RECONSTRUCTION OF MONTHLY AND YEARLY GROUP SUNSPOT NUMBERS FROM SPARSE DAILY OBSERVATIONS

I. G. USOSKIN[1], K. MURSULA[2] and G. A. KOVALTSOV[3]
[1]*Sodankylä Geophysical Observatory (Oulu unit), P.O.Box 3000, FIN-90014 University of Oulu, Finland (e-mail: Ilya.Usoskin@oulu.fi)*
[2]*Department of Physical Sciences, P.O.Box 3000, FIN-90014 University of Oulu, Finland*
[3]*Ioffe Physical-Technical Institute, 194021 St.Petersburg, Russia*

(Received 9 July 2003; accepted 20 August 2003)

Abstract. Some periods before 1820 are poorly covered by sunspot observations. In addition to apparent, long observational gaps, there are also periods when there are only few sparse daily sunspot observations during a long time. It is important to estimate the reliability of the monthly and yearly mean sunspot values obtained from such sparse daily data. Here we suggest a new method to estimate the reliability of individual monthly means. The method is based on comparing the actual sparse data (sample population) to the well-measured sunspot data in 1850–1996 (reference population), and assumes that the statistical properties of sunspot activity remain similar throughout the entire period. For each sample population we first found those months in the reference population that contain the same data set, and constructed the statistical distribution of the corresponding monthly means. The mean and standard error of this distribution represent the mean and uncertainty of a monthly mean sunspot number reconstructed from sparse daily observations. The simple arithmetic mean of daily values can be adequately applied for months which contain more than 4–5 evenly distributed daily observations. However, the reliability of monthly means for less covered months has to be estimated more carefully. Using the estimated, new monthly values, we have also calculated the weighted annual sunspot numbers.

1. Introduction

While the sunspot numbers (SNs) form the longest series of routine solar observations, some periods are not well covered by observational data. In addition to long observational gaps when sunspot activity is unknown, there are periods when observations were very sparse. Such periods raise the problem how to reconstruct average SN values from sparse daily observations. Usually the monthly mean sunspot number R_m is computed as a simple arithmetic mean of all available daily SN values R_d, i.e., $R_m = \langle R_d \rangle$. However, such a method is uncertain when only few (in the extreme only one) R_d values are available within a month. For example, Hoyt and Schatten (1998) noted that traditional monthly SN values can be reliably estimated only if there are more than 4 daily observations evenly distributed within a month. In this paper we discuss in detail a new statistical method (Usoskin, Mursula, and Kovaltsov (2003) to form the monthly (and yearly) means from isolated daily observations. The advantage of this method is that it allows not

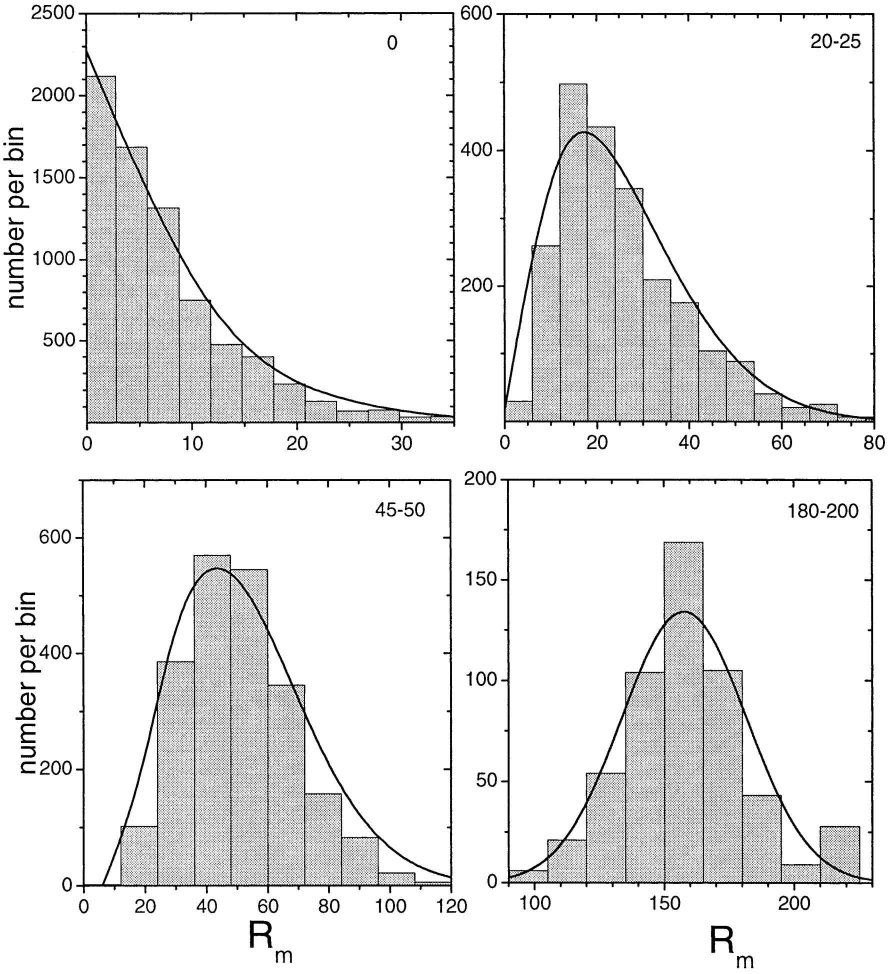

Figure 1. Samples of histogram distributions of monthly R_m together with the rescaled best-fit Poisson distribution functions. The four panels depict the cases of at least one R_d equal to zero or in the interval [20–25], [45–50], [180–200], respectively.

only to calculate the monthly SN value but also to estimate its uncertainty. The method is based on the statistical properties of sunspot activity during the recent, well-covered period, and on the assumption that these properties remain the same throughout the entire period of sunspot observations since 1610. Since the method deals with individual daily SN values, which are not available in the Wolf sunspot number series, we study here the daily group sunspot numbers (GSNs) as presented by Hoyt and Schatten (1998).

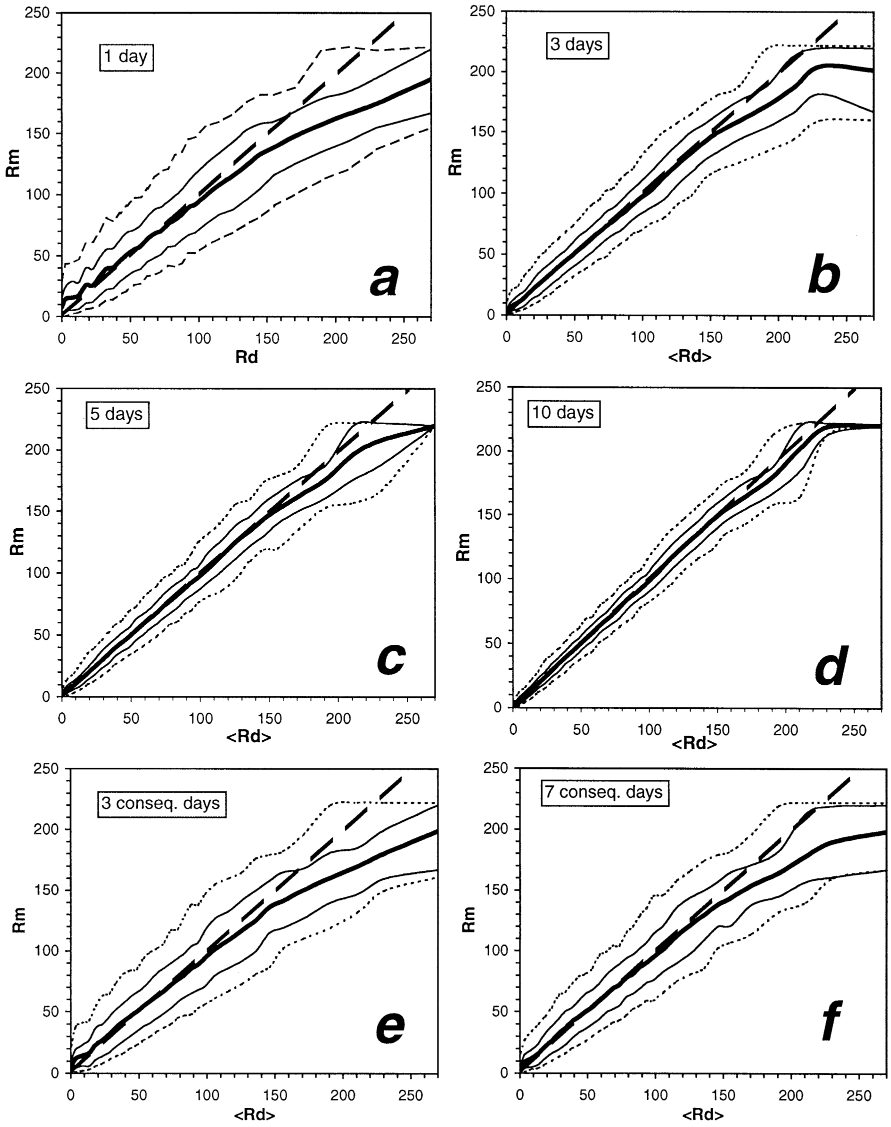

Figure 2. The quality of monthly sunspot numbers calculated from the arithmetic mean of sparse daily values. The horizontal and the vertical axes correspond to the arithmetic mean $\langle R_d \rangle$ and the actual R_m, respectively. Panels (a–d) correspond to 1, 3, 5, and 10 daily observations taken randomly, and panels (e–f) to 3 and 7 days taken consecutively, within a month. *Thick solid*, *thin solid*, and *thin dotted lines* depict the mean, 68% and 95% confidence intervals of the R_m vs. $\langle R_d \rangle$ distribution. *The thick dashed line* denotes the diagonal $R_m = \langle R_d \rangle$.

2. Reconstruction of Monthly Sunspot Numbers

First, we analyzed all daily group sunspot numbers for the period 1850–1996 when the data are reliable and contain no observational gaps. We call this data set (more than 53 000 daily values) the reference population. Then, given one isolated daily sunspot value R_d from the poorly covered sample period, we selected from the reference data set all the days with a daily value close to R_d. The width of the bin for included R_d values was chosen as a compromise between sufficient statistics and resolution: the width of the bin is 5 below 100, 10 for 100–160, 20 for 160–240, and the last bin includes all sunspot values larger than 240. Then we collected the actual monthly means R_m corresponding to these selected days of the reference population. (If more than one appropriate daily value was found within a month, the corresponding R_m value was counted as many times.)

Figure 1 shows samples of histograms of the collected R_m values for R_d equal to zero and within three bins. The histogram distributions are apparently not Gaussian but can be transformed to the Poisson form after scaling the x-axis, i.e., the R_m values. Since the GSN value is the number of sunspot groups G multiplied by a factor of 12.08 (Hoyt and Schatten, 1998), the real statistics behind GSN is the statistics of sunspot groups (rather than sunspot numbers) which have much smaller values. Therefore, if R_g is reduced to G by dividing by a factor $k = 12$, the statistics of $G = R_g/k$ follow the Poisson distribution:

$$f(G, \mu) \sim \frac{\mu^G e^{-\mu}}{G!}, \tag{1}$$

where G is a non-negative integer and μ is the mathematical expectation of the mean. Figure 1 shows the best fit Poisson distributions after rescaling G back to R_g. One can see that these distributions correspond well to the Poisson shape (after rescaling) and approach the Gaussian distribution when increasing R_d.

From such distributions we have computed each monthly mean R_m and its uncertainty σ_m corresponding to one daily R_d value in a month (Figure 2(a)). The usual assumption that $R_m = \langle R_d \rangle$ (thick dashed line) leads to a significant overestimate of the monthly value for $R_d > 100$. If more than one daily observation was available in a month we can still apply the above procedure by looking for the given set of R_d values in the reference population. The corresponding R_m vs. $\langle R_d \rangle$ plots are shown in Figure 2(b–d). The deviation between R_m and $\langle R_d \rangle$ is still significant for three daily observations within a month for $\langle R_d \rangle > 150$, but is small for five observation days and negligible for ten days, in agreement with Hoyt and Schatten (1998). (The horizontal plateau for $\langle R_d \rangle > 220$ is due to lack of statistics for high SN values.)

In the above discussion we assumed that the observational days are taken randomly within the month, i.e., that there are no preferred dates of observations and that individual daily SN values are not correlated. However, it is quite common that daily observations are consecutive and form a single period of a few consecutive

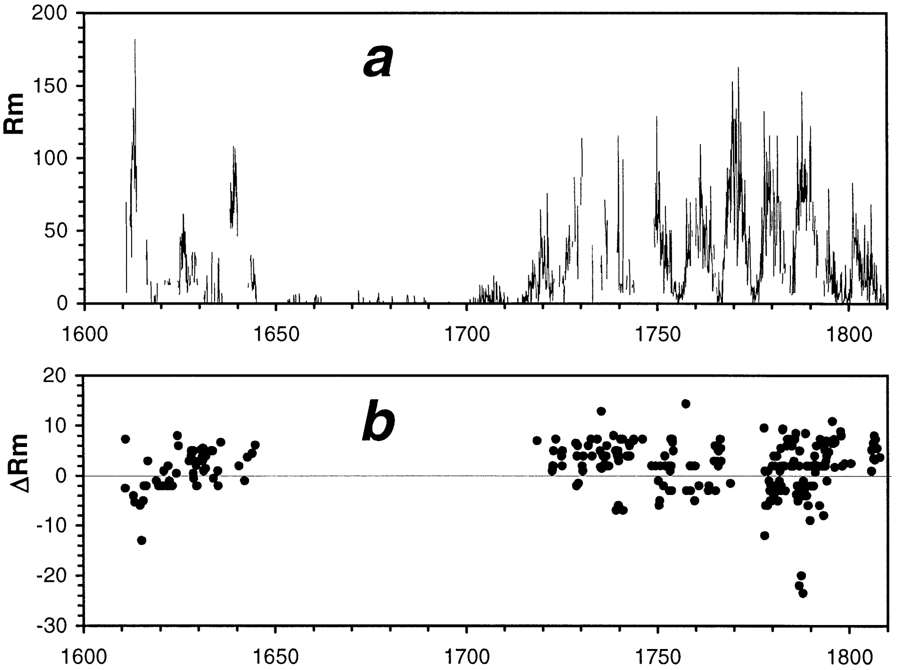

Figure 3. (a) Monthly group sunspot numbers, reconstructed as described in the text. (b) The difference ΔR_m between the formal and newly calculated monthly sunspot numbers.

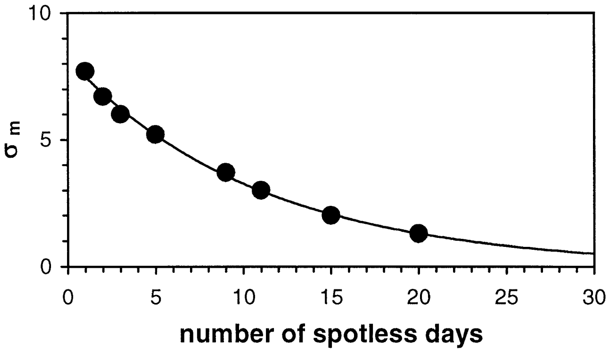

Figure 4. Statistical error of the monthly mean group sunspot number as a function of the number of spotless days within a month from the reference period. The curve depicts the best-fitting exponential.

observational days within a month. In such a case, the individual daily measurements cannot be regarded as random and independent, but the above method can still be applied by looking for the same set of consecutive R_d values. Note that in this case the quality of the R_m reconstruction (see Figures 2(e) and 2(f)) is very close to the single daily observation (Figure 2(a)) because the consecutive observations are strongly correlated.

Thus, using the method illustrated by Figures 1 and 2, one can reconstruct a monthly mean R_m from sparse (or even from a single) daily observations R_d and estimate its uncertainties. Applying this method to all those individual months from the period 1610–1820 that contain five or less separate or 10 or less consecutive daily observations, we have reconstructed the monthly GSN values shown in Figure 3(a). For other months (> 5 evenly distributed or > 10 consecutive daily observations in a month), we took $R_m = \langle R_d \rangle$. The standard error of the mean can be defined in this case as

$$\sigma_m = \sigma_d / \sqrt{n_d - 1}, \tag{2}$$

where σ_d and n_d are the standard deviation and the number of daily R_d values within the month, respectively (Hoyt and Schatten, 1998). The differences between the formal R_m values (Hoyt and Schatten, 1998) and the newly reconstructed monthly values are shown in Figure 3(b). The periods when the reconstruction is clearly different from the formal R_m definition are 1610–1645 and 1710–1810, with the difference remaining typically within ± 10. Only a few months in 1780s show a large difference of about -25. Note also that for most months the difference is positive, indicating that the arithmetic average exaggerates the monthly value, as suggested by Figure 2.

3. Reconstruction of Yearly Sunspot Numbers

The traditional way to obtain yearly sunspot numbers R_y is to compute the arithmetic mean of monthly values R_m, i.e., it is a two-step averaging of daily values $R_y = \langle R_m \rangle = \langle\langle R_d \rangle\rangle$. We note that R_y computed in this way is different from R_y computed directly from all R_d values within the year because the two-step arithmetic averaging (when all monthly values are taken with equal weights) breaks the error propagation if months are not fully covered by daily observations. Therefore, strictly speaking, it is more accurate to calculate the yearly SN from the daily values $R_y = \langle R_d \rangle$ or as a weighted average of monthly values. The weighted average is defined as (Agekyan, 1972; Usoskin, Mursula, and Kovaltsov, 2003)

$$R_w = \frac{1}{w} \sum_{m=1}^{12} w_m R_m, \tag{3}$$

where individual statistical weights are $w_m = 1/\sigma_m^2$, and $w = \sum w_m$. The uncertainty in R_w is determined as follows:

$$\sigma_w = \begin{cases} (\sigma^* + \sigma)/2 & \text{if} \quad \sigma^* < \sigma, \\ \sigma^* & \text{if} \quad \sigma^* > \sigma, \end{cases} \tag{4}$$

where

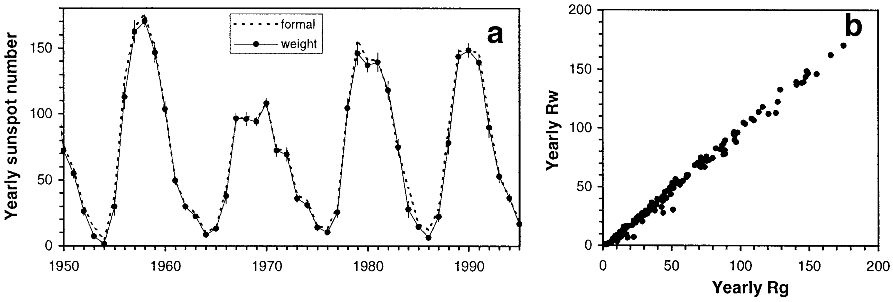

Figure 5. (a) Formal (*dotted curve*) and weighted (*solid curve with error bars*) yearly group sunspot numbers for 1950–1996. (b) Scatterplot of weighted vs. formal yearly group sunspot numbers for 1850–1996.

$$\sigma = 1/\sqrt{w} \tag{5}$$

is the expected mean error and

$$\sigma^* = \sqrt{\frac{1}{(n-1)w}\sum w_m (R_m - R_w)^2} \tag{6}$$

is the actual mean error of R_w.

The monthly values of R_m and σ_m can be taken either as reconstructed above for months with few observational days or directly from daily values otherwise. A special case is when all daily SN values within a month are equal to zero, leading to $\sigma_m = 0$. This makes it impossible to formally apply Equation (3) since the corresponding statistical weight approaches infinity ($w_m \to \infty$). We analyzed those months of the reference population that contain zero values and found a dependence of the observed σ_m on the number of spotless days within a month which is shown in Figure 4. This relation is nearly exponential and predicts that $\sigma_m = 0.51$ for 30 spotless days within a month. This value can be interpreted, e.g., as an observational threshold for the sunspot number or as an error of rounding to integer. Accordingly, we have set all values of $\sigma_m \to 0$ to $\sigma_m = 0.51$ and then applied the above described weighted averaging technique. This leads to $\sigma_w = 0.15$ for a spotless year. We have checked that the applied technique yields the same values, within 1σ error, as the formal yearly GSNs (Hoyt and Schatten, 1998) for the reference population (see Figure 5). The standard deviation between the weighted and formal yearly GSN values is 3.3.

Another problem arises when a year contains only one month of sunspot observations. There are eight years (1614, 1623, 1640, 1731, 1734, 1738, 1746, and 1748) with only one monthly R_m in the GSN series. In these cases we used the same statistical approach which was used for months with few daily observations (see Section 2). For a given R_m and σ_m we found, using the reference population, yearly R_y values for all years containing one monthly SN within the range of $R_m \pm \sigma_m$. Then a histogram of these collected R_y values was constructed in a similar way

TABLE I

Yearly group sunspot numbers: formal values R_g as well as the new, weighted R_w and their uncertainties σ_w.

Year	R_g	R_w	σ_w	Year	R_g	R_w	σ_w	Year	R_g	R_w	σ_w
1610	36.0	15.7	20.7	1650	0.0	0.0	0.15	1690	0.0	0.0	0.15
1611	34.2	54.3	6.4	1651	0.0	0.0	0.15	1691	0.0	0.0	0.15
1612	92.7	94.5	7.5	1652	4.0	0.1	0.5	1692	0.0	0.0	0.15
1613	109.6	86.3	8.0	1653	0.8	0.3	0.3	1693	0.0	0.0	0.15
1614	121.0	115.0	24.0	1654	0.7	0.1	0.2	1694	0.0	0.0	0.15
1615	80.3	20.8	18.6	1655	0.5	0.1	0.2	1695	0.1	0.1	0.15
1616	20.1	17.5	4.1	1656	0.6	0.1	0.2	1696	0.0	0.0	0.15
1617	2.3	0.1	0.2	1657	0.2	0.1	0.15	1697	0.0	0.0	0.15
1618	1.3	0.2	0.3	1658	0.0	0.0	0.15	1698	0.0	0.0	0.15
1619	15.0	13.2	2.0	1659	0.0	0.0	0.15	1699	0.0	0.0	0.15
1620	15.0	13.0	1.1	1660	2.0	0.7	0.5	1700	0.4	0.1	0.15
1621	15.0	15.0	0.2	1661	0.8	0.2	0.3	1701	0.5	0.1	0.15
1622	15.0	16.9	0.4	1662	0.0	0.0	0.15	1702	0.6	0.2	0.15
1623	15.0	13.0	9.0	1663	0.0	0.0	0.15	1703	2.7	2.0	1.3
1624	9.6	11.6	1.6	1664	0.0	0.0	0.15	1704	4.1	1.9	1.0
1625	42.4	34.5	3.9	1665	0.0	0.0	0.15	1705	5.5	3.8	0.9
1626	26.6	16.7	2.0	1666	0.0	0.0	0.15	1706	3.2	1.0	0.7
1627	16.5	15.0	1.7	1667	0.0	0.0	0.15	1707	5.3	2.7	1.6
1628	23.2	23.2	3.0	1668	0.0	0.0	0.15	1708	2.8	0.9	0.7
1629	18.7	17.4	2.3	1669	0.0	0.0	0.15	1709	1.6	0.5	0.4
1630	0.0	4.2	1.9	1670	0.0	0.0	0.15	1710	0.4	0.1	0.2
1631	4.4	3.2	1.4	1671	0.9	0.3	0.4	1711	0.0	0.0	0.15
1632	0.0	0.0	0.15	1672	0.4	0.1	0.15	1712	0.0	0.0	0.15
1633	14.3	0.6	0.7	1673	0.0	0.0	0.15	1713	0.3	0.1	0.2
1634	3.0	0.1	0.4	1674	0.2	0.1	0.2	1714	0.9	0.2	0.3
1635	4.3	0.1	0.3	1675	0.0	0.0	0.15	1715	3.6	1.9	0.6
1636	–	–	–	1676	1.8	0.6	0.5	1716	9.2	3.6	1.6
1637	–	–	–	1677	0.3	0.1	0.2	1717	17.5	15.2	1.6
1638	68.7	68.8	4.7	1678	0.2	0.1	0.15	1718	9.0	5.1	1.8
1639	76.8	69.2	5.1	1679	0.0	0.0	0.15	1719	33.9	26.3	4.4
1640	15.0	17.0	11.0	1680	0.8	0.2	0.3	1720	23.4	18.5	2.5
1641	–	–	–	1681	0.0	0.0	0.15	1721	21.4	7.4	2.6
1642	47.3	16.6	12.9	1682	0.0	0.0	0.15	1722	11.1	10.6	0.8
1643	17.6	15.8	3.1	1683	0.0	0.0	0.15	1723	4.5	8.8	1.1
1644	11.6	3.6	1.8	1684	1.4	0.4	0.4	1724	15.6	11.7	0.5
1645	0.0	0.0	0.15	1685	0.0	0.0	0.15	1725	12.8	8.3	2.4
1646	0.0	0.0	0.15	1686	0.6	0.1	0.15	1726	36.2	33.2	2.3
1647	0.0	0.0	0.15	1687	0.1	0.1	0.15	1727	36.5	41.5	2.2
1648	0.0	0.0	0.15	1688	0.5	0.2	0.15	1728	64.3	60.6	7.4
1649	0.0	0.0	0.15	1689	0.2	0.1	0.15	1729	24.0	8.8	6.4

TABLE I

Continued.

Year	R_g	R_w	σ_w	Year	R_g	R_w	σ_w	Year	R_g	R_w	σ_w
1730	69.7	79.9	17.5	1760	45.5	41.5	4.0	1790	65.1	66.5	13.4
1731	0.0	6.0	9.3	1761	68.5	68.2	5.8	1791	43.2	43.0	4.4
1732	18.0	12.9	12.3	1762	46.2	38.3	4.0	1792	42.0	21.1	13.3
1733	0.0	0.0	0.15	1763	34.2	28.1	2.3	1793	41.0	14.4	9.7
1734	0.0	7.3	10.2	1764	30.5	21.0	2.4	1794	30.2	27.1	7.1
1735	18.3	22.1	2.8	1765	8.4	1.9	1.7	1795	15.7	19.8	2.7
1736	48.6	53.9	1.8	1766	3.7	0.3	0.5	1796	13.7	2.3	1.8
1737	24.0	26.0	4.9	1767	33.9	10.9	4.9	1797	7.7	6.0	1.1
1738	17.0	25.0	13.0	1768	71.4	67.1	5.7	1798	4.8	0.9	0.9
1739	52.5	39.7	9.9	1769	98.5	91.5	8.4	1799	5.6	1.9	1.6
1740	9.3	13.2	0.6	1770	97.6	84.0	8.6	1800	11.0	2.3	1.5
1741	57.7	10.7	10.7	1771	79.4	67.3	9.1	1801	51.1	48.5	3.4
1742	16.1	11.7	3.3	1772	66.2	62.4	4.5	1802	35.3	32.6	3.2
1743	8.3	11.8	1.4	1773	32.4	30.3	3.3	1803	18.6	14.3	2.6
1744	–	–	–	1774	25.8	16.5	3.3	1804	21.6	22.6	5.0
1745	–	–	–	1775	5.6	1.8	0.9	1805	25.6	6.2	3.0
1746	0.0	7.3	10.2	1776	14.0	5.2	2.0	1806	13.3	15.5	2.3
1747	–	–	–	1777	38.3	25.5	4.4	1807	5.0	0.2	0.5
1748	61.0	63.0	17.5	1778	72.0	40.9	4.4	1808	3.5	0.6	0.5
1749	63.2	61.2	5.7	1779	80.8	92.2	5.5	1809	1.2	0.1	0.2
1750	58.0	48.9	2.7	1780	55.0	63.1	6.5	1810	0.0	0.0	0.1
1751	33.7	32.0	2.4	1781	71.1	63.8	5.5	1811	0.3	0.1	0.1
1752	29.0	27.4	3.6	1782	32.9	25.2	4.8	1812	4.0	0.7	0.6
1753	23.9	18.9	4.5	1783	21.1	20.7	4.0	1813	9.1	3.3	1.5
1754	8.8	12.3	1.4	1784	4.8	8.6	3.1	1814	10.4	12.2	1.4
1755	4.7	1.6	0.8	1785	16.0	14.9	3.9	1815	16.9	15.0	0.6
1756	7.3	6.8	1.2	1786	63.3	47.5	4.6	1816	30.8	28.0	1.7
1757	24.8	15.9	2.1	1787	89.2	71.5	4.6	1817	28.1	24.7	1.9
1758	40.7	34.5	3.0	1788	82.5	87.6	4.1	1818	21.7	19.9	1.8
1759	49.5	46.1	2.7	1789	79.7	68.5	7.3	1819	19.2	19.0	2.1

as in the analysis of monthly values. Figure 6 shows some samples of histograms for yearly values (cf., Figure 1) which again depict the approximate Poisson shape. From these histograms one can find the yearly average and its uncertainty for those years when the standard averaging method (Equations (3)–(6)) cannot be applied.

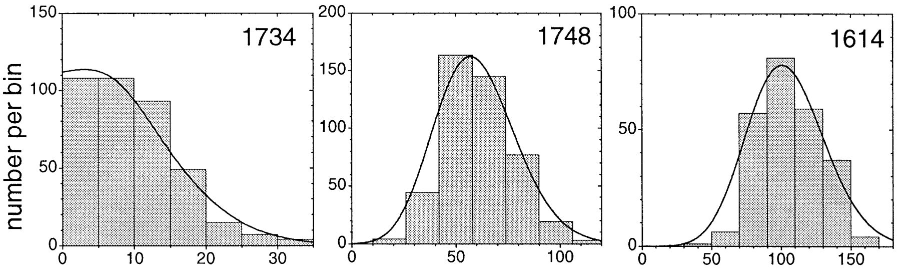

Figure 6. Samples of histogram distributions of yearly R_y together with the rescaled best fitting Poisson distributions for three such years which only have one month of sunspot observations: 1734 ($R_m = 7.3 \pm 7.7$), 1748 ($R_m = 63 \pm 19$), and 1614 ($R_m = 115 \pm 26$).

The weighted yearly GSN values are shown in Figure 7(a) together with the formal yearly GSN series (Hoyt and Schatten, 1998). The weighted yearly GSN values R_w and their standard errors σ_m are also given in Table I for the period 1610–1819 when there are significant gaps in the data. The difference between the two annual curves (Figure 7(b)) becomes significant for those years which are poorly covered by sunspot observations. A few yearly values are modified quite significantly, by more than 30. The new, weighted yearly values are mostly below the corresponding formal values. This is particularly true for the time interval after 1750. We note that the weighted yearly sunspot values are also reduced in 1792–1794 and depict a minimum in 1793, thus confirming the existence of the lost solar cycle in 1790s, as discussed in great detail by Usoskin, Mursula, and Kovaltsov (2003).

4. Conclusions

We have presented a new method, based upon statistical properties of sunspot activity during the last 150 years, which allows to estimate the monthly sunspot number value and its uncertainty from sparse (or even single) daily sunspot observations. The fact that the method can also evaluate the errors in the monthly SN values allows to apply the method of weighted averaging to calculate the yearly sunspot number value from monthly data. We have presented the reconstructed monthly and yearly group sunspot numbers for the period 1610–1810 (reconstructed monthly group sunspot numbers and their uncertainties can be requested from the authors). The method provides a basis for more rigorous studies of the statistical features of sunspot activity during early times when good data coverage was not yet routine.

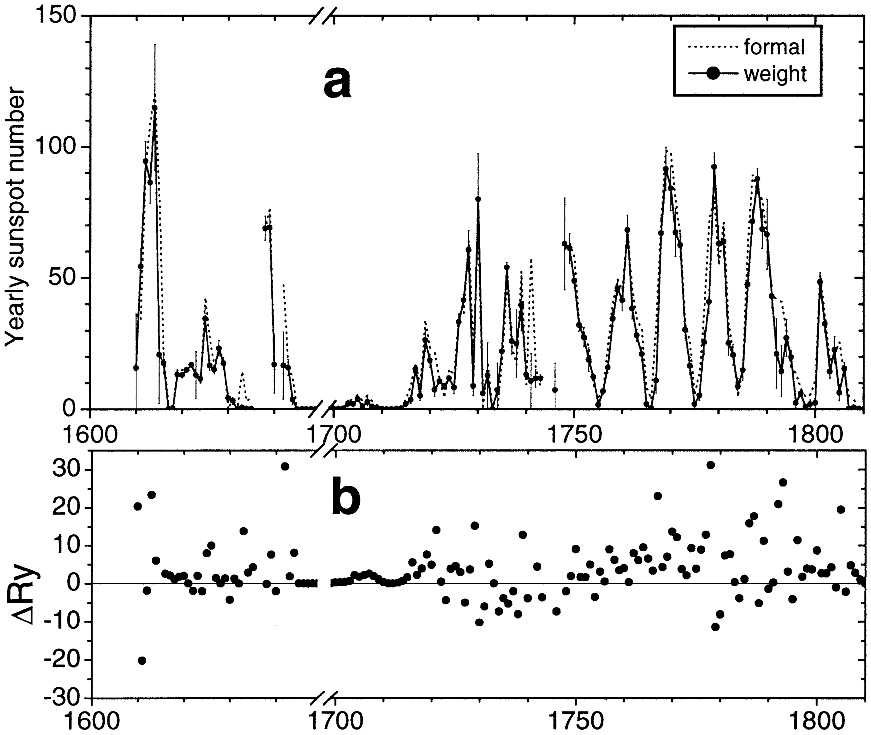

Figure 7. (a) Yearly group sunspot numbers calculated as the formal arithmetic mean (*dotted curve*) and the weighted average (*solid curve with dots*). The latter is given with the estimated uncertainties. (b) The difference ΔR_y between the formal and new, weighted yearly sunspot numbers.

Acknowledgements

The financial support by the Academy of Finland is gratefully acknowledged. GAK was partly supported by the program 'Non-stationary Processes in Astronomy' of Russian Academy of Sciences. We thank the referee for valuable comments.

References

Agekyan, T. A.: 1972, *The Basics of the Errors Theory for Astronomers and Physicists*, Nauka, Moscow (in Russian).

Hoyt, D. V. and K. Schatten: 1998, *Solar Phys.* **179**, 189.

Usoskin, I. G., Mursula, K., and Kovaltsov, G. A.: 2003, *Astron. Astrophys.* **403**, 743.