Université Paris 7 – Denis Diderot
ufr biologie et sciences de la nature

Année :

N° :

Thèse

pour l'obtention du diplôme de

Docteur de l'université Paris 7
Spécialité : Biomathématiques

présentée et soutenue publiquement par

Dominique Soudant

le X décembre 1997

Application de modèles dynamiques bayésiens aux séries temporelles de *Dinophysis* à Antifer (Normandie, France)

Directeur de thèse : Monsieur le professeur Guy Thomas

Jury

Alain-Jacques Valleron	Président
Serge Frontier	Rapporteur
Pierre Legendre	Rapporteur
Guy Thomas	
Alain Ménesguen	

Résumé :
Le *Dinophysis* est une microalgue toxique à l'origine d'épidémies diarrhéiques. En dépit de nombreuses études, l'écologie et la biologie de ce genre phytoplanctonique restent mal connues. Afin de déterminer les conditions d'apparition de *Dinophysis* cf. *acuminata* sur le site d'Antifer (Normandie, France), des prélèvements d'eau de mer ont été effectués quasi-journellement pendant les périodes estivales des années 1987 à 1993. Les séries temporelles résultantes concernent la concentration de la microalgue et les facteurs chimiques, physiques et météorologiques. L'analyse de ces données à l'aide d'outils statistiques classiques (e.g. régression) s'est avérée difficile (e.g. instabilité temporelle des résultats). Une caractéristique commune de ces méthodes est de supposer les relations étudiées invariables dans le temps.

Les modèles dynamiques bayésiens (mdb) sont des modèles de séries temporelles qui intègrent l'hypothèse de relations variables dans le temps. Le modèle de régression linéaire dynamique a été appliquée aux « données d'Antifer ». L'hypothèse de variation temporelle des relations s'est avérée pertinente en écologie phytoplanctonique, à travers les variations de valeurs des coefficients de la régression. De plus, ces variations ont pu être interprétée et ainsi aboutir à un schéma partiel d'explication des concentrations de *Dinophysis* à Antifer. Le pourcentage de variabilité expliqué par une régression dynamique est supérieur à celui de son équivalent statique. Enfin, certaines variables non-significatives en régression statique sont apparues significatives en régression dynamique. Au total, les mdb se sont avérés être une approche originale et féconde du traitement des séries temporelles écologiques. Leurs extensions laissent entrevoir la possibilité de nombreuses collaborations entre biostatistique et écologie.

Mots clés : modélisation, statistique, dynamique, séries temporelles, bayésien, *Dinophysis*, baie de Seine.

Discipline : biomathématiques

ufr biologie et sciences de la nature
2, place Jussieu
tour 54, 4$^{\text{ième}}$ étage
75 251 Paris cedex 05

Avant propos

Je tiens à témoigner toute ma reconnaissance à,

– Alain-Jacques Valleron, directeur de la formation doctorale de biomathématiques (Paris vi et vii), qui m'a accordé sa confiance ;

– Guy Thomas, qui a dirigé ce travail avec compétence ;

– Benoît Beliaeff, qui a assuré avec rigueur et efficacité l'encadrement local ;

– Robert Poggi, qui a accepté de m'accuellir au sein de son service ;

– Patrick Lassus, qui nous a confié les « données d'Antifer » et qui m'a apporté ses lumières sur *Dinophysis* ;

– X, Y, Z pour me faire l'honneur de leur présence au sein de ce jury ;

– Philippe Gros, pour la pertinence de ses commentaires et suggestions ;

– Catherine Belin, qui a toujours répondu à mes questions concernant le réseau de surveillance phytoplanctonique (rephy) ;

– les personnes que j'ai été amené à connaître au sein de l'ifremer, pour leur gentillesse et leur amitié.

Cet avant propos est également l'occasion de remercier les membres de ma famille et mes amis pour leur soutien, et pour certain, leur patience.

Résumé

Le *Dinophysis* est une microalgue toxique à l'origine d'épidémies diarrhéiques. En dépit de nombreuses études, l'écologie et la biologie de ce genre phytoplanctonique restent mal connues. Afin de déterminer les conditions d'apparition de *Dinophysis* cf. *acuminata* sur le site d'Antifer (Normandie, France), des prélèvements d'eau de mer ont été effectués quasi-journellement pendant les périodes estivales des années 1987 à 1993. Les séries temporelles résultantes concernent la concentration de la microalgue et les facteurs chimiques, physiques et météorologiques. L'analyse de ces données à l'aide d'outils statistiques classiques (e.g. régression) s'est avérée difficile (e.g. instabilité temporelle des résultats). Une caractéristique commune de ces méthodes est de supposer les relations étudiées invariables dans le temps.

Les modèles dynamiques bayésiens (MDB) sont des modèles de séries temporelles qui intègrent l'hypothèse de relations variables dans le temps. Le modèle de régression linéaire dynamique a été appliquée aux « données d'Antifer ». L'hypothèse de variation temporelle des relations s'est avérée pertinente en écologie phytoplanctonique, à travers les variations de valeurs des coefficients de la régression. De plus, ces variations ont pu être interprétée et ainsi aboutir à un schéma partiel d'explication des concentrations de *Dinophysis* à Antifer. Le pourcentage de variabilité expliqué par une régression dynamique est supérieur à celui de son équivalent statique. Enfin, certaines variables non-significatives en régression statique sont apparues significatives en régression dynamique. Au total, les MDB se sont avérés être une approche originale et féconde du traitement des séries temporelles écologiques. Leurs extensions laissent entrevoir la possibilité de nombreuses collaborations entre biostatistique et écologie.

Mots clés : modélisation, statistique, dynamique, séries temporelles, bayésien, *Dinophysis*, baie de Seine.

Abstract

Dinophysis is a toxic microalga producing diarrheic shellfish poisoning. Despite many studies, its ecology and biology remain largely unknown. With the aim to determine occurrence conditions of *Dinophysis* cf. *acuminata* in the Antifer harbour, sea water was sampled daily for 3 to 4 months during the summers of 1987 to 1993. Available time series are those of the concentration of the toxic microalga and of chemical (e.g. nutrients), physical (e.g. temperature, salinity) and meteorological (e.g. wind speed and direction) factors. Using classical statistical tools (e.g. regression) to analyse these data rose some difficulties (e.g. instability of results). These methods have in common to suppose constant relationships.

Dynamic Bayesian models (DBM) are time series models assuming time varying relationship. Dynamic linear regression model has been applied to the Antifer time series. The relevance of time varying parameters values in phytoplankton ecology has been shown. Interpretation of parameter evolution has been found and led us to explain, at least in part, the evolution of *Dinophysis* concentration in Antifer harbour. Percentage variation explained by dynamic regression was greater than that of static regression. Finally, some non-significant variables with static regression appeared to be significant with dynamic regression. The use of DBMs revealed an original and prolific method for ecological time series. Their extensions may provide extensive collaboration between biologist and biostatististician.

Keywords: modelling, statistic, dynamic, time series, Bayesian, *Dinophysis*, Seine bay.

Sommaire

Chapitre 1

Introduction

En 1983, plusieurs milliers de cas de diarrhée suite à la consommation de moules ont été signalés en Bretagne sud (Lassus et al., 1983 ; Paulmier et Joly, 1983 ; Pierre et Lassus, 1983). De manière concomitante, le dinoflagellé *Dinophysis* était présent dans l'eau de mer et les hépatopancréas des bivalves. Cette microalgue était déjà associée à des épisodes diarrheïques dans le monde (Chili, Japon) et en particulier en Europe (Angleterre, Espagne, Irlande, Norvège, Pays-Bas, Pays de Galles) (anonyme, 1983 ; Marcaillou-Le Baut et Lassus, 1983). En France, le groupe « *Dinophysis acuminata* » s'est toujours révélé dominant lors des efflorescences côtières (Sournia et al., 1991). Un « test-souris » mis au point au Japon pour évaluer la toxicité des moules a été adapté aux besoins du réseau de surveillance phytoplanctonique (rephy) (Marcaillou-Le Baut et al., 1985) de l'Institut Français de Recherche pour l'Exploitation de la MER (ifremer). Les laboratoires côtiers effectuent toute l'année des prélèvements d'eau sur les zones conchylicoles pour la surveillance de routine. En régime alerte, généralement de mai à septembre, les tests-souris sont exécutés sur des échantillons de coquillages lorsque la concentration cellulaire des *Dinophysis* dans l'eau est jugée alarmante pour la toxicité des bivalves, *i.e.* entre 200 et 2000 cellules par litre selon les zones. Un test positif conduit à l'arrêt de la commercialisation des coquillages dans la zone de prélèvement. L'interdiction est levée après deux tests négatifs. La toxine diarrhéique (*Diarrhetic Shellfish Poisoning*, dsp) a été identifiée en France comme étant essentiellement l'acide okadaïque (Kumagai et al., 1986). Les analyses chimiques ont permis de montrer la présence de ce composé dans les coquillages des secteurs frappés d'interdiction. À l'étranger, des réseaux du même type ont également été mis en place à la suite de phénomènes de toxicité (e.g. Boni et al., 1992). Les microalgues productrices de dsp appartiennent principalement à l'espèce *Dino-*

*physis** (Marcaillou-Le Baut, 1990 ; Séchet et al., 1990 ; Honsell et al., 1992).

L'importance des pertes économiques lors des premiers événements toxiques et l'impératif de protection de la santé publique ont motivé depuis plusieurs années de nombreuses études sur *Dinophysis*. La biologie de la microalgue reste cependant mal connue en grande partie à cause de l'impossibilité actuelle de cultures stables malgré de multiples tentatives. *In situ* on observe des occurrences du dinoflagellé au sein de masses d'eaux caractérisées par des conditions de température, salinité et profondeur particulières (Delmas et al., 1992). Diverses approches mathématiques ont été tentées pour traiter les données environnementales à des fins d'explication de la dynamique du *Dinophysis*. Des modélisations dynamiques ont été réalisées (Boutibonnes, 1987 ; De Cremoux, 1988 ; Ménesguen et al., 1990 ; Chapelle, 1991). Ces modèles ont permis de tester des hypothèses et d'en émettre d'autres. Toutefois, le choix de certains paramètres sont discutables et les augmentations brutales de concentration de la microalgue sont mal restituées. Des dynamiques de nature chaotique ont été recherchées (Cazelles et al., 1991 ; Beltrami et Cosper, 1993). Bien que ces modèles simulent des augmentations brutales de concentrations, leur intérêt explicatif est limité par leur simplicité. De plus, ils nécessitent un grand nombre de données qui n'ont pas été disponibles. Les méthodes statistiques (descriptive, inférentielle, régression, analyse de données, analyse de variance) ont en général permis de confirmer les caractéristiques principales d'apparition de *Dinophysis* (Bensaker, 1986 ; Mer, 1986 ; Faure et Zhang, 1990 ; Le Batteux et Nizard, 1990 ; Soudant, 1990 ; Cazelles et al., 1991 ; Bouksim et al., 1992). Ces études concernant différents lieux et différentes années offrent rarement des résultats concordants, en dehors de l'importance de la température et de la salinité. Les méthodes statistiques sus-citées ont en commun de supposer structurellement des relations stables entre la variable étudiée et les variables explicatives.

West et Harrison (1997) utilisent l'expression « *Dynamic models* » pour désigner la classe des modèles dynamiques bayésiens (mdb). En de nombreux domaines et en particulier en écologie, les modèles dynamiques désignent des modèles déterministes de dynamique (e.g. dynamique de l'effectif d'une population). Ici, le terme « dynamique » se rapporte aux modifications temporelles des relations entre une variable à expliquer et des variables explicatives. Si de tels changements sont suspectés, alors le modèle statique $\mathcal{M}$ (e.g. une régression) établi à partir des informations disponibles avant le temps t peut s'avérer plus ou moins rapidement inadapté aux données obtenues après le temps t. Les mdb sont des modèles intégrant l'hypothèse de relations variables dans le temps et permettent ainsi la prise en compte de ce type de changement structurel. Le qualificatif « bayésien » permet de lever l'ambiguïté du caractère dynamique tout

∗. *Prorocentrum lima* est une espèce benthique également productrice de dsp (Marr et al., 1992)

en précisant la nature probabiliste et statistique de ces modèles. Par la suite, les expressions « modèles dynamiques bayésiens » et « modèles dynamiques » sont utilisées comme synonymes. Les caractéristiques des MDB sont les suivantes (HARRISON et STEVENS, 1976 ; WEST et HARRISON, 1997) :

- une forme paramétrique de type espace d'état, par opposition à une forme fonctionnelle (e.g. $ax^2 + bx + c = 0$) ;
- une représentation probabiliste des informations concernant les paramètres ;
- des prédictions dérivées de distributions de probabilités ;
- une définition séquentielle.

Le théorème de BAYES (BAYES, 1763 ; 1958) intervient dans la définition séquentielle des modèles(cf. annexe A). Un premier travail d'application des modèles dynamiques bayésiens en écologie phytoplanctonique a déjà été réalisé (SOUDANT, 1993). L'objectif était la prédiction des concentrations en *Dinophysis* cf. *acuminata* sur le site d'Antifer, avec pour motivation ultime l'obtention d'un système opérationnel d'aide à la décision intégrable dans le réseau de surveillance phytoplanctonique mis en place par l'IFREMER. Bien que l'établissement de ce modèle n'ai pas abouti, les résultats se sont avérés suffisamment encourageants pour susciter un travail de plus grande ampleur.

L'objectif du présent travail est d'évaluer l'intérêt de l'utilisation des modèles dynamiques bayésiens en écologie phytoplanctonique par leur application à des données ayant trait à l'espèce *Dinophysis acuminata*. L'état des connaissances concernant la biologie et l'écologie de la microalgue est présenté au chapitre 1. Le chapitre 2 introduit les modèles dynamiques bayésiens. Une part importante est consacrée à la présentation de modèles simples permettant la compréhension des principes de la méthode. Cette partie est illustrée par des applications à des données simulées. Les modèles ayant été appliqués à des séries temporelles de concentrations en *Dinophysis* sont présentées. Les résultats obtenus ont été concrétisés par deux articles inclus dans le chapitre 4. Une application non-publiée clôt cette section. L'intérêt de l'approche « modèles dynamiques bayésiens » et ses perspectives sont discutés dans les chapitres 5 et 6.

Chapitre 2

Bilan des connaissances sur le genre *Dinophysis*

Les épisodes toxiques associés à la présence des dinoflagellés ont entraîné de nombreuses études sur cette classe taxonomique. En France, le groupe *acuminata* est le plus fréquemment présent et le plus étudié (Sournia et al., 1991). Bien que des progrès considérables aient été réalisés, les informations concernant ces espèces restent parcellaires. Ainsi, en l'absence d'information sur un sujet donné pour une espèce donnée, il est fait référence à une autre espèce taxonomiquement proche. L'objet de ce chapitre est de donner un résumé des principales connaissances concernant le genre *Dinophysis*. Des informations plus complètes et plus détaillées pourront être trouvées dans des ouvrages spécialisés (e.g. Taylor, 1987 ; Sournia et al., 1991 ; Berland et Lassus, 1997).

2.1 Morphologie et identification

Le genre *Dinophysis* (environ 200 espèces) appartient à la classe des Dinophycées, ordre des Dinophysiales et famille des Dinophysiaceae. Les cellules mesurent de 30 à 100 μm. Visuellement, on notera simplement l'importance des deux plaques latérales alvéolées, constituant l'essentiel de l'enveloppe cellulosique externe, et la présence de deux flagelles, l'un supérieur, l'autre latéral (Fig. 2.1). La structure cellulaire du groupe présente une très grande diversité. Certaines espèces possèdent des particularités qui permettent de les identifier aisément. En revanche, un nombre important d'entre elles sont proches par la taille et la forme de *Dinophysis acuminata*. Les caractéristiques morphologiques des espèces de ce groupe présentent une variabilité importante et, de plus, sont influencées par

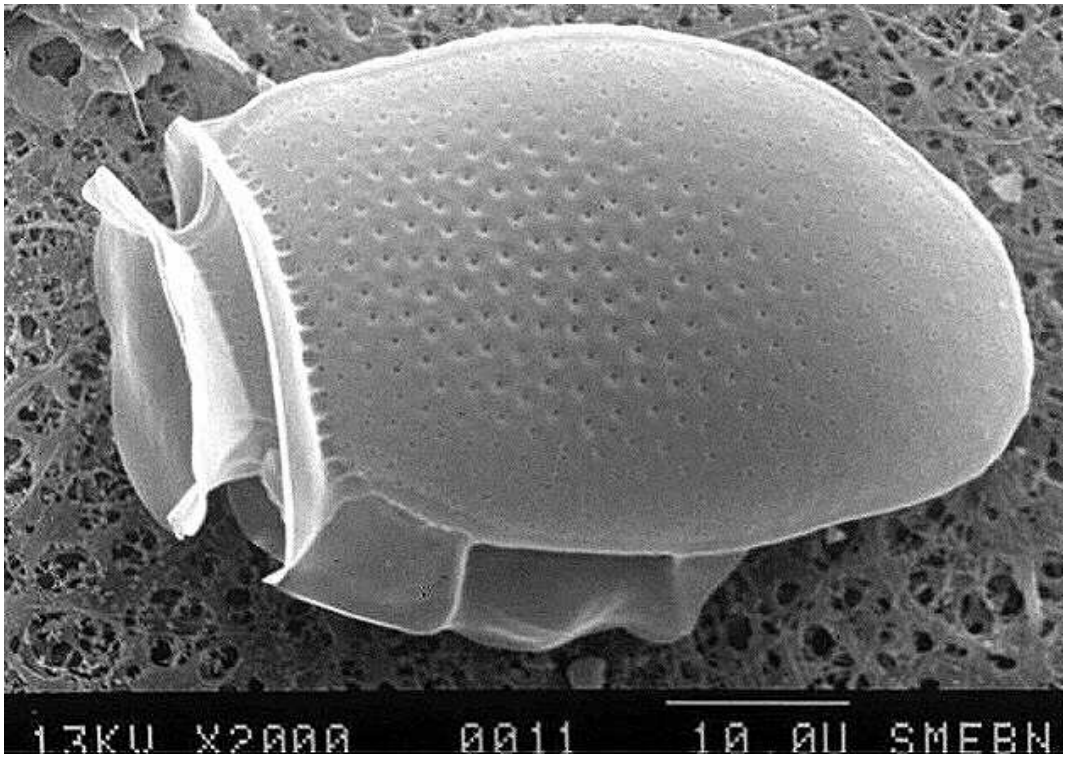

FIG. 2.1 - Dinophysis *cf.* acuminata *: une cellule en vue latérale gauche (microscope à balayage électronique). Dans la bande noire au bas de l'image, la ligne blanche horizontale représente 10 µm. (photographie : P. LASSUS (IFREMER) et A. BARREAU (université de NANTES)).*

les conditions de température et de salinité. Ainsi, l'identification de *Dinophysis acuminata* est difficile et les confusions nombreuses. De ce fait, dans le cadre du REPHY, les dénombrements phytoplanctoniques concernent le genre *Dinophysis* dans sa totalité. Des méthodes numériques et biologiques de reconnaissance automatique ont fait l'objet de recherches. Un système d'analyse d'image couplé à l'utilisation de colorants dans la préparation à analyser permet la détection des cellules de *Dinophysis*. Cependant, la durée de ce procédé est trop importante pour que son utilisation dans le contexte de réseaux de surveillance soit envisagée (MAESTRINI et al., 1997). En revanche, l'analyse d'image a permis de distinguer les différences morphologiques de certaines espèces (TRUQUET et al., 1996). Enfin, les résultats obtenus par système expert basé sur un réseau de neurones ont été jugés très encourageants, bien que les performances soient inférieures à celles obtenues par identification visuelle (CULVERHOUSE et al., 1996). Le résultat le plus probant de l'approche biologique d'identification est l'élaboration d'une sonde nucléique permettant la détection de *Dinophysis acuminata* à partir d'un seuil de concentration de 30 cellules par litres. La spécificité de cette sonde vis à vis de la microalgue a permis sa mise sur le marché (PUEL et al., sous

presse).

2.2 Biologie

Les principales difficultés rencontrées à l'étude du genre *Dinophysis* proviennent de l'absence de cultures stables. Dans le meilleur des cas, des cultures ont pu être conservées quelques mois. Ces résultats sont dus à un mode de nutrition qui n'a pas encore été clairement identifié. Lorsqu'une espèce phytoplanctonique contient des chloroplastes elle est considérée comme autotrophe, c'est-à-dire qu'elle utilise les éléments nutritifs minéraux présents dans l'eau et effectue la photosynthèse. En l'absence de chloroplastes, elle est considérée comme hétérotrophe, c'est-à-dire tirant sa subsistance de la consommation de matières organiques, issues par exemple de la prédation vis à vis d'autres organismes marins. *Dinophysis acuminata* contient des chloroplastes (Hallegraeff et Lucas, 1988), mais certaines de leurs caractéristiques sont atypiques. Les tentatives de mise en culture ont suggéré un comportement autotrophe (Durand-Clement et al., 1988 ; Sampayo, 1993). Sa capacité photosynthétique a été démontrée par la suite (Berland et al., 1994). Toutefois, l'origine des chloroplastes contenus dans les cellules de *Dinophysis* est inconnue. Une hypothèse est que ces chloroplastes appartiendraient à d'autres organismes marins (cf. Maestrini et al., 1997). Un comportement prédateur (phagotrophie) a été observé chez certaines espèces considérées comme autotrophes (Ishimaru et al., 1988). L'hypothèse d'un mode de nutrition autotrophe et hétérotrophe (*i.e.* mixotrophe) a été suggérée (Granèli et al., 1993) et s'avére le plus vraisemblable (Jacobson et Andersen, 1994).

Dinophysis se multiplie par division cellulaire et son taux de reproduction est relativement faible (0,25 à 0,96 division par jour) (Maestrini et al., 1997). L'hypothèse d'une reproduction sexuée a été soulevée (MacKenzie, 1992). Une reproduction de ce type a été observée chez un genre phylogénétiquement proche : *Prorocentrum* (Bhaud et Soyer-Gobillard, 1988 ; Faust, 1993b). Mais Faust (1993a) observe également chez *Prorocentrum lima* un mode de reproduction non-sexué. En ce qui concerne *Dinophysis*, certaines observations sont en faveur d'une reproduction sexuée. Des cellules, auparavant identifiées comme des espèces distinctes, correspondraient à un stade sexuel (Maestrini et al., 1997). Ce mode de reproduction a été associé à la formation de kystes (McLachlan, 1993). Ces derniers sont une forme de résistance à des conditions environnementales défavorables. Ils ont été observée chez plusieurs espèces phytoplanctoniques, par exemple *Alexandrium* sp. (Turgeon et al., 1990). Des modifications de cellules semblables à des processus d'enkystement ont été observées chez certains *Dinophysis* (Moita et Sampayo, 1993). Les corpuscules identifiés comme des kystes présentaient en leur sein de nombreuses petites cellules flagellées (cf. Maestrini

et al., 1997). L'hypothèse selon laquelle ces cellules seraient des formes immatures des cellules de *Dinophysis* habituellement observées a été émise. Cependant, le processus de désenkystement n'a jamais été observé à partir de sédiments prélevés dans des zones d'apparitions de la microalgue (Larrazabal et al., 1990). Ces derniers éléments soulignent le peu de connaissances établies concernant le cycle biologique de *Dinophysis*.

2.3 Ecologie

Dinophysis est rencontré dans toutes les mers du globe et en particulier dans les zones côtières (Smayda, 1990). Seuls les phénomènes concernant le Japon, l'Europe du nord et la France seront évoqués ici. Les concentrations cellulaires dépassent rarement les cent mille cellules par litre. Le port d'Antifer (Normandie, France) est une exception où des concentrations supérieures ont souvent été observées pour *Dinophysis* cf. *acuminata*. Les espèces responsables d'intoxications n'ont jamais été dominantes à une exception prés (cf. Maestrini et al., 1997). Au Japon, *Dinophysis fortii* est principalement mis en cause. En fonction des espèces et des situations géographiques, deux scénarios d'apparition des microalgues toxiques sur le littoral coexistent : un développement côtier et un autre au large, les cellules étant progressivement ramenées à la côte. Un courant océanique est à l'origine de ce déplacement de masses d'eaux. L'importance de son effet est tel qu'il est à la base d'un modèle prédictif (cf. Sournia et al., 1991). Un second courant océanique intervient du printemps à l'automne en entraînant *Dinophysis* du sud vers le nord le long des côtes ouest des îles nippones. D'autres phénomènes physiques (*upwelling*, marée) ont une influence sur la répartition spatiale à une échelle plus fine. En Europe du nord, *Dinophysis acuminata* est souvent présent à la fin de l'été, les concentrations maximales étant à proximité des côtes. Des observations rapportent que les concentrations diminuent avec l'augmentation de la force du vent (Sournia et al., 1991). D'une manière générale, les turbulences peuvent perturber certains processus biologiques (Berdalet et Estrada, 1993) ou physiologiques (e.g. capacité de prélèvement des nutriments). En Suède et en Norvège un courant océanique est suspecté d'entraîner des accumulations d'algues dans les zones littorales. En France, un schéma d'apparition géographique et temporel se répète depuis plusieurs années (Lassus et al., 1988). Les premières alertes débutent à la fin du printemps en Bretagne sud. Durant l'été, ce sont les côtes normandes qui sont à leur tour concernées par des efflorescences de *Dinophysis* cf. *acuminata*. Jusqu'à présent, la microalgue ne s'est pas manifesté sur la zone littorale comprise entre le Finistère nord et la pointe du Cotentin. En revanche elle est signalée en Charente Maritime et sur les côtes méditerranéennes. Dans les pertuis charentais, l'hypothèse d'un développement au large puis d'une accumulation à la côte semble très vraisem-

blable (Delmas et al., 1993). Des phénomènes physiques intervenants à échelle locale sont supposés être responsables des variations de densité dans les baies et les estuaires (De Cremoux, 1988 ; Lassus et al., 1988, 1993). Bien que la répartition horizontale de la biomasse phytoplanctonique ait été beaucoup étudiée et reconnue comme hétérogène (Cassie, 1962 ; Platt, 1972, 1975 ; Harris et Smith, 1977 ; Ledbetter, 1979 ; Mackas et al., 1985 ; Aitsam, 1994), celle de *Dinophysis* n'a jamais fait l'objet de recherches.

La répartition verticale de *Dinophysis* est liée à la température et la salinité de l'eau de mer (Delmas et al., 1992). Plus précisément, les concentrations maximum de la microalgue se trouvent souvent dans la zone de décroissance brutale de la température entre les eaux de surface et les eaux plus profondes (thermocline). Elle correspond également à une zone de discontinuité de la salinité (halocline) et des nutriments (nutricline). Elle est généralement située entre trois et cinq mètres de profondeur durant la période estivale. L'épaisseur de la zone de concentration maximum de *Dinophysis* peut être de quelques dizaines de centimètres (Gentien et al., 1992). La profondeur et l'épaisseur de cette zone sont variables dans la dimension horizontale et dans le temps. Cette caractéristique complique considérablement l'échantillonnage de la microalgue. De plus, les flagelles dont est doté *Dinophysis* lui permettent de se déplacer dans son milieu. Cette capacité natatoire a beaucoup été observée et étudiée, en particulier par Kamykowski (e.g. Kamykowski et Zentara, 1977 ; Kamykowski, 1995). Les vitesses atteintes sont négligeables en comparaison des mouvements de masses d'eaux induits par les phénomènes physiques. Toutefois, il est généralement admis qu'elles permettent un déplacement vertical. Au Japon, ce déplacement n'a pas été observé pour *Dinophysis fortii*. En France, des migrations de *Dinophysis* sp. et *Dinophysis acuminata* ont été observées, respectivement, en baie de Vilaine (Durand-Clement et al., 1988) et dans le port d'Antifer (Lassus et al., 1990). Les déplacements durant 24 heures montrent une plongée la nuit et une remontée vers la surface le jour. Une plongée peut être observée en journée lors des heures de très forte intensité lumineuse. Écologiquement, la présence en surface pourrait correspondre à un « choix » d'intensité lumineuse « optimisant » l'activité photosynthétique et la plongée nocturne à une recherche de nutriments, dont les concentrations sont plus fortes dans les couches plus profondes. Toutefois, les procédures expérimentales mises en œuvre pour observer *in situ* les déplacements actifs de *Dinophysis* sont sujettes à caution. En l'absence de confirmation formelle, il est recommandé d'utiliser avec précaution cette possibilité de migration pour expliquer les phénomènes (Maestrini et al., 1997).

Les relations entre nutriments (sels nutritifs) et *Dinophysis* sont difficiles à cerner. Il est généralement admis que les dinoflagellés peuvent se développer dans des eaux pauvres en sels nutritifs. Les occurrences de *Dinophysis acuminata* ont souvent été observées après des efflorescences de diatomées* ayant épuisé le milieu

∗. Les deux grands groupes de phytoplancton sont les diatomées et les dinoflagellés.

en nutriments (Brockman et al., 1977 ; Lassus et al., 1983). Il reste que la zone d'accumulation des microalgues toxiques est proche des eaux profondes plus riches en sels nutritifs que les eaux de surface. Les corrélations entre concentrations en nutriments et concentration en *Dinophysis* ne sont pas stables géographiquement et temporellement (signe, significativité). Ces absences de corrélations sont peut être à mettre sur le compte de la faible proportion que représente la microalgue au sein des biomasses phytoplanctoniques dans lesquelles elle est rencontrée. En baie de Vilaine, des modèles numériques intégrant l'hypothèse de consommation des sels nutritifs ont permis d'obtenir des simulations vraisemblables des concentrations en *Dinophysis*, en dépit de certaines faiblesses (paramètres incertains, mauvaise adaptation aux pics de concentrations) (Boutibonnes, 1987 ; De Cremoux, 1988 ; Ménesguen et al., 1990 ; Chapelle, 1991). Ces modèles ont également permis de tester des hypothèses évoquées dans la littérature, comme la migration verticale. Le manque de données biologiques induit par l'absence de culture a limité l'exploitation de ces modèles. L'élément le plus stable des conditions de développement de la microalgue toxique reste la stratification thermohaline des eaux et sa stabilité dans le temps (Delmas et al., 1992). Au Japon, la température et la salinité sont à la base d'un modèle prédictif des concentrations en *Dinophysis fortii*. Toutefois, il faut signaler que les concentrations maximales de cette espèce sont observées dans des gammes de température et de salinité plus étroites en comparaison avec celles d'autres espèces de dinoflagellés toxiques.

Dans un objectif prédictif, les espèces phytoplanctoniques apparaissant avant le genre *Dinophysis* ont été recherchées. Une nouvelle fois, c'est au Japon que les résultats les plus significatifs ont été obtenus. Par exemple, les premières apparitions de *Dinophysis acuminata* et *Gonyaulax spinifera* apparaissent régulièrement un mois avant *Dinophysis fortii* en baie de Mutsu (Sournia et al., 1991). Ces données sont utilisées conjointement à d'autres observations hydrologiques et parallèlement aux modèles mathématiques. En France, comme il a déjà été signalé, les efflorescences de *Dinophysis* suivent généralement le déclin des diatomées (Pierre et Lassus, 1983). Les espèces accompagnatrices des occurrences des dinoflagellés toxiques ont également été recherchées afin de mieux cerner leurs besoins physiologiques. Le groupe *Prorocentrum* est souvent associé à *Dinophysis*. *Prorocentrum micans* domine souvent les assemblages phytoplanctoniques présentant *Dinophysis acuminata*. Cependant, ces relations n'ont pas été plus approfondies.

Au sein des communautés phytoplanctoniques, il y a également des organismes zooplanctoniques. Ces derniers consomment le phytoplancton. Si *Dinophysis* était également consommé, son faible taux de reproduction ne lui permettrait pas de se maintenir plusieurs semaines dans son milieu. Il est apparu que ces dinoflagellés sont peu sensibles à ce broutage. Deux mécanismes éventuellement coexistants sont envisagés. Le premier est la limitation ou l'arrêt de l'activité de nutrition des brouteurs par l'émission de toxines dans l'eau de mer : ce phénomène (al-

lélopathie) est connu en écologie entre plantes terrestres, mais également entre espèces phytoplanctoniques (Lewis, 1986). Les toxines diarrhéiques sont considérées comme potentiellement allélopathiques (Wright, 1994). Une partie importante de l'acide okadaïque produit par *Prorocentrum lima* a été retrouvée dans son milieu de culture (Rausch de Traubenberg et Morlaix, 1995). La mise en présence de copépodes (genre de zooplancton brouteur) et d'une communauté phytoplanctonique enrichie en *Dinophysis acuminata* a montré une faible consommation de la microalgue toxique puis, pour certains, un arrêt de sa consommation (Carlsson et al., 1996). Les individus n'ayant pas cessé de consommer le dinoflagellé sont morts. Cependant, la caractéristique de l'allélopathie est un effet à distance qui n'a pas été observé. Le second mécanisme de défense est appelé « la stratégie du termite » (Maestrini et al., 1997). Si la consommation d'une cellule *Dinophysis* entraîne le décès du brouteur ou l'arrêt de consommation de la microalgue, alors ce phénomène est bénéfique pour l'ensemble des individus de l'espèce. L'un et l'autre de ces mécanismes nécessitent de plus amples recherches.

À l'issue de ce rapide survol des connaissances concernant *Dinophysis*, il apparaît que beaucoup d'éléments sont du domaine de la spéculation et d'autres inconnus. De plus, les résultats des recherches soulèvent de nouvelles interrogations, comme la nature animale ou végétale de la microalgue (Maestrini et al., 1997). En ce qui concerne le sujet du présent travail, il faut retenir l'association des occurrences du dinoflagellé toxique avec des conditions de température et de salinité particulières, et l'absence apparente de corrélation avec les concentrations en nutriments.

Chapitre 3

Modèles dynamiques bayésiens

3.1 Introduction

Ce chapitre suit largement le plan de l'ouvrage de West et Harrison paru en 1989 et réédité en 1997 (West et Harrison, 1997). L'objectif recherché est de fournir le plus d'éléments possibles pour une compréhension et une application rapide des modèles dynamiques bayésiens. De ce fait, les démonstrations des résultats ne sont pas données et le lecteur intéressé pourra se référer à West et al. (1985) et West et Harrison (1997).

Historiquement, le modèle Harrison-Stevens (Harrison et Stevens, 1971) est considéré comme le premier mdb. C'est un modèle de prédiction à court terme dont la définition bayésienne est fondée sur le filtre de Kalman (Kalman, 1960 ; Meinhold et Singpurwala, 1983). En 1976, Harrison et Stevens définissent la classe des modèles dynamiques linéaires gaussiens et introduisent plusieurs de ses modèles. Cette classe de modèles est généralisée aux cas non-linéaires et non-gaussiens en 1985 (West ; West et al.). Finalement, West et Harrison publient conjointement « *Bayesian forecasting and dynamic models* » en 1989. Cet ouvrage essentiel décrit la théorie des modèles dynamiques. Pole et al. (1994) présentent cette théorie de façon plus accessible et l'illustrent par des applications très détaillées et réalisées avec le logiciel bats, distribué avec l'ouvrage. Les mdb restent en constante évolution (e.g. Bolstad, 1995 ; West, 1995). La réédition de l'ouvrage de West et Harrison (West et Harrison, 1997) présente les développements les plus marquants depuis 1989. Les auteurs fournissent une importante bibliographie principalement méthodologique. Ces références sont souvent des thèses, rapports de recherche d'universités et actes de colloques, parfois difficiles à obtenir. Les mdb ont surtout été appliqués en économie (Johnston et Harrison, 1980 ; Ameen et Harrison, 1985 ; Migon et Harrison, 1985 ; West et al., 1985 ; Pole et al., 1994 ; West et Harrison,

1997), mais également en démographie (POLE et al., 1994), en environnement (YOUNG et YOUNG, 1992), en climatologie (WEST, 1995), en médecine (SMITH et WEST, 1983 ; WEST, 1995) et en biologie animale (BOLSTAD, 1988b).

Les modèles présentés ici sont univariés et normaux (gaussiens). Le modèle de tendance polynomiale (*Polynomial Trend Model*, PTM) d'ordre un est le plus simple des modèles dynamiques. L'étude de sa version à variances constantes permet d'aborder les principales caractéristiques des MDB. Le modèle d'ordre deux permet l'introduction des notations matricielles et ainsi la définition du modèle d'ordre n. La généralisation de ce modèle conduit au modèle linéaire dynamique général. Le modèle de régression linéaire dynamique et les modèles multiprocess ont fait l'objet d'applications présentées au chapitre 4. Enfin, les MDB sont replacés dans le contexte des modèles classiques de série temporelle et du filtre de KALMAN.

3.2 Modèle de tendance polynomiale d'ordre un

Lorsqu'une mesure est effectuée, elle est généralement entachée d'une erreur de mesure. Cette erreur de mesure laisse supposer l'existence d'une vraie valeur inobservable. Un des objets de la statistique est l'estimation de cette valeur. Une série temporelle est une série de mesures réalisées à différents moments dans le temps. Ces mesures ne sont pas moins exemptes d'erreurs et laissent ainsi supposer l'existence d'une série temporelle inobservable des vraies valeurs. Cette série est appelée le niveau moyen de la série temporelle. Un modèle de tendance polynomial d'ordre un permet l'estimation de cette série temporelle inobservable.

3.2.1 Définition

Soit pour tout t, $t = 1, 2, \ldots$, Y_t la série temporelle des observations, ν_t la série temporelle des erreurs de mesure et μ_t la série temporelle du niveau moyen. La représentation mathématique de la décomposition de la série temporelle des observations en la série temporelle du niveau moyen et la série temporelle des erreurs s'écrit pour tout t, $t = 1, 2, \ldots$,

$$Y_t = \mu_t + \nu_t. \tag{3.1}$$

C'est l'**équation d'observation**. Par hypothèse, les erreurs ν_t sont indépendantes les unes des autres et distribuées selon une loi normale de moyenne 0 et de variance V_t, notée $N[0\,; V_t]$. Dans un PTM d'ordre un, l'évolution de μ_t est contrôlée par l'**équation d'évolution** pour tout t, $t = 1, 2, \ldots$,

$$\mu_t = \mu_{t-1} + \omega_t, \tag{3.2}$$

où les ω_t, appelées erreurs d'évolutions, sont indépendantes les unes des autres et distribuées selon une loi normale $N[0\,;W_t]$. L'équation 3.2 définissant récursivement les niveaux moyens μ_t, $t = 1, 2, \ldots$, il est nécessaire de disposer d'informations concernant μ_0. Ces informations sont incluses dans l'ensemble des informations disponibles avant la première observation, ensemble noté D_0. Les informations concernant μ_0 prennent la forme de la distribution de la variable « μ_0 conditionnellement à D_0 », notée $(\mu_0|D_0)$. Cette distribution est $N[m_0\,;C_0]$. En particulier, D_0 contient m_0, C_0 et toutes les valeurs des variances V_t et W_t. Enfin, les séquences d'erreurs ν_t et ω_t sont indépendantes l'une de l'autre et indépendantes de $(\mu_0|D_0)$. La définition formelle de ce modèle est la suivante :

Définition 1 *Pour tout t, le modèle de tendance polynomiale linéaire dynamique d'ordre un est défini par :*

$$\begin{array}{lll} \textit{Équation d'observation :} & Y_t = \mu_t + \nu_t, & \nu_t \sim N[0\,;V_t], \\ \textit{Équation d'évolution :} & \mu_t = \mu_{t-1} + \omega_t, & \omega_t \sim N[0\,;W_t], \\ \textit{Information initiale :} & (\mu_0|D_0) \sim N[m_0\,;C_0], & \end{array}$$

où m_0 et C_0 sont fixés, et les séquences d'erreurs ν_t et ω_t sont indépendantes, mutuellement indépendantes et indépendantes de $(\mu_0|D_0)$.

(Ici et par la suite la notation $\sim$ signifie « distribuée selon ».).

En pratique, l'inclusion des séquences des variances W_t et V_t et des constantes m_0 et C_0 dans l'ensemble D_0, signifie que c'est l'utilisateur du modèle qui doit les spécifier. Les valeurs m_0 et C_0 reflètent sa connaissance *a priori* de la série temporelle. Par exemple, supposons l'étude de la série temporelle des moyennes journalières de température d'un être humain. Les valeurs inférieures à 32 °C et supérieures à 42 °C sont exceptionnelles. Considérons que ces deux valeurs sont les limites de l'intervalle de confiance à 99 % des températures distribuées selon une loi normale. Alors, la moyenne et la variance au temps $t = 0$ sont,

$$\begin{aligned} m_0 &= (42 + 32)/2 = 37, \\ C_0 &= ((42 - 37)/\epsilon_{0,995})^2 = ((32 - 37)/\epsilon_{0,995})^2 = 3{,}79, \text{ à } 10^{-2} \text{ près,} \end{aligned}$$

où $\epsilon_{0,995} = 2{,}57$ est la valeur d'une loi normale pour la probabilité 0,995. La spécification des séquences des valeurs des variances V_t et W_t peut également s'appuyer sur la connaissance *a priori* de la série temporelle. Par exemple, l'amplitude des erreurs de mesures ou bien une relation de type moyenne-variance peuvent être connues.

La définition 1 implique l'existence des variables aléatoires $(Y_t|\mu_t, D_{t-1})$ et $(\mu_t|\mu_{t-1}, D_{t-1})$, où $D_{t-1} = \{D_{t-2}, Y_{t-1}\}$ est l'ensemble des informations disponibles au temps $t - 1$. En considérant la conditionnalité à μ_t et μ_{t-1} comme implicite, ces variables s'écrivent respectivement $(Y_t|D_{t-1})$ et $(\mu_t|D_{t-1})$. Littéralement, leurs distributions sont celles de variables aléatoires au temps t conditionnellement à l'ensemble des informations disponibles au temps $t - 1$. Ce sont

des distributions *a priori* ou encore des distributions de prédictions. Les variables aléatoires $(Y_t|D_{t-1})$ et $(\mu_t|D_{t-1})$ sont à la base de la procédure séquentielle d'estimation.

3.2.2 Procédure séquentielle d'estimation

Soit au temps $t-1$ la variable aléatoire $(\mu_{t-1}|D_{t-1})$ distribuée selon la loi normale $N[m_{t-1}\,;C_{t-1}]$, dont les paramètres sont supposés connus. Les étapes de l'estimation sont les suivantes :

A priori La distribution de la variable *a priori* $(\mu_t|D_{t-1})$ est $N[a_t\,;R_t]$, où $a_t = m_{t-1}$ et $R_t = C_{t-1} + W_t$. L'addition de la variance W_t reflète un accroissement de l'incertitude concernant le niveau de la série temporelle.

Prédiction $(Y_t|D_{t-1})$ est la variable aléatoire de prédiction distribuée selon $N[f_t\,;Q_t]$, où $f_t = a_t$ et $Q_t = R_t + V_t$. L'augmentation de variance représente la prise en compte de l'erreur de mesure.

A posteriori L'observation de la vraie valeur de Y_t permet la mise à jour de l'*a priori* $(\mu_t|D_{t-1})$ en la variable aléatoire *a posteriori* $(\mu_t|D_t)$, distribuée selon $N[m_t\,;C_t]$, où $m_t = a_t + A_t e_t$ et $C_t = R_t - A_t^2 Q_t$, avec $A_t = R_t/Q_t$ et $e_t = Y_t - f_t$.

Le calcul de la distribution de l'estimation de Y_t ne fait pas partie de la procédure, car ses paramètres ne sont pas nécessaires aux autres étapes. Cette distribution est $(Y_t|D_t) \sim N[g_t\,;P_t]$, où $g_t = m_t$ et $P_t = C_t + V_t$.

L'étape « *a posteriori* » est la plus importante de la procédure (cf. annexe A). L'information contenue dans l'observation Y_t est prise en compte en modifiant la moyenne et la variance de la distribution *a priori* de μ_t. L'élément central de cette prise en compte est la quantité A_t, appelée coefficient adaptatif. A_t est la proportion de variabilité induite par le niveau moyen (*i.e.* R_t) dans la variabilité de la prédiction (*i.e.* Q_t). Ainsi, la moyenne *a posteriori* m_t est égale à la moyenne *a priori* a_t corrigée de la part de l'erreur de prédiction $A_t e_t$ imputable à un changement dans le niveau moyen. L'interprétation du calcul de la variance *a posteriori* C_t suit la même logique : elle est égale à la variance *a priori* moins la part de la variance due au niveau moyen dans la variance de prédiction. La variance étant un opérateur quadratique le coefficient adaptatif est élevé au carré. Les informations concernant les paramètres sont représentées par des distributions de probabilités. La moyenne est l'estimation de la variable considérée. Un intervalle de confiance peut être associé à cette estimation. Par exemple, les bornes de l'intervalle de confiance de la moyenne de $(\mu_t|D_t)$ sont $m_t \pm \epsilon_{1-\alpha/2}\sqrt{C_t}$, au risque de première espèce α fixé, et où $\epsilon_{1-\alpha/2}$ représente la valeur d'une loi normale $N[0\,;1]$ pour la probabilité $1-\alpha/2$. Les séries temporelles des bornes

supérieures et inférieures des intervalles de confiance forment une enveloppe de confiance.

Pour tout t, $t = 1, 2, \ldots$, la moyenne f_t de la distribution de prédiction est égale à la moyenne *a priori* a_t, elle même égale à m_{t-1}. L'égalité de ces moyennes signifie que la prédiction pour le temps t est le niveau estimé pour le temps $t-1$. En revanche, pour tout t, $t = 1, 2, \ldots$, les variances *a posteriori*, *a priori* et de prédiction sont ordonnées tel que $C_{t-1} < R_t < Q_t$. Ainsi, le modèle prédit pour le temps t le niveau estimé pour le temps $t-1$ et augmente la variance, c'est-à-dire l'incertitude. Les prédictions pour un horizon de k unité(s) de temps, $k \geq 1$, sont données par la distribution de $(Y_{t+k}|D_t) \sim N[f_t(k)\,;Q_t(k)]$ avec $f_t(k) = m_t$ et

$$Q_t(k) = C_t + \sum_{i=1}^{k} W_{t+i} + V_{t+k}.$$

Ainsi les prédictions de $t+1$ à $t+k$ forment une droite de pente nulle et de valeur m_t dont la variance augmente linéairement avec k. De ce fait, capacité prédictive de ce modèle est limitée.

3.3 Modèle de tendance polynomiale d'ordre un à variances constantes

La définition du modèle à variances constantes est celle d'un PTM d'ordre un avec pour tout t, $t = 1, 2, \ldots$, la variance d'observation $V_t = V$ et la variance d'évolution $W_t = W$. L'équivalent statique de ce modèle est $Y_t = \mu + \nu_t$, ce qui implique $(Y_t|\mu, V) \sim N[\mu\,;V]$. La description de la série est limitée à la droite $Y_t = \mu$, qui est la moyenne unique et constante des observations. Le PTM d'ordre un à variances constantes est un modèle où la moyenne est supposée **localement** constante. Sa simplicité permet de comprendre en profondeur certaines caractéristiques communes à tout les MDB.

3.3.1 Illustration

La série temporelle étudiée est une série simulée à partir des équations 3.1 et 3.2 (page 14) avec $\mu_0 = 10$. Les réalisations des lois normales centrées des erreurs d'observation et d'évolution sont générées selon la formule $[(\sum_{i=1}^{12} X_i) - 6] \times \sigma$, où X_i est la réalisation d'une loi uniforme sur l'intervalle $[0\,;1]$ et σ est la variance de la loi normale considérée. Les variances d'observation et d'évolution sont respectivement $V = 2$ et $W = 1$. L'intérêt de la simulation est de disposer de la série temporelle inobservable des niveaux moyens. Le modèle appliqué à ces données simulées a pour variance d'observation et d'évolution celles utilisées pour la simulation. Au temps 0, la moyenne du niveau moyen m_0 est fixée à 0 et sa variance C_0 à 20. Les vraies conditions initiales étant $m_0 = 10$ et $C_0 = 1$,

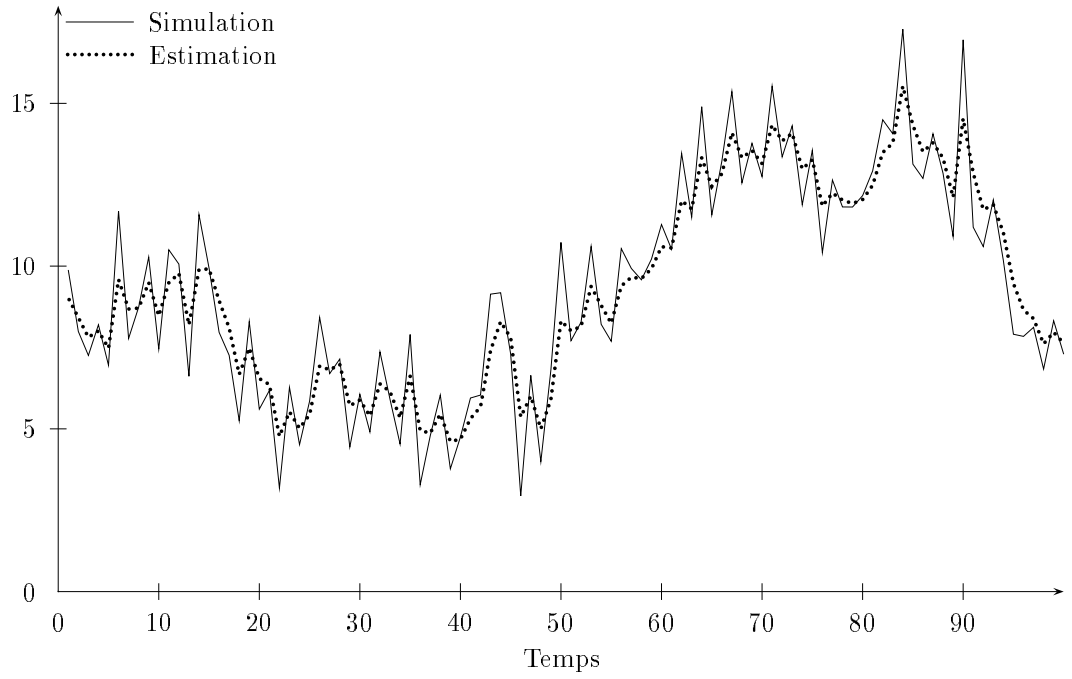

FIG. 3.1 - *Simulation et estimation des observations d'un modèle de tendance polynomiale d'ordre un à variances constantes.*

les spécifications du modèle utilisé reflètent une connaissance plutôt imprécise d'un utilisateur fictif. Malgré ces spécifications vagues, les estimations des observations et des niveaux moyens sont proches des valeurs simulées (FIG. 3.1 et 3.2). L'enveloppe de confiance à 90 % des estimations des μ_t inclut 93 % des valeurs simulées. Comme $\mu_0 = 10$, les premières valeurs simulées sont proches de 10 et la spécification $m_0 = 0$ entraîne une erreur de prédiction importante au temps $t = 1$ (cf. TAB. 3.1). L'impact de la spécification de m_0 est réduit car la proportion $A_1 = 0{,}91$ de l'erreur e_1 est prise en compte lors du calcul de la moyenne *a posteriori* m_1. Le coefficient adaptatif ne dépend ni des observations ni de m_0. Des spécifications différentes de m_0, toutes choses égales d'ailleurs, auraient entraîné des erreurs e_1 qui auraient toutes été prises en compte à hauteur de 91 % dans l'étape *a posteriori*. La variance initiale C_0 étant également corrigée de façon importante lors de cette étape, l'impact de sa spécification imprécise est limité.

3.3.2 Convergences et limites

Dans le tableau 3.1, les variances R_t, Q_t, C_t, P_t et le coefficient adaptatif A_t ne subissent quasiment plus de modifications au-delà de la cinquième unité de temps. WEST et HARRISON (1997, pages 44-45) ont montré les valeurs limites

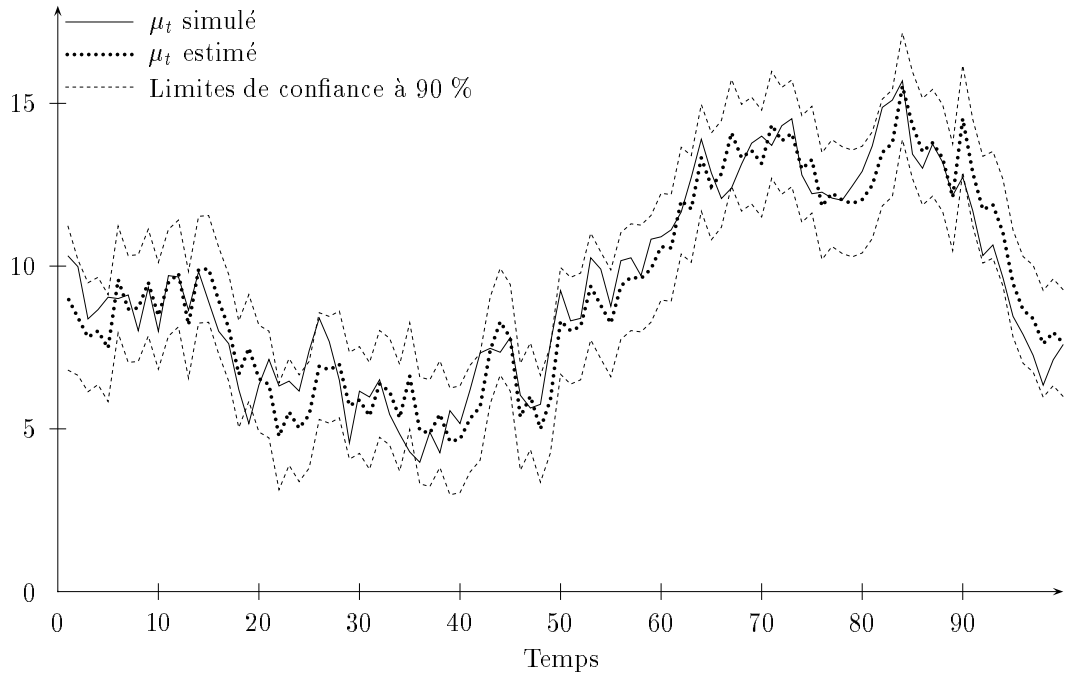

FIG. 3.2 - *Simulation et estimation du niveau moyen d'un modèle de tendance polynomiale d'ordre un à variances constantes.*

suivantes :

- $A_t \to A = r(\sqrt{1 + 4/r} - 1)/2$ avec $r = W/V$;
- $C_t \to C = AV$;
- $R_t \to R = C/(1 - A)$;
- $Q_t \to Q = V/(1 - A)$.

Les valeurs limites de C_t, R_t et Q_t dépendent de la valeur limite de A_t. La valeur limite de A_t est une fonction croissante du ratio r, qui est le ratio signal sur bruit du modèle. Quelques valeurs de cette fonction sont données dans le tableau 3.2.

Lorsque la convergence est atteinte le calcul de la moyenne *a posteriori* de μ_t se limite à l'addition de la proportion A de l'erreur e_t à la moyenne *a priori* de μ_t (cf. étape « *a posteriori* », page 16). Si la variance d'évolution (W) est supérieure à la variance d'observation (V) alors l'utilisateur du modèle suppose que la plus grande part de la variabilité est due à l'évolution du niveau moyen. Le coefficient A est compris entre $(\sqrt{5}-1)/2$ et 1 et ainsi une proportion importante des erreurs de prédiction est prise en compte lors de la mise à jour de $(\mu_t|D_{t-1})$. Inversement, si W est inférieure à V la variabilité est supposée être due aux erreurs de mesure, A est compris entre 0 et $(\sqrt{5} - 1)/2$ et une faible part des

erreurs est prise en compte lors de l'étape « *a posteriori* ». Supposons ρ la vraie valeur du ratio signal sur bruit d'une série temporelle et α celle du coefficient adaptatif. Si r est supérieur à ρ, alors A est supérieur à α et la prise en compte de l'erreur est trop importante. Le modèle est sur-adaptatif. Inversement, le modèle est sous-adaptatif si r est inférieur à ρ. Enfin, le rapport r peut être proche de ρ sans que les variances spécifiées soient voisines des vraies variances. Dans ce cas les estimations des moyennes des distributions seront vraisemblables, mais les estimations des variances seront erronées, et avec elles les intervalles de confiance des moyennes. La spécification des variances W et V est donc cruciale.

3.3.3 Spécification des variances W et V

Dans les résultats précédents, R_t tend vers $C/(1-A)$. Comme $R_t = C_{t-1}+W$, une autre forme de la valeur limite de R_t est $C + W$. L'égalité de ces deux valeurs limites amène $W = CA/(1-A)$. De cette façon, W peut être vue comme une quantité proportionnelle à C. Etant donné la rapidité de la convergence du modèle, West et Harrison (1997, page 51) propose d'adopter ce point de vue pour spécifier la séquence des valeurs de la variance d'évolution. Soit le facteur d'escompte $\delta = 1-A$, $\delta \in]0\,;1[$. Les variances d'évolution sont spécifiées pour tout t, $t = 1, 2, \ldots$, telles que $W_t = C_{t-1}(1-\delta)/\delta$*. Cette spécification implique des valeurs différentes des W_t : la variance d'évolution n'est plus constante. Toutefois, le phénomène de convergence est toujours présent et les valeurs limites sont les mêmes. De plus, les valeurs de W_t convergent vers $W = C(1-\delta)/\delta$. Comme C_t converge vers $V(1-\delta)$ et que le ratio signal sur bruit r est égal à $(1-\delta)^2/\delta$, une autre forme de W est rV. Ainsi l'hypothèse de constance de la variance d'évolution est à l'origine du phénomène de convergence.

Une analyse bayésienne classique (West et Harrison, 1997, chapitre 17) appliquée aux modèles dynamiques permet de définir une procédure séquentielle d'estimation de la variance V constante. L'utilisation de cette analyse suppose en particulier que la structure imposée à la séquence des W_t soit du type $W_t = VW_t^\diamond$, où $W_t^\diamond$ est la variance de ω_t lorsque $V = 1$. Si le ratio signal sur bruit est supposé constant, alors $W_t^\diamond = r$ et ce cas revient à l'utilisation d'un facteur d'escompte. Un modèle utilisant cette procédure est un modèle à variance d'observation V constante et inconnue.

3.3.4 Modèle à variance V constante et inconnue

Soit $\phi = 1/V$ la précision inconnue et constante des mesures. La variable aléatoire $(\phi|D_t)$ est distribuée selon une loi Gamma $G[n_t/2\,;d_t/2]$, où n_t est un

*. En pratique la spécification $\delta = 1$ est possible. Elle implique des variance d'évolution nulles, parfois utilisées.

TAB.3.1: *Éléments de calcul de l'application d'un modèle de tendance polynomiale d'ordre un à variances constantes.*

Temps	Distribution *a priori*		Distribution de prédiction		Observation	Erreur de prédiction	Coefficient adaptatif	Distribution *a posteriori*		Distribution d'estimation	
t	a_t	R_t	f_t	Q_t	Y_t	e_t	A_t	m_t	C_t	g_t	P_t
0								0	20	–	–
1	0	21	0	23	9,88	9,88	0,91	9,02	1,83	9,02	3,83
2	9,02	2,83	9,02	4,83	7,99	-1,03	0,59	8,42	1,17	8,42	3,17
3	8,42	2,17	8,42	4,17	7,26	-1,16	0,52	7,81	1,04	7,81	3,04
4	7,81	2,04	7,81	4,04	8,20	0,38	0,51	8,01	1,01	8,01	3,01
5	8,01	2,01	8,01	4,01	6,95	-1,05	0,50	7,48	1,00	7,48	3,00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
95	11,03	2,00	11,03	4,00	7,91	-3,12	0,50	9,47	1,00	9,47	3,00
96	9,47	2,00	9,47	4,00	7,84	-1,63	0,50	8,65	1,00	8,65	3,00
97	8,65	2,00	8,65	4,00	8,12	-0,53	0,50	8,39	1,00	8,39	3,00
98	8,39	2,00	8,39	4,00	6,84	-1,55	0,50	7,61	1,00	7,61	3,00
99	7,61	2,00	7,61	4,00	8,32	0,71	0,50	7,97	1,00	7,97	3,00
100	7,97	2,00	7,97	4,00	7,29	-0,67	0,50	7,63	1,00	7,63	3,00

Les valeurs sont arrondies à 10^{-2} près.

TAB. 3.2 - *Valeurs limites du coefficient adaptatif d'un modèle de tendance polynomiale d'ordre un à variances constantes en fonction de valeurs du ratio* $r = W/V$.

r	0,01	0,05	0,13	0,26	0,5	0,9	1	1,6	3,1	8
A	0,10	0,20	0,30	0,40	0,50	0,60	0,62	0,70	0,80	0,90

Les valeurs de A sont arrondies à 10^{-2} prés.

entier positif et d_t est strictement supérieur à 0. La moyenne de cette loi est n_t/d_t. L'estimation de la variance V au temps t est $S_t = d_t/n_t$. Ce sont les paramètres d_t et n_t qui sont estimés séquentiellement. L'introduction de ϕ induit la définition suivante :

Définition 2 *Pour tout t, le modèle de tendance polynomiale linéaire dynamique d'ordre un à variance d'observation V constante et inconnue est défini par :*

$$\begin{array}{lll} \textit{Équation d'observation :} & Y_t = \mu_t + \nu_t, & \nu_t \sim N[0\,;V], \\ \textit{Équation d'évolution :} & \mu_t = \mu_{t-1} + \omega_t, & \omega_t \sim T_{n_{t-1}}[0\,;W_t], \\ \textit{Information initiale :} & (\mu_0|D_0,V) \sim T_{n_0}[m_0\,;C_0], & \\ & (\phi|D_0) \sim G[n_0/2\,;d_0/2] & \end{array}$$

où m_0, C_0, n_0 et d_0 sont fixés et les séquences d'erreurs ν_t et ω_t conditionnellement à V sont indépendantes, mutuellement indépendantes et indépendantes de $(\mu_0|D_0,V)$.

$T_{n_0}[m_0\,;C_0]$ désigne une distribution de STUDENT à n_0 degrés de liberté dont le mode est m_0 et le paramètre d'échelle C_0. Le mode et le paramètre d'échelle sont respectivement un paramètre de position et un paramètre de dispersion. Dans une loi normale ces deux paramètres sont la moyenne et la variance.

La procédure séquentielle d'estimation est également modifiée. Soit au temps $t-1$ les variables aléatoires $(\mu_{t-1}|D_{t-1})$ distribuée selon la loi $T_{n_{t-1}}[m_{t-1}\,;C_{t-1}]$ et $(\phi|D_{t-1})$ selon la loi $G[n_{t-1}/2\,;d_{t-1}/2]$. Les paramètres de ces distributions sont supposés connus. Les étapes sont les suivantes :

A priori $(\mu_t|D_{t-1}) \sim T_{n_{t-1}}[a_t\,;R_t]$, où $a_t = m_{t-1}$ et $R_t = C_{t-1} + W_t$.

Prédiction $(Y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t\,;Q_t]$, où $f_t = a_t$ et $Q_t = R_t + S_{t-1}$, avec $S_{t-1} = d_{t-1}/n_{t-1}$.

A posteriori $(\mu_t|D_t) \sim T_{n_t}[m_t\,;C_t]$, où $m_t = a_t + A_t e_t$ et $C_t = (S_t/S_{t-1})(R_t - A_t^2 Q_t)$, avec $A_t = R_t/Q_t$ et $e_t = Y_t - f_t$. La distribution de $(\phi|D_{t-1})$ est mise à jour en $(\phi|D_t)$. Cette variable aléatoire est distribuée selon une loi $G[n_t/2\,;d_t/2]$, où $n_t = n_{t-1} + 1$ et $d_t = d_{t-1} + S_{t-1}e_t^2/Q_t$.

La distribution de l'estimation de Y_t est $(Y_t|D_t) \sim T_{n_t}[g_t\,;P_t]$, où $g_t = m_t$ et $P_t = C_t + S_t$. La distribution de prédiction à k unité(s) de temps est identique à celle du PTM d'ordre 1 (cf. page 17) à l'exception des variances d'observation et d'évolution. V_{t+k} est remplacée par l'estimation la plus récente de V, c'est-à-dire S_t. Les matrices W_{t+k} sont toutes égales à $W_{t+1} = C_t(1-\delta)/\delta$. L'inférence statistique s'effectue de manière classique. Ainsi, les bornes de l'intervalle de confiance du mode de $(\mu_t|D_t)$ sont $m_t \pm t_{n_t,1-\alpha/2}\sqrt{C_t}$, au risque de première espèce α fixé et où $t_{n_t,1-\alpha/2}$ représente la valeur d'une loi de STUDENT à n_t degrés de liberté et pour la probabilité $1-\alpha/2$.

La quantité n_t est augmentée d'une unité à chaque nouvelle observation. Elle peut être vue comme une mesure de la précision de l'estimation. La spécification d'une valeur importante pour n_0 signifie que l'utilisateur a une connaissance initiale précise de la valeur constante de V. Dans la procédure de mise à jour de d_t, la quantité $(S_{t-1}/Q_t)e_t^2$ est ajoutée à d_{t-1}. Le rapport S_{t-1}/Q_t est le pourcentage de la variance d'observation S_{t-1} dans la variance de prédiction Q_t. La mise à jour de d_{t-1} revient à la prise en compte du pourcentage du carré de l'erreur de prédiction attribuable à l'erreur de mesure, et d_t peut être vu comme une somme de carrés d'erreurs. L'estimateur $S_t = d_t/n_t$ est alors une somme de carrés d'erreurs divisée par le nombre d'observations (augmenté de n_0), ce qui est le calcul habituel de la variance. En pratique, au temps $t = 0$ l'utilisateur a une idée de la valeur de la variance S_0 et de la précision qu'il lui accorde en fixant n_0. À partir de ces deux valeurs, il peut calculer $d_0 = S_0 \times n_0$.

Illustration

Les données utilisées sont celles simulées pour l'illustration du PTM d'ordre un à variances constantes. Les conditions initiales sont $m_0 = 0$, $C_0 = 20$, $\delta = 0{,}5$, $n_0 = 1$ et $S_0 = d_0 = 25$. Cette valeur est 12,5 fois plus élevée que la vraie valeur valeur de V et implique ainsi une connaissance imprécise de la variance d'observation. Au début de la série temporelle les estimations des niveaux moyens du modèle à variance V constante et inconnue (FIG. 3.3) sont plus éloignées des valeurs simulées que les estimations du modèle à variances constantes (FIG. 3.2, page 19). Ceci est dû aux conditions initiales $m_0 = 0$, qui entraîne une erreur de prédiction importante au temps $t = 1$, et $S_0 = 25$, qui induit un coefficient adaptatif $A_1 = 0{,}62$. Cette valeur est inférieure à celle du modèle à variances constantes (*i.e.* 0,91) : la part de l'erreur prise en compte lors de l'étape *a posteriori* est moins grande et l'estimation *a posteriori* du niveau moyen est moins proche des données. Cette première erreur de prédiction a également pour effet d'augmenter la variance d'observation du temps $t = 0$, où elle est égale à 25, au temps $t = 1$ où elle est supérieure à 30 (FIG. 3.4). En moins de 10 unités de temps les niveaux moyens estimés selon chacun des modèles sont comparables. Le rapprochement de la courbe des estimations du niveau moyen et de la courbe des valeurs des simulations provoque une baisse brutale des estimations de la variance

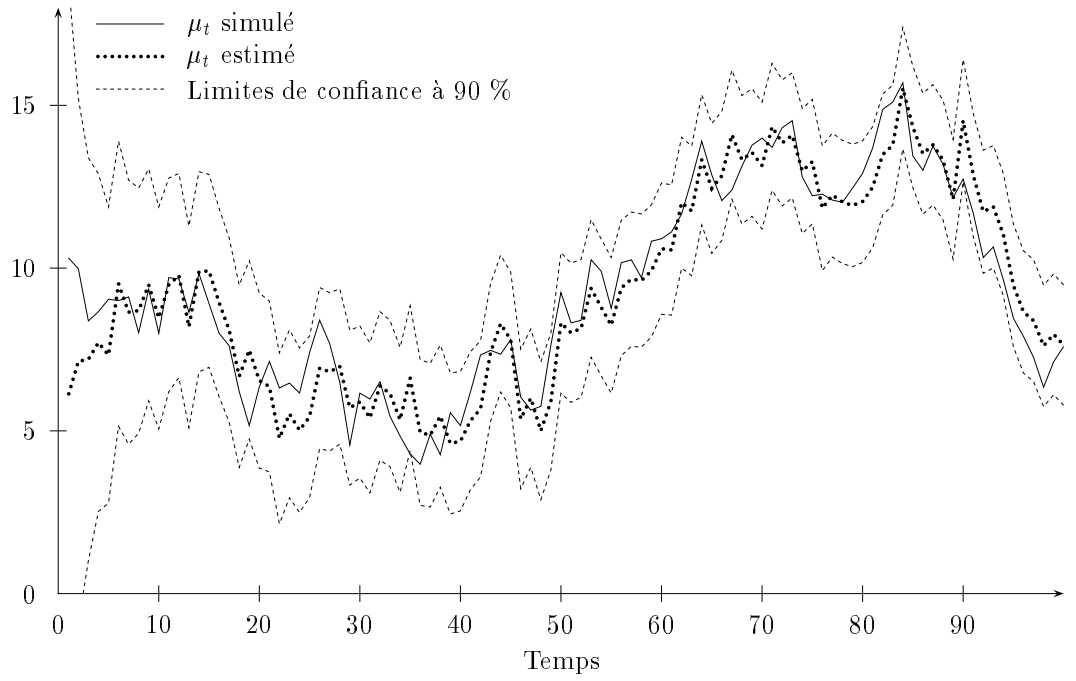

FIG. 3.3 - *Application d'un modèle de tendance polynomiale d'ordre un à variance V constante et inconnue.*

d'observation. Durant ces dix unités de temps les estimations de V restent très supérieures à la vraie valeur $V = 2$. Comme les variances C_t dépendent des estimations de la variance d'observation (cf. étape *a posteriori*, page 22), les intervalles de confiance des moyennes m_t sont également importants. Dans la suite de la série, les estimations de μ_t restant proches des valeurs simulées, les valeurs de S_t et les amplitudes des intervalles de confiance des estimations du niveau moyen diminuent. Enfin, bien que les S_t convergent vers 2, la dernière estimation de la variance d'observation est $S_{100} = 2{,}45$. WEST et HARRISON (1997, page 359) ont montré que la probabilité que la valeur estimée S_t tende asymptotiquement vers la vraie valeur V est 1. Si la série simulée comportait plus de données, l'écart de la dernière valeur estimée à la vraie valeur serait moins important. La vitesse de convergence dépend de la distance entre les « vraies » conditions initiales et celles spécifiées par l'utilisateur. Si les valeurs de m_0, C_0 et S_0 étaient moins éloignées des vraies valeurs, alors la capacité adaptative des MDB aurait permis d'obtenir plus rapidement des estimations cohérentes avec les observations. Toutefois, cette dernière remarque n'est vraie que si le facteur d'escompte est également proche de sa vraie valeur.

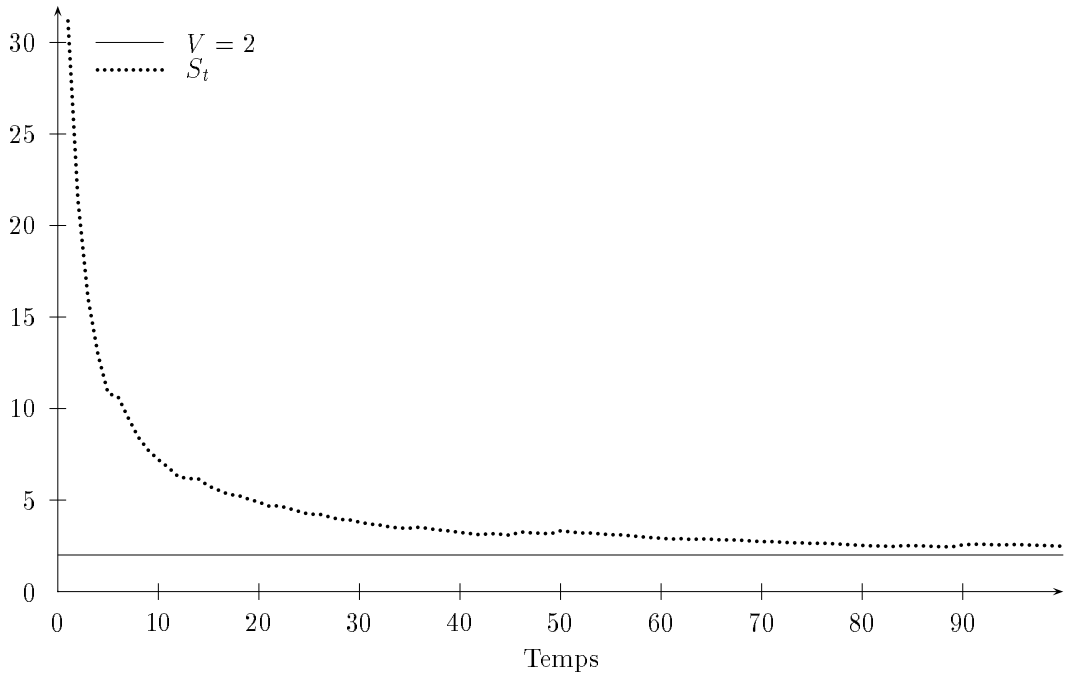

FIG. 3.4 - *Estimation de la variance d'observation constante et inconnue d'un modèle de tendance polynomiale d'ordre un.*

3.3.5 Spécification du facteur d'escompte et validité du modèle

L'analyse des erreurs de prédiction apporte une information sur la validité de la valeur spécifiée de δ. Supposons une série temporelle théorique dans laquelle la vraie valeur du facteur d'escompte est Δ. Si δ est supérieur à Δ alors le modèle est sous-adaptatif. Les changements dans la série temporelle sont trop lentement pris en compte et les erreurs de prédiction présentent des périodes importantes sans changement de signe : les erreurs sont corrélées positivement. Si δ est inférieur à Δ alors le modèle est sur-adaptatif et les erreurs de prédiction sont corrélées négativement, ce qui se traduit par une alternance de signes positifs et négatifs. La corrélation entre les erreurs de prédiction peut être appréciée visuellement. Des procédures statistiques permettent également d'évaluer le degré d'autocorrélation d'une série temporelle (cf. GOURIEROUX et MONFORT, 1990). Nous avons principalement utilisé le test des séquences qui présente les avantages d'être non-paramétrique et simple à mettre en œuvre (SIEGEL, 1956, cf. annexe B.1).

WEST et HARRISON (1997, page 58) proposent trois quantités permettant de juger la pertinence du modèle et la spécification du facteur d'escompte. Ce sont la moyenne des valeurs absolues des erreurs (*Mean Absolute Deviation*, MAD), la moyenne des carrés des erreurs (*Mean Square Error*, MSE) et la log-vraisemblance

TAB. 3.3 - *Statistiques de validité des modèles dynamiques bayésiens.*

δ	MAD	MSE	LL	LL *a posteriori*	Test des séquences
0,9	2,07	6,75	-236,53	-225,61	$p < 10^{-9}$
0,8	1,75	5,29	-224,40	-208,41	$p < 10^{-3}$
0,7	1,65	4,82	-220,47	-199,14	$p < 0.05$
0,6	1,63	4,69	-220,12	-192,61	NS
0,5	1,65	4,73	-221,77	-186,86	NS
0,4	1,69	4,88	-224,88	-180,76	$p < 10^{-2}$
0,3	1,75	5,13	-229,40	-173,24	$p < 10^{-2}$
0,2	1,83	5,47	-235,91	-162,63	$p < 10^{-3}$
0,1	1,94	5,91	-247,56	-145,04	$p < 10^{-4}$

NS : non significatif au risque de première espèce α =0,05 ; p : niveau de signification. Les valeurs sont arrondies à 10^{-2} près.

du modèle (*log-likelihood*, LL), telles que

$$\text{MAD} = \sum_{t=1}^{N} |e_t|/N,$$
$$\text{MSE} = \sum_{t=1}^{N} e_t^2/N,$$
$$\text{LL} = \log(\prod_{t=1}^{N} p(Y_t|D_{t-1})),$$

avec N le nombre d'observation et $p(Y_t|D_{t-1})$ la densité de l'observation Y_t calculée avec la loi de la distribution de la variable aléatoire de prédiction $(Y_t|D_{t-1})$ [†]. Plus MAD est MSE sont petits et LL grand, plus le modèle est proche des données et inversement. Le tableau 3.3 donne les valeurs de ces trois quantités pour différentes spécifications du facteur d'escompte du modèle à variance d'observation V constante et inconnue appliqué aux données simulées avec les conditions initiales de l'illustration de la section précédente (cf. page 23). Nous y avons également porté les résultats de l'application du test des séquences et la log-vraisemblance *a posteriori*, définie comme le logarithme de la vraisemblance de la série temporelle calculée avec les lois des distributions des variables aléatoires d'estimation $(Y_t|D_t)$. Le test du rapport de vraisemblance (KENDALL et STUART, 1977 ; DAGNELIE, 1975, volume 1 ; cf. annexe B.2) montre que les LL *a posteriori* sont

†. Une densité de probabilité n'est pas une probabilité. Si improprement, on accepte de qualifier $p(Y_t|D_{t-1})$ de probabilité, alors la log-vraisemblance peut être interprétée comme le logarithme de la probabilité d'observer la série temporelle Y_t, $t = 1, 2, \ldots$, sous les hypothèses du modèle utilisé.

toujours significativement plus élevées que les LL *a priori*, au risque de première espèce $\alpha = 0{,}05$. La prise en compte de l'observation Y_t dans le calcul des distributions d'estimation est à l'origine de ce résultat. La LL *a posteriori* croit à mesure que le facteur d'escompte diminue. Plus δ est proche de 0, plus le modèle est adaptatif et les estimations voisines des valeurs simulées. Dans une loi de STUDENT la densité maximum est observée pour le mode. Ainsi plus le facteur d'escompte est petit, plus les densités des observations de Y_t sont élevées et avec elles la log-vraisemblance *a posteriori*. Mais la diminution de δ s'accompagne également d'une augmentation des variances, c'est-à-dire d'une diminution de la précision de l'information sur les estimations de Y_t. Le modèle tend a être une simple interpolation linéaire des observations. Or, le modèle a pour objet l'estimation de la série temporelle inobservable du niveau moyen et pas celle des observations. C'est pourquoi la validité du modèle et du choix du facteur d'escompte est appréciée à l'aide de l'analyse des erreurs de prédictions. Les quantités MAD et MSE sont minimales et LL maximale pour la valeur $\delta = 0{,}6$, alors que la vraie valeur est 0,5 ‡. Face à ce résultat, un utilisateur, ne connaissant pas la vraie valeur du facteur d'escompte, pourrait décider de modifier son modèle en spécifiant $\delta = 0{,}6$. Cependant, si le nombre d'observations était très supérieur à 100, les variances et le coefficient adaptatif atteindraient leurs limites et MAD et MSE seraient minimales et LL maximale pour une valeur du facteur d'escompte plus proche de 0,5. Le test des séquences montre que seuls les modèles avec $\delta = 0{,}5$ et $\delta = 0{,}6$ présentent des erreurs de prédiction non-corrélées. L'application du test du rapport de vraisemblance montre que seule la LL du modèle avec $\delta = 0{,}5$ est non significativement différente de celle du modèle avec $\delta = 0{,}6$ au risque de première espèce $\alpha = 0{,}05$. Enfin, WEST et HARRISON (1997, page 50) ont montré qu'il était préférable que le facteur d'escompte spécifié soit légèrement inférieur à la vraie valeur Δ. Ainsi, le modèle ajusté aux informations disponibles au temps t s'adaptera efficacement aux données obtenues après le temps t.

L'utilisation du facteur d'escompte, de la procédure d'estimation séquentielle de la variance et de l'analyse des erreurs peuvent s'adapter à tous les modèles dynamiques bayésiens. Dans les modèles décrits ci-après, seules les différences induites par les définitions des modèles seront soulignées.

3.4 Modèle de tendance polynomiale d'ordre n

L'équivalent statique d'un PTM d'ordre un consiste à décrire la série temporelle par une droite de pente nulle, $Y_t = \mu$. Une droite de pente non nulle $Y_t = \mu + \beta t$

‡. Dans la simulation des données, le rapport $r = W/V$ est égal à 0,5, implique une valeur limite de 0,5 pour le coefficient adaptatif et ainsi la vraie valeur de $\delta = 1 - A$ est également 0,5 (cf. section 3.3.2, page 18).

est l'équivalent statique d'un PTM d'ordre deux. La représentation espace d'état de cette droite prend la forme des équations,

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \omega_{1,t},$$
$$\beta_t = \beta_{t-1} + \omega_{2,t},$$

où pour $i = 1, 2$, $\omega_{i,t} \sim N[0\,; W_{i,t}]$. μ_t reste le niveau moyen de la série et β_t est le changement dans le niveau moyen, c'est à dire la pente de la droite ajustée localement. Soit $\boldsymbol{\theta}_t$ le vecteur des paramètres du modèle et $\boldsymbol{\omega}_t$ le vecteur des erreurs d'évolution, tels que $\boldsymbol{\theta}'_t = (\mu_t \ \beta_t)$ et $\boldsymbol{\omega}'_t = (\omega_{1,t} \ \omega_{2,t})$. La forme matricielle de la représentation espace d'état d'un PTM d'ordre deux est son **équation d'évolution**. Elle s'écrit $\boldsymbol{\theta}_t = \boldsymbol{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$, avec $\boldsymbol{G}$ la matrice d'évolution telle que

$$\boldsymbol{G} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

et où $\boldsymbol{\omega}_t \sim N[\mathbf{0}\,; \boldsymbol{W}_t]$ avec

$$\boldsymbol{W}_t = \boldsymbol{G}\ \mathrm{diag}(W_{1,t}\,; W_{2,t})\ \boldsymbol{G}' = \boldsymbol{G} \begin{pmatrix} W_{1,t} & 0 \\ 0 & W_{2,t} \end{pmatrix} \boldsymbol{G}'.$$

L'**équation d'observation** est identique à celle du modèle d'ordre un et sa forme matricielle est $Y_t = \boldsymbol{F}'\boldsymbol{\theta}_t + \nu_t$, $\nu_t \sim N[0\,; V_t]$, où $\boldsymbol{F}'$ est le vecteur d'observation égal à (1 0). Le PTM d'ordre trois s'obtient en ajoutant au modèle le changement dans la pente sous la forme d'un paramètre supplémentaire γ_t dans le vecteur $\boldsymbol{\theta}_t$, et en modifiant les autres éléments du modèle en conséquence. En généralisant, dans un PTM d'ordre n, $\boldsymbol{F}$ est le vecteur d'observation $n \times 1$ (*i.e.* n lignes et 1 colonne) tel que $\boldsymbol{F}' = (1\ 0\ \ldots\ 0)$, la matrice d'évolution $\boldsymbol{G}$ est $n \times n$ telle que,

$$\boldsymbol{G} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 0 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix},$$

$\boldsymbol{\theta}'_t = (\theta_{1,t}\ \theta_{2,t}\ \ldots\ \theta_{n,t})$ est le vecteur des paramètres, pour tout i, $i = 1, 2, \ldots, n$, $\omega_{i,t} \sim N[0\,; W_{i,t}]$, et $\boldsymbol{\omega}'_t = (\omega_{1,t}\ \omega_{2,t}\ \ldots\ \omega_{n,t})$ est le vecteur des erreurs d'évolution tel que $\boldsymbol{\omega}_t \sim N[\mathbf{0}\,; \boldsymbol{W}_t]$ avec

$$\boldsymbol{W}_t = \boldsymbol{G}\ \mathrm{diag}(W_{1,t}\,; W_{2,t}\,; \ldots\,; W_{n,t})\ \boldsymbol{G}'.$$

La définition du PTM d'ordre n est la suivante.

Définition 3 *Pour tout t, le modèle de tendance polynomiale linéaire dynamique d'ordre n est défini par :*

$$\begin{array}{lll} \textit{Équation d'observation :} & Y_t = \boldsymbol{F}'\boldsymbol{\theta}_t + \nu_t, & \nu_t \sim N[0\,; V_t], \\ \textit{Équation d'évolution :} & \boldsymbol{\theta}_t = \boldsymbol{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim N[\mathbf{0}\,; \boldsymbol{W}_t], \\ \textit{Information initiale :} & (\boldsymbol{\theta}_0 | D_0) \sim N[\boldsymbol{m}_0\,; \boldsymbol{C}_0], & \end{array}$$

où les séquences d'erreurs ν_t et $\boldsymbol{\omega}_t$ sont indépendantes, mutuellement indépendantes et indépendantes de $(\boldsymbol{\theta}_0|D_0)$.

La forme matricielle de la procédure d'estimation est la suivante. Au temps $t-1$ la variable aléatoire $(\boldsymbol{\theta}_{t-1}|D_{t-1})$ est distribuée selon la loi normale $N[\boldsymbol{m}_{t-1}\,;\boldsymbol{C}_{t-1}]$, dont les paramètres sont supposés connus.

A priori $(\boldsymbol{\theta}_t|D_{t-1}) \sim N[\boldsymbol{a}_t\,;\boldsymbol{R}_t]$, où $\boldsymbol{a}_t = \boldsymbol{G}\boldsymbol{m}_{t-1}$ et $\boldsymbol{R}_t = \boldsymbol{G}\boldsymbol{C}_{t-1}\boldsymbol{G}' + \boldsymbol{W}_t$.

Prédiction $(Y_t|D_{t-1}) \sim N[f_t\,;Q_t]$, où $f_t = \boldsymbol{F}'\boldsymbol{a}_t$ et $Q_t = \boldsymbol{F}'\boldsymbol{R}_t\boldsymbol{F} + V_t$.

A posteriori $(\boldsymbol{\theta}_t|D_t) \sim N[\boldsymbol{m}_t\,;\boldsymbol{C}_t]$, où $\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t$ et $\boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{A}_t'Q_t$, avec $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}Q_t^{-1}$ et $e_t = Y_t - f_t$.

La distribution de l'estimation de Y_t est $N[g_t\,;P_t]$ où $g_t = \boldsymbol{F}'\boldsymbol{m}_t$ et $P_t = \boldsymbol{F}'\boldsymbol{C}_t\boldsymbol{F} + V_t$. La distribution de prédiction à l'horizon k, $k \geq 1$, $(Y_{t+k}|D_t) \sim N[f_t(k)\,;Q_t(k)]$ dépend de celle de $(\boldsymbol{\theta}_{t+k}|D_t) \sim N[\boldsymbol{a}_t(k)\,;\boldsymbol{R}_t(k)]$ tel que $f_t(k) = \boldsymbol{F}'\boldsymbol{a}_t(k)$ et

$$Q_t(k) = \boldsymbol{F}'\boldsymbol{R}_t(k)\boldsymbol{F} + V_{t+k},$$

où $\boldsymbol{a}_t(k) = \boldsymbol{G}\boldsymbol{a}_t(k-1)$ et

$$\boldsymbol{R}_t(k) = \boldsymbol{G}\boldsymbol{R}_t(k-1)\boldsymbol{G}' + \boldsymbol{W}_{t+k},$$

avec les valeurs initiales $\boldsymbol{a}_t(0) = \boldsymbol{m}_t$ et $\boldsymbol{R}_t(0) = \boldsymbol{C}_t$.

3.5 Modèle général

Dans le modèle général, le vecteur d'observation et la matrice d'évolution sont quelconques, variables dans le temps et connus quelque soit t.

Définition 4 *Pour tout t, le modèle linéaire dynamique général est défini par :*

$$\begin{array}{lll} \textit{Équation d'observation :} & Y_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \nu_t, & \nu_t \sim N[0\,;V_t], \\ \textit{Équation d'évolution :} & \boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim N[\boldsymbol{0}\,;\boldsymbol{W}_t], \\ \textit{Information initiale :} & (\boldsymbol{\theta}_0|D_0) \sim N[\boldsymbol{m}_0\,;\boldsymbol{C}_0], & \end{array}$$

où les séquences d'erreurs ν_t et $\boldsymbol{\omega}_t$ sont indépendantes, mutuellement indépendantes et indépendantes de $(\boldsymbol{\theta}_0|D_0)$.

La procédure séquentielle d'estimation, les calculs des paramètres de la distribution d'estimation de Y_t et ceux des distributions de prédictions à l'horizon k, $k \geq 1$, restent inchangés.

Nous avons vu que les estimations d'un MDB dépendent uniquement du passé et du présent de la série temporelle et sont ainsi cohérentes avec l'intuition selon laquelle le passé est la cause du futur. En revanche dans le modèle linéaire

généralisé, au temps t les paramètres du modèle dépendent du passé, du présent et du futur de la série temporelle. Ce résultat provient de l'hypothèse implicite d'invariance des paramètres du modèle. Dans certains cas, il peut être admis que les observations des temps $t+1, t+2, \ldots, t+k$ peuvent aider à de meilleures estimations des paramètres au temps t (cf. application de la section 4.4, page 78). Le théorème de Bayes permet la définition d'une procédure pour calculer les paramètres des distributions passées en conservant l'hypothèse de paramètres variables dans le temps. Une distribution ainsi calculée est dite filtrée ou bien lissée. Pour tout k, $1 \leq k < t$, la distribution lissée $(\boldsymbol{\theta}_{t-k}|D_t)$ est distribuée selon $N[\boldsymbol{a}_t(-k)\,;\boldsymbol{R}_t(-k)]$ où

$$\boldsymbol{a}_t(-k) = \boldsymbol{m}_{t-k} + \boldsymbol{B}_{t-k}[\boldsymbol{a}_t(-k+1) - \boldsymbol{a}_{t-k+1}],$$

et

$$\boldsymbol{R}_t(-k) = \boldsymbol{C}_{t-k} + \boldsymbol{B}_{t-k}[\boldsymbol{R}_t(-k+1) - \boldsymbol{R}_{t-k+1}]\boldsymbol{B}'_{t-k},$$

avec

$$\boldsymbol{B}_{t-k} = \boldsymbol{C}_{t-k}\boldsymbol{G}'_{t-k+1}\boldsymbol{R}^{-1}_{t-k+1},$$

et avec $\boldsymbol{a}_t(0) = \boldsymbol{m}_t$, $\boldsymbol{R}_t(0) = \boldsymbol{C}_t$. Si la variance d'observation est constante et connue, alors $(\boldsymbol{\theta}_{t-k}|D_t)$ a pour paramètre d'échelle la matrice $(S_t/S_{t-k})\boldsymbol{R}_t(-k)$. L'équation d'observation permet de calculer les paramètres de la distribution de $(Y_{t-k}|D_t)$.

3.5.1 Extensions

Spécifications et modèles

La liberté laissée dans les spécifications de $\boldsymbol{F}_t$ et $\boldsymbol{G}_t$ permet en particulier l'accès aux versions dynamiques du modèle de régression et des modèles « classiques » de série temporelle (Box et Jenkins, 1976), tels que les modèles autorégressifs (ar), moyenne mobile (ma), les modèles arma et les modèles arma intégrés (arima). Les éléments saisonniers et les fonctions de transfert peuvent également être modélisés sans modification de la procédure d'estimation. Tous ces éléments peuvent se combiner *ad libitum*.

Analyse référentielle

Les applications présentées dans les sections précédentes ont souligné l'influence des spécifications initiales sur le comportement des modèles dynamiques bayésiens. Il n'existe aucune technique bayésienne permettant de représenter l'état de complète ignorance des données. En revanche, une procédure séquentielle appelée « analyse référentielle » permet de calculer les valeurs de conditions

initiales à partir des données. L'utilisation de cette analyse objective et reproductible limite l'influence de l'utilisateur du modèle à la spécification du facteur d'escompte (ou aux variances d'estimation et d'évolution). Cette procédure est présentée en annexe C.

Variance d'observation variable dans le temps

L'étude du PTM d'ordre un a montré comment la simple hypothèse de dépendance du présent par rapport au passé permettait d'obtenir un système dynamique. Ce principe est applicable à la variance d'observation qui devient ainsi variable dans le temps. Lorsque cette extension est utilisée seul le rapport signal sur bruit est constant et fixé dans le modèle par la spécification du facteur d'escompte. De manière identique à la variance, il est possible de mettre en place une procédure séquentielle d'estimation du facteur d'escompte supposé constant ou variable dans le temps. Il convient de noter que l'augmentation du nombre de variables peut aller à l'encontre du principe de parcimonie.

Intervention

Parfois, des informations exogènes peuvent avoir une influence sur la série temporelle étudiée sans pouvoir être prises en compte par le modèle. Par exemple, en économie une grève aura une influence considérable et ponctuelle sur la production. Dans les modèles dynamiques bayésiens, une information de ce type est prise en compte par une modification des paramètres de la distribution de prédiction. Pratiquement, le modélisateur spécifie directement les paramètres de position et de dispersion en fonction de sa connaissance de l'impact de l'événement considéré. Des exemples de telles interventions sont donnés dans l'ouvrage de WEST et HARRISON (1997). De façon un peu différente, le concept d'intervention a également été développé pour les modèles statiques de séries temporelles (GOURIEROUX et MONFORT, 1990).

3.5.2 Données exceptionnelles et données manquantes

L'identification des données exceptionnelles est généralement réalisée par l'examen des erreurs standardisées, $u_t = e_t/\sqrt{Q_t}$, et de leur enveloppe de confiance à 90 %, obtenue par $N[0\,;1]$ si V_t est connue, et par $T_{n_{t-1}}[0\,;1]$ si la variance d'observation est constante et inconnue. Visuellement, une observation s'écartant de façon importante de l'enveloppe de confiance peut être désignée comme exceptionnelle. De façon plus formelle, il est possible de décider qu'une observation non-incluse dans une enveloppe de confiance à 95 % ou 99 % est une donnée exceptionnelle. Ces valeurs s'éloignant largement du processus moyen identifié par le modèle peuvent résulter d'importantes erreurs de mesure ou être le reflet d'une réelle variabilité ponctuelle. Idéalement, il est préférable de pouvoir déterminer

l'origine de ces observations. Mais il reste que ces données marginales n'apportent aucune information sur le processus moyen sous-jacent à la série temporelle étudiée et peuvent contribuer à une dégradation locale des performances du modèle. Lorsque la délicate décision de désigner une donnée comme exceptionnelle a été prise, il est d'usage de l'écarter de l'analyse. Pour les séries temporelles, son traitement revient alors à celui d'une donnée manquante. Dans les modèles dynamiques bayésiens, une donnée manquante au temps $t+1$ entraîne l'utilisation de la distribution *a priori* $(\boldsymbol{\theta}_{t+1}|D_t)$ comme distribution *a posteriori* pour le temps $t+1$. Si k données sont successivement manquantes, $k \geq 1$, c'est la distribution de $(\boldsymbol{\theta}_{t+k}|D_t)$ qui est utilisée comme distribution *a posteriori* pour le temps $t+k$.

L'analyse des erreurs peut également montrer des périodes où le modèle semble inadapté. Par exemple, un changement brutal du niveau moyen de la série temporelle peut entraîner une série d'erreurs du même signe. Pour détecter et traiter automatiquement ce genre d'incident et les données exceptionnelles, West et Harrison (1997, chapitre 11) proposent une procédure basée sur l'intervention. Le concept est généralisé avec les modèles multiprocess.

3.6 Modèle multiprocess

Spécifier les paramètres de la distribution de prédiction lors d'une intervention revient à supposer qu'il existe un second processus qui est ponctuellement ou localement plus adapté à la série temporelle étudiée. Un modèle dynamique constitué de plusieurs sous-modèles est un modèle multiprocess. West et Harrison (1997) identifient deux classes de modèles multiprocess. La première suppose une structure de modèle optimale dont un ou plusieurs paramètres sont incertains. Par exemple, si le modélisateur pense que la valeur de δ est incluse dans l'intervalle [0,7 ; 1], il pourra utiliser un modèle multiprocess de classe 1 comportant quatre sous-modèles avec les valeurs de δ égales, respectivement à, 0,7, 0,8, 0,9 et 1. Dans la seconde classe de modèle multiprocess, pour tout t, $t = 1, 2, \ldots$, un ou plusieurs paramètres sont structurellement susceptibles de prendre des valeurs différentes. Ce type de modèle a beaucoup été étudié et étendu (Harrison et Stevens, 1971, 1976 ; Taylor et Thomas, 1982 ; Fildes, 1983 ; Smith et al., 1983 ; Ameen et Harrison, 1985 ; Bolstad, 1986a, 1986b, 1988a, 1988b, 1995). Seule cette dernière classe est abordée par la présentation du cas particulier du modèle Harrison-Stevens à variances constantes et connues, qui fait l'objet d'une application au chapitre 4.

Le modèle Harrison-Stevens est composé de quatre ptm du second ordre. Chacun d'eux est associé à un état. L'écriture i_t, $i_t \in \{1, 2, 3, 4\}$, désigne l'état i au temps t. Les sous-modèles sont décrit par les quadruplés $\{\boldsymbol{F}, \boldsymbol{G}, V_{i_t}, \boldsymbol{W}_{i_t}\}$, où

$\boldsymbol{F}' = (1\ 0)$,

$$\boldsymbol{G} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

et V_{i_t} et $\boldsymbol{W}_{i_t}$ sont respectivement la variance d'observation et la matrice de variances-covariances d'évolution du sous-modèle de l'état i au temps t. Les sous-modèles correspondent aux états :

(1) « routine » ;

(2) « donnée exceptionnelle » ;

(3) « changement de niveau » ;

(4) « changement de pente ».

Le sous-modèle de l'état (1) sert de référence pour la spécification des éléments des autres sous-modèles. Ainsi, les sous-modèles des états « donnée exceptionnelle », « changement de niveau » et « changement de pente » auront des valeurs plus élevées pour les variances, respectivement, d'observation, d'évolution du niveau moyen et d'évolution de la pente. À titre d'exemple, West et Harrison (1997) donnent les spécifications suivantes :

$$\begin{array}{llll}
(1) & V_{1_t} = 1 & \text{et} & \boldsymbol{W}_{1_t} = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.01 \end{pmatrix}; \\
(2) & V_{2_t} = 100 & \text{et} & \boldsymbol{W}_{2_t} = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.01 \end{pmatrix}; \\
(3) & V_{3_t} = 1 & \text{et} & \boldsymbol{W}_{3_t} = \begin{pmatrix} 10 & 0.01 \\ 0.01 & 0.01 \end{pmatrix}; \\
(4) & V_{4_t} = 1 & \text{et} & \boldsymbol{W}_{4_t} = \begin{pmatrix} 1.1 & 1 \\ 1 & 1 \end{pmatrix}.
\end{array}$$

À chacun des états est associée une probabilité d'occurrence avant l'observation de Y_t. Ce sont les probabilités *a priori* de l'état i au temps t conditionnellement à l'état j au temps $t-1$, $\mathrm{P}(i_t|j_{t-1}, D_{t-1})$, pour $(i_t, j_{t-1}) \in \{1,2,3,4\}^2$. Par hypothèse, elles sont indépendantes de l'état j au temps $t-1$, constantes et connues :

$$\begin{aligned}
\pi_{i_t} &= \mathrm{P}(i_t|j_{t-1}, D_{t-1}), \\
&= \mathrm{P}(i_t|D_{t-1}), \\
&= \mathrm{P}(i_t|D_0).
\end{aligned}$$

Ces probabilités sont incluses dans l'ensemble D_0 et leur somme est égale à 1. L'état « routine » est supposé le plus fréquent. Une spécification des π_{i_t} peut être $\pi_{1_t} = 0{,}85$, $\pi_{2_t} = 0{,}07$, $\pi_{3_t} = 0{,}05$ et $\pi_{4_t} = 0{,}03$. Pour $i_t \in \{1,2,3,4\}$, les probabilités *a posteriori* de l'état i au temps t sont $\mathrm{P}(i_t|D_t) = p_{i_t}$, et pour

$(i_t, j_{t-1}) \in \{1,2,3,4\}^2$, les probabilités de l'état i au temps t et de l'état j au temps $t-1$ sont définies par $\mathrm{P}(i_t, j_{t-1}|D_t) = p_{i_t,j_{t-1}}$. Enfin, les conditions initiales du modèle sont données par la distribution de $(\boldsymbol{\theta}_0|D_0)$. Cette distribution est commune aux quatre sous-modèles.

La procédure de mise à jour du modèle HARRISON-STEVENS est la suivante. Au temps $t-1$, pour tout $j_{t-1} \in \{1,2,3,4\}$, la distribution *a posteriori* $(\boldsymbol{\theta}_{t-1}|j_{t-1}, D_{t-1})$ est distribuée selon une loi normale, $N(\boldsymbol{m}_{j_{t-1}}; \boldsymbol{C}_{j_{t-1}})$ avec la probabilité $p_{j_{t-1}}$.

A priori Chaque état j_{t-1} peut être suivi d'un des quatre états i_t. Les 16 successions possibles d'états du temps $t-1$ au temps t induisent autant de distributions *a priori* $(\boldsymbol{\theta}_t|i_t, j_{t-1}, D_{t-1})$. Pour tout $(i_t, j_{t-1}) \in \{1,2,3,4\}^2$, ces distributions sont gaussiennes, de moyennes $\boldsymbol{a}_{i_t,j_{t-1}} = \boldsymbol{G}\boldsymbol{m}_{j_{t-1}}$ et de variances $\boldsymbol{R}_{i_t,j_{t-1}} = \boldsymbol{G}\boldsymbol{C}_{j_{t-1}}\boldsymbol{G}' + \boldsymbol{W}_{i_t}$.

Prédiction Il existe également 16 distributions de prédiction $(Y_t|i_t, j_{t-1}, D_{t-1})$, gaussiennes, de moyennes $f_{i_t,j_{t-1}} = \boldsymbol{F}'\boldsymbol{a}_{i_t,j_{t-1}}$ et de variances $Q_{i_t,j_{t-1}} = \boldsymbol{F}'\boldsymbol{R}_{i_t,j_{t-1}}\boldsymbol{F} + V_{i_t}$. La densité de probabilité de la prédiction $(Y_t|D_{t-1})$ du modèle multiprocess est la somme des 16 densités de prédiction $(Y_t|i_t, j_{t-1}, D_{t-1})$ pondérées par les produits des probabilités $\pi_{i_t} \times p_{j_{t-1}}$. La moyenne f_t et la variance Q_t de $(Y_t|D_{t-1})$ sont, respectivement, la somme des $f_{i_t,j_{t-1}}$ et des $Q_{i_t,j_{t-1}}$, pondérées par $\pi_{i_t} \times p_{j_{t-1}}$.

A posteriori L'observation de Y_t induit le calcul des erreurs $e_{i_t,j_{t-1}}$. Les mises à jour des distributions *a priori* en distributions *a posteriori* s'effectuent de manière classique. Les probabilités $p_{i_t,j_{t-1}}$ de chacun des 16 modèles sont calculées telles que :

$$p_{i_t,j_{t-1}} = c_t \pi_{i_t} p_{j_{t-1}} \frac{\exp(-0{,}5 e^2_{i_t,j_{t-1}}/Q_{i_t,j_{t-1}})}{\sqrt{Q_{i_t,j_{t-1}}}},$$

où c_t est une constante telle que $\sum_{i_t=1}^4 \sum_{j_{t-1}=1}^4 p_{i_t,j_{t-1}} = 1$.

Effondrement C'est une étape propre aux modèles multiprocess qui correspond à la réduction du nombre de distributions *a posteriori* au nombre d'états considérés. Le principe est de supposer l'effet de l'état en $t-1$ comme négligeable pour le temps $t+1$, et d'estimer les paramètres de la loi normale de $(\boldsymbol{\theta}_t|i_t, D_t)$ par approximation du mélange des quatres distributions $(\boldsymbol{\theta}_t|i_t, j_{t-1}, D_t)$, $j_{t-1} \in \{1,2,3,4\}$. L'analyse montre que la moyenne et la variance de cette distribution approchée sont

$$\boldsymbol{m}_{i_t} = \sum_{j_{t-1}=1}^{4} \boldsymbol{m}_{i_t,j_{t-1}} p_{i_t,j_{t-1}}/p_{i_t},$$

et

$$\boldsymbol{C}_{i_t} = \sum_{j_{t-1}=1}^{4} [\boldsymbol{C}_{i_t,j_{t-1}} + (\boldsymbol{m}_{i_t} - \boldsymbol{m}_{i_t,j_{t-1}})(\boldsymbol{m}_{i_t} - \boldsymbol{m}_{i_t,j_{t-1}})']p_{i_t,j_{t-1}}/p_{i_t},$$

avec $p_{i_t} = \sum_{j_{t-1}=1}^{4} p_{i_t,j_{t-1}}$.

La densité de probabilité de l'estimation $(Y_t|D_t)$ du modèle multiprocess est la somme des 16 densités d'estimation $(Y_t|i_t, j_{t-1}, D_t)$ pondérées par les probabilités $p_{i_t,j_{t-1}}$. Comme pour la densité de prédiction, la moyenne g_t et la variance P_t de $(Y_t|D_t)$ sont, respectivement, la somme des $g_{i_t,j_{t-1}}$ et des $P_{i_t,j_{t-1}}$, pondérées par $p_{i_t,j_{t-1}}$. Il faut remarquer que dans le calcul de ces probabilités entrent les variances de prédiction et avec elles les variances d'observation des quatre sous-modèles. De ce fait, dans le modèle HARRISON-STEVENS la spécification de V_{1_t} a une influence sur les valeurs des estimations, influence qu'elle n'a pas dans les modèles polynomiaux (cf. section 3.3.2, page 18). L'importance de cette particularité est soulignée dans l'application présentée au chapitre 4.

Comme dans le modèle général, il est possible de définir les distributions de prédiction à un horizon de k unités de temps. Dans la procédure décrite ci-avant il faut 4^2 distributions pour calculer une prédiction au temps $t+1$. Pour $k = 2$, il faut calculer les 4^3 distributions $(\boldsymbol{\theta}_{t+2}|h_{t+2}, i_{t+1}, j_t, D_t)$. L'augmentation exponentielle du nombre de distributions avec l'augmentation de l'horizon induit essentiellement des difficultés numériques. Le calcul des paramètres des distributions lissées conduit au même problème. La procédure de lissage pour un horizon $k = 1$ est décrite dans l'annexe D. L'hypothèse de variance d'observation constante et inconnue est transposable aux modèles multiprocess.

La modélisation des séries temporelles comportant des irrégularités n'est pas uniquement le fait des MDB. Une importante littérature y est consacrée au sein de laquelle on trouve des approches statiques (MUIRHEAD, 1986 ; DIGGLE et ZEGER, 1989), dérivées de l'analyse de données (JASSBY et POWELL, 1990) ou issues du contrôle de qualité (LAI, 1995) et des applications du filtre de KALMAN et de l'algorithme d'*Expectation-Maximization* (EM) (DEMPSTER et al., 1977 ; LE et al., 1996).

3.7 Modèle de régression linéaire dynamique

Dans un modèle statique de regression linéaire, une variable dépendante Y_t, $t = 1, 2, \ldots$, est en relation avec n variables indépendantes ou covariables $X_{1,t}$, $X_{2,t}, \ldots, X_{n,t}$, $t = 1, 2, \ldots$, selon l'équation,

$$Y_t = \theta_0 + \sum_{i=1}^{n} \theta_i X_{i,t} + \nu_t,$$

où θ_0 est l'ordonnée à l'origine, θ_i, $i = 1, 2, \ldots$, est le paramètre de la $i^{\text{ème}}$ covariable et ν_t est un bruit blanc. La forme dynamique de la régression linéaire a pour équation d'observation,

$$Y_t = \boldsymbol{F}_t' \boldsymbol{\theta}_t + \nu_t,$$

avec $\boldsymbol{F}_t' = (1\ X_{1,t}\ X_{2,t}\ \ldots\ X_{n,t})$ et $\boldsymbol{\theta}_t' = (\theta_{0,t}\ \theta_{1,t}\ \ldots\ \theta_{n,t})$ où $\theta_{0,t}$ est la version dynamique de l'ordonnée à l'origine, *i.e.* le niveau moyen. L'équation d'évolution est

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t.$$

Le modèle de regression linéaire dynamique (*Dynamic Linear Regression Model*, DLRM) n'est qu'un cas particulier du modèle général avec $\boldsymbol{F}_t$ en tant que vecteur des régresseurs et $\boldsymbol{G}_t$ égal à la matrice identité.

Dans la régression dynamique, les paramètres des variables indépendantes sont potentiellement variables dans le temps et avec eux les relations entre la variable dépendante et les covariables. Le risque lié à cette potentialité est qu'une relation constante apparaisse variable. Le choix des variables explicatives est donc très important. En premier lieu, si une relation est connue comme constante, alors il convient de la traiter comme telle en fixant la variance du paramètre considéré à 0. Ensuite, il faut pouvoir interpréter *a posteriori* les variabilités des relations. En effet, une relation pouvant être envisagée comme variable peut s'avérer inexploitable du fait des interrelations entre les covariables. Ce problème relevant de la colinéarité entre régresseurs est le même que celui rencontré en régression statique. Il peut être détecté par l'examen des corrélations entre les paramètres des covariables. Une corrélation tendant vers 1 doit être considérée comme suspecte. Dans ce cas le modèle est sur-paramétré et la solution consiste à oter une variable indépendante du modèle. La variabilité des relations est contrôlée par les rapports signal sur bruit qui dépendent de la spécification des matrices $\boldsymbol{W}_t$. À l'étape *a posteriori*, la matrice de variances-covariances $\boldsymbol{C}_t$ des paramètres dépend du vecteur des régresseurs. Ainsi, l'utilisation du facteur d'escompte génère par la relation $\boldsymbol{W}_t = \boldsymbol{C}_{t-1}(1-\delta)/\delta$ des matrices d'évolution dépendantes des données et variables dans le temps. Plus le facteur d'escompte est petit, plus les relations sont supposées variables. Et plus une relation est variable, plus elle tend à être aléatoire. C'est pourquoi West et Harrison (1997) conseillent l'utilisation d'un facteur d'escompte relativement grand. À titre d'exemple, avec un facteur d'escompte de 0,95 les éléments des matrices $\boldsymbol{W}_t$ représentent 5,26 % (à 10^{-2} près) des élément des matrices $\boldsymbol{C}_{t-1}$. De plus, une relative stabilité des relations est souhaitable afin d'obtenir des prédictions pertinentes. Comme la moyenne de $(Y_t|D_{t-1})$ dépend du vecteur des régresseurs au temps t, elle n'est pas à proprement parler une prédiction. De vraies prédictions peuvent être obtenues si les covariables sont retardées, *i.e.* $\boldsymbol{F}_t' = (1\ X_{1,t-k_1}\ X_{2,t-k_2}\ \ldots\ X_{n,t-k_n})$ avec pour $i = 1, 2, \ldots, n$, $k_i \geq 1$. Il est également possible de construire un

modèle multivarié permettant de prédire conjointement les régresseurs et la variable dépendante, ou d'utiliser les prédictions de modèles séparés à la place des valeurs des covariables. Par ailleurs, ces approches permettent de traiter les données manquantes des variables indépendantes. West et Harrison ont développé la théorie des modèles multivariés et donnent la méthode pour incorporer les prédictions de modèles séparés (West et Harrison, 1997, resp. chapitre 16 et pages 278-279). Si le modèle est utilisé rétrospectivement dans un but explicatif, ces méthodes peuvent également être utilisées, bien que le problème de prédiction soit alors moins crucial. En revanche, il peut être bénéfique de standardiser toutes les variables (*i.e.* soustraire aux données leurs moyennes puis diviser par les racines de leurs variances) afin que les effets des covariables soient clairement séparés du niveau moyen et que les paramètres du modèle soient comparables.

La littérature concernant les modèles de régression dynamique n'est pas très abondante. En 1975, Brown et al. proposent des techniques pour tester et traiter les régressions sous l'hypothèse de relations variables dans le temps. Dans leur article de 1976, Harrison et Stevens abordent rapidement le modèle présenté ci-dessus. O'Hagan (1978) développe un modèle de régression localisée dont un cas particulier serait la régression présentée par Harrison et Stevens (1976). Johnston et Harrison (1980) présentent une application d'un dlrm. En plus de covariables, le modèle utilisé inclut une tendance du second ordre et une dynamique saisonnière. Duncan et Horn (1982) présentent un modèle de régression dynamique sans faire référence à Harrison et Stevens (1976). Johnston et al. (1986) proposent une application de dlrm et une réflexion sur le processus de modélisation. Les principales sources d'informations concernant les dlrm restent les ouvrages de West et Harrison (1997) et de Pole et al. (1994).

3.8 Modèles de série temporelle, filtre de Kalman et mdb

L'expression « modèles de série temporelle » désigne traditionnellement les modèles arima (*AutoRegressive Integrated Moving Average*) introduis par Box et Jenkins (1976). Ces modèles sont beaucoup utilisés en économétrie où les séries temporelles sont courantes. Tous les processus arima peuvent être modélisés par des mdb (Ameen et Harrison, 1985). Plus précisément, les formes dégénérées (*i.e.* après convergence) des modèles linéaires dynamiques de série temporelle (*Time Series Dynamic Linear Models*, tsdlm) à variance d'observation constante ont des équivalents arima. En particulier, les modèles polynomiaux sont des tsdlm et les formes dégénérées des modèles d'ordre 1, 2 et 3 correspondent, respectivement, aux arima(0,1,1), arima(0,2,2) et arima(0,3,3) (West et Harrison, 1997, resp. pages 46-48, 219-222 et 228-229). Ces modèles dynamiques polynomiaux et arima sont eux mêmes équivalents aux lissages exponentiels, res-

pectivement, simple, double et triple (McKenzie, 1984). Les méthodes de pondération exponentielle sont étroitement liées aux modèles dynamiques bayésiens (Ameen et Harrison, 1984) et arima (Box et Jenkins, 1976 ; Chatfield, 1989 ; Gourieroux et Monfort, 1990 ; Berthouex et Box, 1996). Dans le contexte des modèles arima, les modèles autorégressifs conditionnellement hétéroscédastiques (arch) (Engle, 1982) et leur dérivés (Bresson et Pirotte, 1995, pages 579-610) permettent de supposer la variance d'observation variable dans le temps et de l'estimer à partir des données. Jusqu'à présent, il ne semble pas qu'il existe une comparaison de ces modèles avec les mdb à variance V variable dans le temps. Enfin, les modèles de série temporelle admettent une représentation espace d'états (Chatfield, 1989 ; Gourieroux et Monfort, 1990) qui est généralement associée à une mise à jour par le filtre de Kalman.

Originellement, le filtre de Kalman (Kalman, 1960 ; Kalman et Bertram, 1960a, 1960b) est une méthode linéaire provenant de la théorie du contrôle. Une grande partie de la littérature concernant cet outil mathématique a été publiée dans des revues d'ingénierie. Le terme « filtre » y désigne une méthode permettant de dissocier une série temporelle de ses erreurs de mesure. Le filtre de Kalman a été étendu aux cas non-gaussiens (*Extended Kalman Filter*, ekf; voir Jazwinski, 1970) et à variance d'observation variable dans le temps (Zehnwirth, 1988). Le formalisme et le vocabulaire utilisés en ingénierie sont éloignés de ceux utilisés en statistique. Afin de favoriser l'application du filtre de Kalman dans ce domaine, Meinhold et Singpurwalla (1983) ont présenté la méthode sous un aspect bayésien. Le principal intérêt suscité par le filtre de Kalman concerne le traitement des données exceptionnelles, l'estimation des données manquantes et des valeurs de la série temporelle étudiée (Meinhold et Singpurwala, 1989 ; Yatawara et al., 1991 ; Tiwari et Dienes, 1994). Malgré une ressemblance formelle incontestable entre les mdb et le filtre de Kalman, les articles utilisant ce dernier restent difficiles d'accès. En particulier, les conditions initiales sont en général incomplètes ou omises. Toutefois, Shumway et Stoffer (1982) proposent une estimation de ces quantités par l'utilisation de l'algorithme d'*Expectation-Maximization* (em). Enfin, il est parfois fait référence à de *multiprocess Kalman filter models* et, en particulier, il est remarquable que les « bayésiens » Smith et West utilisent cette expression (Smith et West, 1983). Les résultats de Kalman sont à la base du modèle Harrison-Stevens (Harrison et Stevens, 1971). Mais West et Harrison (1997) déclarent « *To say that "Bayesian forecasting is Kalman Filtering" is akin to saying that statistical inference is regression!* ». Le lien avec le filtre de Kalman a été discuté à deux reprises (cf. discussions de Harrison et Stevens, 1976 ; West et al., 1985). L'originalité des mdb réside dans l'utilisation du filtre de Kalman pour l'estimation en temps réel de paramètres aléatoires et variables dans le temps.

Finalement, il est probable que l'utilisation d'une méthode statique couplée

au filtre de KALMAN permettent d'obtenir des résultats peu différents de ceux d'un MDB, en dépit d'éventuelles divergences analytiques (cf. discussion de DAVIS de l'article de WEST et al., 1985). De plus, ces approches ne sont que des méthodologies différentes pour aborder le même problème du traitement des séries temporelles. Ce qui est important, c'est de choisir une méthode qui répond à ses besoins et suffisamment familière pour éviter les principaux écueils. Les modèles de séries temporelles développés en économétrie présentent un formalisme parfois rugueux et le vocabulaire qui leur est associé est teinté de cette origine économique. Le filtre de KALMAN couplé à ces modèles reste un élément exogène avec un formalisme et un vocabulaire propre. Les MDB viennent de la recherche opérationnelle et ont principalement été utilisés en économie. Mais, en tant que généralisation dynamique du modèle linéaire, ils s'appuient sur un vocabulaire et une méthode qui, par le biais de la régression, sont reconnus dans tous les domaines de la science (cf. introduction de TOMASSONE et al., 1983). Enfin, par définition la procédure séquentielle d'estimation fait structurellement partie des MDB.

Chapitre 4

Applications

4.1 Séries temporelles d'Antifer

Afin d'étudier les conditions d'apparition du *Dinophysis*, des prélèvements d'eau de mer ont été effectués dans le port pétrolier d'Antifer (Fig. 4.1). Ce site a été choisi car de tous les points de prélèvement du réseau de surveillance phytoplanctonique il présentait chaque année les plus fortes concentrations en *Dinophysis*. L'échantillonnage a été réalisé durant trois à quatre mois pendant les périodes estivales des années 1987 à 1993. L'eau de mer était prélevée journellement, à marée haute, à l'extrémité du quai et en surface, *i.e.* de 1 à 3 mètres de profondeur. L'échantillonnage et les mesures suivantes ont été réalisés par le laboratoire municipal du Havre : concentration en *Dinophysis* (cell.$(10\ \mathrm{mL})^{-1}$; Utermöhl, 1955), température de l'eau (°C ; sonde Ponselle), salinité (salinomètre par induction Beckman), concentration en chlorophylle *a* ($\mu\mathrm{g.L}^{-1}$), en nitrate (NO_3), en phosphate (PO_4) et en silicate (SIO_2) ($\mu\mathrm{mol.L}^{-1}$, autoanalyseur Technicon). Le tableau 4.1 synthétise les principales informations concernant ces mesures. Il faut ajouter que la concentration en *Dinophysis*, la température et la salinité ont été mesurés au fond (30 mètres) en 1987 et 1988. Cette dernière année, les concentrations en chlorophylle *a*, nitrate, phosphate et silicate ont également été mesurées à 30 mètres de profondeur. En 1989, en plus des résultats de surface, toutes les variables ont été mesurées à 4 mètres de profondeur, à l'exception de la concentration en silicate. Pour ces prélèvements à 4 et 30 mètres de profondeur, les dates de début et fin d'échantillonnage et les proportions de données manquantes sont identiques à celles des mesures de surface présentées dans le tableau 4.1. La concentration en *Prorocentrum* spp. (cell.$(10\ \mathrm{mL})^{-1}$; Utermöhl, 1955) a été mesurée en 1990. Les heures de prélèvement des échantillons d'eau de mer sont disponibles pour les années 1990 et 1991. Enfin, les coefficients de marée au port du Havre sont publiés chaque année dans « Annuaire des marées

TAB.4.1: *Mesures effectuées sur les échantillons d'eau de mer prélevés à ANTIFER de 1987 à 1993. Les dates « Début » et « Fin » correspondent respectivement aux dates de première et de dernière mesures. Lorsqu'une variable a été mesurée, le tableau donne sa proportion de données manquantes (à l'unité prés). Lorsque l'intervalle [Début ; Fin] est différent de l'intervalle de mesure d'une variable, ce dernier est précisé. De plus, le taux de données manquantes est calculé sur la base de ce nouvel intervalle.*

	Années						
	87	88	89	90	91	92	93
Début	03/06	26/05	01/07	01/06	03/06	19/06	03/06
Fin	30/09	30/09	13/09	11/09	30/09	28/09	30/09
Dinophysis	16 %	0 %	1 %	1 %	0 %	62 %	77 %
Température de l'eau	16 %	0 %	1 %	8 %	0 % du 03/06 au 27/09	62 %	77 %
Salinité	17 % du 05/06 au 30/09	0 %	1 %	7 %	0 % du 02/06 au 15/09	62 %	77 %
Chlorophylle *a*		0 % du 26/05 au 11/08	1 %	10 %	0 % du 02/06 au 14/09		
Nitrate		0 % du 26/05 au 18/08	1 %	7 %	0 % du 02/06 au 22/09		
Phosphate		0 % du 26/05 au 17/08	1 %	8 %	0 % du 02/06 au 22/09		
Silicate		0 % du 26/05 au 16/08					

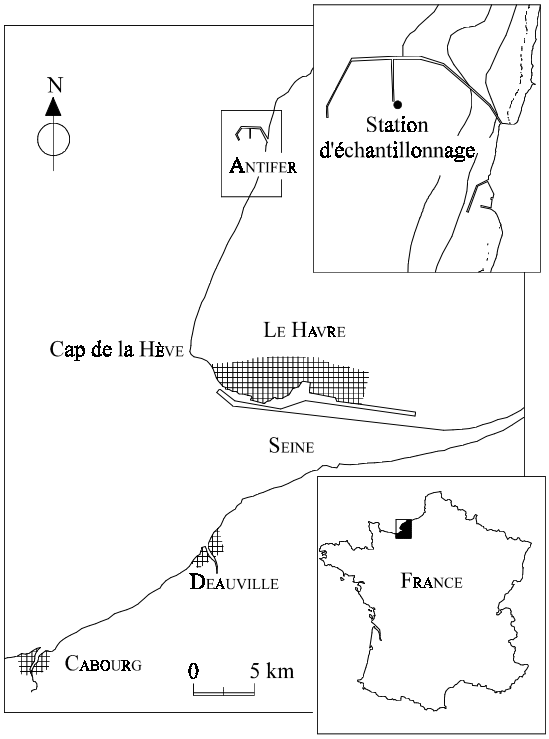

Fig. 4.1 - *Position géographique du site d'Antifer.*

des ports de France » édité par le service hydrographique et océanographique de la marine (shom).

Des données météorologiques ont été fournies par la station du cap de La Hève. De 1987 à 1993, elles concernent les températures maximales, minimales et moyenne de l'air (°C), l'insolation (heure.jour^{-1}), le volume et la durée des précipitations (resp. mm.jour^{-1} et heure.jour^{-1}). Ces données ne présentent pas de données manquantes à l'exception de l'insolation pour les années 1987 et 1988 où les proportions sont respectivement 16 % et 5 % (à l'unité prés). Pour toutes les années, nous disposons du débit de la Seine (m^3.s^{-1}) et, en fréquence trihoraire, de l'état de la mer et de la vitesse (m.s^{-1}) et la direction du vent. Le débit de la Seine ne présentait aucune donnée manquante. La moyenne journalière des données trihoraires de l'état de la mer a été réalisé et compte 1 % de données manquantes en 1991. La nature bidimensionelle des données de vent impose la création d'une variable unidimensionnelle. Nous avons choisi d'effectuer la transformation $[-v_{i,t}\cos(d_{i,t}+\alpha)]$ où $v_{i,t}$ et $d_{i,t}$ sont respectivement la vitesse et la direction du vent lors de la $i^{\text{ème}}$ mesure du jour t et α est une constante. Cette formule est une projection du vecteur de vent sur une droite dont l'orientation est contrôlée par la constante α. Les vents en provenance du sud-ouest semblaient avoir une influence sur la concentration en *Dinophysis* (Lassus et al., 1993). Ainsi, nous avons choisi une projection selon un axe nord-est–sud-ouest en

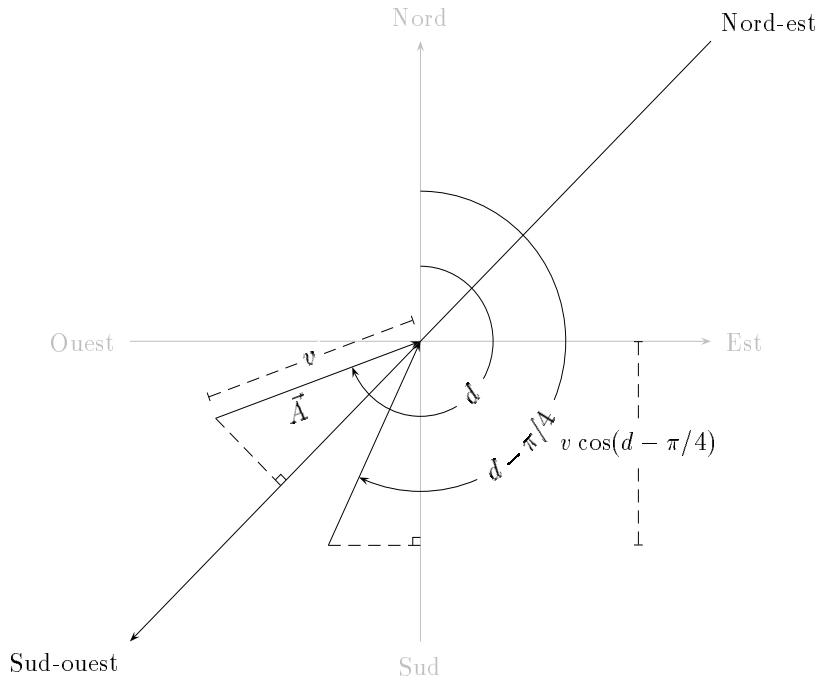

FIG. 4.2 - *Représentation graphique de la transformation des variables liées au vent. Le vecteur de vent $\vec{A}$ est défini par sa direction d (*i.e. *son angle) et sa vitesse v (*i.e. *sa longueur). La valeur de la variable « vent de sud-ouest » correspondant au vecteur $\vec{A}$ est la longueur de sa projection sur l'axe nord-est–sud-ouest égale à $-v\cos(d - \pi/4)$.*

spécifiant $\alpha = -\pi/4$ (cf. FIG. 4.2). La nouvelle variable « vent de sud-ouest » est positive pour les vents de secteur sud-ouest, négative pour les vents de secteur nord-est, et nulle pour les vents de nord-ouest et de sud-est. Elle présente 3 % de données manquantes en 1993.

En conclusion, il faut remarquer que les années 1987, 1992 et 1993 présentent des proportions importantes de données manquantes et les variables environnementales autre que météorologiques se limitent à la température de l'eau et la salinité. Par ailleurs, les données de l'année 1991 comptent seulement sept jours où les concentrations en *Dinophysis* ne sont pas nulles. Enfin, les données de température de l'air sont très corrélées avec la température de l'eau, et de ce fait n'ont pas été utilisées. Les mesures de chlorophylle *a* ont également été écartées, car elles intègrent la biomasse de la microalgue toxique.

4.2 Traitements préliminaires

La première partie du travail a consisté en une meilleure connaissance des données d'ANTIFER. Elle a été réalisée par l'application de méthodes descriptives (e.g. indicateurs de position et de dispersion des distributions des variables disponibles, graphiques) et de méthodes statistiques classiques (e.g. régression, arbre de régression). Dans la droite ligne des travaux précédents (SOUDANT, 1993), les premières applications de modèles dynamiques bayésiens ont eu pour objectif la prédiction des concentrations en *Dinophysis*. Les modèles dynamiques choisis étaient des modèles polynomiaux, des modèles autorégressifs et des régressions comportant uniquement des covariables retardées. Différentes formes d'intervention (cf. page 31) ont été testées. Bien que ces modèles présentaient généralement des performances satisfaisantes, ils se sont révélés mal adaptés à la prévision des augmentations brutales de la concentration en *Dinophysis*. Ces résultats ont eu la vertu de nous rappeler que la compréhension d'un phénomène doit précéder la construction d'un système pour le prévoir. En l'occurrence, les informations disponibles quant à la biologie et l'écologie de la microalgue toxique étaient trop incomplètes pour permettre la réalisation d'un tel système. Dès lors, l'objectif poursuivi a été la détermination des facteurs permettant d'expliquer, au moins en partie, l'évolution temporelle des concentrations en *Dinophysis* à ANTIFER. Les deux articles publiés reproduits dans la section suivante montrent les résultats de notre recherche. La dernière section de ce chapitre présente une application d'un modèle HARRISON-STEVENS.

4.3 Régression dynamique

La régression est un outil communément utilisé pour étudier les relations existantes entre une variable dépendante, ou à expliquer, et un ensemble de covariables indépendantes, ou explicatives. Elle présente par ailleurs l'avantage d'avoir connu une diffusion dans le domaine scientifique dépassant largement le cadre de la statistique. Structurellement, la régression inclut l'hypothèse d'invariance des relations entre la variable dépendante et les variables indépendantes. Des graphiques présentant l'évolution temporelle des concentrations en *Dinophysis* et d'une variable environnementale (e.g. insolation, nitrate), nous ont permis de supposer l'existence de périodes de corrélation et d'indépendance. Ces observations amenaient à :

- questionner l'hypothèse implicite d'invariance des relations de la régression ;
- étudier le comportement de la régression linéaire dynamique dans le contexte phytoplanctonique ;

– interpréter la variabilité des relations entre la variable dépendante et et les variables indépendantes ;

– évaluer l'intérêt de cette application.

Ces points ont été abordés dans les articles publiés présentés ci-après.

4.3.1 *Dynamic linear Bayesian models in phytoplankton ecology*

Les objectifs de cet article sont de présenter l'intérêt d'un modèle de régression linéaire dynamique en écologie phytoplanctonique et de donner les éléments nécessaires à sa mise en œuvre. Afin de favoriser l'introduction des modèles dynamiques bayésiens en écologie, le modèle à variances constantes est choisi pour sa simplicité. La description du modèle se limite à sa définition mathématique et à la procédure séquentielle d'estimation. Les données sont celles de l'année 1988. La sélection des covariables est basée sur des résultats de régression statique.

L'interprétation des paramètres dynamiques montre la capacité du modèle à prendre en compte plusieurs échelles de variabilité temporelle. La pertinence des résultats est soulignée par la mise en évidence de phénomènes d'intensités variables dans le temps. Par exemple, les courant de surface induits par le vent entraînent des accumulations et dispersions de masses d'eau plus ou moins riches en *Dinophysis*. Les difficultés liées à l'interprétation sont également abordées. Enfin, les perspectives offertes par l'utilisation des modèles dynamiques et leurs extensions sont discutées.

Published in Ecological Modelling 99 (1997) 161-169

Dynamic linear Bayesian models in phytoplankton ecology

D. Soudant [1] [a], B. Beliaeff [a] and G. Thomas [b]

[a] *IFREMER, B.P. 1105, 44311 Nantes cedex 03, France*

[b] *INSERM U. 444, 27 rue de Chaligny, 75571 Paris cedex 12, France*

Abstract

As phytoplankton time series show high variabilities which are generated by processes occurring at differing spatial and temporal scales, static regression models may not be adapted to the inherent complexity of the data. The aim of this paper was to consider the advantages of Bayesian dynamic models for phytoplankton time series. Dynamic models allow for time-varying influence of the covariates. The basic assumption is the existence of underlying and unobservable time series for the vector parameter whose distribution is sequentially estimated, allowing on-line analysis. The dynamic linear regression model (DLRM) is described and applied to a time series of the concentration of the marine toxic microalga *Dinophysis* cf. *acuminata*. The evolution in time of the regression parameters shows scales of influence in the environmental factors and provides a segmentation of the time series into significant and non-significant phases. In our application, physical factors accounted for most of the fluctuations in *Dinophysis* cf. *acuminata* concentrations. In particular, the wind parameter exhibited variations which could be interpreted as accumulation and dispersion phenomena. This kind of information may help in understanding the processes underlying the fluctuations in *Dinophysis* cf. *acuminata*

[1] Corresponding author.

concentrations, as long as a sensible interpretation can be found for the parameters evolution.

Keywords: Modelling, Bayesian, dynamic, time series, *Dinophysis*.

1 Introduction

Ecological time series show high variabilities. In phytoplankton ecology, Harris (1980) demonstrated the existence and importance of several temporal and spatial scales. For example, rate of growth changes can be observed on a day-scale; seasonal phytoplankton community changes on a month-scale; and long-term trends on a year-scale. Analysis of chlorophyll *a* series (e.g. Denman and Platt, 1976; Denman et al., 1977) has shown that (i) phytoplankton distribution is driven by turbulences at scales smaller than 1 km, (ii) biological processes overcome turbulent diffusion and create spatial heterogeneity from 1 to 5 km, and (iii) phytoplankton distribution is controlled by water motions (e.g. advection, eddies, upwelling) at scales larger than 5 km. According to Legendre and Demers (1984), this multiplicity of scales has its origin in hydrodynamic forces, considered as environmental factors transmitting variability to living organisms. Moreover, in a turbulent environment, horizontal and vertical diffusion induce strong interrelations between time and space (Boyce, 1974), so that these two dimensions cannot be separated. Therefore, variabilities in phytoplankton time series depend on both temporal and spatial scales.

Red tide occurrences, sometimes associated with shell toxicity, have recently provided sets of incriminated species-abundance time series. The low number of published results attest to the difficulty of analyzing such series. Ménesguen et al. (1990) developed a deterministic dynamic model which provides good results but cannot adjust to abrupt increases in microalgal concentrations. Beltrami and Cosper (1993) looked for chaotic dynamics in concentration fluctuations of a microalga causing red tides. Statistical methods such as static regression or discriminant analysis and time series analysis applied to these series proved unsatisfactory (Mer, unpublished results). These approaches suppose that relationships between the dependent variable and the independent covariates are constant over time. However this hypothesis has not been thoroughly investigated.

Bayesian dynamic models have been used successfully in different fields (West and Harrison, 1989). The aim of this paper was to consider their advantages for phytoplankton time series. These models may be considered as dynamic generalizations of linear models (West et al., 1985). The dynamic aspect resides in a formulation which includes a time-varying relationship hypothesis. We present the dynamic linear regression model (DLRM) and then apply this model

to *Dinophysis* cf. *acuminata* (toxic microalga) concentration time series in 1988 at Antifer, France. The analysis of parameter evolution provides an assessment of the relevance of the assumed time-varying influence of environmental factors.

2 Methods

Consider a time series $Y_t, t = 1, 2, \ldots$ and $\mathbf{F}'_t = (1\ X_{1,t}\ X_{2,t}\ \ldots\ X_{n,t})$ a time dependent vector of regressors. Static linear regression models assume that $Y_t = \mathbf{F}'_t\boldsymbol{\theta} + \nu_t$, where $\boldsymbol{\theta}' = (\theta_0\ \theta_1\ \theta_2\ \ldots\ \theta_n)$ is a vector of time independent parameter and ν_t are independent and identically distributed (iid) in the normal distribution $N[0, V]$. Dynamic models allow for dependence of $\boldsymbol{\theta}$ on time and model the evolution by assuming

$$Y_t = \mathbf{F}'_t\boldsymbol{\theta}_t + \nu_t, \nu_t \overset{iid}{\sim} N[0, V], \tag{1}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim N[\mathbf{0}, \mathbf{W}_t], \tag{2}$$

where equation 1 is the observation equation and ν_t is the observational error, and equation 2 is the evolution equation and $\boldsymbol{\omega}_t$ is the evolutional error. (The notation $\sim$ is used here and elsewhere to denote "distributed as".) This formulation implies that the parameter vector $\boldsymbol{\theta}_t$ is a random variable. To fully specify the model we define the initial distribution $(\boldsymbol{\theta}_0|D_0) \sim N(\mathbf{m}_0, \mathbf{C}_0)$, where $\mathbf{m}_0$ and $\mathbf{C}_0$ are fixed. Moreover, the error sequences ν_t and $\boldsymbol{\omega}_t$ are independent, mutually independent, and independent of $(\boldsymbol{\theta}_0|D_0)$. D_0 is the initial information set, representing all the available relevant information used to specify the model before the first observation. Another representation of equations (1) and (2) is

$$(Y_t|\boldsymbol{\theta}_t, D_{t-1}) \sim N[\mathbf{F}'_t\boldsymbol{\theta}_t, V], \tag{3}$$

$$(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, D_{t-1}) \sim N[\boldsymbol{\theta}_{t-1}, \mathbf{W}_t], \tag{4}$$

where $D_{t-1} = \{D_{t-2}, Y_{t-1}\}$ is the information set at $t-1$.

The parameter vector is sequentially estimated. At $t-1$, let $(\boldsymbol{\theta}_{t-1}|D_{t-1})$, ditributed as $N[\mathbf{m}_{t-1}, \mathbf{C}_{t-1}]$, be the posterior distribution of $\boldsymbol{\theta}_{t-1}$. The estimation steps are the following (West and Harrison, 1989):

- Prior distribution

  Equation (2) allows us to compute the prior distibution $(\boldsymbol{\theta}_t|D_{t-1}) \sim N[\mathbf{a}_t, \mathbf{R}_t]$, where $\mathbf{a}_t = \mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{C}_{t-1} + \mathbf{W}_t$.

- Forecast

  Equation (1) gives the forecast distribution $(Y_t|D_{t-1}) \sim N[f_t, Q_t]$, where $f_t = \mathbf{F}'_t\mathbf{a}_t$ and $Q_t = \mathbf{F}'_t\mathbf{R}_t\mathbf{F}_t + V$. From equation (3) we derive the joint distribution

$$\begin{pmatrix} Y_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| D_{t-1} \sim N\left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} Q_t & Q_t\mathbf{A}'_t \\ \mathbf{A}_tQ_t & \mathbf{R}_t \end{pmatrix}\right],$$

where $\mathbf{A}_t = \mathbf{R}_t \mathbf{F}_t Q_t^{-1}$.

- Posterior distribution

  The prior density is updated to give the posterior density from the joint density: $(\boldsymbol{\theta}_t | D_t) \sim N[\mathbf{m}_t, \mathbf{C}_t]$, where $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t e_t$ and $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t \mathbf{A}_t' Q_t$ with the forecast error $e_t = Y_t - f_t$.

The distribution $(Y_t | D_t)$ is normal, with posterior mean value $\mathbf{F}_t' \mathbf{m}_t$ and variance $\mathbf{F}_t' \mathbf{C}_t \mathbf{F}_t + V$. Statistical inference on dynamic parameters at each time t is performed as in static regression, with parameters non-significantly different from zero if zero is included in the parameter confidence intervals at an α significance level. As DLRM is dynamic, parameter confidence intervals are replaced by time series of parameter confidence intervals.

The performances of the DLRM depend heavily on choosing appropriate values for the variance matrices $\mathbf{W}_t$. West and Harrison (1989, pp. 190-191) suggest specifying $\mathbf{W}_t$ as $\mathbf{W}_t = \mathbf{C}_{t-1}(1-\delta)/\delta$, where $0 < \delta < 1$. δ is the so-called discount factor. Thus, the prior distribution variance matrix of $\boldsymbol{\theta}_t$ is $\mathbf{R}_t = \mathbf{C}_{t-1}/\delta$. Let δ_0 be the "true" value of the discount factor: then if $\delta > \delta_0$ the model is under-adaptive and errors are positively correlated; and if $\delta < \delta_0$ the model is over-adaptive and errors are negatively correlated. Thus, error analysis provides information about the specified discount factor. In practice, West and Harrison (1989, pp. 280) recommend $\mathbf{W}_t$ matrices with small value elements compared to $\mathbf{C}_t$ elements.

Computer programs developed in C on a SUN station are available from the first author.

3 Application to the Antifer series

3.1 Overview of the data

Dinophysis cf. *acuminata* (hereafter refered to as *Dinophysis*) is a toxic microalga which is considered responsible for some diarrheic epidemics. This phytoplankton species has been extensively studied. However, many features of its biology and ecology remain largely unknown (Delmas et al., 1993; Sampayo, 1993; Berland et al., 1995a, 1995b; Maestrini et al., 1996). To study the environmental conditions that may lead to *Dinophysis* blooms, seawater was sampled daily for 3 to 4 months during the summers of 1987 to 1993 at a one-meter depth at the end of the petroleum wharf in Antifer harbour (Fig. 1). The data used here were those for 1988. Figure 2 shows time series of the dependent variable and the independent covariates.

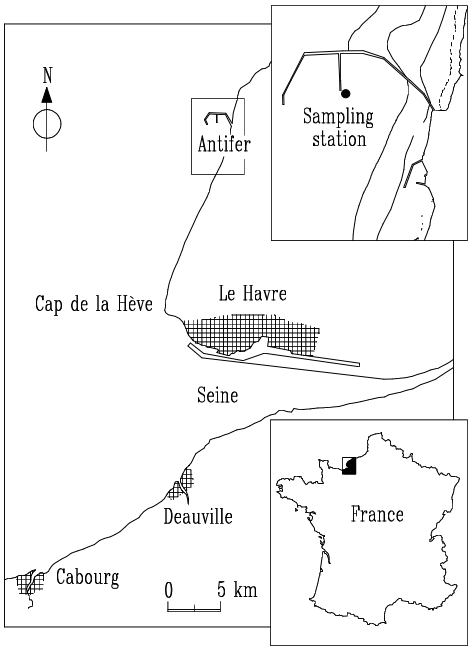

Figure 1: *Geographical location of the Antifer site.*

3.2 Dynamic Linear Regression Model

The dependent variable is $\log(Z_t + 1)$, where Z_t is the *Dinophysis* concentration (cell.(10 mL)$^{-1}$) at day t. The dependent variable and the covariates were standardized to zero mean and unit variance. An initial set of covariates was selected from static regression models, and other covariates were entered stepwise in the correspondent dynamic model. The final DLRM had the following observation equation at time t, $t = 1, 2, \ldots$,

$$Y_t = \theta_{I,t} I_t + \theta_{P,t} P_t + \theta_{T,t} T_t + \theta_{S,t} S_t + \theta_{SW,t} SW_t + \nu_t,$$

where $\nu_t \overset{iid}{\sim} N[0, V]$, and S_t represents Seine flow (m^3.s^{-1}), T_t water temperature (°C), I_t insolation (hour/day), P_t phosphate concentrations (μmol.l^{-1}) and SW_t south-west wind (m.s^{-1}). The wind covariate was computed as the daily mean of eight determinations per day of the variable $[-\alpha_{it} \cos(\beta_{it} - \pi/4)]$, where α_{it} and β_{it} are respectively the speed and direction of the wind on the i^{th} determination of day t. This representation induces a continuous decrease from south-west to north-east, with a zero value for north-west and south-east winds. The constant observation variance V was estimated from the data via an algorithm described by West and Harrison (1989; pp. 117-121). The estimated value was 0.3817; δ was chosen equal to 0.95. We noted that running successive DLRM on the studied

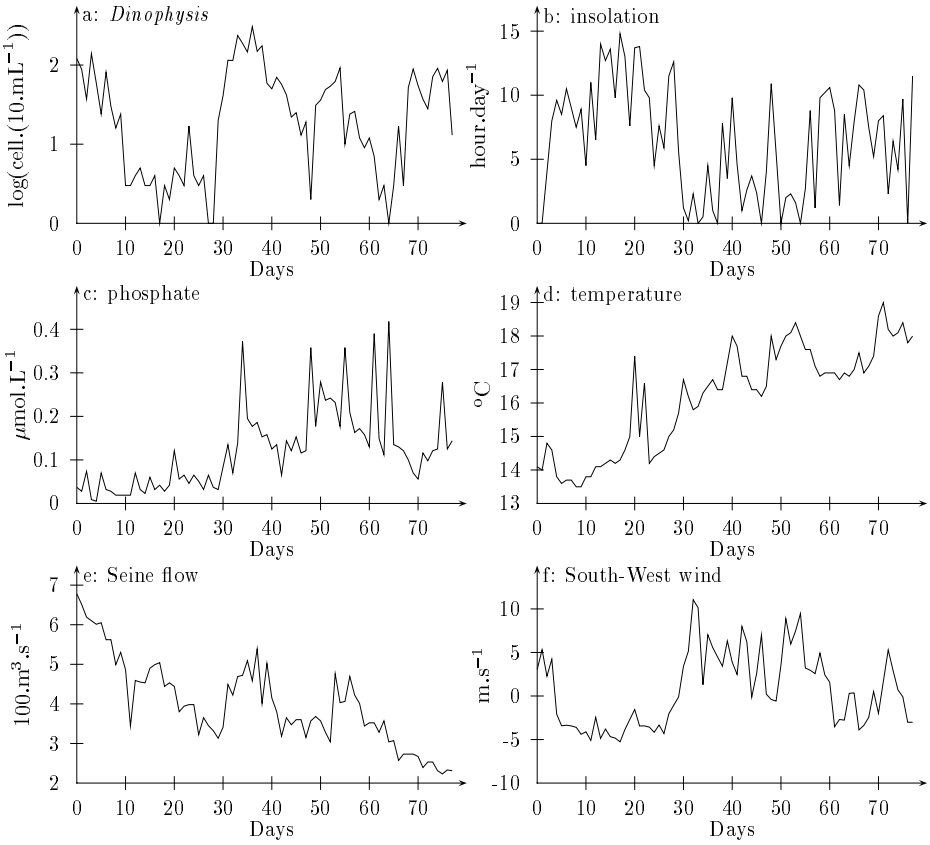

Figure 2: *Observed values of (a):* Dinophysis *concentration, (b): insolation, (c): phosphate concentration, (d): temperature, (e): Seine flow and (f): South-West wind. Time origin is 1 June 1988.*

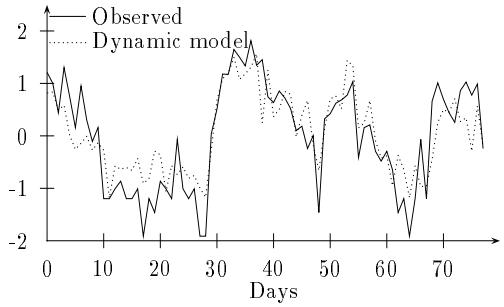

Figure 3: *Observed and DLRM posterior values of* Dinophysis *concentrations at Antifer as from 1 June 1988.*

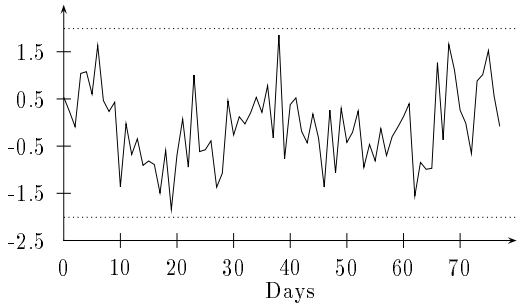

Figure 4: *Standardized residuals of DLRM posterior values and 95% confidence limits as from 1 June 1988.*

time series led to an "optimal" parameter time series. The last vector parameter estimate of the last run was used as the initial condition, $(\boldsymbol{\theta}_0|D_0) \sim N(\mathbf{m}_0, \mathbf{C}_0)$.

3.3 Results

Figure 3 shows DLRM posterior mean value time series. The runs test (Siegel, 1956) was applied to the standardized error $(e_t/\sqrt{Q_t})$ time series (Fig. 4). A random distribution hypothesis cannot be rejected at the $\alpha = 0.05$ level for DLRM posterior value errors, suggesting that the chosen discount factor was satisfactory (see "Methods"). Posterior parameters showed high temporal variations (Fig. 5). Insolation and phosphate parameters (resp. $\theta_{I,t}$ and $\theta_{P,t}$) were generally negative, and those for temperature, Seine flow and south-west wind (resp. $\theta_{T,t}$, $\theta_{S,t}$ and $\theta_{SW,t}$) were positive. From day 10 to day 60, $\theta_{T,t}$ and $\theta_{S,t}$ showed parallel fluctuations and $\theta_{SW,t}$ was negatively correlated with these parameters.

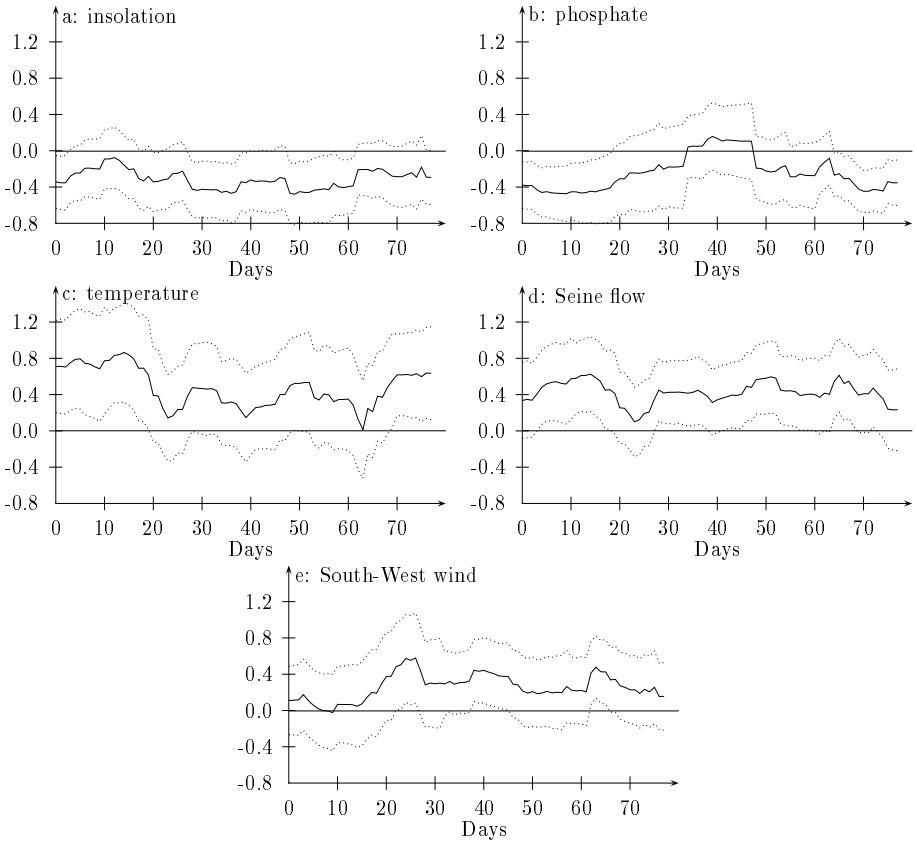

Figure 5: *DLRM posterior parameters and 90% confidence limits of (a): insolation, (b): phosphate, (c): temperature, (d): Seine flow, (e): South-West wind. Time origin is 1 June 1988.*

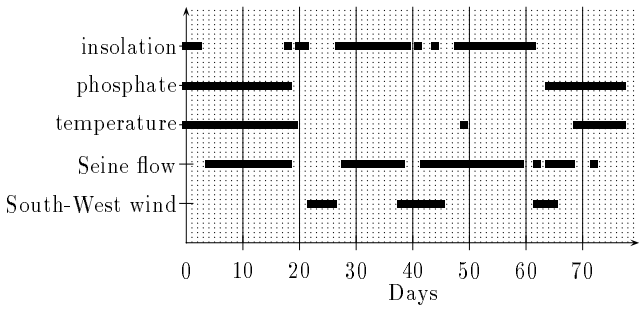

Figure 6: *Significance of parameter as a function of time. Black boxes denote days when the parameter is significantly different from zero. Time origin is 1 June 1988.*

Figure 6 shows days where parameters were significantly different from zero at the $\alpha = 0.1$ level. Temperature and phosphate parameters are significant at the beginning and end of the time series. Conversely, insolation and south-west wind parameters are significant in the middle of the time series.

4 Discussion

When regression parameters were allowed to evolve with time, we observed posterior mean values close to the observations (Fig. 3). The use of dynamic models can be justified if the time-varying influence of factors is reasonable and if the state of knowledge ensures a sensible interpretation of parameter evolutions. Phosphates are nutrition sources for algae. In our application, phosphate parameter interpretation is difficult because of the biological "enigma" posed by *Dinophysis*, especially with respect to its nutrition mode(s) (Maestrini et al., 1996). Moreover, $\theta_{P,t}$ was generally negative but non-significantly different from zero (Fig. 5b), and no particular explanation could be found when it was significant. The significant intervals of the insolation parameter (Fig. 5a) correspond to low insolation values and high concentrations of the toxic microalga (Fig. 2a and b). An explanation for this could come from the ability of *Dinophysis* to move vertically in the water column (Lewandowski and Kaneta, 1987). Thus, high *Dinophysis* abundance values could be the result of the conjunction of sampling at some given depth and time, and low insolation values. Temperature and Seine flow parameters were correlated (Fig. 5c and d). Increasing flow may have induced movements of water masses with different physical and chemical characteristics, as water temperature. Occurrence of *Dinophysis* is frequently associated with persistent stratification (e.g. Delmas et al., 1992). Significant intervals of temperature and Seine flow parameter may reveal the passage of water masses with

perature and Seine flow parameter may reveal the passage of water masses with different *Dinophysis* concentrations. Consequently fluctuations in *Dinophysis* concentrations at Antifer depend more on spatial variability phenomena than on a day-to-day biological variability. The first and last time intervals when $\theta_{SW,t}$ was significant occurred with north-east wind and low *Dinophysis* concentration values, and the middle time interval with south-west wind and high *Dinophysis* concentrations (Fig. 2a, f and 5e). The geographical situation of the sampling site (Fig. 1) suggests that south-west (resp. north-east) wind induced *Dinophysis* cell accumulation (resp. dispersion) phenomena along the coast, particularly in Antifer harbour. The degree of accumulation (resp. dispersion) depended on both wind speed and *Dinophysis* concentration. Thus, wind influence can be considered time-varying on a day scale. Dynamic models actually allow such behaviour to be detected (Fig. 5e). Finally, physical factors accounted for most of the fluctuations in *Dinophysis* concentrations. Moreover, these results suggest that dynamic models can detect the time-varying influences of environmental factors. As changes in parameters depend directly on relational fluctuations between dependent and independent variables, the scales of influence revealed by changes in the parameters can be considered to correspond to those of the covariates.

Various extensions of DLRM could improve model performances. In particular, the counting process and spatial distribution of plants and animals induce variance-to-mean relationships (Taylor, 1961; Frontier, 1972; Kendall, 1995), which in turn, induce heteroscedasticity in the time series. The $\log(x+1)$ transformation that we perform on *Dinophysis* concentration can only provide an approximation to the homoscedastic hypothesis. However dynamic models can be adapted to time series with time-varying observational variance (West and Harrison, 1989; pp. 368-373). This makes dynamic models particulary flexible tools for ecological time series. Besides, sequential estimation of the parameter vector may be generalized at time $t+k$. Thus, time series with missing data in particular can be analyzed. As $\mathbf{F}_t$ is used at the forecast step, the estimate at t is not a prediction, strictly speaking. However joint dynamic models may give covariate predictions. These forecasts or those from other models (e.g. meteorological, hydrodynamical) can be used instead of the regression vector, thus providing true predictions. Morevover, these dynamic joint models can deal with missing covariate data. A corresponding general tool is the dynamic multivariate linear regression model. True predictions can also be obtained using lagged covariates. Autoregressions fall into this type of application. Eventually, the forecaster may modify the parameters directly, e.g. anticipating and incorporating exceptional events in the form of covariates which have an influence at a given threshold. The reader familiar with signal processing methods may have noted a formal likeness between dynamic models and the Kalman filter (Kalman, 1960). Meinhold and Singpurwalla (1983) have shown that the Kalman filter can be considered as a Bayesian method. However, the generalized dynamic model is not based upon the Kalman filter (Harrison and Stevens, 1976; West et al.,

1985). Yet these considerations are well beyond the scope of this paper. In their most general form, dynamic models can be used with non-stationary, non-linear, non-normal time series including those with missing data.

Dynamic Bayesian models can be useful for an understanding of ecological processes. This is illustrated in this work by their ability to detect different influence scales of environmental factors upon phytoplankton dynamics. This approach may satisfy the need for new concepts and methods, as noted in the paper of Harris (1980) and Legendre and Demers (1984). Extensions of these models would allow a more adapted statistical approach to temporal and spatial scales, particularly for ecological characteristics such as non-linearity and non-normality.

However, the scales detected with the dynamic model depend on sampling frequency, and the distinction between spatial and temporal components relies on biological and/or hydrodynamical interpretation. Therefore, studies of spatio-temporal structure are central to understanding the dynamics of toxic phytoplankton species.

Acknowledgements — We are very grateful to P. Gros for critical review of this manuscript. We thank P. Lassus for providing the data and for valuable discussions. L. Giboire is acknowledged for figures.

References

Beltrami, E. and Cosper, E., 1993. Modeling the temporal dynamics of unusual blooms. In: T.J. Smayda and Y. Shimizu (Editors), Toxic phytoplankton blooms in the sea, Elsevier Science, Amsterdam, pp. 731-735.

Berland, B.R., Maestrini, S.Y. and Grzebyk D., 1995a. Observations on possible life cycle stages of the dinoflagellates *Dinophysis* cf. *acuminata*, *Dinophysis acuta*, *Dinophysis pavillardi*. Aquat. microb. Ecol., 9: 183-189.

Berland, B.R., Maestrini, S.Y., Grzebyk, D. and Thomas P., 1995b. Recent aspects of nutrition in the dinoflagellate *Dinophysis* cf. *acuminata*. Aquat. microb. Ecol., 9: 191-198.

Boyce, F.M., 1974. Some aspects of Great Lakes physics of importance to biological and chemical processes. J. Fish. Res. Board Can., 31: 689-730.

Delmas, D., Herbland, A. and Maestrini, S.Y., 1992. Environmental conditions which lead to increase in cell density of the toxic dinoflagellates *Dinophysis* spp. in nutrient-rich and nutrient-poor waters of the French Atlantic coast. Mar. Ecol. Progr. Ser., 89: 53-61.

Delmas, D., Herbland, A. and Maestrini, S.Y., 1993. Do *Dinophysis* spp. come from the open sea along the French Atlantic coast? In: T.J. Smayda and Y. Shimizu (Editors), Toxic phytoplankton blooms in the sea, Elsevier Science, Amsterdam, pp. 489-494.

Denman, K.L., Okubo, A. and Platt, T., 1977. The chlorophyll fluctuation spectrum in the sea. Limnol. Oceanogr., 22: 1033-1038.

Denman, K.L. and Platt, T., 1976. The variance spectrum of phytoplankton in a turbulent ocean. J. Mar. Res., 34: 593-601.

Frontier, S., 1972. Calcul de l'erreur sur un comptage de zooplancton. J. exp. mar. Biol. Ecol., 8: 121-132.

Harris, G.P., 1980. Temporal and spatial scales in phytoplankton ecology. Mechanisms, methods, models and management. Can. J. Fish. Aquat. Sci., 37: 877-900.

Harrison, P.J. and Stevens, C.F., 1976. Bayesian forecasting (with discussion). J. Roy. Statist. Soc. (Ser. B), 38: 205-247.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. of Basic Eng., 82: 34-45.

Kendall, W.S., 1995. A probabilistic model for the variance to mean power law in ecology. Ecol. Model., 80: 293-297.

Legendre, L. and Demers, S., 1984. Towards dynamic biological oceanography and limnology. Can. J. Fish. Aquat. Sci., 41: 2-19.

Levandowski, M. and Kaneta, P.J., 1987. Behaviour in dinoflagellates. In: F.J.R. Taylor (Editor), The biology of dinoflagellates, Blackwell Scientific Publications, Oxford, pp. 360-398.

Maestrini, S.Y., Berland, B.R., Carlsson, P., Granéli, E. and Pastoureaud, A., 1996. Recent advances in the biology of the toxic dinoflagellate genus *Dinophysis*: In Y. Yasumoto, Oshima Y. and Fukuyo Y. (Editors), Harmful algal blooms, Intergovernmental oceanographic commission of unesco, pp. 397-400.

Meinhold, R.J. and Singpurwalla, N.D., 1983. Understanding the Kalman filter. Am. Stat., 37: 123-127.

Ménesguen, A., Lassus, P., De Cremoux, F. and Boutibonnes, L., 1990. Modelling *Dinophysis* blooms: a first approach. In: E. Granèli, B. Sundström, L. Edler and D.M. Anderson (Editors), Toxic Marine Phytoplankton, Elsevier Science publishers, Amsterdam, pp. 195-200.

Sampayo, M.A. de M., 1993. Trying to cultivate *Dinophysis* spp.. In: T.J. Smayda and Y. Shimizu (Editors), Toxic phytoplankton blooms in the sea, Elsevier Science, Amsterdam, pp. 807-810.

Siegel, S., 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill series in psychology, McGraw-Hill, New York, 312 pp.

Taylor, L.R., 1961. Aggregation, variance and the mean. Nature, 189: 732-735.

West, M. and Harrison, P.J., 1989. Bayesian forecasting and dynamic models. Springer series in statistics, Springer-Verlag, New York, 704 pp.

West, M., Harrison, P.J. and Migon, H.S., 1985. Dynamic generalized linear models and Bayesian forecasting (with discussion). J. Am. Stat. Assoc., 80: 73-97.

4.3.2 *Explaining* Dinophysis *cf.* acuminata *abundance in Antifer (Normandy, France) using dynamic linear regression*

Ce second article est plus « appliqué » que le précédent. Le but poursuivi est de trouver des éléments d'explication aux fortes variabilités de concentration en *Dinophysis*. Le modèle à variance d'observation constante et inconnue est présenté simplement. Les détails techniques sont donnés en annexe. La procédure séquentielle d'estimation est illustrée par un exemple sur une série temporelle simulée. Les données utilisées dans l'application sont celles de 1989 et 1990. En supposant que les phénomènes soient identiques d'une année à l'autre, un modèle unique est construit à partir des deux ensemble de données. Les trois covariables « vent de sud-ouest », « salinité » et « coefficient de marée » ont été séléctionnées au regard d'un critère statistique. Les conditions initiales sont spécifiées arbitrairement. Des versions statiques du modèle identifié sont réalisées afin d'effectuer des comparaisons avec les modèles dynamiques.

Les résultats montrent le caractère dynamique de l'influence des trois covariables. L'effet du vent est confirmé. Des données physiques soutiennent cette hypothèse ainsi que celle de l'effet de la marée sur le mouvement des masses d'eau plus ou moins riches en *Dinophysis* et présentes dans le port d'Antifer. Ces éléments permettent d'établir un schéma général d'explication des variations de la concentration en *Dinophysis*. Les pourcentages de variation expliquée par les modèles dynamiques sont supérieurs à ceux des modèles statiques. De plus, ces derniers ne mettent pas évidence l'effet de la covariable « coefficient de marée ».

Published in Marine Ecology Progress Series 156 (1997) 67-74

Explaining *Dinophysis* cf. *acuminata* abundance in Antifer (Normandy, France) using dynamic linear regression

D. Soudant[1,*], B. Beliaeff[1], G. Thomas[2]

[1] IFREMER, B.P. 21105, F-44311 Nantes cedex 03, France

[2] INSERM U. 444, 27 rue de Chaligny, F-75571 Paris cedex 12, France

ABSTRACT: Classical regression analysis can be used to model time series. However, the assumption that model parameters are constant over time is not necessarily adapted to the data. In phytoplankton ecology, the relevance of time-varying parameter values has been shown using a dynamic linear regression model (DLRM). DLRMs, belonging to the class of Bayesian dynamic models, assume the existence of a non-observable time series of model parameters, which are estimated on-line. The aim of this paper was to show how DLRM results could be used to explain variation of a time series of phytoplankton abundance. We applied DLRM to daily concentrations of *Dinophysis* cf. *acuminata*, determined in Antifer harbour (French coast of the English Channel), along with physical and chemical covariates (e.g. wind velocity, nutrient concentrations). A single model was built using 1989 and 1990 data, and then applied separately to each year. Equivalent static regression models were investigated for the purpose of comparison. Results showed that most of the *Dinophysis* cf. *acuminata* concentration variability was explained by the configuration of the sampling site, the wind regime and tide residual flow. Moreover, the relationships of these factors with the concentration of the microalga

*E-mail: dominique.soudant@ifremer.fr

varied with time, a fact that could not be detected with static regression. Application of dynamic models to phytoplankton time series, especially in a monitoring context, is discussed.

KEY WORDS: Phytoplankton · *Dinophysis* · Time series · Regression · Dynamic · Bayesian

INTRODUCTION

To investigate potential relationships between a set of covariates and some observed process, time series are commonly modelled using regression analysis. Regression constant parameters are estimated from the whole data set, assuming constant relationships over time between the dependent variable and covariates. However, these relationships may, in reality, vary over time. For example, the influence of a given covariate can be highly significant during a certain time interval and non-significant the rest of the time. Alternatively, the influence can be significant over the whole time period but subject to large variations. In the first case, the covariate parameter value will be underestimated and thus found non significant; in the second case, large variations will inflate the variance of the estimator and may lead to a conclusion of to non-significance of the covariate. Thus, in classical (i.e. static) regression analysis, dynamic relationships between dependent and independent variables cannot be properly taken into account.

Dynamic Linear Regression Models (DLRMs) belong to the class of Bayesian dynamic models which assume time-varying relationships. The parameters are allowed to evolve with time, and thus the model is adaptable because the values of the estimated parameters and the set of significant covariates may change with time. Dynamic models have been successfully used in the social and economic fields (Pole et al. 1994, West & Harrison 1997). In previous work, we applied a DLRM to the 1988 *Dinophysis* cf. *acuminata* (toxic microalga) time series at Antifer (Soudant et al. 1997). High variabilities of parameters of physical and chemical covariates (e.g. insolation, phosphate) were detected. These results illustrated the relevance of the time-varying influence assumption in phytoplankton ecology.

The aim of this paper was to show how DLRM results can be used to point out some factors explaining, at least in part, the evolution of *Dinophysis* cf. *acuminata* concentrations at Antifer using the 1989 and 1990 time series. Assuming that the same processes determined concentrations of the toxic microalga during both years, a single model was built and applied separately to each data set. Particular attention was given to the adaptability of dynamic models, which allows changes in the set of significant covariates. Static regressions were performed to draw comparisons with DLRM results. Lastly, advantages of Bayesian dynamic models and their extensions are discussed in a monitoring context.

METHODS

Only the general principles of DLRMs are described hereafter. Readers interested in the mathematical elaboration may refer to the appendix and to more specialized papers (West et al. 1985, Pole et al. 1994, West & Harrison 1997).

Let $Y_t, t = 1, 2, \ldots$, denote the dependent variable at time t, and $X_t = \{X_{1,t}, X_{2,t}, \ldots, X_{n,t}\}$, $t = 1, 2, \ldots$, a set of n independent variables, or covariates, measured concomitantly. In the 'static' linear regression model, the dependent variable is related to covariates by assuming

$$Y_t = \theta_0 + \sum_{i=1}^{n} \theta_i X_{i,t} + \epsilon_t,$$

where θ_0 is the intercept, θ_i is the parameter of the ith covariate and ϵ_t, the so-called noise or error term, is a random component, with ϵ_t, $t = 1, 2, \ldots$, independently identically distributed in the normal distribution with mean 0 and variance V. A DLRM assumes a time-varying relationship by allowing covariate parameters to vary with time. Let $\theta_{i,t}$ denote the parameter of the ith covariate at time t. In a DLRM, the regression equation has the form

$$Y_t = \theta_{0,t} + \sum_{i=1}^{n} \theta_{i,t} X_{i,t} + \epsilon_t, \tag{1}$$

where $\theta_{0,t}$ is the dynamic level, i.e. a time-varying intercept. Eq. (1) is the observation equation. Let $\boldsymbol{\theta}'_t = (\theta_{0,t}\ \theta_{1,t}\ \ldots\ \theta_{n,t})$ be the parameter vector. The evolution in time of parameters is modelled as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \tag{2}$$

where $\boldsymbol{\omega}_t$ is an error term with mean $\mathbf{0}$ and variance $\boldsymbol{W}_t$. Eq. (2) is called the evolution equation. As $\boldsymbol{\omega}_t$ is a random variable, Eq. (2) shows that the parameter vector $\boldsymbol{\theta}_t$ is itself a random variable.

The estimated values $\hat{Y}_t$ and $\hat{\boldsymbol{\theta}}_t$ are the respective means of the estimated distributions of the random variables Y_t and $\boldsymbol{\theta}_t$. The parameters of these distributions are estimated sequentially. The following simple artificial example presents the sequential estimation procedure (Fig. 1). Observations were generated with a single covariate, X_t, and without a dynamic intercept as $Y_t = \theta_t X_t + \epsilon_t$, where ϵ_t is an error term (Table 1). The values of X_t and θ_t were chosen and ϵ_t was simulated in the normal distribution with mean 0 and variance 1. The observation equation of the model was $Y_t = \theta_t X_t + \epsilon_t$, i.e. the same as that used to generate the data. The evolution equation was $\theta_t = \theta_{t-1} + \omega_t$, where ω_t is an error term with mean 0 and variance W_t. The procedure estimates a succession of distributions prior to, and posterior to, the current observation. It begins at $t - 1$, with the distribution of θ_{t-1} posterior to the observation of Y_{t-1}. The parameters of this

Table 1: *Simulated data for sequential estimation procedure example.*

Time:	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$
X_t	0	1	2	3	3	2	1	0
θ_t	1	1	1	1	0.5	0.5	0.5	0.5
Y_t	0.00	0.89	2.11	2.91	1.48	0.92	0.39	0.05

distribution are computed after Y_{t-1} has been actually observed. The evolution equation, adding the random variables θ_{t-1} and ω_t, gives the distribution of θ_t prior to the observation of Y_t. The mean and the variance of this distribution are equal to the mean and the variance of the distribution of θ_{t-1} posterior to Y_{t-1}, with W_t added to the variance, as ω_t is centered on 0 and has variance W_t. Thus, the evolution equation implies that $\hat{\theta}_t$ prior to time t is equal to $\hat{\theta}_{t-1}$ posterior to time $t-1$ (Fig. 1A), but with increased uncertainty. The observation equation gives the observation distribution prior to time t, that is before Y_t is actually observed. This is the forecast distribution giving the forecast estimate (Fig. 1B). Then, the actual value of Y_t is observed. With this new information, the distribution of θ_t prior to Y_t is updated giving the distribution of θ_t posterior to Y_t. In our example, this update induced a decrease from 0.98 to 0.68 of the estimated value of the covariate parameter (Fig. 1A), as the actual value was decreased from 1 to 0.5 (Table 1). The estimation of the distribution of θ_t posterior to the observation of Y_t allows a new iteration of the sequential procedure. Beside the procedure, parameters of the posterior forecast distribution are computed using Eq. (1). This distribution gives the on-line fitted value (Fig. 1B).

APPLICATION TO THE ANTIFER TIME SERIES

Data collection

Dinophysis cf. *acuminata* (hereafter referred to as *Dinophysis*) is a microalga producing diarrhetic shellfish poisoning. Despite many studies, some features of its biology and ecology remain largely unknown (Delmas et al. 1993, Sampayo 1993, Berland et al. 1995a, b, Maestrini et al. 1996). As high *Dinophysis* concentrations were observed previously in Antifer harbour (France), sea water was sampled to study the ecological conditions of occurrence. Daily samples were taken at high tide at 1 m depth at the end of the petroleum wharf (Fig. 2) from 1 July to 13 September 1989 and from 1 June to 11 September 1990. The measurements carried out by the municipal laboratory of Le Havre were *Dinophysis* concentration (cells per 10 ml) (Utermöhl 1958), salinity (Beckman induction salinometer), temperature (°C) (Ponselle sonde), nitrate and phosphate

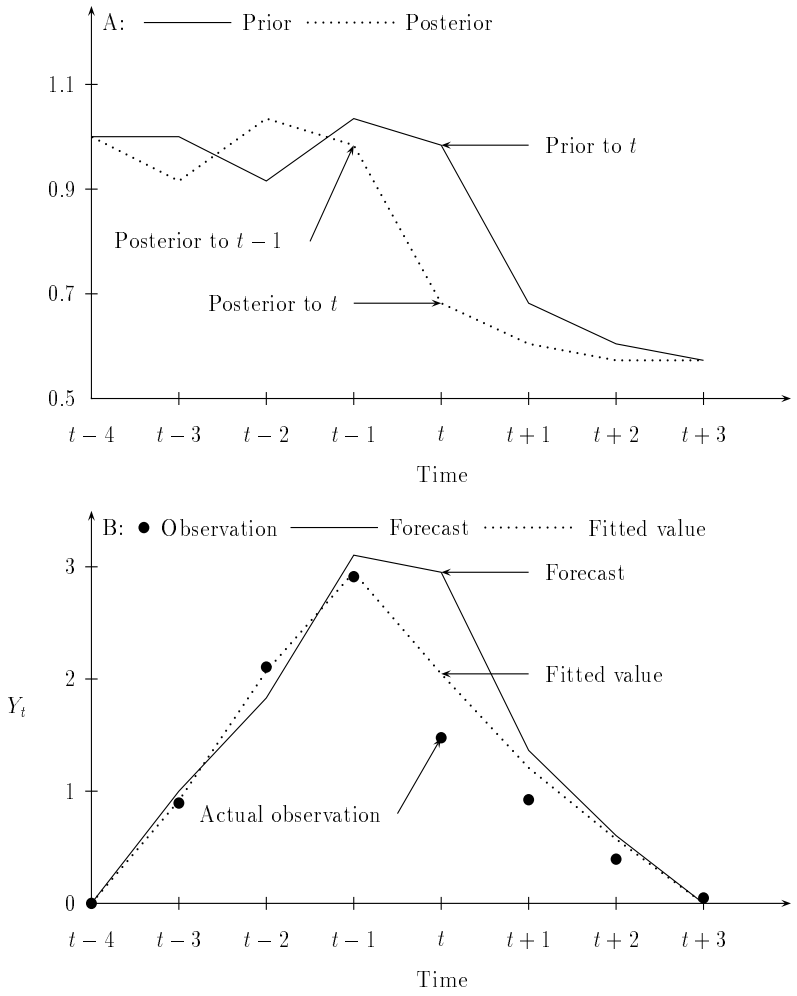

Fig. 1: *Sample DLRM: (A) prior and posterior estimated values of covariate parameter and (B) observation, DLRM forecast, and on-line fitted values.*

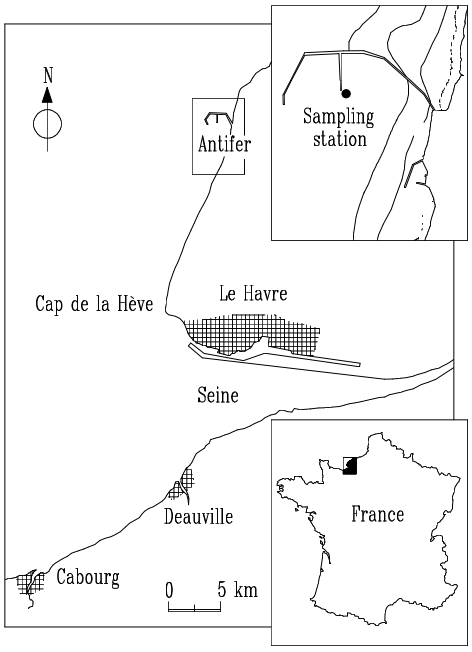

Fig. 2: *Sampling station, Antifer, France.*

concentration (μmol l^{-1}) (Technicon autoanalyser). Insolation (h d^{-1}), rainfall (mm d^{-1}), wind direction and speed (m s^{-1}) and Seine flow (m^3 s^{-1}) were obtained from the Le Havre weather station (Cap de la Hève, cf. Fig. 2). Tide coefficients at Le Havre harbour were drawn from tide tables puplished each year by the French hydrographic and oceanographic navy service (SHOM).

A 'South-West wind' covariate was computed as the daily mean of eight determinations per day of the variable $[-\alpha_{i,t}\cos(\beta_{i,t}-\pi/4)]$, where $\alpha_{i,t}$ and $\beta_{i,t}$ are, respectively, the speed and direction of the wind on the ith determination of day t. This gives a continuous decrease from South-West to North-East, with a zero value for North-West and South-East winds. Finally, the variables were standardized to zero mean, to separate clearly the covariate effects from the dynamic intercept, and unit variance, to allow comparisons between years and between covariate effects. It followed from this standardization that the dynamic intercept of the model at time t was the local mean of the dependent variable. Furthermore, the estimated regression parameters were adimensional.

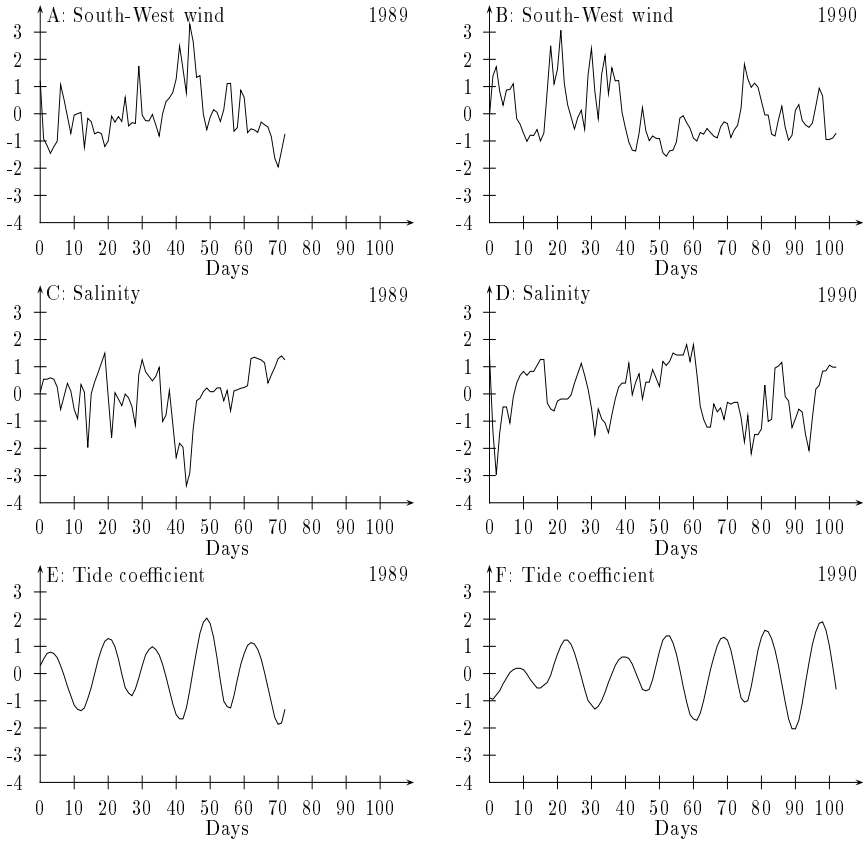

Fig. 3: *Evolution with time of the standardized variables selected as covariates in the model for 1989 and 1990. Day 0 is 1 July for 1989 and 1 forJune 1990.*

Dynamic Linear Regression Model

The dependent variable was $Y_t = \log(Z_t + 1)$, where Z_t was *Dinophysis* concentration on day t. For each year, a model was fitted. The model included a dynamic intercept, and covariates were selected one by one from the set of available variables (see previous section). At each step, the variable which induced the largest model likelihood with a significant gain in likelihood, as assessed using the likelihood ratio test (Kendall & Stuart 1977), was entered in the model. For 1989, only the 'South-West wind' covariate was selected while 'Salinity' and 'Tide coefficient' were selected for the 1990 series. Fig. 3 shows these variables. We used these 3 variables to explain the process underlying the evolution of *Dinophysis* concentration during the 2 years. Thus the final model, common to 1989 and 1990, had the following observation equation at time t, $t = 1, 2, \ldots$,

$$Y_t = \theta_{0,t} + \theta_{SW,t} SW_t + \theta_{S,t} S_t + \theta_{T,t} T_t + \epsilon_t,$$

where $\theta_{0,t}$ represents the dynamic intercept, SW_t 'South-West wind', S_t 'Salinity', T_t 'Tide coefficient' and ϵ_t is an error term. We decided to present results using covariate effects, that is the variable value (i.e. X_t) multiplied by the estimated regression parameter (i.e. $\hat{\theta}_{X,t}$). Confidence intervals at the $\alpha = 0.05$ level were used to test the nullity of the effects: when 0 was between the 2 limits, the effects were considered non-significantly different from 0. Finally, static versions of this model were fitted to data from both years in order to draw comparisons with DLRM results.

RESULTS

There was a succession of peaks of increasing magnitude in *Dinophysis* concentration in 1989 and 1990 (Fig. 4). On-line fitted values were similar to observed values for the 2 years. Figs. 5 & 6 show dynamic intercepts and effects of covariates. For both years, effects were not always significant. At the beginning of the series, and especially in 1990, 95% confidence intervals of effects were initially large and then decreased rapidly. This decrease in uncertainty, with the accumulation of observations and the alternation of time periods when the effects were significant and non-significant illustrated the adaptability of dynamic models.

In 1989 (Fig. 5), the dynamic intercept was significantly different from zero from Day 8 to 21 and from Day 57 to 64, and 'South-West wind' from Day 29 to 34 and from Day 37 to the end of the time series. As 'Salinity' was only significant the last Day of the time series and 'Tide coefficient' was never significant, dynamic intercept and 'South-West wind' effects mainly contributed to the on-line fitted values of the concentration of *Dinophysis*. The dynamic intercept showed a local low concentration of *Dinophysis* in the first interval and a high concentration in the second interval. $\hat{\theta}_{SW,t}$ was always positive. Positive effects corresponded to South-West winds and negative effects to North-East winds. The percentage R^2

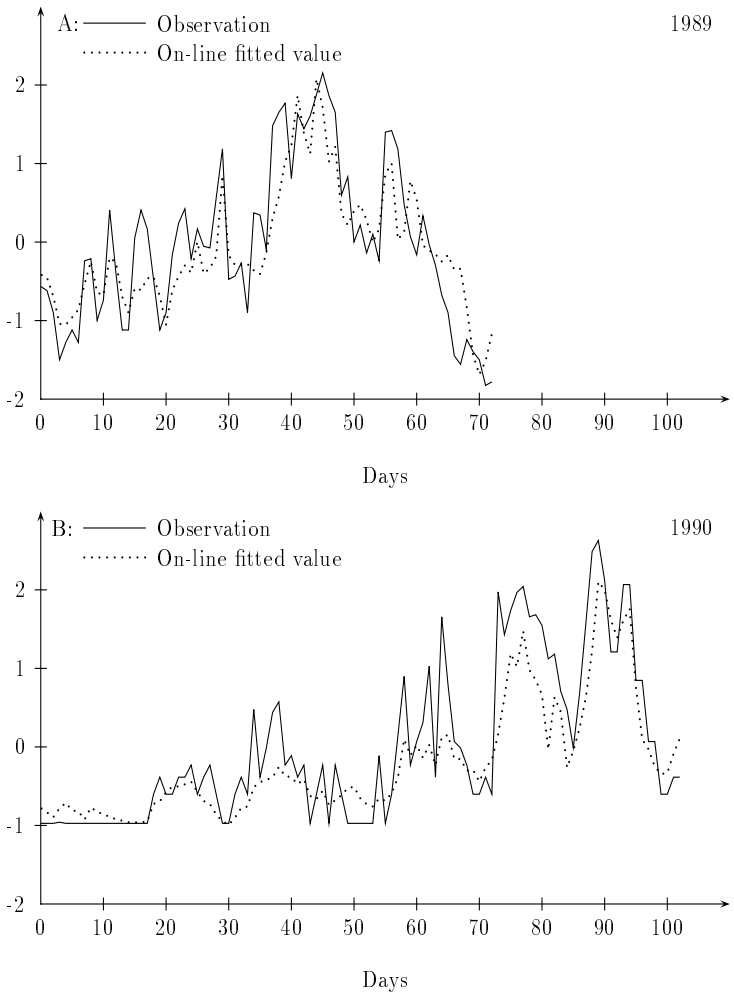

Fig. 4: *Standardized observations of* Dinophysis *concentrations at Antifer and their on-line fitted values. Day 0 is 1 July for 1989 and 1 forJune 1990.*

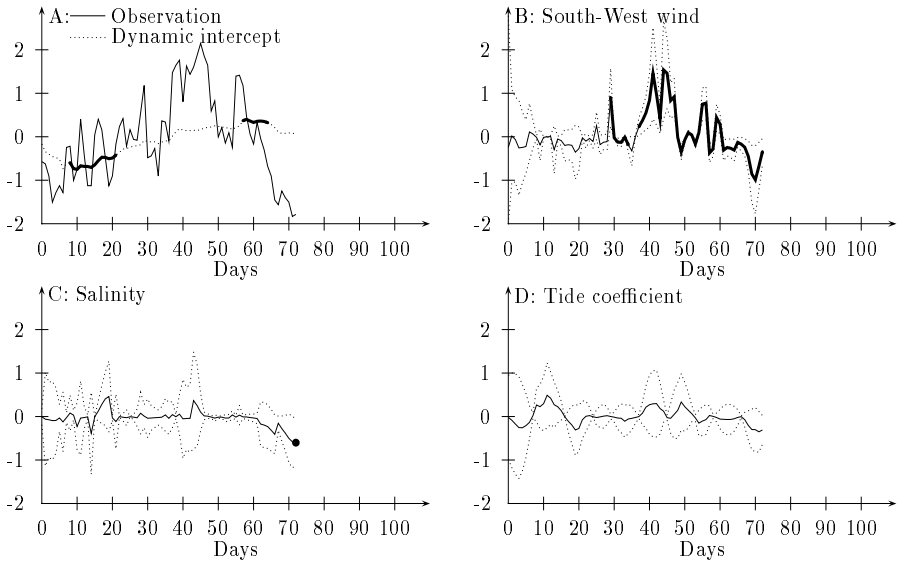

Fig. 5: *DLRM results for 1989. (A) Standardized observations of* Dinophysis *concentrations at Antifer and dynamic local means and (B, C,and D) effects and 95% confidence limits of covariates. Bold lines denote effects significantly different from 0. Day 0 is 1 July 1989. Note that effects are adimensional because of covariate standardization.*

Table 2: *Results of the static regression model for 1989 and 1990. N is sample size, R^2 the percentage of variation explained by the regression, $\hat{\theta}_0$ the estimated value of the intercept, $\hat{\theta}_{SW}$ that of the 'South-West wind' covariate, $\hat{\theta}_S$ that of 'Salinity', and $\hat{\theta}_T$ that of 'Tide coefficient'. ns: non-significantly different from zero at the $\alpha = 0.05$ level (bilateral test); p: significance level.*

	1989	1990
N	73	103
R^2	55.29%	23.79%
$\hat{\theta}_0$	-0.004 (ns)	-0.026 (ns)
$\hat{\theta}_{SW}$	0.621 ($p < 10^{-6}$)	-0.214 ($p = 0.039$)
$\hat{\theta}_S$	-0.213 (ns)	-0.514 ($p < 10^{-5}$)
$\hat{\theta}_T$	0.125 (ns)	-0.159 (ns)

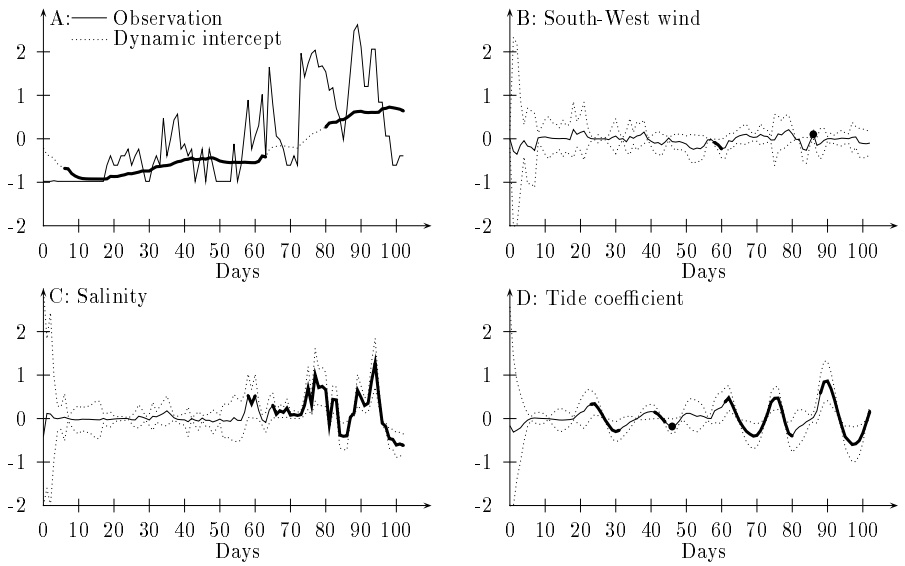

Fig. 6: *DLRM results for 1990. (A) Standardized observations of* Dinophysis *concentrations at Antifer and dynamic local means and (B, C,and D) effects and 95% confidence limits of covariates. Bold lines denote effects significantly different from 0. Day 0 is 1 June 1990. Note that effects are adimensional because of covariate standardization.*

(Draper & Smith 1966) of variation of *Dinophysis* concentration explained by the DLRM was 73.54% and that of static regression was 55.29% (Table 2). In the static regression, the intercept and the estimated values of the parameters of 'Tide coefficient' and 'Salinity' were not significantly different from zero. $\hat{\theta}_{SW,t}$ was highly significant and positive.

In 1990 (Fig. 6), the dynamic intercept was significantly different from zero and negative from Day 6 to 63 and significant and positive from Day 80 to the end of the time series. 'South-West wind' effects were significant from Day 58 to 60 and at Day 86. When the effects of this covariate were significant, its estimated parameter was positive. For 'Salinity', effects were significant from Day 58 to 60 with a positive estimated parameter and from Day 65 to the end of the time series with a negative estimated parameter. Lastly, 'Tide coefficient' effects were significant at Day 46 and during 4 intervals: Days 23 to 31, 41 to 44, 61 to 80 and Day 88 to the end of the time series. $\hat{\theta}_{T,t}$ was positive from the beginning of the time series to Day 56, and then negative to the end of the time series. Dynamic level, 'Salinity' and 'Tide coefficient' explained most of the evolution of *Dinophysis* concentrations. When $\hat{\theta}_{S,t}$ and $\hat{\theta}_{T,t}$ were negative, negative

values (respectivelypositive) of 'Salinity' and 'Tide coefficient' corresponded to positive (respectivelynegative) effects. The percentage of variation explained by the DLRM was $R^2 = 79.57\%$. The R^2 of the static regression was 23.79%. In the static regression, the intercept and the estimated parameter of 'Tide coefficient' were non-significantly different from zero. 'South-West wind' and 'Salinity' parameter estimations were significant and negative.

DISCUSSION

Covariate effects suggested different scenarios to explain *Dinophysis* concentration dynamics. In 1989, the geographical situation of the sampling site (Fig. 2), the location of phytoplankton maximum concentrations in the Seine plume (Ménesguen et al. 1995) and hydrodynamical studies of the Seine bay (e.g. Salomon & Breton 1993) suggested that South-West wind induced *Dinophysis* cell accumulation along the coast, particularly in Antifer harbour. Conversely, North-East wind could provoke cell dispersion. Such transportation phenomena induced by wind has already been observed in the Seine bay (Lagadeuc 1992, Thiébaut et al. 1994). By definition, the 'South-West wind' effect depended on the 'South-West wind' covariate value. 'South-West wind' effect varied also with *Dinophysis* concentration in the water mass subject to the accumulation/dispersion phenomena. As this concentration varied with time, the relationship between 'South-West wind' and the microalgal concentration in Antifer was time-varying. Only the covariates 'Salinity' and 'Tide coefficient' were significant in 1990. Significant negative values of the 'Salinity' estimated parameter suggested that lower surface salinity was accompanied by higher *Dinophysis* concentrations. This result was consistent with the association of *Dinophysis* occurrence with persistent salinity stratification (e.g. Delmas et al. 1992). The establishment of stratification is favoured by small tide coefficients. Greater tide coefficients may provoke water mixing and consequently a decrease of *Dinophysis* concentrations by dilution. A hydrodynamical study has shown that the configuration of Antifer harbour modifies the circulation of water masses (Monbet 1975), so that greater tide coefficients induced a departure of water masses to the North and small coefficients a 'capture' of water masses in the harbour. As for 'South-West wind', relationships between *Dinophysis* concentration and 'Salinity' and between *Dinophysis* concentration, and 'Tide coefficient' varied over time.

Although 'Salinity', 'Tide coefficient' and 'South-West wind' seemed to be important for understanding the evolution of *Dinophysis* concentration at Antifer, some discrepancies appeared between scenarios and results. The 3 covariates were never significant concomitantly. A natural explanation for this observation was related to interdependence among these variables. For example, correlation between 'South-West wind' and 'Salinity' covariates was negative and highly significant ($p < 10^{-7}$) both in 1989 and 1990. One of these 2 covariates may thus mask the influence of the other one. Such correlations seemed to be responsible

for the non-significance of the 'Tide coefficient' estimated parameter in 1989 and, in the static regression results, for the visible inversion of the absolute values of estimated parameters of 'South-West wind' and 'Salinity' for 1989 and 1990 and for the change in the sign of $\hat{\theta}_{SW,t}$ from 1989 to 1990. In 1990, results of the DLRM showed changes in the signs of the estimated parameters of 'Salinity' and 'Tide coefficient'. From our scenarios, these parameters were expected to be negative. The positive parameter values resulted from local positive correlations between the values of *Dinophysis* concentration and 'Salinity' and between those for *Dinophysis* concentration and 'Tide coefficient'. There were only 3 days when the estimated parameter of 'Salinity' was positive, thus we considered this event as fortuitous. South-West wind blew strongly in a chaotic way and 'Salinity' decreased during the first 2 significant intervals of 'Tide coefficient' effects (Fig. 3B and 3D). Then, wind probably induced the 2 first peaks of *Dinophysis* concentration, but these were more correlated with the sinusoid evolution of the tide coefficient. In the static regression, $\hat{\theta}_T$ was not significant, as could have been an average of the significant and non-significant intervals of the dynamic parameter $\hat{\theta}_{T,t}$.

From these results, a general explanation for the evolution of *Dinophysis* concentration was derived as follows: South-West winds draw water masses, possibly stratified and rich in *Dinophysis*, inshore, particularly to Antifer harbour due to the configuration of the site. North-East winds may provoke dispersion of *Dinophysis* cells. Large tide coefficients may induce a decrease of *Dinophysis* concentrations as a consequence of water mass movements and/or dilution. It should be noted that the set of significant covariates is a subset of available variables. A significant serial correlation for the residuals, as the runs test (Siegel 1956) showed us at the $\alpha = 0.05$ significance level for both years, might reflect the absence of at least one key descriptor in the model. As our analysis identified physical factors, this (these) might be biological factor(s). Our explanation illustrated the usefulness of DLRMs as explanatory tools. Dynamic models can also be used as an on-line analysis method for time series as, for instance, the phytoplankton time series issued from monitoring programmes. In this case, data are obtained sequentially and, although not recommended, sampling frequency might be irregular, generating time series with missing data. The sequential definition of dynamic models makes them well suited for such time series analysis. The estimation procedure can manage missing data by forecasting the value at time $t+k$, $k > 1$. Moreover, in ecology, the observational variance is often a function of the mean (Taylor 1961, Kendal 1995), and thus varies in time with the mean. If the variance-to-mean relationship is known, it can be used to specify the sequence of the observational variance. Alternatively, dynamic models can accommodate the assumption of time-varying variance. Finally, the Bayesian model approach of time series modelling can be considered as the dynamic generalization of the linear model, and thus developments of the latter (e.g. multiple linear regression) are adaptable for the former.

DLRM results gave us a more thorough understanding of *Dinophysis* concentration time series in Antifer than did static regression analysis. In particular, time-varying relationships between significant covariates and the concentration of the toxic microalga could not be assessed using static regression. Furthermore, DLRM characteristics and extensions could make dynamic models one of the most efficient tools for analysing time series data, and especially those of monitoring programmes.

Acknowledgements. We express our gratitude to P. Gros for commenting on this manuscript. We thank P. Lassus for providing data and P. Gentien for helpful discussions. L. Giboire is acknowledged for figures. The comments of 3 referees led us to improve our manuscript.

Appendix

The model used is a univariate DLRM with constant and unknown variance V. Let Y_t, $t = 1, 2, \ldots$, denote a time series and $\boldsymbol{X}'_t = (X_{1,t}\ X_{2,t}\ \ldots\ X_{n,t})$, $t = 1, 2, \ldots$, a time dependent vector of variables. The observation is governed by the so-called 'observation equation',

$$Y_t = \boldsymbol{F}'_t \boldsymbol{\theta}_t + \epsilon_t,$$

where $\boldsymbol{F}'_t = (1\ \boldsymbol{X}'_t)$ is the vector of regressors, $\boldsymbol{\theta}'_t = (\theta_{0,t}\ \theta_{1,t}\ \theta_{2,t}\ \ldots\ \theta_{n,t})$ is a vector of time dependent parameters and ϵ_t are observational errors, independently identically distributed in the normal distribution $N(0, V)$. The unknown reciprocal variance or precision is denoted by $\phi = V^{-1}$. At $t-1$, ϕ is distributed in the Gamma distribution $G[n_{t-1}/2, d_{t-1}/2]$. The parameter vector changes through time according to the evolution equation

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim T_{n_{t-1}}[\boldsymbol{0}, \boldsymbol{W}_t],$$

where $\boldsymbol{\omega}_t$ is the evolutional error. (The notation $\sim$ is used here and elsewhere to denote 'distributed as'.) Then we define the initial distributions $(\phi|D_0) \sim G[n_0/2, d_0/2]$ and $(\boldsymbol{\theta}_0|D_0) \sim T_{n_0}(\boldsymbol{m}_0, \boldsymbol{C}_0)$, where n_0, d_0, $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$ are fixed. D_0 is the initial information set, representing all the available relevant information used to specify the model before the first observation, and including all vectors of regressors $\boldsymbol{F}_t$. The error sequences ϵ_t and $\boldsymbol{\omega}_t$ are independent, mutually independent, and independent of $(\boldsymbol{\theta}_0|D_0)$. Lastly, $\boldsymbol{W}_t$ is specified as $\boldsymbol{W}_t = \boldsymbol{C}_{t-1}(1-\delta)/\delta$, where $\delta \in]0, 1[$. δ is the so-called discount factor and controls the model adaptability: if δ is near 0 then the model adaptability is high and if δ is near 1 then the model can only change slowly. The sequential estimation procedure starts at $t-1$. Let us define the information set at time t as $D_t = \{D_{t-1}, Y_t\}$. $(\boldsymbol{\theta}_{t-1}|D_{t-1})$ is distributed as $T_{n_{t-1}}[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}]$ and $(\phi|D_{t-1})$ as $G[n_{t-1}/2, d_{t-1}/2]$. The estimation steps are the following:

Prior. $(\boldsymbol{\theta}_t|D_{t-1}) \sim T_{n_{t-1}}[\boldsymbol{a}_t, \boldsymbol{R}_t]$, where $\boldsymbol{a}_t = \boldsymbol{m}_{t-1}$ and $\boldsymbol{R}_t = \boldsymbol{C}_{t-1} + \boldsymbol{W}_t$.

(Continued on next page)

Appendix (continued)

Prediction. $(Y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t]$, where $f_t = \boldsymbol{F}_t'\boldsymbol{a}_t$ and $Q_t = \boldsymbol{F}_t'\boldsymbol{R}_t\boldsymbol{F}_t + S_{t-1}$ with $S_{t-1} = d_{t-1}/n_{t-1}$.

Posterior. $(\boldsymbol{\theta}_t|D_t) \sim T_{n_t}[\boldsymbol{m}_t, \boldsymbol{C}_t]$, where $\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t$ and $\boldsymbol{C}_t = (S_t/S_{t-1})(\boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{A}_t'Q_t)$, with $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}_tQ_t^{-1}$ and $e_t = Y_t - f_t$. $(\phi|D_t)$ is distributed as $G[n_t/2, d_t/2]$, where $n_t = n_{t-1} + 1$ and $d_t = d_{t-1} + S_{t-1}e_t^2/Q_t$.

The fitted distribution $(Y_t|D_t)$ is distributed as $T_{n_t}[g_t, P_t]$, where $g_t = \boldsymbol{F}_t'\boldsymbol{m}_t$ and $P_t = \boldsymbol{F}_t'\boldsymbol{C}_t\boldsymbol{F}_t + S_t$.
Parameters of the initial ditributions of $(\phi|D_0)$ and $(\boldsymbol{\theta}_0|D_0)$ and the discount factor δ are included in the set D_0 and fixed by the model user. For our model applied to the *Dinophysis* concentration time series (cf. 'Dynamic linear regression model'), the values were $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{C}_0$ equal to the identity matrix, $n_0 = d_0 = 1$ and $\delta = 0.95$ in 1989 and 1990.
Computer programs used to perform DLRM analysis were developed in C on a SUN station and are available from the first author.

LITERATURE CITED

Berland BR, Maestrini SY, Grzebyk D (1995a) Observations on possible life cycle stages of the dinoflagellates *Dinophysis* cf. *acuminata*, *Dinophysis acuta*, *Dinophysis pavillardi*. Aquat Microb Ecol 9:183-189

Berland BR, Maestrini SY, Grzebyk D, Thomas P (1995b) Recent aspects of nutrition in the dinoflagellate *Dinophysis* cf. *acuminata*. Aquat Microb Ecol 9:191-198

Delmas D, Herbland A, Maestrini SY (1992) Environmental conditions which lead to increase in cell density of the toxic dinoflagellates *Dinophysis* spp. in nutrient-rich and nutrient-poor waters of the French Atlantic coast. Mar Ecol Prog Ser 89:53-61

Delmas D, Herbland A, Maestrini SY (1993) Do *Dinophysis* spp. come from the open sea along the French Atlantic coast? In: Smayda TJ, Shimizu Y (ed) Toxic phytoplankton blooms in the sea. Elsevier Science, Amsterdam, p 489-494

Draper NR, Smith H (1966) Applied regression analysis. Wiley series in probability and mathematical statistics. John Wiley and Sons, New York

Kendal WS (1995) A probabilistic model for the variance to mean power law in ecology. Ecol Model 80:293-297

Kendall M, Stuart A (1977) The advanced theory of statistics. Vol. 2. Charles Griffin and Company, London

Lagadeuc Y (1992) Transport larvaire en Manche. Exemple de *Pectinaria koreni* (Malmgren), annélide polychète en baie de Seine. Oceanol Acta 15:383-395

Maestrini SY, Berland BR, Carlsson P, Granéli E, Pastoureaud A (1996) Recent advances in the biology of the toxic dinoflagellate genus *Dinophysis*: the enigma continues. In: Yasumoto T, Oshima Y, Fukuyo Y (ed) Harmful and toxic algal blooms. IOC of UNESCO, p 397-400

Ménesguen A, Guillaud J-F, Aminot A, Hoch T (1995) Modelling the eutrophication process in a river plume: the Seine case study. Ophelia 42:205-225

Monbet Y (1975) Les incidences écologiques de la construction du terminal d'Antifer. Rapport CNEXO-unité littoral

Pole A, West M, Harrison PJ (1994) Applied Bayesian forecasting and time series analysis. Chapman and Hall, New York

Salomon J-C, Breton M (1993) An atlas of long-term currents in the Channel. Oceanol Acta 16:439-448

Sampayo MA de M (1993) Trying to cultivate *Dinophysis* spp. In: Smayda TJ, Shimizu Y (ed) Toxic phytoplankton blooms in the sea. Elsevier Science, Amsterdam, p 807-810

Siegel S (1956) Nonparametric statistics for the behavioral sciences. McGraw-Hill series in psychology. McGraw-Hill, New York

Soudant D, Beliaeff B, Thomas G (1997) Dynamic linear Bayesian models in phytoplankton ecology. Ecol Model 99:161-169

Taylor LR (1961) Aggregation, variance and the mean. Nature 189:732-735

Thiébaut E, Dauvin J-C, Lagadeuc Y (1994) Horizontal distribution and retention of *Owenia fusiformis* larvae (Annelida: Polychaeta) in the bay of Seine. J Mar Biol Assoc UK 74:129-142

Utermöhl H (1958) Zur Vervollkommnung der quantitativen Phytoplankton Methodik. Int Ver Theor Angew Limnol Verh 9:1-38

West M, Harrison PJ (1997) Bayesian forecasting and dynamic models, 2nd edn. Springer series in statistics. Springer-Verlag, New York

West M, Harrison PJ, Migon HS (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion). J Am Stat Assoc 80:73-97

4.3.3 Commentaires

Dans le premier article, la procédure d'estimation de la variance d'observation évoquée dans la section *Dynamic Linear Regression Model* fait référence à celle du modèle à variance d'observation constante et inconnue. De plus, la variance V a été estimée par ce modèle de la même façon et en même temps que les valeurs des conditions initiales. Cependant cette approche implique des valeurs estimées dépendantes du futur de la série temporelle. C'est pourquoi, dans le second article, les valeurs des conditions initiales sont spécifiées arbitrairement. Dans ce même travail, les données manquantes ont été remplacées par les estimations de modèles polynomiaux du premier ordre à variance d'observation constante et inconnue. Les conditions initiales sont $m_0 = 0$, $C_0 = 0$, $d_0 = 0$ et $n_0 = 0$. Pour chaque variable environnementale, les vraisemblances *a priori* pour les valeurs du facteur d'escompte de 0,05 à 1 par pas de 0,05 ont été calculées. Les estimations des modèles ayant la plus forte vraisemblance ont été utilisées.

Lors de la sélection des covariables standardisées du modèle appliqué à l'année 1988, nous avons constaté une corrélation tendant vers -1 entre le niveau moyen (version dynamique de l'ordonnée à l'origine) et le paramètre de la variable « température ». Cette corrélation entraînait des estimations aberrantes de ces paramètres et, par voie de conséquence, nécessitait le retrait du niveau moyen ou de la variable « température » de l'ensemble des éléments du modèle (cf. page 36). L'élimination du niveau moyen d'un modèle dynamique implique une moyenne locale au temps t, $t = 1, 2, \ldots$, toujours égale à zéro, ce qui est une hypothèse très forte. Toutefois, nous avons choisi de conserver uniquement la température en s'appuyant sur deux arguments : (i) la série temporelle d'une covariable est plus riche en information qu'une droite d'équation $y = 1^*$ et, (ii) selon Draper et Smith (1966), en régression statique, la standardisation des variables autorise l'exclusion de l'ordonnée à l'origine. Cependant la régression dynamique n'est pas la régression statique et le niveau moyen n'est pas l'ordonnée à l'origine. Une standardisation autorisant le retrait du niveau moyen d'un MDB doit être dynamique, *i.e.* doit utiliser pour tout t, $t = 1, 2, \ldots$, la moyenne et la variance **locales** de chaque covariable. Ces moyennes et variances peuvent être estimées par des modèles dynamiques de série temporelle ou bien encore par un modèle multiprocess. En l'absence d'une telle standardisation, il est préférable de conserver le niveau moyen. C'est ce que nous avons fait pour le modèle de notre second article, où la corrélation entre le niveau moyen et le paramètre de la variable « température » convergeait également vers -1 pour les années 1989 et 1990. Pour finir, on peut remarquer que la corrélation entre le niveau moyen et le paramètre de la température apporte une preuve supplémentaire du contrôle exercé par ce facteur environnemental sur les concentrations en *Dinophysis*.

$*$. Dans le vecteur des régresseurs $\boldsymbol{F}_t' = (1\ X_{1,t}\ X_{2,t}\ \ldots\ X_{n,t})$ la valeur 1 correspond à l'équivalent d'une covariable pour le niveau moyen. Sa représentation est une droite d'équation $y = 1$.

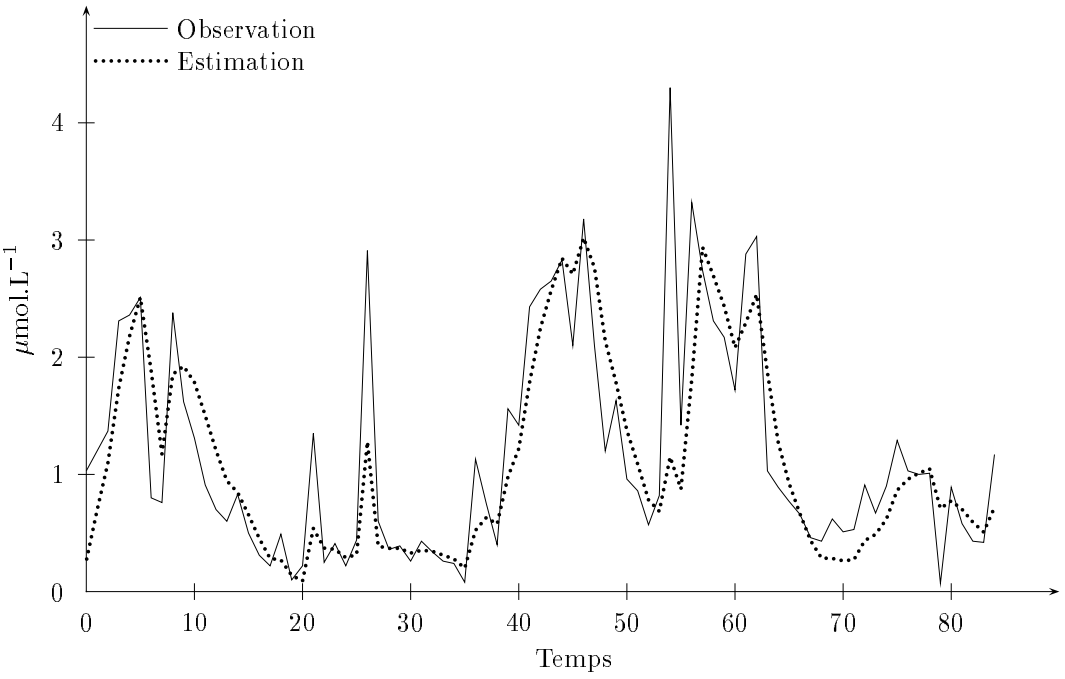

FIG. 4.3 - *Observation et estimation d'un modèle HARRISON-STEVENS des concentrations en nitrate (NO_3) mesurées à ANTIFER à partir du 26 mai 1988.*

4.4 Modèle HARRISON-STEVENS

Pour cette application, prenons le point de vue d'un écologiste désirant extraire la tendance (*i.e.* le niveau moyen) de la série temporelle des concentrations en nitrate à ANTIFER en 1988. Manifestement, cette série présente des données exceptionnelles et des changements de niveau (FIG. 4.3, e.g. resp. jours 26 et 6). L'utilisation d'un modèle HARRISON-STEVENS requiert la spécification de la variance d'observation du modèle « routine » et des conditions initales. Nous avons considéré que la variance des différences premières (*i.e.* $Y_t - Y_{t-1}$) constituait une estimation de la variance du processus étudié. La valeur utilisé pour V_{1_t} est la variance des différences premières sans les données exceptionnelles et les changements de niveau. Ces dernières ont été choisies par la règle de décision suivante : si la valeur absolue d'une différence première standardisée est supérieure à 1,96 alors la valeur est considérée comme exceptionnelle. Les jours 6, 26, 54 et 56 ont ainsi été désignés comme valeurs exceptionnelles. À cette liste nous avons rajouté le jour 8, considérant qu'il représentait un changement de niveau. La valeur de la variance des différences premières est alors 0,27 (à 10^{-2} près). Les éléments du modèle sont obtenus en multipliant ceux du modèle standard donné page 33 par 0,27. En ce qui concerne les conditions initiales, nous avons choisi arbitrairement $\boldsymbol{m}_0 = \mathbf{0}$ et $\boldsymbol{C}_0 = \boldsymbol{W}_{1_t}$. La figure 4.3 montre les observations des concentrations en nitrate et leur estimation par ce modèle HARRISON-STEVENS.

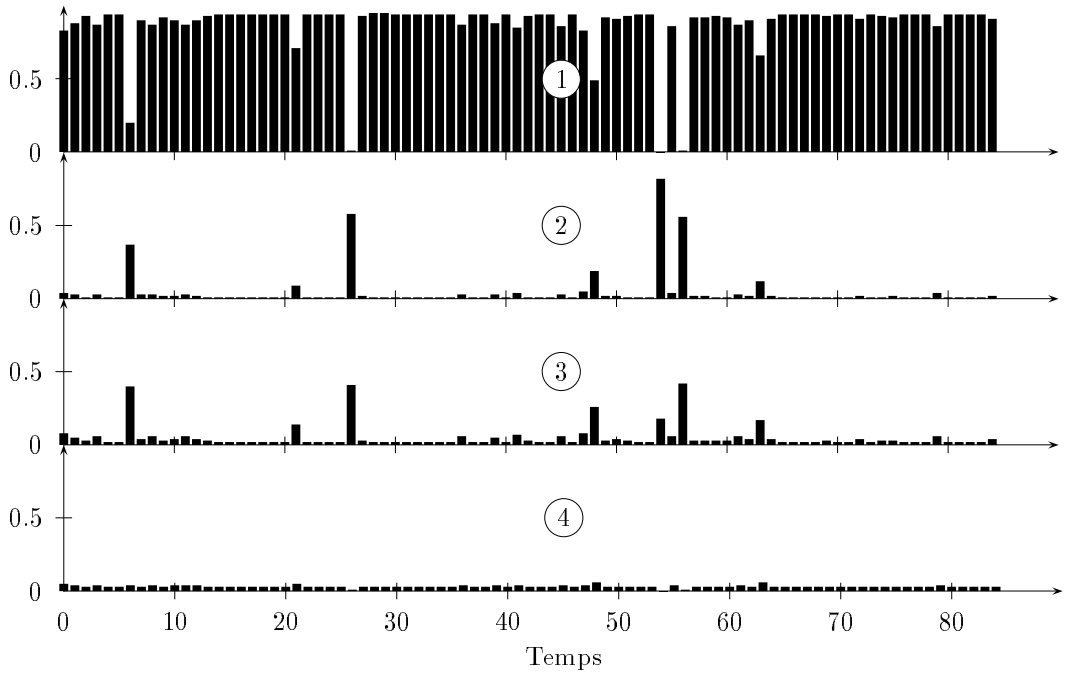

FIG. 4.4 - *Probabilités* a posteriori *des quatre états d'un modèle* HARRISON-STEVENS *appliqué aux concentrations en nitrate (NO_3) mesurées à* ANTIFER *à partir du 26 mai 1988 : 1 « routine », 2 « donnée exceptionnelle », 3 « changement de niveau » et 4 « changement de pente ».*

L'effet multiprocess est manifeste dans la faible prise en compte de la donnée exceptionnelle du jour 56. De manière moins spectaculaire, le caractère particulier du modèle s'exprime dans le traitement de la donnée du jour 26 : c'est également une donnée exceptionnelle qui induit une augmentation ponctuelle de l'estimation. L'estimation d'un modèle polynomial du second ordre aurait augmenté de manière plus marquée, mais surtout serait revenue à son niveau initial beaucoup plus lentement. La figure 4.4 montre les probabilités *a posteriori* des quatre états du modèle. Le modèle « routine » est généralement le plus probable et l'état « changement de pente » a une probabilité toujours inférieure à 0,06 (à 10^{-2} près). Les états « donnée exceptionnelle » et « changement de niveau » présentent en phase des pics de probabilités élevées aux jours 6, 26, 54 et 56. Hormis la date 54 où la probabilité de l'état 2 est 0,82, les résultats du modèle laissent indécis quant à la nature des changements intervenus lors de ces jours. Cette indécision est normale puisque ces probabilités *a posteriori* au temps t sont calculées avec les informations disponibles au temps t et seul le devenir du processus au temps $t+1$ peut permettre de trancher entre les deux états. Par voie de conséquence, lors des changements de niveau des jours 6 et 56 les estimations sont « en retard » d'une unité de temps (FIG. 4.3). L'observation au temps $t+1$ est prise en compte pour le calcul des probabilités lissées (cf. annexe D) présentées figure 4.5. Le modèle

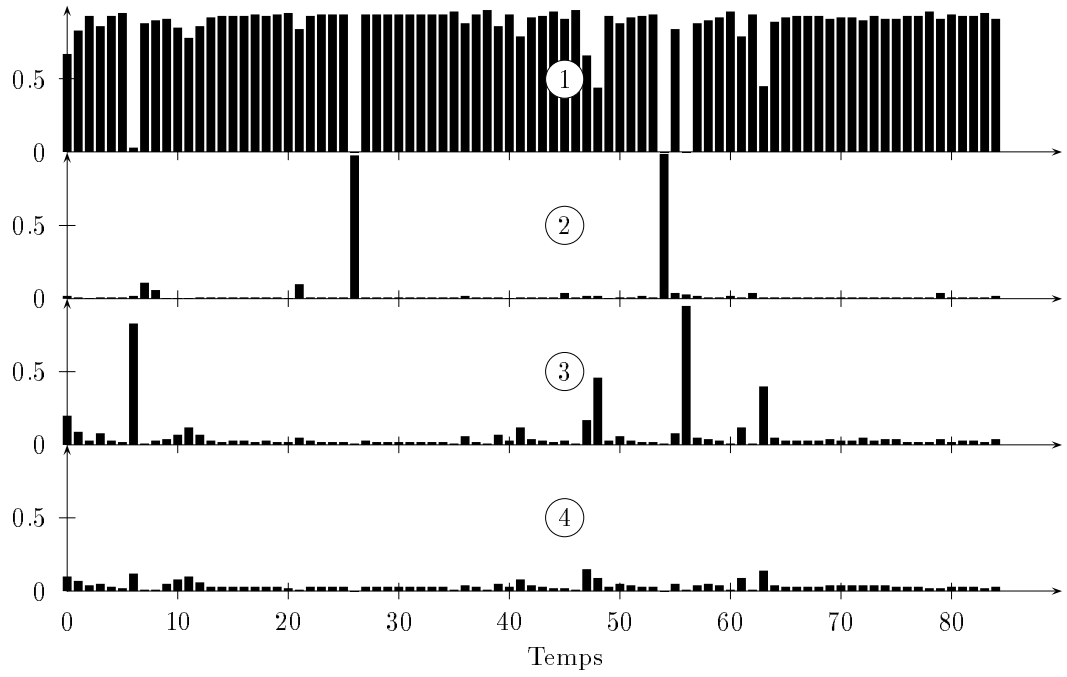

FIG. 4.5 - *Probabilités lissées des quatre états d'un modèle HARRISON-STEVENS appliqué aux concentrations en nitrate (NO_3) mesurées à ANTIFER à partir du 26 mai 1988 : 1 « routine », 2 « donnée exceptionnelle », 3 « changement de niveau » et 4 « changement de pente ».*

« routine » est toujours le plus vraisemblable. La probabilité de l'état « changement de pente » présente des augmentations locales ne dépassant pas 0,16. Les données des jours 26 et 54 sont clairement identifiées comme données exceptionnelles, et celles des jours 6 et 56 comme des changements de niveau. La donnée du jour 8 est considérée comme générée par le processus « routine », et ainsi met en cause la légitimité de son inclusion arbitraire dans l'ensemble des données exceptionnelles. L'estimation lissée (FIG. 4.6) prend en compte l'observation Y_{t+1} pour l'estimation de l'observation au temps t. Les données exceptionnelles sont désormais ignorées et les changements de niveau sont estimés sans retard. Ces estimations lissées sont globalement plus proches des observations que les estimations *a posteriori*. Ce phénomène est particulièrement sensible dans les intervalles de jours [45 ; 50] et [60 ; 65]. Dans ces mêmes intervalles, la probabilité de l'état « changement de niveau » présente des augmentations ponctuelles importantes. Les probabilités des sous-modèles sont une mesure de leur vraisemblance respective, mais également les poids de chacune de leur estimation dans l'estimation de l'observation. Ainsi, les augmentations des probabilités du modèle « changement de niveau » sont à l'origine du meilleur ajustement lors de ces intervalles. Dans les deux cas, les valeurs observées montrent des décroissances marquées qui ne sont pas formellement identifiées comme des changements de niveau. En particulier,

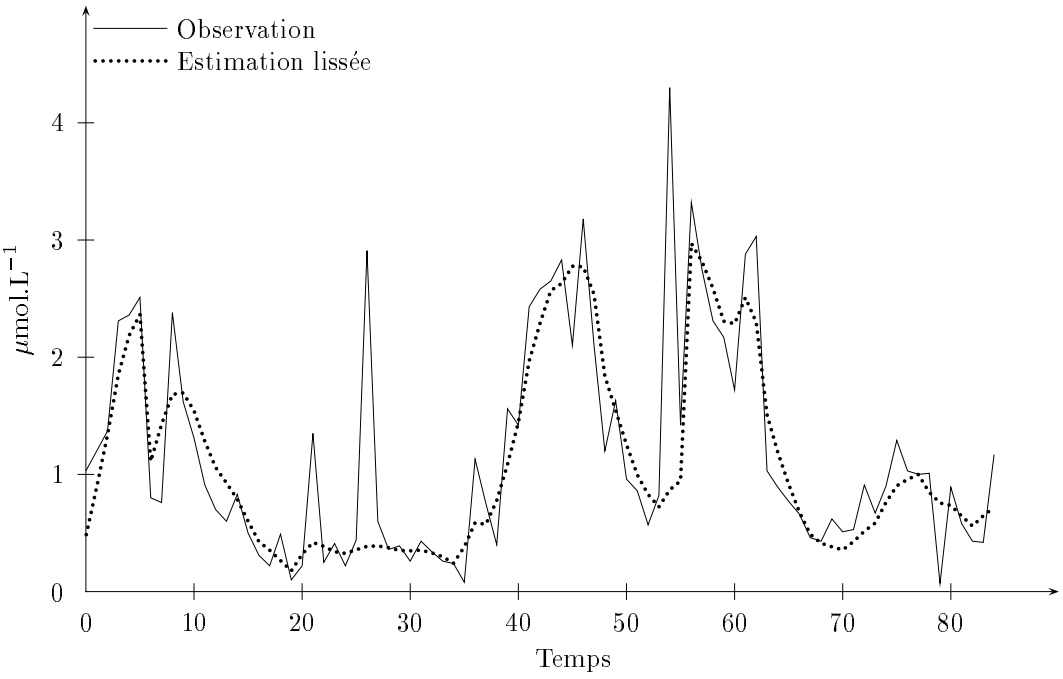

FIG. 4.6 - *Observation et estimation lissée d'un modèle HARRISON-STEVENS des concentrations en nitrate (NO_3) mesurées à ANTIFER à partir du 26 mai 1988.*

du jour 62 au jour 63 la concentration diminue de 2,00 μmol.L^{-1}. Or, du jour 55 au jour 56 la concentration en nitrate augmente de 2,10 μmol.L^{-1} et le jour 56 est désigné comme un changement de niveau. La différence de jugement porté sur ces deux événements possède deux causes qui sont liées. Premièrement, la valeur des estimations aux dates 55 et 62 induisent des écarts avec les dates 56 et 63, respectivement, de 2,38 et $-1{,}27$ μmol.L^{-1} (à 10^{-2} près). Deuxièmement, la valeur spécifiée de la variance d'observation induit des variances des estimations assez importantes pour que l'écart $-1{,}27$ soit plus vraisemblable avec le modèle « routine » qu'avec le modèle « changement de niveau ». Ici, la variance V_{1_t} est trop grande. La variance des différences premières est plus proche de la variance de l'estimation *a posteriori* que de la variance d'observation. Bien qu'il soit analytiquement possible d'établir une relation entre la variance des différences premières et la variance d'observation, il semble judicieux d'utiliser l'hypothèse d'une variance constante et inconnue estimée séquentiellement à partir des données.

En dépit de son caractère non-optimal, cette illustration montre l'efficacité de l'utilisation du modèle HARRISON-STEVENS pour l'extraction de tendance de séries temporelles présentant des données exceptionnelles et des changements de niveau. L'examen des données d'ANTIFER nous ont montré que de tels événements n'étaient pas rares dans les séries temporelles écologiques. De plus, ces séries présentent souvent des données manquantes et/ou une fréquence irrégu-

lière. Ces deux caractéristiques peuvent également être prises en compte par le modèle au prix d'un effort méthodologique. Bien que l'hypothèse de normalité de la distribution d'observation soit « confortable » d'un point de vue technique, elle n'est pas toujours adaptée à ces données généralement positives ou nulles. En particulier, elle peut conduire à des intervalles de confiance comportant des valeurs négatives. Toutefois, les modèles multiprocess peuvent être généralisés à des observations non-gaussiennes. Bolstad (1995) a donné le développement de ce type de modèle sous l'hypothèse d'une observation distribué selon une loi de Poisson. Cette distribution ainsi que les distributions binomiale négative et log-normale, pour lesquelles des développements similaires sont envisageables, seraient sans aucun doute mieux adaptées aux données de l'écologie marine (Bulmer, 1974 ; Daget, 1976 ; Viera Da Silva, 1979 ; El-Shaarawi et al., 1981).

Chapitre 5

Discussion

Dinophysis est une microalgue toxique à l'origine d'épidémie de diarrhées. Les connaissances établies sur la biologie et l'écologie de cette microalgue sont limitées en dépit de nombreuses études. Les tentatives de traitements statistiques des données écologiques ont permis de confirmer les observations *in situ*, mais n'ont pas fourni de nouvelles informations. Toutefois, les méthodes utilisées supposent implicitement des relations constantes dans le temps entre la variable dépendante et les variables explicatives. L'objectif de ce travail est d'appliquer des modèles dynamiques bayésiens en écologie phytoplanctonique et d'évaluer dans ce contexte l'intérêt de leur hypothèse de relations variables dans le temps. Les séries temporelles d'ANTIFER concernent le dinoflagellé toxique *Dinophysis*. Elles sont exceptionnelles en EUROPE par leur fréquence journalière ou quasi-journalière, l'importance de la période de prélèvements (trois à quatre mois) et la répétition de ce protocole pendant sept années. Des modèles bayésiens de régression linéaire dynamique ont été appliqués à ces données.

Intérêts des MDB

L'hypothèse de variabilité des relations entre une variable à expliquer et des variables explicatives s'est avérée pertinente à travers les changements dans les valeurs des paramètres des covariables et l'alternance de périodes significatives et non-significatives. De plus, l'interprétation de la variabilité des paramètres a permis de mieux comprendre l'effet de certaines variables. Cette interprétation a débouché sur un schéma d'explication d'une grande partie de l'évolution des concentrations en *Dinophysis* sur le site d'ANTIFER. Ce schéma repose sur deux phénomènes physiques : les courants de surface induits par le vent et les courant résiduels de marées. L'effet d'accumulation induit par les vents de sud-ouest relevé par LASSUS et al. (1993) a été confirmé et un effet de dispersion des vents de nord-est a été mis en évidence. L'intensité de ces effets dépend de la

concentration en microalgue toxique contenue dans les masses d'eaux soumises aux phénomènes d'accumulation et de dispersion. Comme cette concentration varie dans le temps, les effets varient également dans le temps. L'existence de ces effets est soutenue par l'hydrodynamique de la baie de Seine (Salomon et Breton, 1993), la présence des plus fortes concentrations en phytoplancton au sud-ouest du site d'Antifer (Ménesguen et al., 1995), et par l'observation de transports de larves d'organisme marins dans le panache du fleuve (Lagadeuc, 1992 ; Thiébaut et al., 1994). L'importance de la structure du port d'Antifer entraîne des courants résiduels de marée particuliers et dépendant de la valeur du coefficient de marée. Par faibles coefficients les masses d'eaux sont retenues dans l'enceinte du terminal pétrolier alors qu'elles sont expulsées vers le nord par forts coefficients. Ces mouvements ont un effet sur la concentration en *Dinophysis* au sein du port et cet effet est variable dans le temps pour les mêmes raisons que les effets du vent. Cette hypothèse avait été soulevée par De Cremoux (1988) sur la base des travaux de Monbet (1975). Dans le contexte de ces effets physiques, la salinité est interprétée comme la marque du passage de masse d'eaux. Les faibles salinités de surface sont associées aux fortes concentrations de *Dinophysis*. Elles témoignent d'une stratification de la colonne d'eau qui est favorable au développement de la microalgue (Delmas et al., 1992). Ces faibles salinités de surface pourraient également signaler une origine estuarienne et ainsi des eaux riches en nutriments. L'information apportée par la variation des paramètres a permis une analyse plus fine que les résultats de régressions statiques. Outre le fait que le pourcentage de variation expliquée est supérieur pour la régression dynamique, il est apparu que le paramètre d'une covariable pouvait présenter des périodes de significativité dans un dlrm et ne pas être significatif dans la version statique du même modèle.

Limites et extensions des modèles

L'insatisfaction majeure éprouvée avec les modèles dynamiques bayésiens est que les séries des valeurs des paramètres dynamiques (trajectoires) ne sont pas uniques. En effet, elles dépendent de la spécification des conditions initiales. Dans les modèles que nous avons utilisés, l'influence de ces valeurs semble limitée dans le temps. Il est nécessaire de confirmer ou d'infirmer cette expérience par une étude formelle de l'influence de la spécification des paramètres au temps 0. L'utilisation de la procédure de lissage de la dernière observation vers la première fournit des estimations qui sont bien moins sensibles aux conditions initiales. Cependant, cette méthode suppose que les phénomènes intervenant en fin de série ont une influence sur ceux du début de la série. Cette dépendance ne paraît pas vraisemblable et va à l'encontre de l'intuition selon laquelle le futur dépend du passé. L'analyse référentielle est une autre méthode permettant le calcul de conditions initiales à partir des données. Le calcul de ces valeurs présente les avantages d'objectivité, d'unicité et de reproductibilité, mais n'assure aucune op-

timalité (cf. annexe C). De plus, les données utilisées dans ce calcul sont écartées de l'analyse. Finalement, la possibilité de recherche qui me semble la plus viable est de considérer l'ensemble des trajectoires des valeurs des paramètres et de choisir parmi celles-ci la plus vraisemblable. Cette approche semble réalisable par l'utilisation de *Markov Chain Monte Carlo* (MCMC), dont le couplage avec les modèles dynamiques est présenté par WEST et HARRISON (1997, chapitre 15).

Afin d'améliorer l'adéquation entre les modèles utilisés et la nature des données écologiques, plusieurs extensions sont envisageables. En premier lieu, les séries temporelles comportent souvent des données manquantes, résultantes d'une absence de prélèvement ou d'une fréquence d'échantillonnage irrégulière. Les MDB peuvent structurellement traiter ces absences d'informations de la variable dépendante, et des variables explicatives si l'on utilise un modèle multivarié. En particulier, cette possibilité permettrait de traiter les données du REPHY dont la fréquence d'échantillonnage est variable. Ensuite, la transformation logarithmique des concentrations en *Dinophysis* fournit seulement une approximation satisfaisante à l'hypothèse de variance d'observation constante. Si les variances des mesures sont connues, elles peuvent être utilisées directement dans les MDB. Il en va de même d'une relation du type moyenne-variance. Autrement, la variance V peut être supposée variable dans le temps et être estimée séquentiellement. Les lois de probabilités de POISSON, binomiale négative et log-normale sont en générale mieux adaptées aux données écologiques. Elles peuvent être utilisées pour la distribution d'observation. Enfin, la représentation classique des séries temporelles consiste en un signal et une perturbation aléatoire. De ces deux éléments, seul le signal est riche d'information. De ce fait, on peut envisager un pré-traitement des données sous la forme d'extractions des tendances par un modèle multiprocess. Pour finir, notre problématique se résumant à l'estimation des séries temporelles des paramètres d'un modèle mathématiques, il faut conserver à l'esprit que d'autres approches sont envisageables (cf. O'HAGAN, 1978 ; YOUNG et YOUNG, 1992).

Données et échantillonnage

Un problème récurant dans le traitement de données écologiques est l'existence d'interrelations entre les facteurs environnementaux. Les variables présentent des corrélations significatives sans pour autant être très élevées. Le cas de l'augmentation de la concentration en *Dinophysis* en 1988 aux environs de la date 30 (cf. page 52) est une situation exceptionnelle où la plupart des covariables ont également changé de valeur de façon importante. Ces augmentations concomitantes se traduisent par des corrélations croisées desquelles aucune information ne ressort. Ainsi, la surabondance d'information peut être aussi stérile que son absence. En dehors de ce type de cas extrême, les corrélations peuvent entraîner des biais dans les estimations des paramètres. Les corrélations entre covariables sont directement liées aux corrélations entres paramètres. Leur interprétation

peut parfois être informative, comme cela a été le cas pour les applications de DLRM aux années 1989 et 1990 : la corrélation entre vent et salinité est attendue dans l'hypothèse où le vent est à l'origine de déplacements de masses d'eaux dont les salinités peuvent être différentes. En revanche, lorsque les valeurs absolues des corrélation entre paramètres tendent vers 1, l'estimations des paramètres n'ont pas de sens. Il convient d'ôter une ou plusieurs des variables mises en cause (cf. page 36), comme nous l'avons fait dans le cas particulier d'une corrélation avec le paramètre « niveau moyen » (cf. page 77).

Comme nous l'avons déjà signalé, les séries temporelles d'ANTIFER sont exceptionnelles. Cependant, dans un environnement turbulent tel que la mer ou de grands lacs, les dimensions spatiales et temporelles ne peuvent être séparées (BOYCE, 1974). La variabilité spatiale des facteurs physiques, chimiques et biologiques est contrôlée par les processus hydrodynamiques et est transmise aux organismes vivants (LEGENDRE et DEMERS, 1984). Cela apparaît dans notre travail par une explication de l'évolution des concentrations en *Dinophysis* liée à des phénomènes physiques et à la configuration très particulière du terminal pétrolier (FIG. 4.1, page 43). Il est probable que tout autre site aurait également présenté des spécificités hydrodynamiques locales. Ainsi, la prise en compte de la variabilité temporelle seule est insuffisante pour déterminer les conditions environnementales d'apparition d'une microalgue. Face à ce problème, une première solution est d'établir pour une zone géographique donnée (e.g. l'enceinte du port d'ANTIFER) une procédure d'estimation des valeurs des variables à partir de plusieurs mesures réparties dans cette zone. La répétition dans le temps des mesures et de la procédure permet l'obtention de séries temporelles d'estimations locales. TAGGART (1987) a utilisé avec succès une approche de ce type. La seconde solution consiste à mettre en relation les répartitions spatiales des facteurs physiques, chimiques et biologiques et celle des concentrations en *Dinophysis*. Il est nécessaire d'effectuer d'importantes campagnes d'échantillonnage de la microalgue et d'utiliser une méthode d'estimation géographique, comme le krigeage (MATHERON, 1972). Ce dernier permet l'obtention d'estimations non-biaisées de variables géographiques et de leurs variances mais n'intègre pas la dépendance temporelle entre les répartitions spatiales. Il existe plusieurs méthodes d'estimation de variables géographiques prenant en compte intégralement les dimensions spatiale et temporelle (CRESSIE, 1993). Les plus récentes approches de ce type utilisent souvent des méthodes bayésiennes et des *Markov Chain Monte Carlo* (WALLER et al., 1997). Le manque de données biologiques concernant *Dinophysis* limite l'interprétation des résultats d'un modèle de dynamique. Cependant, couplé à un modèle hydrodynamique cette approche permettrait la cartographie des facteurs environnementaux. Éventuellement, les résultats de ces modèles peuvent être corrigés par assimilation des données issues des échantillonnages par l'utilisation d'un filtre de KALMAN ou d'un modèle dynamique bayésien.

Chapitre 6

Conclusion

Ce travail présente une méthode originale de traitement des séries temporelles : les modèles dynamiques bayésiens. La particularité de ces modèles est de supposer les relations entre la variable dépendante et les variables indépendantes susceptibles de subir des changements. La pertinence de cette hypothèse en écologie phytoplanctonique a été démontrée par l'application de modèles de régression linéaire dynamique à des données concernant la microalgue toxique *Dinophysis*. Les résultats obtenus ont permis de montrer des facteurs importants expliquant en partie l'évolution temporelle du dinoflagellé sur le site d'Antifer (Normandie, France) en 1988, 1989 et 1990. La prédominance des phénomènes physiques nous amène à souligner la nécessité de la prise en compte de la dimension spatiale dans les protocoles d'échantillonnages phytoplanctonique et à esquisser les caractéristiques d'un tel plan. En particulier, cette approche pourrait être utile pour optimiser localement les procédures d'échantillonnage du REPHY et celles envisagées pour l'acquisition de séries temporelles dans le contexte du Programme National « Efflorescences Algales Toxiques » (PNEAT)(Maestrini et al., 1997). Les modèles dynamiques bayésiens et l'ensemble des extensions qu'ils proposent se sont avérés être une approche particulièrement intéressante du traitement des séries temporelles écologiques.

Annexe A

Théorème de Bayes

Soit deux événements A et B, auxquels sont associées les probabilités $\mathrm{P}(A)$ et $\mathrm{P}(B)$, avec $\mathrm{P}(A) \neq 0$ et $\mathrm{P}(B) \neq 0$. La probabilité de l'événement B conditionnellement à l'événement A est :

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(B \text{ et } A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \text{ et } B)}{\mathrm{P}(A)},$$

et symétriquement, la probabilité de l'événement A conditionnellement à l'événement B est $\mathrm{P}(A|B) = \mathrm{P}(A \text{ et } B)/\mathrm{P}(B)$. Ces formules permettent d'écrire :

$$\mathrm{P}(A \text{ et } B) = \mathrm{P}(B \text{ et } A) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A).$$

Au facteur proportionnel $1/\mathrm{P}(B)$ près, la propriété

$$\mathrm{P}(A|B) \propto \mathrm{P}(B|A)\mathrm{P}(A)$$

est connue sous le nom de théorème de Bayes. (La notation $\propto$ signifie « proportionnel à ».). En pratique, on s'intéresse à A. $\mathrm{P}(A)$ représente nos connaissances *a priori* sur A. B représente la réalisation d'une expérience informative sur A. La vraisemblance de cette expérience conditionnellement à nos connaissances *a priori* est $\mathrm{P}(B|A)$. Ce que l'on veut connaître c'est $\mathrm{P}(A|B)$, l'état de nos connaissances *a priori* conditionnellement à la réalisation de l'expérience, c'est-à-dire l'état des connaissances *a posteriori*. Le théorème de Bayes nous permet de calculer $\mathrm{P}(A|B)$ à partir de $\mathrm{P}(A)$ et $\mathrm{P}(B|A)$. Dans ce contexte, ce théorème s'exprime de façon symbolique par :

$$\textbf{\textit{a posteriori}} \propto \textbf{vraisemblance} \times \textbf{\textit{a priori}}.$$

Annexe B

Tests statistiques utilisés

B.1 Test des séquences

Le test des séquences (Siegel, 1956) permet de tester le caractère aléatoire d'une série temporelle présentant uniquement deux types de valeurs. Par exemple, plusieurs jets d'une pièce de monnaie donnent une série de résultats composée de « pile » et de « face », et les signes des erreurs de prédiction d'un modèle donnent une série composée de « + » et de « − ». Le test est basé sur le dénombrement des séquences. Une séquence est définie comme un ou plusieurs résultats du même type encadrés par des résultats de l'autre type. Soit, pour une série temporelle donnée, n_1 le nombre de résultats du premier type, n_2 le nombre de résultat du second type, $N = n_1 + n_2$ le nombre total de résultats et s le nombre de séquences. Les hypothèses testées sont les suivantes.

H_0 : la répartition des deux types de résultat est aléatoire.

H_1 : la répartition des deux types de résultat n'est pas aléatoire.

Si n_1 et n_2 sont inférieurs ou égaux à 20, alors deux tables donnent les valeurs limites inférieures et supérieures de l'intervalle de confiance du nombre de séquence sous l'hypothèse H_0 et au risque de première espèce $\alpha = 0{,}05$. Si le nombre de séquences comptées dans la série temporelle testée n'est pas inclus dans cet intervalle de confiance, alors H_0 ne peut pas être acceptée. Les deux tables des limites inférieures et supérieures sont données dans les ouvrages traitant des tests non-paramétriques (e.g. Siegel, 1956). Si n_1 et n_2 sont supérieurs à 20, alors s est distribué selon une loi normale de moyenne m_s et de variance V_s tels que,

$$m_s = \frac{2n_1n_2}{n_1 + n_2} + 1,$$

$$V_s = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}.$$

La statistique $z = (s - m_s)/\sqrt{V_s}$ est distribuée selon une loi normale $N[0\,;1]$. Si $|z|$ est supérieur à $\epsilon_{1-\alpha/2}$ (la valeur d'une loi normale $N[0\,;1]$ pour la probabilité $1 - \alpha/2$) alors l'hypothèse H_0 ne peut pas être acceptée. Dans le cas où la série étudiée est la suite des signes des erreurs de prédiction d'un modèle, si $z > \epsilon_{1-\alpha/2}$ (resp. $z < -\epsilon_{1-\alpha/2}$) alors les erreurs sont corrélées positivement (resp. négativement).

B.2 Test du rapport de vraisemblance

Le lecteur pourra se référer à Dagnelie (1975) et Kendall et Stuart (1977). Soit deux modèles statistiques différents par la valeur spécifiée d'un paramètre θ, $M(\theta_1)$ et $M(\theta_2)$. À ces modèles sont associées les vraisemblances, V_1 et V_2. Supposons $V_1 > V_2$. Les hypothèses du test sont,

H_0 : les vraisemblances des modèles sont égales.

H_1 : les vraisemblances des modèles différent.

L'hypothèse H_0 est rejeté au risque de première espèce α lorsque la quantité $\chi^2_{obs} = -2\log(V_2/V_1)$ est supérieure ou égale à $x_{1-\alpha}$, où $x_{1-\alpha}$ est la valeur de la fonction de répartition d'une distribution χ^2 à un degré de liberté pour la probabilité $1 - \alpha$. La table de la fonction de répartition d'une distribution χ^2 à un degré de liberté est présente dans les ouvrages statistiques de base (Lellouch et Lazar, 1974 ; Schwartz, 1975 ; Dagnelie, 1975, volume 1). Au risque de première espèce $\alpha = 0{,}05$ l'hypothèse H_0 ne peut être rejetée si la valeur absolue de la différence des log-vraisemblances est inférieure à 1,92.

Annexe C

Analyse référentielle

L'analyse référentielle permet le calcul des paramètres de la distribution initiale d'un MDB à partir des données. Le nombre nécessaire de données est égal au nombre n de paramètres du modèle. La procédure de calcul est séquentielle telle que :

$$\begin{aligned} \boldsymbol{H}_t &= \boldsymbol{G}_t^{-1\prime} \boldsymbol{K}_{t-1} \boldsymbol{G}_t^{-1} \\ \boldsymbol{h}_t &= \boldsymbol{G}_t^{-1\prime} \boldsymbol{k}_{t-1}, \end{aligned}$$

où

$$\begin{aligned} \boldsymbol{K}_t &= \boldsymbol{H}_t + \boldsymbol{F}_t \boldsymbol{F}_t' / V_t \\ \boldsymbol{k}_t &= \boldsymbol{h}_t + \boldsymbol{F}_t Y_t / V_t \end{aligned}$$

avec au temps $t = 1$, $\boldsymbol{H}_1 = \boldsymbol{0}$ et $\boldsymbol{h}_1 = \boldsymbol{0}$. Au temps $t = n$, les paramètres de la distribution initiale $(\boldsymbol{\theta}_t | D_t)$ sont tels que $\boldsymbol{C}_t = \boldsymbol{K}_t^{-1}$ et $\boldsymbol{m}_t = \boldsymbol{K}_t^{-1} \boldsymbol{k}_t$. Si la variance d'observation est constante et inconnue, son estimateur doit être compté dans les paramètres du modèle. Le nombre minimum de données est $n + 1$. Les quantités $\boldsymbol{H}_t$ et $\boldsymbol{h}_t$ sont inchangées avec

$$\begin{aligned} \boldsymbol{K}_t &= \boldsymbol{H}_t + \boldsymbol{F}_t \boldsymbol{F}_t' \\ \boldsymbol{k}_t &= \boldsymbol{h}_t + \boldsymbol{F}_t Y_t \end{aligned}$$

et

$$\begin{aligned} \gamma_t &= \gamma_{t-1} + 1, \\ \delta_t &= \delta_{t-1} + Y_t^2, \end{aligned}$$

avec $\gamma_0 = 0$ et $\delta_0 = 0$. Au temps $t = n + 1$, le calcul de $\boldsymbol{m}_t$ est inchangé et la matrice de variance-covariance des paramètres est $\boldsymbol{C}_t = S_t \boldsymbol{K}_t^{-1}$ où $S_t = d_t / n_t$ avec $n_t = \gamma_t - n$ et $d_t = \delta_t - \boldsymbol{k}_t' \boldsymbol{m}_t$.

L'utilisation des données des n (resp. $n+1$) premières dates dans l'analyse référentielle n'assure pas l'obtention d'estimations valides. En particulier, la matrice

de variances-covariances des paramètres doit être définie positive, et l'estimation de V doit être positive. Des données manquantes ou un problème de colinéarité peuvent entraîner une utilisation de l'analyse référentielle au-delà des n (resp. $n+1$) premières dates.

Les estimations calculées par cette procédure présentent l'avantage d'être totalement indépendantes de l'utilisateur du modèle. L'analyse référentielle assure ainsi la reproductibilité des modélisations réalisées. Toutefois, les estimations calculées de cette manière sont parfois très imprécises. Par exemple, pour le modèle de régression linéaire dynamique appliqué aux données de 1989 présenté dans l'article *Explaining* Dinophysis *cf.* acuminata *abundance in Antifer (Normandy, France) using dynamic linear regression* (cf. page 60), l'analyse référentielle donne des variances estimées des paramètres 10 000 à 100 000 fois plus élevées que celles d'une régression statique. Notre expérience nous ayant montré que ces dernières étaient généralement inférieures à 1, nous avons préféré tenir compte de cette information *a priori* et ne pas utiliser l'analyse référentielle.

Annexe D

Lissage à l'horizon $k = 1$ dans un modèle Harrison-Stevens

Au temps t, les distributions *a posteriori* $(\boldsymbol{\theta}_t|i_t, D_t)$ sont $N(\boldsymbol{m}_{i_t}\,;\boldsymbol{C}_{i_t})$ avec la probabilité p_{i_t}. Pour tout $(i_t, j_{t-1}) \in \{1,2,3,4\}^2$ les distributions lissées sont

$$(\boldsymbol{\theta}_{t-1}|i_t, j_{t-1}, D_t) \sim N(\boldsymbol{a}_{i_t,j_{t-1}}(-1)\,;\boldsymbol{R}_{i_t,j_{t-1}}(-1)).$$

Les paramètres de ces distributions sont tels que

$$\boldsymbol{a}_{i_t,j_{t-1}}(-1) = \boldsymbol{m}_{j_{t-1}} + \boldsymbol{B}_{i_t,j_{t-1}}[\boldsymbol{m}_{i_t} - \boldsymbol{a}_{i_t,j_{t-1}}],$$

et

$$\boldsymbol{R}_{i_t,j_{t-1}}(-1) = \boldsymbol{C}_{j_{t-1}} + \boldsymbol{B}_{i_t,j_{t-1}}[\boldsymbol{C}_{i_t} - \boldsymbol{R}_{i_t,j_{t-1}}]\boldsymbol{B}'_{i_t,j_{t-1}},$$

avec

$$\boldsymbol{B}_{i_t,j_{t-1}} = \boldsymbol{C}_{j_{t-1}}\boldsymbol{G}'_t\boldsymbol{R}^{-1}_{i_t,j_{t-1}}.$$

Les distributions $(Y_{t-1}|i_t, j_{t-1}, D_t)$ sont gaussiennes, de moyennes $f_{i_t,j_{t-1}}(-1) = \boldsymbol{F}'\boldsymbol{a}_{i_t,j_{t-1}}(-1)$ et de variances $Q_{i_t,j_{t-1}}(-1) = \boldsymbol{F}'\boldsymbol{R}_{i_t,j_{t-1}}(-1)\boldsymbol{F} + V_{j_{t-1}}$. La densité de probabilité de $(Y_{t-1}|D_t)$ du modèle multiprocess est la somme des 16 densités $(Y_{t-1}|i_t, j_{t-1}, D_t)$ pondérées par les probabilités $p_{i_t,j_{t-1}}$ calculées lors de l'étape *a posteriori* de la procédure séquentielle d'estimation (cf. page 34). La moyenne $f_t(-1)$ et la variance $Q_t(-1)$ de $(Y_{t-1}|D_t)$ sont, respectivement, la somme des $f_{i_t,j_{t-1}}(-1)$ et des $Q_{i_t,j_{t-1}}(-1)$, pondérées par $p_{i_t,j_{t-1}}$. Les probabilités lissées des quatres états au temps $t-1$ sont $\mathrm{P}(j_{t-1}|D_t) = \sum_{i_t=1}^{4} p_{i_t,j_{t-1}}$. Ces probabilités peuvent être calculées indépendamment de la procédure de lissage.

Bibliographie

Aitsam, A. (1994). Physical and biological background of plankton patchiness. *ICES Coop. Res. Rep.*, 201:3–7.

Ameen, J. R. M. et Harrison, P. J. (1984). Discount weighted estimation. *J. Forecasting*, 3:285–296.

Ameen, J. R. M. et Harrison, P. J. (1985). Normal discount Bayesian model. Dans *Bayesian statistics 2*, Bernardo, J. M., DeGroot, M. H., Lindley, D. V., et Smith, A. F. M., éditeurs, pages 271–298. Elsevier, Amsterdam.

anonyme (1983). Dinoflagellés toxiques et phénomènes d'eaux colorées sur les côtes françaises pendant l'été 1983. *Rev. Trav. Inst. Pêches Marit.*, 47:117–118.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans.*, 53:370–418.

Bayes, T. (1958). Thomas Bayes' essay towards solving a problem in the doctrine of chances. *Biometrika*, 45:293–315. Réimpression de l'article de 1763.

Beltrami, E. et Cosper, E. (1993). Modelling the temporal dynamics of unusual blooms. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 731–735. Elsevier, Amsterdam.

Bensaker, B. (1986). Apprentissage automatique appliqué à l'analyse de données bioclimatiques (détection de *Dinophysis*). Rapport de dea, université du Havre, laboratoire d'analyse et de commande des systèmes, Le Havre.

Berdalet, E. et Estrada, M. (1993). Effects of turbulence on several dinoflagellate species. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 737–740. Elsevier, Amsterdam.

Berland, B. R. et Lassus, P., éditeurs (1997). *Efflorescences toxiques des eaux côtières françaises : écologie, écophysiologie, toxicologie*, volume 13 de *Repéres ocean.* ifremer, Brest.

Berland, B. R., Maestrini, S. Y., Bechemin, C., et Legrand, C. (1994). Photosynthetic capacity of the toxic dinoflagellates *Dinophysis* cf. *acuminata* and *Dinophysis acuta*. *La mer*, 32:107–117.

Berthouex, P. M. et Box, G. E. P. (1996). Time series models for forecasting wastewater treatement plant performance. *Wat. Res.*, 30:1865–1875.

Bhaud, Y. et Soyer-Gobillard, M.-O. (1988). Transmission of gametic nuclei through a fertilization tube during mating in a primitive dinoflagellate, *Prorocentrum micans* Ehr. *J. Cell. Sci.*, 89:197–206.

Bolstad, W. M. (1986a). An efficient algorithm for Harrison-Stevens forecasting using the multiprocess multivariate dynamic linear model. *Commun. Statist.-Simula.*, 15:819–828.

Bolstad, W. M. (1986b). Harrison-Stevens forecasting and the multiprocess dynamic linear model. *The American Statistician*, 40:129–135.

Bolstad, W. M. (1988a). Estimation in the multiprocess dynamic generalized linear model. *Commun. Statist.-Theory Meth.*, 17:4179–4204.

Bolstad, W. M. (1988b). The multiprocess dynamic linear model with biased perturbations: a real time model for growth hormone level. *Biometrika*, 75:685–692.

Bolstad, W. M. (1995). The multiprocess dynamic Poisson model. *J. Am. Stat. Assoc.*, 90:227–232.

Boni, L., Mancini, L., Milandri, A., Poletti, R., Pompei, M., et Viviani, R. (1992). First cases of diarrhoetic poisoning in the northern Adriatic sea. *Sci. Total Environ.*, supplement:419–426.

Bouksim, H., Jousselin, C., Teisseyre, N., et Vignaux, J.-F. (1992). Étude des relations entre des paramètres météorologiques et la concentration en *Dinophysis* au port d'Antifer. Rapport de projet de statistiques, école nationale de la météorologie, Toulouse.

Boutibonnes, L. (1987). Essai de modélisation des variation saisonnières de *Dinophysis sacculus* en baie de Vilaine. Rapport dero-87.13-mr, ifremer, Brest.

Box, G. E. P. et Jenkins, G. M. (1976). *Time series analysis, forecasting and control*. Holden-Day, Oakland, California, édition révisée.

Boyce, F. M. (1974). Some aspects of Great Lakes physics of importance to biological and chemical processes. *J. Fish. Res. Board Can.*, 31:689–730.

Bresson, G. et Pirotte, A. (1995). *Econométrie des séries temporelles*. Presse universitaire de France, Paris.

Brockman, U. H., Eberlein, K., Hosumbeck, P., Trageser, H., Maeir-Reimer, E., Shöne, H. K., et Junge, H. D. (1977). The development of a natural plankton population in an outdoor tank with nutrient-poor sea water I: phytoplankton succession. *Mar. Biol.*, 43:1–17.

Brown, R. L., Durbin, J., et Evans, J. M. (1975). Techniques for testing the constancy of regression relationship over time (with discussion). *J. Roy. Statist. Soc. (Ser. B)*, 37:149–192.

Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30:101–110.

Carlsson, P., Granèli, E., Finenko, G., et Maestrini, S. Y. (1996). Copepod grazing on a phytoplankton community containing the toxic dinoflagellate *Dinophysis acuminata*. *J. Plankton Res.*, 17:1925–1938.

Cassie, R. M. (1962). Microdistribution and other error components of ^{14}C primary production estimates. *Limnol. Oceanogr.*, 7:121–130.

Cazelles, B., Carrat, F., Chau, N. P., et Mary, J.-Y. (1991). Analyses et prédictions de la dynamique de séries phytoplanctoniques : étude de faisabilité. Contrat ifremer-inserm u. 263, Nantes.

Chapelle, A. (1991). *Modélisation d'un écosystème marin côtier soumis à l'eutrophisation : la baie de Vilaine (sud Bretagne). Étude du phytoplancton et du bilan en oxygène*. Thèse de doctorat de troisième cycle, université Paris vi, Paris.

Chatfield, C. (1989). *The analysis of time series: an introduction*. Chapman and Hall, London, quatrième édition.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics. John Wiley and sons, New-York.

Culverhouse, P. F., Simpson, R. G., Ellis, R., Lindley, J. A., Williams, R., Parisini, T., Reguera, B., Bravo, I., Zoppoli, R., Earnshaw, G., McCall, H., et Smith, G. (1996). Automatic classification of field-collected dinoflagellates by artificial neural network. *Mar. Ecol. Prog. Ser.*, 139:281–287.

Daget, J. (1976). *Les modèles mathématiques en écologie*. Collection d'écologie. Masson, Paris.

Dagnelie, P. (1975). *Théorie et méthodes statistiques,* volumes 1 et 2. Presses agronomiques de Gembloux, Gembloux, seconde édition.

De Cremoux, F. (1988). Recherche des facteurs conditionnant le développement de *Dinophysis* par simulation. Rapport dero-88.07-mr, ifremer, Nantes.

Delmas, D., Herbland, A., et Maestrini, S. Y. (1992). Environmental conditions which lead to increase in cell density of the toxic dinoflagellates *Dinophysis* spp. in nutrient-rich and nutrient poor waters of the French Atlantic coast. *Mar. Ecol. Prog. Ser.*, 89:53–61.

Delmas, D., Herbland, A., et Maestrini, S. Y. (1993). Do *Dinophysis* spp. come from the open sea along the French Atlantic coast. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 489–494. Elsevier, Amsterdam.

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algoritm (with discussion). *J. Roy. Statist. Soc. (Ser. B)*, 39:1–38.

Diggle, P. J. et Zeger, S. L. (1989). A non-gaussian model for time series with pulse. *J. Am. Stat. Assoc.*, 84:354–359.

Draper, N. R. et Smith, H. (1966). *Applied regression analysis.* Wiley series in probability and mathematical statistics. John Wiley and sons, New-York.

Duncan, D. B. et Horn, S. D. (1982). Linear dynamic recursive estimation from the viewpoint of regression analysis. *J. Am. Stat. Assoc.*, 67:815–821.

Durand-Clement, M., Clement, J.-C., Moreau, A., Jeanne, N., et Puiseux-Dao, S. (1988). New ecological and ultrastructural data on the dinoflagellate *Dinophysis* sp. from the French coast. *Mar. Biol.*, 97:37–44.

El-Shaarawi, A. H., Esterby, S. R., et Dukta, B. J. (1981). Bacterial density in water determined by Poisson or negative binomial distributions. *Appl. Environ. Microb.*, 41:107–116.

Engle, R. F. (1982). Autoregressive conditionnal heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, 50:987–1008.

Faure, A. et Zhang, S. (1990). Modèles prédictifs de manifestation d'efflorescences de *Dinophysis acuminata*. Rapport de recherche, université du Havre, cellule de suivi du littoral haut normand et laboratoire d'analyse et de commande des systèmes, Le Havre.

Faust, M. A. (1993a). Alternate asexual reproduction of *Prorocentrum lima* in culture. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 115–120. Elsevier, Amsterdam.

Faust, M. A. (1993b). Sexuality in a toxic dinoflagellate, *Prorocentrum lima*. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 121–126. Elsevier, Amsterdam.

Fildes, R. (1983). An evaluation of Bayesian forecasting. *J. Forecasting*, 2:137–150.

Gentien, P., Lunven, M., Lehaître, M., et duvent, J. L. (1992). *In-situ* depth profiling of particle size. *Deep-Sea Res.*, 42:1297–1312.

Gourieroux, C. et Monfort, A. (1990). *Séries temporelles et modèles dynamiques*. Economica, Paris.

Granèli, E., Anderson, D. M., Maestrini, S. Y., et Paasche, E. (1993). Light and dark carbon fixation by the marine dinoflagellate genera *Dinophysis* and *Ceratium*. *ICES Mar. Sci. Symp.*, 197:274. Résumé uniquement. Un article est sous-presse dans *Aquat. Microb. Ecol.*

Hallegraeff, G. M. et Lucas, I. A. N. (1988). The marine dinoflagellate genus *Dinophysis* (Dinophyceae): photosynthetic, neretic and non-photosynthetic, oceanic species. *Phycologia*, 27:25–42.

Harris, G. P. et Smith, R. E. H. (1977). Observation of small-scale spatial patterns in phytoplankton populations. *Limnol. Oceanogr.*, 22:887–89.

Harrison, P. J. et Stevens, C. F. (1971). A Bayesian approach to short term forecasting. *Oper. Res. Quart.*, 22:341–362.

Harrison, P. J. et Stevens, C. F. (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc. (Ser. B)*, 38:205–247.

Honsell, G., Boni, L., Cabrini, M., et Pompei, M. (1992). Toxic or potentially toxic dinoflagellates from the northern Adriatic sea. *Sci. Total Environ.*, supplement:107–114.

Ishimaru, T., Inoue, H., Fukuyo, Y., Ogata, T., et Kodama, M. (1988). Cultures of *Dinophysis fortii* and *Dinophysis acuminata* with the cryptomonad *Plagioselmis* sp. Dans *Mycotoxins and phycotoxins*, Aibara, K. et al., éditeurs, pages 19–21. Université de Tokyo.

Jacobson, D. M. et Andersen, R. A. (1994). The discovery of mixotrophy in photosynthetic species of *Dinophysis* (Dinophyceae): light and electron microscopical observations of food vacuoles in *Dinophysis acuminata*,

Dinophysis norvegica and two heterotrophic dinophysoid dinoflagellates. *Phycologia*, 33:97–110.

Jassby, A. D. et Powell, M. (1990). Detecting changes in ecological time series. *Ecology*, 71:2044–2052.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. Academic press, New-York.

Johnston, F. R. et Harrison, P. J. (1980). An application of forecasting in the alcoholic drink industry. *J. Oper. Res. Soc.*, 31:699–709.

Johnston, F. R., Harrison, P. J., Marshall, A. S., et France, K. M. (1986). Modelling and the estimation of changing relationships. *The Statistician*, 35:229–235.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82:34–45.

Kalman, R. E. et Bertram, J. E. (1960a). Control system analysis and design via the second method of Lyapunov: I continuous-time systems. *J. Basic Eng.*, 82:371–393.

Kalman, R. E. et Bertram, J. E. (1960b). Control system analysis and design via the second method of Lyapunov: II discrete-time systems. *J. Basic Eng.*, 82:394–400.

Kamykowski, D. (1995). Trajectories of autotrophic marine dinoflagellates. *J. Phycol.*, 31:200–208.

Kamykowski, D. et Zentara, S.-J. (1977). The diurnal vertical migration of motile phytoplankton trough temperature gradients. *Limnol. Oceanogr.*, 22:148–151.

Kendall, M. et Stuart, A. (1977). *The advanced theory of statistics*, volume 2. Charles Griffin and company, London, quatrième édition.

Kumagai, M., Yanagi, T., Murata, M., Yasumoto, T., Kat, M., Lassus, P., et Rodriguez-Vasquez, J.-A. (1986). Identification of okadaic acid as the causative toxin of diarrhetic shellfish poisoning in Europe. *Agric. Biol. Chem.*, 50:2853–2857.

Lagadeuc, Y. (1992). Transport larvaire en Manche. Exemple de *Pectinaria koreni* (Malgrem), annélide polychète en baie de Seine. *Oceanol. Acta*, 15:383–395.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems (with discussion). *J. Roy. Statist. Soc. (Ser. B)*, 57:613–658.

Larrazabal, M. E., Lassus, P., Maggi, P., et Bardouil, M. (1990). Kystes modernes de dinoflagellés en baie de Vilaine-Bretagne sud (France). *Cryptogam. Algol.*, 11:171–185.

Lassus, P., Bardouil, M., Berthomé, J.-P., Maggi, P., Truquet, P., et Le Déan, L. (1988). Seasonal occurence of *Dinophysis* sp. along the French Atlantic coast between 1983 and 1987. *Aquat. Living Ressour.*, 1:155–164.

Lassus, P., Martin, A.-G., Maggi, P., Berthomé, J.-P., Langlade, A., et Bachère, E. (1983). Extension du dinoflagellé *Dinophysis acuminata* en Bretagne sud et conséquence pour les cultures marines. *Rev. Trav. Inst. Pêches Marit.*, 47:122–133.

Lassus, P., Proniewski, F., Maggi, P., Truquet, P., et Bardouil, M. (1993). Wind-induced toxic blooms of *Dinophysis* cf. *acuminata* in the Antifer area (France). Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 519–523. Elsevier, Amsterdam.

Lassus, P., Proniewski, F., Pigeon, C., Veret, L., Le Déan, L., Bardouil, M., et Truquet, P. (1990). The diurnal vertical migration of *Dinophysis* cf. *acuminata* in an outdoor tank at Antifer (Normandy, France). *Aquat. Living Ressour.*, 3:143–145.

Le, N. D., Martin, R. D., et Raftery, A. E. (1996). Modelling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *J. Am. Stat. Assoc.*, 91:1504–1515.

Le Batteux, V. et Nizard, G. (1990). Blooms de *Dinophysis acuminata* : bilan des années 1989-1990 sur le site d'Antifer. Rapport de stage, laboratoire municipal d'hygiène du Havre, Le Havre.

Ledbetter, M. (1979). Langmuir circulation and plankton patchiness. *Ecol. Model.*, 7:289–310.

Legendre, L. et Demers, S. (1984). Towards dynamic biological oceanography and limnology. *Can. J. Fish. Aquat. Sci.*, 42:2–19.

Lellouch, J. et Lazar, P. (1974). *Méthodes statistiques en expérimentation biologique*. Médecine-Sciences. Flammarion, Paris.

Lewis, W. M. (1986). Evolutionary interpretation of allelochemical interactions in phytoplankton algae. *Am. Nat.*, 127:184–194.

Mackas, D. L., Denman, K. L., et Abott, M. R. (1985). Plankton patchiness: biology in the physical vernacular. *Bull. Mar. Sci.*, 37:652–674.

MacKenzie, L. (1992). Does *Dinophysis* (dinophyceae) have a sexual life cycle? *J. Phycol.*, 28:399–406.

Maestrini, S. Y., Berland, B. R., Dalet, C., et Lassus, P. (1997). *Dinophysis* spp. Dans *Efflorescences toxiques des eaux côtières françaises : écologie, écophysiologie, toxicologie*, Berland, B. R. et Lassus, P., éditeurs, volume 13 de *Repéres ocean.* ifremer, Brest.

Marcaillou-Le Baut, C. (1990). Les toxines du poison diarrhéique. *Océanis*, 16:359–373.

Marcaillou-Le Baut, C. et Lassus, P. (1983). Synthèse des connaissances sur les efflorescences estivales de *Dinophysis* et *Gyrodinium*. *Rev. Trav. Inst. Pêches Marit.*, 47:119–121.

Marcaillou-Le Baut, C., Lucas, D., et Le Déan, L. (1985). *Dinophysis acuminata* toxin: status of toxicity bioassays in France. Dans *Toxic dinoflagellates*, Anderson, D. M., White, A. W., et Baden, D. G., éditeurs, pages 485–488. Elsevier, Amsterdam.

Marr, J. C., Jackson, A. E., et McLachlan, J. L. (1992). Occurrence of *Prorocentrum lima*, a DSP toxin-producing species from the Atlantic coast of Canada. *J. Appl. Phycol.*, 4:17–24.

Matheron, G. (1972). Théorie des variables régionnalisées. Dans *Traité d'informatique géologique*, Laffitte, P., éditeur. Masson, Paris.

McKenzie, E. (1984). General exponential smoothing and the equivalent ARMA process. *J. Forecasting*, 3:333–344.

McLachlan, J. L. (1993). Evidence for sexuality in a species of *Dinophysis*. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 143–146. Elsevier, Amsterdam.

Meinhold, R. J. et Singpurwala, N. D. (1983). Understanding the Kalman filter. *J. Am. Stat. Assoc.*, 84:479–486.

Meinhold, R. J. et Singpurwala, N. D. (1989). Robustification of the Kalman filter. *The American Statistician*, 37:123–127.

Mer, G. (1986). Synthèse des connaissances sur les facteur pouvant influencer le developpement des blooms de dinoflagellés toxiques. Traitements mathématiques des données hydrologiques acquises en baies de Vilaine en 1984 et 1985. Rapport dero-86.08-mr, ifremer, Nantes.

Migon, H. S. et Harrison, P. J. (1985). An application of non-linear Bayesian forecasting to television advertising. Dans *Bayesian statistics 2*, Bernardo, J. M., DeGroot, M. H., Lindley, D. V., et Smith, A. F. M., éditeurs, pages 681–696. Elsevier, Amsterdam.

Ménesguen, A., Guillaud, J.-F., Aminot, A., et Hoch, T. (1995). Modelling the entrophication process in a river plume: the Seine case study. *Ophelia*, 42:205–225.

Ménesguen, A., Lassus, P., De Cremoux, F., et Boutibonnes, L. (1990). Modelling *Dinophysis* blooms: a first approach. Dans *Toxic marine phytoplankton*, Granèli, E., Sundström, B., Edler, L., et Anderson, D. M., éditeurs, pages 195–200. Elsevier, Amsterdam.

Moita, M. T. et Sampayo, M. A. (1993). Are there cysts in the genus *Dinophysis*? Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 153–157. Elsevier, Amsterdam.

Monbet, Y. (1975). Les incidences écologiques de la construction du terminal pétrolier d'Antifer. Rapport cnexo-unité littoral, ifremer, Brest.

Muirhead, C. R. (1986). Distinguishing outlier types in time series. *J. Roy. Statist. Soc. (Ser. B)*, 48:39–47.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *J. Roy. Statist. Soc. (Ser. B)*, 40:1–42.

Paulmier, G. et Joly, J. P. (1983). Manifestation de *Dinophysis acuminata* sur le littoral normand. *Rev. Trav. Inst. Pêches Marit.*, 47:149–157.

Pierre, M.-J. et Lassus, P. (1983). Perturbation des écosystèmes en baie de Vilaine : analyse des successions phytoplanctoniques précédant l'apparition d'un dinoflagellé toxique. *Rev. Trav. Inst. Pêches Marit.*, 47:134–148.

Platt, T. (1972). Local phytoplankton abundance and turbulence. *Deep-Sea Res.*, 19:183–187.

Platt, T. (1975). The physical environment and spatial structure of phytoplankton populations. *Mém. Soc. R. Sci. Liège (Sér. 6)*, 7:9–17.

Pole, A., West, M., et Harrison, P. J. (1994). *Applied Bayesian forecasting and time series analysis*. Chapman and Hall, New-York.

Puel, O., Galgani, F., Dalet, C., et Lassus, P. (sous presse). Partial sequence of the 24S rRNA and PCR-based assay for the toxic dinoflagellate *Dinophysis acuminata*. *Can. J. Fish. Aquat. Sci.*

Rausch de Traubenberg, C. et Morlaix, M. (1995). Evidence of okadaic acid release into extracellular medium in culture of *Prorocentrum lima* (Ehrenberg) Dodge. Dans *Harmful marine algal blooms*, Lassus, P., Arzul, E., Erard, E., Gentien, P., et Marcaillou, C., éditeurs, pages 493–498. Loivoisier, Paris.

Salomon, J.-C. et Breton, M. (1993). An atlas of long-term currents in the Channel. *Oceanol. Acta*, 16:439–448.

Sampayo, M. A. (1993). Trying to cultivate *Dinophysis* spp. Dans *Toxic phytoplankton bloom in the sea*, Smayda, T. J. et Shimizu, Y., éditeurs, pages 807–810. Elsevier, Amsterdam.

Séchet, V., Safran, P., Hovgaard, P., et T., Yasumoto. (1990). Causative species of diarrhetic shellfish poisonning (DSP) in Norway. *Mar. Biol.*, 105:269–274.

Schwartz, D. (1975). *Méthodes statistiques à l'usage des médecins et des biologistes*. Médecine-Sciences. Flammarion, Paris, troisième édition.

Shumway, R. H. et Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, 3:253–264.

Siegel, S. (1956). *Non-parametrics statistics for the behavioral sciences*. McGraw-Hill series in psychology. McGraw-Hill, New-York.

Smayda, T. J. (1990). Novel and nuisance phytoplankton blooms in the sea: evidence for a global epidemic. Dans *Toxic marine phytoplankton*, Granèli, E., Sundström, B., Edler, L., et Anderson, D. M., éditeurs, pages 29–40. Elsevier, Amsterdam.

Smith, A. F. M. et West, M. (1983). Monitoring renal transplants: an application of the multi-process Kalman filter. *Biometrics*, 39:867–878.

Smith, A. F. M., West, M., Gordon, K., Knapp, M. S., et Trimble, I. M. G. (1983). Monitoring kidney transplant patients. *The Statistician*, 32:46–54.

Soudant, D. (1990). Étude statistique des conditions environnementales conduisant aux efflorescences à *Dinophysis*. Rapport de stage, institut universitaire de technologie de Vannes, département statistique et traitements informatique des données, Vannes.

Soudant, D. (1993). Application de modèles dynamiques bayésiens à la prévision des efflorescences à *Dinophysis* à Antifer (baie de Seine). Rapport del/93.10, ifremer, Nantes.

Sournia, A., Belin, C., Berland, B. R., Erard-Le Denn, E., Gentien, P., Grzebyk, D., Marcaillou-Le Baut, C., Lassus, P., et Partensky, F. (1991). *Le phytoplancton nuisible des côtes de France*. ifremer-cnrs, Brest.

Taggart, C. T. et Legget, W. C. (1987). Wind-forced hydrodynamics and their interaction with larval fish and plankton abundance: a time-series analysis of physical-biological data. *Can. J. Fish. Aquat. Sci.*, 44:438–451.

Taylor, F. R. J., éditeur (1987). *The biology of dinoflagellates*, volume 21 de *Botanical monographs*. Blakwell, Oxford.

Taylor, P. F. et Thomas, M. E. (1982). Short term forecasting: horses for courses. *J. Oper. Res. Soc.*, 33:685–694.

Thiébaut, E., Dauvin, J.-C., et Lagadeuc, Y. (1994). Horizontal distribution and retention of *Owenia fusiformis* larvae (annelida: Polychaeta) in the bay of Seine. *J. Mar. Biol. Assoc. UK*, 74:129–142.

Tiwari, R. C. et Dienes, T. P. (1994). The Kalman filter and Bayesian outlier detection for time series analysis of BOD data. *Ecol. Model.*, 82:159–165.

Tomassone, R., Lesquoy, E., et Millier, C. (1983). *La régression : nouveaux regards sur une ancienne méthode statistique*, volume 13 de *Actualités scientifiques et agronomiques de l'inra*. Masson, Paris.

Truquet, P., Lassus, P., Honsell, G., et Le Déan, L. (1996). Application of a digital pattern recognition system to *Dinophysis acuminata* and *Dinophysis sacculus*. *Aquat. Living Ressour.*, 9:273–279.

Turgeon, J., Cembella, A. D., Therriault, J.-C., et Beland, P. (1990). Spatial distribution of resting cysts of *Alexandrium* sp. in sediments of the lower St. Lawrence estuary and the Gaspé coast (eastern Canada). Dans *Toxic marine phytoplankton*, Granèli, E., Sundström, B., Edler, L., et Anderson, D. M., éditeurs, pages 239–243. Elsevier, Amsterdam.

Utermöhl, H. (1955). Zur Vervollkommnung der quantitativen Phytoplankton Methodik. *Int. Ver. Theor. Angew. Limnol. Verh.*, 9:1–38.

Viera Da Silva, J. (1979). *Introduction à la théorie écologique*. Collection d'écologie. Masson, Paris.

Waller, L. A., Carlin, B. P., Xia, H., et Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease. *J. Am. Stat. Assoc.*, 92:607–617.

West, M. (1985). Generalized linear models: scale parameters, outlier accomodation and prior distributions. Dans *Bayesian statistics 2*, Bernardo, J. M., DeGroot, M. H., Lindley, D. V., et Smith, A. F. M., éditeurs, pages 531–558. Elsevier, Amsterdam.

West, M. (1995). Bayesian inference in cyclical component dynamic linear models. *J. Am. Stat. Assoc.*, 90:1301–1312.

West, M. et Harrison, P. J. (1997). *Bayesian forecasting and dynamic models*. Springer series in statistics. Springer-Verlag, New-York, seconde édition.

West, M., Harrison, P. J., et Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Am. Stat. Assoc.*, 80:73–97.

Wright, J. L. C. (1994). Diarrhetic shellfish poisons: how they are made by *Prorocentrum lima* and why. *Can. Tech. Rep. Fish. Aquat. Sci.*, 2016:66. Résumé seulement.

Yatawara, N., Abraham, B., et MacGregor, J. F. (1991). A Kalman filter in the presence of outliers. *Commun. Statist.-Theory Meth.*, 20:1803–1820.

Young, P. et Young, T. (1992). Environmetrics methods of nonstationary time-series analysis : univariate methods. Dans *Methods of environmental data analysis*, Hewitt, C. N., éditeur, pages 37–77. Elsevier, London.

Zehnwirth, B. (1988). A generalization of the Kalman filter for models with state-dependent observation variance. *J. Am. Stat. Assoc.*, 83:164–167.

Index des auteurs

Index des sujets

Liste des tableaux

Table des figures

Table des matières