

Les problèmes de l'usage de la statistique en sciences sociales

Claude Marois^a

A l'occasion du colloque sur "La statistique impliquée", une table ronde sur la statistique, une indiscipline impliquée, a permis d'étaler un certain nombre de problèmes sur l'utilisation de cette discipline en sciences sociales et en sciences naturelles. Partant du principe que les disciplines scientifiques doivent contribuer mutuellement à l'essor et à la compréhension de problèmes, il faut accepter que la compétence exclusive disciplinaire devient et deviendra de plus en plus difficile à soutenir. Ce qui ne veut pas dire qu'il faut être tolérant vis-à-vis des erreurs ou des écarts de conduite dans les applications de méthodes ou de techniques empruntées d'une autre discipline. L'objet de cet article est de présenter un bilan sommaire des problèmes posés par l'utilisation de la statistique en sciences sociales et particulièrement en géographie humaine.

1 Démarche scientifique et qualité des données en sciences sociales.

Théoriquement, les chercheurs en sciences sociales et en sciences naturelles franchissent les mêmes étapes dans une démarche scientifique permettant de répondre aux objectifs et aux hypothèses de recherche posées initialement. Cette démarche se compose d'une série d'étapes plus ou moins ordonnées dont la démarche quantitative fait partie. Ce processus de recherche "contrôlé" en terme de qualité comporte successivement la définition du problème, la formulation des objectifs et des hypothèses, la définition d'un cadre méthodologique, la collecte des données, le traitement qualitatif et/ou quantitatif des données et l'analyse des résultats. Il y a la recherche de rigueur et de systématisation et les critères de la méthode scientifique sont les mêmes pour les chercheurs en sciences sociales et ceux des sciences naturelles. Cependant, il y a quelques différences qu'il vaut la peine d'énumérer brièvement pour mieux comprendre la recherche en sciences sociales. La première différence fondamentale est l'accent mis sur l'étude du comportement humain et de ses institutions. Fréquemment, ceci implique des interactions entre le chercheur et les sujets de son étude. par exemple, la collecte des données peut s'effectuer via une communication entre individus dont l'un est un observateur et l'autre observé. La qualité de l'information est à la fois dépendante des deux puisque dans le premier cas il collecte des observations, des impressions, des commentaires qu'il

^aDépartement de Géographie. Université de Montréal, C.P. 6128 Succursale "A" Montréal, QcH3C

reçoit du répondant et l'autre livre au chercheur ses opinions, ses perceptions, les informations demandées. Dans d'autres situations, la collecte des données passe par les statistiques dites "secondaires" i.e. dépendantes des données de recensement ou autres bases de données issues d'organismes publics. Dans ce cas il faut s'assurer de la qualité et des caractéristiques de la base de données : cela exige une bonne connaissance des sources. Le chercheur en sciences sociales est régulièrement confronté à des problèmes de définitions de variables, de disponibilité de données, de problèmes de comparaison de variables issues de sources différentes, de comparabilité de données dans le temps et dans l'espace, de changement dans les découpages géographiques à une même échelle etc... Une autre différence touche à la difficulté de mesurer des phénomènes sociaux : ces mesures sont souvent de facture qualitative i.e. des informations sous forme de commentaires, d'opinions et d'évaluations. Aussi, il n'est pas toujours possible d'obtenir des mesures très précises et d'échelles de mesure différentes dont certaines ont des propriétés mathématiques "limitées" (nominales et ordinales). Alors, la nature de l'information est souvent non-paramétrique et n'est pas toujours catégorique ou objet de catégorisation formelle. Dès lors, les techniques statistiques sont souvent plus puissantes que les données et le choix de ces techniques dépend entre autres et dans une large mesure de caractéristiques et des propriétés mathématiques des variables à l'étude. En effet, les méthodes statistiques paramétriques comportent des présomptions qui gênent leurs applications en sciences sociales parce qu'elles sont conçues pour des variables de type intervalle et de rapport. Or les données sont souvent de type ordinal ou nominal alors cela oblige le chercheur à s'orienter dans les statistiques non-paramétriques.

La qualité et la fiabilité des données est donc une préoccupation importante en sciences sociales sans exclure bien sûr qu'elle le soit pour les sciences naturelles. Par ailleurs, il faut préciser que la géographie humaine tient une place spécifique dans ce domaine parmi les sciences sociales en raison de la nature particulière des données qu'elle traite. L'information géographique est tri-dimensionnelle puisque chaque variable est localisée par un système de référence ou de localisation. Cette localisation peut être considérée implicitement (A) ou explicitement (B) dans le traitement numérique de données. En effet le géographe essaie de mettre en évidence la structure spatiale d'un phénomène. Cet objectif vise essentiellement la recherche des caractéristiques spatiales d'un lieu ou d'une distribution, les facteurs explicatifs de ces caractéristiques (au sens géographique), de mettre en relation plusieurs répartitions géographiques dans l'espace etc...

A. Matrice géographique

	V_1	V_i
n_1		
n_j		

B. Matrice géographique

	X	Y	V_1	V_i
n_1				
n_j				

V=variable

X, Y= -coordonnées cartésienne de "n" -observation

2 Démarche scientifique et démarche statistique

Entreprendre la recherche vise fondamentalement à l'acquisition de connaissances à l'intérieur d'un processus de recherche qui ne laisse rien au hasard. A l'intérieur de ce processus, il y a des étapes relatives à la démarche quantitative qui composent en parties la démarche de recherche. En d'autres mots, la démarche quantitative est non seulement subordonnée au cadre de la recherche défini par le chercheur, mais également une composante importante de la démarche scientifique. Par exemple, la formulation d'une hypothèse de recherche ne doit pas être confondue avec l'hypothèse statistique. L'hypothèse de recherche est le résultat d'un questionnement théorique i.e. d'une réflexion théorique sur le sujet ou l'aboutissement d'un état de la question. Or, l'hypothèse statistique est sous-jacente à l'hypothèse de recherche et constitue une supposition ou une proposition concernant la valeur d'un paramètre ou concernant le type de distribution d'une population ou concernant le type de relation statistique entre deux ou plusieurs variables. L'hypothèse statistique est formulée dans le cadre d'une technique statistique comportant des procédures formelles permettant de tester statistiquement des hypothèses. En d'autres mots, c'est l'Hypothèse statistique qui est testée i.e. l'hypothèse nulle. Par exemple, la démarche courante utilisée en statistique est celle d'un test d'hypothèse et le modèle quantitatif comporte en général les étapes suivantes :

- formulation de l'hypothèse nulle ($H - 0$) et l'hypothèse alternative ($H - 1$) ;
- choix d'un test approprié satisfaisant les conditions de la recherche ;
- détermination de la taille de l'échantillon ;
- choix d'une distribution d'échantillonnage ;
- cueillette de données et application de la méthode.

Alors, cette démarche quantitative s'inscrit à l'intérieur d'une démarche scientifique comportant plusieurs types de traitement afin de répondre aux objectifs et aux hypothèses de recherche. C'est pour cela qu'il faut bien comprendre la différence entre les deux démarches.

3 La qualité des applications de la statistique en sciences sociales

L'utilité de la statistique est tout à fait incontestable dans les sciences sociales. Malgré les réticences et les objections, son introduction s'est fait graduellement et a permis un bond important dans le domaine du traitement de données. Cette discipline en apparence sans objet rassemble une gamme de techniques et d'outils mathématiques permettant de réduire une matrice de données en déformant le moins possible les données d'origine, en facilitant l'interprétation et en contribuant à une meilleure compréhension du phénomène à l'étude.

En géographie, l'introduction des méthodes quantitatives au début des années 60 a bouleversé les pratiques de recherche et amené de nouvelles méthodes comme l'échantillonnage probabiliste. Les géographes ont intégré la démarche quantitative dans la démarche scientifique. De plus, l'accès à ces techniques a été facilité par

le développement de l'informatique et l'entrée de systèmes conviviaux. Toutefois, cet encouragement à l'utilisation de l'outil a ouvert la voie à des abus remarqués à plusieurs reprises par les statisticiens. Plusieurs questions sont soulevées par cette situation.

- Les statisticiens ont-ils une compétence exclusive dans le champ de la statistique ?

Plusieurs statisticiens considèrent que plusieurs usagers n'ont pas la compétence nécessaire en statistique pour appliquer les techniques dans leurs recherches. Certains postulent que l'utilisation des techniques implique une compétence et une reconnaissance institutionnelle. Alors, est-ce qu'il faut être détenteur d'un diplôme en statistique pour faire bon usage de la discipline ? Sinon, quel est le niveau requis d'apprentissage et de compétence pour être un bon usager de la statistique ? La qualité des applications étant souvent fonction de la formation académique des usagers et aussi des "essais" plus ou moins réussis de techniques plus ou moins comprises par eux. Néanmoins, le problème semble persister mais à une échelle moins grande car la formation académique dans le domaine de la statistique est de plus en plus répandue. De plus, ceci devrait accroître la qualité des applications statistiques et les statisticiens devraient "surveiller" et critiquer les écarts de conduite.

- Jusqu'où le statisticien peut-il faire acte d'indiscipline ?

Quel est le rôle du statisticien lorsqu'il travaille avec des chercheurs d'autres disciplines ? L'un de ces rôles est certainement dans le cas échéant à titre de consultant dans la construction du modèle statistique. Le dialogue entre l'usager et le statisticien devient possible et nécessaire car on ne peut plus prétendre que la statistique est un champ exclusif des statisticiens car les usagers ont de plus en plus une formation plus ou moins poussée dans le domaine. Dans le cas où l'usager n'y connaît rien, le rapport de dépendance envers le statisticien est grand : alors, comment peut-il y avoir un dialogue pour la mise au point du modèle statistique ? Plus souvent qu'autrement, le statisticien bâtit tant bien que mal un modèle statistique en étant presque en rupture de dialogue avec le chercheur dépossédé et pourtant l'expert de la question. Il y a là une situation où il serait tenté de redéfinir le problème, les hypothèses ou les questions posées dans un domaine où il n'est pas nécessairement compétent à moins **d'en faire un champ exclusif d'applications et son objet de recherche**. Alors, peut-être pourrait-il faire acte d'indiscipline impliquée dans le cours de l'investigation en acceptant qu'il n'est pas le seul détenteur de la logique formelle, de la rigueur scientifique et de la recherche de systématisation.

Il ne faut pas confondre une démarche scientifique et une démarche statistique, car le chercheur sur le sujet possède sa problématique, connaît l'art de poser les bonnes questions et l'état des questions. Cette démarche est indépendante de la statistique : seul l'usager est l'expert dans son domaine, celui qui a développé un créneau lui permettant d'être à la fine pointe. En revanche, le statisticien pourrait développer **une compétence dans la discipline** en apportant une contribution entre autres par l'introduction de nouvelles techniques ou de nouvelles applications à la condition bien sûr d'en connaître l'objet, la problématique et l'état de la question.