

# UN MODÈLE À HASARD PROPORTIONNEL POUR LA PRÉVISION DU RISQUE D'OCCURRENCE DE TEMPÊTES MAGNÉTIQUES EXTRÊMES

Malika Chassan <sup>1</sup> & Jean-Marc Azaïs <sup>2</sup> & Guillaume Buscarlet <sup>3</sup> & Norbert Suard <sup>4</sup>

<sup>1</sup> *Institut de Mathématiques de Toulouse - Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, Cedex 9, France.*

*E-mail : malika.chassan@math.univ-toulouse.fr*

<sup>2</sup> *Institut de Mathématiques de Toulouse*

*E-mail : jean-marc.azais@math.univ-toulouse.fr*

<sup>3</sup> *Thales Alenia Space, 26 Avenue Jean François Champollion, 31100 Toulouse, France*

<sup>4</sup> *CNES, 18 Avenue Édouard Belin, 31400 Toulouse, France*

**Résumé.** Cet article s'intéresse à l'intensité d'occurrence des orages magnétiques de magnitude extrême. Les perturbations ionosphériques engendrées par de tels orages en font des événements redoutés par les systèmes de navigation par satellite et une estimation précise de leur fréquence est nécessaire. Pour cela, nous avons mis en place un modèle inspiré à la fois du modèle à hasard proportionnel de Cox et de la Théorie des Valeurs Extrêmes (TVE). Le nombre d'orages par cycle solaire est modélisé par un processus de Poisson non-homogène. L'intensité de ce processus est exprimée comme le produit d'un risque de base et d'un facteur de risque. Contrairement au modèle de Cox classique, le risque de base est le paramètre d'intérêt : il correspond à l'intensité d'occurrence au cours d'un cycle solaire d'intensité moyenne et permet de prendre en compte la non-stationnarité du phénomène étudié. Le facteur de risque prend en compte l'indice d'activité solaire du cycle. Les orages extrêmes sont trop rares pour permettre une estimation précise des paramètres du modèle. Nous utilisons donc les orages de niveau haut (et pas seulement extrême) pour estimer les paramètres. Ensuite, comme dans la méthode des excès (ou Peaks Over Threshold), l'intensité estimée est extrapolée aux tempêtes extrêmes. Le résultat majeur de cette étude est de mettre en évidence une forte corrélation entre l'intensité d'apparition d'orages extrêmes et l'instant d'apparition dans le cycle solaire. Le modèle et les résultats s'avèrent robustes. Nous proposerons aussi une prévision de l'intensité d'occurrence pour le cycle solaire actuel.

**Mots-clés.** Événements extrêmes, Hasard proportionnel, Processus de Poisson non-homogène.

**Abstract.** The aim of this paper is to give a precise estimation of the extreme magnetic storms frequency per time unit (year) throughout a solar cycle. An innovative approach based on a proportional hazard model is developed. Based on the Cox model, this method includes non-stationarity and covariate influence. The model assumes that

the number of storms during a cycle is a non-homogeneous Poisson process. The intensity of this process can be expressed as the product of a baseline risk and a risk factor. In the Cox model, the baseline risk is a nuisance parameter. In our model, it is a parameter of interest that will be estimated. The risk factor depends on a covariate, the solar activity index. As in Extreme Value Theory (EVT) and especially in Peaks Over Threshold (POT) modeling, all the high level events are used to make estimations and the results are extrapolated to the extreme level events. This study highlights a strong correlation between the occurrence intensity of magnetic storms and their position on the solar cycle. The model can be used to forecast occurrence intensity for the current 24th solar cycle.

**Keywords.** Extreme events, Proportional hazard, Non-homogeneous Poisson process.

## 1 Introduction

Les couches hautes de l'atmosphère terrestre sont soumises aux rayonnements solaires et ionisées par ces derniers, formant ainsi la ionosphère (de 85 km à 800 km d'altitude environ). Au-delà de la ionosphère, se situe la magnétosphère, zone contenant le champ magnétique terrestre. Lors de fortes tempêtes solaires ou d'importantes CME (Éjection de Masse Coronale), les vents solaires ne sont pas totalement déviés par le champ magnétique et peuvent perturber les courants électriques de la ionosphère. Lorsque celle-ci est fortement perturbée, on parle d'orages ou de tempêtes ionosphériques (ou géomagnétiques).

Ces orages perturbent la propagation des ondes radioélectriques, impactant ainsi un grand nombre de systèmes de télécommunication, notamment les systèmes de positionnement par satellite. Une bonne estimation de la fréquence d'apparition de tels orages est donc nécessaire. Cet article s'intéresse plus particulièrement aux orages d'intensité extrême. Le but est donc d'estimer la fréquence (ou intensité) par unité de temps (année) des orages ionosphériques extrêmes. De nombreux jeux de données permettent d'étudier l'activité géomagnétique de la ionosphère. Nous en avons choisi deux en particulier, l'indice *ap* et l'indice *aa*. Ces deux indices présentent l'avantage d'être disponibles depuis longtemps (1868 pour *aa*) et ne sont pas restreints à certaines zones du globe.

L'activité géomagnétique est en lien direct avec l'activité solaire. Ainsi, on peut observer des cycles dans l'activité de la ionosphère qui correspondent aux cycles solaires (cycles d'environ 11 ans). Ces cycles sont aussi observables dans les deux jeux de données sélectionnés. La méthode proposée doit donc prendre en compte cette saisonnalité pour que l'intensité estimée puisse être une fonction du temps. Les orages magnétiques extrêmes étant rares, la Théorie des Valeurs Extrêmes (TVE) semblait adaptée à la résolution de notre problème. Mais la nature des observations (variables discrètes bornées) ne permet pas d'appliquer directement la TVE. Ainsi, nous proposons un modèle s'inspirant à la fois de la TVE et du modèle à hasard proportionnel de Cox. Le modèle à hasard proportionnel permet d'obtenir une intensité dépendant du temps et de prendre en compte

l'influence de covariables. Comme pour la TVE, et plus particulièrement la méthode des excès (Peaks Over Threshold), nous utilisons les événements de niveau haut (moins rares) pour estimer les paramètres du modèle et extrapolons ensuite les résultats aux événements extrêmes.

Dans la section 2, nous présentons plus en détail les données disponibles. La section 3 est dédiée à la présentation du modèle et des modifications apportées au modèle de Cox. Les estimateurs des paramètres du modèle sont présentés en section 4 et les résultats en section 5.

## 2 Présentation des données

Plusieurs indices géomagnétiques sont disponibles pour étudier le niveau d'activité de la ionosphère. Comme l'a montré Rifa (2011), tous ces indices sont fortement corrélés. Ainsi, dans cet article, deux indices seront utilisés : l'indice *ap* (pour "planetary amplitude") et l'indice *aa* (pour "antipodal amplitude"). Ces deux indices sont calculés à partir de l'indice *K* (variation maximale du champ magnétique (en nano-Tesla) sur une période de 3 heures). L'indice *K* est sur une échelle logarithmique, il est converti en indice *a* dont l'échelle est linéaire. L'indice *ap* est une moyenne pondérée des indices *a* fournis par 13 observatoires répartis sur le globe, tandis que l'indice *aa* est calculé à partir des mesures de deux observatoires (en Angleterre et Australie). Nous avons choisi l'indice *ap* pour son caractère global et la taille du jeu de données disponible (depuis 1932, soit 7 cycles solaires). L'indice *aa* est quant à lui moins global mais couvre 12 cycles.

Le principal obstacle à une application directe de la TVE réside dans le fait que chaque valeur de *ap* est un entier de l'ensemble  $\{0, 2, 3, 4, 5, 6, 7, 9, 12, 15, 18, 22, 27, 32, 39, 48, 56, 67, 80, 94, 111, 132, 154, 179, 207, 236, 300, 400\}$ . Bien que dans ses travaux, Anderson (1970) étende certains résultats de convergence à des variables discrètes, il suppose que le support de la distribution est non borné, ce qui n'est pas le cas ici. L'indice *aa* est lui aussi une variable à support discret borné.

Nous avons donc dû proposer une nouvelle méthode adaptée aux jeux de données disponibles et permettant d'estimer la fréquence par unité de temps d'orages extrêmes. Le modèle développé est décrit dans la prochaine section.

## 3 Description du modèle

En analyse de survie, la probabilité d'observer l'événement d'intérêt au temps  $t$ , appelée le risque instantané  $\lambda(t)$ , est définie en utilisant le temps de survie  $T$  (temps d'occurrence de l'événement) par la probabilité d'observer cet événement dans l'intervalle infinitésimal  $[t, t + dt]$  :

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{1}{dt} \mathbb{P}(t \leq T < t + dt | T \geq t), \quad (1)$$

où  $\mathbb{P}(A)$  est la probabilité de l'événement  $A$ .

Notre modèle s'inspire du modèle de Cox où le risque instantané  $\lambda(t)$  s'exprime comme le produit d'un risque de base et d'un facteur de risque géant l'influence de covariables  $(X_1, \dots, X_p)$  :

$$\lambda(t, X_1, \dots, X_p) = \lambda_0(t) \exp\left(\sum_{i=1}^p \beta_i X_i\right),$$

Dans notre modèle, l'événement d'intérêt est l'occurrence d'une tempête et les individus sont les cycles solaires complets. Nous considérons une seule covariable, l'indice d'activité solaire du cycle (notée  $X_j$  pour le cycle  $j$ ). Le paramètre  $\beta$  quantifie son influence. Les différences entre notre modèle et le modèle de Cox sont les suivantes :

- la variable d'intérêt est  $N(t)$ , le nombre d'orages à l'instant  $t$  et l'on suppose qu'il suit un processus de Poisson non-homogène ;

- les observations ne sont pas censurées (seuls les cycles complets sont considérés) ;

- la variable  $D_j$  (longueur du cycle  $j$  en années) est incluse en facteur, ainsi l'unité de mesure est le nombre d'orages par an et non par cycle ;

- $\lambda_0(t)$  n'est pas considéré comme un paramètre de nuisance mais comme le paramètre d'intérêt à estimer ;

- les estimations sont faites en considérant tous des événements de niveau haut et une extrapolation aux événements de niveau extrême ( $ap = 400$ ) est appliquée grâce au paramètre  $P_{400}$ , la probabilité qu'un orage de niveau haut devienne un orage extrême. On suppose donc que le niveau atteint par un orage est indépendant de l'instant d'apparition. Un test du Khi-2 montre que cette hypothèse est acceptable.

On modélise donc  $N_j(t)$ , le nombre d'orage à l'instant  $t$  pour le cycle  $j$  par un processus de Poisson non-homogène d'intensité  $\lambda_j(t)$  telle que :

$$\lambda_j(t) = \lambda_0(t) D_j \exp(\beta X_j),$$

ce que l'on peut aussi écrire :

$$N_j([a, b]) \sim \mathcal{P} \left( \int_a^b \lambda_j(t) dt \right),$$

avec  $\mathcal{P}(\xi)$  la loi de Poisson de paramètre  $\xi$ .

Dans tout l'article,  $\lambda_0(t)$  représente l'intensité de base pour les orages de niveau haut,  $\lambda_j(t)$  correspond au risque instantané prenant en compte le facteur de risque  $D_j \exp(\beta X_j)$  (cycle  $j$ ). L'intensité extrapolée aux orages extrêmes (niveau 400) sera notée  $\lambda_{400,j}(t)$ , i.e.  $\lambda_{400,j}(t) = \lambda_j(t) \times P_{400}$ . Lorsque les indices  $j$  ne sont pas notés, c'est que l'on considère un cycle théorique, d'activité solaire moyenne.

### 3.1 Pré-traitements des données

Afin de définir précisément les objets manipulés (orages, niveaux haut et extrême) nous avons appliqué des pré-traitements aux jeux de données. Premièrement, un processus de "declustering", afin de ne considérer qu'un seul événement lorsqu'une tempête présente plusieurs périodes de forte activité séparées par des périodes plus calmes. Ce processus impose de fixer deux paramètres : un niveau bas (le niveau au-dessus duquel on considère qu'un orage est de niveau haut) et une longueur  $r$  (le nombre maximal de mesures entre deux excès du niveau bas pour qu'ils appartiennent au même orage).

Deuxièmement, une renormalisation temporelle ("time warping") dans le but de pouvoir analyser tous les cycles solaires ensemble (chacun étant de durée différente). La position d'un orage dans le cycle est représentée par un réel de  $[-0.5, 0.5]$  où  $-0.5$  correspond au début du cycle,  $0.5$  à la fin et  $0$  au milieu du cycle (pic de l'activité solaire).

## 4 Estimation

$P_{400}$  étant indépendant de la position dans le cycle, on utilise la fréquence empirique :

$$\widehat{P}_{400} = \frac{\#\{\text{orages de niveau 400}\}}{\#\{\text{orages de niveau haut}\}}.$$

Pour  $\beta$ , on utilise le fait que :

$$N_j = N_j([-0.5, 0.5]) \sim \mathcal{P} \left( \left[ \int_{-1/2}^{1/2} \lambda_0(s) ds \right] D_j \exp(\beta X_j) \right).$$

Comme dans le modèle de Cox classique, on peut vérifier que  $N_j = N_j([-0.5, 0.5])$  est une statistique exhaustive pour  $\beta$ .  $\beta$  est estimé par maximum de vraisemblance dans le modèle linéaire généralisé Poissonien.

Pour l'intensité de base  $\lambda_0(t)$ , on utilise un estimateur à noyau Gaussien. La fenêtre optimale est déterminée par validation croisée. On fournit un intervalle de confiance pour chacun des estimateurs. La qualité des estimateurs de  $\beta$  et  $\lambda_0(t)$  est testée par simulation.

## 5 Résultats

La figure 1 montre l'intensité estimée pour les orages extrêmes avec le jeu de données *ap*.

Figure 1: Intensité instantanée estimée pour les orages extrêmes par extrapolation avec IC, pour un cycle d'activité solaire moyenne. La ligne horizontale correspond à la fréquence empirique des orages extrêmes.

Le processus de declustering demande de fixer un niveau bas. La sensibilité du modèle à un changement de niveau bas est testée pour différents valeurs du paramètre et s'avère très faible. Nous appliquons notre modèle à un autre jeu de données, l'indice  $aa$  et cette fois encore, les résultats sont stables.

Enfin, nous utilisons notre modèle pour donner une prévision de l'intensité des orages magnétiques extrêmes pour le cycle solaire actuel (24<sup>ème</sup>).

## 6 Conclusion

Cette étude met en évidence la forte corrélation entre la probabilité d'occurrence d'un orage extrême et la position dans le cycle solaire. L'influence de l'indice d'activité du cycle est quantifiable grâce au facteur de risque  $D_j \exp(\beta X_j)$ . De part sa flexibilité, notre modèle peut aussi être utilisé pour fournir des prédictions d'intensité pour un cycle solaire quelconque et pour les orages de différentes intensités.

## Bibliographie

- [1] Anderson, C.W. (1970), *Extreme value theory for a class of discrete distributions with applications to some stochastic processes*, J. Appl. Probability, 7:99-113.
- [2] Rifa, E. (2011), *Études des relations entre indices solaires et géomagnétiques*, Intership report.
- [3] Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, London.
- [4] Aalen, O.O. and Borgan, Ø. and Gjessing, H.F. (2008), *Survival and event history analysis. A process point of view*, Springer, New York.