Segmentation de données directionnelles

Serge REBOUL, Mohammed BENJELLOUN

Laboratoire d'Analyse des Systèmes du Littoral (EA 2600)
Université du Littoral Côte d'Opale, 50 Rue Ferdinand Buisson, B.P. 699, 62228 Calais Cedex, France
serge.reboul@lasl.univ-littoral.fr, mohammed.benjelloun@lasl.univ-littoral.fr

Résumé – Nous proposons dans cette communication un estimateur MAP pour la segmentation et l'estimation simultanées d'un processus aléatoire stationnaire par morceaux, suivant une distribution de Von Mises. L'estimateur MAP se présente sous la forme d'une fonction de contraste pénalisée composée d'un terme d'attache aux données et d'un terme de régularisation. La segmentation est alors réalisée en minimisant la fonction de contraste par un algorithme de recuit simulé. Nous présentons l'implantation numérique de cet estimateur dans le cas de la distribution circulaire de Von Mises, l'évaluation des performances de la détection et son utilisation pour la segmentation de la direction du vent au sens de l'écart type angulaire et de la direction moyenne.

Abstract – We introduce in this communication a MAP estimate for the simultaneous estimation and segmentation of a piecewise random process which has a Von Mises distribution. This MAP estimate is a penalized contrast function with two term. The first term measures the fidelity to the observation and the second term is a parameter of regularization. The instant of changes are estimated by minimizing the penalized contrast function with a simulate annealing algorithm. We propose in this communication the numerical computation of this estimate for the circular Von Mises distribution, we evaluate the performance of the segmentation and the application of the method for the segmentation of the wind direction in the standard deviation and mean direction sense.

1 Introduction

Supposons que nous ayons un ensemble de mesures indépendantes identiquement distribuées de la direction bidimensionnelle, notée α_1, α_2,... ,α_n. Ces mesures appelées directionnelles, angulaires, peuvent être représentées comme des points sur la circonférence d'un cercle unité. Nous nous intéressons à la recherche des instants de changements, quand ils ont lieu, de la direction préférentielle et des variations autour de cette direction d'un ensemble de données temporelles ordonnées (suivant le temps croissant). Ces instants de changements étant inconnus, on parle de segmentation du signal ou de détection de rupture. Dans l'application qui nous intéresse, on veut réaliser un système de surveillance des rejets atmosphériques d'un site industriel. Dans notre approche, la mesure ou la prédiction de l'émission d'un polluant atmosphérique passent par la modèlisation de sa dispersion. Une étude préalable réalisée sur les modèles de dispersions nous a conduit à considérer l'évolution des paramètres atmosphériques, et plus particulièrement l'évolution de la vitesse et de la direction du vent, comme un paramètre prépondérant pour les modèles de simulation. En effet le transport et la dispersion des polluants atmosphériques sont principalement dus aux diffusions turbulentes inhérentes aux mouvements atmosphériques. Pour obtenir une estimation réaliste de la dispersion, il est important d'avoir une description précise des turbulences atmosphériques. L'écart type de la direction horizontale du vent σ_α, aussi appelé l'écart type angulaire, et la vitesse moyenne $\bar{\rho}$, donnent une estimation précise de la dispersion latérale (perpendiculaire à la direction du vent $\bar{\alpha}$) de la concentration d'un panache de fumée [1]. Cependant l'évolution perpétuelle des conditions atmosphériques nécessite de segmenter les signaux issus des capteurs en zone stationnaire. Le but de notre travail est d'estimer la direction moyenne et l'écart type de la déviation angulaire de la direction du vent dans les zones stationnaires du signal angulaire. On suppose que le processus est stationnaire par morceaux et c'est dans les zones stationnaires que l'on estime ses paramètres statistiques. On dispose de toutes les données simultanément et la segmentation est différée ou hors ligne, les changements sont détectés simultanément sur l'ensemble du signal. L'estimateur MAP de la position des ruptures de stationnarités et des paramètres statistiques du processus aléatoire est réalisé sous la forme d'une fonction de contraste pénalisée à minimiser. On recherche par un algorithme de recuit simulé la séquence de rupture dans le signal qui minimise cette fonctionnelle. Deux termes composent cette fonction de contraste, un terme d'attache aux données et un terme de régularisation. L'expression du terme d'attache aux données est obtenu à partir du logarithme de vraisemblance de la distribution paramétrique de Von Mises. Les paramètres statistiques de la distribution sont estimés sur le signal réel pour la séquence de rupture testée. L'expression du terme de régularisation est obtenu à partir de la loi a priori de la séquence de rupture. La résolution de la segmentation est guidée par la valeur du paramètre de régularisation associé au nombre de rupturs. La fonction d'attache aux données est exprimée dans la section 2 pour une distribution de Von Mises et il est proposé dans la section 4 d'autres formulations de ce terme qui permettent son implementation numérique. Le paramètre de régularisation est exprimé dans la section

2 et c'est dans la section 4 que l'on montre son influence sur la résolution de la segmentation. Les performances de la segmentation sont présentées et discutées dans la section 5. La dernière section est consacrée à la conclusion.

2 Formulation du problème [2]

Le problème est de détecter les changements dans la distribution statistique de la direction du vecteur vent. Nous supposons pour cela que la distribution du processus qui la décrit dépend d'un paramètre $\mathbf{p}\alpha$. Le problème consiste alors à détecter les changements de $\mathbf{p}\alpha$. Les changements peuvent affecter la moyenne et la variance du processus. Soit $A = (A_{i\geq 1})$ un processus réel de dimension d non stationnaire représentant la direction du vent. On suppose que A est stationnaire par morceaux. Alors il existe des instants de changements $(t_k, k \geq 0)$ tels que $(A_{t_k}, ..., A_{t_{k+1}})$ est stationnaire pour tout $k \in \mathbb{N}$. Le problème de la détection des changements est vu ici comme la segmentation globale de $\underline{A} = (A_1, ..., A_n)$ en segments stationnaires à partir de sa réalisation $\underline{\alpha} = (\alpha_1, ..., \alpha_n)$. On introduit le vecteur aléatoire $\underline{RA} = (RA_1, ..., RA_n)$ défini par :

$$\begin{cases} RA_i &= 1 \text{ si il existe } k \text{ tel que } i = t_k \\ RA_i &= 0 \text{ sinon,} \end{cases}$$

dont on notera la réalisation $\underline{r\alpha} = (r\alpha_1, ..., r\alpha_n)$. Pour une configuration donnée $\underline{r\alpha}$ de S_α segments, on a $\underline{\mathbf{p}\alpha} = (\mathbf{p}\alpha_\mathbf{1}, ..., \mathbf{p}\alpha_{\mathbf{S}_\alpha})$ la séquence de paramètres statistiques telle que $\mathbf{p}\alpha_\mathbf{k}$ soit le paramètre dans le segment k. Alors $\underline{\mathbf{p}\alpha}$ peut être estimé simultanément avec $\underline{RA}$ par l'estimateur MAP obtenu en maximisant la distribution a posteriori de $\underline{RA}$.

$$\begin{aligned} &[\widehat{\underline{r\alpha}}, \widehat{\underline{\mathbf{p}\alpha}}(\widehat{\underline{r\alpha}})] \\ &= \underbrace{argmax}_{(\underline{r\alpha}, \underline{\mathbf{p}\alpha})} Pr(\underline{RA} = \underline{r\alpha} / \underline{A} = \underline{\alpha}; \underline{\mathbf{p}\alpha}). \end{aligned} \quad (1)$$

Si on sait calculer, sans erreurs d'estimation, le vecteur de paramètre statistique $\underline{\mathbf{p}\alpha}$ associé à la séquence de rupture $\underline{r\alpha}$ testée, on peut rechercher la séquence de ruptures réelle en minimisant une fonction de contraste pénalisée de la forme :

$$U_{\underline{\alpha}}[\underline{r\alpha}] = V_{\underline{\alpha}}(\underline{r\alpha}, \underline{\mathbf{p}\alpha})) + ln(\Pi(\underline{r\alpha})), \quad (2)$$

où $V_\alpha(.)$, le terme d'attache aux données, est le logarithme négatif de la fonction de vraisemblance de $\underline{\alpha}$ et $\Pi(\underline{r\alpha})$ la probabilité a priori d'avoir une configuration de rupture $\underline{r\alpha}$. En pratique le vecteur de paramètre $\underline{\mathbf{p}\alpha}$ est inconnu et il faut l'estimer sur le signal réel. Alors, la fonction de contraste pénalisée est donnée par :

$$U_{\underline{\alpha}}[\underline{r\alpha}] = V_{\underline{\alpha}}(\underline{r\alpha}, \widehat{\underline{\mathbf{p}\alpha}}(\underline{r\alpha}))) + ln(\Pi(\underline{r\alpha})), \quad (3)$$

où $\widehat{\underline{\mathbf{p}\alpha}}(\underline{r\alpha})$ est le vecteur de paramètre statistique, estimé sur le signal réel $\underline{\alpha}$, pour une séquence testée $\underline{r\alpha}$. On définit $\underline{RA}$ comme une séquence de variables aléatoires indépendantes de Bernoulli. La probabilité d'avoir une séquence de rupture donnée $\underline{r\alpha}$ est donnée par :

$$\Pi(\underline{r\alpha}) = \lambda_\alpha^{S_\alpha}(1-\lambda_\alpha)^{n-S_\alpha}, \quad (4)$$

où λ_α est la probabilité d'avoir une rupture sur $\underline{\alpha}$.
Soit la distribution de Von Mises de la variable aléatoire circulaire α donnée par [3] :

$$h_A(\alpha) = \frac{1}{2\pi I_0(\gamma)} e^{\gamma cos(\alpha - \mu_\alpha)}, \quad (5)$$

où $I_0(\gamma)$ est la fonction de Bessel modifiée du premier type, d'ordre zéro. Le paramètre μ_α est la direction moyenne, le paramètre γ est un paramètre de concentration. L'expression globale de la fonctionnelle à minimiser est donnée par :

$$[\widehat{\underline{r\alpha}}] = \underbrace{argmin}_{(\underline{r\alpha}) \in \{0,1\}} \quad (6)$$

$$\sum_{k=1}^{S_\alpha} n_k \left(log(I_0(\gamma_k)) - \gamma_k \sum_{i=t_{k-1}}^{t_k} cos(\alpha_i - \hat{\mu}_k) \right) + \quad S_\alpha \beta_\alpha$$

$$\text{avec } n_k = t_k - t_{k-1} \text{ et } \beta_\alpha = \ln\left(\frac{1-\lambda_\alpha}{\lambda_\alpha}\right).$$

La fonction d'attache aux données est le logarithme négatif de la fonction de vraisemblance de la distribution de Von Mises. Elle est estimée sur l'ensemble des données pour une séquence de rupture testée.

3 Définition du paramètre de segmentation

Considérons l'expression 2 associée au processus $\underline{\alpha}$ où les paramètres $\underline{\mathbf{p}\alpha}$ calculés pour une séquence $\underline{r\alpha}$ donnée sont connus. La segmentation consiste à rechercher la séquence qui minimise l'expression 2. Dans cette expression, l'erreur d'estimation $\widetilde{\underline{\mathbf{p}\alpha}}$ liée au calcul de $\underline{\mathbf{p}\alpha}$ sur le signal réel, tel que $\widehat{\underline{\mathbf{p}\alpha}} = \underline{\mathbf{p}\alpha} + \widetilde{\underline{\mathbf{p}\alpha}}$, est supposée nulle. Le calcul du paramètre β_α est d'abord décrit dans le cas idéal ($\widetilde{\underline{\mathbf{p}\alpha}} = 0$) pour être ensuite discuté dans le cas réel ($\widetilde{\underline{\mathbf{p}\alpha}} \neq 0$).

Cas idéal

Le terme d'attache aux données de l'expression 2 est minimum quand la séquence testée est la séquence recherchée. La valeur du paramètre β_α est choisie pour que la valeur de l'expression 2 soit supérieure quand le signal est sous ou sur-segmenté.
La sur-segmentation est obtenue quand la séquence testée $\underline{\mathbf{p}\alpha}^{Su}$ contient la séquence recherchée $\underline{\mathbf{p}\alpha}^{Re}$ plus une rupture supplémentaire. On a alors :

$$U_{\underline{\alpha}}\left[\underline{r\alpha}^{Re}\right] = V_{\underline{\alpha}}^{min}\left(\underline{r\alpha}, \underline{\mathbf{p}\alpha}\right) + f(S_\alpha) \quad (7)$$

$$U_{\underline{\alpha}}\left[\underline{r\alpha}^{Su}\right] = V_{\underline{\alpha}}^{min}\left(\underline{r\alpha}, \underline{\mathbf{p}\alpha}\right) + f(S_\alpha + 1). \quad (8)$$

Or on veut $U_{\underline{\alpha}}\left[\underline{r\alpha}^{Re}\right] < U_{\underline{\alpha}}\left[\underline{r\alpha}^{Su}\right]$. Une solution au problème est que $f(S_\alpha)$ soit une fonction linéaire du nombre de ruptures de paramètre $\beta\alpha$. Cette condition est remplie dans le cadre de la loi de Bernouilli.
La sous-segmentation est obtenue quand la séquence $\underline{r\alpha}^{So}$ contient la séquence recherchée moins une rupture. On a alors $U_{\underline{\alpha}}\left[\underline{r\alpha}^{Re}\right] \leq U_{\underline{\alpha}}\left[\underline{r\alpha}^{So}\right]$ si β_α rempli la condition suivante :

$$0 \leq \beta_\alpha \leq V_{\underline{\alpha}}(t_{k+1} - t_{k-1}, \mathbf{p}\alpha_{\mathbf{k+1}}^{\mathbf{So}}) \quad (9)$$
$$- \left\{ V_{\underline{\alpha}}(t_k - t_{k-1}, \mathbf{p}\alpha_{\mathbf{k}}^{\mathbf{Re}}) + \mathbf{V}_{\underline{\alpha}}(\mathbf{t_{k+1}} - \mathbf{t_k}, \mathbf{p}\alpha_{\mathbf{k+1}}^{\mathbf{Re}}) \right\}.$$

C'est la plus petite valeur de β_α, calculée à partir de l'expression précédente qui va donner la résolution de la segmentation. On montre dans le paragraphe suivant que l'expression 12 de $V_{\underline{\alpha}}(..)$ peut être exprimée sous la forme d'une fonction analityque des paramètres statistiques dans les zones stationnaires. De plus, étant donné l'expression qui lie $\mathbf{p}\alpha_{\mathbf{k+1}}^{\mathbf{So}}$ à $\mathbf{p}\alpha_{\mathbf{k}}^{\mathbf{Re}}$ et $\mathbf{p}\alpha_{\mathbf{k+1}}^{\mathbf{Re}}$, obtenue dans le cas d'une variable aléatoire circulaire par la formule suivante,

$$\mu_\alpha = \mu_{\alpha 1} + arctg\left(\frac{\frac{n-l}{n} sin(\mu_{\alpha 2}-\mu_{\alpha 1})(1-v_2)}{\frac{l}{n}(1-v_1)+\frac{n-l}{n}cos(\mu_{\alpha 1}-\mu_{\alpha 2})(1-v_2)}\right) \quad (10)$$

$$v = 1 - \left\{\frac{l}{n}cos(\mu_\alpha - \mu_{\alpha 1})(1-v_1) + \frac{n-l}{n}cos(\mu_\alpha - \mu_{\alpha 2})(1-v_2)\right\} \quad (11)$$

, on peut calculer β_α pour des valeurs de l, n, $\mathbf{p}\alpha_{\mathbf{k}}^{\mathbf{Re}} = \mathbf{p}\alpha_{\mathbf{1}}$ et $\mathbf{p}\alpha_{\mathbf{k+1}}^{\mathbf{Re}} = \mathbf{p}\alpha_{\mathbf{2}}$ qui fixent la sensibilité à la détection. Ces valeurs définies par l'utilisateur, soit $\mathbf{p}\alpha_{\mathbf{1}} = (\mu_{\alpha 1}, \mathbf{v_1})$ et $\mathbf{p}\alpha_{\mathbf{2}} = (\mu_{\alpha 2}, \mathbf{v_2})$, sont respectivement les valeurs des paramètres $\mathbf{p}\alpha_{\mathbf{1}}$ sur l échantillons et $\mathbf{p}\alpha_{\mathbf{2}}$ sur (n-l) échantillons.

Cas réel

Cependant la valeur du paramètre β_α calculée théoriquement ne tient pas compte de l'erreur d'estimation introduite lors du calcul des fonctions d'attaches aux données sur le signal réel. Cette erreur d'estimation peut provoquer une diminution de la fonction d'attache aux données dans le cas sur-segmenté, on a alors la détection d'une rupture qui n'existe pas (fausse alarme). Dans le cas sous-segmenté, la diminution de la fonction d'attache aux données peut être inférieure à β_α, on a alors une omission (une rupture n'est pas détectée). On peut donc conclure que les performances de la détection de ruptures, associées à la valeur de β_α, évoluent comme les performances de la détection d'une rupture. Enfin β_α peut être considéré comme un seuil de détection des ruptures, auquel il faudra associer une probabilité d'omission et de fausse alarme.

4 Définition de la fonction d'attache aux données

L'expression 6 du logarithme de vraisemblance de la distribution de Von Mises n'est pas utilisable en pratique. En effet il n'existe pas d'estimateur pour le paramètre de concentration γ et la fonction de Bessel modifiée est tabulée. Pour remédier à ce problème, la première solution que nous proposons consiste à simplifier l'expression 6 avec l'approximation $\gamma \approx \frac{1}{s^2}$ [3] pour :

$\hat{v}_k = 1 - \frac{1}{n_k}\sum_{i=t_{k-1}}^{t_k} cos\,(\alpha_i - \hat{\mu}_{k_\alpha})$
$\hat{s}_k^2 = ln\frac{1}{(1-\hat{v}_k)^2}$
et
$\hat{\bar{c}} = \frac{1}{n_k}\sum_{i=t_{k-1}}^{t_k} cos(\alpha_i)$, $\hat{\bar{s}} = \frac{1}{n_k}\sum_{i=t_{k-1}}^{t_k} sin(\alpha_i)$,
$\hat{\mu}_0 = Arctan\left(\frac{\hat{\bar{s}}}{\hat{\bar{c}}}\right)$.

v_k est la variance circulaire, s_k la variance circulaire standard et μ_k la direction moyenne. Le logarithme négatif de la distribution de Von Mises peut alors être approximée par la fonction suivante :

$$V_{\underline{\alpha}}(n_k, v_k) = n_k\left(log\left(I_0\left(\frac{1}{s_k^2}\right)\right) - \left(\frac{1-v_k}{s_k^2}\right)\right). \quad (12)$$

Cette solution n'est pas complètement satisfaisante car il reste dans cette expression la fonction de Bessel tabulée. Considérons n échantillons et une rupture à la position r qui sépare deux zones stationnaires de variance v1 et v2 cf figure 4 . La décroissance de la fonction d'attache aux

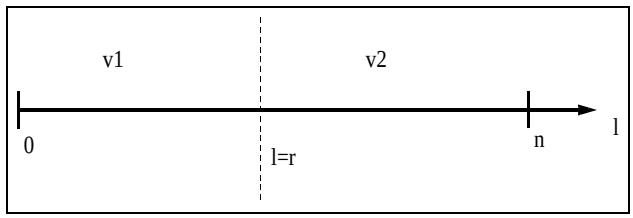

FIG. 1: Configuration de rupture.

données provoquée par l'ajout d'une rupture à la position l est donnée par :

- $l \leq r$

$$\Delta V_{\underline{\alpha}}(l) = V_{\underline{\alpha}}(n, v) - \left\{V_{\underline{\alpha}}(l, v_1) + V_{\underline{\alpha}}(n-l, v')\right\}$$

- $l \geq r$

$$\Delta V_{\underline{\alpha}}(l) = V_{\underline{\alpha}}(n, v) - \left\{V_{\underline{\alpha}}(l, v") + V_{\underline{\alpha}}(n-l, v_2)\right\},$$

où v, v', v" sont calculés avec l'expression 10 à partir des échantillons considérés et de la position de la rupture. Pour qu'une fonction $V_{\underline{\alpha}}(.)$ puisse remplacer le logarithme de vraisemblance de la distribution de Von Mises, il faut que la décroissance de la fonction d'attache aux données soit maximum en l=r, soit :

- $l \leq r$ on veut $\frac{\partial \Delta V_{\underline{\alpha}}(l)}{\partial l} > 0$
- $l \geq r$ on veut $\frac{\partial \Delta V_{\underline{\alpha}}(l)}{\partial l} < 0$

Nous proposons pour $V_{\underline{\alpha}}(..)$ la fonction suivante :

$$V_{\underline{\alpha}}(n_k, v_k) = 2 * n_k * \left(\sqrt{1-(1-v_k)^2} - 1\right), \quad (13)$$

dont l'implantation numérique est facilement réalisable.

5 Expérimentation

On montre dans la section 4 que les performances de la segmentation sont liés aux performances de la détection d'une rupture sur n échantillons, pour une position r du changement associé à l'évolution de la variance circulaire de v1 vers v2. On peut donc considérer dans notre experimentation deux hypothèses :

- l'hypothèse H_0 de détection d'une rupture quand il n'y en a pas, ce qui correspond au cas de la sur-segmentation.
- l'hypothèse H_1 de détection d'une rupture quand elle existe, ce qui correspond au cas de la sous segmentation.

On visualise sur la figure 2 les courbes COR (courbes des caractéristiques opérationnels du récepteur) de la détection, obtenues par simulation numérique. Les courbes en trait plein sont obtenues avec la fonction d'attache aux données que nous proposons. Les courbes en pointillés sont données pour une fonction de contraste calculée à partir

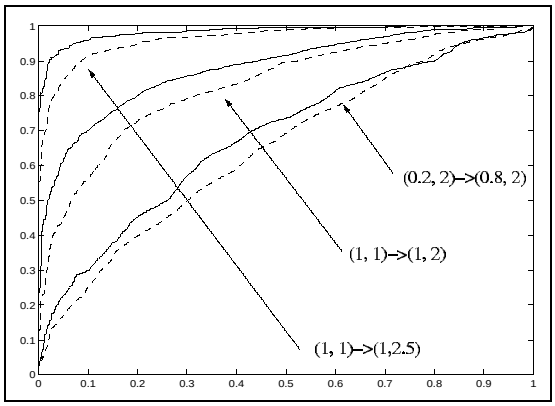

FIG. 2: Courbes COR de la détection des changements de (γ_1, μ_1) -> (γ_2, μ_2)

du logarithme de vraisemblance de la distribution de Von Mises. Les performances de la détection sont meilleures dans le cas de la fonction proposée. On visualise sur la figure 3 la probabilité de détecter et de localiser convenablement la rupture en fonction de la valeur du paramètre β_α. La convention de représentation est la même que pour la figure 2. Les performances sont meilleures pour la fonctionnelle proposée et elles augmentent plus rapidement quand la contrainte de localisation, donnée en nombre de pixels autour de la position réelle, est diminuée. Dans ce

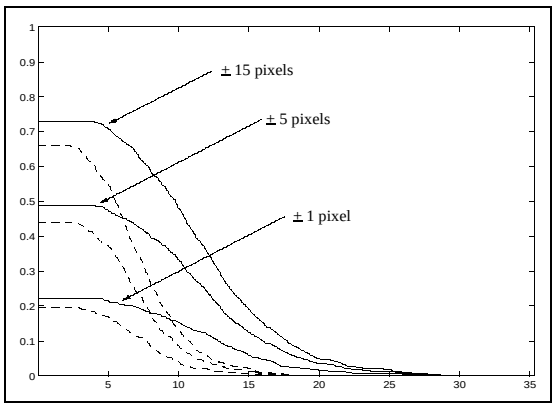

FIG. 3: Probabilité de détection et de localisation.

contexte la détermination de la valeur du paramètre β_α reste délicate. Nous proposons de calculer ce paramètre à partir de la configuration de rupture dans le signal qui donne sa plus petite valeur. Cette configuration définie par les valeurs v1,v2,n,r (cf : figure 1) permet de calculer β_α à partir des expressions 10 et 13. On constate expérimentalement qu'en prenant 70 % de sa valeur théorique les performances de la détection sont maximums. Cependant pour des faibles valeurs de β_α le signal est sur-segmenté. On visualise figure 4 un exemple de segmentation obtenu sur données réelles. Les résultats obtenus montrent la bonne adéquation du modèle statistique avec les données réelles et la faisabilité de la méthode de segmentation.

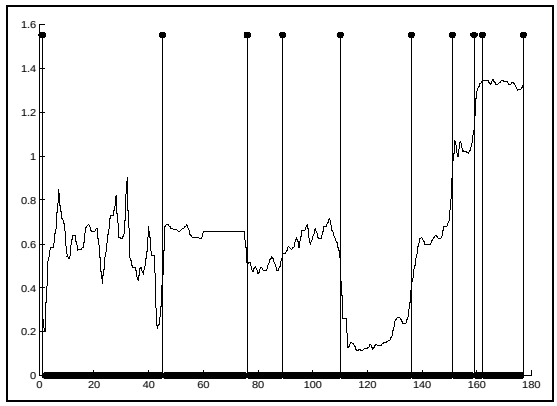

FIG. 4: Segmentation de la direction du vent.

6 Conclusion

Nous présentons dans cette communication un estimateur MAP pour la segmentation et l'estimation simultanée d'un processus aléatoire stationnaire par morceaux, suivant une distribution de Von Mises. Nous montrons que la valeur du paramètre de régularisation guide la résolution de la segmentation et peut être définie comme un seuil de détection des ruptures. Nous proposons une fonction d'attache aux données utilisable en pratique et dont les performances de détection des ruptures sont meilleures que le logarithme de vraisemblance de la distribution de Von Mises. Les perspectives de ce travail sont dans la recherche d'une implantation en ligne de la méthode, en travaillant sur des fenêtres de traitement de tailles variables.

Références

[1] R. U. Weber, "Estimators For The Standard Deviations Of Lateral, Longitudinal And Vertical Wind Components," *Atmospheric Environement*, vol. 32, pp. 3639–3646, Aug. 1998.

[2] M. Lavielle, "Optimal segmentation of random processes," *IEEE Transaction on Signal Processing*, vol. 46, no. 5, pp. 1365–1373, 1998.

[3] K. Mardia and P. Jupp, *Directional Statistics*. Wiley Series In Probability and Statistics, John Wiley and Sons, 1999.

[4] J. Tourneret, "Detection and estimation of abrupt changes contaminated by multiplicative gaussian noise," *Signal Processing*, vol. 68, pp. 259–270, 98.

[5] S. Reboul, D. Brige, and M. Benjelloun, "Segmentation de données directionnelles par fusion bayésienne," in *CIFA2000*, pp. 538–543, July 2000.

[6] K. Ghosh, S. Jammalamadaka, and M. Vasudaven, "Change-point problems for the von mises distribution," *Journal of Applied Statistics*, vol. 26, no. 4, pp. 423–434, 1999.