Centre de Montpellier

LA STATISTIQUE A L'ORSTOM

UNE INDISCIPLINE IMPLIQUÉE

Francis LALOË

Laboratoire d'hydrobiologie et océanographie tropicales

Document ORSTOM Montpellier, 1990, n° 2

FRANCIS LALOË

LA STATISTIQUE A L'ORSTOM
UNE INDISCIPLINE IMPLIQUÉE

Montpellier ORSTOM 1990

Les opinions exprimées dans ce document
n'engagent que la responsabilité de leurs auteurs

Introduction

Si la statistique est une discipline scientifique, de nature mathématique, il n'existe pas à l'ORSTOM de programme dont l'objet principal soit de faire des découvertes dans ce domaine; et ceci ne semble pas devoir être remis en cause.

L'utilité de la statistique est tout à fait reconnue pour ce qui concerne le "traitement des données". Des méthodes statistiques sont présentes dans une très grande quantité des études. Nombre de disciplines ont "intégré" des outils statistiques, parfois sous un vocable spécifique, permettant l'analyse sous une forme adéquate des sources de variation dont l'étude est une caractéristique de leur domaine d'intérêt. L'utilisation de ces outils dans de telles conditions relève bien entendu du domaine technique, et leur adaptation, lorsqu'elle s'avère nécessaire au vu des avancées réalisées dans le cadre des programmes, justifie le recrutement au sein des disciplines concernées. Ces avancées sont de plus en plus obtenues grâce au développement des techniques de traitement automatique de l'information, et c'est ainsi que les profils de recrutement font de plus en plus fréquemment référence, et la plupart du temps de façon très confuse, à des spécialités "statistique et/ou informatique et/ou mathématique".

En l'absence de problématique clairement statistique, et en présence d'une incontestable expertise statistique au sein des diverses disciplines, la situation apparaît très claire et logique. Elle l'est d'autant plus que l'abondance de logiciels permet l'utilisation sans formation excessive des outils statistiques, accroissant ainsi encore les risques de confusion entre statistique et informatique.

En fait l'utilisation massive de l'informatique s'est traduite dans la plupart des cas par une "augmentation" de la masse d'information disponible, mais elle s'est également traduite par une modification de la nature de cette information. Ceci a bien sur entraîné une évolution implacable des approches, évolution non nécessairement toujours positive si

on admet que la qualité d'une information n'est pas proportionnelle au nombre de données contenues dans un fichier.

P. Renaud a montré comment l'informatique "science du traitement automatique de l'information" s'immisce dans les problématiques de recherche; j'essaierai ici de le faire en ce qui concerne la statistique en tant que "science de la qualification de l'information à partir de l'analyse de la variabilité". Une évaluation d'une information peut être abordée selon plusieurs niveaux.

- Il peut s'agir "simplement" d'une mise en forme faite à partir d'hypothèses données a priori sur la nature de la variabilité, la forme pouvant alors être décrite par l'estimation de paramètres dont la définition est donnée dès le départ.

- Il peut aussi s'agir de la recherche et de l'identification de sources de variation, et donc de la recherche de paramètres permettant une mise en forme.

- Il peut alors s'agir d'une remise en cause d'hypothèses données a priori et d'émission de nouvelles hypothèses sur la forme des choses observées, sur lesquelles on désire faire porter l'information disponible. Il peut donc s'agir d'une immixtion dans les disciplines lorsque les formes qu'elles savent reconnaître ne suffisent plus à appréhender celles des choses qui s'imposent à l'observation. Il s'agit donc d'opérations de recherches indisciplinées et impliquées.

Ces différents niveaux imbriqués offrent un plan naturel de présentation qui sera utilisé avec l'aide d'un certain nombre d'exemples

I La description dans un cadre prédéfini.

Ce premier niveau ne doit en aucun cas être vu comme un parent pauvre. Il constitue dans tous les cas une première étape qu'on ne doit pas a priori considérer comme devant être dépassée. En effet il est normal qu'on recoure à l'outil statistique pour décrire une variabilité dont la source est identifiée (ce qui n'implique pas que la cause le soit). Et il

n'est en rien scandaleux qu'un traitement satisfasse à cette demande. C'est d'ailleurs dans ce contexte que la grande majorité de l'activité statistique se déroule, soit du fait des représentants des disciplines eux mêmes, soit de celui de personnes de formation statistique plus spécifique.

Exemple a. Echantillonnage pour la pêche artisanale sénégalaise.

Le problème est de discuter des qualités d'un plan d'enquêtes pour l'acquisition de données sur l'activité et les résultats de la pêche artisanale. Le recours aux techniques de sondage est incontournable, et les halieutes ont mis au point un plan stratifié avec plusieurs niveaux d'observation dans chaque strate. Au vu de la quantité d'estimations à fournir et au vu des contraintes logistiques, il apparaît que ce plan est en définitive , sinon le meilleur possible, au moins difficilement perfectible à l'aide des outils usuels des techniques de sondage. Le statisticien peut néanmoins expliciter en ses propres termes la nature du plan, les divers estimateurs utilisés, et surtout analyser leur précision en fournissant des estimateurs de leur variance. L'analyse de la variabilité aux divers niveaux dans chaque strate est indispensable pour apprécier cette précision et pour discuter des sources de biais pouvant affecter les estimations.

Exemple b. Plans d'expérience en génétique quantitative.

Il s'agit là d'un des domaines les plus approfondis. L'analyse de la variabilité inter et intra variétale ou raciale permet l'estimation de paramètres génétiques (héritabilité par exemple) à la base de l'amélioration des plantes ou des animaux. Les techniques statistiques (analyse de variance de modèles à effets fixes ou aléatoires, ou de modèles mixtes) font l'objet d'enseignements très lourds dans les divers lieux de formation. Le recours à des plans d'expérience signifie qu'on se place dans des conditions permettant d'optimiser la connaissance des aspects qu'on désire étudier. et les hypothèses sur la forme de la variabilité ne peuvent guère être mieux satisfaites.

Exemple c. Analyse des puissances de pêche.

L'étude de l'impact de l'activité d'une ou de plusieurs flottes de pêche sur une ressource exploitée conduit à rechercher une évaluation de cette activité qui contienne le plus d'information possible sur cet impact. Si le nombre de bateaux, et le nombre de jours d'activité de chacun d'eux, contient de l'information, il est souhaitable d'évaluer les performances de chacun d'entre eux, en retirant éventuellement l'impact des variations spatio-temporelles. L'analyse de variance a été proposée, à partir du logarithme des rendements, d'un modèle incluant un effet "bateau" et l'effet d'un facteur combinant un facteur temps et un facteur spatial. Cette analyse est connue sous le nom de "modèle de Robson". L'absence de plan d'expérience se traduit par des déséquilibres imposant le recours a des programmes informatiques couteux en temps, et a conduit à l'écriture de ces programmes. Leur utilisation permet une estimation "satisfaisante" des puissances de pêche.

Exemple d. Echo-intégration.

L'écho-intégration permet des estimations de biomasses, à partir des échos renvoyés par des poissons à la suite d'émissions sonores. Plusieurs problèmes se posent pour évaluer la précision de ces estimations. Outre les problèmes de mesure eux-mêmes, l'extrapolation des données obtenues tout au long du trajet d'un bateau, pour estimer une biomasse totale ou bien une densité en un point situé en dehors du trajet ne peut se faire à partir de moyennes simples de l'ensemble des résultats. Certains points contiennent plus d'information que d'autres. La géostatistique offre un cadre d'analyse intéressant, même si un certain nombre de problèmes liés au fait qu'une campagne n'est pas "instantanée" paraissent difficiles à résoudre.

Exemple e. Cartographie.

La cartographie, lorsqu'elle est réalisée à partir d'un fond de carte avec des entités surfaciques (un découpage administratif par exemple), peut être vue comme une édition de résultats d'une comparaison multiple de moyennes. Représenter deux entités par des couleurs différentes indique en effet qu'il existe une différence significative entre ces entités pour la quantité étudiée. Lorsqu'on dispose de valeurs exactes, à partir de recensements exhaustifs par exemple), il n'y a guère de problème. Lorsqu'on dispose de données extrapolées (à partir des résultats d'un sondage), ou qu'on désire présenter des résultats se référant à une "population parentale théorique", il est encore possible de cartographier de façon identique à partir de ces données. Dans le cas où il n'existe pas de "fond de carte" donné a priori, on peut recourir à des méthodes de géostatistique (variables régionalisées, surfaces de tendance...)

Exemple f. Regression logistique.

Comme son nom l'indique, il s'agit d'une regression. Elle est utilisée lorsqu'on désire ajuster des données d'une variable qualitative de type 0-1 à partir d'une fonction logistique d'une combinaison linéaire de variables explicatives quantitatives ou qualitatives. Dans ce dernier cas, il s'agit plutôt d'une analyse de la variance, et la méthode devrait en fait être nommée "modèle linéaire logistique". C'est une méthode particulièrement utile permettant de faire un ajustement par maximum de vraisemblance dans le cas où la loi de la variable "dépendante" est binomiale.

Ces exemples illustrent bien la diversité d'utilisation des techniques statistiques. Ils font apparaître que des problèmes rencontrés dans des domaines n'ayant a priori que peu de points communs, sont abordés - au sein même des disciplines concernées - par les mêmes outils. Ceci n'est pas un hasard et correspond de fait à des analyses de variabilité de nature semblable.

En revenant sur ces exemples, on observe que la géostatistique intéresse les océanographes, les géographes, mais aussi les hydrologues qui désirent étudier la variabilité spatiale des précipitations, les pédologues et les géologues qui désirent cartographier à partir de prélèvements. Des

demandes nombreuses sont exprimées dans tous ces domaines, et plusieurs logiciels ont été réalisés à des niveaux divers au sein même de notre institut.

De même le modèle linéaire est-il omniprésent. On le trouve pleinement utilisé dans les plans d'expériences, mais le modèle de Robson est un modèle linéaire, ainsi que l'analyse des surfaces de tendances (régression polynomiale en les coordonnées de chaque point). Il en est de même, au moins en première analyse, du kriegeage qui consiste à rechercher des estimateurs "moyennes pondérées" lorsque les observations ne suivent pas des lois indépendantes. La regression logistique est également un modèle linéaire correspondant à un ajustement selon un critère autre que les classiques moindres carrés. Enfin, l'information contenue dans certaines cartes rappelle étrangement celle présentée à l'issue d'analyse de variance de modèles à un facteur.

Les techniques de sondages sont également très fréquemment présentes. Elles ont de nombreux points communs avec la planification des expériences; en particulier, lorsqu'on dispose d'observations réalisées à plusieurs niveaux, il existe une analogie évidente avec les plans hiérarchisés dont l'analyse comporte certaines difficultés, généralement peu connues, et sur lesquelles nous insisterons plus loin.

La statistique mérite donc bien l'appellation de "transverse" qui lui est souvent donnée.

II Mise en forme d'une information impliquant une reformulation.

Exemple a. Echantillonnage pour la pêche artisanale sénégalaise. (suite).

L'étude de la variabilité aux divers niveaux d'observation dans les diverses strates met en évidence des hétérogénéités intrastrates, ainsi que l'impossibilité d'en affiner le découpage due à une augmentation parallèle de la variabilité de leurs effectifs qui deviendraient alors incalculables. Ce phénomène provient de l'existence d'une variabilité de l'impact de l'activité de chaque unité de pêche, laissant alors entrevoir une caractéristique qu'il conviendrait de mieux prendre en compte pour l'étude plus générale de la pêche artisanale.

Exemple b Plan d'expérience en génétique quantitative. (suite).

Mais en contrepartie, le contrôle de l'expérience, en station de recherche, peut conduire à analyser une variabilité de nature bien différente de celle qu'on peut rencontrer dans les exploitations agricoles. Il se peut que l'existence d'une forte variabilité de l'environnement se traduise par l'expression d'interactions de type "génotype-milieu", dont on connaît l'existence, mais qui peuvent être négligées si la technologie utilisée par les exploitants entraine une homogénéisation du milieu, et donc permet de négliger les interactions auxquelles il participe. Il se peut alors qu'il faille également prêter attention à la variabilité de l'environnement et étudier les capacités d'adaptation des divers génotypes étudiés à cette variabilité, c'est-à-dire reconsidérer dans une certaine mesure les critères d'optimisations dont la définition doit être en relation avec le type d'agriculture concernée.

Exemple c Analyse des puissances de pêche. (suite).

L'utilisation d'un programme spécifique conduit à "oublier" la méthode statistique utilisée, et donc à ne pas retirer toute l'information qu'il conviendrait, ni toute la critique de cette information. Il n'y a pas ainsi d'analyse des résidus, ni de discussion sur la variabilité spatio-temporelle. L'analyse des résidus peut indiquer l'importance des "coups nuls", lorsqu'une distribution bimodale apparaît et conduire par exemple à rechercher une puissance de pêche exprimée en deux composantes; une capacité à trouver du poisson et une capacité à réaliser de belles captures lorsqu'on en a trouvé. Ceci peut aider à une meilleure description des tactiques de pêche. Le modèle peut aussi être utilisé pour l'analyse des variations spatio-temporelles. Peut-on les décrire par simple addition d'un effet spatial et d'un effet temporel, ou existe-t-il effectivement une interaction spatio-temporelle? Mais en définitive il convient surtout de regarder quelle est la qualité de l'ajustement obtenu. Il arrive que seulement 5% de la variabilité totale soit expliquée par de tels modèles; dans de telles conditions la principale conclusion qu'on en peut tirer est peut-être bien qu'il est inutile de tenir compte de ces puissances de pêche, soit parce que les unités ont des impacts sur la ressource ayant tous la même loi de distribution, soit parce que la variabilité entre diverses opérations de pêche n'est pas ou mal prise en compte par un tel modèle. Ce pourra être le cas si chaque unité de pêche (bateau) peut faire varier sa puissance de pêche, par exemple en changeant de tactique.

Exemple d Echo-intégration. (fin)

L'observation des données indique dans certains cas que des événement "ponctuels" (rencontre avec une ou deux très grosses concentrations) peuvent contribuer de façon très importante à l'estimation d'une biomasse totale. L'idée de bons sens est alors d'analyser séparément les données par type de dispersion. Les recherches se poursuivent en ce sens, mais il se peut que les contraintes (ou les possibilités) des instruments de mesure ne permettent que partiellement d'atteindre les objectifs fixés. Il est possible de présenter les résultats sur une colonne d'eau en autant de tranches de profondeur qu'on le désire, mais au-delà d'un certain niveau de discrimination, qu'il est très aisé techniquement de dépasser, l'apport d'information devient insignifiant, même si la discrimination entre types de dispersion n'est toujours pas totalement satisfaisante. Il semble, à l'heure actuelle que c'est au niveau de l'acquisition de l'information que les progrès les plus significatifs puissent être attendus. La statistique, si elle peut s'avérer utile, pourrait l'être dans l'analyse des signaux reçus, ou alors beaucoup plus en aval, dans la comparaison des résultats avec des données d'origines différentes.

Exemple e Cartographie (suite).

Lorsque les données qu'on désire résumer sous forme de carte sont issues d'un plan de sondage à plusieurs niveaux, et que les entités surfaciques correspondent à un niveau intermédiaire, le problème est d'estimer la part de variabilité à ce niveau intermédiaire. Ceci correspond au problème traité dans l'analyse de variance d'un modèle hiérarchique à effets aléatoires. Si on désire par exemple cartographier une quantité par départements, avec des observations faites dans des échantillons sélectionnés dans un nombre limité de cantons de chaque département, le "danger" qu'il convient d'éviter est de présenter des différences interdépartementales qui peuvent n'être dues qu'à une variabilité intercantonale et intra départementale. Pour l'interprétation de telles cartes il convient donc de revenir aux caractéristiques du plan de sondage pour une analyse complète de la variabilité observée. Le problème est alors très semblable à ceux traités dans l'analyse des plans d'expérience.

Exemple e regression logistique. (fin)

Cet exemple a pu apparaître comme complètement présenté dans le précédent paragraphe. On peut poursuivre en évoquant le problème général traité dans ce cas particulier; l'estimation par maximum de vraisemblance en présence de variables dont la distribution n'est pas normale, ou lorsqu'on désire s'affranchir d'une telle hypothèse. Le terme "logistique" est peut-être tout aussi impropre que celui de "regression". De fait ce terme provient de ce que, dans l'expression du logarithme de la vraisemblance d'une observation, la fonction de lien, déduite de la relation entre un paramètre d'une loi appartenant à une famille de distributions particulière, et l'espérance de cette loi, a une forme logistique. En ce sens, la regression logistique constitue un cas particulier du "modèle linéaire généralisé" qui permet de traiter les distributions normales (moindres carrés classiques) ou les distributions de Poisson, ou les distributions Gamma. Le problème de la forme des distributions est très évoqué au sein de l'institut. Il l'est tout particulièrement chez les hydrologues (recherche de la distribution de valeurs extrêmes, analyse des écarts à des modèles de précipitation...).

On peut donc confirmer la transversalité de l'outil statistique. Une telle qualité conduit à poser la question de l'emploi du terme "outil". Il s'agit d'un terme nécessaire mais peut-être non suffisant. Il est nécessaire car l'utilisation de la statistique implique une compétence technique comme c'est le cas pour l'utilisation d'un marteau. Il sera suffisant si tout le monde sait à quoi peut servir l'outil en question. Ce peut-être le cas pour un marteau, mais çà ne l'est pas pour la statistique. Un très grand nombre de questions posées dans des domaines très divers peuvent être abordées à partir des mêmes approches. La statistique en est une. Si elle est un outil permettant d'affiner des réponses, elle dépasse cette qualité en permettant de préciser les questions.

III Reformulation des hypothèses initiales concernant la chose étudiée et immixtion dans les domaines disciplinaires.

A force d'être précisées, des questions peuvent être modifiées. En évaluant l'impact d'une source de variation, on peut conclure qu'il n'y en a pas. Inversement, en étudiant la variabilité, on peut identifier des sources de variations qui n'étaient pas prévues au programme. Et c'est ainsi que le statisticien participe à la définition de problématiques, intervenant alors dans les disciplines. Il fait alors acte d'indiscipline impliquée.

Exemples a et c Echantillonnage pour la pêche artisanale sénégalaise et analyse des puissances de pêche. (fins)

L'analyse de la variabilité conduit dans les deux cas à montrer, ou au moins suspecter, que certaines unités de pêche peuvent modifier leur tactique, et donc la répartition de l'impact de leur activité de pêche sur les diverses composantes d'une ressource exploitable. Une telle caractéristique peut se traduire par une considérable difficulté de description. En effet, les approches classiques reposent sur l'existence de situations d'équilibre, et une condition nécessaire à cette existence est une homogénéité tactique de chaque unité de pêche. C'est pourquoi un modèle analysant un effet bateau est "logique", mais c'est aussi pourquoi une description tenant compte de variabilité tactique "intra unité de pêche" semble a priori exclue. Une telle variabilité, si elle est effectivement importante, conduit alors à un aspect informel de la pêche. C'est-à-dire à un très faible pouvoir explicatif à partir d'un modèle conçu pour décrire des unités supposées "homogènes", ou bien a une observation de strates d'effectifs très variable du fait de la possibilité qu'ont les unités de pêche de changer de strate. Tout ceci peut conduire à l'adoption d'un cadre descriptif différent, tenant compte de cette caractéristique, et ne reposant plus sur une quelconque hypothèse d'équilibre.

Exemple b (fin)

Les résultats obtenus grâce à la génétique quantitative sont notoires. Mais il faut les replacer dans le contexte de l'évolution parallèle de la technologie agricole, qui a conduit à pouvoir négliger les interactions de type génotype-milieu du fait d'un contrôle de plus en plus efficace du milieu. Deux coups d'oeil suffisent à sentir qu'un modèle supposant un faible impact du milieu aura des chances de mieux décrire une culture de maïs en France qu'une culture de mil au Sénégal. Les conditions dans lesquelles se réalise l'agriculture ne peuvent être ignorées. Mais d'une certaine manière une telle ignorance aura des conséquences moins graves en France, l'un des pays où la génétique quantitative a été développée avec succès dans son contexte agricole donné, que dans la zone tropicale où la biologie des espèce est peut-être la même, mais où les conditions agricoles sont certainement différentes.

Exemple e Cartographie. (fin) La statistique n'est pas absente de ce domaine, mais les progrès récents et en cours, tant au niveau de l'édition que de celui de l 'accès aux données (SIG) permettent des synthèses dont la qualité de l'expression cartographique,

accessible à tout utilisateur grâce à des produits logiciels conviviaux, pourrait devenir "excessive" lorsqu'une quasi perfection dépend plus de la précision "électronique" du crayon que de celle des données obtenues sur un terrain. Cette amélioration et diffusion de l'outil pourrait justifier une attention particulière en ce qui concerne la traduction de la qualité effective de l'information.

Discussion.

L'implication discutée jusqu'ici à partir d'une approche d'une discipline "qui n'en est pas tout à fait une" ne peut guère être considérée comme une sorte d'association de type pluridisciplinaire classique. En fait l'approche statistique présentée ici ne peut se faire indépendamment d'une ou de plusieurs autres disciplines; parce qu'elle se déroule dans des programmes "d'UR", où on ne peut pas ne pas en rencontrer. Ces programmes ont en général été conçus autour de problématiques disciplinaires. Ces problématiques consistent à "construire des objets", donc à rechercher des choses présentant une variabilité selon les sources préférées des disciplines. Cette variabilité n'est au demeurant souvent considérée que par référence aux résultats d'autres études menées selon la même problématique, le problème étant alors de donner un nouveau point observé qu'on pourra éventuellement replacer dans le contexte général.

Vue sous un angle statistique, une problématique disciplinaire est un ensemble d'hypothèses sur la nature de la distribution de l'information qu'on collecte à partir des choses observées. La problématique engendre un filtre d'observation orientant cette observation en fonction des sources de variation privilégiées. Ceci peut conduire à négliger, ou même nier d'autres sources, et c'est ainsi qu'une approche monodisciplinaire consiste à considérer que les résultats des autres disciplines sont des données, ou alors que l'impact de variabilités "extra-problématiques" sera nul ou au pire indépendant de celui des sources étudiées. Lorsque, à l'évidence, de telles hypothèses ne peuvent raisonnablement être faites, on peut être amené à "inventer" des phénomènes pouvant figer le contexte occulté par le filtre d'observation. C'est ainsi que l'halieute biologiste peut quelque peu idéaliser l'aménageur dont le pouvoir serait de rendre constant l'impact de la pêche sur la ressource, et que l'halieute "socio-économiste" suppose que la ressource est conforme à la volonté du pêcheur: s'il décrit la dynamique interne d'une société qui se tourne vers la pêche, il est évident que la ressource ne peut être une contrainte et qu'elle est encore incomplètement exploitée. Pour les uns le

poisson ne saurait expliquer le pêcheur, pour les autres, c'est le pêcheur qui ne saurait expliquer le poisson.

Dans le cadre d'une invitation au service, il est proposé au statisticien de rechercher à partir des données disponibles la façon de les résumer de la façon la plus courte, tout en gardant un maximum de l'information contenue dans ces données. Mais ceci doit se faire bien sûr sans remettre en cause la définition initiale de la problématique, c'est-à-dire sans remettre en cause la forme de la chose étudiée. On demande de montrer qu'un carré peut être décrit par la longueur d'un de ses cotés, et d'estimer cette longueur, mais on ne demande pas de trouver qu'on n'a pas affaire à un carré. Ceci correspond à la recherche de résumé suffisant dans le cas ou la nature des distributions sous-jacentes est connue (par exemple la moyenne et la variance calculées selon les formules classiques constituent un résumé exhaustif de l'information contenue dans un échantillon issu de tirages dans des lois normales indépendantes de mêmes espérance et variance; dans un contexte beaucoup plus général, un résumé est suffisant si les différences entre ensembles d'information se résumant de façon semblable n'apportent plus aucune précision supplémentaire). La collaboration peut se dérouler dans ces limites, et c'est alors une très bonne chose. Mais si on aboutit à la conclusion selon laquelle la chose étudiée n'est vraiment pas un carré, l'intervention prend un tout autre aspect et les choses se compliquent en ce sens que la discipline déteste admettre ce résultat avant de l'avoir intégré dans sa problématique (lorsque qu'une distribution qu'on croyait normale ne l'est pas, on ne sait plus ce que pourra être un résumé suffisant sans préciser cette distribution, donc redéfinir la forme de l'objet étudié).

L'inadéquation d'une forme pour décrire un objet, peut être décelée, selon une approche statistique, lorsqu'en décidant malgré tout de présenter un carré, on se sent obligé de signaler, souvent par des "anecdotes" que certaines choses ne se sont malgré tout pas déroulées de la même façon que dans tel autre cas à l'issue duquel on avait décrit pourtant exactement le même carré. C'est alors qu'un paradoxe apparaît, concernant le jugement global porté sur l'information et sur la pertinence de l'outil statistique. En effet, si on conclut qu'un carré ne suffit pas à présenter une information, c'est qu'on dispose d'information permettant de le montrer; mais par ailleurs, comme on ne sait pas restituer cette information, on garde l'impression selon laquelle on n'en a justement pas.

De même peut-on lire ou entendre fréquemment qu'une inadéquation avec les "règles" de la statistique interdit l'exploitation de ce dont on dispose, qui n'est donc qu'un bruit. Mais ces "règles" statistiques n'en sont pas; elles ne sont que le produit du filtre que définit la problématique, filtre qui établit la façon de faire une typologie, en définissant qui est l'individu et ce qu'est la relation d'équivalence fondée sur l'homogénéité des individus d'une même classe pour ce qui concerne la source de variation considérée. C'est ainsi qu'on peut vouloir réunir en une même classe toutes les unités de pêche qui ont à une constante près le même impact sur les différentes composantes d'une ressource; c'est-à-dire celles qui adoptent une même tactique. Mais si les unités changent de tactique, c'est qu'elle peuvent changer de classe. Si ces changements se font indépendamment de la perception de l'état (physique et économique) de la ressource, on pourra espérer que l'impact pourrait rester globalement stable, mais si tel n'est pas le cas, alors le poisson explique le pêcheur, et la pêche devient informelle, non parce que les règles statistiques sont non satisfaites, mais parce que les hypothèses sur la distribution des quantités observées sont erronées.

On accuse donc d'impertinence l'une des approches dont on aurait en définitive le plus besoin dans un tel contexte.

Il est bien évident que la nécessité de reformuler des hypothèses peut être admise sans collaboration avec la statistique. C'est même le cas le plus fréquent, ne serait-ce que parce qu'il n'y a que très peu de statisticiens. Mais la recherche de nouvelles formes semble s'imposer de plus en plus fréquemment et ceci est certainement en relation avec le changement de nature de l'information évoqué en introduction. Une information nouvelle en effet n'entre pas toujours dans les cadres habituellement considérés. La statistique, en tant que "science de la qualification de l'information à partir de la variabilité" est alors d'un intérêt évident parce qu'il s'agit d'une science sans objet ni problématique.

La situation dans laquelle on se trouve lorsqu'on ressent informel un objet est "pré-pluridisciplinaire" puisqu'il s'agit d'une situation dans laquelle aucune discipline ne reconnait sa chose descriptible. Une attitude peut consister en des analyses séparées de la chose étudiée à partir de laquelle plusieurs objets "monodisciplinaires" sont reconstruits. Ainsi

lorsque la discipline qui sait décrire des carrés s'intéresse à un cercle, elle peut honnêtement conclure qu'elle a affaire à un carré informel pouvant devenir un objet d'étude. D'autres peuvent détecter qu'un carré ne saurait suffire parce qu'on voit bien qu'il faut plus de quatre côtés pour décrire la chose en question. C'est ainsi qu'on peut entamer la recherche du nombre de côtés du cercle. Il s'agit en fait de situations dans lesquelles une approche statistique s'impose d'une part mais où d'autre part les disciplines ne sont souvent pas à même de le déceler.

Les disciplines n'acceptent guère la pluridisciplinarité que de façon temporaire, c'est-à-dire juste le temps nécessaire pour approprier la source de variation qui s'est nouvellement imposée. Et c'est ainsi, pour hâter cette appropriation, que lorsqu'on admet que s'avérerait utile un individu ayant une formation lui permettant de traiter les données d'une manière supposée satisfaisante, c'est-à-dire en gros un "informaticien-statisticien-mathématicien-modélisateur", il est également bien évident qu'il s'agit aussi et avant tout d'un hydrologue ou d'un géographe ou d'un économiste (pour le secteur informel) ou encore d'un halieute ou écologiste (la liste n'est bien entendue pas exhaustive) qui doit être recruté au sein d'une des six premières commissions (notons tout de même le recrutement au sein de la CS 7 d'une biostatisticienne pour l'halieutique). Et c'est également ainsi qu'une constante quasi générale du discours pluridisciplinaire est l'oubli, ou le refus, d'évoquer l'intérêt d'une approche statistique, sauf à dire qu'il existe des problèmes d'origine statistique. Si ces problèmes sont au moins en partie traités, ce qui paraît naturel, selon une approche statistique, et que des solutions "satisfaisantes" sont apportées, c'est-à-dire que les disciplines reconnaissent de nouvelles formes, il est logique que l'approche statistique, restant sans objet, soit absente de la liste de collaboration, même si on reconnaît qu'un individu d'appartenance peu claire a pu jouer un rôle. Ceci est d'autant plus logique qu'en participant à la construction d'un objet, on adhère évidemment à la discipline qui en prend possession.

Pour terminer cette présentation on peut aborder une source de variation dont tout le monde parle, et que tout le monde désire intégrer à son approche: le temps.

La prise en compte du temps s'impose de fait par l'observation de l'imbrication de plusieurs sources. Ainsi le poisson explique le pêcheur et

le pêcheur explique le poisson. Ceci implique que le pêcheur peut être inconstant et on ne peut plus admettre acquis et fixé son comportement en supposant que d'autres disciplines l'on démontré ou sauraient le faire. Pour résumer le pêcheur, il faut l'indicer par le temps et donc ne plus se référer à des situations d'équilibre, même à terme.

La référence aux systèmes dynamiques a ainsi fait son apparition dans une multitude de problématiques disciplinaires, laissant ainsi entrevoir que le processus d'appropriation de sources nouvelles de variations par diverses disciplines a, au moins une fois, connu une étape d'achèvement. C'est là un "événement" très important. On pourrait penser qu'un système dynamique est une association de sources de variations jusqu'alors étudiées séparément, et que son étude devrait naturellement se dérouler dans un cadre pluridisciplinaire. En fait on observe plutôt une appropriation du concept de système par chacune des disciplines, qui entend bien continuer parler de ses sources privilégiées, en acceptant seulement de se placer dans le cadre valorisant de la "complexité". Et le cloisonnement maintenu entre les disciplines conduit à appeler nomades des concepts communs. La référence de plus en plus générale aux "systèmes complexes" pose un problème technique et théorique.

D'un point de vue théorique la notion de complexité repose sur des résultats et exemples spectaculaires où il a été montré que des processus "chaotiques" peuvent fort bien être engendrés par des systèmes simples d'équations, systèmes extrêmement sensibles, au point que d'infimes perturbations les font changer de façon imprévisible. L'existence de tels systèmes montre qu'il existe des cas où un acharnement thérapeutique conduisant à compliquer de plus en plus des modèles, et à mesurer toujours plus de données, ne saurait permettre une amélioration permettant de prédire (au sens du physicien) les états futurs du système étudié. Il existe même des recherches d'outils pour la détermination, à partir de processus observés, de systèmes les plus simples possibles permettant d'en rendre compte. L'intérêt de ces résultats est évident, mais il convient d'insister sur le fait que s'ils permettent en tant que contre exemples la remise en cause d'un certain scientisme, ils n'autorisent pas, en tant qu'exemples, à en recréer un autre qui justifierait la recherche en toutes circonstances de ces attracteurs étranges dont la ressemblance avec le processus étudié ne peut "valider" le système sous-jacent. Au demeurant dans la plupart des cas, la "complexité" renvoie à un sens plus large et

l'appellation "système dynamique" est largement préférable. Il convient également d'insister sur le fait que la notion de modèle prête à confusion. Il y a d'une part le modèle reposant sur des lois, supposé décrire exactement l'état et la dynamique d'un système et donc supposé permettre une prédiction exacte du futur. C'est l'existence de tels modèles qui est contestable et contestée par la complexité. Il y a d'autre part le modèle "statistique", beaucoup moins ambitieux, qui consiste à projeter l'information contenue dans un espace sur un sous-espace et à séparer la variabilité en deux parties, celle appartenant à ce sous-espace, et celle (résiduelle) appartenant à un autre sous-espace orthogonal au premier. La confusion provient de ce que le sous-espace "modèle" ressemble à un modèle de type "physique" et que trop souvent la présentation du modèle statistique est réduite au compte-rendu de la seule projection sur ce sous-espace. La notion de complexité ne remet pas en cause le modèle statistique, non parce qu'il serait meilleur, mais parce qu'il n'a pas la prétention du modèle "physique". En particulier, lorsqu'il est utilisé pour prédire, ce n'est pas pour rechercher des valeurs exactes, mais des lois de distribution de ces valeurs, la qualité de la prédiction étant largement déterminée à partir de la description de la variabilité résiduelle. C'est ce manque d'ambition qui permettrait peut-être de ne pas pousser à l'infini le comptage du nombre de cotés d'un cercle.

D'un point de vue technique, l'étude des processus et des séries chronologiques est un des domaines les plus difficiles. On observe un décalage croissant entre le contenu théorique et l'utilisation des méthodes mises au point et que tout un chacun peut mettre en oeuvre à l'aide de logiciels adaptés. Quelles que soient leurs qualités ces méthodes sont souvent sous-exploitées, le problème étant alors que les résultats qui en sont tirés sont bien présentés comme des produits "maximaux". Ainsi les modèles ARMA constituent une classe générale de modèles stationnaires souvent utilisée. Mais les résultats présentés sont presque toujours des modèles AR (autorégressifs). La raison en est simple; d'une part les modèles AR sont beaucoup plus clairement interprétables et d'autre part un modèle général ARMA peut (presque) toujours être suffisamment ressemblant à un modèle AR pour que ce dernier lui soit préféré. Encore ne s'agit-il là que de modèles stationnaires, et tout le monde connaît la difficulté d'identifier des périodes, et quels "artefacts" peuvent provenir d'une mauvaise identification. L'amélioration des compétences concernant l'étude des processus pourrait être une priorité au niveau de l'institut.

Conclusion.

Tout ce qui précède ne donne peut-être pas une idée claire du statut de la statistique à l'ORSTOM, ni de ce qu'il devrait être.

Dans la situation actuelle, les statisticiens sont recrutés par les six premières commissions scientifiques (le recrutement par BAP des IT n'empêche pas que ces commissions se sentent responsables des profils et lieux naturels d'adhésion). Cette situation n'est certes pas absurde, d'autant plus que la CS 7 est parfois conviée dans les jurys, mais elle se traduit par des aspects négatifs indéniables.

- L'absence de communication entre personnes ayant des démarches statistiques et utilisant les mêmes outils.

- Une mauvaise identification de ce qu'est la statistique, évidente dans l'énoncé de nombreux profils de recrutements.

- Une difficulté de diffusion des outils et de mise en place de formation pourtant très souvent demandée.

- Une difficulté d'identification de thèmes de recherche pouvant conduire à des collaborations avec des laboratoires spécialisés, en particulier universitaires. Ceci constitue l'une des raisons de la difficulté de recruter des statisticiens à l'ORSTOM.

On pourrait conclure qu'il faudrait que les statisticiens soient systématiquement recrutés au sein de la CS 7 et affectés au sein d'une mission technique ou d'un laboratoire propre. Mais une attitude aussi tranchée se traduirait par au moins un autre effet tout aussi négatif: la disparition de cette "indisciplinarité" et en conséquence celle de cette implication dans les problématiques des autres disciplines. La statistique ne pourrait alors plus être qu'appliquée, ou isolée.

En définitive il convient de trouver un moyen terme. Il faut qu'un statisticien puisse être dans une UR, de façon stable mais non nécessairement définitive ni à plein temps. Il convient aussi que son domaine soit clairement reconnu et que les recrutements même s'ils ne sont pas tous faits dans le cadre de la CS 7, soient réalisés en collaboration

avec cette dernière qui doit être associée à la définition des profils. Si de telles collaborations peuvent être mises en place, le choix de la CS d'adhésion ne devrait plus poser de problèmes.

Enfin, il semble nécessaire que des lieux de rencontre et de formation existent. Il existe déjà un thème statistique au LIA à Bondy, et une cellule de biométrie et statistique est en court de création au centre de Montpellier.

Cette cellule serait une structure légère, conçue comme un lieu d'accueil, de mise en commun, de réflexion méthodologique, de formation, d'animation scientifique et d'ouverture vers l'extérieur. Elle est d'ores et déjà identifiée (avec le CNRS et le CIRAD) comme partenaire privilégié au sein de l'Unité de Biométrie de Montpellier qui associe l'ENSA.M, l'INRA et l'USTL.

Les moyens de cette cellule devraient se limiter à un personnel permanent de deux à trois personnes (appartenant par ailleurs à des "programmes d'UR". Elle devrait disposer d'un local et d'un budget lui permettant d'acquérir des produits logiciels, de couvrir d'éventuels frais de calcul, de gérer une bibliothèque spécialisée et des missions d'animation scientifique.