The Astrophysical Journal, 886:7 (14pp), 2019 November 20 https://doi.org/10.3847/1538-4357/ab4990

Uncertainty Quantification in Sunspot Counts

Sophie Mathieu[1], Rainer von Sachs[1], Christian Ritter[1], Véronique Delouille[2], and Laure Lefèvre[2]
[1] Université Catholique de Louvain, ISBA, LIDAM, Louvain-la-Neuve, Belgium
[2] Royal Observatory of Belgium, Solar physics and Space Weather Department, Brussels, Belgium
Received 2019 August 13; revised 2019 September 26; accepted 2019 September 26; published 2019 November 13

Abstract

Observing and counting sunspots constitutes one of the longest-running scientific experiments, with first observations dating back to Galileo (around 1610). Today the sunspot number (SN) time series acts as a benchmark of solar activity in a large range of physical models. An appropriate statistical modeling, adapted to the time series' complex nature, is, however, still lacking. In this work, we provide the first comprehensive uncertainty quantification analysis of sunspot counts. We study three components: the number of sunspots (N_s), the number of sunspot groups (N_g), and the composite N_c, defined as $N_c := N_s + 10N_g$. Those are reported by a network of observatories around the world and are corrupted by errors of various types. We use a multiplicative framework to provide, for these three components, an estimation of their error distribution in various regimes (short-term, long-term, minima of solar activity). We also propose a robust estimator for the underlying solar signal and fit density distributions that take into account intrinsic characteristics such as overdispersion, excess of zeros, and multiple modes. The estimation of the solar signal underlying the composite N_c may be seen as a robust version of the International Sunspot Number (ISN), widely used as a proxy of solar activity. Therefore, our results on N_c may help characterize the uncertainty on ISN as well. Our results pave the way for a future monitoring of the observatories in quasi-real time, with the aim of alerting the observers when they start deviating from the network and preventing large drifts from occurring.

Unified Astronomy Thesaurus concepts: Sunspot groups (1651); Sunspot number (1652); Sunspots (1653); Model selection (1912); Mixture model (1932)

1. Introduction

1.1. The International Sunspot Number (ISN)

On white-light images, sunspots are visible as dark areas. They correspond to regions of locally enhanced magnetic field and act as an indicator of changing solar activity over time. They have been observed and counted since the invention of the telescope at the beginning of the seventeenth century. As such, the counting of sunspots constitutes one of the "longest-running scientific experiment[s]" (Owens 2013). In 1848, J. R. Wolf from Zürich Observatory created an index, denoted N_c, of solar activity by summing up the total number of sunspots N_s with 10 times the total number of sunspot groups N_g on a daily basis:

$$N_c = 10N_g + N_s. \tag{1}$$

Figure 1 displays smoothed averages of the median value of these three quantities across a set of 21 observatories (also called "stations") chosen for the present study (see Section 2). Modeling the statistics of N_c is far from trivial, as this quantity jumps from 0 to 11 when a sunspot appears on the Sun ($N_s = 1$, $N_g = 1$, and thus $N_c = 11$). By construction, each active region appears thus twice in N_c. The multiplication factor in Equation (1) was introduced by J. R. Wolf to put the number of groups on the same scale as the number of spots. Indeed, during this historical period, a group contained on average 10 spots (Izenman 1985). Note that in recent solar cycles the average number of spots per group is rather around six.

The index N_c, or rather the formula behind it, is at the basis of the ISN. The ISN is distributed through the World Data Center Sunspots Index and Long-term Solar Observations (WDC-SILSO).[3] The N_c values from each observing station in the SILSO network are collected and rescaled, i.e., multiplied by a factor k, to compensate for their differing observational qualities. The N_c values are then combined on a monthly basis to produce the ISN (Clette et al. 2007), which constitutes the international reference for modeling solar activity over the long term. Despite the fact that it is arguably the most intensely used times series in all of astrophysics (Hathaway 2010), its historical part suffers from a number of errors and inconsistencies that have been partly addressed by the recalibration of the ISN in 2015 (Clette & Lefèvre 2016). Even the most recent part (1981–now) lacks proper error modeling and uncertainty quantification; see Section 1.2 below.

1.2. ISN: Origin and Computation

In order to place our work in context, we first describe how the ISN is currently obtained at the WDC-SILSO center. N_s and N_g are entered through an interface (www.sidc.be/WOLF) and stored in a database. The main processing is described in Clette et al. (2007), and we summarize a more recent version of it in Figure 2. The processing uses a pilot station, here the Locarno station, as a reference. It compares the values obtained by a station i to the pilot station via a scaling factor k_i, often referred to as the "k-coefficient":

$$k_i(t) = \frac{\text{pilot}(t)}{Y_i(t)}, \tag{2}$$

where $Y_i(t)$ is the composite index of station i, observed at time t (expressed in days), and pilot(t) is the value of the pilot station. The monthly scaling factors are computed from a sigma-clipping mean of Equation (2), i.e., values differing by more than two standard deviations from the mean are eliminated from the computation process. This processing still

[3] http://www.sidc.be/silso/

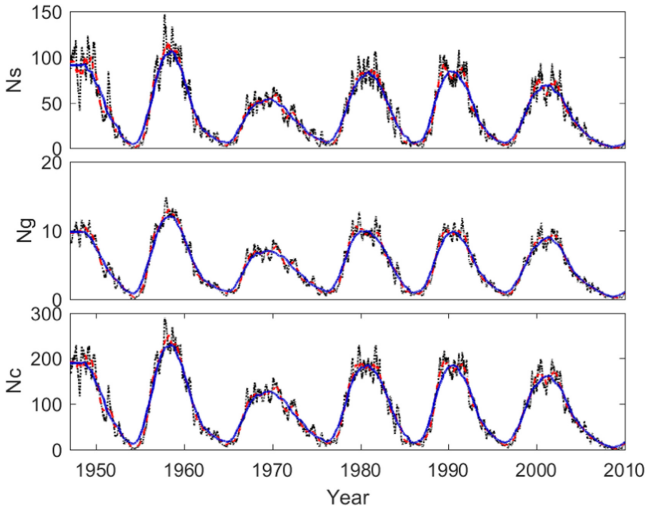


Figure 1. Time evolution during 1947–2013 of the median values across 21 observing stations (see Table 1) for the sunspot counts: (top) N_s, (center) N_g, and (bottom) N_c. The data are averaged over 81 days (black dotted line), 1 yr (red dashed line), and 2.5 yr (blue solid line).

suffers from its historical heritage, summarized in Table 1 of Dudok de Wit et al. (2016). For example, this table shows that between 1926 and 1981 (when the sunspot collection center was in Zurich) there were several standard observers, and no pilot station. As an index derived from count data, N_c (or N_s and N_g) does not necessarily follow a Gaussian distribution (Vigouroux & Delache 1994; Usoskin et al. 2003; Dudok de Wit et al. 2016). A processing based on sigma-clipping is thus not fully adapted, but still undoubtedly better than what was done during the Zurich era. Finally, some steps in the processing date back from the mid-nineteenth century (when J. R. Wolf introduced the sunspot index) and have not been upgraded when the collection and preservation center was moved from Zurich to Brussels in 1981. There were two reasons for this: (1) the new curators of the ISN wanted to keep the uniformity of the series, and (2) the numerical tools available at that time were limited.

The WDC-SILSO team is currently working on improving the ISN computation and coordinates an important community effort to correct past errors. Such effort includes, among others, work by a team from the International Space Science Institute (ISSI) on recalibration of the SN,[4] organization of sunspot workshops,[5] and editorial work for a Solar Physics topical issue on SN recalibration (Clette et al. 2016).

1.3. Previous Works on SN Uncertainty Quantification

Long-term analyses started with models of the shape of the sunspot number time series (Stewart & Panofsky 1938; Stewart & Eggleston 1940). They pursued the works by M. Waldmeier himself (Waldmeier 1939), who tried to understand the solar cycle and predict upcoming cycles. Later on, Morfill et al. (1991) investigated the short-term dynamical properties of the SN series using a Poisson noise distribution superimposed on a mean cycle variation. Vigouroux & Delache (1994) also use a Poisson distribution to approximate the dispersion of daily values of the SN at different regimes of solar activity. Usoskin et al. (2003) develop a reconstruction method for sparse daily values of the SN and model the monthly number of groups corresponding to a certain level of daily values by a Poisson distribution. Schaefer (1997) emphasizes the need for error bars on the AAVSO sunspot series[6] (Foster 1999), and more recent results in Dudok de Wit et al. (2016) present a first uncertainty analysis of the short-term error, through time domain errors and dispersion errors among observing stations, still assuming a Poisson distribution. In Dudok de Wit et al. (2016), however, the authors uncover the presence of overdispersion in the SN and approximate the SN by a mix of a Poisson and a Gaussian distribution in an additive framework. Although non-Poissonian, this additive model fails to capture some of the characteristics of sunspot data. Chang & Oh (2012), on the other hand, use a multiplicative model to simulate sunspot counts in view of assessing the dependency of correction factors on the solar cycle.

1.4. Motivation and Contribution

Our goal in this work is to go beyond the above-mentioned historical heritage by developing a comprehensive uncertainty quantification model for the count data N_s, N_g, and N_c. These quantities are subject to different types of errors and do not behave exactly like Poisson random variables: (1) they experience more dispersion than the Poisson distribution, (2) they are not independent from one day to another (since sunspots can last from several minutes to several months on the Sun), and (3) they exhibit a large number of zeros owing to periods of minimal solar activity.

Our contribution is twofold. First, we develop robust estimators for the physical solar signal, denoted "true" signal in this paper, underlying N_s, N_g, and N_c. We propose a model for their densities that takes into account characteristics such as overdispersion and large number of zero counts. Our processing and estimators are robust to missing values and do not require filling in missing observations, contrarily to previous studies.

Second, we propose an uncertainty model that is motivated by first studies in Dudok de Wit et al. (2016) and that works within a *multiplicative* framework. Our model distinguishes three error types. The *short-term error* accounts for counting errors and variable seeing conditions from one station to another (e.g., weather, atmospheric turbulence), whereas the *long-term error* provides an overall bias in the number of spots (e.g., gradual ageing of the instrument or observer). Finally, a third error type aims at modeling inaccuracies occurring at solar minima and helps differentiating true from false zero counts. As an illustration, the short-term variations coming from the solar variability and the observational errors are clearly visible in Figure 1, superimposed on the approximate seasonality of the 11 yr solar cycle.

In future prospects, our work paves the way for a more robust definition of the ISN. Indeed, the analysis of the different error types allows studying the stability of observatories involved in the computation of ISN. Our study lays the ground for a future monitoring of all active stations within the SILSO network in quasi-real time, with the aim of (1) defining a stable reference of the network, (2) alerting the observers when they start deviating from the network, and (3) preventing

[4] http://www.issibern.ch/teams/sunspotnoser/
[5] https://ssnworkshop.fandom.com/wiki/Home
[6] https://www.aavso.org/category/tags/american-relative-sunspot-numbers

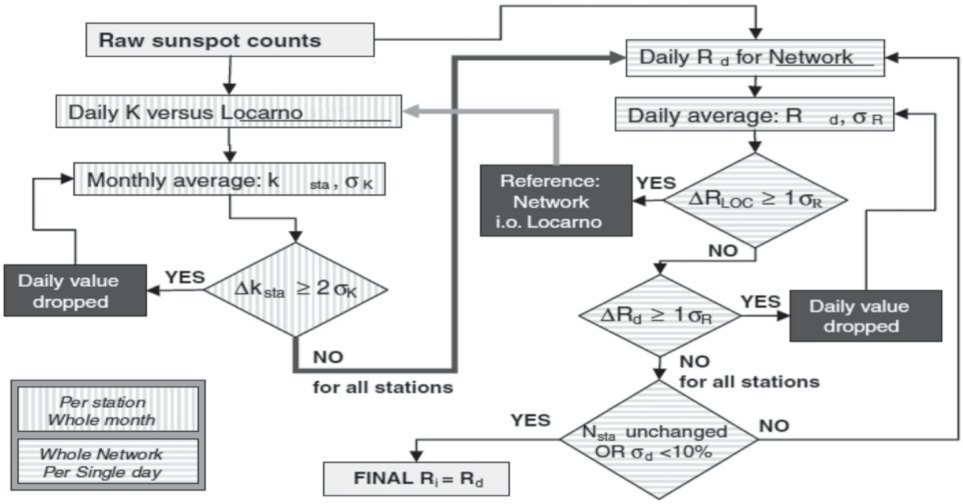


Figure 2. Flowchart of the WDC-SILSO data import procedure, illustrating the succession of hierarchical tests applied to raw observing reports (adapted from Clette et al. 2007).

Table 1
Main Characteristics of the Subset of Stations

ID	Name	Location	Prof. versus Amateur	Team versus Individual	Observing Period	Level	% Obs.	% Obs. Period
A3	Athens Obs.	Athens (Greece)	Prof.	team	1949–1995	1.039	30.16	44.01
BN-S	WFS Obs.*	Berlin (Germany)	Am.	team	1965–2013	1.179	23.50	32.74
CA	Catania Obs.	Catania (Italy)	Prof.	team	1950–2019	1.039	61.87	64.80
CRA	Cragg†	Australia	Am	indiv.	1947–2009	0.904	72.43	77.44
FU	Fujimori	Nagano (Japan)	Am	indiv.	1968–2019	1.055	45.73	67.32
HD-S	Hedewig*	Germany	Am	indiv.	1967–2013	0.931	25.42	36.96
HU	Public Observatory	Hurbanovo (Slovakia)	Am	team	1969–2019	1.004	35.452	52.80
KH	KOERI	Kandilli (Turkey)	Prof.	team	1967–2019	0.968	48.81	51.38
KOm	Koyama	Tokyo (Japan)	Am	indiv.	1947–1996	1.052	40.18	54.84
KS2	Kislovodsk Mountain Obs.	Kislovodsk (Russia)	Prof.	∼indiv.	1954–2019	1.057	85.96	95.98
KZm	University of Graz	Kanzelhohe (Austria)	Prof.	team	1944–2019	1.110	74.23	74.24
LFm	Luft	New York (USA)	Am	indiv.	1944–1988	0.985	34.06	54.68
LO	Specola Solare	Locarno (Switzerland)	Prof.	∼indiv.	1958–2019	1.260	68.27	81.68
MA	Manila Obs.	Manila (Philippines)	Prof.	team	1971–1988	1.023	20.85	78.69
MO	Mochizuki (Urawa)	Saitama (Japan)	Am	indiv.	1978–2019	1.073	35.51	66.09
PO	Observatory	Postdam (Germany)	Prof.	team	1955–1999	0.991	22.12	29.73
QU	PAGASA weather Bureau	Quezon (Philippines)	Prof.	∼indiv.	1957–2019	0.829	45.46	53.83
SC-S	Schulze*	Germany	Am	indiv.	1960-2007	0.943	23.32	33.16
SK	Skalnate Pleso Obs.	Vysoke Tatry (Slovakia)	Prof.	team	1950–2012	0.992	37.95	40.75
SM	San Miguel Obs.	Buenos Aires (Argentina)	Prof.	team	1967–2013	1.220	39.09	56.34
UC	USET	Uccle (Belgium)	Prof.	team	1949–2019	0.991	57.00	59.64

Note. Main characteristics on the set of 21 stations used in this study: acronym (ID), last name of observer or name of station, location, type of observatory (professional versus amateur), type of observer (individual or team), observing period, averaged scaling factor with respect to the network over the studied period (level), and percentage of observations on the full period studied and on their observing period (% Obs. period). Note that ∗ and † symbols represent stations or observers from the SONNE and AAVSO networks, respectively. They are two distinct networks of observing stations that are not members of the WDC-SILSO network and hence are not used to produce the ISN.

large drifts from occurring. A new ISN could then be defined from a stable reference rather than a single pilot station and could benefit from the robust estimators and procedures (including rescaling of observing stations; see Section 4) developed in this work.

Our paper is structured as follows. Section 2 introduces the data set considered. The uncertainty model is presented in Section 3, while Section 4 details the preprocessing of the data. Section 5 provides the estimators (or proxy) for the "true" solar signal underlying N_s, N_g, and N_c, as well as their densities. Finally, Section 6 displays our results on quantification of the different error types, as well as a first stability analysis that takes into account both short- and long-term variability.

2. Data

Similarly to what is done in Dudok de Wit et al. (2016), we study a subset of 21 stations, whose main characteristics are listed in Table 1. The period under study goes from 1947 January 1 until 2013 December 31. It ranges from the

maximum of solar cycle (SC) 18 until the ascending phase of SC 24[7] and covers thus almost six solar cycles.

Table 1 summarizes properties of the stations such as their location, name, type (amateur vs. professional, individual vs. team), observing period, percentage of missing values, and mean scaling factor (level) with respect to the network over the period studied. The procedure that was used to compute the mean scaling factors will be described in Section 4. These mean scaling factors may be viewed as an indication of the general level of counts recorded by the station as compared to the median of the network. Thus level = 1 corresponds to a station that observes the same number of spots as the median of the network. For example, Locarno (LO) with a level of 1.26 observes in general around 20% more spots than the others.

The location of the observatories gives an indication of the weather conditions of the stations and might explain part of the missing values. Moreover, the type of observatory usually impacts the quality of observations and the length of observing periods: an individual might experience less short-term variability than a team (alternating the observer from one week to another), and/or amateurs may have shorter observing periods than professionals.

Our data set contains the daily number of spots N_s, groups N_g, and the composite N_c observed in each of the 21 stations. As Table 1 indicates, the data present an important amount of missing values due to weather conditions preventing stations from observing, periods of instrument maintenance or definitive closures, or births of new observatories.

3. Model

In this section, we present step by step our uncertainty model. It characterizes the observations of the stations (either for the number of spots N_s, groups N_g, or composite N_c) in a multiplicative framework and involves different types of observing errors, as well as a quantity generically denoted by $s(t)$, for N_s, N_g, and N_c. $s(t)$ is a latent variable representing the "true" solar signal, i.e., the actual number of spots N_s, groups N_g, or composite N_c lying on the Sun. It cannot be directly observed, as the counts of the stations are corrupted by different error sources. Our goal is to estimate the distribution of the "true" solar signal and of the errors degrading it. In particular, we are interested in the mean and the variance of these distributions, but also in higher-order moments since the estimated densities are far from Gaussian.

The mean of $s(t)$, denoted by $\mu_s(t)$, will be estimated in Section 5 based on the entire network (to be robust against errors of an individual station) and will be used as a proxy for $s(t)$ in the remaining part of the article. Since our model is multiplicative, a good estimation of $\mu_s(t)$ is the key to get access to the multiplicative errors; see Equation (6). Moreover, a precise estimation of the mean level of an individual station is required for future monitoring, and this depends on the accuracy of the estimation of $s(t)$.

We use a model that is conditional on the latent $s(t)$ and decomposes the observations along two regimes: when $s(t) = 0$ (solar minima) and when $s(t) > 0$ (outside periods of minima); see Section 3.1. This allows introducing, outside of minima, a model with short-term observing errors and long-term drifts; see Section 3.2. A specific error model is then developed for periods of solar minima in Section 3.3, and the complete model is shown in Section 3.4. Finally, Section 3.5 introduces the Hurdle model in order to fit distributions exhibiting an excess of zero values, as is the case here owing to the presence of solar minima and observing errors.

3.1. Conditional Model

The observed counts are studied in two distinct situations: when there are sunspots ($s > 0$) and when there are none ($s = 0$). This separation is motivated by the idea that the absence or the presence of sunspots is led by a series of phenomena involving complex dynamo processes in the solar interior, and which can be modeled by a latent variable with two states. This analysis will lead to a better understanding of the observations and allows differentiating the "true" zeros of the counting process from the "false" zeros that occur when a station reports zero sunspot count in the presence of one or more spots on the Sun.

Let $Y_i(t)$ represent either the number of spots, groups, or composite actually observed (raw, unprocessed data). The index $1 \leqslant i \leqslant N$ denotes the station, and $1 \leqslant t \leqslant T$ is the time. The probability density function (pdf) of $Y := Y_i(t)$ may be decomposed as

$$\begin{aligned} \boldsymbol{P}(Y=0) &= \overbrace{P(Y=0|s(t)>0)P(s(t)>0)}^{1} \\ &\quad + \underbrace{P(Y=0|s(t)=0)P(s(t)=0)}_{2} \\ \boldsymbol{P}(Y \geqslant y) &= \overbrace{P(Y \geqslant y|s(t)>0)P(s(t)>0)}^{3} \\ &\quad + \underbrace{P(Y \geqslant y|s(t)=0)P(s(t)=0)}_{4} \text{ for } y>0. \end{aligned} \tag{3}$$

Terms "1" and "3" in Equation (3) represent the short-term error in the presence of one or more sunspots. Term "1" reflects a situation where no sunspots are reported while there are actually some spots on the Sun ("false" zeros or observational errors due, e.g., to a bad seeing) and leads to an excess of zeros in short-term error distribution.

Term "2" captures the "true" zeros (no sunspot and no sunspot reported), while term "4" reflects a situation where the station reports some sunspots when there are no sunspot on the Sun. Term "4" is neglected outside of solar minima periods. Together, these two terms form the distribution of the error at minima, which has an excess of "true" zeros and a tail modeling the errors of the stations and the short-duration sunspots.

3.2. Short-term and Long-term Errors

Results in Dudok de Wit et al. (2016) evidence a short-term, rapidly evolving, dispersion error across the stations that accounts for counting errors and variable seeing conditions. We define a similar term allowing a possible station dependence, and we denote it $\epsilon_1(i, t)$. Assuming $\mathbb{E}(\epsilon_1(i, t)) = 0$, where $\mathbb{E}$ is the expectation sign, our interest lies in modeling its variance and its tail to study the short-term variability of the stations.

Next, we introduce $\epsilon_2(i, t)$ to handle station-specific long-term errors such as systematic biases in the sunspot counting process. We want to estimate its mean, $\mu_2(i, t)$, and detect whether this mean experiences sudden *jumps* or *drifts* on longer timescales.

[7] https://en.wikipedia.org/wiki/List_of_solar_cycles

Both $\epsilon_1(i,\ t)$ and $\epsilon_2(i,\ t)$ are multiplicative errors, as an observer typically makes larger errors when $s(t)$ is higher (Chang & Oh 2012). Assembling these two types of errors, we propose the following noise model outside of solar minima:

$$Y_i(t) = (\epsilon_1(i,\ t) + \epsilon_2(i,\ t))s(t) \text{ when } s(t) > 0. \quad (4)$$

3.3. Errors at Solar Minima

Let ϵ_3 denote the error occurring during minima of solar activity, when there exist extended periods with no or few sunspots. We assume the error ϵ_3 to be significant when there are no sunspots ($s(t) = 0$) and otherwise negligible in order to not interfere with the errors ϵ_1 and ϵ_2. ϵ_3 captures effects like short-duration sunspots and nonsimultaneity of observations between the stations. At solar minima, the model becomes

$$Y_i(t) = \epsilon_3(i,\ t) \text{ when } s(t) = 0. \quad (5)$$

3.4. General Model

Combining the three error types, we may write our uncertainty model in a compact and generic way as follows:

$$Y_i(t) = \begin{cases} (\epsilon_1(i,\ t) + \epsilon_2(i,\ t))s(t) & \text{if } s(t) > 0 \\ \epsilon_3(i,\ t) & \text{if } s(t) = 0. \end{cases} \quad (6)$$

We assume the random variables (r.v.) ϵ_1, ϵ_2, and ϵ_3 to be continuous, and the r.v. s, ϵ_1, ϵ_2, and ϵ_3 to be jointly independent. Although the "true" number of counts $s(t)$ is discontinuous, its product with a continuous r.v. $(\epsilon_1 + \epsilon_2)$ remains continuous. This is consistent with the fact that, after the preprocessing, the observed data $Y_i(t)$ may be modeled by a continuous r.v.

3.5. Excess of Zeros

All terms in Equation (6) exhibit an excess of zeros, that is, an unusual local peak in the density at zero, due to solar minima periods. As the solar minimum is an important part of a solar cycle, the zeros must be properly treated. Specific models such as the zero-altered (ZA) or the zero-inflated (ZI) two-part distributions may be used for this purpose (Zuur et al. 2009; Colin Cameron & Trivedi 2013). The main difference between both models is that the ZI distribution allows the zeros to be generated by two different mechanisms, contrarily to the ZA model, which treats all zeros in the same way.

As "true" and "false" zeros do not appear together in a single term of Equation (6), we find it appropriate to work with the ZA two-part model (also called the "Hurdle" model) and denote its density by $f(x)$. In this model, the zero values are modeled by a Bernoulli distribution $f_0(x) = b^{1-x}(1-b)^x$ of parameter b. Nonzero values follow a distribution described generically by $f_1(x)$, either another discrete distribution (in case of modeling the counts $\mu_s(t)$ in Section 5) or a continuous distribution for ϵ_1 and ϵ_3 in Section 6:

$$f(x) = \begin{cases} f_0(0) = b & \text{if } x = 0 \\ (1 - f_0(0))\dfrac{f_1(x)}{1 - f_1(0)} = (1-b)\dfrac{f_1(x)}{1 - f_1(0)} & \text{if } x > 0. \end{cases} \quad (7)$$

The ZA distribution will be used to model the estimated densities of $\mu_s(t)$ in Section 5 and those of $\epsilon_1(i,\ t)$ and $\epsilon_3(i,\ t)$ in Section 6.

4. Preprocessing

Due to the different characteristics of the observing means (telescope aperture, location, personal experience, etc.), each station has its own scaling. These differences mainly impact the count of small spots, which cannot be observed with low-resolution telescopes, and the splitting of complex groups, where the personal experience of the observer matters. A preprocessing is thus needed to rescale all stations to the same level when comparing stations on the short term and at solar minima. It is also required to compute a robust estimator of the solar signal based on the entire network. For the analysis of long-term errors, however, the preprocessing will not be applied, as it would suppress long-term drifts that we want to detect. Our proposed preprocessing is robust to missing values and proceeds in two steps.

First, we compute the "timescale," that is, the duration of the period where the scaling factors are assumed to be constant. It is a trade-off between short periods and long periods: the former tends to standardize the observations of the stations, thereby suppressing any differences between the observers, whereas the latter may be too coarse to correct for important changes in observers and instruments. A statistically driven study based on the Kruskal–Wallis (KW) test (Kruskal & Wallis 1952) shows that the appropriate timescale varies with the stations and with the type of counts N_s, N_g, and N_c; see the Appendix for a full description of the test. This timescale may also evolve over time when a station is constant over several months before suddenly deviating from the network. However, to avoid introducing potential biases between the stations, we use the *same* timescale, generically denoted by $\tau^\star$, for all stations over the entire period studied. The selected values of $\tau^\star$ are 8 months for N_s, 14 months for N_g, and 10 months for N_c. We note that that these periods are close to the 12-month period chosen by J. R. Wolf to compute the historical version of the scaling factors.

Second, having determined the timescale $\tau^\star$, we compute the scaling factors using ordinary least-squares regression (OLS) as follows. Recall that $Y_i(t)$ represents either the number of spots, groups, or composite actually observed in a station i, $1 \leqslant i \leqslant N$ at time t, $1 \leqslant t \leqslant T$ (daily values). For convenience, we rearrange the time by an array of two indices $t = (t_1,\ t_2)$, where $1 \leqslant t_1 \leqslant \tau^\star$ and $1 \leqslant t_2 \leqslant T/\tau^\star$. Thus, t_1 corresponds to the index of an observation inside a block of length $\tau^\star$ and t_2 is the index of the block.

Let $\boldsymbol{Y}_{i,t_2} \coloneqq [Y_i((t_1,\ t_2))]_{1 \leqslant t_1 \leqslant \tau^\star}$ denote the vector of the daily observations in station i on block t_2 of length $\tau^\star$ and $\boldsymbol{X}_{i,t_2} \coloneqq [\underset{1 \leqslant i \leqslant N}{\text{med}}\ Y_i((t_1,\ t_2))]_{1 \leqslant t_1 \leqslant \tau^\star}$ be the vector containing the daily values of the median of the network, also of length $\tau^\star$. The scaling factors are computed using the slope of the $\text{OLS}(\boldsymbol{Y}_{i,t_2} | \boldsymbol{X}_{i,t_2})$ regression:

$$\kappa_i(t_2) = (\boldsymbol{X}_{i,t_2}^T \boldsymbol{X}_{i,t_2})^{-1} \boldsymbol{X}_{i,t_2}^T \boldsymbol{Y}_{i,t_2}. \quad (8)$$

The new definition of the scaling factors in Equation (8) is a robust version of a ratio between the observations of the stations and the median of the network. It is similar to the definition of the historical k in Equation (2), where the median of the network replaces the single pilot station as the reference level. The rescaled data, denoted Z_i in the sequel, are defined as

$$Z_i(t) = \frac{Y_i(t)}{\kappa_i(t)},$$

Figure 3. Block diagram of the T procedure defining the solar signal estimator $\hat{\mu}_s(t)$, where $M_t = \underset{1 \leqslant i \leqslant N}{\text{med}}\, Z_i(t)$ is the median of the network. An Anscombe transform is first applied on the median, and missing values are imputed. Then, a fast Fourier transform (FFT) is used to convert the signal to a power spectrum in the frequency domain, followed by an attenuation from a Wiener filter. A step function cancels the amplitude of the frequencies corresponding to the periods inferior to 7 days (low-pass filter). The threshold at 7 days is selected from Dudok de Wit et al. (2016; see Figure 5). It is the smallest visible timescale of the signal, corresponding to the weekly shift of some observatories. Finally, an inverse FFT and an inverse Anscombe transform are applied to the signal.

where the reference appears now in the denominator. In a sense, the ratio in Equation (2) is inverted in order to limit the problem of dividing by zero whenever the stations observe no spots. We explored other methods such as orthogonal regression (also called total least-squares) and OLS($\boldsymbol{X}_{i,t_2}|\boldsymbol{Y}_{i,t_2}$) (Feigelson & Babu 1992). We choose the OLS($\boldsymbol{Y}_{i,t_2}|\boldsymbol{X}_{i,t_2}$) method since it leads to the smallest Euclidean and total variation distances between the median of the network and the individual stations.

5. Solar Signal Estimation

5.1. Choice of the Estimator

To use Equation (6), we need an estimate of a proxy for $s(t)$. We choose this proxy to be the mean of $s(t)$, denoted $\mu_s(t)$. We propose as a robust estimator for $\mu_s(t)$ a *transformed* version of the median of the network:

$$\hat{\mu}_s(t) = T(M_t), \tag{9}$$

where $M_t = \underset{1 \leqslant i \leqslant N}{\text{med}}\, Z_i(t)$ represents the median of the network and T denotes a transformation composed of an Anscombe transform and a Wiener filtering (Davenport & Root 1968). This filtering is applied in order to clean the data from very high frequencies, which can lead to instabilities in the subsequent analysis. The generalized Anscombe transform stabilizes the variance (Murtagh et al. 1995; Makitalo & Foi 2013). It is written as

$$A(x) = \frac{2}{\alpha}\sqrt{\alpha x + \frac{3}{8}\alpha^2}. \tag{10}$$

It is commonly applied in the literature to Gaussianize near-Poissonian variables. It is needed here, as the Wiener filtering performs better on Gaussian data. Similarly to Dudok de Wit et al. (2016), pp. 14–15, we fix $\alpha = 4.2$ in Equation (10). This is the optimal value found for the composite N_c. Before applying the Wiener filtering, missing values of the median of the network are imputed using the algorithm described in Dudok de Wit (2011). Only 49 values are imputed, which represents 0.2% of the total number of values on the period studied. The Wiener filtering is then applied on the transformed and complete set of median values and suppresses the highest frequencies of the signal. Finally, the imputed missing values were reset to NaN ("not a number") in $\hat{\mu}_s(t)$. The block diagram of the procedure is described in Figure 3.

Among other tested estimators (based on the mean, the median of the network, or a subset of stations), with or without application of T, the estimator proposed in Equation (9) turns out to be the most robust to outliers.

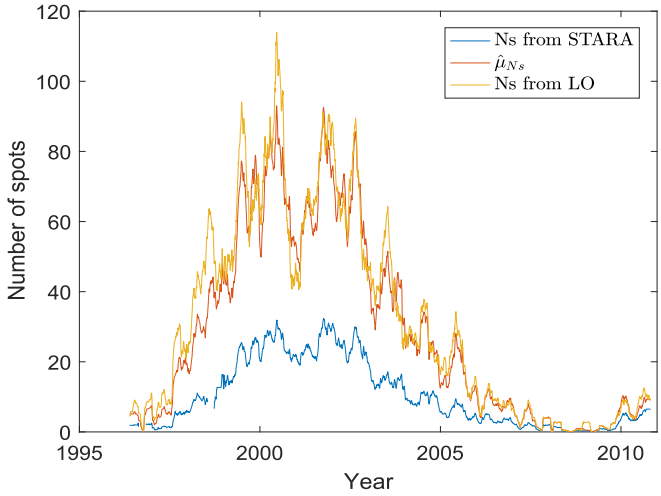


Figure 4. Comparison between the SN obtained from STARA and that from our procedures, for the period 1996 May to 2010 October. The number of spots obtained from the STARA catalog is represented in blue, the actual (unprocessed) number of spots observed in Locarno (LO) is represented in yellow, and $\hat{\mu}_{N_s}$ is plotted in orange. The three quantities shown are averaged over 81 days.

5.2. Comparison with Space Data

To test the quality of our estimator $\hat{\mu}_s$, we compare it with a sunspot number extracted from satellite images of the Sun. We expect less variability when N_s, N_g, and N_c are retrieved from satellite images using automated algorithms, as the rules to count spots and groups are clearly defined. Nevertheless, the measurements are biased by these rules, and the most complex cases, e.g., at maxima, most often require either human intervention or a specific procedure in the algorithm. In any case, a measure of the "true" number of spots and groups does not exist.

As exercise for this comparison, we use the sunspot Tracking And Recognition Algorithm (STARA) sunspot catalog (Watson & Fletcher 2010), regrouping observations from 1996 May to 2010 October (solar cycle 23). This number is extracted using an automated detection algorithm from the images obtained by the MDI instrument on *Solar and Heliospheric Observatory*. It has a lower scaling than our estimator for the number of spots, $\hat{\mu}_{N_s}$, as expected since the definition of a spot in the detection algorithm excludes the pores (spots without penumbra).

We compare three quantities on the period where STARA data are available (1996–2010): N_s (STARA), $\hat{\mu}_{N_s}$, and N_s as recorded by the Locarno station. These are shown in Figure 4. We test the level of variability by computing the mean value of a moving standard deviation over a window of 81 days. It is equal to 14.07 for N_s (STARA) rescaled on $\hat{\mu}_{N_s}$ (5.38 for N_s (STARA) without scaling) against 15.68 for $\hat{\mu}_{N_s}$ and 27.13 for Locarno. As expected, N_s (STARA) experiences less variability

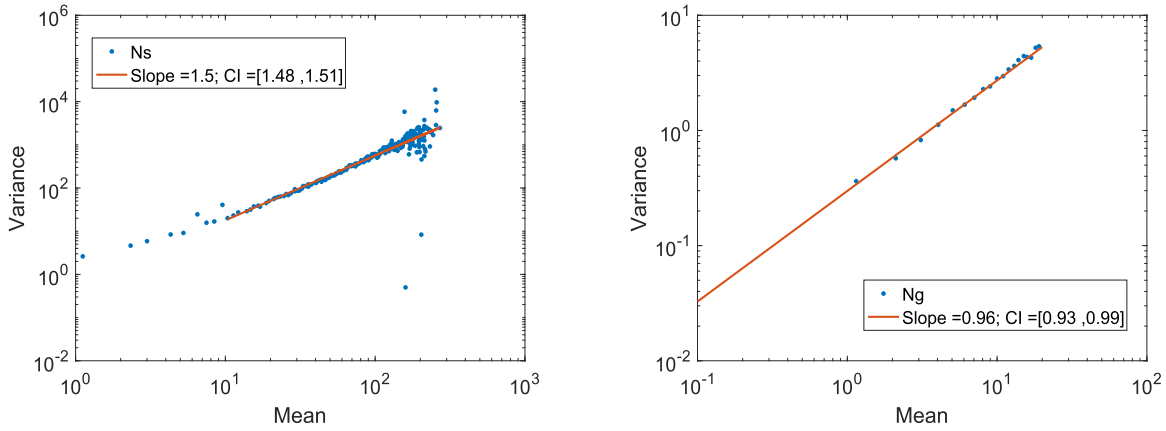


Figure 5. Estimation of the conditional mean–variance relationship for N_s (left) and N_g (right). The red line is a linear fit of the points (shown on a log–log scale), starting at $N_s > 9$ and $N_g > 0$, respectively. In both plots, the legend shows the value of the fitted slope together with its confidence interval at 95%. The value of the intercept is -1.21 for N_g and -0.57 for N_s. The same fit starting at $N_s > 0$ (not shown here) leads to a slope of 1.25 and $\mathrm{CI}_{95\%} = [1.23, 1.28]$.

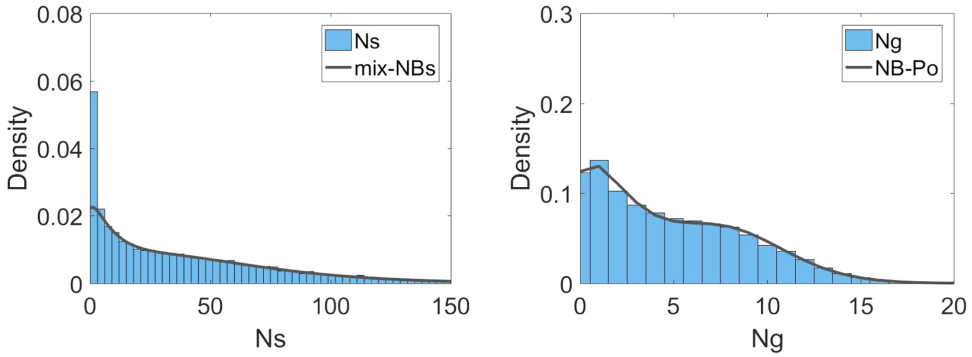


Figure 6. Left: histogram of $\hat{\mu}_{N_s}(t)$ values, computed with a bandwidth (binning) equal to 3, and estimated density for nonzero values of $\hat{\mu}_{N_s}(t)$ (shown by the black line). The complete density is modeled by a ZA mixture of generalized NBs. For the zero values, the MLE value of the Bernoulli parameter is equal to $b = 0.1$. For nonzero values, the MLE values of the parameters in Equation (12) are $r_1 = 1.25$, $p_1 = 0.11$, $r_2 = 2.39$, $p_2 = 0.04$, and $w_1 = 0.32$. Right: histogram of $\hat{\mu}_{N_g}(t)$ values, computed with a bandwidth equal to 1, and corresponding density fitted by MLE (shown in black line). The density is modeled by a mixture of an NB and a Poisson distribution as defined in Equation (13). The fitted parameter values are $\mu_2 = 8.62$, $r_1 = 1.65$, $p_1 = 0.37$, and $w_1 = 0.36$.

than a single station, but its variability is comparable to the one of our estimator.

Nevertheless, satellite images of the Sun have only been available since 1980, and data extracted from those images cannot be traced back until the seventeenth century. Gathering space observations during several decades also requires the use of different satellites and instruments, as instruments age in space. These instruments need calibrations that create additional inaccuracies to the extracted numbers. We thus conclude that $\hat{\mu}_s(t)$ is a more robust estimator of the solar activity, and it will be used as a proxy for $s(t)$ in the sequel.

5.3. Solar Component for $\mathrm{N_s}$ and $\mathrm{N_g}$

We present here the statistical modeling of the number of spots N_s and the number of groups N_g. We do this *separately* for each component since their physical origins are driven by different phenomena: the groups convey information about the dynamo-generated magnetic field in the solar interior, whereas the emergence of individual spots would rather come from fragmented surface flux and agglomeration of small magnetic fields (Thomas & Weiss 2008). Together, the analysis of the spots and groups helps us to better understand the composite N_c and the solar activity, which is not satisfactorily described by only one of the two numbers (Dudok de Wit et al. 2016). In the remainder of this paper, we define a specific notation for the generic $\hat{\mu}_s(t)$ from Equation (9): $\hat{\mu}_{N_s}(t)$ for the number of spots, $\hat{\mu}_{N_g}(t)$ for the number of groups, and $\hat{\mu}_{N_c}(t)$ for the composite.

The authors in Dudok de Wit et al. (2016) showed that the numbers of spots and groups experience more overdispersion than actual Poisson variables. In order to estimate how far the distribution of the "true" $s(t)$ departs from a Poisson distribution, we regress the conditional variance $\mathrm{Var}(\mathrm{Z}_i(\mathrm{t})|\hat{\mu}_s(\mathrm{t}) = \mu)$ versus the conditional mean $\mathbb{E}(Z_i(t)|\hat{\mu}_s(t) = \mu)$ by OLS; see Figure 5. Whereas in a Poisson context the slope of the fit should be close to 1, for $N_s > 10$, our fit shows a slope of 1.5, with confidence interval (CI) $\mathrm{CI}_{95\%} = [1.48, 1.51]$. This points to overdispersion and the need for a generalization of a Poisson pdf. On the contrary, the same plot for $N_g > 0$ displays a slope of 0.96, with $\mathrm{CI}_{95\%} = [0.93, 0.99]$, confirming the validity of a Poisson process assumption. Note that values <11 are excluded from the fit of N_s, as they seem to indicate a different regime. This change in the alignment may indicate the presence of a multimodal distribution; see Figure 6.

Count data with overdispersion are widely modeled by the negative binomial (NB) distribution in the literature (Colin Cameron & Trivedi 2013; Rodriguez 2013) or by its generalization

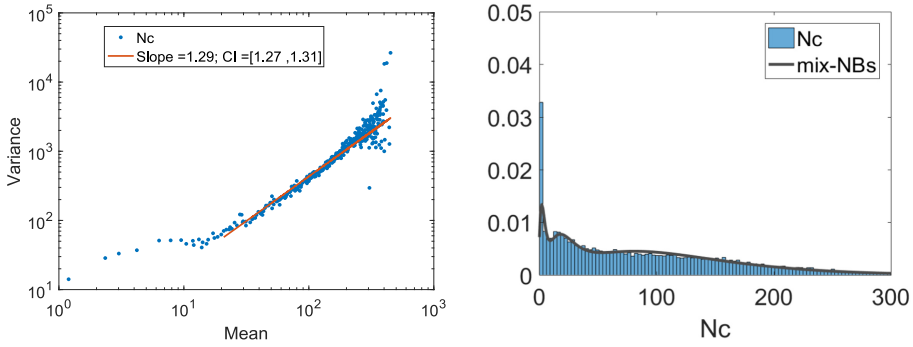


Figure 7. Left: conditional mean–variance relationship for N_c, shown on a log–log scale. Right: histogram of $\hat{\mu}_{N_c}(t)$ values, with a binning equal to bw = 3. The estimated density outside of zero values is shown by a black line. It is modeled as a mixture of three NB distributions (see Equation (14)), with MLE parameter values equal to $r_1 = 3.18$, $p_1 = 0.48$, $r_2 = 4.02$, $p_2 = 0.15$, $r_3 = 3.05$, $p_3 = 0.02$, $w_1 = 0.08$, and $w_2 = 0.19$. The Bernoulli parameter of the density f_0 at zero is equal to $b = 0.07$.

(Jain & Consul 1970):

$$g(x, r, p) = \frac{\Gamma(r + x)}{\Gamma(r)\Gamma(x + 1)} p^r q^x, \tag{11}$$

where $r > 0$, $p \in (0,1)$, $q = (1 - p)$ and Γ is the gamma function.

A visual inspection of the histogram of estimated values $\hat{\mu}_{N_s}(t)$ in the left panel of Figure 6 reveals a local maxima in the distribution around 20–40. We refer to these local maxima as *modes* in the remainder of the article. The underlying density of $\hat{\mu}_{N_s}(t)$ may thus be multimodal, as suspected from the left panel of Figure 5. Such pdf's are classically modeled by a mixture model. As the density shows a typical excess of zeros as well, it requires the use of a ZA distribution defined in Equation (7). We thus fit the complete pdf of the estimated number of spots, $\hat{\mu}_{N_s}(t)$, by a ZA mixture of generalized NB distributions. The density at zero, $f_0(x)$, is represented by a Bernoulli distribution, whereas the density outside zero, $f_1(x)$ in Equation (7), is identified by a mixture of NB distributions:

$$f_1(x, r_1, r_2, p_1, p_2) = w_1 g_1(x, r_1, p_1) + (1 - w_1) g_2(x, r_2, p_2), \tag{12}$$

where g_1, g_2 are NB densities and w_1 is the mixture weight.

Similarly, the histogram of $\hat{\mu}_{N_g}(t)$ exhibits a clear excess in the range of 1–3 compared to a Poisson-like distribution centered between 5 and 8. The pdf of $\hat{\mu}_{N_g}(t)$ shows thus two modes: one around 1–3, and one around 5–8. Such a pdf may be modeled by a mixture of NB and Poisson distributions:

$$f(x, r_1, p_1, \mu_2) = w_1 g(x, r_1, p_1) + (1 - w_1) \frac{\mu_2^x}{x!} e^{\mu_2}, \tag{13}$$

where $\mu_2 > 0$ and, as above, w_1 is the mixture weight.

The fit of these parametric densities is shown in Figure 6 by a black line superimposed on the histograms. All the fits in this article are computed using the maximum likelihood estimation (MLE). The nature of the different modes in the pdf of $\hat{\mu}_{N_s}$ and $\hat{\mu}_{N_g}$ will be discussed in Section 5.5.

5.4. Solar Component for N_c

We now use Equation (9) to estimate the μ_{N_c}, the "true" value of the composite $N_c = N_s + 10N_g$. Again looking at the conditional mean–variance relationship, we observe in the left panel of Figure 7 an overdispersion with a slope of 1.29 and $\text{CI}_{95\%} = [1.27, 1.31]$ for $N_c > 20$. As a compound of both quantities, N_c experiences less overdispersion than N_s and more than N_g. A visual inspection of the histogram of $\hat{\mu}_{N_c}(t)$ values in the right panel of Figure 7 indicates an excess of zeros and several modes, most probably coming from the modes observed in the pdf's of $\hat{\mu}_{N_g}$ and $\hat{\mu}_{N_s}$. We thus find it appropriate to approximate the density of $\hat{\mu}_{N_c}(t)$ by a ZA mixture of three NB distributions, where the density outside zero values, $f_1(x)$ in Equation (7), is identified with

$$f_1(x, r_1,\ldots,r_3, p_1 ,..,p_3) = \sum_{i=1}^{3} w_i g_i(x, r_i, p_i), \tag{14}$$

where w_i are the mixture weights and $\sum_{i=1}^{3} w_i = 1$. The fit of f_1 is represented in the right panel of Figure 7 by a black line.

A statistical analysis (not presented here) shows that the distribution of the ISN is statistically close to the distribution of $\hat{\mu}_{N_c}$. The uncertainty analysis for $\hat{\mu}_{N_c}$, presented in the remainder of the article, remains thus valid for the ISN.

5.5. Conditional Correlation

Due to the physical nature of the data, the local maxima for the densities of $\hat{\mu}_{N_s}$ and $\hat{\mu}_{N_g}$ are not independent. We therefore look at the conditional correlation $\text{Corr}(N_s, N_g|\hat{\mu}_{N_c} = s)$ with the goal to better understand the nature of the modes observed in these two densities, and thus also in the density of $\hat{\mu}_{N_c}$.

Figure 8(a) shows the conditional correlation for different values of $\hat{\mu}_{N_c}$ between 0 and 400. Note that even when $\hat{\mu}_{N_c} = s$, the value of the composite $N_s + 10N_g$ for a particular station may be larger (resp. smaller) than s. Our analysis highlights three regimes of activity:

Minima: $\hat{\mu}_{N_c} \in [0, 11]$. Here, the number of spots and groups oscillates between 0 and 1. As the number of spots equals exactly the number of groups, the correlation is high.

Medium activity: $\hat{\mu}_{N_c} \in [12, 60]$. The correlation progressively decreases, because the number of spots increases faster than the number of groups and then stabilizes. This regime is characterized by the development of smaller spots without penumbra or with a small penumbra. Figure 8(b) shows the bivariate boxplot of N_s and N_g when $\hat{\mu}_{N_c} = 40$. For $N_g = 1$ or $N_g = 2$, we observe values of N_s as high as 40. We clearly observe groups containing a large number

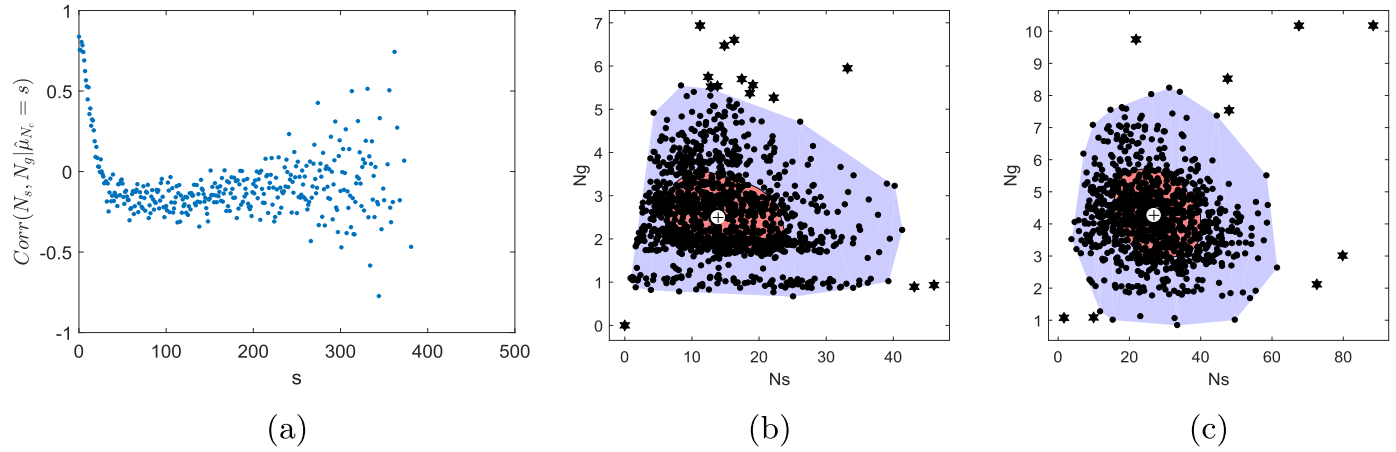


Figure 8. (a) Conditional correlation of N_s and N_g: Corr(N_s, $N_g|\hat{\mu}_{N_c} = s$) for $s \in [0, 400]$. Bivariate boxplot (also called "bagplot") of N_s and N_g when (b) $\hat{\mu}_{N_c} = 40$ and (c) $\hat{\mu}_{N_c} = 70$. The white cross represents the depth median (Rousseeuw et al. 2012). The bag contains 50% of the observations, and it is represented by a polygon in red. The fence (not represented) is obtained by inflating the bag by a factor three. The observations that are outside of the bag but inside of the fence are indicated by a light-gray loop. Outliers are represented by a black star. The correlation is indicated by the orientation of the bag.

of spots, as well as groups, composed of fewer spots, that appear progressively as the penumbra grows and that indicate a transition toward groups with fewer but larger spots. The effect of this transition from small to larger spots is observed in Figure 5 from Clette & Lefèvre (2016).

High activity: $\hat{\mu}_{N_c} > 60$. Figure 8(c) shows the bivariate boxplot of N_s and N_g when $\hat{\mu}_{N_c} = 70$. The plot has a potato shape around ($N_s = 30$, $N_g = 4$). We now observe all kinds of groups. The correlation between groups and spots slightly increases as the number of groups begins to grow as well.

The *medium* and *high* regimes are reflected in the estimated densities of $\widehat{\mu}_{N_s}$ and $\widehat{\mu}_{N_g}$ in Figure 6. The first mode of $\hat{\mu}_{N_g}$, ranging from 1 to 3, corresponds to the *medium* regime, while the second mode, ranging from 5 to 8, reflects the *high* regime. The two distinct regimes provide another justification for the use of a multimodal distribution to characterize the pdf of $\hat{\mu}_{N_g}$. Similarly, there is also a mode in the distribution of $\widehat{\mu}_{N_s}$ around 20–40 that comes from the transition between the *medium* and the *high* regime. The mode is correctly represented by a mixture model. The study of the conditional correlation constitutes the first step toward retrieving the distribution of N_c from its composites N_s and N_g. However, this task is challenging and goes beyond the scope of the article because (1) the distributions of N_s and N_g are complex mixtures and (2) the number of spots is nontrivially correlated to the number of groups.

6. Distribution of Errors

We are now in a position to analyze the error distribution in sunspot counts, the modeling of the distributions of ϵ_3, ϵ_1, and ϵ_2. To do so, we separate minima from nonminima regimes. We also consider two timescales: short-term periods, that is, timescales smaller than one solar rotation (27 days), and long-term periods. Section 6.1 estimates error at solar minima, i.e., when $s(t) = 0$. Section 6.2 analyzes short-term variability of the preprocessed observations when $s(t) > 0$. For the study of long-term error in Section 6.3, we use raw data that did not undergo any preprocessing, in order to be able to detect sudden jumps and/or large drifts in the time series. The correct timescale for the long-term period is also determined in this section, based on a statistically driven procedure. Finally, Section 6.4 compares the characteristics of the different stations based on the error analysis.

6.1. Error at Minima

The study of solar minima periods is complex, as the data show a large variability and dichotomy. Observed values of the error at minima, ϵ_3, are defined as counts made by the stations when the proxy for $s(t)$, defined in Equation (9), is equal to zero:

$$\hat{\epsilon}_3(i, t) = Z_i(t) \text{ when } \hat{\mu}_s(t) = 0, \tag{15}$$

where $Z_i(t)$ corresponds to N_s, N_g, or N_c, and where the generic $\widehat{\mu}_s(t)$ has to be replaced by $\widehat{\mu}_{N_s}(t)$ for N_s, $\widehat{\mu}_{N_g}(t)$ for N_g, and $\widehat{\mu}_{N_c}(t)$ for N_c.

A visual inspection of the histogram of N_s (resp. N_g) in the left (resp. middle) panel of Figure 9 shows an important amount of "true" zeros together with two modes around one and two. Similar modes occur around 11 and 22 in the distribution of N_c in the right panel of Figure 9, as expected. These modes represent short-duration sunspots. Due to the nonsimultaneity of the observations between stations, the proxy for $s(t)$ might be equal to zero even if some spots appear shortly (from several minutes to several hours) on the Sun. These modes can be represented by a t-location-scale (t-LS) distribution, which is a generalization of the Student t-distribution. This distribution has three parameters to accommodate for asymmetry and heavy tails: the location μ, scale $\sigma > 0$, and shape $\nu > 0$ (see Taylor & Verbyla 2004; Evans et al. 2000). Its pdf is defined as

$$g(x, \mu, \sigma, \nu)_{t-\mathrm{LS}} = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(\nu + \frac{(x-\mu)^2}{\sigma^2}\right)^{-\left(\frac{\nu+1}{2}\right)}. \tag{16}$$

The large proportion of zero values for $\hat{\epsilon}_3$ requires the use of a ZA model as in Equation (7). We choose a ZA mixture of t-LS

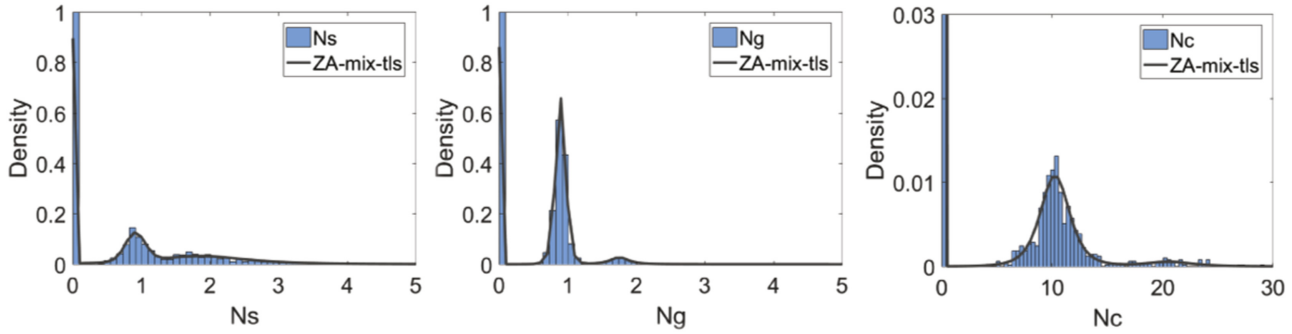


Figure 9. Truncated histograms of $\hat{\epsilon}_3$ for N_s (left), N_g (middle), and N_c (right). The solid line shows the fits using a ZA mixture of t-LS distributions, defined in Equation (17). The values of the Bernoulli parameter in Equation (7) are equal to $b = 0.9$ (left), $b = 0.86$ (middle), and $b = 0.96$ (right). They represent the proportion of "true" zeros. The parameter values for the t-LS fit are (left, for N_s) $\mu_1 = 0.91$, $\sigma_1 = 0.14$, $\nu_1 = 31.16$, $\mu_2 = 1.85$, $\sigma_2 = 0.71$, $\nu_2 = 2.09$, $w_1 = 0.6$; (middle, for N_g) $\mu_1 = 0.89$, $\sigma_1 = 0.07$, $\nu_1 = 6.89$, $\mu_2 = 1.75$, $\sigma_2 = 0.14$, $\nu_2 = 1.33$, $w_1 = 0.09$; and (right, for N_c) $\mu_1 = 10.24$, $\sigma_1 = 1.37$, $\nu_1 = 3.89$, $\mu_2 = 20.57$, $\sigma_2 = 2.33$, $\nu_2 = 1.93$, $w_1 = 0.08$. The bin width (bw $= 0.0917$) is the same for the histograms of both N_s and N_g. It is related to the sample size and the data range of N_s by a simple rule proposed by Scott (1979). The bin width of the histogram of N_c (right) is equal to bw $= 0.4192$ and is also computed by Scott's rule. Note that the right panel is enlarged: the value at zero is 0.96 and not 0.03.

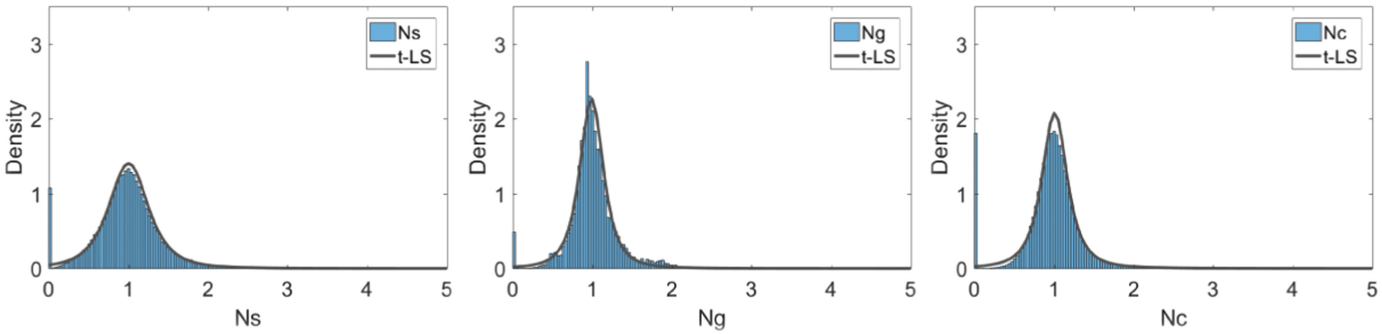


Figure 10. Histograms of $\hat{\tilde{\epsilon}}$ for N_s (left), N_g (middle), and N_c (right). The solid line shows the fits using a t-LS distribution defined in Equation (16). The values of the Bernoulli parameter in Equation (7) are equal to $b = 0.04$ (left), $b = 0.02$ (middle), and $b = 0.06$ (right). They represent the proportion of false "zeros," i.e., stations reporting no sunspot where there are some. The parameter values for the t-LS fit are (left, for N_s) $\mu = 1$, $\sigma = 0.26$, $\nu = 2.8$; (middle, for N_g) $\mu = 0.99$, $\sigma = 0.16$, $\nu = 2.33$; and (right, for N_c) $\mu = 1.01$, $\sigma = 0.17$, $\nu = 2.12$. The bin widths (bw) of the histograms are computed using Scott's rule. For N_s and N_g they are the same (bw $= 0.0328$) and for N_c it is equal to bw $= 0.0433$.

for the complete distribution of $\hat{\epsilon}_3$. The density outside of zero, $f_1(x)$ in Equation (7), is thus identified by such a mixture of t-LS distributions:

$$f_1(x, \mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2) = w_1 g(x, \mu_1, \sigma_1, \nu_1)_{t-\mathrm{LS}} + (1 - w_1) g(x, \mu_2, \sigma_2, \nu_2)_{t-\mathrm{LS}}, \quad (17)$$

where, as before, w_1 is the mixture weight. The histograms and fitted distributions for $\hat{\epsilon}_3$ are shown in Figure 9. The visual closeness between the histogram and the fitted distribution was used as a criterion to select the best pdf among a few intuitive candidates, while the parameters of the distribution are estimated via MLE.

In the previous figures, where the error at minima is represented for all stations combined, outliers defined as $\hat{\epsilon}_3(i, t) > 2$ are not visible for N_g and N_s. A separate analysis (not presented here) shows that the percentage of outliers in each station is low (inferior to 0.5% for N_s). Some stations also observed a high maximal value at minima (e.g., a value of 35 was recorded in QU [Quezon, Philippines] for N_s). This extreme value for minima may correspond to a transcription error that might be verified in the future, before being encoded in the SILSO database.

6.2. Short-term Variability

When the proxy for $s(t)$, defined in Equation (9), is different from zero, the short-term error $\tilde{\epsilon}$ may be estimated using

$$\hat{\tilde{\epsilon}}(i, t) = \frac{Z_i(t)}{\hat{\mu}_s(t)} \text{ when } \hat{\mu}_s(t) > 0. \quad (18)$$

To select the best distribution, we proceed as follows. Different densities are fitted to the values of $\hat{\tilde{\epsilon}}$, outside of zero, using MLE.[8] Then, the AIC criterion is used to choose the best pdf, which in this case is a t-LS distribution.

As we observe an excess of zero, we need a ZA t-LS distribution to represent the complete distribution of $\hat{\tilde{\epsilon}}$. Figure 10 shows the histogram and the fitted pdf of $\hat{\tilde{\epsilon}}$ outside of zero. For the latter, the mean is close to 1, indicating that on average the stations are aligned with $\hat{\mu}_s(t)$. The histogram exhibits a probability mass at zero representative of "false" zeros, that is, of stations that do not observe any sunspot when there are actually some on the Sun. The histogram also shows a tail on the right-hand side, caused by outliers. This asymmetry requires a t-LS rather than a Gaussian distribution to be fitted.

[8] We use the function "'allfitdist.m," last modified in 2012, in Matlab R2016b.

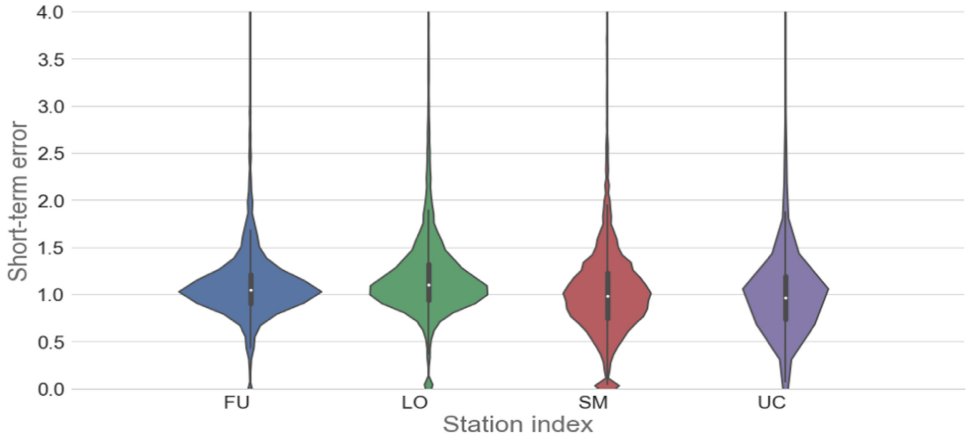


Figure 11. Truncated violin plots of the estimated short-term variability $\widehat{\tilde{\epsilon}}$ for N_s in four stations (FU, LO, SM, and UC). A violin plot (Hintze & Nelson 1998) combines a vertical boxplot with a smoothed histogram represented symmetrically to the left and right of the box. The white dot in the center of the violin locates the mean of the distribution. The thick gray bar shows the interquartile range, and the thin gray bar depicts the interdecile range. The bin width (bw = 0.05) is the same for all stations and is computed with Scott's rule.

The violin plots of four different stations are shown in Figure 11 for the number of spots N_s, where the differences between the stations are the most visible. The mean of the Locarno station (LO), the current reference of the network, is slightly higher than the three other means (and higher than the means of all other stations), around 1.19. This results from its particular way of counting: large spots (with penumbra) count for more than small spots without penumbra.

Another characteristic feature is how the error is distributed around the mean. A violin plot may be seen as a pdf with the x-axis of the density drawn along the vertical line of the boxplot. For example, the pdf of the short-term error of LO is concentrated around the mean, but the entire distribution is shifted upward, unlike the pdf of Uccle (UC), which has much lower values. UC is a professional observatory. Different observers record from one week to another the number of spots, groups, and composite on the Sun. As their experience and methodology slightly change, the shift of observers probably increases the short-term variability of the station. Usually a team of observers experience more variability than a single person, like in FU (Fujimori, Japan). This station has remarkable short-term stability.

Similarly, the San Miguel (SM) station shows the typical shape of a professional observatory. On the other hand, the LO station shows an $\widehat{\tilde{\epsilon}}$ distribution almost characteristic of a single observer: that is because until recently LO had one dominant main observer.

6.3. Long-term Variability

A generic estimator for the long-term error $\epsilon_2(i, t)$ may be defined by

$$\hat{\mu}_2(i, t) = \left(\frac{Y_i(t)}{M_t}\right)^{\star} \text{ when } M_t > 0, \tag{19}$$

where the $\star$ denotes the smoothing process, $Y_i(t)$ are the raw observations, and $M_t = \underset{1 \leqslant i \leqslant N}{\text{med}}\, Z_i(t)$ is the median of the network. The T transform from Equation (9) is not required here, as we apply a moving average (MA) of length L defined below.

This length L should be larger than what is considered as short-term, that is, periods inferior to one solar rotation (27 days). Long-term, on the other hand, is usually defined as periods above 81 days (Dudok de Wit 2011), beyond which the effects of the solar rotation and of the sunspot's lifetime are negligible. The midterm temporal regime corresponds to periods between 27 and 81 days. To select the long-term scale for a given station i, we make the assumption that, for all t belonging to a window of length L, we have

$$\mathbb{E}(\epsilon_2(i, t)) = \mu_2(i, t) \simeq C_i, \tag{20}$$

where C_i is a constant, which might differ from 1. Having $C_i = 1$ means that the station i is at the same level as the median of the network. We test different lengths L ($L > 27$ days) for the MA window and select the long-term regime as the shortest length for which the above assumption is valid. We consider thus $\hat{\mu}_2(i, t)$s of Equation (19) in sliding windows of length L over the total period (1947–2013). We apply a nonparametric equivalent of the t-test (the Wilcoxon rank sum test; Bridge & Sawilowsky 1999) on the $\hat{\mu}_2(i, t)$s to test whether Equation (20) is verified within each window. Longer windows correspond thus to a stronger smoothing but also contain more values to test. As a result of this procedure, we define the long-term regime as all scales above 81 days, as we found this to be the shortest length such that the constant assumption on $\mu_2(i, t)$ is not violated more than roughly 10% of the time. This ties in with what solar physicists consider as the long-term regime.

Depending on our interest in detecting long-term drifts or jumps, different window lengths may be chosen in Equation (19) (some well above 81 days). Indeed, drifts require long smoothing periods (several months, or even years) to be observed, whereas jumps might be oversmoothed by such long smoothing and hence need a smaller MA window.

Figures 12(a)–(d) represent the long-term drifts associated with N_s in four stations starting from 1960. Figures 12(e)–(h) show the scaling factors $\kappa_i(t_2)$s for the same stations used at short-term and minima regimes. We do not represent years before 1960 because FU and SM show too few observations in that period. FU and UC appear relatively stable, unlike stations

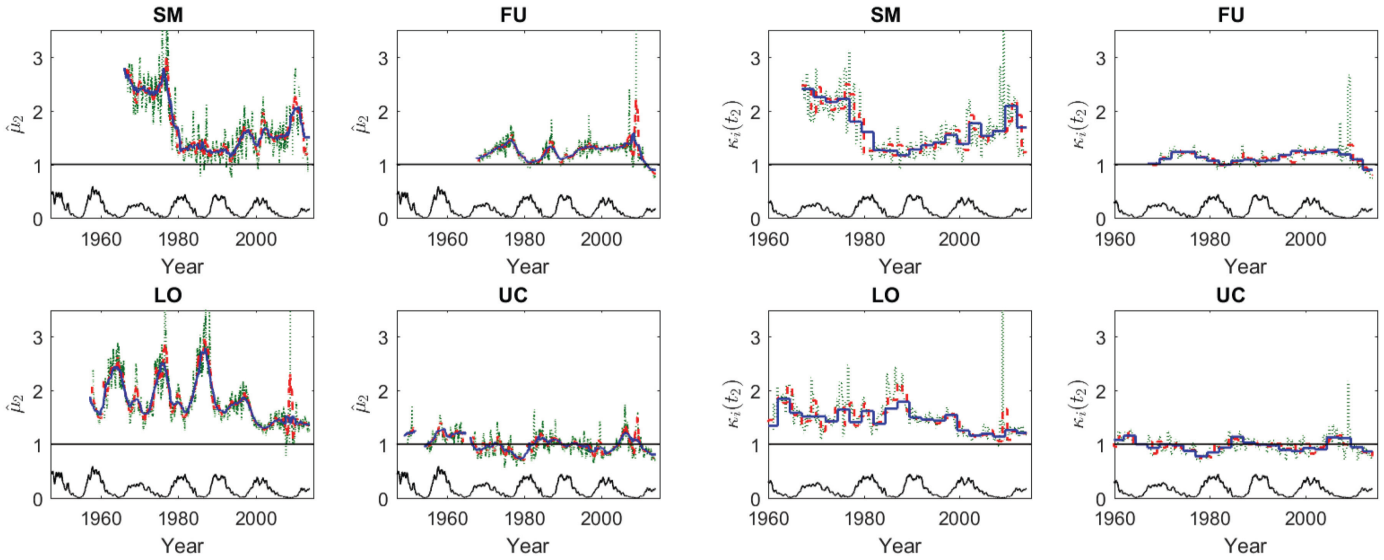


Figure 12. (a–d) Estimation of $\hat{\mu}_2(i, t)$ for N_s in four stations (SM, FU, LO, and UC). $\hat{\mu}_2(i, t)$ is computed with different MA window lengths: 81 days (green dotted line), 1 yr (red dashed line), and 2.5 yr (blue solid line). (e–h) Estimation of the scaling factors for N_s in the same stations. The $\kappa_i(t_2)$s, with $1 \leqslant t_2 \leqslant T/\tau$, are computed using the OLS($\boldsymbol{Y}_{i,t2}|\boldsymbol{X}_{,t2}$) regression in Equation (8) on a block of $\tau = 81$ days (green dotted line), 1 yr (red dashed line), and 2.5 yr (blue solid line). The solar cycle is represented in black at the bottom of the figures for N_s.

LO and SM, which display severe drifts. Bias in the counting process is also larger during solar minima when there are short-duration sunspots. This effect is clearly visible in LO. Indeed, erroneous encoding of counts leads to much higher relative errors during minima than during the remaining part of the solar cycle. Some jumps are also visible on the graphs with the smallest MA length (81 days) in green. This scale is more appropriate to observe the jumps, while longer scales only highlight the long-term drifts of the stations.

We emphasize here the strong link between the preprocessing and the long-term analysis. Indeed, the scaling factors presented in Section 4 are a rough estimate of the long-term error, inspired by the historical procedure of J. R. Wolf. This rough estimation is required to rescale the stations to the same level. This rescaling is used to compute the median M_t of Equation (19). Contrarily to the piecewise constant κ_is computed in Section 4, the $\hat{\mu}_2(i, t)$s are smooth over time and hence are more adapted to a future monitoring of the stations.

6.4. Comparing Stations with Respect to Their Stability

In previous sections, we presented separately the estimations of the short-term error $\hat{\tilde{\epsilon}}(i, t)$, the long-term error $\hat{\mu}_2(i, t)$, and the error at minima $\hat{\epsilon}_3(i, t)$. All three types of errors are needed to assess the quality and stability of one station. It is more important for a station to have a low variability (low interquantile range) than to be aligned on the mean on the network. Indeed, as seen in Section 4, it is easy to rescale a station on the mean of the network.

Figure 13 displays a visual representation of long-term against short-term error for each station. It shows the long-term versus short-term empirical interquantile range on a 2D plot and thus characterizes the stability of the stations outside of minima. Stations in red are the teams of observers. They usually experience more short-term variability than an individual. We see that MO (Mochizuki, Japan), FU, and KOm (Koyama, Japan) have low variability in both the short and long term. They correspond to long individual observers

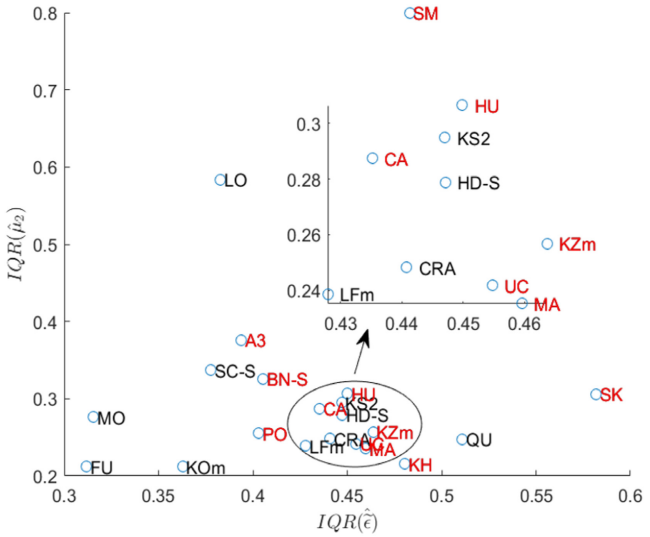


Figure 13. Scatter plot showing the interquantile range of the estimated short-term error $\hat{\tilde{\epsilon}}(i, t)$ and the interquantile range of the estimated long-term error $\hat{\mu}_2(i, t)$, station by station. Stations in red represent the teams of observers; the others are single observers.

with stable observation practices. On the other hand, the LO station shows a poor long-term stability, while its short-term variability is remarkably low for a professional observatory. As mentioned earlier, this is due to the fact that there is a main observer. UC shows a large variability in the short term (due to many observers) but an interesting long-term stability, as already noticed in Figure 11. SM experiences the most severe long-term variability of the network. It also has a large short-term variability, characteristic of a team of observers. QU shows a large short-term variability and a low long-term variability level. Although it seems that it is a single observer, it appears there was a move from one place to another during the observing period, and maybe a change of instrument that would impact the short-term variability. This surprising behavior will prompt SILSO to ask for more metadata.

7. Conclusion and Future Prospects

In this article, we propose the first comprehensive uncertainty model in a *multiplicative* framework for counting spots, groups, and composite on the Sun. Our approach is robust to missing values and was applied on 66 yr of data (1947–2013). We presented several parametric models for the density of the "true" N_s, N_g, and N_c, as well as for the density of their error distribution at minima, short term, and long term. This error quantification allows proposing a first classification of the 21 stations of our pool based on their stability. It shows that the observatories are affected differently by the various types of errors: some are stable with respect to the network at short term but experience large drifts, and vice versa. The analysis highlights the hazards of using a single pilot station as the unique reference of the network.

We intend to use the error models presented in Section 6 for a parametric monitoring of all stations of the network, with a particular focus on new stations. Data from newborn observatories can be recorded for several months. Their distributions may then be compared to the density of the short-term error (or the error at minima if we are in a minima period) of the entire network obtained in this paper. If the stations experience similar errors, they may be included in the network. Otherwise, the stations might need to improve or correct their observing procedure before entering the SILSO network.

A nonparametric monitoring that aims to detect in quasi-real time the long-term drifts of the stations in the network is also under development. An example of a classical monitoring procedure is the CUSUM chart (Koshti 2011). It is frequently used to control the production quality in industry. The chart accumulates the deviations of the mean value above a reference level in a statistic. If the value of the statistic exceeds a predefined threshold depending on the standard deviation of the process, the process is considered out of control and an alert is given. This simple method based on the two first moments of the distribution is obviously not adequate to control heavily skewed variables such as the number of spots. More complex methods need to be developed that will strongly depend on the models of the data.

The work presented here enhances our comprehension of the ISN and its error. It is part of a larger project that aims at improving the quality of the ISN. We started this project a few years ago when a revised version of the ISN was published (Clette & Lefèvre 2016), and we will pursue with the future monitoring of the stations to provide a yet missing quality-control procedure for the ISN. As the ISN is used as a benchmark in several fields of astrophysics and space physics, this is a much-needed task.

This work benefited from highlighting discussions with T. Dudok de Wit. The first author gratefully acknowledges funding from the Belgian Federal Science Policy Office (BELSPO) through the BRAIN VAL-U-SUN project (BR/165/A3/VAL-U-SUN). We also want to thank the International Space Science Institute (ISSI-Bern) and the members of the "Recalibration of the sunspot Number Series" team for providing support.

Appendix
Timescales of the Preprocessing

This appendix details the statistical procedure selecting the timescales of the preprocessing described in Section 4. It is composed of three steps. First, the daily scaling factors are computed using

$$\kappa_i((t_1, t_2)) = \frac{Y_i((t_1, t_2))}{\mathrm{med}_{1 \leqslant i \leqslant N} Y_i((t_1, t_2))}, \tag{21}$$

where, as in Section 4, we rewrite the time by an array of two indices $1 \leqslant t_1 \leqslant 30$ and $1 \leqslant t_2 \leqslant T/30$, corresponding, respectively, to the day and the month of the observation.

Second, the nonparametric KW test (Kruskal & Wallis 1952) is applied on blocks of 30 factors, since the "k-coefficients" of Equation (2) are currently estimated on a monthly basis at the WDC-SILSO. Let $\boldsymbol{\kappa}_{i,t_2} = [\kappa_i((t_1, t_2))]_{1 \leqslant t_1 \leqslant 30}$ denote the vector of the daily factors on 1 month. The test assesses whether the $\boldsymbol{\kappa}_{i,t_2}$s of consecutive months are statistically different. The procedure starts by comparing the distribution of the first month of the period studied, $\boldsymbol{\kappa}_{i,1}$, to the distribution of the second month, $\boldsymbol{\kappa}_{i,2}$. If the test shows that both distributions are significantly different, the next two distributions $\boldsymbol{\kappa}_{i,2}$ and $\boldsymbol{\kappa}_{i,3}$ are tested. Otherwise, the distribution of the first two months $[\boldsymbol{\kappa}_{i,1}\, \boldsymbol{\kappa}_{i,2}]$ is compared to the distribution of the third month $\boldsymbol{\kappa}_{i,3}$. The algorithm is iterated until the end of the period, for each station. Note that the KW test performs well when comparing two or more independent samples of unequal sizes. The correlation of the data is thus neglected in this procedure. Despite the presence of correlations between consecutive days, the correlation between consecutive months is low. The test provides thus a station-specific segmentation, shown in Figure 14 for N_s. The length of the segments indicates the number of consecutive blocks of scaling factors that come from the same distribution. We assume that these factors are constant within each segment.

In the last step, the global timescales for each index are defined from the segmentations of the individual stations. The length of the most frequent segment is first selected in each station. Then, a global scale is estimated from the median of the most frequent lengths by station, for N_s, N_g, and N_c.

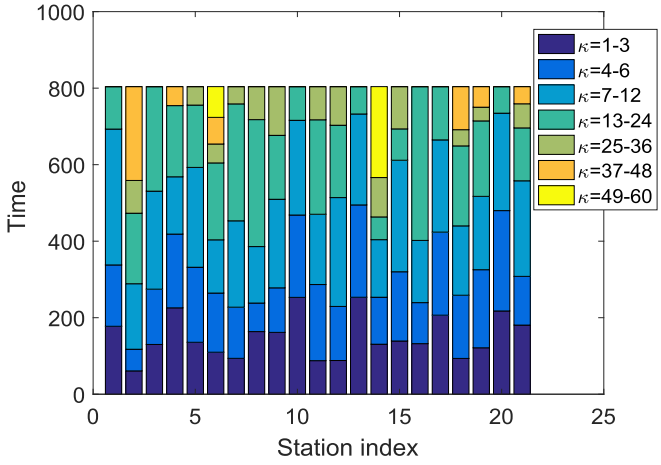


Figure 14. Bar chart representing the results of the KW test applied to the scaling factors for N_s. The x-axis represents the stations indexed from 1 to 21, and the y-axis shows the total period studied expressed in months (1 unit $\approx$ 30 days). The y-axis is not ordered in time, for readability purposes, but it is ordered with respect to the length of the segments. The colors of the chart correspond to the number of blocks that may be grouped into a single factor ("$\kappa = 5$" means that a single scaling factor may be computed for a period of 5 months).

ORCID iDs

Sophie Mathieu https://orcid.org/0000-0003-2105-9733
Laure Lefèvre https://orcid.org/0000-0003-1005-7353

References

Bridge, P., & Sawilowsky, S. 1999, Journal of Clinical Epidemiology, 52, 229
Chang, H.-Y., & Oh, S.-J. 2012, JASS, 29, 97
Clette, F., Berghmans, D., Vanlommel, P., et al. 2007, AdSpR, 40, 919
Clette, F., Cliver, E. W., Lefévre, L., et al. 2016, SoPh, 291, 2479
Clette, F., & Lefèvre, L. 2016, SoPh, 291, 2629
Colin Cameron, A., & Trivedi, P. K. 2013, Regression Analysis of Count Data (2nd ed.; Cambridge: Cambridge Univ. Press)
Davenport, W. B., & Root, W. L. 1968, Random Signals and Noise (New York: McGraw-Hill)
Dudok de Wit, T. 2011, A&A, 533, 29
Dudok de Wit, T., Lefèvre, L., & Clette, F. 2016, SoPh, 291, 2709
Evans, M., Hastings, N., & Peacock, B. 2000, Statistical Distributions (3rd ed.; New York: Wiley)
Feigelson, E., & Babu, G. 1992, ApJ, 397, 55
Foster, G. 1999, JAVSO, 27, 177
Hathaway, D. H. 2010, LRSP, 7, 1
Hintze, J. L., & Nelson, R. D. 1998, The American Statistician, 52, 181
Izenman, A. 1985, The Mathematical Intelligencer, 7, 27
Jain, G., & Consul, P. 1970, SIAM J. Appl. Math., 21, 501
Koshti, V. V. 2011, International Journal of Physics and Mathematical Sciences, 1, 28
Kruskal, W., & Wallis, W. 1952, J. Am. Stat. Assoc., 47, 583
Makitalo, M., & Foi, A. 2013, ITIP, 22, 91
Morfill, G., Scheingraber, H., Voges, W., & Sonett, C. 1991, in The Sun in Time, ed. C. Sonett, M. Giampapa, & M. E. Matthews (Tucson, AZ: Univ. Arizona Press), 30
Murtagh, F., Starck, J.-L., & Bijaoui, A. 1995, A&AS, 112, 179
Owens, B. 2013, Natur, 495, 300
Rodriguez, G. 2013, https://data.princeton.edu/wws509/notes/c4addendum.pdf
Rousseeuw, P., Ruts, I., & Tukey, J. 2012, The American Statistician, 53, 382
Schaefer, B. 1997, JAAVSO, 26, 47
Scott, D. W. 1979, Biometrika, 66, 605
Stewart, J., & Eggleston, F. 1940, ApJ, 91, 72
Stewart, J., & Panofsky, H. 1938, ApJ, 88, 385
Taylor, J., & Verbyla, A. 2004, Statistical Modelling, 4, 91
Thomas, J., & Weiss, N. 2008, Sunspots and Starspots (1 ed.; Cambridge: Cambridge Univ. Press)
Usoskin, G., Mursula, K., & Kovaltsov, G. A. 2003, SoPh, 218, 295
Vigouroux, A., & Delache, P. 1994, SoPh, 152, 267
Waldmeier, M. 1939, MiZur, 14, 470
Watson, F., & Fletcher, L. 2010, in IAU Symp. 273, Physics of Sun and Star Spots, ed. D. P. Choudhary & K. G. Strassmeier (Cambridge: Cambridge Univ. Press), 51
Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. 2009, Mixed Effects Models and Extensions in Ecology with R (Berlin: Springer)

Nonparametric monitoring of sunspot number observations

Abstract

Solar activity is an important driver of long-term climate trends and must be accounted for in climate models. Unfortunately, direct measurements of this quantity over long periods do not exist. The only observation related to solar activity whose records reach back to the seventeenth century are sunspots. Surprisingly, determining the number of sunspots consistently over time has remained until today a challenging statistical problem. It arises from the need of consolidating data from multiple observing stations around the world in a context of low signal-to-noise ratios, non-stationarity, missing data, non-standard distributions and many kinds of errors. The data from some stations experience therefore severe and various deviations over time. In this paper, we propose the first systematic and thorough statistical approach for monitoring these complex and important series. It consists of three steps essential for successful treatment of the data: smoothing on multiple time-scales, monitoring using block bootstrap calibrated CUSUM charts and classifying of out-of-control situations by support vector techniques.

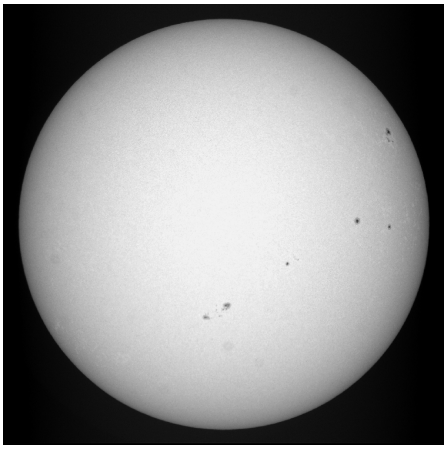
This approach allows us to detect a wide range of anomalies (such as sudden jumps or more progressive drifts), unseen in previous analyses. It helps us to identify the causes of major deviations, which are often observer or equipment related. Their detection and identification will contribute to improve future observations. Their elimination or correction in past data will lead to a more precise reconstruction of the world reference index for solar activity: the International Sunspot Number.

Keywords: Statistical process control; Support vector machine; Correlation; Missing data; Control chart; Block bootstrap

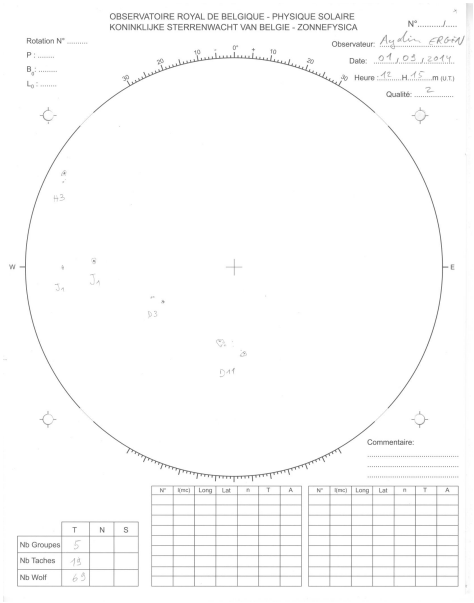
1 Introduction

The main objective of this paper is to provide an efficient monitoring for the sunspot numbers. Until recently, such a monitoring could not be developed with the available statistical methods due to the complexity of the series. Moreover, existing modeling and error quantification of the data were lacking. Advances in statistical process control for panel data combined with a thorough uncertainty model for the sunspot numbers (Mathieu et al., 2019) allow us to construct an appropriate monitoring method for these important data.

1.1 Overview



(a) White light image



(b) Drawing

Figure 1: (a) White (visible) light image of the Sun taken by the Uccle solar equatorial table (USET) on 1st September 2014. (b) The corresponding drawing of the sunspots based on the projected image. Nineteen individual sunspots (N_s) and five groups of sunspots (N_g) are identified in this image. Credit: USET data/image, Royal Observatory of Belgium, Brussels.

Sunspots appear as dark areas in visible light images of the Sun due to their relative lower temperatures. They correspond to regions with a high local magnetic field, which

inhibits the convection of heat coming from the solar interior. They are represented in Figure 1a. Figure 1b also displays the corresponding drawing that is performed by an observer based on the projected image of the Sun. Those drawings are then used to count the number of individual spots, N_s and groups of sunspots, N_g on a daily basis. Both numbers vary with time following a cycle of approximately eleven years that is directly related to the magnetic activity of the Sun (Hathaway, 2010). This cycle is represented for N_s in the lower panel of Figure 2, below. The numbers of spots and groups are the building blocks of a composite, $N_c = N_s + 10N_g$, which is at the basis of the International Sunspot Number (ISN), the reference for modelling long-term solar activity. The multiplication factor was introduced to put N_g on the same scale as N_s since a group contained on average ten spots when the ISN was constructed (Izenman, 1985). This index contains information about N_s and N_g, as it appears that only one of those quantities cannot fully describe the solar activity. Both numbers are thus required. The ISN is nowadays one of the most intensely used time-series in astrophysics (Hathaway, 2010). It enters into models of e.g. the Earth climate (Haigh, 2002; Ermolli et al., 2013) and in space weather predictions (Temmer et al., 2001; Wang and Colaninno, 2014).

Although astronomers started observing sunspots in the beginning of the seventeenth century, it remains surprisingly difficult to arrive at an accurate daily determination of their numbers. Three main difficulties stand out: observability, resolution and interpretation. For Earth-based observatories, the Sun cannot be observed when there are clouds. Instruments with different resolutions may give rise to different counts of sunspots. Distinguishing sunspots and groups of sunspots, which differ essentially by size and shape of their appearance, requires experience and even experts sometimes disagree. Different observers may thus vary in skill and their skills may vary over time. There are also intrinsic sources of variability. Some sunspots are only visible over short periods. Their number may thus change during the day. Moreover, the sunspot activity itself is subject to substantial variabilities, with the most prominent example being the eleven-year solar cycle. The ISN is therefore a weighted consensus among all observatories, also called "stations" participating in the effort.

Due to the multiple sources of variability, we are facing a panel of non-stationary data with many deviating patterns. Those patterns have been partially studied on the short-term in the previous works by Morfill et al. (1991), Vigouroux and Delache (1994) and Dudok de Wit et al. (2016). More recently, Mathieu et al. (2019) developed a comprehensive uncertainty model that reveals drifts that span over several years in the data. Assuring

the quality of the series therefore calls for an automated tool for supervising the observations in quasi real-time. This procedure should be adaptive to the non-normality and the autocorrelation of the data. It should monitor the stations and send alerts when they start deviating to prevent the occurrence of large drifts in future observations. Owing to methodological advances in the sunspot numbers uncertainty modeling and in statistical process control (SPC), it is now possible to develop such a method.

1.2 Univariate dynamic screening system

Many different monitoring procedures have been developed in the SPC literature. Those methods cannot be directly used here however for two main reasons: (1) the mean and variance of the stations change over time (due to e.g. the eleven-year solar cycle) and (2) some stations are deviating in their entire observing period and hence do not have non-deviating or in-control (IC) periods. All stable stations should therefore be used together to judge if a particular station is deviating. To this end, we propose in the following a method based the dynamic screening system developed by Qiu and Xiang (2014). To the best of our knowledge, this method is the only one that can be adapted to the particular characteristics of our data: the non-normality, autocorrelation and non-stationarity of the data as well as the absence of IC periods in all series.

The method of Qiu and Xiang (2014) is based on extensions of the classical CUSUM (Page, 1961) chart. It is composed of two steps. First, the regular patterns (i.e. the mean and the variance) of the data are estimated on a subset of IC series. Second, the data are standardized by these patterns and monitored by a CUSUM chart designed by a block bootstrap method. This procedure constructs, without any parametric assumption, a control scheme that is valid for non-normally distributed and serially correlated data. We can estimate the regular patterns of the data locally in time, since we have at each time point a collection of stable series at our disposal. The method can therefore detect shifts in the mean level of each series, where the means change over time. It can thus accommodate the intrinsic quasi-periodic variations in the sunspot numbers that are related to the solar cycle.

Our method is not in the spirit of a multivariate monitoring. We rather face the situation of a panel with multiple observations of the same phenomena and are interested in monitoring the individual behavior (and errors) of each particular station. Although the ISN is obtained by combining the observations of the panel, the final aim is not to monitor this index directly but to compute it from a subset of non-deviating series, which will be selected by the proposed method.

In the following, we use and extend the work of Qiu and Xiang (2014), to bridge the gaps between the method and the specific requirements of our problem. Those gaps are two-fold. (1) The method of Qiu and Xiang (2014) —as all other methods that we encountered in the literature— cannot be used without knowing a priori which stations are in-control. This information is not available for our data, where *all* stations are expected to contain several kinds of deviations (jumps, oscillating shifts, etc) in their observation period. (2) The method operates with a control chart which sends an alert when a deviation is detected, yet without providing any information about the nature of the shift. Such information is however crucial for us, since it allows to further investigate the causes of the shifts. Although several methods have been developed to automatically predict the size of a shift after an alert (see for instance Cheng et al. (2011) and the references therein), they are not adapted to data which are simultaneously non-normally distributed, serially correlated and contaminated by strong noise.

1.3 Aims

In the following, we propose a nonparametric monitoring that is tailored to the complex features of the sunspot numbers: (a) the missing values, (b) the strong noise, (c) the complex autocorrelation structure and (d) the non-normality. Our method extensively exploits the information contained in the panel to establish a robust IC reference from the network. This allows us to monitor the stations without prior information on their stability. We complete the method by a support vector machine (SVM) procedure that efficiently predicts the size and the shape of a shift once an alert has been raised. Although we could manually build a library with typical shapes and sizes to be compared to the deviations, we select the automatic SVM approach instead.

The control scheme is then applied on past observations to study the deviations of the sunspot numbers. The procedure automatically detects major deviations identified recently by hand in some stations. It also unravels many other deviations, unseen in previous analyzes. In particular, small and persistent shifts that are difficult to identify manually are detected by the method. The precise information about the deviations predicted by the SVM procedures allows us to determine the causes of some prominent deviations. This sets the ground for a future enhancement of the quality of the series. Moreover, the monitoring procedure provides the possibility to be used in real-time to preserve the long-term stability of the stations. It also paves the way to a future redefinition of the International Sunspot Number based on several stations that are stable over time.

This article is structured as follows. In Section 2, we present the main properties of the data and their model. The methods are explained in Section 3. This includes the complete monitoring scheme as well as the SVM procedures to predict the size and shape of the deviations. In Section 4, we apply the proposed method on the sunspot data at different scales, in order to detect both high- and low- frequency shifts and discuss the results on actual stations. In a final section 5, we give some concluding remarks and perspectives. Supplementary materials provide some more details about the monitoring scheme as well as more examples of monitored stations.

2 Data

The data and their specific features are first presented in this section. Then, we introduce the uncertainty model associated to the sunspot numbers. The component of the model that will be monitored in this paper is finally presented alongside with its estimating procedure.

2.1 Presentation of the dataset

The period under study embraces the most recent part of the series and extends from January 1, 1981 till December 31, 2019. It covers thus three complete (eleven-year) solar cycles. The data are composed of the daily observations of a network of 278 Earth-ground observatories disseminated across the world. The records contain the number of spots N_s, groups N_g and composite N_c. They are distributed through the World Data Center Sunspots Index and Long-term Solar Observations (WDC-SILSO)[1]. In the following, we denote by t, $t \in 1, ..., T$, the date-time of the observation and represent the index of the stations by i, $i \in 1, ..., N = 278$.

The data have complex features that are described below. Those should be taken into account in the design of the monitoring procedure.

1. As studied in Mathieu et al. (2019) and previous works, the data are by nature non-normally distributed.

2. The data contain around 70% of missing values over the period studied. Those are mainly caused by the non-overlapping observing periods of the stations. Indeed, some stations started observing only recently while older stations stopped their activity well before 2019. The stations also contain various percentages of missing values

[1]The data are available at the following link: `http://www.sidc.be/silso/`

(ranging from 15% to 75%) over their active observing period, mainly due to weather conditions that prevent the observation of the Sun.

3. The stations are also correlated across the panel and along time since a sunspot may stay from several minutes up to several months on the solar surface.

4. Due to the observing conditions and the solar variability, the series experience a wide range of deviations that vary in shape and size.

2.2 Uncertainty model

Let $Y_i(t)$ represent either the number of spots, groups or composite observed in station i at time t. The observations may be decomposed into a common solar signal, generically denoted by $s(t)$, corrupted by three types of station-dependent errors (Mathieu et al., 2019) in a *multiplicative* framework:

$$Y_i(t) = \begin{cases} (\epsilon_1(i,t) + \epsilon_2(i,t) + h(i,t))s(t) & \text{if } s(t) > 0 \\ \epsilon_3(i,t) & \text{if } s(t) = 0. \end{cases} \quad (1)$$

- $s(t)$ is a latent variable representing the actual number of spots (N_s), groups (N_g) or composite ($N_c = 10N_g + N_s$) of the Sun. This latent variable cannot be directly observed but its mean will be estimated based on the observations of the network and later used as a proxy for $s(t)$.

- ϵ_1 is a short-term error, which is prevailing at scales that are lower than 27 days (i.e. one solar rotation). It typically represents counting errors. We assume that $\mathbb{E}(\epsilon_1(i,t)) = 0$ where $\mathbb{E}$ denotes the expectation sign.

- ϵ_2 denotes a long-term error, which corresponds to scales between 27 days and eleven year (one solar cycle). We are interested in estimating and monitoring its mean, denoted by $\mu_2(i,t)$, which represents the bias of the stations.

- h is defined at time-scales equal to or longer than eleven years. It corresponds to the background level of the stations (accounting e.g. for differences of instruments or counting methodologies of the stations). For identification purpose, we assume that $\mathbb{E}(\epsilon_2(i,t) + h(i,t)) = 1$.

- ϵ_3 is an additive error capturing effects like short-duration sunspots during solar minima, i.e. periods of minimal activity in the eleven-year solar cycle.

The errors ϵ_1, ϵ_2 and h vary on different time-scales and are multiplicative quantities since an observer typically makes larger errors when $s(t)$ is higher (Chang and Oh, 2012). The random variables ϵ_1, ϵ_2, ϵ_3 and h are assumed to be continuous and ϵ_1, ϵ_2, ϵ_3, h and $s(t)$ to be jointly independent. Note that ϵ_1, ϵ_2 and ϵ_3 would be equal to zero and h be equal to one for a station that would be – in absence of any measurement errors – perfectly aligned with the solar signal.

2.3 Long-term bias

h is not the target of our monitoring procedure since it models the intrinsic level of the stations (and not an error). Similarly, we are not interested in monitoring the short-term error ϵ_1, which does not affect the long-term stability of the data, nor the error at minima ϵ_3, which only corresponds to a small part of the solar cycle. Our monitoring aims at the component ϵ_2 and more specifically its mean μ_2. To this end, we isolate the long-term error from the other components of the model in the step-wise approach (described below) that is similar to those of Mathieu et al. (2019).

We first divide the observations by scaling factors to roughly compensate for different observing conditions: $Z_i(t) = \frac{Y_i(t)}{\kappa_i(t)}$. These piece-wise constant scaling factors $\kappa_i(t)$ are computed as the slope of the ordinary least-squares regression between the observations of the stations and the median of the observations ($\underset{1\leq i\leq N}{\text{med}}\, Y_i(t)$) on periods of 8 months for N_s, 14 months for N_g and 10 months for N_c. These values are selected by a statistical-driven study based on the Kruskal-Wallis test (Kruskal and Wallis, 1952) that is completely described in the section 6.3 of Mathieu et al. (2019).

Afterward, we compute M_t, a robust proxy for $s(t)$ based on the median of the rescaled observations:

$$M_t = \underset{1\leq i\leq N}{\text{med}}\, Z_i(t). \tag{2}$$

Motivated from (1), the observations Y are then divided by M_t to remove the main influence of the solar signal. They are also smoothed by a moving-average (MA) filter, represented by a $\star$ in the following equation. This smoothing process untangles eh, $eh = \epsilon_2 + h$, from the short-term error ϵ_1:

$$\widehat{eh}(i,t) = \left(\frac{Y_i(t)}{M_t}\right)^{\star} \quad \text{when } M_t > 0, \tag{3}$$

where $\widehat{eh}$ denotes the estimator of the mean of eh, which is used as a proxy for eh. To analyze the various deviations of the data, different MA-filter window lengths may be used

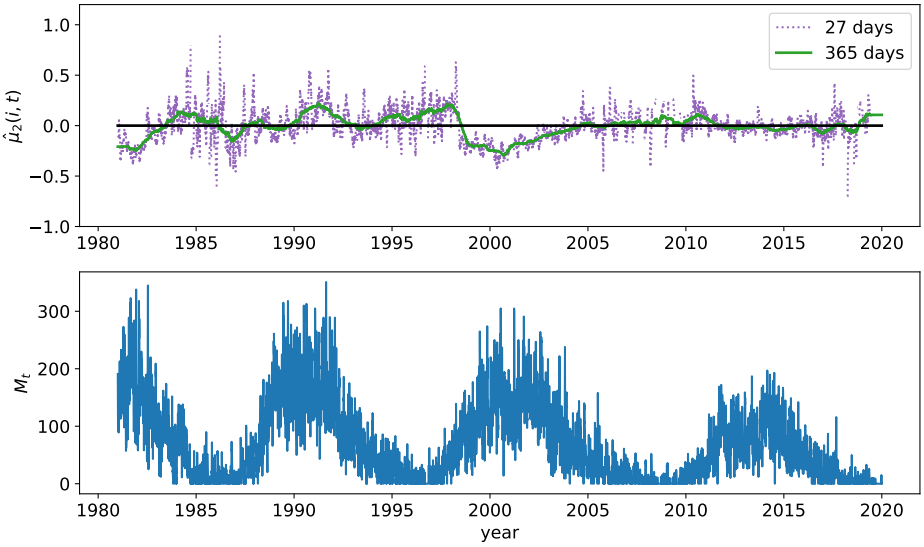


Figure 2: Long-term bias, $\hat{\mu}_2(i,t)$ for N_c, in the Kanzelhöhe Observatory (Austria) over its observing period. The $\hat{\mu}_2$s are smoothed on 27 days (dotted line) to allow the detection of high-frequency shifts and smoothed on 365 days (plain line) to emphasize the low-frequency deviations. M_t is also represented in the lower plot as an estimation of the actual N_c. This figure clearly shows the eleven-year solar cycle (Hathaway, 2010) that is intrinsic to the signal.

in (3). The low-frequency shifts such as persisting drifts are first studied at a yearly scale (i.e. with a window length of 365 days). Then, a window of length equal to 27 days will also be used to examine the high-frequency deviations such as sudden jumps. This value of 27 days corresponds to a physical scale of the data: one solar rotation. It appears to be sufficiently high to overcome the effects of the short-term regime, as demonstrated in Mathieu et al. (2019).

Finally, the levels of the stations are separated from the mean of the long-term, $\mu_2(i,t)$, by applying once again a MA smoothing process denoted by $\star\star$:

$$\hat{\mu}_2(i,t) = \widehat{eh}(i,t) - \widehat{eh}^{\star\star}(i,t), \tag{4}$$

where the MA-filter window length should be larger than those of (3). It is selected here at eleven years, a physical value that is larger than the time-scales of the long-term error ϵ_2 considered here. It also seems appropriate since the location of the observatories or their telescope are unlikely to change much over time. Since we removed the solar signal ($s(t)$) from the long-term error, we assume that the ϵ_2s are independent across the stations. The

main factors that impact those errors (e.g. the location of the station, the instrument or the counting methodology) are indeed intrinsic to each station.
All previously-mentioned quantities are represented in Appendix A. They illustrate the different stages of the computation of the long-term bias, $\hat{\mu}_2(i,t)$. Those $\hat{\mu}_2$s smoothed on 27 and 365 days are also represented in Figure 2 for a particular station.

3 Ingredients of the method

In this section, the complete monitoring procedure depicted in Figure 3 is explained. It is intentionally presented in a generic framework to allow the application of the method on the number of spots N_s, groups N_g as well as composites N_c. There are three phases: (I) estimation of the in-control (IC) parameters of the data, (II) construction and use of the monitoring procedure and (III) identification of out-of-control patterns.
Phase I contains two steps. At first a subset of stations is selected from the panel which follows closely the median signal M_t mentioned in Section 2. This pool of stations is then used as a proxy for IC series in the nomenclature of Qiu and Xiang (2014). They are used to determine the IC patterns (mean and variance) of the data and to provide the basis of the block-bootstrap procedure in Phase II. After standardizing all series by the IC patterns, the CUSUM control chart is calibrated in phase II by a block bootstrap procedure from the pool of IC series. The scheme is then applied to the data for the monitoring. In Phase III, support vector machine (SVM) procedures predict the shifts size and shape on sub-series detected as out-of-control by the CUSUM, for easier problem diagnostic.

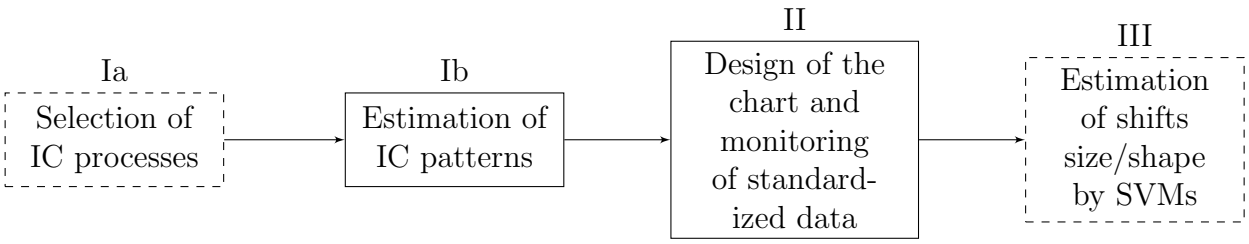


Figure 3: Pipeline of the procedure. The dashed blocks represent the new ingredients that we add to monitor the sunspot numbers.

3.1 Phase I: Estimation of the IC longitudinal patterns

In this phase I, we automatically construct a subset of IC stations from the panel and estimate the IC patterns of the data.

3.1.1 Phase Ia: Selection of the IC processes

In a first stage, we need a subset (pool) of stations whose observations follow closely the median signal M_t. In order to find them, we calculate a stability criterion on each station. This criterion is based on a robust version of the mean squared error (MSE) of $\hat{\mu}_2$:

$$STB(i) = \underset{1\leq t\leq T}{\text{med}}[\hat{\mu}_2(i,t)]^2 + \underset{1\leq t\leq T}{\text{iqr}}\,\hat{\mu}_2(i,t), \tag{5}$$

where $\underset{1\leq t\leq T}{\text{iqr}}\,\hat{\mu}_2(i,t)$ and $\underset{1\leq t\leq T}{\text{med}}\,\hat{\mu}_2(i,t)$ denote respectively the interquartile range (IQR) and the median of the $\hat{\mu}_2(i,t)$ over the time for a given station i. Using these values, we can then cluster the stations and choose the cluster with the lowest values to form what we call the pool of IC stations. For this purpose, we use the k-means clustering (Lloyd, 1957; MacQueen, 1967) with two clusters. As these two clusters can be highly unbalanced, the clustering in two groups is performed recursively until the smallest cluster contains at least 25% of the stations. Since we cluster 1D data (one value per station), other methods based on the ordering or the distribution of the observations could be used. We choose however k-means as we prefer to use an automatic and general-purpose procedure instead.
The pool contains deviations that will be called *disparities* in the following, to be distinguished from the deviations that are supposed to be actually detected by the method. The *disparities* are expected to be typically smaller and less frequent than the deviations occurring in the stations not comprised in the pool. They should be included in the design of the chart otherwise the scheme would be over-sensitive.

The pool suffers in addition from deviations that are of similar magnitude as those of the out-of-control (OC) processes. To cope with this and preserve the detection power of our scheme, we also apply a Shewhart chart (Shewhart, 1931) with *adaptive* confidence intervals on the data. We remove the IC observations that do not fall into one standard deviation around the cross-sectional mean ($\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_2(i,t)$). This step removes around 8.8% of the IC observations. Note that we also tested the method with two standard deviations instead of one and the monitoring results were similar (in this case, we only remove around 0.9% of the IC data). We emphasize that this adaptive Shewhart chart would not be a substitute for our control scheme: it only removes the largest deviations at each time without taking

into account the history of the observations. Therefore, contrarily to our method, it cannot detect the small and persistent shifts.

3.1.2 Phase Ib: Estimation of the mean and the variance of the IC series

We denote by $\mu_0(t)$ and $\sigma_0^2(t)$ respectively the mean and the variance of the $\hat{\mu}_2$ of the pool. Those are estimated by the empirical mean and variance using nearest neighbours (K-NN) regression method:

$$\begin{aligned}\hat{\mu}_0(t) =& \frac{1}{\Delta(t)} \sum_{t'=t-\Delta(t)/2}^{t+\Delta(t)/2} \frac{1}{N_{IC}} \sum_{i_{ic}=1}^{N_{IC}} \hat{\mu}_2(i_{ic}, t') \quad s.t. \quad K = \Delta(t) N_{IC} \\ \hat{\sigma}_0^2(t) =& \frac{1}{\Delta(t)} \sum_{t'=t-\Delta(t)/2}^{t+\Delta(t)/2} \frac{1}{N_{IC}} \sum_{i_{ic}=1}^{N_{IC}} \left(\hat{\mu}_2(i_{ic}, t') - \hat{\mu}_0(t)\right)^2 \quad s.t. \quad K = \Delta(t) N_{IC},\end{aligned} \tag{6}$$

where i_{ic} denotes the index of a station of the pool. With K-NN regression, the temporal window $\Delta(t)$ can be adjusted to compensate the missing values of the stations, such that $\hat{\mu}_0(t)$ and $\hat{\sigma}_0^2(t)$ are always computed on the same number (K) of observations. For appropriate use in the CUSUM chart statistics (to be defined in (8) below), the data must be standardized by the IC mean and variance (as in (7) below). Hence the number of nearest neighbors K is selected to obtain the "best" standardization of the complete panel, in the sense that their empirical mean becomes close to zero and their empirical variance close to one. Then, $\Delta(t)$ is chosen in time direction such that $K = \Delta(t) N_{IC}$.

3.2 Phase II: Monitoring

We now turn our attention to monitoring the entire panel. As a reminder, we are analyzing long-term biases denoted by $\hat{\mu}_2$. Using the IC mean and standard deviation $\hat{\mu}_0(t)$ and $\hat{\sigma}_0(t)$, we standardize the (IC and OC) stations to be able to use common monitoring criteria:

$$\hat{\epsilon}_{\hat{\mu}_2}(i,t) = \frac{\hat{\mu}_2(i,t) - \hat{\mu}_0(t)}{\hat{\sigma}_0(t)}. \tag{7}$$

Let us now focus on one station (drop the index i). We would like to detect indications of patterns which may relate to problems at the station. This includes persistent or gradual deviations (shifts or trends) and oscillating patterns as they may occur when the observatory is used by a rotating pool of observers each of whom has their own particular way of

working. A method for accumulating small and gradual deviations is to aggregate them over time. A well known method for doing so in the context of statistical process control is the cumulative sum (CUSUM) chart (Page, 1961). The two-sided CUSUM chart applied on the residuals writes as:

$$\begin{aligned} C_j^+ &= max(0, C_{j-1}^+ + \hat{\epsilon}_{\hat{\mu}_2}(t) - k) \\ C_j^- &= min(0, C_{j-1}^- + \hat{\epsilon}_{\hat{\mu}_2}(t) + k), \end{aligned} \tag{8}$$

where $j \geq 1$, $C_0^+ = C_0^- = 0$ and $k > 0$ is the allowance parameter (Qiu, 2013).
This chart gives an alert if $C_j^+ > L^+$ or $C_j^- < L^-$, where L^- and L^+ are the control limits of the chart. Since the distribution of the residuals is almost symmetric, we use $L = L^+ = -L^-$.

High deviations may affect the series. Those lead to high values of the CUSUM statistics which may stay in alert for longer periods than the actual durations of the shifts. Therefore, in case of too high (resp. too low) values, we set the chart to a maximal value $2L$ (resp. $-2L$). Hence $|C_j^+|, |C_j^-| \leq 2L$.

3.2.1 Design of the chart

As it is clear from the nature of the data, the series to be monitored have a considerable degree of autocorrelation even when they are in control. We therefore need a method for determining the control limit of the chart that takes autocorrelation into account. The block bootstrap (BB) method does this. It is based on constructing a bootstrap reference distribution by resampling blocks of data and thereby preserving the autocorrelation of the series.
The control limit (L) is adjusted here by a searching algorithm that is explained in details in Appendix B.1 (in the supplementary material). It works as follows. A target shift size, δ_{tgt}, is first estimated on the OC series as explained in Appendix B.2. The allowance parameter is specified to $k = \delta_{tgt}/2$. For an initial value of the control limit, the actual IC average run length, ARL_0, is then evaluated on IC data that are sampled from the pool by the BB procedure. If the actual ARL_0 is inferior (resp. superior) to the pre-specified ARL_0, the control limit of the chart is then increased (resp. decreased). This algorithm is iterated until the actual ARL_0 reaches the pre-specified ARL_0 at the desired accuracy.

As theoretically demonstrated in Lahiri (1999), BB methods using non-overlapping blocks and random block lengths are more variable than those based on overlapping blocks

and constant lengths. Therefore, we select the popular moving BB (MBB) (Kunsch, 1989; Liu and Singh, 1992) to obtain the best performances.
Since the BB preserves the serial correlation of the data inside the blocks, the length of the blocks should be selected appropriately. Large blocks usually model the autocorrelation of the data properly but at the same time do not represent well the variance and the mean of the series. And conversely. Using the method described in Appendix B.3, the block length is selected here as the first value such that the MSE of the empirical autocorrelation of the $\hat{\mu}_2$ becomes stable. This value intuitively corresponds to the smallest length which is able to represent the main part of the autocorrelation of the series.

The data also contain missing values. Among them, the large gaps prevail since the smoothing process in (3) removes the shortest gaps of the series. As the observing conditions could be different after a large amount of missing values (different weather conditions or instruments), we restart the scheme after each gap ($C_j^+ = C_j^- = 0$). Blocks composed only of missing values are not used to design the chart. This may happen when some stations contain few observations on the period studied here, either because they are ancient and stopped observing at the beginning of the period or because they have started observing only recently.

3.3 Phase III: Estimation of the sizes and shapes of the shifts using SVMs

The CUSUM gives an alert when a deviation is detected in the data but does not provide information about the characteristics (shape and size) of the shift. Such information is however valuable to assign possible causes to the shift or to adapt the type of alerts that is sent back to the observers. To that end, Cheng et al. (2011) appended a support vector regression (SVR) to the CUSUM. This method is designed to predict, after each alert, the magnitude of shifts in independent and identically normally distributed data that only experience jumps. In the following, we extend Cheng et al. (2011) and design a method that is effective to detect the sizes *and* the shapes of the deviations in the sunspot number data. This is achieved by a SVM classifier (SVC) (Burges, 1998) in addition to a SVR on top of the chart.

3.3.1 Input vector

When an alert is triggered, the m most recent observations of the stations are fed into the SVR and SVC which then predict the size and shape of the deviation at the origin of the alert, as explained in the next subsection. In particular, the SVR prediction model writes as:

$$\hat{\delta} = f(V_{t'}) = f(\hat{\epsilon}_{\hat{\mu}_2}(t' - m + 1), \hat{\epsilon}_{\hat{\mu}_2}(t' - m + 2), ..., \hat{\epsilon}_{\hat{\mu}_2}(t')), \tag{9}$$

where t' denotes the time of the alert and $V_{t'}$ represents the input vector, i.e. a sequence containing the last m observations of the series.

The length m of the input vector should thus be sufficiently large to contain the starting point of most of the deviations while maintaining the computing efficiency of the method. Large shifts are often quickly detected by the chart (short OC run length) while the smallest shifts are identified only after a certain amount of time (long OC run length). Therefore, the latter require larger input vectors than the former. m is selected here as an upper quantile of the OC run length distribution for a shift size equal to δ_{tgt}, as explained in Appendix B.4. Hence, m should be sufficiently large to allow the identification of shift sizes that are superior or equal to δ_{tgt}.

As the SVM procedures do not support missing values, we have to impute them. Missing observations occurring at the beginning of $V_{t'}$ are simply replaced by the first valid observation encountered, while the "intermediate" gaps are filled by a linear interpolation. However, when there are too many of them, the analysis makes no sense. We decide to only analyze input vectors which have at least 20% of non-missing values.

3.3.2 Support vector regression

The support vector machine (SVM) (Vapnik, 1998) is a supervised machine-learning procedure, here used as a robust classifier and regressor to predict the shape and size of the deviations. The method has a strong theoretical basis that takes root in the optimization theory. It is able to perform efficiently non-linear classification or regression using a kernel trick that implicitly maps the data into a high dimension where the non-linear problem becomes linear. We only introduce the SVR in the following, since the SVC can be expressed with a similar framework. Smola and Schölkopf (2004) may also be consulted for more detailed explanations.

We denote by $\{ \boldsymbol{x}^j, \delta^j | j = 1, 2, ..., M \}$ the M training pairs. $x \in \mathcal{R}^m$ represents a training input, i.e. a series of m observations that contains a deviation and $\delta \in \mathcal{R}$ is its corresponding output, the size of the deviation. The SVR aims at estimating the continuous

regression function relating the deviating observations to the size of the shift, $f(\boldsymbol{x})$, based on the training pairs. This function writes as:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b, \tag{10}$$

where ϕ is the non-linear function mapping the input data into the high dimensional feature space, where the regression may be expressed into a simpler linear problem. The coefficients $\boldsymbol{w}$ and b are then estimated during the training by solving the following optimization problem:

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 + \lambda \frac{1}{M} \sum_{j=1}^{M} L_\epsilon(\delta^j, f(\boldsymbol{x}^j)), \tag{11}$$

where

$$L_\epsilon(\delta^j, f(\boldsymbol{x}^j)) = \begin{cases} |\delta^j - f(\boldsymbol{x}^j)| - \epsilon & |\delta^j - f(\boldsymbol{x}^j)| \geq \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

In the objective function of (11), the parameter λ represents a trade-off between misclassification and regularization whereas ϵ in the loss function of (12) is equivalent to an approximation accuracy, i.e. errors below ϵ are neglected. This optimization problem may be rewritten with Lagrange multipliers as a dual problem and easily solved in the input space thanks to the introduction of a kernel function $K(\boldsymbol{x^j}, \boldsymbol{x^k}) = \phi(\boldsymbol{x^j})\phi(\boldsymbol{x^k})$. This kernel function is an important hyper-parameter of the method that should be carefully selected. After testing different kernels, we choose the radial basis function, to obtain the best prediction results.

3.3.3 Creation of the training and testing sets

The training and testing sets are constructed by simulations since only a limited amount of (unlabelled) observations are available. As the SVM procedures predict the characteristics of the shift after an alert has been raised by the CUSUM, we generate sets of series that will be first monitored by the control scheme before reaching the SVMs. When an alert will be triggered by the CUSUM, the m last values of the series will be assembled and used as an input vector for the SVMs. Hence, we create series that are initially longer than m. Those are randomly sampled from the IC data by BB. To ensure the efficiency and the generalization of the predictions, we then add various deviations with different sizes and shapes on top of the series.

Shift sizes The magnitudes of the shifts, δ, are first randomly sampled from two half-normal

distributions (Evans et al., 2000) supported by $[-\infty, ..., -\delta_{tgt}]$ and $[\delta_{tgt}, ..., \infty]$ respectively. We select the scale parameter of the half-normals equal to 3.5, a value that is sufficiently high to reproduce the highest values/deviations observed in the data.
Shift shapes: For each δ, a series x_{ic} of length T' is generated from the IC pool by the BB. Here, we choose $T' = 500$. Three types of general deviations are then artificially constructed on top of the series:

1. jumps: $x(t) = x_{ic}(t) + \delta$;
2. drifts with varying power-law functions: $x(t) = x_{ic}(t) + \frac{\delta}{T'}(t)^a$, where a is randomly selected in the range $[1.5, 2]$;
3. oscillating shifts with different frequencies: $x(t) = x_{ic}(t) \sin(\eta\pi t)\delta$, where η is randomly selected in the range $[\frac{\pi}{m}, \frac{3\pi}{m}]$.

These deviations are selected to visually correspond to the deviations observed in the data.

Time of the shift: In the data, the shifts may happen not immediately but after an initial IC period. Therefore, we also start the monitoring after a random delay in the range $[m, 3m/2]$, to train the methods at identifying shifts appearing anywhere within the input vector. Note that the SVMs as well as the control chart should be started after m observations are gathered.

The SVR is trained on these constructed sets to predict the size of the deviations in the continuous range $[-\infty, ..., -\delta_{tgt}, \delta_{tgt}, ..., \infty]$. In practice, we observe that the SVR generalizes well and can make predictions on $\mathbb{R}$ even if it was only trained in a smaller range of interest. The SVC also learns on the same sets to identify three different shapes: jumps, drifts and oscillating shifts. If a wide range of deviations are simulated, only three classes are therefore involved in the classification problem.

4 Monitoring the composite sunspot index Nc

In this section, we use the previously-described scheme to solve the monitoring problem of the sunspot numbers. We do so for the composite $N_c = N_s + 10N_g$ (the same approach also works for the two components, N_s and N_g and is presented in the supplemental material). We first study the low-frequency deviations on data that have been smoothed on a year to extract and analyze low-frequency patterns such as trends and persistent shifts. Then,

we examine on data that have been smoothed on 27 days (one solar rotation) the higher frequency patterns such as sudden jumps. The section ends with an example of a monitoring at multiple frequencies applied to a particular observatory. Simulations comparing our control scheme to a purely univariate method are also presented in the supplementary material.

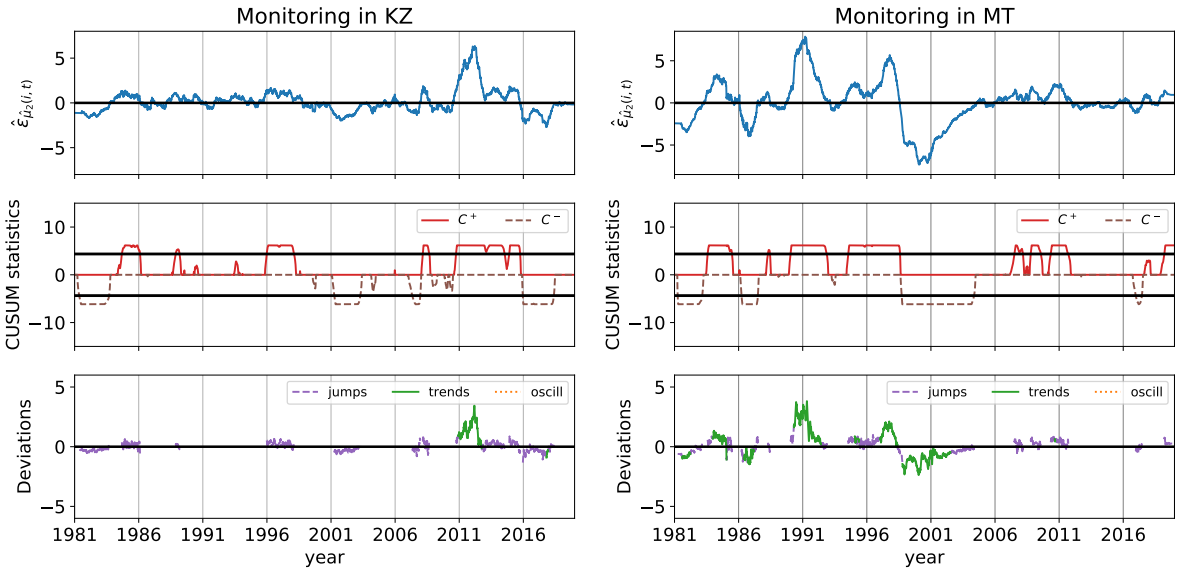
4.1 Lower frequency monitoring

In the first step of low-frequency monitoring of N_c, we smooth the long-term bias ($\hat{\mu}_2$) with a window length of one year as described in Section 2.3. As explained in Section 3.1.1, the network of stations is first reduced to a pool of 119 in-control (IC) stations. In the next step, we extract the IC mean and standard deviation using the K-NN regression described in Section 3.1.2. The selection mechanism finds $K = 4600$ for this step. The resulting mean and standard deviation are then used to standardize all series.

In the second stage, we use the block bootstrap method described in Section 3.2.1 to calibrate the CUSUM chart at an average run length of 200. This requires choosing the block length first. In our situation, a choice of two solar rotations (54) appears appropriate. It is longer than the lifetime of most sunspots but not too long for practical use. The calibration then leads to a control limit of $L = 19$ and a target shift size of $\delta_{tgt} = 1.5$.

Finally, the support vector method for extracting and classifying out-of-control patterns is deployed. It is composed of a SVR to predict the size of the shifts and a SVC to classify the shape of the encountered deviations. We obtain them by creating a set of artificial series of 500 values generated from the IC pool by the BB. These give us series as we would observe in reality including correlations. We then artificially add jumps, trends, and oscillating shifts to them as described in Section 3.3.3. These series are then fed to the CUSUM chart which identifies the out-of-control observations. When an alert is triggered by the chart, an input vector containing the m last observations of the series is assembled. This input vector is then analyzed by the SVR and SVC for predicting the characteristics of the shift. In our case, we harvested 63000 such series from the IC pool and we enriched them with artificial patterns. We then calibrated SVR and SVC models by splitting this set into a training set (80%) and a testing set (20%).

The length of the input vector is specified here at $m = 80$, as explained in Section 3.3.1. This value of 80 corresponds to the 90-th quantile of the OC run length distribution, for a shift of size δ_{tgt}. The other parameters of the support vector machines are automatically selected from a searching interval to obtain the best prediction results. Those are evaluated



(a) (b)

Figure 4: (a) Upper panel: the residuals $\hat{\epsilon}_{\hat{\mu}_2(i,t)}$ for N_c smoothed on one year from the KZ over the period studied (1981-2019). In addition to their disparities, the residuals also contain the actual deviations of the station, which have been removed for the design of the chart as explained in Section 3.1.1. Middle panel: the (two-sided) CUSUM chart statistics applied on the residuals in *square-root scale*. The control limits of the chart are represented by the two horizontal thick lines. Lower panel: the characteristics of the deviations predicted by the SVR and SVC after each alert. (b) Similar figure for MT over the same period.

using the mean absolute percentage error (MAPE) for the regression and the accuracy for the classification problem, see Appendix C. With this method, the regularization parameter λ of the classifier and regressor is set to 13 and the accuracy error ϵ of the SVR is fixed at 0.001. The performances of the SVMS are presented in Appendix C.1. Overall, they are sufficient to achieve our goals: identify the origins of the deviations.

Figure 4 shows results for two stations labeled KZ[2] and MT[3]. The observatory of KZ is rather stable and belongs to the IC pool. It relies on a stable team of well trained observers.

[2]The Kanzelhöhe Observatory in Austria.

[3]The National Observatory of Japan, in Mitaka (Tokyo)

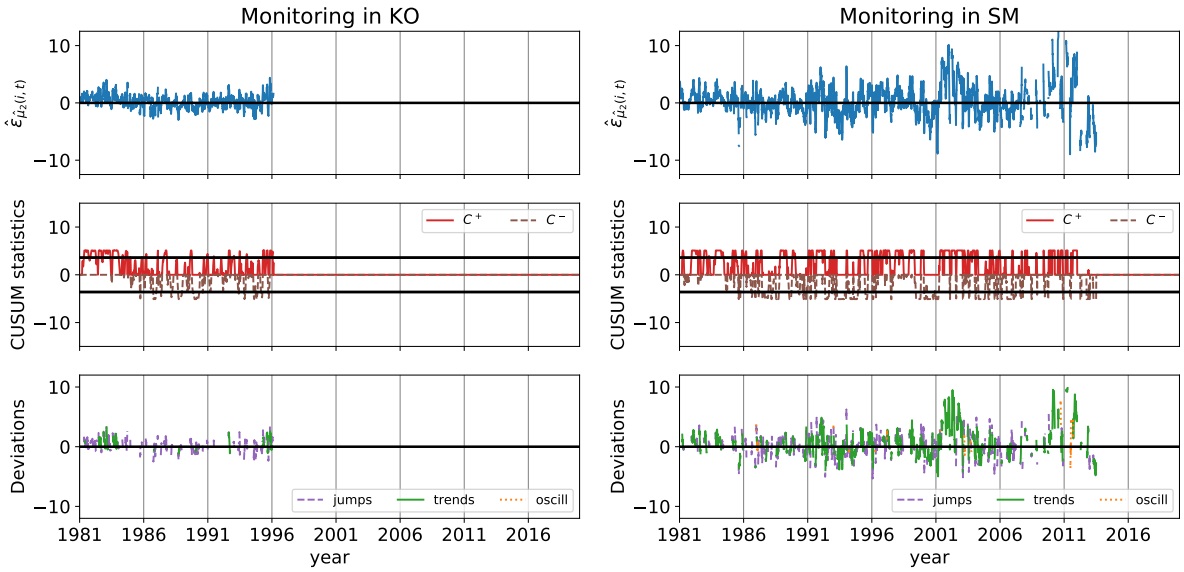
The observatory MT is less stable and shows a severe downward trend around 1998. The cause of this downward trend could be traced to the replacement of visual counts from the direct optical solar image by automatic computerized counts based on digital images from a CCD camera. Given the image sensor technology then available, the spatial resolution of the images was limited, and many small spots that were fully detected in earlier visual observations were not detected anymore by the new equipment.

4.2 Higher frequency monitoring

The method described above can also be applied to the biases ($\hat{\mu}_2$) smoothed on a shorter time window such as the duration of a solar cycle (27 days). Here the selection of the IC pool yields 100 stations and the number of nearest neighbors comes out to $K = 2400$. This number is smaller than before since we are working at a higher frequency. The block bootstrap and SVMs with the same settings as above can be used to calibrate the CUSUM chart. For $\delta_{tgt} = 1.4$, the control limit of the chart is selected at $L = 13$ to obtain an average run length of 200. The length of the input vector is fixed here at $m = 70$, which corresponds to the 90-th quantile of the OC run length distribution.

Figure 5 shows the methodology applied to KO[4] and SM[5]. KO was a Japanese observatory run by a single dedicated observer whose records (which stopped during 1996) were very stable. On the contrary, SM is a severely OC station that experiences large known deviations (Mathieu et al., 2019, Figure 12). The large variations observed in SM are likely caused by the rotation of several observers involved in the counting process. In some countries, the public observatories have also an educational function. Their team of regular observers are usually small and are often completed by student or amateur astronomers that are frequently replaced, which causes large variations. Unfortunately, we could not identify more precisely the origin of the deviations since their observations stopped several years ago and we have not succeeded in contacting them yet. The lack of information is a common problem we face when investigating past deviations in stations that are now inactive and is therefore worth mentioning.
As we see in the figure, the biases vary a lot at 27 days, a scale which is close to the short-term regime. The actual monitoring should be based on a larger scale otherwise some stations such as SM would receive almost constant alerts. If a particular deviation is detected at higher scale (such as one year), it might be interesting however to analyze it

[4]Name of the observer known to the authors, kept for privacy.
[5]The observatory of San Miguel in Argentina.



(a) (b)

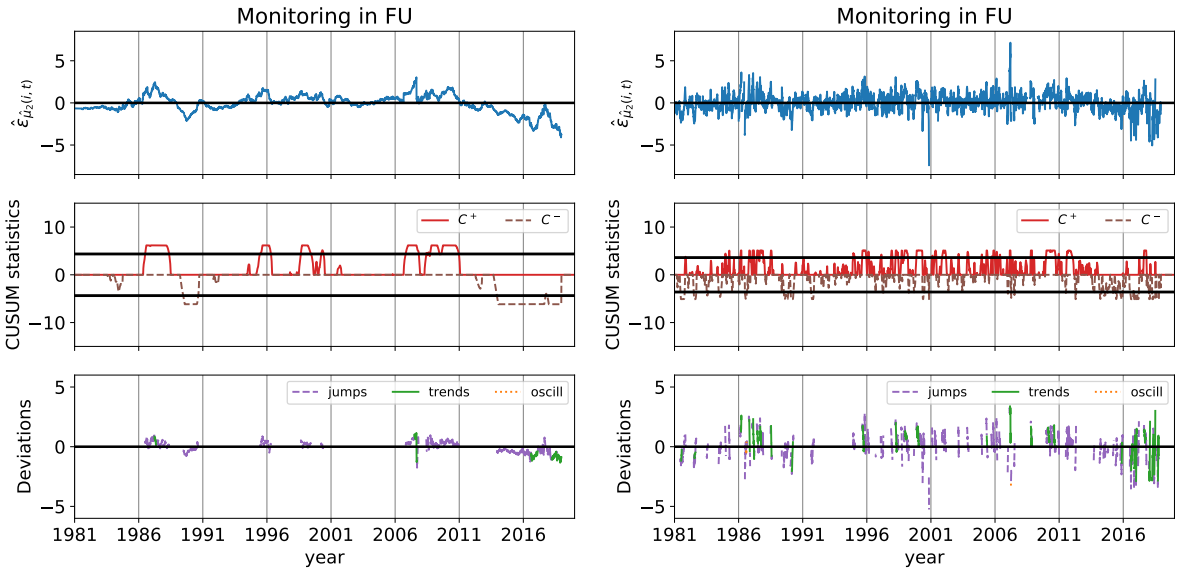
Figure 5: (a) Upper panel: the residuals $\hat{\epsilon}_{\hat{\mu}_2(i,t)}$ for N_c smoothed on 27 days for KO over the period studied (1981-2019). In addition to their disparities, the residuals also contain the actual deviations of the station, which have been removed for the design of the chart as explained in Section 3.1.1. Middle panel: the (two-sided) CUSUM chart statistics applied on the residuals in *square-root scale*. The control limits of the chart are represented by the two horizontal thick lines. Lower panel: the characteristics of the deviations predicted by the SVR and SVC after each alert. (b) Similar figure for SM over the same period.

at 27 days, to better identify its origin.

4.3 Monitoring at multiple frequencies

Figures 4 and 5 display instances of a stable IC station included in the pool and a typical out-of-control observatory for the high- and low- frequency monitoring respectively. To better grasp the motivations of a monitoring at multiple frequencies, the method is applied to the data smoothed on 27 days and one year of FU[6] in Figure 6. The FU station is composed of a single dedicated observer in Japan, who has observed without interruption

[6]Name of the observer known to the authors, kept for privacy.



(a) (b)

Figure 6: (a) Upper panel: the residuals $\hat{\epsilon}_{\hat{\mu}_2(i,t)}$ for N_c smoothed on 365 days from FU over the period studied (1981-2019). In addition to their disparities, the residuals also contain the actual deviations of the station, which have been removed for the design of the chart as explained in Section 3.1.1. Middle panel: the (two-sided) CUSUM chart statistics applied on the residuals in *square-root scale*. The control limits of the chart are represented by the two horizontal thick lines. Lower panel: the characteristics of the deviations predicted by the SVR and SVC after each alert. (b) Similar figure for the values of $\hat{\epsilon}_{\hat{\mu}_2(i,t)}$ smoothed on 27 days in FU.

since 1968 until today, producing one of the longest individual series. His observations are included in the IC pool but yet suffer from recent deviations. In particular, the upward deviation (which looks like a spike) reported in 2007 in FU (Clette, 2013) as well as the downward drift occurring after 2014 are well identified in Figure 6a. Figure 7 shows a zoom of Figure 6a on the time period from 2007 to 2008. After a progressive upward shift, the station experiences a rapid downward trend over five days. This trend, which looks like a jump in the whole period view, it thus correctly classified by the SVC. Even by taking a closer look on the figure however, it remains difficult to precisely identify the origin of the shifts on data that are smoothed on a year. By looking at a smaller scale of 27 days in Figure 6b, we can better characterize the shift in 2007 as a short event and pinpoint its

location. After investigations, this deviation appears to be related to a small over-count that appeared in early 2007 (three groups were reported in FU while most of the network only observed two groups) while the drift might be associated to the health condition of the observer. Note that the long-term biases are not defined (i.e. set to missing values) when the median of the network is equal to zero, see (3). This regime corresponds to those of the variability at minima, represented by ϵ_3. Due to the smoothing procedure of (3), the deviations that appeared close to solar minima, such as the jump in FU, are thus particularly visible.

As shown in the figures, the monitoring and the SVM procedures can cope with a large variety of shifts ranging from small and persistent deviations to large oscillating shifts. The procedures automatically detect major deviations recently discovered by hand as mentioned above. More identified prominent deviations as well as results for other stations are shown in Appendix E. In addition, the chart also unravels many other shifts, typically smaller, that are otherwise difficult to identify.

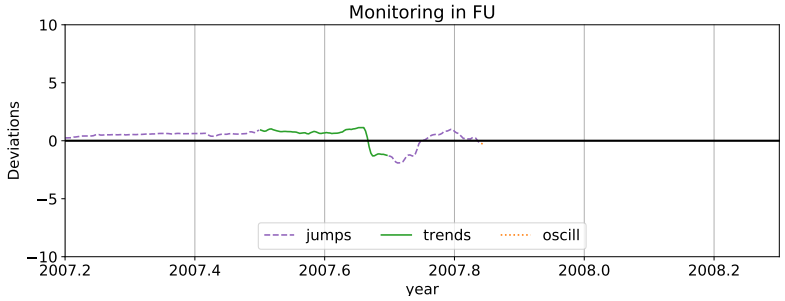


Figure 7: Sizes and shapes of the deviations (taken from Figure 6a) predicted by SVMs in FU over 2007-2008.

Note that the figures represent past observations, which have not been monitored by any control scheme. Consequently, the stations may stay in alert for long consecutive periods. If this method is applied on future observations, any major deviation will be promptly corrected. Therefore we expect better results for data that have already been monitored.

5 Conclusion and perspectives

We presented a nonparametric control scheme to monitor challenging and important datasets related to the observations of sunspots across a wide network of stations. The approach

allows us to deal with the missing values, autocorrelations, and non-normality of the data, to detect and classify station-related anomalies on different frequencies. The procedure is based on a particular choice of methods for smoothing, robust anomaly detection, and anomaly classification. Other methods exist for these steps; yet we believe that our choices are particularly suitable for the problem at hand.

The features of our approach are smoothing on multiple frequencies, CUSUM charting, SVM classification and detailed graphical displays. They also include an automatic pre-selection of an in-control pool and the powerful calibration of the chart using block bootstrap procedures. The associated advantages are robustness, flexibility, automation, and guided interpretation of results. The method allows us to detect and identify the causes of major deviations that occurred in the series. We have seen that monitoring on at least two time-scales is essential to capture these anomalies. Some patterns first attract interest on a long-term scale but it is at the short-term scale that their potential root-causes can be suggested. The method also identifies a wide range of deviations unseen in previous analyses. Most of them have not been related to specific causes yet but will soon be investigated.

This automated method allows us and the researchers at the Royal Observatory who are in charge of producing the International Sunspot Number to have a harmonized view across the network of stations. It provides a way to give specific and targeted advice to the observers. As demonstrated in this paper, the method also delivers easy to interpret graphical displays which facilitate root cause analysis of deviations. The complete re-examination of past data of the whole panel has just started. When they will be finished, these analyses will allow us to arrive at a cleaner data stream and to release an improved version of the International Sunspot Number. Additionally, the implementation of the method in the continuous surveillance of future observations will lead to a faster detection and identification of inconsistencies, their elimination by better observer training or equipment maintenance, and finally to a more precise determination of the sunspot numbers in the future.

The control scheme can also be applied in general to monitoring other panels of time-series. It has been used in Mathieu (2021) to monitor the photovoltaic energy production in Belgium, as one example.

SUPPLEMENTARY MATERIAL

Python package (codes) The subset of data and the codes that we used in this paper are available at `https://github.com/sophiano/SunSpot`.

Figures related to the uncertainty model The Appendix A of the supplementary material contains figures displaying the different quantities appearing in the computation process of the long-term bias.

Algorithms The pseudo-algorithms to design the CUSUM chart, to select the target shift size, to choose the block length and the length of the input vector are explained in the supplementary material in Appendix B.

Performance criteria The performances of the support vector procedures for the high and low frequency monitoring are displayed in the supplementary material in Appendix C.

Simulations A comparison between the proposed scheme and a purely univariate control chart is also provided in the supplementary material in Appendix D.

Additional figures Additional analyzes and figures are also provided in the supplementary material in Appendix E.

References

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–267.

Chang, H.-Y. and Oh, S.-J. (2012). Does correction factor vary with solar cycle? *Journal of Astronomy and Space Sciences*, 29(2):97–101.

Cheng, C.-S., Chen, P.-W., and Huang, K.-K. (2011). Estimating the shift size in the process mean with support vector regression and neural network. *Expert Systems with Applications*, 38(8):10624–10630.

Clette, F. (2013). Private Communication: Talk presented at 3rd Sunspot Number Workshop (Tucson, USA).

Dudok de Wit, T., Lefèvre, L., and Clette, F. (2016). Uncertainties in the sunspot numbers: estimation and implications. *Solar Physics*, 291(9-10):2709–2731.

Ermolli, K., Matthes, K., Dudok de Wit, T., Krivova, N., Tourpali, K., Weber, M., Unruh, Y., Gray, L., Langematz, U., Pilewskie, P., Rozanov, E., Schmutz, W., Shapiro, A., Solanki, S., and Woods, T. (2013). Recent variability of the solar spectral irrandiance

and its impact on climate modelling. *EGU publication : Atmospheric Chemistry and Physics*, 13:3945–3977.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*. Wiley, New-York, 3rd edition.

Haigh, J. (2002). The effects of solar variability on the Earth's climate. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*, 361(1802):95–111.

Hathaway, D. H. (2010). The solar cycle. *Living Reviews in Solar Physics*, 7:1.

Izenman, A. (1985). J.R Wolf and the Zurich sunspot relative numbers. *The Mathematical Intelligencer*, 7(1):27–33.

Kruskal, W. and Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics*, 27(1):386–404.

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In Lepage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 225–248. Wiley, New-York.

Lloyd, S. (1957). Least squares quantization in PCM. Technical Report RR-5497 5497, Bell Lab.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. and Neyman, J., editors, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. University of California Press, California, United States.

Mathieu, S. (2021). *Statistical analysis and monitoring of time-series panels, with a particular focus on sunspot counts*. PhD thesis, Université catholique de Louvain (Belgium).

Mathieu, S., von Sachs, R., Delouille, V., Lefevre, L., and Ritter, C. (2019). Uncertainty quantification in sunspot counts. *The Astrophysical Journal*, 886(1):7.

Morfill, G., Scheingraber, H., Voges, W., and Sonett, C. (1991). Sunspot number variations -stochastic or chaotic. In Sonett, C., Giampapa, M., and Matthews, editors, *The Sun in Time*, pages 30–58. University of Arizona Press, Tucson, United States.

Page, E. S. (1961). Cumulative sum charts. *Technometrics*, 3(1):1–9.

Qiu, P. (2013). *Introduction to Statistical Process Control*. CRC Press, Taylor and Francis Inc, 1st edition.

Qiu, P. and Xiang, D. (2014). Univariate dynamic screening system: an approach for identifying individuals with irregular longitudinal behaviour. *Technometrics*, 56(2):248–260.

Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand, 1st edition.

Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Temmer, M., Veronig, A., Hanslmeier, A., Otruba, W., and Messerotti, M. (2001). Statistical analysis of solar Hα flares. *Astronomy and Astrophysics*, 375(3):1049–1061.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New-York, 1st edition.

Vigouroux, A. and Delache, P. (1994). Sunspot numbers uncertainties and parametric representations of solar activity variations. *Solar Physics*, 152(1):267–274.

Wang, Y.-M. and Colaninno, R. (2014). Is Solar Cycle 24 producing more coronal mass ejections than Cycle 23? *The Astrophysical Journal Letters*, 784(L27):1–7.

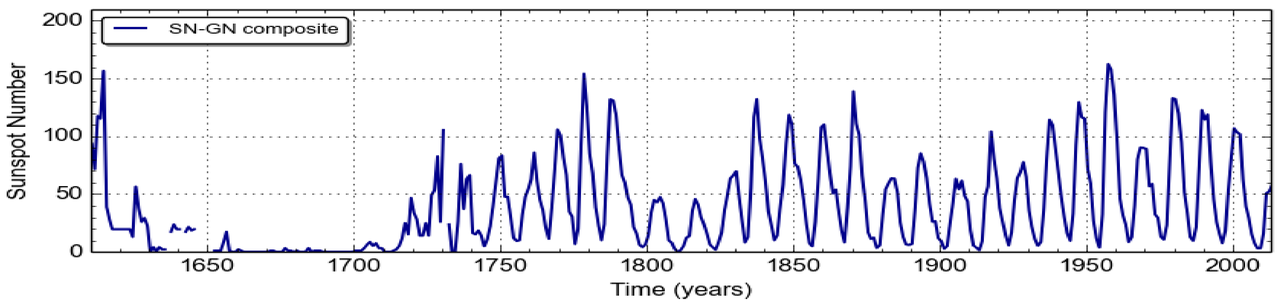
Uncertainty Quantification in Sunspot Counts

Sophie Mathieu [1], Rainer von Sachs [1], Véronique Delouille [2] and Laure Lefèvre [2]

1. Institute of Statistics, Bio-statistics and Actuarial sciences, Université catholique de Louvain, Louvain-la-Neuve, Belgium
2. Solar physics and space weather department, Royal Observatory of Belgium, Brussels, Belgium

E-mail: soph.mathieu@uclouvain.be

The Sunspot Number time series: a benchmark in space science



1. Introduction

Sunspots are dark areas on the sun corresponding to regions of locally enhanced magnetic field and act as an indicator of the solar activity. They have been counted since the invention of the telescope in the 17th century. The count of spots from each observing stations are later combined on a monthly basis at the Royal Observatory of Belgium to produce the International Sunspot Number (ISN) [1]. While the time series of the ISN acts as a benchmark in a large variety of physical sciences, as of today it lacks proper uncertainty quantification and modeling.

We build upon the work in [3], which presents a first uncertainty analysis of time domain errors and dispersion amongst the stations assuming a Poisson distribution. In this poster, we propose a more comprehensive error model that accounts for all types of errors known to the experts, taking into account the zero-inflated and overdispersed nature of the data.

2. Model of Interest

We propose the noise model for the count of spots N_s

$$Y_i(t) = (\varepsilon_1(t) + \varepsilon_2(i,t))s(t) + \varepsilon_3(t),$$

where $Y_i(t)$ is the N_s recorded by station (i.e. observatory) i at time t and

$s(t)$	true number of sunspots (integers)
$\varepsilon_1(i,t) \sim (0, \sigma_1^2(t))$	dispersion error across stations
$\varepsilon_2(i,t) \sim (\mu_2(i,t), \sigma_2^2(i))$	long term bias
$\varepsilon_3(t)$	error at minima : when $s(t) = 0$ (integers)

We assume that all terms are non-negative and jointly independent.

- **Short-term (< 27 days or a solar rotation)**
  As ε_1 is dominant at short term, we set $\mu_2(i,t) = 1$.
  The short-term variability is i.d. among the stations, with $\widetilde{\varepsilon}(t) := \varepsilon_1(t) + \varepsilon_2$

$$Y(t) = \begin{cases} \widetilde{\varepsilon(t)}s(t) & \text{if } s(t) > 0 \\ \varepsilon_3(t) & \text{if } s(t) = 0 \end{cases}$$

- **Long-term (> 27 days or a solar rotation)**
  We look at the long-term regime by applying a low pass-band filter on the time series, typically a MA with a window larger than 27 days ($\star$ denotes the smoothing process). $\varepsilon_2(i,t)$ is dominant in the long-term regime

$$Y_i^\star(t) = \begin{cases} \varepsilon_2(i,t)s(t)^\star & \text{if } s(t) > 0 \\ \varepsilon_3(t)^\star & \text{if } s(t) = 0 \end{cases}$$

  By analogy with the analysis of variance models, the identification constraint of the model is

$$\prod_{i=1}^{N} \mu_2(i,t) = 1,$$

  leading to the following estimator of the long-term bias

$$\widehat{\mu_2}(i,t) = \frac{Y_i^\star(t)}{\left(\prod_{i=1}^{N} Y_i^\star(t)\right)^{1/N}}. \quad (1)$$

3. Data

Fig. 1: Actual network of observing stations.

Characteristics of our dataset

- Period from January 1st, 1947 till December 31, 2013
- Subset of 21 stations
- Scaling
  Due to different characteristics of the observing means (telescope, location, etc.), a pre-processing is needed to rescale all stations to the same level.
  We use a criteria of stability in time with respect to the median of the network to select a pool Γ of Q 'good' stations.
  med_i denotes the median of $Y_i(t)$ over the pool Γ.
  For each station i, we define a **yearly** scaling factor k_i that is constant over a year:

$$k_i = \frac{1}{T}\sum_{t=1}^{T} \frac{\text{med}_{i\in\Gamma}}{Y_i(t)},$$

  where we choose T equal to one year.

4. Solar signal estimation

We define a proxy for the true number of spots as :

$$\widehat{\mu_s(t)} = \underset{i\in\Gamma}{\text{med}}\, Y_i(t),$$

The PDF of $\widehat{\mu_s(t)}$ for N_s may be approximated by a zero-altered generalized negative binomial (ZANB).

A ZA distribution models the zero values by a Bernoulli distribution $f_0(x)$ and non-zero values with a PDF $f_1(x)$ to be specified and defined with respect to a different discrete point measure [5, 2]:

$$f(x) = \begin{cases} f_0(0) & \text{if } x = 0 \\ (1 - f_0(0))\frac{f_1(x)}{1 - f_1(0)} & \text{if } x > 0 \end{cases} \quad (2)$$

Here $f_1(x)$ is a generalized negative binomial.

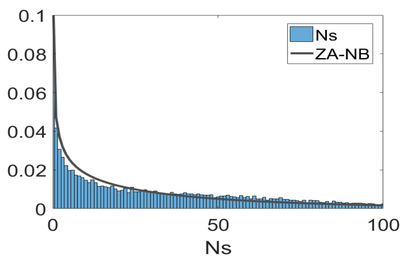


Fig. 2: Histogram of $\widehat{\mu_s(t)}$ for the count of spots N_s.The black line represents the fit of the distribution. The parameters values are $pbern = 0.115$, $p = 0.016$, $r = 0.602$ for the ZA-NB.

5. Short-term variations

When the median of the pool is different from zero, we have access to estimated values of $\widetilde{\varepsilon}$ by taking:

$$\widehat{\widetilde{\varepsilon}}(i,t) = \frac{Y_i(t)}{\widehat{\mu_s(t)}}$$

The PDF that fits best the distribution is a ZA t location-scale (t LS) [6, 4], where the density function $f_1(x)$ of Eq. 2 is a t LS. Such distribution allows the modeling of r.v. with heavier tails than the normal distribution.

The density of a t-Location-Scale is defined (for $\nu > 0$ and $\sigma > 0$) by

$$f(x,\mu,\sigma,\nu)_{tLS} = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(\nu + \frac{(x-\mu)^2}{\sigma}{\nu}\right)^{-(\frac{\nu+1}{2})}$$

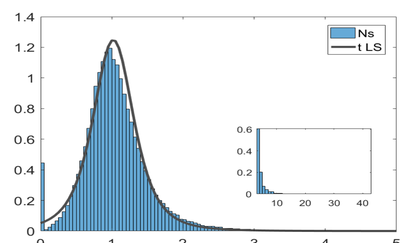


Fig. 3: Histogram of $\widehat{\widetilde{\varepsilon}}$ for the count of spots N_s. The continuous line shows the fit using a t LS distribution, with parameters values equal to $\mu = 1.02$ (mean), $\sigma = 0.30$ (standard deviation), and $\nu = 3.13$ (shape factor). The enclosed box represents a zoom on outliers with values larger than 3.

6. Errors during solar minima

Observed values of ε_3 are defined as counts made when the median of the pool (a proxy for s(t)) is equal to zero.

$$Y(t) = \varepsilon_3(t) \text{ when } \widehat{\mu_s}(t) = 0$$

Its PDF may be described by a ZANB for the N_s.

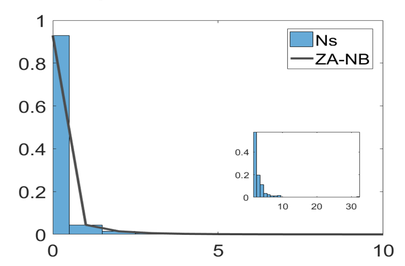


Fig. 4: Histogram of $\widehat{\varepsilon_3}$ for the counts of spots N_s. The continuous line shows the fit using a ZANB distribution, with parameters values equal to $pbern = 0.93$, $p = 0.4$, $r = 0.07$. The enclosed box represents a zoom on outliers with values larger than 1.

7. Long-term drifts

A moving average (MA) on 54 days was applied as a low-pass filter in order to ensure that the denominator in Eq (1) is non-zero, even in periods of solar minima.

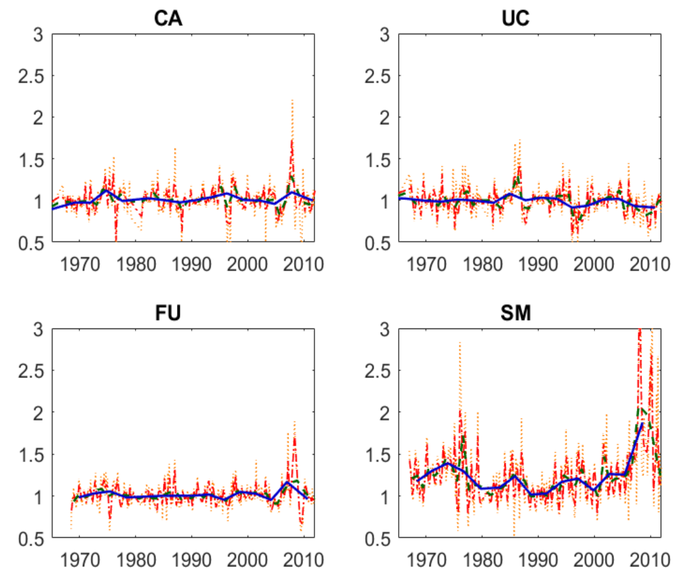


Fig. 5: Estimation of the long-term drifts $\widehat{\mu_2}(i,t)$ of N_s in four stations (CA, FU, UC and SM). $\widehat{\mu_2}(i,t)$ is shown averaged over 27 days (orange dotted line), 81 days (red dash-dot line), 1 year (green dashed line) and 2.5 years (blue plain line).

Fig. 5 represents the long-term drifts associated to four stations for the period studied. (We only represent it from 1970). Stations CA, FU, and UC are included in the pool Γ and are relatively stable, unlike the last station, SM, which displays severe drifts.

8. Summary

Estimated PDF

	$\widehat{\mu_s(t)}$	$\widetilde{\varepsilon}$	ε_3
N_s	ZANB	ZA t−LS	ZANB

The best fit for the short-term error $\widetilde{\varepsilon}$ was obtained with the Matlab function `allfitdist.m`, while for $\widehat{\mu_s(t)}$ and ε_3 different distributions were tested manually.

Our model takes into account:

- Multiplicative and additive framework
- Incorporates prior information on all types of error
- Excess of zeros
- Over-dispersion

Key results

- Short-term error distribution
  → Detection of daily outliers
- Estimation of long-term drifts
  → Quality control of the stations

9. Discussion

This study paves the way for a more comprehensive statistical monitoring of the stations. Such monitoring should include the definition of a robust and reliable pool of reference stations possibly evolving over time, and the triggering of alert in real-time when a station begins to drift or if a break-point is observed.

An iterative procedure may be devised to redefine the pool of stations Γ from this analysis. Indeed, once we have estimates for $\mu_2(i,t)$ and the daily outliers $\widetilde{\varepsilon}$, it is possible to iterate the process by first recomputing the k_i using the median over a more stable set of stations. And afterward reevaluating the different errors using a proxy $\widehat{\mu_s}(t)$ defined on more stable stations.

References

[1] F. Clette, L. Lefèvre, M. Cagnotti, S. Cortesi, and A. Bulling. The Revised Brussels-Locarno Sunspot Number (1981 - 2015). *Solar Physics*, 291:2733–2761, November 2016.

[2] A. Colin Cameron and Pravin. K. Trivedi. *Regression Analysis of Count Data*. Cambridge university press, 2 edition, 2013.

[3] T. Dudok de Wit, L. Lefèvre, and F. Clette. Uncertainties in the Sunspot Numbers: Estimation and Implications. *Solar Physics*, 291:2709–2731, November 2016.

[4] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. John Wiley and Sons, 3 edition, 2000.

[5] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. *Mixed effects models and extensions in ecology with R*. Springer, 2009.

[6] J. Taylor and A. Verbyla. Joint modelling of location and scale parameters of the t distribution. *Statistical Modelling*, 4(2):91–112, 2004.

Acknowledgements

The first author gratefully acknowledges funding from the Belgian Federal Science Policy Office (BELSPO) through the BRAIN VAL-U-SUN project (BR/165/A3/VAL-U-SUN).

Estimation and Modelling of Sunspot Counts

Sophie Mathieu [1], Rainer von Sachs [1], Christian Ritter [1], Véronique Delouille [2] and Laure Lefèvre [2]

1. Institute of Statistics, Bio-statistics and Actuarial sciences, Université catholique de Louvain, Louvain-la-Neuve, Belgium
2. Solar physics and space weather department, Royal Observatory of Belgium, Brussels, Belgium

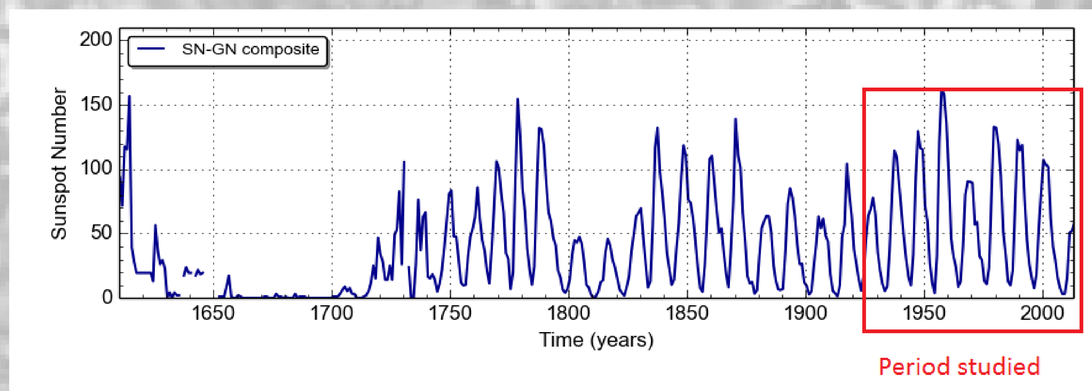
E-mail: soph.mathieu@uclouvain.be

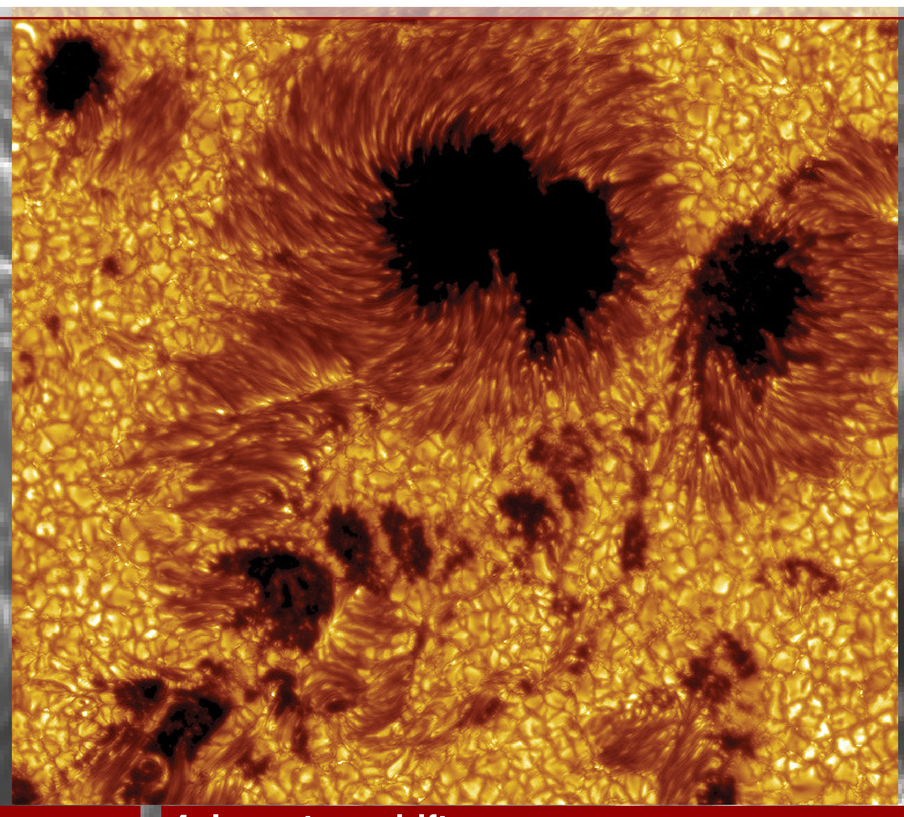
Introduction

We provide the following main contributions:

- A robust estimator for the unobserved number of spots (N_s) and a model of its density (taking into account the overdispersion and large number of zero counts).
- An uncertainty model motivated by first studies in [2] in a **multiplicative** framework. The model distinguishes **three** types of errors
- Similar results (not shown here) are proposed for the number of sunspot groups N_g and composite $N_c = 10N_g + N_s$ at the basis of the Internal Sunspot Number (ISN)
- The study is robust to missing values (do not require to fill-in missing observations)

The ISN time series: a benchmark for the solar activity



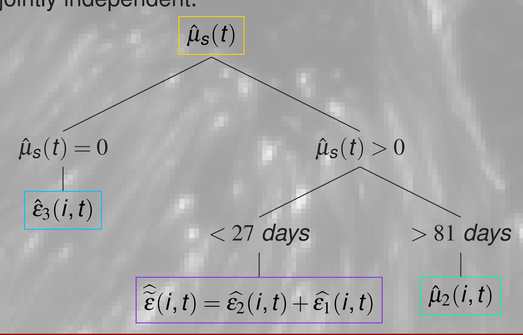


Model

$$Y_i(t) = \begin{cases} (\varepsilon_1(i,t) + \varepsilon_2(i,t))s(t) & \text{if } s(t) > 0 \\ \varepsilon_3(i,t) & \text{if } s(t) = 0. \end{cases} \quad (1)$$

$Y_i(t)$	counts of spots (N_s) recorded by station i, $1 \leq i \leq 21$, at time t (rescaled to adjust for variable instruments and seeing conditions in the different stations)
$s(t)$	true (i.e. without errors) number of sunspots (discontinuous r.v.)
$\varepsilon_3(i,t)$	error at minima : when $s(t) = 0$ (continuous r.v.)
$\varepsilon_1(i,t)$	short-term variability (continuous r.v.)
$\varepsilon_2(i,t)$	long term bias (continuous r.v.)

We assume s, ε_1, ε_2 and ε_3 to be jointly independent.



1. Solar signal estimation

$$\hat{\mu}_s(t) = T(\underset{1 \leq i \leq N}{\text{med}}\, Y_i(t)) \quad (2)$$

T is a *transformed* version of the median of the network composed of:

- Anscombe transform [4] (variance stabilisation)
- Wiener filtering (clean data from high frequencies)

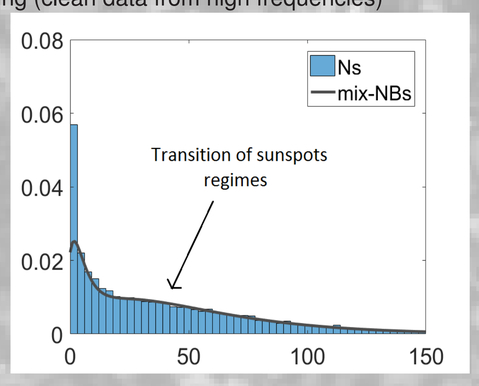


Fig. 1: Histogram of $\widehat{\mu_s}(t)$ for N_s. The density $\widehat{\mu_s}(t)$ may be approximated by a zero-altered mixture of generalized negative binomials [1]. The black line represents the fit of the distribution outside zero.

Simulation

A simulation of N_s was designed to study :

- the origin of the solar variability
- the effects of the filtering procedure
- (future) the impacts of the rescaling of the data
- (future) the influence of missing values on a future quality control of the stations

The algorithm is based on statistical distributions (instead of solar dynamo).
The simulation highlights:

- The important solar variability which guides us to apply the filtering
- The presence of an excess around $N_s \approx 40$ (mainly hidden by the solar variability in the unfiltered median).
  This excess probably corresponds to the different regimes of sunspots growing (from small spots to larger spots with penumbra)

2. Errors at solar minima

$$\hat{\varepsilon}_3(i,t) = Y_i(t) \text{ when } \hat{\mu}_s(t) = 0 \quad (3)$$

We mainly observe at minima:

- True zeros (no sunspots and no sunspots are reported)
- Short-duration sunspots (local excesses around 1 et 2)

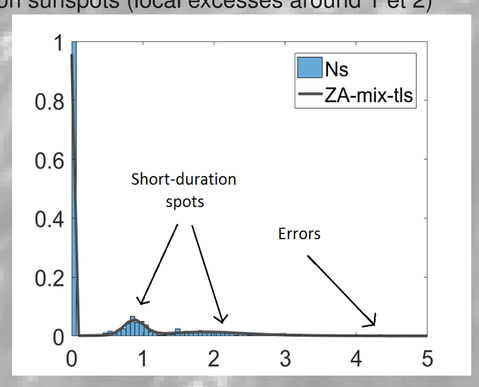


Fig. 2: Histogram of $\hat{\varepsilon}_3$ for the count of spots N_s. The black line represents the fit of the density by a mixture of t location-scale [3] distributions.

3. Short-term variations

$$\widehat{\tilde{\varepsilon}}(i,t) = \frac{Y_i(t)}{\hat{\mu}_s(t)} \text{ when } \hat{\mu}_s(t) > 0 \quad (4)$$

Usually a team of observers experience more variability than a single person (due to the shift of observers with different experiences and methodologies).

FU (Fujimori, Japan)	Individual observer
LO(Locarno, Switzerland)	Professional obs. with one main observer
SM (San Miguel, Argentina)	Team of observers
UC (Uccle, Belgium)	Team of observers

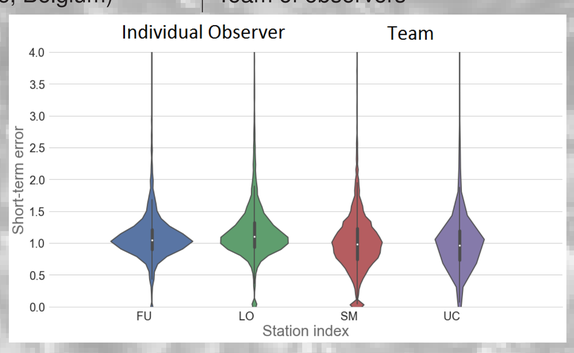


Fig. 3: Truncated violin plots of the estimated short-term variability $\widehat{\tilde{\varepsilon}}$ for N_s in four stations (FU, LO, SM and UC). A violin plot combines a vertical box-plot with a smoothed histogram represented symmetrically to the left and right of the box.

4. Long-term drifts

Let the $\star$ denote a moving average (MA) on windows of different lengths

$$\hat{\mu}_2(i,t) = \left(\frac{Y_i(t)}{\underset{1 \leq i \leq N}{\text{med}}\, Y_i(t)} \right)^{\star} \text{ when } \underset{1 \leq i \leq N}{\text{med}}\, Y_i(t) > 0 \quad (5)$$

- Severe drifts in SM and LO
- FU and UC appear relatively stable
- Bias in the counting process is larger during solar minima:
  higher relative errors during minima than during the remaining part of the solar cycle
- Some **jumps** are visible with the MA length of 81 days while longer scales highlight the **drifts**

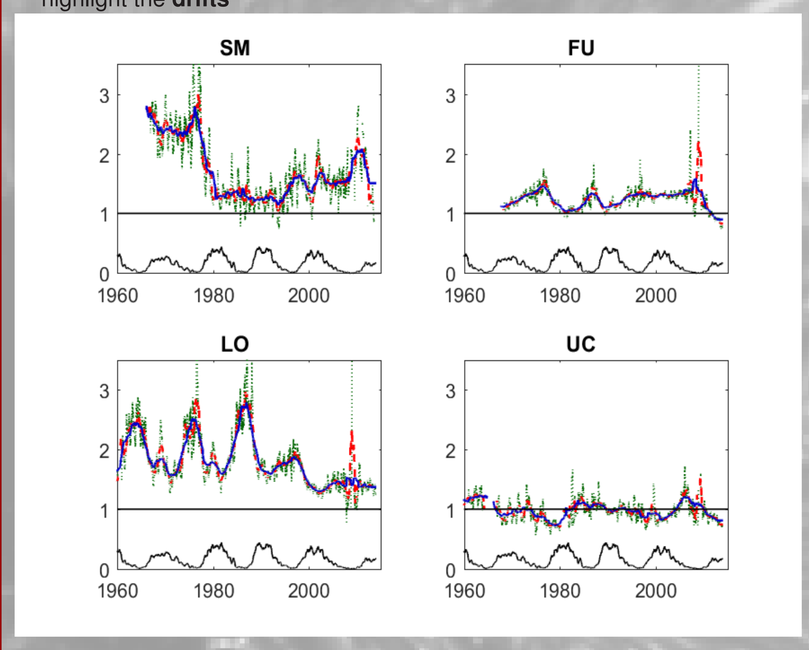


Fig. 4: Estimation of $\hat{\mu}_2(i,t)$ for N_s in four stations (SM, FU, LO and UC). $\hat{\mu}_2(i,t)$ is computed with different MA window lengths: 81 days (green dotted line), 1 year (red dashed line) and 2.5 years (blue plain line)

Conclusion

Key results

- Short-term variability
- Estimation of long-term drifts

Toward a monitoring

With aim to alert the stations in quasi real-time when they start deviating from the network to prevent large drifts observed in the series (see Fig.4).

References

[1] A. Colin Cameron and Pravin. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 2 edition, 2013.

[2] T. Dudok de Wit, L. Lefèvre, and F. Clette. Uncertainties in the sunspot numbers: estimation and implications. *Solar Physics*, 291:2709–2731, November 2016.

[3] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. John Wiley and Sons, 3 edition, 2000.

[4] M. Makitalo and A. Foi. Optimal inversion of the generalized Anscombe transformation for Poisson-gaussian noise. *IEEE Transactions on Image Processing*, 22(1), 2013.

Acknowledgements

The first author gratefully acknowledges funding from the Belgian Federal Science Policy Office (BELSPO) through the BRAIN VAL-U-SUN project (BR/165/A3/VAL-U-SUN).