Prévision des ventes avec les réseaux de neurones : une étude de cas*

Chris Chatfield**

Université de Bath, UK

Julian Faraway

Université du Michigan, USA

RÉSUMÉ

A partir des données de Chatfield-Prothero, plusieurs réseaux neuronaux (RN) sont élaborés et les prévisions en résultant sont comparées avec celles des modélisations de Box-Jenkins et de Holt-Winters. Les résultats montrent que les modèles RN présentent des limites et qu'il n'est pas raisonnable de les utiliser de façon automatique. En définitive, l'analyste prudent doit utiliser les compètences traditionnelles en modèlisation pour sélectionner un bon modèle RN en choisissant, par exemple, les variables d'entrées et une structure appropriée. Le critère BIC est recommandé pour comparer les diffèrents modèles. Des précautions doivent être prises lorsqu'on met au point un modèle RN et lorsqu'on l'utilise pour établir des prévisions. Malheureusement, même avec de telles précautions, les prévisions RN se révèlent d'une précision décevante dans notre étude de cas, ce qui est peut-être dû à la brièveté des séries utilisées.

INTRODUCTION

Au cours des dernières années, les chercheurs ont commencé à étudier les réseaux de neurones (RN) dans le but de savoir s'ils pouvaient résoudre des problèmes statistiques (par exemple Ripley, 1993, 1996 ; Cheng et Titterington, 1994), en particulier ceux liés à la modélisation des séries temporelles et à la prévision (Hill *et al.*, 1994). Dans cet article, nous ne prétendons pas présenter une revue de la littérature scientifique en informatique de plus en plus pléthorique. D'ailleurs, cette dernière semble ignorer les problèmes statistiques ou, du moins, est ambigüe à leur sujet (voir par exemple Chatfield (1993) et les commentaires de Faraway et Chatfield (1995) sur Tang *et al.* (1991)). Malgré certaines prétentions, les résultats empiriques des modèles RN sont plutôt mitigés (Chatfield, 1993 ; Hill *et al.*, (1996), Section 3). L'utilité de ces modèles dans la prévision des séries temporelles n'est toujours pas prouvée (Gorr, 1994).

Il est important de savoir si le succès des modèles RN dépend (a) du type de données, (b) de l'habileté de l'analyste à sélectionner le modèle RN adéquat et/ou (c) des méthodes numériques utilisées pour adapter le modèle et calculer des prévisions. La maitrise du point (a) s'est faite par la mise en œuvre de comparaisons de prévisions à grande échelle, telles que la compétition Santa Fe (Gershenfeld et

* Ce travail fût réalisé alors que Julian Faraway était l'hôte de l'Université de Bath.

** Adresse : School of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK. E-mail : cc@maths.bath.ac.uk

Weigend, 1994) ou l'étude de Hill *et al.* (1996). Quant aux points (b) et (c), ils peuvent être appréciés grâce à des études de cas. C'est le but du présent article.

I. – L'ANALYSE DE BOX-JENKINS DES DONNÉES DE CHATFIELD-PROTHERO

Les principales séries temporelles retenues dans cet article correspondent aux données mensuelles des ventes analysées par Chatfield et Prothero (1973a), complétées par six observations supplémentaires données dans Chatfield et Prothero (1973b). Le graphique supérieur de la figure 1 montre que les données ont une tendance ascendante associée à une variation saisonnière dont la taille est plus ou moins proportionnelle au niveau de la moyenne locale (on parle de saisonnalité multiplicative).

L'approche classique de Box-Jenkins (par exemple Harvey, 1993 ; Box *et al.*, 1994) implique les 4 étapes suivantes : (i) transformer les données si cela est jugé nécessaire ; (ii) différencier les données de manière à les rendre stationnaires ; (iii) après exa-

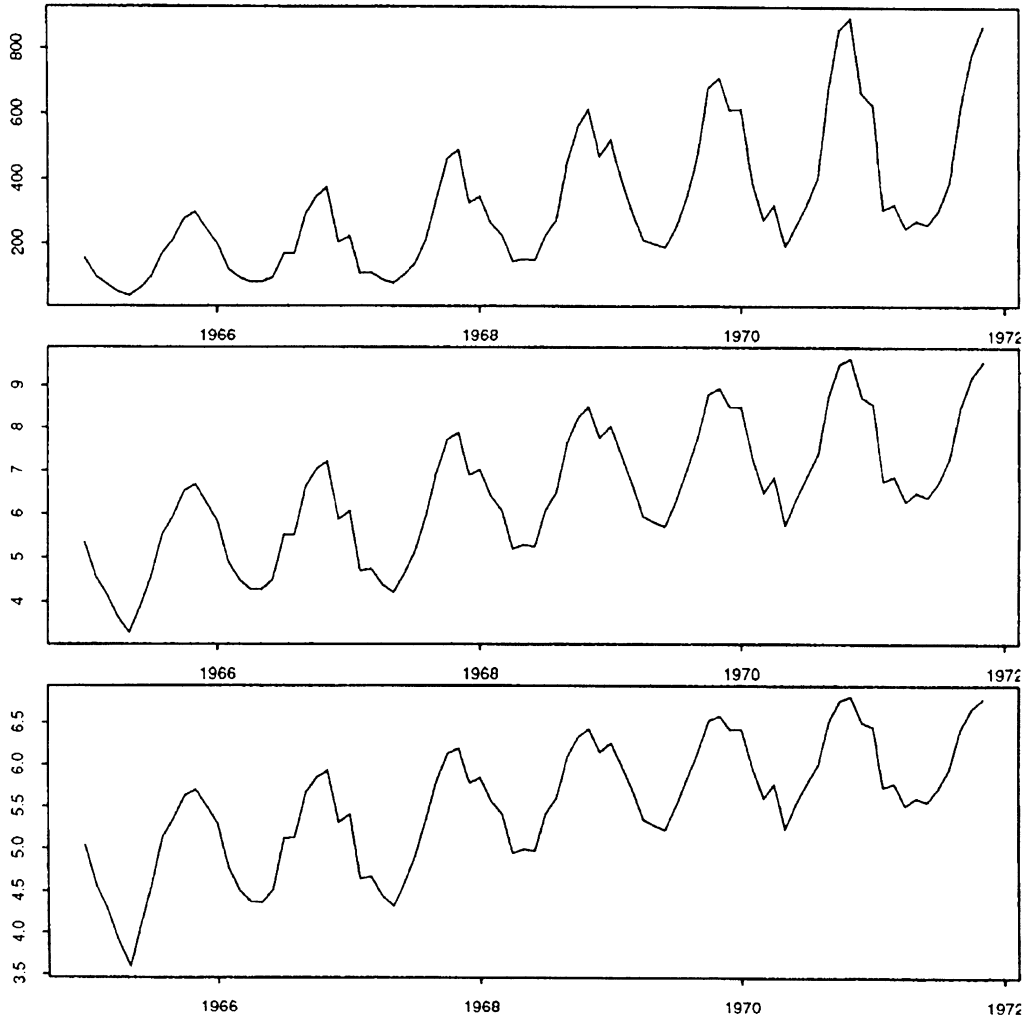

Figure 1. – Données Chatfield-Prothero. Graphe supérieur : ventes mensuelles de la compagnie X de janvier 1965 à novembre 1971 ; graphe intermédiaire : données brutes ; graphe inférieur : logarithmes

men de la fonction d'autocorrélation des séries différenciées, appliquer un modèle saisonnier ARIMA (SARIMA) approprié aux données ; (iv) effectuer des contrôles du diagnostic ; (v) essayer des modèles alternatifs si nécessaire. Cette procédure peut ne pas être directe et le débat entre Chatfield et Prothero (1973a) et Box-Jenkins (1973) puis la réponse de Chatfield et Prothero (1973b) est révélateur de la forte divergence d'opinion quant à la meilleure façon de modéliser les données dans la figure 1. Le principal motif de désaccord était de savoir si c'était les données brutes ou leurs logarithmes ou encore leurs racines cubiques qui devaient être utilisées. Si le but premier d'une transformation de données est de rendre l'effet saisonnier additif, il est difficile de choisir entre les logs ou les racines cubiques (voir figure 1).

Chatfield et Prothero (1973a) ont retenu les logarithmes, donc, une différence saisonnière et une différence non saisonnière, pour rendre les séries stationnaires. Ils ajustent ainsi 4 modèles SARIMA distincts aux logarithmes nommés A, B, C et D et présentés de la façon suivante (voir par exemple Box *et al.*, 1994, p. 333) :

- $A : (1,1,0) \times (0,1,1)_{12}$
- $B : (1,1,0) \times (1,1,0)_{12}$
- $C : (0,1,1) \times (1,1,0)_{12}$
- $D : (0,1,1) \times (0,1,1)_{12}$

Ces 4 modèles fournissent des ajustements aux données à peu près similaires mais des prévisions suffisamment différentes pour qu'il soit difficile de choisir le modèle le plus approprié. Dans la discussion qui suit, les modèles SARIMA ont été également ajustés par des racines cubiques. Chatfield et Prothero (1973b) ont aussi proposé de modéliser les données non transformées en utilisant deux différences saisonnières dans le but d'éliminer l'effet saisonnier multiplicateur. Dans le but d'être complet, nous avons également analysé un modèle, nommé *E*, pour les données non transformées :

- $E : (0,0,1) \times (0,2,1)_{12}$

La plupart des résultats indiqués dans cet article ont été obtenus en ajustant un modèle aux données des 6 premières années puis des prévisions ont été établies sur les 11 dernières observations. Les prévisions seront calculées de 2 façons différentes. Dans un premier temps, toutes les prévisions seront établies à partir de la période 72, en utilisant les données des 6 premières années seulement. Ces prévisions seront appelées multi-pas (MP). Dans un second temps, les prévisions seront établies en avançant pas à pas (1P), les données observées étant introduites une par une. Par exemple, la valeur à l'instant 74 est calculée en utilisant la valeur observée à l'instant 73. Il convient de souligner que les paramètres du modèle ne sont pas re-estimés à chaque pas, lorsqu'on calcule des prévisions une par une (alors que, pour Holt-Winters, les estimations locales de tendance et de saisonnalité sont automatiquement réactualisées).

Pour chacun des modèles, en utilisant les données jusqu'à la date *T*, nous avons déterminé les statistiques suivantes :

1) S = somme des carrés des résidus jusqu'à la date T (les résidus sont les erreurs de prévisions d'un pas en avant).

2) $\widehat{\sigma} = \sqrt{S/(n-p)}$ = écart type résiduel estimé, où n correspond au nombre d'observations utilisées pour ajuster le modèle, et où p représente le nombre de paramètres estimés. Ainsi, pour le modèle *A*, T est égal à 72 et donc n = 72 – 13 = 59 (car 13 observations ont été éliminées en différençant) et p = 2 (puisqu'il y a un paramètre non saisonnier auto-régressif et un paramètre de moyenne mobile saisonnier).

3) AIC correspond au critère d'information de Akaike (= $n \ln(S/n) + 2p$)

4) BIC correspond au critère d'information Bayésien (= $n \ln(S/n) + p + p \ln n$)

5) SS_{MP} = somme des carrés des prévisions multipas en avant effectuées, à la date T, des observations de la date T+1 jusqu'à la fin des séries. Ce sont les prévisions hors échantillon.

6) SS_{1P} = somme des carrés des prévisions un pas en avant des observations de la date T+1 jusqu'à la fin des séries.

Pour davantage d'information sur les critères AIC et BIC, il est recommandé de consulter, par exemple, Priestley (1981). Il nécessaire simplement de retenir que le BIC pénalise l'introduction de paramètres additionnels et conduit donc, généralement, à des modèles au nombre de paramètre plus faible que le AIC.

Les différentes statistiques ci-dessus ont été calculé pour T = 72 et pour les 4 modèles SARIMA ajustés non seulement aux logarithmes mais aussi

aux racines cubiques et aux données non transformées. Le logiciel S-plus a été utilisé. Les résultats pour les modèles sélectionnés sont donnés dans le tableau 3 de la section 4. Les valeurs ont été étalonnées pour les raisons données en section 4. Notons également que toutes les prévisions des modèles utilisant les logarithmes et les racines cubiques ont été retransformées dans les unités d'origine, pour une comparaison équitable. La discussion relative aux résultats est reportée en section 4.

Sans utiliser les logarithmes, un autre moyen d'établir des prévisions pour des données révélatrices de multi-saisonnalité est d'utiliser une version multiplicative du lissage exponentiel de Holt-Winters. La méthode fournit des prévisions dont la pertinence est comparable à celles tirées de l'approche de Box-Jenkins, en particulier pour les séries dont les variations sont dominées par des tendances et des variations saisonnières – voir par exemple Chatfield et Yar (1988). Les résultats obtenus suivant Holt-Winters sont fournis également dans la section 4.

Il est important de souligner que la version multiplicative de Holt-Winters est, par construction, non linéaire ; une prévision n'est pas une fonction linéaire des observations passées. Les modèles ARIMA (linéaires) ajustés aux logarithmes ou aux racines cubiques des données impliquent un modèle non-linéaire pour les données d'origine. Ceci nous conduit à essayer un modèle RN sur les données non transformées pour regarder si la flexibilité non linéaire du modèle RN est en mesure de saisir la saisonnalité multiple.

II. – LES RÉSEAUX DE NEURONES

La brève explication ci-dessous a pour but de rendre cette article compréhensible. Cependant le lecteur pourra également se reporter aux articles de Ripley (1993), Gorr *et al.* (1994, section 2) et/ou Chatfield (1996, section 11.4). Il peut également être profitable de lire une introduction ayant une perspective scientifique en informatique telle que Hertz, Krogh et Palmer (1991) et Gershenfeld et Weigend (1994).

Cet article se focalise sur une forme très courante de RN (artificiels) appelé RN sans boucle de retour avec une couche cachée. Dans la prévision de séries temporelles, nous voulons prévoir les observations futures en utilisant une fonction des observations passées. Un élément déterminant dans les RN est que cette fonction n'a pas besoin d'être linéaire; ainsi un RN peut être envisagé comme une sorte de modèle de régression non linéaire.

La figure 2 représente une architecture typiquement applicable aux prévisions de séries temporelles à partir de données mensuelles. La valeur à la date t doit être prévue en utilisant les valeurs des périodes -1 à -12, ces dernières sont considérées comme des entrées, tandis que les prévisions constituent les sorties. L'exemple illustré inclut une couche cachée de deux neurones. De plus, il existe un terme constant en entrée qui pour des raisons pratiques prend la valeur un. Chaque entrée est connectée aux 2 neurones (cachés), eux-mêmes connectés aux sorties. Il existe aussi une connexion directe de l'entrée constante à la sortie. La « force » de chaque connexion est mesurée par un poids. Une valeur numérique est calculée pour chaque neurone dans les 2 étapes. Tout d'abord une fonction linéaire des entrées est déterminée, laquelle est ensuite transformée en appliquant une fonction appelée fonction d'activation, qui est par construction non linéaire. Une fonction couramment utilisée est la fonction logistique qui donne des valeurs comprises entre 0 et 1. Une opération analogue peut être appliquée aux valeurs des neurones et à l'entrée constante afin d'obtenir la prévision. Il faut cependant noter que la fonction logistique ne doit généralement pas être utilisée au stade du résultat dans la prévision des séries temporelles car elle contraint cette dernière dans l'intervalle (0, 1).

L'introduction d'une constante d'entrée unitaire, connectée à chaque neurone située dans la couche cachée ainsi qu'à chaque sortie, évite d'introduire séparément ce que les informaticiens appellent un biais pour chaque unité. Les biais deviennent simplement parties intégrantes de la série de poids (les paramètres).

Pour le modèle RN de la figure 2, la procédure de calcul d'une prévision de x_t (la sortie) en utilisant les observations retardées, x_{t-1} à x_{t-12}, comme entrées peut être expliquée de la manière suivante. Soit w_{c1} et w_{c2} les poids des connexions entre l'entrée constante et les deux neurones cachées, et $w < 0$ le poids de la connexion directe entre l'entrée constante

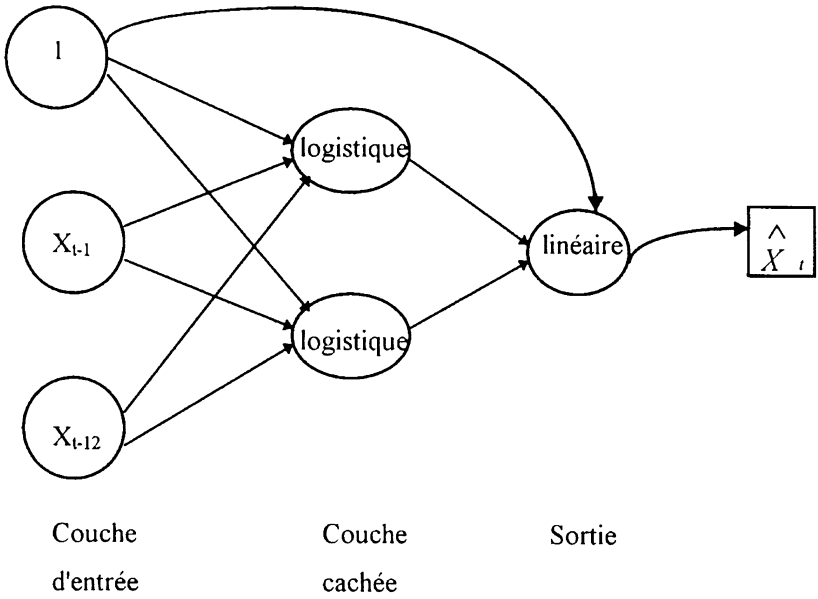

Figure 2. – Architecture d'un réseau de neurones type pour la prévision de séries temporelles avec une couche cachée de deux neurones. La sortie (la prévision) dépend des valeurs retardées au temps (t-1) et (t - 12)

et la sortie. Soit les poids $\{w_{ih}\}$ qui représentent les poids pour les autres connexions entre les deux entrées retardées (i = 1 pour x_{t-1} ou i = 2 pour x_{t-12}) et les deux neurones cachés (h = 1 ou 2) et soit w_{10} et w_{20} les poids pour les connexions entre les deux neurones et la sortie. Les notations des deux neurones cachés peuvent être permutées sans changer de modèle. On prend maintenant la somme linéaire des entrées à chaque neurone et on applique une fonction d'activation, ϕ, pour déterminer les valeurs, appelées z_1 et z_2, pour les deux neurones. Dans notre exemple :

$$z_h = \phi(w_{ch} + w_{1h}x_{t-1} + w_{2h}x_{t-12}) \qquad (1)$$

où h = 1, 2.

Pour finir, on calcule la somme linéaire des valeurs des deux neurones et le poids de la connexion directe avec l'entrée constante (ceci implique que la fonction d'activation à l'étape résultat est la fonction identité).

$$\hat{x}_t = (wco + \sum_h w_{ho}z_h) \qquad (2)$$

On utilise généralement la notation RN(j_1, ..., j_k ; h) pour désigner le RN avec les délais j_1, ..., j_k et h neurones dans la couche cachée. Ainsi, la figure 2 représente un modèle NN(1, 12 ; 2).

Les poids utilisés dans un modèle RN sont estimés, à partir des données, en minimisant la somme des carrés des erreurs de prévision à l'intérieur de l'échantillon, soit $s = \sum_t(\hat{x}_t - x_t)^2$, sur la première partie des séries temporelles. Cette dernière est appelée la série de formation dans le jargon des RN. Dans notre étude de cas, ce sont les données des 6 premières années. La minimisation n'est pas facile du fait que la fonction objectif présente souvent plusieurs minima locaux et que le nombre de poids peut être important. Plusieurs algorithmes numériques ont été proposés afin de sélectionner les poids mais ces algorithmes d'essai sont susceptibles d'exiger des centaines d'itérations pour converger, et encore peuvent-ils converger vers un minimum local. Les valeurs de départ choisies pour les poids peuvent être cruciales et il est conseillé d'essayer plusieurs séries de valeurs de départ afin de voir si des résultats significatifs sont obtenus.

La dernière partie de la série temporelle, la séquence test, est gardée en réserve de manière à ce que des prévisions véritables hors échantillon puissent être établies et comparées avec les observations réelles. L'équation (2) propose effectivement une prévision « un pas en avant », puisqu'elle utilise les valeurs observées réelles de toutes les variables retar-

dées comme entrées. Si des prévisions multi-pas en avant sont requises, il est alors possible de procéder d'une des deux manières suivantes. Premièrement, on peut élaborer une nouvelle architecture avec plusieurs sorties, $\widehat{x}_t, \widehat{x}_{t+1}, \widehat{x}_{t+2}, \ldots, \widehat{x}_t, \widehat{x}_{t+1}, \widehat{x}_{t+2}, \ldots$ où chacune des sorties aurait des poids séparés pour chaque connexion aux neurones. Deuxièmement, on peut réintroduire la prévision « un pas en avant » afin de replacer la valeur retardée de un comme une des variables d'entrée ; ainsi, la même architecture peut être utilisée pour construire la prévision « deux pas en avant » et ainsi de suite. Cette dernière approche itérative a été adoptée en raison de sa simplicité numérique et du nombre moins important de poids devant être estimés.

Malgré cela, le nombre de paramètres dans un modèle RN est manifestement plus important que dans les modèles de séries temporelles classiques, et il est donné, pour un modèle à couche cachée unique, par $p = (n_i + 2)n_u + 1$ où n_i représente le nombre de variables d'entrées (en excluant la constante) et n_u le nombre de neurones cachés (ou unités). A titre d'exemple, l'architecture de la figure 2 (où n_i et n_u sont égaux à deux) contient neuf connexions et a donc neuf paramètres (poids). Le nombre conséquent de paramètres signifie qu'il existe un réel danger que l'algorithme d'ajustement surexploite les données et réalise un ajustement artificiellement bon qui ne conduit pas à de meilleures prévisions. Lorsque l'on compare les modèles RN, il est important d'utiliser des critères tels que le critère d'information d'Akaike (AIC) pour prévenir l'ajustement de paramètres faux, en pénalisant l'ajustement de paramètres supplémentaires, plutôt que simplement comparer la bonne qualité d'ajustement (Ripley 1995).

La modélisation RN est non-paramétrique par définition et il a été suggéré que l'ensemble du processus global puisse être totalement automatisé par l'outil informatique « de façon que les personnes puissent, avec peu de connaissances sur les prévisions ou les RN, établir rapidement des prévisions de qualité » (Hoptroff, 1993). Cet aspect « boîte-noire » peut être vu comme un avantage ou un désavantage. Il est certain que des boîtes noires sont susceptibles de donner des résultats aberrants et que les modèles RN n'en sont pas à l'abri. C'est pourquoi Gershenfeld et Weigend (1994, p. 7) déclarent « qu'il y a eu un échec général des approches « boîte-noire » simplistes ; dans toutes les études réussies (dans la compétition Santa Fe), une analyse exploratoire des données précédait l'application de l'algorithme ». Les résultats de notre étude de cas démontrent aussi qu'un bon modèle RN pour des séries temporelles doit être sélectionné en combinant les compétences générales en modélisation avec une connaissance des analyses de séries temporelles et des problèmes spécifiques rencontrés avec les modèles RN. Notre étude de cas s'attachera : (1) au choix du nombre de variables d'entrées, (2) au choix du nombre de neurones dans la couche cachée, (3) à la procédure numérique d'estimation des poids en incluant le choix des valeurs de départ, (4) au critère de sélection du meilleur modèle.

III. – L'AJUSTEMENT DES MODÈLES RN AUX DONNÉS DES VENTES

Le logiciel que nous avons utilisé pour ajuster les réseaux neuronaux aux séries temporelles peut être consulté à l'adresse Internet suivante : *http ://www.stat.lsa.umich.edu/~faraway/* ; vous y trouverez les détails complets sur l'installation et l'utilisation de ce logiciel. Nous avons utilisé plusieurs fonctions S-plus de Venables et Ripley (1994) (incluant *nnet* et *nnet.Hess*) en conjonction avec certaines fonctions écrites par le second de ces deux auteurs. La fonction *nnet* () utilise l'algorithme de Broyden-Fletcher-Goldfarb-Shannon qui est très proche de la méthode de Newton. Les premières dérivées de la fonction objectif sont calculés en utilisant l'algorithme de rétropropagation. Les valeurs de départ sont soit choisies de façon aléatoire dans une certaine région soit spécifiées par l'utilisateur. La connaissance de S-plus est requise pour une utilisation fructueuse du logiciel. Nous sommes conscients des développements approfondis de la recherche en matière d'ajustement des modèles RN, tels que les méthodes Bayésiennes, mais celles-ci se situent encore au niveau expérimental. Nous pensons donc qu'il est plus approprié d'utiliser le logiciel commercialement disponible et d'évaluer son comportement dans le cas général.

Nous avons décidé de focaliser notre attention sur les modèles RN à une seule couche cachée. L'utilisation de deux couches cachées ou plus, avec tous les

paramètres supplémentaires que cela implique, semble injustifiée pour des séries si brèves. Nous avons jugé opportun d'utiliser la fonction d'activation logistique au niveau de la couche cachée puisque c'est la plus couramment utilisée. Nous nous sommes immédiatement heurtés à plusieurs problèmes pratiques. Le choix par défaut de la fonction d'activation du logiciel, à la fois aux stades du résultat et de la couche cachée, est la fonction logistique. L'impossibilité de spécifier la fonction identité d'activation au stade résultat donnait donc des résultats aberrants, puisque des données non-tranformées se situaient entre 36 et 895. En outre, les valeurs de départ utilisées dans l'algorithme sont hors échelle, l'algorithme d'ajustement n'a donc pu converger de façon satisfaisante. Aussi, il nous a fallu réétalonner les données en les divisant par 100. Tous les chiffres ultérieurs, y compris les prévisions, font référence à ces données. Il apparaît que les poids de départ doivent varier tout au long d'un intervalle raisonnable, ni trop large ni trop étroit, comparé à celui des données. Diviser par 1000 plutôt que par 100 a aussi conduit à des problèmes de convergence. Une procedure d'essai-erreur peut être nécessaire pour sélectionner une échelle adaptée.

Un autre problème est apparu. En raison de redémarrages consécutifs de l'algorithme d'ajustement, avec différents points de départ aléatoires pour les poids, plusieurs minima locaux ont été trouvé (ou même des points selle). Par exemple, on aboutit à 5 minima locaux distincts pour le modèle RN(1, 12 ; 2) et l'ajustement ainsi que l'exactitude des prévisions sont donnés au tableau 1. Nous avons vérifié que la Hessienne est positive pour les 5 minima (même si cela peut s'avérer difficile puisque les minima tendent à être plat) afin de s'assurer qu'ils représentent bien des minima locaux. Comme l'algorithme converge souvent vers un point selle, il est réellement important de vérifier la Hessienne. Même si nous relançons plusieurs fois l'algorithme à partir de différents points de départ, il n'est pas garanti que le cinquième modèle du tableau 1 (qui a la valeur S la plus petite) fournisse le meilleur ajustement global. Tous les modèles évoqués ci-dessous sont le résultat de réajustements du modèle au moins 50 fois à partir des différents points de départ aléatoires en retenant les meilleurs minima.

Tableau 1. – Ajustement et exactitude des prévisions de 5 minima locaux pour le modèle RN(1,12 ; 2)

Prévisions		
S	SS_{MP}	SS_{IP}
247	70,5	70,5
13,8	21,3	18,5
13,8	21,3	18,5
10,7	14,1	17,8
10,5	16,9	30,0

Le lecteur averti remarquera que les deuxième et troisième minima donnés ci-dessus ont des ajustements et des prévisions identiques, on peut donc espérer qu'ils correspondent au même modèle. Il n'en est rien et les poids ajustés de ces deux modèles démontrent hélas qu'ils ne sont pas stables pour les différents minima. Le tableau 2 donne les poids ajustés pour les deux modèles correspondant aux deuxième et troisième minima du tableau 1. De forts écarts apparaissent particulièrement dans les poids relatifs donnés par les deux neurones cachés losrqu'on les connecte à la sortie. Cette instabilité montre qu'il est généralement imprudent d'essayer d'interpréter des poids individuels. Même au niveau du minimum global, il est possible que les poids du modèle correspondant puissent être changé sans modifier notablement l'ajustement ou les prévisions (cas d'une régression multicolinéaire).

Nous avons ensuite comparé plusieurs modèles RN présentant des nombres différents de neurones dans la couche cachée ainsi que des choix différents de variables d'entrées (retardées). Tous les modèles

Tableau 2. – Comparaison des poids de deux minima locaux différents.
(notation définie section 3)

w_{c1}	w_{11}	w_{21}	w_{c2}	w_{12}	w_{22}	w_{co}	w_{1o}	w_{2o}
– 0,40	0,01	0,11	– 30,6	– 12,0	– 8,3	– 14,9	37,6	32,9
– 0,41	0,01	0,12	– 12,1	– 2,3	– 3,3	– 14,4	36,7	7,9

ont été ajusté en utilisant les données des six premières années. Ils ont été alors utilisés pour prévoir les onze dernières observations. Quand plusieurs modèles au nombre de paramètres différents sont comparés, la moyenne résiduelle au carré tout au long de la période d'ajustement ne fournira pas une comparaison équitable. Elle sera biaisée au profit de modèles avec de nombreux paramètres. Les critères de sélection d'un modèle tels que le AIC et BIC, définis dans la section 2, fournissent une comparaison plus équitable. Pour les modèles RN, le nombre de paramètres p est le nombre de poids, tandis que n le nombre d'observations effectives dépend du retard maximum. Les résultats, pour un certain nombre de modèles sélectionnés, sont fournis dans le tableau 3. Les symboles sont définis dans la section 2. Le tableau 3 inclut également les valeurs correspondantes aux modèles de Box-Jenkins, comme décrit dans la section 2. Notons que les chiffres ont été standardisés pour les rendre comparables aux valeurs RN ajustées aux données divisées par 100.

Un modèle linéaire autorégressif a aussi été ajusté aux données non-transformées. Les résultats sont présentés dans le tableau 3. Finalement, la méthode de prévision de Holt-Winters a été retenue avec l'utilisation du logiciel AUTOCAST. Il n'est pas possible de donner les valeurs AIC et BIC pour cette méthode puisqu'elle n'est pas fondée sur un modèle à paramètres constants. Les valeurs de S pour les tendances réduites et linéaires sont respectivement 14,6 et 13,8, alors que les valeurs correspondantes pour SS_{MP} sont respectivement 6,1 et 10,4. Nous n'avons pas calculé les prévisions 1P pour Holt-Winters car elles ne sont pas comparables avec celles des modèles RN ou ARIMA ; en effet, les estimations du niveau, de la tendance et de la saisonnalité sont automatiquement réactualisées après chaque observation.

En examinant tous ces résultats, le lecteur constatera que nous avons utilisé une large gamme de choix pour les variables d'entrées retardées afin d'évaluer les sélections qui sont susceptibles d'être faites par des analystes des séries temporelles non-expérimentés ou confirmés. Il est également utile de garder à l'esprit que les quelques dernières observations dans les séries sont beaucoup plus basses que ce que l'on était susceptible d'attendre d'une simple extrapolation des séries. Ainsi, le modèle RN(1-4 ; 2), qui représente le type même de modèle susceptible d'être essayé par quelqu'un sans grande expérience dans les séries temporelles, présente un ajustement très faible (le BIC est catastrophique). Il donne, cependant, les meilleures prévisions 1P ; probablement parce-qu'il réagit rapidement aux changements à la fin des séries. L'analyste légèrement plus confirmé, réalisant qu'il s'agit de données annuelles, est susceptible d'utiliser tous les retards jusqu'au 12 voire jusqu'au 13. Le modèle RN(1-13 ; 2) contient de nombreux paramètres et présente une valeur S petite qui pourrait conduire toute personne inexpérimentée à croire qu'il s'agit d'un bon modèle, spécialement à la vue du critère AIC. Cependant, la valeur du BIC fournit un point de vue différent et la performance prédictive est faible. L'expérience acquise grâce à l'utilisation d'autres méthodes de prévision suggère qu'il est raisonnable d'inclure le décalage 12 sans les décalages intermédiaires et les modèles à décalage 1,12 et 1, 12, 13 ont une meilleure performance prédictive, surtout si peu de neurones cachés sont utilisés. Notons que les entrées aux décalages 1, 12 et 13 sont dus à l'utilisation d'une différence saisonnière et non saisonnière dans la modélisation de Box-Jenkins. D'autres combinaisons de variables d'entrée ont été essayées mais les résultats (non présentés ici) étaient plus décevants.

Nous avons aussi envisagé la possibilité d'essayer des modèles RN correspondant le plus possible aux versions non-linéaires des modèles de Box-Jenkins de la section 2. Toutefois, nous n'avons pas poursuivi. En effet, bien que les modèles AR et ARI puissent être envisagés comme des modèles RN avec des fonctions d'activation linéaires, nous ne savons pas faire face à des structures de type MA dans le cadre d'une formulation RN ; or 4 des 5 modèles de Box-Jenkins (modèles A,C,D,E) contiennent des termes MA. En outre, la modélisation RN aurait un avantage injuste et critiquable si elle pouvait tirer partie des résultats de l'analyse antérieure de Box-Jenkins. Le modèle RN équivalent au modèle B de la section 2 impliquerait des variables d'entrées aux retards 1, 2, 12, 13, 14, 24 et 25. Il est peu probable qu'un tel choix soit fait spontanément par le modélisateur !

Si le modèle est choisi sur la base de la minimisation du critère AIC ou de l'écart type (ou $\hat{\sigma}$ est comparable au R^2 ajusté), alors le modèle RN(1-13 ; 2) sera retenu et les prévisions seront médiocres. Bien sûr, des statisticiens expérimentés devineraient intuitivement que l'utilisation de trop d'entrées ne peut pas conduire à de bons résultats et n'aurait vraisem-

Tableau 3. – Comparaison de divers modèles RN, de divers modèles de Box-Jenkins et un simple modèle de régression linéaire. Toutes les statistiques sont étalonnées de la même façon et les données transformées ont été convenablement reconverties.

Retards	Nombre de neurones cachés	Nombre de paramètres	Ajustement S	$\hat{\sigma}$	AIC	BIC	prévisions SS_{MP}	SS_{IP}
RN utilisant données non transformées								
1-4	2	13	20,0	0,60	– 57,1	– 15,3	21,3	2,7
1-13	2	31	1,6	0,24	– 152,0	– 56,9	134,0	79,8
1,12	1	5	13,8	0,50	– 78,3	– 62,8	21,0	18,2
1,12	2	9	10,5	0,45	– 86,4	– 58,6	16,9	30,0
1,12,13	1	6	12,4	0,48	– 79,8	– 61,3	19,0	12,0
1,12,13	2	11	9,5	0,44	– 85,9	– 52,0	21,3	22,3
RN utilisant racines cubiques								
1-4	2	13	21,7	0,63	– 51,8	– 10,0	150,0	9,4
1,12,13	1	6	13,3	0,50	– 76,0	– 57,5	15,2	8,7
1,12,13	2	11	9,9	0,45	– 83,1	– 49,2	22,3	10,8
RN utilisant logarithmes								
1-4	2	13	27,0	0,70	– 36,8	5,1	146,0	13,7
1,12,13	1	6	14,0	0,51	– 72,8	– 54,4	13,3	7,5
1,12,13	2	11	12,3	0,51	– 70,3	– 36,5	15,2	6,8
Box-Jenkins sur logarithmes								
Modèle A		2	12,6	0,47	– 84,6	– 80,5	17,5	4,0
Modèle B		2	11,2	0,50	– 60,9	– 57,2	13,8	4,8
Modèle C		2	11,8	0,51	– 60,9	– 57,2	17,2	5,2
Modèle D		2	13,5	0,49	– 83,2	– 79,1	20,3	4,0
Box-Jenkins sur racines cub.								
Modèle A		2	11,7	0,46	– 88,8	– 84,7	18,5	4,2
Box-Jenkins sur données non transformées								
Modèle A		2	14,9	0,52	– 74,8	– 70,7	16,7	5,4
Modèle B		2	18,2	0,63	– 42,4	– 38,7	13,4	11,2
Régression linéaire avec retard 1, 12 et 13								
Régression linéaire		4	12,5	0,48	– 83,7	– 71,4	21,0	12,8

blablement jamais recours à un tel modèle. Par opposition, le meilleur modèle RN retenu par le critère BIC est le modèle RN(1, 12 ; 1). Il est relativement cohérent, contient beaucoup moins de paramètres et il établit des prévisions raisonnables. En résumé, dans ce contexte, l'utilisation du critère de l'AIC est insuffisante pour pénaliser les paramètres supplémentaire.

Tableau 4. – Poids dans le modèle RN(1-13 ; 2) pour les connexions des 13 valeurs retardées aux 2 neurones cachés.

	Retard												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	– 28,3	54,1	– 54,8	26,5	– 27,3	9,4	31,4	– 12,8	4,5	– 19,9	– 37,9	94,9	– 28,5
2	0,2	0,0	0,1	0,0	0,4	0,1	– 0,7	0,2	0,2	– 0,5	1,5	– 0,5	– 0,3

Les modèles de Box-Jenkins donnent généralement de meilleurs BIC que les modèles RN et de nettement meilleures prévisions 1P. Le modèle *E* sur les données non transformées donne un BIC médiocre mais de relativement bonnes prévisions MP. Le modèle de régression linéaire donne un meilleur BIC que tous les autres modèles. Pourtant les meilleures prévisions MP sont établies par Holt-Winters !

Il est naturel de se demander si les meilleures modèles RN trouvés précédemment (ceux qui incluent les retards 1, 12 et peut être 13 mais excluent les retards intermédiaires) auraient pu être découverts sans la connaissance des spécialistes des séries temporelles, issue par exemple de l'expérience tirée de la modélisation de Box-Jenkins. Afin de répondre à cette question, il est utile de se reporter au tableau 4 qui présente les poids des connexions entre les 13 valeurs retardées et les 2 neurones du modèle RN(1-13 ; 2). Il n'y a aucune indication que ces poids sont « petits » pour les retards 2-11. Aussi, dans ce cas, un examen des poids ne s'avère pas utile. Aux vues de l'instabilité dans l'estimation des poids mentionnés précédemment, ce résultat est généralisable.

Pourquoi le fait d'avoir trop de neurones cachées entraîne-t-il une performance faible ? Considérons le modèle RN(1, 2, 12) avec 2 ou 4 neurones cachées. La prévision varie sur les 3 retards; il n'est pas possible de la tracer directement mais on peut la considérer suivant des directions particulières. Dans la figure 3, la réponse un pas prévue est tracée selon la direction de la première composante principale des valeurs aux retards 1, 12 et 13 car la valeur du retard 1 change. Il convient de remarquer que le modèle RN(1, 2, 12 ; 4) s'écarte de la linéarité pour parvenir à un meilleur ajustement sur les données centrales, mais ceci a pour effet d'augmenter les prévisions très rapidement, d'une façon instable, pour les valeurs les plus élevées. Pour être clair, obtenir un meilleur ajustement peut entraîner des prévisions plus faibles.

Une question intéressante (à laquelle nous n'essayons pas de répondre) est de savoir si prendre les logarithmes ou les racines cubiques devrait compter comme un paramètre supplémentaire. Recalculer AIC, BIC et σ sur cette base rend le modèle sur les données non-tranformées moins pertinent.

Le lecteur statisticien a pu noter que nous n'avons rien dit sur la structure d'erreur sous-jacente des modèles RN ajustés dans cette section. Un composant standard de la procédure de modélisation de Box-Jenkins est une analyse résiduelle dont le but est de mettre à jour d'éventuelles déviations par rapport aux hypothèses selon lesquelles les erreurs sont non corrélées et normalement distribuées avec une variance constante. Les praticiens des modèles RN considèrent rarement de tels problèmes mais les résidus devraient réellement faire l'objet d'un contrôle pour s'assurer que le critère d'erreur quadratique est raisonnable. Notre analyse suggérait que les résidus pour les meilleurs modèles RN présentent des queues lourdes par rapport à la distribution normale, mais pas suffisamment pour nous faire considérer un critère d'erreur alternatif. Pour quelques modèles RN faiblement ajustés, il existe une structure de corrélation dans les résidus comme l'on pouvait s'y attendre.

Un lecteur nous a suggéré que les résultats peuvent être la conséquence de notre incapacité à éliminer la tendance avant l'ajustement des modèles RN. La raison pour laquelle cela serait nécessaire ne nous apparaît pas clairement puisqu'un modèle ARI a, par exemple, un équivalent RN. Savoir comment évacuer cette tendance n'apparaît pas non plus clairement (en

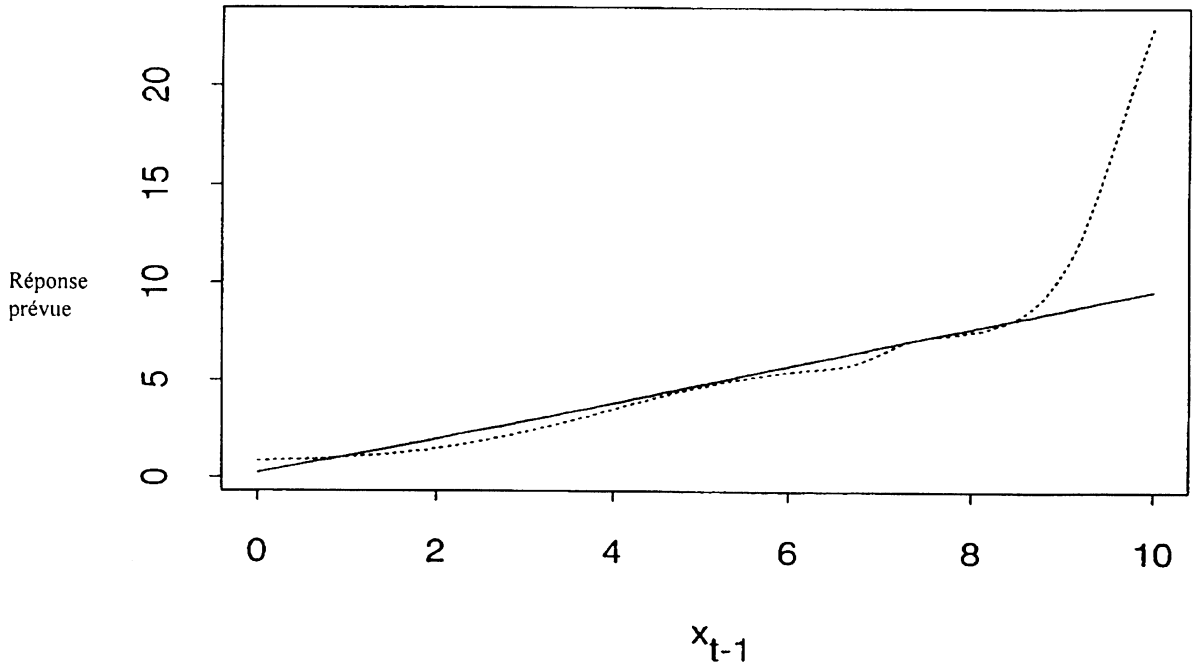

Figure 3. – Réponse prévue pour les modèles RN(1, 12, 13 ; 2) (ligne continue) et RN(1, 12, 13 ; 4) (ligne en pointillés) selon la direction de la première composante principale des valeurs retardées 1, 12 et 13 contre la valeur au retard 1.

appliquant une tendance linéaire globale ou locale ?, en utilisant des moyennes mobiles courtes ? en différenciant ?). Les méthodes de Box-Jenkins et Holt-Winters sont destinées à venir à bout des tendances, ainsi que de la saisonnalité ; la modélisation RN ne sera pas compétitive, à moins que le traitement de la tendance fasse partie intégrante de la procédure. Il peut être profitable d'appliquer la modélisation RN à des données différenciées convenables comme dans l'approche de Box-Jenkins ; ceci devrait faire l'objet d'un autre article.

Finalement, sur tous les modèles considérés dans cette section, l'estimation à l'intérieur de l'échantillon de l'écart-type d'erreur (c'est-à-dire $\widehat{\sigma}$) est nettement inférieur à celui prédit, c'est-à-dire $(SS_{1P}/12)^{1/2}$. Les meilleures valeurs de $\widehat{\sigma}$ se situent aux alentours de 0,1 ; alors que la meilleure estimation de l'écart-type prédit est 0,15 pour le modèle RN(1, 2, 12 ; 2), mais se situe davantage autour de 0,20. Ceci est un phénomène répandu dans le domaine de la prévision des séries temporelles où l'ajustement à l'intérieur de l'échantillon est généralement meilleur que la prévision hors échantillon. L'excès d'optimisme causé par le choix du meilleur modèle parmi les nombreux ajustés et le fait de considérer le modèle sélectionné comme le vrai, a été largement exposé dans les travaux sur les modèles en milieu incertain (Chatfield, 1995a).

DISCUSSION

Une autre étude de cas utilisant des données de compagnies aériennes de Box *et al.*, (1994, Séries G) a été réalisée par Faraway et Chatfield (1995). Grosso modo, les résultats concernant les difficultés d'ajustement des modèles RN ont été qualitativement similaires.

Nous ne prétendons pas que l'analyse d'une ou deux séries suffise à comparer les capacités prédic-

tives des modèles RN et des modèles de Box-Jenkins. Les données des ventes sont par exemple relativement différentes des séries temporelles utilisées dans la compétition Santa Fe qui étaient beaucoup plus longues (plusieurs milliers d'observations) et dans plusieurs cas manifestement non linéaires. Ici les modèles RN semblent généralement plus appropriés, cependant pour une série temporelle financière (données de taux de change), il existe une «différence cruciale entre la formation et la performance des séries de test» (Gershenfeld et Weigend, 1994, p. 40). Les prévisions hors échantillon, tirés du modèle RN ajusté, n'étaient pas meilleures que celles obtenues suivant une démarche aléatoire. Tout ce que l'on est en mesure de dire concernant la performance prédictive, sur la base de notre expérience, est que les modèles RN ne sont pas clairement meilleurs que les solutions alternatives. En outre, pour des séries aussi courtes que celles des données de Chatfield-Prothero, ils peuvent être réellement inférieurs. Il est bien sûr décevant d'avoir à faire état de résultats plutôt négatifs mais, il est essentiel pour le progrès scientifique d'accepter de publier des résultats aussi bien positifs que négatifs (Chatfield, 1995b).

Une critique possible à adresser aux modèles RN est que, même s'ils fournissent des prévisions satisfaisantes, ils ne donnent qu'un faible aperçu de la structure même des données, il peut donc être difficile d'interpréter un modèle RN. Lorsqu'il y a seulement deux entrées, il est possible de dessiner la surface de prévision comme dans la figure 4 de Faraway et Chatfield (1995), mais lorsqu'il y a plus de deux entrées il est alors plus difficile « d'analyser l'intérieur » d'un modèle RN. D'autres techniques pour explorer les données, tels que l'utilisation de modèles additifs généralisés, sont traités par Faraway et Chatfield (1995).

Nous sommes conscients que les modèles RN peuvent être élaborés de façons variées, en permettant par exemple des connexions en saut de couche, l'addition de couches cachées, l'affaiblissement des poids, et des rétro-connexions telles que dans les RN récurrents, mais le prix d'une plus grande souplesse est d'augmenter la probabilité de se dévoyer. Avec des séries aussi courtes, nous pensons qu'il y a déjà suffisament de choix à effectuer.

Pour conclure, nous pensons que notre étude de cas nous permet d'établir les remarques générales suivantes :

1) Il existe de multiples raisons de se tromper avec les modèles neuronaux (comme pour d'autres techniques statistiques sophistiquées d'ailleurs) et il peut être tout particulièrement dangereux d'adopter une approche « boîte noire ». Grand soin doit être porté au choix (1) d'une série appropriée de variables d'entrées, (2) d'une architecture appropriée, (3) de fonctions d'activation appropriées, (4) d'une procédure numérique appropriée pour l'ajustement d'un modèle RN. En particulier, il est nécessaire de choisir des poids de départ judicieux pour l'algorithme de formation et il peut être nécessaire de réétalonner les données en premier lieu.

2) Ajouter des unités cachées augmente le nombre de paramètres dans un modèle RN. Ceci peut conduire à une amélioration de l'ajustement mais également à une détérioration des prévisions hors échantillon.

3) En comparant des modèles avec des nombres de paramètres différents, l'utilisation du critère AIC n'est pas suffisant pour pénaliser l'addition de paramètres suppmémentaires. Aussi, le critère BIC est nécessaire pour s'assurer que l'amélioration de l'ajustement à l'intérieur de l'échantillon ne se fait pas aux dépends de prévisions hors échantillon moins exactes.

BIBLIOGRAPHIE

Box G.E.P. et Jenkins G.M. (1973), Some Comments on a Paper by Chatfield and Prothero and on a Review by Kendall, *Journal of the Royal Statistical Society*, 136, 337-345.

Box G.E.P., Jenkins G.M. et Reinsel G.C. (1994), *Time Series Analysis, Forecasting and Control*, Englewood Cliffs, NJ : Prentice-Hall, 3e édition.

Chatfield C. (1993), Neural Networks : Forecasting Breakthrough or Passing Fad ?, *International Journal of Forecasting*, 9, 1-3.

Chatfield C. (1995a), Model Uncertainty, Data Mining and Statistical Inference (with discussion), *Journal of the Royal Statistical Society*, 158, 419-466.

Chatfield C. (1995b), Editorial : Positive or Negative ? *International Journal of Forecasting*, 11, à paraître.

Chatfield C. (1996), *Time Series Analysis*, Londres, Chapman & Hall, 5e édition.

Chatfield C. et Prothero D.L. (1973a), Box-Jenkins Seasonal Forecasting : Problems in a Case Study (with

discussion), *Journal of the Royal Statistical Society*, 136, 295-336.

Chatfield C. et Prothero D.L., (1973b), A Reply to a Paper by Box and Jenkins, *Journal of the Royal Statistical Society*, 136, 345-352.

Chatfield C. et Yar M. (1988), Holt-Winters Forecasting : Some Practical Issues, *The Statistician*, 37, 129-140.

Cheng B. et Titterington M. (1994), Neural Networks : a Review from a Statistical Perspective (with discussion), *Statistical Science*, 9, 2-54.

Faraway J. et Chatfield C. (1995), Time Series Forecasting with Neural Networks : A Case Study, University of Bath Statistics Group Research Report n° 95 :06.

Gershenfeld N.A. et Weigend A.S. (1994), The Future of Time Series : Learning and Understanding, in *Time Series Prediction*, A.S. Weigend and N.A. Gershenfeld Eds., Reading, MA : Addison-Wesley, 1-70.

Gorr W.L. (1994), Editorial : Research Prospective on Neural Network Forecasting, *International Journal of Forecasting*, 10, 1-4.

Gorr W.L., Nagin D. et Szczypula J. (1994), Comparative Study of Artificial Neural Network and Statistical Models for Predicting Student Grade Point Averages, *International Journal of Forecasting*, 10, 17-34.

Harvey A. (1993), *Time Series Models*, Hemel Hempstead, UK : Harvester Wheatsheaf, 2e édition.

Hertz J., Krogh A. et Palmer R. (1991), *Introduction to the Theory of Neural Computation*, Redwood City, CA : Addison Wesley.

Hill T., Marquez L., O'Connor M. et Remus W. (1994), Artificial Neural Networks Models for Forecasting and Decision Making, *International Journal of Forecasting*, 10, 5-15.

Hill T., O'Connor M. et Remus W. (1996), Neural Network Models for Time Series Forecasts, *Management Science*, à paraître.

Hoptroff R. (1993), The Principles and Practice of Time Series Forecasting and Business Modelling Using Neural Nets, *Neural Computing and Applications*, 1, 59-66.

Priestley M. B. (1981), *Spectral Analysis and Time Series*, London, Academic Press.

Ripley B. (1993), Statistical Aspects of Neural Networks, in *Chaos and networks – Statistical and Probabilistic Aspects*, O. Barndorff-Nielsen, J. Jensen and W. Kendall Eds., London : Chapman & Hall, 40-123.

Ripley B.D. (1995), Statistical Ideas for Selceting Network Architectures, in *Neural Networks : Artificial Intelligence and Industrial Applications*, B. Kappen and S. Gielen Eds., Berlin, Springer, 183-190.

Ripley B.D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, Cambridge University Press.

Tang Z., de Almeida C. et Fishwick P.A. (1991), Time Series Forecasting using Neural Networks versus Box-Jenkins Methodology, *Simulation*, 57, 303-310.

Venables W.N. et Ripley B.D. (1994), *Modern Applied Statistics with S-Plus*, New-York, Springer-Verlag.