

UNE MODÉLISATION DE DONNÉES SPATIO-TEMPORELLES PAR MODÈLES AR SPATIAUX

Aude ILLIG *

RÉSUMÉ

Dans cette étude, nous montrons que les modèles AR spatiaux peuvent être des modèles adaptés à l'étude de données spatio-temporelles de températures dans l'océan Pacifique et que ces modèles permettent des calculs simples et rapides des prédicteurs qui conduisent ensuite à la prévision de certains phénomènes océanographiques. Afin d'identifier les paramètres de modèles AR(p) quadrantaux, nous établissons pour des innovations différences de martingales, les propriétés asymptotiques d'estimateurs construits à l'aide des équations de Yule-Walker. De ces propriétés, nous déduisons un test statistique pour déterminer l'ordre p et obtenons ensuite des estimations pour les coefficients autorégressifs et la variance des innovations.

Mots-clés : Modèles AR spatiaux, équations de Yule-Walker, sélection d'ordre, estimation des coefficients autorégressifs, prédiction, données spatio-temporelles.

ABSTRACT

In the context of space-time data, we study sea surface temperatures in Pacific Ocean with spatial AR models. These models allow straightforward computations of predictors and then give predictions of some oceanographic phenomena. For quadrantal-type spatial AR(p) model identification, asymptotic properties of estimators derived from the Yule-Walker equations are obtained for martingale differences innovations. Then, from these properties, a statistical test to determine the order p is built and estimates of the autoregressive coefficients and the innovations variance are also obtained.

Keywords : Spatial AR models, Yule-Walker equations, order selection, autoregressive coefficients estimates, prediction, space-time data.

1. Introduction

Ces dernières années, l'utilisation de modèles spatiaux pour l'analyse de données récoltées sur un réseau s'est largement développée notamment en géologie, en économie ou encore en océanographie. Les modèles les plus largement étudiés sont les champs de Markov et de Gibbs (voir [3]) qui fournissent

* CEREMADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16 France.

des algorithmes fréquemment utilisés pour la restauration d'images. Parmi les modèles spatiaux figurent aussi les modèles AR quadrantaux (voir [10], [5], [2], [3]) qui étendent à un réseau la notion de processus AR temporel et les modèles dérivés SAR (Simultaneously specified Autoregressive model) et CAR (Conditionally specified Autoregressive model) (voir [11], [1]) qui sont de plus en plus couramment utilisés. En ce qui concerne l'étude de données spatio-temporelles, les modèles de Markov bilatéraux ne sont pas bien adaptés à la structure causale du temps. Des modèles semi-causaux permettraient de prendre en compte la causalité temporelle mais admettent une représentation causale infinie et de très nombreuses observations sont alors nécessaires pour en obtenir une bonne approximation. Quant aux modèles SAR et CAR, ils permettent de modéliser une structure spatiale sur un réseau non nécessairement régulier mais sont majoritairement employés dans le cadre de maillages finis car l'inférence sur le modèle infini conduit à une vraisemblance dépendant de données non observées (voir [11]). Nous montrons ici, par l'étude d'un cas concret, que les modèles AR causaux et plus particulièrement les modèles AR quadrantaux stationnaires étudiés dans [9], peuvent être adaptés à l'étude de certaines données indexées par le temps et l'espace. Outre la relativement faible dimension paramétrique du modèle, nous montrons que les modèles AR quadrantaux, de par leur linéarité, présentent aussi l'avantage de fournir un algorithme à la fois simple et rapide pour calculer les prédictions. Plus précisément, nous étudions des données de températures mensuelles le long de l'équateur à la surface de l'océan Pacifique. Dans cette étude, nous ajustons des modèles AR quadrantaux sur plusieurs ensembles de données issus de [8] et calculons des prédictions d'abord à la frontière du domaine d'observation puis progressivement plus éloignées dans le but de détecter certains phénomènes océanographiques tel que El Niño. Les prédictions que nous obtenons permettent, dans les cas que nous avons envisagés, de prédire les phénomènes réellement observés et nous confortent dans l'utilisation de telles modélisations.

Dans le but d'identifier les paramètres du modèle, nous nous intéressons préalablement aux propriétés asymptotiques d'une famille d'estimateurs construits à partir des équations de Yule-Walker et convenablement normalisés afin d'éviter les problèmes de biais du théorème de la limite centrale (voir [4]). Pour la définition de ces estimateurs, nous nous référons à l'article [7] et pour les preuves de ces propriétés nous nous appuyons sur l'article [10] de Tjøstheim. Les estimateurs considérés dans cet article, ont été introduits et étudiés dans [7] pour des champs ARMA quadrantaux d'innovations indépendantes et identiquement distribuées. Nous étudions ici la consistance et la normalité asymptotique des estimateurs de [7] dans le cas particulier de champs autorégressifs mais pour des innovations vérifiant l'hypothèse de différence de martingales quadrantales de [10]. Nous observons ensuite que l'estimateur des moindres carrés des coefficients autorégressifs, étudié dans [10], fait partie de la famille d'estimateurs considérée et permet alors d'estimer le vecteur des coefficients autorégressifs lorsque l'ordre du modèle est connu. Dans le cas où l'ordre du modèle est inconnu, les propriétés asymptotiques de l'ensemble de la famille d'estimateurs nous permettent de construire une

procédure pour le déterminer. Pour cela, nous introduisons tout d'abord un estimateur consistant de la variance du champ d'innovations et construisons enfin pour déterminer l'ordre du modèle, un test statistique basé sur la normalité asymptotique de la fonction d'autocorrélation partielle spatiale généralisée empirique dont la variance asymptotique dépend de la variance des innovations. Une validation de cette procédure, d'identification est effectuée sur des données simulées dans [6].

2. Modèle AR spatial quadrantal

Nous rappelons la définition de modèle AR(p) quadrantal de Tjøstheim (voir [9], [10]) et étudions ensuite, dans le cas particulier de champs autorégressifs, les propriétés asymptotiques des estimateurs introduits dans [7] pour l'étude du cas général ARMA. Puis, nous appliquons ces propriétés à la construction d'un test permettant de déterminer l'ordre p du modèle.

2.1. Description du modèle

Les champs que nous considérons sont des champs indexés par le réseau \mathbb{Z}^d ($d \geq 2$) muni de l'ordre partiel usuel. Ainsi, pour deux point s et t de \mathbb{Z}^d , nous notons $s \preceq t$ lorsque pour tout $i = 1 \dots d$, $s_i \leq t_i$. Par ailleurs, pour deux points a et b de \mathbb{Z}^d tels que $a \preceq b$, $a \neq b$, nous définissons les ensembles suivants :

$$\begin{aligned} S[a, \infty] &= \{x \in \mathbb{Z}^d \mid a \preceq x\}, \\ S[a, b] &= \{x \in \mathbb{Z}^d \mid a \preceq x \preceq b\}, \\ S\langle a, b \rangle &= [a, b] \setminus \{a\}. \end{aligned}$$

Enfin, pour tout entier n , nous notons \mathbf{n} l'élément de \mathbb{Z}^d dont toutes les composantes sont égales à n .

Comme défini dans [10], un champ $X = (X_t)_{t \in \mathbb{Z}^d}$ à valeurs réelles est appelé champ AR quadrantal d'ordre p dans \mathbb{N}^d s'il vérifie pour tout t dans \mathbb{Z}^d ,

$$X_t - \sum_{j \in S\langle \mathbf{0}, p \rangle} \phi_j X_{t-j} = \epsilon_t \quad (1)$$

avec $(\epsilon_t)_{t \in \mathbb{Z}^d}$ un champ centré, stationnaire au second ordre et de variance $\sigma^2 > 0$ tel que pour tous s et t dans \mathbb{Z}^d , $s \neq t$:

$$\mathbb{E}(\epsilon_t \epsilon_s) = 0.$$

Les paramètres du modèle (1) sont l'ordre p , la variance σ^2 du bruit $(\epsilon_t)_{t \in \mathbb{Z}^d}$ et les coefficients autorégressifs $(\phi_j)_{j \in S\langle \mathbf{0}, p \rangle}$.

Afin d'étudier les propriétés asymptotiques de certains estimateurs des coefficients autorégressifs d'un modèle AR quadrantal, nous supposons que le champ autorégressif X est causal i.e. qu'il admet une écriture unilatérale de la forme suivante :

$$X_t = \sum_{j \in S[\mathbf{0}, \infty]} \psi_j \epsilon_{t-j}, \quad (2)$$

Pour tout $z = (z_1, \dots, z_d)$ dans \mathbb{C}^d et tout $j = (j_1, \dots, j_d)$ dans \mathbb{N}^d , nous notons z^j le produit $z_1^{j_1} \dots z_d^{j_d}$ et définissons $\phi(z) = 1 - \sum_{j \in S(\mathbf{0}, p]} \phi_j z^j$ le polynôme autorégressif. Une condition suffisante pour que le champ (1) soit causal (voir [9] Théorème 1.5) est alors que le polynôme autorégressif ϕ n'admette pas de zéros dans le polydisque unité fermé $\overline{D}(0, 1)^d$ de \mathbb{C}^d . Désormais, nous supposons cette condition vérifiée et les champs AR que nous considérons sont donc des champs stationnaires au second ordre.

2.2. Définition des estimateurs

Dans le but d'estimer les coefficients autorégressifs et de déterminer les ordres d'un modèle ARMA(p, q) quadrantal, des estimateurs des coefficients solutions des équations de Yule-Walker généralisées ont été introduits dans [7] et leurs propriétés asymptotiques y sont étudiées pour $p \neq 0$ et pour des bruits indépendants et identiquement distribués. Nous rappelons ci-dessous la définition de ces estimateurs dans le cas autorégressif et introduisons par ailleurs des estimateurs $\hat{\sigma}_\lambda^2$ de la variance σ^2 lorsque $\lambda \succeq p$.

Considérons un champ AR(p) causal quadrantal X de fonction de covariance $\gamma(\cdot)$ et vérifiant la relation (1). Pour tout $\lambda \in S(\mathbf{0}, \infty]$, introduisons le vecteur de coefficients $\Phi_\lambda = (\phi_{\lambda, i})_{i \in S(\mathbf{0}, \lambda]}$ ordonné par l'ordre lexicographique et solution, lorsqu'elle existe, des équations de Yule-Walker,

$$\mathbb{E}(Y_{\lambda, t} X_{t-j}) = 0 \quad \forall j \in S(\mathbf{0}, \lambda] \quad (3)$$

où

$$Y_{\lambda, t} = X_t - \sum_{i \in S(\mathbf{0}, \lambda]} \phi_{\lambda, i} X_{t-i}.$$

En ordonnant les termes de $S(\mathbf{0}, \lambda]$ par l'ordre lexicographique, les équations (3) se résument par l'équation matricielle suivante :

$$\Gamma_\lambda \Phi_\lambda = \gamma_\lambda \quad (4)$$

avec pour tous $i, j \in S(\mathbf{0}, \lambda]$,

$$\Gamma_\lambda(j, i) = \gamma(j - i)$$

et

$$\gamma_\lambda(j) = \gamma(j).$$

Bien que les coefficients de la matrice Γ_λ et du vecteur γ_λ ne dépendent pas explicitement de l'indice λ , nous avons choisi des notations dépendantes de λ pour garder en mémoire que les matrices et les vecteurs sont indexés par les éléments de $S(\mathbf{0}, \lambda]$.

Remarque 2.1. — Un champ AR(p) causal vérifie les équations de Yule-Walker (4) pour $\lambda = p$ et Φ_λ est égal au vecteur des coefficients autorégressifs $(\phi_j)_{j \in S(\mathbf{0}, p]}$.

Remarque 2.2. — Lorsque X est un champ AR(p) séparable (voir [2]) et $\lambda \succeq p$, la matrice Γ_λ est inversible et l'équation (4) admet une solution unique.

En effet, si par exemple X est un champ $\text{AR}(p_1) \times \text{AR}(p_2)$, on a

$$\Gamma_{(\lambda_1, \lambda_2)} = \Gamma_{\lambda_1} \otimes \Gamma_{\lambda_2}$$

où pour $i = 1, 2$ Γ_{λ_i} désigne la matrice des covariances d'ordre λ_i d'un processus $\text{AR}(p_i)$. Or, la matrice Γ_{λ_i} est inversible si $\lambda_i \geq p_i$. On en déduit donc que pour $\lambda \succeq p$, $\Gamma_{(\lambda_1, \lambda_2)}$ est inversible et que

$$\Gamma_{(\lambda_1, \lambda_2)}^{-1} = \Gamma_{\lambda_1}^{-1} \otimes \Gamma_{\lambda_2}^{-1}$$

Plus généralement, le Lemme 2.3 conduit à la détermination des coefficients $(\phi_{\lambda, j})_{j \in S(\mathbf{0}, \lambda]}$ solution des équations de Yule-Walker (3) lorsque $\lambda \succeq p$.

LEMME 2.3. — Si $(X_t)_{t \in \mathbb{Z}^d}$ est un champ $\text{AR}(p)$ causal de bruit blanc faible $(\epsilon_t)_{t \in \mathbb{Z}^d}$, alors pour tout $\lambda \succeq p$,

$$Y_{\lambda, t} = \epsilon_t \text{ p.s.}$$

Preuve. — Dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$, définissons pour tout $t \in \mathbb{Z}^d$,

$$H_{\lambda, t} = \text{CL}\{X_{t-j}, j \in S(\mathbf{0}, \lambda)\},$$

où $\text{CL}\{\cdot\}$ désigne l'ensemble des combinaisons linéaires de $\{\cdot\}$. Par définition de $Y_{\lambda, t}$,

$$Y_{\lambda, t} = X_t - P_{H_{\lambda, t}} X_t,$$

où $P_{H_{\lambda, t}} X_t$ est la projection dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ de X_t sur $H_{\lambda, t}$. Démontrons alors que

$$P_{H_{\lambda, t}} X_t = \sum_{j \in S(\mathbf{0}, p]} \phi_j X_{t-j}.$$

Puisque $\lambda \succeq p$, on en déduit que $S(\mathbf{0}, p] \subseteq S(\mathbf{0}, \lambda]$ et donc que

$$\sum_{j \in S(\mathbf{0}, p]} \phi_j X_{t-j} \in H_{\lambda, t}.$$

De plus, pour tout $i \in S(\mathbf{0}, \lambda]$,

$$\mathbb{E}(X_{t-i}(X_t - \sum_{j \in S(\mathbf{0}, p]} \phi_j X_{t-j})) = \mathbb{E}(X_{t-i} \epsilon_t)$$

car $(X_t)_{t \in \mathbb{Z}^d}$ vérifie (1). Puisque $(X_t)_{t \in \mathbb{Z}^d}$ est causal, pour tout $i \in S(\mathbf{0}, \lambda]$,

$$\mathbb{E}(X_{t-i}(X_t - \sum_{j \in S(\mathbf{0}, p]} \phi_j X_{t-j})) = \sum_{k \in S[\mathbf{0}, \infty]} \psi_k \mathbb{E}(\epsilon_t \epsilon_{t-i-k}).$$

Or, l'hypothèse de différence de martingales implique que pour tout $i \in S\langle \mathbf{0}, \lambda \rangle$ et tout $k \in S[\mathbf{0}, \infty]$, $\mathbb{E}(\epsilon_t \epsilon_{t-i-k}) = 0$. Par suite, pour tout $i \in S\langle \mathbf{0}, \lambda \rangle$,

$$\mathbb{E}(X_{t-i}(X_t - \sum_{j \in S\langle \mathbf{0}, p \rangle} \phi_j X_{t-j})) = 0.$$

□

Remarque 2.4. — D'après l'unicité de la projection $P_{H_{\lambda,t}} X_t$, la démonstration précédente permet d'en déduire lorsque $\lambda \succeq p$ que pour tout $j \in S\langle \mathbf{0}, p \rangle$,

$$\phi_{\lambda,j} = \phi_j$$

et que pour tout $j \in S\langle \mathbf{0}, \lambda \rangle \setminus S\langle \mathbf{0}, p \rangle$,

$$\phi_{\lambda,j} = 0.$$

D'autre part, le champ X étant un champ AR causal, par application du Lemme 2.3, il vérifie pour tout $\lambda \succeq p$ l'équation suivante :

$$\mathbb{E}(Y_{\lambda,t} X_t) = \sigma^2 \tag{5}$$

que l'on peut réécrire sous la forme

$$\sigma^2 = \gamma_\lambda(\mathbf{0}) - \Phi_\lambda^T \gamma_\lambda \tag{6}$$

où V^T désigne le vecteur ligne associé à tout vecteur colonne V .

Enfin, pour une observation $\{X_t, t \in S[\mathbf{1}, \mathbf{N}]\}$ du champ, nous considérons l'estimateur $\hat{\Phi}_\lambda$ de Φ_λ solution, lorsqu'elle existe, de l'équation matricielle

$$\hat{\Gamma}_\lambda \hat{\Phi}_\lambda = \hat{\gamma}_\lambda \tag{7}$$

avec pour tous $i, j \in S\langle \mathbf{0}, \lambda \rangle$,

$$\hat{\Gamma}_\lambda(j, i) = \frac{1}{N^d} \sum_{t \in S[\mathbf{1}+\lambda, \mathbf{N}]} X_{t-i} X_{t-j}$$

et

$$\hat{\gamma}_\lambda(j) = \frac{1}{N^d} \sum_{t \in S[\mathbf{1}+\lambda, \mathbf{N}]} X_t X_{t-j}.$$

Remarque 2.5. — Dans le cas où $\lambda = p$, l'estimateur $\hat{\Phi}_p$ des coefficients autorégressifs est l'estimateur des moindres carrés étudiés par Tjøstheim (voir [9], [10]). Il diffère de l'estimateur de Yule-Walker qui présente en dimension deux un biais lors de l'étude de la normalité asymptotique (voir [4]).

Introduisons par ailleurs pour tout $\lambda \succeq p$, l'estimateur $\hat{\sigma}_\lambda^2$ de la variance σ^2 du bruit :

$$\hat{\sigma}_\lambda^2 = \hat{\gamma}_\lambda(\mathbf{0}) - \hat{\Phi}_\lambda^T \hat{\gamma}_\lambda \tag{8}$$

où

$$\hat{\gamma}_\lambda(\mathbf{0}) = \frac{1}{N^d} \sum_{t \in S[1+\lambda, \mathbf{N}]} X_t^2.$$

2.3. Propriétés des estimateurs

De la même manière que Tjøstheim dans [10], nous supposons dans cette section que le bruit $(\epsilon_t)_{t \in \mathbb{Z}^d}$ est une différence de martingales quadrantes sur l'espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$. Pour cela, introduisons pour tout $t \in \mathbb{Z}^d$, la filtration naturelle $\{\mathcal{F}_t, t \in \mathbb{Z}^d\}$ associé au champ $(\epsilon_t)_{t \in \mathbb{Z}^d}$. En d'autres termes, pour tout $t \in \mathbb{Z}^d$,

$$\mathcal{F}_t = \sigma\{\epsilon_s, s \preceq t\},$$

où $\sigma\{\cdot\}$ désigne la tribu engendrée par la famille de variables aléatoires $\{\cdot\}$. Notons enfin pour tout $u \in \mathbb{Z}$ et tout $i = 1 \dots d$,

$$\mathcal{F}^i(u) = \vee_{t|t_i \leq u} \mathcal{F}_t$$

où pour deux tribus \mathcal{F} et \mathcal{G} , $\mathcal{F} \vee \mathcal{G}$ désigne la tribu engendrée par la réunion des tribus \mathcal{F} et \mathcal{G} . Alors, $(\epsilon_t)_{t \in \mathbb{Z}^d}$ est une différence de martingale quadrante si pour tout $t = (t_1, \dots, t_d) \in \mathbb{Z}^d$,

$$\mathbb{E}(\epsilon_t | \vee_{i=1}^d \mathcal{F}^i(t_i - 1)) = 0 \quad p.s.$$

En procédant de la même manière que dans la preuve du Théorème 4.1 de [10], nous obtenons, sous les mêmes hypothèses, la consistance des covariances empiriques vers les covariances théoriques. Les équations (4) et (7) d'une part et les équations (6) et (8) d'autre part conduisent alors au résultat suivant.

THÉORÈME 2.6. — *Soit $(X_t)_{t \in \mathbb{Z}^d}$ un champ $AR(p)$, $p \neq \mathbf{0}$ causal dont le bruit $(\epsilon_t)_{t \in \mathbb{Z}^d}$ est un champ strictement stationnaire de différences de martingales quadrantes tel que pour tout $t \in \mathbb{Z}^d$,*

$$\mathbb{E}(\epsilon_t^4) < \infty$$

et

$$\mathbb{E}(\epsilon_t^2 | \vee_{i=1}^d \mathcal{F}^i(t_i - 1)) = \sigma^2 \quad p.s.$$

Pour tout $\lambda \in S(0, \infty]$ tel que Γ_λ est inversible,

$$\hat{\Phi}_\lambda \xrightarrow[N \rightarrow \infty]{p.s.} \Phi_\lambda$$

et si de plus $\lambda \succeq p$,

$$\hat{\sigma}_\lambda^2 \xrightarrow[N \rightarrow \infty]{p.s.} \sigma^2.$$

Pour obtenir la normalité asymptotique de l'estimateur $\hat{\Phi}_\lambda$, on se ramène comme dans [10] pour l'étude de la normalité asymptotique de l'estimateur

des moindres carrés, à l'étude du vecteur $N^{d/2}\hat{\Gamma}_\lambda(\hat{\Phi}_\lambda - \Phi_\lambda)$ dont la j ème composante est donnée par

$$\frac{1}{N^d} \sum_{t \in S[1+\lambda, \mathbf{N}]} X_{t-j} Y_{\lambda, t}.$$

Cette étude peut être menée identiquement à celle du Théorème 4.2 de [10] grâce au Lemme 2.3.

THÉORÈME 2.7. — *Si $(X_t)_{t \in \mathbb{Z}^d}$ est un champ AR(p) $p \neq \mathbf{0}$ vérifiant les hypothèses du Théorème 2.6 et tel que Γ_λ est inversible, alors,*

$$N^{d/2}(\hat{\Phi}_\lambda - \Phi_\lambda) \xrightarrow[N \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 \Gamma_\lambda^{-1}).$$

Preuve. — Par application du lemme 2.3, $Y_{\lambda, t} = \epsilon_t$ et la j ème composante de $N^{d/2}\hat{\Gamma}_\lambda(\hat{\Phi}_\lambda - \Phi_\lambda)$ est

$$\frac{1}{N^{d/2}} \sum_{t \in S[1+\lambda, \mathbf{N}]} X_{t-j} \epsilon_t.$$

Le procédé de Cramer-Wold pour établir la normalité asymptotique conduit alors à étudier pour toute famille de réels $(r_u)_{u \in S[0, \lambda]}$, la convergence en loi des statistiques

$$\frac{1}{N^{d/2}} \sum_{t \in S[1+\lambda, \mathbf{N}]} \sum_{u \in S[\lambda]} r_u X_{t-u} \epsilon_t.$$

Enfin, en remplaçant dans le terme précédent X_{t-u} par son expression unilatérale (2), aux effets de bords près (voir [4]), on est alors ramené pour tout $K \in \mathbb{N}^d$ grand, à l'étude de la convergence en loi de termes de la forme

$$\frac{1}{N^{d/2}} \sum_{t \in S[1+\lambda, \mathbf{N}]} \sum_{u \in S[0, \lambda]} r_u \sum_{v \in S[0, \mathbf{K}]} \psi_v \epsilon_{t-u-v} \epsilon_t. \quad (9)$$

Le terme (9) est identique au terme (4.17) considéré dans [10] et vérifie donc les mêmes propriétés de martingale spatiale qui permettent, par application du Lemme 4.4. de [10], d'en obtenir la normalité asymptotique. \square

2.4. Procédure d'identification du paramètre p

Lorsque le paramètre p est connu, nous disposons d'un estimateur $\hat{\Phi}_p$ des coefficients autorégressifs Φ_p du modèle AR(p) ainsi que d'un estimateur $\hat{\sigma}_p^2$ de la variance σ^2 du bruit. Dans le cas où l'ordre p du modèle est inconnu, nous allons utiliser pour le déterminer les propriétés asymptotiques de la composante d'indice λ des estimateurs $\hat{\Phi}_\lambda$ lorsque $\lambda \succeq p$ et construire un test statistique analogue à celui employé pour la sélection de l'ordre d'un processus AR temporel.

La Remarque 2.4, combinée au Théorème 2.7, permet en particulier d'obtenir pour $\lambda \succeq p$ et $\lambda \neq p$ que

$$N^{d/2} \hat{\phi}_{\lambda,\lambda} \xrightarrow[N \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 w_\lambda)$$

où w_λ est le terme d'indice (λ, λ) de la matrice Γ_λ^{-1} . Si l'on note \hat{w}_λ le terme d'indice (λ, λ) de la matrice $\hat{\Gamma}_\lambda^{-1}$, la consistance des covariances empiriques et le Théorème 2.6, permettent d'en déduire que pour $\lambda \succeq p$ et $\lambda \neq p$,

$$N^{d/2} \frac{\hat{\phi}_{\lambda,\lambda}}{\hat{\sigma}_\lambda \sqrt{\hat{w}_\lambda}} \xrightarrow[N \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1). \quad (10)$$

Cette dernière propriété permet de construire un test statistique pour tester $H_0 : p = p_0$ contre $H_1 : p \succeq p_0, p \neq p_0$. En effet, sous H_0 , pour tout $\lambda \succeq p_0$ et $\lambda \neq p_0$ $\phi_{\lambda,\lambda} = 0$ et la propriété (10) est vérifiée. Sous H_1 , il existe j dans $S(\mathbf{0}, p] \setminus S(\mathbf{0}, p_0]$ tel que $\phi_{\lambda,j} \neq 0$. Quitte à surévaluer la valeur de l'ordre p mais sans modification des coefficients autorégressifs d'après la Remarque 2.4, on peut supposer que $j = \lambda$ et adopter la règle de décision suivante : on se donne une constante $K \in \mathbb{N}$ grande, un risque α et on décide d'accepter H_0 si pour tout $\mathbf{K} \succeq \lambda \succeq p_0, \lambda \neq p_0$,

$$|\hat{\phi}_{\lambda,j}| < u_{1-\alpha/2} \frac{\hat{\sigma}_\lambda \sqrt{\hat{w}_\lambda}}{N^{d/2}} \quad (11)$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ d'une loi normale $\mathcal{N}(0, 1)$.

Remarque 2.8. — Pour tout λ , l'intervalle défini par la relation (11) ne dépend pas de la valeur p_0 testée. Ainsi, la connaissance de ces intervalles permet de tester simultanément différentes valeurs p_0 .

Comme nous le verrons dans la section suivante, ce test statistique peut être mis en oeuvre de manière graphique en dimension 2 et la procédure de détermination du paramètre p s'avère plus pratique que la méthode décrite dans [7] appliquée au cas autorégressif.

3. Analyse de données océanographiques

Dans cette section nous appliquons les résultats précédents à une étude de données océanographiques spatialisées. Sur plusieurs ensembles de données, nous présentons tout d'abord la version graphique du test permettant de déterminer l'ordre du modèle puis calculons les estimations des coefficients autorégressifs et une estimation de la variance. Enfin, le modèle ajusté, nous calculons selon un schéma bien précis des prévisions et les comparons aux données réellement observées afin d'évaluer la pertinence du modèle.

3.1. Description des données

Les données que nous considérons ici sont des données de températures mensuelles au 1er janvier dans l’océan Pacifique (voir [8] SST data) relevées en surface au moyen de balises, de bateaux et plus récemment de satellites. Ces données sont disponibles du 1er janvier 1950 au 1er janvier 1999 le long de l’équateur tous les deux degrés de 150°E à 80°W (longitude du continent Sud Américain à l’équateur). Elles sont donc situées sur une grille régulière (voir tableau 1) de taille 65 × 600 avec pour chaque mois, 65 observations le long de l’équateur et pour chaque degré, 600 observations mensuelles.

TABLEAU 1. — Grille des données globales

Jan 1999	(1,600)	(2,600)	(64,600)	(65,600)
Déc 1998	(1,599)	(2,599)				(64,599)	(65,599)
	⋮						⋮
Mois n° t	(1, t)			(x,t)			(65, t)
	⋮						⋮
Fév 1950	(1,2)	(2,2)	(64,2)	(65,2)
Jan 1950	(1,1)	(2,1)	(64,1)	(65,1)
	150°E	152°E		180°		82°W	80°W
	Degré n° x						

Afin d’utiliser une modélisation AR quadrantale stationnaire, nous travaillons à partir des données de variations annuelles (voir [8] SST anomaly data) obtenues en retranchant aux données décrites ci-dessus les phénomènes annuels et semi-annuels au moyen de processus moyenne mobile classiques dans le but de détecter par exemple les grandes périodes chaudes appelées El Niño ou les phases froides de La Niña.

Dans cette étude nous nous intéressons plus particulièrement à la prédiction des phases chaudes de El Niño qui consistent en un réchauffement de la température de l’océan à l’équateur au niveau du continent Sud-américain au début des mois d’été jusqu’au mois de mai ou juin de l’année suivante, passant par une période de chaleur maximale au mois de décembre. Selon son intensité, El Niño peut engendrer de nombreux effets : perturbations climatiques importantes, diminution du nombre de poissons, dommages des récifs de corail... Dans le but d’anticiper ces effets, de nombreux scientifiques ont travaillé à la prévision de ce phénomène au moyen de modèles océan-atmosphère basés sur la corrélation du phénomène avec le comportement

des températures de surface, de la pression atmosphérique, des alizés ou encore des précipitations. Ces modèles déterministes ou non (voir [8], [12]) sont performants au niveau de la qualité de leurs prévisions. Sans prétendre obtenir la même qualité de prévision, nous effectuons ici des prédictions sur les températures de surface pour l'année 1993 où le phénomène ne s'est pas produit et pour l'année 1997 où le phénomène a pu être observé. Au moyen de modélisation AR spatiale, nous calculons une prédiction à 6 mois du mois de juillet au mois de décembre et comparons ensuite ces prédictions au phénomène qui a été réellement observé. Pour ajuster le modèle aux données nous effectuons un apprentissage sur une période de deux ans précédant le mois de juillet correspondant à un maillage de taille 65×24 .

3.2. Ajustement d'un modèle AR

La première étape pour ajuster un modèle $AR(p)$ est la détermination du paramètre p du modèle. Pour chacune des années 1993 et 1997, nous calculons tout d'abord les estimateurs $\hat{\phi}_{\lambda, \lambda}$ pour des valeurs de λ dans le rectangle $S(\mathbf{0}, \mathbf{15})$ et construisons dans chaque cas, au risque $\alpha = 5\%$, une grille 15×15 où au point λ apparaît le symbole * si la relation (11) est vérifiée et le symbole \circ sinon. Les graphiques obtenus correspondent aux figures 1 et 2 et permettent d'après le Remarque 2.8 de tester différentes valeurs de p . Si l'on se restreint aux petites valeurs de $p = (p_x, p_t)$, l'observation de la figure 1 pour l'année 1993 conduit à choisir $p = (5, 2)$. Pour l'année 1997, le choix n'est pas aussi facile : au vu de la figure 2, il est possible de choisir $p = (4, 2)$ mais les \circ aux points de coordonnées $(5, 3)$ et $(6, 4)$ inciteraient plutôt à choisir $p = (5, 3)$ ou encore $(6, 4)$.

L'ordre p étant déterminé, le vecteur $\hat{\Phi}_p$ donne une estimation des coefficients autorégressifs. Les estimations obtenues sont résumées dans les tableaux 2, 3, 5 et 6. Par ailleurs, dans les tableaux 2 et 3, nous avons noté entre parenthèses l'approximation fournie par le Théorème 2.7. pour écarts-types des paramètres estimés. Dans le cas de l'année 1997, nous avons effectué les calculs pour les valeurs de p suivantes : $p = (4, 2)$, $p = (5, 3)$ et $p = (6, 4)$. Si l'on observe les coefficients autorégressifs en bordure (coefficients en gras) dans le tableau 5 pour $p = (5, 3)$, on constate qu'ils sont majoritairement voisins de zéro comparativement aux autres coefficients du tableau. Quant aux coefficients non situés en bordure, ils sont globalement proches des coefficients autorégressifs du tableau 3 obtenus pour $p = (4, 2)$. De même, le tableau 5 peut être vu comme proche d'un sous-tableau du tableau 6 où les coefficients en bordure sont également majoritairement voisins de zéros. Au vu de la Remarque 2.4, nous choisissons finalement $p = (4, 2)$. Nous remarquons ici que l'application de la méthode pour déterminer les ordres d'un modèle ARMA décrite dans [7] conduit, si l'on souhaite ajuster un modèle AR, au même choix de p sur les deux bases d'apprentissage considérées dans cet article.

Enfin, les estimations $\hat{\sigma}_p$ de l'écart-type du bruit sont données dans le tableau 4.

UNE MODÉLISATION DE DONNÉES SPATIO-TEMPORELLES

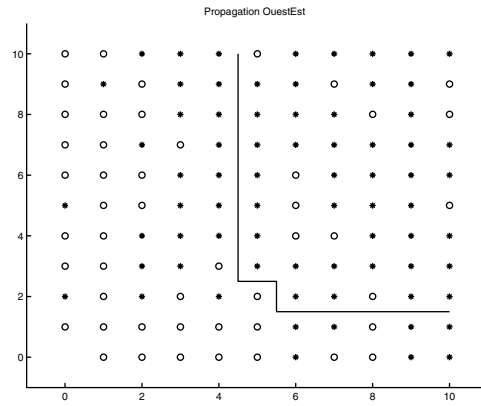


FIG 1. — Sélection de l'ordre p : 1993

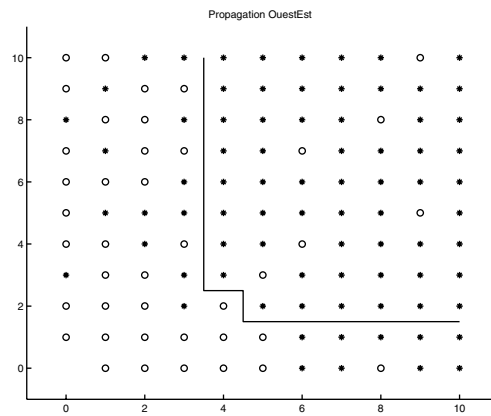


FIG 2. — Sélection de l'ordre p : 1997

TABEAU 2. — Estimation des coefficients autorégressifs : 1993

2	0.1316 (0.0263)	-0.2365 (0.0525)	0.2249 (0.0584)	-0.0737 (0.0616)	-0.1784 (0.0598)	0.1390 (0.0327)
1	0.4945 (0.0255)	-0.8653 (0.0512)	0.4423 (0.0600)	-0.0512 (0.0638)	-0.1201 (0.0616)	0.1166 (0.0340)
0		1.7182 (0.0257)	-0.9201 (0.0509)	0.1380 (0.0594)	0.3039 (0.0575)	-0.2666 (0.0314)
\hat{i}_t/\hat{i}_x	0	1	2	3	4	5

UNE MODÉLISATION DE DONNÉES SPATIO-TEMPORELLES

TABLEAU 3. — Estimation des coefficients autorégressifs : 1997 – $p = (4, 2)$

2	0.1865 (0.0272)	-0.3177 (0.0538)	0.2026 (0.0620)	-0.1641 (0.0602)	0.0929 (0.0322)
1	0.5356 (0.0259)	-0.9021 (0.0518)	0.5557 (0.0627)	-0.2313 (0.0623)	0.0463 (0.0346)
0		1.7199 (0.0255)	-1.1953 (0.0495)	0.7185 (0.0540)	-0.2548 (0.0300)
i_t/i_x	0	1	2	3	4

TABLEAU 4. — Estimation de l'écart-type

$\hat{\sigma}_p$	Ouest-Est
1993	0.0715
1997	0.0674 ($p = (4, 2)$)
	0.0634 ($p = (5, 3)$)
	0.0604 ($p = (6, 4)$)

TABLEAU 5. — Estimation des coefficients autorégressifs : 1997 – $p = (5, 3)$

3	0.1049 (0.0273)	-0.1720 (0.0533)	0.0769 (0.0613)	-0.0148 (0.0643)	-0.0692 (0.0600)	0.0721 (0.0330)
2	0.1483 (0.0289)	-0.2375 (0.0559)	0.1429 (0.0646)	-0.1169 (0.0679)	0.0684 (0.0633)	-0.0032 (0.0352)
1	0.5028 (0.0263)	-0.8348 (0.0520)	0.4706 (0.0634)	-0.1077 (0.0677)	-0.1640 (0.0632)	0.1435 (0.0351)
0		1.6808 (0.0255)	-1.0774 (0.0505)	0.4735 (0.0598)	0.1929 (0.0570)	-0.2891 (0.0309)
i_t/i_x	0	1	2	3	4	5

3.3. Prédiction par modélisation AR

Les modèles étant ajustés aux données, nous allons effectuer une prédiction du mois de juillet au mois de décembre pour les années 1993 et 1997. Lorsque les valeurs $\{x_{(u,v)-j}, j \in S(\mathbf{0}, p)\}$ sont connues, nous calculons la prédiction $\hat{x}_{(u,v)}$ en un point (u, v) du maillage selon la formule suivante :

$$\hat{x}_{(u,v)} = \hat{\varepsilon}_{(u,v)} + \sum_{j \in S(\mathbf{0}, (p_x, p_t))} \hat{\phi}_{p,j} x_{(u,v)-j} \quad (12)$$

où $\hat{\varepsilon}_{(u,v)}$ est une réalisation d'une loi normale $\mathcal{N}(0, \hat{\sigma}_p^2)$.

TABLEAU 6. — Estimation des coefficients autorégressifs : 1997 – $p = (6, 4)$

4	-0.1417 (0.0278)	0.1928 (0.0540)	-0.1194 (0.0619)	0.0611 (0.0648)	-0.0014 (0.0648)	0.1147 (0.0619)	-0.1083 (0.0337)
3	0.1605 (0.0286)	-0.2657 (0.0555)	0.1589 (0.0641)	-0.0470 (0.0675)	-0.0740 (0.0677)	0.0355 (0.0647)	0.0291 (0.0357)
2	0.1420 (0.0289)	-0.2049 (0.0564)	0.1002 (0.0658)	-0.0839 (0.0696)	0.0494 (0.0699)	-0.0323 (0.0665)	0.0335 (0.0363)
1	0.5392 (0.0260)	-0.8758 (0.0520)	0.5055 (0.0638)	-0.1591 (0.0683)	-0.1526 (0.0683)	0.1348 (0.0649)	0.0200 (0.0360)
0		1.6753 (0.0258)	-1.0823 (0.0500)	0.4868 (0.0586)	0.2101 (0.0601)	-0.3201 (0.0572)	0.0110 (0.0319)
i_t/i_x	0	1	2	3	4	5	6

Remarque 3.1. — Dans la formule (12), la prédiction $\hat{x}_{(u,v)}$ est optimale lorsque $\hat{\epsilon}_{(u,v)}$ est égal à zéro. Cependant, nous avons ici choisi un bruit $\hat{\epsilon}_{(u,v)}$ non nul dans le but d’obtenir plusieurs prédictions sur un même échantillon d’apprentissage. Il est ainsi possible, selon les démarches employées par les climatologues, de considérer la prédiction la plus alarmiste dans l’éventualité d’un phénomène El Niño.

Dans le cas où certaines valeurs de l’ensemble $\{x_{(u,v)-j}, j \in S(\mathbf{0}, p)\}$ ne sont pas connues, la prédiction est alors calculée selon la formule (12) en remplaçant ces valeurs manquantes par une estimation décrite ci-dessous. Plus précisément, nous disposons en pratique de l’échantillon d’apprentissage $\{x_{(u,v)}, u = 1 \dots 65, v = 1 \dots 24\}$ correspondant aux 24 mois précédant le mois de juillet de l’année d’étude. Pour calculer la prédiction au mois de juillet le long de l’équateur $\{\hat{x}_{(u,v)}, u = 1 \dots 65, v = 25\}$ à partir de la formule (12), nous devrions disposer de $x_{(u,25)}$ pour $u = 1 \dots p_x$. De même, pour le mois suivant, il serait nécessaire de connaître $x_{(u,26)}$ pour $u = 1 \dots p_x$. Par itération jusqu’au mois de décembre, nous devrions donc disposer de $x_{(u,v)}$ pour $u = 1 \dots p_x$ et $v = 25 \dots 30$. Ces quantités, utiles à la prédiction, sont indisponibles. Afin d’en obtenir des prédictions, nous avons choisi pour $u = 1$ de procéder par ajustement d’un modèle AR temporel de paramètre p_t sur les données $\{x_{(u,v)}, v = 1 \dots 24\}$ obtenant ainsi par prédictions successives les valeurs manquantes $\{\hat{x}_{(1,v)}, v = 25 \dots 30\}$. Lorsque $u = 2, \dots, p_x$, pour obtenir les prédictions $\{\hat{x}_{(u,v)}, v = 25 \dots 30\}$ nous ajustons à u fixé un modèle AR spatial de paramètre $(u - 1, p_t)$ sur les données $\{x_{(\tilde{u},v)}, \tilde{u} = 1 \dots u, v = 1 \dots 24\}$. Enfin, les prédictions restantes sont calculées par applications successives de la formule (12). Le tableau 7 résume le schéma de construction de la prédiction pour les six mois le long de l’équateur.

TABLEAU 7. — Schéma de construction

30	30	30
↑	↑	↑
AR(p_t)	AR($(u - 1, p_t)$)	AR((p_x, p_t))
↓	↓	↓
25	25	25
1	$2 \leftarrow u \rightarrow p_x$	$p_x + 1 \longleftrightarrow 65$
24		
↑		
Base d'apprentissage		
↓		
1		
$1 \longleftrightarrow 65$		

Les représentations graphiques de ces prévisions sont données par les figures 3 et 5 où la carte de gauche représente les données observées et la carte de droite est formée de la base d'apprentissage suivie de la prévision à six mois située dans le cadre noir. Si l'on observe les cartes de prédictions obtenues pour les années 1993 et 1997, on constate que la prédiction obtenue coïncide avec le phénomène observé : observation du phénomène El Niño simplement pour l'année 1997. Plus précisément, sur la figure 3, on observe une diminution de la température après le mois de juin 1993 sur la prédiction, permettant ainsi de prédire que le phénomène ne se produira pas en 1993. Pour l'année 1997, les résultats ne sont pas aussi nets mais la figure 5 permet d'envisager un phénomène El Niño puisque la température de l'eau donnée par la prédiction reste anormalement chaude jusqu'au mois de décembre.

Pour détecter le phénomène El Niño, il est également possible de calculer l'indice NINO3 qui consiste en la moyenne des données mensuelles de 150°W à 90°W . Sur les graphiques des figures 4 et 6, nous avons représenté l'indice calculé sur les données réelles et une estimation de l'indice du mois de juillet au mois de décembre sur les prédictions de températures. Pour l'année 1993, l'indice estimé est très proche de l'indice observé. En revanche, pour l'année 1997, on observe que l'estimation de l'indice est plus éloignée de l'indice réel mais reste suffisamment positive pour s'attendre à la manifestation de El Niño.

Au niveau de l'indice NINO3, les estimations obtenues par les océanographes sont plus précises que celles que nous obtenons à partir de la moyenne des prédictions par modélisation AR entre 150°W et 90°W . Cela s'explique en partie par le fait que les scientifiques connaissent de mieux en mieux les différents phénomènes localisés dans cette zone au niveau de l'interface océan-atmosphère et disposent par ailleurs d'équations physiques permettant de décrire en partie leurs interactions. Cependant, cet exemple simple encourage l'utilisation de modélisation AR spatiale comme outil d'étude de données

UNE MODÉLISATION DE DONNÉES SPATIO-TEMPORELLES

spatio-temporelles. Les procédures d'identification des paramètres du modèle, basées sur les estimateurs calculés à partir des formules (7) et (8), peuvent être mises en oeuvre très simplement et le calcul des prédictions par la formule (12) est très rapide. Dans un contexte où aucune information supplémentaire sur le phénomène observé n'est disponible, les modèles ARMA spatiaux offrent une voie possible d'étude de données spatialisées.

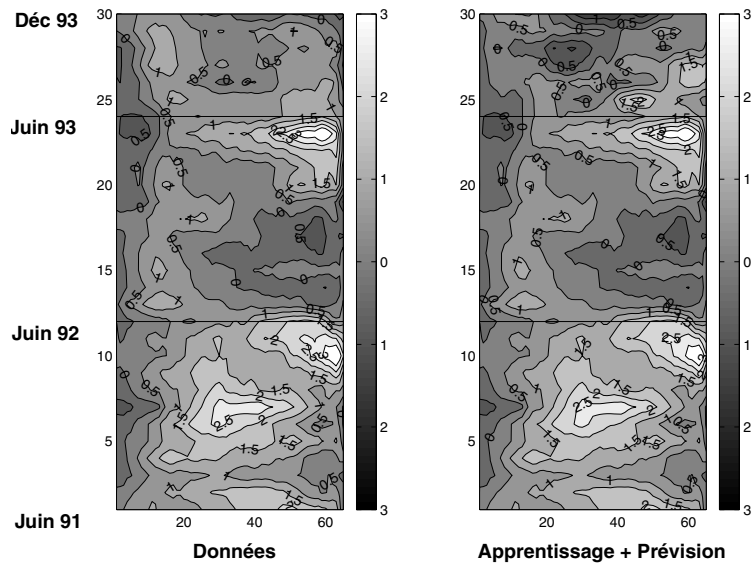


FIG 3. — Courbes de niveau des températures variant entre -3°C et 3°C : 1993

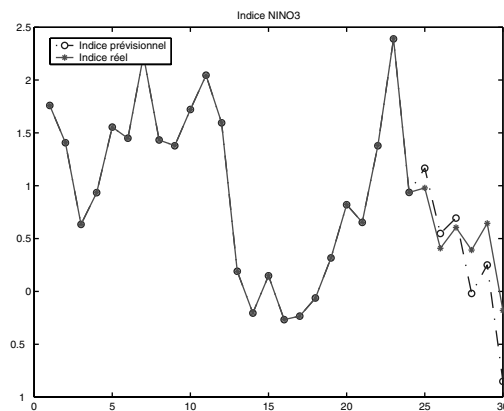


FIG 4. — Indice pour l'année 1993

UNE MODÉLISATION DE DONNÉES SPATIO-TEMPORELLES

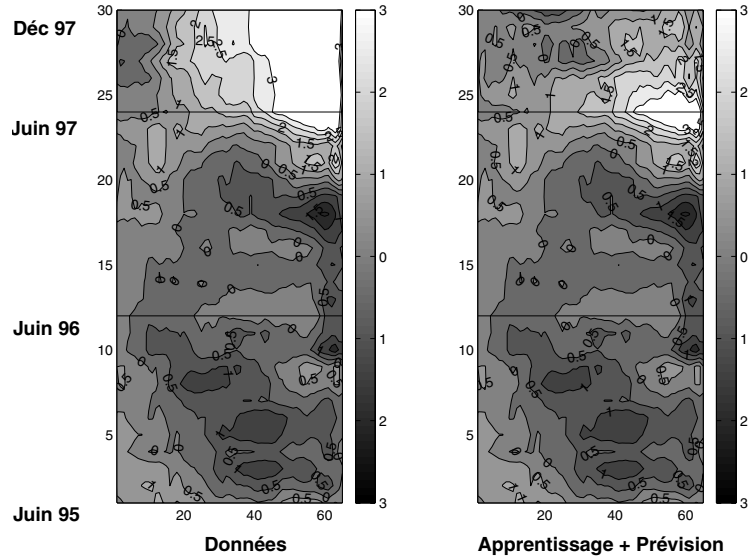


FIG 5. — Courbes de niveau de températures variant entre -3°C et 3°C : 1997



FIG 6. — Indice pour l'année 1997

Remerciements : Je remercie vivement Boris Dewitte et Sérèna Illig du LEGOS de Toulouse pour tous les renseignements qu'ils m'ont communiqués sur les données de températures de surface ainsi que pour toutes les explications qu'ils m'ont fournies sur les méthodes employées par les océanographes qui étudient le phénomène El Niño.

Références

- [1] CRESSIE Noel A. C. (1993) *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.
- [2] ETCHISON T., PANTULA S. G., and BROWNIE C. (1994). Partial autocorrelation function for spatial processes. *Statist. Probab. Lett.*, 21(1) :9-19.
- [3] GUYON X. (1995). *Random fields on a network*. Probability and its Applications (New York). Springer-Verlag, New York.
- [4] HA E. and NEWTON H. J. (1993). The bias of estimators of causal autoregressive processes. *Biometrika*, 80(1) :242-245.
- [5] HUANG D. and ANH V. V. (1992). Estimation of spatial ARMA models. *Austral. J. Statist.*, 34(3) :513-530.
- [6] ILLIG A. (2004). *Étude asymptotique de certains estimateurs dans des modèles ARMA spatiaux*. PhD thesis, Institut National des Sciences Appliquées de Toulouse.
- [7] ILLIG A. and TRUONG-VAN B. (2006). Asymptotic results for spatial ARMA models. À paraître dans *Comm. Statist. Theory Methods* 35(4).
- [8] REYNOLDS R. W. and SMITH T. M. (1994). Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, 7 :929-948.
- [9] TJØSTHEIM D. (1978). Statistical spatial series modelling. *Advances in Appl. Probability*, 10(1) :130-154.
- [10] TJØSTHEIM D. (1983). Statistical spatial series modelling. II. Some further results on unilateral lattice processes. *Adv. in Appl. Probab.*, 15(3) :562-584.
- [11] WHITTLE P. (1954). On stationary processes in the plane. *Biometrika*, 41 :434-449.
- [12] XUE Y., LEETMAA A., and JI M. (2000). Enso prediction with markov-models : The impact of sea level. *J. Climate*, 13 :849-971.