

Statistique et hydrologie à travers quelques exemples appliqués à des problématiques des pays en voie de développement

Hélène Lubès-Niel^a Hayet Khodja ^a Robert Sabatier ^b

Introduction

Le présent article a pour objectif d'illustrer l'implication de la statistique dans des programmes hydrologiques tournés vers le développement en coopération. Le séminaire "Statistique impliquée" (Laloë, 1992) poursuivait ce même but étendu à des disciplines scientifiques très diverses, et mettait d'abord en lumière des problématiques de recherche pour lesquelles le recours à des domaines variés de la statistique, plus ou moins traditionnels, pouvait apporter des éléments de réponse, voire des solutions aux questions posées par une spécialité donnée. C'est cette même volonté de donner la priorité au problème et à la démarche suivie plutôt qu'à l'outil lui-même qui a conditionné la rédaction du papier présenté ici. Le choix des illustrations qui vont suivre montre souvent que les techniques statistiques même les plus simples contribuent très significativement à "résoudre" des problèmes dont la complexité relève de la réalité naturelle des phénomènes hydrologiques pour lesquels tous les instruments d'observation actuels ne donnent qu'une image tronquée et souvent déformée.

Quatre thèmes hydrologiques au cœur des problématiques affichées dans les pays en voie de développement ont été retenus : les effets de la déforestation sur le comportement hydrologique, la prévision des crues, la variabilité spatiale et temporelle des précipitations, et la caractérisation des variations des régimes de pluie. Ils sont représentés ici au travers de quatre programmes de recherche qui s'inscrivent dans des contextes géographiques

^aLaboratoire d'Hydrologie, ORSTOM, B.P. 5045, 34032 Montpellier cedex 1, France

^bLaboratoire de Physique Moléculaire et Structurale, Faculté de Pharmacie, Université Montpellier I, 34060 Montpellier cedex 1, France

différents et ont tous mis en oeuvre une démarche statistique. Les trois premiers problèmes ont été traités par des hydrologues plus ou moins au fait des techniques statistiques et de leur pratique, tandis que le dernier a fait l'objet d'un dialogue entre hydrologues et statisticiens praticiens (Escoufier, 1992).

Les synthèses des trois premières études ont été rédigées à partir d'ouvrages ou d'articles de référence. Il s'agit essentiellement d'une compilation de citations de ces documents dont nous espérons n'avoir pas dénaturé le sens global. La dernière étude qui a fait l'objet d'un récent article (Khodja *et al*, 1997) est présentée de manière plus détaillée.

Les effets de la déforestation sur le comportement hydrologique (Fritsch, 1990)

A l'origine de cette étude il faut placer l'idée généralement répandue que la forêt exerce un effet modérateur sur les écoulements, et que par conséquent son élimination peut considérablement modifier les termes du bilan hydrologique. Pourtant de nombreux travaux menés dans des pays Anglo-Saxons ne corroborent pas ce qui pourrait passer pour une évidence.

Les effets du défrichement de la forêt amazonienne, et de la mise en culture, sur le comportement d'unités hydrologiques ont donc été analysés grâce au suivi hydrologique d'une expérimentation sur des petits bassins versants de Guyane Française, en zone tropicale humide, pendant près de huit années. L'objectif était de quantifier les modifications du comportement hydrologique provoquées par la suppression de la forêt naturelle amazonienne, et induites par des utilisations ultérieures de l'espace, à finalités agricoles, sylvicoles et pastorales.

Le dispositif expérimental soigneusement élaboré était composé de dix petits bassins versants d'une superficie comprise entre 1 et 2 hectares.

Le protocole expérimental comprenait une première phase de deux ans dite "Epoque Verte" durant laquelle les observations hydrologiques ont porté sur le milieu naturel forestier. Ensuite la couverture générale de huit bassins versants, appelés bassins expérimentaux, a fait l'objet de diverses manipulations : défrichement en "Epoque Rouge" et implantation ultérieure d'aménagements agricoles, sylvicoles ou pastoraux en "Epoque Jaune", tandis que deux bassins appelés bassins témoins, ont été conservés en forêt naturelle durant tout le suivi.

Pour évaluer les effets provoqués par une modification des caractéristiques d'un bassin versant, deux sources d'informations distinctes ont dû

être confrontées :

- les débits réellement observés sur le bassin expérimental après l'application du traitement (défrichement, nouvel aménagement),
- les débits que l'on aurait observés dans les mêmes conditions climatiques, si le système n'avait pas été modifié. Le résultat recherché est obtenu par différence entre ces deux termes.

Les écoulements effectivement mesurés pendant les "Epoques Rouge" et "Jaune" de l'expérimentation ont été donc comparés aux écoulements qui auraient été observés dans le même temps si le bassin considéré avait conservé sa couverture forestière originelle. La méthode de comparaison adoptée repose sur le principe des bassins appariés (paired watershed technique) qui fait largement appel à la régression linéaire. Pour chaque bassin expérimental, la meilleure régression a été recherchée dans les deux bassins témoins en utilisant les données de "l'Epoque Verte". Deux variables ont été considérées, dans un premier temps la variable "écoulement total de la crue", puis la variable "écoulement à l'échelle décadaire" (somme des crues individuelles dont la durée se mesure en heures) moins sensible à la trop grande variabilité des phénomènes mesurés au niveau de la crue.

La précision et la robustesse de ces modèles ont été testées par calages et validations croisées ; cette précision s'est avérée variable et fonction de la typologie du bassin basée sur des critères pédo-morphologiques. Les résultats de régression ont ensuite permis de reconstituer le comportement hydrologique "forestier" des bassins expérimentaux pendant toute la durée des essais à partir des données des bassins qui ont continué d'être observés dans les conditions originelles de forêt primaire.

L'utilisation de ces modèles a conduit aux conclusions suivantes.

En "Epoque Rouge" les volumes écoulés ont été supérieurs à ceux qui auraient pu être observés en forêt naturelle, dans des proportions allant de 60% à 130% selon le bassin versant considéré.

En "Epoque Jaune" les écoulements ont diminué tout en restant supérieurs à ceux du milieu naturel pendant les premières années. Un retour à des valeurs comparables à celles du milieu naturel a pu être noté sur des bassins présentant au moins quatre années d'observation dans cette période.

Le travail dont seulement quelques traits viennent d'être soulignés est avant tout hydrologique mais ses conclusions reposent sur les résultats de l'analyse statistique. Il utilise la régression linéaire et différentes méthodes

de validation comme un instrument de mesure plus ou moins précis des effets d'une modification du milieu sur des variables représentatives des aspects du comportement hydrologique potentiellement affectés par cette modification.

La prévision des crues (Cappelaere *et al.* 1996)

Le bassin amazonien s'étend sur plus de 6 millions de km² et apporte 209000 m³/s en moyenne à l'océan Atlantique.

Les apports combinés des tributaires méridionaux et septentrionaux, aux régimes différents, associés à l'effet régulateur des zones d'inondation, génèrent à l'aval de Manaus, un hydrogramme de l'Amazone à crue unique et étalée d'avril à juillet. Les populations et leurs activités, aussi bien rurales qu'urbaines ou semi-urbaines, subissent régulièrement des effets très importants et prolongés, dont la gravité varie en fonction de la sévérité de la crue : celle-ci détermine en effet les surfaces inondées, les hauteurs de submersion, ainsi que les durées au dessus d'une hauteur donnée. A Manaus même, ce sont 25000 personnes qui sont régulièrement touchées pendant des périodes de 3 à 5 mois. Selon l'importance de la crue des risques dramatiques sont encourus par la population. La prévision des crues a donc été jugée indispensable par les autorités.

Les objectifs de prévision qui ont été définis portent sur des délais de 10, 15, 20, 30, 40 et 60 jours avec une bonne précision sur les cotes prédites, de quelques dizaines de centimètres seulement pour un délai de 60 jours. Pour élaborer le modèle de prévision, les hauteurs journalières de 39 sites limnimétriques répartis sur les cours de l'Amazone et de ses formateurs étaient disponibles.

Dans le cadre d'une étude de prévision des crues, le modèle le plus satisfaisant pour l'hydrologue est le modèle de propagation des débits. Mais le manque de fiabilité des courbes hauteurs-débits n'a pas permis de recourir à une représentation de nature physique du processus de propagation, et c'est donc vers une modélisation de type statistique que s'est orienté ce travail à partir de quinze années d'observations concomitantes aux stations de mesure disponibles.

Des équations de prévision de 10 à 60 jours ont été établies par une méthode de régression ascendante pas à pas qui a permis, pour chaque délai, de sélectionner, parmi un grand nombre de variables potentiellement explicatives de l'évolution à venir de la crue à Manaus, le sous-ensemble à la fois aussi réduit et aussi significatif que possible au regard de cette

prévision. Une variable désigne l'association d'une station et d'un décalage temporel qui représente l'antériorité du jour d'observation de la donnée par rapport au jour d'émission de la prévision.

Parmi les quinze crues observées de 1978 à 1992, dix (soit 630 observations journalières au moins) ont été choisies pour constituer l'échantillon de calage, et cinq (340 observations journalières au moins) pour l'échantillon de validation. Ce choix a été effectué de telle sorte que ces deux échantillons soient aussi représentatifs que possible de l'ensemble des crues observées.

La précision des prévisions, si elle décroît bien sûr avec le délai reste tout à fait acceptable jusqu'à 60 jours. L'écart quadratique moyen est inférieur à 10 cm pour un délai de 10 jours, et est de l'ordre d'une quarantaine de centimètres pour le délai de 60 jours.

A côté de ce modèle de prévision portant sur le déroulement des phases de montée vers le maximum et d'amorce de la décrue, une prévision du seul débit maximum de la crue a pu être réalisée au cours de la montée à partir de la simple connaissance de la hauteur observée à Manaus et de la date à laquelle cette observation est effectuée. Etablir un modèle à partir des seules données à Manaus présente le grand avantage de permettre l'exploitation de la très longue série observée à cette station, à savoir les 90 années de 1903 à 1992. Deux modèles linéaires distincts ont été élaborés, de la forme :

$$H_{max} = f[H(j), j]$$

où H représente la hauteur observée à Manaus, et j le jour julien de cette observation, pris dans la période 15 mars-30 mai.

Le premier modèle est constitué par une fonction f purement linéaire des deux variables j et $H(j)$:

$$H_{max} = aH(j) + bj + c$$

appliqué à l'ensemble de la période de définition de j , c'est-à-dire du 15 mars au 31 mai. Le deuxième modèle fait intervenir j non pas comme variable directe du modèle, mais comme variable de contrôle des valeurs des paramètres de la fonction f , celle-ci devenant donc une fonction linéaire de $H(j)$, à paramètres variables a_j et c_j :

$$H_{max} = a_j H(j) + c_j$$

Le premier modèle est dit modèle global tandis que le deuxième est appelé "polylinéaire" puisque constitué d'une juxtaposition de petits sous-modèles linéaires partiels, définis pour une date j donnée de l'année.

Les deux modèles sont établis par régression linéaire, le premier directement sur l'ensemble des données correspondant à l'intervalle de variation de j , de la mi-mars à la fin mai, le second séparément pour chaque valeur de j , donc pour chaque jour de la période considérée.

Un troisième modèle, noté N , véritablement non-linéaire a été réalisé. La fonction f est constituée dans ce cas par un réseau de neurones de type perceptron multi-couche. On écrit alors :

$$H_{max} = N[H(j), j]$$

Les résultats obtenus par cette approche neuronale sont de meilleure qualité que ceux issus des modèles linéaires classiques.

Ces modèles ont été intégrés dans un logiciel opérationnel exploité par les prévisionnistes afin d'effectuer les prévisions en temps réel. Il aurait été préférable bien évidemment de procéder à une modélisation déterministe du phénomène de propagation des crues, car la limite majeure de la modélisation statistique est de soumettre la validité du modèle aux seuls événements dont les caractéristiques sont celles des observations constituant l'échantillon de calage. Toutefois la fraction du bassin Amazonien (72%) qui contribue au niveau de la crue à Manaus, produit toujours un limnigramme fortement lissé par rapport à ceux de chacun des formateurs, et comme le soulignent Molinier *et al.* (1996), une des caractéristiques principales des régimes de l'Amazone et de ses affluents est leur remarquable stabilité interannuelle. Ces conditions offrent à la démarche adoptée une certaine garantie pour une utilisation en prévision.

Les modèles de nature statistique qui ont été établis permettent de réaliser une prévision satisfaisante des niveaux à Manaus, jusqu'à un délai de deux mois. La prévision du maximum de la crue peut également être réalisée avec une précision tout à fait acceptable en utilisant les seules observations à Manaus.

La variabilité spatiale et temporelle des précipitations (Lebel et Le Barbé, 1997, Amani et Lebel, 1997)

Un autre domaine de la statistique mis au service de l'hydrologie va maintenant être évoqué, celui de la géostatistique.

Tous les modèles hydrologiques nécessitent une bonne connaissance de leur variable d'entrée, la pluie. Or l'estimation de la lame d'eau précipitée sur une surface est un problème difficile. La pluie n'est mesurée que ponctuellement et les réseaux de mesure ne permettent pas toujours d'estimer correctement la lame d'eau spatiale.

L'étude de la distribution spatiale et temporelle des précipitations au Sahel a été abordée par l'expérience EPSAT-Niger (*Estimation des précipitations par satellite au Niger*), puis poursuivie dans le cadre du projet HAPEX-Sahel (*Hydrological-Atmospheric Pilot Experiment-Sahel*). Le programme HAPEX-Sahel avait pour finalité une meilleure compréhension des interactions entre le climat et les conditions de surface à travers un suivi expérimental intense de toutes les composantes régissant ces interactions. HAPEX-Sahel a montré en particulier la nécessité d'obtenir des estimations de pluie les plus précises possibles dans cette région, et notamment sur les divers sites spécialement instrumentés pour l'étude. A cet effet des algorithmes statistiques permettant d'estimer la lame d'eau sur une zone, et de fournir un écart-type d'estimation ont été utilisés. Différentes échelles de temps ont été considérées : quelques minutes, l'événement, la décade, le mois et la saison des pluies.

Les techniques de géostatistique (krigeage) ont constitué l'outil d'analyse, particulièrement adapté ici à la représentation conceptuelle du mouvement des systèmes précipitants convectifs. Elles ont permis de quantifier l'organisation spatiale des champs de pluie pour chaque échelle de temps au moyen d'une fonction de corrélation spatiale. Des analyses approfondies ont ensuite conduit à établir des liaisons théoriques entre ces différents variogrammes.

Le fait de disposer d'une modélisation de la distribution des pluies adaptée à la fois à l'échelle de l'événement et à l'échelle de plusieurs événements est aujourd'hui un apport considérable pour comprendre comment l'humidité du sol et la végétation répondent à la distribution de ces pluies sur des échelles de temps allant de l'événement jusqu'à la saison des pluies toute entière. Cette approche géostatistique a permis aussi au niveau de l'événement et du cumul de plusieurs événements de quantifier les incertitudes d'estimation de la pluie en fonction de la hauteur d'eau précipitée et de la surface concernée, et d'en déduire des critères concernant la structure du réseau de mesures au sol. Il est à noter que ces incertitudes calculées de manière théorique, ont été validées expérimentalement (Lebel et Amani, à paraître).

Ces études ne font qu'accréditer les propos de Obled (1986) concernant la pluie en tant qu'entrée privilégiée du système hydrologique : "Ce n'est pas par des considérations hydro- et thermodynamiques que nous appréhenderons sa distribution sur la superficie d'un bassin, pas plus que par des expériences en laboratoire. Et les données recueillies au cours d'un épisode seront le seul point de départ objectif, même s'il apporte avec lui quelques sources d'erreur, pour appréhender l'entrée du système". Obled

préconisait alors l'utilisation de la géostatistique en ces termes, "Quant aux hydrologues, faute de reconnaître aisément dans la Géostatistique une continuité avec le savoir déjà acquis et les approches statistiques classiques, ils reculent devant un investissement intellectuel qui leur semble excessif, et se privent d'une approche qui, sans être révolutionnaire, est à coup sûr enrichissante". Les travaux récents qui viennent d'être menés au cours des expériences EPSAT- Niger et HAPEX-Sahel ont relevé ce défi.

La caractérisation des variations des régimes de pluie (Khodja et al. 1998)

Une place particulière est accordée à ce dernier exemple dont une présentation complète fait l'objet d'un article accepté pour publication dans la Revue de Statistique Appliquée.

L'Afrique de l'ouest non sahélienne a connu autour des années 1970 une baisse des précipitations qui coïncide avec une sécheresse marquée dans les régions sahéliennes. C'est ce constat qui est à l'origine du programme ICCARE (*Identification et Conséquences d'une variabilité du Climat en Afrique de l'ouest et centrale non sahélienne*) qui s'est fixé pour objectif d'identifier "la variabilité du climat" en Afrique de l'ouest et centrale (cf. figure 1) non sahélienne en vue d'analyser ses conséquences sur la gestion des ressources en eau.

L'étude devait être conduite avec pour unique source d'information les données hydro-pluviométriques des réseaux de mesure des pays concernés. Seule une analyse de ces données devait donc permettre de caractériser le phénomène ressenti autour des années 1970 et tout d'abord au niveau de la pluie. La diminution des précipitations témoigne d'une variabilité qui peut concerner d'autres aspects du régime pluviométrique que les seuls totaux précipités : la fréquence des pluies, la sévérité des saisons sèches, et la durée des saisons des pluies. Le recours aux techniques statistiques pour détecter une "variabilité climatique" à partir de données hydrologiques a été préconisé il y a déjà plus de trente ans par l'Organisation Mondiale de la Météorologie (WMO, 1966). Les techniques en question sont des tests qui s'appliquent aux séries chronologiques des variables choisies pour caractériser la variabilité recherchée.

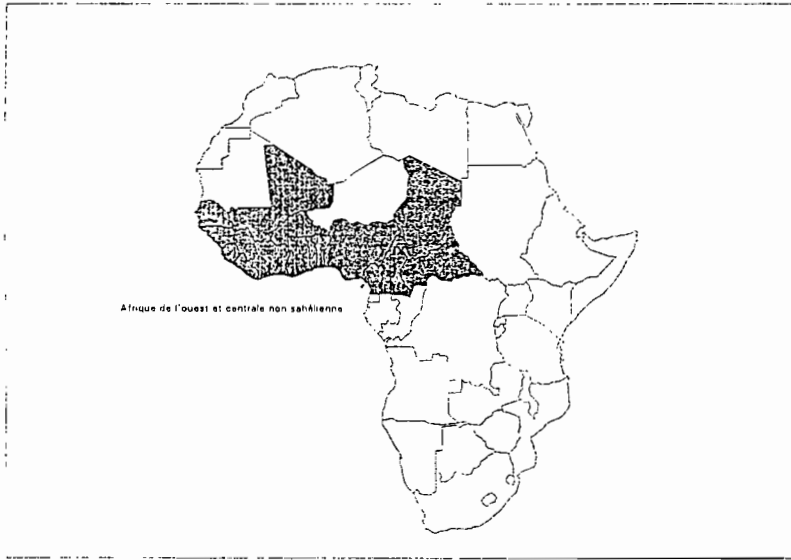


figure 1

Présentation de la zone d'étude

Dans le cadre d'ICCARE l'hypothèse émise sur la nature de la variabilité climatique pressentie sur la zone d'étude est celle d'une rupture qui se traduit par un saut brutal de la moyenne des séries. Des tests ou des procédures statistiques dits "de rupture" ont donc été mis en oeuvre. Il s'agit du test de Pettitt (Pettitt, 1979), du test de Buishand (Buishand, 1982, 1984), de la procédure bayésienne de Lee et Heghinian (Lee et Heghinian, 1977), et de la procédure de segmentation des séries hydrométéorologiques

de Hubert et Carbonnel (Hubert et al., 1989). A l'exception du test de Buishand qui ne conclut que sur l'existence ou non d'une rupture, les autres procédures proposent une estimation de la date de rupture.

Ces tests ont donc été appliqués sur chaque variable définie (pluie annuelle précipitée, nombre annuel de jours de pluie, hauteur des pluies en saison sèche, durée de la saison des pluies) en chaque point de mesure de la zone d'étude qui couvre seize pays, du Sénégal à l'ouest à la République de Centrafrique à l'est. Ils donnent des résultats souvent concordants quant à l'existence de ruptures en moyennes dans les années 1960 et 1970, mais ils restent cependant difficiles à interpréter et à synthétiser dans leur ensemble.

Pour faciliter cette lecture globale, des analyses des correspondances multiples (ACM) ont été réalisées à partir d'une réponse du type "rupture" ou "non-rupture", ou encore "série homogène" ou "année de rupture 19..", donnée par une synthèse des résultats des différents tests appliqués site par site et variable par variable. L'ensemble de ces réponses traité par des ACM a permis d'obtenir un résumé "objectif" d'information dont l'interprétation a débouché sur une organisation spatiale des réponses issues du traitement statistique des séries chronologiques (Gautier *et al*, 1998).

Cette répartition a été confirmée par une analyse spatio-temporelle généralisée au cas multivarié. Cette approche tridimensionnelle a permis d'une part de détecter des ruptures climatiques caractérisées par plusieurs variables pluviométriques, et d'autre part de répartir spatialement les stations ayant subi des changements similaires sur une période commune. Ce travail confié à des statisticiens a bénéficié d'un formalisme rigoureux qui est ici rapidement décrit.

Cette approche s'est déroulée en deux temps.

1 Analyse multidimensionnelle des données

Une analyse exploratoire multidimensionnelle a tout d'abord conduit à constituer, sur la base de l'ensemble des données disponibles, des groupes de stations caractérisés par des comportements pluviométriques similaires vis à vis de l'ensemble des variables.

L'information brute a ainsi été résumée en utilisant la méthode STATIS (Structuration des Tableaux A Trois Indices de la Statistique) (Lavit, 1988). Il s'agit d'une analyse conjointe de K tableaux de données quantitatives (un tableau par station), relevées en différentes occasions, où p mêmes variables (variables pluviométriques) ont été mesurées sur les mêmes N individus (années), et pour lesquels est recherchée entre autre, une structure

commune appelée interstructure.

Trois groupes géographiques à l'intérieur desquels les stations ont des comportements similaires ont ainsi été identifiés par STATIS. Globalement, et pour simplifier, ils s'organisent autour des noyaux suivants :

groupe I: Togo, Bénin et sud de la Côte d'Ivoire,

groupe II: Cameroun, Centrafrique et Burkina Faso, et

groupe III: Guinée Conakry, Guinée Bissau, Mali et Sénégal.

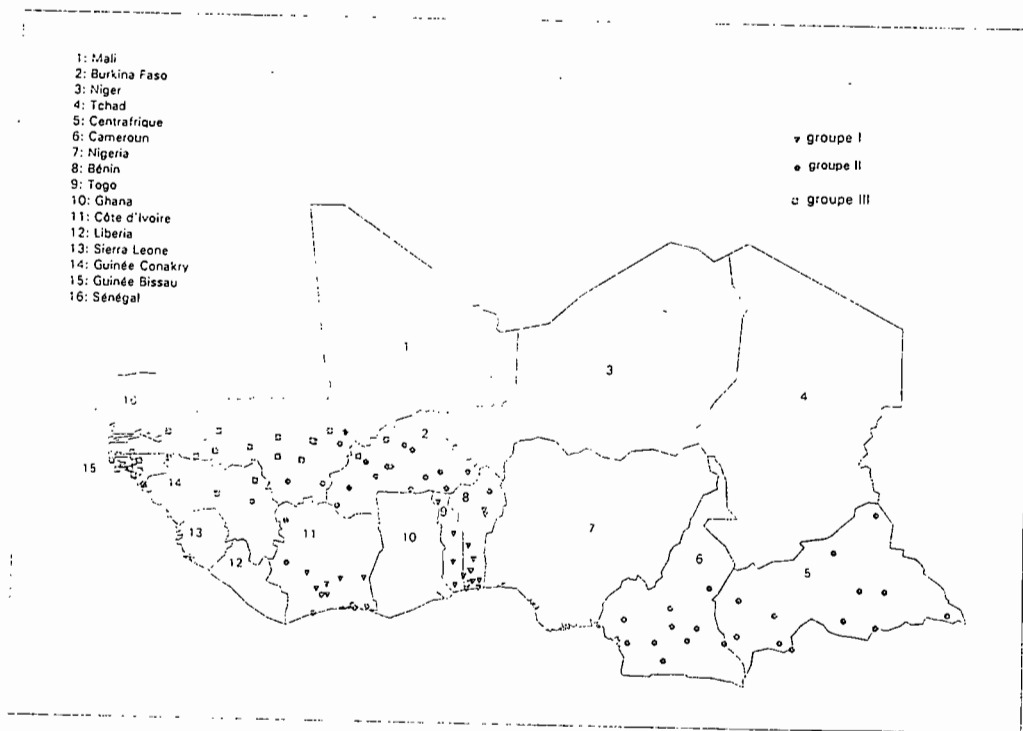


figure 2

Localisation des groupes homogènes

La distribution spatiale des stations fait apparaître assez nettement une répartition nord-sud d'une part et est-ouest d'autre part. Cependant,

le Ghana et le Nigeria dont les données n'ont pas été disponibles font véritablement défaut à l'étude.

La carte (figure 2) obtenue résume donc les comportements pluviométriques de l'ensemble des stations de la zone étudiée entre 1950 et 1980, tels qu'ils ont été décrits par les quatre variables hydrologiques retenues.

Mais cette organisation repose sur un faible pourcentage d'inertie expliquée par les deux premiers axes de l'ACP mise en œuvre dans STATIS. Aussi avant de poursuivre l'analyse, un test de permutation s'appuyant sur l'analyse des corrélations canoniques a permis de vérifier que les groupes obtenus par STATIS étaient significativement différents vis à vis des quatre variables étudiées. Le principe du test et son application aux données sont présentés en détail par Khodja *et al* (1998), nous en retraçons ici les lignes essentielles.

Le test de permutation est ici utilisé pour étudier l'indépendance de deux tableaux de données :

- Y est le tableau croisant en lignes les stations classées par groupes et en colonnes les variables pluviométriques et dans lequel Y_{ij} est la moyenne, calculée sur l'ensemble des années, de la variable j pour la station i .

- X est le tableau croisant en lignes les stations classées par groupes et en colonnes les indicatrices des groupes, ($X_{ij} = 1$ si la station i appartient au groupe j , $X_{ij} = 0$ sinon)

Il s'agit donc de tester l'hypothèse nulle H_0 : Les tableaux sont indépendants contre l'hypothèse alternative H_{-1} : Les tableaux sont liés.

L'hypothèse H_0 se traduit encore en termes de permutabilité des lignes de l'un des deux tableaux.

En référence au principe de Mardia (1971), le test de permutation est mis en œuvre en ne réalisant explicitement aucune permutation, mais en ne faisant intervenir que le calcul des premiers moments de la statistique choisie (Mardia, 1971, Kazi-Aoual *et al.*, 1994), en l'occurrence celle de l'Analyse des Corrélations Canoniques adaptée à l'étude de l'indépendance de deux tableaux de données. Cette statistique semble en effet être la plus appropriée aux données car son approche symétrique permet l'étude des relations linéaires entre les groupes de variables observés sur les mêmes individus en accordant aux deux tableaux des considérations égales.

L'hypothèse H_0 a été rejetée confirmant l'existence de différences significatives entre les trois groupes.

Il faut souligner que les groupes identifiés par l'analyse statistique sont

tout à fait cohérents avec la répartition sur cette région des régimes pluviométriques dont les caractéristiques premières sont le nombre de saisons des pluies (1 ou 2) et la quantité annuelle précipitée.

2 Test de rupture en moyenne et estimation du point de rupture

Le test de changement en moyenne et l'estimation du point de rupture s'effectuent sur les séries des moyennes intergroupes pour chaque variable.

Test de rupture en moyenne

Dans le cadre du modèle multilinéaire gaussien, on pose $x_i = \mu_i + \varepsilon_i$, $1 \leq i \leq N$ avec

$$x_i = \begin{pmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{pmatrix}, \quad \mu_i = \begin{pmatrix} \mu_i^1 \\ \mu_i^2 \\ \vdots \\ \mu_i^p \end{pmatrix}, \quad \varepsilon_i = \begin{pmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \\ \vdots \\ \varepsilon_i^p \end{pmatrix}$$

avec $\varepsilon_i \sim \mathcal{N}_p(0, \Sigma)$ où N est le nombre d'observations, et p le nombre de variables.

On désire tester $H_0 : \mu_1 = \mu_2 = \dots = \mu_N$ contre $H_1 : \mu_1 = \mu_2 = \dots = \mu_r \neq \mu_{r+1} = \mu_{r+2} = \dots = \mu_N$ dans le cas où r et Σ sont inconnus.

La statistique bayésienne suivante:

$$S = u'V^{-1}u \text{ où } u' = \frac{1}{(a'a)^{1/2}} \sum_{i=1}^{N-1} i(x_{i+1} - \bar{x}),$$

avec $a' = 1/2(-(N-1), -(N-3), -(N-5), \dots, (N-5), (N-3), (N-1))$, $V = \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})'$ et \bar{x} étant la moyenne empirique de l'échantillon total (x_1, x_2, \dots, x_N) , répond au problème posé (Srivastava, Carter, 1983).

L'hypothèse H_0 d'égalité des moyennes est ainsi rejetée au niveau α si :

$$S \geq c_\alpha \text{ où } c_\alpha = 1 - \left(1 + p(N - p - 1)F_{p, N-p-1}^\alpha\right)$$

avec $\alpha = \Pr(F_{p, N-p-1} \geq F_{p, N-p-1}^\alpha)$, $F_{p, N-p-1}$ étant la distribution d'une loi de Fisher à p et $N - p - 1$ degrés de liberté.

Estimation du point de rupture

Si l'hypothèse H_0 est rejetée, la position de la rupture est recherchée en se plaçant toujours dans le cadre du modèle multilinéaire gaussien. On considère alors les échantillons (x_1, x_2, \dots, x_r) et $(x_{r+1}, x_{r+2}, \dots, x_N)$. Le point de rupture est celui qui maximise les distances de Mahalanobis es-

timées entre ces échantillons et données par :

$$D_r^2 = (\bar{x}_r - \bar{x}_{N-r})' \hat{\Sigma}^{-1} (\bar{x}_r - \bar{x}_{N-r})$$

où \bar{x}_r , \bar{x}_{N-r} et $\hat{\Sigma}$ sont les estimateurs du maximum de vraisemblance, sous l'hypothèse H_1 , de μ_1 , μ_N et Σ respectivement.

En notant: $\hat{\Sigma} = \sum_{i=1}^r (x_i - \bar{x}_r)(x_i - \bar{x}_r)' + \sum_{i=r+1}^N (x_i - \bar{x}_{N-r})(x_i - \bar{x}_{N-r})'$,

l'estimateur du point de rupture est alors celui qui maximise la statistique $T_r^2 = N^{-1}r(N-r)(\bar{x}_r - \bar{x}_{N-r})' \hat{\Sigma}^{-1} (\bar{x}_r - \bar{x}_{N-r})$.

Les résultats obtenus

Les techniques mises en œuvre ont permis de conclure à l'existence d'une rupture en moyenne sur les groupes I, II et III respectivement en 1962, 1969 et 1967 et c'est le groupe III qui présente la rupture la plus significative.

Les résultats de cette approche viennent ainsi confirmer ceux des études univariées, à savoir l'existence de ruptures en moyennes dans les années 1960 et 1970 en Afrique de l'ouest. Les ruptures détectées en 1967 et 1969 sur les groupes II et III confirmeraient le début d'une période déficitaire, tandis que celle du groupe I en 1962 (Togo, Bénin et quelques stations du sud de la Côte d'Ivoire) traduirait un déficit peu significatif.

Discussion

Les résultats des tests univariés ont été difficiles à interpréter de par la multitude et la variété des réponses obtenues par variable et/ou par site. Deux facteurs au moins viennent compliquer leur analyse, l'un relatif aux performances même de ces tests, l'autre lié à la nature des conditions physiques existant au niveau de chaque station. Une étude par simulation de la puissance et de la robustesse des tests de rupture (Lubès-Niel *et al*, 1998) révèle en effet que ce n'est qu'à partir d'un taux de rupture en moyenne supérieur à 15%, et dans les conditions d'application requises par les tests (qui ne sont pas toujours respectées) que la puissance estimée dépasse 40%. Par ailleurs sans mettre en cause les tests eux-mêmes les résultats auxquels ils aboutissent peuvent être induits par des particularismes naturels locaux prévalant sur les effets à caractère régional. De fait une réalité de variation "climatique" peut être masquée par ces comportements individuels prédominants.

Face à ces réserves, l'approche multivariée directe qui a bien sûr elle-même ses propres limites – le test de détection de rupture en moyenne n'échappe pas aux risques de première et de deuxième espèce comme tout

autre – présente tout de même l'avantage de faire ressortir des traits dominants, et de résumer toute l'information disponible sous la forme d'une synthèse cartographique simple des variations pluviométriques à l'échelle régionale.

Conclusion

Le rôle des statistiques en hydrologie vient d'être présenté à travers quelques applications réalisées au sein de l'ORSTOM.

Les outils de la statistique qui ont été évoqués, régression linéaire simple ou multiple, tests d'hypothèse, analyse multidimensionnelle, géostatistique, ne représentent pas un inventaire exhaustif des techniques de cette discipline qui sont utilisées en hydrologie. Inversement la prévision des crues, la caractérisation des variabilités spatiales et/ou temporelles ne sont pas les seuls thèmes de l'hydrologie qui sollicitent la statistique. A ce propos il serait impardonnable d'oublier de citer les risques hydrologiques dont l'étude est spécialement de nature statistique : la démarche statistique "impliquée" passe par l'analyse d'observations passées pour inférer une loi de probabilité théorique à partir de laquelle les probabilités d'un événement à risque (crue, sécheresse...) sont estimées (Lubès *et al.*, 1994).

Les exemples qui ont été présentés montrent que pour l'hydrologie, le recours à la statistique est le plus souvent incontournable. Dans le cas de la prévision des crues de l'Amazonie, le phénomène de propagation est régi par des lois physiques et donc sa représentation par des équations de régression est due à l'absence des données nécessaires à une modélisation déterministe. Dans le cas de l'étude des pluies sahéliennes, les données dont dispose l'hydrologue ne permettent pas une représentation physique de la distribution spatiale des précipitations qui du reste ne peut être abordée que par des physiciens de l'atmosphère. Il en est de même pour l'étude des variations "climatiques" quand seules des données hydro-pluviométriques sont disponibles. L'utilisation de la régression linéaire s'est avérée bien adaptée, enfin, pour mesurer les effets du défrichement sur le comportement hydrologique des bassins versants.

Toutefois la statistique est affaire de statisticiens et l'hydrologie est affaire d'hydrologues. Aussi la culture statistique acquise par les hydrologues devrait être utilisée plus efficacement à travers un dialogue avec des statisticiens de façon à déboucher sur une vraie recherche en matière de statistique appliquée à l'hydrologie (Masson, 1992).

Références bibliographiques

- Amani A., Lebel T., 1997. Lagrangian kriging for the estimation of Sahelian rainfall at small time steps. *J. Hydrol.* 192 : 125-157.
- Buishand T. A., 1982. Some methods for testing the homogeneity of rainfall records. *J. Hydrol.* 58 : 11-27.
- Buishand T. A., 1984. Tests for detecting a shift in the mean of hydrological time series. *J. Hydrol.* 73 : 51-69.
- Cappelaere B., Lubès-Niel H., Berkhoff C., Thépaut H., Guyot J. L., Molinier M., De Oliveira E., Rodrigues M., 1996. Préviation des crues de l'Amazonie. IAHS Publ. no. 238. 355-366.
- Escoufier Y., 1992. Quels statisticiens pour l'ORSTOM ? Séminfor 5, Statistique Impliquée, Cinquième Séminaire Informatique de l'ORSTOM, Montpellier 2, 3 et 4 septembre 1991. 433-444.
- Fritsch J. M., 1990. Les effets du défrichement de la forêt amazonienne et de la mise en culture sur l'hydrologie de petits bassins versants. Opération ECEREX en Guyane Française. Thèse Université Montpellier II. 392 pages.
- Gautier F., Lubès-Niel H., Sabatier R., Masson J. M., Paturel J. E., Servat E., 1998. Etude de la variabilité du régime pluviométrique de l'Afrique de l'Ouest non sahélienne entre 1950 et 1989 par une analyse multivariée des résultats de tests statistiques d'homogénéité de séries chronologiques. A paraître, *Hydrological Sciences Journal*.
- Hubert P., Carbonnel J. P., Chaouche A., 1989. Segmentation des séries hydrométéorologiques - Application à des séries de précipitations et de débits de l'Afrique de l'Ouest. *J. Hydrol.* 110 : 349-367.
- Kazi-Aoual F., Hitier S., Sabatier R. et Lebreton J.D., 1994. Refined approximations to permutation tests for multivariate inference. *Computational statistics and Data Analysis* 20, 643-656.
- Khodja H., Lubès-Niel H., Sabatier R., Masson J. M., Servat E., Paturel J.E., 1998. Analyse spatio-temporelle de données pluviométriques en Afrique de l'Ouest. Recherche d'une rupture en moyenne. Une alternative intéressante : Les tests de permutations. A paraître, *Revue de Statistique Appliquée*.
- Laloë F; 1992. Séminfor 5, Statistique Impliquée, Cinquième Séminaire Informatique de l'ORSTOM, Montpellier 2, 3 et 4 septembre 1991. 464 pages.
- Lavit C., 1988. Analyse conjointe de tableaux quantitatifs. Editions Masson.
- Lebel T., Le Barbé L., 1997. Rainfall monitoring during HAPEX-Sahel. 2. Point and areal estimation at the event and seasonal scales. *J. Hydrol.* 188-189 : 97-122.

- Lebel T., Amani A. Rainfall estimation in the Sahel. What is the ground truth ?
Submitted to Journal of Applied Meteorology.
- 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* 19(4) : 503-506.
- Lubès H, Masson J.M., Raous P., Tapiou M., 1994. SAFARHY, Logiciel de calculs statistiques et d'analyse fréquentielle adapté à l'évaluation du risque en hydrologie. ORSTOM, Université Montpellier II.
- Lubès-Niel H., Masson J. M., Paturel J. E., Servat E., 1998. Variabilité climatique et statistiques. Etude par simulation de la puissance et de la robustesse de quelques tests utilisés pour vérifier l'homogénéité de chroniques. A paraître, *Revue des Sciences de l'Eau*.
- Mardia K.V., 1971. The effect of non normality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 58, n 1, 105-122.
- Masson J. M., 1992. Hydrologie et Statistiques. Séminfor 5, Statistique Impliquée, Cinquième Séminaire Informatique de l'ORSTOM, Montpellier 2, 3 et 4 septembre 1991. 461-464.
- Molinier M., Guyot J. L., De Oliveira E., Guimaraes V., 1996. Les régimes hydrologiques de l'Amazone et de ses affluents. IAHS Publ. no. 238. 209-222.
- Obled C., 1986. Introduction au krigeage à l'usage des hydrologues. Deuxièmes Journées Hydrologiques de l'ORSTOM à Montpellier, 16-17 septembre 1986. 174-220.
- Pettitt A. N., 1979. A non-parametric approach to the change-point problem. *Appl. Statist.* 28(2) : 126-135.
- Srivastava M.S., Carter M., 1983. An Introduction to Applied Multivariate Statistics. Elsevier Science Publishing.
- W. M. O. (World Meteorological Organization), 1966. Climatic change by a working group of the commission for climatology. W. M. O. 195, TP 100, Tech. note no. 79. WMO Geneva, Switzerland.

Société Française de Biométrie

De l'observation à l'analyse,
implication de la biométrie
dans les pays
de développement

1998

Textes présentés lors des journées réalisées par
la Société Française de Biométrie
à Carcassonne, les 29 et 30 mai 1997

N° 15