COMPARAISON "EXPÉRIMENTALE" DES PRÉVISIONS DE DÉBITS MENSUELS OBTENUS, D'UNE PART AU MOYEN DE MÉTHODES DE RÉGRESSION, D'AUTRE PART EN UTILISANT UN MODÈLE DÉTERMINISTE

par J.-M. MASSON
Centre Hydrométéorologique de Montpellier

Prévoir le débit d'un cours d'eau un certain temps avant qu'il ne se produise est un des objectifs de l'hydrologue. Le but de cette opération est de fournir des éléments de décision pour gérer au mieux les installations hydrauliques.

La prévision se fait au moyen de modèles ajustés sur des observations. Suivant les situations géographiques et le nombre des données historiques que possède l'hydrologue, les possibilités de prévision sont plus ou moins grandes. Cependant, quels que soient les modèles employés, c'est uniquement la précision des résultats obtenus en prévision qui peut mesurer la valeur d'une méthode.

Dans les lignes qui suivent, nous ne pourrons malheureusement pas résoudre, mais simplement poser un certain nombre de problèmes liés à l'emploi de certains modèles pour la prévision. Ces problèmes concernent d'ailleurs aussi bien des méthodes vulgarisées et automatisées comme la régression multiple, que des méthodes plus subjectives comme les modèles déterministes. Nous illustrerons ces problèmes par quelques exemples de prévision sur des débits mensuels.

Les conditions d'une bonne prévision et la définition des termes prévisionnels

Le débit à prévoir dépend pour une part de grandeurs connues au moment de la prévision (*les préviseurs*) et pour une autre part de grandeurs qui se réaliseront entre l'instant où on émet la prévision et la réalisation (*les termes prévisionnels*).

Par exemple le débit moyen du mois prochain peut dépendre pour 60 % du débit du mois actuel qui est alors le préviseur principal et pour 20 % de la pluie du mois prochain, qui est alors le terme prévisionnel principal. Les 20 % non expliqués, dus aux facteurs non pris en compte et à diverses erreurs, constituent ce qu'il est convenu d'appeler *l'erreur*. Ces pourcentages peuvent être déterminés non seulement grâce aux coefficients de corrélations multiples et partiels d'une régression ajustée sur des données observées, mais aussi après ajustement d'un modèle déterministe (modèle à effets retardés par exemple).

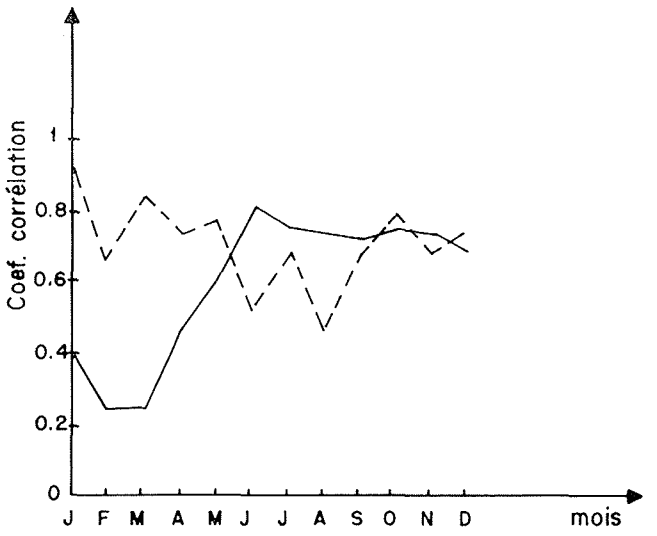
Dans l'état actuel de nos connaissances, c'est-à-dire tant que les prévisionnistes de la Météorologie nationale ne sauront pas émettre des prévisions quantitatives de pluie valables à un mois d'échéance, une prévision sur les débits mensuels est d'autant meilleure que la part des préviseurs dans l'explication de la variabilité des débits est importante.

Ainsi, pour les bassins de haute montagne, les débits des mois d'été, qui sont les plus importants, sont principalement dus à la fonte des neiges. Le stock neigeux est donc le préviseur essentiel à prendre en compte pour la prévision des débits mensuels d'été dans cette situation.

Dans un régime pluvial comme celui du Massif Central, la contribution des termes prévisionnels dans l'explication des débits est souvent importante, avec, pour conséquence, des prévisions peu précises parce que ne concernant qu'un petit pourcentage des débits.

En matière de débits mensuels, un moyen de mettre en évidence les mois où les prévisions seront les meilleures consiste à corréler les débits d'un mois d'une part avec le

Article published by SHF and available at http://www.shf-lhb.org or http://dx.doi.org/10.1051/lhb/1971012



1/ Variation du coefficient de corrélation selon le mois

principal des termes prévisionnels : la pluie du mois, et d'autre part avec le préviseur qui intègre un peu l'influence de tous les autres : le débit du mois précédent ou le débit moyen des dix derniers jours de ce mois.

La comparaison, pour un même mois, des deux coefficients de corrélation, donne rapidement la réponse; quand le terme prévisionnel a plus de poids que le préviseur, les prévisions seront médiocres.

La figure 1 montre pour la Sioule à Pont-du-Bouchet, que cela arrive surtout pendant les mois d'hiver.

Utilisation des régressions c'est-à-dire de modèles linéaires pour la prévision

Nous allons souligner ici quelques problèmes particuliers soulevés par l'utilisation des régressions linéaires comme modèles de prévision, sans nous étendre sur les principes de la régression, car il existe d'excellents manuels sur ce sujet [1].

1. Fluctuations d'échantillonnage.

L'équation de régression est ajustée sur une série limitée d'observations qui ne constitue qu'un échantillon de la population des variables considérées. Cet échantillon est plus ou moins représentatif de la population. Il peut s'agir par exemple d'années particulièrement sèches ou au contraire d'années particulièrement humides.

Les méthodes statistiques permettent de tenir compte des fluctuations d'échantillonnage non seulement pour déterminer les limites de l'espace où il y a α chances sur 100 de trouver les véritables plans de régression, mais aussi pour déterminer l'intervalle de confiance à accorder à une prévision isolée en fonction d'un seuil de probabilité α que l'on se donne.

Mais le calcul des intervalles de confiance fait intervenir le test t de Student ou le test F de Fischer et implique donc qu'un certain nombre de conditions soient vérifiées.

Ainsi les erreurs entre les débits observés et ceux calculés par l'équation de régression doivent être indépendantes : indépendantes de la grandeur des variables utilisées dans la régression (cette indépendance s'appelle homoscédasticité), indépendantes des instruments de mesure utilisés, indépendantes entre elles, etc.).

Avec le nombre limité des observations dont l'hydrologue dispose pour faire ses prévisions, la vérification de toutes ces conditions est difficile sinon impossible. Il peut tout au plus montrer qu'elles ne sont pas systématiquement mises en défaut. Il pourra alors remédier à certaines dépendances par des transformations de variables ou l'adjonction de nouvelles variables (termes quadratiques, produits de deux variables, etc.).

Mais la dépendance des erreurs entre elle est difficile à corriger. Cela n'a pas beaucoup d'importance si le nombre p de préviseurs est petit bien que, théoriquement, les n erreurs (n = nombre d'observation), qui ont :

$$n - (p + 1)$$

degrés de liberté, ne puissent être rigoureusement indépendantes. Mais pour une régression multiple, quand le rapport :

$$(n - p)/p$$

est petit, les conséquences ne sont pas négligeables.

Les hypothèses sur les erreurs (Σ) qui sont faites si on utilise les tests statistiques peuvent être résumées, en écriture matricielle symbolique, par l'expression :

$$\Sigma \sim N(0, Is^2)$$

où I est une matrice unité et s^2 la variance résiduelle.

En conclusion, *parce qu'il est toujours difficile de vérifier rigoureusement les hypothèses faites sur les erreurs, les intervalles de confiance annoncés autour des prévisions faites par les modèles de régression doivent être considérés avec prudence.* Ceci est plus particulièrement vrai quand le nombre de préviseurs est important par rapport au nombre d'observations.

Ainsi sur l'exemple de la figure 2 (Sioule à Pont-du-Bouchet) nous avons dessiné l'intervalle de confiance calculé à partir des 22 observations qui nous ont servi à calculer l'équation de régression (à 9 paramètres) utilisée pour la prévision des débits du mois de juin. C'est un intervalle approximatif extrapolé à partir de quelques points calculés pour des valeurs particulières du vecteur des préviseurs.

Pour 23 prévisions que nous avons faites en nous servant de l'équation de régression, 3 réalisations seulement sont tombées dans l'intervalle de confiance alors que normalement 68 % auraient dû s'y trouver.

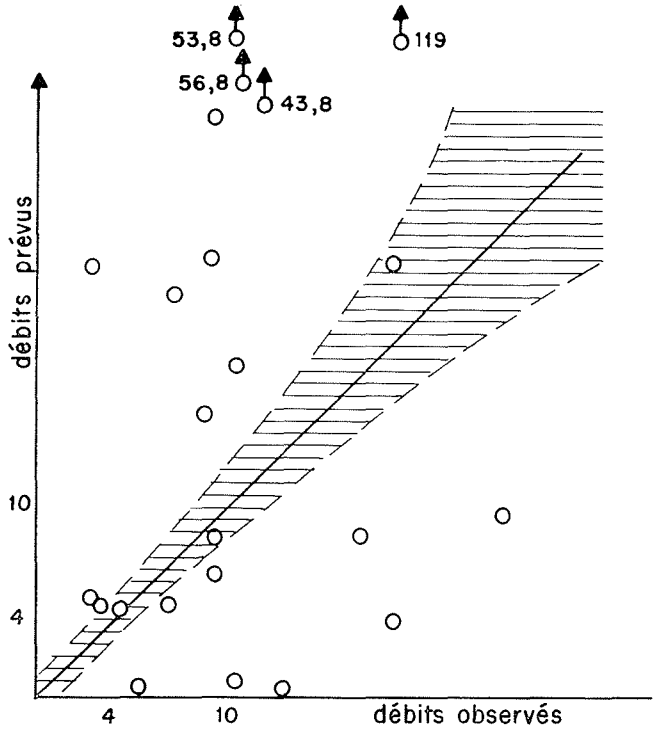
2. Choix du vecteur des préviseurs.

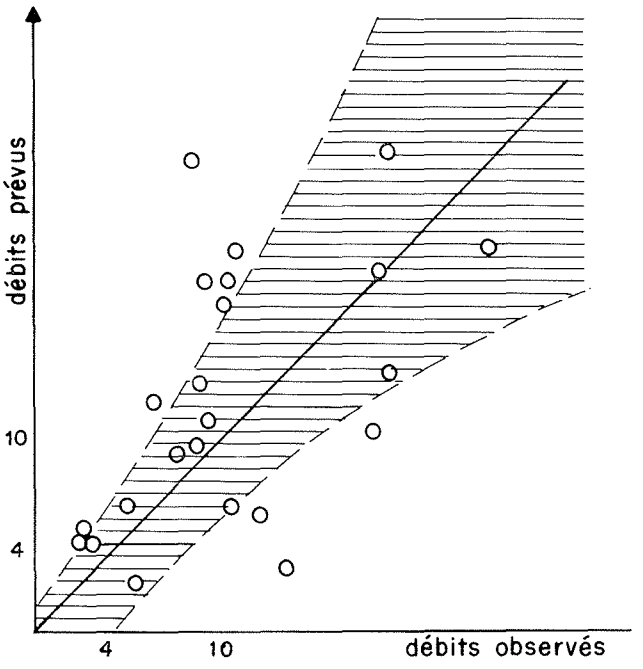
Les variables physiques et climatiques qui ont une influence sur les débits sont nombreuses. Nous allons examiner les méthodes qui nous permettent de décider de faire entrer dans la régression tel ou tel préviseur plutôt que tel autre dont l'influence est négligeable.

Les méthodes classiques font intervenir le test F partiel,

SIOULE A PONT DU BOUCHET

Comparaison de prévisions aux réalisations
pour le débit moyen de juin
régression multiple

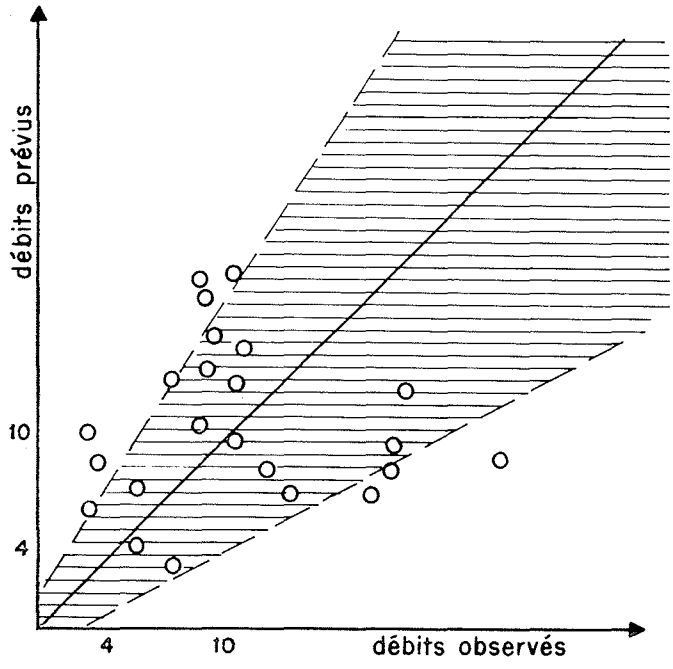


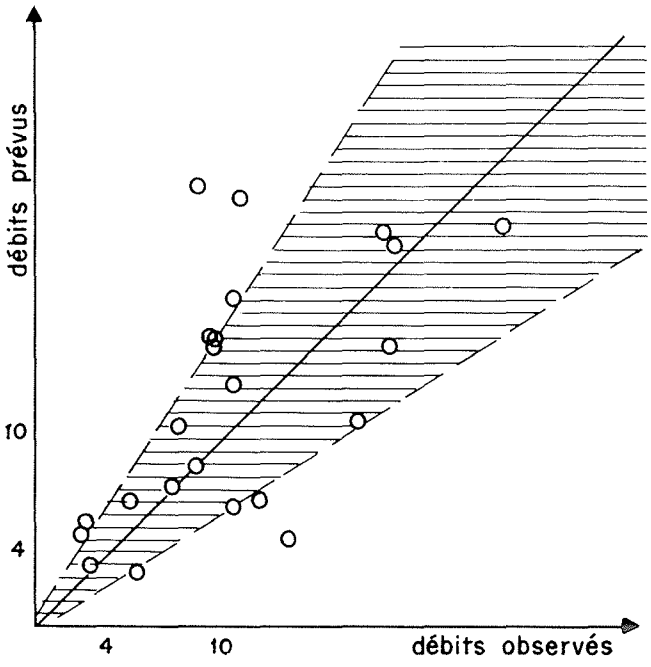


Sélection des préviseurs par la méthode stepwise

2/ En l'absence des termes prévisionnels

3/ En présence des termes prévisionnels





4/ Sélection des préviseurs par la méthode des composantes principales

5/ Régression simple avec le débit du mois précédent

qui, pour être valable, nécessite l'acceptation des hypothèses du paragraphe 1.

Mais les difficultés augmentent quand les préviseurs ne sont pas indépendants, ce qui est relativement fréquent en hydrologie.

L'estimation du vecteur b des coefficients de régression à partir du tableau X des préviseurs observés et du vecteur Y des réalisations observées s'obtient par :

$$b = (X'X)^{-1} X'Y$$

avec :

X' : matrice transposée de X;

$(X'X)^{-1}$: matrice inverse de X'X.

Quand les préviseurs ne sont pas indépendants, la matrice X'X est presque singulière, son déterminant est presque nul et l'inversion présente des difficultés avec pour conséquence des erreurs importantes.

Si β est le vecteur des véritables coefficients de régression et si on appelle L la distance entre ce vecteur et son estimation b telle que :

$$L^2 = (b - \beta)' (b - \beta)$$

On démontre que l'espérance mathématique $E(L^2)$ a pour expression :

$$E(L^2) = s^2 \sum_{i=1}^{i=p} 1/\lambda i$$

où λi sont les valeurs propres de la matrices X'X.

Quand les préviseurs sont corrélés, la matrice X'X possède une ou plusieurs valeurs propres presque nulles et la distance entre les β et leur estimation b tend à devenir très grande :

$$1/\lambda i \to \infty$$

Dans ces conditions, les estimations b sont exagérées en valeur absolue et instables. Que valent les tests statistiques faits sur de telles estimations ?

Aussi, parmi les méthodes classiques de sélection des préviseurs, nous rejetons la méthode « en arrière » [1] parce qu'elle commence par une régression multiple sur l'ensemble des préviseurs à traiter. Nous lui préférons les méthodes « en avant » qui commencent par une régression simple, et plus particulièrement la méthode Stepwise [1] qui à chaque étape en avant, fait les tests F partiels sur toutes les variables. Cependant cette méthode n'élimine pas les préviseurs corrélés ni l'inflation des coefficients de régression et des « statistiques » qui résulte de leur emploi.

Cependant nous avons obtenu généralement de bons résultats quand ces méthodes étaient employées pour sélectionner les préviseurs *en présence* des termes prévisionnels, ce qui confirme des résultats semblables obtenus ailleurs [2]. Cette constatation n'est pas encore expliquée de manière satisfaisante. Ainsi la figure 3 montre les résultats de prévisions effectuées à partir d'une équation de régression ajustée dans ces conditions. Les données sont les mêmes que pour la figure 2 mais ici plus de la moitié des réalisations tombent dans l'intervalle de confiance des prévisions.

Nous préférons quand même employer des méthodes plus explicites. Ainsi les composantes principales, appliquées sur la matrice X'X nous permettent de déterminer combien de préviseurs sont réellement indépendants. Nous choisissons un préviseur par composante retenue : celui qui a le plus fort poids, et parmi ceux-ci, considérés comme indépendants, ce sont les méthodes classiques qui nous disent ceux qu'il faut retenir en définitive.

Ainsi la figure 4, sur le même exemple que les figures 2 et 3, montre les résultats obtenus en prévision avec une équation de régression ajustée par ce moyen.

Remarquons que pour l'exemple de la figure 2 qui donne de très mauvais résultats en prévision, le coefficient de corrélation multiple obtenu sur l'échantillon est de 0,99 tandis que pour les exemples 3 et 4, qui donnent de bien meilleurs résultats en prévision, les coefficients de corrélation multiple ne sont respectivement que de 0,80 et de 0,65.

Le coefficient de corrélation multiple n'est donc pas un critère suffisant pour juger de la validité d'une équation de régression utilisée en prévision.

L'exemple ci-dessus montre aussi que, quelle que soit la méthode de sélection employée, l'équation de régression retenue n'est pas forcément le meilleur modèle. Il existe généralement des modèles différents donnant des résultats équivalents. Quelquefois aussi il n'existe pas de modèle satisfaisant quand on obtient des débits négatifs par exemple. Dans ce cas on peut, comme pour les modèles déterministes, ajouter des contraintes [5].

Enfin, nous testons actuellement la méthode de la Ridge Regression [3] qui, par addition d'un même scalaire à chacun des termes diagonaux de la matrice X'X, diminue considérablement la variance des coefficients de régression, mais apporte un certain biais. Le vecteur des coefficients de régression biaisés b^* est donc obtenu par :

$$b^* = (X'X + KI)^{-1} X'Y$$

où I est la matrice unité et K un scalaire supérieur à zéro.

Pour une valeur de K donnée, on démontre que l'espérance mathématique de la distance $L^2(K)$ dépend de deux termes qu'on peut représenter par $\gamma_{1(K)}$ et $\gamma_{2(K)}$:

$$E(L^2(K)) = \gamma_1(K) + \gamma_2(K)$$

$\gamma_1(K)$ est la variance des estimations de β. C'est une fonction à décroissance monotone de $K = 0$ à $K = \infty$ dont la dérivée à l'origine a pour valeur $-\infty$ quand la matrice X'X a quelques valeurs propres voisines de zéro, c'est-à-dire quand des préviseurs sont corrélés.

Ce terme décroit donc très rapidement avec de faibles valeurs de K.

$\gamma_2(K)$ est le biais apporté par l'emploi de b^* à la place de β. C'est une fonction à croissance monotone de $K = 0$ à $K = \infty$. La dérivée à l'origine est nulle. Elle croît donc lentement pour de faibles valeurs de K.

La somme de ces deux fonctions passe donc par un minimum qui correspond au choix du vecteur le plus stable pour les coefficients de régression.

Un algorithme permet de déterminer rapidement la valeur à prendre pour K.

Cette méthode, utilisée dans d'autres disciplines (identification des systèmes) relègue au rang des accessoires, la notion de coefficient de corrélation multiple habituellement seule prise en compte.

Remarquons enfin que, bien souvent, une régression simple ajustée sur le préviseur le plus important donne d'aussi bons résultats en prévision qu'une régression prenant en compte plus de préviseurs (fig. 5).

Tableau 1

Régressions obtenues automatiquement par la méthode Step Wise
Comparaison des réalisations et des prévisions

Mois	Coefficient de régression	Coefficient de corrélation	Ecart relatif moyen	Intervalle de confiance 60 % échantillon		Intervalle de confiance 68 % population	
				dans	hors	dans	hors
1	0,80	0,39	0,50	11	12	13	10
2	2,17	0,53	0,53	13	10	14	9
3	0,63	0,51	0,52	10	13	12	11
4	0,38	0,68	0,88	8	15	9	14
5	0,49	0,55	0,61	11	12	12	11
6	0,06	0,28	2,54	2	21	2	21
7	— 0,01	— 0,04	3,82	9	14	11	12
8	0,63	0,54	0,74	13	10	13	10
9	0,88	0,63	0,73	9	14	9	14
10	0,14	0,40	1,67	9	14	9	14
11	0,37	0,46	0,72	11	12	13	10
12	0,53	0,53	0,57	7	16	8	15

Tableau 2

Modèle déterministe. Comparaison Prévisions-Réalisations (57-67)

Mois	Moyenne observée	Ecart-type observé	Moyenne prévue	Ecart-type prévu	Coefficient de corrélation	Coefficient de régression	Ordonnée à l'origine	Nombre de points
Janvier.	28,8	11,77	26,54	3,79	0,719	2,23	— 30,4	10
Février.	21,65	7,41	24,2	2,94	0,408	1,02	— 3,19	10
Mars.	22,9	8,17	17,14	2,34	0,315	1,1	4	10
Avril.	21,7	9,06	14,86	3,37	0,589	1,58	— 1,87	10
Mai.	17,74	8,06	12,64	3,12	0,765	1,97	— 7,25	10
Juin.	11,14	5,37	8,49	3,67	0,761	1,11	1,7	10
Juillet.	5,66	3,45	5,29	2,17	0,80	1,28	— 1,05	11
Août.	50,7	3,08	3,54	1,43	0,166	0,357	3.8	11
Septembre.	6,65	6,49	3,90	2,20	0,324	0,944	2,97	11
Octobre.	9,66	10,77	6,7	4,98	0,540	1,16	1,83	11
Novembre.	15,75	10,54	14,98	6,46	0,652	1,06	— 0,17	11
Décembre.	25,42	14,5	23,32	4,09	0,624	2,21	— 26,15	11

Utilisation de modèles déterministes (c'est-à-dire non linéaires) pour la prévision

Avec les modèles de régression, nous avons supposé que le bassin versant se comportait comme un système linéaire interposé entre les préviseurs ou leur transformée (racine carrée ou logarithme) et les débits. Cette supposition est-elle plus ou moins logique que celle qui consiste à représenter le comportement du bassin par un modèle non linéaire formulable mathématiquement et qu'on explique en comparant le bassin à un réservoir rempli par les pluies et soumis à certaines règles de gestion, fonction de la température, de la saison, etc. ?

Pour notre part, nous sommes bien incapables d'apporter une réponse à cette question.

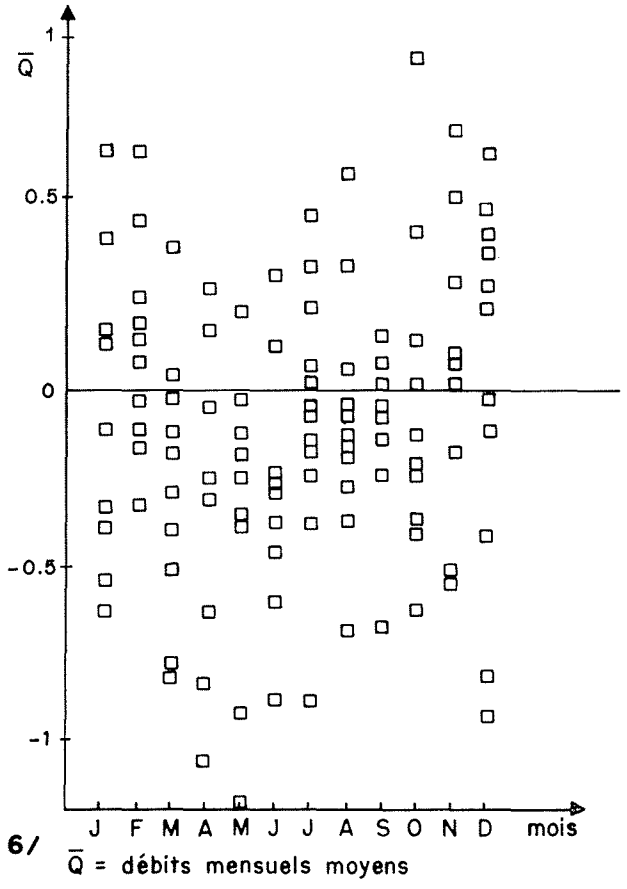
Que les moyens actuels permettent de mieux ajuster les premiers modèles que les seconds, c'est certain, mais est-ce une raison pour abandonner les seconds ?

En ce qui concerne le modèle déterministe utilisé pour la prévision des débits mensuels, on peut le résumer en deux fonctions. D'abord une fonction de production non linéaire par laquelle la pluie mensuelle est transformée en eau disponible pour l'écoulement. Nous avons utilisé la méthode Thornthwaite [6]. Le seul ajustement qui est fait à ce stade par rapport à la méthode citée en référence consiste à ajuster la capacité maximale du réservoir (pouvoir de rétention du bassin) de manière à ce que, sur la période d'observation, le volume d'eau disponible égale le volume d'eau écoulé.

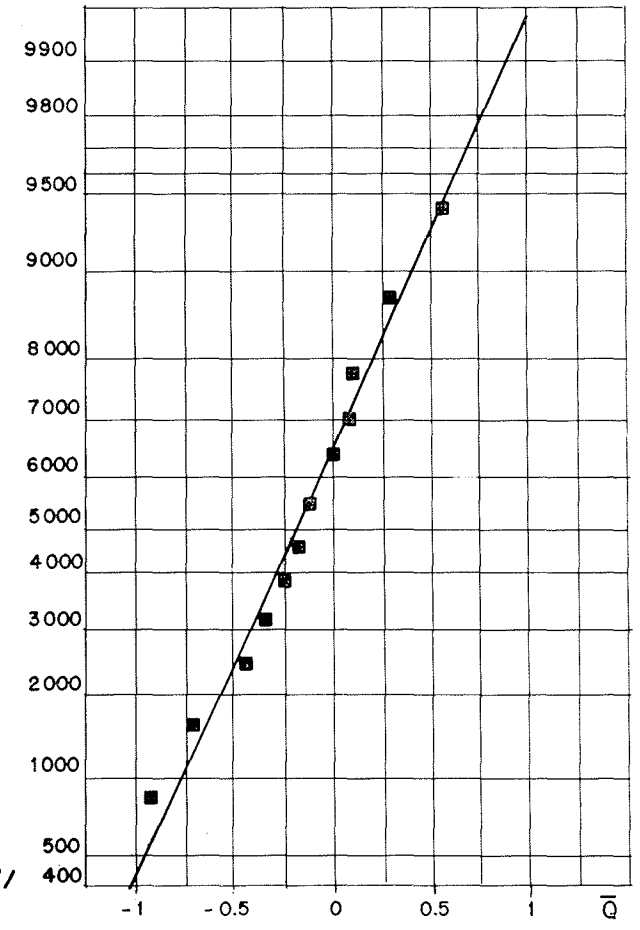
Ensuite une fonction de modulation linéaire, absolument comparable à une régression est utilisée pour représenter le système qui lie l'eau disponible d'un mois aux débits du même mois et à ceux des mois suivants. Cet ajustement peut se faire par les moindres carrés exactement comme pour la régression. Mais pour que les mois à fort débit n'aient pas un poids excessif, il est préférable de travailler sur les valeurs centrées et réduites par les moyennes et les écarts types mensuels. Pour notre part, nous avons utilisé une méthode d'optimisation (Rosembrock) la fonction à minimiser étant la somme des erreurs relatives. L'ensemble des résultats que nous avons obtenus en utilisant le modèle en prévision sur onze années est tout à fait satisfaisant, puisque le coefficient de corrélation obtenu sur toutes les valeurs (126) centrées et réduites à l'échelle mensuelle, est de 0,68.

Nous ne sommes pas surpris de ce résultat. Notre modèle, qui ne comporte qu'un seul paramètre pour la fonction de production (si on considère la méthode Thornthwaite comme une transformation des variables) et 8 paramètres pour la fonction de transfert a été en effet ajusté sur vingt-deux ans soit 264 observations. Une régression a 5 ou 6 paramètres, valable pour un mois seulement, n'est ajustée elle que sur 22 points.

Les tableaux 1 et 2 montrent les résultats des comparaisons faites mois par mois entre des prévisions et les réalisations correspondantes. Pour un tableau les prévisions



6/ $\bar{Q}$ = débits mensuels moyens



7/

Vérification de la loi de distribution des erreurs.
Modele déterministe mensuel.

ont été faites à partir d'équations de régression sélectionnées automatiquement, pour l'autre tableau les résultats sont ceux du modèle déterministe. Au vu des résultats d'un seul modèle, il nous semble en effet difficile de porter un jugement de valeur.

Il reste à savoir calculer un intervalle de confiance. Si nous vérifions (aussi grossièrement que pour les corrélations) que nos erreurs sont normalement distribuées, nous pouvons, pour les prévisions futures déterminer un intervalle de confiance expérimental même si ces erreurs ne sont pas indépendantes, car nous avons la possibilité d'en tenir compte.

Nous avons effectué un certain nombre de vérifications sur nos résultats : la figure 6, où les erreurs sont exprimées par rapport au débit mensuel moyen $\overline{q}$, montre qu'elles sont relativement indépendantes de la saison. Cependant, comme il arrive souvent quand on utilise d'autres méthodes d'ajustement que les moindres carrés, leur moyenne n'est pas nulle et a pour valeur — 0,175. Par contre, la figure 7 montre, sur tous les mois mélangés, que la distribution est gaussienne. Le coefficient d'autocorrélation avec retard de 1 a pour valeur 0,02. L'écart-type des erreurs, qui a pour valeur 0,47, permet donc de calculer l'intervalle de confiance d'une prévision P. Au seuil de probabilité 90 % cet intervalle de confiance a pour expression :

$$P \pm 0{,}175\,\overline{q} \pm 1.65 \times 0{,}479\,\overline{q}$$

Faut-il plus de données pour ajuster ainsi un modèle déterministe que pour calculer des régressions mois par mois ?

Non, si nous considérons que nous ne pouvons pas valablement ajuster une régression sur moins de 20 points, c'est-à-dire vingt années d'observation de débits mensuels. Avec ce volume d'observations si nous utilisons dix années pour ajuster le modèle non linéaire (120 points) il nous en restera autant pour déterminer l'intervalle de confiance.

Conclusion

Dans ce rapport nous avons seulement voulu juger différentes méthodes à partir des résultats obtenus sur quelques dizaines de prévisions faites en utilisant le même modèle. Nous pensons que c'est un moyen plus objectif que celui qui consiste à montrer un certain nombre de résultats obtenus avec un modèle dont la formulation change après chaque réalisation.

Les exemples cités sont trop limités pour permettre de tirer des conclusions définitives. Tout au plus pouvons-nous dire qu'en matière de prévision, le choix d'un modèle plutôt que d'un autre doit être décidé par des comparaisons objectives, mais que ce n'est pas une affaire de doctrine. Il reste suffisamment de problèmes à résoudre dans l'utilisation de l'une ou l'autre des méthodes pour y consacrer encore des investissements.

Bibliographie

[1] DRAPER and SMITH. — Applied Regression Analysis. *John Wiley and Sons.*

[2] GUILLOT et LUGIEZ. — La prévision des débits et des crues dans les services d'exploitation d'E.D.F. *E.D.F.-D.T.G.*

[3] HOERL and KENNARD. — Ridge Regression: Biased estimation for non orthogonal problems. *Technometrics* vol. 12, n° 1, (février 1970), p. 55-82.

[4] MASSON. — La prévision des débits mensuels. *G.H.L.N.H.-E.D.F.*, note (7/70).

[5] ROSENBERG. — Problèmes de corrélations multiples avec contraintes en Hydrologie. *Bulletin de l'A.I.H.S.*, XV, (39/1970), p. 47 à 54.

[6] THORNTHWAITE and MATHER. — Instructions and tables for computing potential évapotranspiration and the water balance. Centerton New-Jersey (1957).

Discussion

Président : M. P. CASEAU

M. le Président remercie M. MASSON et ouvre la discussion.

« Votre modèle, observe M. DUJARDIN (S.O.G.R.E.A.H.) comporte une fonction production et une fonction modulation. Vous avez réglé successivement les paramètres de la première puis de la seconde. Nous avons fait un modèle très semblable et procédé de la même manière. Il me semble que dans votre modèle la fonction production est trop simplifiée, ce qui fait que vous avez beaucoup de difficultés pour régler la fonction modulation. Nous avons établi une fonction production à trois paramètres et réglé une loi liant ces trois paramètres $f(a, b, c) = 0$. Puis nous réglons la fonction modulation à deux paramètres α et β en utilisant la loi précédente pour déterminer en même temps les trois inconnues α, β et $f(a, b, c)$. Le problème réside, comme l'a signalé M. ROCHE, dans le choix du critère de réglage (maximum, écart-type ou autre).

M. le Président signale qu'au récent Congrès de Düsseldorf de l'I.F.A.C., consacré à « l'identification », un mémoire de M. GODIN exposait une méthode analogue à celle décrite par M. MASSON et montrait l'intérêt, lorsqu'on inverse des matrices de corrélation d'ajouter à cette matrice une unité multipliée par un coefficient λ faible. On ajuste ensuite ce coefficient pour optimiser simultanément le biais et l'écart quadratique. La théorie de l'identification a fait en automatisme beaucoup de progrès qui pourraient vraisemblablement être utilisés dans les modèles hydrologiques.

M. GUILLOT demande à M. MASSON de préciser la consistance de l'échantillon sur lequel a été calculé le coefficient de corrélation de 0,76, qu'il a évoqué comme indice de qualité de la méthode déterministe qu'il emploie.

Le modèle a été ajusté sur la période 1935-1956, répond M. MASSON, et les résultats visés ci-dessus correspondent à la période 1957-1967. L'échantillon qui a servi à calculer la corrélation est constitué par $11 \times 12 = 132$ valeurs mensuelles.

M. GUILLOT fait remarquer que, dans ces conditions, la validité explicative du modèle « déterministe » de M. MASSON est très décevante, puisque le coefficient de 0,78 est peu supérieur à la covariation saisonnière résultant du mélange dans l'échantillon des mois d'hiver abondants et des mois d'été secs. Il serait curieux de connaître la valeur des douze corrélations calculées pour chacun des mois de la série de onze ans.

M. DUJARDIN donne quelques résultats obtenus avec le modèle dont il a parlé au début de la discussion. Nous étudions, dit-il, une méthode de réglage automatique qui n'est pas encore opérationnelle et actuellement notre réglage est empirique. Nous avons appliqué notre modèle à un ensemble de dix-neuf bassins, ce qui implique le réglage de dix-neuf séries de paramètres; les résultats étaient ensuite repris sur un modèle de propagation Muskingum pour calculer le débit à l'aval du modèle global. On a ainsi calculé les débits afférents à seize années comprenant les trois années qui ont servi au réglage; on a calculé le pourcentage d'erreur :

$$E = 100\,(Q \text{ calculé} - Q \text{ vrai})/Q \text{ vrai}$$

sur les débits mensuels. Ce pourcentage d'erreur est le même pour les mois d'étiage et les mois de hautes eaux. On trouve que pour :
— 40 % des mois l'erreur est $\leqslant 10\,\%$;
— 75 % des mois l'erreur est $\leqslant 20\,\%$;
— 98 % des mois l'erreur est $\leqslant 40\,\%$.

M. GUILLOT pense qu'un tel résultat est probablement satisfaisant mais pour pouvoir en juger, dit-il, il faudrait pouvoir comparer la variabilité avant prévision et après prévision, afin de se rendre compte si celle-ci est diminuée ou augmentée.

M. PREISSMANN se range à cet avis et souligne que ce que l'on voudrait savoir c'est la part de la variabilité des débits du mois qui est « expliquée » par rapport à la variabilité totale autour de la moyenne.

Nous conclurons sur cette remarque parfaitement justifiée, dit M. le Président, qui remercie les conférenciers et tous les participants à la discussion qui ont rendu particulièrement vivante cette séance.

Abstract

An 'experimental' comparison between monthly discharge forecasts obtained by regression methods and with a deterministic model

Tests have been carried out with several models adjusted in different ways with a view to determining the monthly discharge of the river Sioule at Pont-du-Bouchet in the Massif Central, involving an area of 1,170 km². Records for a period of about twenty years were used to adjust the models, and twenty other years were considered in comparing the predicted and actual discharge data.

The predicted monthly discharge data depend both on quantities known at the time of making the forecast ("predictors") and on quantities associated with events occurring between the time of making the forecast and the instant at which the flow takes place ("prediction terms"). The reliability of the forecast depends on the relative importance of these two types of quantity, but can readily be ascertained by means of a few simple correlations.

Use of linear models : Regressions.

Although these models are now automatic and in general use, they still give rise to a number of problems.

First among these is the independence of errors, which it is often difficult to check and sometimes impossible to achieve, and which is liable to bias statistical checks and leave no alternative but to view the calculated confidence intervals with caution.

Conventional methods of selecting a vector for the "predictors" are based on statistical checks requiring the same caution as the above.

In addition, where the "predictors" are not independent, unstable regression coefficients with exaggerated values result. Thus, automatic methods have been found to give poor results for prediction. Experimentally, however, better results are obtained where the "prediction terms" are available when selecting the "predictors". Satisfactory results have also been obtained by making use of the main components.

The methods of the future (e.g. the Ridge Regression method of system identification), however, ignore the multiple and partial multiple correlation coefficients and other conventional statistical quantities, and only take the stability of the regression coefficients into account.

Use of deterministic model (Thornthwaite).

This model features a production function which is none other than Thornthwaite's balance method with only one parameter adjusted to make the average available water equal to the discharge.

The modulation function is a delayed-effect model adjusted by means of an optimization method (Rosenbrock) in which the sum of the relative errors is made a minimum.

The predicted data supplied by this model are equivalent to those given by regression methods, and with the necessary data for a twenty-year period available, an experimental confidence interval can be calculated for both cases around the data supplied by the deterministic model.

To conclude, therefore, the experimental choice of a prediction method depends on objective comparisons in which all the models to be compared start off with equal chances. However, there are still very many improvements to be made to all the different models available for use.