# Prévision de séries temporelles industrielles

*Antonio Piras, Alain Germond, Thomas Czernichow et Antonio Munoz*

*La modélisation d'une série temporelle peut se décomposer en un certain nombre de grandes étapes fondamentales. D'abord l'analyse de la série, puis la sélection des variables explicatives, le choix de la structure du modèle, son estimation et enfin sa validation. Nous allons dans cet article présenter ce qu'implique chacune de ces parties pour effectuer la prévision de séries temporelles en milieu industriel lorsque l'on choisit d'utiliser des réseaux de neurones artificiels.*

## 1 Introduction

La prévision de séries temporelles en milieu industriel a ceci de particulier qu'elle s'inscrit dans un processus d'optimisation des ressources de la production ou de la distribution (par exemple d'énergie électrique, Fig.1). Dans tous les cas, elle constitue un maillon important sur lequel repose une partie des décisions prises en aval. Ceci implique de devoir s'assurer du bon fonctionnement des modèles de prévision quoi qu'il puisse arriver. Malgré que les réseaux de neurones artificiels soient devenus des outils de modélisation importants pour la prévision de séries temporelles, il faut constater qu'aux regards du nombre d'applications publiées, il n'existe quasiment aucune méthode concrète permettant de servir de guide à l'identification d'un modèle. Cet article propose une série d'étapes à suivre afin de modéliser une série temporelle, et pour chacune de ces étapes, il décrit un certain nombre d'options qui s'offrent aux ingénieurs. Ces étapes sont d'abord l'analyse et la sélection des variables explicatives, puis le choix et l'estimation du modèle, suivi de l'estimation de l'intervalle de confiance de la prévision, pour finir par la validation du système. Ces étapes forment la structure de l'article et la lecture peut donc être suivie du début jusqu'à la fin.

***Alain Germond*** est ingénieur él. dipl. et Dr ès sciences tech. de l'École polytechnique fédérale de Lausanne (EPFL). Il est actuellement Professeur à l'EPFL et directeur du laboratoire des réseaux d'énergie électrique (LRE http://lrewww.epfl.ch), EPFL-DE-LRE, 1015 Lausanne.

***Antonio Piras***, est ingénieur él. dipl. de l'École polytechnique fédérale de Zurich (ETHZ) et Dr ès sciences tech. de l'EPFL. Il est actuellement assistant au LRE est associé chez Prediction Partner Sàrl (http://www.prediction-partner.ch), société spécialisée en problèmes de prévision, PSE-EPFL, 1015 Lausanne.

***Thomas Czernichow***, est ingénieur inf. et math. dipl. de l'ENSEEIHT et Dr de l'Institut National des Télécommunications, Paris. Il est actuellement associé chez Prediction Partner Sàrl, 67, rue Barrault, Paris.

***Antonio Munoz*** est ingénieur industr. dipl. et Dr ing. industr. de l'Universidad Pontificia Comillas (http://www.iit.upco.es), Madrid. Il est actuellement MER et chargé de cours à l'Instituto de Investigatión Tecnológica, Santa Cruz de Marcenado, 26, 28015 Madrid

## 2 Sélection des variables explicatives

Il est souvent intéressant d'effectuer d'abord une analyse linéaire de la série à l'aide des outils statistiques classiques. Ajuster par exemple un modèle de régression linéaire permet d'une part d'effectuer une comparaison avec de futurs modèles, et d'autre part sert de point de départ pour la sélection des retards de chaque série explicative.

### 2.1 Valeurs étranges et pré-traitement des données

L'identification et le rejet des valeurs aberrantes (« outlier ») est indispensable [Czernichow/Munoz 97]. D'une part elles perturbent l'estimation du modèle en obligeant l'algorithme d'estimation à fournir une sortie de fait incohérente. Ceci peut apparaître par exemple lorsqu'une valeur étrange est démesurément grande. Un algorithme d'estimation au sens des moindres carrés donnera alors un poids encore plus important à l'erreur faite par le modèle en ce point. D'autre part, en phase d'utilisation, une valeur aberrante pourra induire une réponse absurde du système, avec des conséquences parfois dramatiques dans la chaîne de décision.

Le but d'un pré-traitement est soit de transformer les variables originales à travers une normalisation, de façon à pouvoir être utilisée convenablement pendant la phase d'estimation, soit de former une combinaison linéaire ou non-linéaire (éventuellement une codification binaire) des variables originaires pour créer des nouvelles variables. Pour certaines séries temporelles composées d'une suite de valeurs horaires, par exemple la consommation d'électricité, il est souhaitable de normaliser chaque heure indépendamment, vu que la variance est fonction de l'heure.

### 2.2 Sélection de variables

Le but des algorithmes de sélection de variables consiste à identifier les séries explicatives du processus que nous voulons modéliser, et aussi pour chacune de ces séries les retards dans le temps nécessaires. La sélection de variables représente une des difficultés majeures de l'estimation d'un modèle. Il faudra chercher suffisamment de variables afin d'expliquer convenablement le processus à modéliser, mais sans en inclure trop, parce qu'il aurait pour conséquence d'augmenter artificiellement la complexité du modèle en diminuant sa capacité de

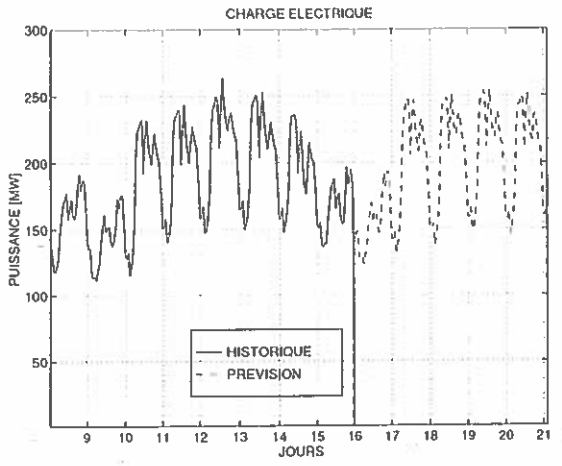


**Fig. 1:** Exemple de prévision de la consommation électrique. Le cycle hebdomadaire est fort visible

généraliser dans les prévisions futures. On distingue trois types de séries explicatives qui impliquent l'utilisation de modèles différents : le passé de la série à prévoir (appelée auto-régressive), le passé des résidus du modèle (appelée moyenne mobile), et le passé, éventuellement le futur, d'autres séries explicatives dites variables exogènes.

Dans le domaine industriel s'ajoute à ceci la difficulté de ne pouvoir éventuellement pas mesurer ou prévoir une variable explicative. Par ailleurs, effectuer l'estimation du modèle avec des valeurs mesurées (de température par exemple), et effectuer ensuite la prévision avec des valeurs prévues, peut avoir des conséquences importantes ou se révéler simplement impossible (par exemple utiliser la température du mois prochain pour effectuer de la prévision à moyen terme de la consommation d'électricité).

### 2.3 L'approche linéaire

Les algorithmes de sélection de variables pour le cas linéaire sont basés sur l'utilisation de tests statistiques mesurant l'augmentation (ou la diminution) d'un indicateur de qualité de prévision lorsqu'un sous-ensemble de variables est introduit (ou supprimé) [Thompson 78].

### 2.4 L'approche neuronale

La plupart des techniques neuronales fonctionnent de la même manière et construisent des indicateurs de qualité de chaque variable d'entrée. Par exemple, l'algorithme du *Optimal Brain Damage* (OBD [Le Cun et al. 90]) est basé sur le calcul des saliences, qui sont une mesure de l'influence du poids (un paramètre du modèle) sur l'erreur d'estimation ou de prévision. La salience peut être utilisée pour effectuer la sélection de variables, en considérant comme critère la somme des saliences provenant de l'entrée d'un modèle. Par ailleurs, dans le cas des modèles récurrents (qui apparaît dès que l'on modélise des processus ayant une composante MA), la salience peut être très compliquée à calculer. Néanmoins, comme l'OBD, il fonctionne en pratique très bien, et donne de bons résultats.

Une autre manière implicite d'effectuer la sélection de variables, est de réduire la dimension du processus par projection non linéaire. L'idée, très proche de l'Analyse en Composantes Principales (ACP), consiste en la projection des données de l'espace d'entrée sur un sous-espace d'information maximum. Le vecteur obtenu est une fonction non linéaire des entrées originelles. Cette ACP non linéaire peut être effectuée en construisant un encodeur-décodeur, constitué d'un réseau, appelé diabolo, ayant une couche cachée de dimension plus petite que la dimension de l'espace d'entrée, et entraîné à reconstituer l'entrée en sortie. Les sorties des neurones cachés sont ensuite utilisées comme entrées d'un deuxième modèle. Cette technique offre une solution élégante au problème de la collinéarité (deux variables, ou plus, d'entrée corrélées), mais permet difficilement d'analyser l'importance de chaque variable d'entrée. Une bonne explication de cette technique peut-être trouvée dans [Fogelmann 94].

Une autre approche se base sur l'analyse des dérivées de la sortie par rapport aux neurones d'une couche particulière. Appliquée à la couche d'entrée, ou aux entrées récurrentes, cette technique permet d'effectuer une sélection des variables [Czernichow/Munoz 95]. Les dérivées du modèle sont le reflet de l'influence que le réseau donne à chaque entrée.

## 3 Modèles et Estimation

### 3.1 Choix du modèle

Le type de réseau que l'on va choisir n'est pas la question la plus importante. En effet, la plupart des modèles utilisés dans la littérature ont des capacités d'approximateurs de fonctions universels. L'optimisation de l'architecture du modèle, l'optimisation de ses paramètres, ont par contre une forte influence dans la qualité des résultats. Chaque type de réseau de neurones possède des propriétés particulières qu'il faudra considérer lors de ces optimisations. Nous renvoyons le lecteur à [Bishop 95] pour une description plus complète des différents modèles existants.

Pour obtenir plusieurs sorties (en général un horizon de plusieurs pas de temps) on peut soit construire plusieurs modèles avec plusieurs entrées et une sortie (MISO) qui est rebouclée dans le système pour prévoir des horizons plus grands, ou utiliser un modèle avec plusieurs entrées et plusieurs sorties (MIMO).

Dans le cas de la prévision de la charge électrique différents types de segmentations ont été expérimentés [Czernichow et al 96]. Certains auteurs ont proposé une séparation en fonction des jours : un modèle spécifique pour les samedis, un modèle pour les dimanches et un modèle pour les jours de semaine. D'autres proposent de faire un modèle pour chaque jour de la semaine. La séparation en sous problèmes journaliers trouve sa justification dans le fait que chaque type de jour de la semaine correspond à un problème spécifique avec ses variables explicatives, et sa dynamique.

La possibilité naturelle de laisser en sortie les 24 heures du jour suivant a été beaucoup testée et implique un grand nombre de paramètres dans le modèle à cause de la taille de la couche de sortie (24 pour 24 points par jour). D'autres auteurs ont vou-

lu effectuer une séparation par saisons, arguant du fait que l'été et l'hiver avaient des dynamiques différentes, et étaient issus d'un processus différent. On trouve aussi des subdivisions mixtes avec une première découpe en un sous-problèmes par saisons (été, hiver et transitions), puis en groupe de trois jours différents à l'intérieur de chaque groupe. On trouvera enfin des modèles différents suivant la tranche horaire à l'intérieur de la journée. Des contraintes météorologiques peuvent aussi nécessiter des modélisations différentes : certains pays ayant des zones à fortes divergences de climat vont devoir utiliser une prévision pour chaque zone géographique.

D'autres techniques dites hybrides ont aussi été expérimentées. Les auteurs de [Yuan/Fine 93] présentent une KSOM (carte de Kohonen) utilisée pour trouver des jours typiques (voir Fig. 2), qui servent à classer la base en sous-groupes, de façon à ensuite estimer différents modèles. Cette idée de séparation automatique, mais à l'aide de modèles hybrides est reprise par [Piras et al.], qui sépare automatiquement l'ensemble en deux groupes, un pour chaque saison. Les deux modèles sont ensuite fusionnés par une sommation floue, permettant une transition continue entre les modèles.

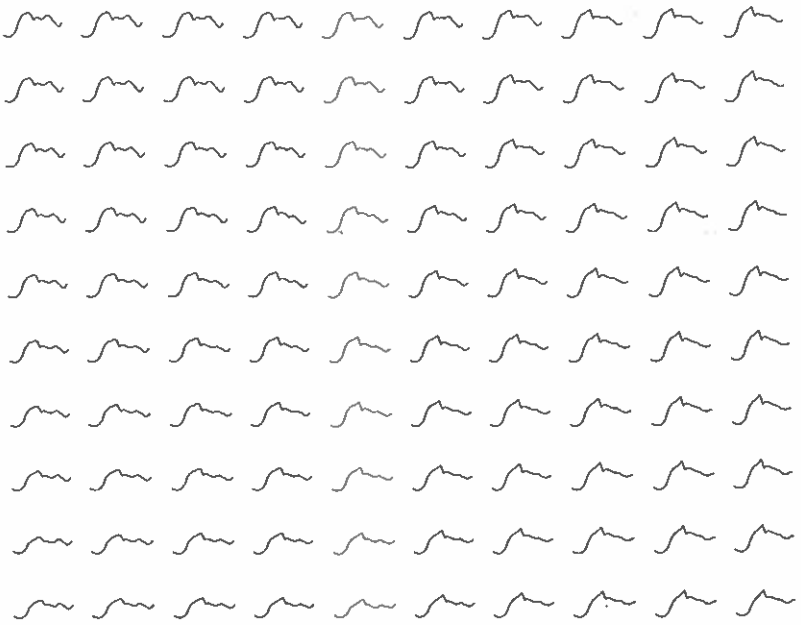

**Fig. 2 :** Exemple de classification de courbe de charge journalière de la consommation électrique. On note que le voisinage topologique entre les courbes est préservé.

Certaines modifications de l'architecture des réseaux ont été proposées sans permettre de vraies conclusions sur leur utilité. Par exemple, on trouvera dans la proposition d'ajouter des connexions linéaires entre l'entrée et la sortie et l'utilisation de réseaux non entièrement connectés. Cela permet de construire des sous-blocs dédiés. On espère ainsi permettre la prise en compte de plusieurs types de corrélations. Une approche novatrice, propose de trouver l'architecture par algorithme génétique. Elle démontre à quel point l'architecture du réseau, en dehors de toute considération d'apprentissage, est fondamentale pour le problème.

De cette foison de modèles et de structures, on retiendra qu'il est indispensable d'adapter les modèles aux problèmes. La plupart des compétitions de prévision montrent que les meilleurs résultats sont obtenus par les équipes qui connaissent bien les problèmes et les données, plutôt que par celles qui connaissent bien les modèles. La décomposition en sous-problèmes, en affectant un modèle différent à chacun, a le défaut de diminuer d'autant les bases de données de chaque sous-partie. Les capacités de généralisation des réseaux de neurones étant en partie proportionnelles à la taille de ces bases, il est indispensable que le découpage effectué ait une justification en terme de dynamique, de mode de fonctionnement.

### 3.2 Optimisation de sa structure

Une fois choisi le type de réseau il reste à décider de son architecture, et notamment du nombre de connexions. Les réseaux de neurones possèdent en générale une capacité d'approximation « trop » grande. Il faut donc souvent les « brider » en diminuant le nombre de connexions utilisées. Ceci peut se faire en utilisant la théorie de la régularisation [Bishop 95], ou la validation croisée, ou encore pouvoir mesurer la qualité d'un modèle pour un problème donné, et prendre le meilleur.

Le travail de Akaike [Akaike 74] s'est dirigé vers la sélection de modèles linéaires, avec un critère permettant de décider si un modèle est meilleur qu'un autre. D'autres indicateurs existent, et prennent aussi en compte le nombre de paramètres, la taille de la base d'apprentissage, et la variance estimé du bruit de la série. Ces critères servent à estimer les capacités de généralisation du modèle, i.e. l'estimation de l'erreur du modèle sur une base nouvelle. Il a été étendu aux modèles neuronaux assez récemment [Moody 94].

### 3.3 Optimisation de ses paramètres

Mis à part cette séparation en architectures, une distinction fondamentale vient du type d'algorithme d'apprentissage utilisé pour l'estimation des poids : supervisé ou non supervisé. Dans le cas supervisé, l'algorithme utilise des paires contenant une entrée et la sortie désirée correspondante du réseau. Le but étant d'ajuster les poids de façon à diminuer l'écart global, pour tous les éléments, entre la sortie désirée et la sortie du réseau répondant au stimuli de l'entrée correspondante. Cela est généralement fait à l'aide d'une fonction qui mesure l'adéquation du réseau à la tâche, appelée fonction de coût.

Il existe un nombre considérable de méthodes d'optimisation, et la descente de gradient n' en est qu'un exemple. Malgré cela, le calcul de ce gradient occupe la très grande majorité des outils d'optimisation neuronale. Ceci pour des raisons de facilité d'utilisation et d'efficacité pratique. Parfois même, le gradient n'est pas une bonne direction, et il est préférable d'utiliser des méthodes de second ordre, utilisant la matrice Hessienne, afin de modifier la direction. Il existe par ailleurs des techni-

ques dites quasi-newton, plus rapides, calculant une estimation de ce Hessien. Le calcul du gradient se complique considérablement si l'on utilise des modèles récurrents (comme c'est le cas lorsque l'on utilise la série des résidus comme variable explicative). De même, pour chaque architecture il faudra disposer d'une procédure permettant de le calculer.

### 3.4 Batch contre Stochastique

Les algorithmes dits batch, attendent de calculer tous les gradients de la base d'apprentissage avant de modifier les poids avec la somme, ou la moyenne, des gradients. Les algorithmes stochastiques tirent des exemples au hasard, dans la base d'apprentissage, en effectuant à chaque fois le calcul du gradient, et la modification des poids. On peut aisément imaginer des versions intermédiaires, où l'on va calculer les gradients d'une sous partie de la base d'apprentissage avant de modifier les poids. En pratique, si l'on n'utilise pas de technique de second ordre, les algorithmes dits stochastiques fonctionnent mieux. Par ailleurs, si la série n'est pas stationnaire, une modification globale des paramètres sur toute la base n'a pas grand sens.

## 4 Validation et intervalle de confiance

Une fois effectué tout le cycle d'estimation il reste bien sûr à valider ces résultats. Une méthode pratique acceptable consiste à garder une base de données, qui n'a jamais servi à la modélisation, et à tester les modèles sur celle-ci. Ceci en assumant que cette base est suffisamment grande et statistiquement représentative. Si les résultats ne sont pas satisfaisants, il faudra envisager de recommencer une ou plusieurs étapes précédentes, en changeant la base de validation afin de ne pas tenter d'améliorer le modèle en fonction de celle-ci.

Pour certains problèmes l'intervalle de confiance de la prévision est parfois plus important que la valeur même de la prévision. Il représente une bonne mesure du risque qu'il y a d'utiliser la prévision. Il permet aussi de générer des scénarios en fonction de l'endroit où l'on se trouve de l'intervalle. Nous pouvons supposer par exemple que les résidus issus du modèle sont Gaussiens. Ainsi, en calculant la moyenne (supposée nulle) et l'écart type des résidus sur la base de validation, nous pouvons inférer sur les bornes des confiance de la prévision. Ceci suppose entre autre chose, que la variance soit constante dans le temps. Si ce n'est pas le cas, il est possible d'utiliser un modèle linéaire ou neuronale [Czernichow/Munoz 97] qui l'estime.

On dispose alors d'un système d'équations permettant d'estimer à la fois l'espérance conditionnelle de la série et sa variance locale. Il faut noter que l'estimation du deuxième modèle pose autant sinon plus de problèmes que le premier. Entre autre, il est bon de refaire une estimation du premier modèle après estimation du second afin d'éliminer les valeurs aberrantes.

## 5 Conclusion

La prévision de séries temporelles est un domaine passionnant de recherche. Pour élaborer des modèles performants, il faut pouvoir comprendre précisément les problèmes dans leurs particularités. On est donc naturellement conduit à l'analyse de domaines très variés comme les séries économiques, financières, industrielles, etc. La prévision de la plupart de ces séries réelles complexes reste encore une tâche améliorable. Les améliorations peuvent être faites à tous les stades de l'estimation du modèle que l'on utilise. Nous avons tenté de montrer en quoi l'application de techniques d'estimation était différente lorsque l'on traitait des séries réelles. Notre expérience nous a permis de mesurer le fossé qui existe entre le modèle théorique et l'utilisation industrielle : Il faut toujours modifier et adapter.

**Références**


[Akaike 74] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol. 19, pp. 716–723, 1974.

[Bishop 95a] C. M. Bishop, Neural Networks for Pattern Recognition: Oxford University Press, ISBN 0 19 853864 2, 1995.

[Bishop 95b] C. M. Bishop, "Chap. 9 Learning and Generalization," in Neural Networks for Pattern Recognition: Oxford University Press, ISBN 0 19 853864 2, 1995, pp. 332–384.

[Czernichow et al. 96] T. Czernichow, A. Piras, K. Imhof, P. Caire, Y. Jaccard, B. Dorizzi, A. Germond, "Short Term Electrical Load Forecasting with Artificial Neural Networks," Int. Journal of Eng. Int. Syst., vol. 4, pp. 85–99, 1996.

[Czernichow/Munoz 95] T. Czernichow, A. Munoz, "Variable Selection through Statistical Sensitivity Analysis: Application to feedforward and recurrent networks.," Institut National des Télécommunications Paris, Dept. Signal et Images, Technical Report 95-07-01, 1995.

[Czernichow/Munoz 97] T. Czernichow, A. Munoz San Roque, "A Probability Estimation Based Criteria For Model Evaluation," in Proceedings of ICANN'97, Lausanne, pp. 1029–1034, 1997.

[Fogelman 94] F. Fogelman Soulie, "Neural Network Architectures for Pattern Recognition," in From Statistics to Neural Networks: Theory and Pattern Recognition Applications., Springer-Verlag, Ed., 1994, pp. 243–262.

[Le Cun et al. 90] Y. Le Cun, J. S. Denker, S. A. Solla, "Optimal Brain Damage," in Proceedings of NIPS, pp. 598–605, 1990.

[Moody 94] J. Moody, "Prediction Risk and Architecture Selection for Neural Networks," in From Statistics to Neural Networks: Theory and Pattern Recognition Applications., Springer-Verlag, Ed., 1994.

[Piras et al.] A. Piras, A. Germond, B.Buchenel, K.Imhof, Y.Jaccard, "Heterogeneous Artificial Neural Networks for Short Term Load Forecasting," IEEE Transactions on Power Systems, Vol.11, No.1, pp.397–402.

[Thompson 78] M. L. Thompson, "Selection of Variables in Multiple Regression. Part I: A Review and Evaluation," Int Stat. Review, vol. 46, pp. 1–19, 1978.

[Yuan/Fine 93] J. L. Yuan, T.L. Fine, "Forecasting Demand for Electric Power," in Proceedings of NIPS-5, pp. 739-746, 1993.