



HYDROLOGIE STATISTIQUE

Enoncé et Corrigé du Contrôle du 21 Janvier 2000
Promos 3HSEE et DEA-STE 1999/2000
Enseignant : R. Ababou

Exercice III

III. ANALYSE STATISTIQUE MULTIVARIEE : CORRELATION MULTIPLE ET A.C.P (Débits Mensuels en 6 Stations)

Références : Ce problème est tiré de l'étude corrélatrice des débits mensuels de 1950-1972 en 6 stations hydrométriques pyrénéennes (Naguilhes etc) par *D. Duband et al.* (EDF), avec comme objectif, entre autres, la recherche de redondances entre stations. Ce problème a été repris et remanié lors d'un Bureau d'Etude d'Hydrologie Statistique : voir l'énoncé du B.E.3 1997/98, Section "Analyse Multivariée", dans le polycopié d'Hydrologie Statistique (R.Ababou).

Une partie des données et des résultats obtenus (corrélations et composantes principales) sont résumés dans les **Tableaux ci-joints**. A partir de ces quelques tableaux, on demande de brèves réponses aux questions suivantes (qualitatives) :

1. Est-ce que l'étude corrélatrice porte sur les débits moyens mensuels ? En d'autres termes : quelles sont les variables hydrologiques réellement analysées ici ?
2. Que remarquez-vous concernant la CP1 (interprétation et conséquences par rapport à l'ACP) ?
3. Que remarquez-vous concernant les 6 "valeurs propres" (conséquences par rapport à l'ACP) ?
4. Au vu de la matrice de corrélation, et des autres informations disponibles, peut-on voir s'il existe des redondances significatives entre stations (2 à 2, par groupes) ?

REponses DU III :

1. Préciser les variables hydrologiques réellement analysées ici (débits...).

La réponse triviale est qu'il s'agit des écoulements mensuels (et non pas des hauteurs d'eau !) en 6 stations hydrométriques. Ces stations correspondent aux exutoires de 6 bassins versants (BV) différents.

Il convient cependant de préciser aussi les points suivants concernant la moyenne et la normalisation des débits.

Moyenne mensuelle :

La variable étudiée est un débit *moyen* mensuel, pour le mois de mai des années 1950 à 1972 incluses. Autrement dit, la variable étudiée (en chaque station) est la *moyenne interannuelle* du débit du mois de mai sur 23 années, considérées comme 23 observations indépendantes de la variable « débit mensuel du mois de mai ». Le débit mensuel est sans doute calculé à partir de la somme des débits journaliers du mois considéré.

Normalisation par l'aire

En fait, la variable étudiée ici n'est pas un débit proprement dit (Q en m^3/s) mais un débit par unité d'aire (A) du Bassin Versant, ou débit spécifique (q), que l'on peut exprimer selon les cas en litre/s/km², en mm/an, ou en mm/mois comme ici. Soit, pour la station numéro « j » :

$$q_j \text{ (mm / mois)} = 2592 \frac{Q_j \text{ (m}^3 / \text{s)}}{A_j \text{ (km}^2)}$$

où le facteur 2592 transforme des m³/s par km² en millimètres de lame d'eau mensuelle (mm/mois) en supposant un mois de 30 jours.

Normalisation par l'écart-type

Enfin, remarquer que les débits étudiés seront à nouveau normalisés si l'on procède à une ACP en variables réduites $(x-m)/\sigma$. Or, que les variables «brutes» soient des débits Q ou des débits spécifiques q , on obtient dans les deux cas les mêmes variables réduites. En effet :

$$Q_j \rightarrow \tilde{Q}_j = \frac{Q_j - \bar{Q}_j}{s_{Qj}} ;$$

$$q_j \rightarrow \tilde{q}_j = \frac{q_j - \bar{q}_j}{s_{qj}} = \frac{\frac{Q_j}{A_j} - \frac{\bar{Q}_j}{A_j}}{\frac{s_{Qj}}{A_j}} = \frac{Q_j - \bar{Q}_j}{s_{Qj}}$$

En résumé, on a le choix entre deux types de normalisation pour les débits [nouvelles notations] :

- I. Normalisation par l'aire drainée (A) : débit brut $Q \rightarrow$ débit spécifique $q = Q/A$;
- II. Normalisation par la variabilité (σ_Q) : débit brut $Q \rightarrow$ débit réduit $q = (Q - m_Q) / \sigma_Q$.

2. Première analyse des résultats de l'ACP : commentaires sur la CP1, interprétation, conséquences.

Cette question concerne la CP1, ou 1^{ère} Composante Principale, encore notée Z_1 . Les résultats fournis dans le Tableau donnent deux indications différentes sur la CP1 :

- ♦ Dans le 1^{er} sous-tableau, 3^{ème} ligne, on trouve les coefficients p_{1j} exprimant la CP1 en fonction des variables réduites ($j=1, \dots, 6$) ;
- ♦ Dans le 2nd sous-tableau, on trouve les valeurs propres λ_i des six CP_i ($i=1, \dots, 6$) et en particulier la valeur propre λ_1 de la CP1.

Cette dernière donnée (λ_1) montre que la CP1 fournit 70% de l'information totale ou

du contenu en variance des 6 stations (ce qui sera vu plus en détail en réponse à la question 3). Mais cette remarque, bien que juste, risque de masquer le fait que l'information apportée par la CP1 ne permet pas de discriminer entre les différentes stations ou groupes de stations, comme on va le voir maintenant.

On s'intéresse donc ici aux coefficients de la CP1 (p_{1j}). Ces coefficients sont aussi appelés «cosinus directeurs» de la CP1 [terminologie de *D. Duband, op.cit.*].

Le 1^{er} sous-tableau montre que les p_{1j} sont tous du même ordre, soit en moyenne :

$$p_{1j} \approx 0.41 \approx 1/\sqrt{6} .$$

Ceci implique, à peu près :

$$CP1 \approx \frac{\sum_{j=1}^{j=6} x_j}{\sqrt{\sum_{j=1}^{j=6} 1_j}} = \frac{\sum_{j=1}^{j=6} x_j}{\sqrt{\sum_{j=1}^{j=6} s_{x_j}^2}} = \frac{\sum_{j=1}^{j=6} x_j}{\sqrt{6}}$$

où x_j représente la variable réduite étudiée (ici le débit réduit). Rappelons que $\text{Var}(x_j) = 1$ par construction, que les λ_j sont les $\text{Var}(CP_j)$, et que l'on a (en variables réduites) :

$$\Sigma \lambda_j = \Sigma \text{Var}(CP_j) = \Sigma \text{Var}(x_j) = K = 6 .$$

On a donc ici une confirmation expérimentale d'un phénomène souvent observé en hydrologie statistique, du moins lorsqu'il s'agit d'étudier des variables préalablement réduites et/ou suffisamment homogènes (comme ici) :

La 1^{ère} Composante Principale est souvent (comme ici) une moyenne des K variables réduites étudiées. On peut alors l'interpréter comme un « facteur de taille » ou une sorte de variable globale, peu discriminante mais globalement représentative de l'ensemble des K variables.

Les coefficients obtenus montrent donc que la CP1 est peu discriminante entre les 6 stations.

En conséquence, bien que la CP1 ait un poids important puisqu'elle explique ici 70% de la variance totale, elle ne peut pourtant pas être utilisée pour rechercher des proximités de comportement ou des redondances entre stations. Pour cela, il faudra plutôt utiliser les CP suivantes (CP2, CP3,...).

Pour une analyse plus approfondie, voir l'Annexe « **Relations entre CP variables réduites** ». On montre dans cette annexe que :

$$r_{ziX_j} = \sqrt{I_i} A_{ij},$$

et on en déduit que la CP1 (Z_1), facteur de taille, est corrélée pareillement avec toutes les variables X_j ($j=1, \dots, K$). On vérifie ici que le coefficient de corrélation de la CP1 avec chacune des $K=6$ variables réduites est en effet à peu près constant (de +80% à +85%).

3. Suite de l'analyse des résultats de l'ACP : commentaires sur les 6 valeurs propres, interprétation, conséquences.

On continue l'analyse précédente en mettant maintenant l'accent sur l'ensemble des CP (et pas seulement la CP1).

On remarque d'abord que les 6 valeurs propres λ_j données dans le Tableau vérifient presque exactement, aux erreurs d'arrondi près, la relation $\sum \lambda_j = 6$. En effet, de façon générale, pour une ACP « réduite », on a l'identité :

$$\text{Trace}(\Lambda) = \sum_{j=1}^{j=K} I_j = \sum_{j=1}^{j=K} s_{Z_j}^2 = K$$

où K est le nombre de variables ($K=6$ ici).

Noter de plus que les CP ont été classées par ordre de λ_j décroissantes, c'est-à-dire

par variances décroissantes puisque $\lambda_j = \sigma_{Z_j}^2$. Or on constate en cumulant les variances (λ_j) dans la table ci-dessous que les 3 premières valeurs propres « expliquent » 98% de la variance totale, soit : $(\lambda_1 + \lambda_2 + \lambda_3)/6 \approx 98\%$.

K = 6	CUMUL DES VARIANCES
70% de la variance dans CP1	$I_1 / 6 \approx 70.1\%$
20% de la variance dans CP2,3	$(I_1 + I_2) / 6 \approx 88.8\%$ $(I_1 + I_2 + I_3) / 6 \approx 98.1\%$
Moins de 2% de la variance dans les CP 4,5,6.	$(I_1 + I_2 + I_3 + I_4) / 6 \approx 99.2\%$ $(I_1 + I_2 + I_3 + I_4 + I_5) / 6 \approx 99.7\%$ $(I_1 + \dots + I_6) / 6 \approx 100\%$

On peut faire alors les choix suivants, selon les objectifs poursuivis :

- I. Pour des objectifs de synthèse des débits, par exemple reconstitution de données, extrapolation spatiale et cartographie, il conviendrait de garder les trois 1ères (CP1, CP2, CP3) ;
- II. Pour des objectifs d'analyse typologique, redondances ou proximités entre stations hydrométriques, il conviendrait de visualiser les résultats dans le plan des (CP2, CP3), la CP1 n'étant que peu discriminante.

NB: Les applications de type (I) et (II) ne sont pas développées dans cet exercice (voir ailleurs : Cours et Bureau d'Etudes).

4. Peut-on analyser les redondances entre stations à partir de la matrice de corrélation (2 à 2,...) ?

Cette dernière question porte sur la possibilité d'utiliser la seule matrice de corrélation des débits aux six stations (non diagonalisée, sans ACP) comme outil de diagnostic sur les redondances entre stations (ou groupes de stations ?).

Les redondances entre stations prises 2 à 2 apparaissent clairement à partir des entrées de la matrice symétrique de corrélation R_{XX} .

Celle-ci est identique à la matrice de covariance C_{xx} des variables réduites. En bref, on a ici :

$$(C_{xx})_{ij} = (R_{xx})_{ij} = \rho_{x_i x_j}$$

Cette matrice symétrique est disponible au bas du Tableau de données fourni ci-joint. En distinguant approximativement les corrélations «fortes» ($\rho \geq 90\%$), «modérées» ($50\% \leq \rho < 90\%$) et «faibles» ($\rho < 50\%$), on peut faire les constations suivantes.

(X1,X2) (X3,X4) (X5,X6)	Ces stations sont fortement corrélées (positivement), avec $\rho > 90\%$.
(X3,X6) (X4,X6)	Ces stations ne sont que faiblement corrélées (positivement) : $\rho < 50\%$.
(...) (...) (...),...	Toutes les autres stations sont modérément corrélées (positivement) avec $\rho \approx 50\%$ à 70% .

On voit donc que les débits des 6 stations sont tous positivement corrélés entre eux¹, et que 3 couples de stations (ci-dessus) sont fortement corrélés.

On pourrait en conclure (selon la finalité) qu'il y a des redondances significatives dans ce réseau de stations, suffisantes peut-être pour justifier la suppression de 3 stations appartenant chacune à l'un des 3 couples ci-dessus : par exemple supprimer les stations (1,3,5) , ou bien (2,4,6) , mais pas (1,2,3) .

On peut comparer cette conclusion avec les résultats de l'ACP montrant que les trois premières CP (qui sont indépendantes) sont largement suffisantes pour expliquer les fluctuations des débits de toutes les stations (qui sont plus ou moins corrélés). (...).

Enfin, il convient de noter que la matrice de corrélation multiple (non diagonalisée) permet aussi de définir différents types de coefficients de corrélation, tels que le coefficient de corrélation « total », le

coefficient de corrélation multiple de la variable $X_1 / (X_2, X_3, X_4, X_5, X_6)$, etc. (Ceci pourrait éventuellement servir à analyser des groupes de stations sans passer par l'ACP, ce qui reste à approfondir...).

¹ Si une corrélation négative apparaissait, il faudrait l'interpréter, et essayer de comprendre en particulier si une forte corrélation négative peut être interprétée comme une redondance (ou non).

ANNEXE :

RELATIONS ENTRE COMPOSANTES PRINCIPALES (CP) ET VARIABLES REDUITES EN ANALYSE MULTIVARIEE

On explicite ici la relation « algébrique » entre les variables étudiées (X_j) et leurs composantes principales (CP_i ou Z_i), en particulier dans le cas où les X_j sont des variables réduites $X_j \leftarrow (X_j - M_j) / \sigma_j$. On en déduit des relations statistiques, telles que la corrélation entre CP_i et X_j , qui permet de mieux interpréter les coefficients reliant les CP_i aux X_j , et en particulier la CP_1 qui apparaît comme un simple « facteur de taille » dans la plupart des cas [cf. Exercice III].

Diagonalisation et changement de base (rappels) :

Le changement de base (P^T) transforme le vecteur des variables étudiées (X), de covariance C_{XX} , en un vecteur de variables (Z) dites *Composantes Principales* (CP) dont la covariance est diagonale ($C_{ZZ} = \Lambda$). La matrice P est orthogonale ($PP^T = P^T P = I$), et les relations de passage sont :

$$(P^T) : \underline{X}^{(n)} \rightarrow \underline{Z}^{(n)} = \underline{P}^T \underline{X}^{(n)}$$

$$(P) : \underline{Z}^{(n)} \rightarrow \underline{X}^{(n)} = \underline{P} \underline{Z}^{(n)}$$

où l'on a utilisé l'exposant " (n) " pour représenter les observations $(n) = (1), \dots, (N)$. Noter que la matrice orthogonale \underline{P} contient *en colonnes* les vecteurs propres de la matrice \underline{C}_{XX} (et sa transposée \underline{P}^T contient les mêmes vecteurs propres *en lignes*).

Le 1^{er} système $Z = P^T X$ exprime les CP_i (Z_i) en fonction des variables de départ (X_j), soit :

$$CP_i^{(n)} \equiv Z_i^{(n)} = \sum P_{ji}^{(n)} X_j^{(n)} \quad (\text{somme sur } j).$$

Si l'on pose $A = P^T$ par commodité, alors ce système devient

$$\underline{Z} = \underline{A} \underline{X},$$

et il s'écrit en notations indicielles :

$$CP_i^{(n)} \equiv Z_i^{(n)} = \sum A_{ij}^{(n)} X_j^{(n)} \quad (\text{somme sur } j).$$

Les $A_{ij} = P_{ji}$ sont donc les coefficients des composantes principales, également appelés *cosinus directeurs des CP* (terminologie de D. Duband).

Noter que les relations ci-dessus sont valables tant en ACP « brute » qu'en ACP « réduite » ; pour plus de détails voir l'Exercice II et l'Annexe Diagonalisation.

Corrélations croisées entre les CP_i et les X_j (réduites)

Les cosinus directeurs $A_{ij} = P_{ji}$ peuvent être interprétés « algébriquement » comme les projections des variables X_j sur les CP_i . Par exemple le coefficient $A_{12} = P_{21}$ est le coefficient de X_2 dans la 1^{ère} CP, et il représente donc la projection de X_2 sur la CP_1 .

Cependant, une interprétation statistique des cosinus directeurs est également possible. Plaçons-nous dans le cas de l'ACP « réduite », et calculons la matrice de covariance croisée \underline{C}_{ZX} des CP réduites (vecteur \underline{Z}) avec les variables réduites (vecteur \underline{X}). On obtient :

$$\begin{aligned}\underline{\underline{C}}_{ZX} &= \langle \underline{\underline{Z}} \underline{\underline{X}}^T \rangle = \langle \underline{\underline{Z}} (\underline{\underline{P}} \underline{\underline{Z}})^T \rangle = \langle \underline{\underline{Z}} \underline{\underline{Z}}^T \underline{\underline{P}}^T \rangle \\ &= \langle \underline{\underline{Z}} \underline{\underline{Z}}^T \rangle \underline{\underline{P}}^T = \underline{\underline{\Lambda}} \underline{\underline{P}}^T = \underline{\underline{\Lambda}} \underline{\underline{A}}\end{aligned}$$

où Λ est la matrice diagonale (λ_j) qui contient les valeurs propres de C_{xx} , et qui représente aussi la matrice de covariance C_{ZZ} des CP (non corrélées entre elles par construction). A partir de cela, on obtient la covariance croisée de la CPi (Z_i) avec la variable réduite X_j :

$$C_{Z_i X_j} = \Lambda_{ii} \mathbf{d}_{ij} A_{ij} = \mathbf{I}_i A_{ij} \quad (\text{sans sommation}).$$

D'où finalement le coefficient de corrélation croisé de la CPi (Z_i) avec la variable réduite X_j :

$$r_{Z_i X_j} = \frac{C_{Z_i X_j}}{\mathbf{s}_{Z_i} \mathbf{s}_{X_j}} = \frac{\mathbf{I}_i A_{ij}}{\sqrt{\mathbf{I}_i} \sqrt{1}} = \sqrt{\mathbf{I}_i} A_{ij} \quad (\text{sans sommation}).$$

A l'inverse, ceci nous permet d'interpréter les cosinus directeurs A_{ij} de la façon suivante :

$$A_{ij} = \frac{r_{Z_i X_j}}{\sqrt{\mathbf{I}_i}} = \frac{r_{Z_i X_j}}{\sqrt{\mathbf{s}_{Z_i}^2}} \quad (\text{sans sommation}).$$

Conséquences pour la CP1 (exemple) :

Lorsque les cosinus directeurs A_{1j} sont tous du même ordre, la 1^{ère} CP (Z_1) apparaît comme un « facteur de taille » global [voir *Exercice III*]. Or la relation ci-dessus montre que la CP1 est un facteur de taille si elle est corrélée pareillement avec toutes les variables X_j ($j=1, \dots, K$).

Par exemple, dans l'Exercice III, avec $K=6$ variables réduites, on peut constater que le coefficient de corrélation de la CP1 avec chacune des 6 variables est à peu près le même, entre +80% et +85% :

$$A_{1j} \approx 1/\sqrt{K} \approx 1/\sqrt{6} \quad \Rightarrow \quad \rho_{Z_1, X_j} = \sqrt{\lambda_1} A_{1j} \approx \sqrt{4.2} / \sqrt{6} \approx +0.83 \approx +83\%, \forall j = 1, \dots, 6.$$

où l'on remarque, incidemment, que le coefficient constant A_{1j} est proche de $1/\sqrt{K} \approx 1/\sqrt{6}$ (à suivre...).

Tableaux de Résultats d'Analyse Multivariée pour l'Exercice (III):

RESULTATS PARTIELS DE CORRELATION MULTIPLE ET A.C.P SUR 6 DEBITS MENSUELS PYRENEENS**P=6 variables analysées :**

Débits mensuels du mois de mai aux stations (1,2,3,4,5,6)

N=23 observations :

Années 1950 à 1972 incluses.

	1: Naguilhes	2: Lanoux	3: Izourt	4: Gnioure	5: Caillaouas	6: Bleu
Moyenne (mm)	377.7	298.5	394.2	401.2	327.2	176.7
Ecart-type (mm)	93.3	71.7	78.9	85.3	115.0	97.8
Coeffs.CP1 réduite	0.452	0.428	0.388	0.414	0.401	0.360

	CP1	CP2	CP3	CP4	CP5	CP6
Valeurs Propres l j	4.208	1.123	0.554	0.070	0.025	0.021

MATRICE DE CORRELATION	1: Naguilhes	2: Lanoux	3: Izourt	4: Gnioure	5: Caillaouas	6: Bleu
1: Naguilhes	1					
2: Lanoux	0.963	1				
3: Izourt	0.636	0.600	1			
4: Gnioure	0.703	0.682	0.965	1		
5: Caillaouas	0.671	0.578	0.473	0.535	1	
6: Bleu	0.646	0.528	0.312	0.357	0.919	1

Les CPj sont calculées ici en termes des variables réduites; les coeffs d'une CPj "réduite" sont aussi appelés "cosinus directeurs".