DETECTION DE FUITES DANS UN RESEAU DE DISTRIBUTION D'EAU POTABLE PAR COMPARAISON ENTRE LES CONSOMMATIONS PREVUES ET MESUREES

T. Denoeux*, R. Sobral*, J.F. Depierre**

* Lyonnaise des Eaux - Dumez
Laboratoire d'Informatique Avancée de Compiègne
Technopolis, rue du Fonds Pernant
60200 Compiègne

** Société Parisienne des Eaux
11 Bld Brune
75014 Paris

Introduction

Gérer le mieux possible les ressources en eau, c'est-à-dire répondre à l'ensemble des besoins au moindre coût, est désormais un objectif prioritaire, tant pour les collectivités locales que pour les industriels. L'un des éléments de cette gestion consiste à accroître le rendement des réseaux de distribution, en intervenant au plus vite lorsqu'un défaut d'étanchéité a été constaté. Dans le cas de la Ville de Paris, le problème est rendu particulièrement difficile par le fait que, les canalisations étant placées en égout, les fuites rejoignent les effluents d'eaux usées, et ne sont donc pas détectables tant qu'elles n'atteignent pas des proportions importantes.

Une expérience de détection de fuites a été menée conjointement par la Société Parisienne des Eaux et le Laboratoire d'Informatique Avancée de Compiègne, avec pour objectif d'évaluer la possibilité de déceler rapidement des fuites moyennes à importantes, en comparant les débits mesurés avec des débits prévus. L'intérêt de cette approche dépend évidemment de la fiabilité de la prévision, qui conditionne l'importance des fuites susceptibles d'être détectées.

La méthodologie qui a été mise en oeuvre repose sur la modélisation de séries chronologiques par réseaux de neurones, technique qui a récemment été appliquée avec succès au problème de la prévision de consommation d'eau journalière [1] [2]. Après une présentation générale de cette technique, nous montrerons l'application qui en a été faite dans le cadre de cette étude. Les résultats obtenus seront ensuite présentés, et situés par rapport aux performances de quelques méthodes de référence. Finalement, l'intérêt de cette approche en réponse au problème posé sera discuté.

L'utilisation des réseaux de neurones en prévision

La prévision de séries chronologiques est un problème relativement ancien, qui est classiquement traité par la méthode de Box et Jenkins [3]. Cette méthode consiste à proposer, à partir d'une analyse des données, un modèle stochastique de type ARIMA (autoregressive integrated moving average model):

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)\,(1 - B^d)\,Z_t = (1 - \theta_1 B - \ldots - \theta_q B^q)\,A_t$$

où: Z_t est la valeur observée à l'instant t

A_t est l'observation d'un bruit blanc à l'instant t

B est un opérateur de retard défini par $B\,Z_t = Z_{t-1}$

$B^d\,Z_t = Z_{t-d}$

$(\phi_i)_{i=1,p}$ et $(\theta_i)_{i=1,q}$ sont des paramètres du modèle

Les paramètres de ce modèle sont ensuite estimés par la méthode du maximum de vraisemblance ou des moindres carrés. Le modèle peut ensuite être complété pour prendre en compte des variables exogènes (cf. [4] pour une application de cette méthode à la prévision de la consommation d'eau).

Très récemment, une approche radicalement différente a été proposée, consistant à utiliser les propriétés d'apprentissage des réseaux de neurones multicouches [5]. Un tel réseau consiste en un ensemble de neurones formels interconnectés. Un neurone formel est un automate qui, en fonction de ses entrées $(i_k)_{k=1,n}$ se place dans un état caractérisé par une activation a:

$$a = f(\sum_k W_k.i_k)$$

f étant une fonction d'activation, et $(W_k)_{k=1,n}$ un ensemble de coefficients appelés poids synaptiques. La sortie de l'automate est une fonction g de son activation (g étant généralement prise égale à la fonction d'identité).

Parmi les différentes architectures possibles, la plus étudiée repose sur une organisation en couches: chaque neurone d'une couche c est connecté à chaque neurone de la couche suivante c+1, mais à l'intérieur d'une même couche les neurones ne sont pas connectés entre eux. Entre la couche d'entrée et celle de sortie sont situés un certain nombre de couches cachées, qui ne communiquent pas directement avec l'environnement du système. Un algorithme d'apprentissage pour un tel réseau a été proposé récemment sous le nom d'algorithme de rétropropagation du gradient [5]. Cet algorithme permet de calculer les poids du réseaux de manière à ce que, des entrées étant présentées au système, les sorties obtenues soient les plus proches possibles (au sens des moindres carrés) des sorties désirées. Des travaux théoriques récents suggèrent qu'il est possible, par ce moyen, d'approximer avec une précision

quelconque n'importe quelle fonction de $\mathbb{R}^p$ dans $\mathbb{R}^q$ susceptible d'être rencontrée dans la pratique, à condition de mettre suffisamment de neurones dans une seule couche cachée [6]. Il est donc possible de modéliser la fonction de transfert d'un système recevant en entrée les consommations passées, et des informations météorologiques, et dont la sortie est la consommation du jour suivant.

Les avantages attendus de cette approche par rapport à la méthode de Box et Jenkins sont [1]:

(1) une mise en oeuvre plus facile, notamment en ce qui concerne l'introduction de variables exogènes (par exemple la pluie et la température dans le cas de la consommation d'eau potable)

(2) une meilleure prise en compte des non-linéarités (par exemple entre la pluie et la consommation d'eau).

Application au problème posé

Le problème posé consiste à prévoir la consommation d'eau pour les prochaines 24 heures, avec une résolution de 15 minutes, en utilisant les consommations antérieures et, au besoin, des informations météorologiques. Nous disposions pour mettre au point la méthode de 6 mois de données environ (entre juillet 1989 et janvier 1990), avec des interruptions liées à des problèmes de fiabilité des capteurs et du système de transmission (fig. 1).

Plusieurs approches reposant sur les techniques évoquées précédemment étaient a priori envisageables, la première étant la construction d'un réseau recevant, en entrée:

- les consommations, par quarts d'heures, pour les p jours précédents,
- les valeurs de pluie et de températures pour les q jours précédents,

et ayant 96 neurones de sortie correspondant aux 96 valeurs de consommation à prévoir. Avec une couche cachée de r neurones, un tel réseau aurait donc au minimum: $(96p+2q + 2)r$ connexions, se qui représente, pour $p=q=2$ et $r=5$, 985 paramètres à ajuster. Cette approche n'est pas viable, pour deux raisons:

(1) le quantité relativement faible de données disponibles ne permet pas de caler un modèle possédant autant de degrés de liberté;

(2) dans la l'hypothèse même d'un nombre de données suffisant, le temps d'apprentissage d'un tel réseau serait pratiquement prohibitif (l'algorithme de rétropropagation du gradient est basé sur une optimisation par descente de gradient, technique typiquement assez lourde en temps de calcul).

La seconde approche qui a été envisagée consistait à mettre en oeuvre un réseau effectuant une prévision de consommation pour le quart d'heure suivant uniquement, à partir des consommations observées au cours des p quarts d'heures précédents, et des informations météorologiques. La prévision aurait ainsi été réitérée 96 fois, de manière à obtenir les 96 valeurs de consommations à prévoir. Cette méthode ayant été testée, nous avons observé une

dégradation rapide des résultats, due au cumul des erreurs (pour p=96, la 96-ième valeur prévue n'est plus obtenue qu'à partir d'une seule valeur mesurée, et de 95 valeurs prévues). Cette méthode a donc également été écartée.

L'approche qui a finalement été retenue s'appuie sur l'invariance de la forme des courbes journalières de consommation, pour un jour de la semaine donnée. Cette propriété peut être mise en évidence en examinant la dispersion autour de la moyenne des courbes journalières lissées (par un filtre médiane d'ordre 2), centrées et réduites. Trois formes bien distinctes sont en fait discernables, correspondant aux jours de semaine, au samedi et au dimanche (fig. 2). Cette remarque permet de ramener le problème initial de la prévision de 96 valeurs de consommations sur 15 minutes, à un problème plus simple de prévision de 2 coefficients de la courbe de consommation (moyenne et écart-type). Une variante, qui a également été étudiée, consiste à normaliser les courbes journalières par le minimum et le maximum de consommation sur la journée (ce qui a pour effet de ramener les valeurs entre 0 et 1), et à faire une prévision de minimum et de maximum. Cette variante est intéressante car, en effectuant la prévision pendant le creux de consommation (vers 4 heures du matin), on dispose déjà d'une très bonne estimation du minimum.

Les réseaux considérés pour la prévision de la moyenne, de l'écart-type et du maximum se composent d'une couche d'entrée de 16 neurones, d'une couche cachée de 5 neurones, et d'un neurone de sortie. Les 16 entrées du système sont, dans le cas de la prévision de la moyenne:

- les consommations moyennes $\mu(j-8)$, $\mu(j-7)$, $\mu(j-1)$ mesurées aux jours j-8, j-7 et j-1 (j étant le jour pour lequel on établit la prévision),
- les températures atmosphériques moyennes T(j-3), T(j-2), T(j-1) ,
- les précipitations H(j-3), H(j-2), H(j-1) ,
- le jour de la semaine codé par 6 variables booléennes.

L'activation du neurone de sortie est interprétée comme la consommation moyenne $\hat{\mu}(j)$ prévue au jour j.

L'apprentissage des poids dans ces réseaux a été réalisé par l'algorithme de rétropropagation du gradient à pas variable [7], avec introduction dans la fonction objectif d'un terme de décroissance des poids [8].

Une fois effectuées les prévisions de moyenne [$\hat{\mu}(j)$], d'écart-type [$\hat{\sigma}(j)$] et de maximum [$\hat{M}(j)$], le vecteur $\hat{C}(j)$ des consommations prévues au jour j est obtenu de 3 façons différentes:

méthode 1: $\hat{C}_1(j) = \hat{\sigma}(j)\, N(j-1) + \hat{\mu}(j)$ si j est un mardi, mercredi, jeudi ou vendredi

$\hat{C}_1(j) = \hat{\sigma}(j)\, N(j-7) + \hat{\mu}(j)$ sinon

(N(j-n) étant le vecteur des valeurs de consommation lissées, centrées et réduites, au jour j-n)

méthode 2: $\hat{C}_2(j) = [\hat{M}(j) - \hat{m}(j)]\ N'(j-1) + \hat{m}(j)$ si j est un mardi, mercredi, jeudi ou vendredi

$\hat{C}_2(j) = [\hat{M}(j) - \hat{m}(j)]\ N'(j-7) + \hat{m}(j)$ sinon

(N'(j-n) étant le vecteur des valeurs de consommation, lissées et ramenées entre 0 et 1, au jour j-n, et $\hat{m}(j)$ le minimum des valeurs de consommation estimé pour le jour j, obtenu en prenant la 96-ième valeur de consommation mesurée au jour j-1)

méthode 3: $\hat{C}_3(j) = \alpha * \hat{C}_2(j) + (1 - \alpha) * \hat{C}_1(j)$

(α étant un vecteur de k-ième composante $\left(\frac{k-1}{95}\right)^2$ et * l'opérateur de multiplication matricielle terme à terme)

La méthode 3 revient à combiner les prévisions obtenues par les méthodes 1 et 2, en donnant davantage de poids à la prévision 2 en début de journée, et à la prévision 1 en fin de journée. L'intérêt d'une combinaison des deux méthodes est apparue en observant les résultats, les prévisions obtenues par la méthode 2 étant généralement meilleures en début de journée (à cause de la très bonne estimation du minimum), et moins bonnes en fin de journée. La pondération indiquée ici est la meilleure obtenue expérimentalement.

Enfin, l'intérêt d'un réajustement de la prévision en cours de journée a également été étudié. L'idée consiste à comparer les moyennes des consommations mesurées et prévues au cours des 2 dernières heures, et à utiliser cette information pour recaler la prévision pour les 2 heures à venir, les autres informations prises en compte étant le jour de la semaine, et la situation dans la journée. L'apprentissage du meilleur recalage a également été effectué par réseau de neurone. Notons que la mise en oeuvre, en mode opérationnel, d'une telle procédure de recalage n'est pas sans danger, car on risque ainsi de s'adapter à une modification brutale des profils de consommation, liée, par exemple, à une fuite. Il faut donc n'effectuer le recalage que si l'écart entre la prévision et la mesure au cours des deux dernières heures n'est pas trop important, ou vérifier a posteriori (toutes les 24 heures par exemple) qu'à force de dériver progressivement, on n'aboutit pas à une situation trop atypique pour être expliquée par des facteurs calendaires ou météorologiques.

Résultats

Les données disponibles ont été réparties en deux groupes, l'un pour l'apprentissage (juillet, première quinzaine d'août, décembre et janvier), l'autre pour le test (deuxième quinzaine d'août, septembre, octobre, première quinzaine de novembre).

Les résultats obtenus pour chacune des 4 méthodes décrites ci-dessus (méthodes 1, 2, 3 et 3 avec réajustement toutes les 2 heures) sont indiqués dans le tableau 1, et comparés par rapport à 7 méthodes de référence:

référence 1: $\hat{C}(j) = C(j-1)$

référence 2: $\hat{C}(j) = C(j-1)$ si j est un mardi, mercredi, jeudi ou vendredi
$\hat{C}(j) = C(j-7)$ sinon

référence 3: $\hat{C}(j) = C(j-1)$ si j est un mardi, mercredi, jeudi, vendredi

$$= \left(N(j-7) \cdot \frac{\sigma(j-7) \cdot \sigma(j-1)}{\sigma(j-8)}\right) + \frac{\mu(j-7) \cdot \mu(j-1)}{\mu(j-8)} \quad \text{sinon}$$

référence 4: $\hat{C}(j) = C(j-1)$ si j est un mardi, mercredi, jeudi, vendredi

$= N'(j-7) .[(M(j-7)+M(j-1)-M(j-8)) - (m(j-7)+m(j-1)-m(j-8))]$
$+ M(j-7)+M(j-1)-M(j-8)$ sinon

référence 5: $\hat{C}(j) = \hat{\sigma}_r(j)\, N(j-1) + \hat{\mu}_r(j)$ si j est un mardi, mercredi, jeudi ou vendredi
$\hat{C}(j) = \hat{\sigma}_r(j)\, N(j-7) + \hat{\mu}_r(j)$ sinon
($\hat{\sigma}_r(j)$ et $\hat{\mu}_r(j)$ étant resp. l'écart-type et la moyenne des consommations pour le jour j, estimés par régression linéaire)

référence 6: $\hat{C}(j) = [\hat{M}_r(j) - \hat{m}(j)]\, N'(j-1) + \hat{m}(j)$ si j est un mardi, mercredi, jeudi ou vendredi
$\hat{C}(j) = [\hat{M}_r(j) - \hat{m}(j)]\, N'(j-7) + \hat{m}(j)$ sinon
($\hat{M}_r(j)$ étant le maximum des consommations pour le jour j, estimé par régression linéaire)

référence 7: $\hat{C}(j) = \sigma(j)\, N(j-1) + \mu(j)$ si j est un mardi, mercredi, jeudi ou vendredi
$\hat{C}(j) = \sigma(j)\, N(j-7) + \mu(j)$ sinon

Les méthodes de référence 1 à 4 sont des méthodes triviales. Les méthodes de référence 5 et 6 permettent de situer les résultats par rapport à ce que l'on obtient par régression linéaire, ou, de manière équivalente, avec un réseau de neurones sans couche cachée. La référence 7 indique, en quelque sorte, les limites de la méthodologie adoptée: elle correspond au cas hypothétique ou l'écart-type et la moyenne des consommations pour le jour j serait connues exactement au jour j-1.

Les 11 méthodes sont évaluées dans le tableau 1 selon 7 indices de performance, définis de la façon suivante. Appelons $\varepsilon_{i,j}$ l'erreur commise le jour j, au quart d'heure i, et p le nombre de jours utilisés pour le test (soit 31). Nous avons:

$$\text{moy} = \frac{1}{96p}\sum_{i,j}|\varepsilon_{i,j}|$$

$$\max(\text{moy}) = \max_j\left(\frac{1}{96}\sum_{i}|\varepsilon_{i,j}|\right)$$

n(k,S) = nombre de jours où, pendant au moins k quarts d'heures consécutifs, l'erreur a été supérieure à S (en valeur absolue), tout en conservant le même signe.

Il apparaît, à l'examen du tableau 1, que:

- la méthode 1 a des performances supérieures ou égales à celles des 6 premières méthodes de référence, selon tous les critères;
- la méthode 2 seule ne donne pas de bons résultats, mais, combinée avec la méthode 1, elle en améliore les performances selon 4 critères sur les 7;
- l'introduction du réajustement toutes les 2 heures améliore sensiblement les performances de la méthode 3, qui ne sont plus très éloignées de celles de la méthode de référence 7 (méthode non déterministe).

La supériorité évidente de la méthode 3 sur la stratégie de la persistance (référence 2) est illustrée par l'exemple de la figure 3. La figure 4 représente les courbes de fréquences cumulées des $|\varepsilon_{i,j}|$ pour la méthode 3, et les références 2 et 7. Ces courbes situent, d'une manière graphique, les performances de la meilleure méthode obtenue (sans réajustement), par rapport à une méthode triviale, et aux limites absolues de l'approche choisie.

Enfin, il est toujours souhaitable de disposer, pour toute prévision, d'une indication concernant sa fiabilité [9]. Dans notre cas, celle-ci peut être indiquée en traçant, de part et d'autre de la courbe représentant $\hat{C}(j)$, les courbes représentant $\hat{C}(j) \pm \varepsilon_{80}$, ε_{80} étant le vecteur des seuils d'erreur non dépassés par 80 % des $\varepsilon_{i,j}$, pour les 96 valeurs de i (fig. 5). En mode opérationnel, le fait que les mesures sortent de la bande définie par les deux courbes extrêmes devrait attirer l'attention du personnel responsable. De plus, la largeur de cette bande (environ ± 200 m^3/h) donne une idée de l'ordre de grandeur des fuites susceptibles d'être détectées par cette méthode.

Conclusion

Une méthodologie pour la prévision des courbes journalières de consommation a été décrite. Cette méthodologie est basée sur l'apprentissage, par réseaux de neurones, des relations liant les moyennes, écart-types et maxima de ces consommations aux valeurs mesurées

antérieurement, en fonction du jour de la semaine et de paramètres météorologiques. L'apport de cette méthodologie par rapport à des méthodes simples de référence a été mis en évidence. De plus, il a été démontré qu'une amélioration sensible des résultats peut être obtenue en réajustant la prévision toutes les deux heures, compte-tenu des erreurs passées.

Les résultats issus d'une prévision parfaite de la moyenne et de l'écart-type des consommations sur la journée montrent que les limites de la méthode ne sont pas encore atteintes. L'introduction de données supplémentaires devrait permettre, dans l'avenir, d'améliorer sensiblement les résultats.

Pour l'instant, il apparaît que des fuites de l'ordre de 200 m^3/h (soit environ 10 % du débit moyen) sont susceptibles d'être mises en évidence, de manière pratiquement instantanée, par ce type de méthode qui, notons le, ne nécessite qu'un seul débitmètre par sous-réseau.

Références

[1] S. Canu, R. Sobral, R. Lengellé, Formal neural network as an adaptative model for water demand, Proc. INNC '90, Vol. II, Kluwer Academic Publishers, 1990.

[2] R. Sobral, S. Canu, Application des réseaux de neurones artificiels à la prévision: la consommation d'eau, Actes Neuro-Nîmes '90, EC2, 1990.

[3] G.E.P. Box, G.M. Jenkins, Time series analysis, forecasting and control, Holden-Day, 1976.

[4] J. Quevedo, G. Cembrano, A. Valls, J. Serra, Time series modelling of water demand. A study on short-term and long term predictions, in Computer Applications in Water Supply, Vol 1, B. Coulbeck & C-H Orr Ed., Research Studies Press Ltd, 1987.

[5] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in Parallel Distributed Processing, Vol. I, Rumelhart D.E., McClelland J. and the PDP research group, MIT press, 1986.

[6] R. Hecht-Nielsen, Neurocomputing, Addison-Wesley, 1990.

[7] F.M. Silva, L.B. Almeida, Speeding up back-propagation, in Advanced Neural Computers, Eckmiller R. Ed., North-Holland, 1990.

[8] S.J. Hanson, L.Y. Pratt, Comparing biases for minimal network construction with back-propagation, in Advances in Neural Information Processing Systems 1, Touretzky D.S. Ed., Morgan Kaufmann, 1989.

[9] T. Denoeux, T. Einfalt, G. Jacquet, Determination in real time of the reliability of radar rainfall forecasts, Journal of Hydrology, 122 (1991), 353-371.

Figures et tableau

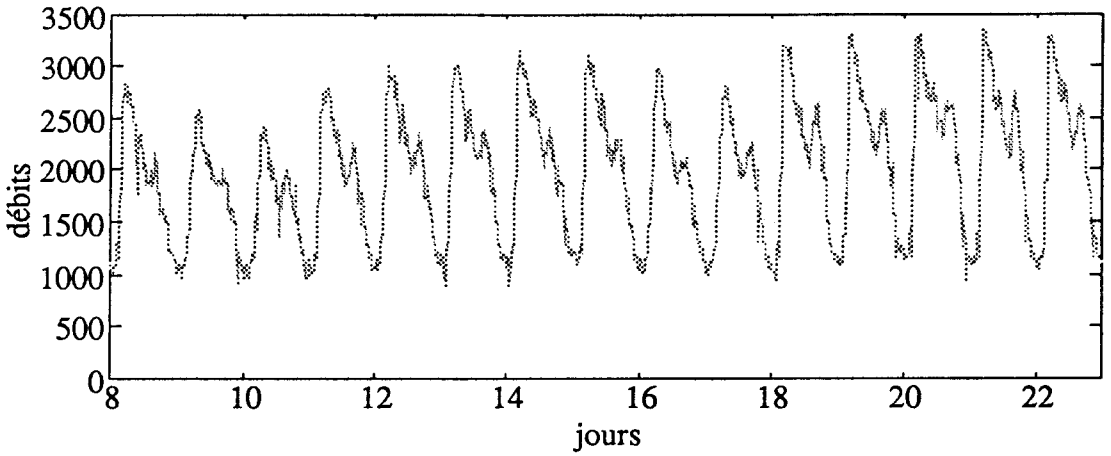

Fig. 1 - Données brutes du 25/08/89 au 08/09/89

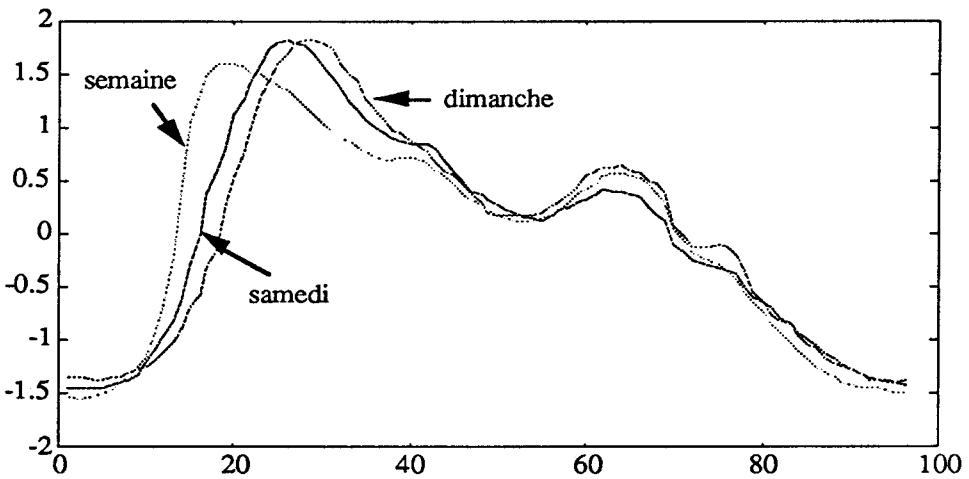

Fig. 2 - Courbes moyennes de consommations, lissées, centrées et réduites

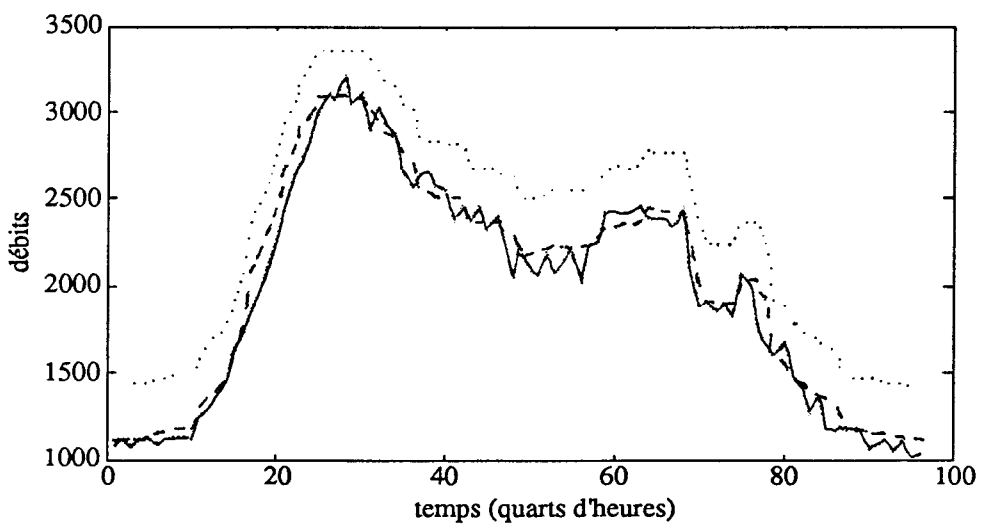

Fig. 3 - Mesures (-) et prévisions (méthode 3: -- / référence 2: ..), le 8/10/89

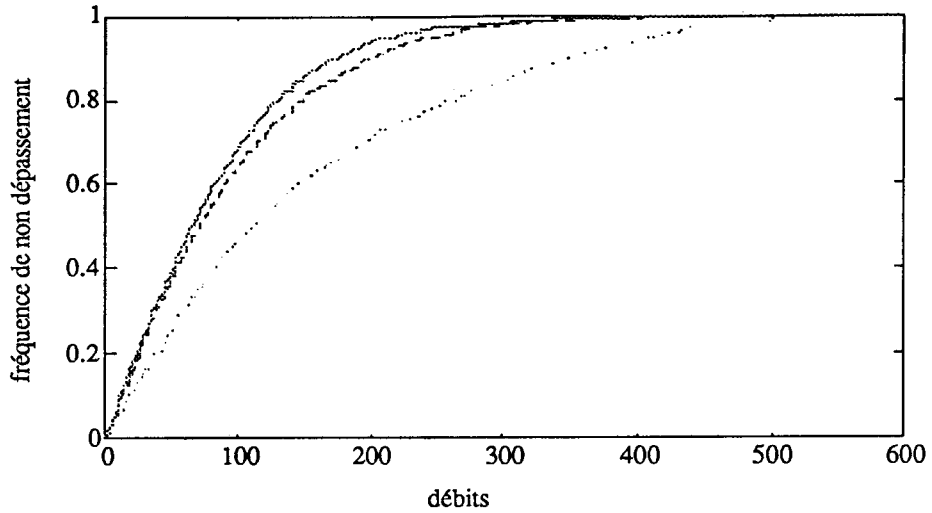

Fig. 4 - Courbes de fréquences cumulées des erreurs, pour la méthode 3 (--), et les références 2 (..) et 7 (-)

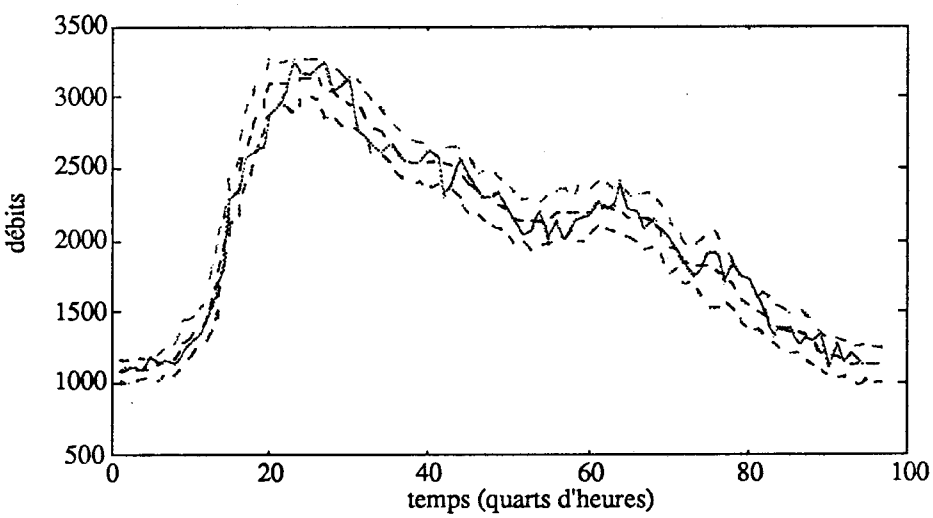

Fig. 5 - Un exemple de prévision, avec intervalle de confiance à 80 %

Tableau 1 - Performances de prévision

Méthode	moy	max(moy)	n(5,200)	n(5,150)	n(5,250)	n(3,200)	n(7,200)
réf. 1	151.7	269.7	22	24	19	27	20
réf. 2	152.8	357.2	13	16	11	23	10
réf. 3	97.9	137.3	7	15	2	16	3
réf. 4	100.6	138.8	5	13	2	17	4
réf. 5	96.3	141.1	5	16	1	15	2
réf. 6	103.8	176.2	9	14	3	17	4
réf. 7	84.5	113.8	1	4	0	6	0
méth. 1	95.7	136.3	5	12	1	14	2
méth. 2	103.6	170.6	8	14	4	19	4
méth. 3	93.7	136.2	4	13	2	12	2
méth. 3 + corr.	86.7	112.1	1	6	0	12	0