ON THE RELIABILITY OF MONTHLY/YEARLY MEANS CALCULATED FROM SPARSE DAILY SUNSPOT NUMBERS

Ilya G. Usoskin[1], **Kalevi Mursula**[2], **and Gennady A. Kovaltsov**[3]

[1]Sodankylä Geophysical Observatory (Oulu unit), POB 3000, FIN-90014 University of Oulu, Finland
[2]Department of Physical Sciences, FIN-90014 University of Oulu, Finland
[3]Ioffe Physical-Technical Institute, 194021 St.Petersburg, Russia

ABSTRACT

Some periods before 1850 are poorly covered by sunspot observations. In addition to apparent 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/yearly mean values obtained from 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 employs two assumptions: (i) statistical properties of sunspot activity are similar throughout the entire period and (ii) individual sparse daily observations are distributed randomly in time. First, for each sample population we found months in the reference population containing the same data set and then constructed the statistical distribution of the corresponding monthly means. From this distribution we calculated the weighted mean and its standard error which gives the uncertainty of a monthly mean sunspot number reconstructed from sparse daily observation. The simple arithmetic mean 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 should be estimated more carefully. Using the estimated monthly values, we can also calculate the weighted annual sunspot numbers.

Key words: Sunspot activity; Solar cycle.

1. INTRODUCTION

While the sunspot numbers (SN) 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 gets uncertain when only few (in the extreme only one) R_d values are available within a month. E.g., (Hoyt & 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 et al. 2003a) to form the monthly mean from isolated daily observations. The advantage of this method is that it allows not only to calculate the monthly SN value but also to estimate its un-

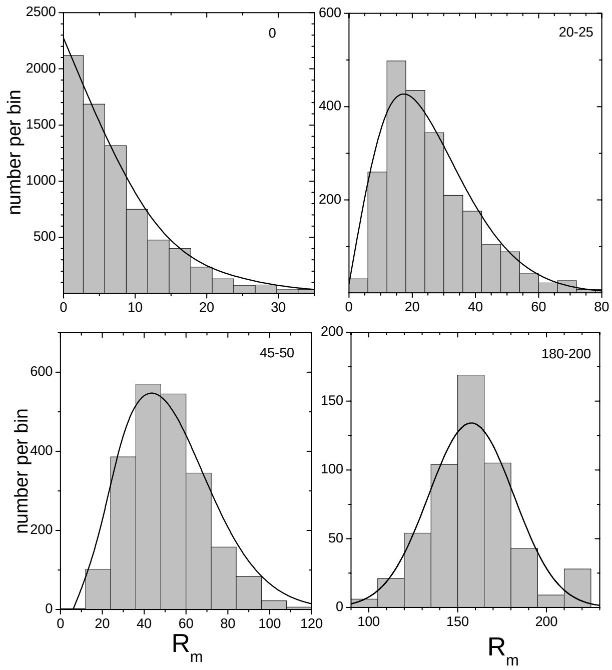

Figure 1. Samples of histogram distributions of monthly R_m together with the rescaled best-fit Poisson distribution functions. The four panels depict the cases of at least one R_d equal to zero or from the interval [20–25], [45–50], [180–200], respectively.

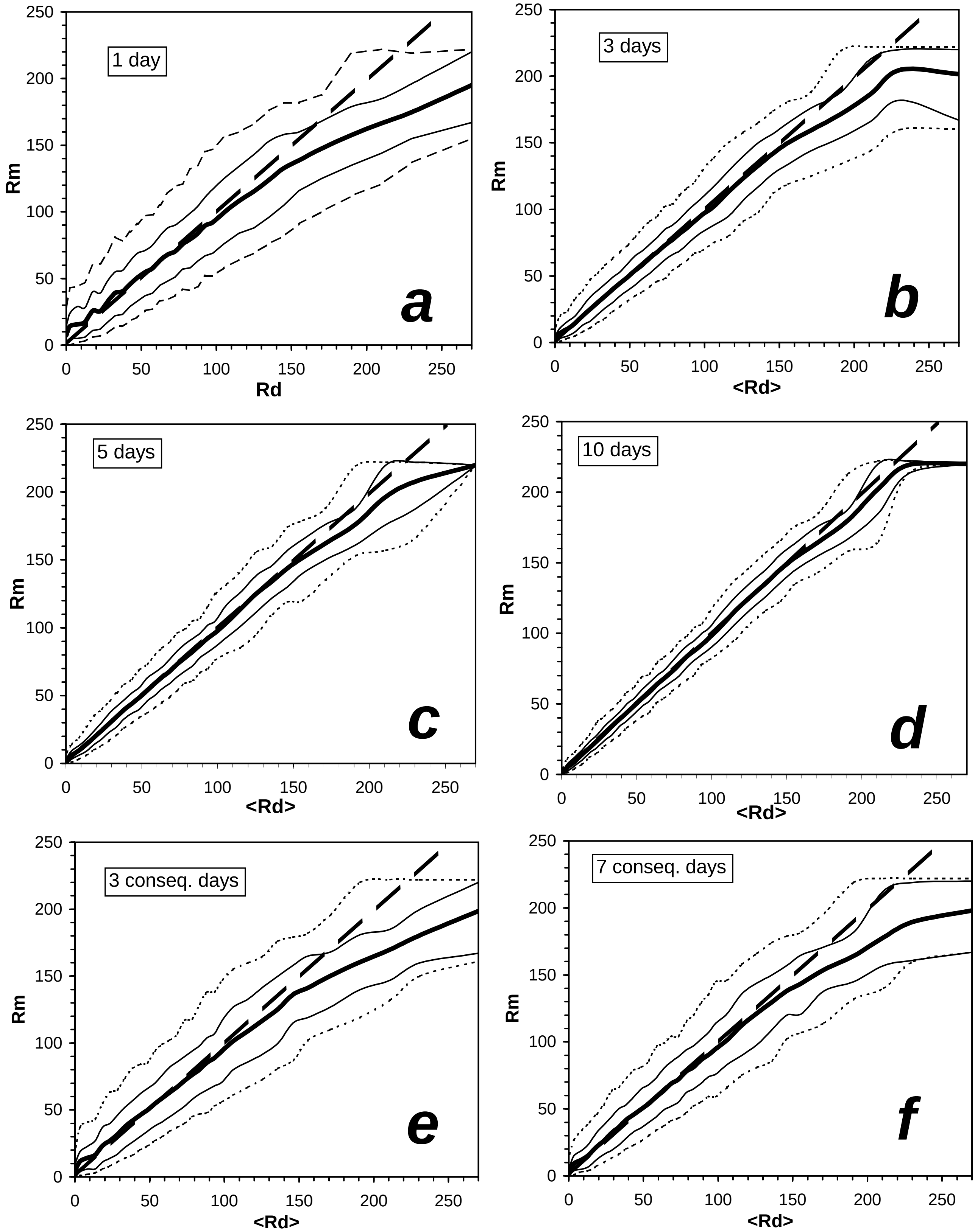

Figure 2. *The quality of monthly sunspot numbers calculated from the arithmetic mean of sparse daily values. The horizontal and the vertical axes correspond to the arithmetic mean $\langle R_d \rangle$ and the actual R_m, respectively. Panels a–d correspond to 1, 3, 5 and 10 daily observations taken randomly, and panels e–f to 3 and 7 days taken consequently, 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. Thick dashed line denotes the diagonal $R_m = \langle R_d \rangle$.*

certainty. 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 (GSN) as presented by (Hoyt & Schatten 1998).

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 53000 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 were 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).

Fig. 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., R_m values. Since the GSN value is the number of sunspot groups G multiplied by a factor of 12.08 (Hoyt & 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:

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

where G is an integer 0,1,2, ... and μ is the mathematical expectation of the mean. Fig. 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 the monthly mean R_m and its uncertainty σ_m corresponding to one daily R_d value in a month (Fig. 2a). 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

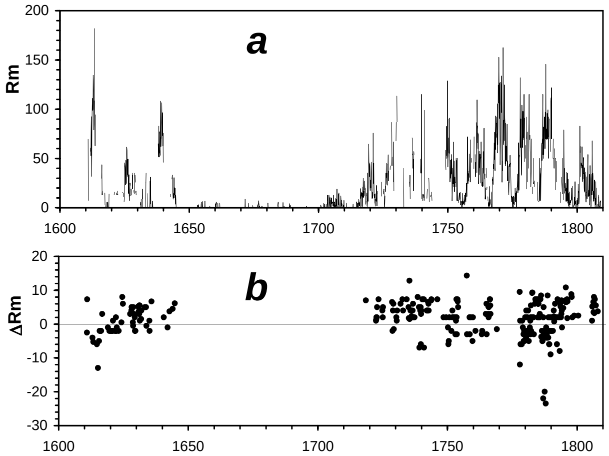

Figure 3. a) Monthly group sunspot numbers, reconstructed as described in the text. b) The difference ΔR_m between the formal and newly calculated monthly sunspot numbers.

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 Fig. 2b–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 & Schatten 1998). (The horizontal plateau for $\langle R_d \rangle > 220$ is because of lack of statistics for high SN values.)

In the discussion above we assumed that the observational days are taken randomly within the month. However, it is quite common that daily observations are consequent and form a single period of a few consequent 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 consequent R_d values. Note that in this case the quality of the R_m reconstruction (see Fig. 2e,f) is very close to the single daily observation (Fig. 2a) because the consequent observations are strongly related to each other.

Thus, using the method illustrated by Figs. 1-2, 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 less than five separate or less than 10 consequent daily observations, we have reconstructed the monthly GSN values shown in Fig. 3a. For other months (>4 evenly distributed or >10 consequent 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

$$\sigma_m = \sigma_d / \sqrt{n_d - 1}, \qquad (2)$$

where σ_d (cf., Hoyt & Schatten 1998) and n_d are the standard deviation and the number of daily R_d values within the month. The differences between the formal R_m values (Hoyt & Schatten 1998) and the newly reconstructed monthly values are shown in Fig 3b. 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 1780's show a large difference of about -25. Note also that for most months the difference is positive, indicating that the arithmetic average exaggerates the montly value, as suggested by Fig. 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 month are not fully covered by daily observations. 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 annual average is defined as (see, e.g., Appendix in Usoskin et al. 2003a)

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

where individual (monthly) weights are $w_m = 1/\sigma_m^2$, and $w = \sum_m w_m$. The R_m and σ_m values as reconstructed above were used for months with few observational days. Otherwise daily values and Eq. 2 were used to calculate the mean and error. These weighted yearly GSN values are shown in Fig. 4a together with the formal yearly GSN series (Hoyt & Schatten 1998). The difference between the two annual curves (Fig. 4b) is also mostly limited within ±10. However, a number of yearly values are modified quite significantly, by more than 30. In particular, the new weighted yearly sunspot values are reduced during 1792–1794, depicting a minimum in 1793 and confirming the existence of the lost solar cycle in 1790's, as discussed in great detail in (Usoskin et al. 2003a; Usoskin et al. 2003b).

4. CONCLUSIONS

We have presented the new method, based upon statistical properties of sunspot activity during the last 150 years, that 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. 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.

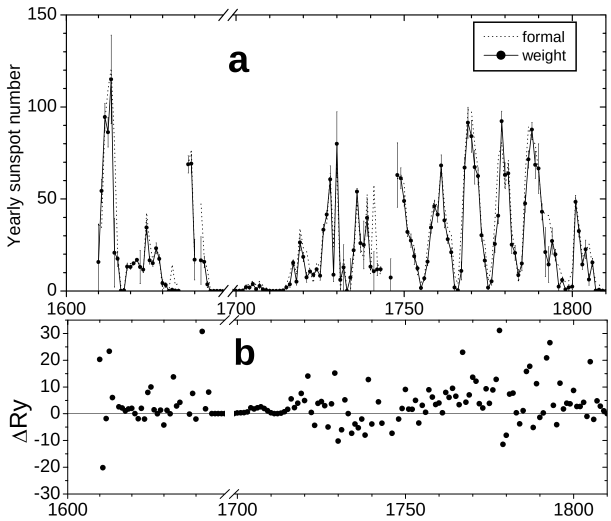

Figure 4. a) Yearly group sunspot numbers calculated as the formal arithmetic mean (dotted curve) and the weighted average (solid curve with dots). The latter is given with the estimated uncertainties. b) The difference δR_y between the formal and newly calculated yearly sunspot numbers.

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.

REFERENCES

Hoyt D. V., & K. Schatten (1998): Solar Phys., 179, 189

Usoskin, I.G., Mursula, K., Kovaltsov, G.A. (2003a): A&A, 403, 743-748.

Usoskin, I.G., Mursula, K., Kovaltsov, G.A. (2003b), The lost sunspot cycle: Reanalysis of sunspot statistics, this volume.