Modélisation et estimation du nombre de taches solaires

Sophie MATHIEU[1], Rainer VON SACHS[1], Véronique DELOUILLE[2], Laure LEFÈVRE[2], Christian RITTER[1]

[1]Université catholique de Louvain, ISBA Louvain-la-Neuve, Belgique

[2]Observatoire Royal de Belgique, département de physique solaire
Bruxelles, Belgique

soph.mathieu@uclouvain.be, rainer.vonsachs@uclouvain.be
v.delouille@oma.be, laure.lefevre@oma.be, christian.ritter@uclouvain.be

Résumé – La série temporelle du nombre de taches solaires constitue la plus longue expérience scientifique encore en cours (400 ans). Elle est utilisée dans de nombreux modèles physiques comme indicateur de l'activité solaire. Cependant, le traitement statistique utilisé pour construire cette série n'est pas adapté à la nature complexe des données. Nous présentons ici un estimateur robuste pour le nombre de taches ainsi qu'un modèle pour sa densité qui tienne compte des caractéristiques inhérences à ces données, à savoir: la surdispersion, l'excès de zéros, et la présence de plusieurs modes. Nous proposons également une simulation du nombre de taches basée sur une approche statistique, afin de mieux comprendre la nature des données et l'effet des différentes sources de bruits sur les procédures d'estimation, et par la suite, sur les procédures de contrôle de qualité.

Abstract – The time series of the sunspot counts constitutes the longuest-running scientific experiment (400 years). It acts as a benchmark in a large variety of physical models and is used as a proxy for the solar activity. It lacks however a statistical processing suited to its complex nature. We present here a robust estimator for the sunspot counts and a model for its density that takes into account intrisic characheristics such as over-dispersion, excess of zeros, and multiple modes. We also develop a statistics-based simulation for the sunspot number that aims at a better understanding of the nature of the data, the effect of the noise on the estimation methods and (in future work) on quality control procedures.

1 Introduction

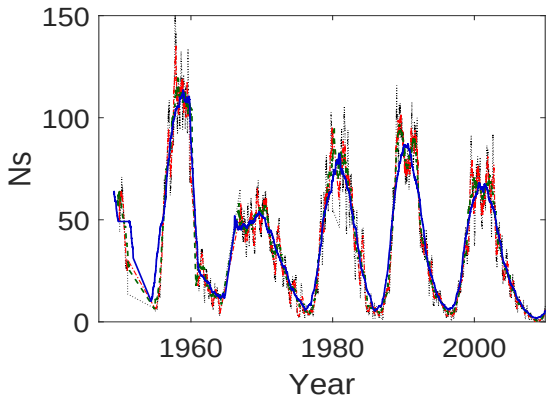

FIGURE 1 – Nombre de taches solaires observées dans la station Uccle, moyennées sur 81 jours (pointillés noirs), 162 jours (tirets-points rouges), sur un an (tirets verts) et 2.5 ans (ligne bleue).

On appelle 'taches solaires' des zones sombres visibles à la surface du soleil. Elles correspondent à des régions où le champ magnétique est localement élevé et sont associées à l'activité solaire, que ce soit à long terme (cycle de 11 ans [5]), ou à court-terme pour la météorologie de l'espace [4]. Ces taches sont observées en Occident depuis l'invention du téléscope au XVIIe siècle, et la série temporelle du nombre de taches qui en résulte constitue actuellement la plus longue expérience scientifique encore en cours [9]. Dans cette expérience, le nombre total de taches sur la face visible du soleil, dénoté N_s, est combiné chaque jour au nombre de groupes de taches, N_g, dans un composite $N_c := 10N_g + N_s$. Une région active apparait donc deux fois dans N_c car une tache isolée est aussi considérée comme un groupe. Le poids attribué au nombre de groupes N_g dans le composite N_c est lié au nombre moyen de taches par groupe observé par l'observateur Wolf [5]. Les N_cs fournis par un ensemble d'observatoires (aussi appelés 'stations') répartis à travers le monde sont ensuite combinés chaque mois afin d'obtenir le nombre de taches international, ou 'International Sunspot Number'(ISN). Depuis 1981, l'Observatoire Royal de Belgique (ORB) est responsable de la production et du contrôle de qualité de l'ISN. Cette série est largement utilisée en sciences naturelles, notamment dans les modèles climatiques pour modéliser le forçage solaire [4]. Fig.1 représente le nombre de taches observées par la station d'Uccle (ORB).

Les procédures pour calculer l'ISN ont évolué au cours du temps [3]. Cependant, la procédure actuelle est toujours basée sur des méthodes statistiques du XIXe siècle. L'ISN est calculé au moyen d'une moyenne robuste (rejet des observations > ou

< 2 σ) des observations individuelles de différentes stations, corrigées par rapport à une station de référence unique. Ce traitement ne tient pas compte de la nature complexe des données qui sont profondément non-gaussiennes et ont de nombreuses (environ 40%) valeurs manquantes, dues principalement aux conditions météorologiques qui empêchent les stations d'observer le soleil. De plus, la distribution de ces données montre une surdispersion par rapport à des données de comptage poissonniennes, des modes multiples et un excès de valeurs en zéro.

Notre contribution consiste à analyser une série temporelle complexe longue de 66 ans et contenant un panel de 21 stations individuelles : en section 2, nous expliquons le pré-traitement appliqué sur les données, ainsi que la nature des erreurs qui corrompent les observations. Un estimateur robuste pour le nombre de taches est présenté en section 3. En section 4, nous proposons un modèle pour la densité du signal estimé, qui tient compte de la nature complexe des données. Finalement la section 5 présente une approche statistique originale pour simuler une série temporelle qui reproduit les caractéristiques principales des données. Cette simulation permettra de mieux comprendre l'origine de la variabilité présente dans les données ainsi que l'impact de certaines procédures sur les observations.

2 Pré-traitement

Nous analysons la période du 1er janvier 1947 au 31 décembre 2013, comprenant le maximum du cycle 18, les cycles solaires 18 à 23 et la phase ascendante du cycle 24 [1], dans 21 stations réparties à travers le monde. En raison des différences intrinséques entre les stations (puissance des instruments, expérience des observateurs, ...) un pré-traitement est nécessaire pour aligner les stations au même niveau. Chaque station est donc multipliée par un facteur correctif évalué tous les huit mois. Ce facteur a été calculé en utilisant une régression par moindres carrés, entre le nombre de taches observées dans la station et la médiane journalière du réseau. La médiane est préférable à une simple moyenne du réseau car les observations sont très variables d'un jour à l'autre, mais aussi au cours d'une même journée. Une partie de la varibilité peut être expliquée par la composante solaire du signal, cf. section 5. Dans [7], nous avons mis en évidence une variabilité importante des stations, avec au moins trois types d'erreurs propres aux observateurs : une erreur à court-terme due au conditions d'observation et aux erreurs de comptage, un biais à long-terme et enfin une erreur spécifique due aux périodes de minima dans le cycle solaire, caractérisées soit par une absence de tache, soit par un petit nombre de taches de courte durée de vie.

1. https://en.wikipedia.org/wiki/List_of_solar_cycles

3 Estimation du signal solaire

Nous dénotons par $Y_i(t)$ le nombre de taches observées dans une station i, $1 \leq i \leq N$, au temps t exprimé en jours. Nous proposons d'utiliser une médiane *filtrée* du réseau de stations comme estimateur pour le nombre réel de taches solaires au temps t, $s(t)$. Appelons cet estimateur $\hat{\mu}_s(t)$:

$$\hat{\mu}_s(t) = \widetilde{\text{med}}_{1\leq i\leq N} Y_i(t) \tag{1}$$

où $\text{med}_{1\leq i\leq N} Y_i(t)$ représente la médiane journalière du réseau et où $\widetilde{med}$ représente la médiane filtrée, obtenue en deux étapes.

Dans la première étape, nous appliquons une transformation d'Anscombe généralisée [6] sur la médiane du réseau :

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

Cette étape est largement utilisée dans la littérature pour transformer des variables aléatoires (v.a.) de type Poisson en v.a. approximativement Gaussiennes. Nous fixons $\alpha = 4.2$ car dans [3] nous trouvons que cette valeur minimise la variance de N_c. La deuxième étape est le filtrage à proprement parler, qui se décompose comme suit. Nous appliquons tout d'abord une FFT sur la médiane transformée. Ensuite, nous atténuons une partie du signal en utilisant un filtre de Wiener (fonction en escalier) choisi pour annuler les amplitudes des fréquences qui correspondent à des périodes inférieures à 27 jours (i.e. une rotation solaire).

Finalement, nous appliquons une FFT inverse et une transformation d'Anscombe inverse pour obtenir la médiane filtrée $\hat{\mu}_s(t)$.

4 Distribution du signal solaire estimé

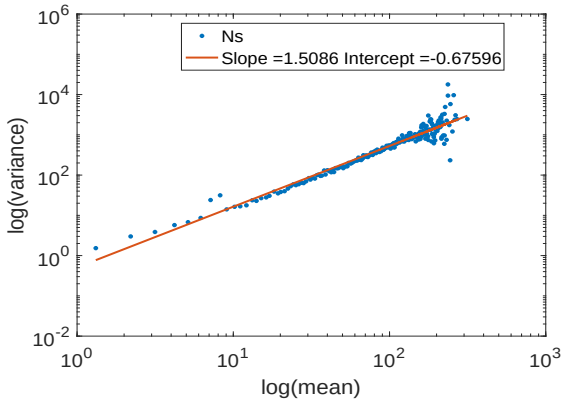

FIGURE 2 – Estimation de la variance du nombre de taches en fonction de la moyenne dans un graphe log-log.

Dans une distribution de Poisson, typique des comptages, la moyenne est égale à la variance. Afin de vérifier cette relation pour le comptage du nombre de taches solaires, nous

avons calculé la variance conditionnelle $\mathrm{Var}(Y_i(t)|\hat{\mu}_s(t) = \mu)$, cfr. Fig.2. Sur le graphique, la pente estimée est proche de 1.5, ce qui indique une surdispersion par rapport à une variable de Poisson. Une telle surdispersion est généralement modélisée par des variables de type binomiale négative [2].

Une inspection visuelle de la distribution des valeurs estimées de $\hat{\mu}_s(t)$ indique également un excès de valeurs nulles et la présence de plusieurs modes. Pour correctement modéliser l'excès de zéros, il est approprié de modéliser la distribution de $\hat{\mu}_s(t)$ par un modèle de Hurdle [2] :

$$f(x) = \begin{cases} f_0(0) & \text{if } x = 0 \\ (1 - f_0(0))\frac{f_1(x)}{1-f_1(0)} & \text{if } x > 0 \end{cases} \quad (3)$$

où les valeurs nulles sont représentées par une distribution de Bernoulli $f_0(x)$ et où les valeurs différentes de zéro sont modélisées par une seconde densité $f_1(x)$, définie par rapport à une mesure discrète. En raisons des modes multiples, de la surdispersion et de la nature de données de comptage, nous proposons d'identifier $f_1(x)$ à une mixture de distributions binomiales négatives généralisées :

$$\begin{aligned} f_1(x, r_1, p_1, r_2, p_2, \lambda) =& \lambda \frac{\Gamma(r_1 + x)}{\Gamma(r_1)\Gamma(x + 1)} p_1^{r_1} q_1^x \\ &+ (1 - \lambda) \frac{\Gamma(r_2 + x)}{\Gamma(r_2)\Gamma(x + 1)} p_2^{r_2} q_2^x \end{aligned} \quad (4)$$

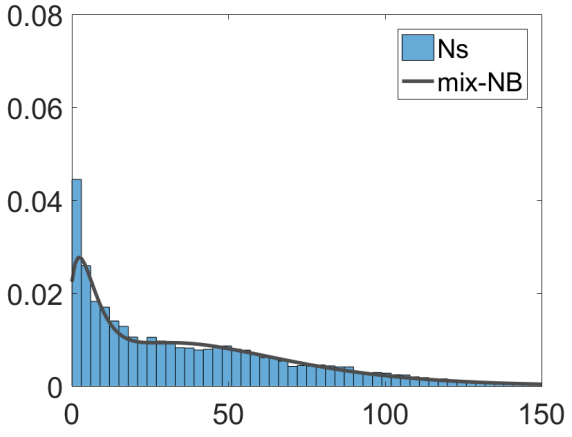

FIGURE 3 – Histogramme de $\hat{\mu}_s(t)$. La ligne noire représente l'estimation de la densité telle que décrite par l'équation 4. Les valeurs des paramètres estimés par maximum de vraisemblance sont $r_1 = 1.93$, $p_1 = 0.20$, $r_2 = 2.86$, $p_2 = 0.04$, $\lambda = 0.32$ and $f_0(0) = 0.06$.

L'histogramme de $\hat{\mu}_s(t)$ ainsi que son modèle estimé sont représentés sur la Fig.4. L'estimation de la distribution a été réalisée par la méthode du maximum de vraisemblance. La distribution des valeurs estimées $\hat{\mu}_s(t)$ contient toujours des caractéristiques discrètes, telles que l'excès de valeurs en zéro, et ce malgré le pré-traitement et la procédure de filtrage qui devraient pourtant rendre ces données continues. Pour reproduire cette persistence de propriétés discrètes, même après filtrage, nous proposons une approche statistique pour simuler les données.

5 Simulation

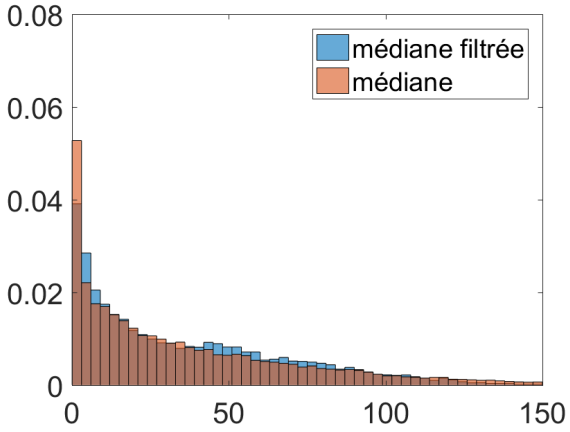

FIGURE 4 – Histogramme de la médiane du réseau $\mathrm{med}_{1\leq i\leq N} Y_i(t)$ et de la médiane filtrée $\widetilde{med}_{1\leq i\leq N} Y_i(t)$ obtenu en utilisant les données simulées.

Pour mieux comprendre la nature des données et l'origine de la variabilité journalière, nous proposons de simuler le mécanisme d'émergence des taches et des groupes de taches sur la surface du soleil de manière statistique (en utilisant des distributions telles que la Poisson et la Binomiale), par opposition à une approche basée sur la physique. L'algorithme est facile à comprendre et permet de simuler rapidement des données qui ont les mêmes caractéristiques statistiques que les observations.

L'idée centrale de notre simulation est de séparer le soleil en 27 secteurs, (une rotation solaire=27 jours) et de compter l'ensemble des taches apparaissant sur la face visible du soleil (égale à 11 secteurs car les observateurs ne comptent pas les taches qui apparaissent à la limite du disque). Chaque jour un nouveau secteur apparait sur la face visible du soleil et un autre secteur disparait.

De nouveaux groupes (s'additionnant aux groupes survivants des rotations précédentes) sont créés sur le secteur, en suivant une loi de Poisson de moyenne égale à l'intensité solaire : $Ng_{\text{eff}} \sim \mathcal{P}(\widetilde{s}(t))$. En pratique $\widetilde{s}(t)$ est une version hautement filtrée de la médiane du réseau appliquée sur le nombres de groupes observés N_g, et indique le niveau d'activité solaire.

Le nombre total de taches est ensuite simulé suivant une loi Binomiale négative dépendant de Ng_{eff} et du nombre moyen de taches par groupe (égal à 7 sur 1947-2013) : $Ns_{\text{tot}} \sim \mathcal{NB}(7Ng_{\text{eff}}, 0.25)$. Ces taches sont distribuées selon une loi uniforme aux groupes Ng_{eff}. Elles représentent le contenu maximal en taches Ns_{max} que les nouveaux groupes du secteur peuvent contenir. Ce contenu ne sera pas modifié avant

que le secteur ne réapparaisse sur la face visible du soleil. Notons que les groupes sont créés avant les taches car leurs mécanismes de formation sont dépendants.

Le contenu effectif des groupes (en taches) Ns_{eff} évolue de jour en jour. Une fois que de nouveaux groupes sont formés, ils grandissent d'abord pendant quelques jours (de deux à six en fonction de la taille du groupe [8]) avant de progressivement disparaître pendant le reste de la rotation solaire, ou plus longtemps s'ils survivent. En effet, les taches se forment plus rapidement (i.e. en quelques jours) qu'elles ne s'évanouissent (en moyenne quelques mois), cfr [8].

Initialement le contenu effectif d'un groupe Ns_{eff} est tiré aléatoirement selon une loi Binomiale $(Ns_{\text{eff}} \sim \mathcal{B}_i(Ns_{max}, 0.5))$. Ensuite, le nombre de taches grandit dans les groupes selon une Binomiale négative $(\mathcal{NB}_i(Ns_{max}, 0.45))$, sans jamais pouvoir excéder Ns_{max}. Enfin, les taches disparaissent lentement suivant une Binomiale $(\mathcal{B}_i(Ns_{\text{eff}}, 0.95))$. L'algorithme simule donc chaque jour le contenu effectif (et potentiel) en taches et groupes des 27 secteurs. Le nombre total de taches $(N'_s(t))$ peut donc être obtenu en additionnant chaque jour le nombre de taches effectif des 11 secteurs visibles du soleil. On simule enfin la variabilité associée aux stations à l'aide d'une Binomiale dépendant du nombre de taches simulées $Ns'(t)$ corrompues par du bruit log-normal $\mathcal{B}_i(N'_s(t)\ \exp(0.3\text{Norm}(0,1)), p < 1)$. La probabilité d'observer des taches est légèrement inférieure à un, car les stations peuvent ne pas voir toutes les taches.

Cette simulation ne vise pas à modéliser des phénomènes physiques mais à mieux appréhender la nature des données et des traitements qui y sont appliqués. Elle contient plusieurs paramètres fixés de façon arbitraire et/ou ajustés sur les observations et idéalise les stations qui n'ont par exemple pas de dérives à long-terme (même si elle pourrait être raffinée à l'avenir en utilisant les résultats de [7]). Cependant, elle permet de comprendre l'importante variabilité journalière entre les stations : les observations sont réparties sur une plage de 24h et il existe des erreurs d'observation. Un estimateur filtré semble donc plus approprié pour estimer le nombre réel de taches solaires.

Elle permet également d'étudier la procédure de filtrage. En effet, en appliquant une médiane filtrée sur les observations on observe un excès de valeurs aux alentours de $N_s = 50$ (fig.4), qui semble absent lorsqu'on utilise la médiane simple. Nous avons donc simulé des données pour voir si ce plateau était réel : cet excès autour de 50 apparaît presque à chaque fois même si l'emplacement exact ainsi que la forme du plateau (plus ou moins étendu) varie d'une simulation à l'autre (fig. 4).

Cette structure est présente dans le signal d'entrée de la simulation $(\widetilde{s}(t))$ et semble être masquée par la variabilité importante présente dans la médiane (simple) du réseau, et au contraire mise en évidence par le filtrage. Notons que, au niveau physique, cette structure pourrait être liée à différents régimes de croissance des taches, c'est à dire au passage des petites taches sans pénombres aux taches plus développées avec pénombre [1].

Dans le futur, la simulation pourrait nous permettre de vérifier qu'un pré-traitement n'induit pas de biais, d'analyser l'impact des valeurs manquantes sur la qualité de notre estimateur, ou encore d'analyser l'effet de procédures de contrôle de qualité.

6 Conclusion

Nous proposons une méthode originale pour analyser le nombre de taches solaires observées en tenant compte des caractéristiques complexes de ces données : la surdispersion, l'excès de zéros et les modes multiples. Nous considérons ici une période de 66 ans, mais notre méthodologie est applicable sur la série complète du nombre de taches s'étalant sur plus de 400 ans. Elle servira de base pour corriger à l'avenir la procédure statistique qui définit l'ISN [2]. Enfin, une simulation simple nous permet de mieux comprendre la nature des données et nous donne un outil puissant pour produire des données semblables aux observations. Elle nous permettra de tester les futures procédures de contrôle de qualité.

Références

[1] F. Clette and L. Lefèvre. The New Sunspot Number : Assembling All Corrections. *Solar Physics*, 291 :2629–2651, 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] K. Ermolli, K. Matthes, T. Dudok de Wit, N.A. Krivova, K. Tourpali, and all. Recent variability of the solar spectral irrandiance and its impact on climate modelling. *EGU Publication : Atmospheric Chemistry and Physics*, 13 :3945–3977, 2013.

[5] A.J. Izenman. J.R Wolf and the Zurich sunspot relative numbers. *The Mathematical Intelligencer*, 7 :27–33, 1985.

[6] M. Makitalo and A. Foi. Optimal Inversion of the Generalized Anscombe Transformation for Poisson-Gaussian Noise. *IEEE Transactions on Image Processing*, 22(1), 2013.

[7] S. Mathieu, R. von Sach, V. Delouille, and L. Lefevre. Uncertainty Quantification in Sunspot Counts. *IEEE DSW2018*, 2018.

[8] J. Murakozy, T. Baranyi, and A. Ludmany. *Sunspot Group Development in High Temporal Resolution*. Springer, 2014.

[9] B. Owens. Long-term research : Slow science. *Nature*, 495(7441) :300, March 2013.

2. Dans le cadre du projet BRAIN.be VAL-U-SUN, `www.sidc.be/valusun`