Minutes
(ROB meeting 23 October)

Sophie Mathieu

November 12, 2019

1 Plan

- Presentation of Shreya:
  her background, her previous work in India about tilted angles, the subject of her thesis about the reconstruction of the past series
- Feedback of the conferences
  The most recurrent comment was to develop an automated algorithm to extract and count the spots and groups. I am working on this algorithm when I have some time (at the end of the day).
- Questions and methodology of the monitoring
- The next meeting is fixed on 16 January (a skype may be planned before if needed).
- The methodology should be ready and tested on the data before the next summer.

2 Second article

The second article will be submitted to an applied statistical journal. Would *Technometrics* be too ambitious? We could try but need to be prepared to quickly submit to another journal in case of rejection (and have the back-up plan ready). The goal is to develop a generic methodology that can be applied to other data. Then, at the end of the article, the methodology will be applied on the sunspot numbers as an example. The procedure is thus driven from our data but should be applicable in other fields as well. Therefore, different datasets or situations where we may apply our monitoring should be added to the paper.
Note that in the first place, the monitoring will be applied on each station (i.e. we do not aim at supervising the mean or the median over the whole

network).
However, as extensions of our work, we should think about a) the definition of an adaptive pool of IC station that may replace in the future the single reference station (LO) to compute the ISN. b) This article is the groundwork for a monitoring of the mean of the stations (which may be used to detect unusual solar activity).

Christian's comment: Qiu could become one of our reviewers. We need to contact him (done!) and be aware of what he and his team is working on at the present time. The papers Qiu et al. (2017) and Qiu and Xiang (2014) are the most relevant for our current work and are summarized in the dropbox. Other summaries will be added on the dropbox soon. Otherwise, he may point out that he has already done what we propose. (Right now we try to prepare summaries of relevant work he and his group has done.)

- Other data sets:

Our data have complicated features such as the correlation, the missing values, the heterogeneous variability and the non-normal distribution. We have also data from a panel of stations.
We need to find other datasets that share at least some features of our data. Promising fields are healthcare (Qiu applies his monitoring to the individual cholesterol level of patients to prevent the occurrence of strokes), finance (personal income distribution ?) or physics.

- What are our contributions ?

- Guidelines to select an IC set in a generic way

- A solution when the individuals have time-varying levels that we do not want to detect by our monitoring
  It appears that this has already been done by Qiu Qiu et al. (2017) in Section 2.5.2.
  The author first removed the mean IC function $\mu(t)$ (same for all stations) from the data. Then, they computed the individual level using a weighted least square estimator. The individual levels are then subtracted from the data and the residual are divided by the IC standard deviation $\sigma(t)$.

- A methodology valid for 'imperfect' IC stations (compare our several-stage procedure with 'perfect' stations in IC and stations that are not during this IC)

- Development of a chart that is robust to missing values
  It appears that this has already been done by Qiu Qiu et al. (2017) but via a more heavy formalism. One possibility hear could also be to

simplify the methodology (and design a simple EWMA chart robust to missing values)?

- Results for the block bootstrap shall be showed (they were simply mentioned by Qiu). We may also provide studies about the impacts of the block bootstrap with respect to a parametric procedure.

- Development of a Python package for the monitoring

• Critical look on what we propose.

What is the advantage of using a control chart to monitor the different stations with respect to a statistical test (similar to a Shewhart chart)? The test may write as:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 - \mu_2 > 0 \end{aligned} \tag{1}$$

where μ_1 is the mean of the IC residuals and μ_2 represents the mean of the residuals in the monitoring part.

- The CUSUM chart is simple and statistically accepted by the community.

- The procedure is fully automated and detects shifts without human intervention

- The method is easy to implement and to explain (to the users and the observers)

- The chart has been calibrated to detect small shifts that can be hard to detect 'by hand' (or using the Shewhart chart)

- The chart automatically adapts itself to the sizes of the shifts (small shifts are detected after a longer period than large shifts)

A posteriori comment by Rainer: It would be good to clearly summarize what are advantages/shortcomings of either of the three basic charts we have been looking on, i.e. Shewhart, CUSUM and EWMA and see which is most suited for us in which situation.
The CUSUM and EWMA charts have similar performances. They are usually used in phase II SPC. Although the EWMA chart has a simpler formalism which is easier to understand, it lacks the theoretical optimal property of the CUSUM. The Shewhart chart does not use the history of the process to detect a shift contrarily to the EWMA and CUSUM. This chart is simple and makes time-point decisions in a framework similar to the hypothesis testing. The chart detects relatively well large shifts but does not perform well to detect small and persistent shifts. It is often used in phase I SPC.

3 Monitoring procedure

3.1 Step 1: Estimation of the regular longitudinal pattern

$i \in 1, ..., N = 21$ is the index of the station
$i_{ic} \in 1, ..., N_{IC}$ is the index of the IC station
$i_{oc} \in 1, ..., N_{OC}$ is the index of the OC station
$t \in 1, ..., T$ represents the time (one observation per day)
$\hat{\mu}_2$ denotes the variable to monitor (the estimation of the long-term error ϵ_2)

$$\hat{\mu}_2(i,t) = \left(\frac{Y_i(t)}{\underset{1 \leq i \leq N}{\text{med}}\, Y_i(t)} \right)^{\star} \quad \text{when} \quad \underset{1 \leq i \leq N}{\text{med}}\, Y_i(t) > 0, \tag{2}$$

Note that we apply the monitoring on the $\hat{\mu}_2(i,t)$ which is a ratio. This can be a problem as a ratio may have heterogeneous variability. Therefore, in a generic methodology, we may apply a logarithmic transformation on the data to stabilize their variance. However, this transformation is obviously not relevant for the sunspot numbers containing many zeros and hence has not been applied on our data.

3.1.1 Selection of the IC stations

The first step is to select a pool of IC stations.
We select a subset of N_{IC} (yet to be defined) stations from the 21 observatories using a stability criterion. Here, we use the mean squared error (MSE) with respect to the median of the network:

$$\begin{aligned} MSE[\hat{\mu}_2](i) &= Bias(\hat{\mu}_2(i,t))^2 + Var(\hat{\mu}_2(i,t)) \\ MSE[\hat{\mu}_2](i) &= [\frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_2(i,t) - 1]^2 + \frac{1}{T}\sum_{t=1}^{T} \left(\hat{\mu}_2(i,t) - \frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_2(i,t) \right)^2 \end{aligned} \tag{3}$$

- How many stations should we include in the (IC) pool?
  If we include too many stations, we may include stations with imperfect stability. While if we include too few stations, we may badly estimate the IC parameters $\mu_0(t)$ and $\sigma_0(t)$. The chart will also be too sensitive and trigger many alerts.

We propose to order the MSE of the stations in Equation 3 and find the largest consecutive difference. We may then select N_{IC} as the number of stations before the occurrence of the largest difference in MSE. Unfortunately, the largest difference appears between the most unstable stations. Since we observe a significant difference between the stations 9-10 and 9 is

close to $N/2$, we select $N_{IC} = 9$ stations.
We may also draw the histogram of the MSE and select all stations within 1σ or use a clustering algorithm to define the IC and OC pools.
Note that simply adding more stations will not give us more informations about the sunspot numbers. A larger pool, composed of different stations around the world, observing the Sun at different times will indeed give us more informations. But stations close in location and observing at the same time will not help.

- Is the MSE a relevant criterion to estimate our pool of IC stations ? We may select the IC stations using only the bias or the variance. Then, our pool of IC stations will change since some stations are more variable but aligned in general with the network and conversely. It may also be interesting to compute the MSE of the rescaled stations, i.e. stations aligned with the network $\hat{\mu}_2(i,t) - \hat{\mu}_1(i,t)$.

It appears to be a reasonable choice. The MSE should probably be computed on the rescaled stations $\hat{\mu}_2(i,t) - \hat{\mu}_1(i,t)$.

3.1.2 Estimation of the mean and the variance of the IC process

Since each station has its own level due to differences of methodology and/or instrument, we first remove the mean level of the stations, denoted $\hat{\mu}_1(i,t)$:

$$\hat{\mu}_2(i,t) - \hat{\mu}_1(i,t) = \hat{\mu}_2(i,t) - \frac{1}{(2\Delta_1)} \sum_{\tau=t-\Delta_1}^{t+\Delta_1} \hat{\mu}_2(i,\tau) \tag{4}$$

Then, we follow the methodology of Qiu. We compute the empirical mean, noted $\hat{\mu}_0(t)$, and the empirical variance, $\hat{\sigma}_0^2(t)$, of the IC stations at each time.

$$\begin{aligned} \hat{\mu}_0(t) &= \frac{1}{(2\Delta)} \sum_{\tau=t-\Delta}^{t+\Delta} \frac{1}{N_{IC}} \sum_{ic=1}^{N_{IC}} \hat{\mu}_2(i_{ic},\tau) \\ \hat{\sigma}_0^2(t) &= \frac{1}{(2\Delta)} \sum_{\tau=t-\Delta}^{t+\Delta} \frac{1}{N_{IC}} \sum_{i_{ic}=1}^{N_{IC}} (\hat{\mu}_2(i_{ic},\tau) - \hat{\mu}_0(t))^2 \end{aligned} \tag{5}$$

- Which length $2\Delta_1$ should we select to estimate $\mu_1(i,t)$?
- Which length 2Δ should we select to estimate $\mu_0(t)$ and $\sigma_0(t)$?

There are two possibilities to select the lengths: a) using background knowledge or b) using the simulations. From background knowledge, $2\Delta_1$ equals to one year appears as a reasonable choice (cf Kruskal-Wallis test in the previous paper). The lower bound for $2\Delta_1$ is 27 days.

We may also choose a first value for $2\Delta_1$ to compute the MSE in Equation 3 and a second value for $2\Delta_1$ to rescale the stations in Equation 4.

- Is it interesting to estimate $\mu_0(t)$ and $\sigma_0(t)$ using the panel ?
  We may standardize the stations with $\hat{\mu}_1(i,t)$ and $\hat{\sigma}_1(i,t)$, i.e. neglecting the panel. The panel would then only be used to adjust the control limits of the chart.

Probably better to keep the panel (to be tested by simulations).

3.2 Step 2: Monitoring of the longitudinal pattern of the observations

Using the estimated mean and variance for the IC process, the observations are standardized:

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

We omit the index i in the remainder as we monitor each individual separately.
Then, we apply a classical control chart such as the CUSUM chart on the residuals:

$$\begin{aligned} C_j^+ &= max(0, C_{j-1}^+ + \hat{\epsilon}_{\hat{\mu}_2}(t) - k) \\ C_j^- &= min(0, C_{j-1}^- + \hat{\epsilon}_{\hat{\mu}_2}(t) + k) \end{aligned} \tag{6}$$

with $j \geq 1$, $C_0^+ = C_0^- = 0$.
The chart gives an alert if:

$$C_j^+ > h^+ \text{ or } C_j^- < h^-.$$

- How to fix the hyper-parameters of the chart ?
  To choose hyper-parameters of the chart (such as N_{IC}, K, Δ, Δ_1, etc.), the ARL_0 is fixed to a certain value. Then, the performance of the chart is evaluated with different chart designs and the most powerful design is selected. The ARL_1 criterion is often used as a performance measure. This criterion is however mainly sensitive to the shift size δ, to the threshold k and to ARL_0. If these parameters are kept constant, the value of ARL_1 will not always help us to select a design or to fix a hyper-parameter.

The ARL_1 appears as a good criterion but we may look for others in the literature. Note that the ARL_0 value is usually high (meaning that the rate of false positives is low) in most practical applications. This is related to the cost of interrupting a production process for a false alert, which is usually

large. Here however, we may have a higher rate of false positive as it does not cost anything to detect an alert. Giving to many alerts may however annoyed or discouraged the observers.

3.3 Estimation of the shift sizes

- How to estimate the shift sizes ?

The CUSUM chart is adapted to detect shifts of sizes $\delta = 2k$. To estimate the shift size, we may compute the difference between the mean of the IC residuals and the mean of the OC residuals:

$$\begin{aligned}\hat{\delta} &= \overline{\epsilon}_{\hat{\mu}_2}(i_{oc}, t) - \overline{\epsilon}_{\hat{\mu}_2}(i_{ic}, t) \\ \hat{\delta} &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_{OC}}\sum_{i_{oc}=1}^{N_{OC}}\epsilon_{\hat{\mu}_2}(i_{oc}, t) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_{IC}}\sum_{i_{ic}=1}^{N_{IC}}\epsilon_{\hat{\mu}_2}(i_{ic}, t)\end{aligned} \tag{7}$$

We obtain $\hat{\delta} = 0.75$ for $N_{IC} = 9$.
A lower band for $\hat{\delta}$ may be computed using Equation 7 with the most stable OC station (CA) and a upper band for $\hat{\delta}$ may be estimated using the least stable OC station (i.e. LO). They are respectively equal to $\hat{\delta}_l = 0.13$ and $\hat{\delta}_l = 3$.
From these values, we may define another estimator $\hat{\hat{\delta}}$ of δ:

$$\begin{aligned}\hat{\delta}_u &= \frac{1}{T}\sum_{t=1}^{T}\epsilon_{\hat{\mu}_2}(LO, t) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_{IC}}\sum_{i_{ic}}^{N_{IC}}\epsilon_{\hat{\mu}_2}(i_{ic}, t) \\ \hat{\delta}_l &= \frac{1}{T}\sum_{t=1}^{T}\epsilon_{\hat{\mu}_2}(CA, t) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_{IC}}\sum_{i_{ic}}^{N_{IC}}\epsilon_{\hat{\mu}_2}(i_{ic}, t) \\ \hat{\hat{\delta}} &= \frac{\hat{\delta}_l + \hat{\delta}_u}{2}\end{aligned} \tag{8}$$

where $\hat{\hat{\delta}} = 1.88$ for $N_{IC} = 9$.

3.4 Monitoring of the sunspot numbers

Given the complexity of our data, it may be relevant to distinguish three types of stations (not only IC and OC stations). A first set of 'truly' IC stations may be considered to estimate $\hat{\mu}_0(t)$ and $\hat{\sigma}_0^2(t)$. Then, a larger set of stations may be used to adjust the control limit of the chart. This second set may contain the 'truly' IC stations and some other stations, slightly more unstable.

This little change of design has a crucial impact of the performances of the chart. This new design may be used to render the chart less sensitive to the shifts (if needed in practice). Indeed, to reach the rate of false positive desired ($ARL_0 = 500$), we need to raise the control limits of the chart with respect to the simple design with only IC and OC stations.

- How to evaluate the ARL_0 of the chart?
  The ARL_0 is the IC average run length, i.e. the average time to an alert when the process is IC. It can be evaluated on the same set of IC stations that were used to compute $\hat{\mu}_0(t)$ and $\hat{\sigma}_0^2(t)$. Or it can be evaluated on a subset of the IC stations (overlapping or not overlapping with the initial set of IC stations).

Maybe interesting but requires much simulations.

- How to reset the chart after a missing value ?

In our case, long gaps usually correspond to periods of maintenance or changes of instrument/location. These gaps may therefore affect significantly the future observations. While short gaps are most probably caused by random weather conditions and are unlikely to impact the forthcoming observations. Therefore, the last value of the chart may be carried forward for small gaps, while the chart may be reset after long gaps. To conciliate both situations, the chart may slightly decrease each missing day. In a generic methodology, if the gap are completely random and unlikely to modify the observing procedure, the last value may simply be carried forward.

3.5 Suggestions/future works

- We may use the EWMA chart instead of the CUSUM.
- We may test if using a rectangular window instead of another kernel affects the $\mu_2(i,t)$
- We may also test the effect of using the mean or the median in the estimation of $\mu_0(t)$ and $\sigma_0(t)$.
- Another criterion such as the ARL_1 may be used to evaluate the performance of the chart
- We may investigate the relation between the chart and the block bootstrap in the literature

References

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

Qiu, P., Zi, X., and Zou, C. (2017). Nonparametric dynamic curve monitoring. *Technometrics*, 60(3):386–397. 2