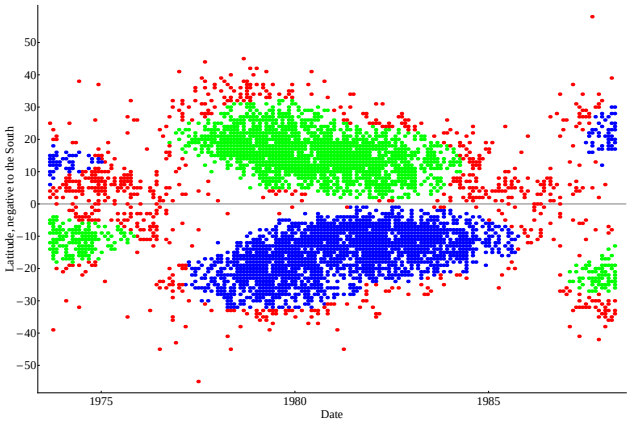

Fig. 1 Clusters isolated by the DBSCAN algorithm for Cycle 21 ($r = 0.07$, $m = 80$). Green and blue colors indicates the two clusters of each cycle, while red is for sunspot groups which do not belong to any cluster. Here and below time is measured in years and latitude is measured in degrees.

We learn from Fig. 1 that large clusters isolated by the algorithm reproduce the visual impression of the wings of the butterfly reasonably well. The amount of points which was identified as noise is rather small, about 5% of the total number of spots.

We checked that increasing r or m reduces the quality of the pattern recognition. In particular, an increase of r for a given m (up to $r = 0.12$) prevents the algorithm from separating the northern and southern wings of the butterfly (Fig. 2). On the other hand, an increase of m for a given r (up to $m = 160$) reduces the sizes of the clusters and considers many spots which belonged to the cluster before as noise (contrary to the visual impression; Fig. 3).

We applied the above method of choosing the governing parameters r and m to several other contemporary cycles from number 14 to 22 and arrived at the conclusion that the method successfully separates cycles and wings of the butterfly shapes.

The cluster identification based on the DBSCAN algorithm is not completely perfect though. It would be desirable to include the halo of red points (noise) surrounding each wing in the clusters as far as possible.

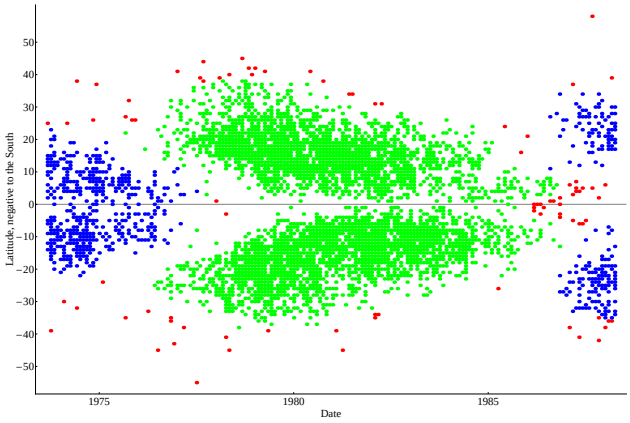

Fig. 2 The DBSCAN algorithm fails to separate the northern and southern wings of the butterfly for large r ($r = 0.12$, $m = 80$) for Cycle 21. Green and blue colors indicate the clusters in the adjacent cycle and red is for sunspot groups which do not belong to any cluster.

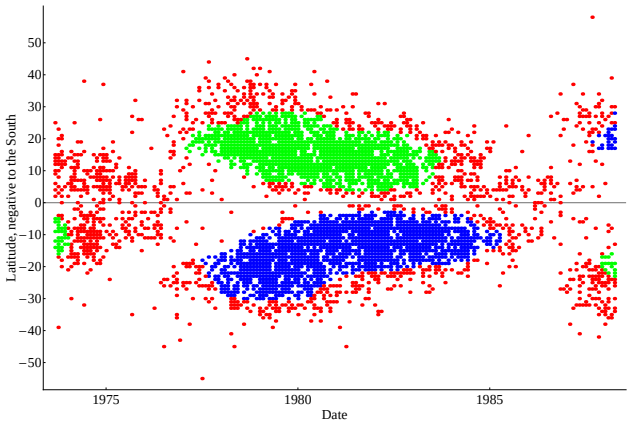

Fig. 3 The DBSCAN algorithm artificially enhances the noise for large m ($r = 0.07$, $m = 160$) for Cycle 21. Green and blue colors indicate the clusters in adjacent cycles and wings while red is for sunspot groups which do not belong to any cluster.

3.2 The C-means clustering algorithm

We tried to improve the results of the DBSCAN algorithm for contemporary cycles in order to attribute as many points as possible in any of the wings of a butterfly. For this purpose we also used the widely exploited C-means clustering algorithm (e.g. Bezdek 1981).

The C-means algorithm is designed to isolate clusters of elliptical shape (which is realistic for the wings of butterflies) provided the number of clusters and approximate position of the cluster centers are known. We presume that a butterfly has two wings of elliptical form and take the position of the centers from the results of the DBSCAN algorithm for the centers of mass of each cluster.

The C-means algorithm works as follows. We consider a cluster isolated by the DBSCAN algorithm as a realization of a 2D Gaussian random quantity. The coordinates of all the spots belonging to the cluster are considered inde-

pendent measurements of that random quantity. Following standard methods, we calculate the mean position μ_i^c and the covariance matrix a_{ij}^c ($i,j = 1,2$ where the first coordinate is time and the second is latitude) of the 2D random quantity which correspond to each cluster. Here c enumerates the clusters. (We recall that the diagonal elements of a_{ii}^c are variances of the temporal and latitudinal coordinates.)

Then we re-define the distance of a spot x^p which belongs to the given cluster from the center of this cluster, μ^c, as

$$r(x^p, \mu^c) = a_{ij}^{-1}(x_i^p - \mu_i^c)(x_j^p - \mu_j^c), \quad (1)$$

where a_{ij}^{-1} is the inversion of a_{ij} and p is the number of an individual spot in the c-th cluster.

The following step of the algorithm consists of a redistribution of the spots among the clusters (for a constant number of clusters K and given quantities μ_i^c and a_{ij}^c) by minimizing the function

$$S = \sum\sum r(x^h, \mu^k), \quad (2)$$

where the first sum is taken over the clusters ($k = 1 \ldots K$) while the second one is taken over the spots which are included in the k-th cluster.

Then the algorithm corrects the values of μ_i^c and a_{ij}^c according to the spots which are included in the cluster at this stage and perform the above steps once more. The algorithm iterates until the positions of the cluster centers (μ_i^c) and their shapes (a_{ij}^c) are not changing significantly anymore. The process usually takes 10–15 iterations.

While separating the spots into the clusters, we take into account the possibility that a spot can be very remote from any cluster. Then the algorithm considers it noise. More precisely, we attribute a spot at the point x^p to the c-th cluster if

$$r(x^p, \mu^c) < u, \quad (3)$$

where u is a number of standard deviations at which a point can be remote from the cluster center to be still attributed to this cluster. If a spot is located at a comparable distance from several clusters we compare with u the weighted quantity $r(x^p, \mu^c)\sum_k r(x^p, \mu^k)^{-1}$. The number of points attributed as noise decays with u. We checked that for $u = 2.5$ the number of points attributed as noise is about 2% for all contemporary and historical cycles investigated. Taking into account that about several percent of sunspot groups violate the Hale polarity law (Khlystova & Sokoloff 2009) we consider $u = 2.5$ an appropriate value.

We show the clusters isolated by the C-means clustering algorithm for Cycles 21 and 22 in Figs. 4 and 5, respectively. On the one hand, we conclude from these figures that almost all sunspot groups are contained in two clusters for each cycle. One of the cluster is located almost exclusively in the northern hemisphere while the other one is located almost entirely in the southern hemisphere. These are the wings of the butterfly shapes with their opposite polarities. On the other hand, the algorithm identifies some of the spots at the very end of the wings as noise. Several spots which we would visually attribute to the past cycle are actually associated with the upcoming one.

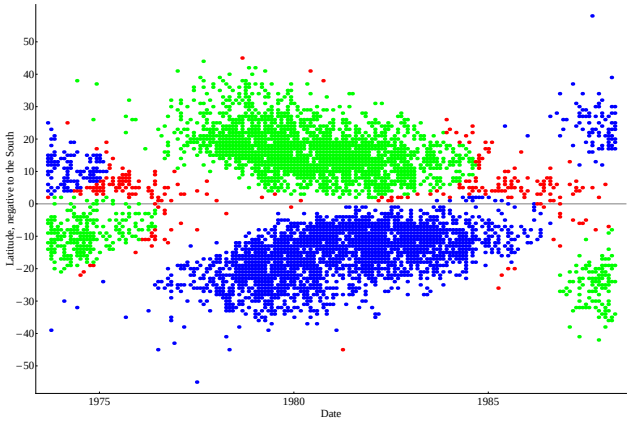

Fig. 4 Clusters (plotted in green and blue) isolated by the C-means clustering algorithm for Cycle 21. Red shows the sunspot groups attributed to noise.

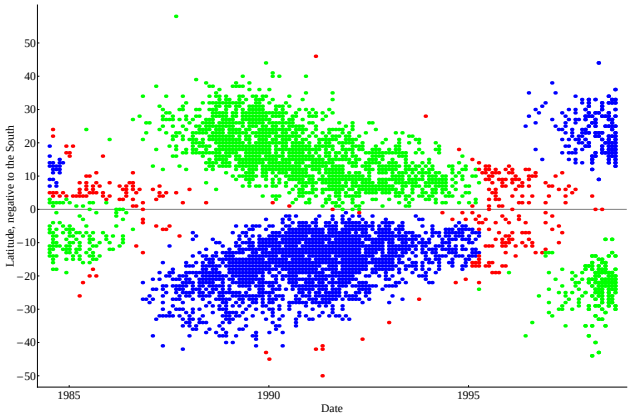

Fig. 5 Clusters (plotted in green and blue) isolated by the C-means clustering algorithm for Cycle 22. Red shows the sunspot groups attributed to noise.

4 Some modifications

Here we verify that the method suggested is stable under moderate variations in the approach.

First, we demonstrate in Fig. 6 that the method gives similar results if we use all individual sunspot groups instead of only the positions at the time of the first observation of a group. For a comparison of the method applied to the two sets, we choose the domain where the spots identified as noise are located. We see that these sets of points are slightly different in the two panels, but the main shape of the clusters remains stable.

Secondly, we check how stably our method works when data sets contain less information since this is important for historical data. We first excluded in a random way two thirds of the sunspot groups from the data of Cycle 21, then we exclude from Cycle 21 the months missing at the same phase of Cycle 2 in the Staudacher data. The resulting data set thus has the same scarcity of data as Cycle 2. Then we apply the cluster identification. As we can see in Fig. 7, the shape of the clusters as well as the positions of points asso-

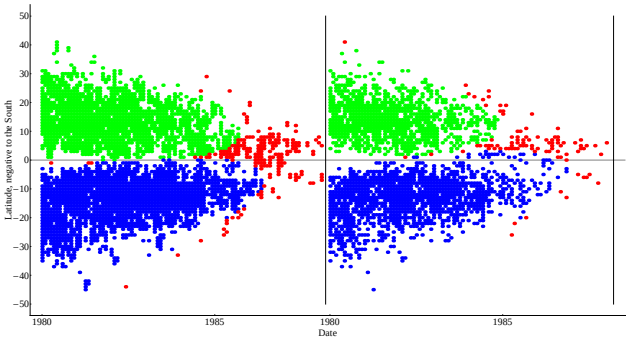

Fig. 6 Comparison of the identification of the wings of the butterfly diagram of Cycle 21, once with all individual sunspot groups (left) and once only for the time of the first observation of a group (right).

ciated with noise remain stable after the significant decrease in the number of points.

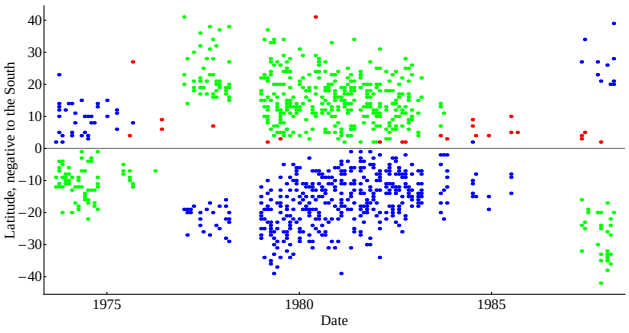

Fig. 7 Clusters isolated by the C-means clustering algorithm for Cycle 21 after two thirds of points were thrown out in a random way and some months of observations were ignored, mimicking the scarcity of data during Cycle 2.

We appreciate that there is a variety of particular cluster analysis methods similar to the one employed above. We verified that the results obtained remain stable if we use the so-called EM-algorithm (Dempster et al. 1977) instead of the similar C-means algorithm (Fig. 8, to be compared with Fig. 5; the critical probability required by this method is chosen as 85%). We see that both methods give similar results, while Fig. 8 exhibits a somewhat different set of noise points than Fig. 5. Note that the EM-algorithm was previously used by Petrovay & Abuzeid (1991) for their analysis.

5 Historical cycles

We applied the above methods to the historical data concerning solar activity in the 18th century as observed by Johann Staudacher. The amateur astronomer in Nuremberg collected about 1000 drawings of the solar disk and a number of additional annotations in a book covering 1749–1799. The drawings were digitized and measured by Arlt (2008)

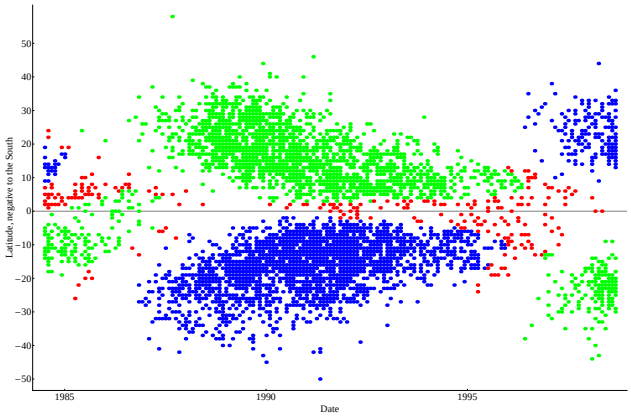

Fig. 8 Clusters identified for Cycle 22 by the EM-algorithm. Points for which the probability of belonging to a cluster is lower than 85% are shown in red.

and Arlt (2009). The difference to modern data is that the information concerning the polarity of sunspot groups is not available so one cannot exploit Hale's polarity law to separate the wings of the butterfly shapes. There is also the peculiarity that the shape of the sunspot distribution in the butterfly diagram for Cycle 1 indicates that the magnetic configuration could have been that of a quadrupolar field rather than a dipolar one at that time.

When applying the DBSCAN algorithm, we have to take into account that the cycle in the 18th century was weaker than the contemporary ones, and much fewer spots were noticed with Staudacher's small telescope. We thus have to choose appropriate values for r and m required for the DBSCAN algorithm. We tried several combinations and found that $r = 0.2$ and $m = 200$ are preferable for Cycle 2, while $r = 0.18$ and $m = 100$ are preferable for Cycles 3 and 4. These values of m mean that the density of the spots in a circle should be about 3 times higher than the mean spot density in the diagram. This choice allowed us to separate the northern and southern wings of the butterflies (Fig. 9). We were unable to separate the wings of the butterfly for Cycle 1. This result supports the visual impression that the symmetry properties of the magnetic configuration in Cycle 1 was quite different from the one in later cycles. We consider the butterfly diagram at Cycle 1 as just one cluster. The choice of $r = 0.18$ and $m = 100$ occurs adequate for such an interpretation.

Then we improved the results of DBSCAN using the C-means algorithm and presumed that Cycle 1 contains just one cluster while the butterfly for Cycle 2 contains two wings (Fig. 10). Figure 11 shows the results for Cycle 3 and Cycle 4.

6 Conclusion and discussion

Based on our results we conclude that the cluster analysis methods provide a reasonable tool to separate solar cycles as well as the wings of butterfly diagrams in an algorithmic

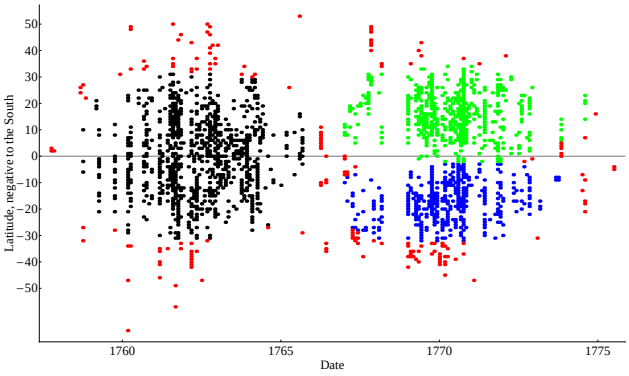

Fig. 9 Clusters isolated by the DBSCAN algorithm for Cycles 1 and 2. Green and blue are for the two wings of the butterfly of Cycle 2 which presumably have opposite polarities. The single cluster obtained for Cycle 1 is given in black. Red is for the spots considered noise.

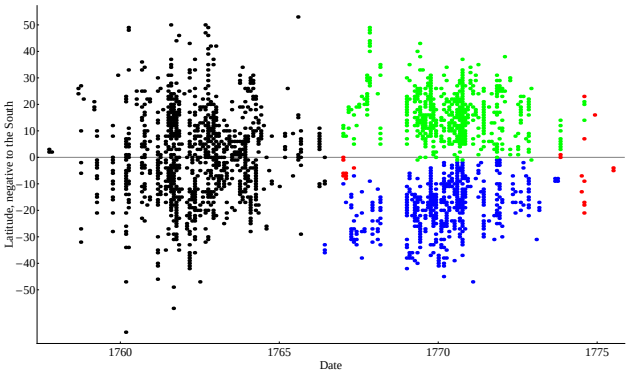

Fig. 10 Clusters isolated by the C-means algorithm for Cycles 1 and 2. The color coding is the same as in the previous figure.

way. The separation of the wings and cycles using the above algorithms does not use any information about the polarity of sunspot groups. The methods can thus be applied even if the polarity is unknown.

The analysis supports the concept that Cycle 1 as it was recorded by Johann Staudacher is quite different from later ones. It is the only cycle among the cycles under investigation where the method fails to separate the wings of the butterfly. This fact can be understood as an indication that the solar magnetic field had a quadrupolar symmetry at that time (Sokoloff et al. 2010) rather than a dipolar symmetry as in contemporary cycles. Note that stellar dynamo theory provides (Moss et al. 2008) a possibility to generate such magnetic configurations apart from traditional dipolar ones for various dynamo models.

Note that the C-means algorithm gives, as a by-product, a possibility to estimate the migration rate for the wings in the butterfly diagram. It simply corresponds to the inclination of the ellipses, which approximate the clusters obtained, with respect to the time axis. Estimates for the migration rate obtained for Cycles 21 and 22 are given in Table 1. This approach gives no pronounced migration for Cycle 1, again in accordance with a quadrupolar configuration

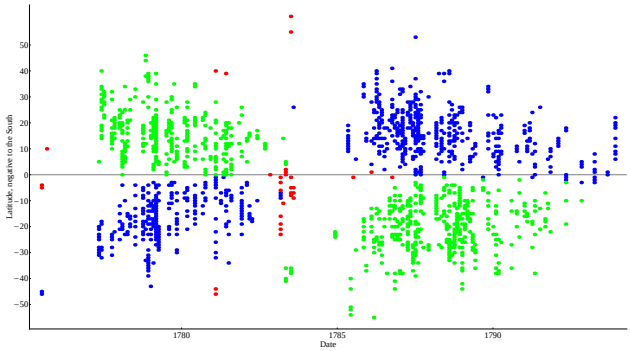

Fig. 11 Clusters isolated by the C-means algorithm for Cycles 3 and 4. The color coding is the same as in Fig. 9.

(Moss et al. 2008). The formal application of this method then means that the activity wave of this cycle is almost a standing one. Note however that the shape of the wings of Cycle 2 (Fig. 10) is quite different from the one of contemporary cycles. It looks plausible that the beginning of Cycle 2 demonstrates something like a poleward propagating activity wave which can be regarded as yet another manifestation of the exceptional nature of solar activity in the 18th century. Note that the ellipses which we obtain for the Cycles 3 and 4 look unrealistically steep as well, and we did not include the corresponding estimates in Table 1.

Table 1 The migration rate in $^\circ\mathrm{year}^{-1}$ for the individual wings in the butterfly diagram.

Cycle	21	22
Northern hemisphere	3.8	4.7
Southern hemisphere	3.4	4.0

Acknowledgements. Partial financial support from RFBR under the grants 09-05-00076 and 10-02-00960-a is acknowledged.

References

Arlt, R.: 2008, Sol. Phys. 247, 399
Arlt, R.: 2009, Sol. Phys. 255, 143
Bezdek, J.C.: 1981, *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York
Brouwer, M.P., Zwaan, C.: 1990, Sol. Phys. 129, 221
Dempster, A.P., Laird, N.M., Rubin, D.B.: 1977, J. Roy. Stat. Soc. B39, 1
Hale, G.E., Nicholson, S.B.: 1925, ApJ 62, 270
Khlystova, A.I., Sokoloff, D.D.: 2009, Astron. Rep. 53, 281
Li, K.J., Yun, H.S., Gu, X.M.: 2001, ApJ 122, 2115
Moss, D., Saar, S.H., Sokoloff, D.: 2008, MNRAS 388, 416
Mouradian, Z., Soru-Escaut, I.: 1993, A&A 280, 661
Petrovay, K., Abuzeid, B.K.: 1991, Sol. Phys. 131, 231
Ribes, J.C., Nesme-Ribes, E.: 1993, A&A 276, 549
Richardson, R.S.: 1948, ApJ 107, 78
Sander, J., Ester, M., Kriegel, H.-P., Xu, X.: 1998, Data Mining and Knowledge Discovery 2, 169
Sokoloff, D., Khlystova, A.I.: 2010, AN 331, 82

Sokoloff, D., Arlt, R., Moss, D., Saar, S.H., Usoskin, I.: 2010, in: A.G. Kosovichev, A.H. Andrei, J.-P.Rozelot (eds.), *Solar and Stellar Variability: Impact on Earth and Planets*, IAU Symp. 264, p. 111