STATISTIQUE EN SCIENCES SOCIALES
Une démarche plus qu'un outil

Marie PIRON
Statisticienne, Département SUD, recrutée en 1994 pour l'ensemble des UR

ITINERAIRE

Au départ, ma double formation en mathématiques et en géographie m'a permis de travailler plus particulièrement sur le traitement de données spatialisées. Par la suite, j'ai été davantage confrontée au traitement statistique de données d'enquêtes, en portant une attention particulière à la validité du choix de la méthode. Tout ceci, ponctué d'expertises et de formations principalement au sein de l'Association pour le Développement et la Diffusion de l'Analyse des Données.

Double cursus en mathématiques et en géographie
Dans le cadre d'un mémoire en géographie quantitative, j'ai étudié la régionalisation de l'espace agricole français au travers des systèmes et structures des exploitations agricoles en utilisant les méthodes d'analyse de données, outils privilégiés pour traiter l'espace géographique dans toute sa complexité et sa multidimensionnalité. Pour aller plus loin dans la compréhension de ces méthodes, j'ai passé un DEA de statistiques théoriques et appliquées.

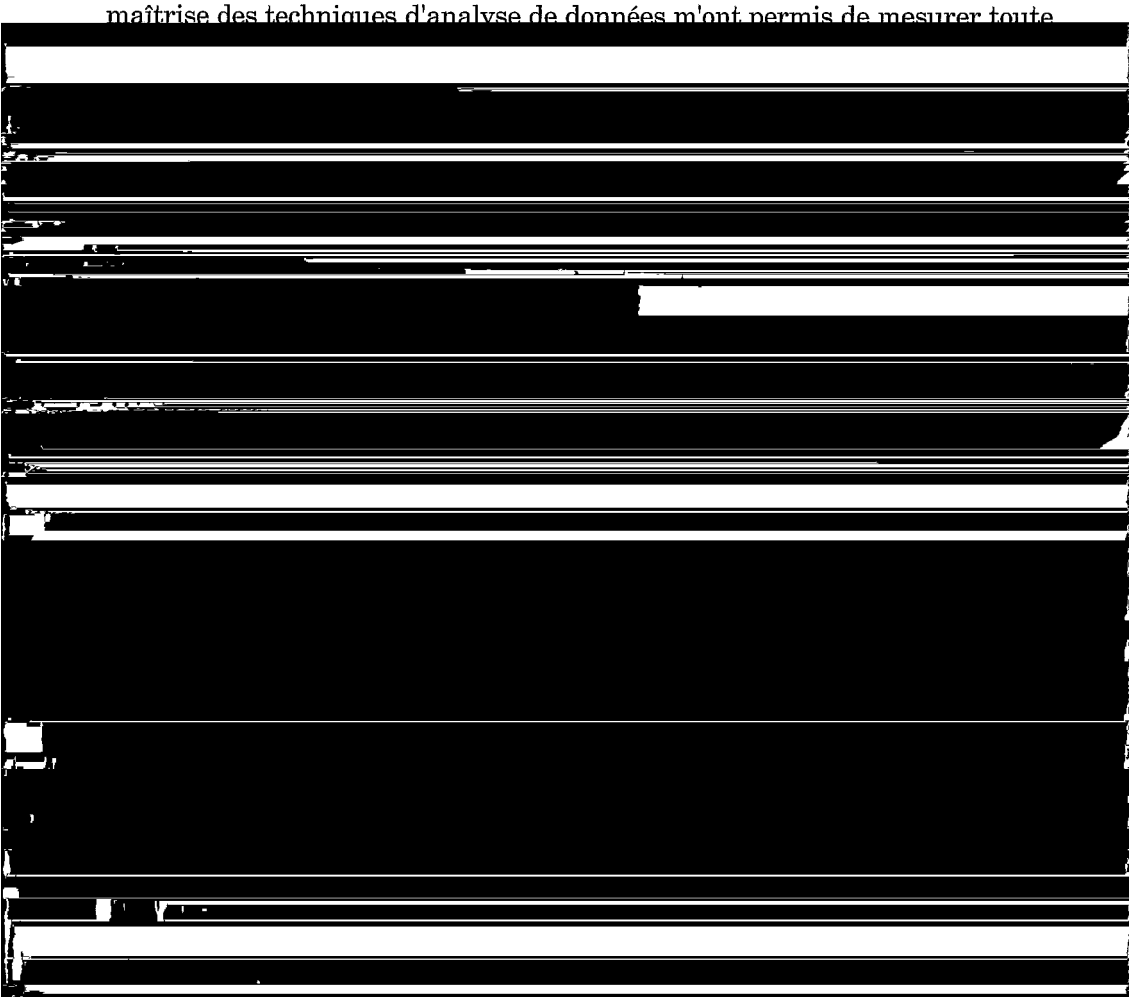
Enquêtes à plusieurs niveaux d'observation et analyse spatiale
Puis l'intégration dans un programme de recherche de l'ORSTOM pour étudier les enjeux des extensions urbaines à Ouagadougou m'a conduite durant deux années au Burkina Faso. J'y ai mené une enquête à passages répétés et à plusieurs niveaux emboîtés d'observation : la parcelle (unité d'échantillonnage), les ménages résidant sur la parcelle, tous les individus composant ces ménages. Ce type d'enquête, relevant d'une structure hiérarchique de l'information, est de plus en plus courant. Son exploitation pose le problème de la mise en relation d'informations définies sur différents niveaux d'observation.
Dans le cadre d'une thèse en analyse des données sous la direction de J.-

exploratoire, qui trouve un champ d'application privilégié dans le traitement de données intégrées d'un système d'information géographique. Maîtrise des processus d'agrégation et/ou de ventilation spatiales, mise en évidence des structures spécifiques aux différents niveaux d'information et des relations entre ces niveaux, réflexions sur la pertinence des découpages géographiques et sur l'effet dû à l'agrégation spatiale, constituent les principales applications de cette méthode qui permet d'aller plus avant dans l'exploitation combinée d'informations correspondant à des découpages spatiaux différents.

Systèmes d'enquêtes socio-économiques et observatoires

Par ailleurs, ce travail de thèse m'a initiée à la méthodologie des enquêtes et des systèmes d'information. Au Centre d'Etudes de l'Emploi, qui m'a accueillie durant deux années, j'ai prolongé la réflexion sur la mise en relation d'informations issues de niveaux emboîtés dans le cadre d'un programme sur "l'organisation et les conditions de travail des salariés", en confrontant le point de vue de l'entreprise et de celui du salarié. Je me suis également intéressée aux analyses de parcours dans le cadre d'une étude sur "l'insertion des handicapés dans le monde du travail" et au traitement de données issues d'observatoires notamment sur "l'évolution de 1990 à 1993 des salariés en Contrats-Emploi-Solidarité".

La pratique du terrain, la conception de systèmes complexes d'enquêtes et la maîtrise des techniques d'analyse de données m'ont permis de mesurer toute



mis en lumière. Une certaine approche de la statistique implique aussi l'évolution d'un concept, la découverte de phénomènes latents, la proposition de nouvelles hypothèses de travail et contribue, en cela, à une approche qualitative des faits sociaux.
Mon expérience m'a fait prendre conscience d'un certain nombre d'écueils et de pièges qui peuvent conduire à une perception de la statistique comme un ensemble d'outils et de techniques appliquées sans une réflexion méthodologique globale.

Une certaine confusion entre statistique et quantitatif
Les recherches en sciences sociales analysent des faits qui par nature sont complexes et multidimensionnels. L'approche qualitative est par conséquent fondamentale. Aujourd'hui ces recherches s'accompagnent de plus en plus de chaînes du traitement de l'information bien identifiées, au travers des différents modes d'accès à l'information tels que les enquêtes, les recensements, les tableaux de statistiques.
L'exploitation de ces bases de données nécessite le recours au traitement statistique alors souvent utilisé pour quantifier des phénomènes. Or, on a sans doute trop tendance à assimiler traitement statistique et approche quantitative, et cette image de la statistique me semble trop étroite. Est-il vraiment nécessaire de vouloir tout chiffrer, mesurer, borner, modéliser, estimer... ?
Par abus, l'outil statistique est parfois considéré comme un recours, un refuge, une boîte noire produisant l'événement. On cherchera à se conforter derrière un indicateur ou un modèle statistique qui "généralise ou prouve tout", et pourtant qui ne décrit rien ou n'explique rien s'il est mal adapté par méconnaissance de l'outil utilisé ou du domaine auquel il est appliqué.
Parallèlement, pour certains, quantifier un phénomène, cela peut être réduire et figer les faits sans tenir compte de toute une dimension intuitive et complexe de la réalité. C'est aussi perdre le caractère individuel de l'objet d'observation qui en fait toute sa richesse.
Cette confusion entre quantitatif et statistique fait référence ici à ce qu'était la statistique, il y a quelques années, avant l'avènement de l'informatique, lorsqu'elle reposait principalement sur la théorie des probabilités. Depuis, la puissance de calculs des ordinateurs a permis d'abord le développement des méthodes d'analyses de données multidimensionnelles, puis des méthodes de validation par ré-échantillonnage qui se substituent progressivement aux calculs analytiques des estimations (souvent non satisfaisants car ils reposent sur des hypothèses contraignantes).

Deux aspects de la statistique se sont alors nettement distingués et opposés pendant un moment : la statistique descriptive et exploratoire et la statistique inférentielle et confirmatoire[1]. Et c'est sur ces deux aspects qu'il faut justement jouer et composer.

Nécessité d'une approche descriptive multidimensionnelle

Accéder rapidement à un esprit de synthèse est l'un des intérêts de la statistique lorsque l'on est face à de vastes recueils de données.

Traditionnellement, le traitement statistique en sciences sociales repose sur une description élémentaire des données (moyenne, médiane, dispersion, pourcentage, corrélation). Des méthodes plus élaborées viennent parfois compléter ces premiers résultats pour approfondir ou expliquer un phénomène en fonction d'une ou de plusieurs autres variables (modèle log-linéaire, régression multiple ou logistique, analyse de la variance, etc.) suivant des hypothèses formulées a priori. Cette démarche est adéquate lorsque le recueil de données ne comporte qu'une dizaine de variables déjà structurées et que l'on cherche à quantifier un phénomène connu ou à valider une hypothèse à partir d'un modèle existant. Dans ce cas le plan de sondage est conçu comme un plan d'expérience et l'on se rapproche davantage d'une problématique telle que celle rencontrée en agronomie ou pour des essais cliniques par exemple.

Mais lorsqu'il s'agit de vastes bases d'informations relevant d'un domaine complexe ou inconnu, comme c'est souvent le cas en sciences sociales, les indicateurs descriptifs usuels ainsi que les tableaux statistiques ne suffisent pas toujours pour comprendre les mécanismes d'un phénomène. D'ailleurs la transcription littérale de ces indicateurs et tableaux (parfois rébarbative aussi bien dans la rédaction que dans la lecture) traduit souvent une errance sur les détails au détriment d'une approche synthétique. Il est alors nécessaire d'avoir recours aux méthodes descriptives multidimensionnelles.

Parce qu'elles font intervenir tout le réseau d'interrelations entre les variables, ces méthodes permettent d'extraire les structures de l'information contenue dans des ensembles volumineux de données. En réduisant l'information (qui en fait élimine les fluctuations et l'aléatoire contenus dans les données), elles donnent ainsi lieu à une présentation claire et rapide des résul-

---

1 - *la statistique descriptive et exploratoire* : les conclusions ne portent exclusivement que sur les données étudiées. Elle permet par des résumés et des graphiques plus ou moins sophistiqués de décrire la population d'étude, d'établir des relations entre les variables sans faire jouer de rôle privilégié à une variable particulière. Elle s'appuie essentiellement sur des notions élémentaires tels que des indicateurs de moyenne et de dispersion et sur des représentations graphiques. La statistique exploratoire fait surtout référence aux techniques descriptives multidimensionnelles (analyse en composantes principales, analyse des correspondances, classification) dont le principal objectif est d'extraire des structures à partir des données.

- *la statistique inférentielle et confirmatoire* : les conclusions obtenues à partir des données étudiées vont au delà de ces données. Elle permet d'extrapoler, c'est-à-dire d'étendre les caractéristiques d'un échantillon à la population, et de valider ou d'infirmer, à partir de tests ou de modèles, des hypothèses formulées a priori ou après une phase exploratoire. Elle s'appuie sur des tests statistiques, des modèles probabilistes. La statistique confirmatoire fait surtout référence aux méthodes explicatives et prévisionnelles destinées à expliquer puis à prévoir, suivant des règles de décision, une variable privilégiée à l'aide d'une ou de plusieurs variables explicatives (régressions multiples et logistiques, analyse de la variance, analyse discriminante, segmentation, etc.).

tats qui se prête plus aisément à l'interprétation. Par ailleurs, alors que la statistique descriptive usuelle s'intéresse aux caractéristiques globales de groupes d'individus, ces méthodes, parce qu'exploratoires, permettent une approche par niveau allant de l'émergence de formes globales définies par des groupes d'individus à la position d'un individu dans ces groupes. Il est ainsi possible d'organiser, de structurer et de synthétiser la base d'information pour comprendre ou découvrir des formes sociales, pour déceler certaines relations déterminantes a priori non soupçonnées, pour suggérer de nouvelles hypothèses.

Or, ces méthodes restent encore trop souvent absentes ou interviennent comme un complément sophistiqué à la suite du traitement. En fait, d'une part elles contiennent, en elles-mêmes, leur propre procédure de validation et, d'autre part elles permettent d'éviter certains piétinements et d'orienter la suite du traitement vers un choix pertinent des tableaux croisés, des corrélations ou vers une approche statistique confirmatoire si tel était l'objectif initial ou si tel le suggèrent les premiers résultats descriptifs. Le choix des modèles n'est plus fait de façon aveugle en fonction des hypothèses de base.

L'ensemble de ces opérations implique simultanément un gain de productivité, une amélioration de la qualité des résultats et l'accès à de nouvelles informations. Intervenant au début de la chaîne du traitement statistique, elles définissent une nouvelle méthodologie intégrant cette dimension qualitative encore trop peu reconnue à la statistique.

Les méthodes d'analyses multidimensionnelles constituent par conséquent des outils d'exploration à la mesure des vastes recueils de données. Elles sont particulièrement bien adaptées pour des systèmes d'information compliqués tels que les observatoires, les systèmes d'enquêtes à plusieurs niveaux d'observation, les systèmes d'information géographique. Et l'on connaît l'enthousiasme actuel pour ces systèmes d'information devenus opérationnels avec le développement des moyens de stockage informatiques. L'un des objectifs, sur un plan méthodologique, est que ces systèmes ne restent pas d'excellents conservatoires de données mais puissent être exploités et analysés en tenant compte de toute la dimension relationnelle (dimension sociale, temporelle, géographique, passage du micro au macro) qui lie les objets observés dans ces systèmes.

Nécessité d'une réflexion méthodologique

Il faut cependant reconnaître qu'il y a de quoi se perdre parmi les indicateurs, les tests, les lois de probabilités, les modèles, les techniques d'analyse

terprétation amène à réfléchir sur la donnée et conduit à effectuer d'autres traitements. Aussi, ce n'est pas une confrontation ou une comparaison successive des outils statistiques qui permet d'obtenir des résultats satisfaisants mais la nécessité du contrôle permanent de l'ensemble du processus du traitement de la donnée.

Enfin, une réflexion méthodologique globale dans le traitement de l'information revêt une grande importance.

Il arrive en effet que des enquêtes soient parfois lancées sans avoir au pré-

jeunes) dans un contexte de crise économique et de politiques d'ajustement structurel. L'objet est d'étudier les innovations qu'un tel contexte suscite. Ma participation à ce programme s'effectue sous forme de missions.
Par ailleurs, je vais être affectée à Abidjan dans le cadre du groupement interdisciplinaire en sciences sociales (GIDIS-CI) intégrant entre autres une équipe de l'UR 55 travaillant sur le thème "Crises, ajustements, recompositions". Plusieurs programmes complémentaires, aux approches disciplinaires différentes, sont développés, intéressant aussi bien le monde ur-

mun une base de données qu'il s'agira d'exploiter en fonction des objectifs de chacun mais aussi en fonction de la problématique qui les réunit.
Ces deux groupes de recherche, de grande envergure, s'articulent autour d'une même problématique globale à la fois méthodologique et thématique. Cela représente une complémentarité analogique non négligeable tant dans la réflexion méthodologique que dans son développement opérationnel.

"*Introduction aux méthodes d'analyse des données*" ; (1992) ; Support de cours ; 65 p ; avec C. Mullon

"*Analyse statistique d'un système d'échelles*" ; (1992) ; Réseau A.D.O.C. ; document de travail n°4 ; éd. Orstom ; 211 p.

"*Les enjeux des extensions urbaines à Ouagadougou 1984-1990*" ; (1992) ; Compte-rendu de fin d'étude ; Orstom-CNRST ; 365 p ; avec E. Lebris *et al*,

"*Enquête clientèle*" ; (1991) ; Tome 1 : "Méthodologie", Tome 2 : "Entreprises", Tome 3 : "Particuliers" ; (Demande de l'EDF) ; Rapport d'étude ADDAD ; avec J.P. Fénelon et *al*.

"*Les centres culturels français en Afrique : évaluation de l'action des CCF dans les pays du champs*" ; (1991) ; Secrétariat Permanent des Etudes, des Evaluations et des Statistiques ; n° 5 évaluations ; éd. ministère de la coopération et du développement ; 244 p ; avec J. Bonnemoy *et al*.