Chapitre 4

PREDICTION ET MODELISATION DES SERIES TEMPORELLES

4.1 Introduction

La caractérisation, et *a fortiori*, l'anticipation des phénomènes physiques reposent sur l'analyse des variables qui se prêtent à l'observation. Dès lors qu'un phénomène est régi par un système d'équations déterministes connues, il est en principe possible de résoudre ce système pour prédire le résultat d'une expérimentation connaissant les conditions initiales. En revanche, lorsque ces équations sont inconnues (ou méconnues), il est nécessaire d'approximer les lois qui gouvernent l'évolution du phénomène physique à partir d'un historique de ses observations antérieures. Les séries temporelles[1] apparaissent donc naturellement dans les domaines scientifiques où la compréhension des phénomènes observés est souhaitable, sinon primordiale. Le désir d'anticiper l'évolution des phénomènes physiques a donc motivé l'emploi de méthodes adaptatives capable d'"apprendre" à extrapoler le futur d'après l'examen du passé.

Dans ce chapitre, nous portons notre attention sur la prédiction des phénomènes dont les équations sous-jacentes sont inconnues; les lois qui gouvernent l'évolution d'un système doivent être inférées. Par exemple, le mouvement d'un pendule ou le rythme des saisons portent en eux le potentiel pour prédire leur comportement futur au vu de leurs oscillations, sans avoir connaissance des mécanismes physiques sur lesquels ils reposent. Nous emploierons les termes "compréhension" et "apprentissage" pour désigner les notions complémentaires et essentielles à l'analyse des séries temporelles. La *compréhension* se réfère à la connaissance mathématique explicite des lois physiques de façon à expliquer entièrement l'évolution du système, tandis que l'*apprentissage* désigne l'émulation des variations temporelles des grandeurs physiques observées. Dans les deux cas, l'objectif

[1] Nous emploierons indifféremment les termes série/suite chronologique/temporelle pour désigner une séquence de valeurs indexées par le temps.

poursuivi est d'expliquer les observations; nous ne considérons pas à ce stade le problème du contrôle des systèmes, étroitement lié à la prédiction.

L'analyse des séries temporelles s'articule autour de trois activités complémentaires : la prédiction, la modélisation et la caractérisation. L'objet de la *prédiction* est d'anticiper fidèlement l'évolution à court terme du système; la *modélisation* s'intéresse à la description du système afin d'expliquer son évolution à long terme. Les deux objectifs ne sont pas identiques : trouver des équations présentant les propriétés du système sur le long terme n'est pas nécessairement la meilleure approche pour obtenir des prédictions correctes à court terme. Inversement, un modèle adapté au court terme peut s'avérer parfaitement inexploitable sur des intervalles temporels plus vastes. Enfin, la *caractérisation* tente, sans connaissance *a priori*, de déterminer les propriétés fondamentales telles que le nombre de degrés de liberté du système.

Dans un premier temps, un état de l'art des méthodes prédictives linéaires puis non-linéaires est présenté. Outre les méthodes statistiques traditionnelles, les fondements théoriques des réseaux de neurones sont exposés. L'accent est mis sur les nouvelles perspectives qu'offrent le DRNN à la lumière des faiblesses d'autres techniques plus classiques. Les sections suivantes sont vouées à la prédiction des suites pour lesquelles un banc-d'essai est disponible dans la littérature. Parmi celles-ci, on compte la suite des *sunspots*[2] et une grande variété de processus *chaotiques* simulés par ordinateur. Une analyse exhaustive de la prédictabilité des *sunspots* est menée dans la double perspective de souligner les propriétés nouvelles du DRNN et de valider les résultats théoriques obtenus au Chapitre 3. Ensuite est abordé le problème de la modélisation des processus chaotiques déterministes dont on sait qu'ils ne sont pas prédictibles à moyen et long terme. La tâche qui incombe au DRNN est de reproduire fidèlement le comportement du processus chaotique grâce à un sous-ensemble de variables d'état. Une fois collectées les réponses successives du réseau opérant en bouclage fermé, l'attracteur du réseau de neurones est comparé à celui du processus chaotique.

4.2 Les principes statistiques de la prédiction

Les réseaux connexionistes occupent une place centrale parmi les outils statistiques de prédiction. Ils sont aussi le thème d'un nouveau courant de recherche vivant et fécond, à la confluence de l'analyse des séries temporelles et de la théorie des systèmes dynamiques, mais qui entraîne trop souvent une technicité de langage accessible aux seuls spécialistes. Ce paragraphe est une rapide synthèse des développements contemporains, qui trouvent ici une expression claire, et constituent une culture indispensable aujourd'hui aux praticiens des réseaux connexionistes. Le DRNN sera mis en perspective avec les modèles statistiques linéaires et les approches neuronales traditionnelles, afin de mettre l'accent sur ses caractéristiques propres.

L'approche standard en prédiction consiste à construire un modèle sous-jacent qui gé-

[2] Les sunspots figurent parmi les suites notoires sur lesquelles il convient d'évaluer les capacités des modèles prédictifs. C'est pourquoi l'analyse des sunspots est effectuée dans ce chapitre en § 4.4 et non dans le chapitre consacré à l'astro-climatologie.

nère, à un instant donné t_0, la continuation $\{x_t\}_{t>t_0}$. Parce que les lois explicites qui régissent l'évolution du système observé ne sont pas connues en général, elles doivent être inférées d'après les régularités présentes dans la suite des observation passées.

Les débuts de l'analyse "moderne" des séries temporelles remonte à 1927 lorsque Yule inventa la technique auto-régressive dans l'intention de prédire le nombre annuel de taches sombres du soleil, les sunspots. Le paradigme des modèles linéaires alimentés par un bruit blanc additif a prévalu durant la moitié de siècle qui suivit. Toutefois, il s'avère inadéquat pour certaines suites déterministes parfois très simples. Par exemple, la suite obtenue par itération de la fonction logistique possède un spectre de puissance impossible à modéliser avec un modèle linéaire. La conscience des limitations dramatiques des modèles linéaires, dans certains cas de figure, est à l'origine de l'engouement récent pour des techniques prédictives non-linéaires issues de l'analyse des systèmes dynamiques.

La puissance accrue du matériel informatique conjuguée à la facilite de stockage de l'information a permis dans les années 1990 l'émergence d'algorithmes complexes dédiés au traitement des données multi-dimensionnelles. Outre les techniques de reconstruction locale dans l'espace des phases, les réseaux connexionistes participent de cette nouvelle vague d'algorithmes capables d'explorer adaptivement des bases de données bien plus vastes que par le passé.

Afin de replacer les réseaux de neurones dans le cadre des techniques prédictives, il est bon de rappeler les résultats incontournables de la théorie de la prédiction linéaire avant de donner enfin une tournure concrète à cette étude, dessinée pour l'instant de façon toute théorique.

4.2.1 Les modèles linéaires

Les modèles linéaires ont la double particularité d'être solidement ancrées à la théorie des processus aléatoires, et de fournir un cadre d'étude simple. En contre partie, la linéarité est souvent inappropriée à la modélisation, lors même que le système est gouverné par des équations analytiques d'une très grande simplicité. Notre intention dans ce chapitre est d'apporter un éclairage nouveau sur les limitations des modèles linéaires. Sauf avis contraire, les suites que nous évoquons sont stationnaires et ergodiques au second ordre.

L'origine des modèles linéaires remonte au début du siècle [Yule 27]. A cette époque, les modèles prédictifs se fondaient principalement sur une vision purement mathématique en ce sens que les processus aléatoires étaient modélisés[3] par

$$X_t = f(t) + \epsilon_t \tag{4.1}$$

où ϵ_t est la contribution stochastique, et $f(t)$ est un polynôme, une fonction rationnelle, ou une combinaison linéaire de fonctions sinusoïdales[4].

Or l'examen attentif des cycles irréguliers qu'exhibe la séquence des enregistrements annuels des taches sombres sur le soleil, encore appelée la suite de Wolfer ou les "sunspots",

[3] Il faut distinguer X_t, la variable aléatoire, de x_t, la valeur de la suite.

[4] Lorsque les fréquences w_i caractéristiques de la suite ne sont pas connues, un moyen simple est d'avoir recours à une technique de recherche basée sur une fonction appelée *périodogramme* [Priestley 93].

a conduit Yule à rompre avec l'approche mathématique traditionnelle; il lui semblait irréaliste de vouloir approximer cette séquence par une suite composée de termes strictement périodiques. Yule s'est basé sur l'étude des modèles analogues aux systèmes oscillatoires tels qu'un pendule dans un milieu résistif excité à des instants aléatoires. Le système oscillatoire exhibe des amplitudes et des cycles irréguliers à l'image des sunspots. Or il est notoire que le mouvement d'un pendule est décrit par une équation différentielle du second ordre de la forme

$$\ddot{X}(t) + \alpha_1 \dot{X}(t) + \alpha_2 X(t) = \epsilon(t) \tag{4.2}$$

où $X(t)$ désigne la déviation angulaire du pendule à l'instant t, et $\epsilon(t)$ est une impulsion aléatoire. Transposons ce système du temps continu au temps discret, Eq. 4.2 s'écrit

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} = \epsilon_t \tag{4.3}$$

ϵ_t désigne un bruit blanc, c'est-à-dire un processus aléatoire stationnaire du second ordre dont la densité spectrale de puissance est constante sur tout l'axe des fréquences. Parmi les processus stationnaires, le bruit blanc revêt une importance toute particulière parce qu'il représente l'archétype de beaucoup de phénomènes physiques. Eq. 4.3 ouvre la voie à une modélisation plus raffinée dans la mesure où l'erreur à un instant t se répercute *ad infinitum* sur l'évolution du système. Le modèle exhibe une comportement stochastique intrinsèque en comparaison au modèle additif (4.1). Clairement, le modèle auto-régressif du second ordre AR(2) (4.3) se généralise sans difficulté à l'ordre k quelconque, on obtient le modèle AR(k)

$$X_t + a_1 X_{t-1} + \ldots + a_k X_{t-k} = \epsilon_t \tag{4.4}$$

Une classe complémentaire de modèles linéaires MA(l) est représentée par les moyennes mobiles d'ordre l définies par

$$X_t = \epsilon_t + b_1 \epsilon_{t-1} + \ldots + b_l \epsilon_{t-l} \tag{4.5}$$

Les moyennes mobiles ont des propriétés quelque peu différentes mais conservent un comportement pseudo-aléatoire. Notons qu'un modèle auto-régressif est modélisable par un modèle MA(∞), et vice-versa. L'étape suivante consiste naturellement à mêler les deux modèles en un; on obtient ainsi un modèle ARMA(k, l) du type

$$X_t + a_1 X_{t-1} + \ldots + a_k X_{t-k} = \epsilon_t + b_1 \epsilon_{t-1} + \ldots + b_l \epsilon_{t-l} \tag{4.6}$$

Il est utile d'interpréter (4.6) comme l'équation d'un filtre linéaire à coefficient constants auquel on applique un "signal" d'entrée ϵ_t. La transformée en Z de (4.6) s'écrit

$$X(z) = \frac{B(z)}{A(z)} E(z) = \frac{b_0 + \ldots + b_l z^{-l}}{a_0 + \ldots + a_k z^{-k}} E(z) \tag{4.7}$$

$X(z)$ est le produit d'une fonction rationnelle et de la transformée en Z, probablement continue, de son entrée. Le nombre de pôle et de zéro de la fraction rationnelle $B(z)/A(z)$ déterminent le nombre de degrés de liberté du système, c'est-à-dire le nombre d'états passés dont se souvient le système (cf. Eq. 4.6). Notons que X_t est stable si et seulement si les zéros de A(z) sont de module inférieur à 1.

Nous avons supposé le processus ϵ_t blanc et non corrélé, sa densité spectrale de puissance $h_\epsilon(w)$ vaut $\sigma_\epsilon^2/2\pi$, où σ_ϵ^2 représente la variance de ϵ_t. La densité spectrale de puissance de X_t, $h_X(w)$, s'obtient aisément grâce à (4.7),

$$h_X(w) = \frac{\sigma_\epsilon^2}{2\pi}\left|\frac{A(e^{-iw})}{B(e^{-iw})}\right|^2 \tag{4.8}$$

D'après (4.8), les modèles ARMA ont un densité spectrale de puissance *rationnelle*. Un résultat important [Azencott 86] établit la réciproque : tout processus à spectre rationnel est un processus ARMA stationnaire centré.

Retenons que l'évaluation des paramètres $\{a_i\}$ et $\{b_i\}$ des modèles AR, MA, et ARMA reposent sur la fonction d'autocorrelation de la suite observée ou peut être déterminée grâce au spectre (4.8) par une fonction rationnelle *ad hoc*. L'ordre du modèle est généralement estimé par une heuristique [Akaike 74]. Le lecteur désireux de se plonger dans les méandres de la modélisation linéaire consultera au choix les ouvrages [Azencott 86, Box 76, Priestley 81] pour un développement plus approfondi des modèles linéaires.

Il faut mentionner que les techniques AR et ARMA sont extrapolées de méthodes déjà anciennes, puisqu'elles remontent au début du siècle.

Les limitations des modèles linéaires

Nous avons vu que les coefficients ARMA, le spectre de puissance, et la fonction d'autocorrelation contiennent essentiellement la même information caractérisant un système linéaire alimenté par un processus blanc non corrélé. Ainsi, la classe des modèles linéaires se prête à la modélisation des suites temporelles lorsque la fonction d'autocorrelation (ou indifféremment le spectre de puissance) est représentatif des caractéristiques statistiques inhérentes au comportement de la suite.

- Cette simplicité, certes attrayante, s'avère rédhibitoire pour des non-linéarités parfois excessivement simples. En effet, des suites dont les spectres de puissance sont très similaires peuvent être générées par des systèmes aux propriétés radicalement différentes. C'est le cas de certains systèmes déterministes.

- Dans le même registre, l'autocorrelation de X_t avec X_{t-j} ($j \in \mathbf{Z}$) ne présente qu'un aspect de la distribution des couple (X_t, X_{t-j}). D'autres aspects vitaux peuvent être omis (l'espérance conditionnelle $\mathrm{E}(X_t|X_{t-j})$ se réduit à une simple fonction linéaire dans le cas où la séquence $\{X_t\}$ est gaussienne).

- Enfin, l'évolution "libre" des processus ARMA (ϵ_t constant) ne permet pas un comportement périodique stable indépendant des valeurs initiales. De surcroît, le signal généré est caractérisé par un nombre fini de fréquences.

Prenons deux exemples classiques afin de mettre à jour les limites des modèles linéaires.

– Considérons la suite générée par

$$x_{t+1} = rx_t(1 - x_t) \tag{4.9}$$

Popularisée dans le cadre des modèles écologiques idéalisés, ce modèle simple possède une dynamique complexe [Ott 93]. Imaginons une famille d'insectes se reproduisant par la ponte de z_n œufs en moyenne, au cours de l'année n. Chaque insecte pond en moyenne r œufs par an mais dévore les œufs en cas de pénurie alimentaire. Si $\overline{z}$ désigne la taille critique de la population, à partir de laquelle la nourriture a été entièrement consommée et tous les insectes dévorent leurs œufs, le nombre effectif d'œufs pondus par insecte s'approxime alors par la relation $r(1 - z_n/\overline{z})$. Cela conduit à la relation $z_{n+1} = rz_n(1 - z_n/\overline{z})$ identique à (4.9) après le changement de variable $x_t = z_t/\overline{z}$. Cette équation décrit par ailleurs un certain nombres de réactions chimiques et de flots hydrodynamiques. Le paramètre r agit de manière qualitative sur le comportement de la suite, variant du point fixe au chaos déterministe. Par exemple pour $r = 4$, chaque itération détruit un bit d'information; du fait de la non-inversibilité de (4.9) et la symétrie de la *densité naturelle invariante*, les prédécesseurs de x_t au nombre de deux sont équiprobables. En d'autres termes, la connaissance de x_t à ϵ près ne permet la connaissance de son successeur, x_{t+1}, qu'avec une précision de 2ϵ. La croissance exponentielle de l'incertitude est une caractéristique du chaos déterministe.

– Le deuxième exemple est tout aussi simple. Considérons la suite générée par

$$x_{t+1} = 2x_t (\text{mod } 1) \tag{4.10}$$

L'action du modulo se comprend aisément en considérant x_t sous forme binaire (i.e. $x_t = 0d_1d_2\ldots = (d_1.2^{-1}) + (d_2.2^{-2}) + \ldots$); chaque itération décale la décomposition binaire d'un rang vers la gauche ($d_i \leftarrow d_{i+1}$). Il résulte que le bit le plus significatif d_1 est éliminé, cédant la place à d_2. L'implémentation physique de (4.10) est la boule de billard èvoluant sans dissipation d'énergie; x_t peut être vu comme les positions successives sur une droite quelconque dans le plan du billard que la boule croise après chaque rebond [Moore 91].

L'évolution de ces deux systèmes uni-dimensionnels déterministes est entièrement déterminée par la condition initiale x_0 comme l'est le processus AR piégé sur une orbite périodique stable. Cependant, ces systèmes élémentaires (une parabole et une droite) génèrent chacun une suite de valeurs possédant une *densité* spectrale inhérente au modèle lui même, par opposition à un spectre discretisé. Dans le contexte des modèles ARMA, c'est le bruit externe en entrée du système qui pourvoit le modèle d'une densité spectrale, alors que dans le cas présent, cette composante est inhérente à ces systèmes. Il faut conclure de cela que la non-linéarité est essentielle dans les systèmes déterministes pour produire un comportement d'apparence aléatoire.

Historiquement, un seuil important a été franchi avec l'avènement en 1980 des modèles semi-linéaires de Tong et Lim [Tong 80] après le règne des modèles linéaires durant plus d'un demi-siècle. Le modèle *auto-régressif à seuils* (TAR)[5] qu'ils proposent est globalement non-linéaire : l'idée consiste à compartimenter l'espace des phases et d'appliquer dans chaque région un modèle ARMA. Toutefois, cette approximation "linéaire par morceaux" nécessite un grand nombre de partitions lorsque la surface présente des non-linéarités quadratiques (e.g. la fonction logistique). Un autre moyen naturel d'étendre Eq. 4.6 fait appel à un polynôme d'ordre supérieur; ces modèles portent le nom de suites de Volterra [Volterra 59].

Les modèles TAR, les suites de Volterra, les fonctions splines, radiales, etc. [Casdagli 89, Casdagli 92] élargissent considérablement le champs des fonctionalités pour modéliser les suites temporelles mais au prix d'une plus grande complexité dans l'adaptation des paramètres. Pour que des modèles non-linéaires soient utiles, un moyen d'exploiter la non-linéairité des données s'impose pour guider la construction du modèle. Or le manque de compréhension de ce problème a sérieusement freiné leur extension.

4.2.2 La prédiction neuronale

Les réseaux de neurones ont récemment émergé sur le terrain de la prédiction et de la reconnaissance de suites temporelles [Connor 94, Gershenfeld 93, Lapedes 87, Mozer 93, Wan 93a, Weigend 90]. Cet engouement tient à la versatilité des réseaux non-bouclés à 3 couches capables d'approximer une fonction non-linéaire arbitraire [Cybenko 89] et à l'efficacité de la descente du gradient [Rumelhart 86]. De surcroît, les promesses qu'offrent les réseaux de neurones se fondent essentiellement sur leur capacité à émuler des caractéristiques non-anticipées des suites temporelles, sans rien connaître des équations sous-jacentes qui les gouvernent. Ces perspectives ont contribué à l'essor des réseaux connexionistes jusqu'à nos jours, et constituent la principale motivation de cette étude.

En dépit de la versatilité des modèles connexionistes classiques, une classe de suite qui ne se soumet pas à la représentation neuronale statique est exhibée dans ce paragraphe; nous en déduirons la nécessité d'introduire une récurrence temporelle interne au sein du modèle lui-même.

Les fondations théoriques de la prédiction neuronale

Considérons le processus d'apprentissage d'un réseau de neurones statique, i.e. le réseau ne possède aucune mémoire interne. Pour un vecteur d'entrée $\mathbf{x}(k)$ le réseau génère une sortie

$$\mathbf{y}(k) = \mathcal{N}(\mathbf{w}, \mathbf{x}(k)) \tag{4.11}$$

où $\mathbf{w}$ représente les paramètres du modèle, i.e. le vecteur de poids. Cette représentation statique exclue le DRNN mais englobe les réseaux à délais *non-bouclés* (e.g. TDNN, FIR, etc.) si l'on considère leur représentation dépliée. Nous avons vu au § 2.4.1 que l'adaptation du vecteur de poids est realisé de manière à minimiser l'erreur quadratique moyenne calculée sur la base d'apprentissage

[5] On retrouvera le modèle TAR de Tong et Lim dans l'examen comparatif des techniques prédictives appliquées au sunspots, pp. 101.

$$\mathrm{Min}_{\mathbf{w}} \frac{1}{T} \sum_{k=1}^{T} \|\mathbf{d}(k) - \mathcal{N}(\mathbf{w}, \mathbf{x}(k))\|^2 \tag{4.12}$$

où $\mathbf{d}(k)$ représente le vecteur des sorties désirées. Selon nos hypothèses de stationnarité et d'ergodicité au sens large [Charbit 90], la moyenne statistique converge vers l'espérance mathématique

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \|\mathbf{d}(k) - \mathcal{N}(\mathbf{w}, \mathbf{x}(k))\|^2 = \mathbf{E}(\|\mathbf{D} - \mathcal{N}(\mathbf{w}, \mathbf{X})\|^2) \tag{4.13}$$

L'espérance est établie sur le couple des variables aléatoires $\mathbf{D}$ et $\mathbf{X}$. Si $\mathcal{N}$ était linéaire, la matrice de poids optimale $\mathbf{w}^\star$ serait aisément obtenue par l'approche de Wiener [Priestley 81]. Mais en raison de la non-linéarité du modèle, il n'existe pas de formule explicite pour $\mathbf{w}^\star$. Afin de poursuivre cette analyse, nous supposons que $\mathcal{N}$ implémente une fonction *universelle*, hypothèse partiellement justifiée par un résultat notoire qui stipule que les réseaux de neurones multi-couches sont capables d'approximer n'importe quelle fonction *continue*[6] [Cybenko 89]. Aussi, si T est suffisamment grand, (4.12) devient

$$\mathrm{Min}_{\mathcal{N}} \mathbf{E}(\|\mathbf{D} - \mathcal{N}(\mathbf{X})\|^2) \tag{4.14}$$

Les variables aléatoires considérées sont à carré intégrable. La solution $\mathcal{N}^\star$ n'est autre que la projection de $\mathbf{D}$ sur le sous-espace d'Hilbert engendré par les variables du vecteur $\mathbf{X}$, c'est-à-dire l'espérance conditionnelle

$$\mathcal{N}^\star = \mathbf{E}[\mathbf{D}|\mathbf{X}] \tag{4.15}$$

Ce résultat est intuitif. Il faut cependant garder à l'esprit qu'un réseau de neurones ne peut qu'approcher $\mathcal{N}^\star$, et donc introduit un *biais* dans la solution. Ce biais est d'autant plus grand que le nombre de neurones est petit. De plus l'apprentissage sur un nombre limité d'exemples introduit une *variance* parmi les l'ensemble des solutions vers lesquelles le réseau peut converger. Cette variance augmente avec la taille du réseau. Un réseau de grande taille est caractérisé par un faible *biais* mais une *variance* élevée et vice versa. Il s'agit d'un compromis bien connu dans la théorie de la régression [Geman 92].

Puisque le processus est stationnaire, le prédicteur optimal est l'espérance conditionnelle de X_t connaissant la suite infinie des observations passées $X_{t-1}, X_{t-2}, \ldots$, soit

$$\hat{X}_k = \mathbf{E}[X_k | X_{k-1}, X_{k-2}, \ldots] \tag{4.16}$$

Or, son calcul directe est trop complexe; il exige la connaissance de la loi conjointe et il est souvent impossible de disposer de toutes les valeurs de $\{x_{t-j}\}_{j=1,\ldots,\infty}$. Nous sommes donc naturellement amenés à substituter $\mathbf{x}(k)$ et $\mathbf{y}(k)$ par

[6] Ce résultat ne dit pas, en revanche, combien de neurones sont nécessaires pour approximer une fonction donnée...

$$\begin{aligned} \mathbf{x}(k) &= [x_{k-1}, x_{k-2}, \ldots, x_{k-M}]^T \\ \mathbf{y}(k) &= x_k \end{aligned} \tag{4.17}$$

dans (4.11) pour obtenir l'estimateur optimal étant donné le passé *limité* $\{X_{t-j}\}_{j=1,\ldots,M}$. Ainsi, $\mathcal{N}^\star$ est simplement donné par

$$\mathcal{N}^\star = \mathbf{E}[X_k | X_{k-1}, X_{k-2}, \ldots, X_{k-M}] \tag{4.18}$$

Enfin, notons que les prédictions itérées sont fournies en réinjectant les estimations en entrée du réseau lorsque l'entrée n'est plus disponible. Ainsi au-delà de l'itération M, le réseau opère en *bouclage fermé* et implémente un système dynamique autonome dont la stabilité est assurée par les bornes imposées aux activations des neurones.

4.2.3 L'auto-régression non-linéaire

La théorie des systèmes dynamiques présente de profondes similitudes avec l'analyse des séries temporelles. Or, nous allons voir que l'auto-régression non-linéaire (NAR) est justifiée pour une certaine classe de processus, mais qu'elle est en revanche, inadaptée lorsque la séquence des observations ne permet pas de reconstruire le vecteur d'etat interne du système.

Les systèmes linéaires

Nous allons rompre avec les méthodes et le langage de l'analyse des séries temporelles pour adopter une représentation des systèmes linéaires dans l'espace des variables qui a fait le succès des filtres de Kalman dans les années 1960. Considèrons, par exemple, le modèle AR(2)

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} = \epsilon_t \tag{4.19}$$

En posant $X_t^{(2)} = X_t, X_t^{(1)} = -a_2 X_{t-1}^{(2)}$, (4.19) peut s'écrire sous forme vectorielle

$$\begin{bmatrix} X_t^{(1)} \\ X_t^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & -a_2 \\ 1 & -a_1 \end{bmatrix} \begin{bmatrix} X_{t-1}^{(1)} \\ X_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \epsilon_t \tag{4.20}$$

Le scalaire X_t s'obtient par

$$X_t = [0, 1] \begin{bmatrix} X_t^{(1)} \\ X_t^{(2)} \end{bmatrix} \tag{4.21}$$

La paire d'équations (4.20) et (4.21) est équivalente à (4.19), à la différence près que la représentation vectorielle est d'ordre 1, c'est-à-dire un *processus de Markov*. La méthode se généralise aux modèles ARMA (4.6) par l'addition d'un bruit η_t, on obtient le système

$$\begin{aligned} \mathbf{z}_{t+1} &= \mathbf{F}\mathbf{z}_t + \mathbf{b}\epsilon_t \\ x_t &= \mathbf{c}^T\mathbf{z}_t + \eta_t \end{aligned} \tag{4.22}$$

où $\mathbf{b}$ et $\mathbf{c}$ sont des vecteurs $M \times 1$ et $\mathbf{F}$ une matrice $M \times M$. (4.22) s'interprète comme l'enregistrement bruité d'un processus expérimental gouverné par la première équation. Si nous cherchons à évaluer $\mathbf{z}_t$ (problème de filtrage), ou dans notre cas $\mathbf{z}_{t+m}$ (problème de prédiction), au moyen d'une fonction borelienne *optimale* des observations $x_t, x_{t-1}, \ldots, x_0$ telle que l'approximation à l'instant t de $\mathbf{z}_{t+m}$ soit donnée par $\hat{\mathbf{z}}_{t+m,t} = f(x_t, x_{t-1}, \ldots, x_0)$, nous sommes amenés à prendre l'espérance conditionnelle

$$\hat{\mathbf{z}}_{t+m,t} = \mathbf{E}[\mathbf{z}_{t+m}|x_t, x_{t-1}, \ldots, x_0] \tag{4.23}$$

Sous l'hypothèse que les variables $\{\epsilon_t, \eta_t\}$ constitue une suite Gaussienne, le Théorème de Kalman [Bouleau 88, Kalman 61] affirme que

$$\hat{\mathbf{z}}_{t,t} = \mathbf{F}\hat{\mathbf{z}}_{t-1,t-1} + \mathbf{K}(x_t - \mathbf{c}^T\mathbf{F}\hat{\mathbf{z}}_{t-1,t-1}) \tag{4.24}$$

où la matrice $\mathbf{K}$ est connue sous le nom de "gain de Kalman"; $\hat{\mathbf{z}}_{t+m,t}$ s'obtient selon la formule

$$\hat{\mathrm{x}}_{t+m,t} = \mathbf{F}^m\hat{\mathrm{x}}_{t,t} \tag{4.25}$$

Ainsi, il est possible de déduire de la séquence d'observations passées, un estimateur optimal de l'état du système m cycles suivants.

A quelle condition est-il possible de reconstituer le vecteur d'état, $\mathbf{z}_t$, au moyen d'une fenêtre temporelle de taille M sur les observations x_t? Pour cela, le système doit être *observable* [Kailath 80]. Eliminons le bruit dans (4.22) et considérons la matrice $M \times M$, dite *matrice d'observabilité*

$$\mathbf{O} = \begin{bmatrix} \mathbf{c}^T \\ \mathbf{c}^T\mathbf{F} \\ \vdots \\ \mathbf{c}^T\mathbf{F}^{M-1} \end{bmatrix} \tag{4.26}$$

Il vient,

$$[x_t, x_{t+1}, \ldots, x_{t+M-1}]^T = \mathbf{O}\mathbf{z}_t \tag{4.27}$$

Aussi, le système est observable si et seulement si $\mathbf{O}$ est inversible. Dans ce cas, x_t est donné par la relation

$$x_t = \mathbf{c}^T\mathbf{F}^M\mathbf{O}^{-1}[x_{t-M}, x_{t-M+1}, \ldots, x_{t-1}]^T \tag{4.28}$$

Maintenant, lorsque $\mathbf{O}$ n'est pas inversible, alors le système n'est pas observable et il est impossible de reconstruire l'état $\mathbf{z}_t$ à partir d'une séquence de M observations x_t du système. C'est exemple souligne les limites qui se posent aux méthodes nécessitant l'emploi d'une fenêtre temporelle même dans le cas simple des systèmes linéaires.

Les sytèmes non-linéaires

Poursuivons l'analyse précédente dans le domaine non-linéaire car l'écrasante majorité des suites observées sont régies, de toute évidence, par des systèmes non-linéaires. Une suite temporelle peut être perçue comme la partie observable d'un système dynamique complexe, régi par un système d'équations aux différences aux variables inconnues, i.e. qui ne se prêtent pas à la mesure. C'est un cas fréquent en météorologie par exemple (e.g. équations de Lorenz) mais aussi observé dans de nombreux domaines de la physique expérimentale (e.g. mouvement d'une particule sous l'action d'un champs électromagnétique, résonance du laser, etc.).

L'état d'un système déterministe est l'information nécessaire à la description de son entière évolution dans le futur. Un paradigme utile à garder à l'esprit est un système d'équations différentielles de dimension finie. Abandonnons un court instant le domaine linéaire pour une représentation des plus générales d'un système dynamique

$$\begin{aligned} \frac{d\mathbf{z}(t)}{dt} &= \mathbf{G}(\mathbf{z}(t)) \\ x(t) &= h(\mathbf{z}(t)) \end{aligned} \tag{4.29}$$

où $\mathbf{G}$ désigne une fonction non-linéaire qui gouverne les transitions d'états, h est la fonction d'observation et $\mathbf{z}(t)$ est le vecteur d'état à l'instant t.

Sous certaines conditions sur la forme des équations, il existe une trajectoire *unique* ayant pour condition initiale $\mathbf{z}_0$. Supposons que seule une des variables du vecteur d'état, x_t, soit mesurable (e.g., la température en un point), nous avons vu que dans le cas où $\mathbf{G}$ et h sont linéaires, le filtre de Kalman [Kalman 61] est l'estimateur optimal du système [Gershenfeld 93, Priestley 93]. Que dire lorsque ces fonctions sont non-linéaires?

L'idée originale de Yule était d'améliorer les prédictions en incorporant dans le modèle les précédentes valeurs de la suite. Il existe, en fait une, lien plus étroit la fenêtre temporelle et la dynamique sous-jacente. Ce lien a été mis en évidence par Ruelle [Ruelle 85], Packard et al. [Packard 80] et Takens [Takens 80] dans les années 1980. En effet, il s'avère qu'un vecteur de délai de longueur suffisante permet de recouvrer la structure géométrique du système étudié; en d'autres termes, il est possible d'anticiper exactement l'évolution du système grâce à un vecteur de délais

$$[x(t-\tau), x(t-2\tau), \ldots, x(t-M\tau)] \tag{4.30}$$

Le résultat établi par Takens [Packard 80, Sauer 93, Takens 80] stipule que pour un grand nombre de systèmes déterministes exempts de bruit, il existe un *difféomorphisme* Ψ (i.e. fonction bijective et différentiable) entre $\mathbf{x}(t)$ et $\mathbf{z}(t)$, pour quasiment tous $\mathbf{G}(), h(), \tau$, dès lors que M est deux fois plus grand que la *dimension fractale*[7] de l'attracteur [Ott 93]. Aussi,

$$\mathbf{z}(t-\tau) = \Psi\left(x(t-\tau), x(t-2\tau), \ldots, x(t-M\tau)\right) \tag{4.31}$$

[7] La dimension fractale est une mesure du nombre effectif de degrés de liberté du système dynamique [Ott 93].

En echantillonnant x aux instants $\Delta t = \tau$, il vient l'auto-régression

$$x(t) = h \circ (\tau \mathbf{G} + \mathbf{I}) \circ \Psi \left(x(t-\tau), x(t-2\tau), \ldots, x(t-M\tau) \right) \tag{4.32}$$

Ce résultat fondamental, analogue aux systèmes linéaires, montre qu'il existe une auto-régression non-linéaire capable de prédire très précisément les observations futures. Puisque le théorème ne dit pas comment trouver la fonction non-linéaire sous-jacente, on cherchera à approximer la composée $h \circ (\tau \mathbf{G} + \mathbf{I}) \circ \Psi$ par un réseau de neurones alimenté par un vecteur de délais. On comprends mieux, dès lors, pourquoi l'usage d'une fenêtre temporelle en entrée d'un réseau neuronal statique est devenue très en vogue ces dernières années [Farmer 88, Lapedes 87, Wan 93a, Weigend 90].

Quand bien même la fréquence d'échantillonnage des observations est égale ou supérieure à $1/\tau$, la technique du fenêtrage temporel pose toutefois deux problèmes difficiles :

- Comment accommoder habilement la taille de la fenêtre d'entrée, M, aux régularités temporelles présentes dans la suite? Il est hélas impossible de déterminer par avance la plage temporelle utile, celle qui capture toutes les contingences temporelles du processus. Faute d'information empirique sur le signal observé, la fonction d'autocorrelation nous aiguillera dans ce choix, tout en gardant à l'esprit qu'elle ne visualise que les moments statistiques d'ordre 2. Or l'impossibilité de choisir convenablement M gène considérablement la technique de fenêtrage temporel pour pallier l'absence de mémoire interne des réseaux statiques. Une valeur de M trop grande complique la tâche du réseau et augmente le coût de l'apprentissage. Inversement, une valeur de M trop faible nuit inéluctablement aux performances du modèle car l'information utile est tronquée. L'usage d'une mémoire *spatiale* n'est donc qu'une solution précaire, un palliatif bien utile dans nombre de cas, mais dont les limites sont patentes. Les problèmes intéressants en reconnaissance de séquence, ceux pour lesquels l'information pertinente se repartit sur des intervalles de temps non bornés, ne se soumettent pas à cette représentation spatiale de la mémoire.

- Un réseau non-bouclé n'est pas toujours capable de reconstruire le vecteur d'état du système dynamique qui a généré la suite. En cela, on rejoint le principe *d'observabilité* introduit par Kalman.

  Considérons par exemple une suite issue de l'observation d'un système dynamique unidimensionnel dont les équations sont données par

$$\left\{ \begin{array}{lcl} x_t & = & f_1(z_t) \\ z_t & = & f_2(z_{t-1}) \end{array} \right. \tag{4.33}$$

  Le prédicteur optimal $\hat{x}_{t+1} = f_1(f_2(f_1^{-1}(x_t)))$ requiert l'inversibilité de $f_1()$. Aussi, l'impossibilité de reconstruire l'état sous-jacent de certain systèmes dynamiques est la pierre d'achoppement des réseaux **non-bouclés**.

4.2.4 Le modèle neuronal ARMA non-linéaire

Pour dépasser les limites de l'auto-régression non-linéaire, il faut revenir au problème originel de l'analyse stastique des séries temporelles : prédire, c'est trouver une relation fonctionnelle, probablement non-linéaire, entre les valeurs passées et la valeur courante de telle sorte que l'erreur d'estimation se réduise à un bruit blanc. Un modèle causal idéal, que symbolise la fonction $h()$, pour la séquence $\{X_t\}$ s'écrit

$$X_t - h(X_{t-1}, X_{t-2}, \ldots) = \epsilon_t \tag{4.34}$$

où ϵ_t est un processus blanc centré. Si la séquence des résidus $\{\epsilon_t\}$ ne formait pas un bruit blanc, il y aurait une structure dans la relation entre X_t et son passé dont le modèle ne tiendrait pas compte. Posé en ces termes, le problème est trop général pour être résolu car pour un nombre fini de données, il est impossible de spécifier $h()$ parmi toutes les fonctions non-linéaires possibles. Un moyen de surmonter cette difficulté est de réduire de manière drastique l'espace (infini) des fonctions candidates à un sous ensemble de fonction paramètrables. Nous avons déjà abordé la classe des fonctions linéaires et établi les limites auxquelles nous nous heurtons dans le contexte la modélisation des suites irrégulières ou chaotiques. D'autre part, lorsque $t \to \infty$, il n'est pas possible de disposer de toutes les valeurs de $\{x_{t-j}\}$, il est donc nécessaire de restreindre le nombre d'entrée de $h()$.

Dans cette optique, concentrons nous sur un modèle très général susceptible d'avoir généré la suite de valeurs $\{x_t\}$; considérons l'extension non-linéaire des modèles ARMA que nous avons brièvement esquissés au chapitre précédent,

$$X_t = h(X_{t-1}, \ldots, X_{t-p}; \epsilon_{t-1}, \ldots, \epsilon_{t-q}) + \epsilon_t \tag{4.35}$$

La nouveauté par rapport à (4.34) réside dans le fait que x_t ne dépend pas seulement des perturbations aléatoires au même instant, représenté par le bruit blanc ϵ_t, mais garde une mémoire des perturbations antérieures jusqu'à t-q. On compense ainsi le manque d'observations passées par une mémoire des perturbations.

Le prédicteur optimal associé est donné par l'espérance conditionnelle

$$\hat{X}_t = \mathbf{E}[h(X_{t-1}, \ldots, X_{t-p}; \epsilon_{t-1}, \ldots, \epsilon_{t-q}) | X_{t-1}, \ldots] \tag{4.36}$$

Si le modèle inversible, i.e. $\epsilon_t = f(X_t, \ldots)$, alors Eq. 4.36 devient

$$\hat{X}_t = h(X_{t-1}, \ldots, X_{t-p}; \epsilon_{t-1}, \ldots, \epsilon_{t-q}) \tag{4.37}$$

Il est donc envisageable d'approximer $h()$ par un modèle neuronal, dans la mesure où l'entrée est maintenant de taille finie. Or le problème qui surgit maintenant a trait à la nature même des entrées : on ne connait pas le bruit, ϵ_t. Il nous faut donc se restreindre à une approximation $\mathcal{N}$ du prédicteur optimal $h()$ donnée par

$$\mathcal{N}^\star = h(X_{t-1}, \ldots, X_{t-p}; e_{t-1}, \ldots, e_{t-q}) \tag{4.38}$$

où $e_j = X_j - \hat{X}_j$. Remarquons que $e_t \neq \epsilon_t$ dans la mesure où dans e_t, il y a non seulement la partie imprévisible, ϵ_t, mais aussi l'erreur d'approximation de la fonction $h()$ que fait le réseau de neurones.

La formule (4.38) est une généralisation naturelle des prédicteur linéaires [Box 76, Connor 94] et fournit un cadre formel pour modéliser une grande variété de suites régies par des équations non-linéaires. L'avantage qu'ont les processus NARMA sur les processus NAR se conçoit comme celui qu'ont les processus ARMA sur les processus NAR [Connor 94].

Considérons une écriture équivalente de (4.38) dans laquelle nous substituons e_t par $\hat{X}_t$

$$\hat{X}_t = \mathcal{N}(X_{t-1}, X_{t-2}, \ldots, X_{t-p}; \hat{X}_{t-1}, \hat{X}_{t-2}, \ldots, \hat{X}_{t-q}) \tag{4.39}$$

La recurrence externe permet de dépasser les limites des modèles statiques uniquement alimentés par les observations retardées. Il est, en effet, possible *en théorie* de prédire la suite $\{x_t\}$ générée par le système d'équations :

$$\left\{ \begin{array}{lcl} x_t & = & f_1(z_t) \\ z_t & = & f_2(z_{t-1}, z_{t-2}, \ldots, z_{t-p}) \end{array} \right. \tag{4.40}$$

par 2 réseaux du type (4.39) en cascade

$$\left\{ \begin{array}{lcl} \hat{x}_t & = & \mathcal{N}_1(y_t) \\ y_t & = & \mathcal{N}_2(y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \end{array} \right. \tag{4.41}$$

Une question urgente se pose au vu des ces équations : la connaissance de $\{x_{t-j}\}_{j \leq p}$ et $\{\hat{x}_{t-j}\}_{j \leq q}$ suffit-elle à l'algorithme d'apprentissage pour permettre au réseau d'approximer correctement la variable interne z_t du système, par un variable interne y_t, ainsi que son évolution à chaque itération? Nous tenterons d'apporter quelques éléments de réponse au travers des expérimentations menées sur les processus chaotiques dans la suite de ce chapitre.

Signalons d'autre part que la rétro-propagation standard est désormais inappropriée pour évaluer correctement l'ajustement des poids, et ce, en raison de la recurrence externe introduite par la seconde fenêtre temporelle sur les sorties différées du réseau. Hormis cette difficulté, l'espoir est d'améliorer la capacité prédictive du réseau dans la mesure où plus d'information est explicitement mise à sa disposition.

Paradoxalement, la prédiction basée sur les modèles NARMA n'a pas rencontrer un écho significatif dans la littérature connexioniste, en raison des problèmes inhérents à l'apprentissage des réseaux récurrents. Le charge de calcul supplémentaire que requiert le "credit assignment" n'a probablement pas été jugée nécessaire pour motiver son utilisation. A l'inverse, l'intérêt s'est porté sur les modèles NAR (que l'on retrouve en remplaçant e_t par zéro dans Eq. 4.38) de manière à bénéficier des atouts d'un apprentissage à faible coût.

4.2.5 Modéliser un système dynamique

Consacré au problème spécifique de la commande optimale des systèmes dynamiques, ce paragraphe franchit le pas entre la représentation d'une série temporelle, en tant qu'observation d'un système dynamique, et la commande du système lui-même. Seul le sens de l'entrée externe est quelque peu différent : dans un cas il s'agit d'un bruit de mesure indésirable mais inéluctable, dans l'autre, le terme est précisément la commande du système qui permet d'agir sur ce dernier de façon à diriger sa trajectoire.

Considérons un système dynamique non-linéaire discret qui, à une commande u_t en entrée, fournit en sortie une réponse y_t. Ce système est régi par une équation d'état vectorielle du premier ordre

$$\begin{aligned} \mathbf{z}_{t+1} &= \mathbf{G}(\mathbf{z}_t, u_t) \\ x_t &= h(\mathbf{z}_t, u_t) \end{aligned} \tag{4.42}$$

Il est bien connu en théorie du contrôle que le modèle dit "state-output model" reliant les états passés à l'état présent décrit une grande variété de systèmes dynamiques non-linéaires [Narendra 91, Srinivasan 94]. Le state-output model est défini par

$$\hat{\mathbf{z}}_t = \mathbf{G}(\mathbf{z}_{t-1}, u_{t-1}) \quad \text{et} \quad \hat{x}_t = h(\mathbf{z}_t, u_t) \tag{4.43}$$

où $\hat{\mathbf{z}}_t$ représente l'estimation du vecteur d'état du système et $\hat{x}_t$ l'estimation de l'observation. Les fonctions non-linéaires $\mathbf{G}()$ et $h()$ sont statiques et par conséquent peuvent être approximées par des réseaux de neurones non-bouclés [Cybenko 89]. Si le vecteur d'état du système est mesurable en plus de sa sortie, alors l'apprentissage de $\mathbf{G}()$ et $h()$ est découplé. Le vecteur d'état du système, $\mathbf{z}_t$, peut être utilisé comme sortie désirée du premier réseau, tandis que x_t est utilisé pour adapter le second. N'importe quel algorithme d'apprentissage, tel que la rétro-propagation, fera l'affaire.

Or si, comme à l'accoutumée, le vecteur d'état n'est pas disponible, alors le modèle NARMA suivant est employé

$$\hat{x}_t = f(x_{t-1}, \ldots, x_{t-n}, u_t, u_{t-1}, \ldots, u_{t-m}, \mathrm{W}_f(t)) \tag{4.44}$$

Ce modèle s'appelle série-parallèle, ou "teacher forcing model", dans la littérature. Bien que le système soit dynamique, la fonction non-linéaire à approximer, $f()$, est statique et par conséquent peut être modélisée par réseau non-bouclé. Aussi, la rétro-propagation peut, une fois encore, être employée.

La capacité de représentation du modèle NARMA (4.44) est inférieure à celle du state-output model (4.43) dans la mesure où il ne mémorise pas le vecteur d'état. Le modèle utilise les entrées et sorties retardées du système pour reconstruire les états. Si cette reconstruction'est pas possible, le modèle NARMA est incapable de représenter le système. Plus précisément, il convient seulement aux problèmes pour lesquels $h()$ est inversible par rapport à $\mathbf{z}_t$, faute de quoi la reconstruction à partir des sorties retardées est ambiguë. On retrouve ici le problème identique à celui rencontré avec les series. Srinivasan et al. fournissent un exemple concret [Srinivasan 94]; ils considèrent le système dynamique défini par les équations

$$\begin{aligned} z_t &= a z_{t-1} + b u_t \\ x_t &= z_t^2 \end{aligned} \tag{4.45}$$

Ainsi le prédicteur optimal du système (4.45) est défini par

$$\hat{x}_t = (a\sqrt{x_{t-1}} + bu_t)^2 \tag{4.46}$$

Puisque $\sqrt{x_{t-1}}$ n'est pas unique, aucun modèle NARMA ne peut être employé pour représenter (4.45).

L'impossibilité de reconstruire l'état sous-jacent de certain systèmes dynamiques est la pierre d'achoppement des modèles NARMA. Connor et al. [Connor 94] ont clairement démontré qu'un réseau NARMA à 3 couches de neurones dans lequel la sortie est réinjectée en entrée est équivalent à une réseau qui admet des connexions récurrentes non ajustables de la couche cachée vers elle-même. Il est donc légitime de vouloir munir les réseaux de neurones de connexions totalement récurrentes et adaptables dans l'optique d'étendre leurs fonctionalités et de pallier les faiblesses des modèles NARMA. Dans cet esprit, on peut envisager l'interconnexions de plusieurs sous-systèmes du type

$$\hat{x}_t = f(\hat{x}_{t-1}, \ldots, \hat{x}_{t-n}, u_t, u_{t-1}, \ldots, u_{t-m}, \mathrm{W}_f(t)) \tag{4.47}$$

appelés "output error models". Ici encore, il est crucial de noter que $f()$ dépend désormais de ses précédentes sorties ce qui rend impossible l'application tel quel de la procédure back-propagation [Rumelhart 86]. La récursion au sein du réseau de neurones est nécessaire à l'implémentation de système dynamiques complexes.

Aussi est-il possible, en théorie, de modéliser n'importe quel système dynamique du type 4.42 par l'interconnexion d'un réseau bouclé et d'un réseau non-bouclé tels que

$$\begin{aligned} \mathbf{z}_t &= \mathcal{N}_1(\mathbf{z}_{t-1}, u_t, W_f(t)) \\ \hat{x}_t &= \mathcal{N}_2(\mathbf{z}_t, u_t, W_g(t)) \end{aligned} \tag{4.48}$$

En somme, l'avènement des réseaux récurrents n'est donc pas un vaine tentative visant à complexifier les modèles existant mais bel et bien une nécessité pour aborder certains problèmes complexes ayant trait à la dynamique des systèmes.

4.2.6 Le DRNN : un filtre de Kalman non-linéaire

Nous nous sommes attachés, dans ce chapitre, à établir les points forts et les faiblesses de quelques grandes familles de modèles prédictifs, dans un langage à mi-chemin entre le formalisme pur et la comprehension intuitive des situations concrètes. Le moment est désormais venu de mettre l'accent sur ce qui distingue le DRNN des autres modèles traditionnels, ainsi que sur les problèmes qui demeurent encore ouverts à ce jour.

De la même manière que le modèle neuronal NARMA

$$\hat{x}_t = \mathcal{N}(x_{t-1}, x_{t-2}, \ldots, x_{t-p}; \hat{x}_{t-1}, \hat{x}_{t-2}, \ldots, \hat{x}_{t-q}) \tag{4.49}$$

étend le modèle ARMA au domaine non-linéaire, le réseau dynamique et récurrent est une généralisation des filtres de Kalman. Au lieu de disposer la mémoire spatialement en entrée de la "boite noire", le vecteur d'état $\mathbf{z}_t$ confère une mémoire interne au réseau.

Le DRNN totalement bouclé peut donc être représenté, d'une manière schématique, par deux sous-réseaux NARMA en cascade, $\mathcal{N}_1$ et $\mathcal{N}_2$,

$$\begin{aligned} \mathbf{z}_t &= \mathcal{N}_1(\mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-D}; x_t, \ldots, x_{t-D}) \\ \hat{x}_{t+1} &= \mathcal{N}_2(\mathbf{z}_t) \end{aligned} \tag{4.50}$$

$\mathcal{N}_2$ est alimenté par la sortie de $\mathcal{N}_1$, lequel peut à son tour être décomposé en deux sous-réseaux en cascade, et ainsi de suite. Le vecteur $\mathbf{z}_t$ dans la représentation (4.50) du DRNN regroupe les activations de tous les neurones qui ne recoivent par de signaux du monde exterieur, i.e. tous les neurones excepté les neurones d'entrée.

Désignons par $\tilde{\mathbf{z}}_t$, le vecteur formé de l'activation de tous les neurones aux instants $t, t-1, \ldots, t-D+1$, soit ND valeurs,

$$\tilde{\mathbf{z}}_t = [\mathbf{z}_t, \ldots, \mathbf{z}_{t-D+1}; x_t, \ldots, x_{t-D+1}] \tag{4.51}$$

Nous pouvons réécrire (4.50) sous une forme canonique

$$\begin{aligned} \tilde{\mathbf{z}}_t &= \mathcal{N}_1(\tilde{\mathbf{z}}_{t-1}, x_t) \\ \hat{x}_{t+1} &= \mathcal{N}_2(\tilde{\mathbf{z}}_t) \end{aligned} \tag{4.52}$$

Les termes du passé lointain ont été bannis pour alléger l'expression (4.50) et renouer avec une "représentation Markovienne" du système, en ce sens que les entrées de $\mathcal{N}_1$ et de $\mathcal{N}_2$ ne dépendent plus que des variables à l'instant précédent.

Cette représentation dynamique des réseaux de neurones soulève plusieurs questions essentielles :

– Comment évolue le réseau lorsque le système etudié n'est pas observable?

– Lors même que le système est observable, la technique rudimentaire de la descente du gradient est-elle apte à inférer sa dynamique sous-jacente en l'absence d'une fenêtre temporelle en entrée.

– Quelles sont les conséquences de la décroissance exponentielle du gradient sur l'habileté du modèle à capturer les structures temporelles de la série observée sur une grande echelle de temps (i.e. $\tau >> D$)?

A défaut de résultats theoriques, les expérimentations menées dans les paragraphes suivants chercheront à repondre partiellement à ces questions difficiles. Ainsi, l'emploi de la fenêtre temporelle sera sciemment abandonné dans l'expérimentation des sunspots, pour *forcer* le DRNN à faire usage de sa mémoire interne.

4.3 Conditions expérimentales

Avant de nous tourner vers les expérimentions à proprement parler, nous faisons un détour pour rappeler les techniques qui permettent d'accélérer l'apprentissage et réduire l'effet néfaste des minima locaux. Une mesure destinée à évaluer la quantité de mémoire dont le réseau fait usage implicitement est aussi présentée.

En vue d'optimiser le pas d'apprentissage η automatiquement, η est adapté au fil de l'apprentissage car la meilleure valeur en début d'apprentissage n'est peut-être pas la meilleure par la suite. Notre approche consiste à vérifier si l'adaptation de la matrice de poids décroît effectivement la fonction d'erreur. Faute de quoi, le pas du gradient ramené à la topologie de la surface d'erreur, est trop grand pour permettre une recherche fine. A l'opposé, si plusieurs déplacement réduisent l'erreur consécutivement, il s'agit peut être d'une vallée en pente auquel cas il est judicieux d'accroître la valeur du pas d'apprentissage. Il est préférable, au regard de la littérature [Hertz 91], d'acccroître la valeur de η par un coefficient constant et de le décroître géométriquement pour permettre une chute rapide lorsque cela s'avère nécessaire. Cela conduit au schéma suivant,

$$\eta = \begin{cases} +\alpha & \text{si } \Delta E \leq 0 \\ -\beta\eta & \text{si } \Delta E > 0 \end{cases} \tag{4.53}$$

Dans la pratique, α sera fixé à 10^{-3} pour les premières 100 époques et ensuite réduit à 5×10^{-4}; $\beta = 0.9$ reste constant. Une moyenne mobile (momentum) de coefficient 0.9 dans l'adaptation des poids est employée pour améliorer la robustesse de la technique du gradient et augmenter la vitesse de convergence.

En outre, pour prévenir le risque potentiel d'*overfitting*, il faut s'assurer que l'apprentissage cesse dès que la composante aléatoire de la suite dans la base d'apprentissage a été capturée par le réseau. Une méthode d'élimination des poids consiste à élaguer les connexions jugées inutiles au cours de l'apprentissage [Hertz 91, Weigend 90]. Pour cela, chaque poids est systématiquement diminué par l'entremise d'un coefficient multiplicatif, $\epsilon < 1$,

$$w_{ij}^{new} = (1 - \epsilon) w_{ij}^{old} \tag{4.54}$$

si bien que les connexions pénalisées s'effacent devant les autres lorsqu'elles ne sont plus renforcées. Pour que les w_{ij} grands ne décroissent pas plus rapidement que les petits, ϵ est rendu dépendant de w_{ij} selon

$$\epsilon_{ij} = \frac{\gamma\eta}{(1 + w_{ij}^2)^2} \tag{4.55}$$

où γ est varié au fil de l'apprentissage. Habituellement, nous commençons avec $\gamma = 10^{-3}$ pour garantir l'élimination des connexions superflues et, lorsque l'erreur stagne, sa valeur est réduite à 10^{-4} pour permettre une décroissance à plus long terme. Ces valeurs ont été choisies heuristiquement et appliquées à tous les réseaux de neurones pour garantir un bon fonctionnement de la procédure d'apprentissage. Outre la procédure d'élimination des poids superflus, la progression de l'apprentissage est mesurée à la décroissance de l'erreur sur la base de prédiction, et non sur la base d'apprentissage.

Par ailleurs, un coefficient quantifie la mémoire employée par le réseau. Ce *past information ratio* (PIR) est défini de la manière suivante

$$\text{PIR} = \frac{\sum_{d>0} \sum_{i,j} w_{ij}^d}{\sum_{d} \sum_{i,j} w_{ij}^d} \tag{4.56}$$

Il s'agit du rapport de la somme des poids des connexions retardées par la somme totale des poids. Ce coefficient indicatif mesure, en essence, la quantité d'information issue du passé dont le réseau fait usage.

Enfin, une mesure standard est donné par l'erreur quadratique normalisée (NMSE), ou *normalized mean squared error* :

$$\text{NMSE} = \frac{1}{\sigma^2 T} \sum_{k=1}^{T} (x(k) - \hat{x}(k))^2 \tag{4.57}$$

$x(k)$ désigne la valeur observée de la suite, et $\hat{x}(k)$, la prédiction fournie en sortie du réseau de neurones; σ^2 est la variance de la suite et T le nombre d'exemples dans la base de test. Remarquons que NMSE = 1 correspond à la prédiction de la moyenne non conditionnée.

4.4 La prédiction des sunspots

La suite des sunspots fut la première suite étudiée avec les processus auto-régressifs par Yule en 1927 [Box 76, Yule 27]. Elle figure désormais parmi les suites incontournables avec lesquelless il faut expérimenter. Les divers approches prédictives statistiques menées sur cette suite fournissent des indexes de référence utiles [Priestley 81, Subba Rao 84, Tong 80, Tong 90]. De surcroît, la suite a fait l'objet d'un étude exhaustive à l'aide d'un perceptron à trois couches par Weigend et al. [Weigend 90]. Ces références littéraires et l'enjeu que la prédiction des sunspots représente justifie cette étude exhaustive.

Les sunspots sont des taches sombres maculant le soleil qui témoignent de l'activité de son champ magnétique, lequel influence par des voies intangibles, les conditions météorologiques sur terre. Le cycle solaire des sunspots varie de 7 à 15 années. Les connaissances pratiques et théoriques du problème acquisent par les astro-physiciens ne suffisent pas à établir des équations physiques capables de modéliser correctement ses fluctuations. Par ailleurs, les tentatives menées pour caractériser les traits essentiels de la suite ont échoué en raison du nombre insuffisant de données. Par conséquent, nous cherchons à établir une prédiction sans référence à un physique modèle sous-jacent. Une méthodologie pour identifier la dimension du système d'équations aux différences sous-jacent, dont nous supposons l'existence, est aussi exposée.

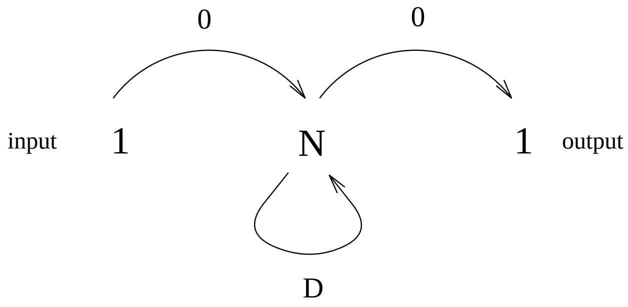

FIG. 4.1 - : Représentation schématique de la configuration du DRNN utilisé pour la prédiction des sunspots. Les flèches représentent une connectivité totale entre les N neurones cachés et D désigne l'ordre des filtres.

4.4.1 Mode opératoire

Généralement, la sélection d'un réseau adéquat se fonde sur des essais préliminaires effectués avec des réseaux candidats, et sur des heuristiques propres à chaque expérimentateur. Ici, nous optons pour une méthodologie rigoureuse; une architecture générique $1 \times N \times 1$, décrite Fig. 4.1, est employée. Les N neurones cachés sont totalement bouclés à l'aide de connexions auto-régressives d'ordre D. C'est à dessein que nous nous limitons à de petits modèles contraints d'exploiter leur mémoire interne pour pallier le manque d'information présentée en entrée. L'accent est mis sur le codage et la capacité de mémorisation plutôt que sur la performance pure.

Le but que nous poursuivons est d'exhiber une solution acceptable qui se traduise en équations formelles. Plutôt que de lancer $N \times (D+1)$ réseaux de neurones parallèlement, il est plus judicieux de commencer avec des petits réseaux et d'accroître progressivement leurs dimensions, c.a.d.

- le nombre de neurones cachés, N.
- l'ordre des filtres, D.

Le nombre de neurones détermine les caractéristiques que peut représenter le réseau, tandis que l'ordre du filtre spécifie la mémoire inhérente au type d'interaction entre les neurones. Dans cette perspective, il est sage de commencer à adapter le réseau $N = 1, D = 0$ pour obtenir une première solution au problème, même très approximative. Vu le nombre de connexions, l'apprentissage est pour ainsi dire immédiat. Ensuite, N est incrémenté systématiquement en conservant les poids précedents jusqu'à $N = 5$ et les 5 réseaux obtenus sont adaptés en parallèle en incrémentant D à chaque étape. En toute logique, la décroissance de l'erreur est monotone. Le plus gros réseau obtenu est $N = 5, D = 5$.

L'effort de calcul est réduit et l'influence des minima locaux dans la quête du minimum globale est limitée. Force est de constater que le DRNN est aussi contraint : 1) de trouver une représentation interne adéquate pour coder la séquence, et 2) de conserver la trace des vecteurs d'état passés en mémoire pour pallier le manque d'information disponible en entrée du réseau.

Concernant les réseaux non-bouclés à délais ou non, il y aura toujours des valeurs que le réseau ne pourra anticiper correctement dès lors que l'information contextuelle relative à cette valeur couvre une période de temps qui outre-passe les limites de la fenêtre temporelle imposée en entrée du réseau. Le DRNN est amené à développer des mécanismes internes pour préserver une image du passé. La récurrence spatiale et temporelle dans le model DRNN favorise la circulation en boucle des signaux à ce dessein. L'information restaurée permet ainsi de fournir en sortie une valeur sensible au passé de la suite.

Finalement, l'erreur normalisée est tracée en 3 dimensions en fonction de N et de D; une chute abrupte est d'abord observée suivie d'une vallée profonde mais de faible pente. Le nombre de neurones cachés au point d'inflexion, lorsque la vallée est atteinte, fournit une valeur indicative du nombre de degrés de liberté nécessaires au réseau pour modéliser la trajectoire de la suite. Cette idée remonte aux premiers jours de la rétro-propagation; l'erreur obtenue par un réseau non-bouclé à 3 couches était tracée en fonction du nombre de neurones au sein de la couche cachée. La taille du réseau minimal donne une indication de la dimension du système d'équations sous-jacent à la suite. Néanmoins, il faut garder à l'esprit que cette mesure est influencée par le choix de la fonction d'activation.

4.4.2 La prédiction

Conformément aux méthodes employées [Priestley 81, Tong 80, Weigend 90], les moyennes annuelles entre 1720 et 1979 sont normalisées. La performance prédictive des réseaux employés est déterminée d'après les résultats obtenus sur une base de test de 59 valeurs correspondant à la période de 1921 à 1979. Les années 1720-1920 constituent la base d'apprentissage et la méthodologie discutée au paragraphe précédent est adoptée. Le réseau $N = 1, D = 0$ est accru jusqu'à $N = 5, D = 0$, ensuite les 5 réseaux sont entraînés indépendamment et l'ordre des filtres est augmenté jusqu'à $D = 5$. Après chaque convergence, les modifications apportées à la structure du réseau augmentent le nombre de paramètres à adapter. Ces paramètres sont initialisés à zéro pour ne pas nuire à la stabilité de la configuration précédente. Une centaine de présentations de la base d'apprentissage suffit d'ordinaire à trouver un autre minimum plus attractif. Il faut compter quelques heures de temps CPU sur une station de travail du type Sparc 10. Remarquons enfin que les solutions obtenues ne vérifient pas toutes les conditions (3.54) mais un régime stable est observé dans tous les cas.

La prédiction à un pas en avant

La mesure de performance s'exprime généralement dans la littérature en terme d'erreur dans l'estimation à un pas en avant de la suite [Priestley 81, Tong 80, Tong 90, Weigend 90]. Ce critère de performance est retenu. Tables 4.1 à 4.5 affichent des informations intéressantes concernant les solutions obtenues pour $1 \geq N \geq 5$ et $0 \geq D \geq 5$. La NMSE calculée sur la base de prédiction est représentée sur un diagramme en 3 dimensions (Fig. 4.2) en fonction de D et N. Bien que la performance généralement augmente avec D, certains réseaux $N = n, D = d$ affichent des performances moindres en comparaison avec le réseau précédent $N = n, D = d-1$. Dans ce cas, la solution à laquelle est arrivé le réseau est conservée pour éviter toute redondance dans les chiffres affichés. Curieusement, la forme de la surface d'erreur est symétrique par rapport à la bissectrice $N = D$. On observe une vallée de faible inclinaison pour $D \geq 2$ et $N \geq 2$ au delà d'une chute abrupte. Le tassement de l'erreur confirme la stagnation de la performance lorsque des paramètres sont ajoutés.

A des fins illustratives, le nombre de degrés de liberté effectifs ainsi que le PIR sont affichés. Le nombre de degrés de liberté effectifs est le nombre de poids qui n'ont pas été éliminés par la procédure d'élagage. Ce nombre augmente de façon monotone avec D et N en moyenne. La meilleure solution obtenue a un PIR de 0.25, i.e. 25% environ des signaux sont retardés. Cela reflète étroitement la quantité de mémoire interne dont fait usage le réseau. De plus, une différence substantielle est observée entre $N = 1$ et $N > 1$, confirmant les résultats médiocres obtenus pour $N = 1$. A une grande valeur du PIR correspond régulièrement de bonnes performances prédictives.

Pour que des comparaisons rigoureuses puissent être faites avec les résultats présentés en [Weigend 90, Tong 90], la base de test a été scindée en deux périodes distinctes correspondant à 1921-1955 et 1956-1979. La raison de cette coupure tient à la variance inégale observée durant ces deux périodes. La variance très forte, mesurée sur l'intervalle 1956-1979, se traduit par une augmentation soudaine des résidus d'erreur durant cette période. Plusieurs architectures affichent des performances très similaires. Le réseau le plus petit en terme de paramètres propres peut être facilement explicité par

$$x_{t+1} = g(6.129 z_t^1 - 3.641 z_t^2 - 3.627) \tag{4.58}$$

où z_t^1 et z_t^2 sont les activations des deux neurones cachés. Le modèle est régi par le système d'équations

$$\begin{aligned} z_t^1 &= g(2.440 z_t^1 - 1.152 z_{t-1}^1 + 0.350 z_{t-2}^1 - 0.115 z_{t-3}^1 + 0.041 z_{t-4}^1 - 0.067 z_{t-5}^1 \\ &\quad -0.066 z_t^2 - 0.355 z_{t-1}^2 + 0.097 z_{t-2}^2 - 0.678 z_{t-3}^2 + 0.094 z_{t-4}^2 - 0.118 z_{t-5}^2 \\ &\quad +3.816 x_t - 0.072) \\ z_t^2 &= g(+1.286 z_t^2 + 0.813 z_{t-1}^2 - 0.171 z_{t-2}^2 - 0.120 z_{t-3}^2 - 0.491 z_{t-4}^2 - 0.180 z_{t-5}^2 \\ &\quad -2.296 z_t^1 - 2.948 z_{t-1}^1 + 0.583 z_{t-2}^1 - 0.342 z_{t-3}^1 - 0.112 z_{t-5}^1 \\ &\quad -0.940 x_t - 0.405) \end{aligned} \tag{4.59}$$

Le réseau se caractérise par seulement 2 neurones cachés entièrement bouclés par des filtres auto-régressifs d'ordre 5. Le nombre total de poids se limite à 30. Des résultats sensiblement meilleurs sont obtenus avec le réseau $N = 5, D = 2$. Une époque de démonstration sur les bases d'apprentissage et de test est présentée pour le modèle $N = 2, D = 5$ de Fig. 4.3 à 4.6 et pour le modèle $N = 5, D = 2$ de Fig. 4.7 à 4.10.

Table 4.6 présente des résultats comparatifs entre une *copie carbone* (i.e. $\hat{x}_t = x_{t-1}$), le modèle TAR [Tong 80, Tong 90], une réseau multi-couches non-bouclé [Weigend 90] et les modèles $N = 5, D = 2$ et $N = 2, D = 5$. Les deux meilleures entrées sont obtenues avec le DRNN. Les prédictions sont comparables sur la période 1921-1955, mais légèrement meilleures sur la période 1956-1979. Il est intéressant de noter que le perceptron à plusieurs couches nécessite environ 30% de paramètres supplémentaires et que le DRNN n'est plus conduit par une fenêtre temporelle en entrée.

La prédiction à plusieurs pas en avant

Nous portons notre attention sur la prédiction à plusieurs pas en avant au départ de 1921. Pour ce faire, le réseau est itéré en bouclage fermé; en d'autres termes, la sortie obtenue par le réseau est systématiquement rétro-propagée en entrée à l'itération suivante. L'erreur normalisée est exhibée Fig. 4.11 pour tous les modèles. Notons que cette erreur n'est pas moyennée comme dans [Weigend 90] pour plusieurs années de départ. L'apprentissage a été mené jusqu'en 1920 de sorte que la prédiction itérative commence en 1921. Le réseau génére un comportement périodique semblable à un oscillateur sinusoïdal, souvent en léger déphasage avec la continuation de la suite. L'augmentation brusque de 1956 témoigne des irrégularités de la suite au-delà de 1956. Paradoxalement, on remarque que les résultats de la prédiction itérative ne s'améliorent pas nécessairement avec des réseaux plus complexes. En effet, l'erreur normalisée affichée Fig. 4.11c est généralement supérieure à celle Fig. 4.11b; les réseaux $N = 2, D = 0$ et $N = 3, D = 0$ ont curieusement obtenu de meilleurs résultats que les autres dans la même catégorie. Les réseaux à un neurone caché $N = 1$ se distinguent par de piètres performances.

En dépit des résultats encourageants obtenus pour la prédiction à court terme, la prédiction itérée n'a pas donné entière satisfaction. Les raisons sont les suivants : contrairement au réseau non-bouclé, la structure générique pour laquelle nous avons opté pour DRNN ne fait plus usage d'une fenêtre temporelle en entrée mais d'une valeur unique. Il résulte que les erreurs commises en sortie sont instantanément reproduites en entrée à l'itération suivante provoquant le cumul géométrique des erreurs. En contraste, la fenêtre temporelle est plus robuste aux erreurs car l'information en entrée est redondante. Une fenêtre temporelle de taille D fournie au réseau des valeurs exactes pour les D itérations suivantes alors que les erreurs ont été amplifiées D fois avec le DRNN. Nous aggravons en quelque sorte la faiblesse notoire [Gershenfeld 93] des réseaux de neurones à prédire la continuation de la suite à long terme. En effet, les réseaux connexionistes procèdent à une approximation *globale* de la surface dans l'espace des phases à l'aide d'une séquence limitée de points. Inversement, les techniques de reconstruction locale de l'espace des phases approximent la géométrie du sous espace attracteur et construisent ainsi une représentation explicite de la dynamique [Gershenfeld 93, Ott 93, Sauer 93].

Comportement du gradient d'erreur rétro-propagé

Fig. 4.12 illustre le comportement des termes $\|\mathbf{y}^n_{n-j}\|_\infty$ moyennés sur la base d'apprentissage pour $n = 1, \ldots, T$. Une représentation logarithmique sied mieux à la décroissance observée en fonction du temps rétrograde j dans le réseau adjoint déplié. Typiquement, la norme du gradient décroît linéairement dès la première passe en oscillant légèrement autour d'une droite jusqu'à $-\infty$. Le critère d'arrêt défini en (3.61) est approprié et l'erreur d'estimation du gradient peut se lire sur la courbe de décroissance. Les simulations corroborent l'étude théorique menée au chapitre précédent. Le nombre de propagations rétrogrades dans l'adjoint varient de 5 à 45, 15 en moyenne avec $\epsilon = 10^{-3}$ dans (3.61), confirmant en cela l'allégation antérieure que le nombre de passes s, défini dans (3.61), est de l'ordre de D

L'histogramme du nombre moyen de rétro-propagations dans le réseau adjoint associé au réseau $N = 2, D = 5$ est tracé Fig. 4.13. La courbe fluctue de 15 à 45, centrée autour de 30, c'est-à-dire, 6 fois l'ordre des filtres $D = 5$. 45 vecteurs d'état passés du réseau, $9 \times D$, suffisent à estimer correctement le gradient d'erreur. On retrouve se résultat en inspectant la courbe Fig. 4.12b. Soulignons toutefois que le réseau $N = 2, D = 5$ présente le plus faible taux de décroissance du gradient parmi les 30 avec lesquels nous avons expérimenté mais néanmoins présente un excellent rapport performance/degrés de liberté. En contraste, le nombre moyen de rétro-propagations dans le réseau adjoint associé au réseau $N = 5, D = 2$ est centré autour de 15, soit seulement 3 fois l'ordre des connexions (voir Fig. 4.12e), mais requiert 45 connexions. De même, l'histogramme de la variation moyenne du nombre de rétro-propagations dans le réseau adjoint est tracé Fig. 4.14, indiquant que $s_{k+1} - s_k$ tombe le plus souvent dans l'intervalle $[-D, +D]$. Cela illustre la difficulté à anticiper la mémoire sollicitée par la procédure d'apprentissage.

4.4.3 Caractériser la dynamique

Concernant les suites temporelles émanant de l'observation d'un phénomène physique, il est fréquent que le nombre de données enregistrées excluent la caractérisation du système d'équations sous-jacent. Il est donc instructif de voir si les réseaux de neurones peuvent aider à caractériser certaines propriétés inhérentes à la suite. Dans cette optique, Weigend et al. ont varié la taille de la fenêtre temporelle appliquée en entrée du réseau afin d'estimer la *dimension de la récursion*, en supposant le processus chaotique. Ils concluent à une dimension de 12.

Les résultats expérimentaux obtenus avec les réseaux dynamiques et complètement bouclés permettent de poursuivre un peu plus loin l'identification du système implicite. D'après Fig. 4.2, il est clair que le seuil de la vallée se situe autour du point $N = 2, D = 2$. Cette solution correspondante portent sûrement les traits essentiels qui caractérisent le système déterministe optimal qui gouverne l'évolution des sunspots. Le nombre de neurones cachés s'interprète comme le nombre de variable du système dynamique tandis que l'ordre des filtres caractérise l'ordre de équations aux différences si bien que nous conjecturons l'existence d'un système d'équations aux différences du second ordre approximant fidèlement le comportement des sunspots.

Ordre du filtre	0	1	2	3	4	5
Nombre initial de paramètres	5	6	7	8	9	10
Nombre effectif de paramètres	4	5	7	8	9	10
PIR	-	10 %	13 %	14 %	14 %	15 %
NMSE base d'apprentissage	0.316	0.209	0.145	0.143	0.143	0.131
NMSE base de test 1921-1979	0.533	0.406	0.289	0.290	0.287	0.286
NMSE base de test 1921-1955	0.462	0.303	0.172	0.169	0.164	0.160
NMSE base de test 1956-1979	1.311	0.999	0.711	0.715	0.706	0.704

TAB. 4.1 - : Résultats de la prédiction à un pas en avant des sunspots en fonction de l'ordre des connexions. Une modèle DRNN de configuration $1 \times 1 \times 1$ a été employé.

Ordre du filtre	0	1	2	3	4	5
Nombre initial de paramètres	11	15	19	23	27	31
Nombre effectif de paramètres	11	15	18	23	26	30
PIR	-	21 %	20 %	24 %	25 %	26 %
NMSE base d'apprentissage	0.309	0.120	0.106	0.113	0.095	0.105
NMSE base de test 1921-1979	0.502	0.230	0.154	0.134	0.128	0.111
NMSE base de test 1921-1955	0.445	0.123	0.115	0.100	0.098	0.091
NMSE base de test 1956-1979	1.236	0.542	0.380	0.330	0.316	0.273

TAB. 4.2 - : Résultats pour la prédiction à un pas en avant en fonction de l'ordre des connexions. Une modèle DRNN de configuration $1 \times 2 \times 1$ a été employé.

Ordre du filtre	0	1	2	3	4	5
Nombre initial de paramètres	19	28	37	46	55	64
Nombre effectif de paramètres	19	20	21	41	47	51
PIR	-	24 %	22 %	26 %	28 %	27 %
NMSE base d'apprentissage	0.306	0.121	0.107	0.107	0.097	0.107
NMSE base de test 1921-1979	0.486	0.216	0.154	0.128	0.131	0.141
NMSE base de test 1921-1955	0.442	0.126	0.116	0.101	0.102	0.103
NMSE base de test 1956-1979	1.196	0.532	0.379	0.316	0.323	0.348

TAB. 4.3 - : Résultats pour la prédiction à un pas en avant en fonction de l'ordre des connexions. Une modèle DRNN de configuration $1 \times 3 \times 1$ a été employé.

Ordre du filtre	0	1	2	3	4	5
Nombre initial de paramètres	29	45	61	77	93	109
Nombre effectif de paramètres	23	31	41	41	39	34
PIR	-	22 %	25 %	25 %	25 %	26 %
NMSE base d'apprentissage	0.306	0.108	0.110	0.107	0.107	0.108
NMSE base de test 1921-1979	0.465	0.136	0.105	0.110	0.113	0.112
NMSE base de test 1921-1955	0.438	0.105	0.091	0.095	0.096	0.095
NMSE base de test 1956-1979	1.145	0.336	0.259	0.271	0.278	0.277

TAB. 4.4 - : Résultats pour la prédiction à un pas en avant en fonction de l'ordre des connexions. Une modèle DRNN de configuration $1 \times 4 \times 1$ a été employé.

Ordre du filtre	0	1	2	3	4	5
Nombre initial de paramètres	41	66	91	116	141	166
Nombre effectif de paramètres	37	53	45	47	38	51
PIR	-	25 %	26 %	27 %	27 %	28 %
NMSE base d'apprentissage	0.305	0.111	0.111	0.106	0.107	0.104
NMSE base de test 1921-1979	0.480	0.130	0.100	0.110	0.111	0.103
NMSE base de test 1921-1955	0.441	0.100	0.093	0.098	0.099	0.094
NMSE base de test 1956-1979	1.182	0.320	0.246	0.271	0.274	0.255

TAB. 4.5 - : Résultats pour la prédiction à un pas en avant en fonction de l'ordre des connexions. Une modèle DRNN de configuration $1 \times 5 \times 1$ a été employé.

Modèle	apprentissage	1921-1955	1956-1979	Nb. param.
$\hat{x}_t = x_{t-1}$	0.289	0.427	0.966	-
TAR	0.097	0.097	0.280	18
MLP $12 \times 3 \times 1$	0.082	0.086	0.350	43
DRNN $N = 2, D = 5$	0.105	0.091	0.273	30
DRNN $N = 5, D = 2$	0.111	0.093	0.246	45

TAB. 4.6 - : Evaluation des modèles prédictifs sur la prédiction des sunspots à un pas en avant en fonction des périodes et des modèles.

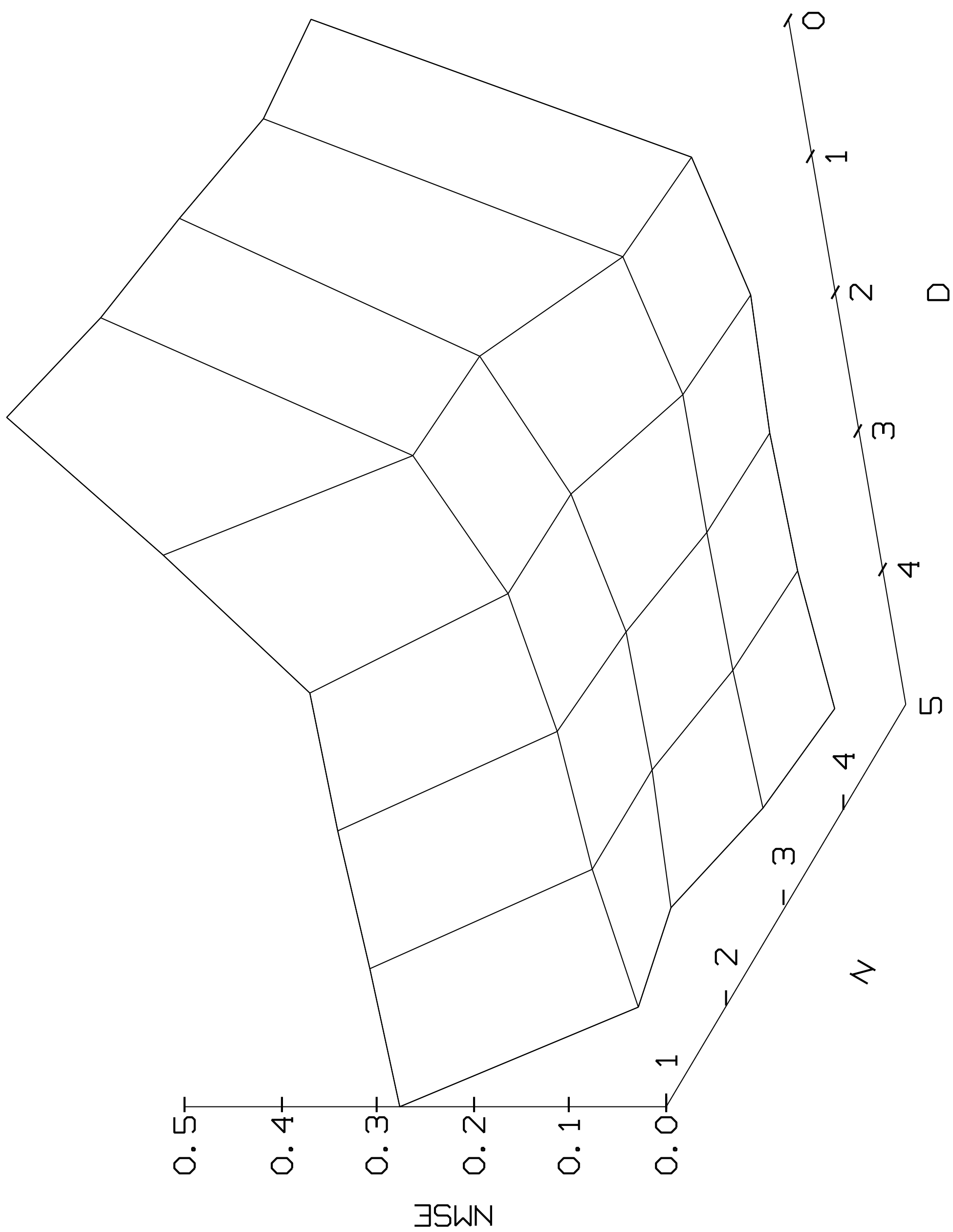

FIG. 4.2 - : NMSE en 3 dimensions pour les réseaux $1 \times N \times 1$ dédiés à la prédiction des sunspots sur la base de test (1921-1979) en fonction du nombre de neurones cachés N et de l'ordre des filtres auto-régressifs D.

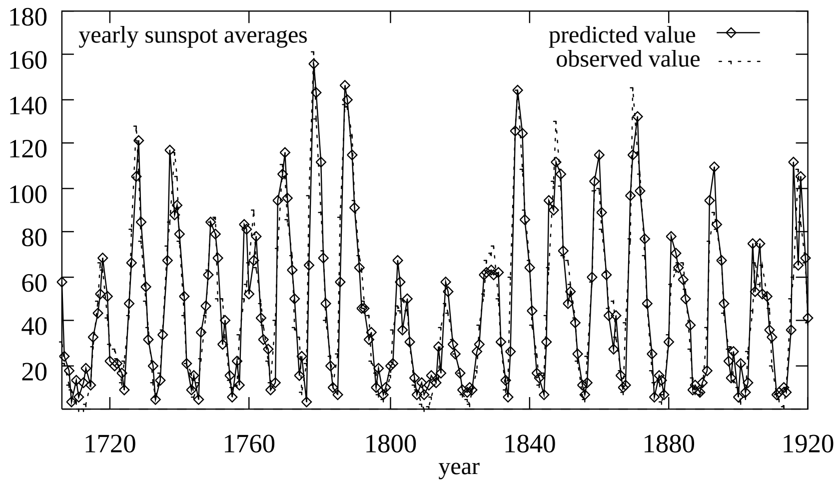

FIG. 4.3 - : Prédiction des sunspots à un pas en avant avec le modèle $N = 2, D = 5$ sur la base d'apprentissage.

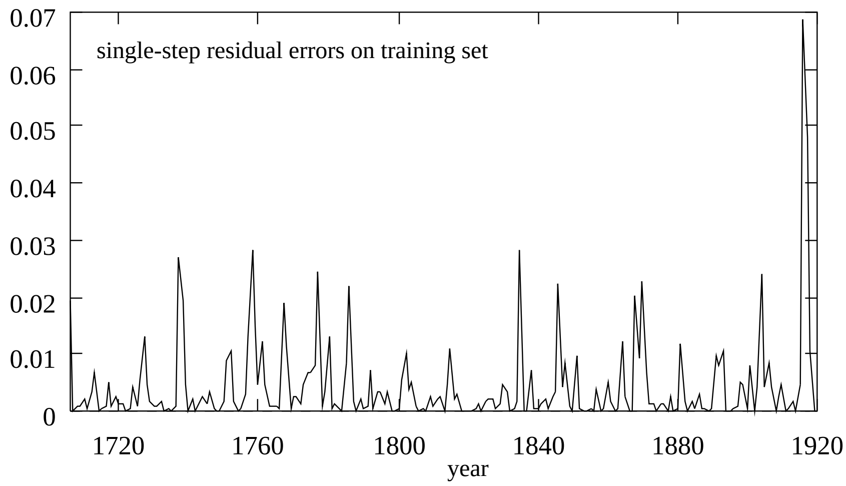

FIG. 4.4 - : Erreurs résiduelles $(\hat{x}_t - x_t)^2/400$ sur la base d'apprentissage pour le modèle $N = 2, D = 5$.

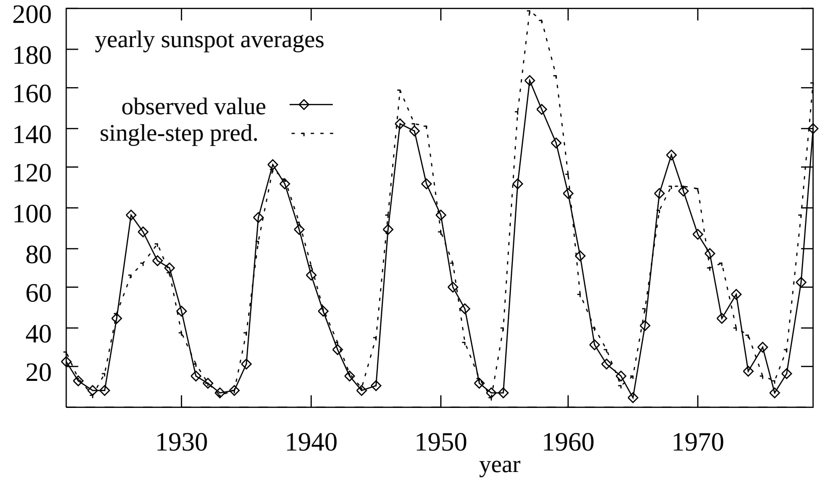

FIG. 4.5 - : Prédiction des sunspots à un pas en avant avec le modèle $N = 2, D = 5$ sur la base de test.

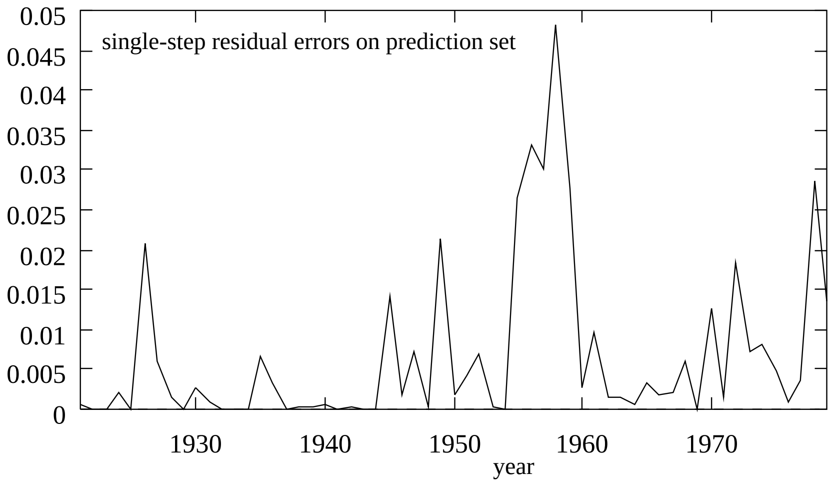

FIG. 4.6 - : Erreurs résiduelles $(\hat{x}_t - x_t)^2/400$ sur la base d'apprentissage pour le modèle $N = 2, D = 5$.

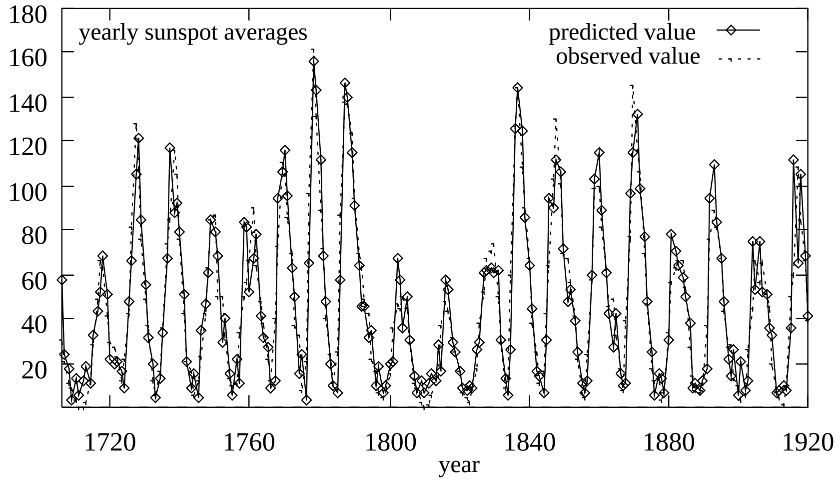

FIG. 4.7 - : Prédiction des sunspots à un pas en avant avec le modèle $N = 5, D = 2$ sur la base d'apprentissage.

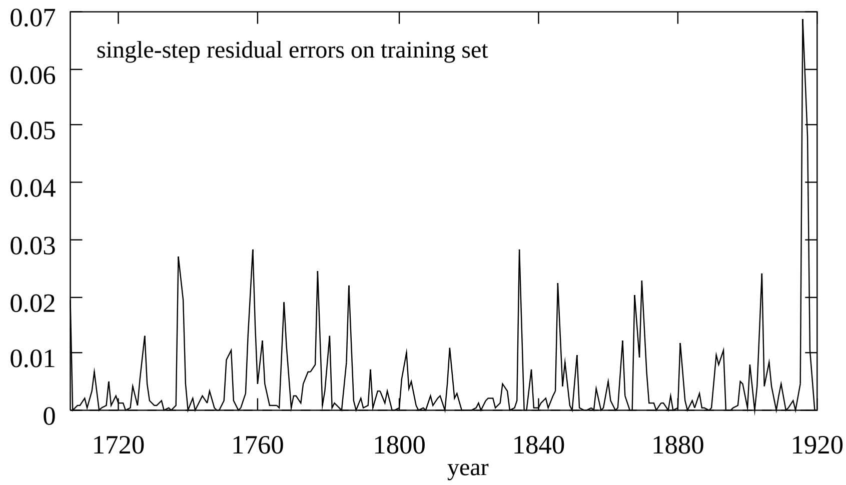

FIG. 4.8 - : Erreurs résiduelles $(\hat{x}_t - x_t)^2/400$ sur la base d'apprentissage pour le modèle $N = 5, D = 2$.

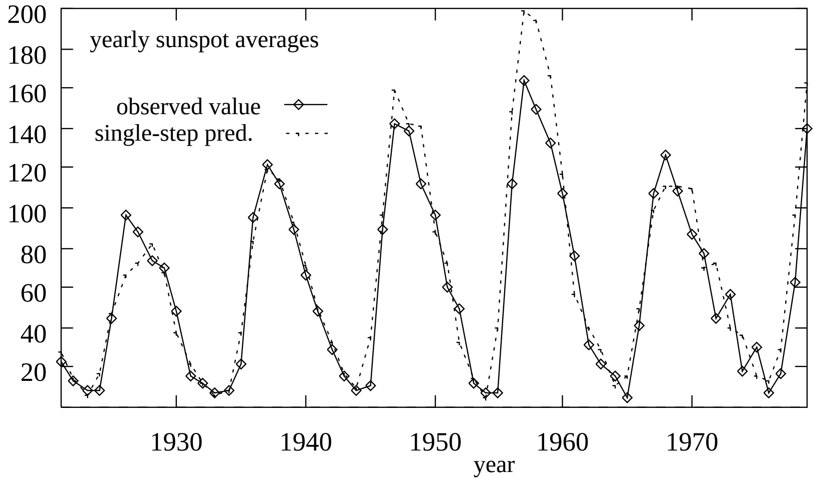

FIG. 4.9 - : Prédiction des sunspots à un pas en avant avec le modèle $N = 5, D = 2$ sur la base de test.

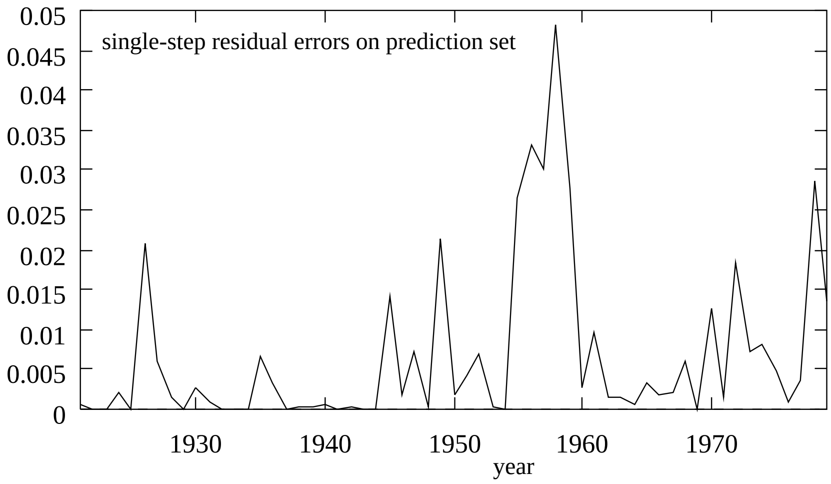

FIG. 4.10 - : Erreurs résiduelles $(\hat{x}_t - x_t)^2/400$ sur la base d'apprentissage pour le modèle $N = 5, D = 2$.

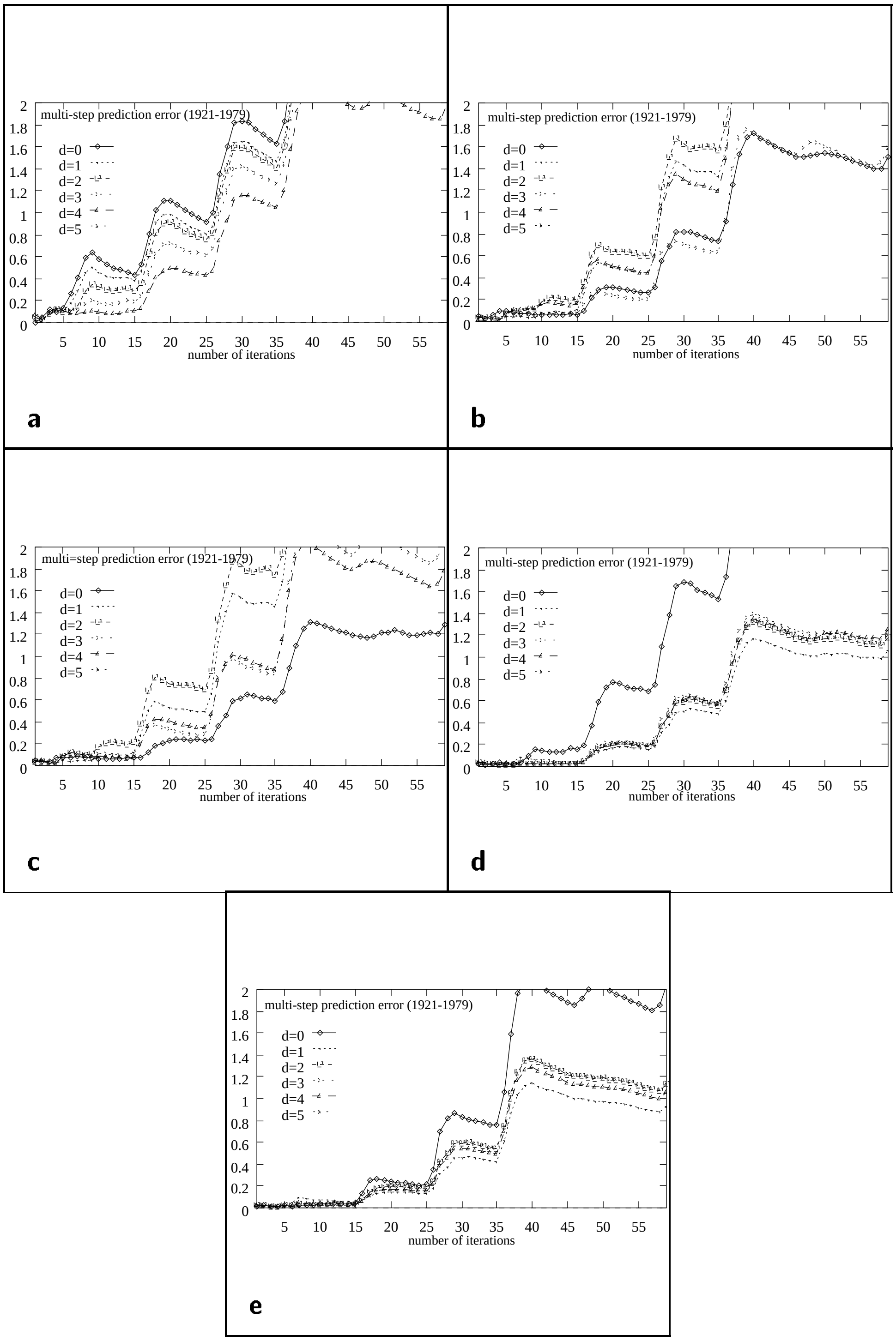

FIG. 4.11 - : NMSE de la prédiction itérée des sunspots en fonction du nombre d'années dans le futur pour les modèles DRNN de la forme $1 \times N \times 1$ avec des connexions d'ordre $D = 1, \ldots, 5$. (a) $N = 1$, (b) $N = 2$, (c) $N = 3$, (d) $N = 4$, (e) $N = 5$.

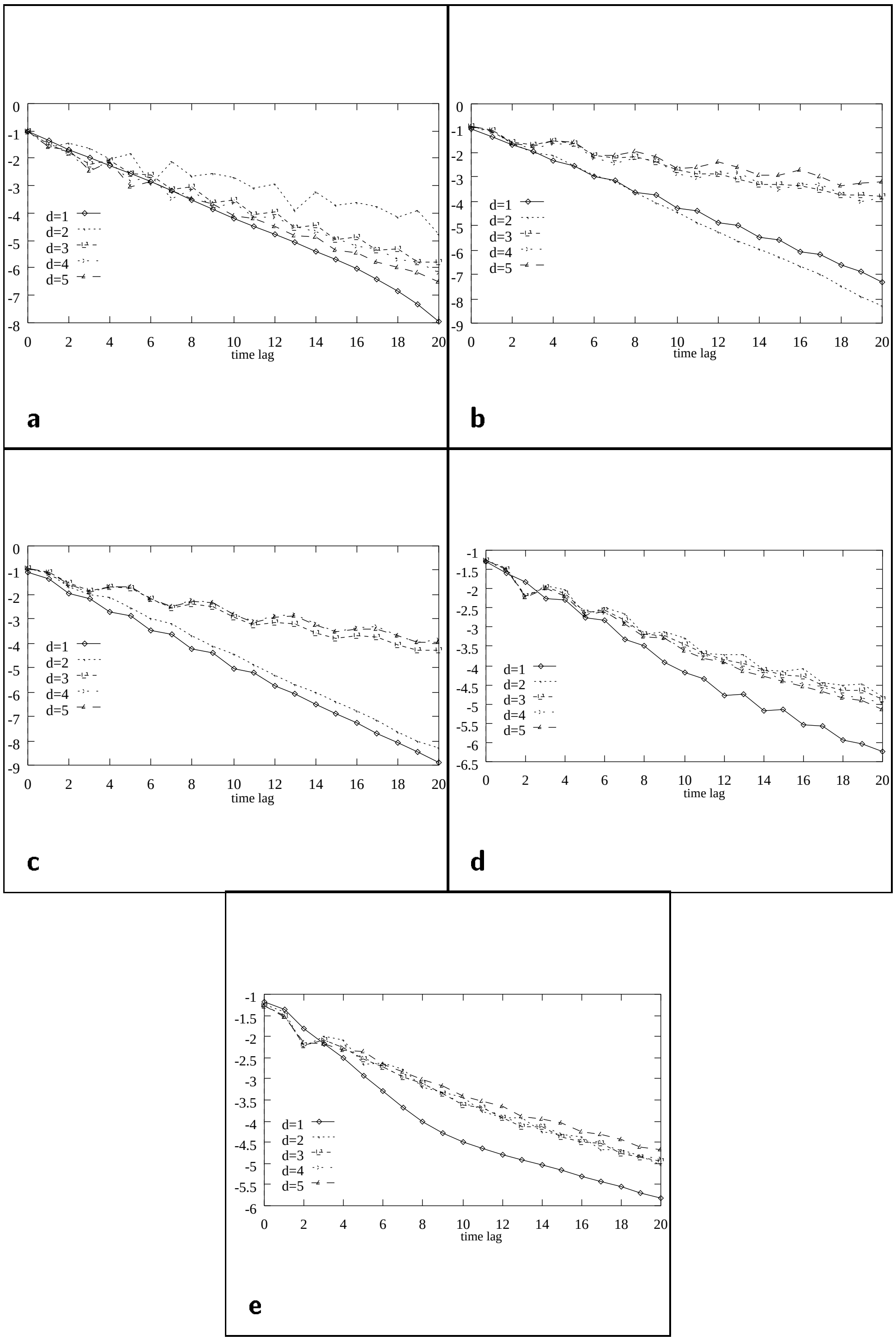

FIG. 4.12 - : $< \mathrm{Log}_{10}(\|\mathrm{y}^n_{n-j}\|_\infty >$, moyenné sur n en fonction du nombre de rétro-propagations j dans le réseau adjoint pour plusieurs modèles DRNN de la forme $1 \times N \times 1$ avec des connexions d'ordre $D = 1, \ldots, 5$. (a) $N = 1$, (b) $N = 2$, (c) $N = 3$, (d) $N = 4$, (e) $N = 5$.

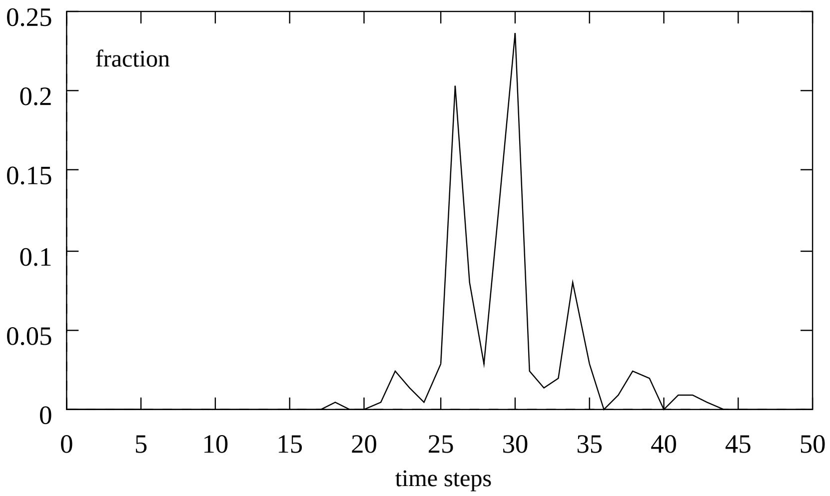

FIG. 4.13 - : Histogramme du nombre moyen de rétro-propagations nécessaires dans le réseau adjoint associé au DRNN $N = 2, D = 5$.

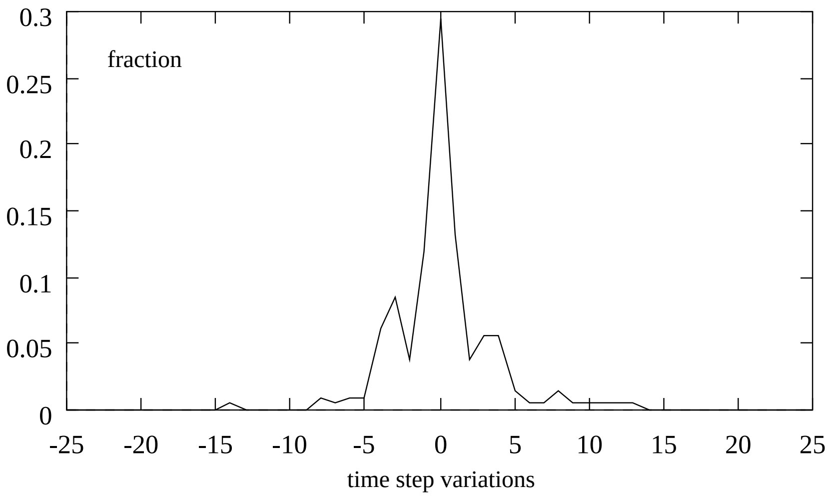

FIG. 4.14 - : Histogramme de la variation du nombre moyen de rétro-propagations nécessaires dans le réseau adjoint associé au DRNN $N = 2, D = 5$ entre deux itérations successives.

4.5 La modélisation des processus chaotiques

L'observation et l'analyse des processus chaotiques remontent au début du siècle lorsque Henri Pointcarré s'intéressa aux orbites "complexes" des astres sous l'effet mutuel de la gravitation. Toutefois, l'attrait pour le chaos est très récent et il y a à cela deux raisons majeures. D'une part les ouvrages mathématiques sont souvent abscons, voire ésotériques, de sorte que les chercheurs dans d'autres domaines n'ont pas facilement accès aux fruits de la recherche mathématique. D'autre part, les résultats concrets et applicables n'ont vu le jour que très récemment. La puissance sans cesse accrue des matériels informatiques, et désormais accessible à tous, permet désormais au "néophyte" d'apporter une contribution à l'étude des phénomènes chaotiques comme en témoigne ce paragraphe.

Dans les exemples qui suivent, des modèles DRNN aux dimensions variables sont entraînés sur des suites temporelles générées par des processus chaotiques. Les résultats des prédictions à court et moyen terme sont exhibés puis analysés. Les attracteurs chaotiques à géométrie fractale sont reconstruits en itérant les réseaux en boulcage fermé.

Mais avant de plonger dans le grand bain de la prédiction des processus chaotiques, il est bon de rappeler brièvement les traits caractéristiques d'une dynamique chaotique.

4.5.1 Le Chaos

L'approche traditionnelle pour anticiper l'évolution d'un système dynamique consiste à établir un modèle mathématique *ad hoc* et à l'intégrer dans l'avenir. Or le nombre de degrés de liberté parfois exorbitant d'un système non-linéaire (e.g. un écoulement turbulent, la météorologie, les grandeurs économiques, etc.) constitue un écueil rédhibitoire à la résolution des équations. Cependant, l'étude des systèmes dynamiques révèle que la dissipation (e.g. la viscosité) réduit le nombre de degrés de liberté effectifs à un nombre restreint. Lorsque le cas se présente, l'évolution du système après un certain temps est piégé dans un sous-espace appelé un *attracteur*. L'attracteur est souvent un objet *fractal* dont la dimension n'est pas nécessairement entière [Ott 93].

L'attribut des processus chaotiques, qui au demeurant définit le chaos, est la dépendance exponentielle aux conditions initiales. Restreignons cette discussions à un système unidimensionnel discret. L'écart infinitésimal entre deux conditions initiales x_0 et $x_0 + \epsilon$ typiquement diverge exponentiellement lorsque le système est itéré en avant. Aussi, pour de grande valeur de n, on a

$$dx_n \sim \exp(hn) dx_0 \tag{4.60}$$

où dx_n désigne l'écart après n itérations. Un indicateur commode de la sensibilité des perturbations des orbites infiniment proches est l'*exposant de Lyapunov*, h, défini par

$$h = \lim_{n \to \infty} \frac{1}{n} \log \mid \frac{dx_n}{dx_0} \mid \tag{4.61}$$

Si $x_{n+1} = M(x_n)$, h s'écrit

$$h = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \log \mid M'(x_k) \mid \tag{4.62}$$

Compte tenu de (4.62), un exposant de Lyapunov positif $h > 0$ indique le chaos.

Aussi, une perturbation de l'état du système, quelqu'en soit l'origine (e.g. erreur de quantification, de modélisation, etc.) provoque la divergence exponentielle de l'état du système par rapport à sa trajectoire sans perturbations. Ce résultat pose une limite fondamentale à la qualité des prédictions dans un futur éloigné.

En revanche, la prédiction à court terme bénéficie de la réduction dimensionnelle. Mais comment trouver la combinaison des degrés de liberté effectifs qui déterminent l'évolution du système dans l'attracteur? En fait il ne s'agit pas tant de savoir quelle variables choisir mais plutôt combien. En effet, le théorème de Takens [Takens 80] affirme qu'un ensemble adéquat d'observations passées permet la reconstruction univoque de l'état du système piégé dans l'attracteur. Nous renvoyons le lecteur au § 4.2.3 pour une formulation précise du théorème de Takens.

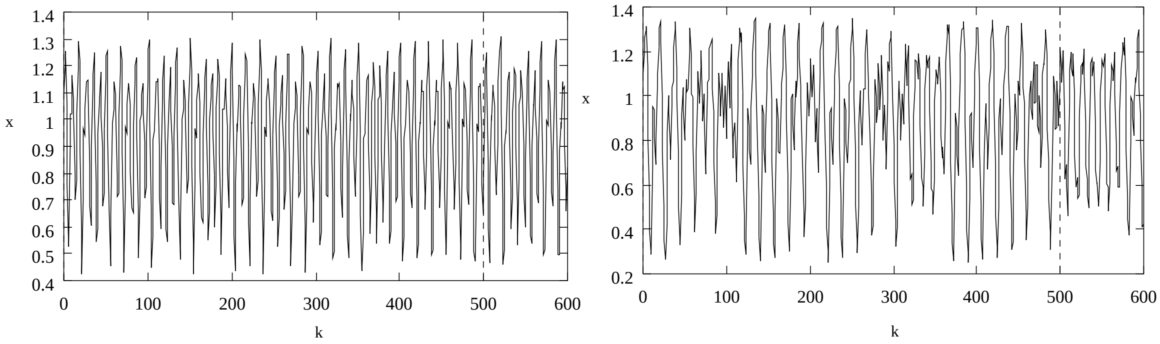

FIG. 4.15 - : Suite de Mackey Glass avec $\Delta = 17$ et $\Delta = 30$.

4.5.2 La suite de Mackey-Glass

Dans ce premier exemple, considérons l'équation différentielle à délais de Mackey-Glass [Mackey 77]

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\Delta)}{1+x(t-\Delta)^{10}} \tag{4.63}$$

Ce modèle est sensé caractériser certains systèmes physiologiques. Sa dynamique, paramétrée par Δ, revêt des formes très diverses variant du point fixe au chaos déterministe en passant par un comportement périodique. L'espace des phases est infini en raison du délai $x(t-\Delta)$. Toutefois, deux valeurs de Δ se distinguent, $\Delta = 17$ et $\Delta = 30$, conduisant à un comportement chaotique avec des attracteurs de dimension finie qualifiés d'*étranges* [Ott 93] de dimension 2.1 et 3.5 respectivement. Pour $\Delta = 17$, la densité spectrale de $x(t)$ présente de nombreux pics témoignant de la quasi-périodicité de la suite obtenue.

La technique de Runge-Kutta du 4^{ieme} ordre à été employée pour simuler le système. Deux bases d'exemples ont été générées pour $\Delta = 17$ et $\Delta = 30$ avec les conditions initiales $x(t) = 0.9$ pour $0 \leq t \leq \Delta$ puis échantillonnées à une fréquence $f = 6$ (Fig. 4.15). Des études récentes [Casdagli 89, Farmer 88, Lapedes 87, Wan 93b] nous ont guidé dans le choix de ces paramètres. Un modèle DRNN de configuration $1 \times 7 \times 1$ avec des filtres d'ordre 4 : 2 : 0 (2 représente l'ordre des filtres employés pour les connexions bouclées des 7 neurones cachés) a été entraîné sur les premiers 500 points de la suite. La sélection de ces paramètres a été faite sur la base du nombre de degrés de liberté pour faciliter la comparaison avec un réseau FIR de configuration $1 \times 15 \times 1$ avec des filtres d'ordre 8 : 2 [Wan 93b]. Le DRNN compte 197 poids ajustables et le réseau FIR 196. La simplicité a prévalu dans le choix de la topologie. En dépit du nombre équivalent de degrés de liberté, les deux réseaux se distinguent par une fenêtre temporelle de taille distincte. Le FIR dispose des 9 dernières valeurs de la suite pour fournir une estimation en sortie tandis que le DRNN se contente de 5 valeurs. L'ordre des connexions retardées en entrée du DRNN est réduit au profit du nombre de connexions récurrentes internes.

Les résultats de la prédiction à un pas en avance sont illustrés Table 4.7 pour différents modèles [Casdagli 89, Wan 93b]. Le DRNN affiche des résultats nettement supérieurs aux techniques traditionnelles et améliore sensiblement les méthodes neuronales non-bouclées.

	NMSE	
modèle	$\Delta = 17$	$\Delta = 30$
Modèle linéaire	-0.57	-0.49
Modèle polynomial	-1.95	-1.40
Modèle rationnel	-1.14	-1.33
loc(1)	-1.48	-1.24
loc(2)	-1.89	-0.42
Fonctions radiales	-1.97	-1.60
RN non-bouclé statique	-2.00	-1.50
RN FIR non-bouclé	-2.31	-1.79
DRNN	-2.33	-2.12

TAB. 4.7 - : Erreur normalisée logarithmique des prédictions de la suite de Mackey-Glass à un pas en avance selon differentes approches : linéaire, polynomiale, rationnelle, reconstruction locale, fonctions radiales et réseaux de neurones (voir [Casdagli 89]).

L'enjeu primordial réside dans la prédiction itérée. Le DRNN est comparé au réseau FIR au-delà de la 500^{ieme} itération. Comme le révèle Fig. 4.16, la prédiction itérée sur les 20 points suivant est remarquablement précise puis se dégrade au-delà.

Une caractéristique fondamentale de la théorie du chaos est l'impossibilité de délivrer des prédictions précises dans le futur éloigné, indépendamment du modèle. En fait, les trajectoires infiniment proches divergent exponentiellement en raison du bruit de quantifications. En moyenne, les prédictions itérées doivent donc diverger exponentiellement de la trajectoire observée. Pour illustrer cela, des prédictions itérées sont fournies au départ de 10 points équidistants de la base de prédiction. L'erreur normalisée *moyenne* est tracée Fig. 4.17 en échelle logarithmique : la linéarité est patente, corroborant l'analyse théorique. On remarquera une pente plus abrupte pour le DRNN par rapport au modèle FIR; les deux courbes se croisent à l'itération 15.

Fig. 4.17 illustre le tracé de la NMSE moyennée en fonction des itérations dans le futur. On remarque une pente linéaire caractéristique des systèmes chaotiques. Il est aussi à noter que passé un certain délai, la NMSE dépasse 0, indiquant que la prédiction est pire que la prédiction constante de la moyenne de la suite. Cela s'explique par le fait que après un certain nombre d'itérations, la prédiction $\hat{x}$ devient statistiquement *indépendante* de la suite observée x. L'erreur mesurée est le double de la variance de la suite et la prédiction linéaire devient “meilleure” au sens statistique, même si la modèle linéaire est incapable de reproduire la dynamique de la suite. Cela s'explique simplement par

$$\begin{aligned} \mathbf{E}[(x-\hat{x})^2] &= \mathbf{E}[(x-m_x)^2] + \mathbf{E}[(m_x-\hat{x})^2] + 2\mathbf{E}[x-m_x]\mathbf{E}[m_x-\hat{x}] \\ &= 2\sigma_x^2 \end{aligned} \quad (4.64)$$

Compte tenu de la taille différente des fenêtres temporelles et sous l'hypothèse du même nombre de connexions dans chaque réseaux, il apparaît que le DRNN délivre de meilleurs prédictions à court terme que le réseau FIR, aux dépends de la précision à long terme. Ce résultat s'interprète intuitivement : dès la 6^{ieme} itération dans le future, le DRNN opère en bouclage fermé. Ainsi toutes les 6 itérations le vecteur d'entrée est entièrement renouvelé par les anciennes sorties du réseau, alors que cette période est de 10 pour le FIR. Il n'est pas étonnant que la qualité des prédictions itérées se dégrade plus rapidement avec le DRNN puisque les erreurs se répercutent en cascade à une fréquence plus élevée. Mais il ne tient qu'à nous d'avoir choisi d'utiliser un fenêtre temporelle plus petite que celle du réseau FIR non-bouclé, alors que ce choix est subordonné à la structure temporelle de la série pour les réseaux statiques.

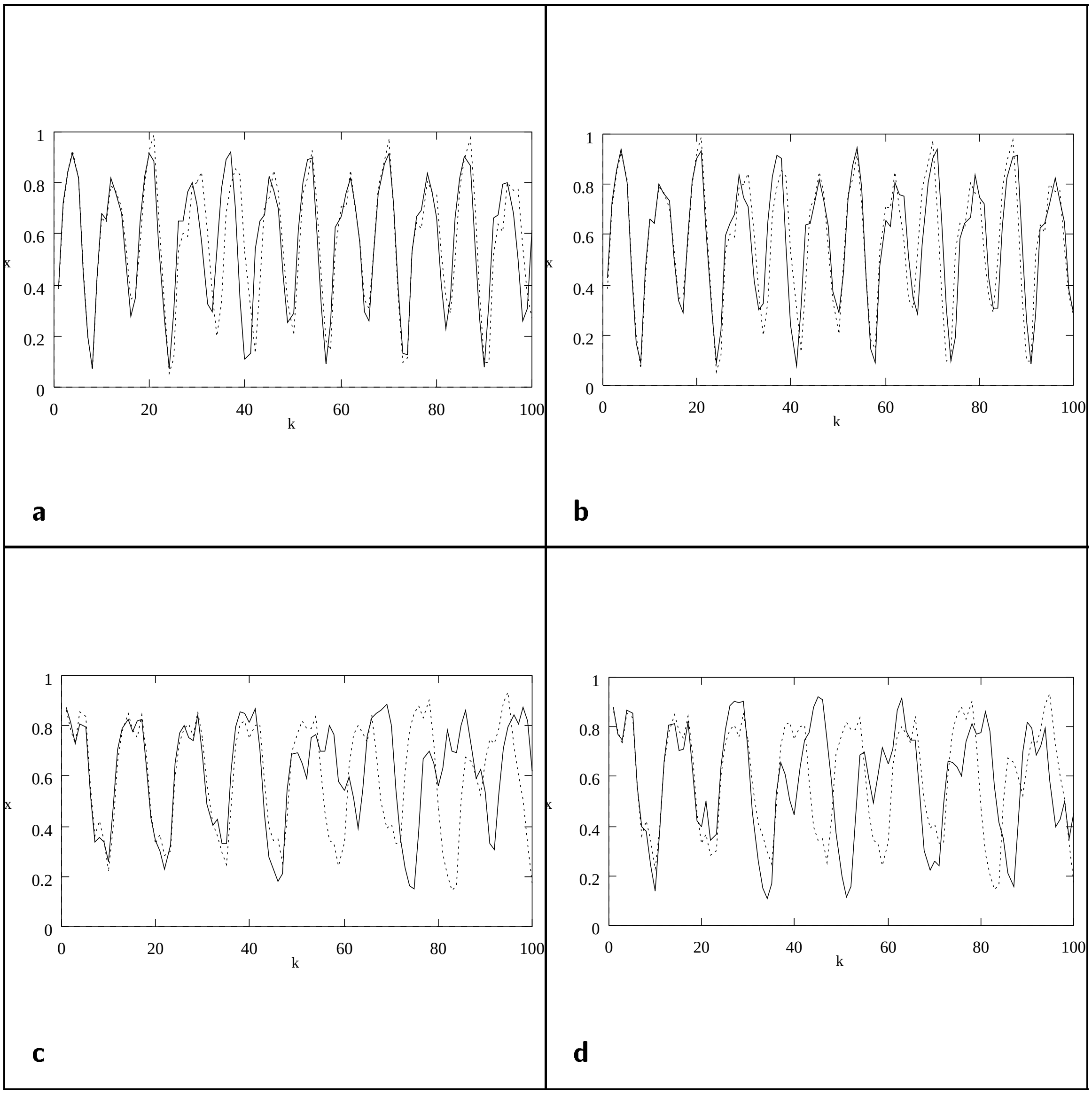

FIG. 4.16 - : Prédiction itérée de la suite de Mackey-Glass : a) réseau FIR, $\Delta = 17$, b) réseau DRNN, $\Delta = 17$, c) réseau FIR, $\Delta = 30$, d) réseau DRNN, $\Delta = 30$. En trait plein, la prédiction; en pointillé, la suite simulée.

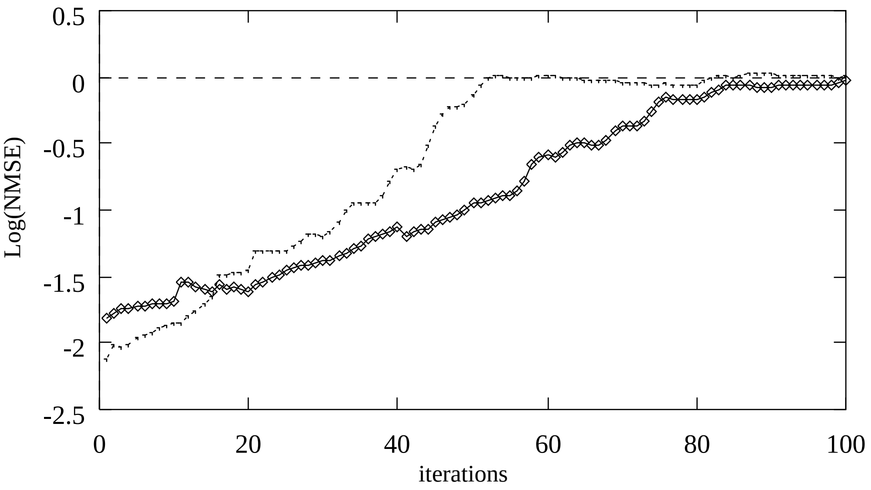

FIG. 4.17 - : Mackey-Glass : logarithme de l'erreur quadratique normalisée des prédictions itérées. La courbe en pointillé représente le DRNN, en trait plein, le réseau FIR.

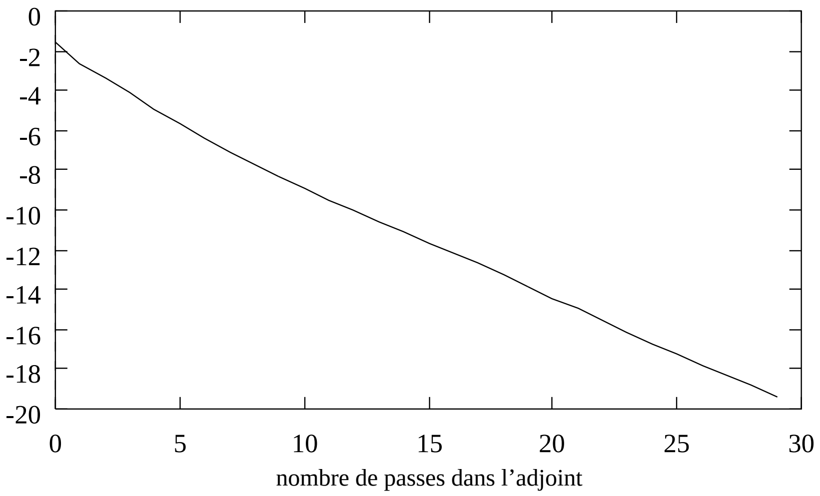

FIG. 4.18 - : Mackey-Glass : illustration du gradient rétro-propagé. $< \mathrm{Log}_{10}(\|y^n_{n-j}\|_\infty >$ est tracé en fonction du nombre de rétro-propagations j dans le réseau adjoint.

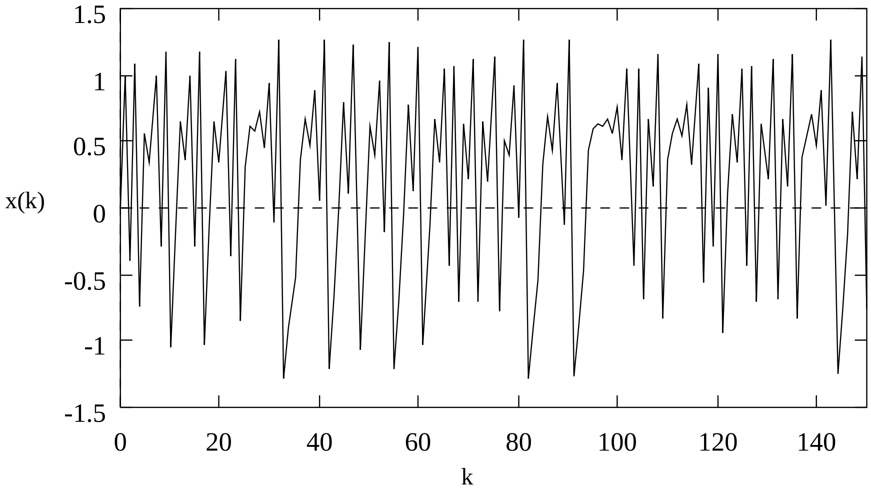

FIG. 4.19 - : Suite de Henon.

4.5.3 La suite de Henon

Dans cet exemple, considérons la suite obtenu en itérant le système bi-dimensionnel

$$\begin{aligned} x(k+1) &= 1.0 - 1.4x^2(k) + 0.3y(k) \\ y(k+1) &= x(k) \end{aligned} \tag{4.65}$$

Le comportement de la suite est illustré Fig. 4.19. L'image de l'attracteur dans l'espace de phases de x en fonction de y est représentée Fig. 4.22. Un agrandissement de l'attracteur révélerait une structure à l'échelle locale constituée d'une multitude de lignes parallèles, et ce quelque soit l'échelle. En fait, l'attracteur de la suite de Henon est constitué d'un ensemble *non-dénombrable* de lignes contiguës : il s'agit d'un attracteur *étrange*. Le calcul numérique montre que la dimension fractale de l'attrateur Fig. 4.22 est de dimension $D = 1.26$, confirmant le qualificatif d'attracteur *étrange* [Ott 93].

Un réseau FIR de dimension $1 \times 12 \times 12 \times 1$ avec des filtres d'ordre 3 : 1 : 1 est entraîné ainsi qu'un modèle DRNN de dimension plus modeste : $1 \times 10 \times 1$ avec des filtres d'ordre 3 : 1 : 0. L'apprentissage est effectué sur 5000 points de la suite et les résultats de la prédiction à un pas en avance sont présentés Table 4.8.

La prédiction itérée, représentée Fig. 4.20, est remarquablement exacte jusqu'à l'itération 10 puis diverge au-delà. La divergence est une conséquence incontournables de la nature chaotique de la suite. Mais la suite issue du réseau itéré en bouclage fermé conserve les propriétés inhérentes au système comme le montre le tracé de l'attracteur dans l'espace des phases Fig. 4.23. Le réseau de neurones a capturé au cours de l'apprentissage la dynamique sous-jacente du système original représenté par (4.65). Il est intéressant de remarquer que le bruit ne vient pas corrompre la dynamique de réseau et donc que l'attracteur est stable. Wan [Wan 93b] a montré à quel point un réseau FIR était robuste en bruitant la suite avec un bruit blanc Gaussien d'un rapport signal/bruit de 7.47 dB.

modèle	NMSE
Modèle linéaire	0.874
RN FIR	0.0017
DRNN	0.0012

TAB. 4.8 - : Erreur normalisée des prédictions de la suite de Henon à un pas en avance.

Le réseau entraîné sur la base d'apprentissage bruitée est en mesure de reconstituer la forme précise de l'attracteur original.

Remarquons enfin la décroissance linéaire du gradient rétro-propagé en Fig. 4.21 indiquant que 10 propagations suffisent en moyenne à obtenir une approximation précise du gradient d'erreur.

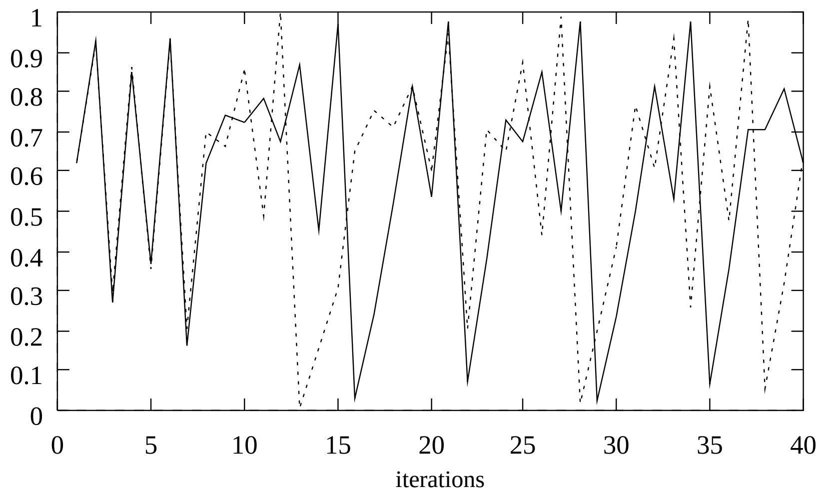

FIG. 4.20 - : Prédictions itérées de la suite de Henon.

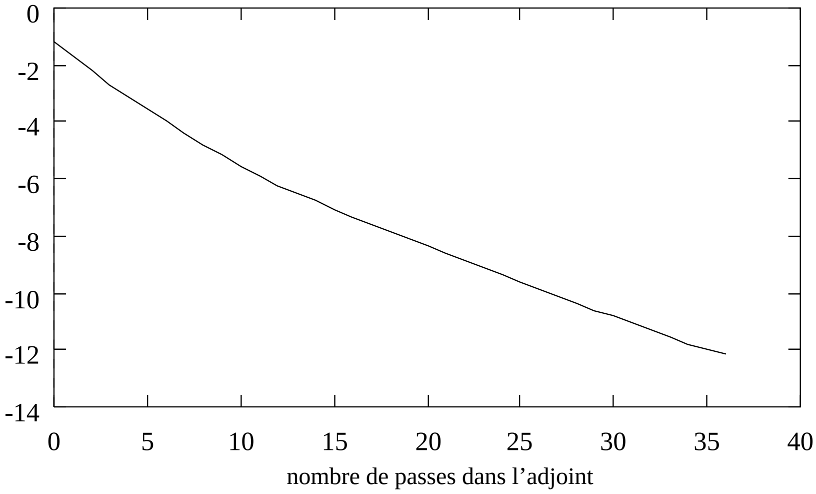

FIG. 4.21 - : Henon : illustration du gradient rétro-propagé. $< \mathrm{Log}_{10}(\|\mathrm{y}^n_{n-j}\|_\infty >$ est tracé fonction du nombre de rétro-propagations j dans le réseau adjoint.

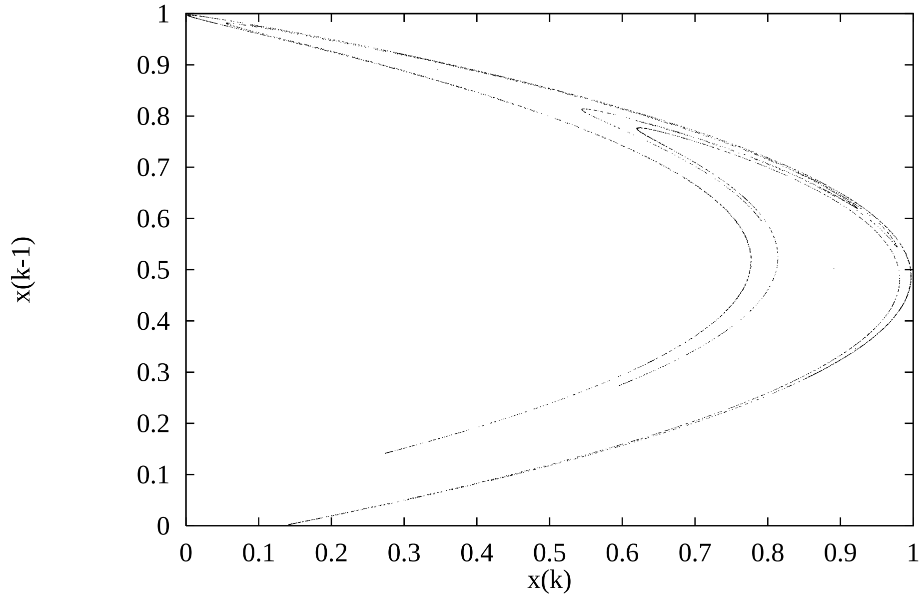

FIG. 4.22 - : Attracteur de Henon : espace des phases de $x(k-1)$ en fonction de $x(k)$.

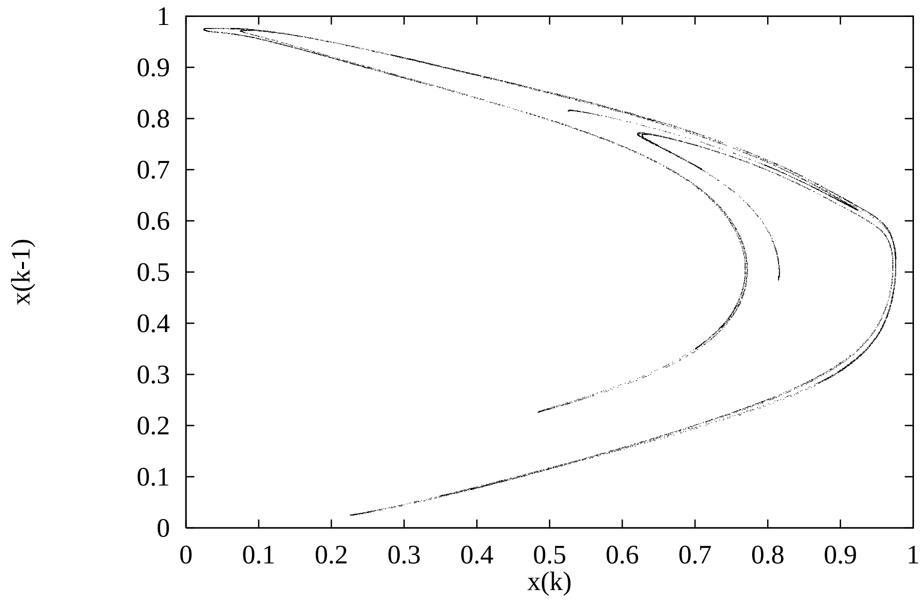

FIG. 4.23 - : Attracteur de Henon reconstruit par un DRNN $1 \times 10 \times 1$ d'ordre $3:1:0$.

4.5.4 Les équations de Lorenz

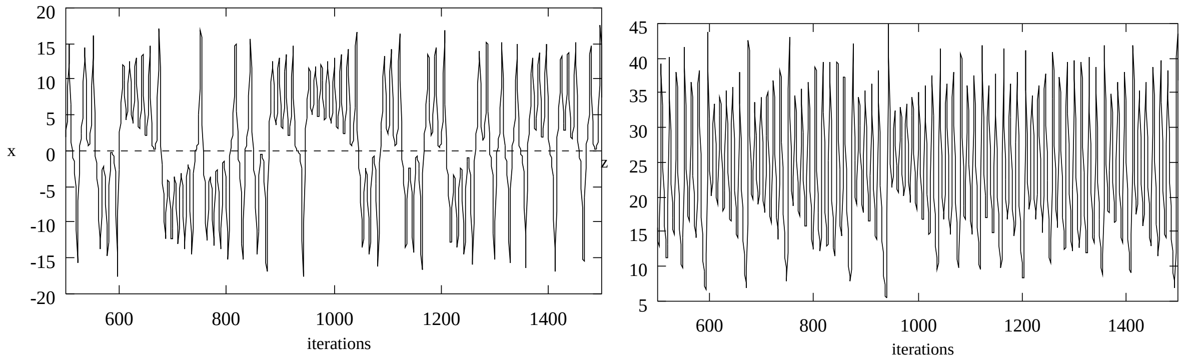

FIG. 4.24 - : Equations de Lorenz (x et y).

Le système de Lorenz est une représentation simplifiée des équations du mouvement d'un fluide en deux dimensions soumis à des échanges thermiques [Lorenz 63, Ott 93]. Ce système d'équations décrit aussi les fluctuations d'un laser infra-rouge [Hubner 93]. Les équations différentielles du système tri-dimensionnel sont les suivantes :

$$\begin{aligned} \frac{dx}{dt} &= -\sigma x + \sigma y \\ \frac{dy}{dt} &= -xz + rx - y \\ \frac{dz}{dt} &= xy - bz \end{aligned} \tag{4.66}$$

Les paramètres prennent leur valeur standard : $\sigma = 10, r = 28, b = 8/3$. Le système est émulé au moyen de la technique Runge-Kutta du 4^{ieme} ordre avec un pas de 0.01. Une suite est constituée en sélectionnant des échantillons avec une période 0.05s. Les variables x et z sont représentées Fig. 4.24. La nature chaotique du processus est manifeste au vu du tracé de z en fonction de x, Fig. 4.26. Le réseau de neurones est alimenté par $x(k)$ et fournit en sortie les estimations $\hat{x}(k+1)$ et $\hat{z}(k+1)$. En délivrant une information tronquée de l'état du système, il incombe au réseau de neurones la lourde tâche d'inférer l'état du système à chaque instant.

Un DRNN de dimension $1 \times 10 \times 2$ avec des filtres d'ordre $30 : 1 : 0$ est entraîné. L'apprentissage est effectué sur 4000 points de la suite et les résultats de la prédiction à un pas en avance sont présentés Table 4.9. Les prédictions du réseau itéré Fig. 4.25 souffrent inévitablement de la divergence des trajectoires.

A la différence des simulations précédentes, le DRNN est ici entraîné à inférer la structure temporelle du processus chaotique, n'ayant à disposition qu'une seule variable d'état. La géométrie fractale de l'attracteur de Lorenz est illustré Fig. 4.26. Le tracé de $\hat{z}$ en fonction de $\hat{x}$, visualisé Fig. 4.27, témoigne de la capacité du modèle neuronal à capturer la relation mutuelle entre les variables du vecteur d'état, nécessaire à la modélisation de la dynamique du système.

Enfin, à titre d'illustration, Fig. 4.28 dévoile la pente observée du gradient à terme de l'apprentissage. La linéarité est manifeste une fois encore.

	NMSE	
modèle	x	z
Modèle linéaire	0.036	0.090
RN FIR	0.0070	0.0095
DRNN	0.0055	0.0078

TAB. 4.9 - : Erreur normalisée logarithmique des prédictions de la suite de Lorenz à un pas en avance.

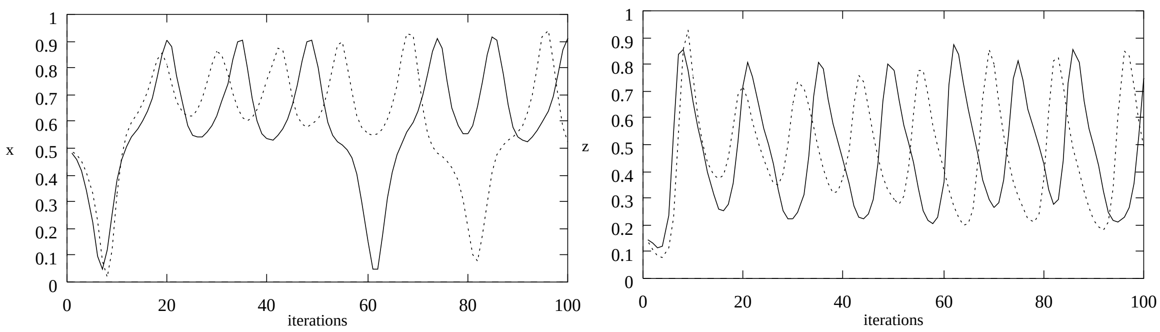

FIG. 4.25 - : Prédictions itérées de la suite de Lorenz.

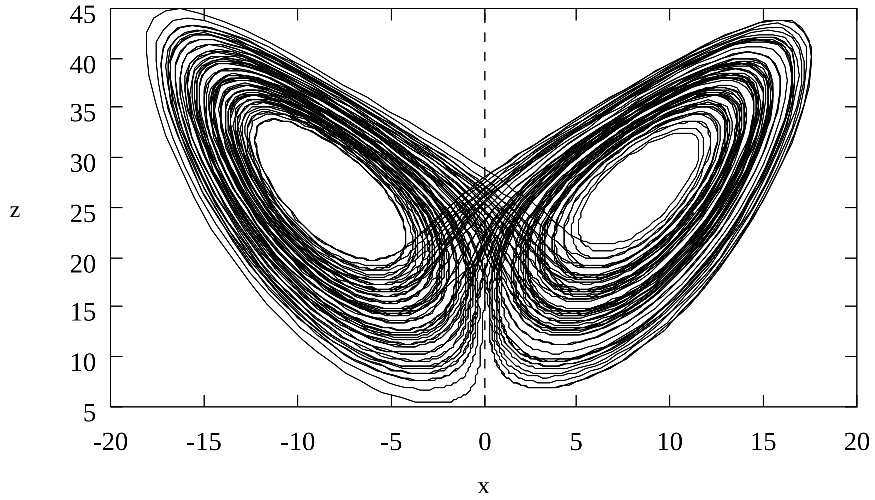

FIG. 4.26 - : Attracteur des équations de Lorenz : espace des phases de z en fonction de x calculé sur 10000 points de la séquence originale obtenus par simulation avec la méthode de Runge-Kutta 4^{ieme} ordre.

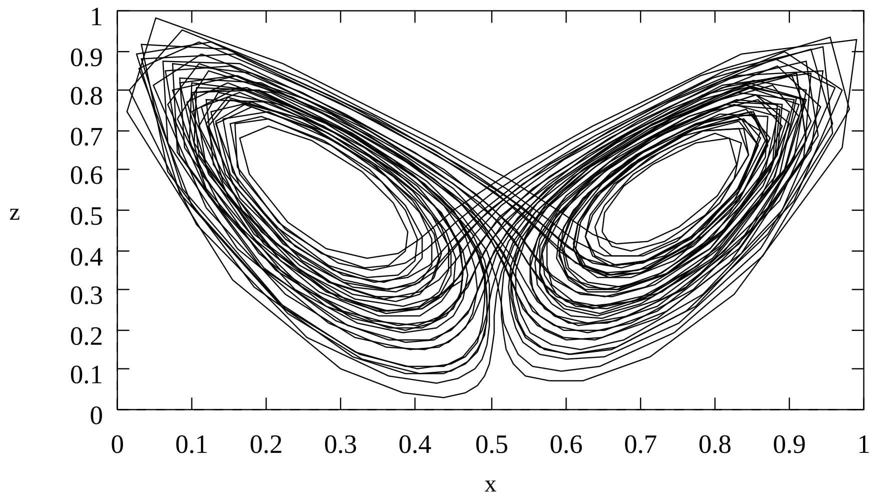

FIG. 4.27 - : Attracteur de l'équation de Lorenz reconstruit avec le DRNN.

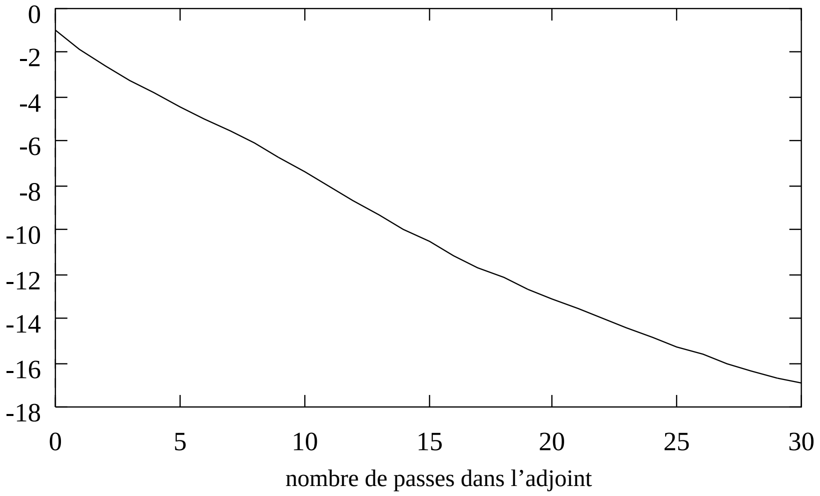

FIG. 4.28 - : Lorenz : illustration du gradient rétro-propagé. $< \mathrm{Log}_{10}(\|\mathrm{y}^n_{n-j}\|_\infty >$ est tracé fonction du nombre de rétro-propagations j dans le réseau adjoint.

4.5.5 La suite d'Ikeda

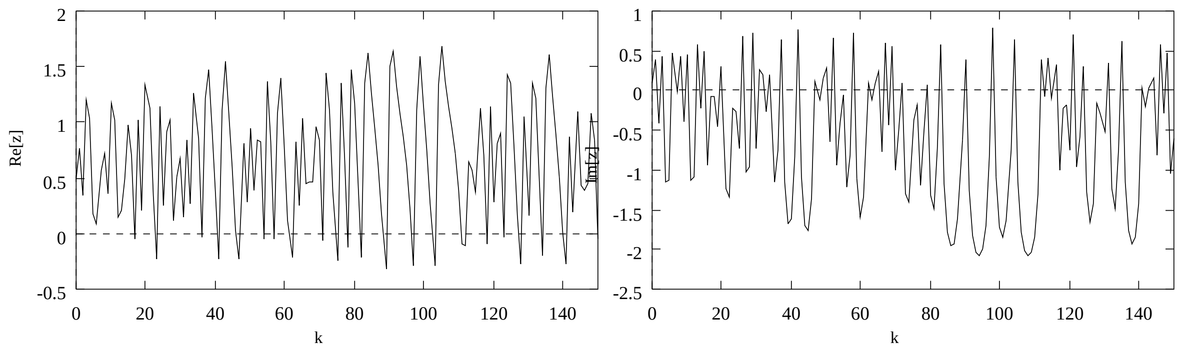

FIG. 4.29 - : Partie réelle et partie imaginaire du nombre complexe issu de l'équation d'Ikeda.

Pour le dernier exemple, nous considérons une équation dans le plan complexe qui émerge de l'étude des interactions d'une onde plane dans un laser [Hammel 85]. L'équation s'écrit

$$z(k+1) = a + bz(k)\exp(i[\phi - c/(1+|z(k)|^2)]) \tag{4.67}$$

avec $a = 1.0, b = 0.9, c = 6.0$ et $\phi = 0.4$. La partie réelle et imaginaire de la suite est illustrée Fig. 4.29 et le tracé de l'espace des phases est représenté Fig. 4.31. A l'instar de l'approche précédente, la tâche dévolue au réseau consiste à reconstruire le vecteur d'état en injectant en entrée du réseau la partie réelle uniquement. Le réseau fournit en sortie les estimations de la partie réelle et imaginaire.

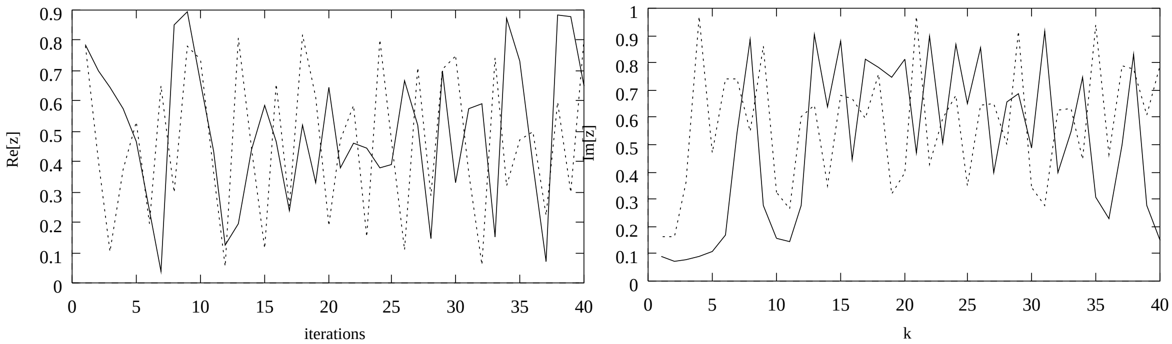

FIG. 4.30 - : Prédictions itérées des équations d'Ikeda : a) partie réelle, b) partie imaginaire.

Un DRNN de dimension $1 \times 15 \times 2$ avec des filtres d'ordre $5 : 1 : 0$ est entraîné sur 10,000 points. La NMSE de la prédiction à un pas en avance, à l'issue de l'apprentissage, est affichée Table 4.10. La reconstruction de l'attracteur par le réseau est fidèle à Fig. 4.31. Une fois encore, le réseau est en mesure de capturer la relation mutuelle entre les variables d'état qu'exige la modélisation de la dynamique du système.

Enfin, l'histogramme du nombre de rétro-propagations du gradient est illustré Fig. 4.33. En règle générale, 20 propagations suffisent à une adaptation exacte des poids.

	NMSE	
modèle	x	z
Modèle linéaire	0.640	0.715
RN FIR	0.0080	0.0150
DRNN	0.0063	0.0134

TAB. 4.10 - : Erreur normalisée logarithmique des prédictions de la suite d'Ikeda à un pas en avance.

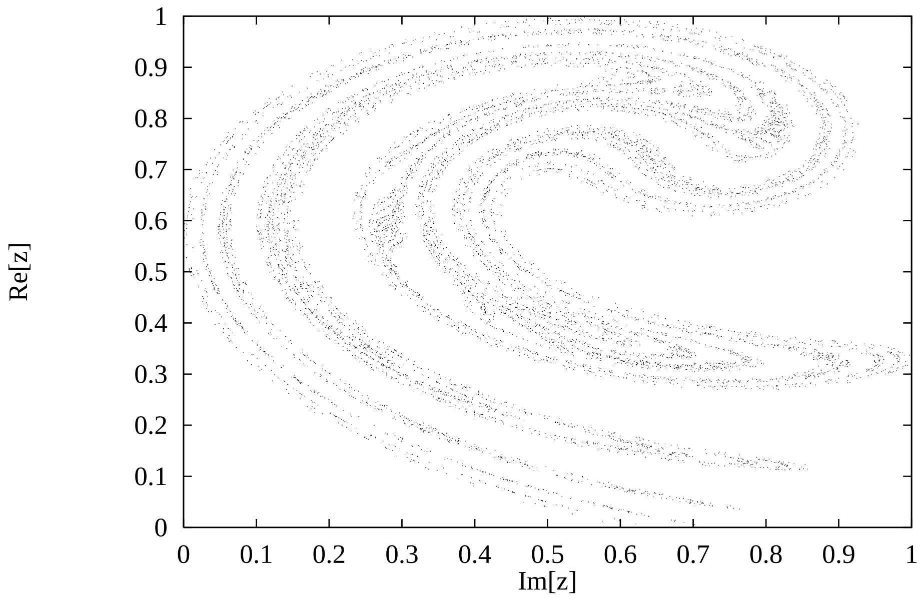

FIG. 4.31 - : Attracteur de la suite d'Ikeda: $Re[z]$ en fonction de $Im[z]$.

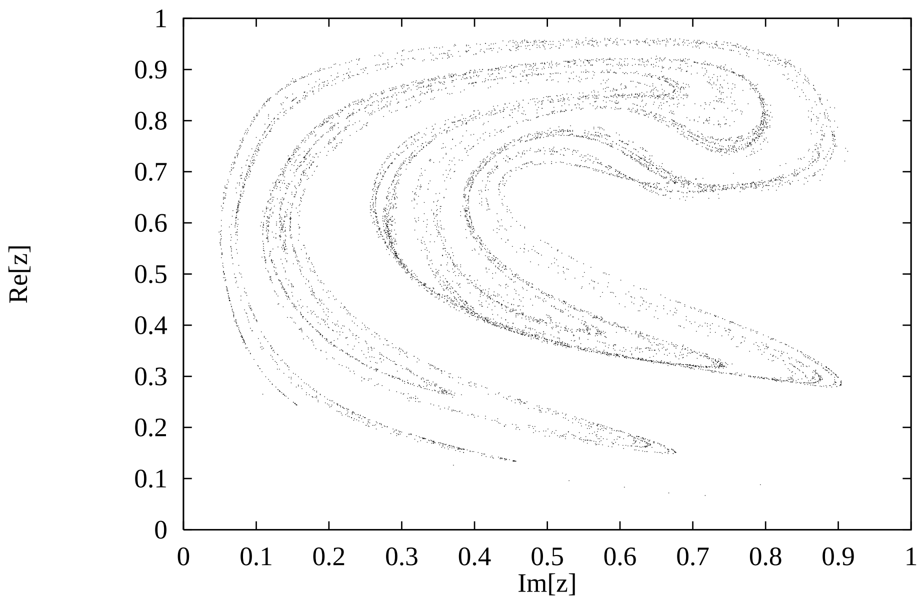

FIG. 4.32 - : Attracteur de l'équation d'Ikeda reconstruit avec le DRNN.

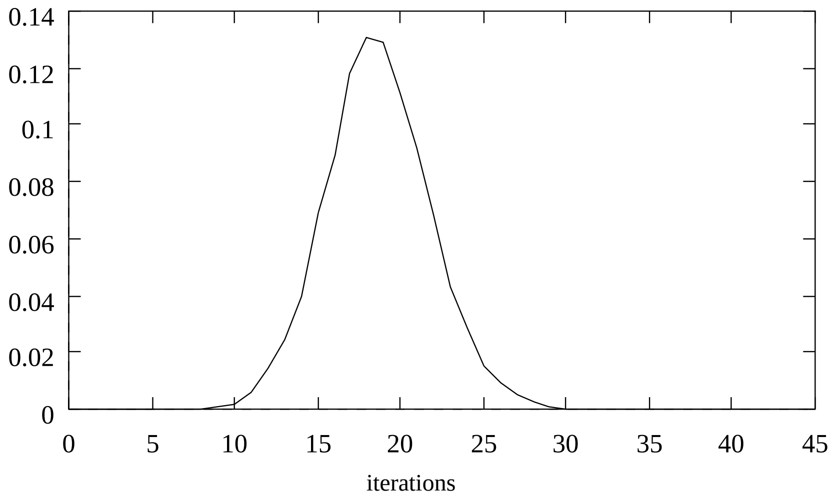

FIG. 4.33 - : Ikeda: histogramme du nombre moyen de rétro-propagations nécessaires dans le réseau adjoint associé au DRNN $1 \times 15 \times 2$ d'ordre $5:1:0$.

4.6 Conclusion

Un état de l'art des méthodes prédictives linéaires puis non-linéaires a été succinctement présenté afin de mettre l'accent sur le potentiel du DRNN à la lumière des faiblesses des autres techniques. Une analyse exhaustive a été menée sur les sunspots dans la double perspective de souligner les propriétés nouvelles du DRNN et de valider les résultats théoriques obtenus au Chapitre 3. Un modèle composé de seulement deux neurones cachés et alimenté par la dernière valeur de la suite est capable de rivaliser avec un modèle non-bouclé de 16 neurones nécessitant une fenêtre temporelle à son entrée. Vu ces résultats surprenants, nous arguons que la suite des sunspots peut être reproduite par l'intégration de deux équations différentielles du second ordre.

Nous avons abordé ensuite la modélisation des processus chaotiques déterministes qui, par définition, ne sont pas prédictibles au-delà de quelques itérations. Le DRNN est capable, dans tous les cas étudiés, de reproduire fidèlement le comportement du processus chaotique en ne disposant que d'un nombre restreint de variables d'états. Une fois collectées les réponses successives du réseau opérant en bouclage fermé, l'attracteur du réseau de neurones est reconstruit. Ce dernier est quasiment identique à celui du processus chaotique. De surcroît, un nombre plus faible de poids qu'il n'en faut à son homologue non-bouclé (le réseau FIR), est nécessaire pour accomplir cette tâche. Enfin, de meilleurs prédictions à court terme sont observées avec le DRNN.