Statistique Impliquée

F.Laloë

Le présent document réunit la majeure partie des communications présentées lors de la cinquième édition de SEMINFOR, le séminaire informatique de l'ORSTOM, réuni à Montpellier du 2 au 4 septembre 1991 sous le titre "Statistique Impliquée". D'une manière générale, l'objectif de ces communications n'était pas de présenter des méthodes ou des résultats nouveaux dans le domaine de la statistique. Il s'agissait plutôt, de la part de scientifiques issus de disciplines très diverses, de présenter des démarches de synthèse d'information en vue de répondre à des questions associées à des problématiques définies dans le cadre de programmes de recherche et relevant de ces diverses disciplines.

Il s'agissait donc en premier lieu de *statistique appliquée.*

C'est à partir de cette idée d'application qu'on peut essayer de mieux caractériser la place de la statistique dans un institut tel que l'ORSTOM, et de discuter le rôle que peuvent y avoir des statisticiens ; statisticiens praticiens, tels que les a décrits Y. Escoufier dans son intervention lors de la table ronde organisée pour clore ces journées.

On peut reprendre, pour ce faire, un élément clé de l'objectif de la biométrie, ou de la démarche du biométricien, tels que les définit J.M. Legay (1976) :

> "notre position est d'accepter les problèmes tels qu'ils sont posés par la pratique de la biologie, de la médecine, de l'agronomie...".

Contrairement aux apparences, la traduction dans les faits de cette déclaration n'est pas évidente. D. Chessel (1992) écrit ainsi à propos de son arrivée au laboratoire de biométrie de l'université Claude Bernard à Lyon :

> "J'imaginais alors que toute université devait être munie d'un tel laboratoire, dit de biométrie, où l'on pensait qu'une formation en mathématique pouvait servir les objectifs d'une discipline expérimentale. J'ignorais, ce qui reste encore assez difficile à comprendre, qu'un tel lieu est en soi un thème de recherches".

Il convient de préciser par quoi identifier une démarche statistique. D'une façon simple et évidente il s'agit déjà de la mise en œuvre d'outils statistiques. Au moment de l'organisation des journées, une classification des différentes communications a ainsi d'abord été envisagée à partir des outils. Il y aurait pu avoir une session "modèle linéaire", une autre "planification des expériences", puis "séries chronologiques" etc...

Mais on peut aussi rechercher une présentation de diverses étapes d'un traitement statistique de l'information et c'est en référence à de telles étapes que les sessions ont été organisées.

Il est assez fréquent d'entendre définir le statisticien par : quelqu'un qui calcule des moyennes et des écarts-types (ou des variances). Et, de fait, on lui demande souvent de calculer l'écart-type associé à une moyenne (la moyenne ne posant quant à elle –de prime abord– pas de problème de calcul).

En s'en tenant à cette définition, on peut se poser la question de savoir en quoi le statisticien pourrait dire qu'une moyenne et un écart-type peuvent constituer une synthèse d'information "satisfaisante" par rapport à une question posée. Une réponse simple et de bon sens vient alors immédiatement : une moyenne et un écart type (ou une variance) sont une synthèse satisfaisante par rapport à une question donnée si elles rendent compte de toute l'information dont on dispose relativement à cette question.

La statistique apporte alors un résultat mathématiquement démontré :

Si $X_1, X_2 \cdots X_n$ sont des variables aléatoires indépendantes distribuées selon des lois normales de mêmes moyenne μ et variance σ^2, alors les variables

$$\overline{X} = \sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

constituent une statistique exhaustive minimale ; parmi toutes les expressions possibles présentant cette qualité, elles sont, de plus, des estimateurs sans biais de variance minimum des paramètres μ et σ^2.

En particulier, toutes les réalisations $x_1, x_2 \cdots x_n$ qui se résument par deux valeurs données $\overline{x}$ et s^2 ont la même probabilité d'occurence. Ceci entraîne que, *conditionnellement aux valeurs prises par les variables $\overline{X}$ et S^2, la probabilité d'occurence d'une réalisation donnée ne dépend pas des paramètres de la loi de distribution.* En d'autres termes, les différences existant entre ces réalisations n'apportent pas d'information sur les valeurs des paramètres μ et σ^2 de la loi des X_i (qualité d'exhaustivité). Le fait que la statistique soit minimale[1] entraîne qu'on ne peut la réduire. Par exemple, la seule moyenne n'est plus une statistique exhaustive, puisque les dispersions des réalisations se résumant par une même valeur moyenne ne sont pas indépendantes de la variance σ^2. Dans cet exemple, la qualité d'exhaustivité minimale de la statistique $\{\overline{X}, S^2\}$ repose sur le fait que les variables $X_1, X_2 \cdots X_n$ ont des lois de répartition normales, indépendantes, de mêmes moyenne et variance. L'exhaustivité minimale n'a de sens qu'en référence à la connaissance, acquise ou supposée, de la nature des lois de distribution des variables aléatoires.

Le concept de statistique exhaustive n'est pas nouveau. Il a été introduit en 1922 par R. Fisher qui écrivait en 1925 (cité par S.F. Arnold 1988) :

> "(sufficient statistic) is equivalent, for all subsequent purposes of estimation, to the original data from which it was derived".

Par la suite ce principe a été étendu pour considérer que toute inférence (estimateurs, tests, intervalles de confiance...) devrait être seulement déduite d'une statistique exhaustive. Ce "principe d'exhaustivité" fait l'objet d'un large consensus chez les statisticiens (Arnold 1988).

[1] Un exposé plus complet pourra être trouvé dans l'article "sufficient statistics" de S.F. Arnold dans Encyclopédia of Statistical Sciences, S. Kotz et N.L. Johnson éditeurs, vol. 9, 1988.

Dans le contexte d'un programme, lorsqu'on recherche des synthèses d'information selon des questions identifiées, on peut s'inspirer de ce principe d'exhaustivité pour considérer que :

> un résumé d'un ensemble d'information est suffisant si les différences possibles entre les ensembles d'information disponible de même nature et se résumant de façon identique n'apportent pas d'éléments de réponse par rapport aux questions posées.

Il devient lors clair que l'information utilisée pour réaliser ces synthèses, le cadre selon lequel les éléments de cette information sont distribués et les questions posées sont des éléments essentiels devant être explicités lors de toute démarche de recherche d'un résumé suffisant.

Les différentes sessions de cette édition de Séminfor ont été définies à partir de ces quelques considérations :

1. la première session, animée par J.D. Lebreton, a naturellement porté sur la **collecte de l'information** ;

2. à l'autre extrêmité de la chaîne, la seconde session, animée par J.P. Chauveau, a porté sur des **indicateurs** et sur leur **pertinence**, c'est-à-dire sur des éléments clés de cadres de synthèse admis au sein de diverses approches disciplinaires et qui en termes statistiques sont considérés comme des éléments incontournables devant figurer dans des synthèses exhaustives ;

3. la troisième session, animée par R. Sabatier, a porté sur l'**analyse descriptive** ; il s'agit là de la description de la distribution des informations, des relations entre variables qualitatives ou quantitatives, pour une meilleure connaissance du cadre probabiliste selon lequel l'information est distribuée ;

4. dans le même ordre d'idées, les communications présentées lors de la quatrième session, animée par J.C. Bergonzini, ont plus particulièrement porté sur des **distributions naturelles** ;

5. La cinquième session, animée par C. Mullon, a porté sur des présentations de **modélisations** et **synthèses** ; Il s'agissait d'exemples de réponses relatives à des questions, réponses faites au travers de modélisations en un sens classique, ou bien en définitive au travers de questions reformulées en fonction de ce qu'on constate savoir dire "de mieux" se rapportant à la question initiale ;

6. En clôture du séminaire, une table ronde a été organisée, animée par M. le Professeur Y. Escoufier, sur la place et le rôle de la statistique et des statisticiens dans un institut tel que l'ORSTOM.

1 collecte de l'information.

La collecte de l'information, telle qu'on peut la décrire à partir des communications présentées, recouvre une grande diversité de situations. Les exposés font toujours référence aux problématiques et aux terrains auprès desquels la collecte est réalisée.

Les questions posées peuvent porter sur les caractéristiques pouvant être utilisées pour la définition d'un système de collecte susceptible de permettre dans de bonnes conditions des interprétations ultérieures. Ainsi, S. Galle recherche si une hypothèse de stabilité temporelle est satisfaite, utilisable pour une répartition raisonnée des sites de mesure d'états hydriques dans une parcelle.

Dans le cas du SIDA, J.L Rey décrit la diversité des questions auxquelles on voudra répondre, chacune d'entre elles conduisant à des stratégies de collecte de natures différentes. L'accent est également mis sur la diversité des situations selon les pays et sur ses conséquences en matière de collecte d'information.

Concernant la collecte d'information de base en géographie rurale, J.C. Roux montre que les cadres sur lesquels on peut habituellement s'appuyer peuvent poser de réelles difficultés face à une situation concrète et ce, aussi bien du point de vue des unités de mesure qui, pour un produit donné dépendent de l'utilisation qui en sera faite, que de celui de la mesure du nombre de personnes actives. En définitive, l'auteur est amené à reposer des questions sur la structure des unités de productions agricoles.

Pour la pêche artisanale au Sénégal, F. Laloë décrit un système de collecte déjà en place, construit à partir d'un cadre de synthèse classique selon lequel cette information devrait être exploitée, et mis en place en bénéficiant d'une solide connaissance de la pêcherie. L'apport de la statistique concerne l'étude même de ce système et l'analyse de la variabilité au sein des données collectées contribue à l'évolution de la nature des synthèses de l'information collectée.

J.P. Minvielle décrit, du double point de vue de l'utilisateur et du concepteur, les caractéristiques d'un logiciel "EMA", susceptible d'accueillir les informations sur les prix dans une région regroupant plusieurs pays, en insistant sur les contraintes liées aux conditions de collecte (comment décrire un produit) et au traitement de données d'origines très diverses.

J.Y. Martin discute de la conception et de la réalisation d'une enquête sur la scolarisation et ses résultats en Guinée, en indiquant la nature des différences et des relations dont l'existence et l'importance devront pouvoir être appréhendées à l'issue de l'opération.

X. Le Roy décrit la collecte et le traitement de l'information nécessaire à la mise en place d'un cadre cartographique de représentation de l'information collectée sur l'activité agricole dans des villages de Côte d'Ivoire.

2 Indicateurs, pertinence.

Les différentes communications regroupées dans cette session ont en commun de discuter d'indicateurs, éléments clés des cadres de représentation au sein de diverses disciplines.

F. Roubaud présente les problèmes associés à l'expression des données économiques dans le cadre de comptes de la nation, dans des contextes où l'économie informelle occupe une place très importante. Les sources de biais associées à

l'évaluation de l'activité informelle, "qui échappe le plus souvent à la collecte statistique", impliquent un effort important et spécifique pour une évaluation, selon des indicateurs communs, d'activités de natures différentes. Cet effort est essentiel en vue d'une possibilité de comparer diverses situations, ou en vue d'évaluer les possibilités d'interventions.

P. Milleville et G. Serpantié montrent comment des indicateurs classiques (de rendements par exemple), adaptés à des conditions de bonne homogénéité de milieu ont dû être complétés dans les cas où l'hétérogénéité de milieu devient très importante, et même "recherchée comme moyen de gestion du risque, ou d'accession à une certaine productivité du travail".

La communication de R. Arvanitis, Y. Chatelin et J. Gaillard porte quant à elle, à partir de bases de données bibliographiques, sur la construction d'indicateurs permettant l'identification de stratégies scientifiques.

P. Livenais, en comparant les résultats de deux recensements au Mexique, montre comment des biais peuvent s'introduire dans l'estimation d'indicateurs clés, tels que la fécondité, rendant extrêmement délicate l'interprétation des résultats issus de ces recensements.

Enfin, B. Maire et F. Delpeuch indiquent comment, à partir d'un ensemble de données d'épidémiologie nutritionnelle, on peut rechercher et définir des indicateurs différents selon divers aspects de la malnutrition, ou encore selon que les questions concernent l'état d'un individu ou d'une population.

3 Analyse descriptive.

L'analyse descriptive est considérée ici comme la recherche dans des ensembles d'information, de types de distributions, d'associations, de coïncidences, de relations... entre les données quantitatives et/ou qualitatives.

En s'appuyant sur des exemples concrets et en utilisant des programmes informatiques adaptés, P. Waniez expose les principes généraux de l'analyse exploratoire.

En utilisant divers outils d'analyse des données, J. Ferraris et A. Samba font une analyse typologique des résultats de sorties de pêche. Cette analyse leur permet de préciser la nature des tactiques et des stratégies des pêcheurs artisans Sénégalais.

H. Robain établit les caractéristiques de sols en Guyane française. L'utilisation d'outils adaptés à l'analyse d'un grand nombre de variables de natures différentes permet des interprétations de la pédogénèse de ces sols.

Y Arnaud et F. Laloë proposent à partir d'une image unique et à l'aide d'une analyse discriminante "classique" un critère de classification de la phase d'évolution d'un nuage dans la zone soudano-sahélienne.

En insistant sur l'adaptation de l'expression cartographique de résultats obtenus à l'aide de divers outils d'analyse des données, M.M. Thomassin propose une description de l'évolution des prix des terres labourables et des prairies permanentes dans les départements français entre 1972 et 1985.

4 distributions naturelles.

Les caractéristiques de la loi de distribution d'une variable ou d'un nombre relativement réduit de variables, peuvent être au centre de certaines études. Les communications présentant cet aspect ont été regroupées dans cette session.

Pour les observations de précipitations journalières, H. Lubès pose ainsi le problèmes liés aux "troncatures", lorsque les observations inférieures à une valeur x_h strictement positive ne permettent pas de décider si elles correspondent à une journée avec ou sans pluie.

Après une présentation sur les lois de distribution de parasites chez leurs hôtes, G. Pichon et C. Mullon présentent deux logiciels sur l'analyse statistique de ces distributions et sur la simulation de relations hôtes parasites.

A partir de la pluie catastrophique observée à Nîmes en octobre 1988, J.M. Masson discute de la possibilité, dans des cadres distributionnels donnés, de mettre en évidence des horsains ("outliers"), c'est-à-dire ne relevant pas de la distribution commune aux réalisations observées en "temps normal".

M. Noirot, selon un critère déterminé à la suite d'une analyse factorielle des correspondances met en évidence des distributions correspondant à un déterminisme génétique simple d'un caractère défini à l'aide de ce critère.

G. Salem et C. Marois indiquent comment des indices d'autocorrélation spatiale peuvent être utilisés pour caractériser la distribution spatiale de l'habitat dans une ville et pour étudier la morphololologie urbaine et la dynamique spatiale.

5 Modélisations, Synthèses.

Les communications présentées dans cette session recouvrent en fait un très large spectre, à l'image des termes mêmes de modèle et de synthèse.

P. Morand et R. Laë utilisent des analyses de tableaux de contingence issus d'enquêtes halieutiques dans le delta central du Niger au Mali, selon des modèles loglinéaires et des régressions logistiques, outils relevant de la panoplie des modèles linéaires généralisés.

K. Simondon présente divers modèles non linéaires de la croissance staturo-pondérale d'enfants, en indiquant leurs qualités respectives selon plusieurs critères d'appréciation.

I. Hurtado, H. Théry, P. Waniez et F. Pelletier présentent un système informatique permettant de suivre l'évolution cartographique des municipios brésiliens et de réaliser des synthèses cartographiques à partir de données issues ou non de traitements statistiques.

C. Mullon, L Bochereau, B. Palagos et G. Pichon montrent comment des problèmes de discriminations peuvent être traités de manière analogue par une approche classique d'analyse des données ou par le recours aux "réseaux de neurones".

M. Pansu, Z. Sallih et P. Bottner présentent un modèle à compartiments pour décrire la cinétique d'évolution de formes du carbone organique dans les sols.

M. Fournier et F. Sondag présentent divers problèmes posés par la mesure de faibles quantités dans des laboratoires d'analyse, en relation avec la capacité de répondre aux questions qui peuvent entraîner le besoin de telles mesures.

En utilisant des ajustements de surfaces de tendances, G. Salem, C. Marois, L. Arreghini et P. Waniez analysent à diverses échelles la densité de population en relation avec les risques dans le domaine de la santé publique.

6 Table ronde sur le thème : place et rôle de la statistique et des statisticiens dans un institut tel que l'ORSTOM.

L'après-midi du 4 septembre a été consacrée à une table ronde, animée M. Y. Escoufier, Vice-Président de l'Université Montpellier II, et réunissant J. Déjardin, biométricien Directeur de recherche de l'ORSTOM, J.D. Lebreton Directeur de recherche au C.N.R.S. et Président de la Société Française de Biométrie, C. Marois, Professeur de Géographie à l'Université de Montréal et J.M Masson, Maître de Conférence au Laboratoire d'Hydrologie Mathématique de l'Université Montpellier II. Les textes reprenant les cinq interventions sont réunis dans le présent document. Ils offrent un large panorama sur les domaines de l'histoire de la biométrie à l'ORSTOM, la formation à la statistique et le métier de statisticien, la modélisation et la nature de l'hydrologie statistique.

7 Discussion.

La présentation rapide des communications qui vient d'être faite est en définitive très arbitraire. Elle donne une idée très certainement biaisée de leur contenu réel, de par la volonté de les décrire en fonction des sessions auxquelles elles ont été affectées. Dans la grande majorité des cas, comme le lecteur pourra rapidement le constater, d'autres choix auraient pu être faits.

Cela peut signifier que le choix d'organisation était mal adapté ; cela peut également signifier qu'il existe une intégration extrèmement forte de la démarche statistique dans l'ensemble du déroulement d'un programme, ce qui n'est guère surprenant si l'on considère une démarche qui fait tout à la fois référence à l'information, aux objectifs et aux synthèses.

Cette forte intégration dans le déroulement d'un programme entraîne la nécessité de préciser l'affirmation initiale selon laquelle la statistique au centre du sujet de SEMINFOR est appliquée. En effet, l'application, au sens strict, n'implique pas d'intégration forte. Ainsi, lorsque l'information devant être traitée est distribuée selon un modèle statistique bien identifié, garanti par un plan d'expérience bien conçu et mis en place, les questions auxquelles on désire répondre correspondent de façon naturelle à des statistiques exhaustives. Dans le cas d'un plan complètement randomisé à un facteur, le test d'égalité de moyennes peut effectivement s'avérer être la meilleure réponse possible à la question posée de savoir si la variable étudiée a ou non la même espérance selon les diverses modalités du facteur étudié. Dans cette situation "optimale", le statisticien ne sait cependant pas répondre de façon définitive, il sait seulement extraire un maximum d'information relative à la question posée et, comme l'indique Legay (1992) à propos de l'apport de la "période fisherienne" à l'expérimentation :

"Si la confrontation d'une hypothèse à la réalité reste la motivation première, elle n'est pas exclusive, car la réalité est bien plus riche. Tout ce qui n'est pas contrôlé, tout ce qui est momentanément rejeté dans l'aléatoire existe"

Mais il n'est pas rare de ne pas pouvoir bien répondre lorsque par exemple une mauvaise randomisation, ou son absence, conduisent à des confusions entre sources de variations telles qu'on ne peut pas interpréter des différences de moyennes comme différences associées à la source de variation étudiée. Cependant, il serait trop facile de prétendre qu'une incapacité à bien répondre ne saurait provenir que d'une mauvaise planification d'expérience ou d'une collecte inadaptée d'information. Il existe des sources de variation, identifiables ou non, pouvant remettre en cause l'interprétation des résultats. Ainsi, Livenais montre que les estimateurs des mêmes taux de fécondité, appliqués aux données issues de recensements successifs n'ont pas la même espérance. Cela veut dire qu'on ne sait souvent pas bien dire quelle est la qualité d'un estimateur, même appliqué aux données obtenues dans le cadre d'un plan de collecte bien raisonné ; mais cela ne pourrait en aucun cas justifier le rejet pur et simple de l'information ainsi acquise. Legay (1992) conclut ainsi, à propos des expériences :

"Oui une expérience est possible, ...elle est un compromis épistémologique établi dans le cadre de notre ignorance"

Les questions auxquelles on désire répondre ne découlent donc généralement pas de façon naturelle de l'information qu'on traite, même si celle-ci a été rassemblée en fonction de ces questions (cf. communications de la session sur la collecte de l'information), et même si les cadres des synthèses qu'on réalise sont définis de façon plus ou moins explicite par ces questions.

L'analyse de la variabilité, celle des distributions conjointes selon lesquelles les données sont réparties (cf. communications des troisième et quatrième parties), permettent alors une critique des questions définies par les problématiques de recherche. Elles permettent donc une critique des cadres de synthèse et des indicateurs qui en constituent l'ossature (cf. communications de la seconde partie). On peut en effet se poser la question de savoir si l'ensemble des situations dont les caractéristiques de variabilité et de distribution sont semblables, et qui conduisent aux mêmes réponses apparaissent **équivalentes**, auquel cas ces réponses correspondent à un ensemble de questions pouvant être qualifié de "suffisant".

En revenant à la notion d'exhaustivité présentée en introduction, il est possible de discuter de ce que peut être l'équivalence évoquée ci-dessus, en se référant à une relation d'équivalence et aux classes qu'elle détermine. En effet il n'y a pas d'expression unique d'une statistique exhaustive minimale. Si une statistique l'est (par exemple $\overline{X}, S^2$ dans l'exemple donné en introduction), alors toutes les statistiques T telles que pour deux ensembles d'observations

$$T(X_{1,(1)} \cdots X_{n,(1)}) = T(X_{1,(2)} \cdots X_{n,(2)})$$

si et seulement si

$$\overline{X_{(1)}} = \sum_{i=1}^{n} X_{i,(1)} = \overline{X_{(2)}} = \sum_{i=1}^{n} X_{i,(2)}$$

$$\text{et } S^2_{(1)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i,(1)} - \overline{X_{(1)}})^2 = S^2_{(2)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i,(2)} - \overline{X_{(2)}})^2$$

sont elles mêmes des statistiques exhaustives minimales. C'est par exemple le cas de la statistique $\{\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\}$. Si l'on définit donc (Lehmann et Scheffé 1950 , 1955, cités par Arnold 1988) une relation d'équivalence par : "deux jeux d'observations sont équivalents s'ils prennent les mêmes valeurs selon une statistique exhaustive minimale", on définit en même temps les classes d'équivalence de cette relation, délimitant une partition exhaustive minimale associée. Cette partition présente l'énorme intérêt d'être unique, conduisant à considérer qu'en définitive "the partition is what is important" (Arnold, 1988).

Lorsqu'on recherche une synthèse par rapport à un certain nombre de questions identifiées, celles-ci forment une hypothèse selon laquelle la représentation qu'elles définissent *ensemble* constitue une "partition exhaustive". Toutes les situations conduisant aux mêmes réponses sont équivalentes.

L'**application** de la statistique peut être considérée "achevée" lorsque la synthèse est réalisée selon les questions initialement posées – *en ayant éventuellement pu mettre en évidence l'intérêt de développements méthodologiques, sujet sortant du domaine de cette édition de SEMINFOR et qui mérite une attention telle qu'il ne pourrait qu'être mal abordé ici* –.

L'**implication** de la statistique intervient lorsqu'elle permet de préciser, de modifier des questions dans le contexte de la recherche de partitions exhaustives.

Mais on se trouve ainsi dans une certaine impasse car il est tout aussi trivial de dire qu'une synthèse est suffisante, que de dire qu'elle ne l'est pas. Dès lors qu'elle existe, elle est suffisante par rapport à elle-même ; en effet, si on énonce toute l'information collectée, on a une statistique exhaustive. Mais comme l'information dont on dispose n'est jamais qu'une sélection faite dans un ensemble d'information accessible, il existe sûrement une question par rapport à laquelle ce qui ne figure pas dans la sélection a un sens. Si on considère cette sélection comme une opération de synthèse, cette dernière n'est pas exhaustive.

En d'autres termes il est trivial de dire "les écarts à la moyenne ont un sens". En d'autres termes également, les opérations pluridisciplinaires conduisent souvent à des annuaires (Couty 1990). Ces annuaires sont des résultats logiques lorsqu'en l'absence de question – ou, ce qui revient au même, en présence d'un refus de ne pas poser une question – la seule synthèse (apparemment) exhaustive est l'absence (apparente) de synthèse.

La situation est donc assez compliquée, et ne peut guère être clarifiée qu'en identifiant des questions clés. Pour ce faire, on peut citer un extrait de l'intervention de G. Winter lors des "journées de septembre" de 1990 concernant le PEO (Projet d'Etablissement de l'ORSTOM) :

"La personnalité scientifique originale de l'ORSTOM (...) s'exprime par des programmes de recherches :

- qui partent des **questions émises par les acteurs du développement**, les traduisent en problématique scientifique et y reviennent de manière inattendue dans une dialectique amont-aval, source d'applicabilité
- qui affrontent ces questions selon diverses **approches thématiques** qui tôt ou tard, ex ante ou ex post, conjuguent **diverses disciplines**
- ...".

Deux types de questions d'importance majeure apparaissent donc, ainsi que, surtout, la nature du lien qui les associent.

Deux exemples, reflétant des contextes agronomiques et halieutiques peuvent illustrer l'apport possible d'une démarche statistique dans le cadre de programmes de recherches. Ces exemples reprennent pour une bonne part les éléments donnés dans des communications présentées dans ce séminaire (Milleville et Serpantié, Ferraris et Samba, Morand et Laë, Laloë).

On peut rechercher la possibilité d'améliorer des plantes en vue d'une meilleure production d'une agriculture. On peut aussi, en vue d'une gestion rationnelle de stocks, rechercher quelle est la distribution de résultats de l'exploitation d'une ressource halieutique, conditionnellement à une activité de pêche donnée.

Dans ces deux situations, chacun des deux types de questions est présent ; production agricole et aménagement des pêches d'une part et d'autre part problématiques scientifiques posées de façon naturelle dans les domaines de la génétique quantitative et de la dynamique des populations.

L'acquisition d'information peut alors être conduite de façon à répondre le mieux possible aux questions directement liées à ces problématiques : héritabilité des caractères liés au rendement des plantes ; mortalité, fécondité, croissance des individus appartenant aux populations exploitées par pêche... Ces informations sont bien sûr des sous ensembles de ceux dont on dispose relativement à l'agriculture et à la pêche.

Dans le cas de l'agriculture, les questions sur "l'optimisation du rendement de la terre en produit utile" conduisent à l'acquisition de connaissances dans un contexte expérimental adapté, en rejetant l'impact de l'hétérogénité du milieu dans une variabilité la plus réduite possible et, surtout, *indépendante* des sources de variation étudiées. Si on recherche une bonne variété d'une plante cultivable, elle sera d'autant plus intéressante qu'elle sera effectivement bonne dans le spectre le plus large possible de conditions de milieu. Ceci conduit à privilégier l'absence d'interaction entre effet de milieu et effet génétique. Une hypothèse d'absence d'une telle interaction est biologiquement fausse, mais elle le sera d'autant moins que le milieu sera peu variable et que les interactions auxquelles il participe ne s'exprimeront pas, ou faiblement.

Dans le cas de l'halieutique, il convient de donner une description de l'impact de l'activité de pêche sur les diverses composantes de la ressource exploitée. Les diverses formes que peut prendre l'activité des pêcheurs font donc l'objet d'une stratification fondée sur l'homogénéité de cet impact. En évoquant un niveau d'exploitation, la question de gestion conduit à émettre une hypothèse sur une

possible stabilisation de l'effet de cette exploitation sur la ressource, c'est-à-dire une hypothèse sur la stabilité des effectifs des strates selon lesquelles les différentes formes d'activité de pêche ont été regroupées.

Dans les deux cas, les hypothèses de stabilité sont remises en cause par l'analyse de la variabilité au sein de l'information disponible sur l'agriculture ou sur la pêche. Pour l'agriculture, non seulement le milieu peut être variable, mais il s'avère *de plus* que les paysans peuvent accroître une hétérogénité naturelle du milieu dans le cadre de la gestion de leur exploitation. Pour l'halieutique, les pêcheurs, *en relation avec l'information dont ils disposent sur l'"environnement"*, peuvent modifier leur activité, donc leur impact sur la ressource, rendant ainsi, dans le cadre d'une même "quantité d'activité", cet impact potentiellement variable.

La mise en évidence de sources de variabilité et d'interactions associées ne conduit pas à une critique de la qualité des problématiques en tant que telles, mais à une discussion de leur qualité en tant que traduction de questions de développement ou bien, en d'autres termes, à une discussion de leur statut au sein d'une partition exhaustive. Si il reste évident que l'acquisition de connaissances sur la biologie des plantes ou sur la dynamique des stocks est essentielle, la qualité de l'information apportée par ces connaissances est largement conditionnée par les contextes agricoles ou halieutiques dont elles sont issues. En effet selon que les hypothèses sur la variabilité implicitement faites en formulant des questions sur le développement sont acceptables ou non, les réponses aux questions scientifiques constitueront *à elles seules* ou non des synthèses exhaustives.

Il est relativement aisé d'exprimer une preuve d'insuffisance, il suffit en effet d'exprimer une *anecdote*. Ce peut être dans le cas de l'agriculture : "telle variété est bien la meilleure, mais les paysans font leurs semis sur une période trop longue" ou dans celui de la pêche : "oui telle espèce n'apparaît plus guère dans les captures, ou telle autre y devient prépondérante, mais c'est parce que les pêcheurs ont réalisé tel report d'activité, qu'on peut interpréter à la lumière de telle opportunité du marché, ou de tel évènement climatique...". Ces anecdotes sont des apports importants d'information ne figurant pas dans les synthèses. Elles sont donc des preuves d'insuffisance des cadres dont ces synthèses relèvent puisqu'elles sont des expressions de différences porteuses d'information.

Si l'on maintient une hypothèse selon laquelle les connaissances apportées aux questions définies par les problématiques scientifiques constituent une synthèse suffisante relativement à la question posée par le développement, c'est-à-dire si l'apport de la recherche ne consiste qu'en une identification de bonnes variétés cultivables ou d'un nombre d'unités de pêche de tel ou tel type conduisant à terme à une maximisation d'un résultat de la pêche, on valide de fait les hypothèses de stabilité sous-jacentes, ce qui peut dans le meilleur des cas diminuer l'efficacité de la réponse à la question de développement et dans le pire des cas conduire à un échec dû à la négation d'une variabilité qui peut constituer un atout essentiel des unités d'exploitation.

Cette discussion porte sur la qualité des réponses aux questions de développement et, dans la mesure où les problématiques pouvaient être issues de ces questions de développement, la discussion porte naturellement sur la qualité de ces questions elles-mêmes. Lorsqu'une cause d'insuffisance de ces questions est identifiée, la recherche de nouveaux cadres de synthèse devient possible, permettant de nouvelles formulations des questions de développement, de nouvelles traductions en problématiques scientifiques.

C'est par sa contribution au déroulement de ce processus dialectique que la statistique, en tant que *discipline de représentation des connaissances*, s'implique.

Références bibliographiques

Arnold S.F., 1988. sufficient statistics. In encyclopédia of statistical sciences, S. Kotz et N.L. Johnson éditors, vol. 9, 72-80.

Chessel D., 1992. Echanges interdisciplinaires en analyse des données écologiques. Document de synthèse pour l'habilitation à diriger des recherches.

Couty P. 1990. La pratique multidisciplinaire à l'ORSTOM Rapport ORSTOM. 41 p.

Fisher R., 1922. On the mathematical foundation of theoritical statistics. Philos. Trans. R. Soc. (Lond.) Ser. A, 222, 309-368.

Fisher R., 1925. Theory of statistical estimation. Proc. Camb. Philos Soc. 22, 700-725.

Legay J.M., 1976. Pour une Biométrie. Statistique et Analyse des Données : 1, 2, 5-11.

Legay J.M., 1992. Une expérience est-elle possible ? In Biométrie et environnement, Lebreton et al. eds. 16 p. Sous presse.

Lehmann E.L. and Scheffé H., 1950, 1955. Completeness, similar regions and unbiased estimation, Sankhya, 10, 305-340 and 15, 219-236.

Winter G., 1990. Le projet de l'ORSTOM (première proposition). Journées d'études, 4-6 sptembre 1990, 75 p.