Modèles autorégressifs à changements de régimes markoviens

Applications aux séries temporelles de vent

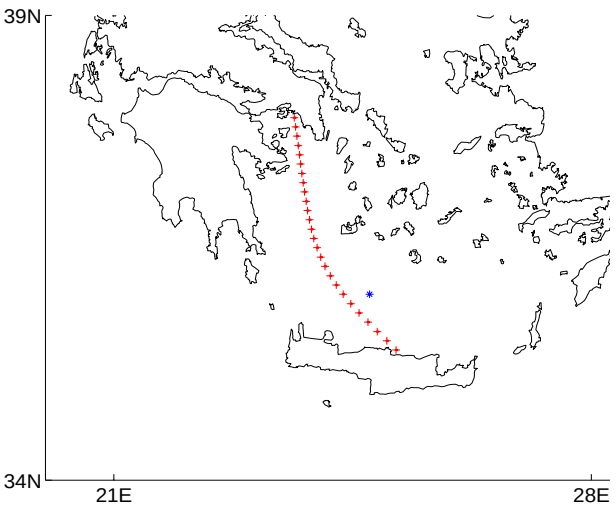
Pierre Ailliot

Motivations

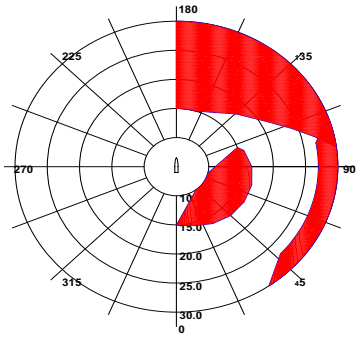
- **Conditions d'états de mer (vent, vagues) influencent...**
  - Evolution d'un trait de côte
  - Faisabilité d'une opération en mer
  - Rentabilité d'une ligne maritime
- **Données disponibles sur des périodes relativement courtes (≈ 50 ans maxi)**
- **Utilisation d'un modèle stochastique afin de simuler de nouvelles séries temporelles d'états de mer**
- **Relations complexes entre les paramètres...**
  - Dans un premier temps, séries temporelles de vent
- **Les vagues sont générées par le vent...**
  - Reconstitution à partir des séries temporelles de vent
  - Filtrage non paramétrique (*Monbet et al., 2003*)

Exemple d'application (projet Egide)

- **Objectif: étudier la rentabilité d'une ligne maritime en Mer Egée pour un bateau donné**



- **Données océano-météorologiques disponibles**
  - **Conditions d'états de mer sur la ligne (3 ans)**
- **...utilisation d'un modèle stochastique afin de simuler de nouvelles séries (500 ans)**
  - **Vent puis vagues**
- **Réponse du navire dans les différents états de mer**
  - **Vitesse maximale du bateau dans chaque état de mer**
  - **Contraintes structurelles, confort des passagers**
- **...développement d'un "simulateur" de traversée**
  - **Entrée: conditions d'états de mer sur la ligne pendant la traversée**
  - **Sortie: traversée normale, retardée ou annulée**

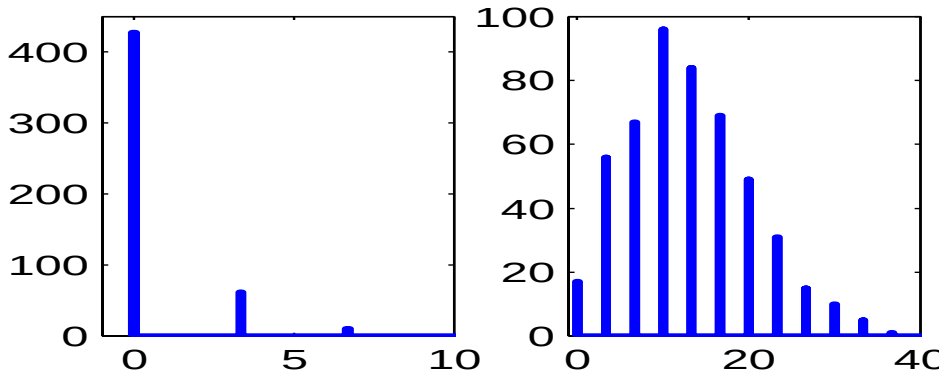


H_s=3m and T_p=7sec

- **Résultats obtenus**

	Données (3 ans)	Simulées (500 ans)
% de traversées annulées	0.00%	0.58%
% de traversées retardées	11.1%	11.2%

- Variabilité de ces quantités



Répartition du pourcentage de traversées annulées (à gauche) et retardées (à droite)

- 8% des années avec plus de 30% de traversées retardées

Plan de l'exposé

1. Définition des modèles MS-AR
2. Etude théorique des modèles MS-AR
3. Modèles en un point fixe
4. Modèle spatio-temporel
5. Perspectives

1. Définition

***Définition*: $\{X_t\} = \{S_t, Y_t\}$ suit un modèle MS-AR si c'est une chaîne de Markov (CM) à espace d'état $\{1 \dots M\} \times Y$ (avec $Y \subset R^d$) telle que**

- $P(S_t | S_{t-1} = s_{t-1}, Y_{t-1} = y_{t-1}, \dots, S_0 = s_0, Y_0 = y_0) = P(S_t | S_{t-1} = s_{t-1})$
  - **CM homogène ou non-homogène**
  - $q_\theta^{(t)}(i,j) = P(S_t = j | S_{t-1} = i)$, $Q_\theta^{(t)} = (q_\theta^{(t)}(i,j))_{i,j \in \{1 \dots M\}}$
  - $\theta \in \Theta$ **avec** Θ **compact de** R^p
  - **Processus non observé**

- $P(Y_t | S_t = s_t, S_{t-1} = s_{t-1}, Y_{t-1} = y_{t-1}, \dots, S_0 = s_0, Y_0 = y_0) = P(Y_t | S_t = s_t, Y_{t-1} = y_{t-1})$
  - $P(Y_t | S_t = s_t, Y_{t-1} = y_{t-1})$**: probabilités d'émission**
  - $P(Y_t \in dy | S_t = s_t, Y_{t-1} = y_{t-1}) = g_\theta(y_t | s_t, y_{t-1}) dy$
  - **Processus observé**

$\Pi_\theta^{(t)}$***: noyau de transition de la CM "complète"*** $\{X_t\} = \{S_t, Y_t\}$

Cas particuliers

- **CM cachées:** $P(Y_t | S_t = s_t, Y_{t-1} = y_{t-1}) = P(Y_t | S_t = s_t)$
- **Modèles autorégressifs (M=1)**

Deux types de modèles MS-AR utilisés pour le vent...

- ***Modèle MS-LAR***
  - **Introduit par *Hamilton (1989)* en économétrie**
    - **Variable cachée représente les cycles économiques (croissance/récession)**
  - **Modèle spatio-temporel pour les champs de vent**
  - ***Définition*:** $Y_t = A_\theta^{(S_t)} Y_{t-1} + B_\theta^{(S_t)} + H_\theta^{(S_t)} \varepsilon_t$
    - $A_\theta^{(s)} \in M_d(\boldsymbol{R})$, $B_\theta^{(s)} \in M_{d,1}(\boldsymbol{R})$ **et** $\Sigma_\theta^{(s)} = H_\theta^{(s)} (H_\theta^{(s)})' \in S_d^+(\boldsymbol{R})$
    - $\{\varepsilon_t\}$ **un bruit blanc gaussien,** ε_t **indépendant de** $Y_{t'}$ **pour** $t' < t$
  - **Probabilités d'émission:**

$$P(Y_t | S_t = s_t, Y_{t-1} = y_{t-1}) \sim N(A_\theta^{(s_t)} Y_{t-1} + B_\theta^{(s_t)}, \Sigma_\theta^{(s_t)})$$

- ***Modèle MS-γAR***
  - $\{Y_t\}$ **à valeurs dans** $\boldsymbol{R}^+$
    - **Intensité du vent, hauteur significative des vagues...**
  - ***Définition*:** $P(Y_t | S_t = s_t, Y_{t-1} = y_{t-1})$ **suit une loi gamma**
    - **de moyenne** $\mu^{(s_t)}(y_{t-1}) = a^{(s_t)} y_{t-1} + b^{(s_t)}$ **avec** $a^{(s)} \geq 0$ **et** $b^{(s)} > 0$
    - **d'écart-type** $\sigma^{(s_t)} > 0$

$$g_\theta(y_t | s_t, y_{t-1})$$

$$= \mu^{(s_t)}(y_{t-1}) / \left((\sigma^{(s_t)})^2 \Gamma\left(\left(\frac{\mu^{(s_t)}(y_{t-1})}{\sigma^{(s_t)}} \right)^2 \right) \right) \left(\frac{y_t \mu^{(s_t)}(y_{t-1})}{(\sigma^{(s_t)})^2} \right)^{\left(\frac{\mu(y_{t-1})}{\sigma^{(s_t)}} \right)^2 - 1} \exp\left(-\frac{y_t \mu^{(s_t)}(y_{t-1})}{(\sigma^{(s_t)})^2} \right) \mathbf{1}_{R^+}(y_t)$$

2. Etude théorique des modèles MS-AR

Estimation

- ***Objectif*: estimer le paramètre inconnu $\theta \in \Theta$ à partir d'une réalisation $\{y_t\}_{t \in \{0 \dots T\}}$ du processus $\{Y_t\}$**

Définition*: *un estimateur du maximum de vraisemblance (EMV) est un maximum de la fonction de vraisemblance

$$L_{T, s_0}(\theta) = \sum_{(s_1, \dots, s_T) \in \{1 \dots M\}^T} \prod_{t=1}^{T} q_\theta^{(t)}(s_{t-1}, s_t) g_\theta(y_t | y_{t-1}, s_t)$$

avec $s_0 \in \{1 \dots M\}$ une condition initiale arbitraire

- **Calcul numérique des EMV?**
- **Qualité des EMV?**

Validation

- **Le modèle permet-il de décrire le phénomène observé?**

Calcul numérique des EMV

Algorithme EM (*Baum et al. (1970), Dempster et al. (1977)*)

***Principe*:** algorithme itératif, partant de $\theta^{(0)} \in \Theta$. A chaque itération:

- ***Etape E (Expectation)*: calcul de la fonction intermédiaire**

$$R(\theta, \theta^{(n-1)}) = E_{\theta^{(n-1)}}[\ln p_\theta(y_1^T, S_1^T | y_0, s_0) | y_0^T, s_0]$$

  - **Expression en fonction des probabilités de lissage $p_\theta(S_t | y_0^T, S_0 = s_0)$**
  - **Algorithme Forward-Backward**

- ***Etape M (Maximisation)*: calcul de**

$$\theta^{(n)} = \text{argmax}_{\theta \in \Theta} R(\theta, \theta^{(n-1)})$$

  - **Selon les modèles, expression analytique ou optimisation numérique**

***Inconvénients*:**

- **Convergence possible vers des extrema locaux**
- **Taux de convergence asymptotique linéaire**

Algorithme quasi-Newton

- Taux de convergence asymptotique super-linéaire
- Nécessite d'évaluer la fonction de vraisemblance et son gradient en un nombre de points importants
- ...se calculent à partir du filtre de prédiction $p_\theta(S_t | y_0^{t-1}, S_0 = s_0)$
- ...qui vérifie une relation de récurrence (algorithme Forward)

Algorithme utilisé en pratique

- Localisation d'un extremum "intéressant"
  - Choix de plusieurs valeurs initiales $\theta^{(0)}$ de manière aléatoire
  - Utilisation de N_1 itérations de l'algorithme EM
- Estimation finale avec l'algorithme quasi-Newton
  - Valeur approchée de la matrice d'information observée

$$I_{T, s_0}^{obs} = -\nabla_\theta^2 \ln(L_{T, s_0}(\hat{\theta}_{T, s_0}))$$

Propriétés asymptotiques des EMV

Bibliographie

- *Baum et al. (1966)*
  - Consistance et normalité asymptotique dans les modèles CMC (Y fini)
- *Leroux (1992)*
  - Consistance des EMV dans les modèles CMC
- *Francq et al. (1998), Krishnamurty et al. (1998)*
  - Consistance des EMV dans les modèles MS-AR
- *Bickel et al. (1998)*
  - Normalité asymptotique des EMV dans les modèles CMC
- *Douc et al. (2004)*
  - Consistance et normalité asymptotique des EMV dans les modèles MS-AR

...conditions vérifiées par les modèles MS-LAR

Consistance des EMV (modèles MS-γAR homogènes)

- **On suppose que le processus $\{Y_t\}$ suit un modèle MS-γAR de paramètres**

  $\theta_0 = (\theta_{S,0}, \theta_{R,0}^{(1)}, \ldots, \theta_{R,0}^{(M)})$

- ***Notation*: $\theta_1 \sim \theta_2$ si les deux paramètres définissent le même modèle, à la numérotation près des états**

Proposition*: *Supposons que les deux conditions ci-dessous sont vérifiées:

***(C1) (stabilité):* $\forall \theta \in \Theta$, *la matrice* Q_θ *est irréductible, le noyau* Π_θ *admet une unique probabilité invariante et la solution stationnaire est ergodique et possède un moment d'ordre* $\kappa > 2$**

***(C2) (identifiabilité):* $\theta_{R,0}^{(i)} \neq \theta_{R,0}^{(j)}$ *si* $i \neq j$**

alors, si* $\bar{P}_{\theta_0}^{Y}$ *désigne la loi stationnaire de* $\{Y_t\}$, *on a

$$\forall s_0 \in \{1 \ldots M\},\ \hat{\theta}_{T,s_0} \to \theta_0 \ \bar{P}_{\theta_0}^{Y} \text{ p.s. quand } T \to \infty$$

***pour la topologie quotient associée à* $\sim$**

Stabilité (modèles MS-γAR homogènes)

- **Résultats existants valables pour les modèles $MS-AR$ fonctionnels de la forme $Y_t = f^{(S_t)}(Y_{t-1}) + \varepsilon_t$**
  - ***Holst et al. (1994), Francq et al. (1998), Yao et al. (2000 et 2001)***

***Proposition**: Soit $\{X_t\} = \{S_t, Y_t\}$ un processus MS-γAR, tel que $\{S_t\}$ soit irréductible et apériodique de probabilité invariante $\pi = (\pi_1, \ldots, \pi_M)$.*

Si l'hypothèse (S1) est vérifiée alors $\{X_t\}$ est géométriquement ergodique

(S1) $\displaystyle\sum_{1 \le i \le M} \pi_i \log(a^{(i)}) < 0$

Si en outre l'hypothèse (S2) avec $\kappa \ge 1$ est vérifiée alors la loi stationnaire de $\{Y_t\}$ admet des moments d'ordre κ

(S2) $\rho(R_\kappa) < 1$ ***avec*** $R_\kappa = (q(i,j)(a^{(j)})^\kappa)_{i,j \in \{1\ldots M\}}$

- **Conditions (C2) et (S2) (avec $\kappa > 2$) impliquent la consistance des EMV**

Normalité asymptotique (modèles MS-γAR homogènes)

- Résultats de Douc et al. (2004) ne s'appliquent pas

$$sup_{\theta \in \Theta, (y_0, y_1, s) \in Y \times Y \times S} g_\theta(y_1 | y_0, s) = \infty$$

Etude de la qualité des EMV par simulation

- Simulation de N=1000 réalisations de longueur T d'un modèle $MS-\gamma AR$
- T équivalent à 22 ans de données de vent ($T \approx 2700$)
- Comparaison à l'écart-type calculé à partir de la matrice d'information

	q(1,1)	q(2,2)	$a^{(1)}$	$a^{(2)}$	$b^{(1)}$	$b^{(2)}$	$\sigma^{(1)}$	$\sigma^{(2)}$
Vraie valeur	0.97	0.96	0.84	0.78	1.06	2.10	1.23	2.30
Biais	0.001	-0.001	-0.003	-0.002	0.020	0.019	0.001	-0.006
Ecart-type	0.028	0.033	0.049	0.058	0.337	0.499	0.126	0.218
$\sigma_{information}$	0.030	0.036	0.055	0.061	0.384	0.604	0.125	0.216

Validation de modèle

Différents critères sont généralement utilisés...

- Interprétabilité de la variable cachée
  - Cycles économiques, types de temps, déplacement des masses d'air...
- Propriétés des résidus
  - Test d'indépendance, variance du résidu

Tests d'adéquation

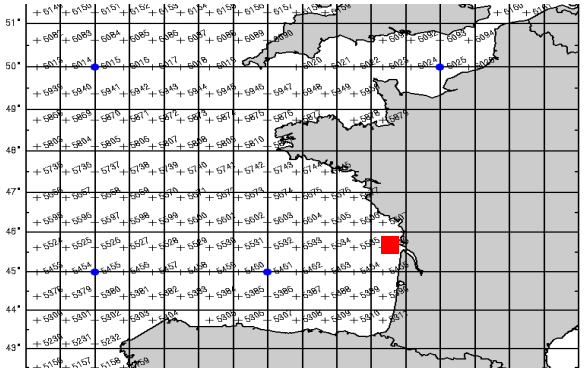
- Choix de différents critères, selon l'application
  - Fonction de répartition marginale, fonction d'autocorrélation...
- Pour chacun de ces critères, choix d'une statistique de test W
  - Exemple: distance de Kolmogorov-Smirnov
- Estimation de la loi de W sous H_0 par simulation

Sélection de modèle

- Première sélection avec le critère $BIC = -2l(\hat{\theta}) + n_{par}\ln(T)$

3. Modèles en un point fixe

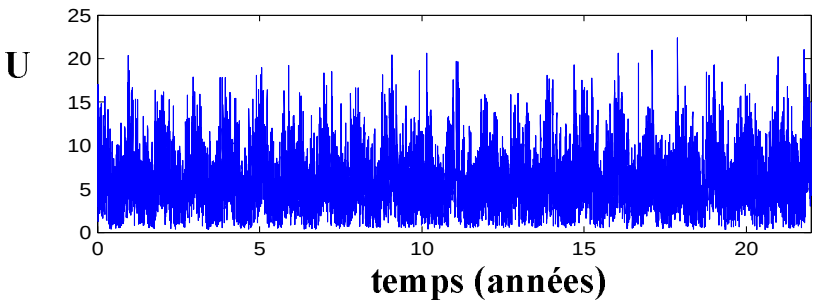
Données utilisées

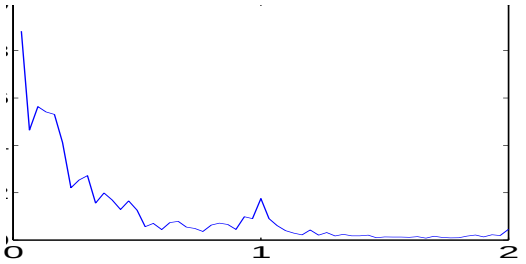


- Produites par OCEANWEATHER
- Données de “hindcast”
  - 22 ans, $\Delta t = 6h$
  - Point étudié: (46.25N, 1.67 E)
- $\{U_t\}$: intensité du vent (ms^{-1}), $\{\Phi_t\}$: direction du vent (degré)

Composantes non stationnaires

- Pas de tendance significative
- Composantes saisonnières
  - Données mois par mois
- Composantes journalières
  - Négligeables en hiver
  - Modèle spécifique en été





Modèles pour l'intensité du vent (mois de janvier)

- **Méthode usuelle (TGP)** ***(Borgman et al., 1991)***
  - ***Hypothèse*****:** $\{V_t\} = \{\Phi^{-1} \circ F_U(U_t)\}$ **est un processus gaussien**
    - F_U **et** Φ **fonctions de répartition de** U_t **et de la loi** $N(0, 1)$
    - **Simulation du processus gaussien** $\{V_t\}$ **par des méthodes exactes**
  - **Permet de décrire la loi marginale et la structure d'ordre 2**
  - **Ne permet pas de décrire l'existence de "type de temps"**

- **Modèle MS-γAR**
  - **Première sélection avec BIC**

M	1	2	3	4	5
BIC	10485	10316	10307	10343	10387

  - **Modèles à 2 ou 3 régimes?**
    - **Interprétabilité de la variable cachée et tests d'adéquation**

Interprétabilité des différents régimes (M=2)

- **Paramètres régissant l'évolution dans les différents régimes**

	$\sigma^{(s)}$	$a^{(s)}$	$b^{(s)}$
Régime 1 (s=1)	**1.37 [0.12]**	**0.79 [0.05]**	**1.46 [0.33]**
Régime 2 (s=2)	**2.40 [0.21]**	**0.77 [0.06]**	**2.24 [0.49]**

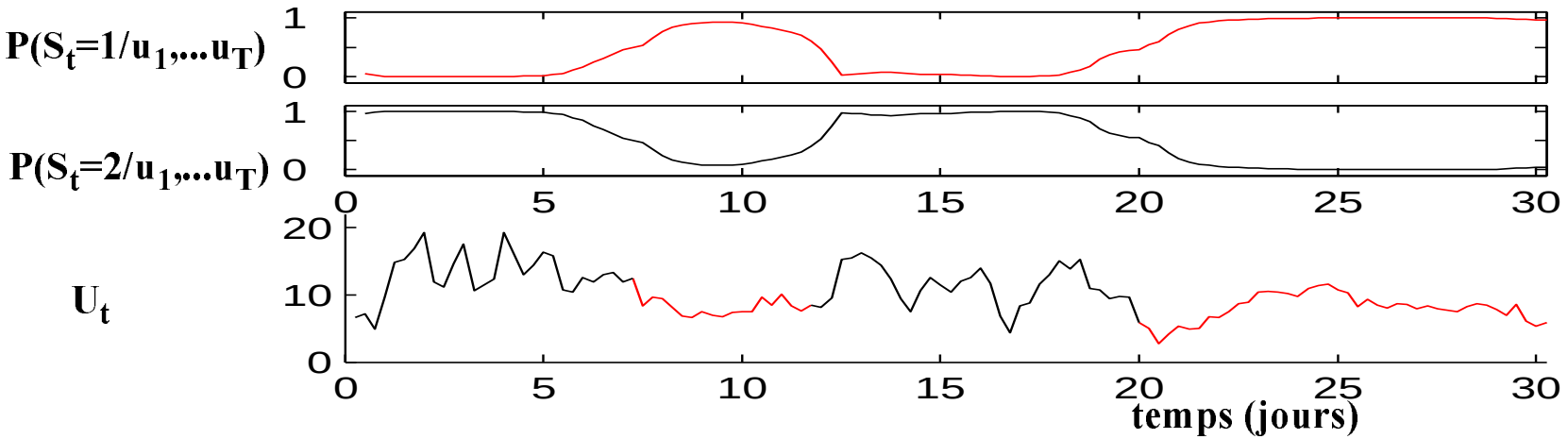
  - **Premier régime: faiblement perturbé, *conditions anticycloniques***
  - **Deuxième régime: volatilité plus importante, *conditions dépressionnaires***

- **Matrice de transition de la CM cachée**

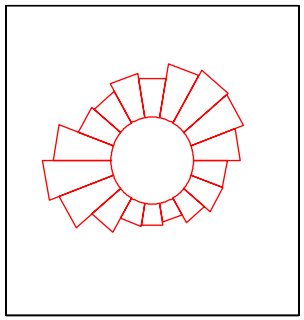
$q(i,j)$	$j = 1$	$j = 2$
$i = 1$	**0.98 [0.03]**	**0.02 [0.03]**
$i = 2$	**0.03 [0.04]**	**0.97 [0.04]**

  - **Temps de séjour moyen:**
    - **$\approx$ 14 jours dans le premier régime**
    - **$\approx$ 7 jours dans le deuxième régime**

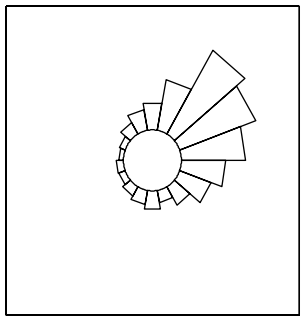
- **Exemple d'évolution des probabilités de lissage**



- **Répartition empirique de la direction du vent dans les différents régimes**



Régime 1

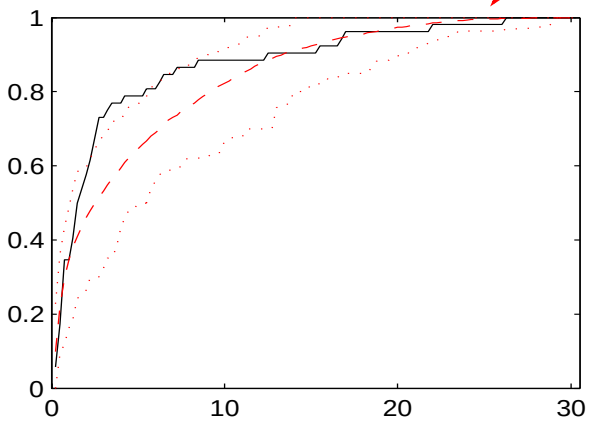


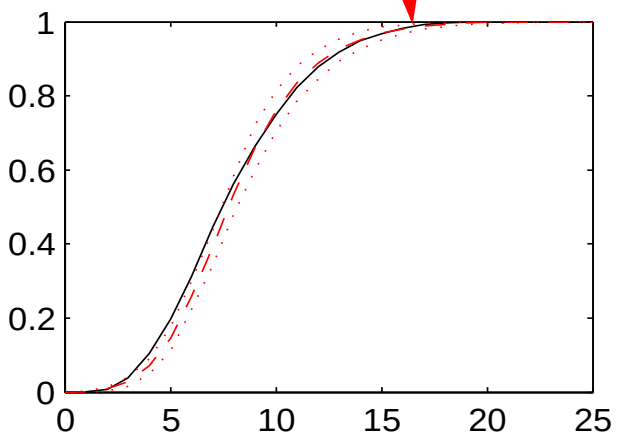
Régime 2

- **Conditions dépressionnaires associées à des vents de Sud-Ouest**

- **Comparaison des modèles TGP et $MS-\gamma AR$**
  - **Valeur entre crochets: limite de la région de rejet au seuil $\alpha = 5\%$**

	TGP	M=2	M=3
Fonct. répart. marginale	0.808 [0.012]	0.000 [0.012]	0.641 [0.024]
Fonct. d'autcorrél.	0.062 [0.012]	0.057 [0.009]	0.086 [0.022]
Fonct. répart. durée tempêtes (14 ms^{-1})	0.053 [0.012]	0.367 [0.032]	0.080 [0.016]
Fonct. répart. durée inter-tempêtes	0.002 [0.006]	0.284 [0.002]	0.053 [0.009]
Fonct. répart. durée calme (7 ms^{-1})	0.124 [0.031]	0.009 [0.031]	0.346 [0.004]





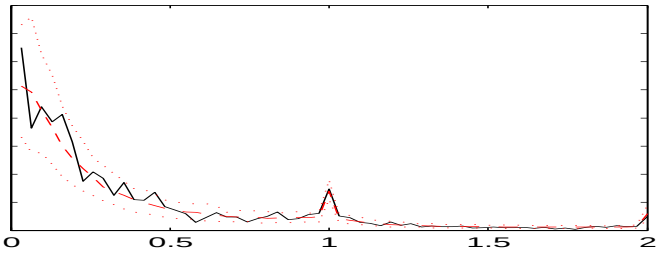
Statistique observée (noir) et correspondant au modèle (rouge)

Deux extensions (chaîne cachée non-homogène)

- **En présence de composantes journalières**

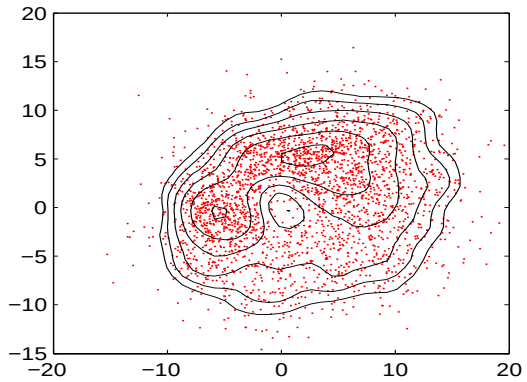
$$q_\theta^{(t)}(i,j) = P(S_t = j \mid S_{t-1} = i) \sim q_{i,j}\exp(\kappa_j \cos(\omega t + \Phi_j))$$

  - $Q = (q_{i,j})$ **une matrice stochastique,** $\kappa_j > 0$, $\Phi_j \in [0, 2\pi[$, $\omega = \pi/2$



- **Pour décrire la relation avec la direction du vent**

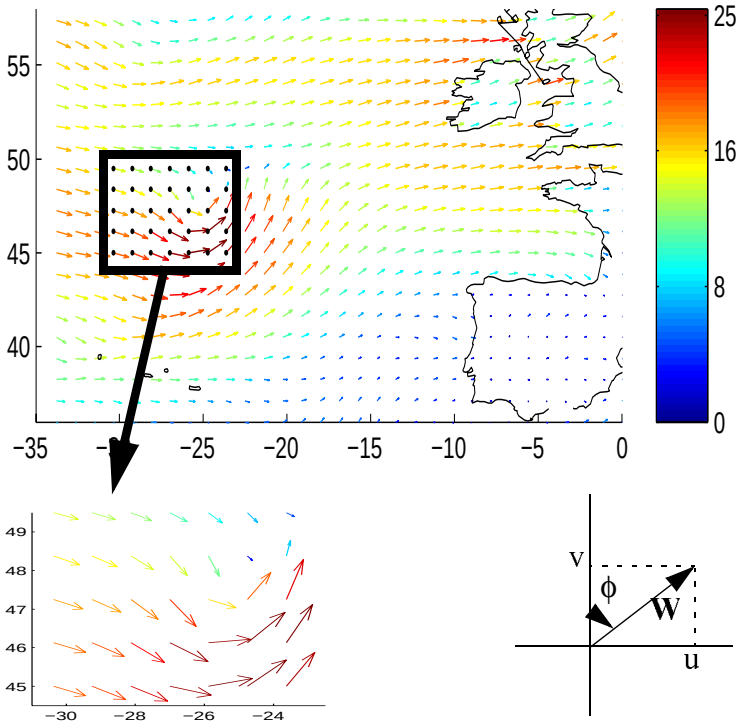
$$q_\theta^{(t)}(i,j) = P(S_t = j \mid S_{t-1} = i) \sim q_{i,j}\exp(\kappa_j \cos(\Phi_t - \Phi_j))$$



4. Modèle spatio-temporel

Données utilisées

- Produites par ECMWF
- Données de “hindcast”
  - Disponibles sur tout le globe
    - $\Delta x = \Delta y = 1.125°, \Delta t = 6h$
  - 11 ans (mois de janvier)
  - Restriction à une zone R_0
    - $600km \times 600km$
    - $N = 35$ points

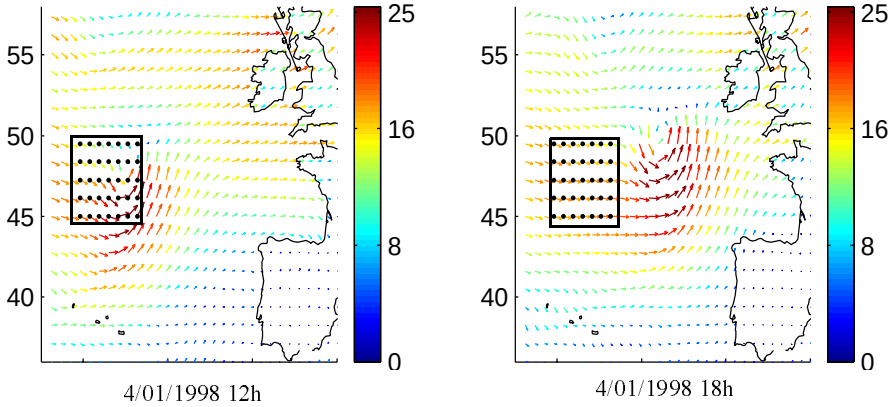


Notations

$$Z_t(R_0) = (u(r_1, t), \ldots, u(r_N, t), v(r_1, t), \ldots v(r_N, t))$$

- u, v : composantes zonale et méridienne
- $R_0 = (r_1, \ldots, r_N)$

Choix du modèle



Les structures météo se déplacent...

- S_t **déplacement entre** $t-1$ **et** t
  - **A valeurs dans** $\{a_1, \ldots, a_M\} \subset Z^2$
  - **Vitesses inférieures à** $150\,kmh^{-1}$

Utilisation d'un modèle $MS-LAR$

$$Z_t(R_0) = A_\theta^{(S_t)} Z_{t-1}(R_0) + B_\theta^{(S_t)} + H_\theta^{(S_t)} \varepsilon_t$$

Paramétrisation et estimation

- **Estimation des valeurs prises par le processus** $\{S_t\}$
  - **Utilisation d'information supplémentaire (champs sur une plus grande zone)**
- **Utilisation de ces déplacements estimés pour...**
  - **Choisir des formes paramétriques pour** $A_\theta^{(s)}$, $B_\theta^{(s)}$, $\Sigma_\theta^{(s)} = H_\theta^{(s)}(H_\theta^{(s)})'$ **et** Q_θ
  - **Obtenir une première estimation de** θ
- **Réestimation des paramètres du modèle**

Paramétrisation de $A^{(s)}$

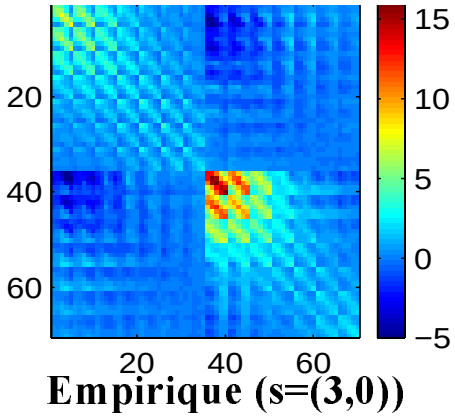
- $Z_{t-1}(R_0) \approx Z_t(R_0 + S_t) + \delta_t$
  - δ_t : **déformation du champ entre les instants** $t-1$ **et** t
- $A^{(s)}$ **fixée, permettant d'extrapoler le champ sur la zone** R_0 **à partir du champ sur la zone** $R_0 + s$

Paramétrisation de $M^{(s)} = (I - A^{(s)})^{-1} B^{(s)}$

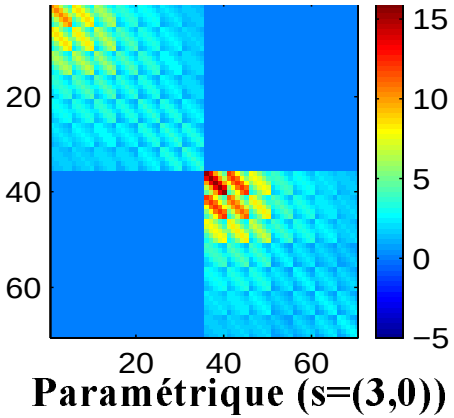
- $M^{(s)} = Fs + G$
- **6 paramètres**

Paramétrisation de $\Sigma^{(s)}$

- **7 paramètres**



Empirique (s=(3,0))



Paramétrique (s=(3,0))

Paramétrisation de l'évolution de la CM cachée

$$q(i,j) = P(S_t = a_j | S_{t-1} = a_i) \sim \exp\left(- \frac{\|a_i - a_j\|^2}{\sigma^2} - (a_j - a_0)' O^{-1} (a_j - a_0) \right)$$

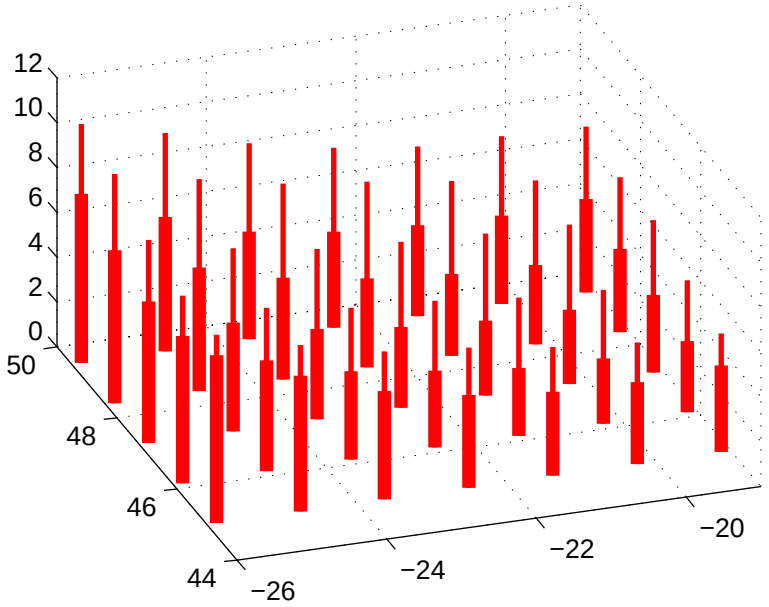
- **6 paramètres**

Estimation des paramètres

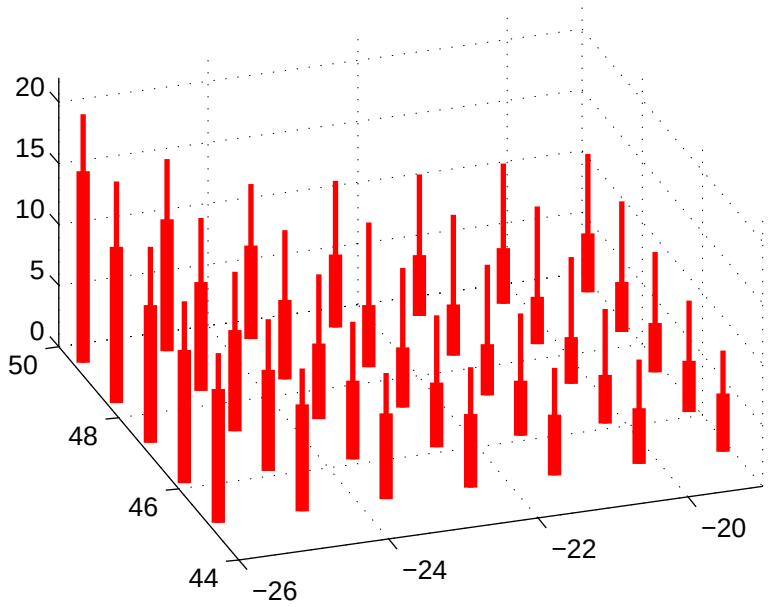
- **Nombre total de paramètres: 19**
  - **Première estimation à partir des déplacements estimés**
  - **EM puis quasi-Newton**
- **Temps de calcul importants...**
  - **Grand nombre d'états pour la CM cachée**
  - **Complexité des probabilités d'émission**

Validation

- ***En prédiction***
  - Comparaison avec un modèle AR(1) (trait fin)



Composante zonale



Composante méridienne

- ***En simulation***
  - Structure d'ordre 2 bien reproduite
  - Lois marginales aux différents points mal reproduites

Perspectives

- **Etude théorique des modèles MS-AR**
  - Normalité asymptotique des EMV dans les modèles $MS-\gamma AR$
  - Propriétés asymptotiques des EMV lorsque $\{S_t\}$ est non-homogène
  - Sélection de modèle
- **Modèles en un point fixe**
  - Tester les modèles sur d'autres paramètres (H_s, T_p,...)
  - Tester les modèles sur des mesures "in-situ"
  - Modèles paramétriques pour les séries directionnelles (Φ, Θ_m,...)
- **Modèle spatio-temporel**
  - Algorithmes plus efficaces pour le calcul des paramètres
  - Tester d'autres paramétrisations
- **Reconstruction de H_s, T_p,... à partir des séries de vent**
- **Boîte à outils (Matlab)**