N O T I O N D'A G R O S Y S T E M E

PREMIERE PARTIE

NOTION DE BASE ET NOTIONS DERIVEES

METHODE D'ETUDE

Chapitre 2

CARACTERISATION D'UN AGROSYSTEME

H= 54970

N AG 1983/1/1

n'existe qu'a
Nouméa

par

B. BONZON et J. DEJARDIN

Fonds Documentaire IRD
010025730

Fonds Documentaire IRD
Cote : B*25730 Ex : unique

Copyright ORSTOM

Septembre 83

NOTE SUR UN PROJET D'OUVRAGE COLLECTIF CONCERNANT LA NOTION D'AGROSYSTEME.

La notion d'agrosystème a été élaborée en 1972 à Adiopodoumé par J. DEJARDIN, C. FILLONNEAU et B. BONZON en réponse aux difficultés rencontrées à l'époque par les agronomes de Côte d'Ivoire dans l'organisation de l'interprétation de leurs données sur les "Interaction sol-plantes fourragères en milieu tropical humide".

Le concept et les notions qu'ils en ont dérivées ont été utilisés en premier par J.C. TALINEAU pour différencier l'évolution de la matière organique du sol sous les graminées et les légumineuses étudiées dans le cadre de ce programme.

Ils ont été mis en oeuvre ensuite par P. de BOISSEZON et B. BONZON pour étudier de façon systématique un certain nombre des relations sol-techniques culturales-manioc dans les conditions d'une étude expérimentale conduite elle aussi à Adiopodoumé (travail qui doit être présenté prochainement).

Ils sont utilisés actuellement de façon très systématique à Nouméa par B. DENIS et B. BONZON dans le cadre du programme multidisciplinaire sur la fertilité et l'évolution sous culture des sols de Nouvelle-Calédonie.

L'expérience acquise par les uns et les autres montrent finalement que ces concepts permettent une plus grande finesse au niveau de l'interprétation scientifique après en avoir facilité l'organisation préalable.

A la suite d'un certain nombre de discussions entre B. DENIS, P. de BOISSEZON et B. BONZON à Nouméa, il est apparu à ces chercheurs qu'il serait certainement intéressant, tant au plan conceptuel qu'aux plans méthodologique et scientifique, de présenter enfin officiellement la notion d'agrosystème dans un ouvrage collectif.

Contactés fin Juin 1983 par B. BONZON pour participer à ce travail, J. DEJARDIN et J.C. TALINEAU ont déjà donné leur accord de principe. C. FILLONNEAU contacté lui aussi à la même époque en Côte d'Ivoire n'a peut être pas encore reçu cette correspondance.

Plus précisément, le projet pourrait comporter deux parties. La première, théorique, exposerait la notion d'agrosystème, les notions que l'on peut y rattacher et les principales méthodes et techniques d'analyses statistiques nécessaires.

La seconde illustrerait la première en s'appuyant sur des résultats tirés des études conduites à Adiopodoumé et Nouméa.

Nombre des chapitres de l'ouvrage pourraient faire l'objet, par ailleurs, de publications préalables.

Le plan général de la première partie est le plus avancé. Elle pourrait comporter quatre chapitres et une annexe.

Le premier chapitre présenterait la notion d'agrosystème et les notions que l'on peut y rattacher.

Les suivants traiteraient successivement :

. le chapitre II , de la caractérisation d'un agrosystème ;

. le chapitre III , de la comparaison de deux ou plusieurs agrosystèmes ;

. le chapitre IV , de l'effet de facteurs de variations contrôlées sur un agrosystème.

L'annexe regrouperait les informations nécessaires à l'application des principaux tests statistiques utilisés pour caractériser un agrosystème, comparer deux ou plusieurs agrosystèmes, étudier l'influence de facteurs de variations sur un agrosystème, l'accent étant mis sur les possibilités d'exécutions automatiques de ces tests.

Rédigée à Nouméa le 06/09/83 et adressé pour critique et compléments à J. DEJARDIN, J.C. TALINEAU, C. FILLONNEAU.

Adressée en même temps pour info à MM. P. FRANQUIN, B. DABIN, J. FORESTIER.

Documents de référence actuels.

1. BONZON B., DEJARDIN J., FILLONNEAU C., GERI M., 1974.

Programme multilocal d'étude des interactions sol plantes fourragères en milieu tropical humide. Organisation générale de l'analyse statistique des résultats. ORSTOM - Adiopodoumé, multig., 14 p, 4 tableaux.

2. TALINEAU J.C., BONZON B., FILLONNEAU C., HAINNAUX G., 1980.

Contribution à l'étude d'un agrosystème praivial dans le milieu tropical humide de la Côte d'Ivoire. 2. Analyse des données relatives à l'état de la matière organique. Cah. ORSTOM, sér. Pédol., vol. XVIII, n° 1, 1980-81 : 29-47.

Pièces jointes : premières ébauches des chapitres 2 et 3 et de l'annexe de la première partie (par B. BONZON, J. DEJARDIN et C. FILLONNEAU si ce dernier est intéressé par le projet).

CARACTERISATION PARTIELLE APPROCHEE D'UN AGROSYSTEME

ASPECTS STATISTIQUES

-§-

CARACTERISATION PARTIELLE APPROCHEE D'UN AGROSYSTEME.

ASPECTS STATISTIQUES.

-§-

Soit donc un agrosystème A dont on désire caractériser la partie à p éléments observés indépendamment les uns des autres sur n sites répartis de façon aléatoire à la surface du champ de référence.

Chaque élément est défini par un "paramètre caractéristique", ou "élémentaire", ou "de base", ou encore plus simplement, par une "caractéristique", vocable plus général car pouvant inclure des aspects qualitatifs. Pour la suite, on ne considérera cependant, ici, que des "caractéristiques mesurables".

Il y a donc p caractéristiques

$$U, V, W, \ldots, X, \ldots, Z$$

et chaque caractéristique fait l'objet de n observations, par exemple pour X :

$$x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n$$

Lorsqu'une observation de X est faite sur chacun des n sites, une autre observation est faite, simultanément ou non, de chacune des (p-1) autres caractéristiques sur chacun des mêmes n sites.

L'ensemble de ces informations peut être rassemblé en une matrice de données $\hat{A}$ de dimensions(p.n) :

$$\hat{A} = \begin{pmatrix} u_1 & u_2 & \ldots & u_i & \ldots & u_n \\ v_1 & v_2 & \ldots & v_i & \ldots & v_n \\ \vdots & & & & & \\ x_1 & x_2 & \ldots & x_i & \ldots & x_n \\ \vdots & & & & & \\ z_1 & z_2 & \ldots & z_i & \ldots & z_n \end{pmatrix}$$

Chaque colonne de $\hat{A}$ caractérise un site, chaque ligne une caractéristique.

Si l'on connaissait la fonction de liaison de chacun des p.(p-1) /2 couples de paramètres (X,U), (X,V), ... etc, et si les conditions de mileu des sites d'observations possibles étaient parfaitement homogènes, le système constitué par la partie de A à p éléments serait défini à partir d'une seule série d'observations des caractéristiques U,V,... etc, c'est-à-dire à partir d'un seul site. Les n vecteurs-colonnes de $\hat{A}$ seraient, en conséquence, identiques.

Dans la réalité, l'intervention de facteurs de variations aléatoires sur l'aire des sites d'observations possibles fait osciller chaque caractéristique X autour de sa valeur moyenne réelle que l'on ne peut, évidemment, qu'estimer en répétant les observations un certain nombre de fois, n ici en l'occurence.

Dans ces conditions d'ailleurs, si l'on peut estimer μ_X, moyenne vraie de X, par

$$\overline{x} = (1/n).\sum_{i=1}^{n} x_i \qquad (1)$$

on sait qu'il faut aussi préciser, en même temps, l'étendue du champ de variation de $\overline{x}$ en estimant l'écart-type de sa distribution, $s_{\overline{x}}$, à partir de

$$s^2_{\overline{x}} = \{ 1 /n.(n-1) \}.\sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad (2)$$

Quant à la détermination des fonctions de liaison des p.(p-1) /2 couples de paramètres (X,U), (X,V),... etc, il s'agit-là du problème fondamental de la caractérisation d'un agrosystème.

Pour chaque couple de paramètres (X,U) deux situations peuvent se présenter :

- ou bien l'on connaît a priori la fonction théorique de la liaison qui les unit, et le problème est alors celui de l'ajustement de cette fonction aux n valeurs observées des couples (X,U). D'une façon générale, il s'agit d'un problème de modélisation ;

- ou bien l'on ne possède aucune information sur cette fonction et il s'agit dans ce cas, plus modestement, d'estimer l'intensité de la liaison puis de préciser une fonction ou un graphique d'estimation de X à partir de U, ou de l'inverse.

D'une façon générale les problèmes à résoudre relèvent alors de la corrélation.

Comme dans l'immense majorité des cas on se trouve dans la seconde situation on rappelera seulement ici l'essentiel de la théorie et surtout des techniques d'analyse de la corrélation nécessaires à la caractérisation approchée de la partie observée de A.

Remarque

Le fait pour chacune des p caractéristiques de pouvoir être liée simultanément aux (p-1) autres, ou du moins à un certain nombre d'entre elles, pose naturellement le problème des liaisons multiples et partielles susceptibles de les unir. D'un point de vue théorique, deux autres séries de problèmes devraient donc être examinées ensuite :

- celle des problèmes relatifs à la caractérisation des liens multiples et partiels entre X d'une part, U, V, W,... etc, d'autre part ;

- celle des problèmes concernant la caractérisation des liens existant entre les ensembles de deux groupes de caractéristiques : par exemple ceux existant entre X et U d'une part, V, W, ... etc, d'autre part (corrélations canoniques).

Comme les conditions d'application de l'analyse multidimentionnelle de p caractéristiques et la conduite des calculs peuvent reposer sur les conditions d'application et les résultats de l'analyse bi-dimentionnelle de ces p caractéristiques prises deux à deux, on se limitera ici, pour l'instant, aux problèmes de la corrélation simple.

S'agissant, enfin, de rappels très orientés, le lecteur aura intérêt à consulter les ouvrages de base de l'analyse statistique et de la statistique multidimentionnelle, notamment ceux de P. DAGNELIE (1969-70, 1975) et J. LEFEVRE (1976).

1 - NOTION DE CORRELATION ET PROBLEMES A RESOUDRE.

1.1. Notions de liaison fonctionnelle, de corrélation et d'indépendance.

Soient donc deux caractéristiques X et U prises parmi les p caractéristiques du système :

1°/ si X et U sont rigidement liées l'une à l'autre, c'est-à-dire si à toute valeur de U l'on sait aussitôt faire correspondre une valeur et une seule de X,

et réciproquement, on dira qu'elles sont liées de façon fonctionnelle : par exemple la surface S d'un cercle et son rayon R.

Dans la réalité et graphiquement, cela se traduira par un nuage de points très étroit à l'allure générale de courbe : dans l'exemple ci-dessus, une parabole que l'on retrouvera aux erreurs près commises sur S et R ;

2°/ Si X et U sont liées de façon moins rigide, en d'autres termes si l'on observe pour un niveau donné de U une certaine variation de X, et réciproquement, les deux caractéristiques seront dites en liaison stochastique ou en corrélation. On dira aussi que leurs répartitions sont liées.

Dans la réalité et graphiquement, cela se traduira par un nuage de points d'autant plus étiré que la liaison sera proche d'une liaison fonctionnelle.

Un diagramme des moyennes liées pourra être tracé à travers le nuage de points. La relation que l'on mettra ainsi en évidence en classant X sur U { X = f (U) } n'est cependant pas la même que celle que l'on obtiendra en classant U sur X { U = g (X) } .

Les deux courbes de régression ainsi dégagées seront d'autant plus éloignées l'une de l'autre et différentes que la liaison s'éloignera d'une liaison du type fonctionnelle et linéaire, un cas particulier - très important - étant celui où les lignes de régression sont des droites ;

3°/ si X et U ne sont pas liées du tout la répartition des valeurs de X sera la même quelle que soit la valeur de U, et réciproquement.

Graphiquement, cette situation devrait se traduire par des droites de régression X = f (U) et U = g (X) parallèles aux axes des U et des X.

Un point à noter au passage : aucune des deux variables n'a été priviligiée.

1.2. Problèmes à résoudre.

Devant caractériser la liaison (X,U) deux premières questions se posent immédiatement. Comment peut-on :

. caractériser l'intensité de cette liaison, son étroitesse ?
. estimer l'une des variables à partir de l'autre ?

Comme les réponses à ces questions dépendent de la normalité des distributions de X et de U d'une part, de la linéarité des fonctions de régression $X = f(U)$ et $U = g(X)$ d'autre part, deux nouveaux problèmes apparaissent. Peut-on et comment :

. vérifier la normalité des distributions des variables ?
. vérifier la linéarité de leur régressions ?

Si l'une des deux régressions n'est pas linéaire - et a fortiori si les deux ne le sont pas - une troisième série de questions se pose :

. est-il possible dans ces conditions de caractériser l'intensité de la liaison ?
. ou : peut-on transformer l'une ou les deux variables pour linéariser la ou les régressions ?

Enfin, si la distribution de X (ou celle de U) n'est pas normale, une septième question devra être examinée :

. est-il possible de transformer X (ou U) pour que sa distribution apparaisse normale ?

Ces questions sont, à l'évidence, très étroitement liées les unes aux autres.

2. CORRELATION DANS LE CAS OU X ET U SONT DISTRIBUEES NORMALEMENT ET LIEES LINEAIREMENT.

2.1. Loi de référence.

. Si l'on fait l'hypothèse que X et U sont distribuées normalement et liées linéairement, on se place dans une situation où l'on dispose d'une loi de référence classique, la loi normale à deux variables.

. Cette loi dépend de cinq paramètres :

- la moyenne μ_x et l'écart-type σ_x de X ;
- la moyenne μ_u et l'écart-type σ_u de U ;
- une constante ρ, appelée coefficient de corrélation de X et U, qui est un nombre compris entre -1 et +1.

. Elle jouit des propriétés suivantes :

- les distributions marginales de X et de U (c'est-à-dire celles de l'ensemble des observations de X d'une part, de U de l'autre) sont normales, de même que les distributions de X à U constant ou de U à X constant ;

- les courbes de régression sont des droites ;

- les écart -types dans les sous-populations liées sont réduits dans la proportion de 1 à $\sqrt{1-\rho^2}$ par rapport aux écart -types σ_x et σ_u.

. Le coefficient de corrélation ρ résume à lui seul le degré de liaison des deux variables :

- si $\rho = 0$, X et U sont indépendantes : les droites de régression sont des parallèles aux axes de coordonnées et la variabilité dans les tranches à U ou à X constant est la même qu'au niveau des distributions marginales ;

- si $\rho = \pm 1$, X et U sont en liaison fonctionnelle linéaire ; les deux droites de régression dont confondues et la variabilité dans les tranches est nulle ;

- si $-1 < \rho < +1$, X et U sont en dépendance stochastique, ou en corrélation, plus ou moins étroite.

. Les relations suivantes existent entre ces cinq paramètres :

- l'équation de la droite de régression de X sur U (ou de X en U) qui permet d'estimer X à partir de U, X étant considérée comme variable dépendante de U, est

$$x(u) - \mu_x = \rho . \frac{\sigma_x}{\sigma_u} . (u - \mu_u) \qquad (3)$$

la quantité

$$\rho . \frac{\sigma_x}{\sigma_u} = \beta_{xu} \qquad (4)$$

désignée par le vocable de "coefficient de regression de X sur U", étant la pente de la droite en question ;

- l'équation de la droite de régression de U sur X (ou de U en X) qui permet d'estimer U à partir de X, U étant considérée comme variable dépendante de X est

$$u(x)-\mu_u = \rho.\frac{\sigma_u}{\sigma_x}.\ (x - \mu_x) \qquad (5)$$

la quantité

$$\rho.\frac{\sigma_u}{\sigma_x} = \beta_{ux} \qquad (6)$$

appelée de la même façon "coefficient de régression de U sur X" étant la pente de cette deuxième droite ;

- les distributions de X dans les sous-populations à U constant ont pour écart-type

$$\sigma_{x(u)} = \sqrt{1-\rho^2}\ .\sigma_x \qquad (7)$$

- les distributions de U dans les sous-populations à X constant ont de la même façon pour écart-type

$$\sigma_{u(x)} = \sqrt{1-\rho^2}.\ \sigma_u \qquad (8)$$

- le coefficient de corrélation ρ est égal à

$$\rho = \frac{\mathcal{M}_{xu}}{\sigma_x.\sigma_u} \qquad (9)$$

où $\mathcal{M}_{xu}$ représente le moment centré d'ordre 1 par rapport à X et U ;

- les coefficients de régression et de corrélation sont liés entre eux par :

$$\rho^2 = \beta_{xu}.\ \beta_{ux} \qquad (10)$$

. Les deux droites de régression passent par le point M $(\bar{x},\bar{u})$

2.2. Estimation des coefficients de corrélation et de régression.

Disposant de n couples (x_i,u_i) d'observations de X et U on peut donc estimer σ_x par s_x calculé à partir de

$$s_x^2 = \{1/(n-1)\}. \sum_{i=1}^{n} (x_i-\bar{x})^2 \qquad (11)$$

σ_u par s_u calculé à partir de

$$s_u^2 = \{1/(n-1)\}. \sum_{i=1}^{n} (u_i-\bar{u})^2 \qquad (12)$$

Mxu par Mxu calculé à partir de

$$Mxu = \{1/(n-1)\}. \sum_{i=1}^{n} (x_i-\bar{x}).(u_i-\bar{u}) \qquad (13)$$

ρ par r_{xu} tel que

$$r_{xu} = \frac{M_{xu}}{\sqrt{s_x^2.s_u^2}} \qquad (14)$$

β_{xu} par b_{xu} tel que

$$b_{xu} = M_{xu}/s_u^2 \qquad (15)$$

β_{ux} par b_{ux} tel que

$$b_{ux} = M_{xu}/s_u^2 \qquad (16)$$

Remarque.

Si l'on applique la méthode des moindres carrés à l'estimation des coefficients des droites de régression de X sur U et de U sur X on retrouve les mêmes estimations de b_{xu} et b_{ux}. En effet, soit par exemple

$$x(u) = a + b_{xu}.\ u \qquad (17)$$

l'équation de la droite de régression de X sur U.

Si liaison il y a entre X et U, l'une des positions les plus probables de cette droite peut être celle qui minimise, globalement, les écarts entre les

valeurs observées x_i et les valeurs estimées $x(u)_i$ données par l'équation de la droite ci-dessus, c'est-à-dire, en fait, qui minimise la quantité

$$Q = \sum_{i=1}^{n} \{x_i - x(u)_i\}^2 = \sum_{i=1}^{n} (x_i - a - b_{xu}.u_i)^2 \qquad (18)$$

Comme x_i et u_i sont connus, le minimun peut être obtenu si l'on a simultanément :

$$\frac{\delta Q}{d_a} = \frac{\delta Q}{db_{xu}} = 0 \qquad (19)$$

c'est-à-dire :

$$\begin{cases} \sum_{i=1}^{n} (x_i - a - b_{xu}.u_i) = 0 & (20) \\ \sum_{i=1}^{n} (x_i - a - b_{xu}.u_i) .u_i = 0 & (21) \end{cases}$$

système de deux équations du premier degré à deux inconnues qui permet d'estimer les deux inconnues a et b_{xu} par

$$a = \bar{x} - b_{xu}.\bar{u} \qquad (22)$$

$$b_{xu} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (u_i - \bar{u}) \Big/ \sum_{i=1}^{n} (u_i - \bar{u})^2 \qquad (23)$$

La droite de régression ainsi définie est bien confondue avec celle donnée par l'équation (3) : elle a la même pente et passe par le même point M $(\bar{x},\bar{u})$.

2.3. Conditions d'utilisation des estimations de r_{xu}, b_{xu} et b_{ux}.

Les quantités r_{xu}, b_{xu} et b_{ux} peuvent toujours être calculées. Leurs estimations reposent néanmoins sur deux hypothèses :

. la normalité des distributions marginales de x et de u ;
. la linéarité des régressions.

La vérification de la première est toujours possible pourvu que le nombre n de couples d'observations soit au moins égal à 5.

Celle de la seconde, qui implique une mise en classes des observations, n'est réalisable, de ce fait, que si n est suffisamment grand (20 à l'extrême minimum).

Un autre problème se pose d'autre part :

. celui du seuil de signification du r_{xu}, b_{xu} ou b_{ux}

En toute logique cette question doit d'ailleurs être réglée avant celle de la linéarité : si r_{xu} n'est pas significativement différent de zéro, un lien non-linéaire peut néanmoins exister entre X et U dont la signification peut être testée si X et U sont normales.

2.3.1. Vérification de la normalité de X et de U.

Tester la normalité de X et de U peut consister en la comparaison de leurs distributions réduites à la distribution normale réduite.

Cette comparaison, qui se conduit de la même façon sur les deux variables, gagne beaucoup en simplicité à s'appuyer sur le test de KOLMOGOROV et SMIRNOV (cf. in DAGNELIE, Théorie et Méthodes Statistiques . Applications Agronomiques, Vol. II, paragraphe 12.2.4, pages 70 et suivantes).

2311 - Principe du test.

Soit U la variable à tester et s_u l'écart-type de sa distribution.

Rangeons les n valeurs observées de U.

Soient u_m la plus petite, u_M la plus grande, u_i la $c^{ème}$.
La fréquence relative cumulée "observée" de u_i est

$$fco(u_i) = c/n \qquad (24)$$

Sa valeur réduite correspondante est :

$$t(u_i) = (u_i - \bar{u})/s_u \qquad (25)$$

Si la distribution de U suit une loi normale, comme on sait faire correspondre une fréquence cumulée théorique fcth (u_i) à toute valeur $t(u_i)$ observée de la variable réduite (cette fréquence est donnée par la fonction de répartition Φt, cf. le paragraphe 1 et la table 1 de l'annexe 1) ont peut donc calculer pour chaque valeur rangée de U l'écart absolu des fréquences relatives cumulées théoriques et observées.

Pour la c^{eme} cet écart est

$$e_c = |fcth(u_i)-fco(u_i)| \qquad (26)$$

KOLMOGOROV et SMIRNOV ont montré que, pour que la courbe de répartition observée puisse être considérée comme normale, ces écarts absolus ne doivent pas dépasser des valeurs limites maximales. Ces valeurs sont fonctions des seuils de probabilité retenus et des effectifs de la population.

D'après DAGNELIE (cf référence ci-dessus), pour des distributions normales définies par leurs moyennes et leur écart-types estimés, u et s_u, les valeurs critiques relatives aux seuils 0,05 et 0,01 sont approximativement pour $n > 5$

$$e_{o,o5} = 0,886/\sqrt{n+1,5} \qquad (27)$$

$$e_{o,o1} = 1,031/\sqrt{n+1,5} \qquad (28)$$

La comparaison des écarts absolus observés e_c à ces limites $e_{o,o5}$ et $e_{o,o1}$ permet donc d'accepter ou non l'hypothèse de la normalité de U.

2312 - Aspects particuliers de la conduite du test.

Le test de KOLMOGOROV et SMIRNOV peut être, comme on vient de le voir, conduit sur les valeurs individuelles rangées de la distribution observée. L'importance des opérations de rangement et des calculs à effectuer (calculs de $t(u_i)$, $fcth(u_i)$, etc) d'une part, la nécessité, de l'autre, de classer ensuite les données pour le test de linéarité, font que l'on a généralement avantage, chaque fois que le test de linéarité est applicable, à réaliser le test précédent au niveau des limites de classes de la distribution observée.

2312.1 - Conduite du test de KOLMOGOROV et SMIRNOV sur des distributions classées.

Soient alors I l'étendue de la distribution observée et k le nombre de classes choisi (cf. plus loin les problèmes relatifs à la détermination de k).

L'intervale de classe i peut être estimé par :

$$i = I/k = (u_M - u_m)/k \tag{29}$$

et les limite des k classes successives par :

pour la première classe :
$$\begin{cases} u_{inf} = u_o = u_m & (30) \\ u_{sup} = u_1 = u_m + i & (31) \end{cases}$$

pour la c^{eme} classe :
$$\begin{cases} u_{inf} = u_{c-1} = u_m + (c-1).i & (32) \\ u_{sup} = u_c = u_m + c.i & (33) \end{cases}$$

pour la k^{eme} classe
$$\begin{cases} u_{inf} = u_{k-1} = u_m + (k-1).i & (34) \\ u_{sup} = u_k = u_M & (35) \end{cases}$$

Les effectifs de U présents dans ces classes sont alors :

$$nu_1 , nu_2 , \ldots , nu_c , \ldots , nu_k ,$$

et les effectifs cumulés correspondants :

$$ncu_1 , ncu_2 , \ldots , ncu_c \ldots , n$$

avec

$$ncu_c = \sum_{c=1}^{n} nu_c \tag{36}$$

A ces effectifs correspondent les fréquences cumulées relatives observées :

$$fcou_1 , fcou_2 , \ldots , fcou_c , \ldots , 1$$

avec

$$fcou_c = ncu_c/n \tag{37}$$

et les limites supérieures réduites de classes

$$tu_1 , tu_2 ,..., tu_c , ...,tu_k$$

telles que

$$tu_c = (u_c - \bar{u}) / s_u \qquad (38)$$

On peut alors appliquer le test de KOLMOGOROV et SMIRNOV en déterminant la fréquence cumulée théorique $fcthu_c$ correspondant à chaque limite supérieure de classe réduite tu_c, et en calculant

$$e_c = \left| fcthu_c - fcou_c \right| \qquad (39)$$

que l'on compare aux limites théoriques données par les formules (27) et (28)

2312.2 - Problème du nombre de classes k à retenir.

Le nombre de classes à retenir peut être approximativement défini par l'une des deux formules suivantes :

$$k_1 = 1/2. \{n/10 - \log_{10}(n)\} \qquad (41)$$

$$k_2 = 10. \log_{10}(n/10) \qquad (42)$$

Ces formules n'aboutissent pas aux mêmes résultats.

Formules possibles	Nombres de classes donnés par les formules (chiffres supérieurs) et à prendre (chiffres inférieurs) en fonction de quelques effectifs.											
	20	26	32	40	51	64	80	100	200	330	400	500
k_1	0,3 -	0,5 -	0,8 -	1,1 -	1,7 -	2,2 2	3,0 3	4,0 4	8,8 8	15,2 15	18,7 18	23,6 23
k_2	3,0 3	4,1 4	5,0 5	6,0 6	7,0 7	8,0 8	9,0 9	10,0 10	13,0 13	15,0 15	16,0 16	16,9 16

La seconde, comme on le remarque sur le tableau ci-dessus, semble convenir aux populations de faibles effectifs. Si on l'adopte, on peut, en principe, essayer de mettre en classe des populations de 20 observations. Les nombre de classes à prendre sont alors de :

3	pour	20	$\leq$	n	<	26
4	"	26	<	n	<	32
5	"	32	<	n	<	40
6	"	40	<	n	<	51
7	"	51	<	n	<	64
8	"	64	<	n	<	80
9	"	80	<	n	<	100

etc...

Remarque.

Une troisième formule empirique donnée par STURGES $k_3 = 1 + (10/3).\log_{10} n$ aboutit à des nombres encore plus élevés de classes que la seconde pour les petits échantillons.

Elle ne semble pas adaptée aux problèmes de l'étude d'un agrosystème.

2312.3 - Problème de la détermination de la fréquence relative cumulée théorique dans le cas d'une réalisation entièrement automatique du test.

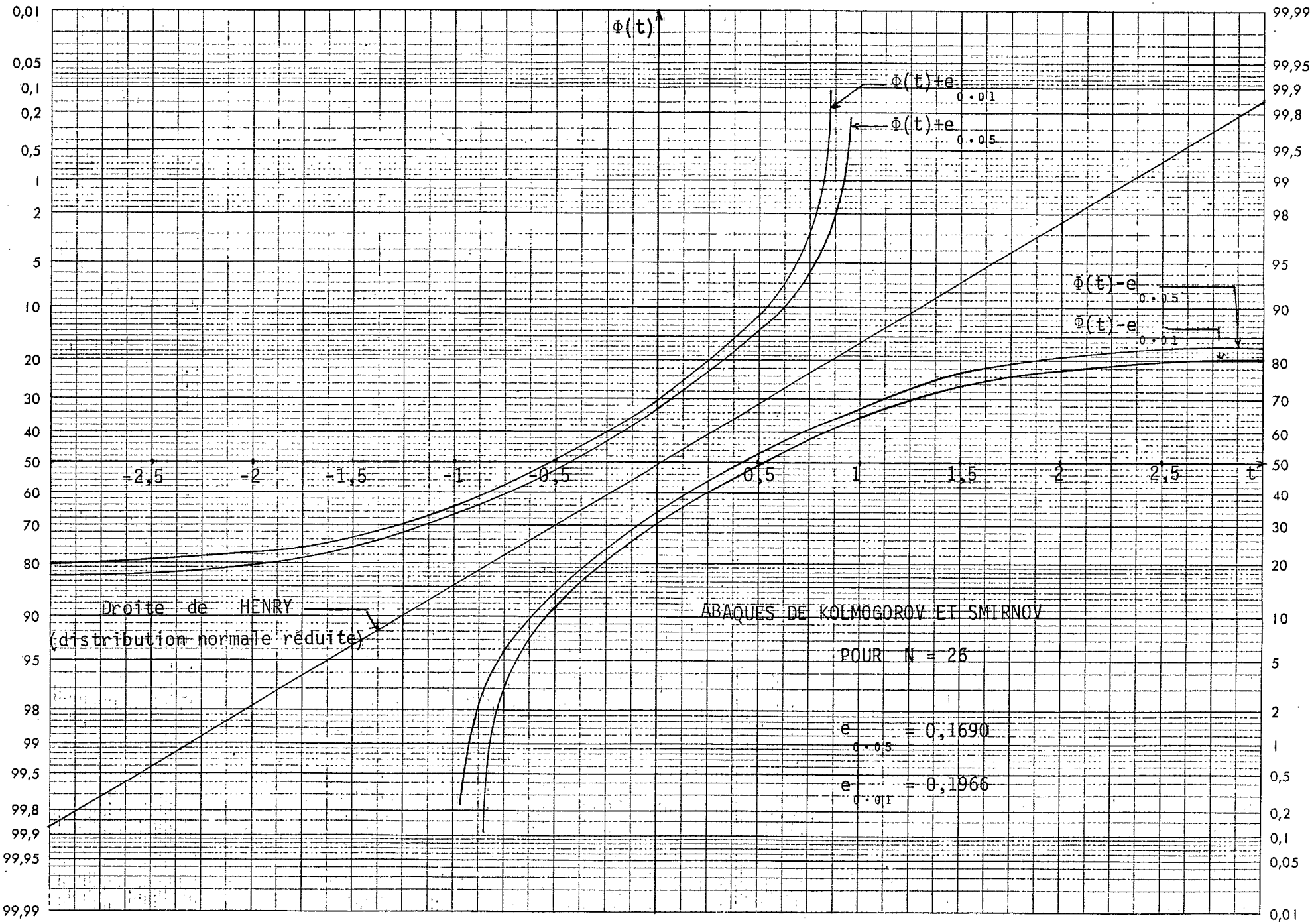
La fréquence relative cumulée théorique fcth est donnée, classiquement, par les tables de la fonction de répartition :

$$\text{fcth} = \Phi(t) = \int_{-\infty}^{t} (1/\sqrt{2\pi}).e^{-t^2/2}.dt \qquad (43)$$

Dans le cas où les moyens de traitement des données le permettent, la "meilleure approximation de HASTINGS" donne directement une valeur suffisamment précise de $\Phi(t)$ (cf annexe 1, paragraphe 1).

2312.4. - Réalisation en série du test dans le cas où l'on ne peut calculer la fréquence relative cumulée théorique.

Dans le cas où les moyens de traitement des données ne permettent pas de calculer les quantités fcth par la meilleure approximation de HASTINGS, mais où


Φ(t)
Φ(t)+e 0.01
Φ(t)+e 0.05
Φ(t)-e 0.05
Φ(t)-e 0.01
Droite de HENRY
(distribution normale réduite)
ABAQUES DE KOLMOGOROV ET SMIRNOV
POUR N = 25
e 0.05 = 0,1690
e 0.01 = 0,1966
t
-2,5
-2
-1,5
-1
-0,5
0,5
1
1,5
2
2,5
0,01
0,05
0,1
0,2
0,5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99,5
99,8
99,9
99,95
99,99

l'on a un nombre assez important de distributions de même effectif n à tester, il peut être avantageux d'établir préalablement un "diagramme de KOLMOGOROV et SMIRNOV" sur du papier probit.

Pour mémoire, l'échelle des ordonnées de ce papier est telle que les fonctions de répartition des distributions normales sont représentées par des droites (droites de HENRY).

A partir de la droite représentative de la fonction de répartition de la loi normale réduite et des valeurs critiques théoriques $e_{0,05}$ et $e_{0,01}$, il est possible , en effet, de tracer sur ce papier les deux courbes, la première située à une distance +e de la droite, la seconde à une distance -e , qui délimitent l'aire du diagramme à l'intérieur de laquelle doivent se situer tous les points de coordonnées $\{ t(u_i), fco(u_i)\}$ ou $\{tu_c , fcou_c\}$ d'une loi réduite observée pour que celle-ci puisse être considérée comme normale.

On remarquera au passage, sur le diagramme type ci-contre, que l'on est en fait beaucoup plus exigeant ici en prenant le seuil 0,05 que le seuil 0,01 , les écarts théoriques admissibles augmentant avec $(1-\alpha)$.

2.3.2. - Signification des coefficients de corrélation et de régression.

D'après la formule (10)

$$r^2_{xu} = b_{xu}.b_{ux}$$

tester si r_{xu} est significativement différent de zéro revient à tester si b_{xu} ou b_{ux} le sont aussi, c'est-à-dire à tester si les droites de régression sont ou non parallèles aux axes de coordonnées.

Deux méthodes peuvent être utilisées.

PREMIERE METHODE.

Considérons la régression X = f(U) et la variance à (n-2) degrés de liberté de la variable dépendante X(U) :

$$s^2_{x(u)} = \sum_n \{x_i - x(u_i)\}^2/n-2 \qquad (44)$$

On démontre que la variance du coefficient de régression b_{xu} peut être estimée par :

$$s^2_{b_{xu}} = s^2_{x(u)} / (n-1).s^2_u \qquad (45)$$

En posant $b_{xu} = b$, $s_{bxu} = s_b$ et $r_{xu} = r$, et en se souvenant que :

$$x(u_i) = \bar{x} + b_{xu}.\ (u_i - \bar{u})$$

$$b_{xu} = r_{xu}.(s_x/s_u)$$

on peut donc estimer la variable $t = b/s_b$

$$t = b/s_b = (r/\sqrt{1-r^2}).\sqrt{n-2} \qquad (46)$$

qui permet à l'aide de la table de STUDENT-FISHER de tester si b est significativement différent de zéro pour (n-2) dégrés de liberté, à différents seuils $(1 - \alpha)$ (cf. le paragraphe 221 et la table 3 de l'annexe 1 pour le test en question)

Comme r_{xu} ne privilégie ni X, ni U, on a de plus

$$b_{xu}/s_{bxu} = b_{ux}/s_{bux} \qquad (47)$$

En d'autres termes, la formule (46) permet de tester à la fois si b_{xu}, b_{ux} et r_{xu} sont significativement différents de zéro.

Réalisation du test.

En réalité, on peut ne pas calculer t : une table des valeurs limites de $|r|$ aux seuils $\alpha = 0.05$, 0.01 et 0.001 a été dressée à partir de la formule suivante déduite de la formule (46) :

$$|r_\alpha| = |t_\alpha| / \sqrt{n-2+t^2_\alpha} \qquad (48)$$

Confer la table 7 à la fin de l'annexe 1.

Par exemple, pour 25 dégrés de liberté, si l'on observe $|r| > 0{,}4869$ celui-ci n'a qu'une probabilité $\alpha = 0.01$ de réapparaître avec la même intensité (mais a, par contre, la probabilité $(1-\alpha) = 0.99$ d'être significativement différent de zéro).

Cependant, lorsque les moyens de calculs le permettent, on peut aussi déterminer directement la probabilité P' (t) correspondant à la valeur observée t (formule 46) à l'aide des formules approchées indiquées à l'annexe 1 paragraphe 222.

Il est facile, ensuite de comparer P' (t) à α = 0.05, 0.01, 0.001, ...etc.

DEUXIEME METHODE

FISHER a montré que la variable

$$z = (1/2).\ L_N\ (1+r)/(1-r) \qquad (49)$$

est distribuée presque normalement quelles que soient ρ la valeur exacte de r et n l'effectif des couples (X,U), et, par ailleurs, que la variance de z est

$$s_z^2 = 1/(n-3) \qquad (50)$$

Par cette méthode ont peut donc aisément déterminer l'intervale de confiance de r ou comparer r à 0 ou à une valeur théorique $r \neq 0$, ou encore comparer deux ou plusieurs valeurs observées de r.

Intervalle de confiance de r.

Soient donc $z = (1/2).\ L_N\ (1+r)/(1-r)$, α le seuil de probabilité choisi, z_m et z_M les limites inférieure et supérieure de l'intervalle de confiance de z. Nous avons :

$$\begin{aligned} z_m &= z - t\alpha.s_z = z - t\alpha/\sqrt{n-3} \\ z_M &= z + t\alpha.s_z = z + t\alpha/\sqrt{n-3} \end{aligned} \qquad (51)$$

et, en utilisant la fonction inverse,

$$r = thz = (e^{2z}-1)\ /\ (e^{2z}+1)\ , \qquad (52)$$

$$\begin{aligned} r_m &= thz_m \\ r_M &= thz_M \end{aligned} \qquad (53)$$

Du fait du caractère quasi-normal de la distribution de z, les limites z_m et z_M peuvent être calculées quel que soit l'effectif n : en effet t_α est égal : 1.96,2.58 ou 3.29 pour les seuils classiques α = 0.05, 0.01 ou 0.001.

Comparaison de r à une valeur théorique r_o.

Soient z et z_o les transformées de r et r_o.

Calculons

$$t = (z-z_o)/s_z=(z-z_o).\sqrt{n-3} \qquad (54)$$

On peut aisément, à l'aide de la table des valeurs de la fonction de répartition (cf. le paragraphe 211 et la table 2 de l'annexe 1), ou grâce à la meilleure approximation de HASTINGS, déterminer la probabilité correspondant à la valeur observée de t ou comparer t observée à t_α.

Remarques.

1 - Lorsque le nombre de couples est faible, une meilleure estimation $\hat{\rho}$ de r peut être obtenue à partir de la formule suivante :

$$\hat{\rho}= r \{1 + (1-r^2)/2(n-3)\} \qquad (55)$$

(travaux de OLKIN et PRATT, 1958, in DAGNELIE 1974)

2 - L'application de la formule (50) implique que l'on ait $n \geqslant 4$

2.3.3. Vérification de la linéarité des régressions.

Vérifier la linéarité des régressions consiste, plus précisément, en l'application de deux tests successifs de non-linéarité, le premier à la régression de X sur U, le second à la régression de U sur X. Ces tests se conduisent naturellement de la même façon dans les deux cas. Nous n'exposerons donc que la démarche à suivre pour tester la non-linéarité de X = f(U), mais les deux tests doivent être réalisés, un lien linéaire testé au niveau de X = f(U) n'excluant pas un lien non-linéaire au niveau U = g(X).

2.3.3.1.Principe du test.

Le test repose sur la comparaison de deux variances de X estimées grâce à

une mise en classes préalable des données sur la variable agissante U, et représentant,

- la première, le carré moyen des écarts à la droite de régression des moyennes par classe de la variable dépendante X,

- la seconde, le carré moyen des écarts par rapport aux moyennes par classe des valeurs observées de X dans les classes.

2.3.3.2. Conduite du test.

Le test comportera les étapes suivantes :

1°/ - la mise en classes de U (k classes). Ce travail peut, en réalité, avoir été déjà entrepris pour le test de normalité de U (cf paragraphe 2.3.2.1) ;

2°/ - le calcul :

a) - des moyennes par classe de U. Soit $\overline{u}_c$ la moyenne de la c^{eme} classe de U ;

b) - des moyennes correspondantes de X, observées $\overline{x}_c$ et estimées $\overline{x}'_c$ à partir de l'équation de la droite de régression

$$\overline{x}'_c = \overline{x} + b_{xu}.(\overline{u}_c - \overline{u}) \qquad (56)$$

c) - des variances s_D^2 et s_E^2 , carrés moyens des quantités Q_D et Q_E à (k-2) et (n-k) dégrés de liberté, éléments de l'équation des sommes des carrés des écarts

$$Q = Q_C + Q_E = Q_R + Q_D + Q_E \qquad (57)$$

établie sur l'identité

$$(x_i - \overline{x}) \equiv (\overline{x}_c - \overline{x}) + (x_{yc} - \overline{x}_c) \equiv (\overline{x}'_c - \overline{x}) + (\overline{x}_c - \overline{x}'_c) + (x_{yc} - \overline{x}_c) \qquad (58)$$

identité dans laquelle x_i, qui est la i^{eme} observation de X, correspond à u_i qui est aussi la y^{eme} observation de U dans la c^{eme} classe de U. Il en résulte que :

$$u_i \equiv u_{yc} \longrightarrow x_i \equiv x_{yc}$$

Ces sommes des carrés des écarts sont donc données par :

$$Q = \sum_{n} (x_i - \bar{x})^2 = \sum_{c=1}^{k} \sum_{y=1}^{nuc} (x_{yc} - \bar{x})^2 \qquad (59)$$

$$Q_C = \sum_{c=1}^{k} nu_c . (\bar{x}_c - \bar{x})^2 \qquad (60)$$

$$Q_E = \sum_{c=1}^{k} \sum_{y=1}^{nuc} (x_{yc} - \bar{x}_c)^2 \qquad (61)$$

$$Q_R = \sum_{c=1}^{k} nu_c . (\bar{x}'_c - \bar{x})^2 = b^2_{xu} . \sum_{c=1}^{k} nu_c . (\bar{u}_c - u^2) \qquad (62)$$

$$Q_D = \sum_{c=1}^{k} nu_c . (\bar{x}_c - \bar{x}'_c)^2 \qquad (63)$$

Les carrés moyens s^2_x, s^2_C, s^2_E, s^2_R, et s^2_D qui leur correspondent représentent différentes estimations de la variance σ^2. Ainsi :

$s^2_x = Q/(n-1)$ est, pour mémoire (cf paragraphe 22), l'estimation de la variance de la distribution marginale de X,

$s^2_C = Q_C/(k-1)$, appelée "variance inter-classe", est une autre estimation de σ^2_x, mais seulement si X n'est pas liée à U ;

$s^2_E = Q_E/(n-k)$, appelée "variance intra-classe", est encore une autre estimation de σ^2_x, mais indépendante,elle, de toute hypothèse sur l'influence de U sur X ;

$s^2_R = Q_R/1$, qui représente la part de la variation de X imputable au "facteur systématique régression linéaire", est aussi une estimation de σ^2_x si la régression linéaire n'est pas significative ;

$s^2_D = Q_D/(k-2)$, enfin, est une dernière estimation de σ^2_x si l'hypothèse de la linéarité est exacte ;

3°/ le calcul et la comparaison à F_α pour (k-2) et (n-k) degrés de liberté du rapport

$$F_{N1} = s^2_D / s^2_E \qquad (64)$$

D'après ce qui précède, si $F_{NL} \geq F_{\alpha}$, l'hypothèse de la non-linéarité peut être acceptée au seuil α et la régression ne pourra donc pas être considérée comme linéaire (pour le test F voir le paragraphe 3 de l'annexe 1 et les tables 4 à la fin de la même annexe).

Remarque.

Une réalisation entièrement automatique du test de F_{NL} est possible à l'aide des formules qui permettent de déterminer la probabilité α correspondant à une valeur observée de F (pour ces formules cf le paragraphe 32 de l'annexe 1) .

Si α observée < 0.05, c'est-à-dire si $(1-\alpha) > 0.95$ on acceptera l'hypothèse de la non-linéarité.

3 - CORRELATION DANS LE CAS OU X ET U SONT DISTRIBUEES NORMALEMENT MAIS NE SONT PAS LIEES LINEAIREMENT.

Au cours des opérations qui viennent d'être décrites deux éventualités peuvent se produire :

- ou bien r_{xu} n'est pas significativement différent de zéro, mais on se demande cependant si les deux variables ne sont pas liées non-linéairement ;

- ou bien r_{xu} est significativement différent de zéro mais l'une des deux régressions, ou les deux, sont significativement non-linéaires.

Dans ces conditions, l'application aux résultats observés de la loi normale à deux variables pour le calcul de r_{xu}, b_{xu}, b_{ux} n'est plus possible.Deux nouveaux coefficients - les rapports de corrélation de X sur U et de U sur X - doivent être utilisés pour caractériser l'intensité de la liaison.

La conduite des calculs est la même dans les deux cas.

3.1 - Rapport de corrélation de X sur U.

Considérons donc la régression $X = f(U)$ et le classement de X sur U.

Nous avons vu au paragraphe précédent (parag.2332) que l'on pouvait décomposer Q, la somme des carrés des écarts de X, en deux composantes Q_C et Q_E telles que

$$Q = Q_C + Q_E$$

Nous avons vu également que $s_{C_2}^2 = Q_C/(k-1)$ et $s_E^2 = Q_E/(n-k)$ représentaient deux estimations indépendantes de σ_x^2, la première si X n'était pas liée à U, la seconde indépendamment de toute hypothèse sur la nature de la liaison.

La comparaison de s_C^2 à s_E^2 permet ainsi de tester la signification de la liaison X = f(U) mise en évidence par le classement, indépendamment de toute hypothèse sur la forme de la liaison.

Si

$$F = s_C^2 / s_E^2 \geqslant F_\alpha \qquad (65)$$

pour (k-1) et (n-k) degrés de liberté, on concluera que les variables X et U sont liées significativement et que la liaison X = f(U) explique

$$100.\frac{QC}{Q} = (\eta_{xu}^2)\% \qquad (66)$$

des variations de X

La quantité $\eta_{xu}^2 = \frac{QC}{Q}$ est appelée "rapport de corrélation".

Elle est toujours positive et comprise entre 0 et 1.

Remarque.

On démontre qu'en théorie $\eta_{xu}^2 \geqslant r_{xu}^2$, l'égalité n'étant réalisée que lorsque la relation X = f(U) est rigoureusement linéaire.

Par l'application de la méthode d'estimation de η_{xu}^2 qui vient d'être exposée, il arrive cependant que l'on observe des valeurs de η_{xu}^2 inférieures à celles des carrés des coefficients de corrélation correspondant. Ceci ne se produit, à notre connaissance, que lorsque n, l'effectif des couples (x_i, u_i) est faible ($n < 40$).

L'explication de cette situation paradoxale réside simplement dans le fait que l'estimation de η_{xu}^2 repose, en partie, sur des décisions empiriques : celles concernant la mise en classe des données (choix du nombre de classes, définition de l'étendue de référence), tandis que l'estimation de r_{xu}^2 ne dépend que des valeurs observées des couples (x_i, u_i).

L'interprétation qu'il semble falloir donner à de telles situations est que η_{xu}^2 est voisin de r_{xu}^2 ou que les deux coefficients ne sont pas significativement

différents de zéro.

3.2. Rapport de corrélation de U sur X.

L'estimation de η^2_{ux}, rapport de corrélation de U sur X doit toujours être faite après celle de η^2_{xu}, que ce dernier soit ou non significatif.

En effet, non seulement la probabilité de mettre en évidence la liaison susceptible d'exister entre les deux variables est aussi élevée, a priori, à partir du classement de U sur X qu'à partir de celui de X sur U, mais surtout cette liaison peut très bien ne pas apparaître significative dans un cas et très hautement dans l'autre, les estimations des rapports de corrélation dépendant, aussi, de la forme réelle de la liaison, c'est-à-dire de celle du graphique des moyennes liées qui la traduit le mieux.

3.3. Linéarisation des relations X = f(U) et U = g(X)

Montrer l'existence d'un lien entre les deux variables sans pouvoir traduire celui-ci par une fonction d'estimation particulière s'est évidemment pas très satisfaisant pour l'esprit, même si la forme générale de la liaison déduite de celle des graphiques des moyennes liées permet d'en décrire les traits essentiels et les conséquences éventuelles sur les variations de A.

Ceci est d'autant plus sensible que les régressions X = f(U) et, ou U = g(X) apparaissent nettement.

La linéarité des régressions est indispensable, par ailleurs, aux analyses de covariance.

S'agissant d'une façon générale des problèmes de la régression curvillinéaire, le lecteur aura intérêt à se reporter aux ouvrages traitant de la question, en particulier à ceux de DAGNELIE déjà cités (1969, paragraphique 2.10 p. 101 et suivantes, et 1970, paragraphe 18.7, p 297 et suivantes).

Le schéma général de la démarche, qui est exposé ci-dessous, n'est donné qu'à titre indicatif.

3.3.1. Principe d'une démarche générale.

Soient donc deux caractéristiques X et U distribuées normalement et liées significativement, mais de façon non-linéaire.

Soient, par ailleurs, T(X) et M(U) deux fonctions mathématiques particulières dont l'application à X et U permet de faire correspondre à tout couple (x_i, u_i) de valeurs observées, un couple (x'_i, u'_i) de valeurs transgénérées telles que :

$$x'_i = T(x_i)$$

$$u'_i = M(u_i), \qquad (67)$$

T(X) et M(U) étant choisies, — soit de façon empirique, soit sur des bases théoriques—,de telle sorte que les relations X' = f(U') ou U' = g(X') puissent apparaître linéaires.

Corrélons alors X' et U' en suivant les étapes décrites au paragraphe 2.

Si, à l'issue des tests, X' et U' apparaissent toujours liées significativement, mais linéairement cette fois-ci, et si d'autre part leurs distributions marginales sont encore normales, alors T(X) et M(U) seront considérées comme des fonctions de linéarisation de la régression.

Leur application aux valeurs observées de X et de U permettra d'estimer X à partir de U, et réciproquement, à l'aide des équations des droites de régressions des données transgérées.

$$x'(u'_i) = \overline{x} + b_{x'u'} \cdot (u'_i - \overline{u'})$$

$$u'(x'_i) = \overline{u'} + b_{u'x'} \cdot (x'_i - \overline{x'}) \qquad (68)$$

et des fonctions inverses des fonctions de transgénération $T^{-1}(X)$ et $M^{-1}(U)$.

3.3.2. Fonctions de linéarisation utilisées couramment.

Pratiquement, un petit nombre seulement de fonctions sont utilisées pour linéariser. Ces fonctions sont les mêmes que celles mises en oeuvre pour répondre à un autre problème : celui de la normalisation des caractéristiques.

On se reportera donc au tableau ci-après présentant ces fonctions de transgénération .

4 - CORRELATION DANS LE CAS OU X ET U NE SONT PAS DISTRIBUEES NORMALEMENT.

Une autre éventualité peut se présenter au cours des opérations de corrélation

de X et U : la non-normalité de l'une des deux caractéristiques, ou même des deux.

Lorsqu'il en est ainsi, le problème se pose évidemment de savoir si l'on peut transformer la ou les variables pour que leurs distributions respectives deviennent normales (sans modifier le caractère linéaire de la liaison si celle-ci semblait déjà l'être, en ou linéarisant en même temps cette régression si elle semblait ne pas l'être).

Ce problème est en réalité étroitement lié au précédent ainsi qu'à celui de l'homogénéisation des variances (l'homogénéité des variances des caractéristiques homologues de deux ou plusieurs agrosystèmes est l'une des conditions requises pour leur comparaison comme on le verra au chapitre suivant).

4.1. Fonctions de transgénération classiques.

Le tableau ci-après regroupe les principales fonctions de transgénération susceptibles d'être utilisées, en précisant les conditions de leur application.

4.2 Règle empirique de choix d'une fonction de transgénération statilisant les variances.

Lorsqu'aucune raison théorique ne permet de choisir une fonction de transgénération particulière et lorsque l'on dispose d'un nombre suffisant de séries de données d'une même caractéristique X, on peut, si une relation linéaire existe entre le logarithme de la moyenne et celui de la variance, utiliser comme fonction de transgénération

$$T(X) = X^{1-b/2} \qquad (b \neq 2) \tag{69}$$

où b est la pente de la droite de régression établie sur les logarithmes de la variance et de la moyenne :

$$\log s_X^2 = b.\log(\bar{x}) + \log k \tag{70}$$

Cas particuliers

$$1°/ \text{ si } b = 1 \quad , s_X^2 = k.\bar{x} \text{ et } T(X) = \sqrt{X} \tag{71}$$

$$2°/ \text{ si } b = 2 \quad , s_X^2 = k.(\bar{x})^2 \text{ et } T(X) = \log_{10} X \text{ (ou } L_N X) \tag{72}$$

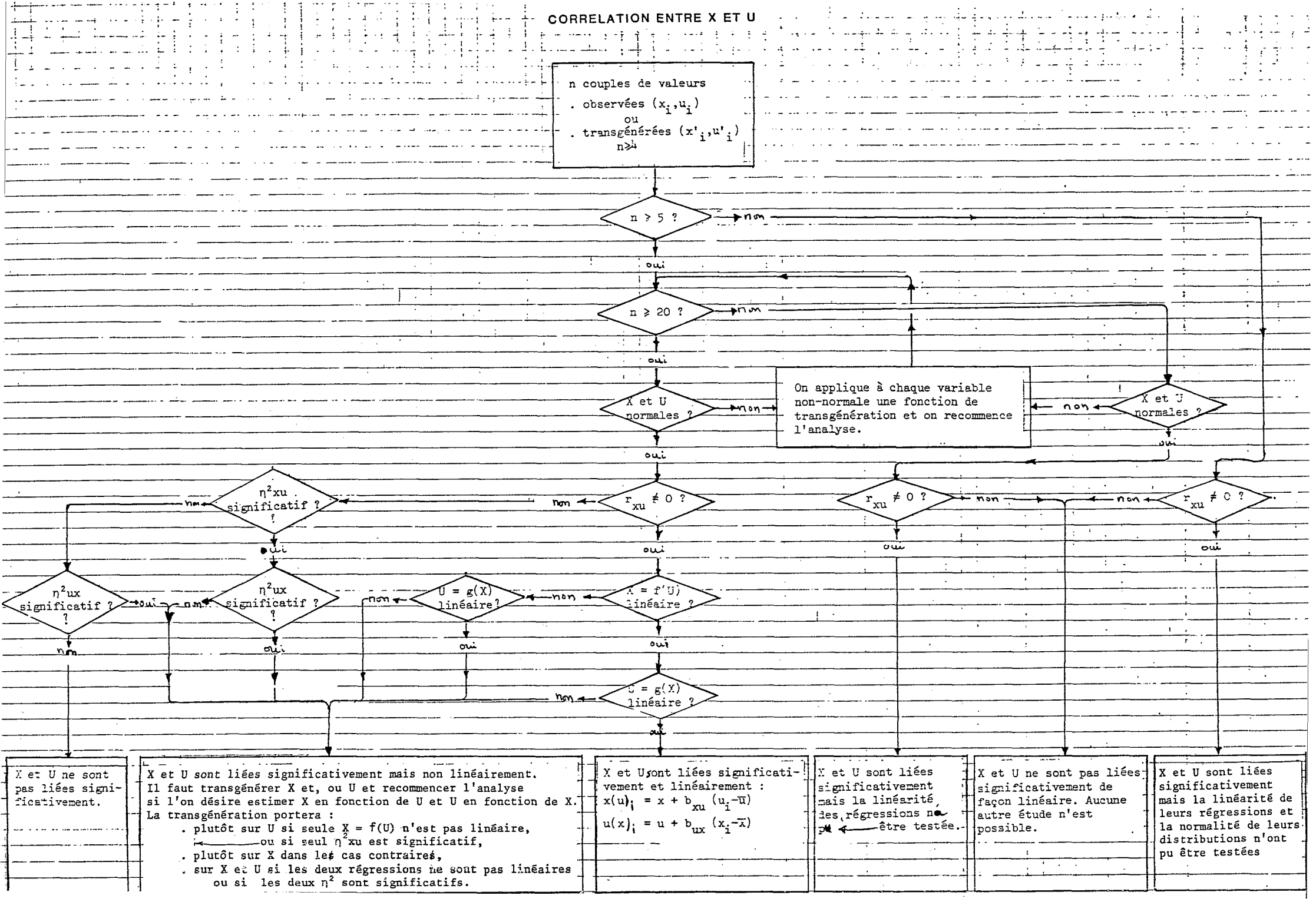
PRINCIPALES FONCTIONS DE TRANSGENERATION $X' = T(X)$ UTILISEES POUR LINEARISER LES REGRESSIONS, NORMALISER LES DISTRIBUTIONS OU STABILISER LES VARIANCES.

Fonctions de transgénération	Conditions d'application
$X' = \sqrt{X}$ Cas particuliers : $X' = \sqrt{X+1/2}$ $X' = \sqrt{X+3/8}$ $X' = \sqrt{X+a}$	X est une variable aléatoire discontinue dont la distribution est proche de celle de l'une des lois de POISSON. Les valeurs observées de X, variable aléatoire discontinue, sont faibles. La valeur de la constante a est recherchée empiriquement
$X' = L_N X$ $X' = \log_{10} X$ $X' = \log_{10} (X+1)$ $X' = \log_{10} (X+a)$	X est distribuée selon une loi log-normale. Ou bien, X est une fonction exponentielle de U (U n'est pas transgénérée simultanément). Ou bien, X est une fonction puissance de U (U est aussi transgénérée par $L_N U$ ou $\log_{10} U$). Les valeurs observées de X sont faibles; $X' = o$ si $X = o$ La valeur de la constante a est recherchée empiriquement.
$X' = \text{arc sin } \sqrt{x/n}$	X est une variable binomiale. Ou bien, X est un rapport compris entre 0 et 1, qui peut être continu.

Remarque :

Pour aucune de ces fonctions on n'observera $\overline{x}' = T(\overline{x})$

CORRELATION ENTRE X ET U



4.3. Remarques sur certains aspests de la non-linéarité.

1 - Du fait de l'influence des modalités de mise en classes des données sur les moyennes liées des caractéristiques et sur les résultats des tests de non-linéarité et de signification du rapport de corrélation (ainsi que sur la valeur de ce dernier), il semble qu'il serait utile de reprendre l'ensemble de ces calculs avec différents nombres de classes encadrant le nombre de classes donné par k_2 (formule 42).

En effet, l'application de fonctions de transgénération peut compliquer par la suite les problèmes d'estimation des valeurs d'une caractéristique en fonction de celles de l'autre.

2 - Lorsque le nombre de classes est suffisant ($n \geq 6$ en pratique), la mise en classe d'une variable sur l'autre peut faire apparaître une classe vide dans la partie centrale de la distribution (au niveau de la quatrième classe, par exemple, dans le cas où $n = 6$). Il y a lieu alors de se demander si le caractère bi-modale de la distribution ne correspondrait pas à un échantillonnage réalisé sur deux populations de la variable utilisée pour le classement.

3 - Dans le même ordre d'idée, l'établissement d'un diagramme des couples observées (x_i, u_i) peut toujours être utile. Une valeur "aberrante" de l'une des deux caractéristiques, voir la présence de deux populations peut apparaître de cette façon.

5. ORGANISATION GENERALE DES CALCULS.

Désirant savoir si un lien pourrait exister entre les deux caractéristiques observées X et U, et, dans l'affirmative, désirant le caractériser par son intensité et par une ou deux fonctions de régression (une si l'on peut considérer que l'une des deux caractéristiques est la variable agissante, deux s'il est impossible de les départager de ce point de vue) on peut résumer les opérations analytiques qui viennent d'être passées en revue par l'organigramme général ci-contre.

CONCLUSION.

Ce tableau permet de se rendre compte, finalement, que si l'on désire caractériser correctement la partie observée de A, le nombre de sites d'observations devra être au minimum de 20.

En réalité, il s'agit-là d'un effectif très faible, car de nombreuses caractéristiques "sol" présentent des coefficients de variation élevés dûs à leur micro-hétérogénéité dans le sol. Les dimensions de cette micro-hétérogénéité du sol sont difficiles à cerner au demeurant.

Par ailleurs, les déterminations analytiques de certaines caractéristiques physiques ou chimiques manquent aussi, parfois, de précision.

Chaque fois que les moyens disponibles le permettront il sera donc préférable de prendre un nombre de sites d'observation très supérieur à 20, et, ou de faire faire en double les analyses chimiques ou physiques les plus imprécises.

Les situations où le nombre de sites d'observations est insuffisant pour caractériser de façon très complète la partie observée de A, sont cependant les plus nombreuses. Elles ne sont pas à rejeter pour autant : chaque fois qu'un ensemble cohérent et homogène d'observations a été constitué, il est toujours intéressant a priori d'en examiner le caractère systématique.