Mécanique des Fluides, Énergétique et Environnement - MF2E 1A

BE - Statistiques

Auteurs :
TRAN Ugo
[redacted] Romain

Responsable du BE :
Pr. Jean-Yves Tourneret
Pr. Corinne Mailhes

Mars 2024

Sommaire

Liste des figures

1 Objectif de l'étude

1.1 Introduction

Pour permettre une meilleure compréhension des phénomènes atmosphérique, de déterminer le climat dans une zone géographique donnée ou encore de prévoir des phénomènes extrêmes, il est important de savoir analyser les données météorologique tel que le vent ici présenté dans ce BE. L'analyse de ces données météorologiques est utile dans des domaines tels que l'agriculture, l'air, la mer, le contrôle de la circulation, calculs d'ingénierie structurelle, études de changements globaux, ressources solaires et éoliennes estimation, etc.

D'après l'article ANALYSIS AND MODELLING OF TIME SERIES OF SURFACE WIND SPEED AND DIRECTION publié en 1999 dans INTERNATIONAL JOURNAL OF CLIMATOLOGY.

L'objectif de notre notre bureau d'étude est de pouvoir analyser des données sur des vitesses de vent et, grâce à l'étude d'un test d'hypothèse de pouvoir déterminer si ces données proviennent de vent fort ou calme.

Pour notre étude, on va modéliser la vitesse du vent grâce à une variable aléatoire Y suivant une loi de Weibull $W(\theta, p)$.

1.2 Méthodologie

Notre étude sera divisé en quatre parties :

- Génération de signaux test
- Estimation statistique
- Détection : Étude des performances d'un test statistique
- Analyse d'un fichier de données contenant des vitesses de mesures du vent

2 Rappels théorique

2.1 Génération d'un signal de loi de Weibull $W(\theta, p)$

On souhaite générer une signal de loi de Weibull $W(\theta, p)$, qui a pour fonction de répartition :

$$F(x;\theta,p) = (1 - \exp(-(\frac{x}{\theta})^p)) \times \mathbb{I}_{\mathbb{R}_+}$$

On sait alors que $Y = F^{-1}(X;\theta,p)$ suit une loi de Weibull $W(\theta, p)$.
Avec $F^{-1}(X;\theta,p) = (-ln(1-X)^{\frac{1}{p}} \times \theta)$

2.2 Estimation statistique

On se donne des observations $(y_1, ..., y_N)$ suivant une loi de Weibull $W(\theta, p)$. On souhaite estimer le paramètre d'échelle θ (p est supposé connu). On va commencer par calculer l'estimateur du maximum de vraisemblance de θ ($\widehat{\theta}_{MV}$) construit à partir des observations $(y_1, ..., y_N)$.
La log-vraisemblance s'écrit alors :

$$\ln L(y_1, ..., y_N;\theta,p) = N \ln(\frac{p}{\theta}) + \sum_{i=1}^{N}[(p-1)\ln(\frac{y_i}{\theta}) - (\frac{y_i}{\theta})^p]$$

Qui admet pour dérivée :

$$\frac{\partial \ln L(y_1, ..., y_N;\theta,p)}{\partial\theta} = -\frac{Np}{\theta}\left[1 - \frac{1}{N\theta^p}\sum_{i=1}^{N} y_i^p\right]$$

Et qui s'annule en :

$$\boxed{\theta_0 = \left(\frac{1}{N}\sum_{i=1}^{N} y_i^p\right)^{\frac{1}{p}}} \tag{1}$$

La vraisemblance admet bien un unique maximum d'après le tableau de variation suivant, avec $f(x) = L(y_1, ..., y_N; x, p)$:

x	Variation de f
$-\infty$	Croissance de f (+)
θ_0	Atteint un maximum en $x = \theta_0$
$+\infty$	décroissance de f (−)

Ainsi $\widehat{\theta}_{MV} = \left(\frac{1}{N}\sum_{i=1}^{N} y_i^p\right)^{\frac{1}{p}}$ est l'estimateur du maximum de vraisemblance.

Pour simplifier l'étude on s'intéresse à $a = \theta^p$. On admettra que l'estimateur du maximum de vraisemblance de a est :

$$\widehat{a}_{MV} = \frac{1}{N}\sum_{i=1}^{N} y_i^p \tag{2}$$

On peut montrer que $\widehat{a}_{MV}$ est un **estimateur non biaisé** et que c'est **l'estimateur efficace** de a.
Commençons par montrer que c'est **un estimateur non biaisé**, on a par linéarité de l'espérance :

$$\mathbb{E}[\widehat{a}_{MV}] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[y_i^p] \tag{3}$$

Montrons que si $Y_i \sim W(\theta, p)$ alors $Z_i = \frac{2}{a}Y_i^p \sim \chi_2^2$. Pour cela, on va calculer la fonction de répartition de Z_i :

$$F_Z(z) = P(Z_i \leq z) = P\left(Y_i \leq \left(\frac{a}{2}\right)^{\frac{1}{p}} z^{\frac{1}{p}}\right) = F_y\left(\left(\frac{a}{2}\right)^{\frac{1}{p}} z^{\frac{1}{p}}\right)$$

F_y étant la fonction de répartition de Y_i. La densité de probabilité de Z_i est alors donné par :

$$f_Z(z) = \frac{\partial F_Z(z)}{\partial z} = \frac{\partial F_y\left(\left(\frac{a}{2}\right)^{\frac{1}{p}} z^{\frac{1}{p}}\right)}{\partial z} = f_Y\left(\left(\frac{a}{2}\right)^{\frac{1}{p}} z^{\frac{1}{p}}\right)\frac{1}{p}\left(\frac{a}{2}\right)^{\frac{1}{p}} z^{\frac{1}{p}-1}$$

Où f_Y est la densité de probabilité de Y_i. Il vient ainsi :

$$\boxed{f_Z(z) = \tfrac{1}{2}\exp\left[-\tfrac{z}{2}\right]}$$

On reconnait bien la densité de probabilité d'une loi χ_2^2, ainsi on a bien $Z_i = \frac{2}{a}Y_i^p \sim \chi_2^2$. Donc $\mathbb{E}[Z_i] = 2$
Par linéarité de l'espérance, il vient :

$$\mathbb{E}[z_i] = \frac{2}{a}\mathbb{E}[y_i^p] \implies \mathbb{E}[y_i^p] = a$$

Finalement, on obtient grâce à l'équation (3) :

$$\boxed{\mathbb{E}[\widehat{a}_{MV}] = a} \tag{4}$$

Ainsi, $\widehat{a}_{MV}$ est bien un **estimateur non biaisé**.
Montrons maintenant que $\widehat{a}_{MV}$ est **l'estimateur efficace** de a, on a :

$$Var[\widehat{a}_{MV}] = \frac{1}{N^2}\sum_{i=1}^{N} Var[y_i^p]$$

Or $z_i = \frac{2}{a} y_i^p \sim \chi_2^2$. Donc $\mathrm{Var}[z_i] = 4$ d'où $\mathrm{Var}[y_i^p] = a^2$. Il vient :

$$\boxed{\mathrm{Var}[\widehat{a}_{MV}] = \frac{a^2}{N}} \tag{5}$$

D'autre part, on a avec $b_n\prime(\theta) = 0$ car l'estimateur est non biaisé:

$$BCR(\theta) = \frac{[1 + b_n\prime(\theta)]}{-\mathbb{E}\left[\dfrac{\partial^2 \ln L(y_1, ..., y_N; a, p)}{\partial a^2}\right]} = \frac{1}{-\mathbb{E}\left[\dfrac{\partial^2 \ln L(y_1, ..., y_N; a, p)}{\partial a^2}\right]}$$

Avec $\ln L(y_1, ..., y_N; a, p) = N \ln\left(\frac{p}{a}\right) + (p-1)\sum_{i=1}^N \ln(y_i) - \frac{1}{a}\sum_{i=1}^N y_i^p$

D'où : $\dfrac{\partial^2 \ln L(y_1, ..., y_N; a, p)}{\partial a^2} = \frac{N}{a^2} - \frac{2}{a^2}\sum_{i=1}^N y_i^p$ donc $\mathbb{E}\left[\dfrac{\partial^2 \ln L(y_1, ..., y_N; a, p)}{\partial a^2}\right] = -\frac{N}{a^2}$

Il vient alors :

$$\boxed{BCR(\theta) = \frac{a^2}{N} = \mathrm{Var}[\widehat{a}_{MV}]}$$

Ainsi, $\widehat{a}_{MV}$ est bien **l'estimateur efficace** de a.

2.3 Détection

On construit maintenant le test de Neyman-Pearson sur ka paramètre $a = \theta^p$.
On fixe α, on rejette H_0 si :

$$\frac{\prod_{i=1}^N \left[\frac{p y_i^{p-1}}{a_1} \exp\left(-\dfrac{y_i^p}{a_1}\right)\right]}{\prod_{i=1}^N \left[\frac{p y_i^{p-1}}{a_0} \exp\left(-\dfrac{y_i^p}{a_0}\right)\right]} > S_{\alpha,0} \Longleftrightarrow \sum_{i=1}^N y_i^p \left(\frac{1}{a_0^p} - \frac{1}{a_1^p}\right) > S_\alpha$$

Avec $S_\alpha = \ln(S_{\alpha,0}) + N[\ln(a_0) - \ln(a_1)]$
Ainsi, la **statistique de test** issue du théorème de Neyman-Pearson associée à ces deux hypothèses s'écrit :

$$\boxed{T(Y) = \sum_{i=1}^N Y_i^p}$$

De plus, pour une probabilité de fausse alarme α, la **région critique** du test (zone de rejet de H_0) est définie par :

$$\boxed{R_\alpha = \left\{ y \in \mathbb{R}^N | T(y) > \lambda_\alpha = \frac{\ln(S_{\alpha,0}) + N[\ln(a_0) - \ln(a_1)]}{\left(\frac{1}{a_0^p} - \frac{1}{a_1^p}\right)} \right\}}$$

Il faut maintenant déterminer λ_α. On a fixé la probabilité de fausse alarme α, on peut donc la calculer. Par définition, on a :

$$\alpha = P(Rejeter\ H_0 |\ H_0\ est\ vraie) = P(T(y) > \lambda_\alpha |\ H_0\ est\ vraie) = P\left(\sum_{i=1}^N \frac{2}{a_0} Y_i^p > \frac{2}{a_0}\lambda_\alpha |\ H_0\ est\ vraie\right) \tag{6}$$

Montrons que $Z_N = \sum_{i=1}^N \frac{2}{a} Y_i^p = \sum_{i=1}^N Z_i \sim \chi_{2N}^2$. Avec $Z_i \sim \chi_2^2$ et $\Phi_{Z_i}(t) = (1-2it)^{-\frac{1}{2}\times 2}$
Pour cela, calculons la fonction caractéristique de Z_N. Par définition, on a :

$$\Phi_{Z_N}(t) = \mathbb{E}\left[e^{itZ_N}\right] = \prod_{k=1}^N \mathbb{E}\left[e^{itZ_k}\right] = (1-2it)^{-N}$$

On reconnait bien la fonction caractéristique d'une loi χ_{2N}^2, on a donc $\sum_{i=1}^N \frac{2}{a} Y_i^p \sim \chi_{2N}^2$.
D'après l'équation (6) on a :

$$\alpha = 1 - G_{2N}\left(\frac{2}{a_0}\lambda_\alpha\right)$$

Finalement on obtient :

$$\boxed{\lambda_\alpha = \frac{a_0}{2} G_{2N}^{-1}(1-\alpha)} \tag{7}$$

De la même manière, on peut calculer la **probabilité de non-détection** β. Par définition, on a :

$$\beta = P\left(Rejeter\ H_1 |\ H_1\ est\ vraie\right) = P\left(\sum_{i=1}^{N} \frac{2}{a_1} Y_i^p < \frac{2}{a_1}\lambda_\alpha |\ H_1\ est\ vraie\right)$$

Ainsi, grâce au même raisonnement on obtient :

$$\boxed{\beta = G_{2N}\left(\frac{2}{a_1}\lambda_\alpha\right)} \tag{8}$$

On peut donc calculer la **probabilité de détection** π :

$$\boxed{\pi = 1 - \beta} \tag{9}$$

3 Analyse de nos résultats

3.1 Génération de signaux test

Dans un premier temps, on a généré K réalisations d'un signal de N échantillons $y = (y_1, ..., y_N)^T$ de loi de Weibull $W(\theta, p)$, où θ est le **paramètre d'échelle** et p est le **paramètre de forme**.
Pour vérifier notre modèle, on l'a comparé sur la figure 1 avec la densité de probabilité d'une loi de Weibull $f(x, \theta, p) = \frac{p}{\theta}\left(\frac{x}{\theta}\right)^{p-1} \exp\left[-\frac{x}{\theta}^p\right]$.

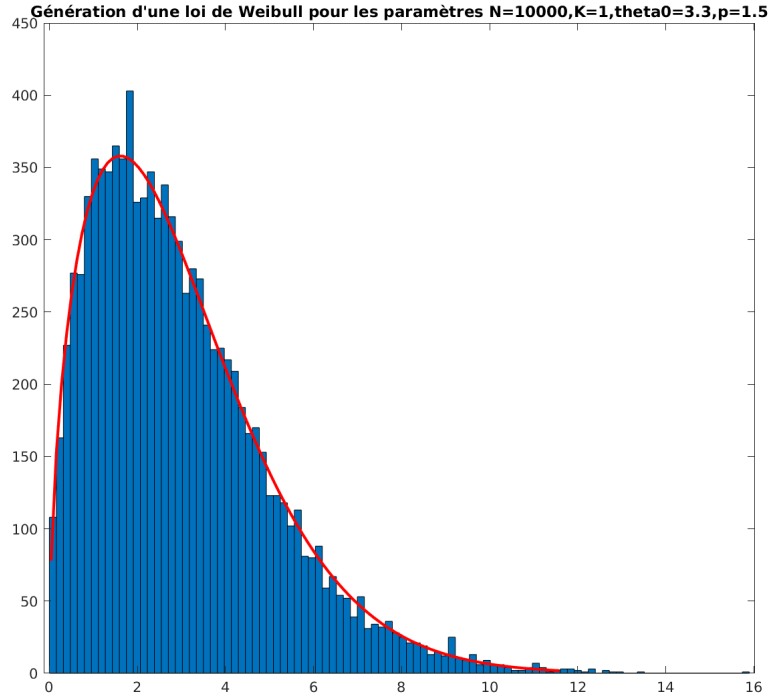


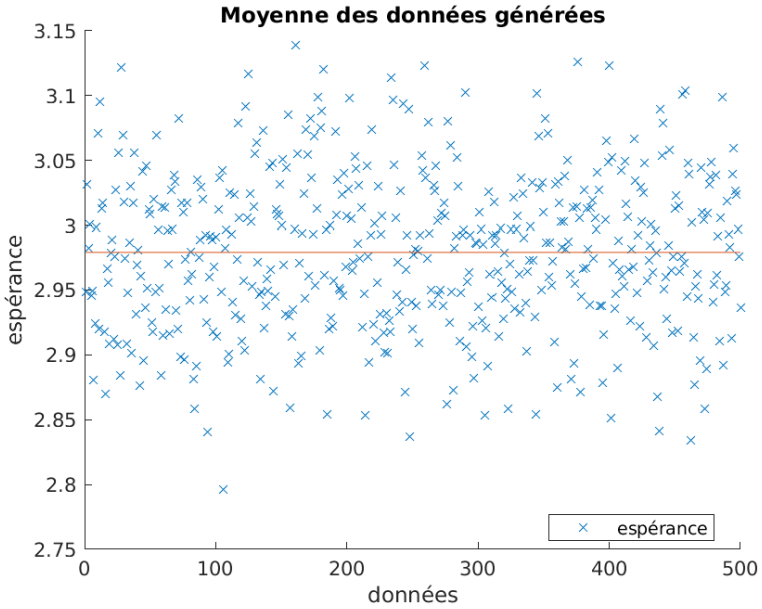
Figure 1: Superposition de la densité de la loi de Weibull avec l'histogramme des données générées.

Avec les valeurs $N = 1000$, $K = 1$, $\theta_0 = 3.3$ et $p = 1.5$, on peut remarquer graphiquement que l'histogramme des données générées est très proches de la densité théorique. On peut essayer de vérifier cela en comparant l'espérance et la variance des données générées avec les valeurs théorique attendu :

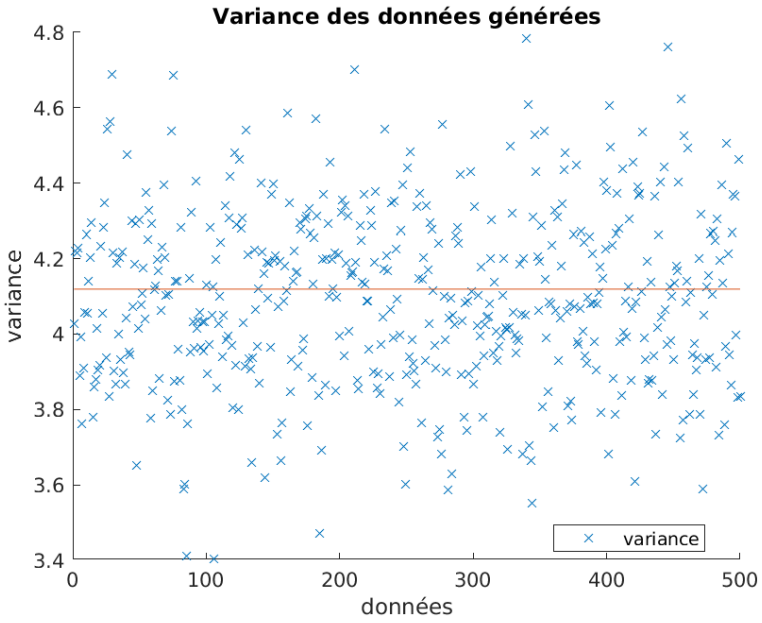
Numérique : $moyenne = 2.99$ et $variance = 4,13$
Théorique : $moyenne = \theta\Gamma(1 + 1/p) = 2.98$ et $variance = \theta^2\Gamma(1 + 2/p) - moyenne^2 = 4.09$

On va maintenant refaire la même chose avec $K = 500$. On a donc une matrice de données et on affiche la moyenne et la variance d'une réalisation du signal :

Numérique : $moyenne = 2.98$ et $variance = 4,05$
On peut représenter les moyennes et variances des colonnes des matrices en fonction des données.



(a) Représentation des moyennes



(b) Représentation des variances

On trouve des résultats proches des valeurs expérimentales précédentes et des valeurs théoriques attendues.

3.2 Estimation statistique

Dans cette section on souhaite en utilisant les observations faites du phénomène aléatoire qui suit une loi de Weibull estimer le paramètre d'échelle θ.

En appliquant l'estimateur $\widehat{a}_{MV} = \frac{1}{N}\sum_{i=1}^{N} y_i^p$ aux différentes réalisations du vecteurs de notre matrice on obtient :

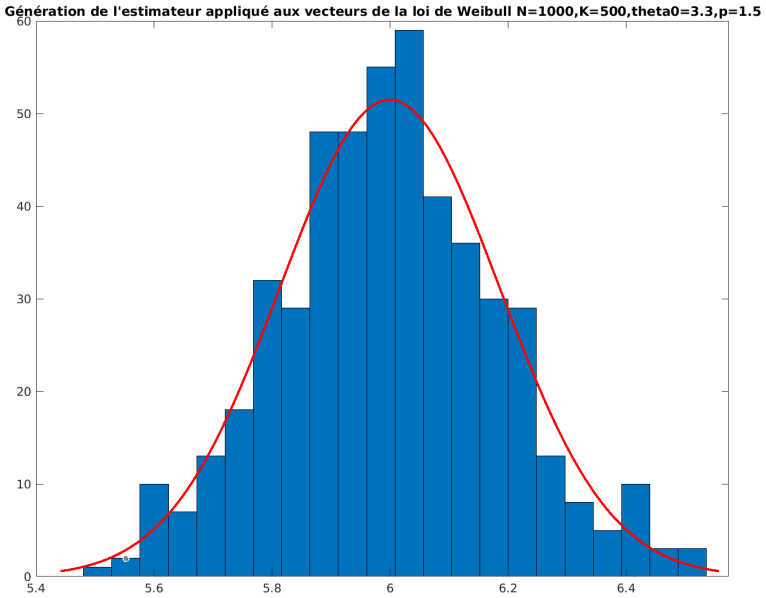


Figure 3: Représentation des estimateurs appliqués aux réalisations des vecteurs pour N=1000

On trouve pour notre estimateur les propriétés suivantes pour N=1000 : $moyenne = 5.99$ et $\theta^p = 5.99$. Ainsi le biais est nul, ce qui est cohérent car $\widehat{a}_{MV}$ est un estimateur non biaisé de $a = \theta^p$. On trouve la variance expérimentale et on la compare a la variance théorique qui tend bien vers 0 quand N tend vers l'infini: $variance_{xperimentale} = 0.038$ et $variance_{thorique} = 0.038$.

On répète la même opération pour N=10000:

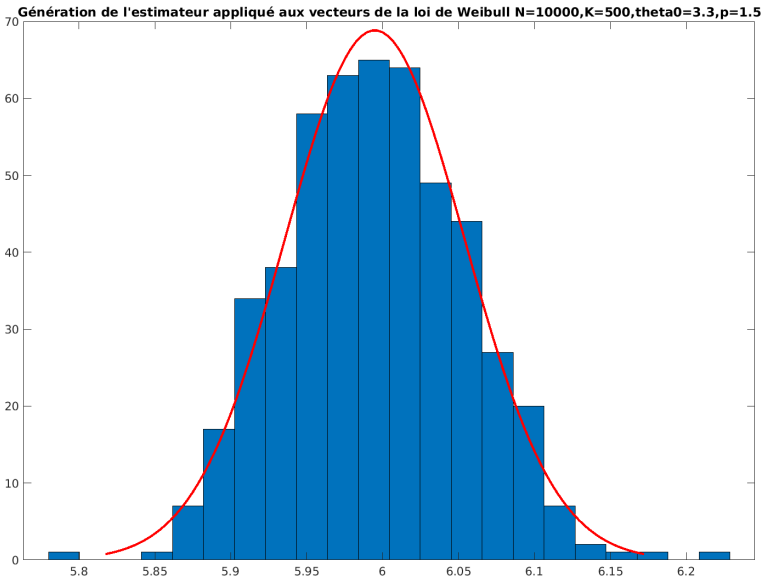


Figure 4: Représentation des estimateurs appliqués aux réalisations des vecteurs pour N=10000

On trouve pour notre estimateur les propriétés suivantes pour N=10000 : $moyenne = 5.99$ et $\theta^p = 5.99$ ainsi le biais est nul, ce qui est cohérent car $\widehat{a}_{MV}$ est un estimateur non biaisé de $a = \theta^p$ comme précédemment. On trouve la variance expérimentale et on la compare a la variance théorique

qui tend bien vers 0 quand N tend vers l'infini: $variance_{xperimentale} = 0.035$ et $variance_{thorique} = 0.036$. On a ici cette fois ci une variance un tout petit peu plus faible.

3.3 Détection

On cherche à étudier dans cette partie les performances d'un test qui détecte un vent fort ou faible. On peut observer pour un risque $\alpha \in \{0.01, ..., 0.99\}$ les courbes COR suivantes en faisant varier N:

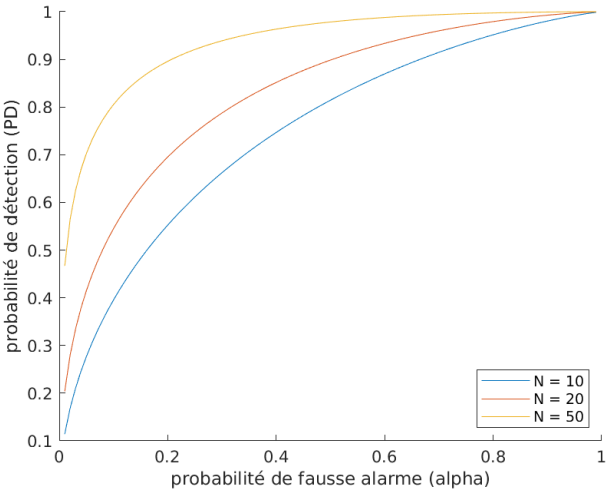


Figure 5: Courbes COR en fonction de N

Et en faisant varier en fonction de θ_1:

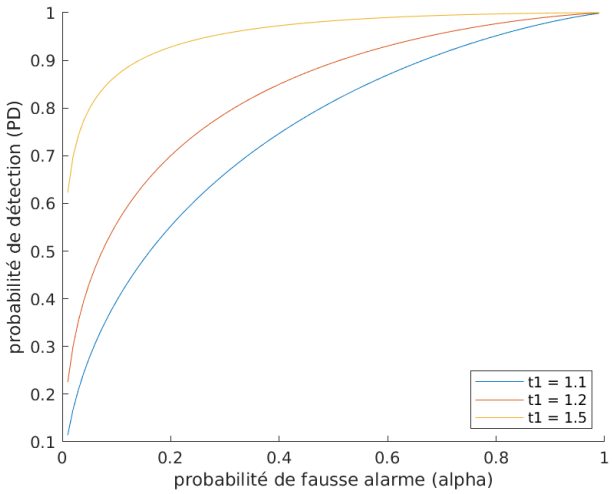


Figure 6: Courbes COR en fonction de θ_1

On observe une augmentation des courbes COR en fonction des augmentations des paramètres N ou θ_1 noté t1 sur la Fig(6).

3.4 Analyse d'un ficher de données

On analyse dans cette partie un fichier de données fourni pour le BE. On peut observer ces données

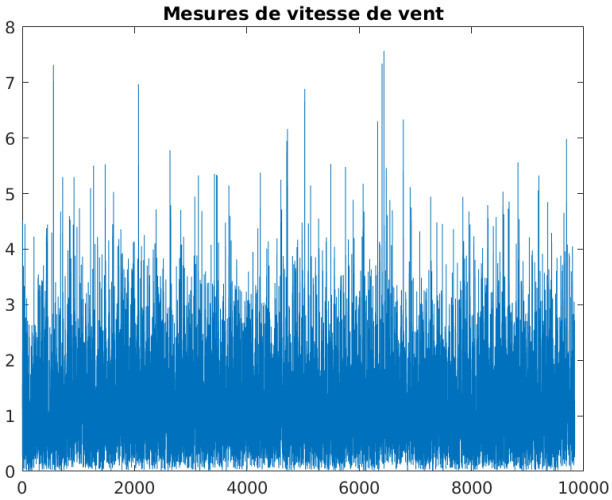


Figure 7: Mesure de la vitesse du vent fourni dans le ficher wind.mat

avant de les filtrer sur une densité d'une loi de Weibull.

À l'aide de la fonction wblfit on détermine a l'aide de la méthode de vraisemblance des estimés des paramètres θ et p, on trouve $\hat{\theta} = 1.2923$ et $\hat{p} = 1,2895$.

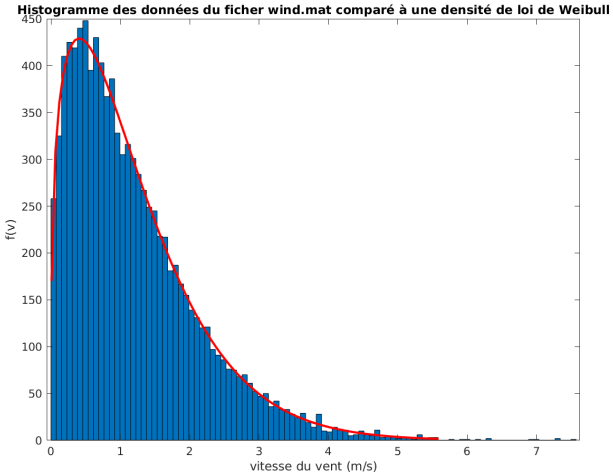


Figure 8: Histogramme des données comparé à une densité d'une loi de Weibull

On compare par la suite sur le même graphique, la fonction de répartition de la loi de Weibull avec ces nouveaux paramètres évaluée aux données du fichier, et la fonction de répartition empirique de ces données. Pour N=100:

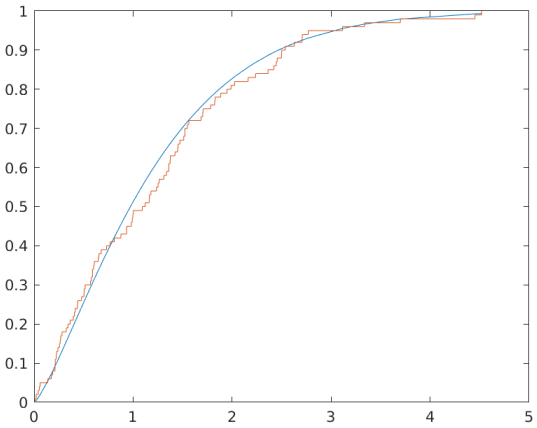


Figure 9: Comparaison entre la fonction de répartition de la loi de Weibull et de sa fonction de répartition empirique pour N=100

On choisit N=100 car pour N beaucoup plus grand on ne peut pas observer de différences entre les courbes car elles sont confondues. Exemple pour N=9839:

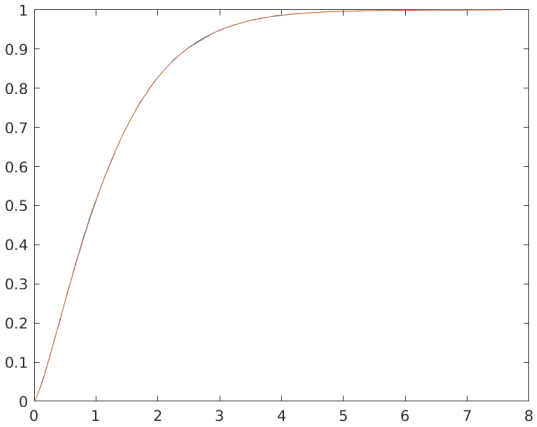


Figure 10: Comparaison entre la fonction de répartition de la loi de Weibull et de sa fonction de répartiton empirique pour N=9839

Soit Dn la valeur de la statistique du test de Kolmogorov,

$$\boxed{Dn = max_i[Ei^+; Ei^-]} \tag{10}$$

Avec $Ei^+ = |\frac{i}{N} - F_W(y_i; \hat{\theta}, \hat{p})|$ et $Ei^- = |\frac{i-1}{N} - F_W(y_i; \hat{\theta}, \hat{p})|$, ou F_W est la fonction de répartition d'une loi de Weibull.

On trouve Dn=0,05 numériquement. En vérifiant à l'aide de la fonction kstest on trouve que H_0 est vérifié. En appliquant ce test à notre ensemble de données on trouve un vent fort.

4 Conclusion

Pour ce BE, nous avons dans un premier temps cherché à estimer les paramètres d'une loi de Weibull. Pour cela, nous avons fixé p afin de trouver une bonne estimation du paramètre θ, pour ensuite déterminer les performances d'un test statistique.
Nous avons après cela utilisé nos résultats sur un jeu de données de vitesses de vents afin d'estimer si ce jeu de données suivait une loi de Weibull.
Finalement nous avons put conclure grâce au test de Kolomogorov si ces données proviennent d'un vent fort ou calme.