A proportional hazard model for the estimation of ionosphere storm occurrence risk

Malika Chassan, Jean-Marc Azaïs, Guillaume Buscarlet, Norbert Suard

A proportional hazard model for the estimation of ionosphere storm occurrence risk

Malika Chassan[1,2] , Jean-Marc Azaïs[1],
Guillaume Buscarlet[3], Norbert Suard[2]
([1]) Institut de Mathématiques de Toulouse, Université Toulouse 3, France
([2]) CNES, Toulouse, France
([3]) Thales Alenia Space, Toulouse, France
malika.chassan@math.univ-toulouse.fr

1 Introduction

Severe magnetic storms are feared events for integrity and continuity of GPS-EGNOS navigation system and an accurate modeling of this phenomena is necessary. Our aim is to estimate the intensity of apparition of extreme magnetic storm per time unit (year).

Our data set, retrieved from [1], consists of 80 years of registration of the so called 3 hours ap index (for "planetary amplitude"). The ap index quantifies the intensity of planetary geomagnetic activity, using data from 13 observatories. Although the equatorial region is not covered by these 13 observatories, they are spread all over the earth and the coverage of the ap index is rather global. The ap index is the linear transformation of the quasi log-scale index Kp, with the same sampling step of 3 hours. The Kp index, and hence the ap index, corresponds to a maximal variation of the magnetic field over a 3 hours period. See [1] for more details on geomagnetic indices. The ap index is available from 1932 to present but for our analysis we will use only the 7 complete solar cycles of the data set, from the 17th (on the general list) which starts on September 1933, to the 23th which ends on December 2008.

There are other data available for the study of the ionosphere magnetic activity, each of them with advantages and disadvantages:

- the aa index (for "antipodal amplitude"). Although this index is available since 1868, it is calculated from only two nearly antipodal geomagnetic stations in England and Australia. Thus, this indice does not take into account all the magnetic activity of the ionosphere.

- the Dst (Disturbance storm time). This index is restricted to the equatorial magnetic perturbation (see Figure 1). Moreover, there are only 57 years of registration available against 80 for the ap index. Nonetheless, this indice gets the advantage to be an unbounded integer contrary to the ap index which lies in a finite set of non consecutive positive integers (see Section 2).

- the raw geomagnetic data are also available for many geomagnetic observatories. Oldest observations date back to 1883 for hourly values and to 1969 for 1 minute values. They consist of the measure by magnetometers of magnetic field variations. The disadvantage of this data is the presence of gaps in the

recording (with gap lengths varying from one month to several years depending on the observatory). The principal disadvantage of these data is the quantity of pre-treatments required.

Since all these index show a strong correlation rate [11], we chose to use only one indice to make our analyses. We opted for the ap index. The main advantage of this data set is the large amount of data. Moreover, there is no gap in the ap index, contrary to raw geomagnetic data. Finally, the ap index is more global than aa index or Dst, as one can see on Figure 1.

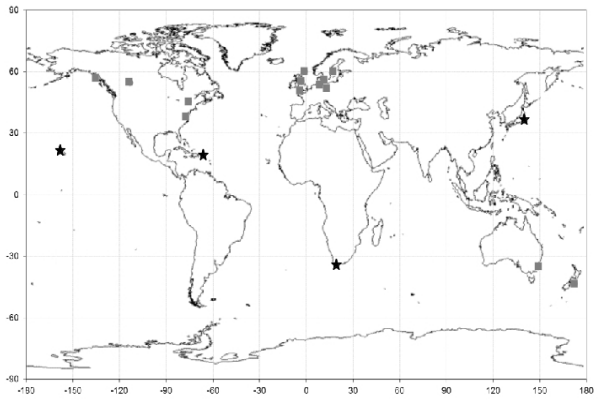

Figure 1: Positions of the observatories for the Dst (★) and the Kp/ap indices (■).

Intensive storms being scarce, classical statistical methods for probability estimation, as empirical frequency, are not precise enough. In many domains, the Extreme Value Theory (EVT) enables to estimate the probability of scarce extreme events. But, because of, among other things, the finite discrete form of our data and the obvious non stationary behavior, the EVT cannot be applied.

In Section 2, we develop the arguments showing that a use of classical EVT is not achievable. In Section 3, we describe our new proportional hazard model. This is the main contribution of this paper. The description of parameter estimators could be found in Section 4. The Section 5 is dedicated to the presentation of applications to our data set.

2 Difficulties to directly apply EVT

As said before, the first obstacle to direct application of EVT is the bounded discrete form of our data. The ap index varies in the set {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}. The application of Extreme Value Theory assumes the continuity of the probability distribution and it is well known that EVT does not apply to discrete finite observations, see for example [5].

The fact that finite discrete data do not enter in the scope of the theory is not the only issue. Indeed, in case of peaks over threshold modeling, one has to choose a threshold. The choice of the optimal threshold is made analyzing the

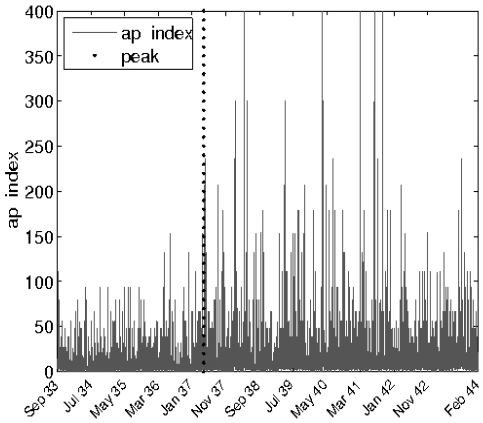

Figure 2: The ap index during the first complete cycle (17^{th} on the general list, from September 1933 to February 1944). The dotted vertical line represents the peak.

behavior of the parameters according to a threshold variation. This is, generally speaking, not possible with discrete data. For example, see the work of Cooley, Nychka and Naveau [8], where the low precision of the measure makes data almost discrete. Here, when the threshold grows up, one can observe a sawtooth behavior of parameters estimators and this makes the threshold selection troublesome.

A second problem is that ap index data obviously show a non stationary pattern. It is well known that the sun activity follows cycles with a duration of about 11 years. Corresponding cycles are observable in ap index behavior and must be taken into account for model assessment. This behavior implies that the probability of a magnetic storm occurrence depends on the position into the cycle. See Figure 2, for example. One can see the first complete solar cycle of the data set. Its middle is indicated by a vertical dotted line. One obviously remarks that strong storms (characterized by a high ap index level) occur principally during the second half of the cycle. Thus, it is not realistic to model this behavior by a standard stationary extreme value model (e.g. with constant parameters). A more efficient approach will be to include non-stationarity in parameters estimation. But once again, for this type of processes, there is no general theory allowing such a modeling.

In various research fields like hydrology, non stationary extreme models are proposed. For example, see the work of Jonathan and Ewans [10]. In this paper, authors want to model the seasonality of extreme waves in the gulf of Mexico. Occurrence rate and intensity of storm peak events vary with season. To model this seasonal effect, the authors have chosen to express the Generalized Pareto parameters as a function of seasonal degree using a Fourier form. But this approach supposes that the classical EVT can be applied, and this is not the case with the data set used in this paper.

3 Model description

In this section, we give a precise definition of what we call a storm, describe data and pretreatments (mostly declustering and time warping). We also describe the model we built and its advantages.

3.1 Storm definition, declustering

Ionospheric perturbations are classified in a standardized way using ap index, according to the Table 1:

Ionosphere Condition	**Kp-index**	**ap index**
Quiet	0-1	<7
Unsettled	2	7 to <15
Active	3	15 to <27
Minor storm	4	27 to <48
Major storm	5	48 to <80
Severe storm	6	80 to <140
Large severe	7	140 to <240
Extreme	8	240 to <400
Extreme	9	≥ 400

Table 1: Relation between Kp, ap and ionosphere activity

We introduce a declustering process of the data in order to consider only one event with the highest intensity even through there are different periods of high intensity separated by less active ones (lower indices). See Chapter 5.3 in [7] for example. This so called Runs Declustering process allows to precisely define what we consider as a storm. We have to set two parameters:

- a *low level*, the threshold above which we consider that a storm begins (typically 111, 132 or 154);

- the run length r, the minimal number of observations below the low level between two events for them to be consider independent.

Thus, two exceedances of the low level separated by less than r measures will be consider to belong to the same cluster (same storm).

Then, for each cluster, we define the storm level as the maximal level reached in the cluster. The first time when this maximum is attained is also saved, it represents the storm date. For a cluster, we define the length of the storm as the number of observations between the first up-crossing and the last down-crossing of the low level.

Durations of magnetic storms are very variable, from 3 or 6 hours for an extreme storm (level 300 or 400) until 90 hours for a low level storm. But, due to this declustering, we consider only one time event (since only the first maximum occurrence time is saved). This is not incoherent since we focus on strong storms, which are brief compared to lower storms but it should be take into account for the probability of occurrence definition.

3.2 Precisions on probability of occurrence

As said before, a storm is now defined by three values: the maximal level, the first time when this maximum is attained and the length of the cluster. This modeling allows to estimate the probability:

$$P_1(t) = \mathbb{P}(\text{a storm of level 400 } \textbf{begins} \text{ at time t})$$

And we want to know the probability:

$$P_2(t) = \mathbb{P}(\text{a storm of level 400 } \textbf{is ongoing} \text{ at time t})$$

In the whole data set, the level 400 is reached 29 times, but only 23 storms of level 400 are counted after the declustering. Among these 23 storms, 17 reach the level 400 only one time and 6 remain at this level two consecutive times.

Hence, we can say that:

$$\begin{aligned} P_2(t) &= P_1(t) + P_1(t-1) \times \mathbb{P}(\text{storm stays at the level 400 two times}) \\ &\simeq P_1(t) \times (1 + \mathbb{P}(\text{storm stays at the level 400 two times})) \\ &\simeq P_1(t) \times (1 + 6/23) \end{aligned}$$

3.3 Data description

After the declustering there are only 23 magnetic storms of level 400. There are not enough individuals to estimate their frequency as a function of the covariates. For the storms of level 300, one counts 44 events and this is still insufficient.

Consequently, we have to use storms of lower levels to estimate the influence of each covariate and extrapolate these results to the extreme level. We will use all the storms of level greater or equal to the *low level* parameter defined in the declustering process to make estimations. For example, if the low level is 111, we call "high level storm" every storm of level 111, 132, 154, 179, 207, 236, 300 or 400. The "extreme level" will be only 400.

The mean probability of occurrence for each high level is given in Table 2.

Level	111	132	154	179	207	236	300	400
Number of storm	182	158	103	84	51	57	44	23
Frequency $\times 10^4$	7.99	6.93	4.52	3.69	2.24	2.50	1.93	1.01
Frequency in year^{-1}	2.33	2.02	1.32	1.08	0.65	0.73	0.56	0.29

Table 2: Number of occurrences and frequency of storms by level

Besides of the 3 hours ap index, we dispose of a covariate representing the solar activity of a cycle. This solar cycle activity characteristic is the maximum of the monthly Smoothed Sunspot Number (monthly SSN). For an easier interpretation of the results, this covariate will be centered. See [2] for more details on the sunspot number.

The lengths of the cycles are also available, we call D_j the length of the j^{th} cycle.

3.4 Time Warping

The durations of the 7 complete solar cycles range from 9.7 to 12.6 years. Thus, in order to analyze all the 7 cycles together, a data warping is applied to each cycle: the position of a storm on a cycle is represented by a number between -0.5 and 0.5 where -0.5 is the beginning of the cycle, 0.5 its end and 0 its middle (peak). In the Figure 3, the dash-dotted line represents the warped time for the first complete solar cycle.

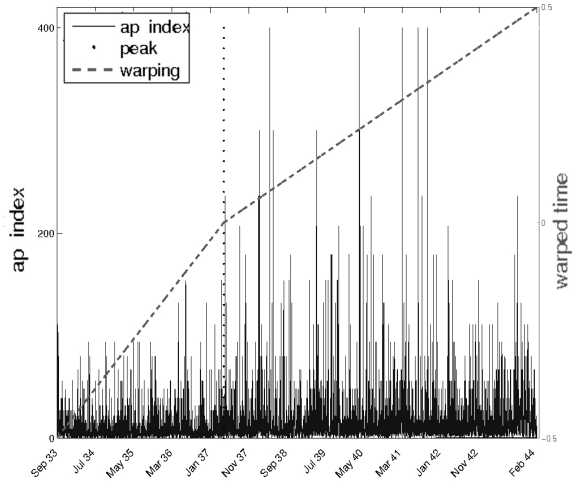

Figure 3: The ap index during the first cycle. The dotted vertical line represents the peak and the dash-dotted line the warped time.

3.5 Proportional hazard model

The model we built is inspired by the Cox model. First introduced in epidemiology, the Cox model is a proportional hazard model which permits to express the instantaneous risk with respect to time and some covariates $(X_1, ..., X_p)$. In epidemiology, these variables are risk factors as well as treatments. The instantaneous risk $\lambda(t, X_1, ...X_p)$ is defined using the occurrence probability in an infinitesimal interval

$$\mathbb{P}\{\text{there exists an event} \in [t, t+dt]\,\} = \lambda(t, X_1, ...X_p)dt$$

In the Cox model, this instantaneous risk is a relative risk with respect to a reference risk $\lambda_0(t)$, often related to a control treatment. The influence of the covariates is modeled by the exponential of a linear combination of them. That is to say:

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

where β_i quantifies the influence of the i^{th} covariate. For more details about the Cox model, see [4].

The model constructed here has undergone meaningful modifications from the Cox model:

- an event (a storm occurrence) may occur several times within a cycle. Hence we use Poisson distributions instead of Bernoulli ones;
- the variable D_j is included as factor, thus the measurement unit is the number of events per time unit and not per cycle;
- $\lambda_0(t)$ is not considered as a nuisance parameter but as a parameter to estimate;
- the estimation is made using all the storms of high level and an extrapolation to the storms of extreme level 400 is applied using the parameter P_{400}, the probability that a high level storm grows into a storm of level 400. The utilization of this parameter assumes that the level reached by a high level storm does not depend on the instant of appearance. A chi-square independence test showed that this assumption is acceptable. For precisions on this test, see Appendix B.

Thus, in the model we developed, the number of observed storms (of high level) during the cycle j at time t, called $N_j(t)$, is supposed to be a non-homogeneous Poisson process of intensity $\lambda_j(t)$ such as :

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

i.e.

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

The basic intensity $\lambda_0(t)$ takes into account the fact that storms occurs more likely during the second half of the cycle. We want to estimate it. Note that only one covariate is used here, the solar activity index X_j and that the parameter β models its influence.

3.6 A model extension

We have seen that there is a strong difference between the two halves of a solar cycle. Thus, we tried to implement a modified model, where the estimation was made separately on every half. Thus, the variable D_j was replaced by $D_{j,1}$ and $D_{j,2}$, the lengths of the first and second half of the cycle, and then, $N_j(.)$ was a non-homogeneous Poisson process of intensity:

$$\begin{aligned} \lambda_{j,1}(t) &= \lambda_0(t) D_{j,1} \exp(\beta_1 X_j) \quad \text{if } t < 0 \\ \lambda_{j,2}(t) &= \lambda_0(t) D_{j,2} \exp(\beta_2 X_j) \quad \text{if } t \geq 0 \end{aligned}$$

But the estimation in this model led to incoherent results. Indeed, because of the presence of a normalization constant different on each half (see Section 4.2), the basic intensity during the first half was higher than during the second one. Hence, this approach was abandoned.

4 Estimation

4.1 P_{400} and β

Since P_{400} is independent of the position in the cycle, the empirical frequency is used

$$\widehat{P_{400}} = \frac{\#\{\text{storms of level } 400\}}{\#\{\text{storms of level} \geq \textit{low level}\}}$$

And noting $m = \#\{\text{storms of level} \geq \textit{low level}\}$ we get the corresponding 95% confidence interval:

$$P_{400} \in \left[\widehat{P_{400}} \pm 1.96\sqrt{\widehat{P_{400}}(1-\widehat{P_{400}})/m}\right]$$

For β, we use the fact that

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

As in the Cox model, we verify the sufficiency of the statistic N_j and β is estimated by its maximum likelihood estimator in a Poisson generalized linear model. A confidence interval is also computed. All the details could be found in Appendix A.

4.2 Basic intensity $\lambda_0(t)$

Here, we use a kernel estimator. Assuming that β is known, we have :

$$\widehat{\lambda_0(t)} = K\sum_{j=1}^{J}\int_{-1/2}^{1/2} dN_j(t-s)\phi(s) = K\sum_{j=1}^{J}\int_{-1/2}^{1/2} N_j(t-s)\phi'(s)ds$$

where J is the number of individuals (cycles) , K a normalization constant and ϕ the kernel, verifying $\phi(\pm 1/2) = 0$ (for the integration by parts) and $\int_{-1/2}^{1/2}\phi(s)ds = 1$.

The bias and the variance of this estimator are calculated using step functions and by passage to the limit. Let ϕ be a step function,

$$\phi(s) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(s)$$

where the $A_i = [t_i, t_{i+1}]$ form a partition of $[-1/2, 1/2]$ (we can assume $t_i < t_{i+1}$ without loss of generality) and the a_i are such that $\int_{-1/2}^{1/2}\phi(s)ds = 1$. Then, for each $t \in [-1/2, 1/2]$,

$$\begin{aligned}\widehat{\lambda_0(t)} &= \sum_{j=1}^{J} K\int_{-1/2}^{1/2} dN_j(t-s)\phi(s) \\ &= K\sum_{j=1}^{J}\Big\{a_1 N_j([t-t_2, t-t_1]) + ... + a_n N_j([t-t_{n+1}, t-t_n])\Big\}\end{aligned}$$

Thus, since $N_j([a,b]) \sim \mathcal{P}\left(Q_i \int_a^b \lambda_0(s)ds\right)$ with $Q_i = D_j \exp(\beta X_j)$ and since $\mathbb{E}\mathcal{P}(\xi) = \mathbb{V}\mathcal{P}(\xi) = \xi$, we get:

$$\mathbb{E}\,\widehat{\lambda_0(t)} = K\sum_{j=1}^{J} Q_j \int_{-1/2}^{1/2} \lambda_0(s)\phi(t-s)ds$$

Similarly, for the variance:

$$\begin{aligned}\mathbb{V}\,\widehat{\lambda_0(t)} &= K^2\sum_{j=1}^{N}\left\{a_1^2\mathbb{V}N_j([t-t_2,t-t_1]) + ... + a_n^2\mathbb{V}N_j([t-t_{n+1},t-t_n])\right\} \\ &= K^2\sum_{j=1}^{N} Q_j \int_{-1/2}^{1/2} \lambda_0(s)\phi^2(t-s)ds\end{aligned}$$

In the case of a kernel concentrated around zero we obtain

$$\mathbb{E}\,\widehat{\lambda_0(t)} \simeq K\sum_{j=1}^{J} Q_j \lambda_0(t)$$

Hence, the choice $K = 1/\sum Q_j$ is convenient and then we get

$$\mathbb{V}\,\widehat{\lambda_0(t)} \simeq \frac{1}{\sum Q_j}\lambda_0(t)\int_{-1/2}^{1/2}\phi^2(s)ds$$

In practice we used for ϕ a Gaussian kernel, i.e.

$$\phi(s) = \frac{1}{\sqrt{2\pi}h}\exp(-\frac{s^2}{2h^2})$$

with h the band width parameter, determined later. Then, using the fact that

$$\phi^2(s) = \frac{1}{2\sqrt{\pi}h}\phi(\sqrt{2}s)$$

where $\phi(\sqrt{2}s)$ is the density function of a normal distribution $\mathcal{N}\left(0,(h/\sqrt{2})^2\right)$ we can say that for h sufficiently small

$$\int_{-1/2}^{1/2}\phi^2(s)ds \simeq \int_{-\infty}^{+\infty}\phi^2(s)ds = \frac{1}{2\sqrt{\pi}h}$$

In order to avoid edge effects, a periodization is applied before the estimation process. The band width parameter h is chosen by cross-validation with minimization of the Integrated Square Error. See [6] or [9] for more details.

Remark: as indicated in Section 3.2, the intensity estimated by the kernel method does not correspond to the intensity we want to evaluate. Indeed, the intensity we estimate correspond to the probability P_1 that a storm of level 400 begins at time t. Hence we apply a correction by multiplying $\widehat{\lambda_0(t)}$ by $29/23$.

Thus, we obtain the approximate confidence interval for $\lambda_0(t)$

$$\lambda_0(t) \in \left[\widehat{\lambda_0(t)} \pm 1.96\sqrt{\frac{1}{\sum Q_j}\frac{\widehat{\lambda_0(t)}}{2\sqrt{\pi}h}}\right]$$

5 Results

5.1 Instantaneous intensity

The graphic in Figure 4 gives the estimation result of $\widehat{\lambda_0(t)}$ for a low level of 111 with the confidence area (i.e. the intensity for all the storms of level greater or equal to 111). The bandwidth parameter is selected by cross validation and is equal to 0.035. As expected, the basic intensity is higher during the second half of the cycle. One can also see a significant increase near of the x-axis zero, highlighting the difference between the two halves of a solar cycle.

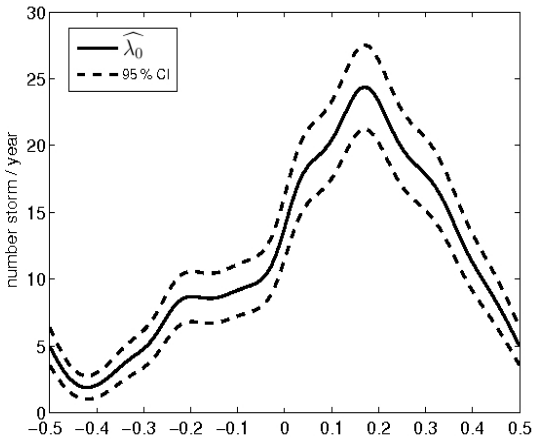

Figure 4: Estimated instantaneous intensity (years^{-1}) of the storms of level greater or equal to 111, for a mean solar activity of 146.7

5.2 P_{400} and β

For $\widehat{P_{400}}$, the obtained results for different low levels are gathered in the Table 3.

Low level	111	132	154
$\widehat{P_{400}}$	0.031384	0.041905	0.059299
95 % C.I.	[0.018477 ; 0.044291]	[0.024765 ; 0.059045]	[0.035266 ; 0.083333]

Table 3: $\widehat{P_{400}}$ (the probability for a high storm to grow into a storm of level 400) and 95 % confidence intervals for each low level

With a low level of 111, the estimation of β gives:

$$\hat{\beta} = 0.0059651 \quad \text{with the 95\% confidence interval} \quad [0.0035873; 0.0083429]$$

Although this value seems to be small, the significance of $\hat{\beta}$ has been shown by a likelihood ratio test. The test of $\beta = 0$ against $\beta = \hat{\beta}$ returns a p-value of 7.02×10^{-7}. Thus, the solar activity index X affects the number of storms

occurring during a cycle. Graphically, the influence of the solar activity index on the number of storms per cycle is observable on Figure 5.

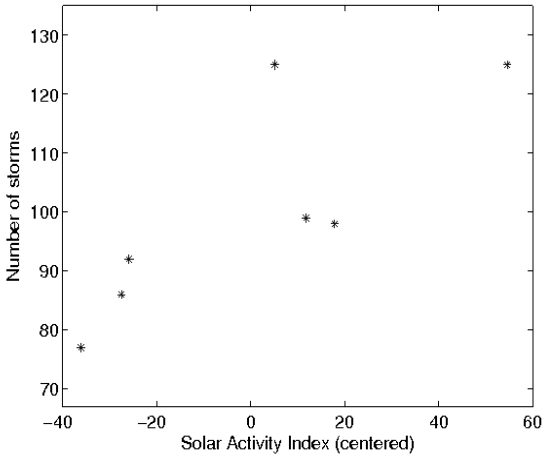

Figure 5: Total number of storms per cycle for a low level of 111 according to the solar activity (centered)

5.3 Instantaneous intensity: extrapolation to level 400 and relative risk

The extrapolation to the storms of extreme level 400 is made by multiplying by $\widehat{P_{400}}$ (with confidence interval). We obtain the final intensity shown in Figure 6. This curve corresponds to intensity of apparition of extreme storms for a solar cycle with a mean solar activity of 146.7. Recall that in the equation:

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

the risk factor is $\exp(\beta X_j)$. Then, using $\hat{\beta}$, we can evaluate the relative risk for a cycle with a given solar activity index. For example compared to the average level of solar activity (146.7), a cycle with a high solar activity of 180 has a relative risk of $\exp(33.3 \times 0.0059651) = 1.22$.

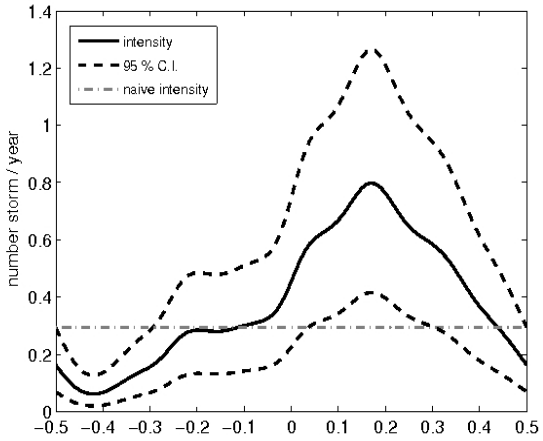

Figure 6: Instantaneous intensity (years^{-1}), with confidence interval, of the storms of level 400 obtained by extrapolation from the low level 111, for a mean solar activity of 146.7. In dash-dotted line the empirical frequency of storms of level 400

5.4 Method stability

The results presented in the previous sections are given for a fixed low level (of 111). This asks the question of the model sensitivity to this parameter. The stability of the employed method can be evaluated by testing the stability to a low level change. The results for two other low levels, 132 and 154, are given in Figure 7. The two last curves seem to be smoother but this is partly due to the bandwidth parameter which is now equal to 0.045 (always selected by cross validation). For more precision, see Figure 8 where the three instantaneous intensity curves are plotted together. One can see that there is no significant difference between the three curves and that the method is rather stable.

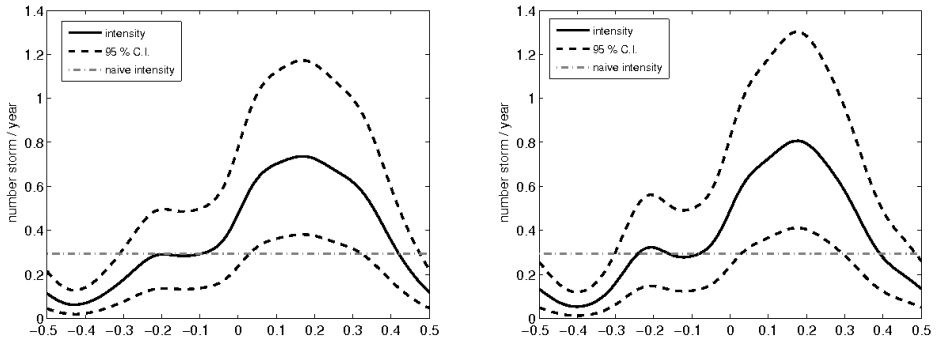

Figure 7: Similar to Figure 6 with a low level of 132 (left) and 154 (right)

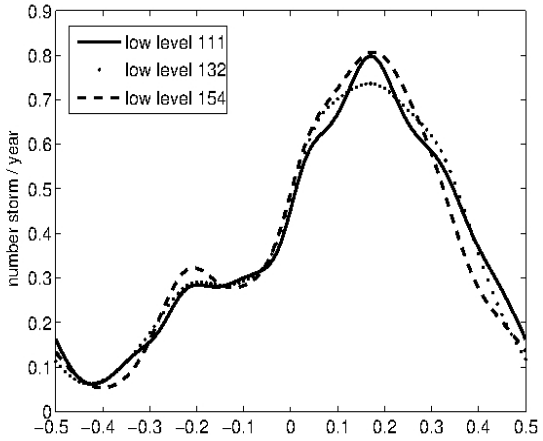

Figure 8: Instantaneous intensity (years^{-1}) of the storms of level 400 obtained by extrapolation from the low levels 111 (plain line), 132 (dotted line) and 154 (dashed line)

5.5 A model extension

In an alternative approach, we consider the gradient of a storm to characterize its strength (instead of the ap index level). Gradients are calculated on one time step (3H) and the storm gradient is defined as the maximal gradient attained during a storm. This approach has been setting up because of the observation of storms with low levels (less than an ap index of 111) but strong effects due to fast variations of the ap index. We have led the same study with this new definition for the storm strength. The extreme gradient level are those greater than 100 and the low one is 35. The estimation of β gives

$$\hat{\beta} = 0.0053499 \quad \text{with the confidence interval} \quad [0.0038128; 0.006887]$$

These values are similar to those obtained with the ap index. The estimated intensity for the storms of extreme gradient is plotted in Figure 9. One can see that the step between the two halves of the cycle is stronger.

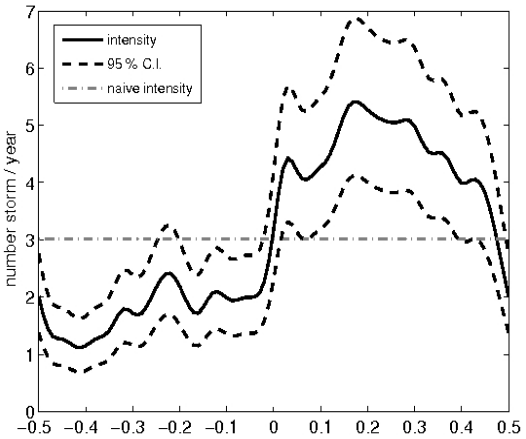

Figure 9: Instantaneous intensity (years^{-1}), with confidence interval, of the storms with extreme gradient (≥ 100) obtained by extrapolation from the low gradient level 35, for a mean solar activity of 146.7. In dash-dotted line the empirical frequency of storms with an extreme gradient

We should precise that the use of the gradient involves one disadvantage. Since the ap index represents a maximum over a 3 hours period, the two values of ap index used for the gradient calculation can be separated by nearly 6 hours or only by few minutes. The real dates of these values are not known and the gradient is calculated using 3 hours time step. However, the calculated gradient gives an approximation of the variation speed of the ap index. Moreover, since the gradient is used analogously to the ap index, the original model is still appropriate here.

6 Conclusion

This study highlights that the intensity of magnetic storm occurrence strongly depends on the position on the solar cycle. The probability is higher during the second half of the cycle. The solar activity also has an influence on this intensity and, giving an activity index, allows to express a relative risk (compared to a cycle with the average level of solar activity 146.7).

The analyses has been performed for different low levels in order to check his stability. The first results are given for a low level of 111 and a comparison is made using two other low levels: 132 and 154. The shape similarity of the three curves attests of the method stability.

The model we built also allows us to make predictions about the current solar cycle. For the beginning date of this 24th cycle, we have chosen December 2008, a date accepted by a panel of experts (although there is no consensus). For the solar activity index, we have used the NOAA prediction with a maximum of 87.9 attained on November 2013 [3]. The end of the 24th cycle is estimated around December 2019 or January 2020. The estimation (from the beginning to present) and the prediction are represented on Figure 10 (plain line).

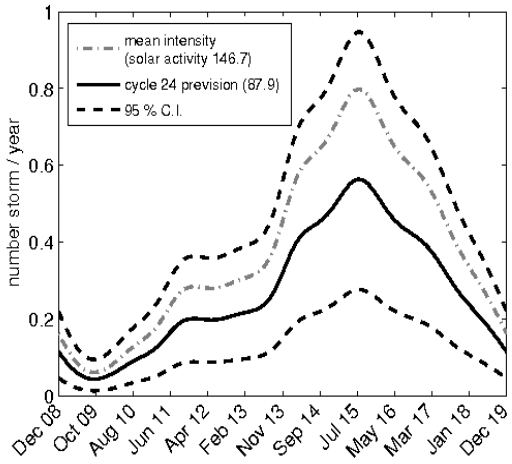

Figure 10: Estimation and prediction of instantaneous intensity (years^{-1}) of the storms of level 400 for the 24th solar cycle, with confidence interval. For comparison, in dash dotted gray, the same intensity for a cycle with a mean solar activity index of 146.7

7 Appendix

A Maximum likelihood estimator of β

The use of N_j instead of $N_j(t)$ for the estimation of β arises the question of the sufficiency of this statistic. Consider only one cycle and the model:

$$N(t) \sim \mathcal{P}\left(\lambda_0(t)dt \; D\exp(\beta X)\right) \qquad \text{for } t \in [-0.5, 0.5]$$

Then, consider $\Delta_1, \Delta_2, ..., \Delta_n$ a partition of [-0.5, 0.5] into n sub segments. For $i = 1...n$, note $N(\Delta_i) = \int_{\Delta_i} dN(t)$ the number of events in Δ_i. Giving that $N(t)$ is a Poisson process we know that the $\{N(\Delta_i), i = 1...n\}$ are independent variables and that $N(\Delta_i) \sim \mathcal{P}\left(\left[\int_{\Delta_i} \lambda_0(s)ds\right] \; D \; \exp(\beta X)\right)$. We note $C_i = \int_{\Delta_i} \lambda_0(s)ds \; D$. Then the Log-likelihood with respect to the counting measure (in which we integrate the weights $1/N(\Delta_i)!$) is

$$-\exp(\beta X)\sum_{i=1}^{n} C_i + \sum_{i=1}^{n}[N(\Delta_i)\log(C_i)] + \beta X \sum_{i=1}^{n} N(\Delta_i)$$

We see that β is linked to the $N(\Delta_i)$ only by the term $\sum_{i=1}^{n} N(\Delta_i)$. Hence there is no loss of information to use the total number of events per cycle for the estimation of β.

We can now compute the maximum likelihood estimator. For the j^{th} cycle, the likelihood with respect to the counting measure with weights $1/N_j!$ is, noting $\alpha = \int_{-1/2}^{1/2} \lambda_0(s)ds$

$$\exp\left(-\alpha \, D_j \exp(\beta X_j)\right)\left(\alpha \, D_j \exp(\beta X_j)\right)^{N_j}$$

and the Log-likelihood for all the J cycles:

$$-\alpha \sum_{j=1}^{J} D_j \exp(\beta X_j) + \log(\alpha) \sum_{j=1}^{J} N_j + \sum_{j=1}^{J} N_j \log(D_j) + \beta \sum_{j=1}^{J} N_j X_j$$

The derivatives in α anb β respectively give :

$$\sum_{j=1}^{J} D_j \exp(\beta X_j) = \frac{\sum_{j=1}^{J} N_j}{\alpha}$$

and

$$\alpha \sum_{j=1}^{J} D_j X_j \exp(\beta X_j) = \sum_{j=1}^{J} N_j X_j$$

Replacing α by the solution of the first equation, we obtain:

$$\sum_{j=1}^{J} D_j X_j \exp(\beta X_j) \sum_{j=1}^{J} N_j = \sum_{j=1}^{J} D_j \exp(\beta X_j) \sum_{j=1}^{J} N_j X_j$$

This implicit equation resolves only numerically (by the secant method).

We can also compute the Fisher information matrix:

$$\begin{pmatrix} \alpha^{-1} \sum_{j=1}^{J} D_j \exp(\beta X_j) & \sum_{j=1}^{J} D_j X_j \exp(\beta X_j) \\ \sum_{j=1}^{J} D_j X_j \exp(\beta X_j) & \alpha \sum_{j=1}^{J} D_j X_j^2 \exp(\beta X_j) \end{pmatrix}$$

The (2,2) coefficient of the inverse matrix of the Fisher information matrix provides the variance of $\hat{\beta}$, used for the construction of a confidence interval.

B Chi-square test:

The chi-square independence test is performed a posteriori. When the instantaneous intensity is estimated, the time interval $[-0.5, 0.5]$ is separated into two parts, of low and high intensity. The intensity threshold for this partition will be the empirical frequency of extreme storms, which is about 0.29 storm per year (horizontal dash-dotted line in Figure 6). The two parts correspond to the times where the instantaneous intensity is respectively below and above this threshold.

Then, the chi-square test is applied to the proportions of extreme level storms for each area and returns a p-value of 0.26, leading to the acceptance of the independence hypothesis. The same test is applied with different thresholds for the partition into two areas (0.40, 0.50 and 0.60) and always leads to the same conclusion.

References

[1] `http://www.ngdc.noaa.gov/stp/GEOMAG/kp_ap.html`.

[2] `http://www.sidc.be/sunspot-data/`.

[3] `http://www.swpc.noaa.gov/ftpdir/weekly/Predict.txt`, Update on April 8, 2013, consulted on April 8, 2013.

[4] O.O. Aalen, Ø. Borgan, and H.F. Gjessing. *Survival and event history analysis. A process point of view*. Springer, New York, 2008.

[5] C. W. Anderson. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probability*, 7:99–113, 1970.

[6] A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

[7] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001.

[8] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *J. Amer. Statist. Assoc.*, 102(479):824–840, 2007.

[9] P. Hall. Large sample optimality of least squares cross-validation in density estimation. *The Annals of Statistics*, 11(4):1156–1174, 1983.

[10] P. Jonathan and K. Ewans. Modeling the seasonality of extreme waves in the gulf of mexico. *Journal of offshore mechanics and Arctic engineering*, 133(2):113–123, 2011.

[11] E. Rifa. Études des relations entre indices solaires et géomagnetiques. *Intership report*.