Parameters of the monitoring

S. Mathieu

December 10, 2020

1 Overview

Table 1: Main parameters of the monitoring procedure

Symbol	Name	Units	Typical values	How to set
-	min time-scale of the rescaling (M_t)	months	8 (Ns), 14 (N_g), 10 (N_c)	Kruskal-Wallis test
-	MA window length (μ_2)	days	≥ 27	physical scale (periodogram)
-	MA window length (h)	days	>8 months	physical scale (periodogram)
K	number of nearest neighbours	-	[10-10000]	algorithm (data)
-	block length	number of obs (here days)	54 [1–100]	algorithm (data)
δ	target shift size	-	1.5 [0.1–2]	algorithm (actual deviations)
k	allowance parameter	-	$\delta/2$	optimal formula
h	control limit	-	[0-30]	searching algorithm
ARL_0	IC average run length	-	200 [50-500]	rate of false positives
m	length of the input vector (SVR and SVC)	number of obs (here days)	70 [10-150]	algorithm (δ)
-	scale (half-normal distribution)	-	3.5 [1-4]	max values of the data
ϵ	accuracy error (SVR)	-	0.001 [0.0001–1]	optimization
λ	regularization parameter (SVR and SVC)	-	10 [1–100]	optimization
-	kernel function (SVR and SVC)	-	rbf	optimization

Main parameters of the monitoring procedure. The table details the symbol, name, unit, typical values as well as the method that may be used to select an appropriate value for each parameter. The first rows are related to the error model and the identification of the long-term error from other components of the model. The second set of rows represent parameters of the control scheme. The last rows account for parameters of the support vector machine procedures.

2 Parameters of the model

The three following parameters are related to the error model that we develop in Mathieu et al. (2019) and the estimation of the long-term error.
As a remainder, let $Y_i(t)$ represent either the number of spots, groups or composite actually 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, in a *multiplicative* framework:

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

ϵ_1 denotes the short-term error representing counting errors and variable seeing conditions; ϵ_2 is a long-term error accounting for systematic bias in the counting. The error ϵ_3 is occurring only at solar minima (when $s(t)$ is equal to zero); it captures effects like short-duration sunspots and the non-simultaneity of the observations across the network of stations. We also introduce $h(i,t)$, which represents the individual levels of the stations. Those typically correspond to the different weather conditions and instruments of the stations.

We aim at monitoring the long-term error of the stations, ϵ_2, which represents the potential bias of the observatories. These errors may be estimated with:

$$\widehat{eh}(i,t) = \left(\frac{Y_i(t)}{M_t}\right)^\star \quad \text{when } M_t > 0, \quad (2)$$

where $eh = \epsilon_2 + h$.

2.1 Time-scale of the rescaling

To estimate the long-term error, we first need to compute M_t, the median of the rescaled observations. The rescaled observations may be obtained by multiplying the observations $Y_i(t)$ by scaling-factors evaluated on a determined period. As explained in Mathieu et al. (2019) section 4, the duration of the period where the scaling-factors are assumed to be constant may be selected by a statistical-driven study based on the Kruskal-Wallis test (Kruskal and Wallis, 1952). This test shows that the minimal time-scale is equal to 8 months for N_s, 14 months for N_g and 10 months for N_c.

2.2 Scale of the long-term

Then, the short-term error ϵ_1 should be untangled from the long-term error and the levels in (1). To this end, a smoothing process by a moving-average (MA) filter, denoted by the $\star$ in (2) is applied on the ratio $Y_i(t)/M_t$. The window length of the MA-filter should be larger than or equal to 27 days, a physical scale of the data corresponding to one solar rotation. This value of 27 days appears to be sufficiently high to overcome the effects of the short-term regime as demonstrated in Mathieu et al. (2019). Different window lengths may then be selected to better highlight the jumps (such as 27 or 54 days) or the drifts (such as 81, 365 or 900 days).

2.3 Scale of the levels

The levels of the stations may finally be separated from the long-term error by applying once again a smoothing process:

$$\hat{\mu}_2(i,t) = \widehat{eh}(i,t) - \widehat{eh}^{\star}(i,t). \tag{3}$$

The scale of the levels should be superior or equal to 8 months for N_s, 14 months for N_g and 10 months for N_c as previously demonstrated by the Kruskal-Wallis test. For identification purpose, we assume that h is a slowly-changing function, which does not vary too much with respect to ϵ_2. Therefore, the MA-filter window length of (3) should also be larger than the window length of (2). We may select the window length at eleven years or one solar cycle, to be much higher than most of the typical scales of the long-term error. This value also seems appropriate since the location of the observatories or their telescope are unlikely to change much over time.

3 Parameters of the control scheme

The following parameters are related to the construction of the control scheme. The method that we used is based on those developed by Qiu and Xiang (2014). It is composed of two phases. In the first phase, the regular patterns, i.e. the mean and the variance of the long-term error, are estimated on a subset of stable or in-control stations, often called 'pool' in the following. Then, the errors of all stations are standardized by these parameters. In the second phase, a CUSUM chart is applied on the standardized errors for the quality control. A support vector machine classifier and regressor (Cheng et al., 2011) are also added on top of the chart to allow the prediction of the size and the form of the shifts after each alert.

As a remainder, the two-sided cumulative sum (CUSUM) (Page, 1961) applied on the standardized errors writes as:

$$\begin{aligned} C_j^+ &= max(0, C_{j-1}^+ + \hat{\epsilon}_{\hat{\mu}_2}(t) - k) \\ C_j^- &= min(0, C_{j-1}^- + \hat{\epsilon}_{\hat{\mu}_2}(t) + k), \end{aligned} \tag{4}$$

where $j \geq 1$, $C_0^+ = C_0^- = 0$ and $k > 0$ is the allowance parameter. The index i of the station is omitted since each station is monitored separately.
This chart gives an alert if $C_j^+ > h^+$ or $C_j^- < h^-$, where h^- and h^+ are the control limits of the chart. Since the distribution of the standardized errors is almost symmetric, we use $h^- = -h^+$.

3.1 Number of nearest neighbours

$$\begin{aligned} \hat{\mu}_0(t) &= \frac{1}{\Delta} \sum_{t'=t-\Delta/2}^{t+\Delta/2} \frac{1}{N'_{IC}} \sum_{i_{ic}=1}^{N'_{IC}} \hat{\mu}_2(i_{ic}, t') \quad s.t. \quad \Delta + N'_{IC} = K \\ \hat{\sigma}_0^2(t) &= \frac{1}{\Delta} \sum_{t'=t-\Delta/2}^{t+\Delta/2} \frac{1}{N'_{IC}} \sum_{i_{ic}=1}^{N'_{IC}} \left(\hat{\mu}_2(i_{ic}, t') - \hat{\mu}_0(t)\right)^2 \quad s.t. \quad \Delta + N'_{IC} = K, \end{aligned} \tag{5}$$

In the first phase of the chart design, the empirical mean ($\hat{\mu}_0(t)$) and standard deviation ($\hat{\sigma}_0(t)$) are estimated on a subset of in-control (IC) stations. These patterns are estimated using K-nearest neighbours estimators. The patterns are thus evaluated across the pool of IC stations and also along a small window in time which includes in total K observations. Hence, the mean and the standard deviation are always computed on the same number of values (K) even if the data contain missing values.
To automatically choose an appropriate value for K, we evaluate the standard deviation of the standardized errors for different values of K. The standard deviation decreases as K augments (contrarily to the mean which increases when K augments). Finally, we select the first value of K such that the standard deviation starts to stabilize, i.e. the 'knee' of the curve.
This parameter depends thus on the data and should be adjusted separately for N_s, N_g and N_c with the procedure previously described.

3.2 Block length

The chart is then designed by the block bootstrap (BB), a procedure which samples blocks of data to generate series with similar autocorrelation as those of the original data. The BB usually preserves the serial correlation of the data inside the blocks. Hence, the length of the blocks is an important parameter. 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. An optimal value for the block length may thus be computed as the first value such that the MSE of the autocorrelation of the errors starts to stabilize. This value intuitively represents the smallest length such as the main part of autocorrelation of the original series is well represented.
By using such a procedure, we select the block length at 54, for N_s, N_g and N_c. Since we have daily observations, the block length may be interpreted in days. 54 days corresponds here to a physical scale equal to two solar rotations. This value seems appropriate since it is larger than the lifetime of most of the sunspots.

3.3 Target shift size

The target shift size, δ, represents the size of the shift that we aim to detect. This parameter is usually defined as the difference between the mean of the errors before the deviation (μ_0) and the mean of the errors after the deviation (μ_1):

$$\delta = \mu_1 - \mu_0,$$

where $\mu_0 = 0$ since the errors are standardized. Hence, δ as the same unit as the standardized long-term error that we monitor. Since $\hat{\mu}_2$ is an error (it is defined as a smoothed ratio between the counts of a station and the counts of the reference), it has no unit. It may be interpreted as follows. Assuming that the counts/observations at time t are equal to $\hat{\mu}_2(i,t)M_t$ (neglecting the short-term error), if a deviation of size δ occurs in the errors, the observed counts are affected as $(\hat{\mu}_2(i,t) + \delta)M_t$.

An appropriate value for δ may be estimated in practice by an algorithm based on the actual deviations of the OC series. In this procedure, a new OC

series is sampled at each step by the block bootstrap and the deviation of the series is estimated using a classical formula (Montgomery, 2004):

$$\hat{\delta} = \begin{cases} k + \frac{C_i^+}{N^+} & \text{if } C_i^+ > h \\ -k - \frac{C_i^-}{N^-} & \text{if } C_i^- < -h, \end{cases} \tag{6}$$

where $\hat{\delta}$ is the estimated shift size expressed in standard deviation units and N^+ (resp. N^-) represents the number of observations where the CUSUM statistics C_i^+ (resp. C_i^-) has been non-zero.
In practice, values of δ around 1.5 appear to be adapted for our data.

3.4 Allowance parameter

The CUSUM statistics are defined as the cumulative sum of the deviations that exceed the allowance parameter. This parameter is usually fixed at $\delta/2$, an optimal relation for normally distributed data (Moustakides, 1986). It is also valid as long as the distribution of the standardized errors is not too far from the normal. This appears to be true for N_s, N_g and N_c.

3.5 Control limit

When the CUSUM statistics exceed the control limits, the errors are considered to be out-of-control (OC) and an alert may be sent to the observers. Hence, the control limits strongly depend on the distribution of the errors and differ for N_s, N_g and N_c. They may be adjusted by a searching algorithm until a pre-specified value of the IC average run length is reached at the desired accuracy.

3.6 IC average run length

The IC average run length, denoted ARL_0, is the mean value of the number of samples collected from the beginning of the process to the occurrence of a false alert. It represents the rate of false positives of the chart (it is similar to the concept of type I error in hypothesis testing context). Large values of ARL_0 are desirable since they reduce the number of false alerts. But at the same time, the chart with high control limits takes more time to detect the deviations. Typical values of the ARL_0 range therefore between 50 and 500, with 200 often selected.

4 Parameters of the support vector machines

The following parameters are directly related to the support vector machine (SVM) procedures.

4.1 Length of the input vector

The length (m) of the input vector represents the m most recent observations of the series that are fed into the SVR and SVC after an alert. The SVR and SVC then predict the size and the form of the deviation at the origin of the alert based on the input vector. m 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 while the smallest shifts may be identified only after a certain amount of time. Therefore, the latter require larger input vectors than the former. Hence, m may be selected as an upper quantile of the run length distribution, computed on data shifted by the smallest shift size that we aim to detect, δ_{min}. The length depends thus on the data and the target shift size. It varies therefore for N_s, N_g and N_c.
Since the sunspots are observed daily, the length of the input vector may be related to a number of days.

4.2 Scale parameter of the half-normal

In the construction of the training and validation sets, the magnitudes of the artificial shifts are randomly sampled from two half-normal distributions (Evans et al., 2000) supported by $[-\infty, ..., -\delta_{min}]$ and $[\delta_{min}, ..., \infty]$ respectively. The scale parameter of the half-normals may be fixed at a value that is sufficiently high to reproduce the largest values/deviations observed in the errors. A scale of 3.5 appears to be sufficient to reproduce the highest deviations of N_s, N_g and N_c.

4.3 Accuracy error

The accuracy error, denoted by ϵ, is a parameter of the support vector regression (SVR). It may be selected to maximize the accuracy of the predictions, measured by the mean absolute percentage error (MAPE) or the normalized root mean squared error (NRMSE). This parameter does not appear to have strong influence on the results however. Therefore, it may be fixed at 0.001 for N_s, N_g and N_c.

4.4 Regularization parameter

The regularization parameter, denoted by λ, represents the trade-off between the mis-classification and the regularization of the SVR and the support vector classifier (SVC). This parameter may be selected to maximize the accuracy of the predictions. Those are measured by the mean absolute percentage error (MAPE) or the normalized root mean squared error (NRMSE) for the SVR and by the accuracy for the SVC. Values around 10 appear to be appropriate for our data.

4.5 Kernel function

The kernel function is another important parameter of the SVR and SVC. It may also be selected to yield the best prediction accuracy. The kernel function is usually chosen among the sigmoid, the radial basis function (rbf), the polynomial or linear kernel. In practice however, the radial basis function appears to be the only kernel that works for our data.

References

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

Evans, M., N. Hastings, and B. Peacock (2000). *Statistical Distributions* (3rd ed.). Wiley.

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

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

Montgomery, D. (2004). *Introduction to Statistical Quality Control* (5th ed.). Wiley.

Moustakides, G. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics 14*(4), 1379–1387.

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

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