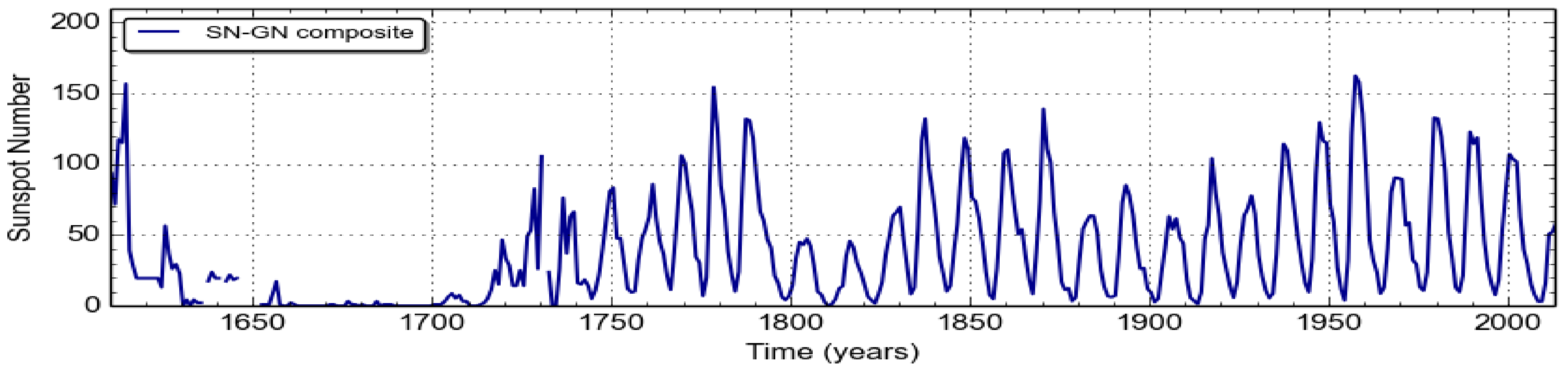
Uncertainty Quantification in Sunspot Counts

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

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

3. Data

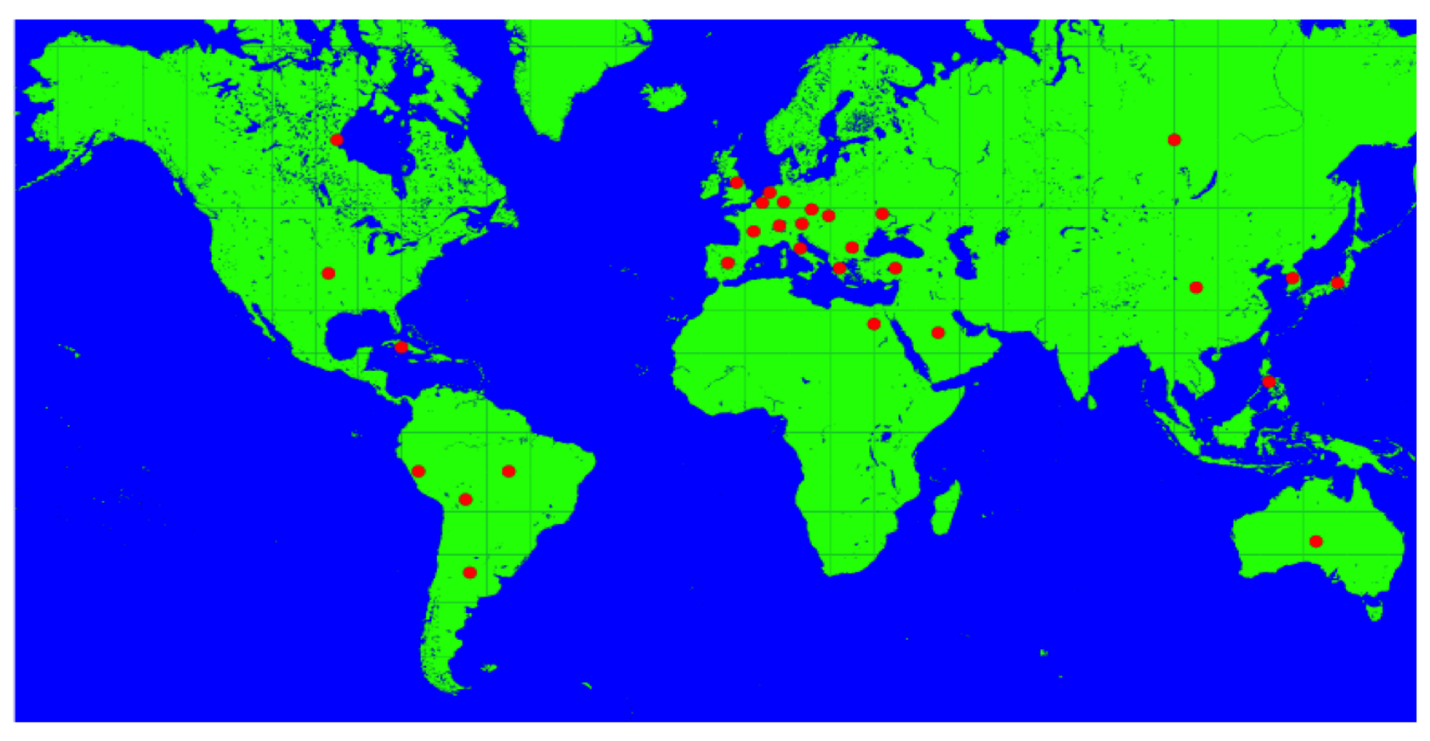


Fig. 1: Actual network of observing stations.

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]:

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

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

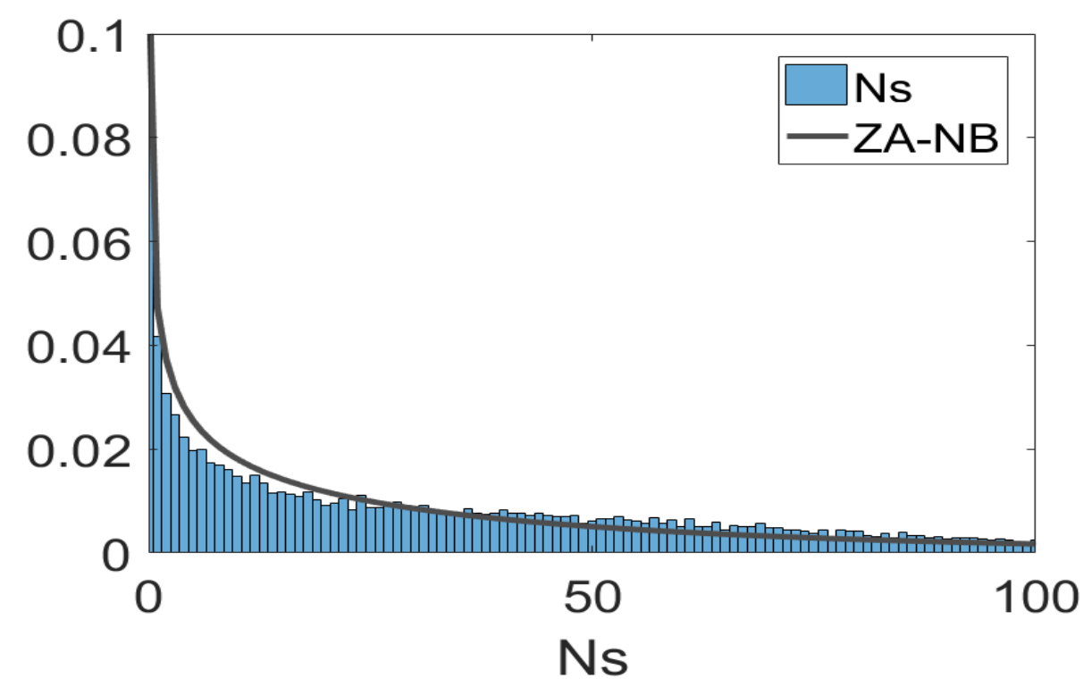


Fig. 2: Histogram of $\widehat{\mu_s}(t)$ for the count of spots N_s.The black line represents the fit of the distribution. The parameters values are $pbern = 0.115$, $p = 0.016$, $r = 0.602$ for the ZA-NB.

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

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

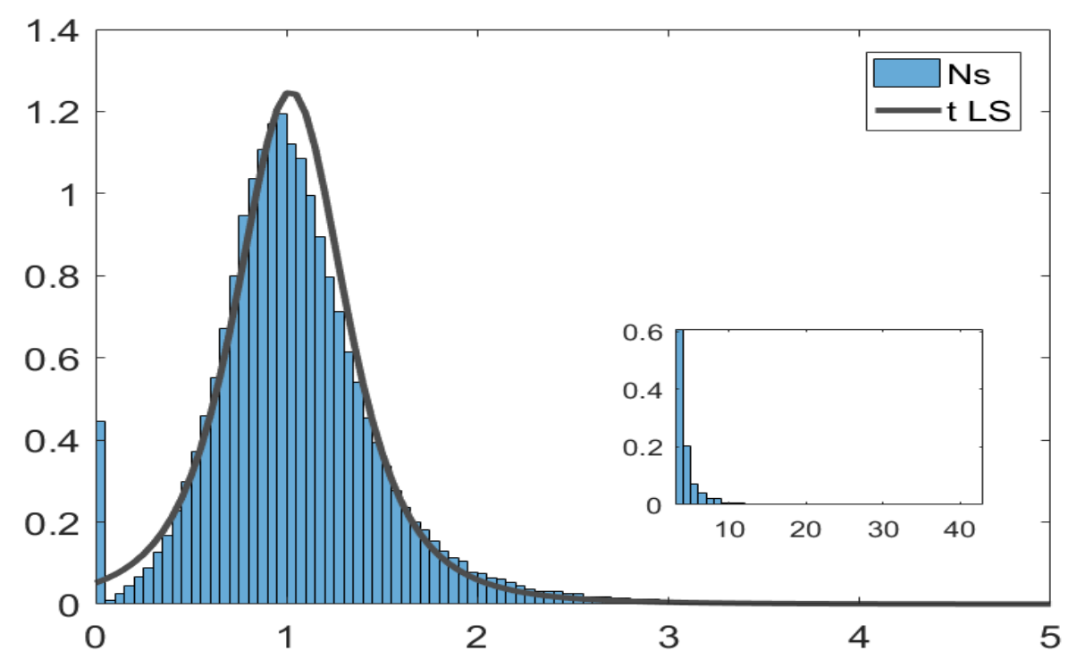


Fig. 3: Histogram of $\widehat{\widetilde{\varepsilon}}$ for the count of spots N_s. The continuous line shows the fit using a t LS distribution, with parameters values equal to $\mu = 1.02$ (mean), $\sigma = 0.30$ (standard deviation), and $\nu = 3.13$ (shape factor). The enclosed box represents a zoom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

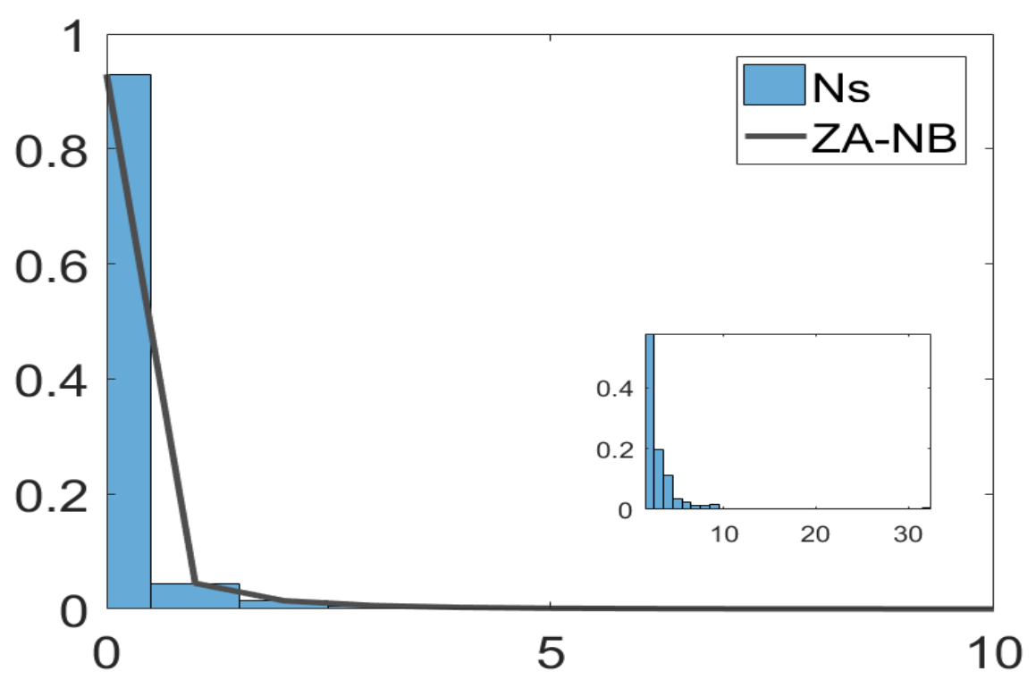


Fig. 4: Histogram of $\widehat{\varepsilon_3}$ for the counts of spots N_s. The continuous line shows the fit using a ZANB distribution, with parameters values equal to $pbern = 0.93$, $p = 0.4$, $r = 0.07$. The enclosed box represents a zoom on outliers with values larger than 1.

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.

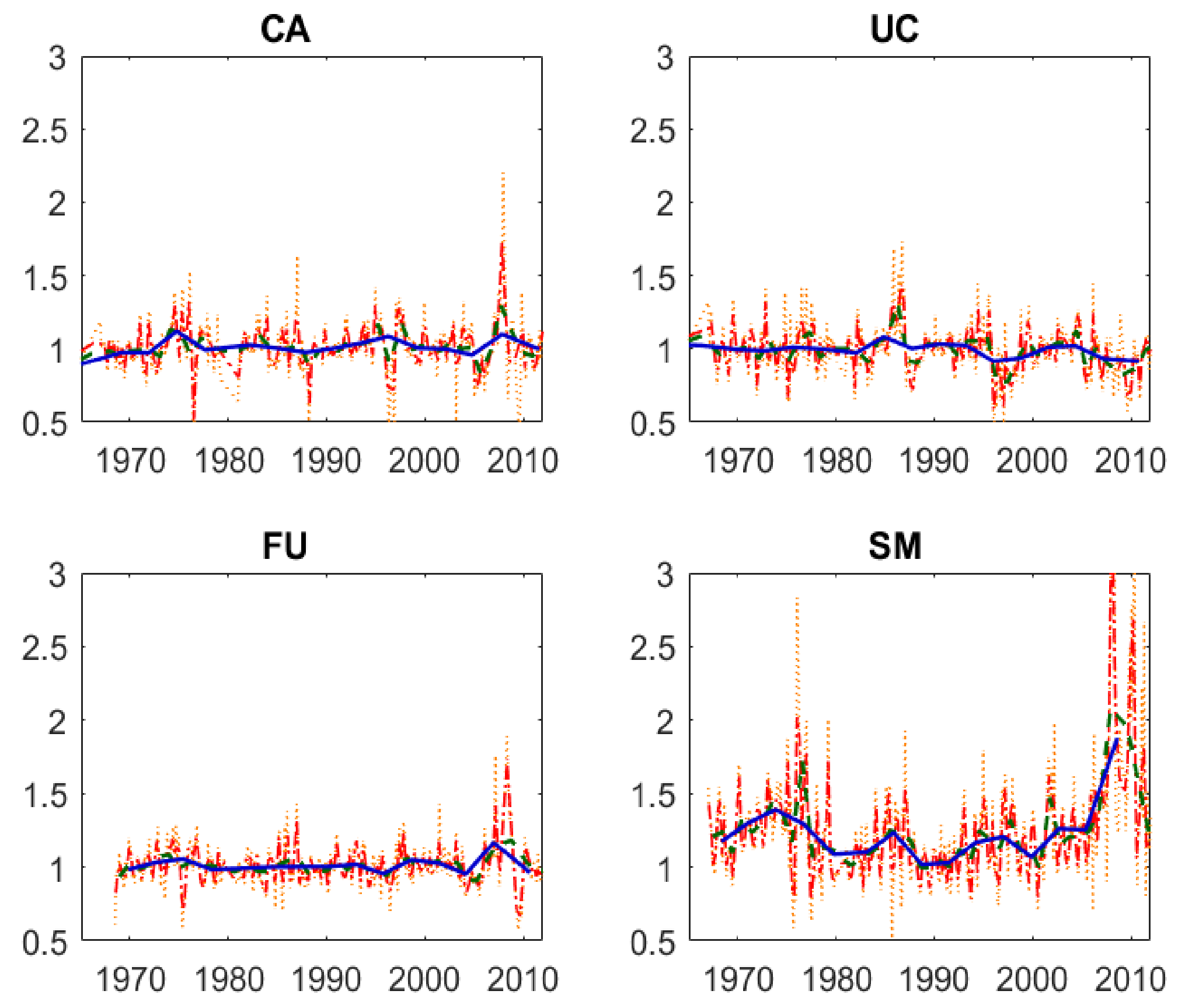


Fig. 5: Estimation of the long-term drifts $\widehat{\mu_2}(i,t)$ of N_s in four stations (CA, FU, UC and SM). $\widehat{\mu_2}(i,t)$ is shown averaged over 27 days (orange dotted line), 81 days (red dash-dot line), 1 year (green dashed line) and 2.5 years (blue plain line).

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).