ORSTOM - DEC
UR 2

ICCARE

RAPPORT N°3

CARACTERISATION DE FLUCTUATIONS DANS UNE SERIE CHRONOLOGIQUE PAR APPLICATIONS DE TESTS STATISTIQUES

ETUDE BIBLIOGRAPHIQUE

ORSTOM-MONTPELLIER

Juin 1994

Caractérisation de fluctuations dans une série chronologique par applications de tests statistiques Etude Bibliographique

---

Programme ICCARE
Rapport n°3

H. LUBES
J. M. MASSON
E. SERVAT
J. E. PATUREL
B. KOUAME
J. F. BOYER

JUIN 1994

INTRODUCTION

La caractérisation d'éventuelles fluctuations d'ordre climatique en Afrique de l'Ouest non Sahélienne repose entre autres analyses sur l'étude de séries chronologiques de données de pluie et de débit à différents sites de mesures, les plus nombreux possibles, sur des périodes les plus longues possibles.

Le présent rapport qui s'inscrit dans la première phase du programme Iccare a pour but de spécifier les traitements statistiques qui sont a priori envisagés pour analyser les séries chronologiques de données hydro-pluviométriques des différents postes répertoriés pour le programme. Ces méthodes seront appliquées pour étudier les précipitations sur le pas de temps de l'année, de la saison (humide et sèche), voire de la décade. Ces traitements pourront également être préconisés pour l'étude de débits moyens annuels et éventuellement d'autres grandeurs de débit à définir.

Les variations au sens le plus général du terme qui seront éventuellement décelées ne seront pas systématiquement qualifiées de climatiques. Une interprétation prudente des résultats devra être donnée, et ce d'autant plus que les séries étudiées seront de courte durée.

Les méthodes statistiques dont il est fait état dans ce rapport concernent l'exploitation d'une série de données et une seule. Ce type d'analyse sera qualifié de ponctuel ou "par site".

Les techniques statistiques relatives à une analyse "régionale" ou "zonale" c'est-à-dire impliquant plusieurs postes feront l'objet d'une deuxième synthèse.

La nature des données traitées et les termes du problème posé étant ainsi rappelés, la première partie de ce rapport est consacrée à une synthèse bibliographique. Dans un deuxième temps, les méthodes retenues sont précisées.

SYNTHESE BIBLIOGRAPHIQUE

La littérature consacrée à l'approche statistique de séries chronologiques de variables hydro-météorologiques est particulièrement abondante.

D'après Kendall et Stuart (1943), l'analyse d'une série temporelle a pour but d'améliorer la compréhension des mécanismes statistiques générateurs de cette série d'observations. Cet objectif ne peut être atteint en considérant une seule série de données. En effet toute série chronologique peut n'être qu'une représentation partielle d'un phénomène complexe générant un nombre substantiel de séries différentes.

Il n'en demeure pas moins qu'une exploitation "par site" s'impose avant de procéder à des interprétations prenant en compte la dimension spatiale des phénomènes générateurs des dites séries ponctuelles.

Tous les auteurs s'accordent pour décomposer une série temporelle typique en quatre parties :

- une tendance
- une périodicité : oscillations plus ou moins régulières autour d'une tendance
- une autocorrélation ou un effet de mémoire : la grandeur d'une observation est dépendante de la magnitude des observations précédentes
- une composante aléatoire, non systématique, irrégulière, c'est-à-dire due au hasard.

Toute série peut être représentée par l'un ou plusieurs de ces constituants.

L'étude des séries temporelles est consacrée en grande partie à isoler et à analyser séparément chacune des composantes constitutives de la série.

La décomposition d'une série est très souvent utile. Toutefois elle ne peut se faire que compte tenu d'hypothèses fortes, et *a priori*, concernant le caractère linéaire ou non du système générateur de ladite série chronologique. Cette phase de modélisation ne paraît pas s'imposer dans le cadre du présent programme. Elle est surtout utilisée dans le cas de variables de type économique.

Il semble préférable de lui substituer la mise en oeuvre de tests statistiques, les plus robustes possibles, spécifiques de l'une ou l'autre des composantes précitées.

Les tests qui vont être présentés sont extraits en grande partie de la note technique n°79 "Climatic change" de l'Organisation Mondiale de la Météorologie (WMO, 1966), et de Kendall et Stuart (1943).

La première catégorie de tests concerne le caractère aléatoire des séries. Dans l'hypothèse où la série est déclarée non aléatoire des tests sont requis pour tenter de caractériser la nature "non aléatoire" présente dans la série. Les tests relatifs à la détection de point de rupture *a priori* à date inconnue termineront cette présentation.

CARACTERE ALEATOIRE DES SERIES

Les tests les plus répandus portent sur la constance de la moyenne de la série tout au long de sa période d'observations.

Ces tests sont en général assez puissants pour faire une distinction entre le caractère aléatoire et le caractère non aléatoire de la série. En revanche tous ne permettent pas d'identifier une alternative à la constance du type tendance, discontinuité brutale, oscillations... Seuls quelques-uns sont relativement puissants vis à vis d'une alternative spécifique, qui le plus souvent relève d'un changement brutal.

Quelques tests ont pour objet la constance de la dispersion de la série, c'est-à-dire qu'ils étudient si la variabilité de la série est uniforme dans le temps.

Les tests non paramétriques ne font pas d'hypothèse sur la nature de la distribution de probabilité de la variable définissant la série des observations. Les tests, paramétriques ou non, sont dits robustes lorsque leurs conditions d'application sont peu strictes (Kotz et al., 1981, vol. 8).

Soit la série chronologique (x_i), i=1,N, les x_i désignent les réalisations de la variable X observées à des pas de temps successifs égaux (Kotz et al., 1981, vol. 9).

L'hypothèse nulle est donc : "la série des (x_i), i=1,N, est aléatoire".

TEST DU RAPPORT DE VON NEUMANN (WMO, 1966 ; BUISHAND, 1982)

Il s'agit du rapport de la moyenne du carré des différences successives des valeurs observées à la variance.

On note V ce rapport :

$$V = \frac{N}{N-1} \frac{\sum_{i=1}^{N-1} (x_i - x_{i+1})^2}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}$$

Pour N grand (N>30), si la série est aléatoire, V est distribué selon une loi normale de moyenne $\frac{2N}{N-1}$ et de variance approximativement égale à $\frac{4(N-2)}{(N-1)^2}$.

Il en résulte que, si l'hypothèse nulle est vraie, la variable :

$$U = \frac{V - \frac{2N}{N-1}}{\frac{2\sqrt{N-2}}{N-1}}$$

possède une distribution normale réduite.

Pour un risque α de première espèce donné, la région d'acceptation de l'hypothèse nulle est comprise entre :

$$(V)_t^- = \frac{2N - 2U_{1-\alpha/2}\sqrt{N-2}}{N-1} \text{ et } (V)_t^+ = \frac{2N + 2U_{1-\alpha/2}\sqrt{N-2}}{N-1}.$$

Aucune hypothèse alternative spécifique n'est associée à ce test.

TEST DES POINTS DE REBROUSSEMENT (KENDALL ET STUART, 1943)

Ce test consiste à compter le nombre de pics et de creux présents dans la série. Le pic est défini comme une valeur qui est plus grande que les deux valeurs qui l'encadrent.

De manière similaire, un creux est une valeur qui est plus petite que ses deux valeurs voisines.

Les pics et les creux constituent les points de rebroussement de la série. Pics et creux peuvent être constitués de plusieurs valeurs égales.

Pour N grand (N>30), sous l'hypothèse nulle, la variable p, nombre de points de rebroussement, suit une distribution normale de

moyenne $\overline{p} = \frac{2(N-2)}{3}$ et de variance $\sigma_p^2 = \frac{16N - 29}{90}$.

Il en résulte que si l'hypothèse nulle est vraie, la variable U = $(p-\overline{p})/\sigma_p$ est une variable normale réduite.

Pour un risque α de première espèce donné, la région d'acceptation de l'hypothèse nulle est comprise entre :

$$\overline{p} - U_{1-\alpha/2}\sigma_p \text{ et } \overline{p} + U_{1-\alpha/2}\sigma_p.$$

Deux hypothèses alternatives sont admises pour ce test, celle de la tendance et celle de la périodicité. Toutefois d'autres tests sont plus appropriés pour arguer d'une tendance.

TEST DES CHANGEMENTS DE SIGNE (KENDALL ET STUART, 1943)

Ce test consiste à compter le nombre de différences premières positives de la série, c'est-à-dire le nombre d'intervalles sur lesquels la série est croissante. Les intervalles où il n'y a ni croissance, ni décroissance sont ignorés.

Pour N grand (N>30), sous l'hypothèse nulle, la variable c, nombre d'intervalles où la série est croissante suit une distribution normale de

moyenne $\overline{c} = \frac{N-1}{2}$ et de variance $\sigma_c^2 = \frac{N+1}{12}$.

Il en résulte que si l'hypothèse nulle est vraie, la variable U = (c-$\bar{c}$)/σ_c est une variable normale réduite.

Pour un risque α de première espèce donné, la région d'acceptation de l'hypothèse nulle est comprise entre :

$$\bar{c} - U_{1-\alpha/2}\,\sigma_c \text{ et } \bar{c} + U_{1-\alpha/2}\,\sigma_c.$$

Ce test a été principalement présenté comme un test dont l'hypothèse alternative est celle d'une tendance spécifiquement linéaire.

TEST DE CORRELATION SUR LE RANG (KENDALL ET STUART, 1943 ; WMO, 1966)

Ce test apparaît comme une amélioration apportée au test précédent. En effet il se propose de calculer le nombre de paires P pour lesquelles $x_j > x_i$, $j > i$, avec i = 1,....., N-1.

Pour N grand, sous l'hypothèse nulle, la variable τ telle que :

$$\tau = 1 - \frac{4Q}{N(N-1)} \text{ avec } Q = \frac{N(N-1)}{2} - P$$

suit une distribution normale de moyenne nulle et de variance égale à

$$\sigma_\tau^2 = \frac{2(2N+5)}{9N(N-1)}.$$

Il en résulte que si l'hypothèse nulle est vraie, la variable U = τ/σ_τ est une variable normale réduite.

Pour un risque α de première espèce donné, la région d'acceptation de l'hypothèse nulle est comprise entre :

$$-U_{1-\alpha/2}\,\sigma_\tau \text{ et } U_{1-\alpha/2}\,\sigma_\tau.$$

L'hypothèse alternative reconnue de ce test est celle d'une tendance.

Lorsque l'on s'intéresse directement à la distribution asymptotique de la variable P, ce test porte le nom de test de Mann-Kendall. Les deux formulations sont équivalentes.

STATISTIQUE DE RANG DE SPEARMAN (WMO, 1966)

La série des (x_i), i = 1, N est transformée en la série équivalente des rangs (k_i), i = 1, N correspondants.

Pour chaque terme de la série est calculée ensuite :

$$d_i = k_i - 1$$

qui désigne la différence entre le rang de l'élément x_i et sa position i dans la série.

Pour N grand (N>30), sous l'hypothèse nulle, la variable :

$$t = r_s\sqrt{\frac{N-2}{1-r_s^2}} \text{ avec}$$

$$r_s = 1 - \frac{6}{N(N^2-1)}\sum_{i=1}^{N} d_i^2$$

suit une distribution t de Student à N-2 degrés de liberté.

Pour un risque α de première espèce donné, la région d'acceptation de l'hypothèse nulle est comprise entre les valeurs théoriques de la variable de Student de probabilité de non-dépassement respectivement égale à $\alpha/2$ et $1-\alpha/2$.

Ce test est largement cité dans la littérature. L'hypothèse alternative est celle de la tendance. La puissance de ce test est comparable à celle du test de corrélation sur le rang.

TEST T DE STUDENT DE LA DIFFERENCE DE DEUX MOYENNES (WMO, 1966 ; CERESTA, 1986)

Ce test paramétrique est appliqué lorsque l'on suppose qu'un changement brutal est intervenu dans la série de telle sorte que celle-ci peut être découpée en deux sous-séries de moyennes significativement différentes.

Si l'on désigne les deux moyennes en question par $\overline{x}_1$ et $\overline{x}_2$ respectivement, et par N_1 et N_2 les nombres de valeurs ayant servi au calcul de chacune d'elles, sous l'hypothèse nulle, la variable :

$$t_d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

suit une distribution t de Student à N_1+N_2-2 degrés de liberté. S_1^2 et S_2^2 sont respectivement les variances estimées des deux sous-séries, les variances théoriques étant supposées égales.

Pour un risque α de première espèce donné, et N_1 et N_2 grands (>30), la région d'acceptation de l'hypothèse nulle est comprise entre les valeurs théoriques de la variable de Student de probabilité de non-dépassement respectivement égale à $\alpha/2$ et $1-\alpha/2$. Un test unilatéral peut être mis en oeuvre si l'on fait une hypothèse *a priori* sur le signe de $\overline{x}_1-\overline{x}_2$.

Ce test est robuste car il peut être appliqué à des données indépendamment de la nature de leur distribution de probabilité à cause du théorème Central Limite.

Il se révèle performant pour détecter une tendance. Il s'avère impuissant vis à vis de toute autre alternative.

TEST DE CRAMER (CRAMER, 1946, PP 389-390 ; WMO, 1966)

Ce test compare les moyennes de sous-séries avec la moyenne de la série totale. Il vérifie que la différence des moyennes est compatible avec l'hypothèse nulle de comportement aléatoire de la série.

On désigne par $\overline{x}$ et s respectivement la moyenne et l'écart-type de la série totale.

$\overline{x}_k$ est définie comme la moyenne de la sous-période de n valeurs, comparée à $\overline{x}$:

$$\overline{x}_k = \frac{\sum_{i=k+1}^{k+n} x_i}{n}$$

Si l'on définit :

$$\tau_k = \frac{(\overline{x}_k - \overline{x})}{s}$$

la variable :

$$t_k = \left[\frac{n(N-2)}{N-n(1+\tau_k^2)}\right]^{1/2} \tau_k$$

suit une distribution t de Student à N-2 degrés de liberté.

Pour un risque α de première espèce donné, et N grand (N>30), la région d'acceptation de l'hypothèse nulle est comprise entre les valeurs théoriques de la variable de Student de probabilité de non-dépassement respectivement égale à $\alpha/2$ et $1-\alpha/2$.

Ce test peut être répété autant de fois que de sous-périodes sont définies. Dans le cas où des périodes se chevauchent une attention toute particulière doit être portée sur l'interprétation du test.

TEST DE CONSTANCE DE LA VARIABILITE (TEST DE BARTLETT, WMO, 1966)

En sus ou à la place d'une inconstance dans la moyenne de la série, peut intervenir une variation dans sa dispersion.

Comme le soulignent Kendall et Stuart (1943), les tests statistiques ordinaires de variation de dispersion sont extrêmement sensibles à la forme de la distribution de fréquence des observations. Ces tests ne sont donc pas robustes. Leur utilisation n'est pas recommandée quand les données s'écartent même peu d'une distribution normale.

Si et seulement si la distribution des observations est normale, le test d'homogénéité des variances dit test modifié de Bartlett peut être appliqué.

Une version simplifiée de ce test est la suivante.

La série est divisée en k sous-périodes égales où $k \geq 2$. Sur chacune d'elles on calcule la variance estimée S_k^2.

On retient la plus grande et la plus petite valeur de ces variances. Le rapport $\frac{S_{max}^2}{S_{min}^2}$ est comparé à des valeurs critiques tabulées.

AUTOCORRELOGRAMME (WMO, 1966 ; CHATFIELD, 1989)

Une autre mesure du caractère aléatoire d'une série chronologique est donnée par le coefficient d'autocorrélation d'ordre 1, et plus généralement par l'autocorrélogramme. En effet on suppose qu'il existe des dépendances significatives entre les termes successifs d'une série non aléatoire.

Le coefficient d'autocorrélation d'ordre k est donné par l'expression :

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x}_1)(x_{t+k} - \bar{x}_2)}{\sqrt{\left[\sum_{t=1}^{N-k}(x_t - \bar{x}_1)^2 \sum_{t=1}^{N-k}(x_{t+k} - \bar{x}_2)^2\right]}}$$

avec $\bar{x}_1$ moyenne des observations (x_i), i = 1, N-k, et $\bar{x}_2$ moyenne des observations (x_i), i = k+1, N.

D'après Chatfield (1989), si une série chronologique est aléatoire, alors pour N grand, $r_k \approx 0$ pour toute valeur de k non nulle. En fait pour une série

chronologique aléatoire, et pour N grand, r_k suit approximativement une distribution normale de moyenne nulle et de variance 1/N.

Il est donc possible de définir une région de confiance contenant pour un seuil de confiance donné, sous l'hypothèse nulle, l'autocorrélogramme. Pour un seuil de confiance 1-α/2 donné, la région de confiance est définie par

$\pm\frac{U_{1-\alpha/2}}{\sqrt{N}}$. U désigne la variable normale réduite.

Une importance particulière doit être donnée au comportement de l'autocorrélogramme pour de faibles valeurs de k, notamment pour k = 1 (WMO, 1966). En effet sur les vingt premières valeurs de r_k, il n'est pas rare qu'une valeur sorte de la région de confiance même lorsque la série est réellement aléatoire. Ceci souligne les difficultés d'interprétation de l'autocorrélogramme, indépendamment des conditions d'application relatives à la linéarité des liaisons, et à la structure des écarts à ces liaisons.

TESTS DE DETECTION DE RUPTURES

Les tests qui vont être présentés ci-après sont plus particulièrement adaptés à la détection de ruptures dans une série chronologique. Une rupture peut être définie de façon générale par un changement dans la loi de probabilité de la série chronologique à un instant donné, le plus souvent inconnu.

TEST DE MANN-WHITNEY (PETTITT, 1979 ; CERESTA, 1986)

Le fondement du test de Mann-Whitney est le suivant (Ceresta, 1986).

La série étudiée est divisée en deux sous-échantillons respectivement de taille m et n.

Les valeurs des deux échantillons sont regroupées et classées par ordre croissant. On calcule alors la somme des rangs des éléments de chaque sous-échantillon dans l'échantillon total. Une statistique est définie à partir des deux sommes ainsi déterminées, et testée sous l'hypothèse nulle d'appartenance des deux sous-échantillons à la même population.

La formulation du test de Mann-Whitney modifié par Pettitt (Pettitt, 1979) est la suivante.

L'hypothèse nulle du test est l'absence de rupture dans la série.

La mise en oeuvre du test suppose que pour tout instant t variant de 1 à N, les séries (x_i), i = 1, t et (x_j), i = t+1, N appartiennent à la même population.

Soit $D_{ij} = sgn(x_i-x_j)$ avec sgn(x) = 1 si x>0, 0 si x = 0, -1 si x<0.

On considère la variable $U_{t,N}$ telle que :

$$U_{t,N} = \sum_{i=1}^{t} \sum_{j=t+1}^{N} Dij$$

Soit K_N la variable définie par le maximum en valeur absolue de $U_{t,N}$ pour t variant de 1 à N-1.

Si k désigne la valeur de K_N prise sur la série étudiée, sous l'hypothèse nulle, la probabilité de dépassement de la valeur k est donnée approximativement par :

$$\Pr ob(K_N > k) \approx 2\exp(-6k^2/(N^3+N^2))$$

Pour un risque α de première espèce donné, si $\Pr ob(K_N > k)$ est inférieur à α, l'hypothèse nulle est rejetée.

Ce test est réputé pour sa robustesse.

TEST DU RAPPORT DE VRAISEMBLANCE (BUISHAND, 1982, 1984)

Ce test fait référence au modèle simple qui suppose un changement de moyenne de la série :

$$x_i = \begin{cases} \mu + \varepsilon_i & i = 1,.....,m \\ \mu + \Delta + \varepsilon_i, & i = m+1,.....,n \end{cases}$$

Les ε_i sont des variables aléatoires normales de moyenne nulle et de variance commune inconnue σ^2. Le point de rupture m et les paramètres μ et Δ sont aussi inconnus.

Plusieurs méthodes statistiques ont été développées pour tester l'hypothèse nulle $\Delta = 0$ contre l'hypothèse alternative $\Delta \neq 0$.

On s'intéresse aux termes de cumul d'écarts suivants :

$$S^*_k = \sum_{i=1}^{k} (x_i - \bar{x}) \text{ pour k = 1,....., N}$$

$$S^*_0 = 0$$

$\bar{x}$ est la moyenne des valeurs $x_1, x_2, ..., x_N$.

S^*_k est tel que :

$$E(S^*_k) = -k(N-m)N^{-1}\Delta,\ \mathrm{k} = 0,\ldots\ldots,m$$
$$E(S^*_k) = -m(N-k)N^{-1}\Delta,\ \mathrm{k} = m+1,\ldots,N$$
$$\mathrm{var}(S^*_k) = k(N-k)N^{-1}\sigma^2,\ \mathrm{k} = 0,\ldots\ldots,N$$

On observe que la moyenne des S^*_k est nulle pour une série homogène ($\Delta = 0$), positive pour $\Delta < 0$ et négative pour $\Delta > 0$. La variance est maximale si k = N/2. Même pour une série purement aléatoire, les valeurs de S^*_k peuvent différer considérablement de zéro, spécialement pour k au voisinage de N/2.

Le test du rapport de vraisemblance porte sur la variable :

$$V = \max\left\{\left|S^*_k\right| / \left[D_x\left\{k(N-k)\right\}^{1/2}\right]\right\} \text{pour } 1 \le k \le N-1$$

avec D_x écart-type de la série.

De grandes valeurs de V conduisent à rejeter l'hypothèse nulle. Les valeurs critiques de la statistique V peuvent être obtenues à partir des valeurs critiques de la statistique W telle que :

$$W = (N-2)^{1/2} V / (1-V^2)^{1/2}$$

qui ont été tabulées par Worsley (1979).

La statistique V est fortement dépendante de l'hypothèse de normalité sur la distribution de la variable étudiée.

STATISTIQUE U (BUISHAND, 1982, 1984)

Le test ici présenté est de nature Bayesienne.

Il fait référence au même modèle de base et aux mêmes termes que le test précédent.

En supposant une distribution *a priori* uniforme pour la position du point de rupture m, la statistique U est définie par :

$$U = \left[N(N+1)\right]^{-1} \sum_{k=1}^{N-1} \left(S_k^* / D_x\right)^2$$

Des valeurs critiques de la statistique U sont données par Buishand (1982) à partir d'une méthode de Monte Carlo.

La statistique U donne moins de poids que la statistique V (test précédent) aux premières et dernières valeurs de la série. En conséquence,

la statistique V est supérieure à la statistique U pour déceler un changement de moyenne en début et en fin de série. Pour tout changement de moyenne survenant au milieu de la série, la statistique U s'avère plus performante.

De plus la statistique U est une statistique robuste qui reste valide même pour des distributions de la variable étudiée qui s'écartent de la normalité.

Remarque : Les tests basés sur les écarts cumulés ont des propriétés optimales dans le cas de changements brutaux de moyenne.

ELLIPSE DE CONTROLE

On rappelle ci-après le modèle de base :

$$x_i = \begin{cases} \mu + \varepsilon_i & i = 1,.....,m \\ \mu + \Delta + \varepsilon_i, & i = m+1,.....,N \end{cases}$$

Les ε_i sont des variables aléatoires normales de moyenne nulle et de variance commune inconnue σ^2. Le point de rupture m et les paramètres μ et Δ sont aussi inconnus. L'hypothèse nulle est $\Delta = 0$.

Soit par ailleurs la définition des variables suivantes :

$$S^*_0 = 0$$

$$S^*_k = \sum_{i=1}^{k} (x_i - \bar{x}) \text{ pour k = 1,....., N}$$

$\bar{x}$ est la moyenne des valeurs $x_1, x_2, ..., x_N$.

S^*_k est telle que :

$$E(S^*_k) = -k(N-m)N^{-1}\Delta,\ \mathrm{k} = 0,.....,m$$
$$E(S^*_k) = -m(N-k)N^{-1}\Delta,\ \mathrm{k} = m+1,....,N$$
$$\mathrm{var}(S^*_k) = k(N-k)N^{-1}\sigma^2,\ \mathrm{k} = 0,.....,N$$

Sous l'hypothèse nulle d'homogénéité de la série, la variable S_k^* suit une distribution normale de moyenne nulle et de variance $k(N-k)N^{-1}\sigma^2$, $\mathrm{k} = 0,.....,N$.

σ^2 inconnue est remplacée par son estimateur à partir de la série étudiée.

Il en résulte que sous l'hypothèse nulle, la variable S_k^* suit une distribution normale de moyenne nulle et de variance approximative

$k(N-k)(N-1)^{-1}D_x^2$, $k = 0,\ldots\ldots,N$ avec $D_x^2 = N^{-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$.

Il est donc possible de définir une région de confiance contenant pour un seuil de confiance donné, sous l'hypothèse nulle, la série des S_k^*. Pour un seuil de confiance 1-α/2 donné, la région de confiance est définie par

$\pm\frac{U_{1-\alpha/2}\sqrt{k(N-k)}}{\sqrt{(N-1)}}D_x$. U désigne la variable normale réduite. Cette région de confiance est appelée ellipse de contrôle. Une relation entre la valeur du seuil de confiance et un nombre maximum admissible de points hors de l'ellipse pour que la série étudiée soit déclarée homogène peut être estimée par simulation, comme cela a déjà été fait pour les résidus d'une régression (Bois, 1971, 1986).

PROCEDURE BAYESIENNE (KOTZ ET AL., 1981, VOL. 1)

Le modèle de base de la procédure est le suivant :

$$x_i = \begin{cases} \mu+\varepsilon & i = 1,\ldots\ldots,\tau \\ \mu+\delta+\varepsilon, & i = \tau+1,\ldots\ldots,N \end{cases}$$

Les ε_i sont indépendants et normalement distribués, de moyenne nulle et de variance σ^2. τ, μ, δ et σ sont des paramètres inconnus, $1 \le \tau \le N-1, -\infty < \mu < \infty, -\infty < \delta < \infty, \sigma > 0$. τ, δ, μ, σ sont indépendants.

τ et δ représentent respectivement la position dans le temps et l'amplitude d'un changement éventuel de moyenne.

L'approche Bayesienne présentée ici est fondée sur les distributions marginales *a posteriori* de τ et δ (Lee and Heghinian, 1977).

Les distributions *a priori* de τ et δ sont :

$p(\tau)=1/(N-1)$, $\tau = 1, 2, \ldots\ldots, N-1$

$p(\delta)$ est normale de moyenne nulle et de variance σ_δ^2.

La distribution *a posteriori* de τ est définie par :

$$p(\tau|x) \propto \left[N/(\tau(N-\tau))\right]^{1/2}\left[R(\tau)\right]^{-(N-2)/2},\ 0 \le \tau \le N-1 \text{ avec}$$

$$R(\tau) = H(\tau) / \sum_{i=1}^{N} (x_i - \bar{x}_N)^2$$

$$= \left[\sum_{i=1}^{\tau} (x_i - \bar{x}_\tau)^2 + \sum_{i=\tau+1}^{N} (x_i - \bar{x}_{N-\tau})^2 \right] / \sum_{i=1}^{N} (x_i - \bar{x}_N)^2$$

$$\bar{x}_N = 1/N \sum_{i=1}^{N} x_i, \ \bar{x}_\tau = 1/\tau \sum_{i=1}^{\tau} x_i, \ \bar{x}_{N-\tau} = 1/(N-\tau) \sum_{i=\tau+1}^{N} x_i.$$

La distribution *a posteriori* de δ est définie à partir de la distribution *a posteriori* de τ, $p(\tau|x)$, et de la distribution conditionnelle *a posteriori* de δ par rapport à τ, $p(\delta|\tau,x)$:

$$p(\delta|x) = \sum_{\tau=1}^{N-1} p(\delta|\tau,x) p(\tau|x)$$

La distribution conditionnelle *a posteriori* de δ par rapport à τ, $p(\delta|\tau,x)$, est une distribution de Student de moyenne $\hat{\delta}_\tau = \bar{x}_{N-\tau} - \bar{x}_\tau$ et de variance $\sigma_{\delta|\tau}^2 = NH(\tau)/[(N-2)(\tau(N-\tau))]$ avec ν = N-2 degrés de liberté.

La fonction densité de probabilité de cette loi de Student est la suivante :

$$p(\delta|\tau,x) = \frac{\nu^{\nu/2}\Gamma((\nu+1)/2)}{\Gamma(1/2)\Gamma(\nu/2)(\sigma_{\delta|\tau}^2)^{1/2}} \frac{1}{(\nu + (\delta - \hat{\delta}_\tau)^2/\sigma_{\delta|\tau}^2)^{(\nu+1)/2}}$$

Le changement éventuel, position et amplitude, correspond au mode des distributions *a posteriori* de τ et δ. La méthode fournit donc la probabilité que le changement se produise au moment τ dans une série où on suppose *a priori* qu'il y a effectivement un changement à un moment indéterminé. De même elle donne une estimation de la probabilité que l'amplitude du changement ait la valeur δ.

PROCEDURE DE SEGMENTATION DES SERIES HYDROMETEOROLOGIQUES

Une procédure de segmentation de séries hydrométéorologiques a été présentée par Hubert et al. (1989).

Le principe de cette procédure est de "découper" la série en m segments (m>1) de telle sorte que la moyenne calculée sur tout segment soit significativement différente de la moyenne du (ou des) segment(s) voisin(s). Une telle méthode est appropriée à la recherche de multiples changements de moyenne.

La segmentation est définie de la façon suivante.

Toute série x_i, $i = i_1, i_2$ avec $i_1 \geq 1$ et $i_2 \leq N$ où $(i_1 < i_2)$ constitue un segment de la série initiale des (x_i), $i = 1,.....N$.

Toute partition de la série initiale en m segments est une segmentation d'ordre m de cette série.

A partir d'une segmentation particulière d'ordre m pratiquée sur la série initiale, on définit :

i_k, $k = 1, 2,....., m$, le rang dans la série initiale de l'extrémité terminale du $k^{ième}$ segment ;

$n_k = i_k - i_{k-1}$, la longueur du $k^{ième}$ segment ;

$\overline{x}_k$ la moyenne du $k^{ième}$ segment, $\overline{x}_k = \frac{\sum_{i=i_{k-1}+1}^{i=i_k} x_i}{n_k}$;

D_m, l'écart quadratique entre la série et la segmentation considérée, $D_m = \sum_{k=1}^{k=m} d_k$

avec $d_k = \sum_{i=i_{k-1}+1}^{i=i_k} (x_i - \overline{x}_k)^2$. Cet écart permet d'apprécier la proximité de la série et de la segmentation qui lui est appliquée.

La segmentation retenue au terme de la mise en oeuvre de la procédure doit être telle que pour un ordre m de segmentation donné, l'écart quadratique D_m soit minimum. Cette condition est nécessaire mais non suffisante pour la détermination de la segmentation optimale. Il faut lui adjoindre la contrainte suivante selon laquelle les moyennes de deux segments contigus doivent être significativement différentes : $\overline{x}_k \neq \overline{x}_{k+1} \forall k = 1,2,....m-1$. Cette contrainte est satisfaite par application du test de Scheffé qui repose sur le concept de contraste (Dagnélie, 1970).

Par conséquent si lors du processus de segmentation d'ordre m+1, aucune segmentation produite n'est valide au sens du test de Scheffé, la segmentation de la série qui est retenue en tant que meilleure segmentation est la segmentation optimale d'ordre m.

D'après les auteurs (Hubert et al., 1989), cette procédure de segmentation peut être regardée comme un test de stationnarité, "la série étudiée est stationnaire" constituant l'hypothèse nulle de ce test. Si la procédure ne produit pas de segmentation acceptable d'ordre supérieur ou égal à 2, l'hypothèse nulle est acceptée. Aucun niveau de signification n'a été attribué à ce test.

METHODES RETENUES DANS ICCARE

Les divers tests qui viennent d'être présentés ne constituent en aucun cas une liste exhaustive des procédures qui ont pour objectif d'analyser "les traits" d'une série chronologique. Toutefois ont été recensés les tests les plus utilisés et les plus argumentés dans la littérature.

Parmi les méthodes présentées au paragraphe précédent, ont été retenus pour l'étude des séries chronologiques du programme Iccare, les tests suivants.

TEST DE CORRELATION SUR LE RANG

Ce test est intéressant du point de vue de son hypothèse alternative qui est celle d'une tendance. De plus ce test s'est révélé satisfaisant pour détecter un changement de moyenne sur des séries aléatoires générées artificiellement avec perturbations (Bonneaud, 1994).

Il a en outre déjà été appliqué dans le contexte africain (Olaniran, 1991).

DETERMINATION DE L'AUTOCORRELOGRAMME

L'estimation de l'autocorrélogramme est incontournable comme première exploitation de toute série chronologique.

TEST DE PETTITT (VERSION MODIFIEE DE MANN-WHITNEY)

La réputation de ce test de détection de rupture dont il existe de multiples applications (Demarée, 1990 ; Sutherland et al., 1991 ; Vannitsem et Demarée, 1991) justifie qu'il soit ici retenu.

STATISTIQUE U DE BUISHAND

La robustesse de ce test et l'originalité de son fondement à partir d'une approche Bayesienne le rendent intéressant.

ELLIPSE DE CONTROLE

L'ellipse de contrôle est un complément graphique original au test de Buishand.

PROCEDURE BAYESIENNE DE LEE ET HEGHINIAN

La procédure Bayesienne de Lee et Heghinian a été appliquée à l'étude de la structure de la saison des pluies en Afrique Soudano-Sahélienne (Chaouche, 1988). Il semble tout à fait justifié de la mettre en oeuvre dans le cadre d'Iccare.

Déjà appliquée à des séries de précipitations et de débits de l'Afrique de l'Ouest (Hubert et Carbonnel, 1993), son utilisation est intéressante dans le programme Iccare.

Parmi les raisons pour lesquelles certaines méthodes n'ont pas été retenues, il faut citer :

- l'absence d'hypothèse alternative précise
- l'existence de tests similaires plus performants ; c'est ainsi que les procédures de détection de rupture ont été jugées plus pertinentes que le test t de Student de différence de deux moyennes ou le test de Cramer
- une hypothèse trop forte de normalité de la variable étudiée, c'est-à-dire un défaut de robustesse
- les conclusions relatives à une étude de simulation de séries aléatoires artificiellement perturbées (Bonneaud, 1994).

REFERENCES BIBLIOGRAPHIQUES

BOIS Ph., 1971. Une méthode de contrôle de séries chronologiques utilisées en climatologie et en hydrologie. Laboratoires de Mécanique des Fluides. Université de Grenoble. "Section hydrologie". 49 p.

BOIS Ph., 1986. Contrôle des séries chronologiques corrélées par étude du cumul des résidus. Deuxièmes journées hydrologiques de l'Orstom. Montpellier. pp 89-100.

BONNEAUD S., 1994. Méthodes de détection des ruptures dans les séries chronologiques. Projet industriel de fin d'études. Institut des Sciences de l'Ingénieur de Montpellier, filière Sciences et Technologies de l'Eau. Laboratoire d'Hydrologie et Modélisation. Université Montpellier II. 40 p.

BUISHAND T. A., 1982. Some methods for testing the homogeneity of rainfall records. Journal of Hydrology, vol. 58, pp 11-27.

BUISHAND T. A., 1984. Tests for detecting a shift in the mean of hydrological time series. Journal of Hydrology, vol. 58, pp 51-69.

Ceresta (Centre d'Enseignement et de Recherche de Statistique Appliquée), 1986. Aide-mémoire pratique des techniques statistiques pour ingénieurs et techniciens supérieurs. Revue de statistique appliquée, vol. XXXIV numéro spécial.

CHAOUCHE A., 1988. Structure de la saison des pluies en Afrique Soudano-Sahélienne. Thèse de l'Ecole Nationale Supérieure des Mines de Paris. 263 p.

CHATFIELD C., 1989. The analysis of time series. An introduction. Fourth edition. Chapman and Hall. 241 p.

CRAMER H., 1946. Mathematical methods of statistics. Princeton University Press, 368 p.

DAGNELIE P., 1970. Théorie et Méthodes Statistiques. Vol 2. Les presses agronomiques de Gembloux. 451 p.

KENDALL S. M., STUART A., 1943. The advanced theory of statistics. Charles Griffin Londres. 2ème volume, 690 p, 3ème volume, 585 p. dans l'édition de 1977.

KOTZ S., JOHNSON N. L., READ C. B., 1981. Encyclopedia of statistical sciences. New York, John Wiley. Vol. 1, pp197-205, vol. 8, pp 157-163, vol. 9, pp 244-255.

LEE A. F. S., HEGHINIAN S. M.,1977. A Shift Of The Mean Level In A Sequence Of Independent Normal random Variables-A Bayesian Approach-. Technometrics, vol. 19, n°4, pp 503-506.

OLANIRAN O. J., 1991. Evidence of climatic change in Nigeria based on annual series of rainfall of different daily amounts, 1919-1985. Climatic change, vol. 19, pp 319-341.

PETTITT A. N., 1979. A non-parametric approach to the change-point problem. Applied Statistics, 28, n°2, pp 126-135.

SUTHERLAND R. A., BRYAN R. B., OOSTWOUD WIJENDES D., 1991. Analysis of the monthly and annual rainfall climate in a semi-arid environment, Kenya. Journal of Arid Environments, vol. 20, pp 257-275.

VANNITSEM S., DEMAREE G., 1991. Détection et modélisation des sécheresses au Sahel. Hydrologie Continentale, vol. 6, n°2, pp 155-171.

WORLD METEOROLOGICAL ORGANIZATION, 1966. Climatic change, by a working group of the Commission for Climatology. World Meteorological Organization, WMO 195, TP 100, Tech. Note n°79 : 78 p.

WORSLEY K. J., 1979. On the Likelihood Ratio Test for a Shift in Location of Normal Populations. Journal of the American Statistical Association, vol. 74, n°366, pp 365-367.

TABLE DES MATIERES

INTRODUCTION 2
SYNTHESE BIBLIOGRAPHIQUE 2
CARACTERE ALEATOIRE DES SERIES 3
TEST DU RAPPORT DE VON NEUMANN 4
TEST DES POINTS DE REBROUSSEMENT 5
TEST DES CHANGEMENTS DE SIGNE 5
TEST DE CORRELATION SUR LE RANG 6
STATISTIQUE DE RANG DE SPEARMAN 6
TEST T DE STUDENT DE LA DIFFERENCE DE DEUX MOYENNES 7
TEST DE CRAMER 8
TEST DE CONSTANCE DE LA VARIABILITE 9
AUTOCORRELOGRAMME 9
TEST DE DETECTION DES RUPTURES 10
TEST DE MANN-WHITNEY 10
TEST DU RAPPORT DE VRAISEMBLANCE 11
STATISTIQUE U DE BUISHAND 12
ELLIPSE DE CONTROLE 13
PROCEDURE BAYESIENNE 14
PROCEDURE DE SEGMENTATION DES SERIES HYDROMETEOROLOGIQUES 15
METHODES RETENUES DANS ICCARE 17
TEST DE CORRELATION SUR LE RANG 17
DETERMINATION DE L'AUTOCORRELOGRAMME 17
TEST DE PETTITT 17
STATISTIQUE U DE BUISHAND 17
ELLIPSE DE CONTROLE 17
PROCEDURE BAYESIENNE DE LEE ET HEGHINIAN 17
PROCEDURE DE SEGMENTATION DES SERIES HYDROMETEOROLOGIQUES 18
REFERENCES BIBLIOGRAPHIQUES 19
TABLE DES MATIERES 21