Controle de séries chronologiques corrélées par étude du cumul des résidus de la corrélation

par Philippe Bois Professeur à l'Ecole Nationale Supérieure d'Hydraulique et de Mécanique de Grenoble (ENSHMG), chercheur à l'Institut de Mécanique de Grenoble (IMG) .

Résumé : Le nombre d'observations hydrométéorologiques indispensables aux études et aux projets d'aménagement devient considérable ; pour utiliser de façon valable cette masse d'informations , il convient d'examiner attentivement la validité des données receuillies à la fin des chaines de mesures et d'acquisition .

La méthode du cumul des résidus de variables chronologiques corrélées permet d'infirmer ou de confirmer de façon précise l'homogénéité des séries , ceci dans le but de déceler des erreurs éventuelles .

Cette méthode présente deux intérêts :

+ un aspect graphique permet à l'utilisateur une interprétation visuelle

+ des seuils de probabilité le renseignent sur les hypothèses de stationnarité ou de non stationnarité

Elle permet donc de déceler des points aberrants ainsi que des périodes hétérogènes provenant , par exemple en hydrologie , soit d'erreurs de mesure , soit de changement d'appareillage ou de phénomènes naturels .

Après des rappels théoriques , on présente des exemples générés stochastiquement dans lesquels on a introduit des erreurs . Enfin , comme il paraissait difficile sinon impossible , d'obtenir analytiquement certaines probabilités , on a évalué les probabilités de dépassement des seuils en générant un grand nombre de séries réputées parfaites , ce qui a permis d'effectuer une certaine synthèse probabiliste .

Cette méthode remplace avantageusement la méthode des doubles cumuls ou double masse puisqu'elle la contient , est plus performante et donne des seuils de probabilité de cohérence des données .

I) Rappels théoriques :

I-1) Définitions :

Nous reprendrons l'article de J. Bernier (1977) en ce qui concerne les notations .

Considérons 2 séries chronologiques corrélées , dont les caractéristiques, calculées sur l'échantillon observé sont les suivantes :

x_i de i = 1 à n Variable de référence de moyenne m_x et d'écart type S_x
y_i de i = 1 à n Variable à tester de moyenne m_y et d'écart type S_y

Soit r le coefficient de corrélation entre ces deux séries observées .

On appellera e_i le résidu de l'observation i :

$$e_i = y_i - m_y - r * S_y/S_x * (x_i - m_x)$$

e_i est l'écart entre la valeur vraie de y_i et son estimée par la corrélation établie sur la série .

Soit $Z_k = e_1 + e_2 + e_3 + ... + e_k$ pour $k < n+1$;

Z_k est le cumul des k premiers résidus .

On sait que : Moyenne des e = 0 d'où : $Z_n = 0$
Variance (e) = Variance (y) * (1 - r^2)
$S^2_e = S^2_y * (1 - r^2)$

Si on trace Z_k en fonction de k , on obtient le tracé du cumul des résidus ; c'est une courbe partant de (0 , 0) aboutissant à (n , 0) où chaque incrément de Z correspond au résidu de l'observation correspondante :

$$Z_k = Z_{k-1} + e_k$$

Si les séries sont homogênes et si les observations sont indépendantes , le tracé de Z_k en fonction de k va donner une courbe oscillant autour de l'axe des abscisses .

Mais , prenons le cas où la série y a été sous estimée pendant une période , par exemple au début de la série , les résidus seront alors plutôt négatifs au début de la série et plutôt positifs vers la fin , dés que l'on aura quitté cette période de sous estimation ;le tracé des Z va mettre en valeur , par effet de cumul , cette hétérogénéité .

La courbe va au début s'éloigner de plus en plus de l'axe des k puis lorsque l'on commencera à retrouver la deuxième période , elle va regagner petit à petit l'axe . Le tracé va avoir alors l'aspect d'un triangle et non d'une oscillation , comme le montrent les figures 1 et 2 .

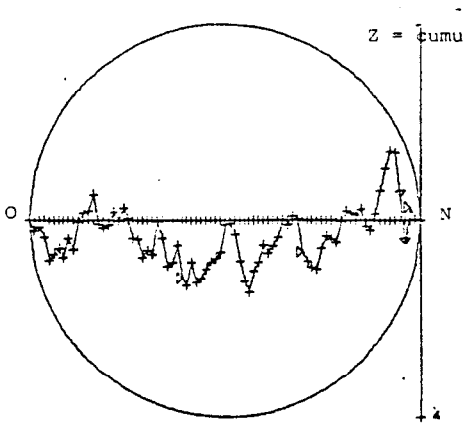

Figure 1 :
Cumul de résidus
de série correcte

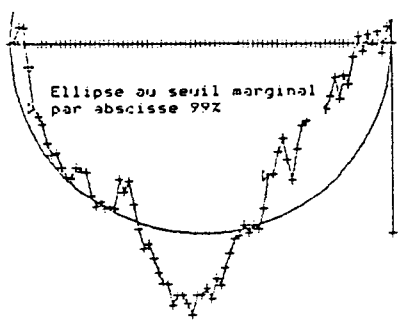

Figure 2 :
Cumul de résidus
de série erronée

I-2) Estimation de la loi de probabilité du cumul des résidus :

Nous venons de voir l'aspect graphique de cette méthode ; une comparaison avec la méthode traditionnelle des doubles cumuls la place nettement en tête d'un point de vue purement qualitatif , c'est à dire au niveau de l'appréciation visuelle . En effet , on remplace un cumul de valeurs par un cumul de résidus .

Mais il est important de savoir à partir de quel écart de Z_k à l'axe des abscisses, indépendamment des échelles choisies, il convient de s'inquiéter sur les données.

Pour cela, on peut, pour une valeur donnée de k, chercher la loi de probabilité de Z_k, dans l'hypothèse où les couples (x_i , y_i) sont binormaux, avec des observations indépendantes et des séries homogènes.

On montre (Bois Ph. ,1976) :

Espérance Mathématique $(Z_k) = E (Z_k) = 0$

Variance $(Z_k) = S^2_y * (1 - r^2) * (k (n - k)/n - k^2 * (m_{xk} - m_x)^2 / (n S^2_x))$

avec m_{xk} = Moyenne des k premiers x_i

Une approximation assez bonne de la variance de Z_k est fournie par :

$$\text{Variance} (Z_k) = S^2_y (1 - r^2) * k (n - k) (n - 1)/n^2$$

c'est à dire que cette variance est une fonction quadratique elliptique de k.

I-3) Détermination de la courbe de contrôle :

Au vu des estimations précédentes, il est aisé de déterminer une limite de contrôle telle que, si pour une abscisse donnée correspondant à une valeur de k, le cumul sort de cette limite, on soit amené à infirmer les hypothèses précédentes, notamment celle d'homogénéité.

On définit ainsi la limite au seuil de confiance C, définie pour chaque valeur de k par les extrémités du segment centré sur l'axe des abscisses et de demi longueur :

Demi-Longueur = t *Variance $^{1/2}(Z_k)$

où t est la valeur de la variable centrée réduite de Gauss de probabilité au dépassement 1 - C/2. Par exemple, pour C= 80 %, (1 - C)/2 = 10 % ; soit, à partir d'une table de la Loi Normale t = 1.28 pour une probabilité de 10 % au non dépassement. On démontre facilement à partir des équations précédentes que les extrémités des segments décrivent une ellipse de grand axe n.

Cela signifie que dans les hypothèses précédentes, pour une valeur donnée de k, il y a une probabilité C pour que le point représentatif de Z_k soit à l'intérieur de ce segment. Mais cela ne veut pas dire que, pour l'ensemble du tracé la probabilité qu'aucun point ne sorte soit C.

Nous avons vérifié nos hypothèses et nos calculs en générant stochastiquement des séries pour diverses tailles d'échantillons et en comptant pour chaque série le nombre de points à l'intérieur de l'ellipse de contrôle au seuil C.

Le tableau suivant montre que les résultats obtenus sont très voisins des résultats attendus malgré quelques hypothèses simplificatrices. Pour chaque classe d'effectifs, 5000 échantillons ont été générés stochastiquement et analysés; la fréquence de non sortie fournie par le tableau concerne l'ensemble des 5000 échantillons pour chaque classe d'effectifs.

Effectif des Echantillons	Seuils C de probabilité d'être à l'intérieur de l'ellipse				
	80.0	90.0	98.0	99.0	99.8
10	81.9	91.2	98.3	99.1	99.82
20	81.0	90.6	98.1	99.1	99.84
30	81.1	90.7	98.1	99.0	99.80
50	80.7	90.3	98.0	99.0	99.80
70	80.1	90.2	98.2	99.1	99.82
100	79.9	90.0	98.2	99.1	99.86
150	80.1	89.7	98.0	99.1	99.80
200	80.0	89.7	98.0	99.1	99.80
500	80.0	90.5	98.2	99.1	99.90
750	82.0	91.1	98.4	99.3	99.90

Pourcentage de points à l'intérieur des ellipses pour 5 seuils C.
Note : pour chaque classe d'effectifs, 5000 échantillons ont été générés et analysés.

Dans les graphiques que nous présentons, nous prendrons, sauf contre indication C = .99 et nous choisirons les échelles de telle sorte que l'ellipse à ce seuil soit représentée par un cercle.

II) Exemples d'hétérogénéités rencontrées classiquement et tracés correspondants :

En Hydrologie, les erreurs assez classiques sont les suivantes :

+ Erreur multiplicative

+ Erreur additive

+ Dérive dans le temps

+ Changement de corrélation

L'erreur multiplicative existe en pluviométrie , c'est , par exemple , l'utilisation d'une éprouvette mal adaptée à la surface réceptrice d'un pluviomètre. Dans ce cas , on montre que l'espérance mathématique de Z_k n'est pas nulle mais décrit en quelque sorte un triangle (cf Figure 3) .

L'erreur additive (cf Figure 4) se rencontre en mesures de température lors d'un déplacement d'une station de mesure . Le tracé présente les mêmes caractères que précédemment .

La dérive dans le temps existe parfois en mesure d'énergie (vieillissement de corps noir) ou peut être provoquée par la dérive d'appareils électroniques ou par d'autres causes (par exemple , le grandissement d'obstacles , type végétation en mesure de durée d'insolation ou même de pluie). Le point représentant l'espérance mathématique de Z_k décrit une sorte d'ellipse (cf Figure 5) .

Quant au changement de corrélation , nous l'avons déja observé lorsque l'observation est de qualité très différente d'une période à l'autre . L'espérance mathématique de Z_k est nulle quelque soit k , mais la variance de Z_k n'a pas la même expression d'une période à l'autre (cf Figure 6) .

Les figures suivantes obtenues par simulation d'erreurs dans des séries générées stochastiquement illustrent ces divers cas de figure .

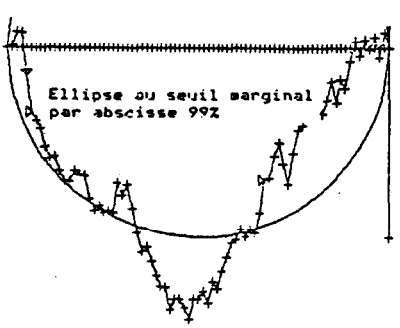

Figure 3 :
Erreur Multiplicative

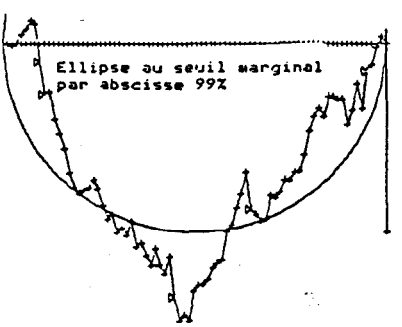

Figure 4 :
Erreur Additive

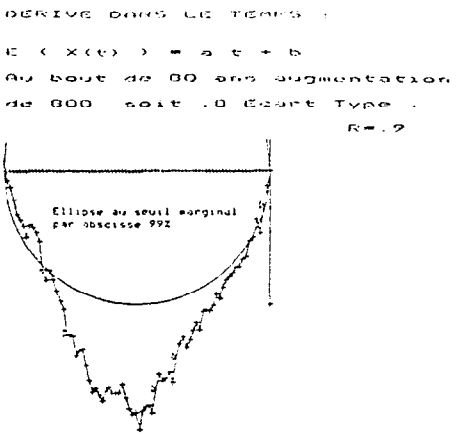

Figure 5 :
Dérive dans le temps

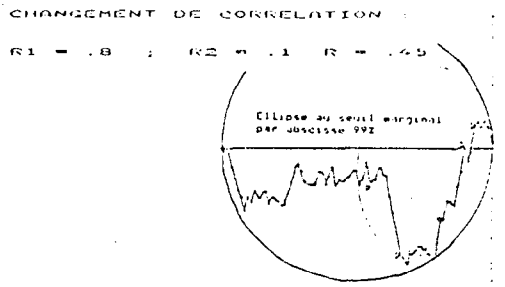

Figure 6 :
Changement de corrélation

III) Evaluation par simulation stochastique des probabilités de sortie globale pour des séries réputées bonnes :

Le calcul permettant de définir l'ellipse de contrôle est fondé sur la probabilité C de ne pas sortir du segment pour une valeur donnée de k ; il est donc interessant d'avoir un renseignement plus global, à savoir, **quelle est la probabilité, pour un seuil donné C qu'aucun point ne sorte de cette ellipse, toutes valeurs de k confondues**.

Si les cumuls étaient indépendants, le calcul serait aisé, puisque ce serait le produit des probabilités, mais ils ne le sont pas puisque l'on passe d'un cumul au suivant en ajoutant une valeur petite par rapport au cumul dès que l'on se trouve un peu éloigné des extrémités.

Le calcul fait appel à des probabilités conditionnelles et nous n'avons pu le résoudre de façon satisfaisante ; J. Bernier (1978) dans son article fournit des bornes.

Comme nous désirions des ordres de grandeur pour des cas rencontrés en Hydrologie, nous avons procédé par une méthode de Monte Carlo. Nous avons pris comme variables de calcul :

+ la taille N de l'échantillon
+ la valeur C du seuil de probabilité

Puis , pour chaque valeur de N , de N=10 à N=750 , nous avons simulé par tirage aléatoire et construction de séries binormales corrélées et réputées bonnes , 5000 échantillons de taille N . Nous avons ensuite compté le nombre de séries où tous les points du cumul restent à l'intérieur de l'ellipse de contrôle au seuil C , avec 5 valeurs différentes de C .

Les résultats de cette simulation sont fournis par la figure 7 . On constate évidemment que cette probabilité est bien inférieure à C , ce qui est d'ailleurs évident puisqu'il y a déja une probabilité 1 - C que le premier point sorte de l'ellipse

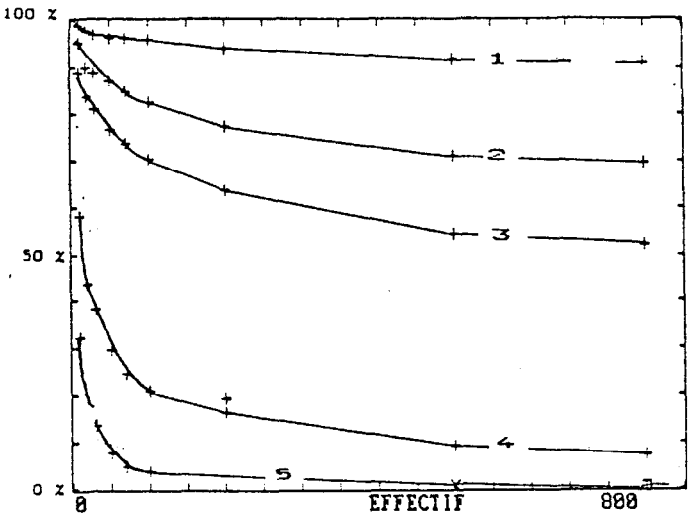

Figure 7 : Récapitulatif des Pourcentages de traçés sans sortie pour 5 seuils C :

1 : Seuil C = 99.8 %
2 : Seuil C = 99.0 %
3 : Seuil C = 98.0 %
4 : Seuil C = 90 %
5 : Seuil C = 80 %

Par exemple, pour une série de 100 valeurs, il y a 82 chances sur 100 qu'il n'y ait pas plus d'un point à l'extérieur de l'ellipse de contrôle à 99 % .

C'est pourquoi nous avons également compté le nombre d'échantillons où au maximum deux points étaient sortis . La figure 8 donne les résultats pour la valeur de C = 99 % .

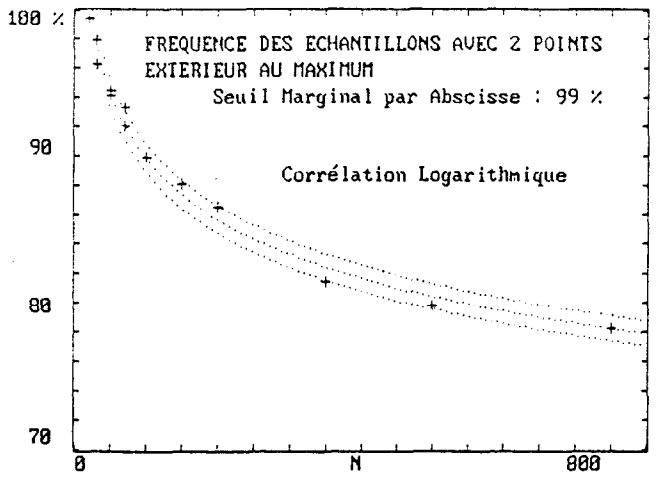

Figure 8 : Fréquence des échantillons avec au maximum 2 points à l'extérieur de l'ellipse de contrôle C = 99 % .

Par exemple, pour une série de 100 valeurs, il y a 82 chances sur 100 qu'il n'y ait pas plus de deux points à l'extérieur de l'ellipse de contrôle à 99 %.

Application pratique :

Dans la pratique , quelques problèmes importants sont à résoudre :

+ Choix de la variable témoin :

C'est évidemment un choix important ; on peut avoir des idées à partir de critiques historiques , de contrôle sur le terrain et dans les archives . Sinon , si l'on n'a aucune certitude de posséder une station de référence , on peut prendre comme variable de référence , non plus une station , mais une combinaison linéaire de stations, ne comprenant évidemment pas la station à tester . Cette combinaison peut être définie, par exemple , par les premières composantes d'une analyse en composantes principales , ce qui aura pour effet d'éliminer en quelque sorte les stations hétérogènes , plutôt moins bien corrélées avec les autres . Le résidu analysé pourra être alors le résidu de la corrélation multiple entre la station à tester et quelques composantes principales .

+ Exploitation des résultats :

Supposons que le tracé mette en évidence une anomalie . On pourra ensuite découper la série en sous-séries , vérifier que ces sous-séries sont homogênes et comparer alors les équations de régression entre la variable à tester et la variable témoin . Cet examen pourra permettre dans certains cas de déceler le type d'erreur :

Si seul le terme constant varie significativement d'une sous série à l'autre , il s'agit plutôt d'une erreur de type additif (décalage de l'origine des mesures , déplacement d'une mesure thermométrique etc ...) .

Si le coefficient de régression est significativement différent , on peut avoir décelé une erreur de type multiplicatif (changement de réceptacle d'un pluviomètre , par exemple) .

Il est alors possible , avec une certaine prudence , d'envisager des corrections .

Conclusions :

L'utilisation de la méthode du cumul des résidus pour le contrôle des séries chronologiques est un outil puissant et partioulièrement bien adapté aux moyens actuels de microinformatique : calculs rapides , tracés aisés et aspect conversationnel .

Cette méthode est sensible ; elle utilise en effet l'information chronologique (comme la méthode des doubles cumuls) , l'information tirée de la corrélation (ce qui n'est pas le cas de la méthode des doubles cumuls) et fournit de plus une appréciation probabiliste sur l'hypothèse de stationnarité .

Bibliographie :

Bernier J. -1977 - Etude de la stationnarité des séries hydrométéorologiques . La Houille Blanche N° 4 - 1977 pp 313-319

Bois Ph. - 1971 - .Une méthode de contrôle de séries chronologiques utilisées en climatologie et en hydrologie . Publication du Laboratoire de Mécanique des Fluides . Université de Grenoble (Mai 1971) . 30 pages .

Bois Ph . - 1976 - . Contribution à l'analyse et à la prévision de variables hydrométéorologiques Applications à la prévision des débits du Niger et des avalanches à Davos . Thèse de Doctorat ès-Sciences . Institut National Polytechnique de Grenoble . Septembre 1976 . 218 pages .

Hinkley D. V. - 1971 - . Inference about the change-point from cumulative sum tests. Biometrika N° 58-3 .

Karlin S. - 1966 - . A first course in stochastic processes . Academic Press . p. 281 Problème 1 .