N° d'Ordre de la thèse : 3139

T H È S E

Présentée

DEVANT L'UNIVERSITÉ DE RENNES I

pour obtenir

le grade de DOCTEUR DE L'UNIVERSITÉ DE RENNES I

Mention Mathématiques et Applications

par

Pierre Ailliot

Institut de Recherche Mathématique de Rennes
Institut Français de Recherche pour l'Exploitation de la Mer
École Doctorale MATISSE

TITRE DE LA THÈSE :

Modèles autorégressifs à changements de régimes markoviens. Applications aux séries temporelles de vent

Soutenue le 15 novembre 2004 devant la Commission d'Examen

COMPOSITION DU JURY :

DELYON Bernard	Président	Université de Rennes 1
DESHAYES Jean	Directeur	Université de Rennes 1
MONBET Valerie	Examinateur	Université de Bretagne Sud
PREVOSTO Marc	Examinateur	IFREMER
RYCHLIK Igor	Rapporteur	Lund University
ROBERT Christian P.	Rapporteur	Université Paris-Dauphine
YAO Jian-Feng	Examinateur	Université de Rennes 1

Remerciements

Je tiens tout d'abord à remercier tout ceux avec qui j'ai eu le plaisir de travailler au cours de cette thèse, et notamment Jean Deshayes, Valérie Monbet et Marc Prevosto.

Mes remerciements vont également à l'ensemble des membres du jury, le président Bernard Delyon, les rapporteurs Christian P. Robert et Igor Rychlik, ainsi que les examinateurs Jean Deshayes, Valérie Monbet, Marc Prevosto et Jian-Feng Yao, pour m'avoir fait l'honneur d'évaluer ces travaux.

Table des matières

Introduction

Motivations

Les conditions d'états de mer interviennent de manière déterminante dans de nombreuses activités humaines et influencent de nombreux phénomènes physiques. Ainsi, par exemple, elles conditionnent le vieillissement des structures "offshore", la faisabilité d'opérations en mer, l'évolution d'une nappe de pétrole ou l'érosion d'une côte. Lorsque l'on veut dimensionner une plate-forme pétrolière ou prévoir l'évolution d'un trait de cote à moyen terme, il est alors nécessaire de caractériser ces conditions d'état de mer.

Pour cela, on dispose de différentes sources de données. Tout d'abord, on peut utiliser les données in-situ, c'est à dire obtenues par des mesures directes. Les deux principaux moyens d'acquisition de ce type de donnée sont les bouées et les satellites. Ils permettent de mesurer le vent et les vagues avec une bonne précision, mais souffrent de plusieurs inconvénients. Ainsi, il y a généralement des données manquantes dans les données issues de bouées (dues aux pannes) et l'échantillonnage temporel des données satellitaires est peu satisfaisant par rapport à l'échelle temporelle des phénomènes météorologiques puisqu'il y a un délai de plusieurs jours entre deux passages successifs d'un satellite au dessus d'un point donné. Les bases de données obtenues sont alors difficiles à exploiter directement. De plus, ces moyens d'acquisition demeurent relativement coûteux à mettre en oeuvre, et les bases de données sont généralement disponibles sur des périodes relativement courtes (quelques années). Une alternative consiste à utiliser des données de "hindcast", c'est à dire obtenues en assimilant les observations in-situ dans les modèles numériques de prévision météorologique. Avec ces méthodes on peut reconstituer de manière relativement précise les conditions d'état de mer sur une période nettement plus longue (jusqu'à 40 ans) qu'avec les moyens de mesure traditionnels. Nous avons choisi d'utiliser ce type de donnée dans cette thèse (cf annexe A).

Selon l'application envisagée, différents types d'analyse statistique peuvent être menés sur ces données. La plus usuelle d'entre elles consiste à caractériser les valeurs extrêmes. Ainsi, par exemple, le dimensionnement d'une plate-forme pétrolière dépend principalement des valeurs extrêmes de la hauteur des vagues et les ingénieurs ont alors besoin de connaître des périodes de retour à 20, 50 ou 100 ans pour ce processus. Pour d'autres applications, la connaissance de ces périodes de retour n'est pas suffisante. Ainsi, l'évolution d'un trait de côte, d'une fissure dans une structure, ou la propagation d'une nappe d'hydrocarbure dépend non seulement des événements extrêmes mais aussi de l'accumulation d'événements d'intensité moins importante. Comme nous l'avons mentionné ci-dessus, les bases de données sont disponibles sur des périodes relativement courtes, et ne restituent donc qu'une partie de la forte variabilité des conditions d'état de mer.

Il est alors naturel de chercher une méthode permettant de simuler des conditions d'état de mer synthétiques ce qui permet d'enrichir de manière artificielle les bases de données disponibles. L'évolution dans le temps et dans l'espace des états de mer ne pouvant être identifiée com-

me un processus déterministe, il faut alors avoir recours à une modélisation probabiliste de ces phénomènes. Le modèle construit est ensuite utilisé pour simuler de nouveaux historiques d'état de mer artificiels, le nombre et la longueur de ces séquences simulées pouvant être choisis selon l'application. Ces séquences artificielles peuvent alors être utilisées en entrée dans le modèle d'évolution du phénomène considéré (par exemple, modèle de transport sédimentaire ou de dérive d'une nappe de polluant) ce qui permet de calculer empiriquement la probabilité qu'un scénario, tel que l'arrivée d'une nappe de pétrole en un endroit donné, se produise. Evidemment, la qualité des résultats obtenus avec ce type de méthode va dépendre fortement du réalisme des séquences simulées, et nous allons donc valider les modèles stochastiques à travers leurs capacités à simuler des séries temporelles réalistes. La plupart des modèles introduits dans cette thèse peuvent aussi être utilisés pour combler des valeurs manquantes dans les bases de données ou pour réaliser des prédictions à court terme.

Au cours de cette thèse, un exemple d'application à la rentabilité d'une ligne maritime a été plus précisément développé dans le cadre d'un projet européen Egide. Des conditions d'état de mer ont alors été simulées sur la ligne maritime reliant les ports du Pirée et de Héraklion en Mer Egée. Ensuite ces séquences simulées sont utilisées en entrée d'un "simulateur de traversée". Ce simulateur permet de calculer, en fonction des conditions rencontrées sur la ligne et des caractéristiques du bateau à passagers considéré, si la traversée se fait dans des conditions normales, ou, lorsque les conditions sont sévères, est retardée voir annulée. On en déduit alors des quantités telle que la répartition des traversées annulées ou retardées au cours d'une période donnée de l'année. La méthode utilisée ainsi que les résultats obtenus sont plus précisément décrits dans *Ailliot et al.* (2003, [4]). D'autres exemples d'applications peuvent être trouvés dans *Brown et al.* (1984, [27]), *Castino et al.* (1998, [31]) (évaluation de la production d'énergie par une éolienne), *O'Caroll et al.* (1984, [91]), *Monbet et al.* (2001, [85]) (planification d'opérations off shore) *Borgman et al.* (1991, [18]), *Waeles et al.* (2004, [118]) (transport sédimentaire).

Notons que ce type d'approche est aussi couramment utilisé en hydrométrie. Des modèles stochastiques sont utilisés pour simuler des séries temporelles de cumuls de précipitations, ce qui permet, par exemple, d'étudier la production d'énergie par une centrale hydro-électrique ou d'évaluer les risques de crue (cf *Guttorp* (1996, [54])).

Paramètres d'état de mer considérés

Un état de mer peut être décrit approximativement par une combinaison de plusieurs paramètres synthétiques. Les variables les plus utilisées dans les applications sont l'**intensité du vent** U, la **direction du vent** Φ, la **hauteur significative des vagues** H_s, la **période des vagues** T_m et la **direction moyenne des vagues** Θ_m. Ces paramètres sont plus précisément définis dans l'annexe A. L'objectif final est de trouver une méthode permettant de simuler l'évolution spatio-temporelle du processus $(U, \Phi, H_s, T_m, \theta_m)$.

Cependant, il existe des relations physiques complexes entre ces différents paramètres et il

semble difficile de trouver un modèle permettant de décrire directement l'évolution de ce processus multivarié. Nous avons alors choisi, dans un premier temps, de nous intéresser uniquement au vent, c'est à dire au processus $Y = (U, \Phi)$. Une première justification de ce choix est pratique : les bases de données relatives au vent sont nettement plus nombreuses et généralement disponibles sur des périodes plus longues que celles relatives aux autres paramètres d'états de mer (cf annexe A). Une autre justification, plus physique, est que la manière dont évolue le processus $X = (H_s, T_m, \theta_m)$ en un point donné est généralement plus difficile à interpréter, son évolution étant liée à la situation météorologique globale. Ainsi, par exemple, la manière dont évolue le processus $X = (H_s, T_m, \theta_m)$ dans le golfe de Gascogne dépend des conditions météorologiques sur tout l'océan Atlantique nord, les vagues pouvant se propager sur des milliers de kilomètres.

Les vagues sont principalement générées par le vent, et, en théorie, l'évolution du processus X peut se déduire de manière déterministe de celle du processus Y. En pratique, une première méthode consiste à utiliser un modèle numérique. De tels modèles sont, par exemple, utilisés quotidiennement par les centres de prévisions météorologiques pour calculer les conditions d'états de mer correspondant aux champs de vents prédits par les modèles météorologiques. Cependant, l'objectif étant de simuler les processus sur une longue durée (de l'ordre de quelques centaines d'années par exemple) l'utilisation de ce type de modèle serait très coûteux en temps de calcul. De plus, ces modèles numériques utilisent les champs de vent sur une zone étendue pour prédire les conditions d'états de mer en un point donné (cf ci-dessus) et les modèles développés dans cette thèse permettent uniquement de simuler les séries temporelles de vent en un point fixé ou sur une zone de taille restreinte. Dans ce cas, la relation entre les processus X et Y ne peut plus être considérée comme déterministe et il semble naturel d'utiliser un modèle stochastique.

De tels modèles ont été développés au cours de cette thèse. Cependant, ce problème est relativement éloigné du thème central et nous avons choisi de les présenter uniquement en annexe sous la forme de deux articles qui ont été publiés dans des actes de conférence. Dans *Ailliot et al.* (2003, [5]) nous présentons une première méthode basée sur une recherche des plus proches voisins dans une séquence d'apprentissage. Elle est utilisée pour reconstruire les conditions d'état de mer sur une ligne maritime en Mer Egée à partir de la connaissance du vent en un point situé au milieu de la ligne. Une méthode plus sophistiquée est présentée dans *Marteau et al.* (2004, [79]). Cette méthode suppose que le processus (Y, X) suit un modèle de chaîne de Markov cachée dans lequel Y désigne le processus observé et X la chaîne de Markov cachée. Le noyau de transition de la chaîne de Markov cachée, ainsi que les probabilités d'émission, sont ensuite estimés de manière non-paramétrique à partir d'une séquence d'apprentissage dans laquelle les deux processus sont observés simultanément. Un algorithme de reconstitution, à savoir l'algorithme de Viterbi, est ensuite utilisé pour calculer les valeurs les "plus probables" prises par le processus X connaissant une réalisation du processus Y.

Plan de la thèse

Différents modèles sont proposés dans cette thèse pour les séries temporelles de vent, et parmi eux les modèles autorégressifs à changements de régimes markoviens jouent un rôle particulier. La thèse est organisée de la manière suivante.

L'étude théorique des modèles autorégressifs à changements de régimes markoviens fait l'objet du **chapitre 1** de cette thèse. Ils ont été introduits il y a 15 ans en économétrie par *Hamilton* (1989, [55]), chaque régime correspondant à un état distinct de l'économie, puis utilisé ensuite dans différents domaines d'applications. Cependant, de nombreux problèmes théoriques les concernant n'ont été résolus que récemment. Dans un premier paragraphe nous définissons ces modèles de manière générale, puis nous introduisons plus précisément les différents modèles qui sont utilisés dans les chapitres suivants pour les séries temporelles de vent. Ensuite nous passons en revue différents problèmes théoriques liés à ce type de modèle : stabilité, calcul puis propriétés asymptotiques des estimateurs du maximum de vraisemblance et enfin sélection de modèle et validation. Pour chacun de ces problèmes, nous rappelons tout d'abord les résultats existants, puis nous vérifions si ils s'appliquent aux modèles développés pour le vent. Nous verrons que ce n'est pas toujours le cas, et nous démontrerons alors différents résultats pour ces modèles spécifiques.

Le **chapitre 2** est consacré à la modélisation des séries temporelles de vent en un point fixe. La méthode usuelle consiste à supposer que le processus observé peut être rendu approximativement gaussien via une transformation simple, et il suffit alors de simuler des processus gaussiens ce qui peut être fait en utilisant des approches paramétriques (modèle ARMA par exemple) ou non-paramétriques. Cependant, nous allons voir que cette méthode ne permet pas de modéliser certaines non-linéarités présentes dans les séries temporelles de vent, et en particulier la manière dont se succèdent les tempêtes. Avant de proposer un modèle plus réaliste, il faut étudier précisément les mécanismes d'évolution du processus considéré. Les conditions de vent sont régies par les champs de pression, qui eux-mêmes dépendent de la position des grandes masses d'air (anticyclones, dépressions...). Il est généralement admis que les différentes configurations peuvent être regroupées en quelques "types de temps". Les données de vent sont alors sujettes à des changements de régimes induits par ces types de temps : par exemple, dans les conditions dépressionnaires les vents évoluent rapidement (forte volatilité) alors que dans les conditions anticycloniques elles évoluent plus lentement. Comme nous ne disposons pas d'observations de cette variable "type de temps", cette dernière est introduite sous la forme d'une variable cachée, que nous supposerons être une chaîne de Markov à espace d'état fini. L'évolution de l'intensité du vent dans chaque type de temps est ensuite décrite par des modèles autorégressifs distincts, et finalement nous proposons donc d'utiliser un modèle autorégressif à changements de régimes pour décrire l'intensité du vent. A notre connaissance, ce type de modèle n'a jamais été proposé dans ce domaine d'application. Par ailleurs, afin de prendre en compte la positivité de l'intensité du vent, des modèles autorégressifs originaux, construits à partir de la loi gamma, sont utilisés. Nous validons ensuite ce modèle selon différents critères,

et en particulier nous vérifions le réalisme des séquences simulées avec ce modèle en utilisant une méthode originale introduite au chapitre précédent. Les résultats obtenus sont sensiblement meilleurs que ceux obtenus avec la méthode usuelle. Ensuite, différentes extensions de ce modèle sont proposées. Dans ces extensions, la chaîne de Markov cachée devient non-homogène, les probabilités de transition dépendant de variables "explicatives". Nous montrons en particulier que ce type de modèle permet de décrire la relation complexe entre l'intensité et la direction du vent, ainsi que les composantes journalières créées par les différences de température entre le jour et la nuit.

L'évolution de certains phénomènes, tels que la dérive d'une nappe de polluants ou l'évolution d'un trait de côte, dépend de l'évolution du vent en plusieurs points simultanément, et il est alors nécessaire de développer un modèle permettant de décrire l'évolution spatio-temporel des champs de vent. Ce problème fait l'objet du **chapitre 3**. L'approche "usuelle" consiste à ajuster un modèle autorégressif. Généralement, une analyse en composantes principales est effectuée au préalable afin de diminuer la dimension de l'espace des observations. Ce type de modèle permet principalement de modéliser la structure du second ordre des champs de vent, mais ne permet pas de reproduire, par exemple, le déplacement de structures météorologiques. Nous proposons alors d'utiliser, une nouvelle fois, un modèle autorégressif à changements de régimes dans lequel la variable cachée représente le déplacement des masses d'air. Nous validons ensuite le modèle proposé en vérifiant sa capacité à réaliser des prédictions à court terme et à simuler des champs de vent artificiels réalistes. Les résultats obtenus en prédiction à court terme sont nettement meilleurs que ceux obtenus avec les modèles autorégressifs, mais par contre les champs de vent simulés avec ce modèle ont des caractéristiques sensiblement différentes des champs de vent initiaux.

Les chapitres 2 et 3 peuvent être lus indépendamment l'un de l'autre, mais font référence au chapitre 1.

Dans l'**annexe A**, nous définissons les paramètres d'état de mer considérés dans cette thèse, et nous décrivons les différentes bases de données qui sont utilisées dans cette thèse. Dans l'**annexe B**, nous avons regroupé différents lemmes techniques sur la loi gamma. Ces lemmes sont utilisés au premier chapitre et permettent d'étudier les propriétés théoriques du modèle utilisé au deuxième chapitre pour décrire l'intensité du vent. Les **annexes C et D** sont deux articles qui ont été publiés dans des actes de conférences, à savoir *Ailliot et al.* (2003, [5]) et *Marteau et al.* (2004, [79]). Les problématiques étudiées dans ces deux articles ont déjà été mentionnées ci-dessus.

1. Modèles autorégressifs à changements de régimes markoviens

Ce chapitre est consacré à l'étude théorique des modèles autorégressifs à changements de régimes markoviens.

Dans le premier paragraphe, nous définissons le modèle de manière générale, puis nous donnons différents exemples. Nous insistons en particulier sur les différents modèles utilisés aux chapitres 2 et 3 pour décrire les séries temporelles de vent. La stabilité de ces modèles est étudiée au paragraphe 1.b, puis dans les deux paragraphes suivants, nous nous intéressons aux estimateurs de maximum de vraisemblance (**EMV**). Le paragraphe 1.c est consacré au calcul numérique de ces estimateurs puis dans le paragraphe 1.d, nous nous intéressons à leurs propriétés asymptotiques. Dans ces différents paragraphes, nous commençons par rappeler les résultats existants, puis nous vérifions si ils s'appliquent aux modèles utilisés au chapitre 2 et 3. Nous verrons que ce n'est pas toujours le cas, et nous prouvons alors certains résultats pour ces modèles spécifiques.

Enfin, au paragraphe 1.e, nous abordons les problèmes de la sélection et de la validation de modèle. Nous proposons en particulier une méthode de validation générale, valable pour toute méthode de simulation, qui permet de tester le réalisme des séquences simulées.

1.a Définitions et exemples

Les modèles autorégressifs à changements de régimes ont été introduits il y a 15 ans par *Hamilton* (1989, [55]) en économétrie, puis ont été ensuite largement utilisés en économétrie et en traitement automatique de la parole (cf *Krolzig* (1997, [69]) et *Douc et al.* (2004, [42]) ainsi que les références citées dans ces ouvrages). Ces modèles sont définis au paragraphe 1.a.1. Nous introduisons aussi dans ce paragraphe diverses notations qui sont utilisées dans la suite de cette thèse.

Ces modèles englobent différents modèles largement répandus dans la littérature, tels que les modèles de mélanges, les chaînes de Markov cachées et les modèles autorégressifs. Ces différents modèles sont définis au paragraphe 1.a.2. Nous décrivons aussi dans ce paragraphe différents modèles autorégressifs à changements de régimes particuliers, à savoir ceux qui sont les plus couramment étudiés dans la littérature et ceux utilisés dans les chapitres 2 et 3.

1.a.1 Définitions

Un processus autorégressif à changements de régimes est un processus bivarié $\{S_t, Y_t\}$ où $\{S_t\}$ désigne une chaîne de Markov d'ordre 1 et pour lequel, conditionnellement à $\{S_t\}$, $\{Y_t\}$ est une chaîne de Markov non-homogène d'ordre r, la distribution conditionnelle de Y_t sachant $\{S_{t'}\}_{t' \leq t}$ et $\{Y_{t'}\}_{t' < t}$ dépendant uniquement de $\{Y_{t'}\}_{t'=t-r}^{t-1}$ et de S_t. Lorsque $\{X_t\}$ désigne un processus quelconque, nous noterons $X_s^t = (X_s, X_{s+1}, \ldots, X_t)$ pour $s \leq t$.

Nous dirons que le processus $\{S_t, Y_t\}$ suit un ***modèle autorégressif à changements de régimes markoviens*** (noté dans la suite ***MS – AR*** pour Markov Switching AutoRegressive model) lorsque

- $\{S_t\}_{t\geq 0}$ est une chaîne de Markov à espace d'état $\boldsymbol{S}$. Dans la suite, cette chaîne pourra être homogène ou non-homogène et nous noterons $R_\theta^{(t)}(s,A) = P_\theta(S_t \in A | S_{t-1} = s)$ le noyau de transition de cette chaîne de Markov. Nous nous placerons, sauf mention contraire, dans le cas où $\boldsymbol{S}$ est un ensemble fini. Nous supposerons alors que $\boldsymbol{S} = \{1 \ldots M\}$, M désignant le nombre de régimes. Nous noterons $q_\theta^{(t)}(i,j) = P_\theta(S_t = j | S_{t-1} = i)$ et μ la mesure comptage sur $\boldsymbol{S}$.
- $\{Y_t\}_{t>-r}$ est un processus à espace d'état $\boldsymbol{Y} \subset \boldsymbol{R}^d$ tel que pour $t>0$, la loi conditionnelle de Y_t sachant Y_{1-r}^{t-1} et S_0^t dépend uniquement de Y_{t-r}^{t-1} et de S_t. Nous supposerons en outre que ces lois conditionnelles admettent une densité $g_\theta(y | Y_{t-r}^{t-1}, S_t)$ par rapport à une même mesure dominante ν sur $\boldsymbol{Y}$.

Nous supposerons que le paramètre θ appartient à un sous ensemble compact et d'intérieur non vide Θ de $\boldsymbol{R}^p$. Le processus $\{S_t\}$ sera appelé ***régime***. Lorsque nous nous intéresserons à l'estimation du paramètre θ, nous supposerons que le processus $\{S_t\}$ n'est pas observé (***processus caché***), et l'inférence sur le paramètre θ devra alors être faite uniquement à partir du ***processus observé*** $\{Y_t\}$ (cf paragraphe 1.c).

Nous noterons $\overline{Y}_t = Y_{t-r+1}^t$ et $X_t = \{S_t, \overline{Y}_t\}$. Le processus $\{X_t\}$ ainsi défini est une chaîne de Markov d'ordre 1 à valeurs dans $\boldsymbol{S} \times \boldsymbol{Y}^r$, dont le noyau sera noté $\Pi_\theta^{(t)}$. Il est caractérisé, pour f une fonction bornée sur $\boldsymbol{S} \times \boldsymbol{Y}^r$, $s_0 \in \boldsymbol{S}$ et $\bar{y}_0 = (y_{-r+1}, \ldots, y_0)$, par

$$\Pi_\theta^{(t)} f(s_0, \bar{y}_0) = \int_{S \times \boldsymbol{Y}^r} f(s, y_{-r+2}, \ldots, y_1) g_\theta(y_1 | \bar{y}_0, s) q_\theta^{(t)}(s_0, s) \mu(\mathrm{d}s) \nu(\mathrm{d}y_{-r+1}) \ldots \nu(\mathrm{d}y_0)$$

Lorsque la chaîne de Markov $\{S_t\}$ est homogène, il en est de même de chaîne de Markov $\{X_t\}$: les noyaux de transition de ces chaînes de Markov seront alors notés R_θ et Π_θ respectivement. Nous noterons $P_{\theta,\pi}$ la loi induite sur $\boldsymbol{S}^{\boldsymbol{N}} \times \boldsymbol{Y}^{\boldsymbol{N}}$ par la chaîne de Markov $\{X_t\}_{t\geq 0}$ lorsque $X_0 \sim \pi$ et $E_{\theta,\pi}$ l'espérance associée à cette loi. Nous serons amenés à plusieurs reprises à faire l'hypothèse que le noyau Π_θ possède une unique probabilité invariante. Des conditions permettant de garantir l'existence et l'unicité de cette probabilité invariante sont données au paragraphe 1.b. Dans ce cas, π_θ désignera cette probabilité invariante et nous noterons $\bar{P}_\theta = P_{\theta,\pi_\theta}$ et $\bar{E}_\theta = E_{\theta,\pi_\theta}$.

Finalement, p_T désignera la densité par rapport à la mesure $(\mu \otimes \nu)^{\otimes T}$ sur $(S \times Y)^T$ pour $T \in \boldsymbol{N}$. En pratique, nous noterons abusivement cette densité p pour les différentes valeurs de T, celle-ci étant généralement évidente.

Lorsque les modèles autorégressifs sont d'ordre $r = 1$, la relation entre les processus $\{Y_t\}$ et $\{S_t\}$ peut être résumée par le graphe d'indépendance conditionnelle représenté sur la figure 1.1. Une définition rigoureuse de ce type de graphe peut être trouvée dans *Durand* (2003, [43]).

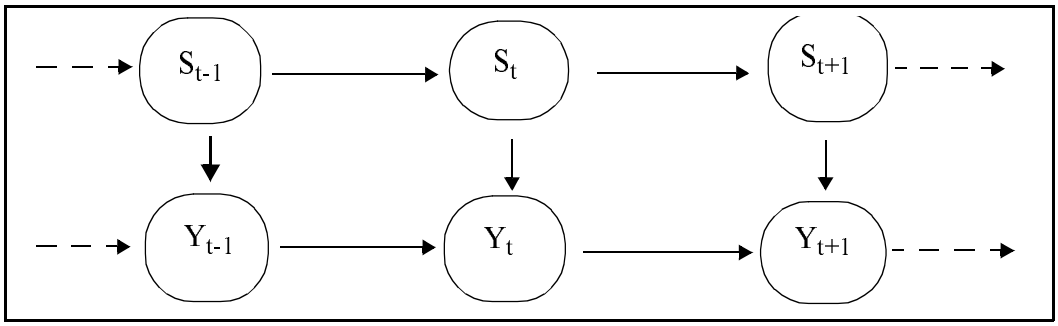


figure 1.1 : Graphe d'indépendance conditionnelle pour le modèle MS – AR d'ordre 1

1.a.2 Exemples

Les modèles autorégressifs à changements de régimes englobent différents modèles largement répandus dans la littérature statistique. Ainsi, lorsque $r = 0$, le processus $\{S_t, Y_t\}$ devient une ***Chaîne de Markov Cachée*** (noté ***CMC*** dans la suite) : la loi conditionnelle de Y_t sachant Y_{1-r}^{t-1} et S_0^t dépend uniquement de S_t (cf figure 1.2). Ce type de modèle a été abondamment étudié et un état des lieux récents des résultats existant pour ce type de modèle peut être trouvé dans *Ephraim et al.* (2002, [46]). Ils se sont révélées être particulièrement utiles dans de nombreux domaines d'applications (cf [46] et *MacDonald et al.* (1997, [76]). Lorsque l'on suppose en outre que le processus $\{S_t\}$ est i.i.d., on obtient les ***modèles de mélange*** (cf figure 1.3).

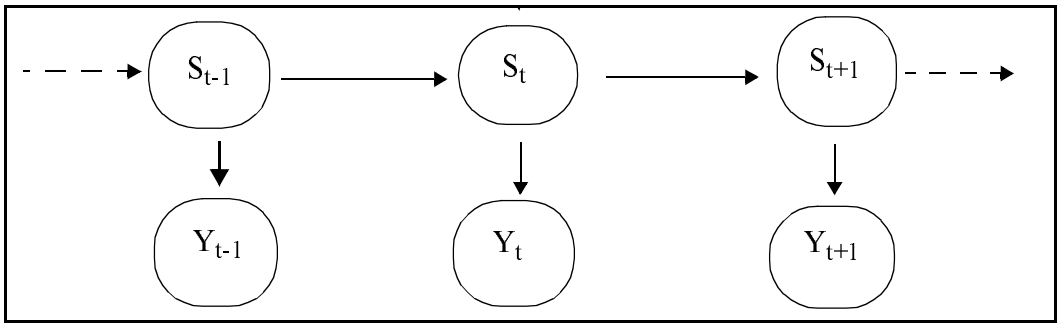


figure 1.2 : Graphe d'indépendance conditionnelle pour le modèle CMC

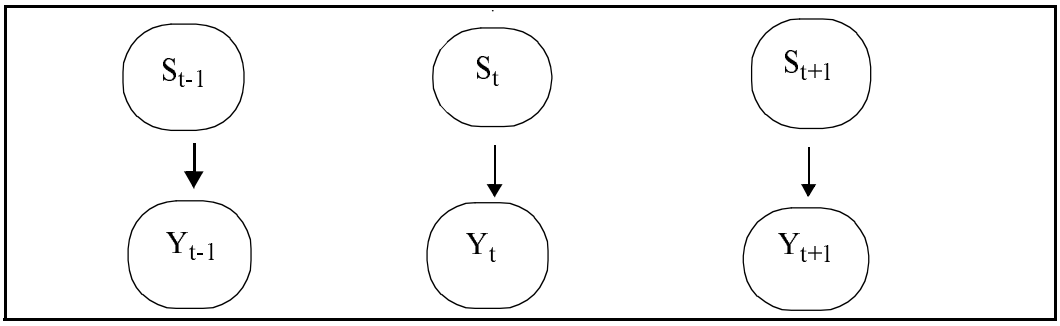


figure 1.3 : Graphe d'indépendance conditionnelle pour les modèles de mélange

Lorsque l'espace $\boldsymbol{S}$ est réduit à un singleton, on obtient les modèles markoviens, qui englobent en particulier les ***modèles autorégressifs***. Les modèles autorégressifs les plus couramment

utilisés dans la littérature s'écrivent sous la forme :

$$Y_t = f(\bar{Y}_{t-1}) + \varepsilon_t \tag{1.1}$$

avec avec f : $\boldsymbol{Y}^r \to \boldsymbol{Y}$ et $\{\varepsilon_t\}_{t \in N}$ un bruit blanc. Nous appellerons ces modèles ***autorégressifs non linéaires* (*NAR*).** Parmi ces modèles, on trouve en particulier le ***modèle autorégressif linéaire* (*LAR*)**, et qui correspondent au cas où la fonction f est une forme linéaire. Plus précisément, nous dirons que le processus $\{Y_t\}_{t>-r}$ suit un modèle LAR d'ordre r lorsque

$$Y_t = \sum_{i=1}^{r} a_i Y_{t-i} + b + \varepsilon_t \tag{1.2}$$

avec $a_i \in M_d(\boldsymbol{R})$, $b \in M_{d,1}(\boldsymbol{R})$ pour $i \in \{1 \ldots r\}$ et $\{\varepsilon_t\}_{t \in N}$ un bruit blanc. Parmi les modèles NAR, on peut aussi citer les réseaux de neurones ou encore certains modèles à seuil que nous décrivons brièvement au chapitre suivant. Notons que dans le cas où le résidu possède une densité par rapport à une mesure dominante, il en est de même pour les lois conditionnelles $P(Y_t | \bar{Y}_{t-1} = \bar{y}_{t-1})$. Par exemple, dans le cas où le résidu $\{\varepsilon_t\}_{t \in N}$ est un bruit blanc gaussien centré de variance Σ alors la loi conditionnelle de Y_t sachant $\bar{Y}_{t-1} = \bar{y}_{t-1}$ est alors une loi normale de moyenne $E[Y_t | \bar{Y}_{t-1} = \bar{y}_{t-1}] = f(\bar{y}_{t-1})$ et de variance $var(Y_t | \bar{Y}_{t-1} = \bar{y}_{t-1}) = \Sigma$.

Cependant, ce type de modèle n'est pas adapté pour décrire certaines séries temporelles et notamment celles qui sont à valeurs dans sous ensemble donné de $\boldsymbol{R}^d$. Ainsi, dans le chapitre 2, nous nous focaliserons sur les séries temporelles correspondant à l'intensité et à la direction du vent, et qui sont à valeurs respectivement dans $\boldsymbol{R}^+$ et $\boldsymbol{R}/(2\pi\boldsymbol{Z})$. Supposons tout d'abord que le processus $\{Y_t\}$ est à valeurs dans $\boldsymbol{R}^+$. Il semble alors naturel de remplacer la loi normale par une loi dont le support est $\boldsymbol{R}^+$. Plus précisément, nous proposons d'utiliser le ***modèle autorégressif gamma* (γ*AR*)** pour décrire l'évolution de l'intensité du vent. On suppose alors que la loi conditionnelle de Y_t sachant que $\bar{Y}_{t-1} = \bar{y}_{t-1}$ est une loi gamma de moyenne

$$\mu(\bar{y}_{t-1}) = E[Y_t | \bar{Y}_{t-1} = \bar{y}_{t-1}] = \sum_{i=1}^{r} a_i y_{t-i} + b$$

avec $a_i \geq 0$ et $b > 0$ des paramètres et de variance σ^2. Si nous notons

$$\gamma(y;\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)} e^{-y/\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} \mathbf{1}_{\boldsymbol{R}^+}(y)$$

la densité de la loi gamma, avec α, β des paramètres strictement positifs, sa moyenne m vaut $\alpha\beta$ et son écart type σ vaut $\beta\sqrt{\alpha}$. On vérifie alors aisément que $\alpha = m^2/\sigma^2$ et $\beta = \sigma^2/m$. La densité de loi conditionnelle de Y_t sachant $\bar{Y}_{t-1} = \bar{y}_{t-1}$ est alors $\gamma(y;\alpha_t,\beta_t)$ avec $\alpha_t = ((\mu(\bar{y}_{t-1}))/\sigma)^2$ et $\beta_t = \sigma^2/\mu(\bar{y}_{t-1})$, c'est à dire

$$p(y_t|\bar{y}_{t-1}) = \frac{\mu(\bar{y}_{t-1})}{\sigma^2\Gamma\left(\left(\frac{\mu(\bar{y}_{t-1})}{\sigma}\right)^2\right)}\left(\frac{y_t\mu(\bar{y}_{t-1})}{\sigma^2}\right)^{\left(\frac{\mu(\bar{y}_{t-1})}{\sigma}\right)^2-1}\exp\left(-\frac{y_t\mu(\bar{y}_{t-1})}{\sigma^2}\right)\mathbf{1}_{R^+}(y_t)$$

A notre connaissance, les propriétés théoriques de ce type de modèle ont été peu étudiées, la littérature existant sur les modèles autorégressifs étant principalement consacrée aux modèles NAR. Dans la suite nous ne nous intéressons pas directement au modèle γAR, mais plus généralement aux modèles autorégressif à changements de régimes dans lequel l'évolution dans chaque régime est décrit par un modèle γAR. Les modèles γAR sont un cas particulier de ce type de modèle, lorsque $M = 1$, et on peut alors déduire des résultats donnés aux paragraphes suivants des conditions assurant la stabilité, ainsi que la consistance et la normalité asymptotique des estimateurs du maximum de vraisemblance pour le modèle γAR.

Notons que le choix de la loi gamma est arbitraire, et que nous pouvons construire le même type de modèle en utilisant une autre distribution à support dans $\boldsymbol{R}^+$ dont les paramètres sont caractérisés par les moments d'ordre 1 et 2 (loi log-normale ou loi de Weibull par exemple). Cependant, comme nous le verrons au chapitre 2, les résultats obtenus avec la loi gamma sont satisfaisants pour les données de vent, et nous avons choisi de nous cantonner à ce modèle dans la suite.

Nous reviendrons au chapitre 2 sur la modélisation des séries temporelles directionnelles, c'est à dire à valeur dans $\boldsymbol{R}/(2\pi\boldsymbol{Z})$. En particulier, différents modèles autorégressifs spécifiquement pour ce type de séries temporelles sont décrits au paragraphe 2.a.3. Enfin, notons que la liste donnée ci-dessus est loin d'être exhaustive et de nombreux autres modèles autorégressifs ont été proposés. On peut ainsi citer les modèles $GARCH$ qui sont décrits plus précisément au paragraphe 2.a.3 et les modèles RCA définis au paragraphe 1.b.1. Une liste plus complète peut être trouvée dans *Grunwald et al.* (1999, [52]) et *Tong* (1990, [115]).

Modèle autorégressif à changements de régimes markoviens

Dans le modèle $MS-AR$ introduit par Hamilton, la chaîne de Markov cachée prend un nombre fini de valeurs, chacune de ces valeurs correspondant à un état de l'économie et l'évolution de la variable observée dans chaque régime est décrit par un modèle autorégressif linéaire avec innovations gaussiennes. Nous dirons que $\{Y_t\}$ suit un modèle ***autorégressif linéaire à changements de régimes markoviens* (*MS – LAR*)** lorsque $\{Y_t\}$ suit un modèle $MS-AR$ dans lequel la variable cachée est une chaîne de Markov homogène et que l'évolution du processus observé dans chaque régime est décrite par un modèle autorégressif linéaire. On a alors, pour $t>0$:

$$Y_t = \sum_{i=1}^{r} a_i^{(S_t)}Y_{t-i} + b^{(S_t)} + h^{(S_t)}\varepsilon_t \qquad (1.3)$$

avec $a_i^{(s)} \in M_d(\boldsymbol{R})$, $b^{(s)} \in M_{d,1}(\boldsymbol{R})$ et $\sigma^{(s)} = h^{(s)}(h^{(s)})' \in S_d^+(\boldsymbol{R})$ pour $i \in \{1\ldots r\}$ et $s \in \boldsymbol{S}$

et $\{\varepsilon_t\}_{t\in N}$ un bruit blanc vectoriel centré réduit. On peut réécrire l'équation (1.3) sous la forme

$$\bar{Y}_t = A^{(S_t)}\bar{Y}_{t-1} + B^{(S_t)} + H^{(S_t)}E_t \tag{1.4}$$

avec

$$\bar{Y}_t = \begin{bmatrix} Y_t \\ \vdots \\ Y_{t-r+1} \end{bmatrix},\ A^{(s)} = \begin{bmatrix} a_1^{(s)} & a_2^{(s)} & \dots & a_{r-1}^{(s)} & a_r^{(s)} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix},\ B^{(s)} = \begin{bmatrix} b^{(s)} \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ H^{(s)} = \begin{bmatrix} h^{(s)} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \text{ et}$$

$E_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ pour $s \in \boldsymbol{S}$. Nous noterons aussi $\Sigma^{(s)} = H^{(s)}(H^{(s)})'$ pour $s \in \boldsymbol{S}$. Cette écriture est

en particulier utilisée au paragraphe 1.b.1 puisqu'elle permet de ramener l'étude de la stabilité des modèles $MS-LAR$ d'ordre $r>1$ à celle des modèles $MS-LAR$ d'ordre 1. Au chapitre 3, nous proposons un modèle $MS-LAR$ pour décrire l'évolution spatio-temporelle des champs de vent. Une attention tout particulière est donc portée à ce modèle dans la suite de ce chapitre.

Un autre type de modèle largement étudié dans la littérature est le modèle ***autorégressif non-linéaire à changements de régimes markoviens*** ***(MS – NAR)*** dans lequel l'évolution du processus observé dans chaque régime est décrit par un modèle autorégressif non-linéaire. On suppose alors que

$$Y_t = f(\bar{Y}_{t-1}, S_t) + H^{(S_t)}\varepsilon_t \tag{1.5}$$

avec $f(.,s) : \boldsymbol{Y}^r \to \boldsymbol{Y}$, $\Sigma^{(s)} = H^{(s)}(H^{(s)})' \in S_d^+(\boldsymbol{R})$ pour $s \in \boldsymbol{S}$ et $\{\varepsilon_t\}_{t\in N}$ un bruit blanc. Nous verrons en particulier qu'il existe de nombreux résultats théoriques pour ce type de modèle (stabilité, propriétés asymptotiques des estimateurs du maximum de vraisemblance...).

Enfin, nous parlerons de modèle ***autorégressif gamma à changements de régimes markoviens d'ordre r*** ***(MS – γAR)*** lorsque l'évolution dans chaque régime est décrite par un modèle γAR. On suppose alors que la loi conditionnelle de Y_t sachant que $\bar{Y}_{t-1} = \bar{y}_{t-1}$ et $S_t = s_t$ est une loi gamma de moyenne

$$\mu(\bar{y}_{t-1}, s_t) = E[Y_t \,|\, \bar{Y}_{t-1} = \bar{y}_{t-1}, S_t = s_t] = \sum_{i=1}^{r} a_i^{(S_t)} y_{t-i} + b^{(S_t)}$$

et d'écart type $\sigma^{(s_t)}$ avec $a_i^{(s)} \geq 0$ et $b^{(s)} > 0$ pour $i \in \{1 \dots r\}$ et $s \in \boldsymbol{S}$. Nous montrons au chapitre 2 que ce modèle est adapté pour décrire l'évolution de l'intensité du vent et nous porterons donc une attention toute particulière à l'étude de ce modèle dans les paragraphes ci-des-

sous. Nous noterons $\bar{a}^{(s)} = (a_r^{(s)}, \ldots, a_1^{(s)})$ de telle manière que $\mu(\bar{y}_{t-1}, s_t) = \bar{a}^{(S_t)}\bar{y}_{t-1} + b^{(S_t)}$ et $\theta_R^{(s)} = (a_r^{(s)}, \ldots, a_1^{(s)}, b^{(s)}, \sigma^{(s)})$ les paramètres servant à décrire l'évolution du processus observé dans le $s^{\text{ème}}$ régime. Nous noterons aussi

$$A^{(s)} = \begin{bmatrix} a_1^{(s)} & a_2^{(s)} & \ldots & a_{r-1}^{(s)} & a_r^{(s)} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \quad \text{et} \quad B^{(s)} = \begin{bmatrix} b^{(s)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{de telle manière que}$$

$E[\bar{Y}_t | \bar{Y}_{t-1}, S_t] = A^{(S_t)}\bar{Y}_{t-1} + B^{(S_t)}$.

Finalement, le modèle $MS-\gamma AR$ sera paramètré par $\theta = (\theta_S, \theta_R^{(1)}, \ldots, \theta_R^{(M)})$ avec

$$\theta_S \in \Theta_S = \left\{ q(i,j) \geq 0 \,|\, \sum_{j=1}^{M} q(i,j) = 1 \right\}$$

et $\theta_R^{(s)} \in \Theta_R^{(s)}$ un sous espace compact de $(\boldsymbol{R}^+)^r \times \boldsymbol{R}^{+*} \times \boldsymbol{R}^{+*}$. Nous noterons aussi $\theta_R = (\theta_R^{(1)}, \ldots, \theta_R^{(M)})$ l'ensemble des paramètres servant à décrire l'évolution du processus observé dans les différents régimes et $\Theta_R = \Theta_R^{(1)} \times \ldots \times \Theta_R^{(M)}$.

Plus généralement, l'espace des paramètres Θ des différents modèles utilisés aux chapitres 2 et 3 peut se décomposer sous la forme $\Theta_S \times \Theta_R$ avec Θ_S l'espace des paramètres servant à décrire l'évolution de la chaîne cachée $\{S_t\}$ et Θ_R l'espace des paramètres servant à décrire l'évolution du processus observé $\{Y_t\}$ dans les différents régimes. Par contre, dans le modèle du chapitre 3, l'espace Θ_R ne peut pas se décomposer sous la forme $\Theta_R^{(1)} \times \ldots \times \Theta_R^{(M)}$, les mêmes paramètres servant à décrire l'évolution du processus observé $\{Y_t\}$ dans les différents régimes.

Il est parfois utile d'utiliser une autre paramétrisation pour la matrice de transition de la chaîne cachée. En effet, avec celle donnée ci-dessus, on peut être amené à résoudre des problèmes d'optimisation sous contraintes lors du calcul des estimateurs du maximum de vraisemblance (cf paragraphe 1.c.3). Afin d'éviter ces problèmes d'optimisation sous contraintes, on peut utiliser les paramètres $l(i,j) = \ln(q(i,j)/q(i,i))$ pour $i \neq j$. Ces paramètres sont bien définis si les coefficients diagonaux $q(i,i)$ sont non nuls, ce qui en pratique ne semble pas contraignant pour les applications envisagées dans la suite de cette thèse. Il est facile de vérifier que cette nouvelle paramétrisation définit une bijection de

$$\Theta_S^* = \left\{ (q(i,j))_{i,j \in \{1\ldots M\}} \,|\, q(i,j) \geq 0, \sum_{k=1}^{M} q(i,k) = 1, q(i,i) > 0 \right\}$$

dans $\boldsymbol{R}^{M(M-1)}$, l'inverse de cette transformation étant donnée par $q(i,j) = e^{l(i,j)} \Big/ \left(1 + \sum_{k \neq i} e^{l(i,k)} \right)$ si $i \neq j$ et $q(i,i) = 1 \Big/ \left(1 + \sum_{k \neq i} e^{l(i,k)} \right)$.

Le modèle $MS - \gamma AR$ ne fait pas partie de la famille des modèles $MS - NAR$, qui est le type de modèle généralement étudié dans la littérature. Dans les paragraphes suivants, nous étendrons différents résultats théoriques valables pour les modèles $MS - NAR$ aux modèles $MS - \gamma AR$.

Dans les différents modèles décrits ci-dessus, la chaîne cachée est supposée homogène. Cependant, nous verrons qu'il est parfois utile de la supposer non-homogène afin de modéliser, par exemple, la relation entre le processus observé $\{Y_t\}$ et certaines variables explicatives (cf 2.c). L'étude théorique de ces modèles est plus complexe et sera peu abordée dans ce chapitre et nous nous placerons donc, sauf mention contraire, dans le cas où la chaîne $\{S_t\}$ est homogène. Lorsque la chaîne est non-homogène, nous noterons le modèle $\boldsymbol{NHMS - AR}$ (**Non-Homogeneous Markov Switching AutoRegressive model**).

1.b Stabilité

L'étude de la stabilité des modèles $MS - AR$ est relativement complexe. Ainsi, même dans le cas le plus simple des modèles $MS - LAR$ à deux régimes, des études numériques ont permis de montrer qu'un modèle avec un ou deux régimes instables peuvent conduire à un modèle stable et que vice-versa, la stabilité des différents régimes ne garantit pas la stabilité du modèle (cf *Holst et al.* (1994, [62])).

Nous supposerons que le paramètre θ est fixé, et, afin de simplifier les notations, nous omettrons ce paramètre dans l'écriture des différents noyaux de transition dans la suite de ce paragraphe. Ainsi, par exemple, nous noterons Π au lieu de Π_θ le noyau de transition de la chaîne de Markov $\{X_t\} = \{S_t, \overline{Y}_t\}$.

Nous nous intéressons ici tout d'abord à chercher des conditions garantissant l'existence et l'unicité d'une probabilité invariante pour ce noyau de transition. Supposons que la chaîne de Markov $\{X_t\}$ possède une probabilité invariante π et notons $\pi^{(S)}$ la trace de cette probabilité sur $\boldsymbol{S}$. On peut vérifier aisément que la chaîne de Markov $\{S_t\}$ est stationnaire sous $\pi^{(S)}$. Dans la suite, nous supposerons que la chaîne de Markov $\{S_t\}$ est irréductible et apériodique et nous noterons $\pi^{(S)} = (\pi_1, \ldots, \pi_M)$ sa probabilité invariante. Nous donnons aussi des conditions garantissant l'existence de moments d'ordre $\alpha \geq 1$ pour la loi stationnaire. Ceci nous servira en particulier au paragraphe 1.d.2 pour montrer la consistance et la normalité asymptotique des *EMV*.

Nous proposons tout d'abord un état des lieux des résultats existants sur la stabilité des modèles $MS - LAR$ et $MS - NAR$ aux paragraphes 1.b.1 et 1.b.2 respectivement, puis nous nous focalisons ensuite sur le modèle $MS - \gamma AR$. Les modèles d'ordre 1 sont étudiés au paragraphe

1.b.3 et les modèles d'ordre supérieur au paragraphe 1.b.4.

Les principaux ouvrages de référence utilisés dans ce paragraphe, en ce qui concerne la stabilité des chaînes de Markov, sont *Meyn et al.* (1993, [84]) et *Robert* (1996, [97]).

1.b.1 Stabilité des modèles *MS-LAR*

Afin de montrer la stabilité des modèles $MS-LAR$, on peut utiliser les résultats existants sur les modèles autorégressifs à coefficients aléatoires (RCA), c'est à dire de la forme

$$Y_t = A_t Y_{t-1} + B_t \tag{1.6}$$

avec A_t et B_t des suites de matrices aléatoires. En effet, les modèles $MS-LAR$ définis par l'équation (1.4) sont des cas particuliers de ces modèles avec $A_t = A^{(S_t)}$ et $B_t = B^{(S_t)} + \Sigma^{(S_t)} E_t$. Des conditions garantissant l'existence et l'unicité d'une solution stationnaire stricte pour les modèles RCA peuvent être trouvées dans *Brandt* (1986, [24]) et *Bougerol et al.* (1992, [20]) lorsque le processus $\{A_t, B_t\}$ est strictement stationnaire et ergodique. Il y est démontré que si les deux conditions suivantes sont vérifiées :

(i) $E[\ln^+(\|B_t\|)] < \infty$ et $E[\ln^+(\|A_t\|)] < \infty$ (avec $\ln^+(x) = max(\ln(x), 0)$)

(ii) $p = inf\left\{ \frac{1}{t} E[\ln(\|A_t A_{t-1} \ldots A_1\|)] \,|\, t \geq 1 \right\} < 0$

alors la série $Y_t = \sum_{k \geq 0} A_t \ldots A_{t-k+1} B_{t-k}$ converge p.s et le processus $\{Y_t\}$ ainsi défini est l'unique solution stationnaire à l'équation (1.6).

Dans les conditions **(i)** et **(ii)** ci-dessus, $\|.\|$ désigne une norme quelconque sur $\boldsymbol{R}^d$ ainsi que la norme matricielle associée. On peut en fait montrer que la quantité

$$p = inf\left\{ \frac{1}{t} E[\ln(\|A_t A_{t-1} \ldots A_1\|)] \,|\, t \geq 1 \right\}$$

qui intervient dans l'hypothèse **(ii)** est indépendante de la norme $\|.\|$ choisie. La proposition suivante est une conséquence immédiate de ce résultat.

Proposition 1.1 (stabilité des modèles MS – LAR)

Supposons que la chaîne de Markov $\{S_t\}$ soit stationnaire. Si les conditions suivantes sont vérifiées:

(i) $E_{\pi^{(S)}}[\ln^+(\|H^{(S_1)} \varepsilon_1 + B^{(S_1)}\|)] < \infty$ *et* $E_{\pi^{(S)}}[\ln^+(\|A^{(S_1)}\|)] < \infty$

$$\textit{(A1)}\ p = \inf\left\{\frac{1}{t}E_{\pi^{(S)}}[\ln(\left\|A^{(S_t)}A^{(S_{t-1})}\ldots A^{(S_1)}\right\|)]\,\middle|\, t\geq 1\right\} < 0$$

alors le processus $\{Y_t\}$ $MS-LAR$ *défini par l' équation (1.4) possède une unique solution strictement stationnaire, qui est donnée par*

$$Y_t = \sum_{k\geq 0} A^{(S_t)}\ldots A^{(S_{t-k+1})}(B^{(S_{t-k})} + H^{(S_{t-k})}\varepsilon_{t-k})$$

Lorsque chaîne $\{S_t\}$ prend un nombre fini de valeurs, la condition **(i)** de la proposition 1.1 est automatiquement vérifiée si $E[\ln^{+}\|\varepsilon_1\|] < \infty$. Par contre, la condition **(A1)** est difficile à vérifier en pratique. En fait, il suffit de montrer qu'il existe $t \geq 1$ tel que $E_{\pi^{(S)}}[\ln(\left\|A^{(S_t)}A^{(S_{t-1})}\ldots A^{(S_1)}\right\|)] < 0$ mais cela peut s'avérer fastidieux. Toutefois, dans le cas des modèles univariés d'ordre 1 (i.e $r = d = 1$), pour lesquels $a^{(s)} = A^{(s)} \in \boldsymbol{R}$, on peut facilement vérifier que cette condition est équivalente à la condition **(A1')** ci-dessous.

$$\textit{(A1')}\ \beta = E_{\pi^{(S)}}[\ln(\left|a^{(S_t)}\right|)] = \sum_{s\in S}\pi_s \ln\left|a^{(s)}\right| < 0$$

On en déduit, en particulier, que si les différents régimes sont stables (i.e $\left|a^{(s)}\right| < 1$ pour $s \in \boldsymbol{S}$) alors le modèle à changements de régimes est stable, ce qui n'est pas vrai dans le cas général (cf *Holst et al.* (1994, [62])). On peut aussi utiliser ce critère pour construire des modèles $MS-LAR$ stables avec un ou plusieurs régimes instables. Il suffit pour cela que le temps moyen passé dans les régimes instables soit suffisamment faible par rapport au temps passé dans les régimes stables.

Dans le cas des modèles d'ordre supérieur à 1 ou multivarié, la condition **(A1)** n'est donc pas satisfaisante. En outre, cette proposition ne garantit pas l'existence des moments d'ordre 2 à la solution stationnaire, ce qui est utile pour montrer la consistance des estimateurs du maximum de vraisemblance (cf *Krishnamurthy et al.* (1998, [68])). Le critère suivant est proposé dans *Holst et al.* (1994, [62]) et démontré dans *Yao* (2001, [125]).

Proposition 1.2 (stabilité des modèles MS – LAR)

Si la condition suivante est vérifiée

$$\textbf{(A2)}\ \rho(N) < 1 \ \textit{avec}\ N = \begin{bmatrix} q(1,1)(A^{(1)}\otimes A^{(1)}) & \ldots & q(1,M)(A^{(M)}\otimes A^{(M)}) \\ \vdots & q(i,j)(A^{(j)}\otimes A^{(j)}) & \vdots \\ q(M,1)(A^{(1)}\otimes A^{(1)}) & \ldots & q(M,M)(A^{(M)}\otimes A^{(M)}) \end{bmatrix}$$

alors la condition ***(A1)*** *est vérifiée. Si on suppose en outre que* $E[\|\varepsilon_1\|^2] < +\infty$, *alors la série* $Y_t = \sum_{k \geq 0} A^{(S_t)} \ldots A^{(S_{t-k+1})}(B^{(S_{t-k})} + H^{(S_{t-k})}\varepsilon_{t-k})$ *converge p.s et dans* L^2. *Le processus* $\{Y_t\}$ *ainsi défini est l'unique solution stationnaire de l' équation (1.4).*

Cette proposition fournit donc un moyen pratique de vérifier la stabilité des modèles $MS-LAR$ et donne aussi une condition suffisante pour que la loi stationnaire possède un moment d'ordre 2.

1.b.2 Stabilité des modèles *MS-NAR*

Des conditions garantissant la stabilité des modèles $MS-NAR$ d'ordre $r = 1$ peuvent être trouvées dans *Francq et al.* (1998, [49]) et *Yao et al.* (2000, [124]), dans les cas particuliers où les fonctions $f(., s) : \boldsymbol{Y} \to \boldsymbol{Y}$ sont lispschitziennes ou sous linéaires pour $s \in \boldsymbol{S}$.

Cas lipschitzien

Supposons tout d'abord que les fonctions $f(., s)$ sont lispschitziennes, c'est à dire qu'il existe des constantes $a^{(s)}$ et une norme $\|.\|$ sur $\boldsymbol{R}^d$, telles que pour $s \in \boldsymbol{S}$ et $y, y' \in \boldsymbol{Y}$, on ait

$$\|f(y, s) - f(y', s)\| \leq a^{(s)}\|y - y'\|$$

Les propositions 1.3 et 1.4 ci-dessous sont démontrées dans *Yao et al.* (2000, [124]).

Proposition 1.3 (Stabilité des modèles MS – NAR dans le cas Lipschitzien)

Si la condition ***(A1')*** *est vérifiée et* $E[\ln^+(\|\varepsilon_1\|)] < \infty$ *alors l'équation (1.5) possède une unique solution stationnaire et ergodique.*

Proposition 1.4 (Existence de moments pour les modèles MS – NAR dans le cas lipschitzien)

Si les conditions suivantes sont vérifiées

(i) *Il existe* $c \geq 1$ *tel que* $E[\|\varepsilon_1\|^c] < \infty$

(A3) $\rho(N_c) < 1$ *avec* $N_c = \begin{bmatrix} q_{11}(a^{(1)})^c & \ldots & q_{1M}(a^{(M)})^c \\ \vdots & q_{ij}(a^{(j)})^c & \vdots \\ q_{M,1}(a^{(1)})^c & \ldots & q_{MM}(a^{(M)})^c \end{bmatrix}$

alors l'équation (1.5) possède une unique solution stationnaire et ergodique, et cette solution possède un moment d'ordre c.

La proposition 1.4 est démontrée également dans *Francq et al.* (1998, [49]) dans le cas particu-

lier où $c = 1$.

Dans *Yao et al.* (2000, [124]), il est montré que :

$$\beta = E_{\pi^{(S)}}[\ln(|a^{(S_t)}|)] = \sum_{s \in S} \pi_s \ln(a^{(s)}) \leq \ln(\rho(N_1))$$

On en déduit que si la condition **(A3)** est vérifiée avec $c > 0$ alors $c \sum_{s \in S} \pi_s \ln(a_s) \leq \ln(\rho(N_c)) < 0$ et la condition **(A1')** est donc vérifiée.

Cas sous linéaire

Supposons maintenant que les fonctions $f(., s)$ sont sous-linéaires, c'est à dire qu'il existe des constantes $a^{(s)}$ et $b^{(s)}$ et une norme $\|.\|$ sur $\boldsymbol{R}^d$, tel que pour $s \in \boldsymbol{S}$ et $y \in \boldsymbol{Y}$, on ait

$$\|f(y, s)\| \leq a^{(s)}\|y\| + b^{(s)}$$

Les proposition suivantes sont démontrées dans *Yao et al.* (2000, [124]).

Proposition 1.5 (Stabilité des modèles MS – NAR dans le cas sous-linéaire)

Supposons que les conditions ***(i)*** *et* ***(ii)*** *ci-dessous sont vérifiées :*

(i) *la variable* ε_1 *possède une densité par rapport à la mesure de Lebesgue et cette densité est strictement positive sur* $\boldsymbol{R}^d$.

(ii) *il existe* $c > 0$ *tel que* $E[\|\varepsilon_1\|^c]$ *soit fini.*

Si de plus la condition ***(A1')*** *est vérifiée alors la chaîne de Markov* $\{X_t\}$ *est géométriquement ergodique.*

Proposition 1.6 (Existence de moments pour les modèles MS – NAR dans le cas sous-linéaire)

Si les conditions ***(i), (ii)*** *de la proposition 1.5 et la condition* ***(A3)*** *sont vérifiées avec* $c \geq 1$, *alors la solution stationnaire possède un moment d'ordre* c.

Si on compare les hypothèses des propositions 1.3 et 1.4 avec celles des propositions 1.5 et 1.6 respectivement, on peut remarquer que des hypothèses plus fortes sont faites sur le résidu dans le cas sous-linéaire. La condition **(i)** de la proposition 1.5 permet de garantir que la chaîne $\{X_t\}$ est irréductible et fortement fellerienne. Elle pourrait être remplacée par d'autres conditions (cf *Yao et al.* (2000, [124])).

1.b.3 Stabilité des modèles *MS-γAR* d'ordre 1

Les propositions données au paragraphe précédent concernent uniquement les modèles $MS - NAR$ et ne s'appliquent donc pas aux modèles $MS - \gamma AR$. Nous allons cependant mon-

trer que les techniques utilisées pour démontrer la proposition 1.5 s'adaptent à ce type de modèle et, afin de simplifier l'exposé, nous considérons dans un premier temps uniquement les modèles d'ordre $r = 1$.

Afin de montrer la stabilité des modèles $MS-\gamma AR$ nous allons utiliser les résultats généraux existants sur la stabilité des chaînes de Markov. Plus précisément, nous allons établir que le noyau de transition de la chaîne de Markov $\{X_t\}$ vérifie une ***condition de dérive*** de la forme **(D1)**.

(D1) *Il existe une fonction* $V \geq 1$, *des constantes* $K < 1$ *et* $L < \infty$, *un ensemble petit* C *tels que*

$$\Pi V(x) = E[V(X_1) \big| X_0 = x] \leq KV(x) + L\mathbf{1}_C(x) \qquad (1.7)$$

Il est montré dans *Meyn et al.* (1993, [84]) que si $\{X_t\}$ est une chaîne de Markov ϕ-irréductible et apériodique de noyau de transition Π et que ce noyau vérifie la condition **(D1)** alors cette chaîne de Markov est V-uniformément ergodique, et en particulier elle est géométriquement ergodique.

Il est parfois plus facile de montrer que le noyau Π^p vérifie une condition de dérive. Plus précisément, notons **(D2)** la condition ci-dessous:

(D2) *Il existe une fonction* $V \geq 1$, *des constantes* $K < 1$ *et* $L < \infty$, *un ensemble petit* C *et un entier* $T > 0$ *tels que*

$$\Pi^T V(x) = E[V(X_T) \big| X_0 = x] \leq KV(x) + L\mathbf{1}_C(x) \qquad (1.8)$$

On peut montrer que si $\{X_t\}$ est une chaîne de Markov ϕ-irréductible et apériodique de noyau de transition Π et que ce noyau vérifie la condition **(D2)** alors la chaîne de Markov $\{X_t\}$ est V-uniformément ergodique, et en particulier elle est géométriquement ergodique (cf *Meyn et al.* (1993, [84]) et *Fonseca* (2000, [48])).
Notons enfin que si l'une des conditions **(D1)** ou **(D2)** est vérifiée avec $V(x) = \|x\|^c$ alors la loi stationnaire possède un moment d'ordre c.

La proposition 1.7 ci-dessous donne en particulier des conditions suffisantes garantissant l'existence et l'unicité d'une loi stationnaire ergodique pour les modèles $MS-\gamma AR$ d'ordre 1. La preuve de cette proposition suit largement celle de la proposition 1.5 donnée dans *Yao et al.* (2000, [124]).

Proposition 1.7 (Stabilité du modèle $MS-\gamma AR$ ***d'ordre 1)***

Si la condition ***(A1')*** *est vérifiée alors le modèle* $MS-\gamma AR$ *est géométriquement ergodique.*

Preuve

L'irréductibilité de la chaîne de Markov $\{X_t\}$ provient de celle de la chaîne cachée et du fait que $g_\theta(y_t|s_t, y_{t-1}) > 0$ pour $y_{t-1}, y_t \in \boldsymbol{R}^+$ et $s_t \in \boldsymbol{S}$. On cherche ensuite à montrer une inégalité de la forme **(D2)**.

Par construction du modèle $MS-\gamma AR$ on a $E[Y_t|Y_0^{t-1}, S_0^{t-1}] = a^{(S_t)}Y_{t-1} + b^{(S_t)}$. En itérant cette relation, on obtient

$$E[Y_t|Y_0, S_0^t] = (a^{(S_t)}\ldots a^{(S_1)})Y_0 + (a^{(S_t)}\ldots a^{(S_2)})b^{(S_1)} + (a^{(S_t)}\ldots a^{(S_3)})b^{(S_2)} + \ldots + b^{(S_t)}$$

En utilisant l'inégalité de Jensen et le fait que $(a+b)^{\frac{1}{n}} \leq a^{\frac{1}{n}} + b^{\frac{1}{n}}$ pour $a, b \geq 0$ et n entier naturel, on en déduit alors que

$$E[(Y_t)^{1/t}|Y_0, S_0^t] \leq (a^{(S_t)}\ldots a^{(S_1)})^{1/t}(Y_0)^{1/t} + (b^{(S_t)})^{1/t} + \sum_{i=2}^{t}(a^{(S_t)}\ldots a^{(S_i)}b^{(S_{i-1})})^{1/t}$$

puis

$$E[(Y_t)^{1/t}|y_0, s_0] \leq E[(a^{(S_t)}\ldots a^{(S_1)})^{1/t}|s_0](y_0)^{1/t} + B_{t,s_0}$$

avec $B_{t,s_0} = E[(b^{(S_t)})^{1/t}|s_0] + \sum_{i=2}^{t} E[(a^{(S_t)}\ldots a^{(S_i)}b^{(S_{i-1})})^{1/t}|s_0] < \infty$. Par ailleurs, comme la chaîne $\{S_t\}$ est irréductible et apériodique, on en déduit que pour toute condition initiale $s_0 \in \boldsymbol{S}$, $\frac{1}{t}\sum_{k=1}^{t}\log(a^{(S_k)}) \to \sum_{s\in S}\pi_s \log(a^{(s)})$ p.s. Or, par hypothèse, cette quantité est strictement négative et on a donc, lorsque $t \to +\infty$:

$$(a^{(S_t)}\ldots a^{(S_1)})^{1/t} \to \prod_{s=1}^{M}(a^{(s)})^{\pi_s} < 1 \; p.s$$

Par le théorème de convergence dominée, on obtient que pour $s_0 \in \boldsymbol{S}$, $lim_{t\to+\infty}E[(a^{(S_t)}\ldots a^{(S_1)})^{1/t}|s_0] < 1$. On en déduit qu'il existe un entier naturel T et une constante $K < 1$ tels que $E[(a^{(S_T)}\ldots a^{(S_1)})^{1/T}|s_0] \leq K$ pour $s_0 \in \boldsymbol{S}$. Si on pose $V(s,y) = y^{1/T} + 1$, on a finalement montré qu'il existe des constantes $K < 1$ et $L < \infty$ tels que pour $(s_0, y_0) \in \boldsymbol{S} \times \boldsymbol{R}^+$ on a:

$$\Pi^T V(s_0, y_0) = E[V(S_T, Y_T) | s_0, y_0] \leq KV(s_0, y_0) + L$$

Afin de conclure, nous allons utiliser le fait que la chaîne de de Markov est fortement fellerienne, c'est à dire que si f est une fonction mesurable et bornée sur $\boldsymbol{X}$ alors Πf est continue. Ceci découle aisément du théorème de convergence dominée, et du lemme B.3 (cf annexe B). On utilise ensuite le même raisonnement que dans *Meyn et al.* (1993, [84]) (lemme 15.2.8). Posons $K' = K + \frac{1-K}{2}$ et $C = \{(s, y) \in \boldsymbol{S} \times \boldsymbol{R}^+ \,|\, K'V(s, y) \leq KV(s, y) + B\}$. On vérifie aisément que C est un ensemble compact de $\boldsymbol{S} \times \boldsymbol{R}^+$, donc petit puisque la chaîne est fortement fellerienne, et que l'inégalité suivante est vérifiée pour $(s_0, y_0) \in \boldsymbol{S} \times \boldsymbol{R}^+$

$$\Pi^T V(s_0, y_0) \leq K'V(s_0, y_0) + L\mathbf{1}_C(s_0, y_0)$$

Par ailleurs, puisque $K < 1$, on a $K' < 1$, et on a donc montré une inégalité de la forme **(D2),** ce qui achève la démonstration.

□

Nous verrons au paragraphe 1.d.2 que l'étude des propriétés asymptotiques des estimateurs du maximum de vraisemblance nécessite de faire des hypothèses sur l'existence de certains moments de la loi stationnaire. La proposition 1.8 ci-dessous donne des conditions nécessaires pour que de tels moments existent.

Proposition 1.8 (Existence de moments pour la loi stationnaire d'un modèle MS – γAR d'ordre 1)

Si la condition ***(A3)*** *est vérifiée avec* $c \geq 1$, *alors la loi stationnaire possède un moment d'ordre* c.

Preuve

Pour démontrer ce résultat nous allons vérifier que, sous les hypothèses de la proposition 1.8, le noyau de transition vérifie une condition de dérive de la forme **(D2)** avec $V(x) = x^c$. Par définition du modèle $MS - \gamma AR$, on a

$$E[(Y_t)^c \,|\, Y_0^{t-1}, S_0^t] = \gamma_c(a^{(S_t)} Y_{t-1} + b^{(S_t)}, \sigma^{(S_t)})$$

avec $\gamma_c(\mu, \sigma)$ le moment d'ordre c d'une loi gamma de moyenne μ et variance σ^2. D'après le lemme B.1, on a:

$$\gamma_c(\mu, \sigma) = \mu^c \left(1 - \frac{e\sigma^2}{2\mu^2} + o_\infty\left(\frac{1}{\mu^2}\right)\right)$$

On en déduit alors qu'il existe des constantes finies D_1 et D_2 telles que

$$E[(Y_t)^c \,|\, Y_0^{t-1}, S_0^t] \leq (a^{(S_t)} Y_{t-1})^c + D_1 (Y_{t-1})^{c-1} + D_2$$

et des constantes finies D_3 et D_4 telles que

$$E[(Y_t)^{c-1} \Big| Y_0^{t-1}, S_0^t] \leq D_3 (Y_{t-1})^{c-1} + D_4$$

En utilisant ces deux relations, on peut montrer, grâce à un raisonnement par récurrence, une majoration de la forme :

$$E[(Y_t)^c \Big| y_0, s_0] \leq E[(a^{(S_t)} \ldots a^{(S_1)})^c \Big| s_0] y_0^c + L_t y_0^{c-1} + M_t$$

avec L_t et M_t des constantes finies. Notons alors $f_{s_0, t} = E_{(s_0)}[(a^{(S_t)} \ldots a^{(S_1)})^c]$. On a

$$f_{s_0, t} = E_{(s)}[(a^{(S_t)} \ldots a^{(S_1)})^c] = \sum_{s_1, s_2, \ , \ldots, s_t} q_{s_0, s_1} (a^{(s_1)})^c q_{s_1, s_2} (a^{(s_2)})^c \ldots q_{s_{t-1}, s_t} (a^{(s_t)})^c$$

On peut vérifier alors que $(N_c)^t \mathbf{1} = f_t$ avec $f_t = (f_{1, t}, f_{2, t}, \ldots, f_{M, t})'$ et $\mathbf{1} = (1, 1, \ldots, 1)'$. En utilisant l'hypothèse **(A3)**, on montre alors qu'il existe un entier $T > 0$ tel que $K = sup\{f_{s, T} \big| s \in \boldsymbol{S}\} < 1$, ce qui implique que pour $s_0 \in \boldsymbol{S}$ et $y_0 \in \boldsymbol{R}^+$

$$E[(Y_T)^c \Big| s_0, y_0] \leq A y_0^c + L_T y_0^{c-1} + M_T \tag{1.9}$$

On conclut alors de la même manière que pour la proposition 1.7, en utilisant le fait que les ensembles compacts sont petits, ce qui permet une inégalité de la forme **(D2)** avec $V(x) = \|x\|^c$.

□

Les propositions 1.7 et 1.8 donnent des conditions garantissant l'ergodicité géométrique et l'existence de moments pour la loi stationnaire respectivement. Les hypothèses faites semblent raisonnables si on les compare à celles des paragraphes 1.b.1 et 1.b.2.

1.b.4 Stabilité des modèles *MS-γAR* d'ordre supérieur à 1

Dans ce paragraphe, nous proposons des critères garantissant la stabilité des modèles $MS - \gamma AR$ d'ordre $r > 1$. Comme dans le cas des modèles d'ordre 1, on peut vérifier aisément que la chaîne de Markov $\{X_t\} = \{S_t, \bar{Y}_t\}$ est irréductible et fortement fellerienne. Il reste ensuite à trouver des conditions sur les coefficients du modèle pour qu'une condition de dérive de la forme **(D2)** soit vérifiée. Posons $\|x\| = \sqrt{x'x}$ la norme Euclidienne sur $\boldsymbol{R}^r$. La proposition 1.9 est une généralisation immédiate de la proposition 1.7.

Proposition 1.9 (Stabilité du modèle MS – γAR)

Si la condition ***(A1')*** *est vérifiée* avec $a^{(s)} = \left\| A^{(s)} \right\|$ pour $s \in \boldsymbol{S}$ *alors le modèle* $MS - \gamma AR$ *correspondant est géométriquement ergodique.*

Preuve:

On utilise le même schéma de preuve que pour la proposition 1.7. Ecrivons tout d'abord que

$$\bar{Y}_t = A^{(S_t)}\bar{Y}_{t-1} + B^{(S_t)} + \zeta_t$$

avec $\zeta_t = \bar{Y}_t - A^{(S_t)}\bar{Y}_{t-1} - B^{(S_t)} = (\varepsilon_t, 0, \ldots, 0)'$

Par définition, on a $E[\varepsilon_t | \bar{Y}_{t-1}, S_t] = 0$ et $var(\varepsilon_t | Y_{t-1}, S_t) = (\sigma^{(S_t)})^2$. En particulier, on a :

$$E[\|\zeta_t\| \,|\, \bar{Y}_{t-1}, S_t] = E[|\varepsilon_t| \,|\, \bar{Y}_{t-1}, S_t] \leq \sigma^{(S_t)}$$

On en déduit que

$$E[\|\bar{Y}_t\| \,|\, Y_{t-1}, S_t] \leq \|A^{(S_t)}\| \|\bar{Y}_{t-1}\| + K$$

avec $K = max_{s \in S}(\|B^{(s)}\| + \sigma^{(s)})$. On raisonne ensuite de la même manière que pour le modèle d'ordre 1.

□

On vérifie aisément que les hypothèses de la proposition 1.9 impliquent la condition **(A1)** puisque

$$E_{\pi_0^{(S)}}\left[\frac{1}{t}\ln(\|A^{(S_t)}A^{(S_{t-1})}\ldots A^{(S_1)}\|)\right] \leq E_{\pi_0^{(S)}}[\ln(\|A^{(S_1)}\|)]$$

En particulier, les conditions de la proposition 1.1, garantissant la stabilité du modèle $MS-LAR$, sont plus faibles que celles de la proposition 1.9.

En s'inspirant des résultats de stabilité existants sur les modèles autorégressifs à seuil (cf *Tong* (1990, [115]), *Gueguan* (1994, [53]), on peut donner d'autres critères garantissant la stabilité des modèles $MS-\gamma AR$. En effet, la structure de ces modèles est relativement proche de celles des modèles $MS-AR$ puisqu'il s'agit de modèles autorégressifs à changements de régimes. La différence provient de la modélisation des changements de régimes : alors que dans les modèles $MS-AR$ ils sont indépendants de la manière dont évolue le processus observé, dans les modèles à seuil, ils dépendent uniquement des valeurs passées du processus observé.

Par exemple, on peut montrer aisément la proposition 1.10 ci-dessous, due à *Hili* (1992, [61]) dans le cadre des modèles autorégressifs à seuils. La preuve de ce résultat est similaire à celle donnée dans [61] et [53] pour les modèles autorégressifs à seuils.

Proposition 1.10 (Stabilité du modèle $MS-\gamma AR$)

Supposons que $max_{s \in S}\sum_i a_i^{(s)} < 1$, *alors le modèle* $MS-\gamma AR$ *est géométriquement ergodique.*

Notons que dans le cas de modèles d'ordre 1, cette condition est nettement plus forte que celle de la proposition 1.7. De plus, si les conditions de la proposition 1.10 sont satisfaites, alors les modèles autorégressifs correspondant aux différents régimes sont stables. Ces conditions pa-

raissent donc très restrictives et pourrait sans doute être améliorées.

Par ailleurs, la généralisation de la proposition 1.8, qui garantit l'existence de moments d'ordre e pour la loi stationnaire des modèles $MS-\gamma AR$ d'ordre $r = 1$, semble délicate. En effet, la preuve de la proposition 1.8 repose sur une majoration de $E[(Y_t)^c \,|\, \bar{Y}_{t-1}, S_t]$, ce qui semble difficile à établir dans le cas des modèles d'ordre $r > 1$.

1.c Calcul des estimateurs du maximum de vraisemblance

De nombreux algorithmes, permettant le calcul numérique des estimateurs du maximum de vraisemblance (noté **EMV** dans la suite) dans les modèles à variables cachées, ont été proposés dans la littérature. Nous avons choisi de nous principalement principalement sur 2 d'entre eux, à savoir les algorithmes EM et quasi-Newton. L'objectif de ce paragraphe est de décrire précisément comment nous avons utilisé ces deux algorithmes en pratique pour estimer les paramètres des différents modèles introduits aux chapitres 2 et 3.

Ce paragraphe est structuré de la manière suivante. Dans la partie 1.c.1 nous définissons plus précisément la fonction de vraisemblance pour les modèles $MS-AR$, puis nous donnons diverses expressions pour cette fonction ainsi que pour son gradient et son hessien. Ces formules seront utilisées à plusieurs reprises dans la suite de ce paragraphe, ainsi qu'au paragraphe 1.d lorsque nous nous intéresserons aux propriétés asymptotiques des EMV.

Dans la partie 1.c.2, nous décrivons le principe général de l'algorithme EM puis nous détaillons sa mise en place pratique de cet algorithme pour les modèles $MS-AR$ utilisés aux chapitres 2 et 3. Dans le paragraphe 1.c.3, nous décrivons rapidement la mise en place des algorithmes du type quasi-Newton dans les modèles $MS-AR$, puis dans le paragraphe 1.c.4 nous discutons les avantages et inconvénients respectifs de ces deux algorithmes. Cette discussion aboutit à un algorithme "hybride", dans lequel les deux algorithmes décrits précédemment sont utilisés successivement. Enfin, dans le paragraphe 1.c.5, nous décrivons brièvement les autres algorithmes d'estimation qui existent dans la littérature.

1.c.1 La fonction de vraisemblance dans les modèles *MS-AR*

Supposons tout d'abord que la chaîne de Markov $\{X_t\}$ est homogène et que le noyau de transition Π_θ possède une unique probabilité invariante. Soit $(y_{1-r}, \ldots, y_T)$ une observation du processus Y_{1-r}^T. La vraisemblance conditionnelle est définie, pour $\theta \in \Theta$, par :

$$L_T(\theta) = \bar{p}_\theta(y_1, \ldots, y_T | \bar{y}_0; \theta) = \sum_{s_0 \in S} p_\theta(y_1, \ldots, y_T | \bar{y}_0, s_0) \bar{p}_\theta(s_0 | \bar{y}_0)$$

avec $p_\theta(y_1, \ldots, y_T | \bar{y}_0, s_0)$ la densité de la loi de y_1^T sachant y_{-1+r}^0 et s_0.

En pratique, il est difficile de travailler directement avec cette vraisemblance, car elle fait intervenir la probabilité $\bar{p}_\theta(s_0 | \bar{y}_0)$ qui dépend de la loi stationnaire et que nous ne disposons pas d'expression analytique pour cette quantité. Il est alors usuel de remplacer cette probabilité par une probabilité arbitraire ς sur $\boldsymbol{S}$ et donc de travailler avec la vraisemblance conditionnelle mo-

difiée définie par

$$L_{T,\varsigma}(\theta) = p_{\theta,\varsigma}(y_1, \ldots, y_T | y_{-1+r}^0) = \sum_{s_0 \in S} p_\theta(y_1, \ldots, y_T | \bar{y}_0, s_0)\varsigma(s_0) \tag{1.10}$$

Nous appellerons alors abusivement un ***estimateur de maximum de vraisemblance*** (EMV) une valeur $\hat{\theta}_{T,\varsigma} \in \Theta$ qui maximise cette vraisemblance conditionnelle modifiée. D'après l'hypothèse de compacité faite sur l'espace des paramètres Θ, ce maximum existe dès que la fonction de vraisemblance est continue. Des conditions assurant la continuité de cette fonction sont données au paragraphe 1.d. Nous noterons aussi $l_{T,\varsigma}(\theta) = \ln(L_{T,\varsigma}(\theta))$ la log-vraisemblance conditionnelle modifiée. Plusieurs choix sont possibles pour ς, mais nous verrons au paragraphe 1.d que ce choix a peu d'influence sur la qualité de l'estimateur $\hat{\theta}_{T,\varsigma}$ si T est suffisamment grand, le choix de ς ne modifiant les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. Dans la suite de ce paragraphe, afin de simplifier les notations, nous choisirons $\varsigma = \mathbf{1}_{s_0}$ avec $s_0 \in \boldsymbol{S}$. La généralisation au cas où ς est quelconque est immédiate.

On peut aussi noter que la formule (1.10) permet de définir une fonction de vraisemblance dans le cas où la chaîne de Markov $\{X_t\}$ ne possède pas une unique probabilité invariante. On peut ainsi définir, par exemple, une fonction de vraisemblance pour les modèles $NHMS-AR$. Dans la suite du paragraphe, la chaîne de Markov cachée $\{S_t\}$ pourra être homogène ou non.

Il existe plusieurs expressions permettant de calculer la fonction de vraisemblance. On peut tout d'abord utiliser la formule (1.11) ci-dessous.

$$L_{T,\varsigma}(\theta) = \sum_{(s_1, \ldots, s_T) \in S^T} p_\theta(s_1, \ldots, s_T, y_1, \ldots, y_T | \bar{y}_0, s_0) \tag{1.11}$$

avec

$$p_\theta(s_1, \ldots, s_T, y_1, \ldots, y_T | \bar{y}_0, s_0) = \prod_{t=1}^{T} q_\theta^{(t)}(s_{t-1}, s_t) g_\theta(y_t | \bar{y}_{t-1}, s_t)$$

la ***vraisemblance des données complètes***. Le calcul de la vraisemblance avec cette expression fait intervenir la somme de M^T termes et il est alors préférable d'utiliser la formule suivante :

$$L_{T,\varsigma}(\theta) = \prod_{t=1}^{T} p_\theta(y_t | \bar{y}_0^{t-1}, s_0) \tag{1.12}$$

avec

$$p_\theta(y_t | \bar{y}_0^{t-1}, s_0) = \sum_{s \in S} p_\theta(y_t | \bar{y}_{t-1}, S_t = s) p_\theta(S_t = s | \bar{y}_0^{t-1}, s_0)$$

Les quantités $p_\theta(S_t = s | \bar{y}_0^{t-1}, s_0)$ sont appelées ***probabilités de prédiction***. Nous appellerons

filtre de prédiction la probabilité $p_\theta(S_t|\bar{y}_0^{t-1}, s_0)$. Nous verrons au paragraphe suivant que ces quantités vérifient une relation de récurrence simple, et peuvent être calculées en effectuant un nombre d'opérations de l'ordre de TM^2, ce qui permet de calculer rapidement la fonction de vraisemblance. Notons que la formule (1.12) se réécrit sous la forme (1.13) ci-dessous.

$$l_{T,\varsigma}(\theta) = \sum_{t=1}^{T} \ln p_\theta(y_t|\bar{y}_0^{t-1}, s_0) \tag{1.13}$$

Nous verrons au paragraphe 1.d que cette formule permet aussi de montrer que la fonction de log-vraisemblance vérifie un théorème de Shannon-McMillan-Breiman, étape nécessaire pour établir la consistance des EMV.

Afin de calculer le gradient $\nabla_\theta l_{T,\varsigma}(\theta)$ de la log-vraisemblance, une première méthode consiste à dériver l'équation (1.13). L'expression obtenue fait intervenir le gradient du filtre de prédiction $\nabla_\theta p_\theta(S_t|\bar{y}_0^{t-1}, s_0)$, et on peut montrer que cette quantité vérifie aussi une relation de récurrence simple (cf *Rynkiewicz* (2000, [101])). Notons que cette formule est aussi utilisée dans *Legland et al.* (2000, [72]) et *Douc et al.* (2001, [41]) pour montrer la normalité asymptotique des estimateurs dans les modèles *CMC*.

Il existe une autre formule permettant de calculer le gradient de la log-vraisemblance : il s'agit de ***l'identité de Fisher***. Cette relation repose sur le fait que

$$\ln p_\theta(y_1^T|\bar{y}_0, s_0) = \ln p_\theta(y_1^T, s_1^T|\bar{y}_0, s_0) - \ln p_\theta(s_1^T|\bar{y}_0^T, s_0) \tag{1.14}$$

En dérivant une première fois l'égalité (1.14) par rapport à θ, puis en intégrant cette nouvelle égalité par rapport à $p_\theta(s_1^T|y_1^T, s_0, \bar{y}_0)$, si cela est possible, on obtient alors l'identité de Fisher :

$$\nabla_\theta \ln p_\theta(y_1^T|\bar{y}_0, s_0) = \sum_{(s_1,\ldots,s_T)\in S^T} \nabla_\theta(\ln p_\theta(y_1^T, s_1^T|s_0, \bar{y}_0)) p_\theta(s_1^T|s_0, \bar{y}_0^T) \tag{1.15}$$

Pour obtenir cette identité, on utilisé le fait que

$$\sum_{(s_1,\ldots,s_T)\in S^T} \nabla_\theta(p_\theta(s_1^T|\bar{y}_0^t, s_0)) = 0$$

Des conditions assurant la validité de la formule (1.15) sont discutées plus précisément à la fin de ce paragraphe. Notons que cette cette formule se réécrit sous la forme

$$\begin{aligned}\nabla_\theta \ln p_\theta(y_1^T|\bar{y}_0, s_0) = &\sum_{(s_1,\ldots,s_T)\in S^T} \nabla_\theta(\ln q_\theta^{(t)}(s_{t-1}, s_t)) p_\theta(s_{t-1}, s_t|s_0, \bar{y}_0^T) \\ &+ \sum_{(s_1,\ldots,s_T)\in S^T} \nabla_\theta(\ln g_\theta(y_t|\bar{y}_{t-1}, s_t)) p_\theta(s_t|s_0, \bar{y}_0^T)\end{aligned}$$

Elle permet d'exprimer simplement le gradient de la log-vraisemblance en fonction des quanti-

tés $p_\theta(s_{t-1}, s_t | s_0, \bar{y}_0^T)$ pour $t \in \{1 \dots T\}$. Cette formule est utilisée au paragraphe 1.c.3 pour évaluer le gradient de la log-vraisemblance. Elle est aussi utilisée dans *Douc et al.* (2004, [42]) pour montrer que la fonction $(1/\sqrt{T})\nabla_\theta l_{T,\varsigma}(\theta)$ vérifie un T.C.L, première étape pour montrer la normalité asymptotique des EMV (cf 1.d.3).

En dérivant l'équation (1.14) une deuxième fois par rapport à θ, si c'est possible, puis en intégrant de nouveau par rapport à $p_\theta(s_1^T | \bar{y}_0^T, s_0)$, on obtient :

$$\begin{aligned}\nabla_\theta^2 \ln p_\theta(y_1^T | \bar{y}_0, s_0) = & \sum_{(s_1, \dots, s_T) \in S^T} \nabla_\theta^2 (\ln p_\theta(y_1^T, s_1^T | \bar{y}_0, s_0)) p_\theta(s_1^T | \bar{y}_0^T, s_0) \\ & - \sum_{(s_1, \dots, s_T) \in S^T} \nabla_\theta^2 (\ln p_\theta(s_1^T | \bar{y}_0^T, s_0)) p_\theta(s_1^T | \bar{y}_0^T, s_0)\end{aligned} \tag{1.16}$$

Notons $I_Y(\theta) = -\nabla_\theta^2 l_{T,s_0}(\theta)$ la matrice d'information observée, ainsi que $I_{S,Y}(\theta) = -\nabla_\theta^2 \ln p_\theta(y_1^T, s_1^T | s_0, \bar{y}_0)$ l'information observée pour les données complètes et $I_{S|Y}(\theta) = -\nabla_\theta^2 (\ln p_\theta(s_1^T | s_0, \bar{y}_0^T)$ l'information conditionnelle observée.

L'équation (1.16) se réécrit alors sous la forme (***formule de Louis*** ou ***principe de l'information manquante***)

$$I_Y(\theta) = E_\theta[I_{S,Y}(\theta) | \bar{y}_0^T, s_0] - E_\theta[I_{S|Y}(\theta) | \bar{y}_0^T, s_0]$$

Le dernier terme est généralement appelé "***information manquante***", et interprété comme la perte d'information due au fait que la chaîne $\{S_t\}$ n'est pas observée. Notons que, puisque $\sum_{(s_1, \dots, s_T) \in S^T} \nabla_\theta^2 (p_\theta(s_1^T | \bar{y}_0^T, s_0)) = 0$, on a:

$$\begin{aligned}E_\theta[I_{S|Y}(\theta) | \bar{y}_0^T, s_0] &= E[\nabla_\theta (\ln p_\theta(s_1^T, y_1^T | \bar{y}_0, s_0)) \nabla_\theta (\ln p_\theta(S_1^T, y_1^T | \bar{y}_0, s_0))' | \bar{y}_0^T, s_0] \\ &\quad - E[\nabla_\theta (\ln p_\theta(S_1^T, y_1^T | \bar{y}_0, s_0)) | \bar{y}_0^T, s_0] E[\nabla_\theta (\ln p_\theta(S_1^T, y_1^T | \bar{y}_0, s_0)) | \bar{y}_0^T, s_0]' \\ &= cov(\nabla_\theta (\ln p_\theta(S_1^T, y_1^T | \bar{y}_0, s_0)) | \bar{y}_0^T, s_0)\end{aligned} \tag{1.17}$$

On en déduit en particulier que l'information manquante est une matrice positive. En pratique, la formule de Louis peut, par exemple, être utilisée pour calculer la matrice d'information observée à partir de l'algorithme E.M (cf paragraphe 1.c.2) ou pour montrer que $(1/T)\nabla_\theta^2 l_{T,\varsigma}(\theta)$ vérifie une loi des grands nombres puis la normalité asymptotique des EMV (cf 1.d et *Douc et al.* (2004, [42])).

La validité des formules (1.15), (1.16) (1.17) repose uniquement sur le fait que les différents termes de l'équation (1.14) sont différentiables, et on peut vérifier que c'est le cas dès que la condition ci-dessous est satisfaite :

- Pour tout $s \in \boldsymbol{S}$, pour tout $(\bar{y}, y') \in \boldsymbol{Y}^r \times \boldsymbol{Y}$, les fonctions $\theta \to q_\theta^{(t)}(s, s')$ et $\theta \to g_\theta(y' | \bar{y}, s)$

sont de classe C^2

Dans le cas où $\boldsymbol{S}$ est infini, il faut pouvoir intervertir les signes $\sum$ (ou $\int$) et ∇_θ. Des conditions garantissant la validité des formules (1.14)-(1.16) dans ce cas peuvent être trouvées dans *Douc et al.* (2004, [42]).

1.c.2 Algorithme EM

Il s'agit d'un algorithme récursif très couramment utilisé pour estimer les paramètres d'un modèle à variables cachées. Il a été introduit initialement dans le cadre des modèles CMC par *Baum et al.* (1970, [10]), puis étendu par la suite aux modèles à variables cachées plus généraux par *Dempster et al.* (1977, [38]) et aux modèles $MS-AR$ par *Hamilton* (1990, [56]). Il s'agit d'un algorithme itératif partant d'une valeur initiale $\theta^{(0)} \in \Theta$ des paramètres. A chaque itération de l'algorithme, il y a deux étapes, à savoir l'étape E (Expectation) et l'étape M (Maximisation). Nous décrivons ci-dessous plus précisément ces deux étapes dans le cas des modèles $MS-AR$. $\theta^{(n)}$ désigne la valeur des paramètres après n itérations.

Etape E

Cette étape consiste à calculer la fonction intermédiaire $Q(\theta, \theta^{(n-1)})$ définie pour $(\theta, \theta^{(n-1)}) \in \Theta^2$ par :

$$Q(\theta, \theta^{(n-1)}) = E_{\theta^{(n-1)}}[\ln p_\theta(y_1^T, S_1^T | \bar{y}_0, s_0) | \bar{y}_0^T, s_0] \tag{1.18}$$

Cette fonction représente l'espérance de la log-vraisemblance complétée $\ln p_\theta(y_1^T, s_1^T | \bar{y}_0, s_0)$ sous la loi conditionnelle $p_{\theta^{(n-1)}}(s_1, \ldots, s_T | \bar{y}_0, s_0)$. La log-vraisemblance complétée est donnée par :

$$\ln p_\theta(y_1^T, s_1^T | \bar{y}_0, s_0) = \sum_{t=1}^{T} [\ln q_\theta^{(t)}(s_{t-1}, s_t) + \ln g_\theta(y_t | \bar{y}_{t-1}, s_t)]$$

Le calcul de la fonction $Q(\theta, \theta')$ fait donc intervenir uniquement les quantités $p_{\theta'}(s_{t-1}, s_t | y_{1-r}^T, s_0)$ et

$$p_{\theta'}(s_t | y_{1-r}^T, s_0) = \sum_{s \in S} p_{\theta'}(s, s_t | y_{1-r}^T, s_0)$$

pour $t \in \{1 \ldots T\}$. Comme dans le cas des modèles CMC, ces quantités peuvent être calculées en utilisant l'algorithme Forward-Backward introduit par *Chang et al.* (1966) [33]. Décrivons plus précisément cet algorithme dans le cas des modèles $MS-AR$. Notons, pour $t \in \{1 \ldots T\}$,

$$\alpha_t(s) = p_{\theta'}(S_t = s, y_1, \ldots, y_t | \bar{y}_0, s_0)$$
$$\beta_t(s) = p_{\theta'}(y_{t+1}, \ldots, y_T | S_t = s, \bar{y}_t)$$

On peut vérifier que

$$p_{\theta'}(s_{t-1}, s_t | y_{1-r}^T, s_0) = \frac{\alpha_{t-1}(s_{t-1}) q_{\theta'}^{(t)}(s_{t-1}, s_t) \beta_t(s_t)}{p_{\theta'}(y_1^T | \bar{y}_0)}$$

avec $p_{\theta'}(y_1^T | \bar{y}_0, s_0) = \sum_{s \in S} \alpha_T(s)$, et donc que la connaissance des quantités $(\alpha_t)_{t \in \{1 \dots T\}}$ et $(\beta_t)_{t \in \{1 \dots T\}}$ permet de calculer la fonction $Q(\theta, \theta')$. On peut montrer que ces quantités vérifient les formules de récurrence ci-dessous :

$$\alpha_0(s) = 1_{\{s_0\}}(s) \text{ et } \alpha_t(s) = g_{\theta'}(y_t | s, \bar{y}_{t-1}) \sum_{s \in S} q_{\theta'}^{(t)}(s_{t-1}, s) \alpha_t(s_{t-1})$$

$$\beta_T(s) = 1 \text{ et } \beta_t(s) = \sum_{s \in S} q_{\theta'}^{(t)}(s, s_{t+1}) g(y_{t+1} | \bar{y}_t, s_{t+1}) \beta_{t+1}(s_{t+1})$$

Lorsque la chaîne de Markov cachée prend un nombre fini de valeurs, les formules ci-dessus permettent d'exprimer simplement α_t et β_t, en fonction de α_{t-1} et β_{t+1}, respectivement à l'aide d'un produit matriciel. La première formule, qui permet de calculer récursivement la séquence $(\alpha_t)_{t \in \{1 \dots T\}}$, donne lieu à l'algorithme Forward et la deuxième formule, qui aboutit au calcul de $(\beta_t)_{t \in \{1 \dots T\}}$ à l'algorithme Backward. Chacun de ces algorithmes nécessite d'effectuer de l'ordre de TM^2 opérations.

En pratique, lorsque la longueur des séquences observées T est grande, on observe des différences d'ordre de grandeur importantes entre les coefficients α_1 et α_T, ce qui pose des problèmes numériques. En effet, notons $N_t = \sum_{s \in S} \alpha_t(s) = p_{\theta'}(y_1^t | \bar{y}_0, s_0)$. Nous verrons au paragraphe suivant, que sous des conditions assez générales, la log-vraisemblance vérifie une loi des grands nombres, c'est à dire que $\frac{1}{T} \ln(p_{\theta'}(y_1^t | \bar{y}_0, s_0))$ converge presque sûrement vers une fonction contraste $H(\theta, \theta')$ et que cette fonction est négative. On s'attend alors à ce que $N_t \approx \exp(tH(\theta, \theta'))$, et en pratique on observe bien une décroissance de α_t à une vitesse exponentielle. Il est alors usuel de travailler avec les quantités normalisées $\bar{\alpha}_t(s) = \frac{\alpha_t(s)}{N_t} = p(s_t | y_{1-r}^t)$ et $\bar{\beta}_t(s) = \frac{\beta_t(s)}{N_t}$. Des formules de récurrence simples pour les quantités $\bar{\alpha}_t$, $\bar{\beta}_t$ et N_t peuvent se déduire aisément de celle données ci-dessus pour α_t et β_t.

Dans le cas où la chaîne cachée est à valeurs continues, les formules précédentes restent valables en remplaçant les signes Σ par des $\int$. Cependant, le calcul analytique des quantités (α_t) et (β_t) est typiquement infaisable, sauf dans certains cas particuliers (cf *Kunsch* (2001, [70]). Dans les autres cas, on peut utiliser une approximation stochastique de $Q(\theta, \theta')$ (algorithme MCEM). Une telle approximation peut être obtenue en utilisant un algorithme MCMC. Pour ce-

la, N réalisations $(s_1^{(i)}, \ldots, s_T^{(i)})$ de la loi $P_{\theta'}(S_1, \ldots, S_T | y_0^T, s_0)$ sont simulées, puis une estimation de la fonction $Q(\theta, \theta')$ est obtenue en moyennant la log-vraisemblance complétée $\ln p_\theta(y_1^T, s_1^{(i)}, \ldots, s_T^{(i)} | \bar{y}_0, s_0)$ correspondant à ces N réalisations. Une description plus précise de cet algorithme peut être trouvée dans *Wei et al.* (1990, [120]), *Booth et al.* (2001, [17]) et *Douc et al.* (2004, [42]).

Etape M

Dans cette étape, la valeur courante des paramètres est réestimée par

$$\theta^{(n)} = \arg max_{\theta \in \Theta}(Q(\theta, \theta^{(n-1)}))$$

Dans certains modèles, tels que les modèles $MS-LAR$ avec innovations gaussiennes, on peut trouver une expression analytique pour $\theta^{(n)}$ (cf *Hamilton* (1990, [56])). Cependant, ce n'est pas le cas pour tous les modèles, comme par exemple pour les modèles $MS-\gamma AR$. On utilise alors généralement un algorithme d'optimisation numérique pour calculer $\theta^{(n)}$. Lorsque la nouvelle valeur des paramètres $\theta^{(n)}$ est telle que $Q(\theta^{(n)}, \theta^{(n-1)}) > Q(\theta^{(n-1)}, \theta^{(n-1)})$, on parle d'algorithme **GEM (Generalized EM)**.

La méthode usuelle consiste à utiliser quelques itérations d'un algorithme d'optimisation du type quasi-Newton pour calculer $\theta^{(n)}$. D'autres approches ont été proposées. On peut citer par exemple les algorithmes ECM dans lequel une succession de maximisations de problèmes plus simples est utilisée (cf *Meng et al.* (1995, [81])). Un exemple d'utilisation de l'algorithme ECM aux modèles $MS-LAR$ peut être trouvé dans *Saxton et al.* (1999, [103]). Pour les modèles que nous avons utilisés en pratique, les algorithmes quasi-Newton sont faciles à mettre en oeuvre et cette solution a été retenue.

Regardons plus précisément la mise en place de l'algorithme GEM dans le cas des modèles $MS-AR$ utilisés dans les chapitres 2 et 3. Dans ces différents modèles, on peut scinder les paramètres sous la forme $\theta = (\theta_S, \theta_R)$ avec $\theta_S \in \Theta_S$ les paramètres servant à décrire le noyau de transition de la chaîne cachée et $\theta_R \in \Theta_R$ ceux décrivant l'évolution du processus observé dans les différents régimes. D'après l'équation (1.18), on a alors

$$Q(\theta, \theta') = Q_S(\theta_S, \theta') + Q_R(\theta_R, \theta')$$

avec

- $Q_S(\theta_S, \theta') = \sum_{(s_1, \ldots, s_T) \in S^T} \ln(q_{\theta_S}^{(t)}(s_{t-1}, s_t)) p_{\theta'}(s_{t-1}, s_t | y_{1-r}^T, s_0)$
- $Q_R(\theta_R, \theta') = \sum_{(s_1, \ldots, s_T) \in S^T} \ln(g_{\theta_R}(y_t | \bar{y}_{t-1}, s_t)) p_{\theta'}(s_t | y_{1-r}^T, s_0)$

On est alors ramené à deux problèmes d'optimisation dans des espaces de plus petite dimension puisque pour calculer $\tilde{\theta} = \arg max_{\theta \in \Theta}(Q(\theta, \theta'))$, il suffit de chercher $\tilde{\theta}_S = \arg max_{\theta_S \in \Theta_S}(Q_S(\theta_S, \theta'))$ et $\tilde{\theta}_R = \arg max_{\theta_R \in \Theta_R}(Q_R(\theta_R, \theta'))$. En effet, on a alors

$\tilde{\theta} = (\tilde{\theta}_S, \tilde{\theta}_R)$.

- **Maximisation de $Q_S(\theta_S, \theta')$:**

Lorsque l'espace $\boldsymbol{S}$ est de cardinal fini et que le modèle est homogène on obtient

$$Q_S(\theta_S, \theta') = \sum_{s, s' \in \boldsymbol{S}} \ln(q_{\theta_S}(s, s')) \sum_{1 \le t \le T} p_{\theta'}(S_{t-1} = s, S_t = s' \,|\, y_{1-r}^T)$$

Si la matrice de transition est paramétrée par ses probabilités de transition c'est à dire si $\theta_S \in \Theta_S = \left\{ q(i,j) \ge 0 \,|\, \sum_{j=1}^{M} q(i,j) = 1 \right\}$, on peut vérifier par un calcul simple que $\tilde{\theta}_S = (\tilde{q}(i,j))_{i,j \in \boldsymbol{S}}$ est donné par

$$\tilde{q}_{i,j} = \frac{\sum_{1 \le t \le T} p_{\theta'}(S_{t-1} = i, S_t = j \,|\, y_{1-r}^T)}{\sum_{1 \le t \le T} \sum_{1 \le k \le M} p_{\theta'}(S_{t-1} = i, S_t = k \,|\, y_{1-r}^T)} \tag{1.19}$$

On peut remarquer que ces formules sont intuitives, et c'est sans doute une des raisons principales de l'algorithme EM. Par contre, pour les modèles non-homogènes utilisés au chapitre 2 ainsi que pour le modèle (homogène) introduit au chapitre 3 nous n'avons pas trouvé d'expression analytique pour $\tilde{\theta}_S$. Nous avons alors utilisé des algorithmes d'optimisation numérique du type quasi-Newton.

- **Maximisation de $Q_R(\theta_R, \theta^{(n-1)})$**

Plaçons-nous dans le cas où $\boldsymbol{S}$ est fini et supposons en outre, comme dans le cas des modèles $MS-\gamma AR$ utilisés au chapitre 2, que l'on puisse décomposer θ_R sous la forme $\theta_R = (\theta_R^{(1)}, \ldots, \theta_R^{(M)})$ avec $\theta_R^{(s)} \in \Theta_R^{(s)}$ les paramètres régissant l'évolution du processus observé dans le régime $s \in \boldsymbol{S}$. On a alors $Q_R(\theta_R, \theta') = \sum_{s \in \boldsymbol{S}} Q_R^{(s)}(\theta_R^{(s)}, \theta'))$ avec

$$Q_R^{(s)}(\theta_R^{(s)}, \theta') = \sum_{1 \le t \le T} \ln(g_{\theta_R^{(s)}}(y_t \,|\, \bar{y}_{t-1}, s)) p_{\theta'}(s_t \,|\, y_{1-r}^T)$$

On est alors amené à maximiser M fonctions dépendant d'un nombre restreint de paramètres. Dans le cas des modèles $MS-\gamma AR$, il n'existe pas d'expression analytique simple pour ces quantités. Nous avons alors utilisé un algorithme quasi-Newton. Malgré ces optimisations numériques, l'algorithme reste simple à mettre en oeuvre et rapide puisque dans les faits on est ramené à M problèmes d'optimisation sur des espaces de petite dimension ($r+2$ dans le cas des modèles $MS-\gamma AR$) et que les fonctions $Q_R^{(s)}(\,.\,, \theta')$, ainsi que leurs gradients, peuvent se calculer rapidement à partir des quantités $p_{\theta'}(s_{t-1}, s_t \,|\, y_{1-r}^T, s_0)$ calculées dans l'étape E.

Dans le cas du modèle $MS-LAR$ introduit au chapitre 3, l'espace des paramètres ne peut

être décomposé sous la forme $\theta_R = (\theta_R^{(1)}, \ldots, \theta_R^{(M)})$, puisque les mêmes paramètres servent à décrire l'évolution du processus observé dans les différents régimes, et nous avons alors utilisé directement un algorithme d'optimisation pour la fonction $Q_R(., \theta')$. Les temps de calculs peuvent alors devenir importants lorsque la dimension de l'espace des paramètres Θ_R est élevée.

1.c.3 Algorithme quasi-Newton

Une alternative naturelle à l'algorithme EM consiste à utiliser un algorithme d'optimisation du type quasi-Newton. Ces algorithmes nécessitent l'évaluation de la fonction $l_{T,\varsigma}(\theta)$ et de son gradient $\nabla_\theta(l_{T,\varsigma}(\theta))$ en un nombre de points qui peut être important. Pour que l'algorithme soit efficace, il faut alors pouvoir évaluer ces quantités rapidement.

La fonction $l_{T,\varsigma}(\theta)$ est généralement estimée à partir de la formule (1.13). Celle ci permet d'exprimer simplement $l_{T,\varsigma}(\theta)$ en fonction du filtre de prédiction $p_\theta(s_t | s_0, \bar{y}_0^{t-1})$ et ces quantités vérifient une relation de récurrence simple. En pratique, on retrouve l'algorithme Forward décrit dans le paragraphe 1.c.2, et la fonction $l_{T,\varsigma}(\theta)$ peut être estimée en utilisant de l'ordre de MT^2 opérations.

Pour calculer la fonction $\nabla_\theta l_{T,\varsigma}(\theta)$, une première approche consiste à utiliser à nouveau la formule (1.13): en la dérivant, on obtient une formule permettant de calculer le gradient $\nabla_\theta(l_{T,\varsigma}(\theta))$ à partir du filtre de prédiction et de son gradient $\nabla_\theta p_\theta(s_t | s_0, \bar{y}_0^{t-1})$, celui-ci pouvant être calculé récursivement. Cette approche est utilisée par exemple dans *Rynkiewicz* (2000, [101]). Une deuxième approche consiste à utiliser la formule de Fisher (1.14). Cette formule permet en effet d'exprimer $\nabla_\theta l_{T,\varsigma}(\theta)$ en fonction des quantités $p_\theta(s_{t-1}, s_t | \bar{y}_{1-r}^{T}, s_0)$. On peut alors calculer la quantité $\nabla_\theta l_{T,\varsigma}(\theta)$ de la même manière que la fonction auxiliaire $Q(\theta, \theta')$ de l'algorithme EM.

En fait les deux approches ont la même complexité et utilisent de l'ordre de $2MT^2$ pour calculer à la fois $l_{T,\varsigma}(\theta)$ et $\nabla_\theta l_{T,\varsigma}(\theta)$. En pratique nous avons choisi d'utiliser la deuxième approche.

1.c.4 Comparaison des algorithmes EM et quasi-Newton

Plusieurs raisons expliquent le succès de l'algorithme GEM. La première d'entre elles est sans doute sa simplicité de mise en oeuvre dans de nombreux modèles. Par ailleurs, cet algorithme possède de bonnes propriétés de stabilité numérique. Ceci provient en partie de la croissance de la log-vraisemblance à chaque itération. En effet, en utilisant l'inégalité de Jensen, on peut montrer que

$$l_{T,\varsigma}(\theta) - l_{T,\varsigma}(\theta') \geq Q(\theta, \theta') - Q(\theta', \theta') \tag{1.20}$$

On en déduit que $L_{T,\varsigma}(\theta^{(n)}) > L_{T,\varsigma}(\theta^{(n-1)})$ si $Q(\theta^{(n)}, \theta^{(n-1)}) > Q(\theta^{(n-1)}, \theta^{(n-1)})$, et cette dernière condition est vérifiée pour les algorithmes GEM. De plus la convergence de cet algorithme vers un extremum local de la fonction de log-vraisemblance a été établie dans le cas des CMC par *Baum et al.* (1970, [10]) puis pour les modèles à variables cachées sous des hypothè-

ses générales par *Wu* (1983, [123]).

Les algorithmes EM et GEM ont cependant certains inconvénients bien connus. Une première limitation de cet algorithme est sa vitesse de convergence parfois lente au voisinage des extrema locaux. Plus précisément, l'algorithme EM définit une transformation M de Θ dans Θ tel que $\theta^{(n+1)} = M(\theta^{(n)})$. Soit θ^* un point fixe de M tel que $\theta^{(n)} \to \theta^*$, on a alors

$$\theta^{(n+1)} - \theta^* \approx \nabla_\theta M(\theta^*)(\theta^{(n)} - \theta^*)$$

Il est par ailleurs montré dans *Dempster et al.* (1977, [38]) que $\nabla_\theta M(\theta^*) = E_\theta[I_{S|Y}(\theta^*)\big|\bar{y}_0^T, s_0]E_\theta[I_{S,Y}(\theta^*)\big|\bar{y}_0^T, s_0]^{-1}$. La vitesse de convergence asymptotique de l'algorithme EM est donc linéaire, avec un taux de convergence qui dépend du rapport entre les matrices d'information manquante et complète. Intuitivement, plus l'information manquante est importante, plus la vitesse de convergence asymptotique de l'algorithme *EM* est lente.

De plus, l'utilisation d'un algorithme GEM au lieu de l'algorithme EM peut encore ralentir la vitesse de convergence. On s'attend à ce que la vitesse de convergence dépende du nombre d'itérations de l'algorithme d'optimisation numérique utilisée dans l'étape M. Lorsque le nombre d'itérations est grand, la nouvelle valeur des paramètres $\theta^{(n)}$ va être proche d'un extremum de la fonction intermédiaire $Q(.\,, \theta^{(n-1)})$ et l'algorithme GEM va avoir un comportement proche de l'algorithme EM, mais de tels algorithmes vont être coûteux en temps de calcul. Numériquement, il semble que pour les premières itérations de l'algorithme GEM, lorsque $\theta^{(n)}$ est éloigné d'un point fixe de M, un faible nombre soit préférable : l'utilisation d'un grand nombre d'itérations modifie peu la valeur de $l_{T,\varsigma}(\theta^{(n+1)})$ et est coûteuse en temps de calcul. Par contre, lorsque $\theta^{(n)}$ se rapproche d'un point fixe, il est plus efficace d'utiliser un plus grand nombre d'itérations de l'algorithme quasi-Newton. En pratique, dans les premières boucles de l'algorithme EM, nous avons utilisé une seule itération de l'algorithme d'optimisation numérique, ce nombre d'itérations augmentant progressivement pour atteindre une dizaine d'itérations lorsque la valeur des paramètres se rapproche d'un point fixe de M. Des résultats théoriques sur la vitesse de convergence de l'algorithme ECM peuvent être trouvés dans *Meng* (1994, [82]) et *Saxton et al.* (1999, [103]). Par contre, à notre connaissance, il n'existe pas de résultat théorique sur la vitesse de convergence de l'algorithme GEM lorsqu'un algorithme du type quasi-Newton est utilisé dans l'étape M et ce problème pourrait être abordé ultérieurement.

La vitesse de convergence asymptotique des algorithmes du type quasi-Newton est super-linéaire, c'est à dire $\left\|\theta^{(n)} - \theta^*\right\| / \left\|\theta^{(n-1)} - \theta^*\right\| \to 0$ quand $n \to \infty$. Il est alors nettement plus efficace d'utiliser ce type d'algorithme au voisinage des extrema locaux. Par contre, l'algorithme EM semble plus efficace lorsque la valeur courante des paramètres est éloignée des maxima de la fonction de vraisemblance.

Une deuxième limitation des algorithmes EM est liée au fait que la fonction de vraisemblance est généralement multimodale. La séquence $\theta^{(n)}$ peut alors converger vers un extremum local "peu intéressant" de la fonction de log-vraisemblance, selon la valeur initiale du paramètre

$\theta^{(0)}$. Il est alors primordial de bien choisir cette valeur initiale. Notons que ce problème est général, et que les algorithmes du type quasi-Newton peuvent aussi converger vers des extrema locaux "peu intéressants". Plusieurs stratégies ont alors été proposées pour choisir $\theta^{(0)}$.

La première approche consiste à utiliser des "informations supplémentaires" afin d'obtenir une première estimation des paramètres. Par exemple, dans le cas des modèles $MS-\gamma AR$ du chapitre 2, la variable cachée représente le "type de temps". On pourra alors utiliser une classification en type de temps, effectuée au préalable afin de réaliser un premier découpage et obtenir ainsi une première estimation des paramètres. De tels découpages pourraient soit être effectués "à la main" en utilisant, par exemple, l'expertise d'un météorologue, ou des méthodes statistiques de classifications. Dans le cas du modèle spatio-temporel introduit au chapitre 3, la variable cachée a une interprétation physique bien précise (déplacement des masses d'air), et une première estimation des valeurs prises par cette variable est obtenue en utilisant les champs de vent sur une zone plus grande.

Lorsque nous ne disposons pas d'informations supplémentaires permettant d'obtenir une première estimation des paramètres, une alternative consiste à exécuter l'algorithme EM avec plusieurs valeurs initiales choisies dans un ensemble de valeurs "raisonnables" en fonction du phénomène observé. Plusieurs stratégies peuvent être utilisées pour choisir ces valeurs initiales, soit sur une grille fixée par le modélisateur, soit aléatoirement. Nous avons utilisé cette dernière solution pour les modèles du chapitre 2 et la manière dont sont choisis les paramètres initiaux est décrite dans ce chapitre. Une autre méthode consiste à utiliser une variante stochastique de l'algorithme EM. Une approximation stochastique de $Q(\theta, \theta^{(n-1)})$ remplace alors la valeur exacte de cette fonction, ce qui peut permettre d'éviter de converger vers des extrema locaux "peu intéressants". Différentes méthodes ont été proposées pour construire cette approximation, chacune de ces méthodes aboutissant à un algorithme différents (SEM, SAEM et MCEM par exemple). Une comparaison expérimentale de ces algorithmes peut être trouvée dans *Celeux et al.* (1995, [32]).

Enfin, notons que l'algorithme EM, au contraire des algorithmes quasi-Newton, ne permet pas de calculer directement la matrice d'information observée $I_Y(\theta) = -\nabla^2_\theta l_{T,s_0}(\theta)$. Or nous verrons au paragraphe 1.d que cette quantité est intéressante puisqu'elle permet d'estimer la variance des EMV. Différentes approches ont été proposées pour calculer une approximation de $I_Y(\theta)$ à partir de l'algorithme EM. Elles reposent toutes sur la formule (1.16) qui permet de lier l'information observée à l'observation des données complètes. Le premier terme de cette équation, à savoir $E_\theta[I_{S,Y}(\theta)|\bar{y}_0^T, s_0]$ est égal à $\nabla^2_{\theta'} Q(\theta', \theta)|_{\theta'=\theta}$ et peut se calculer de la même manière que la fonction $Q(\theta', \theta)$. Le deuxième terme, égal à $E_\theta[I_{S|Y}(\theta)|\bar{y}_0^T, s_0]$, est plus complexe. On y trouve, par exemple, des facteurs de la forme $p_\theta(s_t, s_{t'}|\bar{y}_0^T, s_0)\nabla_\theta g_\theta(y_t|\bar{y}_{t-1}, s_t)\nabla_\theta g_\theta(y_{t'}|\bar{y}_{t'-1}, s_{t'})$ avec $t, t' \in \{1 \dots T\}$ quelconques (cf (1.17)). Dans le cas des modèles CMC ou $MS-AR$, il faut alors calculer les quantités $p_\theta(s_t, s_{t'}|\bar{y}_0^T, s_0)$ pour $t, t' \in \{1 \dots T\}$ et le temps de calcul peut alors devenir important, notamment lorsque la longueur de la séquence d'apprentissage T est grande.

Plusieurs méthodes ont alors été proposées pour calculer une approximation de ce terme en un temps de calcul raisonnable. Une première méthode est décrite dans *Hugues* (1997, [64]). Cette méthode suppose que pour les produits impliquant des termes correspondant à des dates t et t' bien séparées peuvent être considérés indépendants. Une deuxième méthode peut être trouvée dans *Meng et al.* (1995, [81]). Elle repose sur le fait que si θ^* est un point fixe de M alors

$$E_\theta[I_{S|Y}(\theta^*)\big|\bar{y}_0^T, s_0] = \nabla_\theta M(\theta^*) E_\theta[I_{S,Y}(\theta^*)\big|\bar{y}_0^T, s_0]$$

Cette formule permet donc d'exprimer la matrice d'information manquante en fonction du gradient de M et de la matrice d'information complète. Il suffit alors de construire une approximation de $\nabla_\theta M(\theta^*)$ afin d'obtenir une valeur approchée de $E_\theta[I_{S|Y}(\theta^*)\big|\bar{y}_0^T, s_0]$. C'est ce qui est fait dans [81] par différence finie.

Algorithme utilisé en pratique

Finalement, afin de calculer les EMV dans les modèles $MS-AR$, nous avons utilisé un algorithme hybride. Dans un premier temps, afin de localiser un extremum "intéressant" de la fonction de log-vraisemblance, l'algorithme EM est utilisé avec plusieurs valeurs initiales choisies de manière aléatoire. Dans un deuxième temps, l'algorithme quasi-Newton est utilisé, ce qui permet d'accélérer la convergence de l'algorithme au voisinage du point fixe identifié par l'algorithme EM et fournit directement une valeur approchée de la matrice d'information observée. L'algorithme utilisé est plus précisément décrit ci-dessous:

- **Première étape:** localisation d'un maximum intéressant de la fonction de log-vraisemblance. Pour cela, nous avons choisi aléatoirement N_{init} valeurs initiales pour les paramètres dans un ensemble de valeurs "raisonnable" en fonction du phénomène physique considéré. Ensuite, N_1 itérations de l'algorithme EM sont utilisées pour chacune de ces valeurs initiales, puis le meilleur jeu de paramètres est sélectionné. Cette étape n'est pas nécessaire lorsqu'une première estimation des paramètres peut être obtenue par une autre manière, comme c'est le cas au chapitre 3 par exemple.
- **Deuxième étape:** estimation finale des paramètres. Pour cela, l'algorithme EM est tout d'abord à nouveau utilisé avec comme valeur initiale le jeu de paramètres obtenu à l'issu de l'étape précédente. On arrête l'algorithme EM lorsque

$$\left\{ \frac{l_{T,\varsigma}(\theta^{(n)}) - l_{T,\varsigma}(\theta^{(n-1)})}{l_{T,\varsigma}(\theta^{(n)}) + l_{T,\varsigma}(\theta^{(n-1)})} \leq \varepsilon \right\},$$

  puis l'algorithme quasi-Newton est finalement utilisé pour obtenir la valeur finale des paramètres. En pratique, nous avons choisi $\varepsilon = 10^{-4}$.

1.c.5 Autres méthodes d'estimation

Outre les méthodes mentionnées ci-dessus, à savoir l'algorithme EM et ses variantes (GEM,

ECM, MCEM,...) et les algorithmes du type quasi-Newton, une autre approche couramment utilisée consiste à estimer récursivement les paramètres. Il s'agit des algorithmes utilisant une estimation récursive des paramètres. Dans cette méthode, la valeur de $\theta^{(n)}$ est obtenue en utilisant uniquement les n premières observations. Chaque nouvelle observation est ensuite utilisée pour obtenir une nouvelle estimation des paramètres. Dans le cadre des modèles $MS-AR$, une description de cette méthode peut être trouvée dans *Holst et al.* (1994, [62]) et *Rynkiewicz* (2000, [101]). Une comparaison des résultats obtenus avec cet algorithme et les algorithmes EM et quasi-Newton figure dans [101]. Enfin notons que différents auteurs proposent d'utiliser une approche bayésienne (cf *Robert et al.* (1993, [98]), *Ephraim et al.* (2002, [46]) et *Kunsch* (2001, [70]) ainsi que les références citées dans ces papiers). Ces différentes méthodes n'ont pas été testées au cours de cette thèse.

1.d Propriétés asymptotiques des estimateurs du maximum de vraisemblance

Alors que de nombreux auteurs se sont intéressés au calcul numérique des EMV dans les modèles CMC et plus généralement $MS-AR$ (cf paragraphe 1.c), il existe nettement moins de littérature sur l'étude théorique du comportement asymptotique de ces estimateurs, et de nombreux résultats n'ont été démontrés que récemment.

Lorsque l'on veut montrer la consistance des EMV dans un modèle paramétrique quelconque, il est usuel de commencer par vérifier l'identifiabilité des paramètres du modèle. Ce problème est abordé au paragraphe 1.d.1, et nous montrons en particulier que les paramètres du modèle $MS-\gamma AR$ sont identifiables.

Les paragraphes 1.d.2 et 1.d.3 sont consacrés à la consistance et à la normalité des EMV, respectivement. La consistance des EMV a été établie simultanément par *Krishnamurthy et al.* et *Francq et al.* en 1998 ([68] et [49] respectivement). Nous montrons que les hypothèses faites dans [68] pour démontrer ce résultat sont vérifiées par le modèle $MS-\gamma AR$. La normalité asymptotique des EMV a été établie encore plus récemment par *Douc et al.* en 2004 ([42]). Les modèles considérés dans [42] sont très généraux, et notamment l'espace d'état de la chaîne de Markov cachée fini mais seulement compact. Les hypothèses faites par ces auteurs sont alors relativement fortes et en particulier ne sont pas vérifiées par le modèle utilisé $MS-\gamma AR$. Nous proposons alors des hypothèses plus faibles, qui sont vérifiées par le modèle $MS-\gamma AR$, et qui semblent suffisantes lorsque la chaîne de Markov cachée est à espace d'état fini (la démonstration de ce résultat n'est pas donnée dans cette thèse).

Enfin, dans le paragraphe 1.d.4 nous vérifions, à l'aide de simulations numériques, le comportement des EMV lorsque la séquence d'apprentissage est d'une taille comparable à celles disponibles en pratique pour les séries temporelles de vent. En particulier, nous regardons l'évolution du biais et de la variance des estimateurs en fonction du nombre d'années de mesure et nous vérifions que la variance empirique des estimateurs est proche de l'inverse de la matrice d'information observée.

Dans la suite de ce paragraphe, nous supposerons que le processus $\{Y_t\}$ suit un modèle $MS-AR$ de paramètre θ_0 et nous supposerons que θ_0 appartient à l'intérieur du compact Θ. Nous supposerons aussi, sauf mention contraire, que la condition **(K1)** ci-dessous est vérifiée. Des critères pratiques permettant de vérifier cette hypothèse sont donnés au paragraphe 1.b.

(K1) *Pour tout* $\theta \in \Theta$, *la matrice* Q_θ *est irréductible et le noyau* Π_θ *admet une unique probabilité invariante et la solution stationnaire est ergodique.*

1.d.1 Identifiabilité

Soit $\overline{P}_\theta^Y$ la trace de $\overline{P}_\theta$ sur $\boldsymbol{Y}^{\boldsymbol{N}}$. Nous dirons que le paramètre θ_0 est identifiable si la condition **(I)** ci-dessous est vérifiée.

(I) $\theta = \theta_0$ *si et seulement si* $\overline{P}_\theta^Y = \overline{P}_{\theta_0}^Y$

Cette condition est très générale et difficile à utiliser en pratique. Nous proposons dans cette partie de l'expliciter concrètement pour certains modèles $MS-AR$ particuliers.

Il peut exister différentes sources de non-identifiabilité dans les modèles $MS-AR$. Tout d'abord, il est clair que la loi stationnaire est invariante par permutation de la numérotation des états cachés. Plus précisément, dans la suite de ce paragraphe, nous allons supposer, comme c'est le cas dans les modèles $MS-\gamma AR$, que les paramètres peuvent se décomposer sous la forme $\theta = (\theta_S, \theta_R^{(1)}, \ldots, \theta_R^{(M)})$ avec θ_S les paramètres servant à décrire le noyau de transition de la chaîne de Markov cachée et $\theta_R^{(s)}$ les paramètres servant à décrire l'évolution du processus observé dans le régime $s \in \boldsymbol{S}$. Nous noterons alors $\sim$ la relation d'équivalence définie par

$$(\theta_{S,1}, \theta_{R,1}^{(1)}, \ldots, \theta_{R,1}^{(M)}) = \theta^{(1)} \sim \theta^{(2)} = (\theta_{S,2}, \theta_{R,2}^{(1)}, \ldots, \theta_{R,2}^{(M)})$$

si il existe une permutation $\tau \in S_n$ telle que $\theta_{R,1}^{(i)} = \theta_{R,2}^{(\tau(i))}$ et $q_{\theta_{S,1}}(i,j) = q_{\theta_{S,2}}(\tau(i), \tau(j))$ pour tout $i, j \in \boldsymbol{S}$. Il est clair que si $\theta_1 \sim \theta_2$ alors $\overline{P}_{\theta_1} = \overline{P}_{\theta_2}$ et $\overline{P}_{\theta_1}^Y = \overline{P}_{\theta_2}^Y$. Par ailleurs, lorsque les $\theta_R^{(i)}$ ne sont pas distincts deux à deux alors certains des paramètres du noyau de transition de la chaîne cachée peuvent ne pas être identifiables.

L'identifiabilité des modèles CMC a été étudiée par *Leroux* (1992, [73]). Il montre que l'identifiabilité de ces modèles est équivalente à celle des mélanges de produits de densité correspondants aux probabilités d'émission dans les différents régimes. Il utilise ensuite les résultats de *Teicher* (1967, [113]) qui démontre que si une famille de densités est identifiable, alors la famille des produits finis de ces densités est identifiable.

Ces résultats ont ensuite été étendus aux modèles $MS-LAR$ avec innovation gaussienne par *Francq et al.* (1998, [49]) et *Krishnamurthy et al.* (1998, [68]). Notons toutefois que la notion

d'identifiabilité utilisée dans ces deux articles est moins générale et légèrement différente de celle décrite ci-dessus. Par ailleurs, ces résultats ne s'appliquent pas directement au modèle introduit au chapitre 3. Nous discutons plus précisément l'identifiabilité de ce modèle au paragraphe 3.b.3.

Proposition 1.11 (identifiabilité des modèles $MS-\gamma AR$***)***

Soit $\{Y_t\}$ *un processus qui suit un modèle* $MS-\gamma AR$ *de paramètres* $\theta_0 = (\theta_{S,0}, \theta_{R,0}^{(1)}, \ldots, \theta_{R,0}^{(M)})$. *Supposons que la condition* ***(K1)*** *est vérifiée et que* $\theta_{R,0}^{(i)} \neq \theta_{R,0}^{(j)}$ *pour* $i, j \in \boldsymbol{S}$ *et* $i \neq j$. *Alors* $\bar{P}_\theta^Y = \bar{P}_{\theta_0}^Y$ *si et seulement si* $\theta_0 \sim \theta$.

Preuve

Commençons par démontrer le lemme suivant, qui est utilisé à plusieurs reprises dans la suite de ce chapitre.

Lemme 1.1

Soit $\{Y_t\}$ *un processus qui suit un modèle* $MS-\gamma AR$ *de paramètre* θ_0. *Si la condition* ***(K1)*** *est vérifiée alors* $\bar{p}_{\theta_0}[\bar{y}_0, S_1 = s_1] > 0$ *pour tout* $\bar{y}_0 \in (\boldsymbol{R}^{+*})^d$ *et* $s_1 \in \boldsymbol{S}$.

Preuve

Fixons $s_1 \in \boldsymbol{S}$ et $\bar{y}_0 = (y_0, \ldots, y_{-r+1}) \in (\boldsymbol{R}^{+*})^d$. La preuve du lemme repose sur la formule suivante :

$$\begin{aligned}
&\bar{p}_{\theta_0}[\bar{y}_0, S_1 = s_1] \\
&= \sum_{s_{-r+1} \in S} \int \bar{p}_{\theta_0}(\bar{y}_0, S_1 = s_1 | \bar{y}_{-r}, S_{-r+1} = s_{-r+1}) \bar{p}_{\theta_0}(\bar{y}_{-r}, S_{-r+1} = s_{-r+1}) d\bar{y}_{-r} \\
&= \sum_{s_{-r+1}, \ldots, s_0 \in S^r} \int \prod_{k=-r+1}^{0} q_\theta(s_k, s_{k+1}) p_\theta(y_k | \bar{y}_{k-1}, s_k) \bar{p}_{\theta_0}(\bar{y}_{-r}, S_{-r+1} = s_{-r+1}) d\bar{y}_{-r}
\end{aligned}$$

En utilisant l'irréductibilité de la chaîne de Markov cachée (cf **(K1)**), on peut construire un chemin $s_{-r+1}, \ldots, s_0$ tels que $q_\theta(s_k, s_{k+1}) > 0$ pour $k \in \{-r+1, 0\}$.

Soit alors $s_{-r+1}, \ldots, s_0$ un tel chemin et posons alors $q_- = min_{k \in \{-r+1, 0\}} q_\theta(s_k, s_{k+1})$ et $c_{K,k} = inf_{y_{k-r}^{-r} \in [0,K]^{1-k}} p_\theta(y_k | y_{-r+1}^{k-1}, y_{k-r}^{-r}, s_k)$ pour $K > 0$ et $k \in \{-r+1, \ldots, 0\}$. On a alors

$$\bar{p}_{\theta_0}(S_1 = s_1, \bar{y}_0) \geq (q_-)^r \left(\prod_{k=-r+1}^{0} c_{K,k} \right) \int_{]0,K]^r} \bar{p}_{\theta_0}(\bar{y}_{-r}, S_{-r+1} = s_{-r+1}) d\bar{y}_{-r}$$

En utilisant le fait que $y_k > 0$ pour $k \in \{-r+1, \ldots, 0\}$, on peut vérifier aisément que

$p_\theta(y_k | y_{-r+1}^{k-1}, y_{k-r}^{-r}, s_k) > 0$ pour $y_{k-r}^{-r} \in [0, K]^{1-k}$, et donc que que $c_{K,k} > 0 \ \forall K > 0$ par un argument de compacité. Par ailleurs, on a $\int_{(R^{+*})^r} \bar{p}_{\theta_0}(\bar{y}_{-r}, S_{-r+1} = s_{-r+1}) \cdot d\bar{y}_{-r} = \bar{P}_\theta(S_{-r+1} = s_{-r+1})$. Or, si la condition **(K1)** est vérifiée, cette dernière probabilité est strictement positive. Il existe donc $K > 0$ tel que $\int_{]0,K]^r} \bar{p}_{\theta_0}(\bar{y}_{-r}, S_{-r+1} = s_{-r+1}) \cdot d\bar{y}_{-r} > 0$. On conclut alors aisément.

□

Preuve de la proposition

Soit $\theta = (\theta_S, \theta_R^{(1)}, \ldots, \theta_R^{(M)}) \in \Theta$ avec $\theta_R^{(i)} = (\bar{a}^{(i)}, b^{(i)}, \sigma^{(i)})$. Il est évident que si $\theta_0 \sim \theta$ alors $\bar{P}_{\theta_0}^Y = \bar{P}_\theta^Y$. Examinons la réciproque.

Supposons que $\bar{P}_{\theta_0}^Y = \bar{P}_\theta^Y$. On a alors en particulier $\bar{p}_{\theta_0}(Y_1 | \bar{Y}_0) = \bar{p}_\theta(Y_1 | \bar{Y}_0)$, $\bar{P}_{\theta_0}^Y$ p.s, ce qui implique que

$$\sum_{i \in S} \bar{p}_{\theta_0}[S_1 = i | \bar{Y}_0] g_{\theta_0}(Y_1 | \bar{Y}_0, i) = \sum_{i \in S} \bar{p}_\theta[S_1 = i | \bar{Y}_0] g_\theta(Y_1 | \bar{Y}_0, i) \ \bar{P}_{\theta_0}^Y \text{ p.s} \qquad (1.21)$$

En utilisant l'identifiabilité des mélanges de loi gamma (cf *Teicher* (1963, [112])), on en déduit que

$$\sum_{i \in S} \bar{p}_{\theta_0}(S_1 = i | \bar{Y}_0) \delta_{\bar{a}_0^{(i)} \bar{Y}_0' + b_0^{(i)}, \sigma_0^{(i)}} = \sum_{i \in S} \bar{p}_\theta(S_1 = i | \bar{Y}_0) \delta_{\bar{a}^{(i)} \bar{Y}_0' + b^{(i)}, \sigma^{(i)}} \ \bar{P}_{\theta_0}^Y \ p.s$$

Or, d'après le lemme 1.1, on a $\bar{p}_{\theta_0}(S_1 = 1 | \bar{y}_0) > 0$ et $\bar{p}_{\theta_0}(\bar{y}_0) > 0$ pour $\bar{y}_0 \in (\boldsymbol{R}^{+*})^d$. On en déduit qu'il existe un ensemble $\Omega \subseteq (\boldsymbol{R}^{+*})^d$ tel que $\lambda(\Omega^c) = 0$ (où λ désigne la mesure de Lebesgue) et pour lequel

$$\forall \bar{y}_0 \in \Omega, \ \exists i(\bar{y}_0) \in \boldsymbol{S} \text{ tel que } \bar{a}_0^{(1)} \bar{y}_0' + b_0^{(1)} = \bar{a}^{(i(\bar{y}_0))} \bar{y}_0' + b^{(i(\bar{y}_0))} \text{ et } \sigma_0^{(1)} = \sigma^{(i(\bar{y}_0))}$$

Puisque $\lambda(\Omega^c) = 0$, on en déduit qu'il existe $i_1 \in \boldsymbol{S}$ tel que $\bar{a}_0^{(1)} = \bar{a}^{(i_1)}$, $b_0^{(1)} = b^{(i_1)}$ et $\sigma_0^{(1)} = \sigma^{(i_1)}$. On montre de même que pour $k \in \boldsymbol{S}$, il existe $i_k \in \boldsymbol{S}$ tel que $\theta_{R,0}^{(k)} = \theta_R^{(i_k)}$ avec $\theta_{R,0}^{(k)} = (\bar{a}_0^{(k)}, b_0^{(k)}, \sigma_0^{(k)})$ et $\theta_R^{(k)} = (\bar{a}^{(k)}, b^{(k)}, \sigma^{(k)})$. Si l'on suppose en outre que les n-uplets $(\theta_{R,0}^{(k)})_{k \in S}$ sont distincts 2 à 2, on en déduit alors qu'il existe une permutation $\tau \in S_n$ tel que $\theta_{R,1}^{(i)} = \theta_{R,2}^{(\tau(i))}$ pour $i \in \boldsymbol{S}$.

Il reste à vérifier que les coefficients de la matrice de transition sont identifiables. Pour cela, on utilise le fait que $\bar{p}_{\theta_0}(Y_1, Y_2 | \bar{Y}_0) = \bar{p}_{\theta}(Y_1, Y_2 | \bar{Y}_0)$ $\bar{P}^Y_{\theta_0}$ p.s, c'est à dire que

$$\begin{aligned} &\sum_{i,j \in S} \bar{p}_{\theta_0}(S_1 = i | \bar{Y}_0) q_{\theta_{S,0}}(i,j) g_{\theta_0}(Y_1 | \bar{Y}_0, i) g_{\theta_0}(Y_2 | \bar{Y}_1, j) \\ &= \sum_{i,j \in S} \bar{p}_{\theta}(S_1 = i | \bar{Y}_0) q_{\theta_S}(i,j) g_{\theta}(Y_1 | \bar{Y}_0, i) g_{\theta}(Y_2 | \bar{Y}_1, j) \end{aligned} \quad \bar{P}^Y_{\theta_0} \text{ p.s} \qquad (1.22)$$

En invoquant le résultat de *Teicher* (1967, [113] sur l'identifiabilité des mélanges de produits de densité, on en déduit que

$$\begin{aligned} &\sum_{i,j \in S} \bar{p}_{\theta_0}(S_1 = i | \bar{Y}_0) q_{\theta_{S,0}}(i,j) \delta_{\bar{a}_0^{(i)} \bar{Y}_0 + b_0^{(i)}, \sigma_0^{(i)}} \delta_{\bar{a}_0^{(j)} \bar{Y}_1 + b_0^{(j)}, \sigma_0^{(j)}} \\ &= \sum_{i,j \in S} \bar{p}_{\theta}(S_1 = i | \bar{Y}_0) q_{\theta_S}(i,j) \delta_{\bar{a}^{(i)} \bar{Y}_0 + b^{(i)}, \sigma^{(i)}} \delta_{\bar{a}^{(j)} \bar{Y}_1 + b^{(j)}, \sigma^{(j)}} \end{aligned} \quad \bar{P}^Y_{\theta_0} \, p.s$$

puis finalement que $q_{\theta_{S,0}}(i,j) = q_{\theta_S}(\tau(i), \tau(j))$.

□

Dans la preuve précédente, on a montré que si $\bar{P}^Y_{\theta_0}(y_{-r+1}, \ldots, y_1, y_2) = \bar{P}^Y_{\theta}(y_{-r+1}, \ldots, y_1, y_2)$ alors $\theta_0 \sim \theta$. Cela signifie que le paramètre θ_0 peut être caractérisé de manière unique par la loi de $(Y_{-r+1}, \ldots, Y_2)$. Remarquons aussi que la preuve ci dessus s'adapte facilement à d'autres modèles, tels que les modèles $MS-LAR$ avec innovations gaussiennes, par exemple.

1.d.2 Consistance

Nous dirons alors que l'estimateur du maximum de vraisemblance est consistant si $\hat{\theta}_{T,\varsigma}$ converge presque sûrement vers θ_0 pour toute mesure initiale ς lorsque $T \to \infty$. Plusieurs approches ont été développées pour montrer la consistance des EMV dans les modèles CMC et plus généralement $MS-AR$, cependant toute ces preuves reposent sur le même schéma de démonstration, à savoir :

- Prouver un théorème de Shannon-McMillan-Breiman, c'est à dire montrer la convergence de $(1/T) \ln p_{\theta}(Y_1^T | \bar{Y}_0, s_0)$ vers une fonction $H(\theta_0, \theta)$.
- Etablir que cette fonction limite vérifie $H(\theta_0, \theta_0) \geq H(\theta_0, \theta)$ avec égalité si et seulement si $\theta = \theta_0$.

Les premiers résultats de consistance pour les modèles CMC ont été démontrés par *Baum et al.* en 1966 ([9]) dans le cas ou $\boldsymbol{Y}$ et $\boldsymbol{S}$ sont finis. Ces résultats ont été étendus au cas où $\boldsymbol{Y}$ est quelconque mais $\boldsymbol{S}$ fini par *Leroux* (1992, [73]). Afin de montrer que la fonction de log-vraisemblance vérifie une loi des grands nombres, il utilise un théorème ergodique pour les processus sous-additifs démontré dans *Kingman* (1976, [67]). Cette approche a ensuite été généralisée

aux modèles $MS-AR$ par *Francq et al.* (1998, [49]), dans le cas des modèles $MS-NAR$, et *Krishnamurthy et al.* (1998, [68]) dans le cas des modèles $MS-AR$ généraux. Dans ces deux articles, $\boldsymbol{S}$ est supposé fini et il est montré que les différentes hypothèses énoncées s'appliquent aux modèles $MS-LAR$ avec innovations gaussiennes, et donc que les EMV sont consistants pour ce type de modèle. L'approche utilisée dans [68] est décrite plus précisément ci-dessous.

Dans le même temps, *Mevel* (1997, [83]) et *Legland et al.* (2000, [72]) ont utilisé une autre méthode, basée sur l'ergodicité géométrique d'une chaîne étendue incluant le filtre de prédiction, afin de montrer un théorème de Shannon-McMillan-Breiman puis la consistance des EMV dans le cas des modèles CMC lorsque $\boldsymbol{S}$ est fini et $\boldsymbol{Y}$ quelconque. Cette approche a ensuite été étendue pour montrer la consistance faible des EMV dans le cas où $\boldsymbol{S}$ est un espace compact par *Douc et al.* (2001, [41]) mais par contre n'a pas été étendue aux modèles $MS-AR$.

Enfin, plus récemment, la consistance des EMV a été établie dans le cadre des modèles $MS-AR$ généraux, lorsque $\boldsymbol{S}$ est un ensemble compact non nécessairement fini par *Douc et al.* (2004, [42]). La démonstration de ce résultat utilise les propriétés de la chaîne de Markov non homogène $\{S_t\}_{t\geq 0}$ conditionnellement aux observations $\{Y_t\}_{T\geq t\geq 0}$. Une condition de minoration uniforme des noyaux permet de montrer l'oubli exponentiel des conditions initiales de la chaîne $\{S_t\}_{t\geq 0}$ puis une loi des grands nombres pour la fonction de log-vraisemblance. Nous verrons au paragraphe 1.d.3 que cette méthode permet aussi de montrer la normalité asymptotique des EMV, ce que ne permet pas de faire l'approche utilisée dans *Krishnamurthy et al.* (1998, [68]). Il est par ailleurs mentionné dans *Douc et al.* (2004, [42]) que la généralisation des techniques basées sur les chaînes de Markov étendues utilisées dans *Mevel* (1997, [83]), *Legland et al.* (2000, [72]) et *Douc et al.* (2001, [41]) nécessiterait des hypothèses plus fortes. Nous revenons aussi plus précisément sur cette approche dans la suite de ce paragraphe.

Intéressons nous, dans un premier temps, plus précisément au théorème démontré dans *Krishnamurthy et al.* (1998, [68]). Introduisons tout d'abord les différentes hypothèses utilisées dans cet article.

(K2) $\forall s\in \boldsymbol{S}$ *et* $\forall\theta\in\Theta$, *il existe* $\eta>0$ *tel que* $\bar{E}_{\theta_0}[sup_{|\theta-\theta_0|<\eta}\left|\ln(g_\theta(Y_1|\bar{Y}_0,s))\right|]<\infty$

(K3) $\forall s\in \boldsymbol{S}$ *et* $\forall\theta\in\Theta$ *la fonction* $(\bar{y}_0,y_1)\rightarrow g_\theta(y_1|\bar{y}_0,s)$ *est continue sur* $\boldsymbol{Y}^r\times\boldsymbol{Y}$

(K4) $\forall s\in \boldsymbol{S}$ $\bar{P}_{\theta_0}(S_1=s|\bar{Y}_0)>0$, $\bar{P}_{\theta_0}$ *p.s.*

(K5) $\forall s\in \boldsymbol{S}$ *la fonction* $\bar{y}_0\rightarrow\bar{p}_{\theta_0}(S_1=s|\bar{y}_0)$ *est continue sur* $\boldsymbol{Y}^r$

(K6) $\forall\bar{y}_0\in\boldsymbol{Y}^r$, $\forall y_1\in\boldsymbol{Y}$ *et* $\forall s,s'\in\boldsymbol{S}$, *les fonctions* $\theta\rightarrow g_\theta(y_1|\bar{y}_0,s)$ *et* $\theta\rightarrow q_\theta(s,s')$ *sont continues sur* Θ

On vérifie aisément que la condition **(K6)** implique la continuité de la fonction de log-vrai-

semblance et donc que l'estimateur du maximum de vraisemblance est bien défini. Des hypothèses de continuité similaires sont classiquement faites pour établir la consistance des EMV, même dans le cas plus simple des variables i.i.d.

Les hypothèses **(K1)** et **(K2)** permettent de montrer que la fonction de log-vraisemblance vérifie une loi des grands nombres. La preuve de ce résultat est relativement simple. Elle repose sur le fait que $W_{m,n} = \ln(max_{s_t \in S}(p_\theta Y_{m+1}^n \big| Y_{m-r+1}^m, s_t))$ est un processus sous-additif. Il suffit alors d'appliquer le théorème ergodique pour les processus sous-additifs démontré dans *Kingman* (1976, [67]). La deuxième étape consiste à montrer que la fonction limite $H(\theta_0, \theta)$ vérifie les propriétés énoncées au début de ce paragraphe, ce qui s'avère plus difficile puisqu'on ne dispose pas de représentation simple de cette fonction. On aboutit alors à une condition d'identifiabilité moins explicite que celle donnée au paragraphe 1.d.1.

Plus précisément, notons $\xi_t = (\xi_{t,s})_{s \in S}$, avec $\xi_{t,s} = p_{\theta,\varsigma}(S_t = s \big| Y_{t-r}^{t-1})$, le filtre de prédiction. Dans [68], une mesure $\tilde{P}_{\theta_0,\theta}$ est construite de telle manière que le processus (ξ_t, Y_t) soit stationnaire sous $\tilde{P}_{\theta_0,\theta}$ et

$$H(\theta_0, \theta) = \tilde{E}_{\theta_0,\theta}\left[\ln\left(\sum_{s \in S} \xi_{1,s} p_{\theta, \pi_0^{(S)}}(Y_1 \big| \bar{Y}_0, S_1 = s)\right)\right]$$

On peut alors utiliser cette représentation de $H(\theta_0, \theta)$ pour montrer que $H(\theta_0, \theta_0) \geq H(\theta_0, \theta)$ et que si $H(\theta_0, \theta_0) = H(\theta_0, \theta)$, alors, $\tilde{P}_{\theta_0,\theta}$ p.s,

$$\sum_{s \in S} \bar{P}_{\theta_0}[S_1 = s \big| \bar{Y}_0] p_{\theta_0}(Y_1, Y_2 \big| \bar{Y}_0, S_1 = s) = \sum_{s \in S} \tilde{E}_{\theta_0,\theta}[\xi_{1,s} \big| \bar{Y}_0] p_\theta(Y_1, Y_2 \big| \bar{Y}_0, S_1 = s) \quad (1.23)$$

Notons alors $D(\theta_0) = \{\theta \in \Theta | H(\theta_0, \theta) = H(\theta_0, \theta_0)\}$ et $\rho(\theta, \theta_0) = inf_{\theta_0' \in D(\theta_0)}(\|\theta - \theta_0'\|)$. La proposition suivante est démontrée dans [68].

Proposition 1.12 : (consistance des EMV dans les modèles $MS-AR$***)***

Si les conditions ***(K1-K6)*** *sont vérifiées, alors pour toute mesure initiale* ς *sur* $\boldsymbol{S}$, $\rho(\hat{\theta}_{T,\varsigma}, \theta_0) \to 0$ $\bar{P}_{\theta_0}$ *p.s. lorsque* $T \to \infty$.

Afin de montrer la consistance des EMV dans un modèle donné, il faut alors vérifier les conditions **(K1-K6)** puis identifier l'ensemble $D(\theta_0)$. Le cas des modèles $MS-LAR$ avec innovations gaussiennes est traité dans [68]. Il y est montré que si la condition **(K1)** est vérifiée et si la loi stationnaire possède un moment d'ordre 2, alors les conditions **(K2-K6)** sont vérifiées. De plus, dans le cas où les paramètres $\theta_{R,0}^{(i)}$ régissant l'évolution du processus observé dans les différents régimes sont distincts 2 à 2, alors $H(\theta_0, \theta_0) = H(\theta_0, \theta)$ si et seulement si $\theta_0 \sim \theta$.

La proposition 1.13 ci-dessous traite le cas des modèles $MS-\gamma AR$.

Proposition 1.13 : (consistance des EMV dans les modèles $MS-\gamma AR$***)***

Supposons que le processus $\{Y_t\}$ *suit une modèle* $MS-\gamma AR$ *de paramètres* $\theta_0 = (\theta_{S,0}, \theta_{R,0}^{(1)}, \ldots, \theta_{R,0}^{(M)})$ *avec* $\theta_{R,0}^{(i)} \neq \theta_{R,0}^{(j)}$ pour $i \neq j$. *Supposons que la condition* ***(K1)*** *est vérifiée et que la loi stationnaire possède un moment d'ordre* $c > 2$. *Alors les conditions* ***(K2-K6)*** *sont vérifiées. Si de plus la condition* ***(I)*** *est vérifiée alors pour toute mesure initiale* ς *sur* $\boldsymbol{S}$, $\hat{\theta}_{T,\varsigma} \to \theta_0$ $\bar{P}_{\theta_0}$ *p.s. lorsque* $T \to \infty$.

Remarques :

1. Des conditions garantissant l'existence d'une solution stationnaire et ergodique possédant des moments d'ordre $c \geq 1$ pour les modèles $MS-\gamma AR$ sont discutées au paragraphe 1.b.
2. La condition **(I)** est discutée au paragraphe 1.d.1.

Preuve:

Commençons par vérifier que, sous les hypothèses de la proposition 1.13, les hypothèses (**K2-K6)** sont satisfaites. Les conditions de continuité **(K3)** et **(K6)** sont faciles à vérifier et la condition **(K4)** est une conséquence directe du lemme 1.1 Afin de vérifier **(K5)**, il suffit de montrer que la fonction $\bar{y}_0 \in (\boldsymbol{R}^{+*})^r \to \bar{p}_{\theta_0}(\bar{y}_0, S_1 = s_1)$ est continue. En effet, il en est alors de même pour les fonctions $\bar{y}_0 \in (\boldsymbol{R}^{+*})^r \to p_{\theta_0}(\bar{y}_0) = \sum_{s_1 \in S} p_{\theta_0}(\bar{y}_0, S_1 = s_1)$ et $\bar{y}_0 \in (\boldsymbol{R}^{+*})^r \to \bar{p}_{\theta_0}(S_1 = s_1 | \bar{y}_0) = \bar{p}_{\theta_0}(\bar{y}_0, S_1 = s_1)/p_{\theta_0}(\bar{y}_0)$. Or on a

$$\bar{p}_{\theta_0}[\bar{y}_0, S_1 = s_1]$$

$$= \sum_{s_{-r+1}^0 \in S^r} \int_{(\boldsymbol{R}^{+*})^r} \prod_{k=-r+1}^{0} q_\theta(s_k, s_{k+1}) p_\theta(y_k | y_{k-1}^{-r+1}, y_{-r}^{k-r+1}, s_k) \bar{p}_{\theta_0}(\bar{y}_{-r}, S_{-r+1} = s_{-r+1}) d\bar{y}_{-r}$$

En utilisant la condition **(K3),** on vérifie aisément que la fonction intégrée est continue, et on peut majorer ce terme uniformément en $\bar{y}_0$ par une constante sur les ensembles de la forme $[\varepsilon, +\infty[^r$ grâce au lemme B.4 On conclut en utilisant le théorème de convergence dominée.

La condition **(K2)** est plus technique à démontrer. Soit $s \in \boldsymbol{S}$ et $\theta \in \Theta$. Notons $\theta = (\theta_S, \theta_R^{(1)}, \ldots, \theta_R^{(M)})$ avec $\theta_R^{(1)} = (\bar{a}^{(i)}, b^{(i)}, \sigma^{(i)})$. On a

$$\begin{aligned}\ln(g_\theta(Y_1|\bar{Y}_0, s)) &= \ln(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)}) - 2\ln(\sigma^{(s)}) - 2\ln\Gamma\left(\frac{(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})^2}{(\sigma^{(s)})^2}\right) \\ &\quad - \frac{Y_1(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})}{(\sigma^{(s)})^2} + \left(\frac{(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})^2}{(\sigma^{(s)})^2} - 1\right)\ln\left(\frac{Y_1(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})}{(\sigma^{(s)})^2}\right)\end{aligned} \tag{1.24}$$

Nous allons montrer que ces différents termes sont intégrables par rapport à $\bar{P}_{\theta_0}$ lorsque $\bar{E}_{\theta_0}[(Y_t)^c] < \infty$ avec $c > 2$. Il est clair que si cette condition est vérifiée, alors $\bar{E}_{\theta_0}[\left|\ln(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})\right|] < \infty$ et $\bar{E}_{\theta_0}\left[\frac{Y_1(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})}{(\sigma^{(s)})^2}\right] < \infty$. En utilisant le fait que $\ln(\Gamma(x)) \sim_\infty x\ln x$, on vérifie de même que $\bar{E}_{\theta_0}\left[\ln\Gamma\left(\frac{(\bar{a}^{(s)}\bar{Y}_0 + b^{(s)})^2}{(\sigma^{(s)})^2}\right)\right] < \infty$. Il reste à vérifier l'intégrabilité du dernier terme de l'équation (1.24), qui découle facilement de celle de $\ln Y_1$, $Y_k \ln Y_1$ et $Y_k Y_l \ln Y_1$ pour $k, l \in \{-r+1, \ldots, 0\}$. Posons alors $l(\alpha, \beta) = E[|\ln X|]$ pour X une variable aléatoire suivant une loi gamma de paramètres $(\alpha, \beta) \in (\boldsymbol{R}^{+*})^2$. On a

$$\begin{aligned}\bar{E}_{\theta_0}[|\ln Y_1|] &= \bar{E}_{\theta_0}\left[\int_{\boldsymbol{R}^+} (|\ln y_1|\bar{p}_{\theta_0}(y_1|\bar{Y}_0, S_1))dy_1\right] \\ &= \bar{E}_{\theta_0}\left[l\left(\frac{(\bar{a}_0^{(S_1)}\bar{Y}_0 + b_0^{(S_1)})^2}{(\sigma^{(S_1)})^2}, \frac{(\sigma^{(S_1)})^2}{\bar{a}_0^{(S_1)}\bar{Y}_0 + b_0^{(S_1)}}\right)\right]\end{aligned} \tag{1.25}$$

On déduit alors du lemme B.2 qu'il existe des constantes A et B telles que pour $s \in \boldsymbol{S}$ et $\bar{y}_0 \in (\boldsymbol{R}^+)^r$ on ait

$$l\left(\frac{(\bar{a}_0^{(s)}\bar{y}_0 + b_0^{(s)})^2}{(\sigma^{(s)})^2}, \frac{(\sigma^{(s)})^2}{\bar{a}_0^{(s)}\bar{y}_0 + b_0^{(s)}}\right) \le A\left|\ln(\bar{a}_0^{(s)}\bar{y}_0 + b_0^{(s)})\right| + B \tag{1.26}$$

On en déduit que $\bar{E}_{\theta_0}[\ln Y_1] \le \bar{E}_{\theta_0}[A\left|\ln(\bar{a}_0^{(S_1)}\bar{Y}_0 + b_0^{(S_1)})\right| + B] < \infty$. On démontre de même, en utilisant l'inégalité (1.26), que $\bar{E}_{\theta_0}[Y_k \ln Y_1] < \infty$ et $\bar{E}_{\theta_0}[Y_k Y_l \ln Y_1] < \infty$ pour $k, l \in \{-r+1, \ldots, 0\}$.

Enfin, il reste à montrer que $H(\theta_0, \theta) = H(\theta_0, \theta_0)$ si et seulement si $\theta \sim \theta_0$. Ceci se fait à partir de l'équation (1.23) et en utilisant le même raisonnement que dans la preuve de la proposition 1.11.

□

Dans *Douc et al.* (2004, [42]), on peut trouver des conditions garantissant la consistance des EMV pour des modèles $MS-AR$ plus généraux. Ainsi, l'espace d'état $\boldsymbol{S}$ de la chaîne cachée n'est plus supposé fini, mais seulement compact. Pour démontrer ce résultat, les auteurs utilisent les hypothèses (**D1-D3)** ci-dessous. Nous avons choisi de remplacer les signes $\sum$ par des signes $\int$ pour insister sur le fait que ces conditions sont valables dans le cas ou $\boldsymbol{S}$ est continu.

(D1) *Pour tout* $\theta \in \Theta$, *le noyau de transition* Π_θ *est apériodique et positif au sens de Harris.*

(D2)

(a) $inf_{\theta \in \Theta} inf_{s, s' \in S} q_\theta(s, s') > 0$ *et* $sup_{\theta \in \Theta} sup_{s, s' \in S} q_\theta(s, s') < \infty$

(b) $\forall y_1 \in \boldsymbol{Y}, \forall \bar{y}_0 \in \boldsymbol{Y}^r,\ 0 < inf_{\theta \in \Theta} \int_S g_\theta(y_1 | \bar{y}_0, s)\mu(ds)$ *et* $sup_{\theta \in \Theta} \int_S g_\theta(y_1 | \bar{y}_0, s)\mu(ds) < \infty$

(D3)

(a) $b_+ = sup_{\theta \in \Theta} sup_{(y_0, \bar{y}_1, s) \in \boldsymbol{Y}^r \times \boldsymbol{Y} \times S} g_\theta(y_1 | \bar{y}_0, s) < \infty$

(b) $E_{\theta_0}[\left|\ln b_-(\bar{Y}_0, Y_1)\right|] < \infty$ *avec* $b_-(\bar{y}_0, y_1) = inf_{\theta \in \Theta} \int_S g_\theta(y_1 | \bar{y}_0, s)\mu(ds)$

La proposition suivante est démontrée dans [42].

Proposition 1.14 (consistance des EMV dans les modèles $MS-AR$***)***

Si les conditions ***(D1-D3)***, ***(I)*** *et* ***(K6)*** *sont vérifiées alors pour toute mesure initiale* ς *sur* $\boldsymbol{S}$ $\hat{\theta}_{T, \varsigma} \to \theta_0$ $\bar{P}_{\theta_0}$ *p.s.*

Une première différence notable avec la proposition 1.12, outre qu'elle s'applique à des modèles $MS-AR$ plus généraux, est que la méthode utilisée pour montrer que la fonction de log-vraisemblance vérifie une loi des grands nombres permet de construire une représentation explicite de la limite $H(\theta_0, \theta)$ et aboutit à des conditions d'identifiabilité plus facilement interprétables. Toutefois, comme nous l'avons vu dans le cas des modèles $MS-\gamma AR$, les conditions d'identifiabilité données dans *Krishnamurthy et al.* (1998, [68]) sont généralement suffisantes en pratique. Un autre avantage de cette méthode est qu'elle peut être étendue pour montrer la consistance des EMV dans le cas où le processus $\{Y_t\}$ n'est pas stationnaire, c'est à dire lorsque la loi de $X_0 = (S_0, Y_0)$ est différente de la probabilité invariante π_θ. Dans les exemples considérés dans les chapitres 2 et 3, le processus observé sera supposé stationnaire et cette extension n'est donc pas utilisée.

Cependant, pour les modèles utilisés aux chapitres suivants, les hypothèses **(D1-D3)** sont plus restrictives que les hypothèses **(K1-K6)**. Tout d'abord, l'hypothèse **(D1)** est plus forte que

l'hypothèse **(K1)**, et les propositions 1.1-1.4, qui traitent de la stabilité des modèles $MS-LAR$ et $MS-NAR$ lipschitzien, permettent uniquement de montrer la condition **(K1)**. Par contre les propositions 1.5-1.10, valables pour les modèles $MS-NAR$ sous-linéaire et $MS-\gamma AR$, permettent de montrer directement l'hypothèse **(D1)** et donc en particulier l'hypothèse **(K1)**.

Dans le cas où $\boldsymbol{S}$ est fini, l'hypothèse **(D2)(a)** est équivalente à ce que tous les coefficients de transition de la matrice Q_θ soient non nuls. Pour les applications envisagées, cette hypothèse ne semble pas trop restrictive. L'hypothèse **(D2)(b)** est vérifiée par de nombreux modèles comme les modèles $MS-NAR$ avec des innovations possédant une densité strictement positive sur $\boldsymbol{Y}$ et les modèles $MS-\gamma AR$.

On peut montrer que la condition **(D3)** est vérifiée par les modèles $MS-LAR$ avec innovations gaussiennes. Par contre, la condition **(D3)(a)** impose des contraintes relativement fortes sur les coefficients des modèles $MS-\gamma AR$. En effet, on peut vérifier que si $(\bar{a}^{(s)}\bar{y}_0 + b^{(s)})/\sigma^{(s)} < 1$ alors $\lim_{y_1 \to 0} g_\theta(y_1|s,\bar{y}_0) = +\infty$, et donc que la condition **(D3)(a)** n'est pas vérifiée lorsque il existe $s \in \boldsymbol{S}$ tel que $b_0^{(s)}/\sigma_0^{(s)} < 1$. Cependant, dans le cas où $\boldsymbol{S}$ est fini, la condition **(D3)'** ci-dessous semble suffisante:

(D3)'

(a) $E_{\theta_0}[\left|\ln b_+(\bar{Y}_0, Y_1)\right|] < \infty$ *avec* $b_+(\bar{y}_0, y_1) = sup_{\theta \in \Theta,\, s \in S} g_\theta(y_1|\bar{y}_0, s)$

(b) $E_{\theta_0}[\left|\ln b_-(\bar{Y}_0, Y_1)\right|] < \infty$ *avec* $b_-(\bar{y}_0, y_1) = inf_{\theta \in \Theta,\, s \in S} g_\theta(y_1|y_0, s)$

On peut vérifier que la condition **(D3)'** est vérifiée par les modèles $MS-LAR$ avec innovations gaussiennes et les modèles $MS-\gamma AR$ en s'inspirant de la preuve utilisée pour vérifier la condition **(K2)**. La preuve de la consistance des EMV sous les hypothèses **(D1), (D2), (D3'), (I)** et ***(K6)*** et lorsque $\boldsymbol{S}$ est fini n'est pas donnée ici. Elle nécessiterait de reprendre la démonstration de certains lemmes de *Douc et al.* (2004, [42]) et pourrait être faites ultérieurement.

1.d.3 Normalité asymptotique

Les premiers résultats de normalité asymptotique dans les modèles CMC ont été démontrés par *Baum et al.* (1966, [9]) dans le cas ou $\boldsymbol{Y}$ et $\boldsymbol{S}$ sont finis. Ces résultats ont ensuite été étendus par *Bickel et al.* (1998, [16]) dans le cas où $\boldsymbol{Y}$ est infini puis par *Jensen et al.* (1999, [65]) lorsque $\boldsymbol{Y}$ et $\boldsymbol{S}$ sont infinis. Ce dernier cas est aussi traité dans *Douc et al.* (2001, [41]) en utilisant l'ergodicité d'une chaîne de Markov étendue introduite dans *Legland et al.* (2000, [72]).

Dans le cadre des modèles $MS-AR$, la normalité asymptotique a été établie dans *Douc et al.* (2004, [42]) en étendant l'approche utilisée dans *Bickel et al.* (1998, [16]) et *Jensen et al.* (1999, [65]). La preuve de ce résultat, découle d'un théorème central limite pour $\nabla_\theta \ln p_\theta(y_1^T|\bar{y}_0, s_0)$ et d'une loi des grands nombres pour $\nabla^2_\theta \ln p_\theta(y_1^T|\bar{y}_0, s_0)$. Pour démontrer

ces deux théorèmes limites, les auteurs utilisent l'identité de Fisher (1.15) et la formule de Louis (1.16)-(1.17), respectivement, ainsi que l'oubli exponentiel des conditions initiales pour la chaîne de Markov non-homogène $\{S_t\}_{t\geq 0}$ sachant $\{Y_t\}_{T\geq t\geq 0}$ établie sous les hypothèses **(D1-D2)**.

Afin de montrer la normalité asymptotique des EMV, on va supposer qu'il existe $\eta > 0$ tel que les hyppothèses **(D4-D6)** ci-dessous soient vérifiées.

(D4) $\forall \bar{y}_0 \in \boldsymbol{Y}^r$, $\forall y_1 \in \boldsymbol{Y}$ *et* $\forall s, s' \in \boldsymbol{S}$, *les fonctions* $\theta \to g_\theta(y_1|\bar{y}_0, s)$ *et* $\theta \to q_\theta(s, s')$ *sont de classe* C^2 *sur la boule de centre* θ_0 *et de rayon* η.

(D5)

(a) $sup_{\theta\in\Theta} sup_{s,s'\in S} \left\|\nabla_\theta \log q_\theta(s, s')\right\| < \infty$ *et* $sup_{\theta\in\Theta} sup_{s,s'\in S} \left\|\nabla^2_\theta \log q_\theta(s, s')\right\| < \infty$

(b) $\bar{E}_{\theta_0}[sup_{|\theta-\theta_0|<\eta} sup_{s\in S} \left\|\nabla_\theta \ln g_\theta(Y_1|\bar{Y}_0, s)\right\|^2] < \infty$

(c) $\bar{E}_{\theta_0}[sup_{|\theta-\theta_0|<\eta} sup_{s\in S} \left\|\nabla^2_\theta \ln g_\theta(Y_1|\bar{Y}_0, s)\right\|] < \infty$

(D6)

(a) $\int_S sup_{|\theta-\theta_0|<\eta} g_\theta(y_1|\bar{y}_0, s)\mu(ds) < \infty$ *pour* $\nu^{\otimes r} \otimes \nu$ *presque tout* $\bar{y}_0, y_1 \in \boldsymbol{Y}^r \times \boldsymbol{Y}$

(b) $\int_Y sup_{|\theta-\theta_0|<\eta} \left\|\nabla_\theta g_\theta(y_1|\bar{y}_0, s)\right\| \nu(dy_1) < \infty$ *et* $\int_Y sup_{|\theta-\theta_0|<\eta} \left\|\nabla^2_\theta g_\theta(y_1|\bar{y}_0, s)\right\| \nu(dy_1) < \infty$

pour $\nu^{\otimes r} \otimes \mu$ *presque tout* $(\bar{y}_0, s) \in \boldsymbol{Y}^r \times \boldsymbol{S}$

La proposition suivante est démontrée dans *Douc et al.* (2004, [42]).

Proposition 1.15 (Normalité asymptotique des EMV dans les modèles $MS-AR$***)***

Si les conditions ***(D1-D6), (I) et (K6)*** *sont vérifiées alors, pour toute mesure initiale* ς *sur* $\boldsymbol{S}$,

$$T^{1/2}(\hat{\theta}_{T,\varsigma} - \theta_0) \to N(0, I(\theta_0)^{-1}) \ \bar{P}_{\theta_0}\text{-}faiblement.$$

De plus, $\forall s_0 \in \boldsymbol{S}$, $-T^{-1}\nabla^2_\theta \ln p_\theta(Y_1^T|\bar{Y}_0, s_0)\big|_{\theta=\hat{\theta}_{T,\varsigma}} \to I(\theta_0)$ $\bar{P}_{\theta_0}$*-p.s.*

La première partie de cette proposition montre que les EMV sont asymptotiquement gaussiens, et la deuxième partie donne un moyen pratique d'estimer la variance asymptotique des estimateurs à partir de la matrice d'observation observée. Dans la fin de ce paragraphe, nous discutons la validité des différentes hypothèses de la proposition 1.15 pour les modèles utilisés aux

chapitres 2 et 3.

On vérifie aisément que la condition **(D4)** est vérifiée par les modèles $MS-LAR$ et $MS-\gamma AR$. En ce qui concerne la condition **(D5)**, on peut montrer qu'elle est vérifiée par les modèles $MS-LAR$ avec innovations gaussiennes lorsque $E_{\theta_0}[|Y_0|^4]<\infty$ et par les modèles $MS-\gamma AR$ lorsque $E_{\theta_0}[|Y_0|^{\alpha}]<\infty$ avec $\alpha>4$. La preuve de ce résultat est relativement fastidieuse et utilise les même techniques que celles utilisées pour montrer que ces modèles vérifient la condition **(K2)**.

Enfin, la condition **(D6)** est difficile à vérifier en pratique puisqu'elle nécessite le calcul de $\nabla_\theta g_\theta(y'|\bar{y},s)$ et $\nabla_\theta g_\theta^2(y'|\bar{y},s)$. Cependant, cette hypothèse semble inutile lorsque $\boldsymbol{S}$ est fini. En effet, cette hypothèse sert à garantir la validité des formules (1.15), (1.16) et (1.17). Or, dans le cas où $\boldsymbol{S}$ est fini, ces formules sont valides dès que la condition **(D4)** est vérifiée (cf paragraphe 1.c.1). Par ailleurs, nous avons vu que la condition **(D3)** est vérifiée par les modèles $MS-LAR$ avec innovations gaussiennes mais pas par les modèles $MS-\gamma AR$. Là aussi, il semble que la condition **(D3')** proposée ci-dessus soit suffisante, et cette condition est vérifiée par les modèles $MS-\gamma AR$. La preuve de ces conjectures, à savoir la normalité asymptotique des EMV sous les conditions **(D1), (D2), (D3'), (D4), (D5), (I) et (K6)** dans le cas où $\boldsymbol{S}$ est fini, n'est pas donnée dans cette thèse mais pourrait l'être ultérieurement.

1.d.4 Etude des estimateurs du maximum de vraisemblance par simulation

Les résultats donnés ci-dessus permettent de vérifier que les estimateurs du maximum de vraisemblance ont de bonnes propriétés asymptotiques. Cependant, en pratique, il est aussi intéressant d'avoir une idée de la qualité des estimateurs lorsque la longueur de la séquence observée, T, est fixée. Cela permet en effet de donner une idée sur la quantité de données nécessaires pour avoir des estimateurs de qualité "raisonnable". Pour cela nous avons utilisé une méthode de Monte-Carlo. Plus précisément, nous avons réalisé l'expérience décrite ci-dessous:

- Choix de plusieurs valeurs de longueurs de séquences d'apprentissages, T, correspondants à des longueurs usuelles de mesures de séries temporelles de vent, à savoir 5, 10, 15, 20 et 50 ans. En pratique, nous supposerons qu'il y a 122 données dans chaque année et que les séquences correspondants aux différentes années sont indépendantes (cf paragraphe 2.a.1). En pratique, nous avons donc utilisé donc entre 5 et 50 séquences indépendantes de longueur 122 pour calibrer le modèle.
- Pour chacune de ces valeurs de T, nous avons alors simulé N_{sim} = 1000 réalisations du modèle choisi, puis les estimateurs du maximum de vraisemblance correspondant à ces N_{sim} séquences ont été calculées. En pratique, nous avons utilisé l'algorithme quasi-Newton décrit au paragraphe 1.c.3, avec pour valeur initiale la vraie valeur du paramètre. Ceci revient implicitement à supposer que l'EMV est dans un voisinage de θ_0, et permet d'éviter la recherche fastidieuse d'un extremum intéressant de la fonction de log-vraisemblance en utilisant des conditions initiales aléatoires et l'algorithme EM (cf 1.c.4). Pour chacune des séquences simulées, nous avons aussi calculé la matrice d'information observée.

- Les estimations correspondant à chacune de ces séquences ont ensuite été utilisées pour calculer le biais et la variance empiriques des estimateurs pour les différentes valeurs de T.

Les résultats donnés dans les tableaux 1.1-1.2 correspondent à un modèle $MS - \gamma AR$ d'ordre $r = 1$ et avec $M = 2$ régimes. Nous avons choisi pour θ_0 l'EMV correspondant à une série temporelle de vent. Les valeurs sont données dans la première colonne du tableau 1.1. On peut vérifier, en utilisant la proposition 1.7, que ce modèle vérifie la condition **(K1)** et que la loi stationnaire possède des moments d'ordre strictement supérieur à 4 avec la proposition 1.8 (par exemple, on peut vérifier numériquement que la condition **(A3)** est vérifiée avec $\alpha = 10$, et donc que la loi stationnaire possède un moment d'ordre 10), ce qui implique la consistance et la normalité asymptotique des EMV.

Le comportement global des estimateurs est satisfaisant. Ainsi, le biais et l'écart type des estimateurs sont relativement faibles et décroissent lorsque la taille des séquences d'apprentissage augmente. Par ailleurs, la matrice d'information observée permet généralement de fournir une bonne approximation de la variance des estimateurs. Nous avons aussi trouvé que les EMV sont approximativement gaussiens dès que T est suffisamment grand. Cela justifie la construction d'intervalles de confiance pour les estimateurs en utilisant les quantiles de la loi $N(0, 1)$.

	Vraie valeur	10*biais (5 ans)	10*biais (10 ans)	10*biais (15 ans)	10*biais (20 ans)	10*biais (50 ans
$q(1, 1)$	0.97	0.0451	0.0121	0.0066	0.0119	-0.0001
$q(2, 2)$	0.96	-0.0710	-0.0142	-0.0052	-0.0089	0.0046
$a^{(1)}$	0.84	-0.1062	-0.0623	-0.0326	-0.0308	-0.0141
$a^{(2)}$	0.78	-0.1260	-0.0346	-0.0321	-0.0249	-0.0043
$b^{(1)}$	1.06	0.6293	0.3806	0.1952	0.1967	0.0743
$b^{(2)}$	2.10	1.2381	0.3392	0.2585	0.1907	0.0123
$\sigma^{(1)}$	1.23	0.1334	0.0776	0.0479	0.0120	0.0168
$\sigma^{(2)}$	2.30	-0.1301	-0.1170	-0.0765	-0.0643	-0.0672

Tableau 1.1 Evolution du biais des estimateurs en fonction du nombre de données disponibles

Une étude plus précise de ces tableaux montre que les paramètres associés au deuxième régime sont globalement moins bien estimés que ceux correspondant au premier régime. Deux facteurs peuvent expliquer ce comportement. Tout d'abord, l'écart type de la loi conditionnelle

$\sigma^{(2)}$ est plus grand que $\sigma^{(1)}$ et pour un modèle AR usuel, on sait que la variance des estimateurs est proportionnelle à la variance du résidu. Par ailleurs, on a $q(2, 2) < q(1, 1)$, et donc la chaîne cachée passe en moyenne plus de temps dans l'état 1 que dans l'état 2. Pour étudier l'influence de ces différents facteurs, il faudrait réaliser des simulations avec d'autres valeurs θ_0.

	Vraie valeur	5 ans	10 ans	15 ans	20 ans	50 ans
$q(1, 1)$	0.96	0.072 (0.076)	0.041 (0.042)	0.031 (0.034)	0.028 (0.030)	0.017 (0.017)
$q(2, 2)$	0.97	0.136 (0.142)	0.050 (0.050)	0.037 (0.040)	0.033 (0.036)	0.020 (0.020)
$a^{(1)}$	0.84	0.101 (0.130)	0.072 (0.079)	0.051 (0.059)	0.049 (0.055)	0.033 (0.032)
$a^{(2)}$	0.78	0.131 (0.145)	0.082 (0.084)	0.068 (0.071)	0.058 (0.061)	0.037 (0.032)
$b^{(1)}$	1.06	0.779 (1.018)	0.516 (0.557)	0.375 (0.443)	0.337 (0.384)	0.201 (0.239)
$b^{(2)}$	2.10	1.317 (1.498)	0.797 (0.878)	0.599 (0.703)	0.499 (0.604)	0.304 (0.375)
$\sigma^{(1)}$	1.23	0.295 (0.289)	0.190 (0.178)	0.150 (0.149)	0.126 (0.125)	0.078 (0.076)
$\sigma^{(2)}$	2.30	0.491 (0.471)	0.319 (0.310)	0.249 (0.276)	0.218 (0.216)	0.136 (0.137)

Tableau 1.2 Evolution de l'écart type des estimateurs en fonction du nombre de données disponibles. La première valeur correspond à 10 fois l'écart type empirique des estimateurs et la deuxième valeur (entre parenthèse) à 10 fois celui obtenu à partir de la matrice d'information observée (valeur moyenne sur les différentes réalisations)

1.e Sélection de modèle et validation

Dans les paragraphes précédents, nous avons vu comment estimer les paramètres lorsque le modèle est spécifié, c'est à dire lorsque l'ordre des modèles autorégressifs r et le nombre de régimes M sont connus. Cependant, en pratique, lorsque l'on veut utiliser un modèle $MS-AR$ pour décrire un phénomène réel, ces quantités sont inconnues. Il faut alors trouver une méthode permettant de sélectionner le "meilleur modèle". Ce problème est abordé rapidement au paragraphe 1.e.1.

Une fois le "meilleur modèle" sélectionné, il reste ensuite à vérifier son adéquation aux séries temporelles observées. Ce problème est abordé au paragraphe 1.e.2. Après avoir rappelé les méthodes de validation généralement utilisées, nous proposons une méthode générale permettant de tester la capacité d'un modèle à simuler des séquences réalistes. Pour cela, une succession de

tests, utilisant des méthodes de Monte-Carlo pour estimer les lois des statistiques de test, est effectuée.

1.e.1 Sélection de modèle

De nombreux auteurs se sont intéressés récemment au problème de la sélection du nombre de régimes dans les modèles à variables cachées. Nous renvoyons à *Boucheron et al.* (2003, [19]), *Durand* (2003, [43]), *Ephraim et al.* (2002, [46]) pour un état des lieux récent sur cette problématique pour les modèles CMC et à *Krolzig* (1997, [69]) pour des résultats plus spécifiques aux modèles $MS-AR$. En fait, le problème est difficile à résoudre d'un point de vue théorique, et plusieurs auteurs ont alors proposé de tester la performance des différents critères existants sur des données synthétiques. De telles études expérimentales peuvent être trouvées par exemple dans *MacKay* (2002, [77]) et *Durand* (2003, [43]) pour les modèles CMC et dans *Psaradakis et al.* (2003, [94]) pour les modèles $MS-AR$. De ces différentes études, il ressort que le critère BIC a généralement un bon comportement. Ce critère a été proposé initialement par *Schwarz* en 1978 ([111]) puis utilisé dans de nombreux domaines par la suite. Il s'agit d'un critère de log-vraisemblance pénalisée, qui est défini par

$$BIC = -2l(\hat{\theta}) + n_{par}\ln(T) \tag{1.27}$$

avec l la fonction de log-vraisemblance, $\hat{\theta}$ l'estimateur du maximum de vraisemblance, n_{par} le nombre de paramètres du modèle et T la longueur de la séquence observée. Le modèle sélectionné sera alors celui pour lequel la valeur du critère BIC est la plus petite. Notons que le calcul de ce critère est immédiat lorsque la valeur du maximum de la fonction de log-vraisemblance est connue, et les algorithmes décrits au paragraphe 1.c permettent justement de calculer ce maximum. Ce critère est donc facile à utiliser en pratique.

L'étude du comportement théorique de ce critère reste un problème ouvert même dans le cas plus simple des modèles CMC. Afin de vérifier son efficacité sur les modèles spécifiques utilisés dans cette thèse, nous pourrions tester son efficacité sur des données synthétiques comme dans les articles mentionnés ci-dessus. Ces tests n'ont pas été réalisés, mais nous verrons dans le chapitre 2 que ce critère fonctionne correctement sur les données de vent considérées, puisqu'il permet généralement de sélectionner des modèles parcimonieux et qui "s'ajustent bien aux données".

La sélection de l'ordre r des modèles autorégressifs est nettement moins délicate d'un point de vue théorique. On pourra consulter par exemple *Krolzig* (1997, [69]), *Hamilton* (1993, [57]) et *Hamilton* (1996, [58]).

1.e.2 Validation de modèle

Il existe peu de critères permettant de vérifier le bon ajustement d'un modèle $MS-AR$ à des données réelles. Notons tout d'abord que le critère choisi peut dépendre de l'application envi-

sagée. Ainsi, en économétrie, les modèles $MS-AR$ sont généralement utilisés pour identifier puis expliquer les cycles économiques *Hamilton* (1989, [55]), *Krolzig* (1997, [69]). Une attention toute particulière est alors portée à l'interprétabilité du modèle, chaque régime devant correspondre à un état de l'économie. Dans le chapitre suivant, nous montrerons de même, pour chacun des modèles proposés, que les valeurs prises par la chaîne cachée possède une interprétation physique (type de temps). Toutefois, ce type de critère de validation reste purement qualitatif et est donc insuffisant.

Une méthode plus formelle, usuelle pour les modèles AR, consiste à calculer des résidus conditionnels, puis à vérifier qu'ils sont non-corrélés. Cependant, dans le cas des modèles $MS-AR$ l'utilisation de ce type de méthode n'est pas justifiée théoriquement (cf *Krolzig* (1997, [69])). On peut aussi regarder la variance empirique de ces résidus, ce qui permet d'évaluer la capacité du modèle à réaliser des prédictions à un pas de temps. Cette méthode est utilisée au paragraphe 3.c.1 pour valider le modèle spatio-temporel introduit au chapitre 3.

Finalement, la méthode la plus répandue consiste à comparer certaines statistiques calculées à partir des observations avec les quantités théoriques correspondantes du modèle à valider. En général, plusieurs critères sont considérés, tels que la fonction de répartition de la loi marginale et la fonction d'autocorrélation, par exemple. Notons que ce type de méthode n'est pas particulier au modèle $MS-AR$.

Il faut alors trouver une méthode permettant de décider si les écarts observés sont significatifs ou non. Pour certains modèles simples, la loi de la statistique considérée peut être calculée analytiquement, ce qui permet de construire des tests. Par exemple, dans le cas des variables iid on peut comparer les fonctions de répartition en utilisant un des nombreux tests d'ajustement existant (Kolmogorov-Smirnov, Cramer-Von Mises, chi-deux...). Une première limitation de ce type d'approche est que, généralement, seul le comportement asymptotique de la statistique de test est connu, et il est alors valable uniquement lorsque la taille de l'échantillon est "suffisamment" grande. Par ailleurs, lorsque le modèle devient plus complexe, la loi de la statistique considérée n'est plus calculable analytiquement. Dans ce dernier cas, la plupart des auteurs se contentent alors de comparer visuellement les quantités théoriques à leurs versions empiriques. Par exemple, afin de comparer deux fonctions de répartition, on peut tracer les quantiles théoriques en fonction des quantiles empiriques (qqplot). Dans le cadre des modèles CMC, cette méthode est plus précisément décrite dans *MacKay* (2002, [78]). Cependant, ce type de méthode ne permet pas de quantifier la qualité de l'ajustement. Nous proposons ici une méthode générale, dans laquelle des méthodes de Monte-Carlo sont utilisées afin d'estimer la loi des statistiques de test.

Afin de simplifier l'exposé, plaçons-nous tout d'abord dans le cas où le critère considéré est un scalaire, tel que l'espérance mathématique d'une composante de la loi marginale par exemple. Notons θ la valeur du paramètre correspondant au "vrai" processus physique et θ_0 la valeur du paramètre correspondant au modèle de simulation choisi. On veut alors tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Pour cela, on choisit une statistique de test, notée S. Dans

l'exemple ci-dessus on pourra choisir la moyenne empirique. On note p_v le ***niveau de signification*** de ce test, celui-ci dépendant de la forme de la région d'acceptation du test. Dans le cas d'un test bilatéral, pour lequel la région d'acceptation est choisie de la forme $[q_{\alpha/2}, q_{1-\alpha/2}]$, avec α le risque de première espèce et q_a le quantile d'ordre a de la loi de S sous H_0 pour $a \in]0,1[$, le niveau de signification du test est donné par :

$$p_v = 2\min(P_{H_0}(s^{\text{obs}} > S), P_{H_0}(s^{\text{obs}} < S)) \qquad (1.28)$$

avec s^{obs} la valeur de la statistique calculée à partir de la séquence observée. Cette valeur représente la plus petite valeur de α pour laquelle l'hypothèse H_0 serait rejetée avec un risque de première espèce α. Pour calculer cette quantité, il faut connaître la loi de S sous H_0. Or, lorsque le modèle ou la statistique considérée est complexe, ou lorsque la longueur de la série observée est trop courte pour pouvoir utiliser les résultats asymptotiques, cette distribution est inconnue. Il semble alors naturel d'estimer cette distribution en utilisant une méthode de Monte-Carlo. Pour cela, le modèle considéré est utilisé afin de simuler N séries temporelles de même longueur que la séquence observée, puis pour chacune de ces séquences la valeur de la statistique de test est calculée. Ces valeurs sont notées $(s_i^{\text{sim}})_{i \in \{1\ldots N\}}$. Elles sont ensuite utilisées pour évaluer le niveau de signification du test, grâce à l'estimateur empirique suivant :

$$\hat{p}_v = \frac{2}{N}\min(card\{i \in \{1\ldots N\} | s^{obs} > s_i^{\text{sim}}\}, card\{i \in \{1\ldots N\} | s^{obs} < s_i^{\text{sim}}\})$$

Dans la plupart des applications envisagées dans cette thèse, il est souhaitable que le modèle considéré permette de restaurer la loi marginale du processus. Il faut alors tester $H_0 : F = F_0$ contre $H_1 : F \neq F_0$ avec F la fonction de répartition de la loi marginale du processus observé et F_0 celle correspondant au modèle. Notons $\hat{F}(x)$ la fonction de répartition empirique et $F^{obs}(x)$ celle calculée à partir de la séquence observée. Différentes approches pourraient être envisagées pour réaliser un tel test. On peut tout d'abord utiliser une statistique de test scalaire en utilisant une des nombreuses distances existant dans la littérature pour comparer des fonctions de répartition. La plus populaire d'entres elles est sans doute la distance de Kolmogorov-Smirnov, mais il est bien connu que cette distance est plus sensible près du centre que des queues de la distribution. Pour cette raison, le test d'ajustement de Anderson-Darling est généralement préféré : la statistique de test correspondante donne plus de poids aux queues de la distribution. Cependant, il s'agit d'un critère intégré, comme la distance de Cramer-Von-Mises ou la distance du chi-deux, il peut masquer des écarts significatifs pouvant exister entre $F(x)$ et $F_0(x)$ pour certaines valeurs de x. Ici, une approche légèrement différente a été utilisée. Notons $p_v(x)$ le niveau de signification du test dont l'hypothèse nulle est $F(x) = F_0(x)$, c'est à dire

$$p_v(x) = 2\min(P_{H_0}(F^{obs}(x) > \hat{F}(x)), P_{H_0}(F^{obs}(x) < \hat{F}(x))) \qquad (1.29)$$

Nous avons ensuite utilisé comme statistique de test la statistique V définie ci-dessous :

$$V = min_{\{x|t_1 < F_0(x) < t_2\}} p_v(x)$$

avec $0 < t_1 < t_2 < 1$. Le minimum est pris sur l'intervalle $\{x|t_1 < F_0(x) < t_2\}$ parce que la variabilité de $\hat{F}(x)$ est nulle sous H_0 lorsque $F_0(x) = 0$ ou $F_0(x) = 1$. En pratique, nous avons choisi $t_1 = 1 - t_2 = 0.05$. La distribution de la statistique de test p_v est évaluée empiriquement sous H_0 à partir de séquences simulées puis cette distribution empirique est ensuite utilisée afin de construire la région de rejet du test, pour un risque de première espèce α fixé. Cette région est de la forme $[0, v_\alpha]$ avec $P_{H_0}[V < v_\alpha] = \alpha$. En pratique, dans les chapitres suivants, nous avons simulé 500 séquences afin d'estimer la distribution de $F^{obs}(x_i)$ sous H_0 en un nombre fini de point $x_1, \ldots x_P$, puis 500 séquences supplémentaires ont été simulées afin d'estimer la distribution de la statistique de test $min_{i \in \{1 \ldots P\}} p_v(x_i)$ puis la région de rejet.

Enfin, afin d'interpréter les défauts d'ajustement détectés par la méthode décrite ci-dessus, nous avons tracé à plusieurs reprises dans les chapitres suivants, des figures sur lesquelles sont représentées la fonction de répartition calculée à partir des observations, F^{obs}, et la fonction de répartition "théorique" correspondant au modèle F_0, celle-ci étant en pratique estimée en utilisant un grand nombre de séquences simulées. Afin de pouvoir comparer visuellement les quantités $F^{obs}(x)$ et $F_0(x)$ pour une valeur de x fixée, nous avons tracé sur ces figures des intervalles de fluctuation (ou "de pari") à 95%, notés $I(x)$. Ces intervalles sont tels que $P_{H_0}[F^{obs}(x) \in I(x)] = 0.95$. Ils ont été estimés par une méthode de Monte-Carlo. Pour cela, un grand nombre de réalisations du modèle, de la même taille que la série temporelle observée, ont été simulées, puis les valeurs de la statistique considérée ont été calculées pour chacune de ces réalisations, et enfin nous avons utilisé les quantiles empiriques à 2.5% et à 97.5% de ces quantités pour estimer respectivement les bornes inférieure et supérieure de l'intervalle $I(x)$. En pratique, nous avons simulé 100 séquences pour estimer ces quantités.

A titre d'exemple, nous proposons ici de discuter rapidement les résultats obtenus avec un modèle $MS - \gamma AR$ à deux régimes pour l'intensité du vent (cf paragraphe 2.b.4). Pour cet exemple, la valeur de la statistique de test est égale à $V^{obs} = 0$, et la région de rejet, pour un risque de première espèce de 5%, est égale à $[0, 0.0052]$. Le test rejette donc l'hypothèse $H_0 : F = F_0$, et conclut donc à un écart significatif entre la fonction de répartition observée et celle correspondant au modèle à valider. Sur la figure 1.4, nous avons représenté les quantités F^{obs}, F_0 ainsi que l'intervalle de fluctuation correspondant. On peut remarquer que la fonction de répartition observée est au dessus de la borne supérieure de l'intervalle de fluctuation correspondant au modèle pour les faibles valeurs de x, ce qui signifie que le modèle a tendance à simuler trop peu de faibles valeurs. La nullité de la statistique de test signifie qu'il existe un point x_i tel que la valeur de la fonction de répartition empirique observée en ce point, $F^{obs}(x_i)$, est supérieure ou inférieure à toutes les valeurs des fonctions de répartitions empiriques, correspondant aux séquences simulées, en ce point.

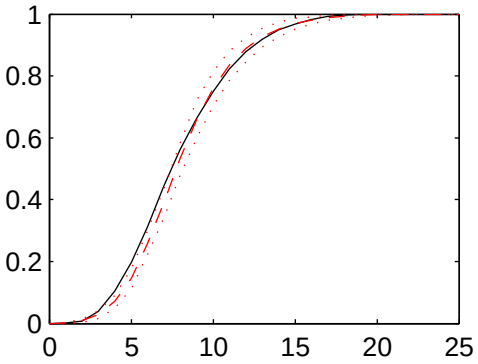


figure 1.4 : Exemple de comparaison d'une fonction de répartition observée (trait continu) avec celle d'un modèle $MS-\gamma AR$. La fonction de répartition "théorique" correspondant au modèle est représenté en tirets, et l'intervalle de fluctuation à 95% correspondant est représenté par les lignes pointillées.

On peut vérifier que si la fonction de répartition observée est à l'intérieur de l'intervalle de fluctuation à 95% pour toutes valeurs de x, alors l'hypothèse $H_0 : F = F_0$ est acceptée avec la méthode de test décrite ci-dessus et un risque de première espèce de 5%. Par contre la réciproque n'est pas vraie, comme le montre la région de rejet obtenue.

La même démarche peut être utilisée pour comparer d'autres statistiques multivariées ou fonctionnelles, telles que les fonctions d'autocovariance, par exemple.

La méthode décrite ci-dessus permet de quantifier l'adéquation de séquences observées à un modèle quelconque, pour peu qu'on sache en simuler des réalisations. On peut alors comparer différents modèles, éventuellement de natures différentes, entre eux. Ainsi, cette méthode est utilisée aux paragraphes 2.b.4, 2.c.1, 2.c.2 et 3.c.2 pour valider les différents modèles $MS-AR$ introduits dans la suite et comparer ces modèles aux autres modèles existants dans la littérature. Cette méthode, peut aussi être utilisée pour sélectionner le "meilleur" modèle $MS-AR$. Cependant, en pratique, cela serait relativement coûteux en temps de calcul puisqu'il faudrait alors simuler un grand nombre de réalisations de chacun des modèles en compétition. En pratique, nous avons alors utilisé le critère BIC défini au paragraphe précédent afin de réaliser une première sélection de modèle, puis la méthode de validation décrite ci-dessus a ensuite été utilisée pour comparer les meilleurs modèles au vu de ce critère (cf 2.b.4).

2. Modèles en un point fixe

Cette partie est consacrée à la modélisation des séries temporelles de vent en un point fixe.

Dans un premier paragraphe, nous faisons une synthèse bibliographique des différents modèles existant pour les séries temporelles de vent en un point fixe. Nous décrivons aussi ceux utilisés pour les processus H_s et Θ_m (hauteur significative et direction moyenne des vagues respectivement), la manière dont évoluent ces paramètres étant généralement proche de celle de U et Φ (intensité et direction du vent respectivement).

Les deux paragraphes suivants sont consacrés à l'utilisation des modèles $MS-AR$ pour les séries temporelles de vent. Dans le paragraphe 2.b, le modèle $MS-\gamma AR$ est utilisé pour modéliser l'intensité du vent, et dans le paragraphe 2.c nous décrivons deux extensions originales de ce modèle, dans lesquels la chaîne de Markov cachée n'est pas homogène.

Pour chacun des modèles proposés, nous justifions tout d'abord leur utilisation avec des arguments physiques. En particulier, pour justifier l'introduction de la variable cachée, nous mettons en évidence l'existence de "régimes météorologiques" (que nous appellerons "type de temps") dans les séries temporelles de vent. La construction de ces modèles s'est avérée particulièrement longue et les nombreuses tentatives infructueuses que nous avons faites ne sont pas rapportées ici. Ensuite, nous expliquons brièvement comment nous avons estimé les paramètres de ces modèles, puis nous vérifions que les différents régimes identifiés correspondent bien à des régimes météorologiques distincts et réalistes. Enfin, ces modèles sont validés en utilisant la méthode de test au paragraphe 1.e.2. Nous comparons aussi les résultats obtenus avec ceux correspondant aux modèles généralement utilisés pour ce type de séries temporelles.

Tous les résultats de cette partie ont été obtenus en utilisant les 22 ans de données de la base AES40 au point (46.25N , 1.667E) . Une carte montrant la position de ce point ainsi que des détails supplémentaires sur la base de données AES40 sont présentés dans l'annexe A.

2.a Principaux modèles existants

Avant de nous focaliser sur les modèles $MS-AR$, nous proposons dans ce paragraphe une synthèse bibliographique des différentes méthodes qui ont été proposées pour décrire et simuler les séries temporelles d'état de mer. Une attention toute particulière est portée à deux d'entre elles, à savoir une méthode utilisant des processus gaussiens transformés, et qui est la méthode "usuelle" dans ce domaine, et une méthode de bootstrap.

Dans une première partie nous nous intéressons aux différentes composantes non-stationnaires qui peuvent être présentes dans les séries temporelles de vent, à savoir les composantes interannuelle (tendance), saisonnière et journalière. Afin de prendre en compte ces non-stationnarités, la méthode la plus usuelle consiste à filtrer ces composantes afin se ramener à une série temporelle stationnaire. La méthode retenue dans la suite de ce chapitre est décrite et argumentée dans le paragraphe 2.a.1.

Ensuite, nous nous focalisons sur la modélisation de la série temporelle stationnaire résiduelle. Nous avons regroupé les modèles existants en trois grandes catégories. La première d'entre

elles regroupe les modèles basés sur l'hypothèse que le processus peut être considéré approximativement gaussien, quitte à lui appliquer une transformation préalable. Elle englobe en particulier la démarche de Box et Jenkins, qui est sans doute la méthode la plus couramment utilisée pour décrire et simuler des séries temporelles, et en particulier les séries temporelles de vent. Elle englobe aussi une autre méthode, non-paramétrique, couramment utilisée dans ce domaine d'application. Nous donnons une description relativement détaillée de cette deuxième méthode au paragraphe 2.a.2, puisque les résultats obtenus avec cette méthode seront comparés à ceux correspondant aux modèles $MS-AR$ dans les paragraphes suivants.

Dans la deuxième catégorie sont regroupés différents modèles paramétriques utilisant l'hypothèse que le processus observé est markovien. Nous y décrivons en particulier différents modèles autorégressifs non linéaires. Ces modèles sont décrits au paragraphe 2.a.3.

Enfin la dernière méthode décrite est une méthode de bootstrap, qui elle aussi suppose que le processus observé est markovien, mais utilise ensuite une estimation non paramétrique du noyau de transition. Certains résultats obtenus avec cette approche sont décrits au paragraphe 2.c.1.

2.a.1 Modélisation des composantes non-stationnaires

Selon les séries temporelles considérées, on peut identifier une ou plusieurs composantes non-stationnaires dans les données de vent et plus généralement d'état de mer, à savoir une tendance liée à l'évolution inter-annuelle du climat, des composantes annuelles induites par la saisonnalité des phénomènes météorologiques et enfin des composantes journalières créées par les différences de température existant entre le jour et la nuit. Chacune de ces composantes nécessite un traitement spécifique, et ce problème est discuté dans la suite de ce paragraphe.

Composante interannuelle

De nombreux météorologues se sont intéressés récemment à la possible existence de tendances dans les phénomènes météorologiques. Les mécanismes physiques pouvant expliquer ce genre de tendance sont diverses: phénomènes cycliques de quelques années tels que El Nino dans le pacifique ou la NAO en Atlantique du nord, évolutions à plus long terme, qui peuvent être "naturelles" ou liées à l'activité humaine... Ces tendances sont généralement difficiles à mettre en évidence d'un point de vue statistique à cause du faible nombre d'années de mesures disponibles par rapport à l'échelle temporelle de ces événements.

Toutefois, certains auteurs ont étudié la possible existence d'une tendance dans les séries temporelles d'état de mer. Ainsi, *Athanassoulis et al.* (1995, [8]) trouve une augmentation de 1 à 4 cm par an de la moyenne annuelle de la hauteur significative des vagues pour la période couvrant les années 1956 à 1975, et ceci en différents points de l'Atlantique nord. Pour obtenir ces résultats, une droite de régression est ajustée aux moyennes annuelles du processus. Les moyennes annuelles correspondant à nos données d'intensité du vent sont représentées sur la figure 2.1 et l'existence d'une telle tendance ne semble pas évidente. De plus, à notre connaissance, ces phénomènes sont encore relativement mal modélisés physiquement. Il nous semble

donc, dans l'état actuel des connaissances, difficile de les inclure de manière réaliste dans nos modèles. Dans la suite, nous avons alors choisi de négliger ces composantes.

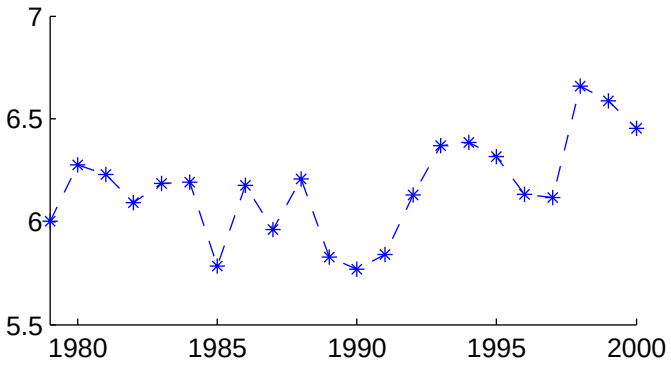


figure 2.1 : Moyenne annuelle de l'intensité du vent pour la période 1979-2000. La moyenne annuelle de l'intensité du vent, en ordonnée, est exprimée en ms^{-1}.

Saisonnalité

Par contre, les composantes saisonnières sont facilement observables sur les séries temporelles de vent considérées. Ainsi, lorsqu'on trace l'évolution de l'intensité du vent U au cours des 22 années pendant lesquels les données sont disponibles, on peut par exemple remarquer que les vents sont généralement plus forts en hiver qu'en été (cf figure 2.2).

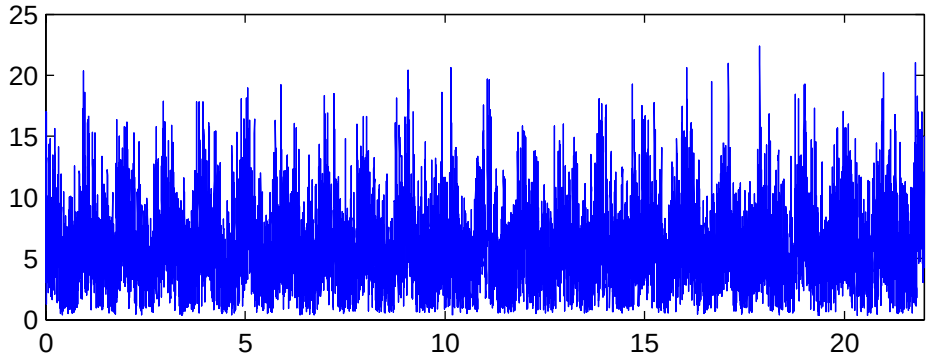


figure 2.2 : Evolution de l'intensité du au cours des 22 années. Le temps, en abscisse, est exprimé en années, et l'intensité du vent, en ordonnée, en ms^{-1}.

Cette saisonnalité peut aussi être observée sur la direction du vent Φ. Ainsi, la distribution marginale de ce processus dépend fortement de la saison (cf figure 2.3). Au mois de janvier, cette distribution est bimodale et il y a 2 directions de vent dominantes, à savoir les vents soufflant du sud-ouest et les vents soufflant du nord-est. Chacun de ces modes correspond à un type de temps (dépressionnaire et anticyclonique respectivement) souvent observé à cette période de l'année. Une description plus précise de ces types de temps est donnée au paragraphe 2.b.1. Part contre, au mois de juillet, on observe principalement des vents de secteur nord-ouest.

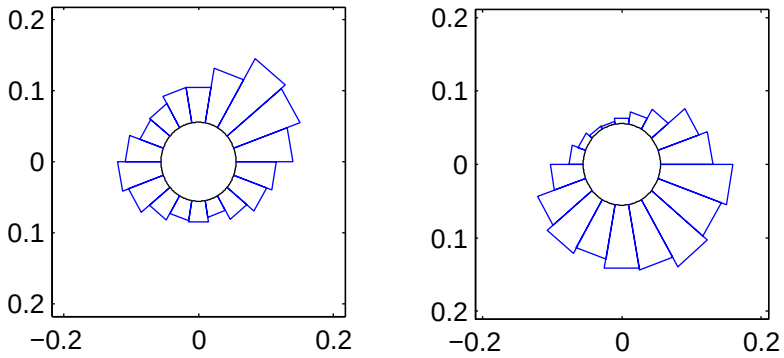


figure 2.3 : Densité de la loi marginale du processus Φ au mois de janvier (à gauche) et juillet (à droite).

Différents modèles ont été proposés pour enlever les composantes saisonnières dans les séries temporelles. Lorsque $\{Y_t\}$ désigne un processus d'état de mer, on trouve généralement la décomposition suivante (cf *Walton et al.* (1990, [119]), *Stephanakos* (1999, [109])):

$$Y_t = m(t) + \sigma(t) Y_t^{stat} \tag{2.1}$$

avec

- m et σ des fonctions (déterministes) périodiques de période un an. Ces fonctions modélisent respectivement la variation de la moyenne et de l'écart type de la loi marginale du processus au cours de l'année.
- $\{Y_t^{stat}\}$ un processus stationnaire.

Généralement, une transformation de Box-Cox est appliquée au préalable sur le processus $\{Y_t\}$ afin d'obtenir un processus dont la loi marginale est approximativement gaussienne. Cette transformation est plus précisément décrite au paragraphe 2.a.2.

Une alternative, couramment utilisée pour les séries temporelles de données météorologiques, consiste à supposer que le processus est stationnaire par morceaux. En effet, il est généralement admis que ce type de séries temporelles, peut être supposée stationnaire mois par mois. Pour chacun des 12 mois de l'année, on suppose alors que les observations collectées au cours des différentes années de mesure sont des réalisations indépendantes d'un même processus stationnaire. Lorsque le nombre d'années de mesure disponible est trop faible pour ajuster un modèle différent pour chaque mois, on peut supposer les processus stationnaires sur une plus longue période. Par exemple, dans *Brown et al.* (1984, [27]), les processus sont supposés stationnaires saison par saison. Selon les applications, on peut être soit intéressé par des statistiques mensuelles (planifications d'opérations offshore ou production d'électricité par une éolienne par exemple), auquel cas on peut utiliser directement les différents modèles pour évaluer les quantités d'intérêts pour les différents mois. Par contre, pour d'autres applications (transport sédimentaire, par exemple), il est nécessaire de simuler plusieurs années de données en continu. Il faut alors développer une méthode permettant de "recoller" les séquences simulées avec les

différents modèles, au début et à la fin de chaque mois. La méthode utilisée peut dépendre du modèle choisi pour décrire l'évolution du processus dans les différents mois. Une méthode générale, valable quelque soit le modèle choisi, est décrite dans *Borgman et al.* (1991, [18]).

L'avantage principal de la première méthode est que, une fois les fonctions m_Y et σ_Y estimées, on peut ajuster un modèle à la série résiduelle Y_{stat} avec relativement peu de données. Par exemple dans *Stephanakos* (1999, [109]), les fonctions m_Y et σ_Y sont estimées avec des données satellitaires, puis un modèle ARMA est ajusté à la série résiduelle Y_{stat} avec seulement 2 années de données de bouée. Au contraire, avec la deuxième méthode, un modèle différent étant ajusté pour chaque mois séparément, il faut généralement avoir une séquence d'apprentissage plus longue.

Cependant, les données utilisées dans cette thèse sont disponibles sur une période relativement longue (22 ans) et nous avons alors choisi d'utiliser la deuxième méthode. En effet, la première méthode est sans doute trop simple pour permettre de décrire de manière suffisamment précise les différences existant entre les mécanismes régissant l'évolution des paramètres considérés aux différentes saisons. Ce modèle permet de modéliser que la moyenne et la variance du processus dépendent de la saison, mais ne permet pas de prendre en compte, par exemple, que la fonction d'autocorrélation décroît plus vite vers 0 en été qu'en hiver ou que le nombre et la nature des types de temps (cf 2.b) peuvent dépendre de la saison. De plus, nous n'avons pas réussi à adapter de manière satisfaisante la première méthode au processus directionnel $\{\Phi_t\}$.

Pour chaque mois, nous disposons alors de 22 séries temporelles indépendantes de longueur 122 pour calibrer les modèles.

Composantes journalières

Les données d'états de mer, et notamment les données de vent, peuvent aussi être sujettes à des variations journalières. Pour les données que nous avons utilisées, ces composantes sont plus marquées en été et aux points proches de la côte. Par exemple, pour le point choisi, qui se trouve à environ 20 km de la côte, le tracé de la fonction d'autocorrélation et surtout du périodogramme permet clairement de mettre en évidence cette composante pour les mois de juillet (cf figure 2.4)

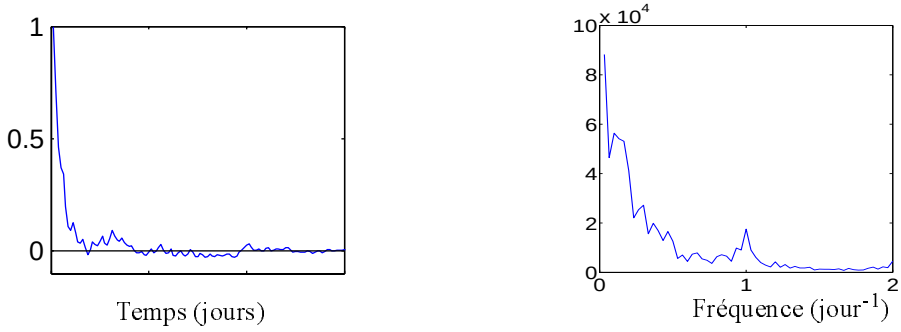


figure 2.4 : Fonction d'autocorrélation empirique (à gauche) et périodogramme (à droite) de l'intensité du vent pour le mois de juillet.

Les valeurs de la moyenne et de l'écart type empirique de loi marginale de U aux différents moments de la journée sont données dans le tableau 2.1 : les vents moyens les plus forts sont observés à minuit et les plus faibles à midi.

	0h	6h	12h	18h
Moyenne	5.03	4.85	4.29	4.93
Ecart type	1.94	1.93	2.38	2.14

Tableau 2.1 Moyenne et écart type de U aux différentes instants de la journée (mois de juillet)

Pour enlever ces composantes journalières, la même méthode que celle utilisée pour enlever les composantes saisonnières est généralement utilisée (cf *Brown et al.* (1984, [27]), *Daniel et al.* (1991, [37])). On suppose alors que

$$Y_t = m(t) + \sigma(t) Y_t^{stat} \tag{2.2}$$

avec

- m et σ des fonctions (déterministes) périodiques de période un jour.
- $\{Y_t^{stat}\}$ un processus stationnaire.

Les valeurs données dans le tableau 2.1 sont alors des estimateurs naturels des fonctions m et σ.

Nous décrivons une autre méthode, utilisant les modèles $MS-AR$, au chapitre 2.c. Dans ce modèle, la chaîne de Markov cachée est supposée non-homogène, la matrice de transition dépendant de l'heure de la journée.

2.a.2 Modèles basés sur les processus gaussiens

Les processus d'état de mer ne peuvent généralement pas être considérés gaussiens. Par exemple, la loi marginale du processus U est à support positif et généralement dissymétrique avec un skewness positif (cf figure 2.5). Cependant, il est généralement possible de transformer ces processus en des processus dont les lois marginales sont approximativement gaussiennes. Si on fait l'hypothèse supplémentaire que ce processus transformé est gaussien, on peut alors utiliser les nombreuses techniques existant pour ce type de processus (modèles paramétriques du type ARMA, méthodes de simulation exacte,...)

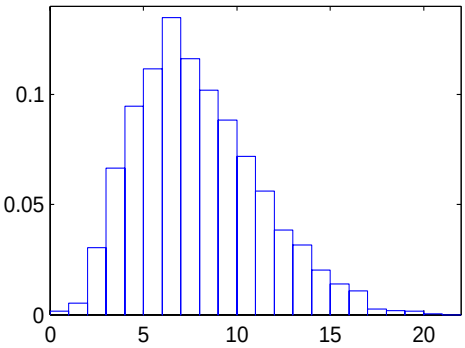


figure 2.5 : Répartition de l'intensité du vent au mois de janvier.

Plus précisément, soit $\{Y_t\}$ un processus multivarié stationnaire à valeurs dans $\boldsymbol{R}^d$ et $\{y_t\}_{t \in \{1 \ldots T\}}$ une réalisation de ce processus. Les techniques décrites dans ce paragraphe supposent qu'il existe un processus gaussien stationnaire $\{X_t\}$, à valeurs dans $\boldsymbol{R}^d$, et une transformation $G: \boldsymbol{R}^d \to \boldsymbol{R}^d$ tel que $Y_t = G(X_t)$. Les algorithmes de simulation utilisant cette démarche peuvent être divisés en trois étapes:

- **ajustement du modèle,** qui consiste à estimer la fonction G et la structure du deuxième ordre du processus gaussien $\{X_t\}$. En pratique, on va choisir G de telle manière que les lois marginales et les structures du second ordre des processus $\{Y_t\}$ et $\{G(X_t)\}$ coincident.
- **simulation du processus gaussien,** dans laquelle des réalisations du processus gaussien $\{X_t\}$ sont simulées.
- **transformation,** où la transformation G est appliquée aux échantillons simulés afin d'obtenir des réalisations du processus initial $\{Y_t\}$.

Méthode de Box et Jenkins

La méthode générale décrite ci-dessus englobe en particulier la méthode de Box et Jenkins (cf *Box et al.* (1976, [23])). Il s'agit sans doute de la méthode la plus couramment utilisée pour simuler les séries temporelles de vent (*Brown et al.* (1984, [27]), *Daniel et al.* (1991, [37]), *Nfaoui et al.* (1996, [90])) et de paramètres d'état de mer (*O'Carroll* (1984, [91]), *Stephanakos* (1999, [109]), *Cunha et al.* (1999, [36]), *Yim et al.* (2002, [126])). Elle est aussi très couramment utilisée dans de nombreux autres domaines d'application. Cette méthode est décrite brièvement ci-dessous.

Plaçons nous tout d'abord dans le cas où $d = 1$ (processus monovarié). $H = G^{-1}$ est choisie dans la famille des transformations de Box-Cox, c'est à dire de la forme:

$$T_\lambda(x) = \frac{x^\lambda - 1}{\lambda}$$

pour $x \geq 0$ et $0 \leq \lambda \leq 1$ et avec la convention $T_0(x) = \ln(x)$. Le paramètre λ est choisi de telle manière que la loi marginale de $T_\lambda(Y_t)$ soit approximativement gaussienne. Différentes métho-

des peuvent être utilisées pour estimer le paramètre λ (cf [27] et [37] par exemple). Généralement, lorsque $Y = H_s$, la transformation $T_0(x) = \ln(x)$ est utilisée (cf [91] et [109] par exemple), la loi marginale de ce processus étant généralement bien décrite par une loi log-normale. Pour le processus $Y = U$, on trouve généralement un transformation de Box-Cox avec λ compris entre 0.5 et 1 (cf [27], [90]). Enfin, d'autres transformations paramétriques ont été proposés. Par exemple, la transformation suivante est utilisée dans [91] pour le processus $Y = U$:

$$H(x) = x - \frac{k}{x}$$

Lorsque le processus $\{Y_t\}$ est multivarié, une transformation de Box-Cox est généralement appliquée indépendamment sur les différentes composantes. Dans [91], une transformation plus sophistiquée est proposée pour le processus $Y = (H_s, T_m)$: la transformation de Box-Cox usuelle est d'abord appliquée au processus H_s, puis une transformation de Box-Cox tenant compte de la spécificité de la loi conditionnelle $P(T_m | H_s = h)$ est ensuite appliquée au processus T.

Ensuite, une fois choisie la transformation H, un modèle $ARMA$ est ajusté à la série temporelle $\{H(y_t)_{t \in \{1...T\}}\}$. Ainsi, pour le processus H_s, les modèles suivants ont été proposés: $AR(1)$ ([91]), $ARMA(2, 2)$ ([109]), $AR(20)$ ([36]) et $ARMA(4, 4)$ ([126]). Pour le processus U, des modèles $AR(1)$ ([114]) ou $AR(2)$ ([37], [89], [90]) sont généralement utilisés, les modèles plus complexes n'apportant pas d'améliorations notables. Ces modèles fournissent en général une bonne approximation de la structure d'ordre 2 du processus $\{H(Y_t)\}$ (cf [109]).

Méthode non paramétrique : *TGP*

Pour les paramètres d'états de mer, une autre approche, basée sur le même principe est aussi couramment utilisée. Il s'agit d'une méthode non paramétrique, dans laquelle la fonction H est choisie en s'inspirant de la transformation des scores normaux, puis le processus gaussien $\{X_t\}$ est simulé en utilisant des techniques de simulation exacte. Cette approche a été proposée initialement dans *Walton et al.* (1990, [119]) afin de simuler des réalisations du processus H_s. Cette méthode a ensuite été étendue aux processus d'état de mer multivariés par *Borgman et al.* (1991, [18]), *Monbet et al.* (2001, [85]). Cette méthode a aussi été utilisée dans d'autres domaines d'application. Par exemple, dans *Gioffre et al.* (2000, [50]) elle est utilisée pour simuler les contraintes exercées par le vent sur un immeuble. Dans la suite, cette méthode sera notée ***TGP (Translated Gaussian Process)***.

Transformation des scores normaux

Le lemme d'inversion permet de transformer une variable aléatoire U de loi uniforme sur $[0, 1]$ en une variable aléatoire de loi quelconque dont la fonction de répartition est connue. Plus précisément, soit Y une variable aléatoire de fonction de répartition F_Y et $F_Y^-(u) = inf\{x; F_Y(x) \geq u\}$ l'inverse généralisée de F_Y. Le ***lemme d'inversion*** nous dit que la

variable aléatoire $F_Y^-(U)$ suit la même loi que Y. Cette transformation est couramment utilisée pour simuler des échantillons i.i.d. d'une loi quelconque, dont la fonction de répartition est connue, à partir d'échantillons de la loi uniforme. En utilisant le même principe, on peut alors transformer une variable aléatoire X de loi $N(0, 1)$ en une variable aléatoire qui admet pour fonction de répartition F_Y. Pour cela, on peut utiliser la ***transformation des scores normaux*** qui est définie par

$$G = F_Y^- \bullet \Phi \tag{2.3}$$

avec Φ la fonction de répartition de la loi $N(0, 1)$.

Si $\{Y_t\}$ désigne un processus monovarié et stationnaire, nous noterons F_Y la fonction de répartition de sa loi marginale. Soit $\{X_t\}$ un processus stationnaire à valeurs réelles qui admet pour loi marginale la loi $N(0, 1)$ et G la transformation définie par l'équation (2.3). On vérifie alors aisément que les processus $\{G(X_t)\}$ et $\{Y_t\}$ ont la même loi marginale. Cette transformation est utilisée au paragraphe 2.b.4 pour simuler des réalisations du processus $Y = U$. La fonction de répartition G a été estimée en utilisant la fonction de répartition empirique de la loi marginale de $\{Y_t\}$. D'autres méthodes d'estimation, plus sophistiquées, pourraient être utilisées. Dans *Borgman et al.* (1991, [18]), des formes paramétriques sont ajustées sur les queues de la distribution ce qui permet de mieux restaurer les valeurs "extrêmes" dans les séquences simulées. Dans le cas des processus monovariés, une autre transformation, non paramétrique, est proposée dans *Rychlik et al.* (1997, [99]). La fonction G est alors choisie de telle manière que les nombres moyens de franchissements de niveaux du processus gaussien transformé coïncident avec ceux de la série observée.

Plusieurs extensions ont ensuite été proposées pour les processus multivariés $Y_t = \{Y_t^{(1)}, \ldots, Y_t^{(d)}\}$ à valeurs dans $\boldsymbol{R}^d$, l'objectif étant de trouver une transformation G telle que le processus $\{G^-(Y_t)\}$ soit approximativement gaussien. Une première méthode consiste à appliquer la transformation des scores normaux indépendamment sur les différentes composantes. On prend alors $G(x_1, \ldots, x_d) = (g_1(x_1), \ldots g_d(x_d))$ avec $g_i = F_{Y^{(i)}}^- \bullet \Phi$ et $F_{Y^{(i)}}$ la fonction de répartition de la loi marginale du processus $\{Y_t^{(i)}\}$ pour $i \in \{1 \ldots d\}$. Cette méthode est utilisée par exemple dans *Borgman et al.* (1991, [18]) pour le processus (H_s, T, Θ_m) et dans *Gioffre et al.* (2000, [50]) pour simuler les contraintes exercées par le vent sur un immeuble en plusieurs points simultanément. Cependant, il est montré dans *Monbet et al.* (2001 ,[85]) que cette transformation ne permet pas de restaurer la loi marginale bivariée du processus $Y = (H_s, T_m)$ du fait de la forte relation existant entre ces deux paramètres. La transformation suivante est alors utilisée:

$$G: \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} F_{Y^{(1)}}^-(\Phi(x_1)) \\ F_{Y^{(2)}|Y^{(1)} = F_{Y^{(1)}}^-(\Phi(x_1))}^-(\Phi(t)) \end{bmatrix} \tag{2.4}$$

où $F_{Y^{(1)}}$ désigne la fonction de répartition de $Y^{(1)}$ et $F_{Y^{(2)}|Y^{(1)}=y_1}$ celle de la loi conditionnelle $P(Y^{(2)}|Y^{(1)}=y_1)$. Si $\{X_t\}$ est un processus stationnaire de loi marginale $N(0, I_2)$ alors le processus $\{G(X_t)\}$ admet la même loi marginale que $\{Y_t\}$.

Dans *Ailliot et al.* (2001, [3]), nous proposons une variante de la transformation (2.4) pour le processus $Y=(H_s, \Theta_m)$. Cette transformation tient compte de la spécificité du paramètre circulaire θ_m. Cette transformation est définie par $G = G_1 \circ G_2$ avec

$$G_1 : \begin{bmatrix} \rho \\ \alpha \end{bmatrix} \rightarrow \begin{bmatrix} F^{-}_{H_s}(R(\rho)) \\ F^{-}_{\Theta_m|H_s=F^{-}_{H_s}(R(\rho))}(\alpha) \end{bmatrix} \tag{2.5}$$

où R désigne la fonction de répartition de la loi de Rayleigh, F_{H_s} celle de la loi marginale du processus H_s et $F_{\Theta_m|H_s=h}$ celle de la loi conditonnelle $P(\Theta_m|H_s=h)$ et G_2 un inverse de la fonction H_2, définie sur $\boldsymbol{R}^{+*} \times [0, 2\pi[$ par:

$$H_2 : \begin{bmatrix} \rho \\ \alpha \end{bmatrix} \rightarrow \begin{bmatrix} \rho\cos(\alpha) \\ \rho\sin(\alpha) \end{bmatrix}$$

Pour le processus $Y=(U, \Phi)$, les résultats obtenus avec cette transformation ne sont pas satisfaisants, et le processus $\{G^{-}(Y_t)\}$ ne peut pas être supposé gaussien. Ceci est sans doute dû à la complexité de la loi marginale bivariée de ce processus (cf figure 2.8) : la forme de cette loi entraîne que la loi conditionnelle $P(\Theta_m|H_s=h)$ dépend fortement de h, et la transformation G^{-} induit alors une déformation complexe de l'espace d'état du processus initial. Nous avons obtenu de meilleurs résultats avec une transformation légèrement différente, à savoir la transformation $G = G_1 \circ G_2$ avec G_1 définie par :

$$G_1 : \begin{bmatrix} \rho \\ \alpha \end{bmatrix} \rightarrow \begin{bmatrix} F^{-}_{U|\Phi=F^{-}_{\Phi}(\alpha)}(R(\rho)) \\ F^{-}_{\Phi}(\alpha) \end{bmatrix} \tag{2.6}$$

où F_Φ désigne la fonction de répartition de la loi marginale du processus Φ et $F_{U|\Phi=\phi}$ celle de la loi conditonnelle $P(U|\Phi=\phi)$. On peut vérifier que, si $\{X_t\}$ désigne un processus stationnaire de loi marginale $N(0, I_2)$, alors le processus $\{G(X_t)\}$ admet la même loi marginale que $\{Y_t\}$, lorsque $G = G_1 \circ G_2$ avec G_1 défini par l'équation (2.5) ou (2.6).

Les résultats obtenus avec cette dernière transformation sont décrits plus précisément dans le paragraphe 2.c.1. En pratique, les fonctions de répartition qui interviennent dans l'équation (2.6) ont été estimées en utilisant la fonction de répartition empirique de la loi marginale du processus $\{Y_t\}$. Nous noterons dans la suite $\hat{G}$ l'estimateur de G correspondant.

Estimation de la structure du second ordre

Supposons maintenant que la fonction G est fixée, il reste alors à estimer la structure du deuxième ordre du processus gaussien $\{X_t\}$. Dans *Borgman et al.* (1991, [18]) et *Monbet et al.* (2001, [85]) la fonction d'autocorrélation de ce processus est estimée par la fonction d'autocorrélation empirique de la série temporelle transformée $\left\{\hat{G}^{-}(y_t)\right\}_{t \in \{1 \ldots T\}}$. Les simulations effectuées au paragraphe 2.b et 2.c ont été obtenues avec ces estimateurs. A notre connaissance, il n'existe pas de résultat théorique sur les propriétés asymptotiques de ces estimateurs.

Les résultats obtenus au paragraphe 2.c montrent que les séquences simulées ont des fonctions d'autocorrélation significativement différentes de celles de la séquence initiale. Dans *Gioffre et al.* (2000, [50]), dans le cas où la transformation G est la transformation des scores normaux appliquée indépendamment sur chaque composante, ce problème est résolu de la manière suivante. On note, pour $k \in \{1 \ldots d\}$, $Z_t^{(k)} = g_k(X_t^{(k)})$ avec g_k la transformation des scores normaux associée à la k^{ieme} composante du processus et $\{X_t\}$ un processus gaussien stationnaire. On peut vérifier que, pour $(h, t) \in N^2$ et $(i, j) \in \{1 \ldots d\}^2$ on a

$$cov(Z_i(t), Z_j(t+h)) = \iint g_i(x_i) g_j(x_j) f(x_i, x_j, \rho_{i,j}(h)) dx_i dx_j$$

avec $f(x_i, x_j, \rho_{i,j}(h))$ la densité d'une loi normale de moyenne $[0, 0]'$ et de matrice de variance-covariance

$$\begin{bmatrix} 1 & \rho_{i,j}(h) \\ \rho_{i,j}(h) & 1 \end{bmatrix}$$

et $\rho_{i,j}(h) = cov(X_i(t), X_j(t+h))$ la fonction d'autocorrélation du processus $\{X_t\}$. La fonction d'autocorrélation du processus $\{X_t\}$ est alors estimée par $(\hat{\rho}_{i,j}(h))_{(i,j) \in \{1 \ldots d\}^2, h \in \{0, \ldots, T\}}$ de telle manière que, pour $(i, j) \in \{1 \ldots d\}^2$ et $h \in \{0, \ldots, T\}$:

$$\iint \hat{g}_i(x_i) \hat{g}_j(x_j) f(x_i, x_j, \hat{\rho}_{i,j}(h)) dx_i dx_j = \hat{\sigma}_{i,j}(h)$$

avec $\hat{\sigma}_{i,j}(h)$ la fonction d'autocorrélation empirique du processus $\{Y_t\}$. Avec cette méthode d'estimation, le processus $\{\hat{G}(X_t)\}$ va avoir comme fonction d'autocorrélation la fonction d'autocorrélation empirique du processus observé. Une procédure itérative permettant de calculer les $\rho_{i,j}(h)$ est décrite plus précisément dans *Popescu et al.* (1998, [93]). Il est par ailleurs mentionné dans *Gioffre et al.* (2000, [50]) que la solution à ce problème n'existe pas toujours.

Ce type de méthode pourrait sans doute s'adapter à la transformation (2.6) ce qui permettrait d'améliorer l'adéquation entre la structure d'ordre 2 des données et celle des séquences simulées.

Simulation du processus gaussien

Enfin, il reste à simuler des réalisations du processus gaussien stationnaire $\{X_t\}$ dont la loi marginale est la loi $N(0, I)$ et dont la fonction d'autocorrélation est connue. Différentes méthodes de simulation exacte ont été proposées pour les processus gaussiens. Dans la suite, nous avons utilisé la méthode décrite dans *Borgman et al.* (1991, [18]) et *Popescu et al.* (1998, [93]). Cette méthode utilise les propriétés de la transformée de Fourrier d'un processus gaussien, et a l'avantage de pouvoir être implémentée rapidement en utilisant la transformée de Fourrier rapide. D'autres techniques de simulations plus sophistiquées peuvent être trouvées dans *Dietrich et al.* (1997, [40])). Les séries temporelles considérées en pratique sont relativement courtes (122 données par mois) et l'utilisation de ces algorithmes ne semble pas justifiée.

Finalement, les méthodes décrites dans ce paragraphe sont relativement simples à mettre en oeuvre, et permettent de simuler rapidement des séquences artificielles qui possèdent une loi marginale et une fonction d'autocovariance proche de celles des données. Cependant, comme nous allons le voir au paragraphe 2.b et 2.c, ces méthodes ne permettent pas de reproduire certaines non-linéarités qui sont présentes dans les données. Il faut alors utiliser des modèles plus évolués.

2.a.3 Modèles markoviens

Dans ce paragraphe, nous avons regroupé différents modèles markoviens qui ont été proposés pour les séries temporelles de vent et plus généralement d'état de mer. Notons tout d'abord qu'il semble physiquement raisonnable de supposer que ces processus sont des chaînes de Markov d'ordre restreint et nous avons vu qu'un modèle autorégressif linéaire d'ordre 1 ou 2 était généralement suffisant pour décrire la structure d'ordre 2 du processus U.

Dans un premier temps, nous nous intéressons aux chaînes de Markov à espace d'état fini, puis dans une deuxième partie, nous décrivons brièvement une autre méthode développée au début de cette thèse afin de simuler le processus bivarié (H_s, θ_m). Elle repose sur une approximation de la série temporelle initiale par une courbe linéaire par morceaux, chaque rupture pouvant alors être interprétée comme le début ou le maximum d'une tempête. Le nouveau processus obtenu est ensuite modélisé en utilisant une hypothèse markovienne.

Enfin les deux parties suivantes sont consacrées aux modèles autorégressifs non-linéaires. Nous décrivons tout d'abord différents modèles qui ont été proposés pour les séries temporelles circulaires, c'est à dire à valeur dans le tore $R/2\pi Z$, puis nous nous intéressons à deux modèles autorégressifs qui ont été proposés pour modéliser les processus H_s et U, et en particulier à un modèle autorégressif à seuil (TAR) et un modèle hétéroscédastique $GARCH$.

Chaînes de Markov à espace d'état fini

Décrivons tout d'abord brièvement le principe général des méthodes décrites dans ce paragraphe.

- lorsque le processus est à espace d'état continu, ce qui est généralement le cas pour les para-

mètres d'état de mer, cet espace est tout d'abord discrétisé en un nombre fini de classes, ce qui permet de se ramener à un processus à espace d'état fini.

- ce nouveau processus est ensuite supposé être une chaîne de Markov dont la matrice de transition est estimée à partir des séquences observées.
- de nouvelles réalisations du processus sont finalement simulées à partir de cette matrice de transition.

Le succès de cette méthode est sans doute principalement dû à sa simplicité et à sa rapidité de mise en oeuvre. La limitation principale est le nombre important de paramètres à estimer lorsque l'espace d'état de la chaîne de Markov est grand, ce qui est le cas si une discrétisation suffisamment fine de l'ensemble des valeurs prises par le processus initial est utilisée. Ainsi, pour l'intensité du vent U, des classes de largeur $2ms^{-1}$ sont sans doute acceptables pour les applications envisagées, ce qui pour les données considérées dans cette thèse aboutirait à une dizaine de classes (les vents observés varient entre $0ms^{-1}$ et environ $20ms^{-1}$, cf figure 2.5). Si on ne fait aucune hypothèse sur la forme de la matrice de transition, il faut alors estimer une centaine de paramètres pour un modèle d'ordre 1 et environ mille paramètres pour un modèle d'ordre 2.

Différents modèles ont été alors proposés pour réduire le nombre de paramètres. Ainsi, dans *Vik* (1981, [117]) les valeurs prises par le processus H_s sont discrétisées en classes de largeur $1m$ puis une chaîne de Markov d'ordre 1 est utilisée pour décrire le processus à espace d'état fini correspondant. Afin de réduire le nombre de paramètres, la matrice de transition est supposée tridiagonale, c'est à dire que seules les transitions aux états voisins sont autorisées. Les coefficients de la matrice de transition sont alors estimés de telle manière que la loi stationnaire de la chaîne de Markov et les durées de persistance moyennes au dessus de certains seuils soient celles calculées à partir de la base de données: on obtient alors un système d'équations qui permet de calculer les coefficients de la matrice de transition. Ce type d'hypothèse est sans doute irréaliste pour l'intensité du vent puisque ce processus évolue rapidement.

Par contre, l'utilisation des chaînes de Markov à espace d'état fini semble plus adaptée pour décrire la direction du vent. En effet, de nombreuses bases de données fournissent ce paramètre directement sous forme de données discrétisées : il est habituel pour les météorologistes de regrouper les directions de vent en "secteur" (en général, 16 classes sont utilisées). De plus, comme nous le verrons dans la suite de ce chapitre, il existe relativement peu de modèles permettant de décrire directement les séries temporelles circulaires. Ainsi, dans *MacDonald et al.* (1997, [76]), une chaîne de Markov du premier ordre est utilisée pour décrire l'évolution de la direction du vent. Ce modèle ne permettant pas de prendre en compte de manière satisfaisante la dépendance temporelle de ce processus, ils proposent alors d'utiliser une chaîne de Markov d'ordre 2. Afin de limiter le nombre de paramètres ils utilisent le modèle de Raftery dans lequel les matrices de transitions sont supposées être de la forme:

$$P(\Phi_t = j | \Phi_{t-1} = i_1, \Phi_{t-2} = i_2) = \lambda_1 q_{i_1,j} + \lambda_2 q_{i_2,j} \tag{2.7}$$

avec $Q = (q_{i,j})$ une matrice stochastique et λ_1 et λ_2 des paramètres positifs vérifiant la contrainte $\lambda_1 + \lambda_2 = 1$. Ce modèle possède un paramètre de plus que le modèle d'ordre 1 et une comparaison basée sur des critères de log-vraisemblance pénalisée (BIC et AIC) montre une meilleure adéquation de ce modèle en comparaison du modèle d'ordre 1. Ces résultats ont été obtenus sur des données de vent à Koeberg (Afrique du Sud).

Processus ponctuels marqués

Une autre approche, utilisant aussi les chaînes de Markov, a été développée au début de cette thèse afin de simuler des réalisations du processus (H_s, Θ_m). La première étape consiste à approcher la série temporelle H_s par un processus linéaire par morceaux, que nous noterons H_{lin}. Une telle approximation est relativement classique pour les données d'état de mer (*Labeyrie* (1990, [71]), *Arena* (2002, [7])). Elle permet de découper la série initiale en une suite "d'événements météorologiques" (tempête ou période de calmes par exemple), ce qui peut être plus facile à modéliser. Dans *Ailliot et al.* (2001, [3]), les instants de rupture sont détectés aux dates auxquelles la série temporelle atteint ses extrema locaux. Cette méthode fonctionne relativement bien pour le processus H_s puisque ces ruptures ont une interprétation météorologique réaliste (arrivée ou maximum de tempêtes) et que la courbe linéaire par morceaux fournit une bonne approximation de la série initiale. Un exemple de découpage obtenu avec la même méthode pour le processus U est donné sur la figure 2.6. On voit nettement apparaître quelques coups de vent au début de la séquence, et chacun d'entre eux est alors décrit par un triangle. Par contre, à la fin de la séquence, de nombreuses ruptures, sans interprétations météorologiques claires, sont détectées, et il faudrait alors utiliser un algorithme de détection de rupture plus sophistiqué. Cette approche n'a pas été approfondie plus en détail pour les données de vent. Des détails supplémentaires peuvent être trouvés dans *Ailliot et al.* (2001, [3]). En particulier un modèle markovien est proposé pour le processus ponctuel marqué correspondant au processus bivarié (dates de rupture, hauteur significative aux dates de rupture). Nous donnons aussi une comparaison des résultats obtenus avec cette méthode et avec la méthode *TGP* décrite au paragraphe 2.a.2.

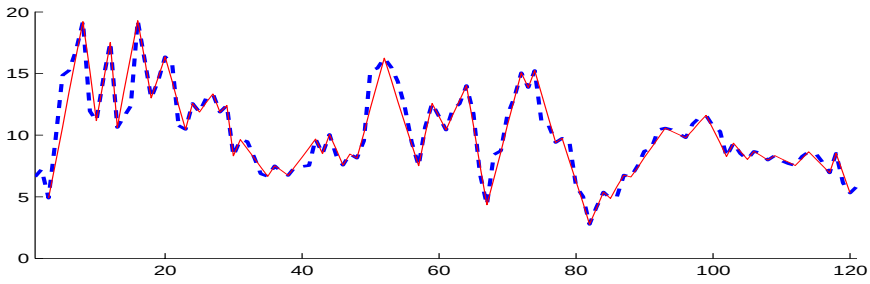


figure 2.6 : Evolution de l'intensité du vent au cours du mois de janvier 1991 (en tiret) et approximation par un processus linéaire par morceaux (trait continu)

Modèles autorégressifs pour les variables circulaires

L'analyse des variables aléatoires circulaires (c'est à dire à valeur dans le tore $R/2\pi Z$) est relativement classique (cf *Mardia* (1972, [75]) par exemple), mais par contre peu d'auteurs se sont intéressés à l'analyse des séries temporelles de variables circulaires. Nous décrivons ici différents modèles autorégressifs paramétriques qui ont été proposés pour ce type de processus. Une description plus détaillée de ces modèles peut être trouvée dans *Breckling* (1989, [25]) et *Fisher et al.* (1994, [47]). Nous revenons au paragraphe 2.c.1 sur l'utilisation de ces modèles pour nos données.

Soit $\{\Phi_t\}$ un processus stationnaire à valeurs dans $R/2\pi Z$. Principalement 3 types de modèles autorégressifs ont été proposés pour un tel processus:

- Les modèles obtenus par "pliage" d'un processus à valeurs réelles (Wrapped Autoregressive model). On suppose alors que $\Phi_t = Y_t \ [2\pi]$ avec $\{Y_t\}$ un processus à valeurs réelles qui suit une modèle autorégressif.
- Les modèles obtenus en utilisant une "fonction lien", c'est à dire une fonction $g : \mathrm{R} \to]-\pi, \pi[$ strictement monotone et vérifiant $g(0) = 0$ (on peut prendre par exemple $g(x) = 2\arctan(x)$). On utilise alors cette fonction pour transformer un processus autorégressif $\{Y_t\}$ à valeurs réelles en un processus $\Phi_t = g(Y_t)$ à valeurs dans le tore.
- Les modèles spécifiant directement la densité de la loi conditionnelle $P(\Phi_t | \Phi_{t-1}, \ldots, \Phi_{t-k})$. C'est par exemple le cas du modèle autorégressif de Von-Mises proposé initialement dans *Breckling* (1989, [25]). La loi de Von-Mises de paramètres (θ, κ) est définie par sa densité :

$$f(\phi) = \frac{1}{2\pi I_0(k)} e^{\kappa \cos(\phi - \phi_0)}$$

pour $\phi \in R/2\pi Z$ avec $\kappa > 0$ la concentration et $\phi_0 \in R/2\pi Z$ la direction moyenne. On définit alors le modèle autorégressif de Von-Mises de la manière suivante: on suppose que $P(\Phi_t | \Phi_{t-1}, \ldots, \Phi_{t-k})$ suit une loi de Von Mises de paramètres (θ_t, κ_t) donnés par

$$\kappa_t e^{i\theta_t} = \kappa_1 e^{i\phi_{t-1}} + \kappa_2 e^{i\phi_{t-2}} + \ldots \kappa_p e^{i\phi_{t-p}} + \kappa_0 e^{i\phi_0}$$

avec $\kappa_0, \kappa_1, \ldots \kappa_p \in R^{+*}$ et $\phi_0 \in R/2\pi Z$ des paramètres. Dans ce modèle, la concentration $\kappa(t)$ évolue au cours du temps (modèle hétéroscédastique). On peut préférer le modèle à concentration constante où $\theta(t)$ est défini de la même manière et $\kappa(t)$ est choisi constant, égal à κ.

Modèles autorégressifs non linéaires pour U et H_s

Dans la fin de ce paragraphe nous décrivons plusieurs modèles autorégressifs paramétriques non-linéaires qui ont été proposés pour les processus U et H_s. Ces modèles ont été introduits récemment afin de modéliser certaines non-linéarités présentes dans les données et qui ne peuvent pas être modélisées par l'approche introduite au paragraphe 2.a.2.

Réseaux de neurones

Différents auteurs ont proposé d'utiliser les réseaux de neurones afin de modéliser l'évolution des processus U (*Stephos* (2000, [110]), *More et al.* (2003, [89])) et H_s (*Monbet et al.* (2001, [86])). D'après les résultats obtenus par ces différents auteurs, il semble que ces modèles permettent d'obtenir des prédictions à court terme légèrement meilleures que celles obtenues avec les modèles autorégressifs linéaires (cf [89]).

Nous avons testé ces modèles sur nos données d'intensité du vent. En utilisant un perceptron multi-couche avec 1 couche cachée et 5 unités (modèle avec 16 paramètres), la diminution de la variance de l'erreur de prédiction à un pas de temps est inférieure à 1% par rapport à un modèle autorégressif linéaire d'ordre 1 (2 paramètres). L'introduction de ce modèle plus complexe ne semble donc pas justifiée d'autant plus que ses paramètres sont difficiles à interpréter. Sur la figure 2.7, nous avons représenté le nuage de point (U_{t-1}, U_t), et la relation entre ces deux variables aléatoires semble approximativement linéaire.

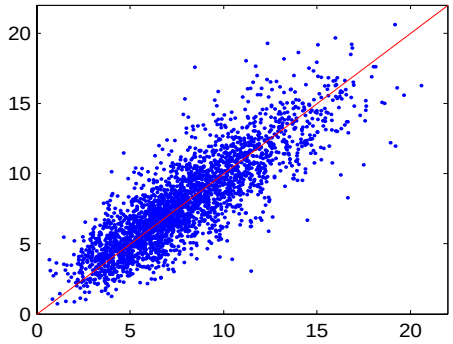


figure 2.7 : Nuage de points (U_{t-1}, U_t).

Modèles à seuil

Dans *Scotto et al.* (2000, [104]) un modèle ***SETAR*** (***Self-Exciting Threeshold AutoRegressive model)*** est proposé pour représenter le processus $Y = H_s$. Ces auteurs supposent que:

$$Y_t = \sum_{i=1}^{r} a_i^{(S_t)} Y_{t-i} + b^{(S_t)} + \sigma^{(S_t)} \varepsilon_t$$

avec $S_t = i$ si et seulement si $Y_{t-d} \in [r_i, r_{i+1}[$. Ici, $r_1 < r_2 < \ldots < r_M$ sont des paramètres du modèle et $\{\varepsilon_t\}$ désigne un bruit blanc gaussien. Il s'agit d'un modèle autorégressif à changements de régimes, dans lequel la valeur du régime à l'instant t_0 dépend uniquement des valeurs passées du processus $\{Y_t\}$. Nous expliquons au paragraphe 2.b.1 pourquoi l'utilisation d'un modèle $MS-\gamma AR$ nous semble mieux adaptée pour décrire l'évolution de l'intensité du vent.

En pratique, le modèle identifié comporte 2 régimes, l'évolution dans chaque régime étant décrite par un modèle $AR(10)$ et $d = 7$, ce qui correspond à un retard de 21 heures. Les auteurs comparent les résultats obtenus avec ce modèle et ceux obtenus avec un modèle

$AR(22)$. Les séquences simulées avec le modèle $SETAR$ semblent avoir des caractéristiques "plus proches" de celles des données que celles simulées avec le modèle $AR(22)$, notamment en ce qui concerne la loi marginale et la fonction d'autocorrélation.

Modèles GARCH

Un modèle autorégressif non linéaire du type ***GARCH*** est proposé dans *Toll* (1997, [114]) pour le processus $Y = U$. Il suppose alors que ce processus est markovien d'ordre r, la loi conditionnelle de Y_t sachant que $(Y_{t-1}, \ldots, Y_{t-r}) = (y_{t-1}, \ldots, y_{t-r})$ étant décrite par une loi gamma

- de moyenne $\mu_t = \sum_{i=1}^{r} a_i y_{t-i} + b$
- de variance $\sigma_t^2 = \alpha + \sum_{i=1}^{p} \lambda_i (y_{t-i} - \mu_{t-i})^2 + \sum_{i=1}^{q} \kappa_i \sigma_{t-i}^2$

Le modèle identifié est un modèle d'ordre $r = 2$ et il est montré dans *Toll* (1997, [114]) que le modèle obtenu permet de décrire l'hétéroscédasticité présente dans les séries temporelles d'intensité du vent. Cette hétéroscédasticité est clairement visible sur le nuage de points montré sur la figure 2.7 : le nuage est plus dispersé pour les vents de forte intensité que pour les vents de faible intensité. Nous verrons dans le paragraphe 2.b.1 que les modèles $MS-AR$ permettent aussi de modéliser cette hétéroscédasticité et pourquoi ce dernier modèle nous semble mieux adapté.

2.a.4 Local Grid Bootstrap

Dans ce paragraphe, nous décrivons une méthode de bootstrap qui permet de simuler des réalisations d'un processus markovien stationnaire multivarié à espace d'état continu. Les techniques de bootstrap ont été introduites par *Efron* (1979, [44]) dans le cas des variables i.i.d. Récemment, différentes extensions ont été proposées pour les séries temporelles (*Bühlmann* (2002, [29]), *Politis* (2003, [92]), *Monbet et al.* (2003, [88])).

Décrivons plus précisément la méthode de "***Local Grid Bootstrap***" (***LGB***) introduite dans [88]. Soit $\{Y_t\}$ un processus stationnaire à valeur dans $\boldsymbol{R}^d$, que nous supposerons être une chaîne de Markov et $\{y_t\}_{t \in \{0, \ldots, T\}}$ une réalisation de ce processus. Afin de simplifier l'exposé, nous nous plaçons dans la suite dans les cadre des chaînes de Markov d'ordre 1, l'extension au cas des modèles d'ordre supérieur étant immédiate. Nous noterons $P(y, A)$ le noyau de transition de cette chaîne de Markov. Le principe de la méthode LGB consiste à construire un estimateur non paramétrique de ce noyau, noté $\hat{P}(y, A)$, puis à simuler des réalisations d'une chaîne de Markov avec ce noyau empirique.

Soit $y \in \boldsymbol{R}^d$. Le noyau de transition $\hat{P}(y, A)$ va être à support fini $G_y^{(T)}$, la probabilité de $y' \in G_y^{(T)}$ étant donnée par :

$$\hat{P}(y, y') = \frac{\hat{p}(y, y')}{\sum_{z \in G_y^{(T)}} \hat{p}(y, z)} \tag{2.8}$$

avec

- $I_y = \{t \in \{0, \ldots, T-1\} | d(y_t, y) < \sigma_T\}$ l'ensemble des point observés dans un voisinage de y pour une certaine distance fixée $d(\,.\,,\,.\,)$.
- $\hat{p}(y, y') = \sum_{t \in I_y} K_d\left(\frac{y' - y_{t+1}}{h_T}\right) K_d\left(\frac{y - y_t}{h_T}\right)$
- $G_y^{(T)} = \{y_{t+1} | t \in I_y\} \cup G$ l'ensemble des points qui peuvent être atteints par la chaîne de Markov à partir de y. Ici G représente une grille fixée contenant un nombre fini de points. L'introduction de cette grille permet d'assigner des probabilités non nulles à des points non observés dans la série initiale.
- K_d est un noyau sur $\boldsymbol{R}^d$ qui vérifie les hypothèses usuelles pour l'estimation de densité par une méthode à noyaux (cf *Monbet et al.* (2003, [88]) et *Silverman* (1986, [106])).
- $h_T = \frac{\sum_{i=1}^{d} s_{i,i}}{d} \left(\frac{4}{T(2d+1)}\right)^{1/(d+4)} h_0$ avec $s_{i,i}$ la variance de la $i^{\text{ème}}$ loi marginale du processus $\{Y_t\}$ et $h_0 > 0$ une constante fixée.

Une discussion sur le choix de σ_T, h_0 et G peut être trouvée dans *Monbet et al.* (2001, [87]). Il y est par ailleurs montré, que sous certaines hypothèses, la chaîne de Markov associée au noyau $\hat{P}$ est positive récurrente. De plus, le noyau de transition et la loi stationnaire de cette chaîne de Markov convergent faiblement vers ceux de la chaîne initiale quand T tend vers $+\infty$.

Notons enfin qu'une généralisation de cette méthode aux processus cyclostationnaires peut être trouvée dans *Monbet et al.* (2003, [88]). Avec cette extension, on peut simuler directement des processus saisonniers, sans enlever les composantes saisonnières au préalable. Il est aussi intéressant de noter que cette méthode peut être utilisée pour simuler des réalisations d'une série temporelle circulaire : il suffit pour cela de choisir une distance $d(\,.\,,\,.\,)$ sur le tore.

Les résultats obtenus avec cette méthode pour le processus (U, Φ), ainsi qu'une comparaison avec ceux obtenus avec les modèles TGP et $MS-AR$ sont décrits au paragraphe 2.c.1.

2.b Modèle MS-AR pour l'intensité du vent

Afin de modéliser l'évolution de l'intensité du vent, nous proposons d'utiliser un modèle autorégressif à changements de régimes markoviens. A notre connaissance ce type de modèle n'a jamais été proposé pour les processus d'état de mer.

Le premier paragraphe est consacré à la justification physique du choix de ce modèle en introduisant la notion de "type de temps". En particulier, nous expliquons pourquoi l'emploi d'un modèle $MS-AR$ nous semble plus adapté que les modèles autorégressifs non-linéaires introduits au paragraphe 2.a.3.

Dans le deuxième paragraphe, nous nous expliquons brièvement comment nous avons ajusté le modèle aux données, puis les deux derniers paragraphes sont consacrés à la validation du modèle identifié. Dans le troisième paragraphe, nous vérifions l'interprétabilité physique de ce modèle, et nous montrons en particulier que les valeurs prises par la variable cachée correspondent à des types de temps réalistes. Enfin, dans le paragraphe 2.b.4, nous montrons, en utilisant la méthode de validation introduite au paragraphe 1.e.2, que les modèles $MS-\gamma AR$ permettent de restituer les principales caractéristiques des données de vent et améliorent sensiblement les résultats obtenus avec la méthode "usuelle", à savoir la méthode TGP.

2.b.1 Description du modèle et justification physique

Nous avons déjà évoqué à plusieurs reprises l'existence de "type de temps" dans le paragraphe 2.a. Il s'agit d'une notion couramment utilisée par les météorologues afin de décrire la climatologie d'une région donnée (cf *Bénichou* (1995,[12])). Ainsi, il est généralement admis que les différentes configurations météorologiques peuvent être regroupées en quelques classes, chacune d'entre elles étant caractérisée par la position des principaux centres d'action (anticycloniques et dépressionnaires) qui gouvernent le climat de la région considérée. Cela se traduit de différentes manières sur les séries temporelles de vent.

Ainsi, sur la figure 2.8, nous avons représenté la loi marginale bivariée du processus $\{U_t, \Phi_t\}$ au mois de janvier. Cette densité est clairement bimodale, avec un mode correspondant à des vents de nord-est relativement faibles et un mode correspondant à des vents de sud-ouest et qui sont de plus forte intensité. Notons que cette bimodalité peut aussi être observée sur la répartition de la direction du vent (cf figure 2.3).

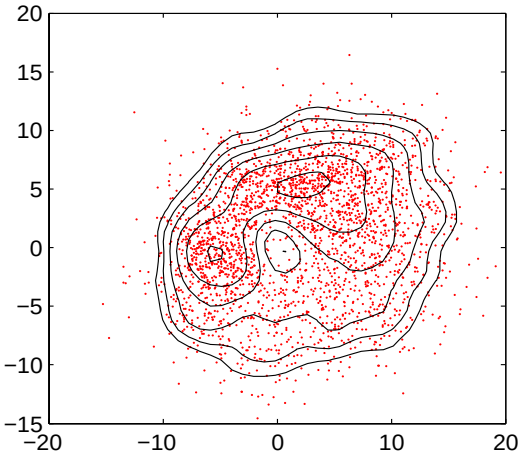


figure 2.8 : Densité empirique de la loi marginale du processus (u, v) au mois de janvier. Les courbes superposées sur le nuage de points représentent certaines lignes de niveau de la densité empirique estimée par une méthode à noyau.

En fait, chacun de ces modes correspond à un type de temps (ou "régime climatologique") couramment observé à cette époque de l'année. Ces deux types de temps sont décrits ci-dessous:

- **le type de temps «ouest perturbé»:** dans ce type de temps, le centre d'action positif est une zone anticyclonique s'étendant du sud des Açores au sud du continent Européen et le centre d'action négatif est une vaste zone dépressionnaire sur le nord-est de l'Atlantique. Ce type de temps est souvent fortement perturbé. Les perturbations contournent les grands centres d'action décrits précédemment et elles ont donc généralement des trajectoires orientées ouest-est. Leurs vitesses de déplacement sont généralement comprises entre 40 et 80 km/h mais peuvent atteindre jusqu'à 130 km/h dans les régimes les plus perturbés. Nous reviendrons sur l'étude de ces déplacements au chapitre 3, lorsque nous décrirons le modèle spatio-temporel. Ces dépressions se succèdent avec une période de 48 heures en général, cette périodicité pouvant s'abaisser jusqu'à 24 heures dans les régimes les plus perturbés. Les vents de sud-est à sud-ouest à l'avant de la perturbation tournent au secteur ouest ou nord ouest à l'arrière. Ces perturbations peuvent amener des vents violents et des mers grosses sur la côte Atlantique française. Finalement, l'évolution du vent dans ce type de temps peut se résumer approximativement de la manière suivante :
  - le processus U évolue rapidement (forte volatilité) lors du passage des groupes de perturbations et peut prendre de fortes valeurs.
  - les vents soufflent généralement de l'ouest-sud-ouest avec une rotation plus ou moins rapide lors du passage des perturbations.
- **le type de temps «nord-est anticyclonique»:** ce régime est généralement caractérisé par la présence d'un anticyclone à l'ouest des îles Britanniques se prolongeant par une dorsale en direction du large ouest du Golfe de Gascogne. Ce type de temps peut durer plusieurs semaines. Il est généralement faiblement perturbé, cependant des coups de vent y sont parfois associés lorsque le gradient de pression se renforce sur la face sud-est de l'anticyclone. L'évolution du processus (U, Φ) dans ce type de temps est caractérisée par
  - une évolution lente du processus U (faible volatilité) et des valeurs généralement moins fortes que dans le type de temps "ouest perturbé".
  - les vents soufflent généralement du nord-est.

Une classification plus fine, utilisant un plus grand nombre de types de temps peut être trouvée dans [105].

Nous proposons d'introduire cette notion de type de temps sous la forme d'une chaîne de Markov cachée à espace d'état fini. Différents auteurs ont déjà proposé d'utiliser des modèles CMC afin d'introduire la notion de régime climatologique pour des séries temporelles de données climatologiques (cf *Hugue et al.* (1999, [63]) et *MacDonald et al.* (1997, [76]), par exemple). Dans ces modèles, les valeurs prises par la variable observée sont supposées être indépendantes conditionnellemment à la chaîne cachée. Nous avons trouvé que ce type de modèle ne permettait pas de reproduire la forte dépendance existant entre deux valeurs successives de l'intensité du vent, et nous proposons alors d'utiliser un modèle $MS-AR$. Il reste ensuite à

choisir la forme des modèles autorégressifs qui vont permettre de décrire l'évolution de l'intensité du vent dans les différents types de temps. Comme ce processus est à valeurs positives, l'utilisation d'un modèle $MS-LAR$ avec innovations gaussiennes n'est pas appropriée. Une première alternative consiste à appliquer une transformation préalable sur nos données, comme dans l'approche de Box et Jenkins décrite au paragraphe 2.a.2. L'inconvénient principal de cette méthode est que la série temporelle transformée ne correspond plus à un paramètre météorologique et que son évolution est alors plus difficile à interpréter. Nous avons alors préféré utiliser un modèle $MS-\gamma AR$. Nous avons choisi d'utiliser la loi gamma parce qu'elle a pour support $\boldsymbol{R}^+$ et qu'elle est caractérisée par ses moments d'ordre 1 et 2. Nous avons déjà mentionné au paragraphe 1.a.2 que ce choix est arbitraire et que d'autres distributions ayant les même caractéristiques (loi log-normale par exemple) pourraient être utilisées. Nous n'avons trouvé ni évidences physiques ni critères statistiques rigoureux permettant de choisir entre ces différentes distributions. Cependant, nous avons obtenu de bons résultats avec la loi gamma (cf 2.b.4), ce qui semble justifier à posteriori le choix de ce modèle.

Le modèle TAR décrit au paragraphe 2.a.3 permet aussi d'introduire la notion de régime météorologique puisqu'il s'agit d'un modèle autorégressif à changements de régimes dans lequel le processus $\{S_t\}$ est endogène, la valeur prise par cette variable aléatoire étant uniquement déterminée par les valeurs passées du processus observé. Au contraire, dans les modèles $MS-AR$, les régimes évoluent de manière autonome : conditonnellement au passé, le régime à l'instant t ne dépend que du régime à l'instant $t-1$. Ce deuxième modèle nous semble plus naturel pour l'intensité du vent puisqu'il semble physiquement difficile de prévoir les changements de types de temps en utilisant l'intensité du vent aux instants précédents.

Le modèle $GARCH$ décrit au paragraphe 2.a.3 permet de décrire l'hétéroscédasticité du processus U induite par les changements de types de temps. Ainsi, il permet de modéliser le fait qu'à certaines périodes le processus évolue lentement (par exemple dans les régimes anticycloniques), ce qui est modélisé par une faible valeur de σ_t, alors qu'à d'autres périodes, lors du passage d'un groupe de perturbations, il évolue plus rapidement, ce qui se traduit par une forte valeur de σ_t. Cependant, dans ce modèle, la notion de type de temps n'est pas directement introduite et ses paramètres semblent plus difficilement interprétables.

2.b.2 Calibration

La phase de calibration consiste à identifier le modèle qui s'ajuste le mieux aux données : il faut alors choisir le nombre de régimes M et l'ordre des modèles autorégressifs r (sélection du "meilleur" modèle), et estimer la valeur des paramètres de ce modèle. Pour cela, nous avons tout d'abord calculer les EMV pour différentes valeurs de M et r. Pour cela, nous avons utilisé l'algorithme décrit au paragraphe 1.c.4. Pour initialiser cet algorithme, nous avons tiré aléatoirement, de manière indépendante, $N_{init1} = 20 \times M^2$ valeurs des paramètres en utilisant les lois suivantes :

- $a_i^{(s)}$ de loi uniforme sur [0, 1]

- $b^{(s)}$ de loi uniforme sur [0, 5]
- $\sigma^{(s)}$ de loi uniforme sur [1.5, 4]
- $Q \sim I_M + R$ avec $R = (r_{i,j})$ des coefficients de loi uniforme sur [0, 1].

Le choix de ces lois a été guidé par la connaissance physique du phénomène, ce qui nous donne notamment un ordre de grandeur pour les valeurs des paramètres $a_i^{(s)}$, $b^{(s)}$ et $\sigma^{(s)}$ qui servent à décrire l'évolution du processus observé dans les différents régimes. Pour la matrice de transition Q, nous avons choisi des matrices aléatoires ayant des valeurs relativement élevées sur la diagonale puisqu'on s'attend à ce que les types de temps aient une durée de persistance moyenne de l'ordre de quelques jours. Pour les données utilisées, il semble que surtout le choix des paramètres régissant l'évolution du processus observé dans les différents régimes détermine la position finale des paramètres, alors que le choix de Q a peu d'influence.

Nous avons ensuite calculé le critère BIC afin d'effectuer une première sélection de modèle (cf paragraphe 1.e.1). Les valeurs obtenues pour les modèles $MS-\gamma AR$ pour $r = 1$ et $1 \leq M \leq 5$ sont données dans le tableau 2.2. Parmi ces modèles, le critère BIC sélectionne le modèle avec $M = 3$, mais on peut remarquer que l'écart avec le modèle à deux régimes est relativement faible. A titre de comparaison, les valeurs du critère AIC pour les différents modèles sont données dans le tableau 2.2. Ce critère décroît avec le nombre de paramètres et sélectionne donc un modèle avec un nombre de régimes supérieur ou égal à 5. Or ces modèles sont physiquement peu réalistes, puisque, par exemple, ils possèdent des matrices de transitions avec de très faibles valeurs sur la diagonale (inférieur à 0.1) ce qui correspond à des régimes avec des durées moyennes de persistance faibles. Nous avons aussi testé ces modèles en simulation avec la méthode introduite au paragraphe 1.e.2, et les résultats obtenus sont moins bons qu'avec les modèles à 2 et 3 régimes.

M	1	2	3	4	5
BIC	10485	10316	**10307**	10343	10387
AIC	10467	10273	10207	10184	10168

Tableau 2.2 Critères AIC et BIC pour les modèles $MS-\gamma AR$ avec $r = 1$ (mois de janvier).

Nous avons aussi calculé la valeur du critère BIC pour les modèles d'ordre $r = 2$ et la comparaison est nettement favorable aux modèles d'ordre $r = 1$. Afin de tester si ceci est dû à la contrainte $a_s^{(2)} \geq 0$, nous avons fait les mêmes calculs pour le modèle $MS-LAR$ avec innovations gaussiennes et la comparaison des critères BIC est toujours nettement favorable aux modèles d'ordre $r = 1$. Nous avons par ailleurs trouvé que la valeur des paramètres obtenus pour ce dernier type de modèle est très proche de celle obtenue pour les modèles $MS-\gamma AR$. En pratique, il semble alors possible d'utiliser les modèles $MS-LAR$ avec innovations gaussiennes afin d'inférer la forme du modèle $MS-\gamma AR$ (sélection de modèle, première estimation de la

valeur des paramètres...). Pour ces modèles, le calcul les EMV est nettement plus rapide, puisque nous disposons d'expressions analytiques pour les maxima dans l'étape M.

2.b.3 Interprétabilité du modèle

Comme nous l'avons vu ci-dessus, le critère BIC sélectionne le modèle $MS-\gamma AR$ à 3 régimes, mais la valeur de ce critère pour le modèle à 2 régimes est relativement proche. Nous proposons d'étudier plus en détail ces deux modèles et dans la suite de ce paragraphe 2.b. Nous allons tout d'abord vérifier l'interprétabilité physique de ces deux modèles.

Modèle $MS-\gamma AR$ à deux régimes

Considérons tout d'abord le modèle à deux régimes. Les valeurs des différents paramètres de ce modèle sont données dans les tableaux 2.3 et 2.5.

	$\sigma^{(s)}$	$a_1^{(s)}$	$b^{(s)}$
Régime 1 (s=1)	1.37 [0.058]	0.79 [0.015]	1.46 [0.105]
Régime 2 (s=2)	2.40 [0.089]	0.77 [0.021]	2.24 [0.224]

Tableau 2.3 Paramètres des modèles autorégressifs dans les différents régimes (modèle $MS-\gamma AR$ avec $M = 2$, mois de janvier). Les valeurs entre crochets correspondent à un écart type estimé à partir de la matrice d'information de Fischer observée.

On peut tout d'abord remarquer que l'écart type de la loi conditionnelle $\sigma^{(1)}$ du premier régime est nettement plus faible que celle du deuxième régime $\sigma^{(2)}$ et la volatilité est donc plus importante dans le deuxième régime que dans le premier. Le deuxième régime va donc permettent de décrire les conditions dépressionnaires. Or nous avons vu au paragraphe 2.b.1 que l'intensité du vent évolue plus rapidement dans les conditions dépressionnaires que dans les conditions anticycloniques, à cause du passage successif de perturbations. On peut alors proposer l'interprétation suivante pour les deux régimes :

- le premier régime décrit les conditions peu perturbées associées principalement aux conditions anticycloniques.
- le deuxième régime décrit les conditions perturbées associées principalement aux conditions dépressionnaires.

Nous proposons ci-dessous différent critères qui confortent cette interprétation. On peut tout d'abord remarquer que les pentes $a_1^{(i)}$ relativement proches dans les deux régimes, mais par contre que l'ordonnée à l'origine du deuxième régime, $b^{(2)}$, est nettement plus grande que celle du premier régime, et on s'attend alors à ce que le moyenne de la loi stationnaire du deuxième régime soit plus grande que celle du premier régime. En effet cette moyenne est donnée, pour $s \in \boldsymbol{S}$, par

$$m^{(s)} = \frac{b^{(s)}}{1 - a^{(s)}} \tag{2.9}$$

Les valeurs numériques sont données dans le tableau 2.4 : on obtient bien une moyenne significativement supérieure dans le deuxième régime. On retrouve donc que les vents sont généralement plus forts dans le régime perturbé. Nous donnons aussi, à titre indicatif, les écarts types $e^{(s)}$ de ces lois stationnaires dans le tableau 2.4. Ils sont donnés, pour $s \in \boldsymbol{S}$, par la formule:

$$e^{(s)} = \frac{\sigma^{(s)}}{\sqrt{1-(a^{(s)})^2}}$$

	Moyenne	écart type
Régime 1	7.11	2.26
Régime 2	9.84	3.77

Tableau 2.4 Moyenne et écart type de la loi stationnaire dans les différents régimes (modèle $MS-\gamma AR$ avec $M = 2$, mois de janvier).

La matrice de transition qui régit l'évolution de la variable cachée possède des valeurs assez fortes sur la diagonale (cf tableau 2.5), ce qui implique que les régimes ont de grandes durées de persistance moyenne, celles-ci étant données, pour $s \in \boldsymbol{S}$, par la formule :

$$d_s = \frac{1}{1 - q(s, s)}$$

En effet, on peut vérifier que les durées de séjour dans l'état s suivent une loi géométrique de paramètre $1 - q(s, s)$, la moyenne de cette loi étant donnée par la formule ci-dessus. Les résultats obtenus, ainsi que la loi stationnaire de la chaîne de Markov $\{S_t\}$ sont donnés dans le tableau 2.6. Le premier régime a une durée de persistance environ égale à 2 semaines et le second de l'ordre de 1 semaine. On trouve donc que les régimes peu perturbés durent en général plus longtemps que les régimes dépressionnaires : ceci est en accord avec la climatologie de la région considérée.

$q(i,j)$	$j = 1$	$j = 2$
$i = 1$	0.98 [0.0074]	0.02 [0.0074]
$i = 2$	0.03 [0.0103]	0.97 [0.0103]

Tableau 2.5 Matrice de transition de la chaîne de Markov cachée (modèle $MS-\gamma AR$ avec $M = 2$, mois de janvier).

	Durées de pers. moyenne	Loi stationnaire
Régime 1	13.5 jours	0.65
Régime 2	7.23 jours	0.35

Tableau 2.6 Durées de persistance moyenne et loi stationnaire de la chaîne de Markov cachée (modèle $MS-\gamma AR$ avec $M = 2$, mois de janvier).

Afin de vérifier l'interprétation des paramètres donnée ci-dessus, on peut aussi calculer la séquence cachée $\{\hat{s}_t\}_{t \in \{1...T\}}$ la "plus probable" connaissant les observations. Différentes notions de chaîne cachée la plus "probable" peuvent être utilisées, la plus courante consistant à prendre :

$$\{\hat{s}_t\}_{t \in \{1...T\}} = \arg max_{s_1^T \in S^T}(p_{\theta,\varsigma}(s_1^T, y_0^t))$$

Comme dans le cas des modèles CMC, cette séquence la plus probable peut être calculée rapidement en utilisant l'algorithme de Viterbi.

Un exemple de découpage obtenu avec cet algorithme est montré sur la figure 2.9. A titre de comparaison, les probabilités de lissage $p_{\theta,\varsigma}(s_t|y_0^t)$ sont représentées sur la figure 2.10 et on peut voir que les découpages obtenus avec cette deuxième méthode sont légèrement différents. Visuellement, ces découpages semblent généralement réalistes, les dates associées au premier régime correspondant effectivement à des périodes ou l'intensité du vent évolue lentement et celles associées au deuxième à des périodes ou l'intensité du vent évolue plus rapidement à cause du passage de perturbations successives.

.

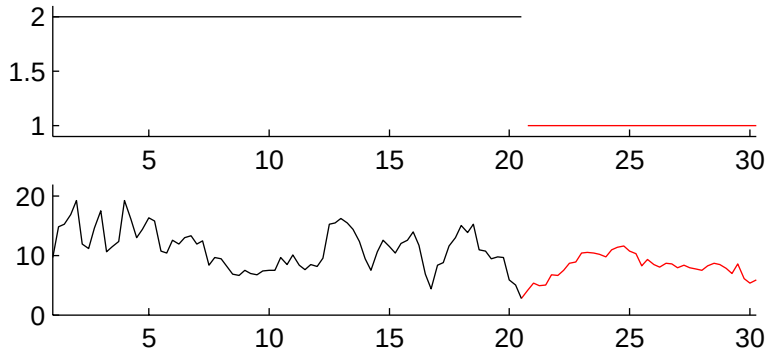


figure 2.9 : Intensité du vent au cours du mois de janvier 1991 (en bas) et type de temps le plus probable correspondant calculé avec l'algorithme de Viterbi (en haut). Le temps, en abscisse, est exprimé en jours (modèle $MS-\gamma AR$ avec $M = 2$).

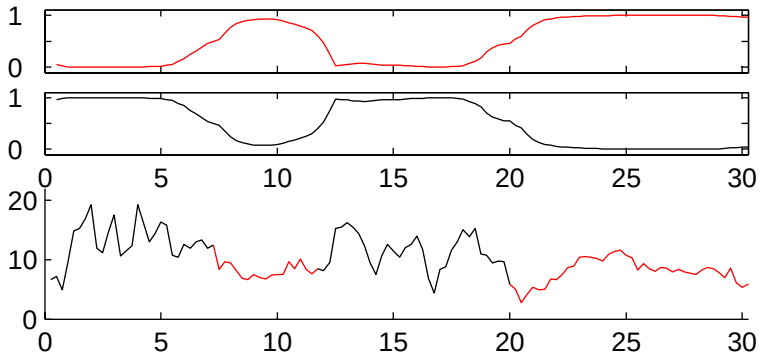


figure 2.10 : Intensité du vent au cours du mois de janvier 1991 (en bas) et probabilités de lissage $P(S_t = 1 | Y_1, \ldots Y_T)$. *(en haut) et* $P(S_t = 2 | Y_1, \ldots Y_T)$ *(au milieu). Le temps, en abscisse, est exprimé en jours (modèle* $MS - \gamma AR$ *avec* $M = 2$*).*

Enfin, en utilisant les découpages effectués ci-dessus, on peut tracer la répartition empirique de la direction du vent dans les différents régimes. Les résultats obtenus sont montrés sur la figure 2.11. La répartition de la direction du vent dans les deux régimes est bien distincte, ce qui confirme que les deux régimes identifiés correspondent bien à des régimes météorologiques bien distincts. On obtient que les conditions perturbées, décrites par le deuxième régime, sont principalement associées à des vents de secteur sud-ouest, et ceci est conforme à la climatologie de la région étudiée. Par contre, le premier régime est associé à toutes les orientations de vent.

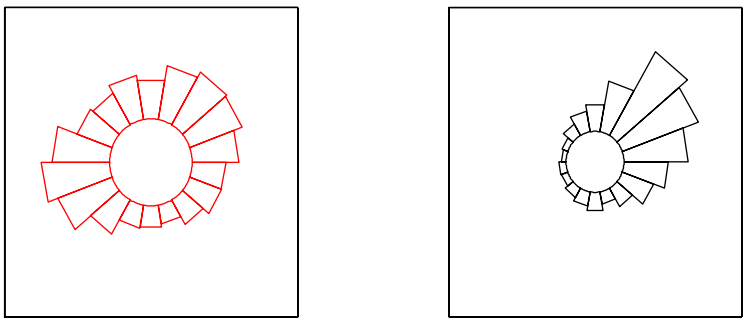


figure 2.11 : Répartition de la direction du vent dans le régime 1 (à gauche) et dans le régime 2 (à droite). (modèle $MS - \gamma AR$ *avec* $M = 2$*, mois de janvier).*

Modèle $MS - \gamma AR$ à trois régimes

Intéressons-nous maintenant au modèle sélectionné par le critère BIC, à savoir le modèle à trois régimes. Au vu des tableaux 2.7 et 2.8 et de la figure 2.13, on peut proposer l'interprétation suivante pour les différents régimes:

- le premier régime est caractérisé par une faible volatilité (cf tableau 2.7) et est principalement associé à des vents de faible intensité (cf tableau 2.8) et soufflant du nord-est (cf figure 2.13). Ce régime correspond donc principalement au type de temps "nord-est anticycloni-

que" décrit au paragraphe 2.b.1.

- le deuxième régime a une volatilité proche de celle du premier régime. Par contre, la moyenne de la loi stationnaire de ce régime montre qu'il est généralement associé à des vents de plus forte intensité, et sur la figure 2.13, on peut voir que ce régime est principalement associé à des vents soufflant du sud-ouest. Ce régime permet donc de décrire les conditions dépressionnaires faiblement perturbées.
- dans le troisième régime, l'intensité du vent varie rapidement et les vents soufflent principalement du secteur ouest-sud-ouest. Ce régime est associé aux conditions dépressionnaires fortement perturbées.

Un exemple de découpage obtenu avec l'algorithme de Viterbi est représenté sur la figure 2.12. On retrouve visuellement l'interprétation donnée ci-dessus.

	$\sigma^{(s)}$	$a_1^{(s)}$	$b^{(s)}$
Régime 1 (s=1)	1.24 [0.051]	0.71 [0.012]	1.55 [0.121]
Régime 2 (s=2)	1.44 [0.073]	0.77 [0.023]	2.73 [0.346]
Régime 3 (s=3)	2.84 [0.165]	0.74 [0.035]	2.18 [0.213]

Tableau 2.7 Paramètres des modèles autorégressifs dans les différents régimes (modèle $MS-\gamma AR$ avec $M = 3$, mois de janvier). Les valeurs entre crochets correspondent à un écart type estimé à partir de la matrice d'information de Fischer observée.

	Moyenne	écart type
Régime 1	5.31	1.76
Régime 2	11.79	2.26
Régime 3	8.41	4.22

Tableau 2.8 Moyenne et variance de la loi stationnaire dans les différents régimes (modèle $MS-\gamma AR$ avec $M = 3$, mois de janvier).

La matrice de transition a des valeurs moins élevées sur la diagonale que celle du modèle avec $M = 2$ (cf tableau 2.9), les régimes ont donc des durées de persistance moyennes plus faibles (cf tableau 2.10). On peut aussi remarquer que les probabilités de transition entre les régimes 1 et 3 sont très faibles.

$q(i,j)$	$j = 1$	$j = 2$	$j = 3$
$i = 1$	0.91 [0.026]	0.08 [0.050]	$1.3\ 10^{-6}$ [0.0325]
$i = 2$	0.12 [0.023]	0.77 [0.015]	0.11 [0.023]
$i = 3$	$9.4\ 10^{-6}$ [0.023]	0.21 [0.046]	0.79 [0.038]

Tableau 2.9 Matrice de transition de la chaîne de Markov cachée (modèle $MS-\gamma AR$ avec $M = 3$, mois de janvier).

	Durée de pers. moyennes	Loi stationnaire
Régime 1	2.85 jours	0.49
Régime 2	1.08 jours	0.34
Régime 3	1.17 jours	0.17

Tableau 2.10 Durées de persistance moyenne et loi stationnaire de la chaîne de Markov cachée (modèle $MS-\gamma AR$ avec $M = 3$, mois de janvier).

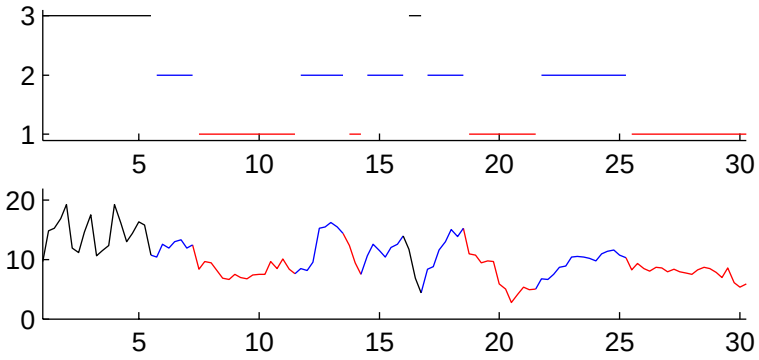


figure 2.12 : Intensité du vent au cours du mois de janvier 1991 (en bas) et type de temps le plus probable correspondant calculé avec l'algorithme de Viterbi (en haut) (modèle $MS-\gamma AR$ avec $M = 3$).

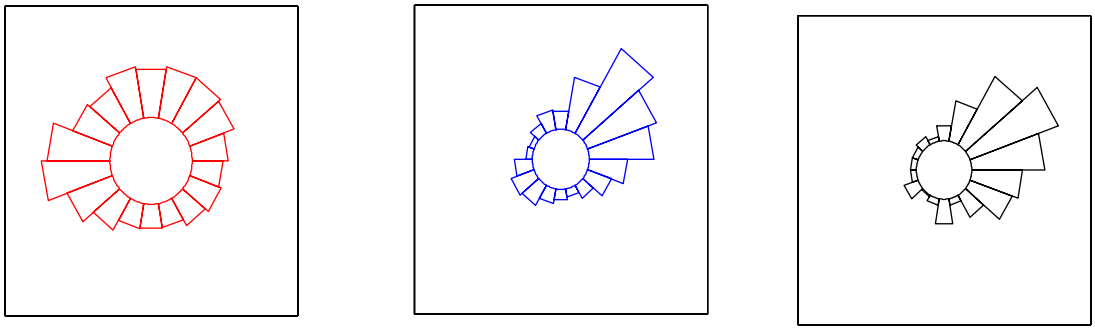


figure 2.13 : Répartition de la direction du vent dans les différents régimes du modèle $MS-\gamma AR$ avec $M = 3$: régime 1 à gauche, 2 au milieu et 3 à droite (mois de janvier).

2.b.4 Validation en simulation

Finalement, les modèles $MS-\gamma AR$ avec $M = 2$ et $M = 3$ semblent largement interprétables, les différents régimes correspondant à des types de temps bien distincts et physiquement réalistes. Afin de choisir entre ces deux modèles, nous allons utiliser la méthode de validation introduite au paragraphe 1.e.2. A titre de comparaison, les résultats obtenus avec la méthode TGP sont aussi donnés.

Afin de comparer ces trois modèles, les critères suivants ont été utilisés:

- F_U **:** fonction de répartition empirique de la loi marginale du processus $\{U_t\}$. Il s'agit sans doute d'un des critères principaux pour les applications, le comportement de la plupart des phénomènes considérés étant fortement conditionné par la répartition de l'intensité du vent.
- C_U: fonction d'autocorrélation empirique du processus $\{U_t\}$. Ce critère permet de vérifier si une certaine forme de dépendance temporelle est bien restituée dans les séquences simulées.
- F_{extr}: fonction de répartition empirique des maxima mensuels, ce qui permet de vérifier si la forte variabilité inter-annuelle du processus est bien restituée.
- $F_{U>2/3U_{max}}$: fonction de répartition empirique des durées de persistance des tempêtes. Nous définissons une tempête grâce à un seuil s_0 : nous appelons alors "tempête" un événement pendant lequel l'intensité du vent dépasse ce seuil s_0. La durée de persistance d'une tempête est alors définie comme le temps de séjour au dessus de ce seuil. En pratique, nous avons choisi pour seuil $s_0 = \frac{2}{3}U_{max}$, avec U_{max} la valeur maximale observée pendant les 22 années de mesure, et qui vaut environ $21 ms^{-1}$. Notons que cette définition de tempête n'est pas celle habituellement utilisée en météorologie : celle-ci correspond à un seuil $s_0 \approx 25 ms^{-1}$.
- $F_{U<2/3U_{max}}$: fonction de répartition empirique des durées d'inter-arrivée entre les tempêtes. Il s'agit de la fonction de répartition des temps de séjour en dessous du seuil $s = \frac{2}{3}U_{max}$. Ce critère permet de vérifier si la manière dont se succèdent les tempêtes est bien restituée dans les séquences simulées.

- $F_{U<1/3U_{max}}$: fonction de répartition empirique des durées des périodes de calme. Il s'agit de la fonction de répartition des temps de séjour en dessous du seuil $s = \frac{1}{3}U_{max}$. Ce critère peut être important pour certaines applications, tel que la programmation d'une opération offshore, par exemple, qui nécessite de longues périodes de calme.

Nous avons ensuite calculé, pour chacun des trois modèles en compétition, la valeur de la statistique de test correspondant à ces différents critères ainsi que la zone de rejet correspondante (cf paragraphe 1.e.2). Les résultats obtenus sont donnés dans le tableau 2.11.

	TGP	M=2	M=3
F_U	0.808 [0.012]	0.000 [0.012]	0.641 [0.024]
C_U	0.062 [0.012]	0.057 [0.009]	0.086 [0.022]
F_{extr}	0.068 [0.008]	0.010 [0.008]	0.371 [0.022]
$F_{U>2/3U_{max}}$	0.053 [0.012]	0.367 [0.032]	0.080 [0.016]
$F_{U<2/3U_{max}}$	0.002 [0.006]	0.284 [0.002]	0.053 [0.009]
$F_{U<1/3U_{max}}$	0.124 [0.031]	0.009 [0.031]	0.346 [0.004]

Tableau 2.11 Comparaison des modèles TGP et MS – γAR avec 2 et 3 régimes. La première valeur correspond à la valeur observée de la statistique de test et la deuxième, entre crochets, à la borne de la région critique pour un risque de première espèce $\alpha = 5\%$ *: l'hypothèse* H_0 *est acceptée si la première valeur est supérieure à la valeur entre crochets.*

Globalement, il semble que les différents modèles permettent de reproduire avec succès les différentes caractéristiques sélectionnées, à deux exceptions près.

Tout d'abord, le modèle TGP ne permet pas de reproduire la fonction de répartition des durées de persistance des temps de séjours en dessous du seuil $14ms^{-1}$ et qui peuvent être interprétées comme les durées séparant deux tempêtes successives. Afin d'analyser ce manque d'adéquation, nous avons représenté sur la figure 2.14 les fonctions de répartition empirique de ces durées de persistance, ainsi que celle correspondant au modèle. Le modèle TGP ne permet pas de reproduire la forme particulière de cette fonction de répartition et en particulier a tendance à simuler trop peu de faibles durées de persistance. Ceci semble indiquer que ce modèle ne permet pas de reproduire le fait que les perturbations arrivent généralement en groupe. En fait, il est fréquent que ce modèle ne permette pas de bien reproduire à la fois les durées de persistances des calmes ou des tempêtes. En effet, pour un processus gaussien stationnaire, centré et monovarié, il est facile de vérifier que les durées de persistance des séjours au dessus d'un seuil s_0 ont la même distribution que les durées de persistance en dessous du seuil $-s_0$. Les processus

gaussiens transformés vont alors avoir le même type de caractéristiques lorsque la transformation utilisée est monotone, ce qui est le cas pour la transformation des scores normaux. Lorsque le processus considéré ne possède pas cette propriété de symétrie, la méthode TGP ne permet alors pas de reproduire à la fois les durées de persistance des calmes et des tempêtes.

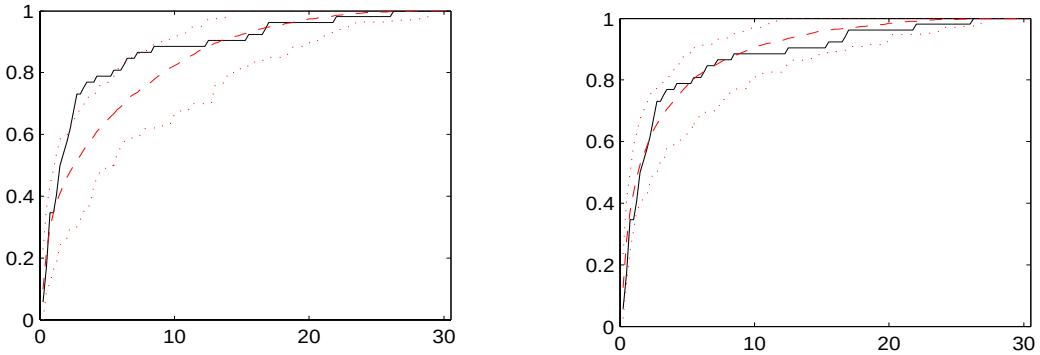


figure 2.14 : Comparaison des fonctions de répartitions des durées d'interarrivée (en jour) des tempêtes (seuil $14ms^{-1}$).—— données, - - séquences simulées, ···· intervalle de fluctuation à 95%. Modèle TGP à gauche et $MS-\gamma AR$ avec $M = 2$ à droite.

Les résultats obtenus avec le modèles $MS-\gamma AR$ à deux régimes sont nettement meilleurs en ce qui concerne ce critère. L'introduction de deux régimes, l'un servant à décrire les conditions peu perturbées et l'autre les conditions perturbées semble donc permettre de reproduire de manière plus réaliste la façon dont se succèdent les coups de vent (cf figure 2.14). Par contre ce modèle ne restitue pas correctement la fonction de répartition marginale du processus. Sur la figure 2.15, nous avons représenté la fonction de répartition empirique ainsi que les fonctions de répartition obtenues avec les modèles $MS-\gamma AR$ pour $M = 2$ et $M = 3$: le modèle à deux régimes a tendance à simuler trop de vent de faible intensité, par contre les résultats obtenus avec le modèle à trois régimes sont largement satisfaisants.

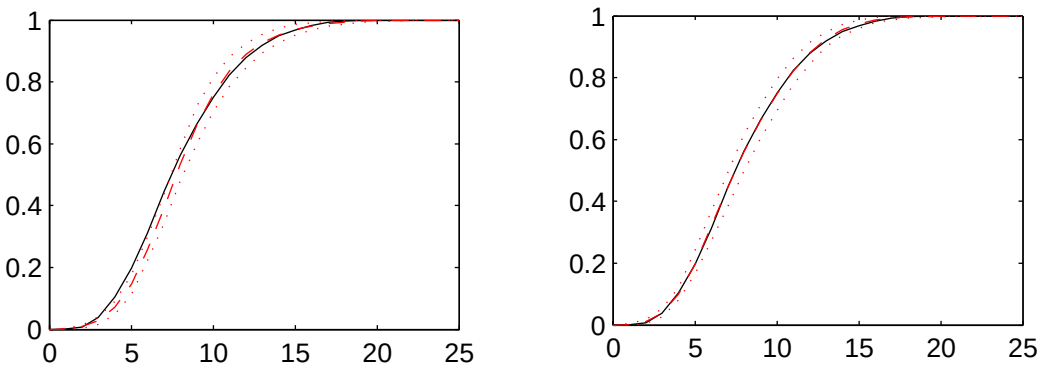


figure 2.15 : Comparaison des fonctions de répartitions de la loi marginale de U. —— données, - - modèle, ···· intervalle de fluctuation à 95%. Modèle $MS-\gamma AR$ avec $M = 2$ à gauche et $M = 3$ à droite.

Finalement, seul le modèle $MS-\gamma AR$ à trois régimes permet de reproduire les différents critères sélectionnés (cf tableau 2.11). On peut par ailleurs noter qu'il s'agit du modèle sélectionné par le critère BIC. Les figures correspondant à ce modèle, pour les différents critères sélectionnés, sont données ci-dessous. Elles permettent de vérifier la bonne adéquation entre les statistiques calculées à partir des données et des séquences simulées. En ce qui concerne les fonctions d'autocorrélation (cf figure 2.16), le pic observé sur la fonction d'autocorrélation empirique $\hat{\gamma}(h)$ pour $h \approx 10$ jours ne semble pas bien restitué. Toutefois cet écart n'est pas significatif au risque de première espèce 5% (cf tableau 2.11) et l'interprétation physique de ce pic n'est pas claire. Le modèle semble aussi permettre de restituer la forte variabilité interannuelle des données (cf figure 2.17), ainsi que les principales caractéristiques des calmes et des tempêtes (cf figure 2.17 et figure 2.18). On peut cependant noter, en comparant les figures 2.15 et 2.18 que les résultats obtenus avec ce modèle semble légèrement moins bons que ceux obtenus avec le modèle à deux régimes en ce qui concerne les durées de séjour au dessous du seuil $14ms^{-1}$.

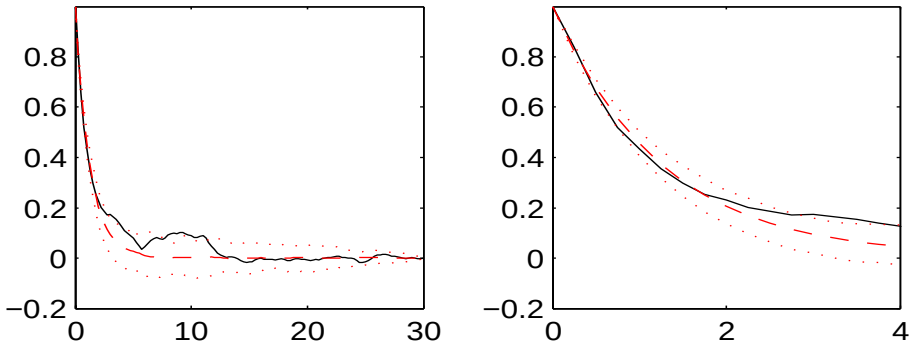


figure 2.16 : Fonction d'autocorrélation. Le temps est exprimé en jour. La figure de droite représente la fonction d'autocorrélation sur les 4 premiers jours uniquement. Modèle $MS-\gamma AR$ avec $M = 3$, —— données, - - modèle, ... intervalle de fluctuation à 95%.

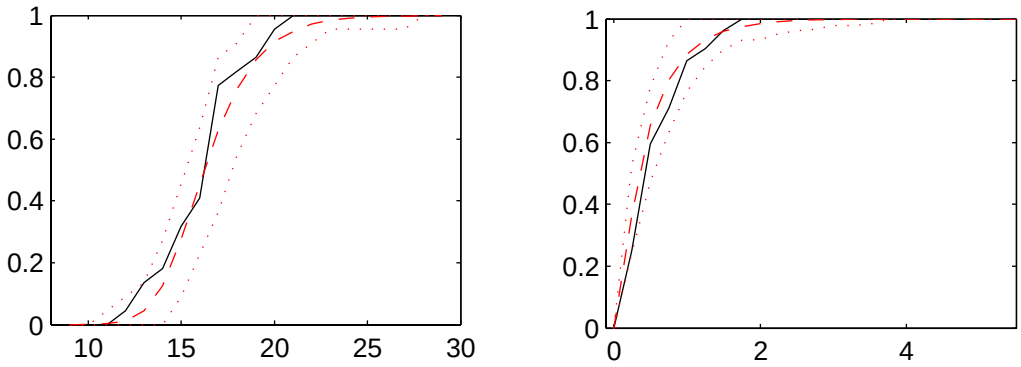


figure 2.17 : Fonction de répartition des maxima annuels à gauche et fonctions de répartitions des durées de persistance (en jour) des tempêtes (seuil $14ms^{-1}$) à droite. Modèle $MS-\gamma AR$ avec $M = 3$, —— données, - - modèle, ... intervalle de fluctuation à 95%.

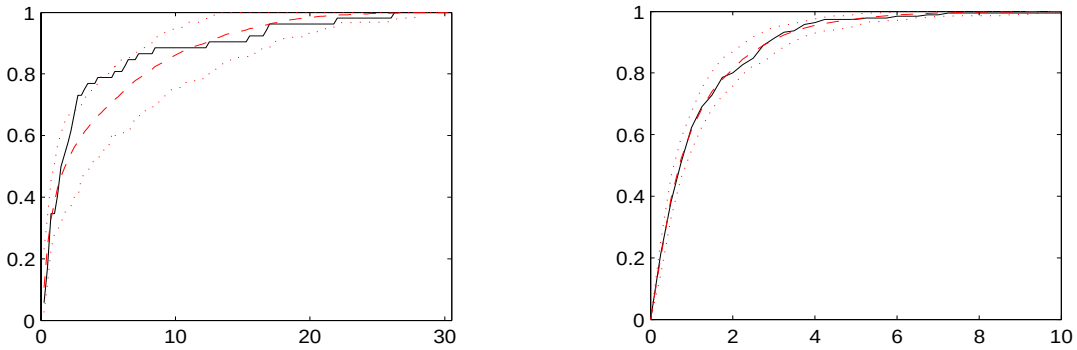


figure 2.18 : Comparaison des fonctions de répartitions des durées de persistance des "périodes de calme": seuil $14ms^{-1}$ *à gauche et* $7ms^{-1}$ *à droite. Modèle* $MS-\gamma AR$ *avec* $M = 3$, —— *données,* - - *modèle,* ···· *intervalle de fluctuation à 95%.*

Conclusion

Finalement, les modèles $MS-\gamma AR$ permettent de décrire les principales caractéristiques du processus U, et améliorent les résultats obtenus avec la méthode TGP. En particulier, nous avons montré que l'introduction de changements de régime permet de mieux décrire la manière dont se succèdent les périodes de calmes et les coups de vent, ce qui peut être important pour certaines applications (transports sédimentaires, fatigue de structure offshore...). Nous revenons sur la comparaison de ces deux modèles au paragraphe 2.c.1, dans le cadre de la simulation du processus bivarié $\{U_t, \Phi_t\}$. Une discussion plus générale est donnée dans ce paragraphe : nous y comparons en particulier les avantages respectifs des approches paramétriques et non paramétriques, les temps de calculs...

Par ailleurs, nous avons vu, sur cet exemple, que le critère BIC permet de sélectionner des modèles parcimonieux, dont les paramètres sont physiquement interprétables, et qui permettent de simuler des séries temporelles d'intensité du vent réalistes. Nous avons testé le modèle $MS-\gamma AR$ sur des données relatives à d'autres saisons et à d'autres régions, avec des climatologies différentes (cf *Ailliot et al.* (2003, [5])), et sur ces différents exemples, le critère BIC s'est avéré être un critère efficace pour effectuer une première sélection de modèle.

2.c Deux exemples d'utilisation de modèle NHMS-AR

Dans cette partie nous décrivons deux exemples d'utilisation des modèles autorégressifs à changements de régimes markoviens non-homogènes ($NHMS-AR$) pour les séries temporelles de vent.

Dans un premier temps, ce modèle est utilisé afin de décrire la relation existant entre les processus $\{U_t\}$ et $\{\Phi_t\}$. Pour cela, nous proposons une extension du modèle $MS-\gamma AR$ décrit au paragraphe 2.b. Dans ce modèle, la matrice de transition $Q^{(t)}$, qui régit l'évolution de la variable cachée à l'instant t, dépend de la direction du vent, l'évolution de l'intensité du vent dans chaque régime étant toujours décrite par un modèle γAR. Dans un deuxième temps, ce modèle est

utilisé afin de décrire les composantes journalières du processus $\{U_t\}$: la matrice de transition $Q^{(t)}$ est alors une fonction périodique de période un jour.

Pour chacun de ces modèles, nous reprenons les différentes parties du paragraphe 2.b. Après avoir décrit plus précisément le modèle en justifiant, si possible, physiquement la construction du modèle, nous décrivons rapidement comment nous avons calibré le modèle, puis nous vérifions que l'interprétabilité physique du modèle obtenu et enfin sa capacité à simuler des séquences réalistes.

2.c.1 Modélisation de la relation entre l'intensité et la direction du vent

Description du modèle

Au paragraphe 2.b, seule la modélisation du processus $\{U_t\}$ a été abordée. Cependant, dans certaines applications, l'évolution du phénomène étudié dépend aussi de la direction du vent, et la connaissance de son intensité n'est alors plus suffisante. Nous avons donc cherché à étendre les modèles décrits aux paragraphes précédents au processus bivarié $\{U_t, \Phi_t\}$.

Modélisation de la direction du vent

Intéressons nous tout d'abord uniquement au processus directionnel $\{\Phi_t\}$. Nous avons essayé de trouver un modèle paramétrique simple pour ce processus, mais ces tentatives se sont révélées infructueuses. Les principales difficultés que nous avons rencontrées sont décrites brièvement ci-dessous.

La première d'entre elles a déjà été mentionnée : il s'agit de la nature de ce paramètre. Les modèles utilisés classiquement pour les séries temporelles à valeurs réelles ne peuvent pas s'appliquer directement à ce processus. Plusieurs modèles spécifiques ont été proposés dans la littérature, et ceux-ci sont décrits au paragraphe 2.a. Plus précisément, à notre connaissance, principalement deux catégories de modèle ont été proposées pour ce type de série temporelle, à savoir les chaînes de Markov à espace d'état fini et différents modèles autorégressifs spécifiques. La première méthode utilise un grand nombre de paramètres et semble donc peu satisfaisante.

La deuxième difficulté est liée à la complexité des phénomènes physiques régissant l'évolution de ce paramètre. Ceci se traduit, par exemple, par la bimodalité de la loi marginale du processus (cf figure 2.3), chaque mode étant associé à un type de temps. Il est clair que les différents modèles autorégressifs pour variables circulaires décrits au paragraphe figure 2.a.3 ne permettent pas de reproduire cette bimodalité. Il semble alors naturel, une nouvelle fois, d'utiliser des modèles à changements de régimes. Tout d'abord, nous avons utilisé un modèle CMC. Ce type de modèle a été proposé dans *MacDonald et al.* (1997, [76]) pour décrire la direction du vent à Koeberg (Afrique du Sud). Dans cet ouvrage, les valeurs prises par le processus $\{\Phi_t\}$ sont discrétisées puis les probabilités d'émission $P(\Phi_t|S_t)$ sont paramétrées par des lois à support fini. Nous avons testé ce type de modèle en supposant que ces probabilités d'émission sont a support continu et suivent une loi de Von-Mises. Le modèle obtenu permet bien d'identifier les diffé-

rents modes de la distribution de Φ mais ne permet pas de restituer la forte dépendance existant entre deux valeurs successives. Cette limitation est aussi notée dans [76]. Afin de remédier à ce problème, nous avons ajusté un modèle $MS-AR$, l'évolution dans chaque régime étant décrite par un modèle autorégressif de Von-Mises (cf paragraphe 2.a.3). Les résultats obtenus avec ce modèle ne sont pas satisfaisants non plus, les mécanismes physiques régissant l'évolution de ce paramètre étant sans doute trop complexes pour être décrits par quelques régimes.

Ainsi, afin de décrire l'évolution de la direction du vent, Breckling (1989, [25]) propose de séparer l'évolution du processus $\{\Phi_t\}$ à différentes échelles de temps, avec une composante "géostrophique", que nous noterons $\{\Phi_t^{geos}\}$, qui est directement associé à la positon des grands centres d'actions (anticyclones et vastes zones dépressionnaires), et une composante résiduelle $\{\Phi_t^{res}\} = \{\Phi - \Phi_t^{geos}\}$ qui correspond au passage d'événements météorologiques de plus faible emprise spatio-temporelle (perturbations, composantes journalières...) et qui font fluctuer la direction du vent autour de la direction géostrophique. Dans [25], la composante géostrophique est identifiée en lissant la série initiale, et il est montré, à l'aide d'une analyse météorologique de la position des principales masses d'air à chaque date, que le processus ainsi obtenu correspond bien à une composante géostrophique physiquement réaliste. Différentes techniques de lissage, adaptées au processus circulaire, sont décrites dans [25]. Dans la suite, nous avons choisi Φ_t^{geos} de telle manière que

$$\rho \begin{bmatrix} \cos(\Phi_t^{geos}) \\ \sin(\Phi_t^{geos}) \end{bmatrix} = \sum_{d \geq k \geq -d} U_{t-k} \begin{bmatrix} \cos(\Phi_{t-k}) \\ \sin(\Phi_{t-k}) \end{bmatrix} \tag{2.10}$$

où ρ désigne une constante de normalisation. En pratique nous avons choisi $d = 4$ et Φ_t^{geos} représente alors la direction associée au vent moyen calculé sur un intervalle de 24 heures centré en t. Des exemples d'évolution de ce processus sont donnés sur la figure 2.19. Visuellement, on peut observer des paliers correspondant à des situations météorologiques où les centres d'action sont à des positions stables et des périodes pendant lesquelles la direction géostrophique évolue plus ou moins lentement, ce qui correspond aux périodes de transition pendant lesquelles la position des centres d'action évolue.Il semble difficile de trouver un modèle paramétrique simple permettant de reproduire ce type de comportement.

Finalement, des différentes méthodes testées pour simuler des réalisations du processus $\{\Phi_t\}$, seule la méthode LGB décrite au paragraphe 2.a.4. a permis d'obtenir des résultats satisfaisants. Les résultats obtenus avec cette méthode sont plus précisément décrits dans la suite de ce paragraphe.

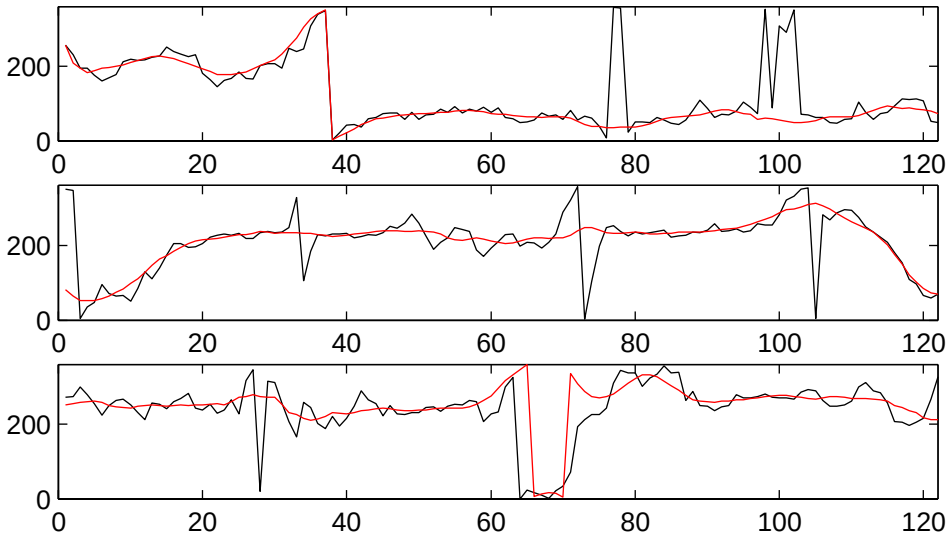


figure 2.19 : Exemples d'évolution des processus Φ (en noir) et Φ^{geos} (en rouge).

Modélisation de la relation entre l'intensité et la direction du vent

Intéressons-nous maintenant au processus bivarié $\{U_t, \Phi_t\}$. Une première approche consiste à utiliser les coordonnées cartésiennes $\{u_t, v_t\}$, ce qui permet de se ramener à une série temporelle à valeurs dans R^2. Cependant, la loi marginale de ce processus est complexe (cf figure 2.8: bimodalité, peu de valeurs proches de l'origine...) et l'interprétation physique de l'évolution de ces paramètres est plus délicate, et nous avons alors préféré travailler directement sur le processus $\{U_t, \Phi_t\}$.

Au paragraphe 2.b, nous avons utilisé un modèle $MS-\gamma AR$ pour décrire le processus $\{U_t\}$ et nous avons montré que les valeurs prises par la variable cachée sont interprétables et que celles-ci correspondent à la notion de "type de temps" couramment utilisée en météorologie. En particulier, nous avons mentionné qu'il existe une forte relation entre la direction du vent et les valeurs prises par la chaîne cachée : ainsi, par exemple, les régimes perturbés sont principalement associés à des vents de sud-ouest, et les régimes peu perturbés principalement à des vents de nord-est (cf figure 2.11 et figure 2.13). Afin de lier l'évolution des processus $\{U_t\}$ et $\{\Phi_t\}$, nous proposons alors d'utiliser la chaîne cachée $\{S_t\}$. Nous allons supposer qu'il s'agit d'une chaîne de Markov non-homogène dont la matrice de transition dépend de la direction du vent. Un modèle chaîne de Markov cachée non-homogène est proposée dans *Hugue et al.* (1999, [63]) afin de prévoir les précipitations locales à partir de variables météorologiques globales. Nous nous sommes inspirés de ce modèle pour construire le modèle décrit dans la suite de ce paragraphe.

Dans la suite de ce paragraphe, $\{\Psi_t\}$ représentera un processus circulaire, qui en pratique sera soit égal à $\{\Phi_t\}$ soit obtenu en appliquant une transformation sur ce processus. Nous allons supposer que $Y = U$ suit un modèle $MS-AR$ dans lequel

- $\{S_t\}$ est une chaîne de Markov non-homogène dont la matrice de transition à l'instant t, $Q_\theta^{(t)}$ dépend uniquement de Ψ_t. En pratique, nous avons utilisé la paramétrisation suivante:

$$q_\theta^{(t)}(i,j) = P(S_t = j \mid S_{t-1} = i) \sim p_{i,j} \exp(\kappa^{(j)} \cos(\Psi_t - \psi^{(j)})) \qquad (2.11)$$

avec $P = (p_{i,j})_{i,j \in S}$ une matrice stochastique, $(\kappa^{(j)})_{j \in S}$ des paramètres strictement positifs et $(\psi^{(j)})_{j \in S}$ des paramètres dans $[0, 2\pi[$. Une justification empirique du choix de cette paramétrisation est donnée ci-dessous.

- l'évolution du processus $\{Y_t\}$ dans chaque régime est décrite par un modèle γAR comme au paragraphe 2.b.

Dans la suite, ce modèle sera noté $NHMS - \gamma AR$. Dans ce modèle, le processus $\{\Psi_t\}$ est considéré comme une variable explicative et le modèle permet alors de décrire l'évolution du processus $\{Y_t\}$ en fonction de celle de ce processus. En particulier, ce modèle ne fait aucune hypothèse sur l'évolution $\{\Psi_t\}$. Lorsque les modèles autorégressifs sont d'ordre $r = 1$, la relation entre les différents processus $\{Y_t\}$, $\{\Psi_t\}$ et $\{S_t\}$ peut être résumée par le graphe d'indépendance conditionnelle représenté sur la figure 2.20..

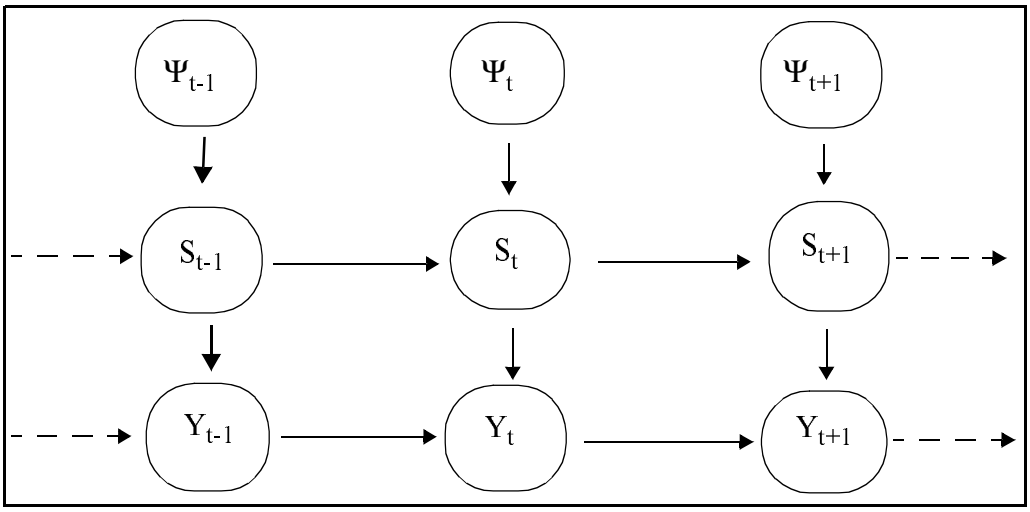


figure 2.20 : Graphe d'indépendance conditionnelle pour le modèle $NHMS - \gamma AR$ d'ordre 1

En particulier, pour $1 \le t \le T$, on a $P(Y_t \mid Y_1^{t-1}, S_1^t, \Psi_1^T) = P(Y_t \mid \bar{Y}_{t-1}, S_t)$ et $P(S_t \mid S_1^{t-1}, \Psi_1^T) = P(S_t \mid S_{t-1}, \Psi_t)$. La première relation implique que l'évolution du processus $\{Y_t\}$ est conditionnellement indépendante de celle de $\{\Psi_t\}$ connaissant $\{S_t\}$ et la deuxième que la variable cachée $\{S_t\}$ est une chaîne de Markov non-homogène. La paramétrisation du noyau de transition de la chaîne cachée est décrite par l'équation (2.11). Ce choix s'est fait en s'inspirant de la relation suivante, obtenue grâce à la formule de Bayes :

$P(S_t = j \mid S_{t-1} = i, \Psi_t = \psi) \sim P(\Psi_t = \psi \mid S_{t-1} = i, S_t = j) \times P(S_t = j \mid S_{t-1} = i)$

Si on fait l'hypothèse supplémentaire que $P(\Psi_t = \psi | S_{t-1} = i, S_t = j)$ suit une loi de Von-Mises de paramètres $(\psi_{i,j}, \kappa_{i,j})$ alors on obtient la paramétrisation

$$q_\theta^{(t)}(i,j) \sim p_{i,j} \exp(\kappa_{i,j} \cos(\Psi_t - \psi_{i,j}))$$

En fait, nous avons trouvé que le modèle plus parcimonieux décrit par l'équation (2.11), dans lequel $\psi_{i,j} = \psi^{(j)}$ et $\kappa_{i,j} = \kappa^{(j)}$ pour $i, j \in \boldsymbol{S}$, donne des résultats comparables, et nous avons donc choisi d'utiliser ce dernier modèle. Finalement, on a donc

$$q_\theta^{(t)}(i,j) = \frac{p_{i,j} \exp(\kappa_j \cos(\Psi_t - \psi_j))}{\sum_{k \in S} p_{i,k} \exp(\kappa_k \cos(\Psi_t - \psi_k))} \tag{2.12}$$

La matrice de transition est donc paramétrée par $\theta_s = ((p_{i,j})_{i,j \in S}, (\psi^{(j)})_{j \in S}, (\kappa^{(j)})_{j \in S})$, et ce modèle possède $2M$ paramètres supplémentaires par rapport au modèle $MS - \gamma AR$ homogène avec le même nombre de régimes.

Calibration

Afin d'estimer les différents paramètres du modèle, nous avons utilisé le même type d'algorithme que celui utilisé paragraphe 2.b, les différents algorithmes d'optimisation utilisés dans ce paragraphe (GEM et quasi-Newton) se généralisant sans problème pour ce modèle (cf paragraphe1.c). Par contre, les temps de calculs sont nettement plus longs. En effet, dans l'étape M de l'algorithme EM, nous n'avons pas trouvé d'expression analytique pour les paramètres θ_S régissant l'évolution de la variable cachée, et nous avons alors dû utiliser une procédure d'optimisation numérique. Pour l'algorithme quasi-Newton, les temps de calculs sont plus grands à cause du nombre plus élevé de paramètres. Afin de limiter les temps de calcul, il semble assez naturel d'utiliser les modèles homogènes afin d'obtenir une première estimation des paramètres, puis d'utiliser cette estimation comme valeur initiale des paramètres. Cependant, en pratique ce choix ne semble pas judicieux (convergence vers un extremum local non optimal). Nous avons alors choisi $N_{init1} = 30 \times M^2$ valeurs initiales de manière aléatoire, tous les paramètres étant choisis indépendamment en utilisant les lois suivantes:

- $a_i^{(s)}$ de loi uniforme sur [0, 1]
- $b^{(s)}$ de loi uniforme sur [0, 5]
- $\sigma^{(s)}$ de loi uniforme sur [1.5, 4]
- $P \sim I_M + R$ avec $R = (r_{i,j})$ des coefficients de loi uniforme sur [0, 1]
- $\kappa^{(s)}$ de loi uniforme sur [0.5, 5]
- $\psi^{(s)}$ de loi uniforme sur $[0, 2\pi]$

Différents processus $\{\Psi_t\}$ ont été testés en entrée du modèle. Tout d'abord, nous avons choisi $\Psi_t = \Phi_{t+k}$ avec $k \in \{-3, \ldots, 3\}$ afin de détecter des possibles décalages temporels existant entre ces deux processus, puis nous avons pris $\Psi_t = \Phi_t^{geos}$, avec Φ_t^{geos} la composante géostrophique définie au début de ce paragraphe. Cette variable pourrait en effet être plus adap-

tée pour prédire le type de temps, puisque plus directement lié à la situation météorologique globale (cf [25]). Afin de choisir le "meilleur" processus $\{\Psi_t\}$, ainsi que le nombre de régimes M, nous avons une nouvelle fois utilisé le critère BIC. Les résultats obtenus sont donnés dans le tableau 2.12. Ce critère sélectionne le modèle avec $M = 3$ et $\Psi_t = \Phi_{t+2}$, et les vérifications effectuées montre que ce modèle est en effet le "meilleur" en simulation.

M	2	3	4	5
$\Psi_t = \Phi_{t-3}$	9805	9778	9898	9952
$\Psi_t = \Phi_{t-2}$	9810	9731	9720	9803
$\Psi_t = \Phi_{t-1}$	9800	9732	9738	9801
$\Psi_t = \Phi_t$	9793	9740	9708	9783
$\Psi_t = \Phi_{t+1}$	9790	9697	9679	9771
$\Psi_t = \Phi_{t+2}$	9788	**9674**	9681	9777
$\Psi_t = \Phi_{t+3}$	9787	9716	9782	9797
$\Psi_t = \Phi_t^{geos}$	9796	9754	9803	9867

Tableau 2.12 Critère BIC pour le modèle $NHMS-\gamma AR$.

Interprétabilité du modèle

Comme dans le cas des modèles $MS-\gamma AR$, on peut vérifier que les différents régimes correspondent à des types de temps réalistes. Les valeurs de ces paramètres sont données dans les tableaux 2.13, 2.15 et 2.16. Les valeurs données entre crochets correspondent aux valeurs diagonales de l'inverse de la matrice d'information observée. A notre connaissance, aucun résultat théorique ne permet de vérifier si ces valeurs correspondent approximativement à la variance des estimateurs, comme c'était le cas dans les modèles $MS-\gamma AR$. Ces valeurs sont donc données à titre indicatif. Les paramètres régissant l'évolution de la variable cachée en fonction de la direction du vent, dont les estimations sont données dans les tableaux 2.15 et 2.16, sont difficilement interprétables directement. Sur la figure 2.21, nous avons alors tracé l'évolution des probabilités de transition $P(S_t = j|S_t = i, \Psi_t = \psi)$ en fonction de ψ et sur la figure 2.22 l'évolution de la probabilité invariante de la chaîne de Markov homogène de noyau de transition $Q(\psi) = (P(S_t = j|S_t = i, \Psi_t = \psi))_{i,j}$ en fonction de ψ.

A partir de ces différents éléments, on peut proposer l'interprétation suivante pour les différents régimes:

- le premier régime correspond à des vents dont l'intensité évolue lentement (type de temps peu perturbé) et il peut être associé avec des vents venant de toutes les directions.
- le deuxième régime est moyennement perturbé, et est généralement associé à des vents de faible intensité ou faiblissant. Il peut aussi être associé à des vents de différents secteurs. Cependant, quand les vents basculent au secteur ouest, la chaîne cachée a tendance à basculer vers le troisième régime.
- le troisième régime correspond à des vents perturbés, et est principalement associé à des vents de secteur ouest. Ainsi, la probabilité de rester dans l'état 3 alors que les vents sont de secteur est est très faible, et les probabilités de passer des régimes 1 ou 2 au régime 3 sont très faibles, excepté quand les vents soufflent de l'ouest.

On peut noter que les régimes identifiés sont nettement différents de ceux identifiés avec le modèle homogène, ce qui peut paraître surprenant. Afin de vérifier l'interprétation donnée ci-dessus, on peut calculer, comme dans le cas des modèles homogènes, la chaîne cachée la "plus probable" connaissant les observations, en utilisant par exemple l'algorithme de Viterbi qui se généralise sans problème au cas non-homogène. Un exemple de découpage obtenu est donné à la figure 2.23.

	$\sigma^{(s)}$	$a_1^{(s)}$	$b^{(s)}$
Régime 1 (s=1)	1.1654[0.0003]	0.8469[0.0002]	1.1745[0.0003]
Régime 2 (s=2)	1.5323[0.0095]	0.5640[0.0011]	1.8588[0.0457]
Régime 3 (s=3)	1.8534[0.0052]	0.7393[0.0012]	3.4436[0.1915]

Tableau 2.13 Paramètres des modèles autorégressifs dans les différents régimes (modèle NHMS – γAR, mois de janvier). Les valeurs entre crochets sont les termes diagonaux de l'inverse de la matrice d'information observée)

	Moyenne	écart type
Régime 1	7.67	2.22
Régime 2	4.36	1.95
Régime 3	14.00	2.52

Tableau 2.14 Moyenne et variance de la loi stationnaire dans les différents régimes (modèle NHMS – γAR, mois de janvier)

$q_{i,j}$	$j = 1$	$j = 2$	$j = 3$
$i = 1$	0.8643[0.0008]	0.1292[0.0008]	0.0065[0.0001]
$i = 2$	0.1828[0.0124]	0.6705[0.0057]	0.1467[0.0208]
$i = 3$	0.2970[0.0173]	0.1819[0.0288]	0.5211[0.0234]

Tableau 2.15 Matrice de transition (modèle NHMS – γAR, mois de janvier). Les valeurs entre crochets sont les termes diagonaux de l'inverse de la matrice d'information observée).

	$\psi^{(s)}$	$\kappa^{(s)}$
Régime 1 (s=1)	4.3215[0.3842]	0.5401[0.2324]
Régime 2 (s=2)	5.6934[1.0392]	0.4373[0.0385]
Régime 3 (s=3)	6.0873[0.0262]	2.1222[0.3022]

Tableau 2.16 Valeur des paramètres ψ_i et κ_i (modèle NHMS – γAR, mois de janvier). Les valeurs entre crochets sont les termes diagonaux de l'inverse de la matrice d'information observée).

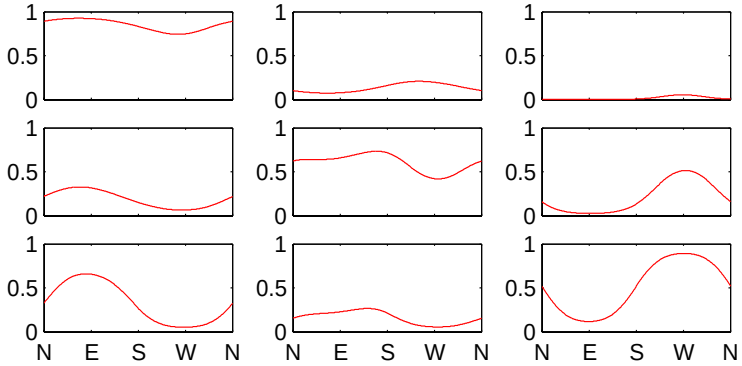


figure 2.21 : Evolution des probabilités de transition $P(S(t) = j|S(t-1) = i, \Phi(t) = \phi)$ en fonction de ϕ. La figure en haut et à gauche correspond à $i = j = 1$, en haut et au milieu à $i = 1$ et $j = 2$...(modèle NHMS – γAR, mois de janvier).

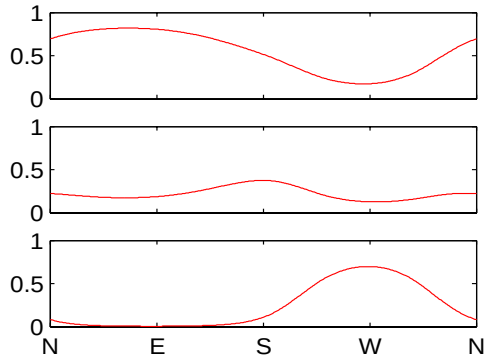


figure 2.22 : Evolution de la loi stationnaire de la chaîne de Markov de noyau de transition $Q(\psi)$ en fonction de ψ (modèle $NHMS-\gamma AR$, mois de janvier). Premier régime en haut, deuxième régime au milieu et troisième régime en bas.

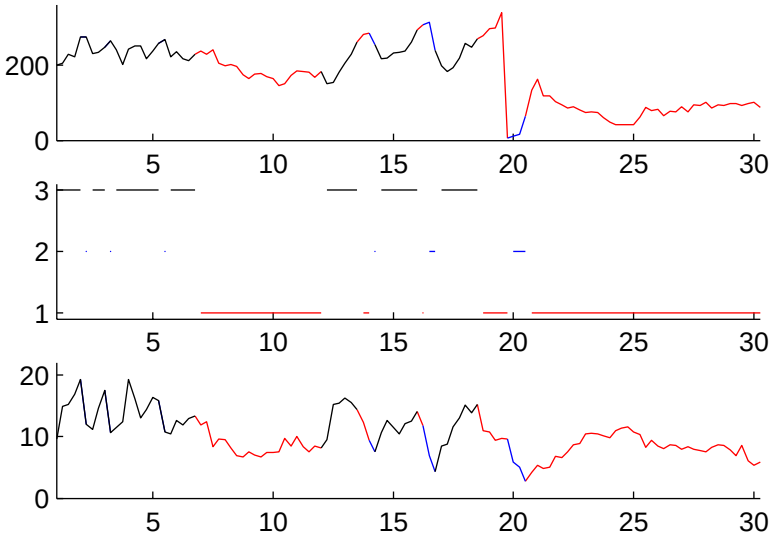


figure 2.23 : Intensité et direction du vent au cours du mois de janvier 1991 (en bas et en haut respectivement) et type de temps le plus probable correspondant calculé avec l'algorithme de Viterbi (au milieu) (modèle $NHMS-\gamma AR$, mois de janvier).

Validation du modèle en simulation

Dans la fin de ce paragraphe, nous proposons finalement de comparer les modèles $NHMS-\gamma AR$, TGP et LGB. Avant de nous focaliser sur les résultats obtenus avec la méthode de validation introduite au paragraphe 1.e.2, nous commençons par une discussion générale sur les inconvénients et avantages respectifs de ces trois méthodes. Une première différence notable provient de la nature des modèles, à savoir paramétrique ($NHMS-\gamma AR$) et non-paramétrique (TGP et LGB). Nous proposons tout d'abord un examen général des avantages respectifs de ces deux approches dans le cadre de la simulation de processus. Un avantage incontestable du modèle paramétrique est son interprétabilité, comme nous l'avons vu au début de ce paragraphe. Il peut ainsi être utilisé afin de mieux comprendre la climatologie de la région considérée. Au

contraire, le pouvoir explicatif des modèles non-paramétriques est inexistant. Un autre inconvénient de ce deuxième type d'approche est la difficulté liée à l'estimation des différents paramètres, puisque cette étape semble difficile à automatiser. Il faut alors choisir les différents paramètres "à la main", et ici cela a été fait de telle manière que les séquences simulées soient "réalistes" en fonction des critères choisis. Cette étape peut s'avérer fastidieuse, un grand nombre de tests pouvant s'avérer nécessaire avant d'obtenir des séquences simulées ayant les propriétés voulues.

Dans le cadre des modèles $NHMS-\gamma AR$, les paramètres sont estimés automatiquement en utilisant les estimateurs du maximum de vraisemblance. Notons toutefois que cette procédure peut s'avérer coûteuse en temps de calcul. Cependant, en ce qui concerne les modèles paramétriques, l'étape la plus fastidieuse est la construction du modèle, cette étape nécessitant en effet une étude précise du phénomène physique considéré. Par ailleurs, le modèle construit permet généralement de décrire uniquement ce phénomène précis, alors que les méthodes non-paramétriques sont plus générales et peuvent être utilisées sans modification pour d'autres types de processus.

En pratique, les temps de calcul en simulation des modèles $NHMS-\gamma AR$ et TGP sont comparables. Ils sont supérieurs pour le modèle LGB, puisque les noyaux de transition sont estimés localement pour chaque nouvelle valeur générée, alors que pour les deux autres modèles considérés l'estimation est faite au préalable. Pour que les temps de calculs de la méthode LGB demeurent raisonnables, une attention particulière doit alors être portée afin d'optimiser l'algorithme de simulation, notamment dans la recherche des plus proches voisins. Les résultats montrés ci-dessous ont été obtenus en utilisant l'algorithme développé dans *Monbet et al.* (2003, [87]). La recherche de plus proche voisin est effectuée en utilisant l'algorithme kd-tree (cf *Bentley* (1975, [13])).

Nous proposons maintenant de comparer ces différents modèles en utilisant la méthode décrite au paragraphe 1.e. Pour cela, les critères suivants ont été choisis:

- F_U, F_Φ et $F_{(U,\Phi)}$ les fonctions de répartition des lois marginales des processus $\{U_t\}$, $\{\Phi_t\}$ et $\{U_t, \Phi_t\}$ respectivement.
- C_u et C_v les fonctions d'autocorrélations des processus $\{u_t\}$ et $\{v_t\}$ respectivement. Nous avons choisi de regarder la structure d'ordre 2 du processus $\{u_t, v_t\}$ au lieu de celle de $\{U_t, \Phi_t\}$ directement. En effet, il faudrait alors prendre en compte la nature du processus $\{\Phi_t\}$. Diverses définitions de la fonction d'autocorrélation d'un processus circulaire peuvent être trouvées dans *Breckling* (1989, [25]), toutefois celles-ci sont relativement difficiles à interpréter.
- F_{extr}, $F_{U>2/3U_{max}}$, $F_{U<1/3U_{max}}$ et $F_{U<2/3U_{max}}$. Ces quantités sont définies au paragraphe 2.b.4.

Les résultats obtenus pour ces différents critères sont donnés dans le tableau 2.17 pour les trois modèles considérés. Ces résultats ont été obtenus en simulant l'équivalent de 1000 fois 22 mois de janvier. En ce qui concerne le modèle LGB, nous avons utilisé des modèles d'ordre

$r = 2$. Par ailleurs, afin de simuler des réalisations du processus $\{U_t, \Phi_t\}$ avec le modèle $NHMS - \gamma AR$, il faut fournir des séquences de direction simulées en entrée du modèle. Pour cela, nous avons utilisé les séquences fournies par l'algorithme LGB. Dans la suite nous utiliserons l'abréviation $LGB + NHMS - \gamma AR$ pour désigner les séquences simulées avec cette méthode.

	TGP	LGB	$LGB + NHMS - \gamma AR$
F_U	0.662 [0.012]	0.432 [0.012]	0.143 [0.004]
F_Φ	0.704 [0.008]	0.392 [0.004]	0.392 [0.004]
$F_{(U,\Phi)}$	0.542 [0.002]	0.094 [0.004]	0.124 [0.002]
C_u	0.000 [0.006]	0.068 [0.016]	0.234 [0.006]
C_v	0.000 [0.002]	0.000 [0.008]	0.000 [0.004]
F_{extr}	0.668 [0.014]	0.182 [0.012]	0.342 [0.004]
$F_{U>2/3U_{max}}$	0.031 [0.008]	0.182 [0.012]	0.292 [0.008]
$F_{U<2/3U_{max}}$	0.016 [0.006]	0.054 [0.006]	0.046 [0.009]
$F_{U<1/3U_{max}}$	0.000 [0.004]	0.118 [0.004]	0.502 [0.006]

Tableau 2.17 Comparaison des modèles TGP et LGB et $LGB + NHMS - \gamma AR$. La première valeur correspond à la valeur observée de la statistique de test et la deuxième à la borne de la région critique : l'hypothèse H_0 est acceptée avec un risque de première espèce $\alpha = 5\%$ si la première valeur est supérieure à la valeur entre crochets.

Le tableau 2.17 montre que la méthode TGP permet de bien restaurer la distribution marginale du processus bivarié. En fait, cette méthode est construite de telle manière que cette distribution marginale ainsi que la structure du second ordre coïncident. Il peut alors paraître surprenant que ce modèle ne permette pas de reproduire les fonctions d'autocovariance. Ce problème pourrait sans doute être résolu en utilisant une méthode d'estimation plus sophistiquée pour la structure du second ordre du processus transformé (cf 2.a.3). Cette méthode ne permet pas non plus de restaurer les durées de persistance des périodes de calme. Les résultats obtenus sont plus précisément décrits sur la figure 2.24, et le modèle semble simuler trop peu de longues durées de persistance. Ce problème est discuté plus précisément au paragraphe 2.b.4.

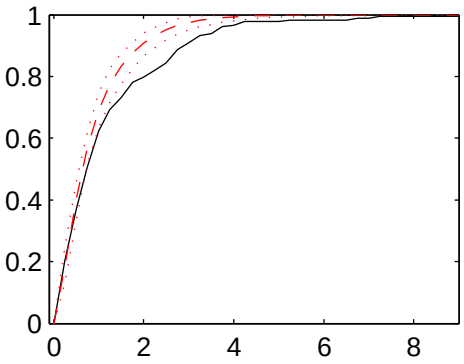


figure 2.24 : Comparaison des fonctions de répartitions des durées de persistance des "périodes de calme": seuil $14ms^{-1}$ *à gauche et* $7ms^{-1}$ *à droite. Modèle TGP,* —— *données,* - - *modèle,* ···· *intervalle de fluctuation à 95%.*

Intéressons nous maintenant à la méthode $LGB + NHMS - \gamma AR$. On peut tout d'abord noter que ce modèle permet de restaurer la loi marginale bivariée du processus, et la variable cachée semble donc à même de capter la relation complexe existant entre les processus $\{U_t\}$ et $\{\Phi_t\}$. Sur la figure 2.25, nous avons représenté la loi bivariée de la série observée ainsi que celle correspondant aux séquences simulées, et il semble que la forme générale de cette distribution soit bien restaurée (bimodalité, peu de valeurs près de l'origine...). Il est cependant difficile de comparer précisément les deux distributions à partir de cette figure, et nous avons alors tracé l'évolution des deux premiers moments de la loi conditionnelle $P(U_t|\Phi_t = \phi)$ avec ϕ sur la figure 2.26. On retrouve sur cette figure que les vents de secteur ouest sont généralement d'intensité plus forte et variable que ceux de secteur est, ce qui est bien reproduit par les séquences simulées.

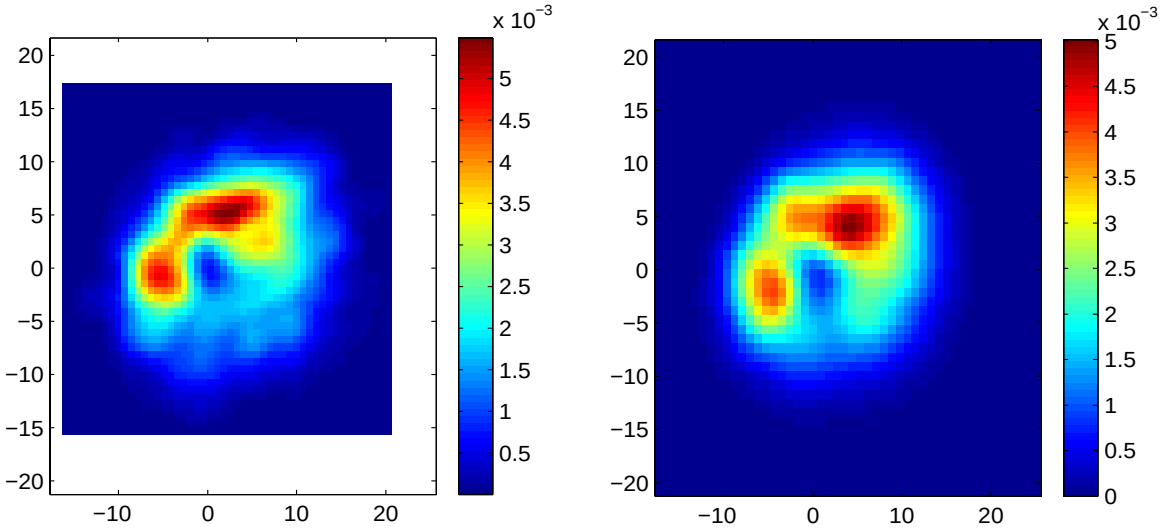


figure 2.25 : Densité empirique de (u, v). *Donnée à gauche, simulée à droite.*

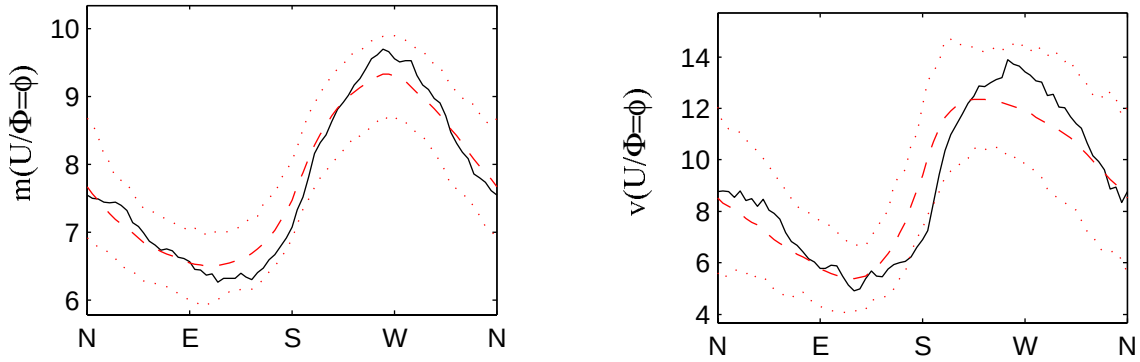


figure 2.26 : Moyenne (à gauche) et variance (à droite) de la loi conditionnelle $P(U_t | \Phi_t = \phi)$. —— *données,* – – *modèle,* ···· *intervalle de fluctuation à 95%.*

Ce modèle permet aussi de reproduire la fonction d'autocorrélation de la composante zonale u, mais pas celle de v. Afin d'interpréter ce manque d'adéquation, nous avons représenté ces fonctions d'autocorrélation sur la figure 2.27. Le modèle sous-estime significativement la forte corrélation existant à 3-4 jours pour la composante v. Nous avons par ailleurs trouvé que cette forte corrélation se retrouve sur plusieurs bases de données, mais son interprétation physique n'est pas claire. Enfin, comme dans le cas des modèles $MS-\gamma AR$ homogènes, les durées de persistance sont bien restituées.

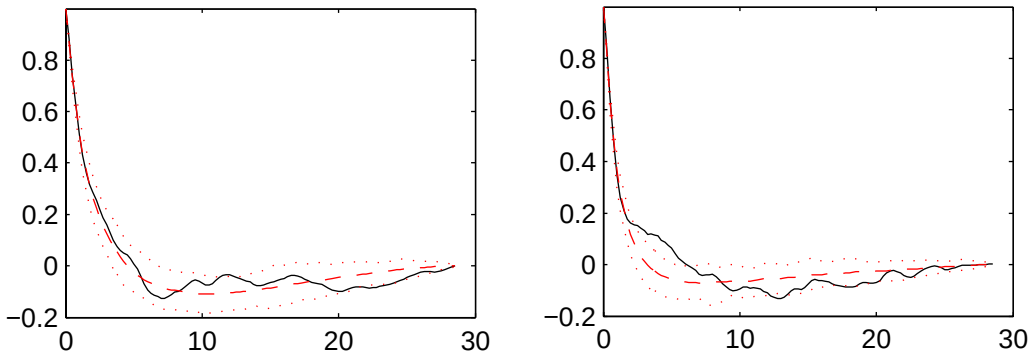


figure 2.27 : Fonction d'autocorrélation de u *(à gauche) et* v *(à droite).* —— *données,* – – *modèle,* ···· *intervalle de fluctuation à 95%.*

Les résultats obtenus avec l'algorithme LGB sont proches de ceux obtenus avec le modèle $NHMS-\gamma AR$, et les différentes caractéristiques choisies sont bien restituées, exceptée la fonction d'autocorrélation de la composante méridienne v. Il semble donc qu'un modèle markovien d'ordre 2 soit à même de reproduire les principales caractéristiques des données. On peut cependant noter que la qualité des simulations est très sensible aux choix des paramètres. De plus, cette méthode a tendance à reproduire des séquences observées dans les données, notamment dans les régions où il y a peu d'observations : ainsi, les tempêtes simulées sont généralement similaires à celles observées. Ce problème est plus précisément discuté dans *Monbet et al.* (2001, [87]). Même si cela semble difficile à quantifier, le modèle paramétrique parait plus adapté pour générer des "nouveaux" événements.

Conclusion

Finalement, l'algorithme TGP est l'algorithme le plus rapide et facile à mettre en oeuvre. Cette méthode permet principalement de bien modéliser la loi marginale du processus et sa structure du deuxième ordre. Cependant, il ne permet pas de reproduire certaines non-linéarités qui peuvent être présentes dans les données, et celles-ci peuvent être importantes pour certaines applications, notamment lorsque la manière dont se succède les événements (tempêtes, périodes de calmes...) est importante. Il est alors préférable d'utiliser l'une des deux autres méthodes (LGB ou $LGB + NHMS - \gamma AR$). En effet, les critères sélectionnés montrent que ces deux méthodes permettent de mieux reproduire la chronologie des événements.

2.c.2 Modélisation des composantes journalières

Dans ce paragraphe, nous décrivons plus brièvement un deuxième exemple d'utilisation des modèles $NHMS - AR$ pour les séries temporelles de vent.

Description du modèle

Les résultats montrés aux paragraphes 2.b ont été obtenus à partir des données relatives aux mois de janvier. Nous avons aussi testé les modèles $MS - \gamma AR$ sur des données relatives aux autres mois. Globalement, nous avons trouvé le même type de résultats pour les données relatives aux mois d'hiver, d'automne et de printemps.

Cependant, pour les mois d'été, le modèle sélectionné avec le critère BIC est le modèle avec $M = 4$ régimes (cf tableau 2.18). Ceci semble peu compatible avec la climatologie de la région, puisque en été principalement un régime climatologique est observé, à savoir le type de temps «nord-ouest anticyclonique». Dans ce type de temps, l'anticyclone des Açores est situé au nord de sa position moyenne et s'étend alors sur le golfe de Gascogne. L'intensité du vent est relativement faible et évolue généralement lentement (faible volatilité). Les vents sont généralement de secteur nord-ouest (cf figure 2.28).

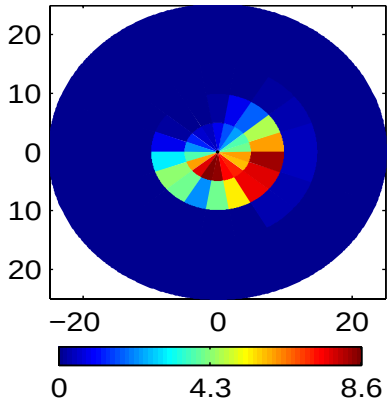


figure 2.28 : Densité de la loi marginale du processus (U, Φ) au mois de juillet: l'intensité et la direction du vent ont été discrétisées (classes de $5ms^{-1}$ et $20°$ respectivement), la couleur de chaque secteur représentant le pourcentage d'observation correspondante.

De plus, les différents régimes du modèle identifié sont peu interprétables, avec des durées

de persistance moyenne faibles (quelques heures) dans les différents régimes et les tests de comparaisons donnent des résultats nettement moins bons que ceux décrits au paragraphe 2.b.4. Ce manque d'adéquation est dû à l'existence de composantes journalières : les différents régimes identifiés permettent alors de décrire l'évolution du processus aux différents moments de la journée.

Au paragraphe 2.a.1, nous avons décrit comment l'existence de ces composantes se traduit sur les données, ainsi que la méthode généralement utilisée pour les modéliser. Dans ce paragraphe, nous proposons une autre méthode. Pour cela, nous allons utiliser un modèle $NHMS-\gamma AR$ dans lequel la matrice de transition de la chaîne de Markov cachée dépend de l'heure de la journée. Plus précisément, nous allons supposer que le processus $\{Y_t\}=\{U_t\}$ suit un modèle $MS-AR$ dans lequel

- $\{S_t\}$ est une chaîne de Markov non-homogène, dont la matrice de transition $Q_\theta^{(t)}$ est une fonction périodique de période un jour. En pratique, nous avons utilisé la paramétrisation suivante:

$$q_\theta^{(t)}(i,j) = P(S_t = j \,|\, S_{t-1} = i) \sim p_{i,j}\exp(\kappa^{(j)}\cos(\omega t + \Psi^{(j)})) \tag{2.13}$$

  avec $P = (p_{i,j})_{i,j\in S}$ une matrice stochastique, $(\kappa^{(j)})_{j\in S}$ des paramètres strictement positifs, $(\phi^{(j)})_{j\in S}$ des paramètres dans $[0, 2\pi[$ et $\omega = \pi/2$ (de telle manière que la périodicité de $Q_\theta^{(t)}$ soit de 1 jour).
- l'évolution du processus $\{Y_t\}$ dans chaque régime est décrit par un modèle γAR comme au paragraphe 2.b.

Ce type de modèle a déjà été proposé pour décrire des processus non-stationnaires par *Macdonald et al.* (1997, [76]). Ils proposent d'utiliser un modèle CMC non-homogène à deux régimes afin de décrire les composantes saisonnières et journalières de la direction du vent à Koeberg (Afrique du Sud). On peut vérifier que la paramétrisation utilisée dans [76] pour la matrice de transition $Q_\theta^{(t)}$ est un cas particulier de celle proposée ci-dessus lorsque $M = 2$.

Calibration

L'estimation des paramètres peut être effectuée en utilisant les mêmes algorithmes qu'au paragraphe 2.c.1. Le tableau 2.18 donne les valeurs du critère BIC pour le modèle $MS-\gamma AR$ (homogène) et $NHMS-\gamma AR$. On peut tout d'abord remarquer que les paramètres supplémentaires introduits dans le modèle $NHMS-\gamma AR$ sont largement justifiés au vu de ce critère. Le modèle sélectionné avec ce critère est le modèle avec cinq régimes et une nouvelle fois, il s'agit du modèle donnant les meilleurs résultats en simulation.

M	1	2	3	4	5	6
Homogène	10043	9816	9786	9781	9786	9797
Non homogène	10043	9631	9549	9451	**9436**	9521

Tableau 2.18 Critère BIC pour les modèles $MS-AR$ et $NH-MSAR$ (mois de juillet).

Nous avons réalisé la même expérience en utilisant les données relatives aux mois de janvier. On a trouvé que l'utilisation du modèle $NHMS-\gamma AR$ n'est pas justifiée au vu du critère BIC. En fait, le modèle identifié ne détecte pas de composantes journalières significatives et le modèle obtenu est alors proche d'un modèle $MS-\gamma AR$ (homogène), la matrice $Q^{(t)}$ étant quasiment constante.

Interprétabilité du modèle

Les paramètres de ce modèle sont difficiles à interpréter, en partie à cause du nombre important de paramètres. Dans le tableau 2.19, nous donnons les paramètres des modèles autorégressifs servant à décrire l'évolution de l'intensité du vent dans les différents régimes. Ensuite, afin de décrire de manière synthétique comment évolue la matrice de transition de la chaîne cachée aux différents moments de la journée, nous avons calculé les lois stationnaires des chaînes de Markov homogènes $\{S_{4k+t}\}_k$ $t \in \{0\ldots3\}$: ces quantités sont notées $\pi^{(t)}$. On peut aisément vérifier que ces chaînes de Markov homogènes admettent comme matrice de transition les matrices $Q^{(t+3)}Q^{(t+2)}Q^{(t+1)}Q^{(t)}$, ce qui permet le calcul des quantités $\pi^{(t)}$ pour $t \in \{0\ldots3\}$. Les résultats sont donnés dans le tableau 2.20. Il reste cependant relativement difficile d'interpréter les différents régimes.

	$\sigma^{(s)}$	$a_1^{(s)}$	$b^{(s)}$
Régime 1 (s=1)	1.35	0.62	2.99
Régime 2 (s=2)	1.33	0.39	1.13
Régime 3 (s=3)	0.74	0.54	2.81
Régime 4 (s=4)	1.08	0.91	0.86
Régime 5 (s=5)	1.01	0.80	0.21

Tableau 2.19 Paramètres des modèles autorégressifs dans les différents régimes (modèle NHMS-AR avec $M = 5$, mois de juillet).

$\pi_i(s)$	00:00h	06:00h	12:00h	18:00h
s=1	0.130	0.100	0.111	0.477
s=2	0.006	0.014	0.418	0.126
s=3	0.265	0.119	0.007	0.210
s=4	0.252	0.295	0.365	0.187
s=5	0.347	0.472	0.099	8.80e-05

Tableau 2.20 Loi stationnaire de la chaîne de Markov $\{S_t\}$ aux différents moments de la journée (modèle NHMS-AR avec $M = 5$, mois de juillet).

Validation du modèle en simulation

Finalement, afin de valider ce modèle, nous avons vérifié le réalisme des séquences simulées. Tout d'abord, sur la figure 2.29, nous avons tracé la fonction de répartition empirique de la loi marginale de la série temporelle observée ainsi que la fonction de répartition théorique correspondant aux modèle identifiés. On peut vérifier visuellement que la fonction de répartition observée est située dans l'intervalle de fluctuation à 95% obtenu pour le modèle, ce qui montre que l'écart entre les deux fonctions de répartition n'est pas significatif (si on réalise un test avec un risque de première espèce de 5%). Ensuite, afin de vérifier si ce modèle permet de restaurer les composantes journalières, nous avons comparé la moyenne et l'écart type de la répartition du vent aux différentes heures de la journée : les résultats obtenus sont donnés dans les tableaux 2.21 et 2.22 respectivement. On peut vérifier à nouveau que les valeurs observées sont situées dans un intervalle de fluctuation à 95% calculé à partir du modèle, et donc que les différences observées peuvent être expliquées par les erreurs d'échantillonnage. Finalement, nous avons comparé les structures d'ordre 2 (cf figure 2.30 pour la fonction d'autocorrélation et le figure 2.31 pour le périodogramme) et les résultats obtenus sont aussi satisfaisants.

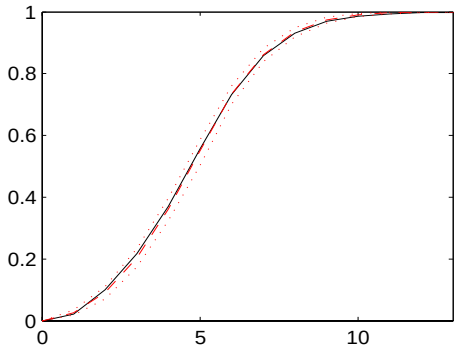


figure 2.29 : Comparaison des fonctions de répartition de la loi marginale de U (modèle $NHMS-AR$ avec $M = 5$, mois de juillet). —— données, - - modèle, ···· intervalle de fluctuation à 95%.

	0h	6h	12h	18h
Moyenne (séquences simulées)	5.02	4.87	4.26	4.98
Quantile empirique à 2.5%	4.84	4.67	4.06	4.77
Quantile à 97.5%	5.23	5.05	4.49	5.24
Moyenne (données)	5.03	4.85	4.29	4.93

Tableau 2.21 Moyenne de la loi marginale de l'intensité du vent aux différentes instants de la journée (modèle NHMS-AR avec $M = 5$, mois de juillet).

	0h	6h	12h	18h
Ecart type (séquences simulées)	1.91	1.92	2.15	2.18
Quantile empirique à 2.5%	1.80	1.81	1.99	2.06
Quantile empirique à 97.5%	2.05	2.09	2.28	2.32
Ecart type (données)	1.94	1.93	2.26	2.14

Tableau 2.22 Ecart type de la distribution marginale de l'intensité du vent aux différentes instants de la journée (modèle NHMS-AR avec $M = 5$, mois de juillet).

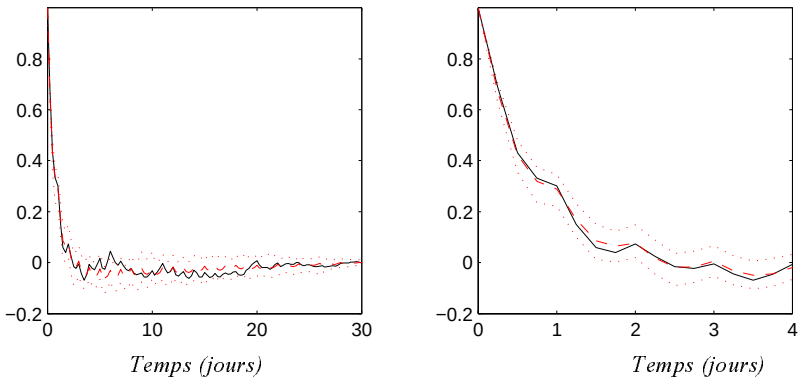


figure 2.30 : Fonction d'autocorrélation. Le temps est exprimé en jours. La figure de droite représente la fonction d'autocorrélation sur les 4 premiers jours. Modèle NHMS-AR avec $M = 5$, mois de juillet, —— données, – – modèle, ···· intervalle de fluctuation à 95%.

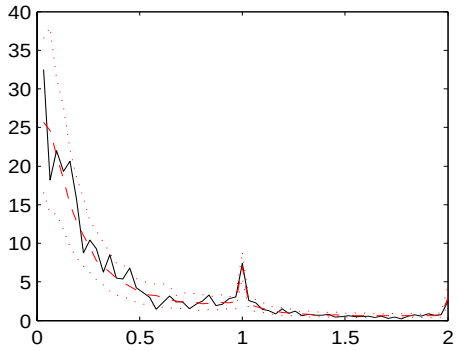


figure 2.31 : Périodogramme. Le temps est exprimé en jour^{-1}. La figure de droite représente la fonction d'autocorrélation sur les 4 premiers jours. Modèle NHMS-AR avec $M = 5$, mois de juillet, —— données, - - modèle, ···· intervalle de fluctuation à 95%.

Conclusion

Le modèle présenté ci-dessus permet donc de modéliser correctement, au vu des critères considérés, les composantes journalières qui sont présentes dans les séries temporelles de vent. Par contre, alors que les modèles identifiés dans les paragraphes 2.b et 2.c.1 sont largement interprétables, nous n'avons pas trouvé de signification météorologique simple pour les différents régimes de ce modèle. Il est aussi intéressant de noter que le critère BIC permet une nouvelle fois de sélectionner un "bon modèle", c'est à dire un modèle permettant de simuler des séquences réalistes. Les expériences effectuées montrent que les autres modèles donnent des résultats nettement moins satisfaisants.

Nous avons donc présenté deux exemples d'application des modèles $NHMS-AR$ aux séries temporelles de vent. Dans le premier exemple, la variable cachée permet de modéliser la relation entre deux processus (l'intensité et la direction du vent), alors que dans le deuxième exemple elle sert à modéliser des composantes non-stationnaires. D'autres applications pourraient être envisagées. Par exemple, on pourrait utiliser ce type de modèle pour décrire les composantes saisonnières, en supposant que la matrice de transition de la variable cachée est une fonction périodique de période un an, ou faire dépendre la matrice de transition de plusieurs facteurs simultanément, tels que la direction du vent et l'heure de la journée par exemple.

Ce type de modèle semble donc être un outil efficace permettant de décrire des phénomènes sujets à des changements de régimes, ceux-ci étant liés à une variable explicative. Les propriétés théoriques de ces modèles, non étudiées au cours de cette thèse, pourront faire l'objet d'études ultérieures.

3. Modèle spatio-temporel

Dans le chapitre précédent, nous nous sommes uniquement intéressés à la modélisation des séries temporelles d'état de mer en un point fixe. Cependant, lorsque l'on veut prévoir l'évolution de certains phénomènes, tel que la propagation d'une nappe de polluant par exemple, il est nécessaire de connaître les paramètres d'état de mer en plusieurs points simultanément. Il peut alors être intéressant de disposer d'un modèle spatio-temporel permettant de décrire l'évolution des champs de vent. Pour cela nous proposons, une nouvelle fois, d'utiliser un modèle $MS-AR$. Ce choix est justifié dans le paragraphe 3.a.

Dans le modèle proposé, la variable cachée est introduite pour décrire le déplacement des structures météorologiques. En pratique, cette variable va prendre un nombre important de valeurs (de l'ordre de quelques centaines). De plus, le processus observé, à savoir le champ de vent, est un processus multivarié dont le nombre de composantes peut aussi être important, selon la zone d'étude choisie. Afin d'obtenir un modèle avec un nombre raisonnable de paramètres, il est alors nécessaire d'utiliser des formes paramétriques parcimonieuses pour décrire l'évolution de la chaîne de Markov cachée ainsi que celle du processus observé dans les différents régimes. La paramétrisation choisie est plus précisément décrite au paragraphe 3.b.

Enfin, au paragraphe 3.c, nous validons le modèle obtenu de deux manières différentes. Nous montrons tout d'abord qu'il améliore nettement les résultats obtenus avec les modèles autorégressifs linéaires en prédiction à court terme, puis nous vérifions sa capacité à simuler des champs de vents réalistes en utilisant la méthode décrite au paragraphe 1.e.2.

3.a Introduction

Alors que de nombreux auteurs se sont intéressés à la modélisation des séries temporelles de paramètre d'états de mer en un point fixe, la modélisation de l'évolution spatio-temporelle de ces paramètres a été peu abordée. Les modèles proposés dans la littérature peuvent être regroupés en deux grandes catégories:

- les approches "***lagrangiennes***", dans laquelle on va suivre le déplacement des entités météorologiques (tempêtes, zones de calmes...). Chaque entité est alors vue comme un événement dont il faut modéliser l'évolution dans l'espace et le temps.
- les approches "***eulériennes***", dans laquelle on cherche à modéliser directement l'évolution du processus en un ensemble de points fixés.

Nous avons choisi d'utiliser une approche eulérienne. Avant de nous focaliser sur cette approche, nous justifions brièvement ce choix au paragraphe 3.a.1.

Ensuite, au paragraphe 3.a.2, nous décrivons plus précisément la zone d'étude : nous avons choisi un domaine situé dans le golfe de Gascogne et pour lequel nous disposons des données en 35 points distincts. Le champ de vent sur ce domaine, à un instant donné, est alors décrit par un vecteur de taille 70, les conditions de vent en chaque point étant décrites par un vecteur bivarié.

L'approche eulérienne "classique" consiste à ajuster un modèle autorégressif linéaire à ce

processus multivarié. Généralement, une analyse en composantes principales est effectuée au préalable, ce qui permet de réduire le nombre de composantes du processus initial et donc le nombre de paramètres du modèle. Nous mettons en évidence, au paragraphe 3.a.3, deux inconvénients de ce type de modèle, à savoir le manque d'interprétabilité des paramètres et son incapacité à reproduire le déplacement des structures météorologiques. Nous proposons alors d'utiliser un modèle $MS-LAR$ avec innovations gaussiennes dans lequel la variable cachée représente le déplacement des masses d'air.

3.a.1 Approches lagrangiennes

Les approches lagrangiennes ont été utilisées pour simuler différents types de phénomènes météorologiques. Ainsi, *Casson et al.* (1998, [30]) s'intéressent aux champs de vent dans les cyclones tropicaux, *Boukhanovsky et al.* (2003, [22]) à la hauteur significative dans les tempêtes en mer de Barens et dans [96] ce sont les champs de vent dans les tempêtes en Mer du Nord qui sont considérés. Ce type d'approche a aussi été utilisé pour simuler les précipitations dans les cellules pluvieuses (*Willems* (2001), [122]).

Afin de mettre en oeuvre ce type de méthode, il faut tout d'abord définir les structures météorologiques d'intérêt. Usuellement, en météorologie, les structures sont définies à partir des champs de pression. On distingue généralement six types de structures météorologiques (cf *Mayencon* (1982), [80]):

- ***les anticyclones:*** zones de hautes pressions, généralement de grande taille et pouvant stationner au même endroit pendant plusieurs jours. Le vent tourne autour du centre dans le sens des aiguilles d'une montre dans l'hémisphère nord.
- ***les dépressions:*** zones de basses pressions. On distingue généralement deux types de dépressions, à savoir les ***perturbations*** qui sont des minima dépressionnaires de petite taille et mobiles (elles peuvent se déplacer à une vitesse pouvant atteindre 150 kmh^{-1}) et les vastes zones dépressionnaires dont la position est approximativement stationnaire. Dans ces structures le vent tourne autour du centre dans le sens contraire des aiguilles d'une montre dans l'hémisphère nord.
- ***les cols:*** zone séparant deux dépressions, associés généralement à des vents faibles.
- ***les talwegs*:** excroissance d'une dépression, les isobares s'emboîtent les uns dans les autres en forme de V. Généralement, le champ de vent est discontinu au voisinage de ces talwegs (intensité et direction). Ils sont généralement associés aux fronts dans les dépressions.
- ***les dorsales*:** excroissance d'un anticyclone, les isobares s'emboîtent les uns dans les autres en forme de U. Les vents y sont généralement faibles.
- ***les marais barométriques:*** région où les isobares sont espacés et désorganisés. Elles sont généralement associées à des vents faibles et de direction variable.

La figure 3.1 représente un exemple de carte isobarique sur laquelle ces différentes structures sont présentes.

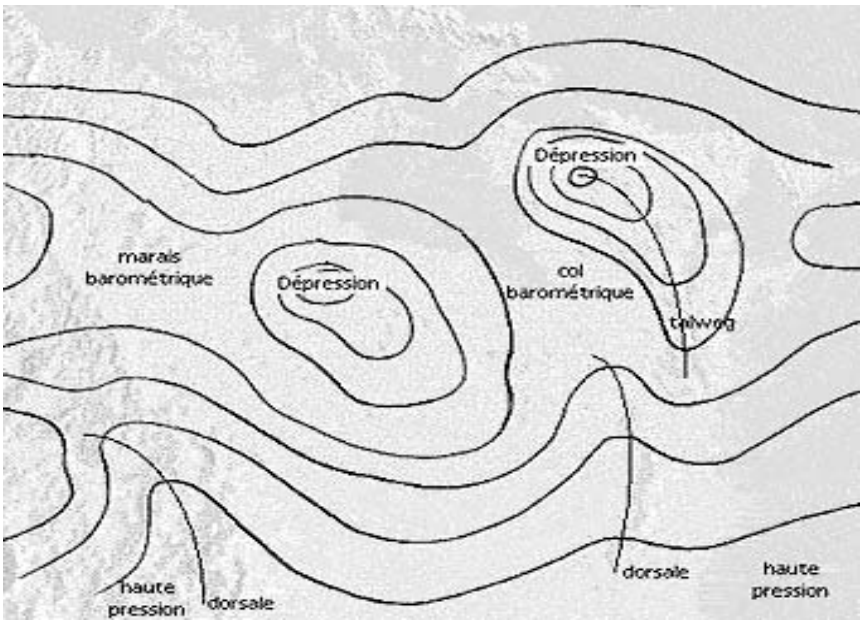


figure 3.1 : Exemple de situation météorologique où les principales structures météorologiques sont présentes. Les lignes noires représentent les isobares (lignes de niveau du champ de pression).

Dans les différents articles mentionnés ci-dessus, seules les dépressions sont considérées. Ceci permet, par exemple, de calculer des périodes de retour pour les vents extrêmes puisque les vents les plus forts sont observés dans ce type de structure météorologique. Cependant, pour certaines applications, les vents de faible ou moyenne intensité sont également importants et il faudrait alors considérer les autres types de structures.

Il faut ensuite trouver une méthode permettant de détecter ces structures dans la base de données disponible. Dans [30] et [96] les tempêtes sont identifiées à la main par un météorologue. Celui-ci fournit alors une nouvelle base de données dans laquelle sont répertoriées les différentes tempêtes ainsi que leurs évolutions au cours du temps. Nous avons cherché à automatiser cette phase d'identification. Dans [22] une tempête est définie comme un ensemble de points pour lequel la hauteur significative des vagues dépasse un seuil donné. Ce type de définition ne convient pas pour les données de vent puisque la répartition de l'intensité du vent dans une perturbation est généralement complexe, avec, par exemple, des vents faibles près du centre (cf figure 3.2).

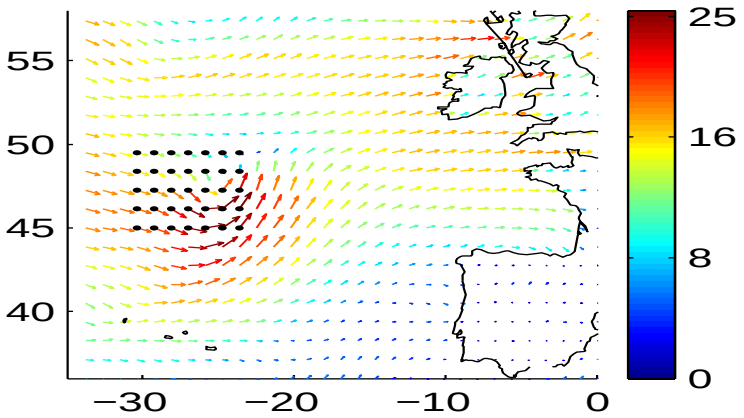


figure 3.2 : Situation du 4 Janvier 1998 à midi. Une dépression est centrée approximativement sur le point de coordonnée 25°O 48°N.

Il faut alors utiliser des méthodes plus sophistiquées. Nous avons testé les techniques de détection automatique de points singuliers dans les champs de vecteur développées par *Corpetti* (2002, [34]) sur les champs de vent ECMWF. Des exemples de résultats peuvent être trouvés dans [34] (page 188). Cette méthode semble fonctionner correctement et le centre des principales structures est généralement bien identifié. Il reste cependant plusieurs problèmes à régler. Ainsi, cette méthode a tendance à détecter un grand nombre de structures, certaines sans interprétation météorologique claire, et il faudrait trouver un moyen permettant de sélectionner uniquement celles qui sont "intéressantes". Par ailleurs, cette méthode ne permet pas de suivre les structures entre 2 instants successifs.

Une fois les structures d'intérêt identifiées, la deuxième étape consiste à trouver un modèle paramétrique permettant de décrire leur apparition dans le domaine d'étude puis leur évolution spatio-temporelle. Pour cela, une étape préliminaire consiste à trouver des formes paramétriques simples permettant de décrire de manière réaliste la répartition des champs de vent dans les structures météorologiques considérées. La méthode la plus usuelle consiste à paramétrer les champs de pressions puis à utiliser ensuite la formule du "vent géostrophique" (3.1) afin de calculer les champs de vents correspondants:

$$W = -\frac{1}{\rho f}(\overrightarrow{grad(P)} \wedge \vec{k}) \tag{3.1}$$

avec W le vecteur vent, P le champ de pressions, ρ la masse spécifique de l'air, f le paramètre de Coriolis et $\vec{k}$ le vecteur unitaire vertical. Avec cette formule le vent obtenu est donc de direction orthogonale au gradient de pression et d'intensité proportionnelle à ce gradient. Cette formule est approximative mais donne tout de même une bonne idée de la forme des champs de vent dans les structures (cf *Mayencon* (1982, [80])). Dans *Breckling* (1989) ([25]), les champs de pression sont choisis de la forme suivante:

$$P(x) = p_0 + \frac{\sum_{j=1}^{J} \frac{d_j}{d(x, c_j)^2}(p_j - p_0)}{1 + \sum_{j=1}^{J} \frac{d_j}{d(x, c_j)^2}}$$

avec $(c_1, \ldots, c_J)$ les positions des centres des J structures météorologiques influant sur la pression au point x, $p_0 \approx 1013 hp$ la pression atmosphérique normale et d_j des paramètres qui dépendent de la masse d'air. En pratique, d_j représente la distance jusqu'à laquelle la structure j a une influence et les valeurs suivantes sont choisies dans [25]: $c_j \approx 1000 km$ si la j^{eme} structure est un anticyclone et $c_j \approx 300 km$ si la j^{eme} structure est une dépression. Des paramétrisations basées sur le même principe peuvent être trouvées dans *Casson et al.* (1998, [30]) pour les cyclones tropicaux. Malheureusement, ces modèles semblent largement trop simples pour décrire de manière réaliste les perturbations de l'Atlantique Nord. En effet, la forme des champs de vent dans ces perturbations est généralement complexe comme on peut le voir sur l'exemple de la figure 3.2 (fortes dissymétries, présence de discontinuités dues aux fronts...).

Dans [96], différents types d'analyses en composantes principales sont testés afin de décrire la forme des champs de vent dans les perturbations. Cependant, cette méthode ne permet pas d'obtenir des paramétrisations parcimonieuses, le nombre de composantes principales à retenir pour obtenir une description suffisamment fine des champs de vent dans les perturbations étant relativement important. Finalement, il semble donc difficile de trouver des modèles paramétriques simples permettant de décrire ces structures. Ceci provient de la complexité et de la forte variabilité de la forme des champs de vent dans ce type de structures.

Il semble donc difficile de trouver un modèle paramétrique permettant de décrire le passage des dépressions dans le golfe de Gascogne. De plus, un tel modèle ne serait pas totalement satisfaisant pour les applications envisagées, puisqu'il ne permettrait pas de décrire, par exemple les périodes de calme, et plus généralement ne semble pas adapter pour simuler les champs de vent sur une grille spatio-temporelle fixe. Dans la suite, nous avons donc choisi d'utiliser une approche eulérienne.

3.a.2 Zone d'étude et notations

Dans cette thèse, nous avons choisi de nous intéresser à une zone comprise entre les latitudes 45°N et 50°N et les longitudes 23°O et 31°O, soit une zone d'environ 600km par 600km. Ce choix est arbitraire puisque nous disposons des données sur tout le globe (cf annexe A). En fait, nous avons choisi cette zone d'une taille "raisonnable" au vu des applications envisagées et de l'échelle des phénomènes météorologiques considérés et relativement éloignée de la côte afin d'éviter les problèmes liés à la déformation des champs de vent au-dessus de la terre.

La zone considérée contient $N = 35$ points de la grille du modèle ECMWF. $R_0 = (r_1^0, \ldots, r_N^0)$ désignera l'ensemble de ces points. Nous noterons (x_i^0, y_i^0) les coordonnées

du point r_i^0. R_1 désignera l'ensemble des points de la grille ECMWF compris entre les latitudes 36° N et 58° N et les longitudes 35° O et 0° O (cf figure 3.3).

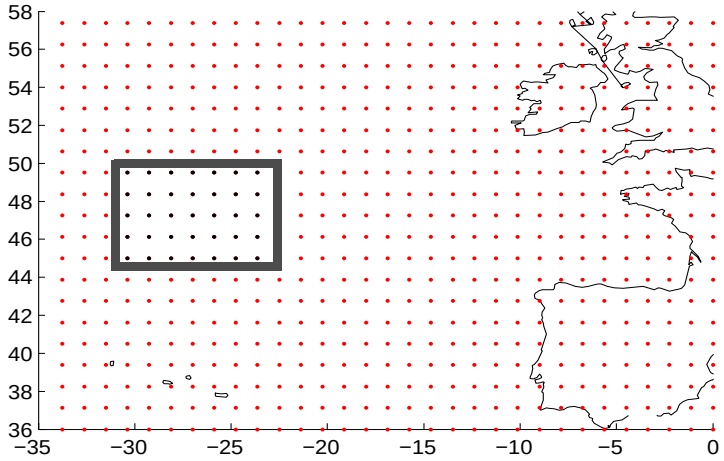


figure 3.3 : Grille du modèle ECMWF. La zone d'étude choisie R_0 est l'ensemble des points situés à l'intérieur du rectangle rouge.

Plus généralement, si $R = (r_1, \ldots, r_n)$ un ensemble de points quelconque de la grille ECMWF, nous utiliserons l'ordre suivant pour numéroter les sites : si (x_i, y_i) représente les coordonnées du point r_i, alors $r_i < r_j$ si $x_i < x_j$ ou $x_i = x_j$ et $y_i < y_j$. Avec cette numérotation, le point r_0^1 est donc situé au sud-ouest de la zone R_0, r_0^5 est au nord-ouest, r_0^{31} au sud-est...Si $t \in \boldsymbol{Z}$, nous noterons $Z_t(R)$ la variable aléatoire $(u_t(r_1), u_t(r_2), \ldots, u_t(r_n), v_t(r_1), \ldots v_t(r_n))$ où $u_t(r_i)$ et $v_t(r_i)$ désignent respectivement les composantes zonale et méridienne du vent au point r_i et à la date t. Nous noterons aussi $u_t(R) = (u_t(r_1), u_t(r_2), \ldots, u_t(r_n))$ et $v_t(R) = (u(r_1, t), \ldots, u(r_n, t))$. Enfin, nous noterons $Y_t = Z_t(R_0)$.

3.a.3 Approches eulériennes

L'objectif est donc de trouver un modèle stochastique décrivant l'évolution du processus $\{Y_t\}$, qui est à valeurs dans $\boldsymbol{R}^{2N}$. Comme au chapitre précédent, nous supposerons que ce processus est stationnaire mois par mois et nous allons nous intéresser uniquement au mois de Janvier. Nous supposerons que les 10 mois de Janvier disponibles dans la base de données ECMWF sont des réalisations indépendantes d'un même processus stationnaire. Sur la figure 3.4, la moyenne empirique du processus $\{Z_t(R_1)\}$ est représentée. Les différences observées entre les différents sites sont dues à la position moyenne des grands centres d'action qui influencent la climatologie de la région à cette époque de l'année. Au sud de la zone, on a généralement des conditions anticycloniques associées à l'anticyclone des Açores: les vents sont généralement faibles et enroulés dans le sens des aiguilles d'une montre. Par contre, au nord de la zone, on a généralement une vaste zone dépressionnaire et donc des vents généralement plus forts. En ce qui concerne la zone d'étude R_0 (entourée par un rectangle sue la figure 3.4) la moyenne semble relativement homogène.

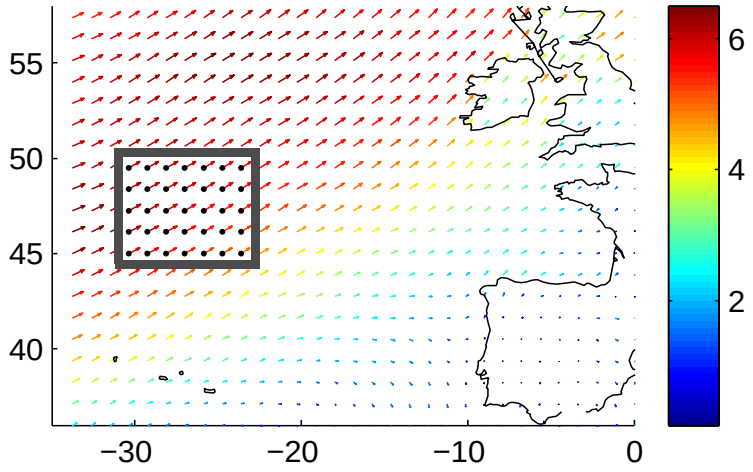


figure 3.4 : Moyenne empirique du champ $\{Z_t(R_1)\}$ (mois de janvier). Les points correspondants à la zone d'étude R_0 sont ceux à l'intérieur du rectangle.

D'après l'hypothèse de stationnarité temporelle, la structure d'ordre 2 du processus $\{Y_t\}$ est caractérisée par sa fonction d'autocovariance Σ définie, pour $h \in \boldsymbol{Z}$, par

$$\Sigma(h) = cov(Y_t, Y_{t+h}) = E[Y_t Y_{t+h}'] - E[Y_t]E[Y_{t+h}]' \tag{3.2}$$

Ces matrices ont la structure par bloc suivante :

$$\Sigma(h) = \begin{bmatrix} \Sigma_{(u,u)}(h) & \Sigma_{(u,v)}(h) \\ \Sigma_{(u,v)}(h)' & \Sigma_{(v,v)}(h) \end{bmatrix}$$

avec $\Sigma_{(u,u)}(h) = cov(u_t(R_0), u_{t+h}(R_0))$, $\Sigma_{(u,v)}(h) = cov(u_t(R_0), v_{t+h}(R_0))$ et $\Sigma_{(v,v)}(h) = cov(v_t(R_0), v_{t+h}(R_0))$. Pour $h \in Z$, nous noterons ρ la fonction d'autocorrélation. Celle-ci est définie, pour $h \in \boldsymbol{Z}$, par $\rho(h) = (\rho_{i,j}(h))_{i,j \in \{1\ldots70\}}$ avec

$$\rho_{ij}(h) = \Sigma_{ij}(h)/(\sqrt{\Sigma_{ii}(0)\Sigma_{jj}(0)}) \text{ pour } i,j \in \{1\ldots2N\}$$

Sur la figure 3.5, les estimations usuelles $\hat{\Sigma}(0)$ et $\hat{\rho}(1)$ de $\Sigma(0)$ et $\rho(1)$ respectivement sont représentées. On retrouve tout d'abord sur ces figures la structure par bloc mentionnée ci-dessus. On peut aussi remarquer que les valeurs correspondant aux points situés au nord-ouest de la zone semblent significativement plus fortes que celles correspondant à ceux situés au sud-est, et il semble donc abusif de supposer que le processus est spatialement stationnaire.

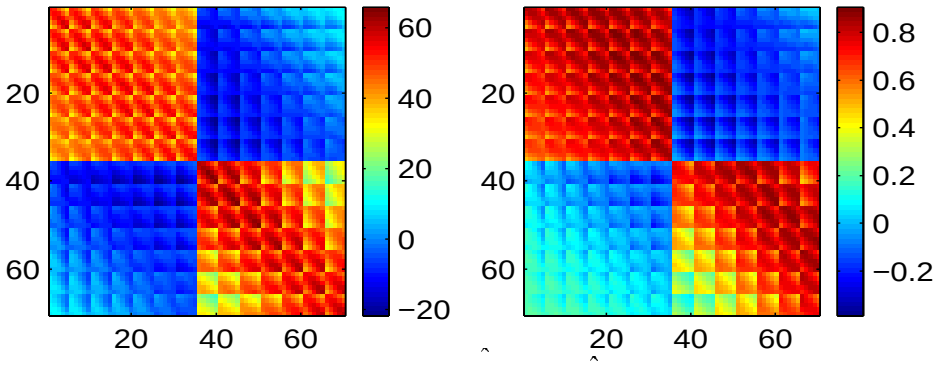


figure 3.5 : Matrice $\hat{\Sigma}(0)$ *et* $\hat{\rho}(1)$ *.*

Sur la matrice $\hat{\Sigma}(0)$, les plus fortes valeurs sont observées sur la diagonale de la matrice, et la corrélation entre les différents points décroît ensuite logiquement avec la distance entre ces points. On peut cependant remarquer que la corrélation entre les points, même éloignés les uns des autres, reste relativement forte. On s'attend alors à ce qu'une analyse en composantes principales (***ACP***) permette de réduire efficacement la dimension de l'espace des observations. Les vecteurs principaux de l'ACP vont alors correspondre aux vecteurs propres de la matrice $\Sigma(0)$. En météorologie, ces vecteurs sont généralement appelés "***Empirical Orthogonal Functions***" (**EOF**). Sur les données ci-dessus, les 5 premiers axes principaux expliquent environ 95% de la variance totale (cf tableau 3.1). L'ACP permet donc de réduire efficacement la dimension de l'espace des observations puisqu'on peut alors décrire approximativement les champs de vent à une date donnée par les composantes principales associées aux quelques axes principaux les plus significatifs. Une approche naturelle, pour modéliser l'évolution spatio-temporelle des champs de vent, consiste alors à ajuster un modèle à la série temporelle correspondant aux composantes principales sélectionnées. Ainsi, dans *Boukhanovsky et al.* (2003, [21]), un modèle autorégressif linéaire multivarié avec innovations gaussiennes est utilisé.

Nombre de composantes	1	2	3	4	5
Pourc. de var. expliquée	43	40	7	3	3
Pourc. cumulé de var. expliquée	43	83	90	93	96

Tableau 3.1 Pourcentage de variance expliquée par les 5 premières composantes principales.

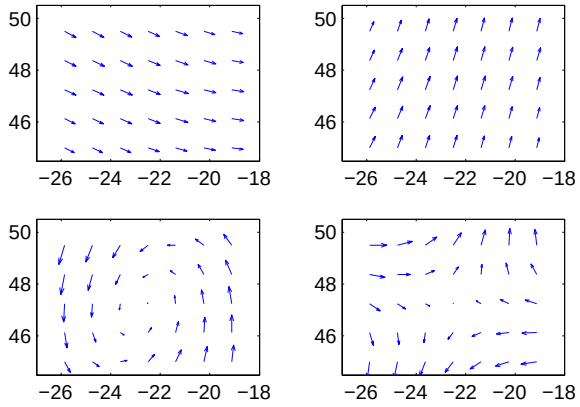


figure 3.6 : Les 4 premiers axes principaux de l'ACP. Le premier est représenté en haut à gauche, le deuxième en haut à droite...

Sur la figure 3.6, nous avons représenté les quatre premiers axes principaux. L'interprétation physique de ces axes n'est pas claire. Ainsi, par exemple, lorsqu'une perturbation passe à proximité de la zone R_0, les champs de vent sont généralement complexes et la manière dont évolue les composantes principales est alors difficile à interpréter. Différentes variantes de l'ACP traditionnelle ont été proposées afin de renforcer l'interprétabilité de la décomposition (cf *Jolliffe et al.* (2002), [66]). Certaines de ces méthodes ont été testées sans succès sur nos données et il nous a alors semblé préférable de chercher à modéliser directement l'évolution du processus $\{Y_t\}$.

Nous avons déjà mentionné que les structures météorologiques, et notamment les perturbations, peuvent se déplacer entre deux instants successifs. Ceci peut se voir de différentes manières sur nos données. Ainsi, la matrice $\hat{\rho}(1)$ (cf figure 3.5) est nettement dissymétrique avec des valeurs plus élevées au dessus de la diagonale. Ceci est dû au fait que les structures météorologiques se déplacent généralement de l'ouest vers l'est. Ce comportement est aussi illustré par l'exemple de la figure 3.7.

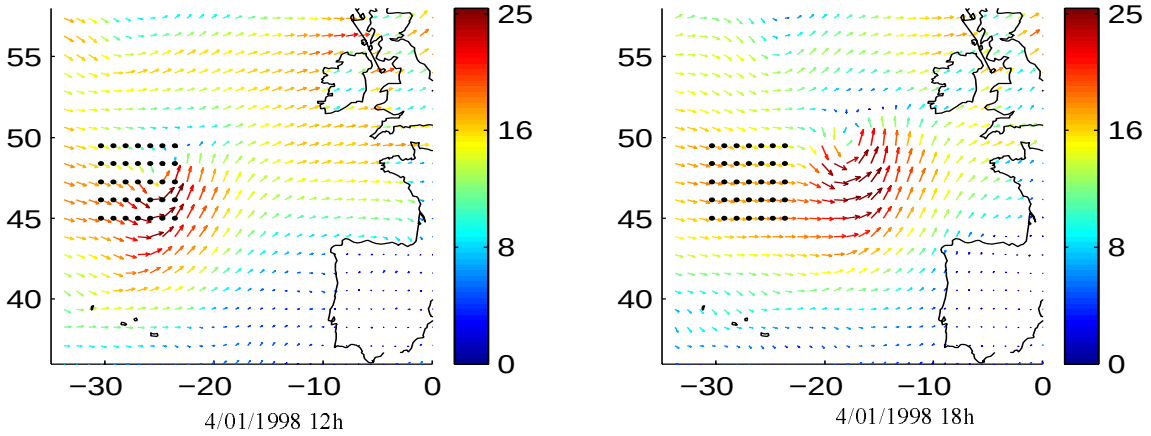


figure 3.7 : Situation du 4 Janvier 1998: la dépression située sur la zone S_0 (points en rouge) à 12 heure se retrouve, légèrement déformée, sur une zone située plus à l'est à 18 heure.

Afin de modéliser l'évolution du processus $\{Y_t\}$ nous avons alors choisi d'introduire ce déplacement sous la forme d'une variable cachée. Plus précisément, nous proposons d'utiliser un modèle $MS-LAR$ avec innovations gaussiennes, c'est à dire de la forme:

$$Y_t = A^{(S_t)} Y_{t-1} + B^{(S_t)} + H^{(S_t)} E_t \tag{3.3}$$

avec $A^{(s)} \in M_d(\boldsymbol{R})$, $B^{(s)} \in M_{d,1}(\boldsymbol{R})$ et $\Sigma^{(s)} = H^{(s)}(H^{(s)})' \in S_d^+(\boldsymbol{R})$ pour $s \in \boldsymbol{S}$, $\{E_t\}$ un bruit blanc gaussien et S_t une chaîne de Markov homogène qui représente le déplacement entre les instants $t-1$ et t de la structure météorologique présente sur la zone R_0 à l'instant $t-1$. Nous supposerons par ailleurs que le processus non observé, $\{S_t\}$, est à valeurs dans un sous ensemble fini $\boldsymbol{S} = \{a_1, \ldots, a_M\}$ de $\boldsymbol{Z}^2$ correspondant à des déplacements sur la grille ECMWF avec une vitesse inférieure à 150 kmh^{-1} (cf *Mayencon* (1982, [80]), p 77).

Le cardinal de $\boldsymbol{S}$ est relativement important (de l'ordre de 300) et le processus observé est à valeurs dans $\boldsymbol{R}^{70}$. Afin d'obtenir un modèle parcimonieux, il faut alors trouver des formes paramétriques simples pour décrire l'évolution du processus observé dans les différents régimes ainsi que celle de la variable cachée. C'est l'objet du paragraphe suivant.

3.b Paramétrisation du modèle

Pour paramétrer efficacement l'évolution du processus observé dans les différents régimes, il faut avoir une idée de la manière dont évoluent les champs de vent conditionnellement au déplacement des masses d'air. Pour cela, nous avons estimé ces déplacements, en utilisant le fait que nous connaissons les champs de vent sur une zone plus large que R_0. La technique utilisée est plus précisément décrite au paragraphe 3.b.1.

Ensuite, au paragraphe 3.b.2, nous nous servons de ces valeurs estimées afin de choisir des formes paramétriques parcimonieuses pour les matrices $A^{(s)}$, $B^{(s)}$ et $\Sigma^{(s)}$ ainsi que pour le noyau de transition de la variable non observée. Ces valeurs sont aussi utilisées afin d'obtenir une première estimation des paramètres du modèle. Dans un deuxième temps, ces paramètres sont réestimés via l'algorithme EM. Cette étape est plus précisément décrite au paragraphe 3.b.3.

3.b.1 Calcul du déplacement des masses d'air

Dans ce paragraphe nous décrivons une méthode permettant de calculer une suite $\{\hat{s}_t\}_{t \in \{1 \ldots T\}}$ d'éléments de $\boldsymbol{S}$ qui représente approximativement le déplacement des champs de vent entre deux instants successifs.

Par définition de $\hat{s}_t$, le champ de vent observé sur la zone R_0 à l'instant $t-1$ doit alors être proche du champ observé sur la zone $R_0 + \hat{s}_t$ à l'instant t, c'est à dire qu'on doit avoir $Z_{t-1}(R_0) \approx Z_t(R_0 + \hat{s}_t)$. Il semble alors naturel de prendre :

$$\hat{s}_t = \operatorname{argmin}\{\| Z_{t-1}(R_0) - Z_t(R_0 + s) \| \mid s \in S\} \text{ pour } t \in \{1, \ldots, T\} \tag{3.4}$$

Cependant, lorsque l'on choisit $\{\hat{s}_t\}_{t \in \{1 \ldots T\}}$ de cette manière on obtient à certaines dates des valeurs physiquement irréalistes. Par exemple, lorsque les champs de vents sont homogènes ou complexes, la fonction $s \to \|Z_{t-1}(R_0) - Z_t(R_0 + s)\|$ peut posséder plusieurs minima locaux et le minimum absolu ne correspond pas forcément à un déplacement réaliste. En particulier, la séquence $\{\hat{s}_t\}_{t \in \{1 \ldots T\}}$ obtenue avec cette méthode présente de nombreuses "ruptures" (cf figure 3.8) alors qu'on s'attend à ce que le processus $\{S_t\}$ évolue lentement.

Afin de filtrer ces valeurs aberrantes, nous avons alors calculé

$$(\hat{s}_1, \ldots, \hat{s}_T) = \operatorname{argmax}\{f(s_1, \ldots, s_T) \,|\, (s_1, \ldots, s_T) \in S^T\} \tag{3.5}$$

avec

$$f(s_1, \ldots, s_T) = \prod_{t=2}^{T} \exp\left(-\frac{\|s_t - s_{t-1}\|}{d_0}\right) \prod_{t=1}^{T} \exp\left(-\frac{\|Z_{t-1}(R_0) - Z_t(R_0 + s_t)\|^2}{\sigma^2}\right) \tag{3.6}$$

Le paramètre $d_0 > 0$ est un paramètre de lissage temporel : lorsque $d_0 \to +\infty$, le premier terme de l'équation (3.6) disparaît et on obtient alors la suite $(\hat{s}_1, \ldots, \hat{s}_T)$ définie par l'équation (3.4), et lorsque $d_0 \to 0$, ce terme devient infini et la séquence $(\hat{s}_1, \ldots, \hat{s}_T)$ devient constante.

En pratique, la maximisation de la fonction f peut être effectuée rapidement en utilisant l'algorithme de Viterbi. En effet, la forme de la fonction à maximiser est similaire à celle de la vraisemblance complète d'une chaîne de Markov cachée, à savoir :

$$p_\theta(s_1, \ldots s_T, y_1, \ldots, y_T) = \prod_{t=1}^{T} p_\theta(s_{t-1}, s_t) \prod_{t=1}^{T} p_\theta(y_t | s_t)$$

Nous avons choisi empiriquement $d_0 = 50$ km et $\sigma = 3 ms^{-1}$. Un exemple de séquence $\hat{s}_t$ obtenue avec ces paramètres est représentée sur la figure 3.8. Sur cette figure, la séquence obtenue avec l'équation (3.4) est aussi représentée.

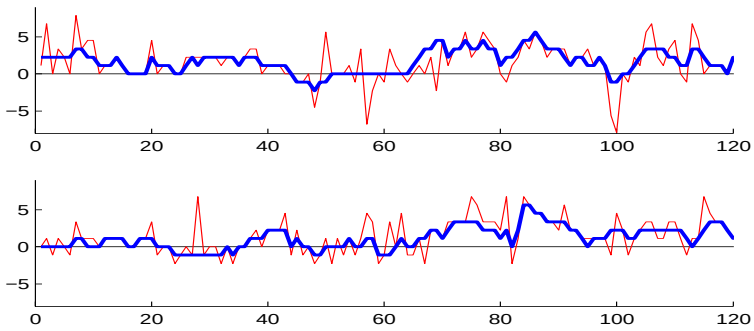


figure 3.8 : Séquences $\hat{s}_t$ obtenues avec les méthode décrites par l'équation (3.4) (trait fin) et l'équation (3.5) (trait épais). La figure du haut représente la composante zonale et la figure du bas la composante méridienne. Mois de Janvier 1992.

Afin de valider cette méthode, nous avons vérifié visuellement le réalisme physique de la séquence construite, et les résultats semblent globalement bons. Un exemple est donné sur la figure 3.9.

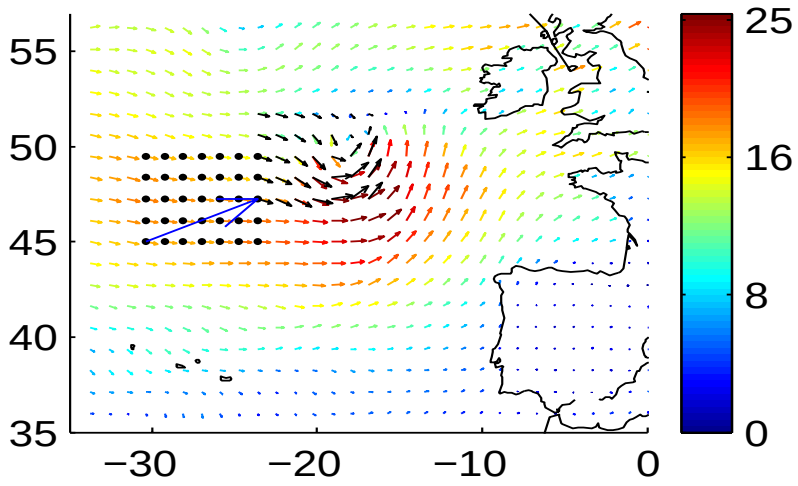


figure 3.9 : Situation du 4 Janvier 1998 (suite): le champ en rouge correspond à la situation à 18 heure, le vecteur bleu représente le vecteur de translation calculé $\hat{s}$ et le champ en noir la situation observée à 12 heure sur la zone R_0 (points en bleu).

La densité empirique de la séquence $\hat{s}$ (cf figure 3.10) montre que les déplacements ont généralement lieu de l'ouest-sud-ouest vers l'est-nord-est, ce qui est conforme à la climatologie de la région. Comme aucun déplacement important vers l'est, le nord et le sud n'a été observé, le domaine $\boldsymbol{S}$ a été réduit afin de limiter les temps de calculs dans la suite. Le domaine retenu est montré sur la figure 3.10 : cet ensemble est de cardinal 147.

.

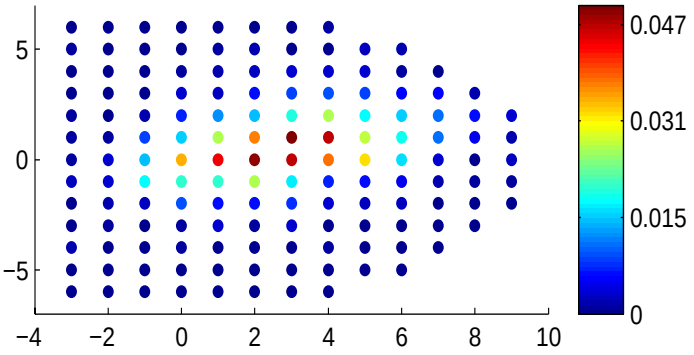


figure 3.10 : Densité empirique de $\hat{s}$.

Comme nous l'avons déjà mentionné dans l'introduction de ce paragraphe, nous allons principalement nous servir de cette séquence afin d'inférer la manière dont évoluent les champs de vent sur la zone R_0 conditionnellement aux déplacements des structures météorologiques. Pour cela nous allons estimer différentes caractéristiques de la loi conditionnelle $p(Y_t | Y_{t-1}, S_t)$. Ainsi, sur la figure 3.11, nous avons représenté un estimateur de la moyenne conditionnelle de

$Z_t(R_1)$ sachant $S_t = s$ pour différentes valeurs de s. Pour cela, nous avons utilisé l'estimateur:

$$\frac{1}{card\{t \,|\, \|\hat{s}_t - s\| \le d_s\}} \sum_{t=1}^{T} 1_{\{\|\hat{s}_t - s\| \le d_s\}} Z_t(R_1)$$

avec d_s le plus petit réel positif d tel que $card\{t \in \{1 \dots T\} \,|\, \|\hat{s}_t - s\| \le d\} \ge n_0$ et n_0 un entier positif fixé (dans la suite $n_0 = 100$).

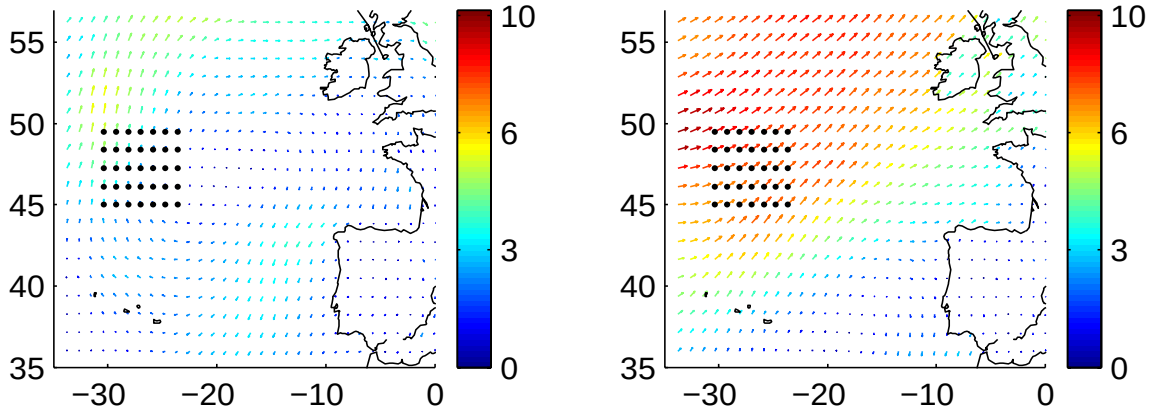


figure 3.11 : Champs de vents moyens pour $s = (0, 0)$ et $s = (4, 1)$.

La figure 3.11 permet de mettre en évidence que les valeurs prises par $\hat{s}$ sont fortement liées à la position des grands centre d'action qui régissent la météorologie de l'Atlantique Nord-Est. Ainsi lorsque $\hat{s} = (0, 0)$ (déplacement nul) la moyenne conditionnelle correspond à une situation où les vents sont généralement faibles et de secteur sud-sud-est sur la zone R_0, un anticyclone étant positionné au centre de la zone R_1 comme le montre l'enroulement des champs de vent dans le sens des aiguilles d'une montre. Pour $\hat{s} = (4, 1)$ (déplacement important vers l'est), les vents sont généralement assez forts, de secteur ouest-sud-ouest, et correspondent à une situation où une vaste zone dépressionnaire est située sur le nord de la zone R_1.

Finalement, la méthode décrite ci-dessus permet de calculer rapidement une première estimation du déplacement des masses d'air et les vecteurs obtenus semblent généralement réalistes. Ces valeurs vont nous servir à inférer des formes paramétriques simples pour les matrices $A^{(s)}$, $B^{(s)}$ et $\Sigma^{(s)}$ qui servent à décrire l'évolution du processus observé dans les différents régimes, puis à obtenir une première estimation des différents paramètres introduits. Dans un deuxième temps, les valeurs prises par le processus "déplacement des masses d'air" seront à nouveau supposées non-observées et les différents paramètres du modèle seront réestimés par maximum de vraisemblance. Il semble donc "inutile" de développer des algorithmes plus sophistiqués pour calculer ces vecteurs de déplacement. Notons enfin qu'une autre méthode, basée sur des techniques de déformation d'images, est proposée dans *Aberg* (2002, [1]).

3.b.2 Paramétrisation des modèles autorégressifs dans les différents régimes

Nous cherchons à écrire un modèle de la forme (3.3) pour modéliser l'évolution des champs

de vent conditionnellement au déplacement des masses d'air. Afin de limiter le nombre de paramètres du modèle, nous proposons dans ce paragraphe des formes paramétriques simples pour $A^{(s)}$, $B^{(s)}$ et $\Sigma^{(s)}$ (avec $s \in S$) et le noyau de transition de la chaîne cachée. Les différents choix effectués sont motivés, et en particulier nous mentionnons les hypothèses qui nous semblent les plus restrictives.

Paramétrisation de $\mathbf{A^{(s)}}$

Par définition du processus S_t, on sait que si $S_t = s$ alors $Z_{t-1}(R_0) \approx Z_t(R_0 + s)$. On va alors choisir $A^{(s)}$ comme une matrice permettant d'extrapoler le champ de vent sur la zone R_0 à partir de la connaissance du champ sur la zone $R_0 + s$.

Notons $\varsigma_t = Z_t(R_0 + S_t) - Z_{t-1}(R_0)$ la déformation du champ de vent entre les instants $t-1$ et t. On va supposer que ς_t est indépendant de S_t et noter $D = cov(\varsigma_t)$. En s'inspirant de l'estimateur des moindres carrés pour les modèles de régressions linéaires, on va choisir les matrices $A^{(s)}$ de la manière suivante, pour $s \in \boldsymbol{S}$:

$$A^{(s)} = cov(Z_t(R_0), Z_t(R_0 + s)) \times [cov(Z_t(R_0 + s)) + D]^{-1} \tag{3.7}$$

En pratique, les différentes matrices de covariance intervenant dans l'équation ci-dessus peuvent être estimées en utilisant l'hypothèse de stationnarité temporelle des champs de vent ainsi que les séquences $\{\hat{s}_t\}$ calculées au paragraphe précédent. On obtient ainsi des estimations des matrices $A^{(s)}$, et dans la suite de ce chapitre, nous allons supposer que les matrices $A^{(s)}$ sont fixées, égales à ces valeurs. L'estimateur $\hat{D}$ de la matrice D est montré sur la figure 3.12.

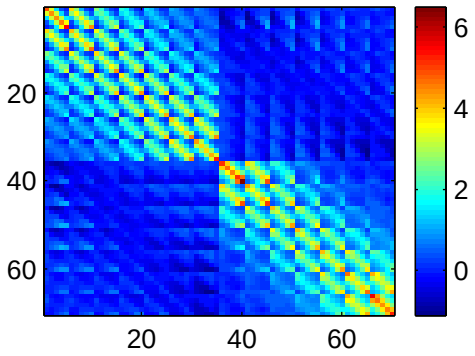


figure 3.12 : Matrice de covariance empirique $\hat{D}$.

Un exemple de champ prédit avec cette méthode est donné sur la figure 3.13. Nous avons tracé le même type de figures pour différentes dates, et nous avons ainsi pu vérifier visuellement que la méthode d'extrapolation linéaire décrite ci-dessus permet généralement de prédire la forme générale du champ de vent sur la zone R_0. Diverses méthodes d'extrapolation non-linéaires ont aussi été essayées. Celles-ci n'ont pas apporté d'améliorations notables sur la qualité de la prédiction. De plus, l'utilisation de ces méthodes aboutirait à un modèle $MS-NAR$ au lieu d'un modèle $MS-LAR$ et ces modèles sont plus difficiles à manipuler en pratique.

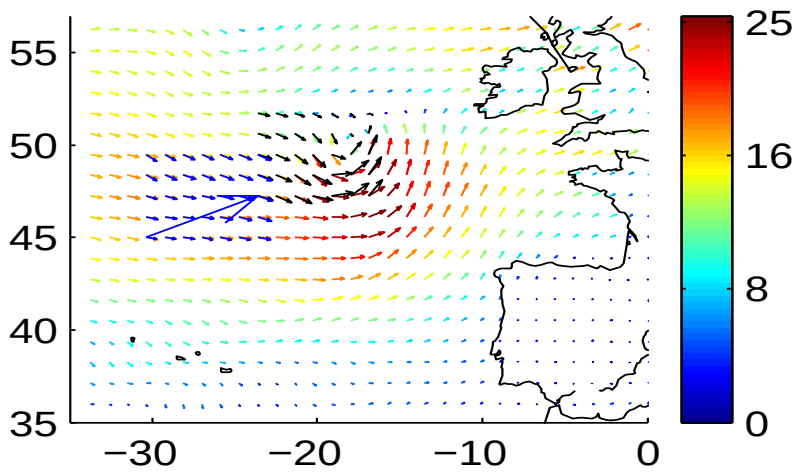


figure 3.13 : A gauche: champ à l'instant t sur la zone S_1 (en couleurs), le champ $z_0(t-1)$ (en noir) est dessiné sur la zone translaté $S_0 + a(t)$, le champ extrapolé $A_{\hat{a}(t)} z_0(t-1)$ est représenté en bleu sur la zone S_0.

Nous avons vu, au chapitre 1, que la stabilité du modèle $MS-LAR$ défini par l'équation (3.3) dépend des matrices $A^{(s)}$ et de la manière dont évolue la chaîne $\{S_t\}$. Avec le choix fait ci-dessus, on peut vérifier numériquement que $\|A^{(s)}\| < 1$ pour $s \in \boldsymbol{S}$ et $\|.\|$ la norme matricielle associée à la norme euclidienne. On en déduit en particulier que le critère **(A1)** de la proposition 1.1 est vérifié. Par contre, si nous choisissons $D = 0$ dans l'équation (3.7), le rayon spectral des matrices $A^{(s)}$ peut être supérieur ou égal à 1, notamment lorsque l'intersection des zones R_0 et $R_0 + s$ est non vide. En effet, il existe alors $i \in \{1 \ldots N\}$ tel que $r_i^0 \in R_0 \cap R_0 + s$ et on peut vérifier alors que les vecteurs de $\boldsymbol{R}^{2N}$ dont toutes les composantes sont nulles sauf la i$^{\text{ème}}$ sont des vecteurs propres associés à la valeur propre 1. Les simulations numériques effectuées montrent que les modèles $MS-LAR$ correspondants sont instables.

Le choix fait ci-dessus suppose implicitement que les matrices de covariance $cov(Z_{t-1}(R_1) | S_t = s)$ et $cov(Z_t(R_0) - Z_{t-1}(R_0 + s) | S_t = s)$ sont indépendantes de s. Cette hypothèse est sans doute abusive. En effet, en utilisant les déplacements calculés au paragraphe 3.b, on peut estimer ces matrices de covariances conditionnelles. Ces estimations montrent qu'il existe une relation entre la matrice de covariance de la déformation des structures. Ainsi, dans les conditions dépressionnaires, qui correspondent à des déplacements importants des masses d'air vers l'est, les déformations sont généralement plus importantes que lorsque les conditions sont anticycloniques. Cependant, ces relations semblent difficiles à modéliser et les tests effectués n'ont pas permis d'apporter d'améliorations notables à la qualité de la prédiction.
Une extension naturelle de ce modèle est présentée dans *Ailliot et al.* (2003, [4]). Ce modèle est plus précisément décrit ci-dessous. Notons

$$\hat{Y}_t^{(1)} = (\hat{u}_t^{(1)}(r_1^0), \ldots, \hat{u}_t^{(1)}(r_N^0), \hat{v}_t^{(1)}(r_1^0), \ldots, \hat{v}_t^{(1)}(r_N^0))$$

le champ obtenu par extrapolation linéaire comme décrit ci-dessus, c'est à dire

$\hat{Y}_t^{(1)} = A^{(S_t)} Y_{t-1}$ avec $A^{(S_t)}$ la matrice définie par l'équation (3.7). Dans [4] la matrice $A^{(S_t)}$ est choisie de telle manière que

$$\hat{Y}_t = A^{(S_t)} Y_{t-1} = (\hat{u}_t(r_1^0), \ldots, \hat{u}_t(r_N^0), \hat{v}_t(r_1^0), \ldots, \hat{v}_t(r_N^0))$$

avec, pour $i \in \{1 \ldots N\}$:

$$\begin{bmatrix} \hat{u}_t(r_i^0) \\ \hat{v}_t(r_i^0) \end{bmatrix} = \Phi^{(S_t)}(r_i^0) \begin{bmatrix} \hat{u}_t^{(1)}(r_i^0) \\ \hat{v}_t^{(1)}(r_i^0) \end{bmatrix} + B^{(S_t)}(r_i^0) \begin{bmatrix} u_{t-1}(r_i^0) \\ v_{t-1}(r_i^0) \end{bmatrix}$$

où $B^{(s)}(r_i^0) \in M_2(R)$ et $\Phi^{(s)} \in M_{N,1}(R)$ représentent des matrices de paramètres. Dans ce modèle, les matrices $A^{(s)}$ sont obtenues comme combinaison linéaire de 2 matrices, la première d'entre elles permettant d'extrapoler le champ sur la zone R_0 à partir de la connaissance du champ sur la zone $R_0 + s$ et donc de modéliser le déplacement des structures météorologiques et la deuxième décrivant l'évolution résiduelle du processus en chaque point via un modèle autorégressif bivarié. Ce type de modèle, dans le cas des champs scalaires et lorsqu'il n'y a pas de changements de régimes, est proposé dans *Cressie* (1993, [35]) (modèle STAR). Lorsque $B^{(s)}(r_i^0) = 0$ pour $i \in \{1 \ldots N\}$, nous retrouvons les matrices $A^{(s)}$ définies par (3.7). Nous avons trouvé que l'utilisation de ce type de modèle n'est pas justifiée pour nos données, l'introduction des nouveaux paramètres ne permettant pas de diminuer significativement l'erreur de prédiction.

Paramétrisation de B(s)

Le choix de $B^{(s)}$ doit permettre notamment de modéliser le fait que la moyenne de la loi stationnaire du régime s, définie par $M^{(s)} = (I - A^{(s)})^{-1} B^{(s)}$, dépend fortement de s (cf figure 3.11). $M^{(s)}$ s'écrit sous la forme $M^{(s)} = (m_u^{(s)}(1), \ldots, m_u^{(s)}(N), m_v^{(s)}(1), \ldots, m_v^{(s)}(N))'$ où $m_u^{(s)}(i)$ et $m_v^{(s)}(i)$ désignent respectivement la moyenne de la composante zonale et de la composante méridienne au point r_i^0. Nous allons supposer que $m_u^{(s)}(i) = m_u^{(s)}$ et $m_v^{(s)}(i) = m_v^{(s)}$ pour $i \in \{1 \ldots N\}$, c'est à dire que la moyenne de la loi stationnaire est identique en tous les points de R_0. Cette hypothèse est sans doute contraignante puisque, par exemple, on s'attend à ce que la moyenne soit plus forte aux points situés au nord de la zone dans les conditions dépressionnaires. Cependant, ceci semble difficile à décrire avec un modèle paramétrique simple.

Si on suppose que les estimateurs des moyennes conditionnelles $E(Y_t | S_t = s)$ donnés au paragraphe 3.b fournissent une estimation raisonnable de $M^{(s)}$, on peut en déduire une estimation $\hat{m}^{(s)}$ de $m^{(s)} = (m_u^{(s)}, m_v^{(s)})$ en moyennant les valeurs obtenues pour les différents points de R_0. On obtient ainsi un champ de vecteur $s \to \hat{m}^{(s)}$ qui est représenté sur la figure 3.14. On retrouve, en particulier, que le vent moyen dans un système météorologique est lié à sa direction de propagation. Par exemple, dans les systèmes se déplaçant vers l'est les vents soufflent généralement dans la même direction, et plus la vitesse de déplacement est grande plus les vents moyens sont forts.

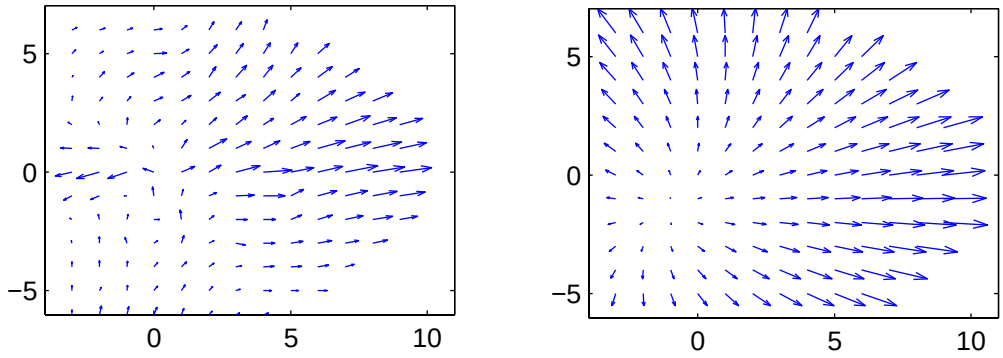


figure 3.14 : champs de vecteurs $s \to \hat{m}^{(s)}$ *(à gauche) et* $s \to \hat{F}s + \hat{G}$ *(à droite).*

Afin de modéliser ce champ de vecteurs en utilisant un nombre restreint de paramètres, nous allons faire l'hypothèse supplémentaire que pour $s \in \boldsymbol{S}$:

$$\hat{m}^{(s)} = Fs + G \tag{3.8}$$

avec $F \in M_2(\boldsymbol{R})$ et $G \in M_{2,1}(\boldsymbol{R})$ des matrices de paramètres. Nous avons obtenu une première estimation de ces matrices, notées $\hat{F}$ et $\hat{G}$, en utilisant les estimateurs des moindres carrés. Sur la figure 3.14., nous avons représenté les champs de vecteur $s \to \hat{m}^{(s)}$ et $s \to \hat{F}s + \hat{G}$. La comparaison de ces champs de vecteur montre que l'hypothèse (3.8) est sans doute simplificatrice mais permet tout de même de modéliser le comportement général de $\hat{m}^{(s)}$.

Choix de $\Sigma^{(s)}$

Il reste à modéliser les matrices de covariances des résidus. Notons que ce terme est important, puisqu'il permet modéliser la partie du champ de vent non prédite par le champ de vent à l'instant précédent et va donc permettre de modéliser, entre autres, l'arrivée de nouveaux "événements" (perturbations par exemple) et la déformation des structures entre deux instants successifs. Ce terme va donc avoir une influence sur le réalisme des séquences simulées avec le modèle.

A partir des séquences $\hat{s}_t$ calculées au paragraphe 3.b et des estimateurs de $A^{(s)}$ et $B^{(s)}$ décrits aux deux paragraphes précédents, nous avons calculé des estimations $\hat{\Sigma}^{(s)}$ de

$$\Sigma^{(s)} = \begin{bmatrix} \Sigma^{(s)}_{(u,u)} & \Sigma^{(s)}_{(u,v)} \\ \Sigma^{(s)}_{(u,v)} & \Sigma^{(s)}_{(v,v)} \end{bmatrix} \text{ pour } s \in \Lambda$$

La figure 3.15 montre la matrice $\hat{\Sigma}^{(s)}$ obtenue pour $s = (3, 0)$. A première vue, la structure de cette matrice semble complexe. Nous allons nous intéresser dans un premier temps uniquement à la diagonale de cette matrice.

Sur la figure 3.15, nous avons représenté la variance empirique de l'erreur de prédiction pour

la composante méridienne aux différents points du domaine R_0, c'est à dire la fonction $r_i^0 \to \hat{\Sigma}^{(s)}_{(v,v)}(i,i)$. Cette figure montre que les erreurs les plus importantes sont commises sur les points situés à l'ouest de la zone, et donc les plus éloignés de la zone $R_0 + s$, et les erreurs les plus faibles sont commises pour les points appartenant à $R_0 \bigcap R_0 + s$.
En fait, cette erreur est due principalement à deux facteurs, à savoir l'erreur d'extrapolation et la déformation des structures entre deux instants successifs. Or, l'erreur d'extrapolation a tendance à augmenter avec la distance entre le point où on veut faire la prédiction et la zone sur laquelle on dispose de l'information.

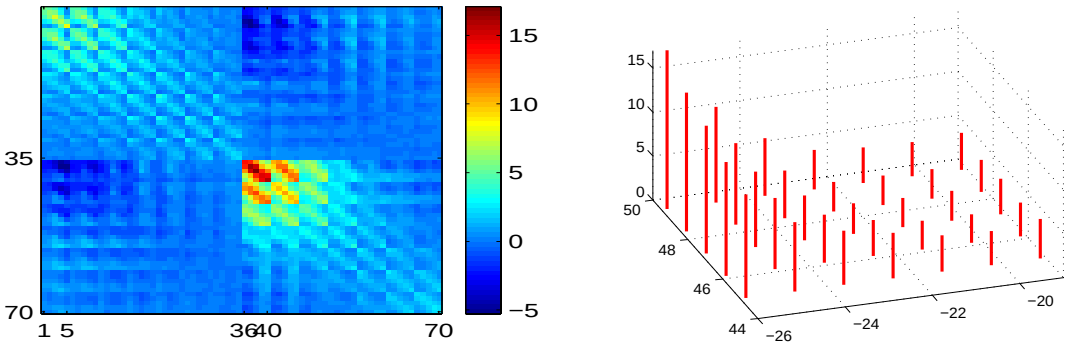


figure 3.15 : Matrice de covariance empirique $\hat{\Sigma}^{(s)}$ pour $s = (3,0)$, et répartition spatiale de l'erreur sur la composante méridienne $r_i^0 \to \hat{\Sigma}^{(s)}_{(v,v)}(i,i)$.

Ce comportement est plus précisément mis en valeur sur la figure 3.16 où est représentée l'évolution de l'erreur $\hat{\Sigma}^{(s)}_{(v,v)}(i,i)$ en fonction de la distance entre le point r_i^0 et la zone translatée $R_0 + s$. Afin de modéliser les coefficients diagonaux de la matrice $\Sigma^{(s)}$, nous avons alors fait les hypothèses suivantes:

$$\begin{aligned} \Sigma^{(s)}_{(u,u)}(i,i) &= f_u(d(r_i^0, R_0 + s)) \\ \Sigma^{(s)}_{(v,v)}(i,i) &= f_v(d(r_i^0, R_0 + s)) \end{aligned} \tag{3.9}$$

avec f_u et f_v des fonctions à valeurs positives définies sur $\boldsymbol{R}^+$ et d la distance euclidienne sur $\boldsymbol{R}^2$. L'équation (3.9) signifie que l'erreur commise en un point $r_i^0 \in R_0$ dépend uniquement de la distance entre ce point et la zone "translatée" $R_0 + s$.

Des estimations empiriques $\hat{f}_u$ et $\hat{f}_v$ de f_u et f_v respectivement peuvent être obtenues à partir des matrices $\hat{\Sigma}^{(s)}_{(u,u)}$ et $\hat{\Sigma}^{(s)}_{(v,v)}$. Celles-ci sont représentées sur la figure .

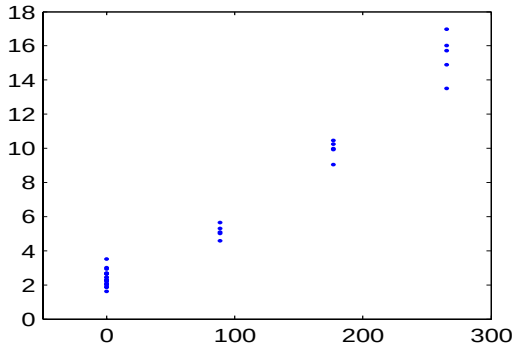


figure 3.16 : Evolution de l'erreur $\hat{\Sigma}^{s}_{(v,v)}(i,i)$ *en fonction de la distance* $d(r_i^0, R_0 + s)$ *(en km) entre le point* r_i^0 *et la zone* $R_0 + s$.

En s'inspirant des modèles couramment utilisés pour modéliser les variogrammes (cf *Cressie* (1993, [35])), nous avons testé différentes formes paramétriques pour les fonctions f_u et f_v et finalement, nous avons retenu le modèle exponentiel, à savoir

$$f_u(d) = \beta_u + \alpha_u e^{-\frac{d}{d_u}}$$

avec $\beta_u \geq 0$, $\alpha_u \geq 0$ et $d_u > 0$ des paramètres (et de même pour f_v). Nous avons obtenu une première estimation de ces paramètres en ajustant les formes paramétriques aux estimations empiriques $\hat{f}_u$ et $\hat{f}_v$ par la méthode des moindres carrés. Les résultats obtenus sont représentés sur la figure .

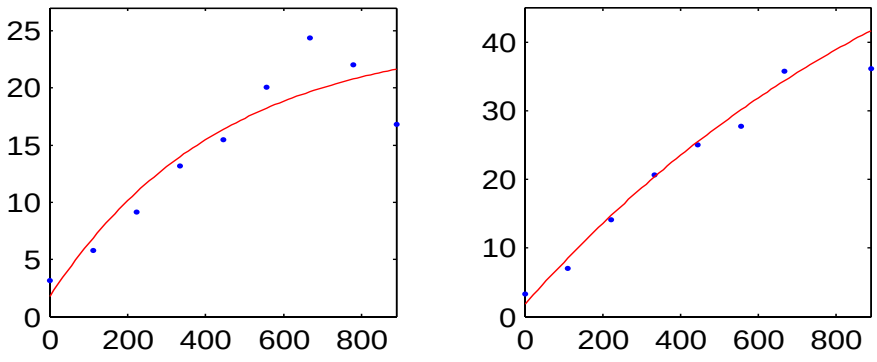


figure 3.17 : Estimations $\hat{f}_u$ *et* $\hat{f}_v$ *(en bleu) et ajustement du modèle exponentiel (en rouge).*

Il reste ensuite à modéliser les matrices de corrélation $\rho^{(s)}$ associées à la matrice de covariance $\Sigma^{(s)}$, c'est à dire définie par $\rho^{(s)}(i,j) = \Sigma^{(s)}(i,j) / \sqrt{\Sigma^{(s)}(i,i)\Sigma^{(s)}(j,j)}$ pour $s \in \boldsymbol{S}$ et $i, j \in \{1, \ldots, 2N\}$. Notons que ces matrices admettent la même structure par bloc que les matrices de covariance $\Sigma^{(s)}$, et s'écrivent donc sous la forme

$$\rho^{(s)} = \begin{bmatrix} \rho^{(s)}_{(u,u)} & \rho^{(s)}_{(u,v)} \\ \rho^{(s)}_{(u,v)} & \rho^{(s)}_{(v,v)} \end{bmatrix}$$

Un exemple de matrice de corrélation empirique est donné à la figure 3.18. Nous allons supposer que

$$\begin{aligned} \rho^{(s)}_{(u,u)}(i,j) &= g_u(\|r_i^0 - r_j^0\|) \\ \rho^{(s)}_{(v,v)}(i,j) &= g_v(\|r_i^0 - r_j^0\|) \\ \rho^{(s)}_{(u,v)}(i,j) &= 0 \end{aligned} \tag{3.10}$$

avec g_u et g_v des fonctions définies de $\boldsymbol{R}^+$ dans $\boldsymbol{R}$, vérifiant $g_u(0) = g_v(0) = 1$. Les deux premières hypothèses signifient que la corrélation entre les erreurs commises sur les composantes zonale et méridienne respectivement en deux points différents dépend uniquement de la distance entre ces deux points. La dernière hypothèse signifie que les erreurs commises sur les deux composantes sont non corrélées, ce qui est sans doute abusif au vu de la figure 3.18. Dans la suite, nous avons choisi :

$$g_u(d) = g_v(d) = e^{-\frac{d^2}{d_0^2}}$$

avec $d_0 > 0$ un paramètre. Ce choix assure que les matrices $\rho^{(s)}$ et donc $\Sigma^{(s)}$ sont définies positives pour $s \in \boldsymbol{S}$ (cf *Cressie* (1993, [35]) p 89). Une première estimation de d_0 a été choisie empiriquement égale à $450km$.

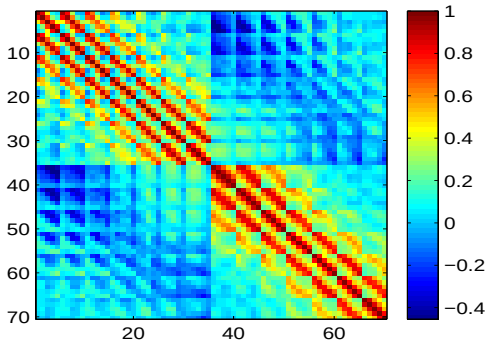


figure 3.18 : Matrice de corrélation empirique $\hat{\rho}^{(s)}$ *pour* $s = (3, 0)$.

Finalement, nous avons modélisé les matrices de covariances du résidu dans les différents régimes avec un total de 7 paramètres. Un exemple de comparaison de la matrice de covariance empirique avec sa version paramétrique est donné à la figure 3.19. Sur cet exemple, qui correspond au déplacement $s = (3, 0)$, la paramétrisation choisie permet de bien reproduire la forme générale des matrices de covariance empirique. Nous avons vérifié visuellement qu'il en était

de même pour les autres valeurs de $s \in \boldsymbol{S}$.

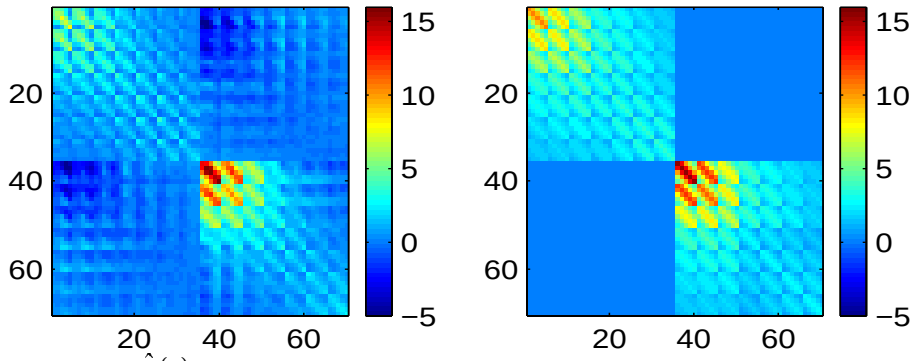


figure 3.19 : Matrice $\hat{\Sigma}^{(s)}$ *à gauche et matrice paramétrique correspondante à droite pour* $s = (3, 0)$.

Paramétrisation de la matrice de transition de la variable cachée

Nous avons déjà mentionné que la chaîne de Markov cachée peut prendre un nombre relativement élevé de valeurs : avec les choix effectués au paragraphe 3.b, nous avons $M = card(\boldsymbol{S}) = 147$. Afin d'avoir un nombre de paramètres raisonnables, nous allons supposer que, pour $s_i, s_j \in \boldsymbol{S}$,

$$q_\theta(i,j) = P(S_t = s_j | S_{t-1} = s_i) \sim \exp\left(-\frac{\|s_i - s_j\|^2}{\sigma^2}\right) \exp((s_j - s_0)' \Sigma^{-1} (s_j - s_0)) \qquad (3.11)$$

avec

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \in S_2^+(\boldsymbol{R})$$

et $\sigma_1, \sigma_2 > 0$, $\rho \in]-1,1[$, $s_0 \in \boldsymbol{R}^2$ et $\sigma > 0$. Le premier terme de l'équation (3.11) permet de modéliser le fait que les probabilités de transition décroissent avec la distance entre les éléments de $\boldsymbol{S}$ et le deuxième terme est introduit afin de modéliser la loi stationnaire de la chaîne de Markov $\{S_t\}$ (une estimation empirique de cette loi est donnée sur la figure 3.10). Avec le choix fait ci-dessus, on peut vérifier que $q_\theta(i,j) > 0$ pour $(i,j) \in \boldsymbol{S}^2$ et donc que la chaîne de Markov est donc irréductible et apériodique.

Notons enfin qu'une extension naturelle consisterait à supposer que la variable cachée est à valeurs continues (dans un sous espace de $\boldsymbol{R}^2$).

3.b.3 Estimation des paramètres

Finalement, afin de modéliser l'évolution du processus $\{Y_t\}$, nous proposons d'utiliser un modèle $MS-LAR$ avec innovations gaussiennes de la forme $Y_t = A^{(S_t)} Y_{t-1} + B^{(S_t)} + \Sigma^{(S_t)} E_t$ avec

- $A^{(s)} \in M_{2N}(\boldsymbol{R})$ pour $s \in \boldsymbol{S}$. Ces matrices sont fixées.
- $B^{(s)} \in M_{2N,\,1}(\boldsymbol{R})$ pour $s \in \boldsymbol{S}$. La fonction $s \to B^{(s)}$ est paramétrée par les matrices $F \in M_2(\boldsymbol{R})$ et $G \in M_{2,\,1}(\boldsymbol{R})$ soit un total de 6 paramètres. Notons $\theta_B = (F, G) \in \boldsymbol{R}^6$.
- $\Sigma^{(s)} \in S^+_{2N}(\boldsymbol{R})$ pour $s \in \boldsymbol{S}$. La fonction $s \to \Sigma^{(s)}$ est paramétrée par $\theta_\Sigma = (\alpha_u, \beta_u, d_u, \alpha_v, \beta_v, d_v, d_0) \in (\boldsymbol{R}^+)^7$.
- $\{S_t\}$ une chaîne de Markov à valeurs dans $\boldsymbol{S} \subset \boldsymbol{Z}^2$ de matrice de transition Q_θ. Ce noyau de transition est paramétrée par $\theta_S = (\sigma_1, \sigma_2, \sigma, \rho, a_0) \in (\boldsymbol{R}^{+*})^3 \times]\text{-}1,1[\times \boldsymbol{R}^2$.

Le modèle proposé possède donc un total de 19 paramètres. Nous noterons alors $\theta = (\theta_R, \theta_S)$ l'ensemble des paramètres avec $\theta_R = (\theta_B, \theta_\Sigma)$ l'ensemble des paramètres servant à décrire l'évolution des champs de vent dans les différents régimes.

Nous avons déjà mentionné que ce modèle vérifie la condition **(A1)** de la proposition 1.1 et que la chaîne de Markov cachée est irréductible et apériodique. En utilisant la proposition 1.1, on en déduit donc l'existence et l'unicité d'une solution stationnaire. Si on veut établir la consistance des estimateurs du maximum de vraisemblance, il reste à vérifier deux choses, à savoir que la loi stationnaire possède un moment d'ordre 2 et l'identifiabilité des paramètres. Pour établir le premier point on pourra utiliser la proposition 1.2. Par contre, l'identifiabilité des paramètres de ce modèle semble difficile à établir. En effet, les mêmes paramètres servent à décrire l'évolution du processus observé dans les différents régimes, et on ne peut alors pas adapter directement le raisonnement utilisé dans la preuve de la proposition 1.11. On peut tout de même appliquer la proposition 1.13, mais on ne sait plus caractériser l'ensemble $D(\theta_0)$ qui représente l'ensemble des limites possibles de la suite des estimateurs du maximum de vraisemblance $\theta_{T,\zeta}$.

Nous avons tout de même calculé les EMV en utilisant l'algorithme GEM. Nous avons pris pour valeur initiale des paramètres les estimations obtenues à partir de la séquence $(\hat{s}_t)_{t \in \{1 \ldots T\}}$. Rappelons que l'étape E requiert de l'ordre de $2TM^2$ opérations, et est donc nettement plus longue à mettre en oeuvre que dans les exemples développés au chapitre précédent. En ce qui concerne l'étape M, on peut écrire la fonction à maximiser sous la forme (cf paragraphe 1.c.2) :

$$Q(\theta, \theta^{(n-1)}) = Q_S(\theta_S, \theta^{(n-1)}) + Q_R(\theta_R, \theta^{(n-1)})$$

On est alors ramené à maximiser indépendamment ces deux fonctions. Pour aucune des deux, il n'existe d'expression analytique pour la valeur réalisant le maximum. Nous avons donc utilisé quelques itérations d'un algorithme quasi-Newton sur les fonctions $Q_S(.\,, \theta^{(n-1)})$ et $Q_R(.\,, \theta^{(n-1)})$ pour réestimer les paramètres à chaque itération de l'algorithme GEM. La maximisation de $Q_S(.\,, \theta^{(n-1)})$ est relativement rapide, mais par contre celle de $Q_R(.\,, \theta^{(n-1)})$ est nettement plus fastidieuse du fait du nombre important d'opérations nécessaires pour calculer cette fonction ainsi que son gradient. Les temps de calculs demeurent tout de même raisonnables, et la convergence de l'algorithme est obtenue après quelques heures de calcul sur un PC avec un processeur Intel Pentium de 1700 MHz et 512 MO de RAM.

3.c Validation du modèle et discussion

Dans le chapitre 2, afin de valider les différents modèles introduits nous avons principalement testé leurs capacités à simuler des séquences réalistes. Une autre méthode de validation usuelle d'un modèle consiste à vérifier son aptitude à réaliser des prédictions à court terme. Dans la première partie de ce paragraphe, nous montrons que le modèle introduit améliore nettement la qualité des prédictions par rapport à un modèle autorégressif linéaire, puis dans une deuxième partie, le modèle est validé en simulation en utilisant la méthode développée au paragraphe 1.e.

3.c.1 Validation en prédiction

Afin de prédire le champ de vent à l'instant $t+k$ à partir de la connaissance des champs de vent jusqu'à l'instant t avec le modèle $MS-AR$ identifié, nous avons calculé l'espérance conditionnelle $\overline{E}_\theta[Y_{t+k}|Y_0^t = y_0^t]$. Nous décrivons ci-dessous comment nous avons calculé cette espérance conditionnelle. Afin de simplifier l'exposé, nous considérons uniquement le cas $k = 1$ (prévision à un pas de temps). La généralisation est immédiate, et on pourra consulter *Krolzig* (1997, [69]) pour des détails supplémentaires.

On peut vérifier aisément que

$$\begin{aligned}\overline{E}_\theta[Y_{t+1}|Y_0^t = y_0^t] &= \sum_{s\in S} \overline{P}_\theta[S_{t+1} = s|y_0^t] E[Y_{t+1}|S_{t+1} = s, Y_t] \\ &= \sum_{s\in S} \overline{P}_\theta[S_{t+1} = s|y_0^t](A^{(s)}Y_t + B^{(s)})\end{aligned} \tag{3.12}$$

Le calcul de l'espérance conditionnelle $\overline{E}_\theta[Y_{t+1}|Y_0^t = y_0^t]$ fait donc intervenir uniquement les probabilités de prédiction $\overline{P}_\theta(S_{t+1} = s|y_0^t)$. Ces probabilités sont données par

$$\overline{P}_\theta(S_{t+1} = s|y_0^t) = \sum_{s_0\in S} \overline{P}_\theta(S_{t+1} = s|y_0^t, S_0 = s_0)\overline{P}_\theta(S_0 = s_0|y_0^t) \tag{3.13}$$

Le calcul de ces quantités est généralement infaisable puisque la distribution $\overline{P}_\theta(S_0 = s_0|y_0^t)$ fait intervenir la loi stationnaire du processus $\{S_t, Y_t\}$ et que celle-ci est inconnue. Nous avons alors remplacé ces probabilités par une loi arbitraire ς dans l'équation (3.13). En pratique, nous avons choisi ς égale à la loi stationnaire de la matrice de transition Q_θ. Il reste alors à calculer les quantités $P_\theta(S_{t+1} = s|y_0^T, S_0 = s_0)$, ce qui peut être fait en utilisant l'algorithme Forward-Backward décrit au paragraphe 1.c.2.

Sur la figure 3.20, nous avons représenté la matrice de covariance empirique du résidu pour le modèle $MS-LAR$ décrit au paragraphe 3.b.2 et celle correspondant à un modèle $AR(1)$ dont les paramètres ont été estimés par la méthode des moindres carrés. La comparaison directe de ces matrices n'est pas aisée. Nous avons alors tracé les erreurs de prédiction sur les composantes zonales et méridiennes aux différents points du domaine R_0 sur la figure 3.21.

Le modèle $MS-LAR$ améliore sensiblement la qualité de la prédiction en tous les points du domaine, l'amélioration étant toutefois plus nette pour les points situés à l'est de la zone R_0, et qui sont donc généralement plus proches de la zone translatée $R_0 + S_t$ puisque le déplacement des structures météorologiques se fait généralement vers l'est. En ces points, l'erreur de prédiction est proche des valeurs obtenues pour la matrice représentant la déformation des structures météorologiques (cf figure 3.12). L'erreur commise est nettement plus importante pour les points situés à l'ouest du domaine. Ces points sont généralement plus éloignés de la zone translatée $R_0 + S_t$, et il semble donc normal qu'on ait plus de difficulté à prédire les conditions de vent en ces points.

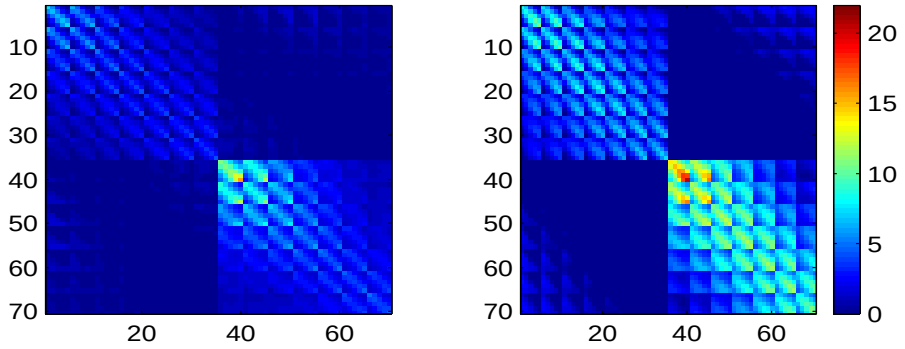


figure 3.20 : Matrice de covariance de l'erreur de prédiction. Modèle $MS-AR$ à gauche et AR à droite.

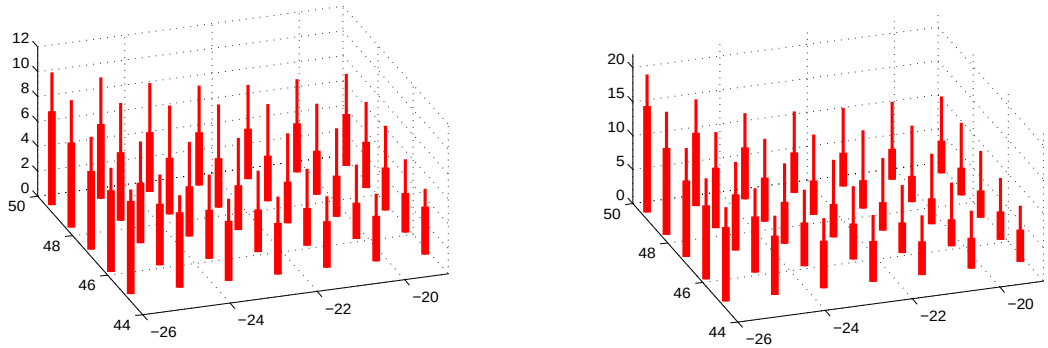


figure 3.21 : Variance de l'erreur de prédiction commise aux différents points de R_0. Composante zonale à gauche et méridienne à droite. Trait épais: modèle MS-AR. Trait fin : modèle AR.

Nous avons aussi calculé les prédictions à plusieurs pas de temps ($k = 2, k = 3, \ldots$) avec les deux modèles. Afin de comparer ces résidus, nous avons calculé la norme de Perron-Frobenius de leurs matrices de covariance empiriques. Cette norme est définie, pour $\Sigma \in M_n(R)$, par $\|\Sigma\|_{PF} = \sqrt{tr(\Sigma'\Sigma)}$. Les résultats obtenus sont donnés dans le tableau 3.2 et on peut voir que le modèle $MS-AR$ améliore nettement les résultats obtenus avec le modèle AR.

	1	2	3	4	5
MS-AR	151	538	952	1263	1498
AR	300	981	1719	2257	2560

Tableau 3.2 Evolution de la norme de P.F en fonction de l'intervalle de prédiction.

Nous avons donc montré que l'introduction de la variable "déplacement des masses d'air" sous la forme d'une variable cachée permet d'améliorer sensiblement la qualité des prédictions par rapport aux modèles $AR(1)$.

3.c.2 Validation en simulation

Afin de tester la capacité de ce modèle à simuler des champs de vents réalistes, nous avons procédé comme au chapitre précédent, à savoir que nous avons simulé un grand nombre de réalisations de ce modèle, puis nous avons comparé certaines propriétés de ces séquences simulées avec celles des données. Dans ce paragraphe, comme au paragraphe 2.c.2, nous avons choisi de montrer uniquement certains résultats graphiques.

Plus précisément, nous avons tout d'abord comparé les séquences simulées aux différents points du domaine R_0 avec les données correspondantes. Pour cela, nous avons utilisé les mêmes statistiques qu'au chapitre précédent, et nous avons comparé les lois marginales, les durées de persistance des périodes de calme et des tempêtes ainsi que les fonctions d'autocorrélation en ces différents points. Certains des résultats obtenus pour le point r_{18}^0 (point situé au centre du domaine R_0) sont montrés sur les figures 3.22-3.25. Ces résultats ne sont pas satisfaisants, sauf en ce qui concerne les durées de persistance des tempêtes qui sont relativement bien restituées par le modèle (cf figure 3.25).

Tout d'abord, le modèle $MS-AR$ ne permet pas de reproduire la complexité de la loi marginale du processus $\{u_t(r_{18}^0), v_t(r_{18}^0)\}$ (cf figure 3.22). Notamment, la fonction de répartition de l'intensité du vent est mal reproduite (cf figure 3.23) avec trop peu de vent de faible intensité. Par contre, la fonction de répartition de la direction du vent est bien restaurée. Afin de remédier à ce problème, on pourrait soit essayer d'appliquer une transformation initiale sur les données dans le but de rendre ces lois marginales approximativement gaussiennes comme au paragraphe 2.a.2 soit tester d'autres types de modèles autorégressifs afin de décrire l'évolution des champs de vent dans les différents régimes.

Les fonctions d'autocorrélation des processus $\{u_t(r_{18}^0)\}$ et $\{v_t(r_{18}^0)\}$ ne sont pas bien restaurées non plus (cf figure 3.24). Les fonctions d'autocorrélations empiriques calculées à partir des données sont complexes. Ainsi celle du processus $\{u_t(r_{18}^0)\}$ décroît nettement moins vite que celle de $\{v_t(r_{18}^0)\}$ et cette deuxième possède de nombreux extrema, dont la répartition semble périodique (période comprise entre 1 et 2 jours d'après une étude spectrale). L'interprétation physique de cette périodicité n'est pas claire.

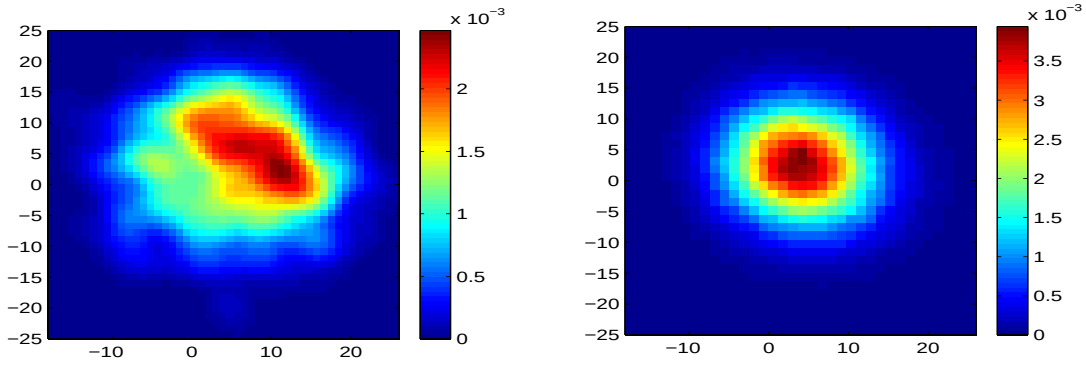


figure 3.22 : Loi marginale bivariée du processus $\{u_t(r_{18}^0), v_t(r_{18}^0)\}$. *Données à gauche, modèle à droite.*

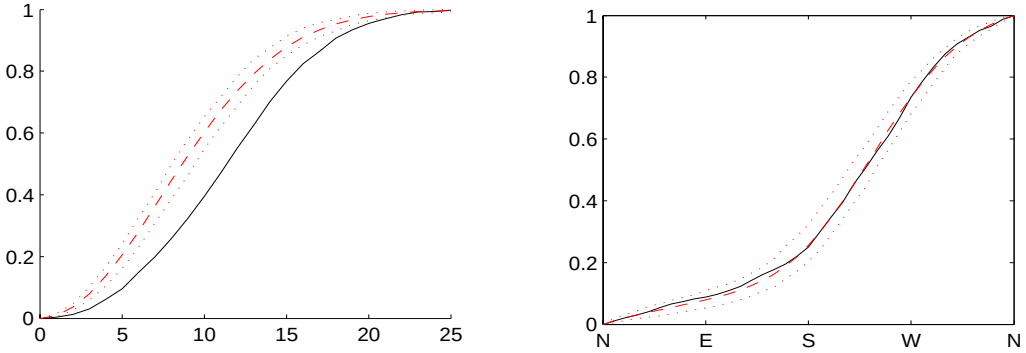


figure 3.23 : Répartition de l'intensité (à gauche) et de la direction du vent (à droite) au point r_{18}^0. —— *données,* – – *modèle,* ···· *intervalle de fluctuation à 95%.*

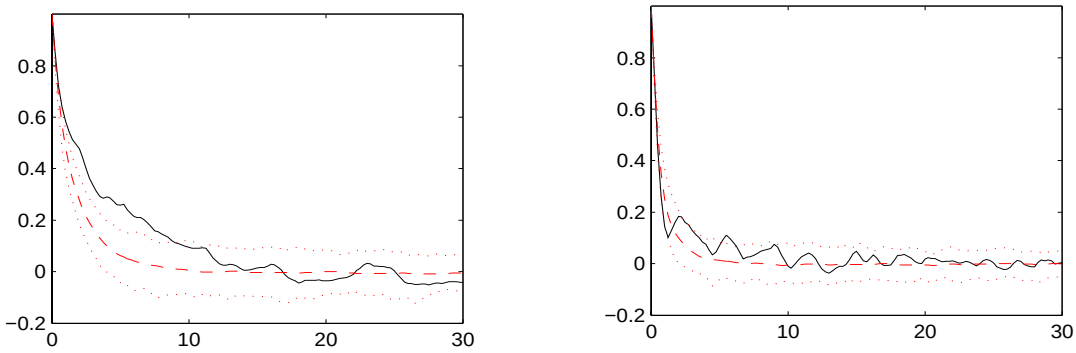


figure 3.24 : Fonctions d'autocorrélation de $\{u_t(r_{18}^0)\}$ *(à gauche) et de* $\{v_t(r_{18}^0)\}$ *(à droite).* —— *données,* – – *modèle,* ···· *intervalle de fluctuation à 95%.*

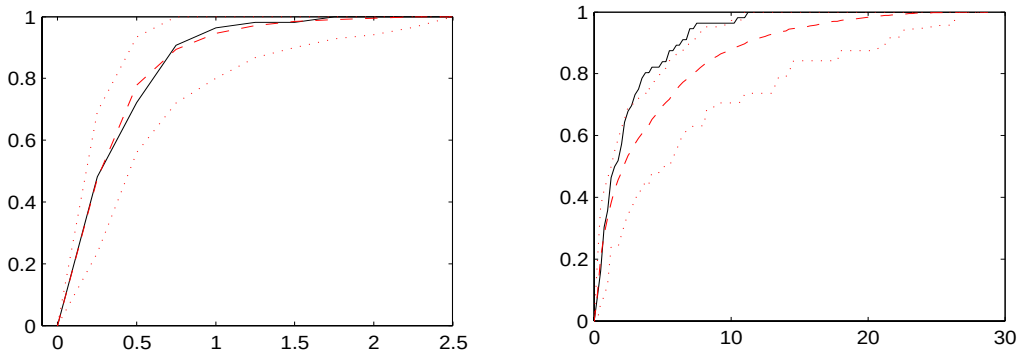


figure 3.25 : Fonctions de répartitions des durées de persistance des tempêtes (à gauche) et des inter-arrivées de tempêtes (à droite) au point r^0_{18} *(seuil* $16ms^{-1}$*). — données, - - modèle, · · · · intervalle de fluctuation à 95%.*

Ensuite, afin de vérifier si le modèle permet de reproduire la relation existant entre les différents points, nous avons comparé les structures d'ordre 2 des champs de vent observés et simulés. Sur les figure 3.26 et 3.27, nous avons représenté les estimations de $\Sigma(0)$ et $\rho(1)$ calculées à partir des séquences simulées et des données. Visuellement, ces matrices ont des caractéristiques proches. Notamment, la comparaison des matrices $\hat{\rho}(1)$ montre que le modèle semble reproduire le déplacement moyen des structures météorologiques vers l'est. On pourrait utiliser la méthode de test générale introduite au paragraphe pour tester si les écarts entre ces matrices est significatif ou non.

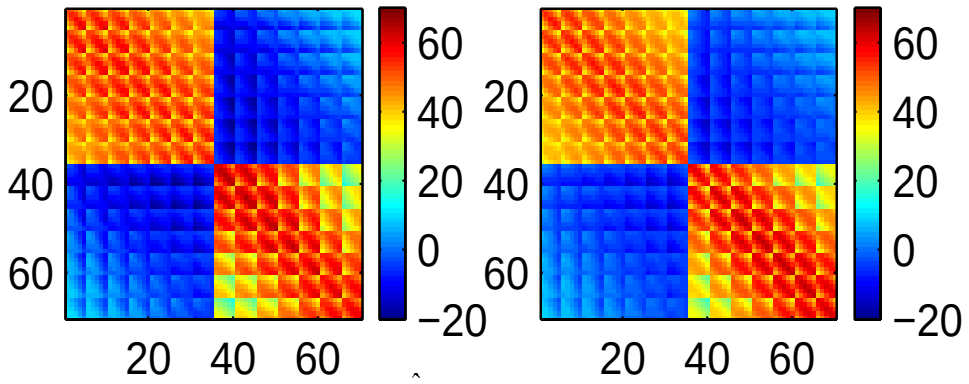


figure 3.26 : Matrice de covariance $\hat{\Sigma}(0)$ *(données à gauche et simulée à droite).*

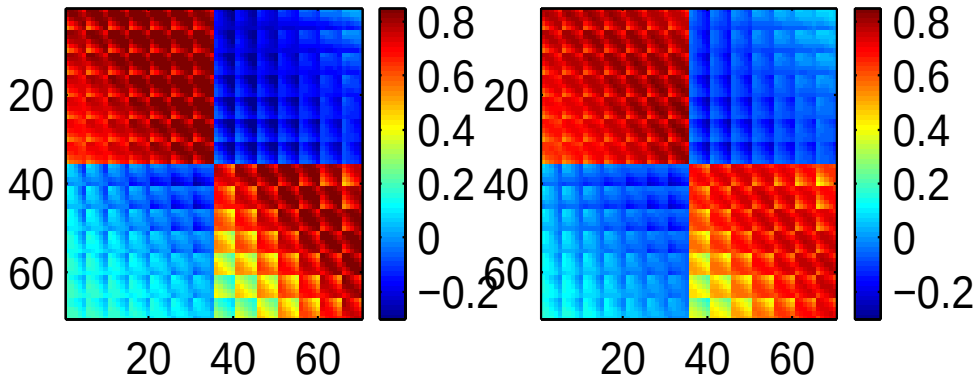


figure 3.27 : Matrice de corrélation $\hat{\rho}(1)$ *(données à gauche et simulée à droite).*

Evidemment, ces critères simples ne permettent pas de tester, par exemple, si la forme des champs de vent simulés sont physiquement réalistes. Par exemple, on peut se demander si le modèle permet de reproduire les formes des champs de vent dans les perturbations. Ces caractéristiques sont plus difficiles à quantifier.

Conclusion

Dans ce chapitre, nous proposons donc un modèle original pour décrire l'évolution spatio-temporel des champs de vent. Il s'agit d'un modèle $MS-AR$ dans lequel la variable cachée représente le déplacement des structures météorologiques. Nous avons montré que l'introduction de cette variable cachée permet d'obtenir des prévisions à un ou plusieurs pas de temps nettement meilleures qu'avec les modèles autorégressifs linéaires.

En revanche, la comparaison des séquences simulées avec ce modèle et de la série temporelle initiale donne des résultats décevants. Ces résultats pourraient sans doute être améliorés en utilisant d'autres paramétrisations. Nous avons ainsi mentionné au paragraphe 3.b.2 différentes limitations de celle qui a été choisie. Nous avons testé diverses paramétrisations alternatives sans succès. Ces tests sont relativement fastidieux, principalement à cause des temps de calculs nécessaires pour estimer les paramètres du modèle et du manque de critère numérique synthétique permettant de quantifier la qualité d'un modèle.

Enfin, dans ce chapitre, nous avons considéré uniquement une zone éloignée de la côte, ce qui permet d'éviter les problèmes liés à la déformation des champs de vent au-dessus de la terre. Ces déformations posent des problèmes notamment lorsque l'on veut calculer le déplacement des masses d'air (cf paragraphe 3.b.1) et estimer les matrices $A^{(s)}$ qui permettent d'extrapoler les champs de vent. Il faudrait alors réfléchir à l'implémentation de ce type de modèle en zone côtière.

Conclusion et perspectives

Dans cette thèse plusieurs modèles originaux sont proposés pour les séries temporelles de vent. Parmi ceux-ci une attention particulière est portée aux modèles autorégressifs à changements de régimes markoviens ($MS-AR$).

Nous proposons tout d'abord d'utiliser ce type de modèle pour décrire l'intensité du vent en un point fixe, la variable cachée étant introduite afin de modéliser les changements de régimes induits par les types de temps. L'originalité du modèle développé provient de la forme des modèles autorégressifs qui servent à décrire l'évolution du processus observé dans les différents régimes. En effet, ceux-ci sont paramétrés en utilisant une loi gamma (modèle $MS-\gamma AR$), ce qui permet de prendre en compte directement la positivité du processus considéré.

Ensuite, deux extensions de ce modèle sont développées, dans lesquelles la chaîne de Markov cachée devient non-homogène (modèle $NHMS-\gamma AR$). Dans la première d'entre elles, la matrice de transition dépend de la direction du vent, ce qui permet de modéliser la relation existant entre le type de temps et la direction du vent alors que dans la deuxième, nous supposons qu'il s'agit d'une fonction périodique, de période un jour, ce qui permet d'introduire les composantes journalières. D'autres extensions pourraient être envisagées par la suite.

Enfin, nous proposons un modèle $MS-AR$ pour décrire l'évolution spatio-temporelle des champs de vent, la variable cachée représentant le déplacement des structures météorologiques. Conditionnellement à ces déplacements, les champs de vent sont supposés suivre un modèle autorégressif linéaire avec innovations gaussiennes (modèle $MS-LAR$). Des formes paramétriques parcimonieuses sont utilisées pour ces modèles autorégressifs.

Dans le chapitre 1, les propriétés théoriques de ces différents modèles sont étudiées. Les résultats existants s'appliquent directement au modèle $MS-LAR$. On dispose ainsi de critères pratiques permettant de vérifier la stabilité de ce modèle ainsi que le bon comportement asymptotique des estimateurs du maximum de vraisemblance. Cependant, avec la paramétrisation choisie pour décrire les champs de vent, l'identifiabilité des paramètres du modèle n'a pu être vérifiée et on ne peut alors pas caractériser explicitement l'ensemble des limites possibles de la suite des estimateurs du maximum de vraisemblance. Ce problème pourrait être étudié plus précisément par la suite.

Par contre, les résultats existants ne s'appliquent pas directement aux modèles $MS-\gamma AR$, et nous démontrons alors différents résultats spécifiques. Tout d'abord, nous proposons des conditions garantissant la stabilité de ces modèles. Elles pourraient probablement être améliorées, notamment en ce qui concerne ceux d'ordre strictement supérieur à 1. Ensuite, nous nous intéressons aux propriétés asymptotiques des EMV. Après avoir prouvé l'identifiabilité des paramètres, nous montrons que ce modèle vérifie les conditions du théorème énoncé dans *Krishnamurthy et al.* (1998), ce qui implique la consistance des EMV. En ce qui concerne la normalité asymptotique, à notre connaissance le seul résultat existant est énoncé dans *Douc et al.* (2004), et les modèles $MS-\gamma AR$ ne vérifient pas les hypothèses faites par ces auteurs. Nous

proposons alors des hypothèses plus faibles, vérifiées par le modèle $MS-\gamma AR$, et qui semblent suffisantes dans le cas où le nombre de régimes est fini. La preuve de la normalité asymptotique des EMV sous ces hypothèses plus faibles n'est pas donnée dans cette thèse, mais pourrait être développée ultérieurement. Nous vérifions ensuite, en utilisant des techniques de Monte-Carlo, le bon comportement de ces estimateurs lorsque la longueur des séquences d'apprentissage est d'une taille comparable à celles des bases de données disponibles en pratique. L'étude théorique des modèles $NHMS-\gamma AR$, non abordée dans cette thèse, pourrait l'être postérieurement.

A la fin du premier chapitre, nous abordons un problème important en pratique, qui est celui de la sélection et de la validation de modèles. A notre connaissance, il n'existe pas de résultats théoriques sur la convergence des critères de sélection du type "log-vraisemblance pénalisée", tel que le critère BIC, pour ce type de modèle. Cependant, diverses études de Monte-Carlo ont permis de montrer le bon comportement de ce critère sur des données synthétiques. Nous avons pu vérifier qu'il en était de même sur les données de vent et que ce critère permet généralement de sélectionner des modèles parcimonieux qui s'ajustent bien aux données. Ensuite, nous proposons une méthode originale permettant de tester l'adéquation d'un modèle à des données. Il s'agit d'une méthode générale puisqu'elle est valable pour n'importe quel modèle pour peu qu'on sache en simuler des réalisations. Elle permet de comparer certaines statistiques de la séquence observée à celles correspondant aux séquences simulées et ainsi de vérifier le réalisme des séquences simulées avec le modèle à valider.

Dans le chapitre 2, nous décrivons plus précisément les résultats obtenus avec les modèles $MS-\gamma AR$ et $NHMS-\gamma AR$ pour des données de vent en un point situé près de la côte vendéenne. Après avoir vérifié l'interprétabilité météorologique des différents modèles obtenus, nous testons le réalisme des séquences simulées avec la procédure introduite à la fin du premier chapitre. Les résultats sont comparés avec ceux correspondant à deux autres modèles de simulation non-paramétriques, à savoir la méthode TGP qui suppose que le processus observé peut être rendu gaussien via une transformation simple et la méthode LGB qui est une méthode de bootstrap pour les séries temporelles. Nous mettons ainsi en évidence que les modèles $MS-AR$ permettent de modéliser certaines non linéarités présentes dans les séries temporelles d'intensité du vent et améliorent ainsi les résultats obtenus avec l'approche usuellement utilisée dans ce domaine (TGP). Les modèles $MS-AR$ semblent notamment être plus à même de reproduire la manière dont se succèdent les tempêtes et les périodes de calme, ce qui peut être important pour certaines applications, telle que l'évolution d'un trait de côte, par exemple. Une étude plus précise de ce type de phénomène pourrait être menée afin de vérifier l'influence de la chronologie des événements sur son évolution. Dans ce chapitre, nous abordons aussi le problème de la modélisation de la direction du vent. Aucun des modèles paramétriques proposés ne donne des résultats satisfaisants. Ce problème particulier pourrait être approfondi ultérieurement.

Enfin, dans le chapitre 3, nous décrivons plus brièvement les résultats obtenus avec le modèle $MS-LAR$ pour des données de champs de vent dans le Golfe de Gascogne. Il est montré que ce modèle permet d'améliorer nettement la qualité des prédictions à court terme par rapport à

un modèle autorégressif linéaire. Par contre, si nous utilisons la méthode de validation du premier chapitre, les résultats obtenus sont nettement moins convaincants, puisque, par exemple, les séquences simulées avec ce modèle ont des lois marginales et des fonctions d'autocorrélation significativement différentes de celles des données. Plusieurs limitations de la paramétrisation choisie sont mentionnées au chapitre 3, et la qualité des séquences simulées pourrait probablement être améliorée en utilisant des paramétrisations plus complexes. De telles paramétrisations pourraient être développées et testées ultérieurement.

Par ailleurs, un problème important en pratique, à savoir le calcul des conditions d'état de mer associées aux séries temporelles de vent simulées, n'a pas été mentionné dans cette thèse et est laissé en annexe. Une méthode de filtrage non-linéaire, utilisant des noyaux non-paramétriques, donne des résultats satisfaisants en pratique. Différents problèmes théoriques soulevés par cette méthode sont actuellement étudiés.

Enfin, la mise en oeuvre des méthodes présentées dans cette thèse a conduit au développement d'une boite à outil sous Matlab. Celle-ci regroupe des programmes permettant d'ajuster, de simuler puis de valider les différentes modèles présentés dans cette thèse. Une version plus élaborée de cette boîte à outil sera bientôt disponible, puis diffusée par l'IFREMER.

Bibliographie

[1] Aberg S. (2002). Modelling and prediction of wind fields using Gaussian Markov random fields and warping. *Master's thesis*, University of Lund.

[2] Ailliot P. (2000). Simulation de processus bivarié correspondant au couple hauteur significative/direction moyenne d'un état de mer. *Rapport de DEA*, Université de Rennes 1.

[3] Ailliot P., Prevosto M. (2001). Two methods for simulating the bivariate process of wave height and direction. *Proc. ISOPE conf.*

[4] Ailliot P., Prevosto M. (2003). Modélisation de l'évolution spatio-temporelle des champs de vent. *Actes des journées de statistiques.*

[5] Ailliot P., Prevosto M., Soukissian T., Diamanti C., Theodoulides A., Politis C. (2003). Simulation of sea state parameters process to study the profitability of a maritime line. *Proc. ISOPE Conf.*

[6] Ailliot P., Monbet V. (2004). A nonhomogeneous Markov switching autoregressive model for wind time series. *Submitted to the Scandinavian Journal of Statistics.*

[7] Arena F., Puca S., Tirozzi B. (2002). A new approach for the reconstruction of significant wave height time series. *Proc. OMAE.*

[8] Athanassoulis G.A., Stephanakos Ch.N. (1995). A non stationary stochastic model for long-term time series of significant wave height. *J. of Geophys. Res.*, vol 100, No C8, 00 16,149-16,162 .

[9] Baum L.E, Petrie T. (1966). Statistical inference for probabilistic function of functions of finite state Markov chains. *Ann. Math. Statis.*, vol 17, pp 1554-1663.

[10] Baum L.E, Petrie T., Soules G., Weiss N. (1970). A maximisation technique occurring in the statistical analysis of probabilistics functions of Markov chains. *Ann. Math. statis.*, vol 41, pp 164-171.

[11] Bellone E. (2000) Nonhomogeneous hidden Markov models for downscaling synoptic atmospheric patterns to precipitation amounts. *PhD Thesis*, University of Washington.

[12] Bénichou P. (1995). *Classification automatique de configurations météorologiques sur l'Europe occidentale*. Monographie, Météo France.

[13] Bentley (1975). Multidimensional binary search trees used for associative searching. *ACM*, vol 18, No 8.

[14] Berliner L.M., Wikle C.K., Cressie N. (2000). Long Lead prediction of Pacific SSTs via bayesian dynamic modelling. *J. of climate* , vol 13, pp 3953-3968.

[15] Besse P., Raimbault N. (2000). Comparisons of split linear fitting of wind curves. *Prépublication.*

[16] Bickel P.J., Ritov Y., Ryden T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, vol 26, No 4, pp1614-1635.

[17] Booth J.G, Hobert J.P., Jank W. (2001). A survey of Monte Carlo algorithms for maximising the likelihood of a two-stages hierarchical model. *Statist. Modelling*, vol 1, pp 333-349.

[18] Borgman L.E., Sheffner N.W. (1991). Simulation of time sequences of wave height, period and direction. *Technical report US Army Corps of Engineers*.

[19] Boucheron S., Gassiat E. (2003). Order estimation and model selection. *Preprint*.

[20] Bougerol P., Picard N. (1992). Strict stationarity of generalized autoregressive processes. A*nn. Probab.* vol 20, pp 1714-1730.

[21] Boukhanovsky A.V., Krogstad H.E., Lopatoukhin L.J. , Rozhkov V.A. (2003). Stochastic simulation of inhomogeneous metocean fields. Part I: annual variability. *International Conference on Computational Science* , Springer Verlag, pp 213-222.

[22] Boukhanovsky A.V., Krogstad H.E., Lopatoukhin L.J., Rozhkov V.A. Athanassoulis G.A., Stephanakos C.N. (2003). Stochastic simulation of inhomogeneous metocean fields. Part II: synoptic variability and rare events. *International Conference on Computational Science*, Springer Verlag, pp 223-233.

[23] Box G.E.P., Jenkins G.M. (1976). *Time series analysis, forecasting and control (revised edn.)* Holden-Day, San Francisco.

[24] Brandt A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Adv Appl. Probab*, vol 18, pp 211-220.

[25] Breckling J. (1989). *The analysis of directional time series: applications to wind speed and direction*. Lecture Notes in Statistics 61, Springer-Verlag: Berlin

[26] Brockwell, P., Davis, R. (1991). *Time series: theory and methods, 2nd edition*. Springer Verlag, New York.

[27] Brown B.G., Katz R.W, Murphy A.H. (1984). Time series models to simulate and forecast wind speed and wind power. *J. of Clim. and Appl. Meteor.*, vol 23, pp 1184-1195 .

[28] Buckle B., Haugh D., Thomson P. (2002). On growth and volatility regime switching models for New Zealand GDP data. *Working paper*.

[29] Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, vol 17, pp 52-72.

[30] Casson E., Coles S. (1998). Extreme hurricane wind speeds: estimation, extrapolation and spatial smoothing. *J. of Wind Engineering and Industrial Aerodynamics*, vol 74-76, pp 131-140.

[31] Castino F. , Festa R., Ratto C.F. (1998). Stochastic modelling of wind velocities time series, *J. of Wind Engineering & Industrial Aerodynamics,* vol 74-76, pp 141-151.

[32] Celeux G., Chauveau D., Diebolt J. (1995). On stochastic versions of the EM algorithm.

R*apport de recherche INRA*.

[33] Chang R.W, Hancock J.C. (1966). On receiver structures for channels having memory. *IEEE trans inform Theory,* vol IT-12, pp 463-468.

[34] Corpetti T. (2002). Estimation et analyse de champs denses de vitesses d'écoulements fluides. *Phd Thesis*, Université de Rennes 1.

[35] Cressie N. (1993). *Statistics for spatial data*. Wiley.

[36] Cunha C. et Guedes Soares C. (1999). On the choice of data transformation for modelling time series of significant wave height. *Ocean Engineeting*, vol 26, pp 489-506.

[37] Daniel A.R., Chen A.A. (1991). Stochastic simulation and forecasting of hourly average wind speed sequences in jamaica. *Solar Energy*, vol 46, No 1, pp 1-11.

[38] Dempster A.P. , Laird N.M. and Rubin D.B. (1977). Maximum Likelihood from Incomplete Data via The EM Algorithm, *J. of Royal Statistical Society*, vol 39, pp 1-38.

[39] Deo M.C, Shridar Naidu C. (1999). Real time wave forecasting using neural networks. *Ocean Engineering*, vol 26, pp 191-203.

[40] Dietrich C.R., Newsman G.N. (1997). Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *Siam J. Sci Comput*, vol 18, No 4, pp 1088-1107.

[41] Douc R., Matias C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernuoulli*, vol 7, No 3, pp 381-420.

[42] Douc R., Moulines E., Ryden T. (2004). Assymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *To appear in the Annals of Statistics*.

[43] Durand, J.B. (2003). Modèles à structure cachée : inférence, estimation, sélection de modèles et applications . *Thèse de doctorat*, Univerité Joseph Fourier.

[44] Efron B. (1979). Bootstrap methods : another look at the Jacknife. *Ann. Statist,* vol 7, pp 1-26.

[45] Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman Hall, New York.

[46] Ephraim, Y., Merhav N. (2002). Hidden markov process. *IEE Transactions on Information Theory*, vol 48, No 6, pp 1518-1569 .

[47] Fisher N.I, Lee A.J. (1994). Time series analysis of circular data. *J.R Statist. Soc. B*, vol 56, No 2, pp 327-339.

[48] Fonseca G. (2000). Stability and estimation of nonlinear time series modes. *PhD Thesis*.

[49] Francq C., Roussignol M. (1998). Ergodicity of autoregressive processes with Markov-

switching and consistency of the maximum-likelihood estimator, *Statistics*, vol 32, pp 151-173.

[50] Gioffre M., Gusella V., Griogriu M. (2000). Simulation of non-Gaussian field applied to wind pressure fluctuations. *Probabilistic Engineering Mechanics*, vol 15, pp 339-345.

[51] Graflund A., Nilsson B. (2002). Dynamic Portfolio Selection: The Relevance of Switching Regimes and Investment Horizon. *Working Papers* 2002:8, Lund University, Department of Economics.

[52] Grunwald G., Hyndman R.J. , Tedesco L., Tweedie R.L. (1999). Non-Gaussian conditional linear AR(1) models. *Australian and New Zealand Journal of Statistics*, vol 42(4), pp 479-495.

[53] Gueguan D. (1994). *Séries chronologiques non linéaires à temps discret*. Econmica.

[54] Guttorp P. (1996). *Stochastic modeling of rainfall*. In M.F. Wheeler (editor): Environmental studies: mathematical, computational and statistical analysis, pp 171-187. New-York : Springer.

[55] Hamilton J.D. (1989). A new approach to economic analysis of nonstationary time series and the business cycle. *Econometrica*, vol 57, pp 357-384.

[56] Hamilton J.D. (1990). Analysis of time series subject to changes in regime, *J. Economet.*, vol 45, pp 39-70.

[57] Hamilton J.D. (1993). Estimation, inference, and forecasting of time series subject to changes in regime. *Handbook of Statistics*, vol 11, edited by G. S. Maddala, C. R. Rao, and H. D. Vinod, North-Holland, pp 231-260.

[58] Hamilton J.D. (1996). Specification testing in markov-switching time-series models. *J. Economet.*, vol 70, pp 127-157

[59] Hardy D.E., Wlaton J.J. (1978). Principal components analysis of vector wind measurements. *J. of Applied Metorology*. vol 17, pp 1153-1162.

[60] Hawkins, D.S, Allen D.M, Stromberg A.J. (2001). Determining the number of components in mixture of linear model. *Computational Statistics & Data Analysis*, vol 38, pp 16-48.

[61] Hili O. (1992). Sur les modèles autorégressifs à seuil. *CRAS*, pp 573-576.

[62] Holst U., Lindgren G., Holst J., Thuvelsholmen M. (1994). Recursive estimation in switching autoregressions with a Markov Regime. *J. of Time Series Analysis*, vol 15, No 5, pp 489-505.

[63] Hugue J.P, Guutorp P., Charles S.P. (1999). A non homogeneous hidden markov model for precipitation occurrence. *Appl. Statis.*, vol 48, part 1, pp 15-30.

[64] Hugues J.P. (1997). Computing the observed information matrix in the hidden Markov model using the EM algorithm. *Statistics & Probability Letters*, vol 32, pp 107-114.

[65] Jensen J.L., Petersen N.V. (1999). Asymptotic normality of the maximum likelihood esti-

mator in state space model. *Ann. Statis.*, vol 27, No 2, pp 514-535.

[66] Jolliffe I.T., Uddin M., Vines S.K. (2002). Simplified EOFs-three alternatives to rotation. *To appear in Climate Res.*

[67] Kingman J.F.C. (1976). *Subadditive processes*. Ecole d'été de Probabilités de Saint-Flour. (1975). Lecture notes in mathematics 539. Berlin, Springer.

[68] Krishnamurthy V., Ryden T. (1998). Consistent estimation of linear and non-linear autoregressive models with markov regime. *J. of Time Series Analysis*. vol 19, No 3, pp 291-307.

[69] Krolzig H.M. (1997). *Markov-switching vector Autoregressions. Modelling, statistical inference and applications to business cycle analysis*. Lecture notes in economics and mathematical systems 454. Springer-Verlag, Berlin.

[70] Kunsch H.R. (2001). State space and hidden Markov models. *Complex Stochastic Systems* (O.E. Barndorff-Nielsen, D.R. Cox and C. Kluppelberg, eds.). Chapman&Hall/CRC Press Boaca Raton, FL, 109-173.

[71] Labeyrie J. (1990). Stationary and transient states of random seas. *Marine Structures*., vol 3, pp 43-58.

[72] Legland F., Mevel L. (2000). Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems*, vol 13, pp 63-93.

[73] Leroux B.G. (1992). Maximum-likelihood estimation for hidden markov models. *Stochastic Process and their Applications*, vol 40, pp 127-143.

[74] Louis T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. soc. B*, vol 44, pp 226-233.

[75] Mardia K.V. (1972). *Statistics of Directional Data.* Academic Press, New York.

[76] MacDonald I. L., Zucchini W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. New York: Chapman and Hall.

[77] MacKay R.J. (2002). Estimating the order of a hidden Markov model. *The Canadian Journal of Statistics*, vol 30, pp 573-589

[78] MacKay R.J. (2002). Assessing the goodness-of-fit of hidden markov models. *Biometrics*, vol 60, pp 444-450.

[79] Marteau P.F, Monbet V., Ailliot P. (2004). NonParametric modeling of cyclo-stationary markovian process. Part II: prediction and dimension reduction, *Proc. ISOPE Conf.*, vol. III.

[80] Mayencon R. (1982). *Météorologie marine*. Editions Maritimes et d'Outre-Mer.

[81] Meng X.L., Rubin D.B. (1995). Maximum likelihood estimation via the ECM algorithm : a general framework. *Biometrika,* vol 80, No 2, pp 267-278.

[82] Meng X.L. (1994). On the rate of convergence of the E.C.M. algorithm. *The annals of sta-*

tistics, vol 22, No 1, pp 326-339.

[83] Mevel L. (1997). Statistique asymptotique pour les modèles de Markov cachés. *Thèse de doctorat*, Université de Rennes 1.

[84] Meyn S.P., Tweedie R.L. (1993). *Markov chains and stochastic stability*. Springer-Verlag, London.

[85] Monbet V., Prevosto M. (2001). Bivariate simulation of non stationary and non gaussian observed processes. Application to sea state parameters. *Applied Ocean Research*, vol 23, pp 139-145.

[86] Monbet V., Iovleff S. (2001). Comparaison de méthodes non linéaires pour la simulation de processus d'état de mer. *Actes des journées de Statistique*.

[87] Monbet M., Marteau P.F. (2001). Continuous Space Discrete Time Markov Models For Multivariate Sea State Parameters Process. *Proc. ISOPE conf.*

[88] Monbet M., Marteau P.F. (2003). The local grid bootstrap for stationary multivariate markov processes. *To appear in J. of Statistical Planning and Inference*.

[89] More A, Deo M.C. (2003). Forecasting wind with neural networks. *Marine structures*, vol 16 , pp 35-49.

[90] Nfaoui H., Buret J., Sayigh A.A. (1996). Stochastic simulaton of hourly average wind speed sequences in Tangiers (Morocco). *Solar Energy*, vol 56, No 3, pp 301-314.

[91] O'Carroll (1984). Weather Modelling for Offshore Operations. *The Statistician*, 33, pp 161-169.

[92] Politis D. N., (2003). The impact of bootstrap methods on time series analysis. *Statistical Science* , vol 18, No 2, pp 219-230.

[93] Popescu R., Deodatis G., Prevost J.H. (1998). Simulation of homogeneous nonGaussian stochastic vector fields. *Prob. Engng. Mech,* vol 13, No 1, pp 1-13.

[94] Psaradakis Z., Spagnolo N. (2003). On the determination of the number of regimes in Markov-switching autoregressive model. *Journal of Time Series Analysis*, vol 24, No 2, pp 237-252.

[95] Rabiner L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, vol 77, pp 257-286.

[96] Risk Engineering, Inc. (2002). Estimation of 10-4 waves using synthetic storms (phase II-North sea). *Technical report*, joint industry project.

[97] Robert C.P. (1996). *Méthodes de Monte Carlo par chaînes de Markov.* Economica, Paris.

[98] Robert C.P., Celeux G., Diebolt J. (1993). Bayesian estimation of Hidden Markov chains: a stochastic implementation, *Statist. Probab. Lett.*, vol 16, pp 77-83.

[99] Rychlik I., Johannesson P., Leadbetter M.R. (1997). Modelling and statistical analysis of ocean-wave data using transformed gaussian processes. *Marine Structures*, vol 10, pp 13-47.

[100] Ryden T., Terasvirta T., Asbrink S. (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econ.*, vol 13, pp 217-244.

[101] Rynkiewicz J. (2000). Estimation de modèles autorégressifs à changements de régime markovien. Cahiers de la MSE No 60.

[102] Sahin A.J, Sen Z. (2001). First-order Markov chain approach to wind speed modelling. *J. of Wind Eng.. and Indust. Aerodyn.*, vol 89, pp 263-269.

[103] Saxton J., Swensen A.R. (1999). ECM-algorithms that converge at the rate of EM. *Discussion papers,* No244, Statistics Norway, research department .

[104] Scotto M.G., Guedes Soares C. (2000). Modelling the long-term time series of significant wave height with non-linear threshold models. *Coastal Engineering,* vol 40, pp 313-327.

[105] SHOM, Instructions nautiques.

[106] Silverman B.W. (1986). *Density estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall.

[107] Smith R.L. (2003). *Environmental statistics*. Preprint.

[108] Soukissian, T.H., Prospathopoulos, A., Diamanti, C. (2002). Wave and wind data analysis of the Aegean Sea. Preliminary results. *The Global Atmosphere and Ocean Systems*. vol 8, No 2-3, pp 163-189.

[109] Stephanakos Ch.N. (1999). Nonstationary stochastic modelling of time series with applications to environmental data. *PhD Thesis*, National Technical University of Athens.

[110] Stephos A. (2000). A comparison of various forecasting techniques applied to mean hourly wind time series. *Renewable Energy,* vol 21, pp 23-35.

[111] Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics*, vol 6, pp 461- 464.

[112] Teicher H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.*, vol 32, pp 244-248.

[113] Teicher H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.*, vol 38, No 4, pp 1300-1302.

[114] Toll R.S.J. (1997). Autoregressive conditional heteroscedasticity in daily wind speed measurements. *Theor Appl. Climatol.*, vol 56, pp 113-122.

[115] Tong H (1990). *Non-linear time series. A dynamical system approach*. Oxford University Press.

[116] Tucker, M.J. (1991). *Waves in ocean engineering: measurement, analysis, interpretation.*

Ellis Horwood Series in Marine Science.

[117] Vik I. (1981). The tile series generating module-Modelling aspects. Ocean Research- operational criteria, CNRD 3-2 task 2.

[118] Waeles B., Le Hir P., Silva Jacinto R. (2004). Modélisation morphodynamique cross-shore d'un estran vaseux. *C. R. Geoscience*, vol 336, pp 1025-1033.

[119] Walton T.L., Borgman, L.E. (1990). Simulation of non-stationary, non-gaussian water levels on the great lakes, *J. of Waterways, Ports, Coastal and Ocean Division*, ASCE, vol 116, No 6.

[120] Wei G.C., Tanner M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm, *J. of the American Statistical Association*, vol 85, pp 669-704.

[121] Wikle C.K., Millif R.F., Nychka D., Berliner L.M. (1999). Spatio-temporal hierarchical bayesian modelling: tropical ocean surface winds. *J. of the American Statistical Association*, vol 96, pp 382-397.

[122] Willems P. (2001). A spatial rainfall generator for small spatial scales. *J. of Hydrology*, vol 252, pp 126-144.

[123] Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Ann. Stat*, vol 11, pp 95-103.

[124] Yao J.F., Attali J.G. (2000). On stability of nonlinear AR processes with markov switching. *Adv. Appl. Prob.*, vol 32, pp 394-407.

[125] Yao J.F. (2001). On square-integrability of an AR process with Markov switching. *Statistics & Prabability Letters*, pp 265-270.

[126] Yim J.Z., Chou C., Ho P. (2002). A study on simulating the time series of significant wave heigth near the keelung harbor. *Proc. ISOPEE Conf.*

[127] Zhang J., Stine R.A. (2001). Autocovariance structure of markov regime switching models and model selection. *Preprint.*

Annexe A : Bases de données

Définition des paramètres océano-météo d'intérêt

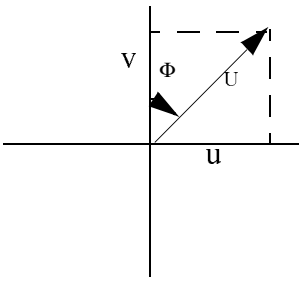
Un état de mer est généralement décrit par plusieurs paramètres synthétiques. Dans cette thèse, seuls les 7 paramètres ci-dessous sont considérés :

Paramètres pour le vent

- U : **intensité du vent en mètre par seconde**. Ce paramètre est à valeurs dans $\mathbf{R}^+$. Il représente usuellement l'intensité moyenne du vent sur une période variant de 10 minutes à 1 heure selon la base de données.
- Φ : **direction du vent en degré**. Ce paramètre est à valeurs dans le tore $\boldsymbol{R}/360\boldsymbol{Z}$. Il représente l'écart angulaire, en degré, entre le nord et la direction moyenne de laquelle vient le vent (sens anti-trigonométrique). Par exemple, $\Phi = 90$ correspond à un vent venant de l'est et $\Phi = 180$ à un vent de sud.
- (u, v) : **composante zonale** du vecteur vent moyen (positive dans le sens ouest-est) et **composante méridienne** du vecteur vent moyen (positive dans le sens sud-nord). Ces deux paramètres sont exprimés en mètre par seconde.

Les paramètres U, Φ, u et v sont liés par la relation

$$(u, v) = \left(U\cos\left(\frac{2\pi}{360}(270-\Phi)\right), U\sin\left(\frac{2\pi}{360}(270-\Phi)\right)\right)$$



Paramètres d'états de mer

- H_s : **hauteur significative en mètre.** Il s'agit d'un paramètre à valeurs dans $\mathbf{R}^+$. Il est défini à partir du spectre d'énergie du champs de vague. Sous certaines conditions (spectre à bande étroite), il est approximativement égal à la moyenne du tiers des plus hautes vagues.
- Θ_m : **direction moyenne de propagation de l'état de mer en degré.** Ce paramètre est à valeur dans le tore $\boldsymbol{R}/360\boldsymbol{Z}$. Il est aussi défini à partir du spectre d'énergie. Intuitivement, il représente la direction moyenne dans laquelle vont les vagues.
- T_m : **période en seconde**. Paramètre à valeurs dans $\mathbf{R}^+$. Il existe plusieurs définitions pour la période d'un état de mer. Selon l'application et la base de données, T_m peut désigner la période pic, la période moyenne...

Cette liste est loin d'être exhaustive, mais il s'agit sans doute des paramètres les plus couramment utilisés en pratique. Une liste plus complète est donnée dans *Tucker* (1991).

Bases de données utilisées

Différentes méthodes peuvent être utilisées pour réaliser des mesures directes (ou "in-situ") des paramètres d'état de mer : satellites, bouées, bateaux... Cependant, ces données ne sont généralement pas directement exploitables. En ce qui concerne les satellites, les données sont disponibles avec une discrétisation temporelle trop large (de l'ordre de quelques jours entre deux mesures successives au même point). Les bouées fournissent des données en continu, mais les bases de données sont généralement disponibles sur des périodes de temps relativement courtes et il peut y avoir de nombreuses données manquantes. Notons enfin que pour le vent, il existe de nombreuses bases de données pour des points situés à terre (sémaphores, aéroports...).

Pour cette thèse, nous avons préféré utiliser des données issues de modèles météorologiques. La qualité de ces données dépend de plusieurs facteurs, et en particulier du modèle lui-même ainsi que de la manière dont sont assimilées les mesures "in-situ". On parle de données de "forecast" quand uniquement les données disponibles jusqu'à l'instant t sont utilisées pour prévoir la situation aux instants t+1, t+2..., de données de "nowcast" lorsqu'à chaque date toutes les données in-situ disponibles jusqu'à cette date sont utilisées et de "hindcast" lorsque les données passées, présentes et futures sont utilisées pour prévoir la situation à un instant donné. Ce sont donc les données de hindcast qui sont les plus fiables. L'évolution en temps et en espace des champs de vent dépend de la climatologie de la région considérée. Les modèles développés dans cette thèse l'ont été principalement pour une région de référence, à savoir le golfe de Gascogne. Il n'est alors pas évident que ces modèles soient adaptés pour d'autres régions. Notons cependant que dans le cadre d'un projet Européen, les méthodes développées ont été aussi testées pour des données en mer Egée (cf *Ailliot et al.* (2004, [5])). Trois bases de données, disponibles à l'IFREMER, permettent de décrire les conditions d'états de mer dans ces deux régions.

- ***La base de données fournie par ECMWF*** (European Centre for Medium-Range Weather Forecast). Elle décrit les champs de vents sur tout le globe avec une donnée toutes les 6 heures depuis 1992. Les données sont fournies sur une grille fixe avec une discrétisation spatiale de 1.125 degré en latitude et longitude (cf figure 1). En particulier, cette base de données couvre les deux régions d'étude (Golfe de Gascogne et Mer Egée). Ce sont des données de nowcast. Elles sont utilisées au chapitre 3.

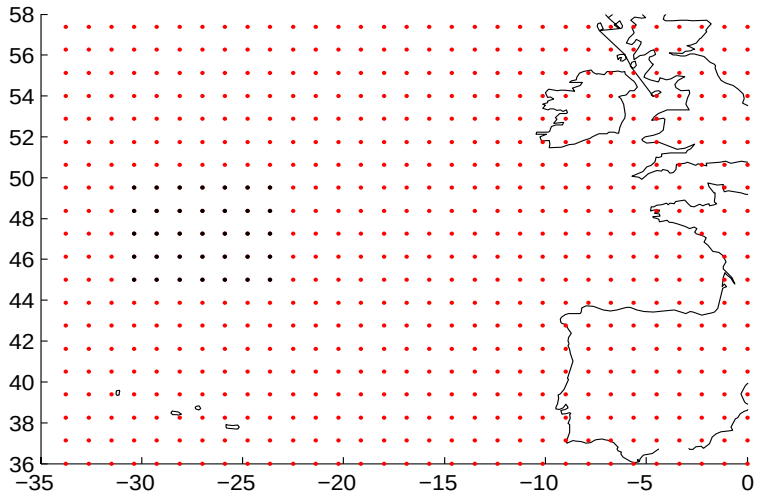


figure A.1 : Grille du modèle ECMWF.

- ***La base de données AES40 fournie par OCEANWEATHER***. Dans cette base de données, on a accès à de nombreux paramètres synthétiques d'états de mer, et en particulier aux 7 paramètres définis ci-dessus. Ces données sont disponibles toutes les 6 heures, depuis 1979, sur une grille fixe le long de la côte Atlantique française (cf figure 2). Il s'agit de données de Hindcast. Elles sont utilisées au chapitre 2.

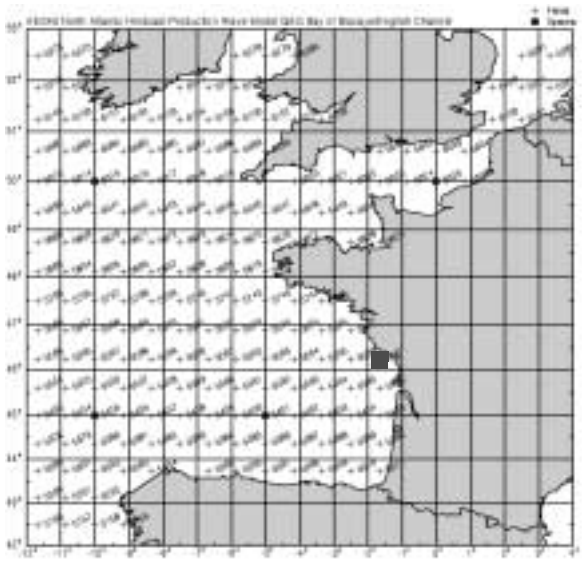


figure A.2 : Grille du modèle AES40. Les données utilisées au chapitre 2 concernent le point de coordonnées (46.25N , 1.667E) *représenté par un carré rouge.*

- ***La base de données fournie par le NCMR*** *(National Center for Marine Research)*. Cette base a été fournie dans le cadre d'un projet Européen Egide. Elle décrit les conditions d'état de mer depuis Septembre 1999 en Mer Egée. En particulier, les 7 paramètres définis ci-dessus sont disponibles, avec une donnée toutes les 3 heures et tous les $1/20$ degré. Ce sont

des données de forecast : tous les jours, à minuit, les données in-situ disponibles sont utilisées pour établir une prévision pour la journée suivante. Un discussion sur la qualité de ces données peut être trouvée dans *Soukissian et al* (2002). Cette base de données est utilisée dans Ailliot et *al.* (2004).

Annexe B : Quelques lemmes techniques sur la loi gamma

Etude de la fonction gamma

La fonction gamma est définie sur $\boldsymbol{R}^{+*}$ par

$$\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt$$

Cette fonction est décroissante sur $[0, x_0]$ puis croissante sur $[x_0, +\infty[$ avec $1 < x_0 < 2$ et $\gamma_- = min_{x \in R^+} \Gamma(x) = \Gamma(x_0) = 0,885\ldots$. Un développement asymptotique de cette fonction au voisinage de $+\infty$ est donné par (formule de Stirling) :

$$\Gamma(x) = \sqrt{2\pi} e^{-x} x^{x-1/2} \left(1 + \frac{1}{12x} + o\left(\frac{1}{x}\right)\right)$$

Notons $\Psi(x) = \dfrac{\Gamma'(x)}{\Gamma(x)}$ la dérivée logarithmique de la fonction Γ. On a alors la relation :

$$\Psi(x+1) = \Psi(x) + \frac{1}{x} \text{ pour } x > 0$$

On en déduit en particulier que:

$$\frac{\Gamma'(x)}{\Gamma(x)} \sim_{\infty} \ln(x)$$

Moments de la loi gamma

La densité de la loi gamma est donnée par

$$\gamma(y;\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)} e^{-y/\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} \mathbf{1}_{R^+}(y)$$

avec α et β des paramètres strictement positifs. Le moment d'ordre $c > 0$ d'une loi gamma de paramètres (α, β) est donné par

$$\gamma_c = \int_0^{+\infty} y^c \gamma(y;\alpha,\beta) dy = b^{\alpha} \frac{\Gamma(c+\alpha)}{\Gamma(\alpha)} \qquad (2.1)$$

En particulier, sa moyenne μ vaut $\alpha\beta$ et son écart type σ vaut $\beta\sqrt{\alpha}$. Les deux premiers moments de la loi gamma détermine donc les paramètres de manière unique. Cette remarque permet de paramétrer la densité de la loi gamma par μ et σ via le changement de variable $\alpha = (\mu/\sigma)^2$ et $\beta = \sigma^2/\mu$. Le lemme ci-dessous donne le comportement asymptotique du moment d'ordre α d'une loi gamma de moyenne μ et de variance σ lorsque $\mu \to +\infty$. Il est utilisé pour démontrer la proposition 1.8, qui donne des conditions garantissant l'existence de moment pour la loi stationnaire du modèle $MS - \gamma AR$.

Lemme B.1

Soit $\gamma_c(\mu, \sigma)$ *le moment d'ordre* $c > 0$ *d'une variable aléatoire de loi gamma de moyenne* $\mu > 0$ *et d'écart type* $\sigma > 0$. *On a alors* $\gamma_c(\mu, \sigma) = \mu^c\left(1 - \frac{c\sigma^2}{2\mu^2} + o_\infty\left(\frac{1}{\mu^2}\right)\right)$.

Preuve:

On utilise le fait que $\gamma_c(\mu, \sigma) = \left(\frac{\sigma^2}{\mu}\right)\frac{\Gamma(c + \mu^2/\sigma^2)}{\Gamma(\mu^2/\sigma^2)}$ et le développement asymptotique de la fonction Γ.

□

Le lemme suivant porte aussi sur la majoration des moments de la loi gamma. Il est utilisé pour démontrer la condition **(K2)** puis la consitance des EMV.

Lemme B.2

Notons $l(\alpha, \beta) = E[|\ln X|]$ *pour* X *une variable aléatoire suivant une loi gamma de paramètres* $(\alpha, \beta) \in (R^{+*})^2$. *On a alors la majoration suivante:*

$\forall \alpha_0 > 0,\ \exists A, B > 0$ *tels que* $\forall \alpha > \alpha_0\ \ \forall \beta > 0\ \ l(\alpha, \beta) \le A\ln(\alpha) + B + |\ln(\beta)|$

Preuve:

Remarquons tout d'abord que $l(\alpha, \beta) = E\left[\left|\ln\left(\frac{X}{\beta}\right) + \ln(\beta)\right|\right] \le l(\alpha, 1) + |\ln(\beta)|$. Par ailleurs, on a :

$$l(\alpha, 1) = -1/\Gamma(\alpha)\int_0^1 \ln(x)x^{\alpha-1}e^{-x}dx + 1/\Gamma(\alpha)\int_1^\infty \ln(x)x^{\alpha-1}e^{-x}dx\ .$$

On en déduit aisément que $l(\alpha, 1) \sim_\infty \frac{\int_0^\infty \ln(x)x^{\alpha-1}e^{-x}dx}{\Gamma(\alpha)} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \sim_\infty \ln(\alpha)$. On conclut alors en utilisant le fait que, d'après le théorème de convergence dominée, la fonction $\alpha \to l(\alpha, 1)$ est continue sur R^{+*}.

□

Majorations pour le noyau du modèle γAR

Le modèle γAR est défini dans le paragraphe 1.a.2. On dit que le processus $\{Y_t\}$ suit un modèle γAR lorsque, conditionnellement à son passé, Y_t suit une loi gamma de moyenne $\mu(\bar{Y}_{t-1}) = \bar{a}\bar{Y}_{t-1} + b$ et de variance σ^2 avec $\bar{a} = (a_1, \ldots, a_r) \in (\boldsymbol{R}^+)^r$, $b \in \boldsymbol{R}^{+*}$ et $\sigma \in \boldsymbol{R}^{+*}$ des paramètres. On a alors

$$p(y_t|\bar{y}_{t-1}) = \frac{\mu(\bar{y}_{t-1})}{\sigma^2\Gamma\left(\left(\frac{\mu(\bar{y}_{t-1})}{\sigma}\right)^2\right)}\left(\frac{y_t\mu(\bar{y}_{t-1})}{(\sigma)^2}\right)^{\left(\frac{\mu(\bar{y}_{t-1})}{\sigma}\right)^2-1}\exp\left(-\frac{y_t\mu(\bar{y}_{t-1})}{(\sigma)^2}\right)\mathbf{1}_{R^+}(y_t)$$

Ce modèle ne vérifie pas les hypothèses couramment faites pour les modèles autorégressifs. En effet, on suppose généralement que (modèle NAR) :

$$Y_t = f(\bar{Y}_{t-1}) + \varepsilon_t$$

avec $\{\varepsilon_t\}$ une séquence i.i.d , et les modèles γAR ne peuvent pas s'écrire sous cette forme. Afin d'étudier la stabilité des modèles NAR, on est généralement amené à faire des hypothèses sur le bruit $\{\varepsilon_t\}$. Par exemple, si on suppose que cette densité est bornée, on vérifie alors aisément que $sup_{\bar{y}_0, y_1} p_\theta(y_1|\bar{y}_0) < \infty$ et on en déduit que la chaîne de Markov associée est fortement fellerienne.

On vérifie aisément que pour les modèles γAR, on n'a plus $sup_{\bar{y}_0, y_1} p_\theta(y_1|\bar{y}_0) < \infty$ dès que $\frac{\mu}{\sigma} < 1$, et il faut alors utiliser des majorations plus fines. On déduit aisément du lemme B.3 ci-dessous que les noyaux de transition associés aux modèles $MS - \gamma AR$ sont fortement felleriens, ce qui sert pour démontrer la stabilité des modèles $MS - \gamma AR$ (cf proposition 1.7).

Lemme B.3

Soit $\{Y_t\}$ *un processus suivant un modèle* γAR *de paramètre* $(a_1, \ldots, a_r, b, \sigma) \in (\boldsymbol{R}^+)^r \times \boldsymbol{R}^{+*} \times \boldsymbol{R}^{+*}$. *Soit* $M > 0$ *et* $\|.\|$ *une norme quelconque sur* $\boldsymbol{R}^r$. *Il existe une fonction* h *définie sur* $\boldsymbol{R}^{+*}$, *intégrable et telle que* $\forall y_1 \in \boldsymbol{R}^{+*}$:

$$sup_{\|\bar{y}_0\| \le M} p(y_1|\bar{y}_0) \le h(y_1)$$

Preuve

Il suffit de montrer que le lemme est vrai avec la norme euclidienne $\|x\| = \sqrt{x'x}$. On vérifie alors que la fonction

$$h(y) = \frac{\mu_+}{\sigma^2\gamma_-}\exp\left(-\frac{y\mu_+}{\sigma^2}\right) max\left(\left(\frac{\mu_+ y}{\sigma^2}\right)^{\left(\frac{\mu_+}{\sigma}\right)^2-1}, \left(\frac{\mu_- y}{\sigma^2}\right)^{\left(\frac{\mu_-}{\sigma}\right)^2-1}\right)$$

avec $\mu_+ = \|\bar{a}\|M + b$, $\mu_- = b$ convient

□

Le lemme ci-dessous est utilisé pour montrer que les modèles $MS-\gamma AR$ vérifient la condition **(K5)** (cf proposition 1.13).

Lemme B.4

Soit $\{Y_t\}$ *un processus suivant un modèle* γAR *de paramètre* $(a_1, \ldots, a_r, b, \sigma) \in (\boldsymbol{R}^+)^r \times \boldsymbol{R}^{+*} \times \boldsymbol{R}^{+*}$. *On a alors,* $\forall \varepsilon > 0$,

$$sup_{y_1 > \varepsilon} sup_{\bar{y}_0 \in (\boldsymbol{R}^+)^d} p(y_1 | \bar{y}_0) \leq max\left(\frac{1}{b}, \frac{1}{b}\left(\frac{\varepsilon b}{\sigma^2}\right)^{\left(\frac{b}{\sigma}\right)^2 - 1}, \frac{1}{\sigma^2}\right)$$

Preuve

Notons $\tilde{\gamma}(x;\mu,\sigma)$ la densité de loi gamma de moyenne $\mu > 0$ et variance σ^2 et fixons $\varepsilon > 0$. Il est facile de vérifier que :

$$sup_{y_1 > \varepsilon} sup_{\bar{y}_0 \in (\boldsymbol{R}^+)^d} p(y_1 | \bar{y}_0) \leq sup_{x > \varepsilon} sup_{\mu > b} \tilde{\gamma}(x;\mu,\sigma)$$

Soit $\mu \geq b$. L'étude des variations de la fonction $x \to \tilde{\gamma}(x;\mu,\sigma)$ montre que

- cette fonction est décroissante sur $\boldsymbol{R}^+$ si $\mu^2 < \sigma^2$, et par conséquent, si cette condition est vérifiée, on a :

$$\begin{aligned} sup_{x \geq \varepsilon} \tilde{\gamma}(x;\mu,\sigma) &\leq \tilde{\gamma}(\varepsilon;\mu,\sigma) \\ &\leq \frac{1}{\mu}\left(\frac{\varepsilon\mu}{\sigma^2}\right)^{\left(\frac{\mu}{\sigma}\right)^2 - 1} \\ &\leq \frac{1}{b}\left(\frac{\varepsilon b}{\sigma^2}\right)^{\left(\frac{\mu}{\sigma}\right)^2 - 1} \\ &\leq max\left(\frac{1}{b}, \frac{1}{b}\left(\frac{\varepsilon b}{\sigma^2}\right)^{\left(\frac{b}{\sigma}\right)^2 - 1}\right) \end{aligned}$$

- elle atteint son maximum en $x = \mu - \frac{\sigma^2}{\mu}$ si $\mu^2 \geq \sigma^2$. Par suite :

$$
\begin{aligned}
sup_{x \geq \varepsilon} \gamma(x, \mu, \sigma) &\leq \gamma\left(\mu - \frac{\sigma^2}{\mu}, \mu, \sigma\right) \\
&= \frac{1}{\sigma^2 \Gamma(\alpha)} (\alpha^2 - 1)^{\alpha^2 - 1} e^{1 - \alpha^2} \qquad \text{avec } \alpha = \left(\frac{\mu}{\sigma}\right)^2 \geq 1 \\
&\leq \frac{1}{\sigma^2}
\end{aligned}
$$

□

Annexe C

Simulation of sea state parameters process to study the profitability of a maritime line

P. Ailliot and M. Prevosto
IFREMER
Brest, France

T. Soukissian and C. Diamanti
NCMR
Anavyssos, Greece

A. Theodoulides and C. Politis
HRS
Piraeus, Greece

ABSTRACT

This paper deals with a method of studying the profitability of a maritime line under the service of a specific ship when the data describing the sea state along it are available only for a short period of time or are missing. As a specific example, the line Piraeus - Heraklion at Aegean Sea is considered, where wave data are available for a very short period (3 years). The method developed is based on the use of another larger wind database (11 years) providing realistic artificial sea-state conditions for a wider time period. This is accomplished in two steps: First, by developing a stochastic simulator of wind conditions in a point corresponding to the most severe conditions of the line (this is performed using a non-homogeneous Markov model) and second, by establishing a simple non-parametric method associating the simulated wind conditions at this point to sea-state conditions along the line. Finally, the profitability of the line is examined by combining the simulated sea-state conditions along the line with the seakeeping behavior of the ship presented in the form of a suite of polar diagrams which provide the operable regime of speed and heading of the ship for each sea-state condition.

KEYWORDS

Time series, Maritime line, Sea state simulation, Wind speed and direction, Switching autoregressive model, Non homogeneous Markov model.

INTRODUCTION

In order to assess the profitability of a maritime line, the knowledge of the sea-state climatology along the line is necessary as well as the limiting operation criteria defining the acceptable parameters (speed, direction) of the ship. For a precise calculation, a very large number of sea-states histories which could be encountered by the ship must be used. Unfortunately, these histories are limited in number by the period of simulation of the numerical models (hind-, now- or fore-cast), which generally covers several years (1 up to 40 years). To overcome this constraint the development of a simulator of sea-state histories, statistically realistic, would be of great interest.

The wave data used in the present work, have been produced by the NCMR (National Center for Marine Research, Athens). They describe sea-state conditions along the line Piraeus-Heraklion during the period from October 1999 to September 2002. On the other hand, the behaviour of ship in waves and the operability limiting criteria are combined in speed polar diagrams corresponding to each particular sea state and defining acceptable and unacceptable regions of ship service. These diagrams have been produced by HRS (Hellenic Register of Shipping). These polar diagrams, in conjunction with the sea state data, will permit us to assess, if the ship service, a given day, is possible and, in the case of positive answer, the duration of the trip. These results will be used in a last step to estimate the profitability of the maritime line.

To evaluate accurately the profitability of the line, results for a longer time period than the three years, where data are available from the NCMR database, are required. So, a stochastic model which describes the sea-state history conditions along the line has been developed. This task was accomplished in two steps: At first, a non homogeneous Markov model (NHMM) has been fitted to the time series of a larger database (11 years, now-cast) of wind speed and wind direction at a point near the line where the most severe conditions were observed. This model has been used to simulate artificial wind conditions at this point. In a second step, a quick and simple algorithm has been developed in order to estimate the sea-state conditions corresponding to the simulated wind data. Finally, these synthetic sea-state conditions have been used in order to assess the operability of the ship and thus, the profitability of the maritime line.

WIND AND WAVE FORECASTING FOR THE AEGEAN SEA

A main part of the POSEÏDON system, Soukissian et al. (1999), is the POSEÏDON forecasting system, which consists of the weather and the wave forecasting system. The weather forecasting system is based on the SKIRON model, developed at the University of Athens (Kallos (1997), Nickovic (1998)). For a detailed description of the system see Soukissian *et al.* (2002). The wave-forecasting model running operationally in NCMR is the WAM-cycle 4 wave model, providing forecasts for the Aegean Sea since October 1999. The complete theory, on which WAM is based, is described in detail in WAMDIG (1988), while the description of the physics and the numerical schemes used by WAM-cycle 4 can be found in Komen et al. (1994). The WAM implementation for the

Aegean Sea is based on a nested version, where the outer nest covers the Mediterranean Sea and the inner nest the Aegean Sea. Thus, the boundary conditions for the Aegean Sea are obtained through the WAM-Mediterranean model. In table 1 the operational characteristics of the WAM model are summarized.

The WAM model runs operationally once a day giving forecasts for 3 hour intervals up to 72 hours ahead. The forecasts presented here correspond to the first 24 hours of the model output for the period from October 1999 to September 2002.

The experience obtained so far by using WAM model is that it can sufficiently describe the trends of the wave climate, but its key shortcoming is the underestimation of the severe sea states.

Table 1. Operational characteristics of the WAM model

Domain of application	Mediterranean Sea (34^0 N – 41^0 N) Aegean Sea (20^0 E – 29^0 E)
Grid	1.25^0 (Mediterranean Sea) 0.05^0 (Aegean Sea)
Time step	180 sec
Spectral discretization	16 discrete directions 30 discrete frequencies (logarithmically) 0.05054 - 0.66264 Hz
Wind input	POSEÏDON weather forecasting system

STOCHASTIC SIMULATION OF SEA-STATE CONDITIONS

We will assume that the WAM model provides an acceptable description of the sea-state conditions along the examined line during the last 3 years. With these data, we could estimate the joint distribution of H_s, T_p and Θ_m, where H_s, T_p and Θ_m represents respectively the significant wave height, the peak period and the mean direction, on the maritime line. With this distribution we could calculate the probability that severe conditions (according to some criteria giving by the polar diagrams) occurs on the line, and thus, for example, the percentage of cancelled or delayed crossings.

However this method would fail if we are interested in estimating more complicated quantities, like the mean time of crossing, where the knowledge of the space-time structure of the sea-state along the line is necessary. On the contrary, if we find an efficient way of simulating artificial sea-state conditions on the line on a long period of time, we will be able to simulate a great number of scenario of exploitation of the line, and so, will be able to provide reliable assessment of the profitability.

Simulation of sea-state conditions

At first, let $(x_1, ..., x_N)$, with N equal 17, be points on the grid of the WAM model lying also along the maritime line Piraeus-Heraklion (Fig. 1). The distance between adjacent points is about 12 nautical miles.

We now use another data base, produced by ECMWF, giving only the wind conditions but having the advantage to contain data for a longer period of time (11 years). In particular, this data base describes wind conditions in the period January 1992-December 2002, with a time step of 6 hours at the point (36N 27.75E), denoted by x_0, which is closed to the line and in the area where the most severe conditions were observed (figure 1).

We first used ECMWF data in order to simulate artificial wind conditions at the point x_0. To perform these simulations, a stochastic model (described below) of the time evolution of the bivariate process $(X(t), \theta(t))$, has been developed (X the wind speed, θ the wind direction).

In a second step, a simple algorithm has been used in order to associate realistic sea-state conditions at the points $(x_1, ..., x_N)$, corresponding to these artificial wind conditions.

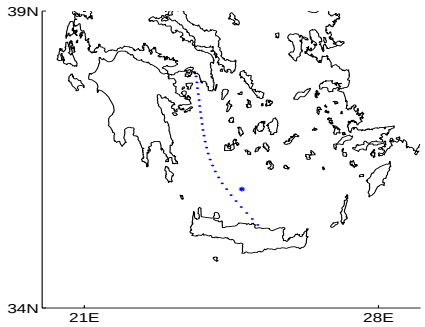


Fig. 1. : Points $(x_1, ..., x_N)$ (dotted) and point x_0 (star)

Stochastic model for the wind

Model for X. Several authors have worked on the problem of modelling and simulating the process X (Brown, 1984). As this process is usually nonstationary (seasonal and diurnal components) and non-Gaussian, a transformation is usually first applied to the data in order to get a stationary and Gaussian time series. Then, an *ARMA* model can be fitted to the transformed time series.

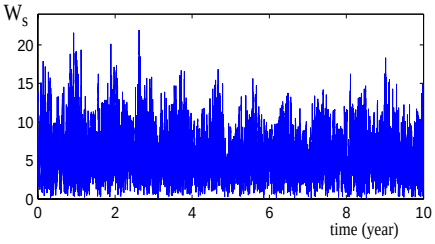


Fig. 2. : Evolution of the wind speed (11 years of data, ECMWF)

Elimination of the nonstationary components. In our data base, no significant daily nonstationarity has been observed but there are important seasonal variations (figure 2) with, for example, more severe storms in winter than in summer. Several models have been proposed in order to describe this non-stationarity (Cunha (1997)). In present work, these models have not been applied, but instead of this, we made the assumption that the data are strictly stationary within each month, that is all months of January are statistically similar, as are months of February, March, ... (in other words, we assume that there is no long-term trend in the data). Thus, a different model has been adjusted for each month and each of them has been fitted by using 11 (11 years) time series, assumed independent, of length 122 (one average month, each 4 hours). This method seems to work satisfactorily when the available data cover a relatively long period of time.

The results shown in the present work were obtained by considering data for the month of January.

Autoregressive model for X. The process X is non-Gaussian (figure 3). Its distribution is defined on the positive real axis and is generally positively skewed. In order to fit an autoregressive model to X, a transformation is usually first applied to the data in order to get a process with a Gaussian distribution (Brown, 1984). In this work, the non-Gaussian character of the data has been directly modelled using a Gamma autore-

gressive model of order k (Gamma AR(k)). More precisely, we will assume that (Toll (1997))

$$X(t) \sim \gamma(\alpha(t), \beta(t)) \tag{1}$$

with density

$$\gamma(x;\alpha, \beta) = \frac{\alpha^{\beta}}{\Gamma(\beta)} e^{-\alpha x} x^{\beta-1} \tag{2}$$

where

$$\alpha(t) = \frac{\mu(t)}{\sigma(t)^2} \text{ and } \beta(t) = \frac{\mu(t)^2}{\sigma(t)^2} \tag{3}$$

and $\mu(t)$ and $\sigma(t)$ represent respectively the conditional mean and standard deviation of $X(t)$ and are given by the following equations:

$$\mu(t) = \alpha_1 X(t-1) + \ldots \alpha_k X(t-k) + \beta \tag{4}$$

$$\sigma(t) = \sigma \tag{5}$$

The parameters have been estimated using numerical maximum likelihood conditionally to the first k observations for each of the 11 sequences.

Validation of the model. In order to check the ability of this model to describe the non-Gaussian character of the process, we have simulated with this model 500 sequences of length 122 (it corresponds to 500 months of January) and we have compared the distribution of these synthetic data with the one of the original data. The two distributions are in a good agreement (figure 3). Other characteristics of these simulated sequences are shown in figure 5 and figure 6. For these simulations, the model with $k = 3$ has been selected, according to the BIC criterion (see below table 2).

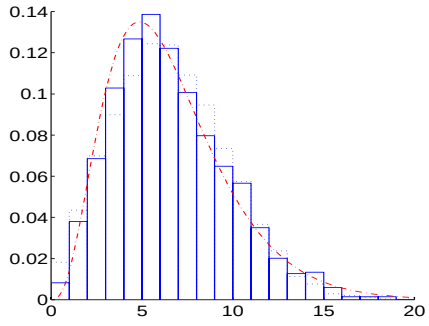


Fig. 3. : Comparison of the data distribution (solid line). Comparison with a gamma distribution (dashdotted line) and with the Gamma AR(3) model (calculated by simulation, dotted line) - Month of January

Switching gamma autoregressive model with Markov regime. The previous model seems to be able to restore the non-Gaussian distribution of the data, but an other important feature of the data is not catch. In fact, a closer look at the data reveals an other important characteristic of the process: depending on the meteorological situation, in some period the wind speed evolves slowly whereas in other ones there is more variability (figure 4). In order to describe this feature, we have used a switching autoregressive model with Markov regime. We assume that there exists a hidden variable, called "weather type", which governs the evolution of the observable process $X(t)$ by modifying the parameters of the Gamma AR model. The "weather type", $W(t)$ will be assumed to have a finite state space $\{1 \ldots M\}$. More formally, let X_t represent the sequence of the process X from time 1 to t : (X_t = ($X(1)$, …, $X(t)$)), and similarly for W_t. A switching Gamma AR(k) with Markov regime is defined by the following assumptions:

$$P(W(t) | W_{t-1}, X_t) = P(W(t) | W(t-1)) \tag{6}$$

$$P(X(t) | X_{t-1}, W_t) = P(X(t) | X(t-1), \ldots, X(t-k), W(t)) \tag{7}$$

The assumption (6) means that the weather type W, which is not observable, is a first order homogeneous Markov chain with regime transitions independent of previous observations.
Let $a_{i,j} = P(W(t) = j | W(t-1) = i)$ denote the transition probabilities, $A = (a_{i,j})$ the $M \times M$ transition matrix, $\pi_j = P(S_1 = j)$ and $\pi = \{\pi_j\}$ the $M \times 1$ vector representing the initial distribution.

Equation (7) states that $X(t)$ evolves like an autoregressive model of order k with parameters depending on the current value of the weather type W. More precisely, we have assumed that

$$P(X(t) | X(t-1), \ldots, X(t-k), W(t) = i) \sim \gamma(\alpha(t), \beta(t)) \tag{8}$$

with $\alpha(t)$ and $\beta(t)$ given by (3) and

$$\mu(t) = \alpha_1^{(i)} X(t-1) + \ldots \alpha_k^{(i)} X(t-k) + \beta^{(i)} \tag{9}$$

$$\sigma(t) = \sigma^{(i)} \tag{10}$$

The maximum likelihood estimates of the parameters have been calculated with the E.M (Expectation Maximization) algorithm, with a numerical optimization in the Maximization step. As the likelihood function can be multi-modal, the E.M algorithm may converge to a local maximum. In order to avoid these local maxima, we run the algorithm several time with different, randomly chosen, initial values.

Model selection. In order to select the best model, which means selecting M the number of regime and k the order of the autoregressive models, we have found that the Bayes Information Criterion (BIC) is useful in the sense that it identifies relatively parsimonious models which fit the data well. This criterion is defined as

$$BIC = -2LL + N \log(T) \tag{11}$$

with LL the log-likelihood of the model, N the number of parameters and T the number of observations.

Table 2. Comparison of the different models

	M=1	**M=2**	**M=3**
k=1	N=3 LL=-2641.1 BIC=5305.8	N=9 LL=-2506.6 BIC=5186.0	N=17 LL=-2543.8 BIC=5210.1
k=2	N=4 LL=-2640.1 BIC=5309.0	N=11 LL=-2559.1 BIC=5197.5	N=20 LL=-2540.0 BIC=5222.4
k=3	N=5 LL=-2619.9 BIC=5295.9	N=13 LL=-2556.4 BIC=5206.4	N=23 LL=-2535.1 BIC=5235.8

This criterion selects the model with $M = 2$ (2 weather types) and $k = 1$.

The conditional standard deviations of this model are $\sigma^{(1)} = 1.56$ and $\sigma^{(2)} = 2.96$. Thus, the first weather type is associated to wind speed evolving slowly whereas the second is associated to periods with a more important variability (stronger conditional standard deviation). In order to visualize it, we have applied the Viterbi algorithm to the data: it permits to calculate the most likely values of the process W according to the observations (figure 4).

Validation of the model. With the selected model we have generated 500 artificial time series of length 122 (each corresponding to a month of January), and we have compared some statistical properties of these artificial sequences with the original ones of the database. We also compared these results to the ones previously obtained with the only use of the Gamma AR(3) model (best model for $M = 1$ according to the BIC criterion; see Tab. 2).

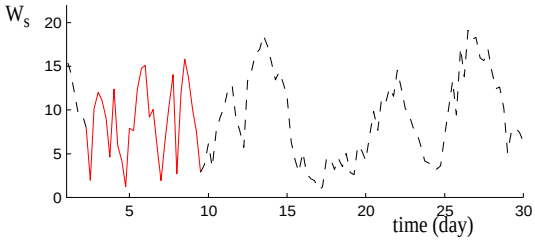


Fig. 4. : Wind speed for the month of January 1995 and corresponding values of W. Dates with $W = 1$ are represented with a dashed line and $W = 2$ with a continuous line

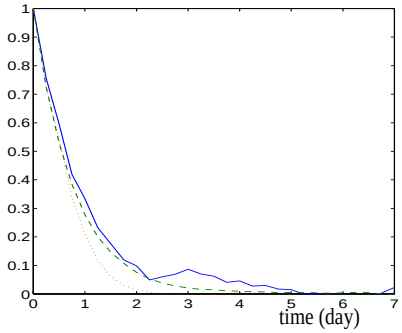


Fig. 5. : Comparison of the autocorrelation functions. Observations (continuous line), Gamma $AR(3)$ model (dotted line), switching Gamma $AR(1)$ model (dashdotted line)

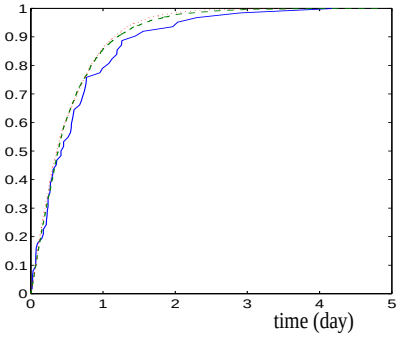


Fig. 6. : Distribution function of the storm duration (Wind speed above $13ms^{-1}$). Observations (continuous line), Gamma $AR(3)$ model (dotted line), switching gamma $AR(1)$ model (dashdotted line)

The statistical properties of the simulated sequences with the two models are close to the properties of the original data; compare e.g., the probability density functions, the autocorrelation functions (figure 5) or the time of duration of the storms (figure 6). For these data, the benefits of the more elaborated model with switching parameters are not obvious (except a better fit of the autocorrelation functions). However, in other works on wind data in the Bay of Biscay, where the difference between the weather type is more significant, this model outperforms more clearly the simple autoregressive model with fixed parameters.

Model for (X, θ)**.** There may exist a strong relation between the wind speed and the wind direction. In order to link the evolution of these two processes, we have used a non-homogeneous Markov model (NHMM) in which the wind direction modify the transition probabilities of the hidden Markov chain which represents the weather type. A NHMM model has also been proposed by Hughes (1999) for relating broad scale atmospheric circulation to local rainfall occurrences. More precisely, we will assume that (7) still holds and that

$$P(\theta(t)|W_t, X_t, \theta_{t-1}) = P(\theta(t)|\theta(t-1)) \quad (12)$$

$$P(W(t)|W_{t-1}, X_t, \theta_t) = P(W(t)|\theta(t), W(t-1)) \quad (13)$$

The assumption (12) means that θ is a first order Markov chain. In order to parameterize $P(\theta(t)|\theta(t-1))$, we could use one of the autoregressive models for circular data proposed in Brecling (1989) or Fisher (1994). However these models are not suitable when the distribution of the process is multi-modal, as it is the case in Aegean sea (figure 8). Here, one of the modes corresponds to wind blowing from the north and the other one from the west. For this application (profitability of the maritime line), a precision of 20° on the wind direction seems sufficient. Thus the wind direction has been classified into 18 sectors $I_1, \ldots, I_{18}$ of equal widths. Let Q be the 18×18 transition matrix of the Markov chain ($Q(i,j) = P(\theta(t+1) \in I_j | \theta(t) \in I_i)$). MacDonald (1997) proposed to use a second order Markov chain model (the Raftery model). This model outperforms the simple first order Markov chain for wind data in Koeberg (South Africa). As the results we have obtained with the first order Markov chains were satisfactory, this more elaborate model has not been tested.

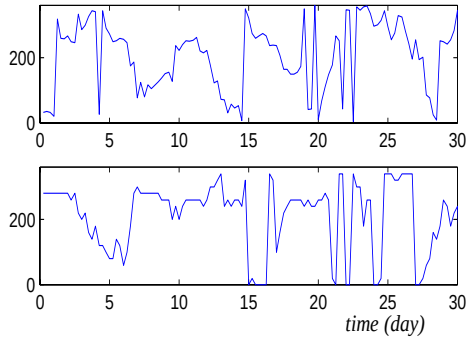


Fig. 7. : Example of time series of wind direction. Data (top) simulated with a first order Markov chain (bottom)

Assumption (13) states that given the history of the weather type up to time $t-1$ and of the wind speed and direction up to time t, the weather type at time t depends only on the previous weather type and the current wind direction. The wind direction at time t is thus used to modify the transition probabilities of the hidden process.

Models using the weather type to link the relation between the wind directions and the wind speed has already been used (for example in Castino, 1998). The weather type is calculated as a deterministic function of the wind direction (the directions are usually divided in 2 or 3 sectors) and a different model is fitted to the wind speed in each of these weather types. These models can be written as special cases of the NHMM by forcing $P(W(t)|\theta(t), W(t-1))$ to be degenerate.

In order to parameterize $P(W(t)|W(t-1), \theta(t))$, we used the fact that:

$$P(W(t)=j|W(t-1)=i,\theta(t)) \sim P(\theta(t)|W(t-1)=i,W(t)=j)P(W(t)=j|W(t-1)=i) \quad (14)$$

Then, we used the following assumption:

$$P(\theta(t)|W(t-1)=i,W(t)=j) = P(\theta(t)|W(t)=j) \sim V(\theta_j,\kappa_j) \quad (15)$$

with density (Von Mises distribution)

$$V(\theta;\theta_j,\kappa_j) = \frac{1}{2\pi I_o(\kappa_j)} e^{\kappa_j \cos(\theta-\theta_j)} \quad (16)$$

Finally we get

$$P(W(t)=j|W(t-1)=i,\theta(t)) \sim \gamma_{i,j} e^{\kappa_j \cos(\theta(t)-\theta_j)} \quad (17)$$

with the constraints $\sum_j \gamma_{i,j} = 1$ to ensure identifiability of the parameters.

The maximum likelihood estimates of the parameters have been calculated with the E.M algorithm (Hughes). This algorithm is computationally expensive because of the numerical optimization used in each step of the algorithm. As it was the case for the Switching Gamma AR(k) model, we used randomly chosen initial values in order to initialize the algorithm, and we used the BIC criterion in order to choose the best model. The model with $k = 1$ and $M = 2$ has been selected.

Validation of the model. In order to validate the model, we have simulated 200 realisations of length 122. At first, we have compared the bivariate distribution of the original data with the one of the simulated sequences (figure 8).

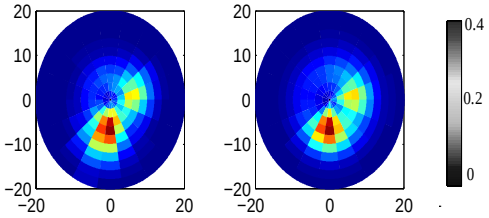


Fig. 8. : Joint distribution of the wind speed and the wind direction. Observation (left), simulated (right)

The observed distribution is bimodale, with 2 prevailing directions (west and north). The simulated distribution is a good approximation of the observed one. Other characteristics, like the autocovariance functions of the wind speed and the time of duration of the storms and the direction in which the wind is blowing in these events, have also been compared and are in a good agreement.

Simulation of sea-state conditions from wind conditions

This part describes a simple non-parametric method linking the temporal evolution of the wind conditions at the point x_0 (input data) to the sea-state conditions all along the line (output data).

Notations. Let

$$Y(x_i,k) = (H_s(x_i,k), T_p(x_i,k), \Theta_m(x_i,k)) \quad (18)$$

where $H_s(x_i,k)$, $T_p(x_i,k)$ and $\Theta_m(x_i,k)$ represent respectively the significant wave height, the peak period and the mean direction of the sea-state at time k for the point $x_i\,(i \in \{1\ldots N\})$ as calculated by the WAM model. The time index k is supposed to be in $\{1,\ldots,T\}$ with T the number of available data during the three years the forecast model was operational.

$Y(k) = [Y(x_1,k)',\ldots,Y(x_N,k)']'$ will represent the sea-state conditions at the different points along the line at time k.

Let $X(k) = (u(k),v(k))$ where $u(k)$ and $v(k)$ denote respectively the meridional and the zonal component of the wind at time k for the point x_0 as given by ECMWF data base.

Algorithm. We need to develop a quick and efficient algorithm which associates artificial sea-state conditions $Y_{sim}(t)$ along the maritime line to the input sequence $X_{sim}(t)$, $t \in \{0,\ldots,T_{sim}\}$, representing simulated wind conditions at the point x_0. The observed conditions X and Y will be used to train the algorithm.

- *Initialisation:* let

$$t_o = \arg min(\{\|X(t)-X_{sim}(0)\| \,|\, t \in \{1\ldots T\}\}) \quad (19)$$

be the date in the training sequence when the wind conditions are the more similar to the current wind conditions and let

$$Y_{sim}(0) = Y(t_0) \quad (20)$$

- *Recursion:* suppose we have already calculated $Y_{sim}(k)$. Let

$$t_{k+1} = \arg min\left(d\left(\begin{bmatrix} X_{sim}(k+1) \\ Y_{sim}(k) \end{bmatrix}, \begin{bmatrix} X(t+1) \\ Y(t) \end{bmatrix} \right) \middle| \, t \in \{1\ldots T\} \right) \quad (21)$$

be the date in the training sequence when the conditions (wind and wave) are the more similar to the current conditions according to the distance d and let $Y_{sim}(k+1) = Y(t_{k+1})$.

The distance d has been chosen as

$$d\left(\begin{bmatrix} X_{sim}(k+1) \\ Y_{sim}(k) \end{bmatrix}, \begin{bmatrix} X(t+1) \\ Y(t) \end{bmatrix} \right) = w_1 \|X_{sim}(k+1)-X(t+1)\| \quad (22)$$

$$+ w_2 \|Z_{sim}(x_{i_0},k+1)-Z(x_{i_0},t+1)\| + w_3 |T_{sim}(x_{i_0},k)-T(x_{i_0},t)|$$

where x_{i_0} is the point of the line closest to x_0, the (w_1,w_2,w_3) are fixed weights, and

$$Z(x_i,t+1) = (H_s(x_i,t)\cos(\Theta_m(x_i,t)), H_s(x_i,t)\sin(\Theta_m(x_i,t))) \quad (23)$$

and similarly for Z_{sim}.

Validation of the model. In order to check the ability of this algorithm to simulate realistic sea-state sequences, we used the data of the first two years (sea-states from POSEÏDON and wind from ECMWF) and we predicted sea-state conditions of the third year using the wind conditions at the point x_0.

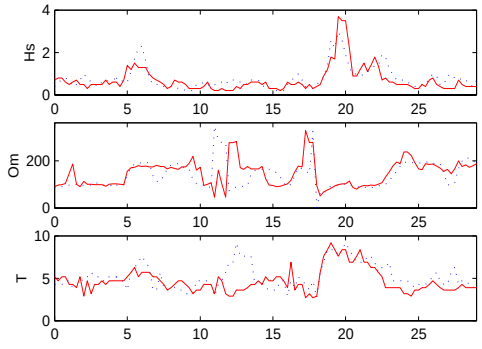


Fig. 9. : Comparison of Y and Y_{sim} for the month of January 2002 at the point x_{i_0}. WAM data (dotted line), and calculated from EC-

MWF data (solid line)

The weights have been chosen by numerical experimentation. It was found that the values $w_1 = 2$, $w_2 = 3$ and $w_3 = 1$ give good results (figure 9).

The comparison of these simulated sea-state conditions with the original data by the POSEÏDON system shows that this very simple algorithm is able to predict efficiently the sea-state conditions at the different points $(x_1, ..., x_N)$ from the wind conditions at x_0 (figure 9). For example, if $\varepsilon(t) = H_s(x_{i_0}, t) - H_{sim}(x_{i_0}, t)$ is the error of prediction on significant wave height for the point x_{i_0}, we get $\bar{\varepsilon} = 0.02$ and $var(\varepsilon) = 0.092$. The comparison at the other points of the line gives also good results.

PROFITABILITY OF THE MARITIME LINE

Polar Diagrams

Ship responses in a seaway, as a result of the wave induced motions, should not exceed specific limiting values in order to allow safe sailing, not only from the structural point of view but also from the crew effectiveness and passenger comfort aspects. These operability limiting criteria (named also seakeeping criteria) define the level of ship responses at which appropriate actions by the shipmaster are to be taken in order to reduce their magnitude and consequently their effects. The most important seakeeping criteria concern vertical and lateral motions and acceleration, rolling, and also phenomena like bow slamming, propeller raising and deck wetness. Proposed limiting values may be found in the literature (e.g. Karpinen, 1987). In the present work, in order to study the profitability of the maritime line, two criteria were used: the vertical acceleration at forward perpendicular below $0.18g$ and the roll angle below 4deg.

A convenient way to depict those combinations of ship speeds and heading angles that lead to exceedance of a certain seakeeping criterion or combinations of criteria, in a specific sea state, is by means of polar diagrams (figure 10). The unshaded areas in a polar diagram show to the captain the operable regime of speeds and headings. It is clear that at certain ship-waves heading angles operability is independent of speed. On the other hand, with speed fixed there are some headings that are operable and others that are not. In practice the shipmaster can be provided on line by a suite of such diagrams. For different locations along the ship route, he uses them in order to design ship route in a manner that ship responses are acceptable (Politis et al., 2002).

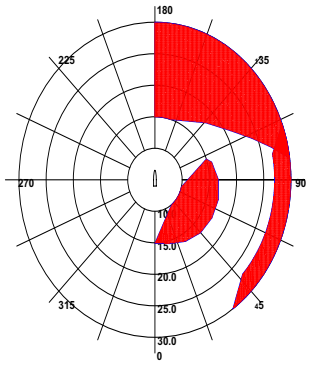


Fig. 10. : Polar diagram for $H_s = 3m$ and $T_p = 7s$.

The polar diagrams presented in this paper were derived by using the seakeeping software SWAN. SWAN is based on a three-dimensional Rankine Panel Method, according to which quadrilateral panels are distributed over the ship hull and part of the surrounding free surface. The free surface conditions implemented in SWAN linearize the wave potential above the double body flow. The numerical solution algorithms were derived using a rational stability analysis that leads to convergent and efficient wave flow simulation, free of numerical dissipation. The theory underlying SWAN is presented in Sclavounos (1996).

Route simulation

In order to simulate the voyages of the ship which leave the port at time d_0, we used the following algorithm. Suppose you know d_k, the time when the ship is at the point x_k:

- calculate the sea state conditions at this time and for this point by temporal linear interpolation;
- use the polar diagrams in order to calculate the maximum speed (we assume that the direction is given) at which the ship can sail in order to reach the next point, if it is possible. If the conditions are too severe, the voyage is cancelled;
- from this maximum speed and the distance between the points x_k and x_{k+1}, calculate d_{k+1}. And so on.

Results

We have simulated the sea-state conditions along the examined line for the equivalent of 500 months of January, and then, for each month, we have calculated the times of voyage from Piraeus to Heraklion for a daily departure at 6 o'clock (figure 11).

Then, we have compared the results obtained with these simulated sequences to the one obtained with the 3 year WAM data (table 3). We have found a higher frequency of cancelled and delayed voyages with the simulated data than given by the WAM database. This is explained by the fact that less severe storms have been observed in January in the period where the WAM model was operational (figure 2, WAM data are available during the last 3 years). The mean delay as calculated with WAM data is 22 minutes and 26 minutes with the simulated data.

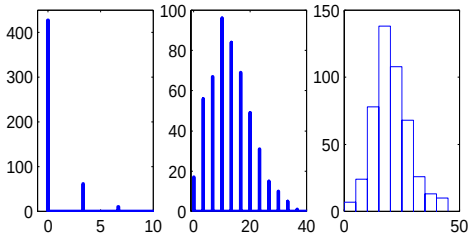


Fig. 11. : Distribution of the percentage of cancelled crossing (left), delayed crossing (middle) and distribution of the mean delay in minute (right) for the 500 simulated months of January

Table 3. Frequencies of delayed and cancelled crossings

	WAM	simulated
cancelled	0.00%	0.58%
delayed	11.1%	13.5%

CONCLUSION

In order to assess the profitability of a maritime line in Aegean sea, we used data produced by a numerical model. As these data were available on a too short period of time, we have developed a stochastic simulator of sea-state conditions along the line. At first, a new model has been proposed to describe and simulate the bivariate time series of wind speed and wind direction. This non homogeneous Markov model introduces a

non-observable variable, the weather type, in order to relate the two processes with a non homogeneous Markov model. It has been applied to a long time series of wind conditions available at a point near the line where the most severe conditions were observed. Then, a simple method has been used in order to associate realistic sea-state conditions to these simulated wind conditions.

Finally, polar diagrams, which give the acceptable maximum speed of the ship for each sea state conditions, have been used to estimate the durations of the voyages of the ship. The results obtained with the original and the simulated data have been compared. This comparison shows that, because of the strong variability of the climatological conditions, the 3 years of data were not sufficient in order to get a reliable estimation of the profitability of the line and demonstrates the interest of the simulation strategy.

ACKNOWLEDGEMENTS

The work has been partially supported by a grant from the EGIDE-PLATON program.

REFERENCES

Breckling, J. (1989). “The Analysis of Directional Time Series: Applications to Wind Speed and Direction”. *Springer*.

Brown, B.G., Katz, R.W., Murphy, A.H. (1984). “Time Series Models to Simulate and Forecast Wind Speed and Wind Power”. *J. of Climate and Applied Meteo.* Vol 23, pp1184-1195.

Castino, F., Festa, R., Ratto, C.F. (1998). “Stochastic Modelling of Wind Velocities Time Series”. *J. of Wind Ener. and Ind. Aero*. 74-76, pp 141-151.

Cunha, C., Guedes Soares, C. (1999). “On the Choice of Data Transformation for Modelling Time Series of Significant Wave Height”. *Ocean Engineering*. Vol 26, pp 489-506.

Fisher, N.I, Lee, A.J. “Time Series Analysis of Circular data”. *J. R. Statist. Soc B.* Vol 56, No2, pp 327-339.

Hugues, J.P, Guttorp, P. (1999). “A Non-Homogeneous Hidden Markov Model for Precipitation Occurence”. *Applied statistics.* Vol 48, part 1, pp 15-30.

Kallos, G., Nickovic, S., Papadopoulos, A., Jovic, D., Kakaliagou, O., Misirlis, N., Boukas, L., Mimikou, N., Sakellaridis, G., Papageorgiou, J., Anadranistakis, E. and Manousakis, M., (1997), "The regional weather forecasting system Skiron: An overview", *Proceedings of the Symposium on Regional Weather Prediction on Parallel Computer Environments, 15-17 October 1997, Athens, Greece.* pp 109-122.

Karpinen, T., (1987), "Criteria for seakeeping performance predictions", *VTT Technical Research Center of Finland, Ship Laboratory*

Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S. and Janssen, P.A.E.M. (1994), "Dynamics and Modelling of Ocean Waves", *Cambridge University Press*.

MacDonal, I.L, Zucchini ,W. (1997). “Hidden Markov and Other Models for Discrete-Valued Time Series”. *Chapman & Hall.*

Monbet, M., Prevosto, M. (2001). “Bivariate Simulation of Non Stationary and Non Gaussian Observed processes. Application to Sea State Parameters”. *Applied Ocean Research*. Vol 23, pp 135-145.

Nickovic, S., Mihailovic, D., Rajkovic, B., and Papadopoulos, A. (1998), "The Weather Forecasting System SKIRON, Vol II: Model description". *ISBN 960-8468-16-7*.

Politis, C., Voutsinas, V., Theodoulides A. (2002), "On line assessment of operability of a RO-RO passenger ship in a seaway", Presented at the "Atmospheric Modeling from Microscale to Global - *5th RAMS workshop and Related Applications, Santorini 2002*.

Sclavounos, P., (1996), "Computation of wave ship interactions", *Advances in Marine Hydrodynamics*.

Soukissian, T.H., Chronis, G. Th. and Nittis, K. (1999), "POSEÏDON: Operational Marine Monitoring System for Greek Seas", *Sea Technology*. Vol. 40, No. 7, pp. 32-37.

Soukissian, T.H., Prospathopoulos, A. and Diamanti, C., 2002, "Wave and wind data analysis of the Aegean Sea. Preliminary results". *The Global Atmosphere and Ocean Systems*. Volume 8, Numbers 2-3, pp. 163 - 189.

Toll, R.S.J (1997). “Autoregressive Conditional Heteroscedasticity in Daily Wind Speed Measurements”. *Theor. Appl. Climatol*. Vol 56, pp 113-122.

WAMDI group: Hasselmann, S., Hasselmann, K., Bauer, E., Janssen P.A.E.M., Komen, G.J., Bertotti, L., Lionello P, Guillaume, A., Cardone, V.C., Greenwood, J.A., Reistad, M., Zambresky, L. and Ewing, J.A. (1988), "The WAM model - a third generation ocean wave prediction model". *J. Phys. Oceanogr*. Vol 18, pp 1775-1810.

Annexe D

Nonparametric Modelling of Cyclo-Stationary Markovian Processes Part II: prediction and dimension reduction

V.Monbet
UBS/SABRES, Vannes
France

P.F. Marteau
UBS/VALORIA, Vannes
France

P. Ailliot
IFREMER/Metocean team, Brest
France

ABSTRACT

This paper deals with spatio-temporal conditional prediction of sea state parameters given nearby observations of the same parameters or given the observations of other sea state parameters at the same geographic point. An algorithm referred as Non Parametric Viterbi (NPV) and based on Hidden Markov Chain theory is proposed. It is shown that this algorithm can be used, for instance, to predict missing values in sea state data networks such as bouy networks or to predict a sea state process given part of the multivariate observed vector. The reduction of the dimension of the representation state space in which the process is described is also of practical use since this allows jointly to reduce the size of the learning data set and to maintain the algorithmic complexity at a tractable level.

KEYWORDS: **Nonlinear Time Series, Markov chains, Prediction, Viterbi algorithm, Non Parametric Estimation.**

INTRODUCTION

We address the non parametric modeling of cyclo-stationary multivariate Markovian processes using a continuous state space and discrete time Hidden Markov Model (HMM) for which all necessary densities functions are approximated using samples. Let us first describe two potential applications of the proposed Viterbi algorithm.

Pittalis[9] and Puca [10] study a neural network approach to reconstruct missing data in spatial bouys networks. Well known drawbacks of Neural Networks models for time series prediction if the large amount of data and the often high computation time required to learn the model. NPV algorithm is an alternative for missing data reconstruction. And it is shown bellow that good prediction can be obtained with quite small learning samples for significant wave height processes.

In paper [1], Ailliot and al. propose a bootstrap method to study the profitability of the maritime line Piraus-Heraklion at Aegean sea. The principle of their algorithm is to simulate a large number of realistic sea state histories and to estimate the profitability of the line given polar diagram of the boat and the simulated sea state parameters. In practice, the wind intensity is simulated at a central point of the line using a parametrical Hidden Markov Model (see [1],[2]). Then, the sea state parameters corresponding to the simulated wind at time t are deduced in searching the nearest neighbor of vector $(W_{sim}(t,x_0), SS_{sim}(t-1,x_0), \cdots, SS_{sim}(t-1,x_L))$ in an hindcast data set. SS_{sim} denotes the sea state parameter simulated vector. For longer lines, prediction methods are good ways to spread out the punctual simulation of wind speed to near by points along the line. This spreading should improve the simulations. Such applications will be studied in a future paper.

The proposed NPV approach to conditional prediction is based on HMM modelling. Hidden Markov model is basically a Markov chain whose internal state cannot be observed directly but only through some probabilistic function. That is, the internal state of the model only determines the probability distribution of the observed variables. Let us consider as example that we have an observed time series of significant wave height H_s and that the corresponding mean wind speed $W_s$is unobserved. In this case W_s is the hidden Markov chain also refered as states, H_s is the observed variable and the HMM is specified by the prior probability distribution of W_s, the transition kernel $P(W_s(t)|W_s(t-1))$ of the Markov chain and the conditional

observation distribution $P(H_s|W_s)$.

Three classical problems listed below have been solved for discrete HMM that make this kind of model quite useful [11].

1. the estimation of the probability that the model generates the observation sequence $\{O_t, t \geq 1\}$: the forward-backward algorithm has been developed;
2. the recovery of the most likely hidden state sequence corresponding to these observations: dynamic programming (Viterbi algorithm [5]) is commonly used;
3. the estimation of the parameters of the HMM (transition matrix, prior state distribution, observation conditional distribution) to better account for the observations: the EM-algorithm or Baum-Welch algorithm have been proposed for this task.

In continuous state space situations, the problem of parameter estimation is much more complex [13]. In this paper we address mainly three issues:

1. the local discretization of the observation and state spaces in which the process is handled.
2. the estimation of the parameters of the HMM model, namely the probability density functions involved in the proposed model according to the proposed discretization of observation and state spaces.
3. the estimation of the hidden state sequence $\{S_t, t \geq 1\}$ conditionally to the observation sequence $\{O_t, t \geq 1\}$ and the model M: $P(\{S_t, t \geq 1\}|\{O_t, t \geq 1\}, M)$.

The proposed model approximates the initial multivariate process (IMP) by decomposing it into a Lower Dimensional cyclo-stationary Markov Chain (LDMC) for which state transitions are hidden. Indeed, state process is indirectly observed through a second stochastic process that generates a multivariate observation from any state of the LDMC. The hidden LDMC state vector coincides with the first coordinates of the state vector of the IMP, while the multivariate observation vector coincides with the last coordinates of the IMP state vector. This decomposition induces a dimension reduction that allows to handle more complex processes, at a computational cost that can be estimated from the data. A Viterbi algorithm referred as Non Parametric Viterbi Algorithm (NPV algorithm) is proposed to extract most likely LDMC state trajectories from sequences of observation vectors. This approach can be used to predict state trajectories when the underlying multivariate dynamical process is partially observed. The method is thus of practical use in case of partially missing or noisy sample data. It can also be used as a bootstrap technique when the dimension reduction of the state space is necessary for complexity management.

In the first section the Hidden Markov Model for cyclostationary process is defined, NPV algorithm is described and shortly discussed. Then, in the second part, this algorithm is tested on real sea state data.

I. HIDDEN MARKOVIAN MODEL FOR CYCLO-STATIONARY PROCESS MODELIZATION

A. Notations, definitions and hypothesis

Let $\{X_t, t \geq 1\}$ be a d-dimensional stochastic process in $(\mathbf{R}^d, \mathcal{B}_d)$ where $\mathcal{B}_d$ is the Borel σ algebra over $\mathbf{R}^d$. We call this process the Initial Multivariate Process (IMP). This process can be decomposed in two dependent lower dimensional processes: $\{S_t, t \geq 1\}$ being a u-dimensional stochastic process in $(\mathbf{R}^u, \mathcal{B}_u)$, and $\{O_t, t \geq 1\}$ being a v-dimensional stochastic process in $(\mathbf{R}^v, \mathcal{B}_v)$, with $u + v = d$, and $\forall t, X_t = (S_t, O_t)$.

We suppose that there exists a positive integer $p < \infty$ such that the stochastic process $\{S_t,\ t > p\}$ with $Y_t = \{S_t, S_{t-1}, ..., S_{t-p+1}\}$ forms a cyclostationary Markov chain on $(\mathbf{R}^{up}, \mathcal{B}_{up})$ with periodic transition probability function $P_t(y, A)$ where $A \subset \mathbf{R}^u$ and $t \in [0, \Pi]$. Furthermore P_t is supposed to vary slowly such that it can be assumed that the process is almost stationary on small time intervals. We assume that for each $t \in [0, \Pi]$ the Markov kernel $P_t(y, A)$ admits a stationary distribution π_t with continuous probability density function $f_{t,Y}$ with respects to Lebesgue measure and such that $t \to \pi_t$ is periodic with period Π.

Furthermore, we suppose that there exists a positive integer $q < \infty$ such that the stochastic process $\{Z_t,\ t > p\}$ where $Z_t = \{O_t, O_{t-1}, ..., O_{t-q+1}\}$ is driven according to a distribution that we suppose also cyclo-stationary and synchronic with the $\{S_t, t \geq 1\}$ process, i.e. with the same period Π. We assume that this distribution remains almost stationary over a small time interval and admits a continuous probability density function that we suppose periodic.

We consider that the $\{O_t, t \geq 1\}$ stochastic process is made available, through measurements or any kind of simulation procedure (for instance by means of LGB resampling [8]). The question that we address is to undercover the underlying Markov process, namely, $\{S_t, t \geq 1\}$ conditionally to the observed $\{O_t, t \geq 1\}$ stochastic process in order to approximate at best the initial process (IMP).

B. Local discretization of the state spaces

Let VY_t be the neighborhood of Y_t defined as the set of $Y_l \in \{\tilde{Y}_\tau | d(\tilde{Y}_\tau, Y_t) \leq \sigma_T/2\}$ where $\tau \in [T - \tau_0, T + \tau_0]\}$. We note $I[VY_t]$ the set of time index such that for all $l \in I[VY_t], Y_l \in VY_t$. Furthermore, we define the image $(VY_t)^+$ of VY_t by $(VY_t)^+ = \{S_{l+1}, l \in I[VY_t]\} \subset \mathbf{R}^u$. We define VZ_t and $I[VZ_t]$ similarly. Finally we define $VY_t|Z_t$ the set $\{\tilde{Y}_l | \tilde{Z}_l \in VZ_t\}$.

The local discretization of the state space is obtained according to the schema presented in Figure 1 : for each observation Z_t, we evaluate the neighborhood VZ_t that defines the set $VY_t|Z_t$. This set defines the local discretization of the state space at step t.

C. Parameters estimation of the HMM model

The density probability functions $f_{t,Y}$ and $g_{t,Y}$ of the cyclostationary distributions as well as the transition probability distributions will be approximated from data using appro-

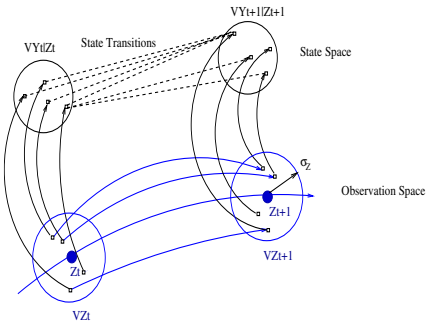


Figure 1. Discretization of the state space given the observations VZ_t and VZ_{t+1}

priate kernel estimates. In particular, we suppose that some learning data $\{\tilde{X}_t = (\tilde{Y}_t, \tilde{Z}_t), t \geq 1\}$ is available to estimate non parametrically the transition probability density functions for the subprocess $\{Y_t, t \geq 1\}$: the dynamics of $\{Y_t, t \geq 1\}$ is not hidden at this stage.

Let K_u and K_v be probability density functions on $\mathbf{R}^u$ and $\mathbf{R}^v$ respectively, K_{up} and K_{vq} probability density functions on $\mathbf{R}^{up}$ and $\mathbf{R}^{vq}$ respectively. Let $\{h_T, T = 0, 1, 2, ...\}$ and $\{h'_T, T = 0, 1, 2, ...\}$ sequences of positive numbers such that $h_T \to 0$ and $h'_T \to 0$ as $T \to \infty$. We suppose that the density kernels K_u, K_v, K_{up} and K_{vq} satisfy the usual conditions.

Estimation of the transition probability density The transition probability function of state S_{t+1} given observation Y_t is estimated non parametrically using a kernel estimate. The kernel estimate is based on the product of three terms. Two first terms measure the distance between respectively the state vectors and the observation vectors of the current neighborhood. And third kernel measure the distance between the current date and the observation date modulo the period.

$$p(S_{t+1}|Y_t) = \sum_{i \in I[VY_t]} K_u\left(\frac{S_{t+1} - \tilde{S}_{i+1}}{h_T}\right) \times K_{up}\left(\frac{Y_t - \tilde{Y}_i}{h_T}\right) \times exp\left(-\left|\frac{(t-i) \bmod \Pi}{\sigma_\Pi}\right|\right) \quad (1)$$

From this density of probability, we define the probability mass for the discrete random variable J taking its values in $I[VZ_{t+1}] = \{k \in \mathbf{N},\ \tilde{Z}_k \in VZ_{t+1}\}$, with probability mass function given by:

$$P(J = k) = \frac{p(S_{t+1} = \tilde{S}_k|Y_t)}{\sum_{j \in I[VZ_{t+1}]} p(S_{t+1} = \tilde{S}_j|Y_t)}, \forall k \in I[VZ_{t+1}] \quad (2)$$

Estimation of the observation probability density Observation probability density functions are estimated similarly as transition probability functions just above.

$$p(Z_t|Y_t) = \sum_{i \in I[VY_t]} K_{vq}\left(\frac{Z_t - \tilde{Z}_i}{h'_T}\right) \times K_{up}\left(\frac{Y_t - \tilde{Y}_i}{h_T}\right) \times exp\left(-\left|\frac{(t-i) \bmod \Pi}{\sigma_\Pi}\right|\right) \quad (3)$$

From this density of probability, we define the probability mass for the discrete random variable L tacking its values in $I[VY_t] = \{k \in \mathbf{N},\ \tilde{Y}_k \in VY_t\}$, with probability mass function given by:

$$P(L = k) = \frac{p(Z_t = \tilde{Z}_k|Y_t)}{\sum_{j \in I[VY_t]} p(Z_t = \tilde{Z}_j|Y_t)}, \forall k \in I[VY_t] \quad (4)$$

D. Estimation of the most likely state sequence

Given an observed sequence ${Z_1}^T = (Z_1, \cdots, Z_T)$, the inference of the most likely state sequence $\hat{Y}_1^T = (\hat{Y}_1, \cdots, \hat{Y}_T)$ is achieved using algorithms that perform the following maximization:

$$\hat{Y}_1^T = \max_{{Y_1}^T} P({Y_1}^T|{Z_1}^T, M) = \max_{{Y_1}^T} P({Y_1}^T, {Z_1}^T, M)$$

The Viterbi algorithm [5] finds the above maximum with a relatively efficient recursive solution: its computational cost is proportional to the number of non-zero transitions probabilities multiplied by the sequence length. First we define $\delta(i,t)$ as:

$$\delta(i,t) = \max_{{Y_1}^{t-1}} P({Z_1}^t, {Y_1}^{t-1}, Y_t = \tilde{Y}_i, M)$$

which can be computed recursively as follows, using the usual Markov conditional independence assumptions:

$$\delta(i,t) = P(Z_t|Y_t = \tilde{Y}_i) \times \max_j P(Y_t = \tilde{Y}_i|Y_{t-1} = \tilde{Y}_j, M) \times \delta(j, t-1) \quad (5)$$

with the initialization:

$$\delta(i,1) = P(Z_1|Y_1 = \tilde{Y}_i)P(\tilde{Y}_i)$$

If we define:

$$\mathcal{I}(i,t) = arg\max_j P(Y_t = \tilde{Y}_i|Y_{t-1} = \tilde{Y}_j, M)\delta(j, t-1)$$

then we obtain the optimal state sequence $\hat{Y}_1^T$ using the following backward recursion:

$$\hat{j}_T = arg\max_i \delta(i,T), \quad \hat{j}_{t-1} = \mathcal{I}(\hat{j}_t, t) \text{ and } \forall t, \hat{Y}_t = Y_{\hat{j}_t} \quad (6)$$

E. Non Parametric Viterbi Algorithm

The proposed Non Parametric Viterbi algorithm can be described as follows.

Initialization step -

- Select the initial observation Z_0, the bandwidth parameters h_T and h'_T, the width σ_Y (respectively σ_Z) of the neighborhood of a given state (respectively of a given observation) and the time window σ_H in which the process is supposed to be stationary.
- From Z_0 (at $t = 0$) determine the set of initial states $VY_0|Z_0$ as specified in Figure 1 at $t = 0$.
- Initialize prior probabilities according to the mass function defined in equation (4).

Step t -

- Discretize the state space using $VY_t|Z_t$ as specified in Figure 1 at time t (we suppose that the state space is discretized at step $t-1$, using $VY_{t-1}|Z_{t-1}$).
- Evaluate observation probabilities at step t according to the mass function defined in equation (4).
- Evaluate state transition probabilities from step $t-1$ to step t according to the mass function defined in equation (2).
- Compute $\delta(i,t)$ according to equation (6) for $i \in I[VY_t|Z_t]$

Stop condition *Step* T -

- At $t = T$ compute the most likely state sequence according to equation (6).

In next section, NPV algorithm is applied and discussed on sea state hindcast data.

II. APPLICATIONS

In order to test NPV algorithm, hindcast data of OCEANWEATHER database at points 5740 to 5745 are used (see map of figure 8). Points 5740 to 5745 are positioned along latitude 47.5N with longitude varying from -7.5W to -3.2W with constant longitude step. The distance between twi consecutive points is about 100 kilometers. At each point, 22 years of data are available with 6 hours time step. Now we present numerical results obtained with NPV algorithm [8].

Spatial context

Here, we show that the NPV algorithm achieves to predict sea state parameters at a geographical point given the same sea state parameters in a nearby area. We predict the significant wave height H_s at Oceanweather points 5741 to 5745 given the observation of H_s at point 5740, denoted H_s^{5740}. All these points are not very far from each other, so that it is convenient to compare the NPV prediction to the prediction obtained by just translating the reference signal H_s^{5740}. Table I reports the evolution of the absolute error between the predicted signal and the observed one with respect to the size of the learning sample when one predict a two years series of H_s^{5745} given H_s^{5740}. It is shown that the absolute error is about constant when the learning sample size increases.

Size (in years)	1	2	3	4	5	10
Err_{NPV}	0.40	0.39	0.38	0.40	0.40	0.40

Table I

Then, we have predicted two years of the time series and the learning sample contains 5 years of data. The given observation is H_s^{5740}. Table I and figure 3 report the absolute error between the predicted H_s signal and the observed one for all the studied points. The error is respectively calculated for NPV prediction and translation prediction.

Prediction	H_s^{5741}	H_s^{5742}	H_s^{5743}	H_s^{5744}	H_s^{5745}
Err_{NPV}	0.13	0.21	0.27	0.35	0.41
Err_{trans}	0.09	0.19	0.29	0.55	0.73

Table II

Figure 2 shows an example of prediction obtain by NPV algorithm for H_s^{5745} given the observation of H_s^{5740}. As reference, the H_s^{5740} signal (dotted line) has been add to observed H_s^{5745} (black solid line) and predicted H_s^{5745} (red solid line) on the figure. We observe that the global evolution of H_s^{5745} is well reproduce.

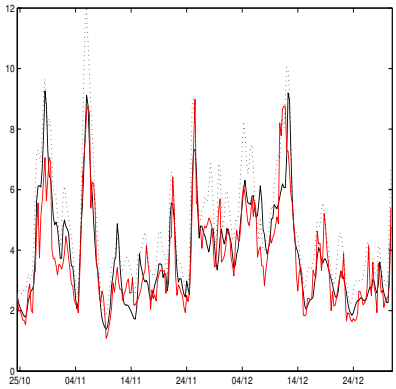


Figure 2. Time series of Hs with respects to the date; dotted line: H_s^{5740}, solid black line: observed H_s^{5745}, solid red line predicted H_s^{5745}

Multivariate context

We consider now a multivariate time series (H_s, T_p, W_s) on a fixed point of the Oceanweather grid, namely point 5740. H_s, T_p, W_s denotes respectively the significant wave height, the peak period and the wind speed. NPV algorithm can be used to predict one component (or more) of the multivariate process given the observation of remaining components. For instance, one can reconstruct W_s given observation of H_s if some sufficient long observation of bivariate procee (H_s, W_s) are available. This procedure allows to reduce the dimension of the state space reconstruction when using bootstrap algorithm such as Local Grid Bootstrap (See [8]) or to couple NPV with a parametric model. The reduction of dimension influences firstly the quality of estimation of

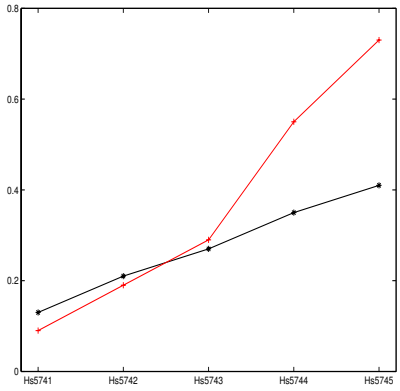


Figure 3. Absolute error between the predicted H_s signal and the observed one; black stars: NPV prediction, red plus: translation prediction

probability density functions: in LGB algorithm the transition probability density functions are defined in R^{up+vq} while in LGB+NPV the probability density functions are defined respectively in R^{up} and R^{vp}. Now, by convergence properties of kernel estimates, it is known that the estimation error increases with the dimension of the definition space and decreases with the sample size [4]. So that we will need larger samples for LGB than for LGB+NPV. Secondly, the reduction of dimension speed up the search for nearest neighboors.

In the example below, we compare both approaches (LGB alone vs. LGB+NPV) to generate time sequences of (H_s, T_p, W_s). The learning sample used is formed by 5 years of Oceanweather data at point 5740. And 15 years time histories are generated. In NPV, the observation is the bivariate process (H_s, T_p) and state W_s is predicted. Figure 4 to 7 compare statistics of LGB simulation and LGB+NPV simulation to the data. The choice of the plotted statistics is driven by applications. Indeed, for the considered applications, it is important to restore the instantaneous marginal and joint distribution of the parameters and the persistence statistics of storms and calm weather. Figure show the estimated statistics, for instance repartition functions or persistence statistic, and the empirical confidence interval. All figures show a good agreement between observed and generated data. And we can observe through confidence intervals that the difference between LGB and LGB+NPV simulation is not significant. The worse result is observed for the joint probability density function of (H_s, W_s). Indeed, figure 5 shows that LGB+NPV generate some 'strange wind' for low H_s. But such H_s is often not determinant for profitability or reliability applications. Furthermore, the global dependence structure seems to match.

CONCLUDING REMARKS

In this paper, a non parametrical version of Viterbi algorithm is proposed to modelize and predict multivariate cyclostationary Markov processes. Several potential applications have been presented such as replacement of missing data in bouy networks or bootstrap estimation of profitabil-

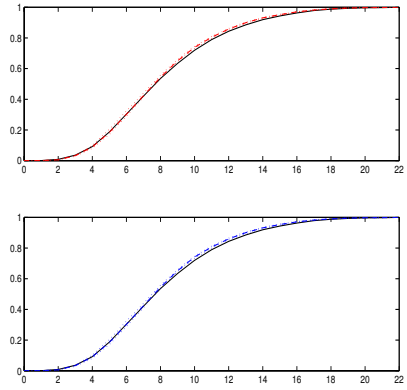


Figure 4. Comparison of distribution functions of observed and generated W_s. Top: LGB; bottom: NPV . Black solid line: observed data; red dashed line: generated data; red doted line: confidence interval

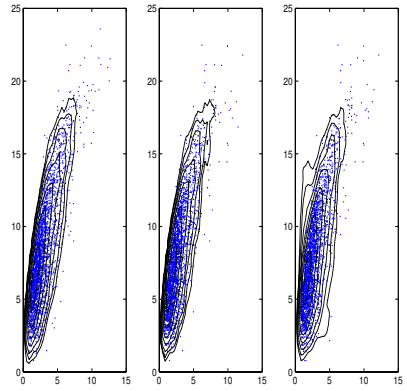


Figure 5. Joint probability density function of couple (H_s, W_s). Right: observed data, center: LGB, left: NPV

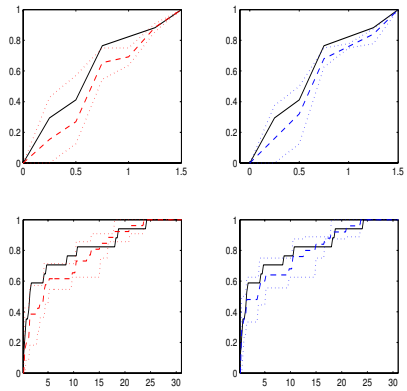


Figure 6. January month - Top: persistence of storms (Left figures: LGB; Right figures: LGB+NPV). Bottom: persistence of calm between storms. Solid line: observed data; dashed line: generated data; doted line: confidence interval

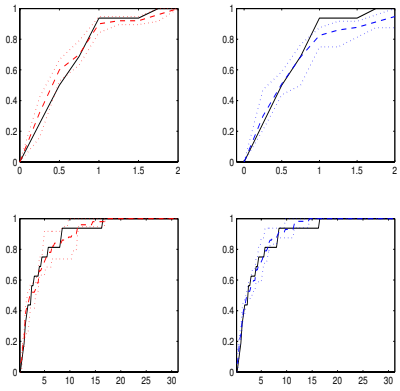


Figure 7. August month - Top: persistence of storms (Left figures: level LGB; Right figures: LGB+NPV). Bottom: persistence of calm between storms. Solid line: observed data; dashed line: generated data; doted line: confidence interval

ity of maritime lines. The algorithm has been discussed and numerical results evaluated against sea state hindcast data have been presented.

NPV algorithm is shown to give good predictions of time histories of different sea state parameters even with quite small learning sample (one year) when the observation and state process are dependants. Other numerical tests that are not presented here have demonstrated that when the dependence between both processes decrease it is necessary to increase the length of the learning series. But the predictions still of good quality.

This algorithm runs quickly and is quite easy to use for various types of multivariate data. There is essentially one critical parameters in the algorithm: the bandwidth parameter used in the non parametrical estimations of the probability density functions. This parameter can be regulated empirically.

REFERENCES

[1] Ailliot, P., Prevosto, M., Soukissian, T., Diamanti, C., Theodoulides, A., Politis C., 2003. Simulation of sea state parameters process to study the profitability of a maritime line. Proc. of ISOPE Conf. 2003.

[2] Ailliot, P., 2004. Switching Autoregressive Models - Application to Wind Simulation. PhD Thesis University of Rennes 1.

[3] Athreya, K.B., Atuncar, G.S., 1998. Kernel estimation for real-valued Markov chains. Sankhya: The Indian J. Stat. 60, Series A, Pt.1, 1-17.

[4] Bosq D., 1998. Non parametric statistics for Stochastic Processes. Springer, New York.

[5] Forney, G.D.Jr., 1973. The Viterbi algorithm Proc. of the IEEE Vol. 61, No. 3, pp. 268-278.

[6] Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic stability. Springer, London.

[7] Monbet V., Marteau P.F., 2001. Continuous Space Discrete Time Markov Models for Multivariate Sea State Parameter Processes. Proc. ISOPE Conf. 2001.

[8] Monbet V., Marteau P.F., 2004. Non Parametric Modelling of Cyclo-Stationary Markovian Processes Part I: Simulation of multivariate sea state processes, Proc. Isope 2004

[9] Pittalis S., Bruschi A., Puca S., Tirozzi B., 2003. Reconstruction of Sea Events and Extreme Value Analysis. Proc. ISOPE 2003.

[10] Puca S., Tirozzi B., Arena G., Corsini S., Inghilesi R., 2001. A neural network approach to the problem of recovering lost data in a network of marine buoys'. Proc. ISOPE Conf. 2001.

[11] Rabiner 1989. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of the IEEE, 1989.

[12] Stephanakos Ch.N. 1999. Non stationary stochastique modelling of time series with applications to environmental data. PhD Thesis.

[13] Thrun S., Langford J.C., Fox D. , 1999. Monte Carlo Hidden Markov Models: Learning Non-Parametric Models of Partially Observable Stochastic Processes Proc. 16th International Conf. on Machine Learning. 1999.

Notations

Mathématiques

$S_n(\boldsymbol{R})$	ensemble des matrices symétriques de taille $n \times n$ à coefficients réels
$S_n^+(\boldsymbol{R})$	ensemble des matrices symétriques positives de taille $n \times n$ à coefficients réels
$\rho(M)$	rayon spectral de la matrice M
$\mathbf{1}_A$	fonction indicatrice sur A
$card(X)$	cardinal de l'ensemble X
$\ln^+(x)$	maximum de $\ln(x)$ et de 0
$\nabla_\theta f(\theta)$	gradient de la fonction f en θ
$\nabla_\theta^2 f(\theta)$	hessien de la fonction f en θ
δ_x	masse de Dirac en x
$N(m, \Sigma)$	loi normale de moyenne m et de matrice de variance-covariance Σ
i.i.d	indépendant et identiquement distribué

Paramètres océano-météo (cf annexe A)

U	intensité du vent
Φ	direction du vent
u	composante zonale du vent
v	composante méridienne du vent
H_s	hauteur significative des vagues
T_m	période des vagues
Θ_m	direction de propagation des vagues

Modèles

$MS-AR$	modèle autorégressif à changements de régimes (cf page 7)
CMC	modèle chaîne de Markov cachée (cf page 9)
LAR	modèle autorégressif linéaire (cf page 10)
NAR	modèle autorégressif non linéaire (cf page 10)
γAR	modèle autorégressif gamma (cf page 10)
$MS-LAR$	modèle autorégressif linéaire à changements de régimes (cf page 11)
$MS-NAR$	modèle autorégressif non-linéaire à changements de régimes (cf page 12)
$MS-\gamma AR$	modèle autorégressif gamma à changements de régimes (cf page 12)
$NHMS-AR$	modèle autorégressif à changements de régimes non-homogènes (cf page 14)
$NHMS-\gamma AR$	modèle autorégressif gamma à changements de régimes non-homogènes (cf 2.c)
$ARMA(p,q)$	modèle autorégressif à moyenne mobile d'ordre (p,q)
TGP	processus gaussien transformé (cf page 64)
LGB	local grid bootstrap (cf page 73)