



Statistiques

TP / TDM

Auteurs :

Jean-Yves Tourneret, Toulouse INP
Corinne Mailhes, Toulouse INP

Références :

1A MF2E
2A SN Apprentis

Version 3.0 du
10 février 2026



Table des matières

1	Notation	1
2	Déroulement des séances	2
3	Le TP/TDM	3
3.1	Génération de signaux test	4
3.2	Estimation statistique	4
3.3	Détection	5
3.4	Analyse d'un fichier de données	7
4	Annexes	9
4.1	Lois de Weibull, du Chi2 et Gamma	9
4.2	Estimation des paramètres de la loi de Weibull	10
4.2.1	Estimateur du maximum de vraisemblance du paramètre d'échelle .	10
4.2.2	Estimateur des moments du paramètre d'échelle	10
4.2.3	Borne de Cramér-Rao du paramètre d'échelle	11
4.2.4	Estimateur du maximum de vraisemblance des deux paramètres . .	11
4.3	Test de Neyman-Pearson pour la détection	12
4.4	Article de Martin (1999)	13

1 Notation

Les élèves se répartiront par binôme ou monôme mais **pas de trinôme**.

La notation tiendra compte de deux aspects :

- les réponses aux questions posées à la fin de chaque séance
- et la qualité des codes fournis représentatifs de la bonne compréhension des notions étudiées durant le TP/TDM.

En particulier, les commentaires fournis après chaque résultat obtenu sont importants et la qualité de ces commentaires est plus importante que le fait de terminer tout le TP/TDM (bien sûr, si vous faites les deux, c'est mieux!).

Plus précisément :

1. A la fin de chaque séance de TP/TDM, 15 min avant la fin, un questionnaire sera distribué, à remplir par binôme (ou monôme).

2. A la fin du TP/TDM, le LiveEditor (.mlx) sera envoyé à **corinne.mailhes@tesa.prtd.fr**, avec un nom identifiant les auteurs (**NOM1-NOM2.mlx**).

Le LiveEditor présentera l'évolution du projet au fur et à mesure des questions

- avec une partie texte destinée à faire comprendre le problème posé,
- comment il a été résolu,
- le code matlab correspondant, commenté de manière claire
- et après le code et l'affichage des résultats sur des figures, des commentaires sont attendus sur les résultats obtenus.

Un nouvel utilisateur doit pouvoir comprendre ce que vous avez souhaité implanter.

Concernant les figures, elles doivent être suffisamment renseignées, claires et lisibles pour que l'utilisateur les comprenne :

- un titre, des axes x et y lisibles, avec un label permettant de comprendre la nature des axes ("Fréquences", "Secondes", "Amplitude", etc.) ;
- attention aux polices de caractères!!! suffisamment grandes pour être lisibles ;
- une légende permettant de comprendre à quoi correspondent les diverses courbes, bien différenciées par des couleurs nettes ;

2 Déroulement des séances

Il y a 4 séances de prévues :

Séance 1 : initiation Matlab, l'essentiel à savoir pour faire des stats ;

Séance 2 : début du TP/TDM - Génération de signaux test et Estimation statistique ;

Séance 3 : Détection ;

Séance 4 : Analyse d'un fichier de données.

Ce séquençement n'est pas strict et pourra être adapté en fonction des avancées.

3 Le TP/TDM

La vitesse du vent peut être modélisée par une variable aléatoire Y de loi de Weibull $\mathcal{W}(\theta, p)$ comme illustré sur la figure ci-dessous issue de [3]. Des propriétés de la loi de Weibull sont rappelées à la fin de cet énoncé. En particulier, la valeur moyenne de cette variable aléatoire est proportionnelle au facteur d'échelle θ , avec un facteur de proportionnalité proche de 1 pour des valeurs de paramètre de forme p entre 1 et 1.5.

Ainsi, si on estime le paramètre d'échelle θ associé à des observations de la vitesse du vent, cette estimation peut nous donner une bonne idée de la valeur moyenne de cette vitesse (et de sa variance aussi qui est proportionnelle à θ^2).

L'objectif de ce BE est tout d'abord d'étudier certaines propriétés d'un estimateur du paramètre θ (pour p connu) construit à partir de N données $y_1, \dots, y_N$ de loi de Weibull $\mathcal{W}(\theta, p)$. Dans un second temps, on s'intéressera à construire un test d'hypothèses pour décider à partir des observations si on est en présence de vent calme ou de vent fort. Enfin, on analysera des données réelles en appliquant tout ce qui aura été construit auparavant.

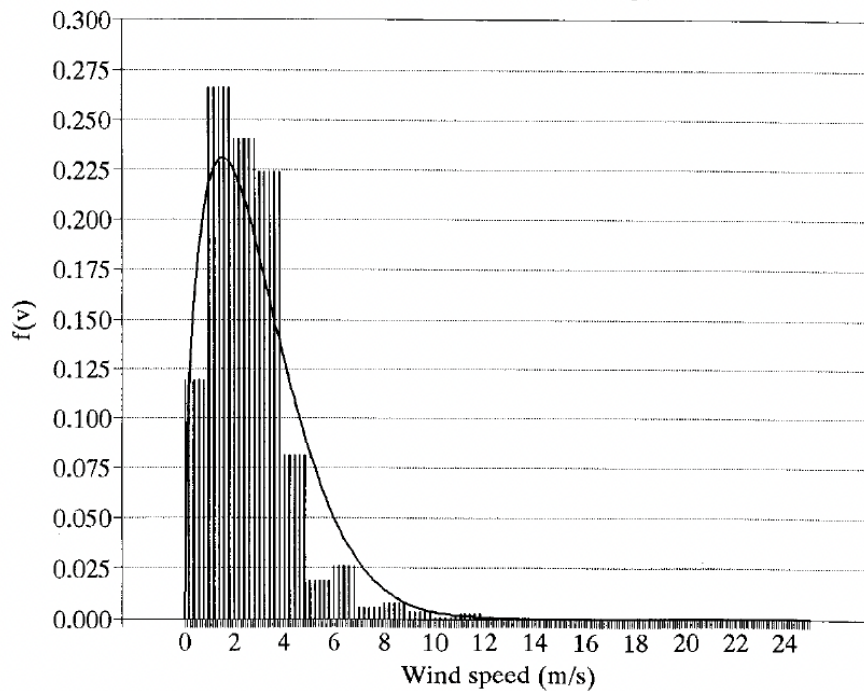


FIGURE 1 – Histogramme de la vitesse du vent comparé à la densité de Weibull (figure extraite de [3]).

Travail à effectuer

3.1 Génération de signaux test

Pour valider les algorithmes d'estimation et de détection développés dans ce TP/TDM, on souhaite générer K réalisations d'un signal de taille N (un N -échantillon) $\mathbf{y} = (y_1, \dots, y_N)^T$ de loi de Weibull $\mathcal{W}(\theta, p)$, dans laquelle θ est appelé paramètre d'échelle et p paramètre de forme (voir en Annexe 4.1 pour les caractéristiques de cette loi). Pour cela, on va utiliser la fonction Matlab¹ `wblrnd` pour générer une matrice $\mathbf{Y}$ de taille $N \times K$, dont chaque colonne contient une réalisation du signal de taille N .

1. Implanter cette génération et générer une matrice de données avec $N = 10000$, $K = 500$, $\theta_0 = 3.3$ et $p = 8$. Tracer les moyennes et variances estimées des colonnes de $\mathbf{Y}$ à l'aide des fonctions `mean` et `var`. Comparer avec les résultats théoriques attendus.
2. Vérifier que l'histogramme des données générées d'une colonne choisie au hasard est en accord avec la densité de la loi de Weibull à l'aide de la fonction `histfit`.

3.2 Estimation statistique

En utilisant les observations faites du phénomène aléatoire qui suit une loi de Weibull, on souhaite estimer le paramètre d'échelle θ (p est supposé connu, ici supposé $p = 8$).

Estimateur du maximum de vraisemblance : on montre que l'estimateur du maximum de vraisemblance s'écrit (voir Annexe 4.2) :

$$\hat{\theta}_{MV} = \left(\frac{1}{N} \sum_{i=1}^N Y_i^p \right)^{1/p}$$

Cet estimateur est asymptotiquement sans biais, convergent et asymptotiquement efficace (son étude est mission impossible...)

1. Implanter l'estimateur $\hat{\theta}_{MV}$ pour chacune des K réalisations des données générées. On obtient alors K valeurs de $\hat{\theta}_{MV}$, notées $\left(\hat{\theta}_{MV}(k) \right)_{k=1, \dots, K}$. Les représenter en les comparant à la valeur théorique attendue.
2. Comparer les propriétés des estimations faites avec les propriétés attendues de l'estimateur (biais, variance).
3. Comment illustrer empiriquement grâce à ces simulations que cet estimateur est convergent ?
4. Comparer à la borne de Cramér-Rao donnée dans l'Annexe 4.2. Commentaires.

1. Pour chaque fonction Matlab, n'hésitez pas à taper `help nom-fonction` pour en comprendre tous les paramètres et à cliquer sur la documentation en ligne correspondante.

Estimateur construit par la méthode des moments Un autre estimateur $\hat{\theta}_{Mom}$ peut facilement être trouvé à partir des caractéristiques de la loi (méthode des moments). Il s'écrit (voir Annexe 4.2) :

$$\hat{\theta}_{Mom} = \frac{1}{\Gamma\left(1 + \frac{1}{p}\right)} \frac{1}{N} \sum_{i=1}^N Y_i$$

5. L'implanter et estimer la valeur de θ pour chacune des K réalisations des données générées. On obtient alors K valeurs de $\hat{\theta}_{Mom}$, notées $\left(\hat{\theta}_{Mom}(k)\right)_{k=1,\dots,K}$. Les représenter en les comparant à la valeur théorique attendue.
6. Comparaison des deux estimateurs : dans un premier temps, comparer les estimations, leurs moyennes et leurs variances. Pour analyser plus en détail et comparer les deux estimateurs en termes de variance, tracer la variance de $\hat{\theta}_{Mom}$ et celle de $\hat{\theta}_{MV}$ en fonction du nombre d'observations du phénomène aléatoire N , ainsi que la borne Cramér-Rao. Qu'en concluez-vous ?

3.3 Détection

Attention : dans le contexte de données de vitesses de vent, on considère $\theta_0 = 0.9$ et $\theta_1 = 1.1$, avec, $p = 1.5$ (différent de la partie précédente).

On cherche à étudier les performances d'un test statistique permettant de détecter si les données y_i correspondent à un vent calme, défini comme ayant une vitesse moyenne inférieure à $v_{seuil} = 1ms^{-1}$ [3] ou à un vent fort. Ceci se traduit sur le paramètre d'échelle par $\theta = \theta_0 < v_{seuil}/\Gamma\left(1 + \frac{1}{p}\right)$ ou $\theta = \theta_1 > v_{seuil}/\Gamma\left(1 + \frac{1}{p}\right)$, θ_0 et θ_1 étant supposés fixés (voir les propriétés de la loi de Weibull en section 4.1).

Prendre cette décision revient à décider entre les deux hypothèses :

$$H_0 : \quad \theta = \theta_0 \quad H_1 : \quad \theta = \theta_1$$

sur la base des observations faites.

Pour choisir entre ces deux hypothèses en donnant des probabilités de fausse alarme et de non détection, il faut construire le **test de Neyman-Pearson**. Les deux hypothèses associées à ce test s'écrivent en se ramenant à un test sur le paramètre $a = \theta^p$ (plus facile à construire) :

$$H_0 : \quad a = a_0 \quad H_1 : \quad a = a_1 > a_0 \quad \text{avec } a_0 = \theta_0^p \text{ et } a_1 = \theta_1^p \quad (1)$$

On montre (voir Annexe 4.3) que la statistique de test issue du théorème de Neyman-Pearson associée à ces deux hypothèses s'écrit

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

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

$$R_\alpha = \{ \mathbf{y} \in \mathbb{R}^N | T(\mathbf{y}) > \lambda_\alpha \}.$$

alors le seuil de décision s'écrit

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

où G_{2N}^{-1} est l'inverse de la fonction de répartition d'une loi du chi2 à $2N$ degrés de liberté (une loi χ_{2N}^2). De même, la probabilité de non-détection (ou risque de 2ème espèce) du test s'exprime sous la forme suivante

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

où G_{2N} est la fonction de répartition d'une loi du chi2 à $2N$ degrés de liberté.

Dans ce qui suit, on souhaite tracer les courbes théoriques de la puissance du test $\pi = 1 - \beta$ en fonction de la probabilité de fausse alarme α (courbes COR), puis retrouver ces courbes par simulations.

1. En utilisant les fonctions `chi2inv` et `chi2cdf`, calculer la puissance théorique `pi_theorique` du test pour un risque $\alpha \in \{0.01, 0.02, \dots, 0.98, 0.99\}$.
 - Tracer les courbes COR obtenues pour les valeurs de a_0 et a_1 correspondant aux valeurs $\theta_0 = 0.9$ et $\theta_1 = 1.1$ et différentes valeurs de N ($N \in \{10, 20, 50\}$). Commenter les résultats obtenus.
 - Tracer les courbes COR obtenues pour a_0 correspondant à $\theta_0 = 0.9$, $N = 50$ et différentes valeurs de a_1 correspondant à différentes valeurs de θ_1 ($\theta_1 \in \{1.1, 1.2, 1.5\}$). Commenter les résultats obtenus.
2. On cherche maintenant à retrouver les courbes COR par simulation car dans d'autres applications, il peut être compliqué d'avoir les expressions théoriques. Pour cela, on va se donner un vecteur de valeurs de seuil et pour chaque valeur de seuil, on va *estimer* la probabilité de fausse alarme et la puissance du test, ce qui va permettre de construire la courbe COR par simulation.
 - Générer K réalisations de signaux de longueur N associés à l'hypothèse H_0 du test (1).
 - Générer K réalisations de signaux de longueur N associés à l'hypothèse H_1 du test (1).
 - Calculer les $2K$ valeurs possibles de la statistique de test T associées à ces $2K$ réalisations et construire un vecteur de seuil couvrant ces $2K$ valeurs possibles (on pourra utiliser les fonctions matlab `min` et `max`).

- Pour chaque valeur de seuil, estimer la probabilité de fausse alarme (à partir des réalisations sous H_0) et la puissance du test (à partir des réalisations sous H_1).
- Tracer les courbes COR estimées en les comparant aux courbes COR théoriques calculées à la question 2 pour les valeurs de a_0 et a_1 correspondant aux valeurs $\theta_0 = 0.9$ et $\theta_1 = 1.1$ et pour $N = 10, 20$ et 50 .
- Etudier l'influence de K . Commenter.

3.4 Analyse d'un fichier de données

On désire dans cette partie analyser un fichier de données contenant des mesures de vitesse de vent.

1. Charger le fichier `wind.mat` et représenter graphiquement les mesures de vitesse de vent contenues dans le vecteur `test`.
2. À l'aide de la fonction `histfit`, comparer l'histogramme des données à celui de la loi de Weibull la plus proche. Commenter.
3. On désire vérifier qu'il est raisonnable de penser que les données du vecteur `test` sont distribuées suivant une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$ à l'aide d'un test de Kolmogorov. Pour cela,
 - À l'aide de la fonction `wblfit`, déterminer des estimées des paramètres θ et p associées à ce vecteur de données (notées $\hat{\theta}$ et $\hat{p}$) obtenues à l'aide de la méthode du maximum de vraisemblance.
 - Représenter sur la même figure la fonction de répartition de la loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$ évaluée aux données du fichier `wind.mat` (rangées par ordre croissant avec la fonction `sort` et notées $y_1 < y_2 < \dots < y_N$) et la fonction de répartition empirique de ces données estimée avec la fonction `ecdf` et représentée à l'aide de la fonction `stairs` (tout d'abord en ne considérant que $N = 100$ données pour mieux voir puis avec la totalité du fichier).
 - Écrire une fonction qui permet de calculer les écarts E_i^+ et E_i^- définis par

$$E_i^+ = \left| \frac{i}{N} - F_W(y_i; \hat{\theta}, \hat{p}) \right|, \quad E_i^- = \left| \frac{i-1}{N} - F_W(y_i; \hat{\theta}, \hat{p}) \right|$$

pour $i = 1, \dots, N$, où N est le nombre de données x_i du vecteur `test` et $F_W(x_i; \hat{\theta}, \hat{p})$ est la fonction de répartition d'une loi de Weibull $\mathcal{W}(\hat{\theta}, \hat{p})$ (fonction `wblcdf`). En déduire la valeur de la statistique du test de Kolmogorov. Pour conclure, on utilisera l'approximation du seuil du test de Kolmogorov dans le cas où $N > 100$:

$$P \left[\sup_x \left| \widehat{F}_N(x) - F(x) \right| > \frac{1.36}{\sqrt{N}} \right] = 5\% \text{ et } P \left[\sup_x \left| \widehat{F}_N(x) - F(x) \right| > \frac{1.63}{\sqrt{N}} \right] = 1\%$$

- Vérifier le résultat de la question précédente en utilisant la fonction `kstest` de Matlab. Ne pas hésiter à utiliser `help kstest` pour connaître les entrées et sorties de cette fonction.
- 4. Appliquer le test de détection fait à la question précédente pour décider si ces données proviennent d'un vent fort ou d'un vent calme.

Références

- [1] A. C. Cohen. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. *Technometrics*, 7(4) :579–588, 1965.
- [2] A. Joarder, H. Krishna, and D. Kundu. Inferences on Weibull parameters with conventional type-I censoring. *Comput. Stat. Data Anal.*, 55(1) :1–11, 2011.
- [3] M. Martin, L. V. Cremades, and J. M. Santabàrbara. Analysis and modelling of time series of surface wind speed and direction. *Journal of Climatology*, 19(2) :197–209, 1999.

4 Annexes

4.1 Lois de Weibull, du Chi2 et Gamma

- **Loi de Weibull** $\mathcal{W}(\theta, p)$ $p > 0, \theta > 0, x \in \mathbb{R}^+$
 - **Paramètres** : θ paramètre d'échelle (scale parameter) et p paramètre de forme (shape parameter).
 - **Densité de probabilité** : $f(x; \theta, p) = \frac{p}{\theta} \left(\frac{x}{\theta}\right)^{p-1} \exp \left[-\left(\frac{x}{\theta}\right)^p\right] \mathcal{I}_{\mathbb{R}^+}(x)$, avec $\mathcal{I}_{\mathbb{R}^+}$ la fonction indicatrice sur $\mathbb{R}^+$
 - **Fonction de répartition** : $F(x; \theta, p) = 1 - \exp \left[-\left(\frac{x}{\theta}\right)^p\right] \mathcal{I}_{\mathbb{R}^+}(x)$
 - **Moyenne** : $\mu = \theta \Gamma \left(1 + \frac{1}{p}\right)$
 - **Variance** : $\theta^2 \Gamma \left(1 + \frac{2}{p}\right) - \mu^2$

$\Gamma(\cdot)$ représente ici la fonction Gamma : $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt$ et sous Matlab, cette fonction est directement accessible par `gamma(z)`.
- **Loi du Chi2 à L degrés de liberté** χ_L^2 $L > 0, x \in \mathbb{R}^+$
 - **Définition** : si $X_1, \dots, X_L$ sont L variables aléatoires indépendantes de même loi normale $\mathcal{N}(0, 1)$ alors la loi de $\sum_{l=1}^L X_l^2$ est une loi du χ_L^2
 - **Densité de probabilité** : $f(x; L) = \frac{1}{2^{\frac{L}{2}} \Gamma(\frac{L}{2})} x^{\frac{L}{2}-1} \exp \left(-\frac{x}{2}\right) \mathcal{I}_{\mathbb{R}^+}(x)$
 - **Moyenne** : L
 - **Variance** : $2L$
- **Loi Gamma** $\mathcal{G}(\nu, \theta)$ $\nu > 0, \theta > 0, x \in \mathbb{R}^+$
 - **Paramètres** : ν paramètre de forme (shape parameter) et θ paramètre d'intensité (on peut aussi utiliser un paramètre d'échelle $b = 1/\theta$, c'est le paramètre utilisé dans Matlab).
 - **Densité de probabilité** : $f(x; \nu, \theta) = \frac{\theta^\nu}{\Gamma(\nu)} \exp(-\theta x) x^{\nu-1} \mathcal{I}_{\mathbb{R}^+}(x)$
 - **Moyenne** : $\frac{\nu}{\theta}$
 - **Variance** : $\frac{\nu}{\theta^2}$
 - **Remarque 1** : Une loi du Chi2 χ_L^2 est une loi Gamma $\mathcal{G}(\frac{\nu}{2}, \frac{1}{2})$
 - **Remarque 2** : Si $X \sim \mathcal{G}(\nu, \theta)$ alors $tX \sim \mathcal{G}(\nu, \frac{\theta}{t})$ pour $t > 0$.

4.2 Estimation des paramètres de la loi de Weibull

4.2.1 Estimateur du maximum de vraisemblance du paramètre d'échelle

Le paramètre de forme p étant supposé connu, la construction de l'estimateur du maximum de vraisemblance du paramètre d'échelle θ d'une loi de Weibull $\mathcal{W}(\theta, p)$ se fait en cherchant le paramètre qui maximise la log-vraisemblance. Toutefois, il peut être préférable de re-paramétriser la densité en posant $a = \theta^p$ (la dérivée est moins longue à écrire et plus facile à calculer!). La log-vraisemblance s'écrit alors

$$\ln L(y_1, \dots, y_N; a, p) = N \ln \left(\frac{p}{a} \right) + (p-1) \sum_{n=1}^N \ln y_n - \frac{1}{a} \sum_{n=1}^N y_n^p$$

et on cherche le paramètre a qui maximise la log-vraisemblance via la dérivée

$$\frac{\partial \ln L(y_1, \dots, y_N; a, p)}{\partial a} = -\frac{N}{a} + \frac{1}{a^2} \sum_{n=1}^N y_n^p$$

ce qui nous donne

$$\hat{a}_{MV} = \frac{1}{N} \sum_{i=1}^N Y_i^p.$$

En appliquant le principe de l'invariance fonctionnelle, on en déduit l'expression de l'estimateur du maximum de vraisemblance du paramètre d'échelle :

$$\hat{\theta}_{MV} = \left(\frac{1}{N} \sum_{i=1}^N Y_i^p \right)^{1/p}.$$

L'étude de son biais et de sa variance n'est pas simple : on se contentera des propriétés asymptotiques de l'estimateur du maximum de vraisemblance : asymptotiquement sans biais et efficace (et convergent).

4.2.2 Estimateur des moments du paramètre d'échelle

La méthode des moments permet de proposer facilement un estimateur du paramètre d'échelle θ . Le moment d'ordre 1 d'une loi de Weibull s'exprime par (voir Annexe 4.1)

$$\mu = \theta \Gamma \left(1 + \frac{1}{p} \right)$$

Un estimateur du paramètre d'échelle peut ainsi être proposé à partir de l'estimateur du moment d'ordre 1 :

$$\hat{\theta}_{Mom} = \frac{1}{\Gamma \left(1 + \frac{1}{p} \right)} \bar{Y}$$

avec $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$, estimateur de la moyenne dont on connaît les propriétés : sans biais et de variance égale à la variance des variables aléatoires divisée par N .

L'estimateur $\hat{\theta}_{Mom}$ est donc sans biais et de variance :

$$\text{Var}(\hat{\theta}_{Mom}) = \frac{1}{\left[\Gamma\left(1 + \frac{1}{p}\right)\right]^2} \theta^2 \frac{\Gamma\left(1 + \frac{2}{p}\right) - \left[\Gamma\left(1 + \frac{1}{p}\right)\right]^2}{N} = \left(\frac{\Gamma\left(1 + \frac{2}{p}\right)}{\left[\Gamma\left(1 + \frac{1}{p}\right)\right]^2} - 1 \right) \frac{\theta^2}{N}$$

L'estimateur des moments proposé est donc aussi convergent.

4.2.3 Borne de Cramér-Rao du paramètre d'échelle

La log-vraisemblance s'écrit :

$$\ln L(y_1, \dots, y_N; \theta, p) = N \ln \left(\frac{p}{\theta^p} \right) + (p-1) \sum_{n=1}^N \ln y_n - \frac{1}{\theta^p} \sum_{n=1}^N y_n^p$$

Sa dérivée seconde est

$$\frac{\partial^2 \ln L(y_1, \dots, y_N; \theta, p)}{\partial \theta^2} = \frac{Np}{\theta^2} - \frac{p(p+1)}{\theta^{p+2}} \sum_{n=1}^N y_n^p$$

En prenant en compte que $Z = \frac{2}{\theta^p} \sum_{i=1}^N Y_i^p$ suit une loi du Chi2 à $2N$ degrés de liberté (voir dans la section 4.3 pour la démonstration), on a

$$E \left[-\frac{\partial^2 \ln L(y_1, \dots, y_N; \theta, p)}{\partial \theta^2} \right] = -\frac{Np}{\theta^2} + \frac{p(p+1)}{2\theta^2} E[Z] = -\frac{Np}{\theta^2} + \frac{p(p+1)}{2\theta^2} 2N = \frac{Np^2}{\theta^2}$$

La variance de tout estimateur non biaisé de θ est toujours supérieure ou égale à la borne de Cramér-Rao telle qu'exprimée ci-dessous

$$\text{BCR}(\hat{\theta}) = \frac{\theta^2}{Np^2}$$

4.2.4 Estimateur du maximum de vraisemblance des deux paramètres

Pour estimer les deux paramètres d'une loi de Weibull par la méthode du maximum de vraisemblance [1, 2], il est préférable de re-paramétriser la densité en posant $a = \theta^p$. La log-vraisemblance s'écrit alors

$$\ln L(y_1, \dots, y_N; a, p) = N \ln \left(\frac{p}{a} \right) + (p-1) \sum_{n=1}^N \ln y_n - \frac{1}{a} \sum_{n=1}^N y_n^p$$

qui admet pour dérivées partielles

$$\begin{aligned}\frac{\partial \ln L(y_1, \dots, y_N; a, p)}{\partial a} &= -\frac{N}{a} + \frac{1}{a^2} \sum_{n=1}^N y_n^p \\ \frac{\partial \ln L(y_1, \dots, y_N; a, p)}{\partial p} &= \frac{N}{p} + \sum_{n=1}^N \ln y_n - \frac{1}{a} \sum_{n=1}^N (y_n^p \ln y_n).\end{aligned}\tag{2}$$

En annulant ces deux dérivées partielles, on montre facilement que p est solution de l'équation

$$\frac{1}{p} = h(p) = \frac{\sum_{n=1}^N (y_n^p \ln y_n)}{\sum_{n=1}^N y_n^p} - \frac{1}{N} \sum_{n=1}^N \ln y_n$$

qui est une équation de point fixe $p = \frac{1}{h(p)}$ qu'on peut résoudre 1) en trouvant une solution initiale suffisamment proche de la solution recherchée ou 2) à l'aide de l'algorithme de Newton-Raphson basé sur la récursion

$$p_{k+1} = p_k - \frac{g(p_k)}{g'(p_k)} \text{ avec } g(p) = p - \frac{1}{h(p)}.$$

4.3 Test de Neyman-Pearson pour la détection

A risque α fixé, le test qui maximise la puissance du test est défini par

$$\text{Rejet de } H_0 \text{ si } \frac{L(y_1, \dots, y_N; p|H_1)}{L(y_1, \dots, y_N; p|H_0)} > S_\alpha$$

En cherchant à maximiser le logarithme du rapport de vraisemblance et en utilisant l'expression de la vraisemblance donnée dans la section précédente, on a :

$$\text{Rejet de } H_0 \text{ si } N \ln \left(\frac{p}{a_1} \right) - \frac{1}{a_1} \sum_{n=1}^N y_n^p - N \ln \left(\frac{p}{a_0} \right) + \frac{1}{a_0} \sum_{n=1}^N y_n^p > S'_\alpha$$

soit

$$\text{Rejet de } H_0 \text{ si } \sum_{n=1}^N y_n^p > \lambda_\alpha \text{ en prenant en compte que } a_1 > a_0.$$

On montre ainsi que la statistique de test issue du théorème de Neyman-Pearson associée à ces deux hypothèses s'écrit

$$T(\mathbf{y}) = \sum_{i=1}^N Y_i^p.$$

Pour une probabilité de fausse alarme α , la région critique du test (zone de rejet de H_0) est définie par

$$R_\alpha = \{ \mathbf{y} \in \mathbb{R}^N | T(\mathbf{y}) > \lambda_\alpha \}.\tag{3}$$

Pour pouvoir relier le seuil de décision λ_α aux risques α et β , il faut connaître la loi de T . Dans un premier temps, on étudie la loi de $Z_i = \frac{2}{a}Y_i^p$, Y_i suivant une loi de Weibull $\mathcal{W}(\theta, p)$. En prenant en compte ce changement de variable, la densité de probabilité de Z_i s'écrit :

$$h(z) = f\left(\left[\frac{a}{2}z\right]^{\frac{1}{p}}\right) \left|\frac{1}{p}\left(\frac{a}{2}\right)^{\frac{1}{p}}z^{\frac{1}{p}-1}\right|$$

avec f la densité de probabilité d'une loi de Weibull $\mathcal{W}(\theta, p)$ avec $a = \theta^p$, ce qui conduit à :

$$h(z) = \frac{1}{2} \exp\left(-\frac{z}{2}\right).$$

On reconnaît la densité de probabilité d'une loi du Chi2 à 2 degrés de liberté, i.e., $Z_i \sim \chi_2^2$. En calculant la fonction caractéristique de $Z = \sum_{i=1}^N Z_i$, on montre ensuite facilement que $Z = \sum_{i=1}^N Z_i = \frac{2}{a} \sum_{i=1}^N Y_i^p = \frac{2}{a} T$ suit une loi du Chi2 à $2N$ degrés de liberté. Ainsi, à risque α fixé, on a :

$$\alpha = P[T(\mathbf{y}) > \lambda_\alpha | H_0] = P\left[Z > \frac{2}{a_0} \lambda_\alpha | Z \sim \chi_{2N}^2\right]$$

Et le seuil de décision s'écrit

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

où G_{2N}^{-1} est l'inverse de la fonction de répartition d'une loi du chi2 à $2N$ degrés de liberté. Et la probabilité de non-détection (ou risque de 2ème espèce) du test s'exprime sous la forme suivante

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

où G_{2N} est la fonction de répartition d'une loi du chi2 à $2N$ degrés de liberté.

4.4 Article de Martin (1999)

(voir page suivante)

ANALYSIS AND MODELLING OF TIME SERIES OF SURFACE WIND SPEED AND DIRECTION

M. MARTÍN, L.V. CREMADES* and J.M. SANTABÀRBARA

Departament de Projectes d'Enginyeria, Universitat Politècnica de Catalunya (UPC), ETSEIB, Av. Diagonal, 647 planta 10, 08028 Barcelona, Spain

Received 7 August 1997

Revised 10 July 1998

Accepted 24 July 1998

ABSTRACT

This work presents a climatological study of winds at a particular site and proposes a simple model to simulate data time series of horizontal surface wind. The model relies on three major hypotheses: (i) speed and direction are treated as independent variables; (ii) wind can be expressed as the sum of a deterministic component (including trend and cyclical variations) plus a probabilistic or stochastic component; and (iii) both components (deterministic and stochastic) are independent. An hourly based 7 year wind series measured at the meteorological station in Almazora, located at the Spanish Mediterranean coast, is analyzed. Only the deterministic component of this series is simulated in accordance with the model proposed. Results indicate that the model is especially suitable for predicting wind at sites in which some cyclic pattern is expected. Copyright © 1999 Royal Meteorological Society.

KEY WORDS: surface wind; deterministic model; time series; wind speed and direction; harmonic analysis; Mediterranean coast

1. INTRODUCTION

Time series analyses of meteorological data have an increasing interest in many fields. Series are especially interesting for a better understanding of atmospheric phenomena in order to model them, to determine the climate of a geographical area, or to forecast possibilities of occurrence of some extreme situations. Therefore, the analysis of meteorological data series is useful in fields such as agriculture, air and sea traffic control, structural engineering calculations, global change studies, solar and wind resources estimation, etc.

Up to the present, this analysis has been carried out from several points of view: (i) simple characterization of the statistical distribution of the data series; (ii) harmonic analysis, modeling from the harmonic analysis; (iii) fitting to a model or probabilistic process such as AR, ARMA or ARIMA; (iv) adjustment to a relatively simple mathematical expression depending on physical magnitudes related to the data series; and (v) models that include differential equations establishing relations between the variables considered.

The aim of this paper is to present an approximation to the time series of wind speed and direction (treated as separated series), mainly studied by means of harmonic analysis, in order to identify trend (if any) and cyclic components. The final goal is to model wind data series, to reproduce the whole of the observed series and predict future values. Some particular applications are:

- (i) To obtain estimations of wind speed and direction at a location where there are historical observations, but current measurements are not available. These estimations could be used as input data for a diagnostic model of air pollutants dispersion in real time.

* Correspondence to: Departament de Projectes d'Enginyeria, Universitat Politècnica de Catalunya (UPC), ETSEIB, Av. Diagonal, 647 planta 10, 08028 Barcelona, Spain.

- (ii) To reduce the volume of data, because the information is summarized in one (or several) mathematical formulae (Coronas and Baldasano, 1984; Phillips, 1984).
- (iii) To check wind measurements obtained at the meteorological stations, in order to detect when observed data abnormally disagree from the expected behavior, hence making a recalibration of sensors necessary.

The proposed model does not directly depend on physical or geographical variables. Due to the complexity of factors and mechanisms that influence and steer the wind, which is highly variable both in space and time, the study has focused on getting a statistical model that describes the time variation of wind values in accordance with some cyclic patterns. This statistical model uses as input only the available wind data and assumes that future behavior should be the same as the one measured in the past. For this reason, the proposed model, apart from the suitability of the fitting obtained, is only valid for the site (meteorological station) to which the wind data series belong.

As an example of application, the methodology is applied to a 7 year wind data series of the surface meteorological station of Almazora (Castellón, Spain).

2. BACKGROUND

Some other works (Ishida, 1990; Brett and Tuller, 1991; Frisch *et al.*, 1991; Gavalda *et al.*, 1992; Schumann, 1992) tried to characterize the cyclic variations of wind speed and direction time series (not only to ground level data) by spectral analysis. Wickle *et al.* (1995) used the information given by the analysis to model the deterministic component of the series. Most of the authors filter the data to emphasize a particular range of frequencies. In this study, the authors work instead with hourly based series in search of all sorts of available cycles of the underlying process, only limited by the discretization of data.

Regarding the stochastic component of the series, Wickle *et al.* (1995) modeled this part by means of an autoregressive AR(1) process. Some other authors (Brown *et al.*, 1984; Breckling, 1989; Daniel and Chen, 1991; Fisher and Lee, 1994) take from the beginning the data as a realization of a stationary random process (AR, ARMA or ARIMA) or transform data to remove non-stationarity before modeling.

3. METHODOLOGY

The first decision taken at the beginning of this work was whether to deal either with a wind module and direction, assumed to be two different and independent time series, or with the components of the horizontal wind vector. One of the advantages of the former is that this is the usual way to record wind data by the traditional sensors (anemometer and vane). Furthermore, this way is more intuitive and it is easier to interpret speed and direction separately than with horizontal components. On the contrary, circular variables statistics (Mardia, 1972) are more complex than linear magnitudes statistics. However, because of its advantages, the first option was chosen.

Following the traditional approach, the model built for both separated series describes the variation of the observed variable as composed of a deterministic component (fully predictable) and a stochastic component (partially predictable). Trend and cyclic variations are included in the first component. Remaining irregular fluctuations are included in the second one. In this work, focus will just be on the deterministic component. Regarding the stochastic component, only stationarity is checked.

Figure 1 shows the scheme of the process followed in this work to analyze and model the wind data series. Previous steps include calculating some basic statistical parameters for both wind speed and direction and obtaining the distribution function from the full series of observations. The distribution function applied to wind speed data is widely used in wind energy assessment studies because it is not necessary to know when the different speeds occur, but the frequency of occurrence and whether they are

high enough. This in fact, (i) represents a static model that cannot describe the time evolution of wind and (ii) assumes that the observations are independent (although they are actually correlated), which implies that variances and confidence intervals of estimators calculated are erroneous. To surmount these difficulties, a dynamic model based on the statistical theory of time series analysis is needed.

Secondly, calculation of autocorrelation coefficients and different data plots allow us to identify patterns, cyclic or not, observed in the time series. Application of Fourier analysis (Jenkins and Watts, 1968) allows the identification of the cyclic patterns (i.e. the frequencies corresponding to the main

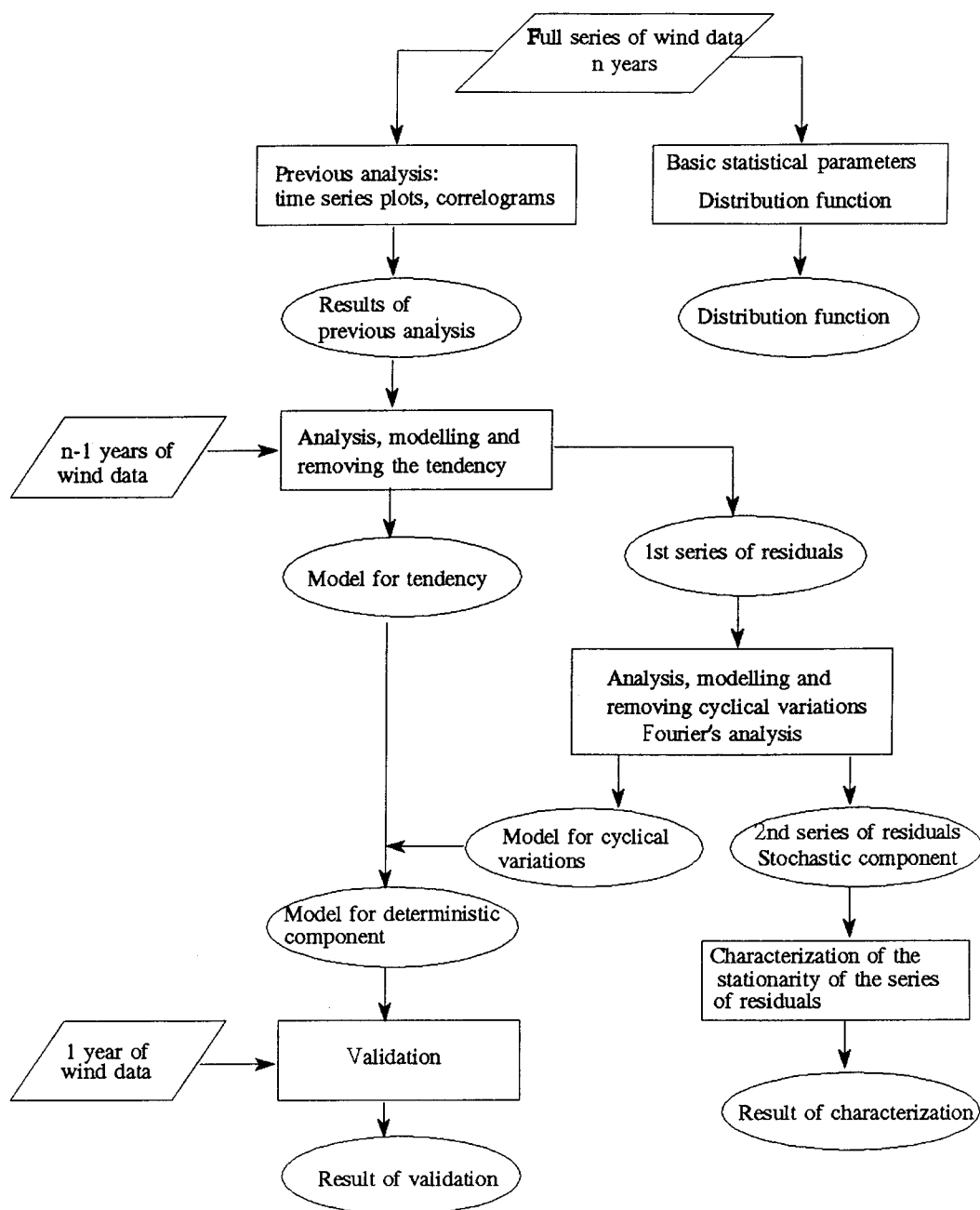


Figure 1. Scheme of the process followed in this work to analyse and model wind data series

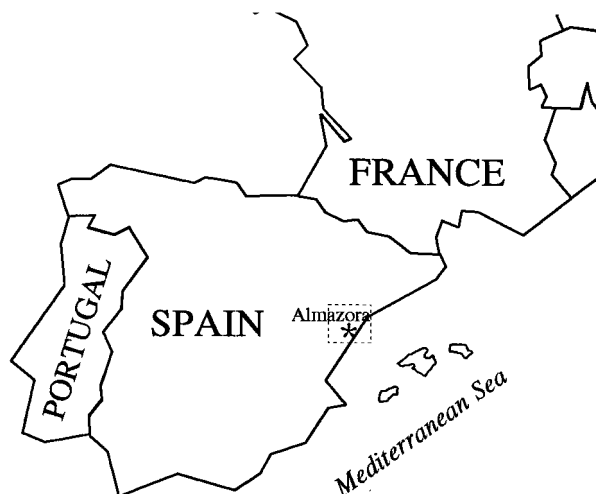


Figure 2. Location of the meteorological station in Almazora (Castellón, Spain)

harmonics) in the individual annual series and in the full series. Relevant frequencies to be considered in the modeling should be the same for the whole set of observations and for the annual subsets.

Then, the authors try to model the deterministic component of the series, studying the speed and the direction separately; first the trend, if any, and after removing it, the cyclic variations. One unique expression for the cyclic variations (valid for all the years, as no cycle longer than 1 year is expected) is extracted from the previously obtained information, whose coefficients are fitted to the series of observations, except for 1 year of data, which is reserved to validate the model. The model obtained (trend + cyclic variations) estimates the most probable value of wind for any time at the site where the data series were measured. Finally, if the series calculated by the model is subtracted from the original data series, a series of residuals is obtained, the stationarity of which is analyzed.

4. EXAMPLE OF APPLICATION: ALMAZORA

Data used in this study consists of 7 years of hourly wind observations (61 368 records) provided by the Instituto Nacional de Meteorología (INM) measured at Almazora (Castellón, Spain). Wind speed was measured as an hourly average in m s^{-1} with a resolution of tenths, and dominant wind direction during every hour expressed as the central angle ($^{\circ}$) of the sector (among 16 possible sectors). Measurement period is from January 1, 1983 to December 31, 1989. Sensors are at the height of 10 m asl.

Almazora town is located at 5 km from the Mediterranean shoreline (see Figure 2). It is in the middle of a great sedimentary plain (Castellón Plain) and close to the Mijares River. Its co-ordinates are $0^{\circ}3'W$ longitude, $39^{\circ}56'N$ latitude and 32 m altitude. The nearest mountains are 10 km away from the town: a coastal mountain range oriented NE–SW and heights up to 1000 m asl, and another range located inland, whose major orientation is NW–SE reaching 2000 m altitude. The Mijares River follows a NW–SE path down to the Castellón Plain.

Performance of the proposed model highly depends on the quality of the wind data series used in its development. Among the factors that influence the quality of data, the following can be mentioned: average period of measurements (frequency), percentage of useful available data over the potential data that could have been measured, and the quality of sensors. Furthermore, data show good agreement with wind climatology expected for that site as stated in the literature (Font Tullot, 1983; Bosque and Vilà, 1989, 1992; Carreras, 1992). Although the length of the series is shorter than desirable, the percentage of useful data is 100% of the total available data.

4.1. Data characterization

Main winds blow from the second and fourth quadrants alternately throughout the year. Prevailing wind directions are NW and WNW, followed by ESE, E, SSE and SE, due to the channeling induced by the Mijares valley and orography. These six directions comprise 64.4% of the observations. Wind speed values $< 1 \text{ m s}^{-1}$ are below the anemometer threshold, and are considered as calms. They represent 11.8% of the whole of available data.

Daily wind variation is very clear, especially from March through to September. It can be seen, for instance, in Figure 3(a), how the hourly mean speed during August is at a minimum at 21:00 h solar time (ST), increases to reach the maximum at 14:00 h ST, and then decreases. Simultaneously, the dominant

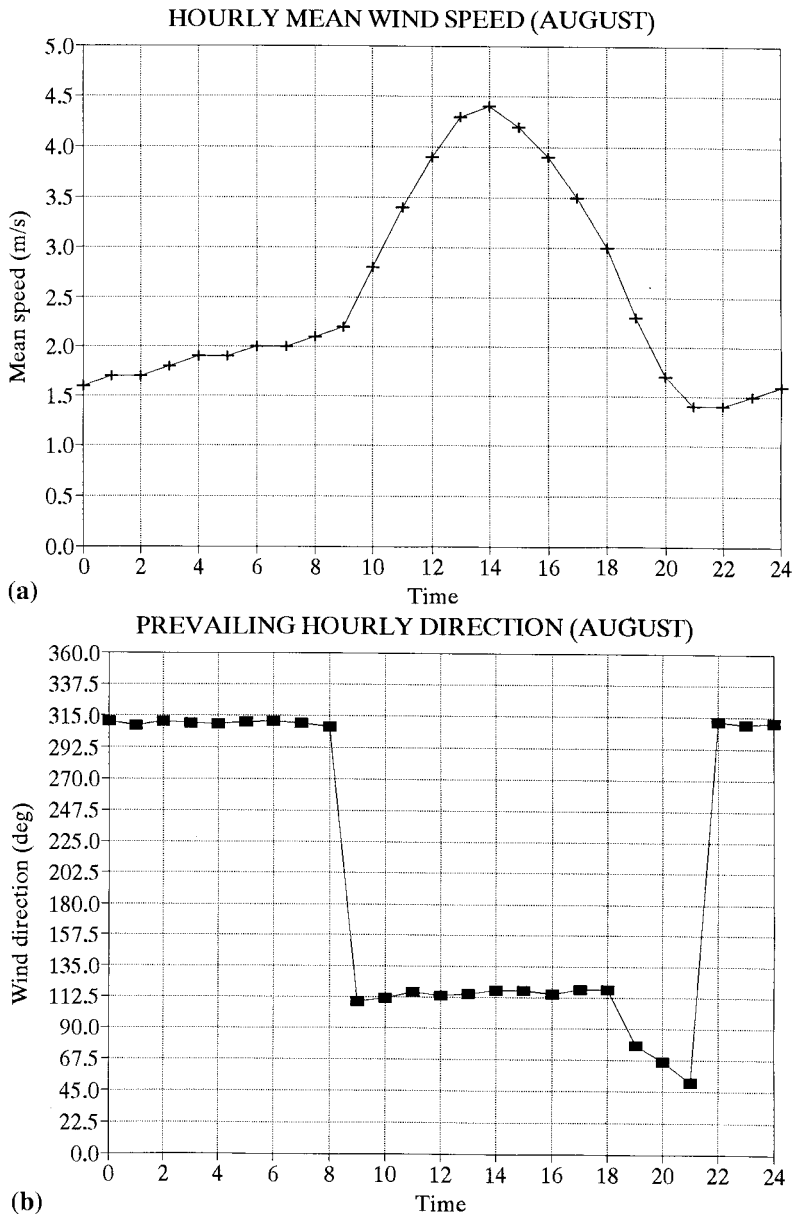


Figure 3. Hourly mean wind speed (a) and most probable direction (b) during the month of August. In (a), calm is considered as a 0.5 m s^{-1} wind speed

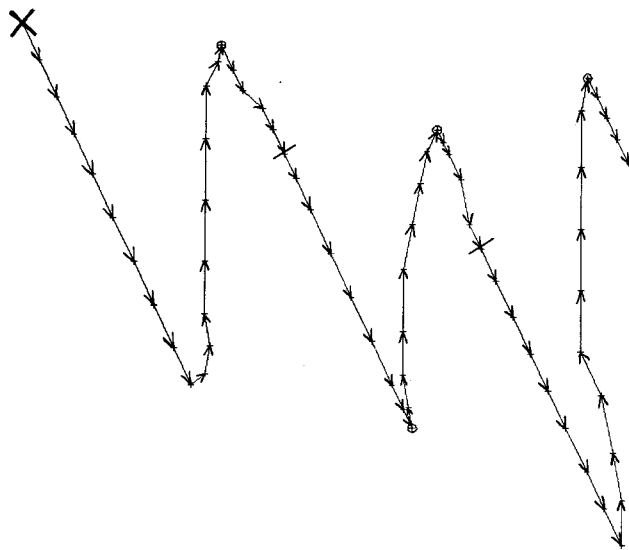


Figure 4. Representation of the hourly wind vector throughout 3 consecutive days from Almazora data series. Consecutive vectors are plotted one after the other. The series starts at the upper left corner (X) following the direction marked by the arrows. The change of day is indicated by a line perpendicular to vectors. Calms are represented by a circle. Vector length is proportional to wind speed

direction is ESE (112.5°) in the second quadrant, from 9:00 h ST until 18:00 h ST. Then it changes to be NNW (315°) in the fourth quadrant, from 22:00 h ST until 8:00 h ST (next day) (see Figure 3(b) and Figure 4). The same mean pattern is observed during most of the year, although the maximum speed and the time vary when the second quadrant winds are dominating. There is strong evidence that wind speed and direction are strongly correlated, as expected with a land–sea breeze situation.

Finally, Figure 5(a) and (b) show the annual and monthly mean speeds for the whole period of record. There is a clear decreasing trend, probably due to the growth of the number of buildings around the meteorological station through the years. These buildings belong to an industrial estate and their heights are lower than the anemometer height. The nearest buildings are about 30 m away from the mast. Nowadays, there are very few free building plots, and the surroundings of the meteorological station can be considered as an isotrope.

The presence of calms leads to the appearance of gaps in the speed and direction data series, which makes the calculation of the statistical properties of the series and the modeling difficult. To solve it, the authors assumed that a calm had an associated wind speed of 0.5 m s^{-1} . Taking into account this hypothesis, mean wind speed from the whole data series is 2.7 m s^{-1} and the standard deviation (S.D.) is 1.8 m s^{-1} . The histogram of the speed series appears in Figure 6, with data grouped in intervals of 1 m s^{-1} . It can be seen that the distribution of frequencies is unimodal and asymmetric, which is in agreement with the Weibull distribution:

$$f(v) = \frac{k}{A} \left(\frac{v}{A} \right)^{k-1} \exp(- (v/A)^k), \quad (1)$$

with $A = 3.3 \text{ m s}^{-1}$ and $k = 1.5$, found by least-square fitting of the cumulative distribution function.

The problem of calms in the case of wind direction is solved by linear interpolation between valid values. However, the direction value that replaces the calm in the series is not the one calculated as before, but the central direction corresponding to the sector to which the calculated direction belongs. In this way, coherence with the rest of the series is kept. In this case, the frequency distribution is bimodal, where the most frequent sectors are ESE (112.5°) and NW (315°) (Figure 7). Since the gap between both modes is 202.5° , the authors have considered this bimodal distribution as two unimodal distributions separated

by 180° (Mardia, 1972). Taking into account that wind direction is a circular variable and that the frequency distribution is bimodal, the mean direction calculated for the first modal component is 121.9° , while for the second one it is 301.9° . For both, the S.D. is 42.1° . A theoretical bimodal Von Mises' distribution has been fitted to the direction data series, making $\mu_\theta = 121.9^\circ$ and calculating λ and κ by means of minimizing graphically the squared differences between the theoretical distribution and the frequency one, i.e.

$$g(\theta) = \frac{1}{2\pi I_0(\kappa)} [\lambda \exp(\kappa \cos(\theta - \mu_\theta)) + (1 - \lambda) \exp(-\kappa \cos(\theta - \mu_\theta))], \quad (2)$$

with $\kappa = 3$ and $\lambda = 0.36$; $I_0(\kappa)$ is the zero-order modified Bessel function.

Comparing observed frequencies with those obtained by Equations (1) and (2), by using the Pearson χ^2 test, the hypotheses that these equations describe the observed wind speed and direction, respectively, are fulfilled.

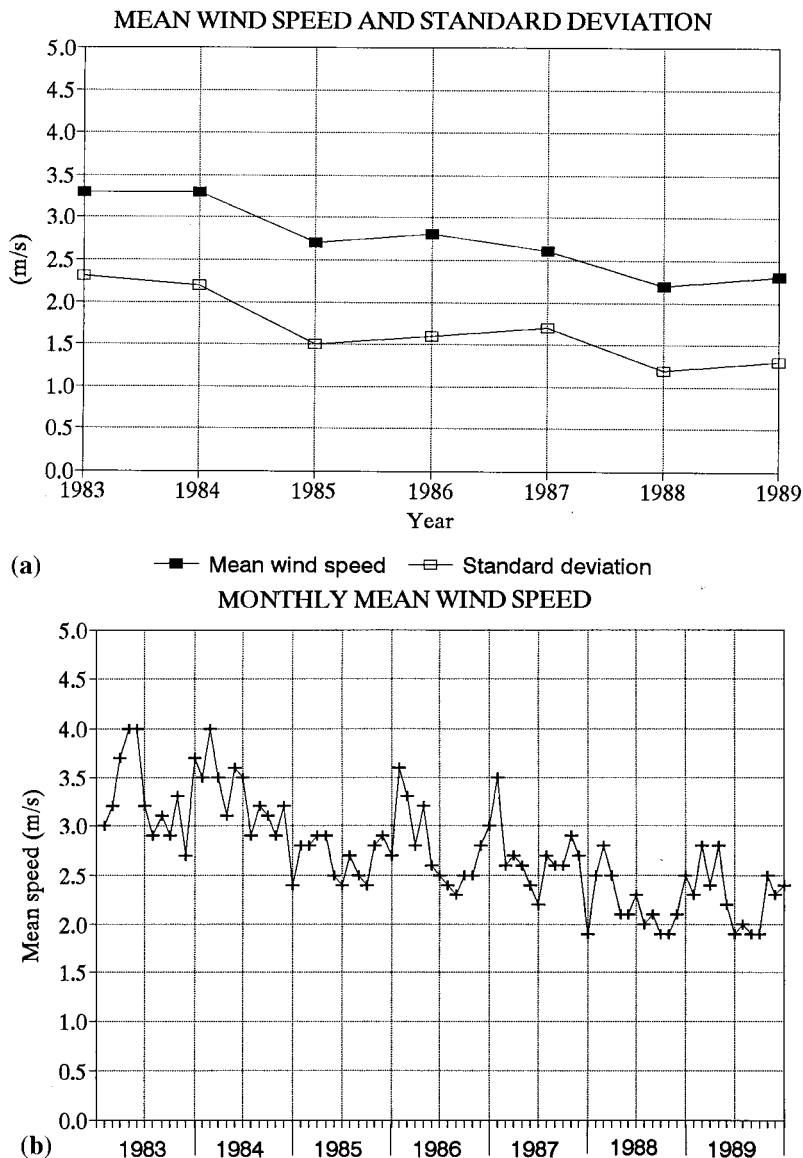


Figure 5. Annual (a) and monthly (b) mean wind speed from January 1983 through to December 1989

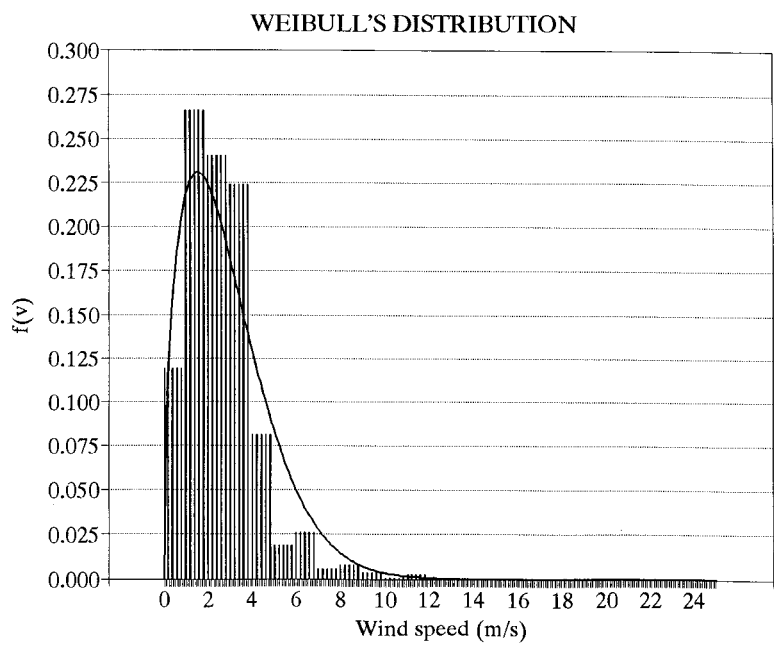


Figure 6. Histogram of wind speed and Weibull distribution that fits the data

Autocorrelation coefficients plots (correlograms) of speed and direction series are shown in Figure 8. The calculation of the line spectrum and the autocorrelation function for the direction series implies that the meteorological wind angle θ as a complex number, in which the x -component of the unit vector with angle θ is the real part and the y -component, the imaginary part. The transformation used is the following:

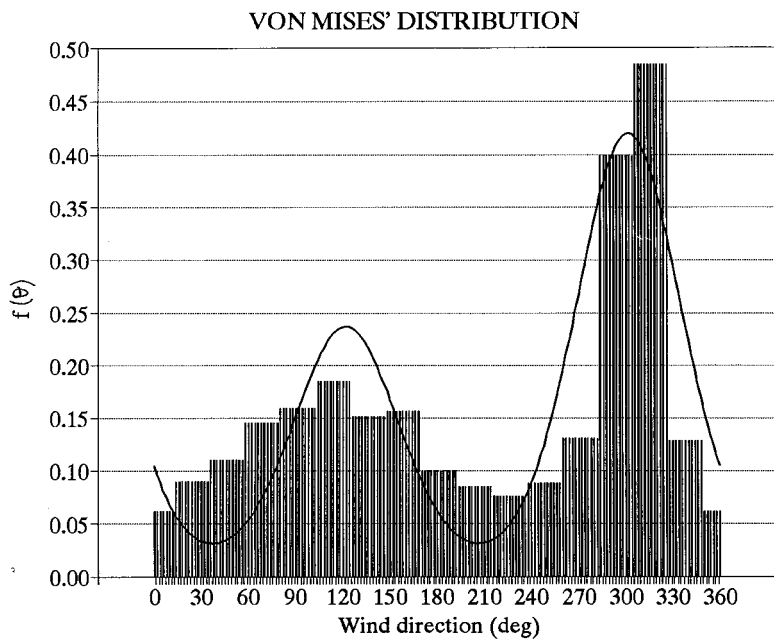


Figure 7. Histogram of wind direction and Von Mises distribution that fits the data

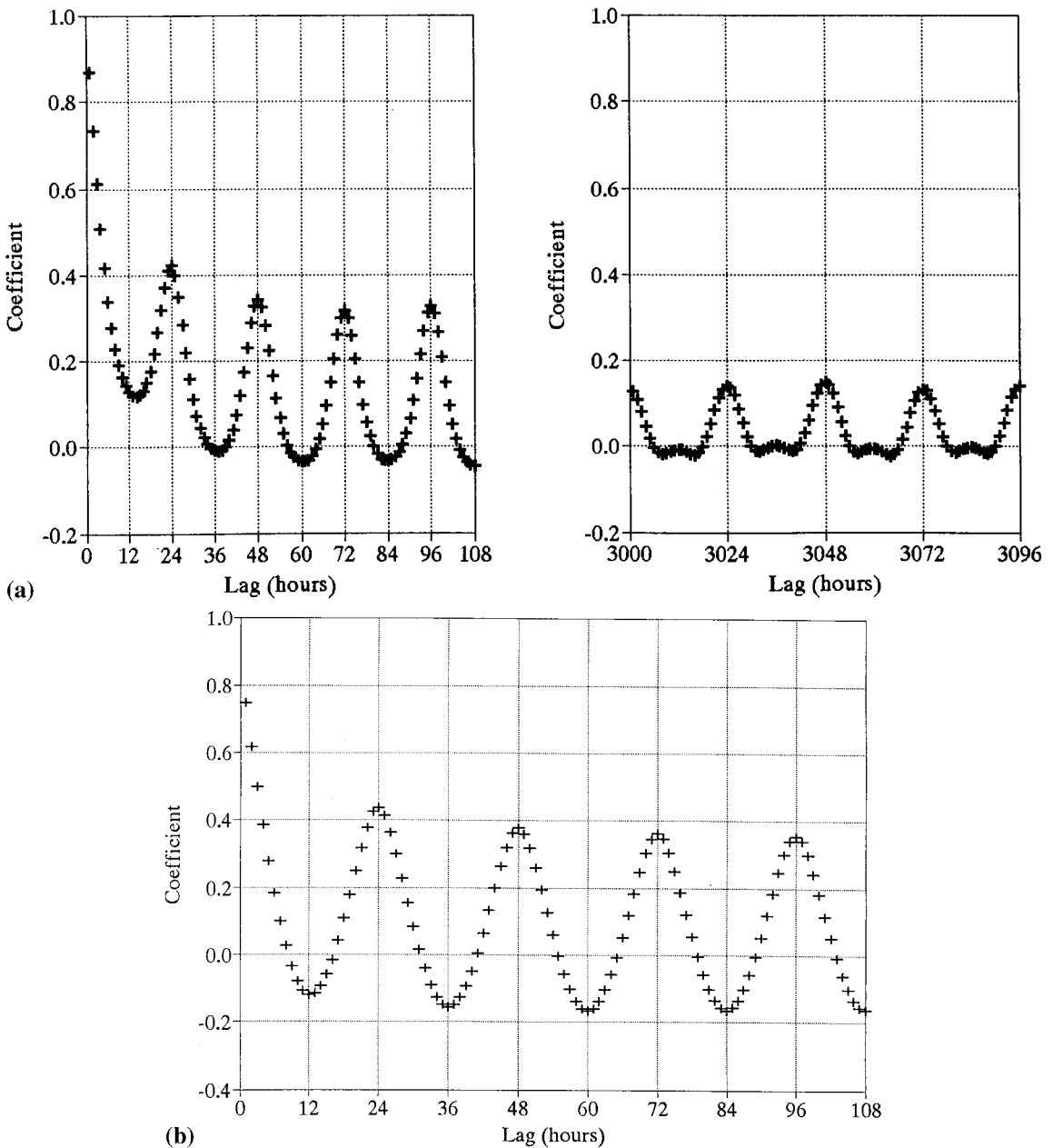


Figure 8. Autocorrelation coefficients for the wind speed (a) and wind direction (b) series

$$\text{Re}(\theta) \equiv \theta_x = -\sin \theta,$$

$$\text{Im}(\theta) \equiv \theta_y = -\cos \theta.$$

(3)

Then, the autocorrelation function and the line spectrum are computed using the discrete Fourier transform, the former supported on the autocorrelation theorem (Press *et al.*, 1989).

First autocorrelation coefficients are large, which indicates the existence of an important short-term correlation. This correlation is usually due to short-span processes of a stochastic nature. After the 12th coefficient, correlograms for both speed and direction have a sinusoidal path and a 24-h period. This

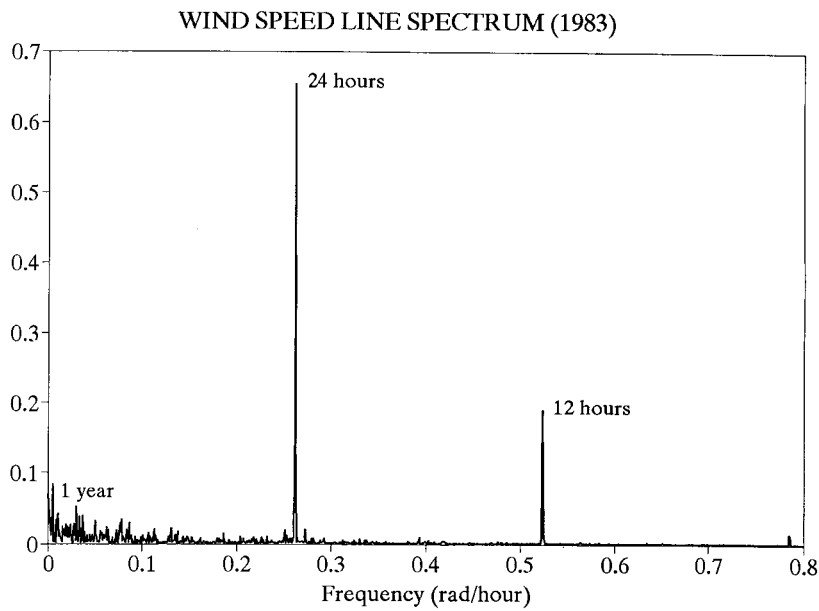


Figure 9. Line spectrum of the wind speed series (1983)

cyclic behaviour is also observed for high time lag values, which proves that the oscillation is well-established and deterministic. In the case of the wind speed, a cycle with a 12-h period is also observed.

Harmonic analysis applied to wind speed and direction series shows that the main frequencies correspond to the 1-year, 24-h and 12-h periods (Figures 9 and 10). The same cycles have also been observed by other authors in surface wind studies (Brett and Tuller, 1991; Gavalda *et al.*, 1992). The 1-year and 24-h periods are the natural earth cycles. The 12-h period for wind speed series is well-defined and corresponds to the daytime and night-time maxima due to the full development of the land–sea

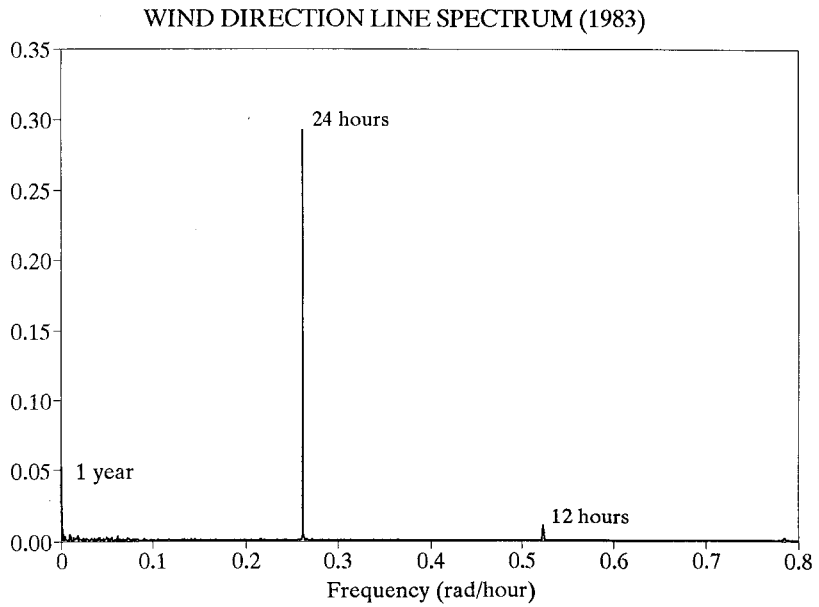


Figure 10. Line spectrum of the wind direction series (1983)

Table I. Verification and validation (bold) for the trend term of the wind speed model

t (year)	1983	1984	1985	1986	1987	1988	1989
Mean speed observed (m s^{-1})	3.3	3.3	2.7	2.8	2.6	2.2	2.3
Mean speed calculated (m s^{-1})	3.6	3.4	3.1	2.9	2.7	2.5	2.3

Table II. Model coefficients c_j for the cyclical term of the wind speed series (Equation (6))

Period	j	Frequency (rad h^{-1})	c_j^a (m s^{-1})
1 year	1	0.0007	$0.086 - 0.061i$
24 h	2	0.2618	$-0.389 + 0.090i$
12 h	3	0.5236	$0.087 - 0.232i$

^a $i = \sqrt{-1}$.

breezes. In the case of wind direction, this 12-h period does not have this physical meaning and actually the peak in the line spectrum is much lower.

4.2. Proposed model

To calibrate the model, just the first 6 years of the data series are used (1983–1988). The last year (1989) is reserved for validation purposes.

In the case of wind speed, v_t , the proposed model is the following:

$$v_t = m_{v,t} + S_{v,t} + z_{v,t}, \quad (4)$$

where $m_{v,t}$ is the trend term, $S_{v,t}$ is the seasonal or cyclical term, and $z_{v,t}$ is the stochastic component.

The trend has been calculated by fitting the mean annual speed values from 1983 to 1988 to a straight line. The expression obtained is:

$$m_{v,t} = 21 - 0.21(t - 1900), \quad (5)$$

where t is the year. Regression coefficient is rather low ($r^2 = 0.88$), but the model is able to predict the mean speed for 1989 (Table I). In order to obtain valid predictions for more than 1–2 years beyond the period of record, it is necessary to recalculate the model coefficients using the new available data.

The cyclical term, $S_{v,t}$ is obtained through harmonic analysis as follows:

$$S_{v,t} = \sum_j [c_j \exp(i\omega_j t) + c_{-j} \exp(-i\omega_j t)], \quad (6)$$

where ω_j with $j = 1, 2, 3$, are the three main frequencies, t is time in h, c_j , c_{-j} are indeterminate coefficients, $\exp(i\omega_j t) = \cos(\omega_j t) + i \sin(\omega_j t)$ and i is the imaginary unit ($\sqrt{-1}$).

Fitting Equation (6), to the wind speed series, excluding trend, by a non-linear least-square technique, the resulting coefficients c_j are those appearing in Table II. Coefficients c_{-j} are not included in this table, since they are complex conjugates of c_j . If this fitted model is evaluated by the percentage of values of the simulated wind speed series that differ by 1 m s^{-1} or less with regard to the actual series, one obtains the

Table III. Verification and validation (bold) for the cyclical term of the wind speed model

Year	1983	1984	1985	1986	1987	1988	1989
Percentage of values with $\Delta v \leq 1 \text{ m s}^{-1}$	54.1	53.2	62.4	63.6	66.2	77.1	72.5

Table IV. Model coefficients c_j for the cyclical term of the wind direction series (Equation (8))

Period	j	Frequency (rad h ⁻¹)	c_j^a	c_{-j}^a
c_0	0	0	0.135–0.121i	—
1 year	1	0.0007	0.207–0.036i	0.057–0.115i
24 h	2	0.2618	0.058–0.351i	0.338+0.043i

^a $i = \sqrt{-1}$.

values shown in Table III. Excluding the first 2 years, the rest of the years have percentages over 60%, which are acceptable if one considers that the proposed model does not take into account the stochastic component of the series. It is remarkable that the simulation improves with time. The reason is that the wind speed data dispersion is decreasing with time, in a similar way to what occurs with the annual mean speed (see Figure 5(a)).

In the case of wind direction, θ_t , no trend was observed in the series analyzed, but a cyclical term exists similar to the one observed for the wind speed. Furthermore, the remaining fluctuations do not depend on the daily variation point, i.e. random fluctuations are additive. So, the model proposed is (remember that θ is expressed as a complex number):

$$\theta_t = S_{\theta,t} + z_{\theta,t},$$

where the cyclical term, $S_{\theta,t}$, is expressed as follows:

$$S_{\theta,t} = c_0 + \sum_j [c_j \exp(i\omega_j t) + c_{-j} \exp(-i\omega_j t)],$$

where ω_j with $j = 1, 2$, are the two main frequencies.

The model coefficients for the deterministic seasonal term of the direction series, obtained by using a similar technique to that used for the speed series, appear in Table IV. As an evaluation of the ability of the proposed model to predict the actual wind direction series, Table V shows the percentages of the difference values between the simulated and observed series, less than or equal to one and two direction sectors. In the case of $\Delta\theta \leq 45^\circ$, all the values, except for 1984, are around 70%. This is a good result if the fact that the model only considers the deterministic component of the series is taken into account.

Finally, the study of the correlograms for the series of residuals shows that the trend and cyclical terms have been successfully removed. Furthermore, computed correlograms from subsets of the series of residuals have the same overall behavior as those for the entire series. This suggests that the autocorrelations of the underlying process are time-independent. Therefore, the remaining stochastic component can be reasonably considered as stationary (at least in a weak sense) and ergodic. Provided that stationarity and ergodicity conditions are met, the autocorrelation function may be used to make inferences about the underlying process of which the residual series is a sample (Chatfield, 1989; Brett and Tuller, 1991; Peña, 1994).

Table V. Verification and validation (bold) for the cyclical term of the wind direction model

Year	1983	1984	1985	1986	1987	1988	1989
Percentage of values with $\Delta\theta \leq 22.5^\circ$	55.1	37.7	51.1	55.8	56.1	59.9	52.7
Percentage of values with $\Delta\theta \leq 45^\circ$	73.4	60.6	68.3	72.0	68.4	72.3	66.2

5. CONCLUSIONS

In principle, it is not possible to assert that wind is produced by only either deterministic or probabilistic phenomena. In accordance with Alvarez (1986), the deterministic hypothesis supports the existence of certain systematic and regular cycles whose behaviour it is intended to establish, but not the cause that originates them. That is why the validity of the model proposed (that only includes the deterministic component) is restricted to the point site of the meteorological station where the wind measurements were collected. The existence of well-defined oscillations related to natural cycles has been established. This confirms the deterministic hypothesis.

The proposed model is intended to reproduce the deterministic component of the wind data series. So, the methodology followed would be specially suitable for describing and predicting wind at sites in which some cyclic patterns are expected, such as land–sea breezes, katabatic–anabatic winds, etc. In the case studied, the deterministic component of the model is enough to produce useful results (within an acceptable error) *ca.* 60–70% of time, for both wind speed and direction.

ACKNOWLEDGEMENTS

The authors would like to thank Dr M. Millán and Rosa Salvador for supplying the data series to produce this paper.

REFERENCES

- Alvarez, N. 1986. *Aplicación del análisis armónico a fenómenos meteorológicos y económicos*, Instituto Nacional de Meteorología, Madrid.
- Bosque, J. and Vilà, J. 1989. *Geografía de España, Vol. 1: Geografía Física*, Planeta, Barcelona.
- Bosque, J. and Vilà, J. 1992. *Geografía de España, Vol. 10: Comunidad Valenciana, Murcia, Glosario, Índice General*, Planeta, Barcelona.
- Breckling, J. 1989. *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, Springer, Berlin.
- Brett, A.C. and Tuller, S.E. 1991. 'The autocorrelation of hourly wind speed observations', *J. Appl. Met.*, **30**(6), 823–833.
- Brown, B.G., Katz, R.W. and Murphy, A.H. 1984. 'Time series models to simulate and forecast wind speed and wind power', *J. Clim. Appl. Met.*, **23**(8), 1184–1195.
- Carreras, C. 1992. *Geografía General dels Països Catalans. Vol. 1: el Clima i el relleu*, Enciclopedia Catalana, Barcelona.
- Chatfield, C. 1989. *The Analysis of Time Series: An Introduction*, 4th edn., Chapman and Hall, London.
- Coronas, A. and Baldasano, J.M. 1984. 'Fourier analysis of meteorological data to obtain a typical annual time function', *Solar Energy*, **32**(4), 479–488.
- Daniel, A.R. and Chen, A.A. 1991. 'Stochastic simulation and forecasting of hourly average wind speed sequences in Jamaica', *Solar Energy*, **46**(1), 1–11.
- Fisher, N.I. and Lee, A.J. 1994. 'Time series analysis of circular data', *J. R. Stat. Soc. B: Methods*, **56**(2), 327–339.
- Font Tullot, I. 1983. *Climatología de España y Portugal*, Instituto Nacional de Meteorología, Madrid.
- Frisch, A.S., Stankov, B.B., Martner, B.E. and Kaimal, J.C. 1991. 'Mid-troposphere wind spectra from long term wind profiler measurements', *J. Appl. Met.*, **30**(12), 1646–1651.
- Gavaldà, J., Massons, J., Camps, J. and Díaz, F. 1992. 'Statistical and spectral analysis of the wind regime in the area of Catalonia', *Theor. Appl. Clim.*, **46**(2–3), 143–152.
- Ishida, H. (1990). 'Seasonal variations of spectra of wind speed and air temperature in the mesoscale frequency range', *Boundary-Layer Met.*, **52**(4), 335–348.
- Jenkins, G.M. and Watts, D.G. 1968. *Spectral Analysis and its Applications*, Holden-Day, San Francisco.
- Mardia, K.V. 1972. *Statistics of Directional Data*, Academic Press, London.
- Peña, D. 1994. *Estadística Modelus y Métodos*, 7th edn., Alianza Editorial, Madrid.
- Phillips, W.F. 1984. 'Harmonic analysis of climatic data', *Solar Energy*, **32**(3), 319–328.
- Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 1989. 'Numerical Recipes', *The Art of Scientific Computing. Fortran Version*, Cambridge University Press, Cambridge.
- Schumann, E.H. 1992. 'Interannual wind variability on the south and east coasts of South Africa', *J. Geophys. Res. D: Atmospheres*, **97**(18), 20397–20403.
- Wickle, C.K., Sherman, P.J. and Chen, T. 1995. 'Identifying periodic components in atmospheric data using a family of minimum variance spectral estimators', *J. Clim.*, **8**(10), 2352–2363.