Arbres de Régression : Modélisation Non Paramétrique et Analyse des Séries Temporelles

Anne-Emmanuelle Badel, □ *Olivier Michel* □ *et Alfred Hero* ◊

□ Laboratoire de Physique (URA 1325 CNRS), École Normale Supérieure de Lyon,

46 allée d'Italie, 69364 Lyon Cedex 07, France

tel : 72 72 85 59 - fax : 72 72 80 80

◊ Department of Electrical Engineering and Computer Science,

University of Michigan, Ann Arbor, MI 48109-2122, USA

e-mail : aebadel@physique.ens-lyon.fr, omichel@physique.ens-lyon.fr, hero@eecs.umich.edu

7 Octobre 1996

1 Introduction

Ces dernières années, un nombre très important d'études et de publications se sont portées sur les systèmes dynamiques non linéaires, et plus particulièrement sur les systèmes chaotiques, dont le comportement "imprédictible" [1] - bien que régi par un système d'équations déterministes - ne peut être expliqué et analysé par des méthodes classiques de traitement du signal. La grande variété des comportements non linéaires [1, 2, 3] semble *ab initio* vouer à l'échec toute tentative pour identifier directement des modèles. Les différentes études menées sur ces systèmes se sont par conséquent focalisées sur la détermination expérimentale de caractéristiques "chaotiques" à partir de séries temporelles observées [4, 5, 6, 7, 8], le problème de la détection étant alors le plus souvent éludé. La sensibilité au bruit de tels systèmes, la diversité des comportements pouvant conduire à des "signatures" de chaos et l'absence de toute méthode expérimentale totalement fondée permettant sa détection nécessitent de mettre en œuvre des outils d'analyse multiples : il est nécessaire de toujours s'attacher à pallier ce manque de méthode strictement adaptée à

[1] Il s'agit ici d'imprédictibilité à long terme; cette notion sera reprise et discutée plus tard.

la détection - et donc à l'estimation - du chaos, par la redondance des représentations que l'on peut en faire, ainsi que par la variété des tests que l'on peut y appliquer. Il faut noter que d'un point de vue plus théorique, de nombreuses études portent sur la détermination analytique de ces mêmes caractéristiques (existence d'une mesure invariante [9], calcul des exposants de Lyapunov [6], dimensions généralisées [10]) pour des systèmes dont les équations de fonctionnement sont parfaitement connues. Ces équations ne sont bien sûr pas toujours identifiables dans les situations expérimentales.

L'objet de cet article est de proposer une nouvelle méthode de représentation d'un système dynamique, la seule hypothèse formulée étant de considérer que le système est observé en régime "stationnaire", au sens où il reste localisé dans une partie bornée de son espace des états et où il n'est pas en régime transitoire. Le travail présenté s'appuie sur une méthode de partition de l'espace des phases du système reconstruit par la méthode des retards [11, 12, 13]. L'objet de cette partition est de diviser l'espace des phases en un ensemble de cellules au sein desquelles l'ensemble des données (ou réalisations du vecteur d'état) apparaît comme un ensemble de réalisations d'un processus aléatoire, n'ayant pas de dynamique propre : l'ensemble des réalisations ne semble pas décrire localement une trajectoire mais ressemble plutôt à un "nuage". Enfin nous étudions comment la partition peut être exploitée dans le contexte de modélisation et d'analyses des séries temporelles.

Dans une première partie, les problèmes liés à la modélisation non linéaire sont soulignés afin de préciser les motivations d'une nouvelle approche ; le contexte dans lequel cette méthode est développée est ensuite décrit brièvement : les méthodes d'estimation des paramètres, indispensables pour une reconstruction correcte de l'espace d'état du système à partir d'une série temporelle observée[2], y sont rappelées. Le principe et la mise en œuvre de l'algorithme de partition récursive finalement adopté, sont détaillés dans la section suivante. Les critères utilisés pour itérer ou arrêter la construction de la partition, basés sur une statistique empirique locale (dans l'espace des phases reconstruit) y sont discutés. La quatrième partie de cette étude est consacrée à la mise en œuvre et à l'application de mé-

[2]Les séries temporelles seront ici considérées comme discrètes que ce soit une propriété intrinsèque ou liée à un échantillonnage.

thodes de prédiction fondées sur l'utilisation des partitions obtenues et considérées alors en tant qu'arbre de régression. L'intérêt de cette approche non paramétrique est illustré dans le cadre de la détermination (sans *a priori*) des paramètres de reconstruction de l'espace des phases des systèmes dynamiques. Enfin dans la dernière partie, nous soulignons le potentiel de cette approche dans le contexte de l'estimation de modèles auto-régressifs à seuils.

2 Description du problème

2.1 Modèles non linéaires

Dire d'un système ou d'un signal qu'il est non linéaire n'est pas suffisant pour caractériser ce dernier tant il est vrai qu'il existe une infinité de comportements ou de propriétés conduisant à sortir du cadre simplement linéaire. Il existe par conséquent de très nombreuses approches permettant de modéliser des comportements non linéaires [1, 2, 3]. Parmi celles-ci, la plus générale sinon la plus naturelle prend comme point de départ un modèle linéaire de type $ARMA(p,q)$ (modèle auto-régressif à moyenne ajustée d'ordres p et q) :

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + \sum_{j=0}^{q} b_j e_{n-j} \qquad (1)$$

pour lequel $\{e_n\}$ est un "bruit" d'entrée de variance σ_e^2 (le plus souvent à distribution normale) et les coefficients $\{a_i, b_j\}$ sont habituellement constants. Soient

$$\underline{\mathbf{X}}_n^{(k)} = [x_{n-1}\ x_{n-2}\ \ldots\ x_{n-k}]^T$$

et

$$\underline{\mathbf{E}}_n^{(k')} = [e_{n-1}\ e_{n-2}\ \ldots\ e_{n-k'}]^T$$

les vecteurs construits à partir des k (respectivement k') valeurs passées de x_n (respectivement e_n). Différents modèles non linéaires ont été proposés correspondant à l'introduction d'une forme de dépendance des coefficients de (1) vis-à-vis de $\underline{\mathbf{X}}_n^{(k)}$ et $\underline{\mathbf{E}}_n^{(k')}$:

- les modèles à seuils pour lesquels

$$\left\{\begin{array}{l} a_i = a_i^{(j)} \\ \\ b_i = b_i^{(j)} \end{array}\right. \quad \text{si } x_{n-d} \in I_j \tag{2}$$

  où $\{I_j\}_{j=1,\ldots,r}$ représente une partition de $\mathbb{R}$ en r segments [1, 2, 3]. Ces modèles ont été introduits pour décrire les comportements de cycles limites que ne peuvent décrire les modèles linéaires [3] ;

- les modèles bilinéaires [1, 2, 3] ont été introduits pour rendre compte d' "explosions" (ou variations brutales et importantes de l'amplitude) et correspondant à des coefficients a_i et b_i combinaisons linéaires des composantes de $\underline{\mathbf{X}}_n^{(k)}$ et $\underline{\mathbf{E}}_n^{(k')}$:

$$\left\{\begin{array}{l} a_i = \sum_{l=1}^{k} \alpha_l x_{n-l} + \sum_{l'=1}^{k'} \alpha_{l'}^{*} e_{n-l'} \\ b_i = \sum_{m=1}^{k} \beta_m x_{n-m} + \sum_{m'=1}^{k'} \beta_{m'}^{*} e_{n-m'} \end{array}\right. \tag{3}$$

  où $\{\alpha_i, \beta_i, \alpha_i^*, \beta_i^*\}$ est un ensemble de coefficients constants. Une généralisation de (3), où les coefficients $\alpha_i, \beta_i, \alpha_i^*, \beta_i^*$ sont des fonctions polynomiales, conduit aux modèles polynomiaux [1, 2, 3] ;

- Dans le cadre des modèles hétéroscédastiques, la dépendance ne s'exprime pas directement sur les coefficients mais sur la variance du bruit. Cela permet de tenir compte des cumulants d'ordre supérieur à 2 : ces derniers sont nuls pour un bruit gaussien de variance constante, une variance non constante implique l'existence de cumulants d'ordre supérieur à 2 non nuls. Ces modèles sont notamment utilisés en économie [3].

Le problème de l'identification des ordres p et q du modèle s'avère très complexe ici. En particulier, il n'est pas indépendant du choix des fonctions exprimant les coefficients $\{a_i, b_i\}$ en fonction de $\underline{\mathbf{X}}_n^{(k)}$ et $\underline{\mathbf{E}}_n^{(k')}$. Cependant il existe quelques discussions [2, 14] pouvant guider le choix du type de modèle non linéaire et permettant de rendre compte au mieux des propriétés du signal étudié :

- représentation des marginales bivariées ou densités de probabilité conjointe de x_n et x_{n-k} [3] : la forme de l'espace occupé par ces marginales bivariées conduit à privilégier un type de modèle, par exemple, un ellipsoïde correspond à un modèle linéaire, une accumulation indique plutôt un modèle polynomial ;

- tests de linéarité ou de validité d'un type de modèle (cf. [14] pour une revue de quelques tests) : par exemple les propriétés des multispectres permettent de détecter la présence d'une non linéarité de type polynomiale [1].

Malgré l'existence de ces tests orientant vers un type de modèle, les choix d'un modèle paramétrique et de l'ordre de ce dernier sont déterminés à partir d'un ensemble d'*a priori* et d'heuristiques et ne font pas, à notre connaissance, l'objet d'études systématiques.

2.2 Approche statistique et systèmes dynamiques

Nous proposons dans cette étude une approche non paramétrique de représentation d'un signal non linéaire. Cependant nous nous placerons toujours dans le cadre de deux hypothèses restrictives fondamentales :

- La série temporelle étudiée $\{x_n\}$ doit pouvoir être considérée comme une observation (ou mesure) d'un système régi par l'équation d'état d'un système dynamique :

$$\underline{\mathbf{X}}_{n+1} = F_{\underline{\mathbf{X}}_n}(\underline{\mathbf{X}}_n) + \epsilon_n \tag{4}$$

  où $\underline{\mathbf{X}}_n$ représente le vecteur d'état du système (cf. [15] et paragraphe 2.3). Ce vecteur d'état n'est pas forcément observable mais supposé de faible dimension. $F_{\underline{\mathbf{X}}_n}$ est une fonction vectorielle $\mathbb{R}^d \rightarrow \mathbb{R}^d$, non linéaire dans le cas général, dépendant de l'état courant $\underline{\mathbf{X}}_n$ et décrivant l'évolution dynamique du système.

- L'ensemble des états qui décrivent le système pendant la durée d'observation correspond à l'évolution du système en régime permanent (toute notion de transitoire sera exclue). De plus, le bruit (d'état au sens où il affecte $\underline{\mathbf{X}}_{n+1}$) est supposé suffisamment faible pour ne pas conduire à une trajectoire quittant le bassin d'attraction associé au cycle limite sur lequel le système évolue (ce qui aurait pour conséquences de réintroduire un régime transitoire).

Ces deux hypothèses appellent les remarques suivantes :

- L'expérience ne donne quasiment jamais accès aux différentes variables correspondant à l'ensemble des degrés de liberté du système mais uniquement à quelques observables. De plus, la dynamique du système n'est connue qu'à partir de données échantillonnées. Il convient donc d'obtenir une représentation (à un petit nombre de degrés de liberté) de la dynamique dans un espace des phases reconstruit à partir des données expérimentales observées. Cette reconstruction fera l'objet du paragraphe 2.3.

- La détermination d'un modèle de type (4) est essentiellement motivée par la volonté de prédire $\underline{\mathbf{X}}_{n+1}$ à partir de l'observation de $\underline{\mathbf{X}}_n$. De telles prédictions nécessitent la connaissance en moyenne temporelle d'un ensemble de trajectoires associées à diverses conditions initiales. Une autre approche consiste à substituer aux moyennes temporelles des moyennes d'ensemble calculées à partir de la distribution de probabilité sur l'espace des états. Cette approche repose sur l'existence d'une mesure invariante sur l'attracteur :

$$\overline{\mu}(\underline{\mathbf{X}}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N-1} \delta[\underline{\mathbf{X}} - F^i(\underline{\mathbf{X}}_0)] \tag{5}$$

  La validité et la pertinence de cette approche sont à la base de la théorie ergodique développée par J.P. Eckmann et D. Ruelle [9].

Il existe peu de situations où la mesure invariante est connue : expérimentalement nous utiliserons comme mesure la mesure naturelle c'est-à-dire la probabilité que $\underline{\mathbf{X}}_n$ se trouve dans un voisinage $V(\underline{\mathbf{X}}_0)$ donné de $\underline{\mathbf{X}}_0$ appartenant à l'attracteur [7, 9]. Les paragraphes 2.3 et 3 proposent une méthode d'estimation de cette distribution de probabilité.

2.3 Reconstruction de l'espace des phases

Les vecteurs d'état sont définis comme l'ensemble des d composantes correspondant aux d degrés de liberté du système. Il est donc théoriquement nécessaire d'avoir autant de mesures indépendantes que le système possède de degrés de liberté. En général, on considérera que l'on ne dispose que d'une série échantillonnée issue de l'observation expérimentale d'une grandeur fluctuante et à partir de laquelle l'espace des phases du système devra être estimé. Une estimation possible consiste à définir l'ensemble des vecteurs :

$$\underline{\mathbf{X}}_n = [x_n \ \dot{x}_n \ \ddot{x}_n \ \ldots]^T \tag{6}$$

à l'aide de différences finies. L'inconvénient de cette méthode est sa forte sensibilité au bruit. H. Whitney [11] a proposé une alternative en considérant les vecteurs:

$$\underline{\mathbf{X}}_n = [x_n \ x_{n-\tau_1} \ \ldots \ x_{n-\tau_{d-1}}]^T \tag{7}$$

où d est la dimension de l'espace estimé et $\{\tau_i\}_{i=1,\ldots,d-1}$ un ensemble de retards. d doit vérifier $d \geq 2D+1$ où D est la dimension de l'espace d'état [11]. N.H. Packard, J.P. Crutchfield, J.D. Farmer et R. Shaw ont appliqué cette méthode dans le cadre des systèmes chaotiques [12]. Un système dynamique sera dit chaotique s'il s'agit d'un système dynamique déterministe dont le comportement est très dépendant des conditions initiales au sens où des trajectoires initiales proches divergent exponentiellement. La validité de l'approche dans le cadre des systèmes chaotiques a été justifiée par F. Takens [13] : pour une série infinie, pour des valeurs quelconques des τ_i et avec la condition suffisante sur la dimension d :

$$d \geq 2D_2 + 1 \tag{8}$$

où D_2 est la dimension de corrélation [16], l'espace constitué par l'ensemble des vecteurs (7) est équivalent topologiquement et dynamiquement (à une transformation difféomorphique[3] près) à l'espace des phases réel au sens où les propriétés géométriques de l'attracteur sont conservées.

Du point de vue expérimental (i.e. à partir de séries temporelles à durée limitée), l'espace des phases est reconstruit par "plongement" suivant [12] en prenant $\tau_i = i\tau$ pour les vecteurs (7) ; le choix de τ n'est alors plus indifférent.

[3] Une fonction f est un homéomorphisme si elle est bijective et continue ainsi que son inverse f^{-1} ; si de plus elle est différentiable, il s'agit d'un difféomorphisme.

- Importance du choix du retard τ :

  Dans le cas des observations non bruitées sur des temps infinis, le choix du retard τ est indifférent [13] mais il devient critique [17, 18] lorsque la durée d'observation est limitée comme le montrent la figure 1 pour le système de Rössler[4] :

  – Si τ est trop faible, toutes les coordonnées sont fortement corrélées : $x_k \approx x_{k+1}$, les vecteurs définis par (7) avec $\tau_i = i\tau$ sont presque colinéaires, l'espace des phases tend à être une droite et la dimension estimée tend vers 1.

  – Si au contraire τ est trop grand, les coordonnées sont presque indépendantes. L'ensemble des vecteurs $\{\underline{\mathbf{X}}_n\}_{n=1,\ldots,N}$ parcourt la quasi-totalité de l'espace des phases : le système est proche d'un bruit vectoriel à coordonnées indépendantes et la dimension estimée tend vers la valeur de la dimension de reconstruction.

  La méthode la plus couramment utilisée consiste à choisir τ à la valeur du premier zéro de la fonction d'auto-corrélation estimée :

  $$C(\tau) = \frac{1}{N}\sum_{n=0}^{N} x_n x_{n+\tau} \tag{10}$$

  Il a aussi été proposé de prendre le premier minimum de l'information mutuelle, cette méthode est exposée et discutée dans [7, 8, 19, 20] et pour les systèmes à forte périodicité, les deux méthodes fournissent des résultats pratiquement identiques [10, 18].

- Choix de la dimension d :

  Soit $d_0 \in \mathbb{N}$ la plus petite dimension vérifiant (8). Si $d > d_0$, d'après le théorème de F. Takens, toute la dynamique du système est prise en considération. En notant $L = \max_n x_n - \min_n x_n$ et $N_{d\tau} = N - (d-1)\tau$, la densité d'observations par unité

[4]Le système de Rössler est défini par :

$$\left\{\begin{array}{l} \frac{dx}{dt} = -y - z \\ \frac{dy}{dt} = -x + ay \\ \frac{dz}{dt} = b + xz - cz \end{array}\right. \tag{9}$$

avec $a = 0,15$; $b = 0,2$ et $c = 10$.

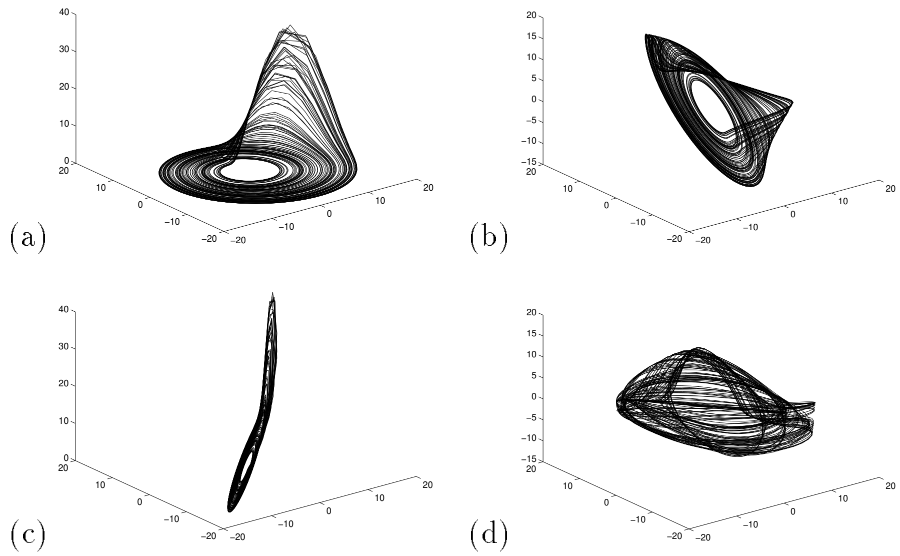


Figure 1: Influence de la valeur choisie pour le retard sur la topologie de l'espace des phases reconstruit à partir de la première composante du système de Rössler (9) : espace des phases issu du système (a) et reconstruction avec le retard optimal τ_{opt} (b), avec un retard trop faible $\frac{\tau_{opt}}{4}$ (c) et avec un trop grand retard $4\tau_{opt}$ (d).

d'hypervolume de dimension d est $\delta = \frac{N_{d\tau}}{L^d}$. Par conséquent, pour une série de N observations, la densité de probabilité pour un vecteur d'observation (7) de se trouver dans un voisinage donné de taille caractéristique l_0 varie typiquement comme $(\frac{l_0}{L})^d$. Il s'ensuit que la variance d'estimation de la densité de probabilité sur une partition du voisinage variant comme l'inverse du nombre de points s'exprime comme une fonction de $(\frac{L}{l_0})^d$: la variance augmente géométriquement avec la dimension de reconstruction. Il est donc expérimentalement recommandé de choisir d le plus faible possible soit $d = d_0$. Différentes méthodes ont été proposées pour déterminer cette valeur minimale :

- l'estimation de la dimension de l'attracteur par l'algorithme de P. Grassberger et I. Procaccia [16] et utilisation de la condition (8),
- l'étude de la dynamique estimée par la reconstruction : différentes techniques comme l'analyse en valeurs singulières [4], l'estimation de la *dimension locale intrinsèque* (LID) [21], la méthode dite des "faux plus proches voisins" [22], la

méthode des champs de vecteurs [23] ou l'analyse en composantes indépendantes [18, 24] étudie l'évolution des voisinages lorsque la dimension de reconstruction augmente d'une unité. La dimension est alors choisie comme la plus petite valeur d pour laquelle deux points proches en dimension d le restent en dimension $d+1$.

Une représentation correcte de l'espace des phases étant obtenue, nous proposons dans le paragraphe 3 de "discrétiser" cet espace afin d'estimer la probabilité d'avoir une observation dans un voisinage de $\underline{\mathbf{X}}_0$ à la date envisagée.

3 Estimation d'une distribution de probabilité sur l'espace des phases

3.1 Partitionner

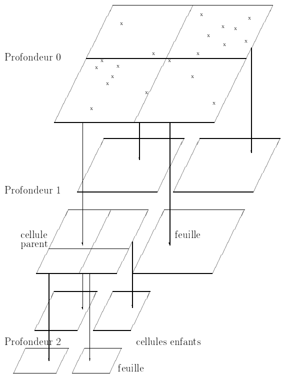
3.1.1 Principe : estimer la densité de probabilité

L'estimation de la distribution de probabilité sur l'espace des phases consiste à évaluer un histogramme d-dimensionnel associé à cet espace. L'objectif est donc d'obtenir un ensemble de cellules qui correspondent à une zone de l'espace des états sur lesquelles on a une probabilité uniforme : la probabilité de trouver une observation dans un voisinage de $\underline{\mathbf{X}}_0$, conditionnée à la présence de l'observation dans la cellule, doit être constante pour tout $\underline{\mathbf{X}}_0$ appartenant à cette même cellule de la partition.

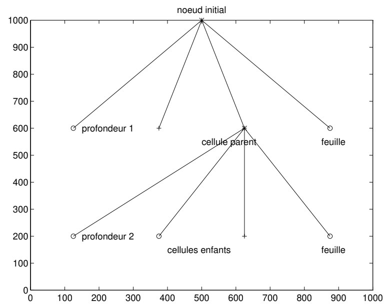
La construction de la partition sera récursive comme l'illustre la figure 2 pour un espace de dimension 2 : soient Π un élément d'une partition de l'espace des phases et $\{\pi_i\}_{i=1,\ldots,k}$ une subdivision de Π en k éléments, on note p_i la probabilité que $\underline{\mathbf{X}}_n \in \pi_i$ si $\underline{\mathbf{X}}_n \in \Pi$. L'information (définie comme l'opposé de l'entropie de Shannon de la distribution) associée à un histogramme en k bins réalisé à partir de N réalisations indépendantes peut s'écrire sous la forme :

$$-S = -\log_2 N = \sum_{i=1}^{k} p_i \log_2 p_i - \sum_{i=1}^{k} p_i \log_2 N_i \tag{11}$$

où $p_i = \frac{N_i}{N}$ est la probabilité que $\underline{\mathbf{X}}_n$ se trouve dans la cellule (le bin) indicée par i et N_i est le cardinal de cette cellule (le nombre de points appartenant au bin considéré).



(a)



(b)

Figure 2: Subdivision de l'espace des données (a) et arbre associé (b) pour un espace des phases de dimension $d = 2$: le nœud initial est subdivisé en 4 sous-ensembles. Si cette distribution est jugée non uniforme, on poursuit la subdivision de chaque cellule. Parmi les subdivisions qui sont réalisées au niveau suivant, seule celle du coin gauche est jugée non uniforme et subdivisée à nouveau.

L'information se décompose donc comme la somme de deux termes :

- l'information permettant de déterminer π_i ;
- l'information nécessaire pour coder $\underline{\mathbf{X}}_n$ sachant que $\underline{\mathbf{X}}_n \in \pi_i$.

Le gain d'information, ou la réduction de l'incertitude apportée par la connaissance de $\{\pi_i, i = 1, \ldots, k\}$ vaut :

$$\Delta = \log_2 k + \sum_{i=1}^{k} p_i \log_2 p_i \tag{12}$$

Il s'agit de la différence entre l'entropie maximale qu'on peut obtenir par une subdivision en k cellules et l'entropie obtenue pour la distribution de probabilité $\{p_i\}_{i=1,\ldots,k}$ définie ci-dessus associée à la partition $\{\pi_i\}_{i=1,\ldots,k}$ de Π. L'objectif de la partition étant d'obtenir l'histogramme d-dimensionnel des évènements observés dans l'espace des états, la fonction Δ définie par (12) doit s'annuler lorsque la cellule Π dont on étudie la partition est homogène (au sens où la distribution des évènements au sein de Π est uniforme). La règle de partition doit par conséquent être choisie telle que Δ soit minimisée sous l'hypothèse H_0

d'indépendance statistique des composantes des vecteurs d'états et de distribution uniforme des observations. Δ est une quantité positive qui s'annule lorsque p_i = constante = $\frac{1}{k}$. Une partition de $\Pi = \{\underline{\mathbf{X}}_i = [x_{i1} \ \ldots \ x_{id}]^T\}_{i=1,\ldots,N}$ en $\{\pi_i\}_{i=1,\ldots,k}$ est facilement obtenue en choisissant un seuil sur chacun des d axes (associés aux d composantes) créant $k = 2^d$ sous-cellules. Sous l'hypothèse H_0, les valeurs des seuils sont choisies autour de la valeur du médian de la distribution marginale sur l'axe considéré. Le médian sur le m-ième axe est défini par :

$$\text{median}\{x_i\} = \begin{cases} x_{\frac{n}{2}m}, & n \text{ pair} \\ x_{\frac{n+1}{2}m}, & n \text{ impair} \end{cases} \tag{13}$$

lorsque les $\{x_{im}\}_{i=1,\ldots,N}$ ont été classés par ordre croissant. On a donc :

$$P\{x_m \leq s_m\} = P\{x_m \geq s_m\} = \frac{1}{2}, \tag{14}$$

où s_m est le seuil du m-ième axe et x_m représente la m-ième coordonnée de $\underline{\mathbf{X}}_n$.

3.1.2 Test de l'homogénéité de la partition

La cellule Π ayant été subdivisée en 2^d sous-cellules, il est nécessaire de juger de l'homogénéité de Π à partir des probabilités respectives des différentes sous-cellules. Ce caractère homogène sera estimé à l'aide du test de χ^2 [25, 26] mesurant l'écart à la distribution uniforme de la distribution constituée par les probabilités estimées des sous-cellules. Il est en effet possible de montrer et de vérifier que pour de faibles écarts à la distribution uniforme (qui est la situation idéale recherchée), les quantités χ^2 et Δ conduisent au même type de critère. On écrit les probabilités associées à une distribution légèrement non uniforme sous la forme :

$$p_i = \frac{1}{k} + s_i \epsilon_i \tag{15}$$

où ϵ_i et s_i ($s_i \in \{-1, 0, 1\}$, le cas $s_i = 0$ correspondant à l'absence d'écart à la distribution uniforme pour la probabilité p_i) sont respectivement la valeur absolue de l'écart à la distribution uniforme et son signe. Un développement limité de Δ relatif à $\epsilon_i \ll p_i$ conduit à l'expression :

$$\Delta = \frac{1}{2\ln(2)}\chi^2 + \frac{k^2}{6\ln(2)}\left[\sum_{i=1}^{k-1}(s_i\epsilon_i)^3 - (\sum_{i=1}^{k-1} s_i\epsilon_i)^3\right] + \ldots \tag{16}$$

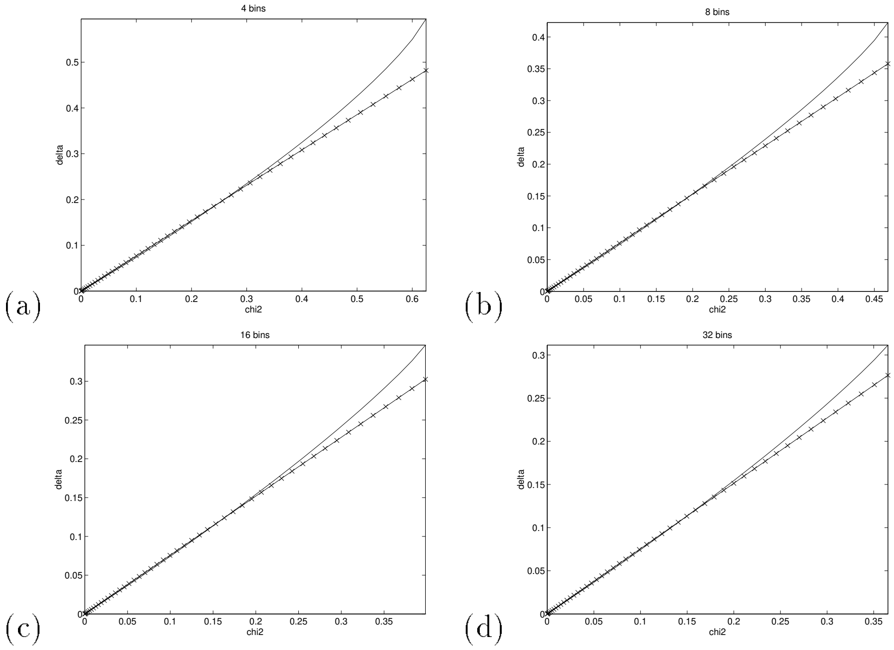


Figure 3: Δ en fonction de χ^2 pour (a) $k = 4$, (b) $k = 8$, (c) $k = 16$ ou (d) $k = 32$ bins.

On considère l'étude numérique d'un cas particulier d'écart à la distribution uniforme (15):

$$p_i = \begin{cases} \frac{1}{k} + (i - \frac{k}{2} - 1)\epsilon & \text{pour} \quad i \in \{1, 2, \ldots, \frac{k}{2}\} \\ \frac{1}{k} + (i - \frac{k}{2})\epsilon & \text{pour} \quad i \in \{\frac{k}{2} + 1, \frac{k}{2} + 2, \ldots, k\} \end{cases} \tag{17}$$

Les conditions que doit vérifier l'ensemble des $\{p_i\}$ ($\sum_{i=1}^{k} p_i = 1$ et $0 \leq p_i \leq 1$) impliquent que $\epsilon \leq \frac{2}{k^2}$. En calculant les différentes distributions de probabilité correspondant à des valeurs croissantes de ϵ (variant de 0 à $\frac{2}{k^2}$), on évalue pour chacune d'elles Δ et χ^2. La figure 3 représentant Δ en fonction de χ^2 illustre ces résultats pour $k = 4$, $k = 8$, $k = 16$ et $k = 32$ soit les subdivisions des cellules en 2^d sous-cellules pour $d = 2$, $d = 3$, $d = 4$ et $d = 5$. Aux faibles valeurs de ϵ (qui correspondent aux faibles valeurs de χ^2 et de Δ), on retrouve le comportement linéaire décrit par (16). Ces considérations établissent l'équivalence entre les deux types de critères : l'un basé sur un gain relatif d'information, l'autre étant le classique test de χ^2 mesurant l'écart à la distribution uniforme.

3.2 Règles de construction

Soit au niveau p une partition de l'espace des phases de dimension d constituée des cellules π_i^p (qui seront qualifiées de "parents"). Une partition plus fine est obtenue en subdivisant chaque cellule en 2^d sous-cellules π_j^{p+1} au niveau $p+1$, ces sous-cellules étant dites "enfants". La subdivision est effectuée en sélectionnant les d seuils sur chacun des axes comme le médian des projections des observations se trouvant dans la cellule conformément aux justifications des paragraphes précédents. Chaque sous-cellule sera alors soit à son tour subdivisée soit considérée comme une cellule de la partition finale. Les différentes subdivisions fournissent des ramifications formant une structure hiérarchique qui sera appelée "arbre". Dans la dénomination "arbre", une subdivision correspond à un "nœud", les liens parent-enfants constituent les "branches" et chaque extrémité qui représente une cellule finale de la partition sera appelée "feuille". La figure 2 montre la construction d'un arbre pour un espace de dimension 2, la structure hiérarchique et les subdivisions successives.

3.3 Critères d'arrêt de la partition récursive

Deux critères seront retenus pour arrêter la partition récursive.

3.3.1 Obtention d'une partition uniforme pour la cellule considérée

La partition récursive est arrêtée si la distribution de probabilité associée à la partition est jugée uniforme *via* le test de χ^2 [25, 26] ou de façon équivalente si Δ est insuffisant (cf. paragraphe 3.1.2). Dans ce cas de figure, la subdivision augmente la complexité de la partition sans qu'un gain substantiel d'informations soit obtenu. Ainsi utiliser un critère d'uniformité de la partition est équivalent à tenir compte du compromis entre le gain d'information faible et la plus grande complexité de la partition. Les résultats présentés dans cet article ont été obtenus en utilisant un test de χ^2 à $(d-1)$ degrés de liberté et 15% comparant $\{p_i\}_{i=1,\ldots,2^d}$ à la distribution uniforme. Cela correspond à itérer la méthode de partition si $\Delta \geq \Delta_{\mathsf{min}}$ où Δ_{min} est déterminé par les relations établies au paragraphe 3.1.2. Les critères ont été maintenus identiques indépendemment de l'ordre de la récurrence (ou profondeur de l'arbre) ou du nombre d'évènements contenus dans la cellule Π considérée.

3.3.2 Non pertinence de l'étude statistique

La partition d'une cellule Π est fondée sur l'estimation du médian de la distribution marginale à partir d'un nombre d'évènements observés fini. Soit une cellule Π dont la dynamique marginale sur le m-ième axe présente une amplitude définie par :

$$d_m = \max_{i, \underline{\mathbf{X}}_i \in \Pi} x_{im} - \min_{i, \underline{\mathbf{X}}_i \in \Pi} x_{im} \tag{18}$$

où x_{im} est la m-ième composante de $\underline{\mathbf{X}}_i$. On note $\mathcal{N}$ le nombre de points contenus dans cette cellule et $\hat{T}_m$ le médian estimé (cf. paragraphe 3.1.1 et équation 13) sur le m-ième axe de la distribution marginale de $\{\underline{\mathbf{X}}_i \in \Pi\}$. La distribution du médian estimé $\hat{T}_m$ est une loi asymptotiquement normale [27, 28] dont la moyenne est égale au médian théorique T_m et de variance [29] :

$$var(\hat{T}_m|\mathcal{N}) = \frac{1}{4\mathcal{N}[f_{x_m}(T_m)]^2} \tag{19}$$

où f_{x_m} est la fonction de densité de probabilité pour le m-ième axe. Sous l'hypothèse H_0 d'indépendance statistique des composantes, $f_{x_m} = \frac{1}{d_m}$ et (19) s'écrit :

$$var(\hat{T}_m|\mathcal{N}) = \frac{d_m^2}{4\mathcal{N}} \tag{20}$$

L'erreur standard sur l'estimation du médian est donc $\frac{d_m}{2\sqrt{\mathcal{N}}}$; cette erreur conduit à une erreur d'estimation des hypervolumes en dimension d de l'ordre de

$$\epsilon = d\frac{d_m^d}{2\sqrt{\mathcal{N}}} \tag{21}$$

(21) est obtenue par développement limité au premier ordre en supposant que les amplitudes des dynamiques marginales et les erreurs standard sur les médians $\hat{T}_m$ sont du même ordre (i.e. $d_m \sim d_{m'} \forall (m, m')$ et $var(\hat{T}_m|\mathcal{N}) \sim var(\hat{T}_{m'}|\mathcal{N})$). Cette expression est utilisée pour contrôler la pertinence statistique de la méthode de partition. En effet, si l'erreur standard (21) devient trop importante par rapport au volume caractéristique de la cellule, les erreurs d'estimation de la partition sont du même ordre que la dimension d'une cellule résultant de cette partition. Afin de limiter l'erreur standard (21), on impose donc

$\epsilon < (cd_m)^d$ avec $c < 1$ d'où $\frac{d}{2\sqrt{\mathcal{N}}} < c^d$ et finalement $\frac{d}{2c^d} < \sqrt{\mathcal{N}}$. Le choix $c = \frac{1}{2}$ conduit par conséquent à stopper l'algorithme de partition si

$$\mathcal{N} < \frac{d^2}{4} 4^d \tag{22}$$

Il est intéressant de noter que ce critère permet de définir une notion de résolution limite de la partition en fonction du nombre d'observations $\mathcal{N}$ dans le voisinage considéré et de la dimension de reconstruction d.

La méthode décrite dans ce paragraphe conduit donc à une partition de l'espace des phases correspondant à un ensemble de cellules homogènes ou dont la subdivision n'apporte pas un gain important d'information. Il est à noter que cette approche justifie la partition proposée par A.M. Fraser pour l'estimation de l'information mutuelle [7, 8, 19, 20].

3.4 Coût de l'algorithme

Le coût algorithmique étant très largement dépendant du signal étudié, nous proposons ici d'en donner une borne supérieure. Soit une cellule Π^p à la profondeur p de l'arbre. L'ensemble $\{\pi_i^{p+1}\}_{i=1,\ldots,2^d}$ de ses enfants est défini en estimant les seuils indépendamment sur chaque axe comme le médian des coordonnées de l'ensemble des N points de l'espace des phases utilisés. Le coût algorithmique pour classer M points est proportionnel à $M \log_2 M$ en utilisant un algorithme de classement optimisé [25]. La profondeur p de l'arbre est constituée au maximum de $(2^d)^p$ cellules Π^p : il s'agit du cas le plus défavorable pour l'algorithme où aucune cellule π_i^l (avec $l < p$) des profondeurs inférieures n'a été considérée comme une cellule de la partition finale. Par conséquent, $d2^{dp}$ classements de $\frac{N}{2^{dp}}$ points en moyenne (en supposant que les points se répartissent de façon à peu près équivalente sur l'ensemble des cellules $\{\pi_i^p\}$) sont nécessaires. Le nombre CP_p d'opérations à effectuer à la profondeur p est donc proportionnel à :

$$CP_p = d2^{dp} \frac{N}{2^{dp}} \log_2 \frac{N}{2^{dp}} = Nd \log_2 \frac{N}{2^{dp}} \tag{23}$$

L'addition des coûts CP_p associés aux différentes profondeurs de la profondeur $p = 0$ à la profondeur $p = P_{max} - 1$ pour un arbre de profondeur maximale P_{max} (on ne subdivise pas les cellules de la profondeur P_{max}) conduit à un coût algorithmique total de :

$$CA_{max} = \sum_{p=0}^{P_{max}} CP_p = Nd \sum_{p=0}^{P_{max}-1} \log_2 \frac{N}{2^{dp}} = NdP_{max} \log_2 \frac{N}{2^{\frac{d(P_{max}-1)}{2}}} \tag{24}$$

Il convient de noter qu'il s'agit du coût dans le cas le plus défavorable, correspondant au cas où, à chaque profondeur de l'arbre, l'ensemble des cellules est effectivement subdivisé en 2^d enfants. Dans les cas rencontrés expérimentalement, comme le montrent les différentes structures d'arbres obtenues (cf. figures 4, 8, 9 ou 10) pour lesquelles à chaque profondeur il existe des cellules terminales, le coût algorithmique CA est nettement inférieur à CA_{max}. Des éléments de comparaison numérique avec le coût d'autres méthodes sont présentés au paragraphe 4.3.

4 Application à la prédiction

4.1 Principe

Dans ce paragraphe, nous proposons d'utiliser la structure d'arbre obtenue et la partition de l'espace des phases qui en est déduite pour classifier de nouvelles observations et obtenir une valeur de prédiction par régression sur les données appartenant à chacune des cellules.

Dans la suite, nous désignerons par les termes "série d'apprentissage" la série temporelle à partir de laquelle une partition de l'espace des phases reconstruit par plongement a été obtenue. À chacune des cellules π_i^p obtenue lors de la partition récursive de l'espace des phases peut être associée une fonction $h_i^p : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$. Afin de simplifier l'exposé, nous nous limitons ici au cas où les cellules considérées sont des "feuilles" (i.e. $p = p_{max}$) et $d' = 1$, ce qui ne limite en rien la généralité de la méthode présentée. Dès lors, on note $h_i^p(\underline{\mathbf{X}}) = h_i(\underline{\mathbf{X}})$. Les fonctions h_i associées à chaque cellule appartiennent à un ensemble de fonctions données *a priori*. Dans le paragraphe 5, elles sont estimées à partir de la série d'apprentissage parmi l'ensemble des fonctions linéaires d'ordre d (cf. équation (39)) s'exprimant comme combinaison affine des d coordonnées vectorielles de $\underline{\mathbf{X}}$. L'intérêt de cette approche et des liens avec les modèles $SETAR$ [1, 2, 3] nous ont conduit à leur consacrer un paragraphe particulier.

Dans les exemples suivants, les fonctions $h_i(\underline{\mathbf{X}})$ sont définies par :

$$h_i(\underline{\mathbf{X}}) = \hat{E}[x^a_{k+1} | \underline{\mathbf{X}}^a_k \in \pi_i] \quad \text{si} \quad \underline{\mathbf{X}} \in \pi_i \tag{25}$$

ou :

$$h_i(\underline{\mathbf{X}}) = \text{median}[x^a_{k+1} | \underline{\mathbf{X}}^a_k \in \pi_i] \tag{26}$$

où x^a_i et $\underline{\mathbf{X}}^a_i$ représentent respectivement les points de la série d'apprentissage et les vecteurs reconstruits pas la méthode des retards à partir de la série d'apprentissage. $\hat{E}$ désigne la valeur estimée de l'espérance mathématique sur la cellule considérée :

$$\hat{E}[x^a_{k+1} | \underline{\mathbf{X}}^a_k \in \pi_i] = \frac{1}{N_i} \sum_{i=1}^{N_i} x^a_{k+1} \quad \text{pour} \quad \underline{\mathbf{X}}^a_k \in \pi_i \tag{27}$$

et le médian a été défini par (13). D'autres choix de h peuvent être plus intéressants et ne seront pas discutés ici.

Ainsi, si

$$\underline{\mathbf{X}}_n = [x_n \ x_{n-\tau} \ \ldots \ x_{n-(d-1)\tau}]^T \tag{28}$$

est le vecteur d'état courant (pour la série temporelle observée jusqu'à la date $t = n\tau$), $h_i(\underline{\mathbf{X}_n})$ défini par (25) ou (26) réalise une prédiction quantifiée :

$$\hat{x}_{n+1} = \sum_i h_i(\underline{\mathbf{X}_n}) I(\underline{\mathbf{X}}_n \in \pi_i) \tag{29}$$

où $I(\bullet)$ est la fonction indicatrice. Par construction, le pas de prédiction est égal à τ (cf. (28)). Pour chaque x_n, l'opération de prédiction peut donc se décomposer en trois étapes :

- construire le vecteur d'état courant $\underline{\mathbf{X}_n}$,
- classifier $\underline{\mathbf{X}_n}$ (on notera π_i la cellule à laquelle appartient $\underline{\mathbf{X}_n}$),
- $\hat{x}_{n+1} = h_i(\underline{\mathbf{X}_n})$

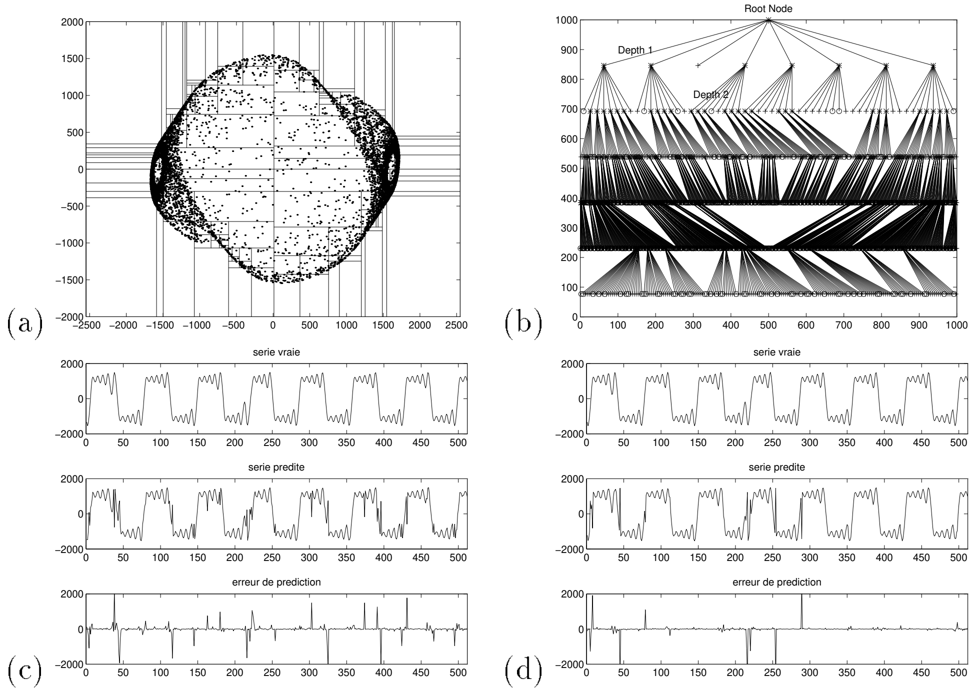


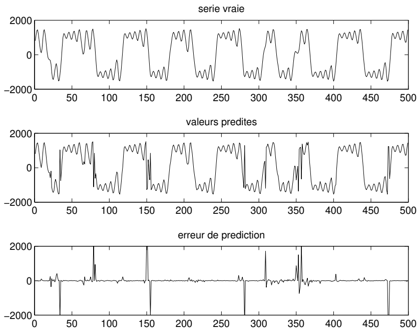
Figure 4: Prédiction à un pas (égal au retard τ de reconstruction) pour le système expérimental de Double Scroll (40, 41). L'espace des phases reconstruit en dimension 2 et la partition induite par l'algorithme sont portés en figure (a). (b) représente l'arbre associé à l'algorithme récursif de partition calculée pour $d = 3$. (c) et (d) présentent les résultats de prédiction à un pas pour h_i définie par (25) et (26). La valeur de τ utilisée est 4 et la série d'apprentissage comprend 8192 points.

La figure 4 présente la prédiction à un pas (égal au retard τ de reconstruction) de $\underline{\mathbf{X}}_n$ pour le système de Double-Scroll défini par (40,41). Dans l'exemple envisagé, les deux méthodes donnent des résultats tout à fait comparables. Ceci s'explique par le fait que dans les cas considérés les deux estimateurs sont sensiblement égaux.

4.2 Comparaison avec la méthode des plus proches voisins

Il est intéressant de comparer les performances de cette approche avec d'autres méthodes de prédiction comme la méthode des plus proches voisins ou la technique des fonctions radiales décrites par M. Casdagli [30]. Du fait même de ses similitudes avec l'approche proposée ici, seule la méthode du plus proche voisin[5] a été l'objet d'une étude comparative.

[5] Il existe de nombreuses variantes de cette approche utilisant des sommes pondérées de k plus proches voisins [30, 31]. Ici seul le premier plus proche voisin est considéré. Ce dernier est choisi tel qu'il soit

(a) 

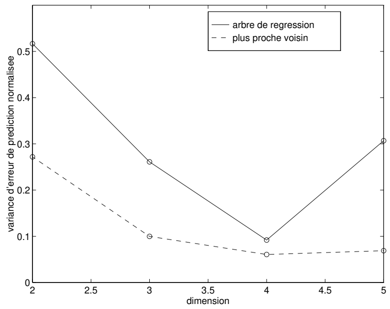
(b) 

Figure 5: Prédiction à un pas (égal au retard τ de reconstruction) pour le système expérimental de Double Scroll (40, 41). (a) présente les résultats obtenus avec la méthode des plus proches voisins. La valeur de τ utilisée est 4 et la série d'apprentissage comprend 8192 points. (b) montre l'évolution de la variance d'erreur de prédiction à un pas (égal au retard de reconstruction) en fonction de la dimension de reconstruction pour les deux méthodes : arbre de régression et plus proche voisin. La variance est calculée sur 8192 points et normalisée par l'énergie du signal selon l'expression (31).

Cette méthode [32, 33] conduit à choisir comme valeur de prédiction $\hat{x}_{n+1}$ définie par :

$$\left\{ \begin{array}{l} \underline{\mathbf{X}}^a_{n_0} = \arg\min_{\{\underline{\mathbf{X}}^a_k\}} \parallel \underline{\mathbf{X}}_n - \underline{\mathbf{X}}^a_k \parallel \\ \hat{x}_{n+1} = x^a_{n_0+1} \end{array} \right. \tag{30}$$

où "$\parallel \bullet \parallel$" est la distance (euclidienne) entre $\underline{\mathbf{X}}_n$ et $\underline{\mathbf{X}}^a_k$ et où $\{\underline{\mathbf{X}}^a_k\}$ décrit la série d'apprentissage. La figure 5(a) présente les résultats obtenus sur une série chaotique expérimentale par cette méthode de prédiction à un pas (égale à τ par construction) avec les mêmes paramètres de reconstruction que pour ceux des figures 4.

Les résultats obtenus par les deux types de méthode (arbre et plus proches voisins) montrent un même ordre de grandeur pour la variance d'erreur de prédiction à un pas (voir la figure 5(b)) :

$$V = \frac{\sum_{i=1}^{L}(e_i - E[e_i])^2}{\sum_{i=1}^{L}(x_i - E[x_i])^2} \tag{31}$$

où l'erreur de prédiction :

$$e_i = |\hat{x}_{i+1} - x_{i+1}| \tag{32}$$

a ici été estimée par (31) pour $L = 500$ réalisations de x_i.

éloigné du point courant de plus d'une longueur de corrélation temporelle.

4.3 Rapide comparaison des coûts de calcul des deux méthodes

Il est à noter que lorsqu'un arbre de régression de profondeur maximale P_{max} est estimé, la prédiction quantifiée est obtenue par au maximum P_{max} opérations de comparaison. Par contre, la méthode des plus proches voisins impose pour chaque point de considérer autant de comparaisons qu'il y a de réalisations dans la série d'apprentissage. De ce fait, la méthode des arbres est plus rapide que la méthode des plus proches voisins. On suppose pour l'étude suivante du coût algorithmique de la prédiction que l'arbre de régression d'une profondeur maximale P_{max} préexiste, on se reportera au paragraphe 3.4 pour l'étude du coût de cette construction. Pour une série d'apprentissage de N_0 points et N_1 points à prédire, la méthode des plus proches voisins nécessite, outre le calcul de N_0 distances entre le point à prédire et l'ensemble des points de la série d'apprentissage, de déterminer parmi ces N_0 distances les k plus petites pour une méthode à k plus proches voisins (ici on prend $k = 1$) soit un coût algorithmique en $N_0(d + 1 + \log_2 N_0)$ pour chaque point N_1 soit au total un algorithme en :

$$CPPV = N_1 N_0(d + 1 + \log_2 N_0) \tag{33}$$

(a) (b)

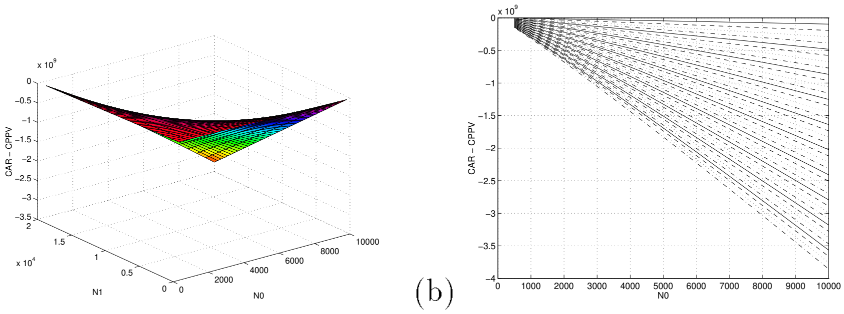


Figure 6: Écart entre le coût algorithmique avec les arbres de régression (CAR_{max}) et celui par la méthode des plus proches voisins ($CPPV$) en fonction des nombres de points N_0 de la série d'apprentissage et N_1 à prédire. Le coût est calculé avec une dimension $d = 2$ et une profondeur $p = 5$ pour l'arbre.

Pour le même nombre de données et de points à prédire, un arbre de régression de profondeur maximale P_{max} ne réclame au plus qu'un coût en $2dP_{max}N_1$ auquel s'ajoute le coût de la construction CA au maximum en $CA_{max} = N_0 d P_{max} \log_2 \frac{N_0}{2^{\frac{d(P_{max}-1)}{2}}}$ soit un algorithme en :

$$CAR_{max} = 2N_1 d P_{max} + N_0 d P_{max} \log_2 \frac{N_0}{2^{\frac{d(P_{max}-1)}{2}}} \tag{34}$$

Le coût de calcul est plus faible avec les arbres de régression comme l'illustre la figure 6 : la différence $CAR_{max} - CPPV$ est toujours négative, ce résultat est valable pour les autres valeurs de dimension et de profondeur que celles utilisées pour la figure 6. Comme au paragraphe 3.4 et pour les mêmes raisons, il convient de noter que le coût estimé ici est un coût maximal et que dans la pratique ce coût est nettement plus faible : $CAR < CAR_{max}$.

4.4 Détermination des paramètres de reconstruction τ et d

La figure 5(b) montre que la variance d'erreur de prédiction V présente un minimum en fonction de la dimension de reconstruction : pour les données expérimentales liées au système expérimental de Double Scroll (40,41), la valeur du minimum est obtenue pour $\hat{d} = 4$. Tant que la dimension d utilisée est inférieure à la "bonne" dimension (cf. paragraphe 2.3), toute la dynamique du système n'est pas considérée : à mesure que d augmente en tendant vers d_0, de plus en plus d'informations sur la dynamique sont considérées et la variance d'erreur de prédiction diminue. Quand d devient supérieure à d_0, les erreurs d'estimation de la densité de probabilité évoquées dans le paragraphe sur le choix de la dimension de reconstruction entraînent une augmentation de la variance d'erreur de prédiction. Le minimum de cette quantité fournit donc la dimension d_0.

Il est alors intéressant de généraliser cette étude à une étude conjointe en fonction de la dimension et du retard de reconstruction : on obtient ainsi une méthode[6] de détermination des deux paramètres de reconstruction de manière simultanée [34]. En pratique, les quantités sont étudiées en les normalisant par τ pour s'affranchir du problème de la dépendance linéaire pour de faibles valeurs du retard. Si τ est trop faible, les coordonnées

[6] Il s'agit de la même philosophie que celle employée lors de l'utilisation de l'algorithme de P. Grassberger et I. Procaccia pour la détermination de la dimension de reconstruction : on emploie un algorithme d'estimation et on exploite son comportement quand les paramètres de reconstruction τ et d varient pour en déduire leurs valeurs τ_0 et d_0.

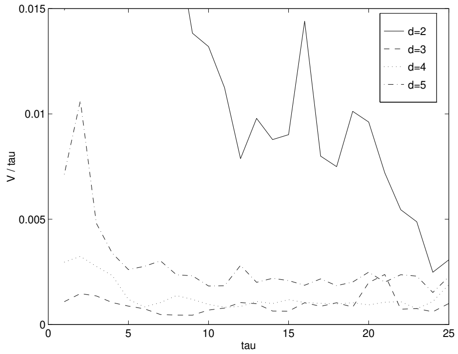


Figure 7: Variance d'erreur de prédiction pour le système de Rössler en fonction de τ et d.

des vecteurs reconstruits sont fortement corrélées : $x_k \approx x_{k+1}$ (cf. paragraphe 2.3). Les vecteurs sont presque colinéaires, la prédiction donne de bons résultats sans que la valeur du retard soit adéquat. Cette étude donne des résultats satisfaisants et conformes à ceux de [18] aussi bien pour $\hat{d}$ que pour $\hat{\tau}$. La figure 7 montre les résultats obtenus ($\hat{d} = 3$ et $\hat{\tau} = 8$) sur le système de Rössler défini par (9).

Cette approche est à comparer à celle proposée par W. Liebert, K. Pawelzik et H.G. Schuster dans [35] qui quantifient la conservation de la proximité des points de l'espace d'état quand la dimension de reconstruction augmente. Si la dimension de reconstruction est insuffisante, l'espace obtenu correspond à une projection de l'espace réel : la proximité de deux points provient soit d'un voisinage réel soit d'un effet dû à la projection. L'étude du rapport des distances calculées en dimension d puis $d+1$ permet de détecter la présence d'artefacts liés à la projection. Si d est insuffisante, la distance en dimension $d+1$ de deux points proches en dimension d à cause de la projection est très supérieure devant la distance en dimension d. Par contre, si d est suffisante, le rapport des distances en dimension d et en dimension $d+1$ sera proche de 1. Ils proposent donc de choisir les paramètres de reconstruction en minimisant le rapport[7] $\frac{W}{\tau}$ où :

$$W = \ln E \left[\prod_{k=1}^{L} \frac{dist^{\tau}_{d+1}(i, j(k,d)) dist^{\tau}_{d}(i, j(k,d+1))}{dist^{\tau}_{d+1}(i, j(k,d+1)) dist^{\tau}_{d}(i, j(k,d))}^{\frac{1}{2L}} \right]$$

[7]Le fait de diviser par τ permet comme pour la méthode proposée ici avec la variance d'erreur de s'affranchir du problème de la dépendance linéaire pour de faibles valeurs du retard.

où $dist^{\tau}_{m}(i, j(k,n))$ est la distance en dimension m du k-ième plus proche voisin en dimension n. Les deux méthodes donnent en général des résultats comparables [34]. Cependant pour les mêmes raisons que pour la prédiction par les plus proches voisins, la méthode de W. Liebert, K. Pawelzik et H.G. Schuster est plus coûteuse en temps de calcul du fait de la nécessité d'évaluer toutes les distances.

5 Lien avec les modèles auto-régressifs à seuils

L'algorithme décrit au paragraphe 3 peut conduire à une structure très complexe d'arbre (associée à une partition très fine), y compris dans le cas de systèmes relativement simples. Par exemple, l'étude en dimension 2 d'un espace des phases constitué d'une succession de comportements linéaires bruités met en évidence ce genre de situation (cf. figure 8). En effet, un comportement linéaire du type $x_n = ax_{n-\tau}$ ($a \neq 0$) en dimension 2 conduit à une partition en 4 cellules dont deux sont pratiquement vides de points. L'algorithme ne s'arrête que sur un critère de contrôle de la pertinence du test statistique (i.e. quand trop peu de points appartiennent à une cellule pour qu'une approche statistique reste pertinente, cf. paragraphe 3.3.2). Afin d'éviter ce type de "divergence" de la méthode de partition, nous proposons d'exprimer, lors de la construction récursive de la partition, les données sur leur base propre. Une décomposition en valeurs singulières (Singular Value Decomposition, notée SVD dans la suite) est effectuée sur chaque cellule Π^p (associée à un nœud de l'arbre à la profondeur p) à l'étape p de la récurrence. En notant $\underline{\mathbf{X}}^{\Pi^p}$ les points appartenant à la cellule Π^p, la matrice de covariance locale :

$$\Gamma(\Pi^p) = E_{\Pi^p}\left[(\underline{\mathbf{X}}^{\Pi^p} - E_{\Pi^p}[\underline{\mathbf{X}}^{\Pi^p}])(\underline{\mathbf{X}}^{\Pi^p} - E_{\Pi^p}[\underline{\mathbf{X}}^{\Pi^p}])^T\right] \tag{35}$$

où $E_{\Pi^p}[\bullet] \stackrel{\text{def}}{=} E[\bullet | \underline{\mathbf{X}}^{\Pi^p} \in \Pi^p]$ est l'espérance mathématique de "$\bullet$" évaluée sur Π^p. $\Gamma(\Pi^p)$ est décomposée sous la forme ${M^p}^T D M^p$ où D est une matrice diagonale contenant les valeurs propres $\{\lambda_p\}$ de $\Gamma(\Pi^p)$ et $M^p = [m_1|m_2|\ldots]$ la matrice de changement de base formée à partir des vecteurs propres m_p associés aux valeurs propres λ_p. La partition $\{\pi_j^{p+1}\}_{j=1,\ldots,k}$

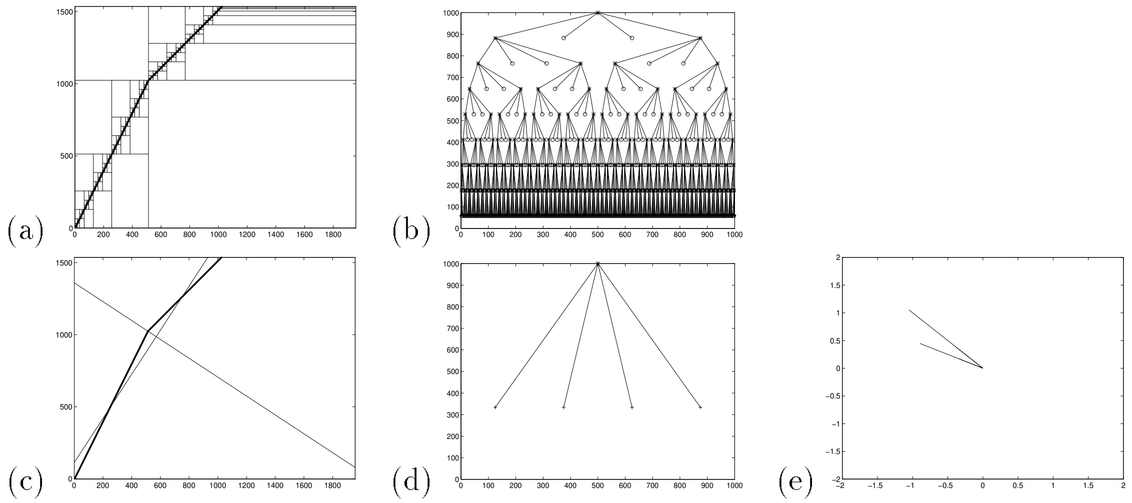


Figure 8: Construction de l'arbre avec (c,d,e) et sans (a,b) SVD sur un espace des phases constitué de deux segments de droite : (a) et (c) présentent les partitions obtenues, (b) et (d) les structures d'arbre qui leur sont associées. (e) est une représentation vectorielle des modèles obtenus sur chacune des feuilles : les directions sont données par les vecteurs associés au modèle de la cellule et la norme est la probabilité d'obtenir la cellule.

de Π^p en k éléments est alors appliquée aux données transformées :

$$\underline{\mathbf{X}}^{\pi_j^{p+1}} = M^p(\underline{\mathbf{X}}^{\Pi^p} - \underline{C}^{\Pi^p}), \tag{36}$$

où $\underline{C}^{\Pi^p} = E_{\Pi^p}[\underline{\mathbf{X}}^{\Pi^p}]$ est le barycentre des vecteurs $\underline{\mathbf{X}}^{\Pi^p}$ de la cellule considérée. Par récurrence (d'une profondeur p à la suivante $p+1$), l'expression d'un vecteur à la profondeur p s'exprime en fonction des données initiales par [27, 28] :

$$\underline{\mathbf{X}}^{\Pi^p} = \mathcal{M}^{p-1}\underline{\mathbf{X}}^{\Pi^0} - \mathcal{C}^{p-1} \tag{37}$$

où $\mathcal{M}^p = \prod_{i=0}^{p} M^i$ et $\mathcal{C}^p = \sum_{i=0}^{p}[\prod_{j=i}^{p} M^i]C^{\pi^i}$.

Soit P_{max} la profondeur maximale de l'arbre (i.e. la partition finale est obtenue après P_{max} récurrences) ; $\underline{\mathbf{X}}^{\Pi^{P_{max}}}$ est par construction un bruit uniforme sur $\Pi^{P_{max}}$. De plus en remarquant que $\underline{\mathbf{X}} = \underline{\mathbf{X}}^{\Pi^0}$, (37) se ré-exprime par :

$$\mathcal{M}^{p-1}\underline{\mathbf{X}} = \underline{\epsilon} + \mathcal{C}^{p-1} \tag{38}$$

(38) est l'équation d'un processus auto-régressif $(d-1)$-dimensionnel avec pour entrée un

bruit $\underline{\epsilon} + \mathcal{C}^{p-1}$. L'identification du modèle sous la forme :

$$x_n = \sum_{i=1}^{d-1} a_i x_{n-i} + \epsilon_n = \underline{\mathbf{A}}^T \underline{\mathbf{X}}_n + \epsilon_n \tag{39}$$

est conduite dans chaque cellule par la recherche du noyau de l'endomorphisme défini par $\mathcal{M}^{p-1}$. Le modèle obtenu (modèle auto-régressif sur chaque élément de la partition défini par un ensemble de seuils) est à rapprocher des modèles $SETAR$ (pour Self Exciting Threshold AutoRegressive) définis par H. Tong [2]. Ces modèles nécessitent une estimation difficile des seuils : fixer les seuils, estimer les modèles associés puis changer ces seuils avant de choisir les "meilleurs" dans l'ensemble de ceux testés. L'avantage de notre méthode est de réduire ce problème d'estimation : les seuils sont fournis directement par la construction de la partition. Il est donc important de noter que cette approche permet une estimation (grossière) des seuils des modèles à seuils sans aucun modèle *a priori* fournissant ainsi une réponse à la discussion de M.B. Priestley [1] et H. Tong [2] concernant les difficultés rencontrées pour estimer les seuils ; ces derniers proposent de ne pas considérer plus d'un seuil pour limiter les problèmes d'estimation. Les résultats des deux algorithmes (sans et avec SVD) pour deux dynamiques linéaires sont présentés en figure 8 : celui utilisant la SVD détecte les deux structures contrairement à celui qui n'effectue pas l'orthogonalisation récursive des données. En figure 8(e) sont indiquées les représentations vectorielles $\underline{\mathbf{A}}$ des modèles associés à chaque cellule ($\sum_{i=1}^{p} a_i x_{n-i} = \epsilon_n \Longrightarrow \underline{\mathbf{A}}^T \underline{\mathbf{X}}_n = \underline{\epsilon}$) : les segments ont pour vecteur directeur les représentations vectorielles $\underline{\mathbf{A}}$ des modèles associés à chaque cellule et pour norme la probabilité de la cellule.

L'utilisation de la SVD fournit une partition différente et tout à fait indépendante de celle construite sans SVD : en l'absence de SVD, les comportements linéaires ne sont pas détectés contrairement à la méthode avec SVD qui détecte à chaque étape de la construction un écart à un modèle linéaire décrit par la réécriture des données dans leur base propre. D'autre part, le test d'homogénéité d'une subdivision est effectué après l'orthogonalisation, ce qui conduit à tester l'homogénéité autour des composantes linéaires. Généralement la structure apparente de l'arbre issu de la construction avec SVD est plus simple : les comportements linéaires sont détectés et enregistrés par l'intermédiaire des

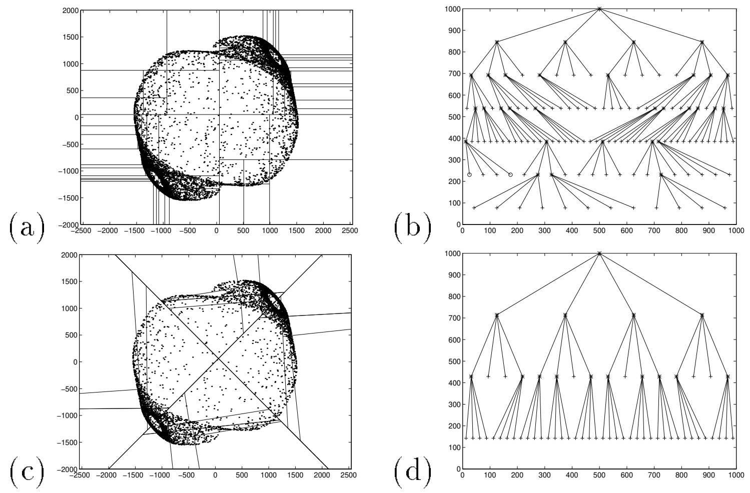


Figure 9: Construction de l'arbre avec ((c) et (d)) et sans SVD ((a) et (b)) sur l'espace reconstruit en dimension $d = 2$ avec le retard $\tau = 4$ pour le système expérimental de Double-Scroll : les partitions obtenues ((a) et (c)) et la structure de l'arbre ((b) et (d)).

matrices de changement de base. L'apport de la SVD est donc double : elle permet d'accéder à un modèle à seuils au sein de chaque cellule et l'obtention de cellules homogènes est plus rapide puisque les structures de type linéaires (dont l'arbre garde l'information par l'intermédiaire de la matrice de changement de base) disparaissent récursivement. La figure 9 illustre cet aspect avec les résultats obtenus avec et sans SVD pour une réalisation expérimentale du système de "Double-Scroll"[8] décrite dans [36].

La figure 10 illustre la "détection" des deux modèles linéaires présents dans le système auto-régressif d'ordre 2 à seuils défini par :

$$x_n = \begin{cases} -0.562x_{n-2} - 3.91 + e_n & si\ x_{k-1} \leq 0 \\ 1.71x_{n-1} - 0.81x_{n-2} + 0.356 + e_n & sinon \end{cases} \tag{42}$$

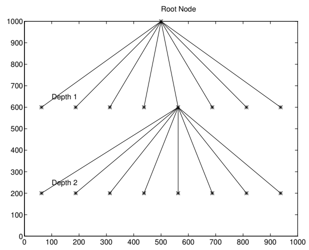
[8] Le système de "Double-Scroll" est décrit par le système d'équations suivantes :

$$\begin{cases} \frac{dx}{dt} = \alpha(y - h) \\ \frac{dy}{dt} = x - y + z \\ \frac{dz}{dt} = -\beta y \end{cases} \tag{40}$$

où

$$h = \begin{cases} m_1 x + m_0 - m_1 & si\ x > 1 \\ m_0 x & si - 1 < x < 1 \\ m_1 x - m_0 - m_1 & si\ x < -1 \end{cases} \tag{41}$$

avec les valeurs des paramètres : $\alpha = 9$, $\beta = \frac{100}{7}$, $m_0 = -\frac{1}{7}$ et $m_1 = \frac{2}{7}$.

(a) 

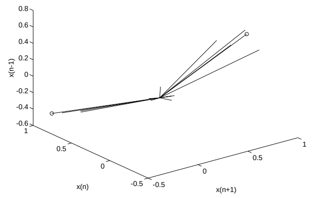
(b) 

Figure 10: Construction de l'arbre avec SVD (a) sur l'espace reconstruit en dimension $d = 3$ pour le système décrit par un modèle à seuils et représentation vectorielle des modèles obtenus sur chacune des feuilles (b), les vrais modèles étant représentés par un 'o'.

6 Conclusions et perspectives

Les arbres de régression fournissent une approche intéressante pour modéliser par une méthode non linéaire et non paramétrique des systèmes dynamiques ou des séries temporelles non linéaires issues de tels systèmes. La construction de ce modèle est justifiée dans le cadre de la théorie de l'information et la méthode récursive employée fournit une structure hiérarchique aux données permettant une utilisation pratique en vue d'applications. Outre la présentation et l'étude de ce modèle ainsi que de sa construction, un certain nombre d'applications ont été envisagées : estimation facilitée des modèles auto-régressifs à seuils, prédiction des séries chaotiques ou détermination des paramètres de reconstruction nécessaires à l'emploi de la méthode des retards pour reconstruire un espace des phases avec une seule série temporelle. L'intérêt de cette méthode est de ne nécessiter que peu d'*a priori* pour être mise en place tout en fournissant des résultats comparables à ceux d'autres méthodes pour un coût de calcul moindre.

D'autres applications doivent être envisagées notamment en lien avec les notions issues de la théorie de l'information : par exemple les notions liées au taux de génération d'entropie du système ou du modèle markovien que l'on peut déduire de cette représentation (à partir des probabilités de transition entre cellules) sont l'objet d'études en cours. D'autre part, la représentation des systèmes par un arbre de régression peut aussi être la base d'une comparaison de ces systèmes *via* cette modélisation. La multiplicité des applications qui peuvent être issues de ce modèle lui confère donc un intérêt certain.

Références

[1] M.B. PRIESTLEY, *Nonlinear and Nonstationary Time Series Analysis*, Academic Press, 1988.

[2] H. TONG, *Non Linear Time Series : a Dynamical System Approach*, Oxford Science Publication, Oxford University Press, NY, 1990.

[3] D. GUÉGAN, *Séries Chronologiques Non Linéaires à Temps Discret*, Economica, 1994.

[4] H.D.I. ABARBANEL, R. BROWN, J.J. SIDOROWICH, L.S. TSIMING, "The Analysis of Observed Chaotic Data in Physical Systems", *Review of Modern Physics*, vol. 65, no. 4, 1993, pp. 1331-1392.

[5] M. CASDAGLI, S. EUBANK, J.D. FARMER, J. GIBSON, "State Space Reconstruction in Presence of Noise", *Physica D*, vol. 51, 1991, pp. 52-98.

[6] J.P. ECKMANN, S. OLIFFSON KAMPHORST, D. RUELLE, S. CILIBERTO, "Lyapunov Exponents from Time Series", *Physical Review A*, vol. 34, no. 6, 1986, pp. 4971-4979.

[7] A.M. FRASER, "Information and Entropy in Strange Attractors", *IEEE Transactions on Information Theory*, vol. 35, no. 2, 1989, pp. 245-262.

[8] R. SHAW, "Strange Attractors, Chaotic Behaviour, and Information Flow", *Z. Naturforschung*, vol. 36a, 1981, pp. 80-112.

[9] J.P. ECKMANN, D. RUELLE, "Ergodic Theory of Chaos and Strange Attractors", *Review of Modern Physics*, vol. 57, no. 3, 1985, pp. 617-656.

[10] O. MICHEL, P. FLANDRIN, "Application of Methods Based on Higher Order Statistics for Chaotic Signal Analysis", *Signal Processing*, vol. 53, no.2, 1996.

[11] H. WHITNEY, "Differentiable Manifolds", *Annals of Mathematics*, vol. 37, no. 3, 1936, pp. 645-680.

[12] N. H. PACKARD, J.P. CRUTCHFIELD, J.D. FARMER, R. SHAW, "Geometry from a Time Series", *Physical Review Letters*, vol. 45, no. 9, 1980, pp. 712-716.

[13] F. TAKENS, "Detecting Strange Attractors in Turbulence", *Lecture Notes in Mathematics*, vol. 898, 1981, pp. 366-381.

[14] D. GUÉGAN, "Detecting Non Linearity : A Review", *Statistique et Analyse des Données* vol. 15, no. 2, 1990, pp. 1-17.

[15] D. RUELLE, *Chaotic Evolution and Strange Attractors*, Cambridge University Press, 1989.

[16] P. GRASSBERGER, I. PROCACCIA, "Characterization of Strange Attractors", *Physical Review Letters*, vol. 50, no. 5, 1983, pp. 346-349.

[17] P. FLANDRIN, O. MICHEL, P. RUIZ, "Chaos et Analyse Non Linéaire du Signal", rapport interne, Laboratoire de Physique ENS-Lyon, 1993.

[18] O. MICHEL, P. FLANDRIN, "Higher Order Statistics for Chaotic Signal Analysis", *Computer Techniques and Algorithms in Digital Signal Processing, Control and Dynamic Systems, Advances in Theory and Applications*, C.L. Leondes, Academic Press, vol. 75, 1996, pp. 105-154.

[19] A.M. FRASER, H.L. SWINNEY, "Independent Coordinates for Strange Attractors from Mutual Information", *Physical Review A*, vol. 33, no. 2, 1986, pp. 1134-1140.

[20] A.M. FRASER, "Reconstructing Attractors from Scalar Time Series: a Comparison of Singular System and Redundancy Criteria", *Physica D*, vol. 34, 1989, pp. 391-404.

[21] A. PASSAMANTE, M.E. FARREL, "Characterizing Attractors Using Local Intrinsic Dimension *via* Higher Order Statistics", *Physical Review A*, vol. 43, no. 10, 1991, pp. 5268-5274.

[22] M.B. KENNEL, R. BROWN, H.D.I. ABARBANEL, "Determining Embedding Dimension for Phase-Space Reconstruction Using a Geometrical Construction", *Physical Review A*, vol. 45, no. 6, 1992, pp. 3403-3411.

[23] D.T. KAPLAN, L. GLASS, "Direct Test for Determinism in a Time Series", *Physical Review Letters*, vol. 68, no. 4, 1992, pp. 427-430.

[24] P. COMON, "Independent Component Analysis", *Proceedings of the International Signal Processing Workshop on Higher Order Statistics*, Chamrousse (France), 1991, pp. 11-120.

[25] W.H. PRESS, B.P. FLANNERY, S.A. TEUKOLSKY, W.T. VETTERLING, *Numerical Recipes in C: the Art of Scientific Computing*, Cambridge University Press, 1988.

[26] R. BARLOW, *Statistics. A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989.

[27] O. MICHEL, "Regression Trees for Phase Space Analysis and Prediction of Non-Linear Time Series", rapport OTAN no. 21B93F, Paris, 1995.

[28] O. MICHEL, A.O. HERO, "Tree-Structured Non-Linear Signal Modeling and Prediction", *ICASSP'95 Proceedings*, Detroit, Michigan, 1995.

[29] A.M. MOOD, F.A. GRAYBILL, D.C. BOES, *Introduction to the Theory of Statistics*, Mc Graw Hill International Editions, Statistics Series, 3rd ed., 1974.

[30] M. CASDAGLI, "Nonlinear Prediction of Chaotic Time Series", *Physica D*, vol. 35, 1989, pp. 335-356.

[31] B. FINKENSTÄDT, P. KUHBIER, "Forecasting Nonlinear Economic Time Series : A Simple Test To Accompany the Nearest Neighbor Approach", *Empirical Economics*, vol. 20, 1995, pp. 243-263.

[32] O. MICHEL, A.O. HERO, A.-E. BADEL, P. FLANDRIN, "Tree based Modeling, Prediction and Analysis of Chaotic Time-Series", *Proceedings of IEEE Workshop on Non Linear Signal and Image Processing*, vol. I, Halkidiki, Greece, 1995, pp. 117-120.

[33] J.D. FARMER, J.J. SIDOROWICH, "Exploiting Chaos to Predict the Future and Reduce Noise", *Evolution, Learning and Cognition*, Ed. Lee, Y.C. (World Scientific), 1988, pp. 277-330.

[34] A.-E. BADEL, O. MICHEL, A.O. HERO, "Arbres de Régression Pour l'Analyse de Séries Chaotiques", *GRETSI'95 Proceedings*, vol. 1, Juan-les-Pins, France, 1995, pp. 169-172.

[35] W. LIEBERT, K. PAWELZIK, H.G. SCHUSTER, "Optimal Embeddings of Chaotic Attractors from Topological Considerations", *Europhysics Letters*, vol. 14, no. 6, 1991, pp. 521-526.

[36] T.P. WELDON, "An Inductorless Double-Scroll Chaotic Circuit", *American Journal of Physics*, vol. 58, no. 10, 1990, pp. 936-941.