JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

EDMOND MALINVAUD

Les grands échantillons de données individuelles et leur exploitation statistique

Journal de la société statistique de Paris, tome 118, n° 1 (1977), p. 2-15

<http://www.numdam.org/item?id=JSFS_1977__118_1_2_0>

I

COMMUNICATIONS

LES GRANDS ÉCHANTILLONS DE DONNÉES INDIVIDUELLES ET LEUR EXPLOITATION STATISTIQUE

Edmond MALINVAUD,

Directeur général de l'I. N. S. E. E.

La possibilité de réunir et d'exploiter des grandes bases de données individuelles a donné une nouvelle dimension au métier de statisticien. La réflexion et l'expérience dégagent des techniques pour l'utilisation efficace des données administratives, pour les fusions de fichiers, pour l'induction à partir des gros échantillons, etc. La statistique mathématique elle-même a réorienté certaines de ces recherches en vue de telles applications.

The opportunity of joining together and exploiting the great individual data sources has given a new dimension to the statistician's profession. Analysis and experience bring out techniques for the utilization of administrative data, for combinations of files, and for the induction from big samples, etc. Mathematical Statistics themselves have re-directed some of their research in view of such applications.

Die Möglichkeit einer Sammlung und einer Ausbeutung von ungeheuren Mengen von individuellen Einzelheiten gab dem Beruf des Statistikers neue Dimensionen. Die Uberlegung und die Erfahrung entwickeln Techniken für eine erfolgreiche Ausbeutung von Verwaltungsresultaten, die Zusammenlegung von Karteien, für Schlussfolgerungen, die von grossen Massen von Gegebenheiten ausgehen, usw. Die mathematische Statistik selbst hat gewisse Untersuchungen zur Verwendung ihrer Methoden auf diesem neuen Gebiet umgestellt.

Le travail des statisticiens a toujours combiné des opérations de nature variée, opérations tantôt simples et répétées, tantôt délicates et plus ou moins complexes, tantôt même purement abstraites. Pour l'exposé de ce soir, elles peuvent être regroupées en trois catégories :

1° constitution de bases de données assez nombreuses pour révéler des permanences statistiques;

2° calculs plus ou moins lourds sur ces données en vue d'en faire apparaître les traits significatifs;

3° recherche des méthodes grâce auxquelles des inductions valables peuvent être établies à partir des données.

Journal de la Société de statistique de Paris, tome 118, n° 1, 1977.

Ces trois types de fonctions subsistent aujourd'hui; mais l'activité des statisticiens s'est beaucoup renouvelée, notamment à la suite de la révolution technique que constitue l'utilisation en grand de l'informatique. Nous avons aujourd'hui la possibilité d'exploiter de grosses bases de données individuelles, potentiellement riches en informations. Nous nous devons de le faire; mais ceci ne va pas sans exiger de nous une nouvelle technicité.

C'est d'elle que je voudrais parler ce soir, pensant particulièrement à ceux d'entre vous dont la formation statistique est ancienne. Vous êtes sûrement curieux de connaître la vie professionnelle de la nouvelle génération des statisticiens qui travaillent dans l'administration et dans l'université. C'est au fond de cela dont il va s'agir.

Pour en donner un aperçu je prendrai successivement les trois catégories d'opérations que j'ai citées tout à l'heure.

I — Bases de données

Les deux grandes sources de données sont d'une part les enquêtes et recensements, d'autre part les informations collectées à l'occasion d'opérations administratives. Il faut les considérer tour à tour. Il faut aussi consacrer quelque temps aux fusions entre bases de données, car celles-ci jouent un rôle de plus en plus important pour améliorer la précision et la pertinence de nos statistiques.

1. *Enquêtes et recensements*

Je peux passer assez rapidement sur les enquêtes et recensements dont la nature n'a pas sensiblement varié. Certes les enquêtes sont plus nombreuses et intéressent des effectifs plus importants. Mais les questionnaires restent soumis à la contrainte très sévère qu'impose un bon accueil au moment de la collecte : en pratique on se heurte à de grandes difficultés quand on demande aux personnes ou entreprises interrogées des informations qu'elles n'ont pas disponibles soit en mémoire, soit dans des documents aisément accessibles. C'est pourquoi les questions posées dans les recensements démographiques n'ont guère varié depuis le début du siècle; c'est pourquoi on recherche toujours la simplicité dans les enquêtes auprès des entreprises; c'est pourquoi on vise à ce que l'interview d'un ménage soit aussi bref que possible, de quelque vingt minutes pour certaines enquêtes à une heure pour les plus complexes.

Toutefois l'informatique a complètement changé les conditions du dépouillement en apportant une souplesse qui manquait autrefois. Pour les recensements démographiques du début du siècle, L. March avait mis au point une chaîne d'exploitation qui était très efficace en fonction des moyens de l'époque mais qui limitait malgré tout assez strictement les tableaux que l'on pouvait obtenir. Aujourd'hui l'informatique aide le statisticien de multiples façons.

Ainsi le chiffrement est maintenant limité aux codes élémentaires correspondant à des questions analytiques tandis que la machine codifie les classements synthétiques : dans les recensements la catégorie socio-professionnelle est déterminée par l'ordinateur à partir des réponses aux questions concernant la profession, le statut, la qualification, l'activité économique.

Plus généralement l'ordinateur assiste et contrôle le chiffrement comme la saisie, c'est-à-dire la mise sur un support propre au traitement automatique. Des informations peuvent être écrites littéralement et codifiées par la machine. Les réponses manquantes, impossibles ou suspectes sont repérées systématiquement et, suivant les cas, soit renvoyées au chiffrement pour investigations complémentaires par émission de « messages d'anomalie » (données sur les grandes entreprises), soit complétées ou corrigées automatiquement de telle manière que les tableaux statistiques ne laissent apparaître ni lacune, ni incohérence sans en être pour cela biaisés.

Par ailleurs on peut concevoir un dépouillement extrêmement riche en reclassant de multiples façons les unités élémentaires. On peut même conserver sur support informatique et réexploiter en fonction de besoins nouveaux non seulement les données telles qu'elles ont été collectées sur les questionnaires, mais surtout des fichiers intermédiaires qui sont beaucoup plus commodes. Sur ces fichiers intermédiaires l'information est structurée non pas en fonction des exigences de la collecte mais en vue des dépouillements statistiques les plus divers; elle a fait l'objet d'un premier classement, éventuellement de calculs qui sont généralement utiles par la suite. Par exemple après un recensement ou une enquête sur les entreprises, un fichier intermédiaire peut contenir non plus toutes les grandeurs directement relevées mais surtout des soldes et des ratios : valeur ajoutée, part des salaires, rapport entre les stocks et la production, productivité du travail, etc.

En somme, les techniques de dépouillement se sont totalement renouvelées; il faut pour les concevoir une grande compétence dans cette spécialité nouvelle qu'est l'analyse informatique.

2. *Données administratives*

Dans la vie moderne chacun d'entre nous doit fournir souvent des informations au fisc, à la sécurité sociale ou à d'autre administrations. Le bon sens suggère que le statisticien a intérêt à tirer parti de ces informations plutôt qu'à chercher à les collecter à nouveau. En effet l'utilisation systématique des sources administratives a bien constitué une des novations majeures de l'après-guerre dans le domaine de la collecte. Il faut toutefois en connaître les servitudes.

Les données administratives sont en principe d'une bonne qualité : elles doivent être exhaustives pour le champ qu'elles recouvrent; elles donnent lieu à des vérifications puisqu'elles constituent une référence pour certaines décisions administratives. Elles sont également déjà rassemblées dans des bureaux de sorte que le statisticien les trouve commodément à un coût faible.

L'expérience prouve néanmoins que la réalité est moins favorable que ce schéma l'impliquerait. Les données ne sont pas toujours celles que le statisticien recherche; le champ couvert et la définition des grandeurs doivent permettre l'application de lois et de règlements plutôt que faciliter la connaissance de la réalité économique et sociale. Il est certes possible de modifier quelque peu les questionnaires administratifs pour mieux répondre aux besoins du statisticien; mais de telles modifications ne peuvent jamais aller très loin. Le contrôle des déclarations faites à l'administration est par ailleurs orienté vers les réponses à certaines questions particulières alors que beaucoup d'autres ne jouent qu'un rôle secondaire et parfois ne sont même pas renseignées; les vérifications visent uniquement à une application efficace des textes (par exemple elles ne concernent pas les revenus situés au-dessous des seuils

d'imposition). De plus l'exhaustivité n'est pas garantie au statisticien qui n'intervient pas au moment de la collecte, mais plus tard après l'utilisation par les bureaux; il doit exiger que sa base de données concerne tous les documents y compris ceux qui seraient encore à l'examen en raison de retards ou de litiges. Enfin les documents et procédures administratives ne sont pas toujours uniformes dans l'ensemble de la France, ce qui complique évidemment beaucoup les dépouillements statistiques.

De tous ces facteurs il résulte que, si la constitution de bases de données administratives a un coût global relativement faible, elle exige en revanche énormément de temps de cadres. La situation varie d'ailleurs d'un cas à l'autre. L'exploitation statistique des déclarations annuelles de salaires (D. A. S.) est relativement simple; effectuée depuis longtemps déjà, elle donne satisfaction dans l'ensemble. Également ancienne, la statistique des déclarations des entreprises pour le calcul des bénéfices industriels et commerciaux (B. I. C.) a exigé beaucoup d'efforts et reste d'une qualité variable selon les années. Nous mettons en place actuellement un système pour l'utilisation des fichiers de paie des administrations en vue de l'établissement de statistiques sur les effectifs et les salaires dans la fonction publique : la diversité des pratiques appliquées pour le paiement par les administrations complique beaucoup notre tâche. Bien qu'elle ait été envisagée depuis longtemps et que nous y travaillons actuellement de façon active, la mobilisation statistique des données collectées par la sécurité sociale n'a pas encore été systématiquement réalisée.

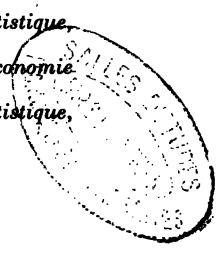
3. *Fusions de fichiers*

Pour étudier de nombreux phénomènes il s'avère nécessaire de regrouper au niveau individuel des informations obtenues séparément. Ce peut être dans le cadre d'une même exploitation, comme quand il s'agit de regrouper tous les salaires reçus pendant une année par un même individu mais versés par divers employeurs (salariés II); il faut alors procéder à une opération automatique de regroupement des déclarations en fonction des individus concernés. Le plus souvent ces informations proviennent de deux sources distinctes; pour les réunir il faut procéder à une fusion de fichiers. Je peux donner quelques exemples faisant apparaître comment nous pouvons ainsi constituer des bases de données plus riches que celles résultant de chacune des sources prises isolément.

Dans de nombreux cas, connaître la situation d'un individu ou d'une entreprise à un moment particulier ou durant une période particulière ne suffit pas; il faut connaître deux ou plusieurs situations, distinctes dans le temps, pour procéder à une analyse longitudinale. La base de données doit alors rapprocher pour chaque individu les informations obtenues à l'aide de collectes successives de même nature : par exemple les déclarations annuelles de salaires successives pour un échantillon d'individus (salariés III), les réponses successives des entreprises sur les prévisions et réalisations d'investissements dans les enquêtes de l'I. N. S. E. E., les bilans et comptes successifs des entreprises dans les centrales de bilans (1)

Dans d'autres cas on souhaite rapprocher pour une même unité les informations

1. Sur ces trois types de fusions de fichier, voir :
— B. Grais, « Les salaires : indicateurs rapides et observations approfondies », *Économie et statistique*, juillet-août 1974.
— V. Thollon-Pommerol, « Les entreprises prévoient difficilement leurs investissements », *Économie et statistique*, mars 1974.
— « Les diverses centrales des bilans sont regroupées en un comité de liaison », *Économie et statistique*, octobre 1975.



venant de sources distinctes soit pour réaliser un contrôle mutuel, soit plus souvent pour que ces informations se complètent et permettent des analyses en profondeur que chaque source prise isolément ne permet pas. Ainsi notre collègue Claude Gruson inspire un projet de fusion des enquêtes réalisées par l'I. N. S. E. E. auprès des chefs d'entreprise avec les informations collectées à leur sujet par la centrale des risques de la Banque de France (autorisation et utilisation des crédits bancaires). A l'I. N. S. E. E. nous réalisons maintenant régulièrement pour les grandes et moyennes entreprises la fusion des B. I. C. avec les résultats des enquêtes annuelles d'entreprises faites par les statisticiens du ministère de l'Industrie et des autres ministères techniques : nous vérifions ainsi l'exhaustivité de chacune des deux sources et la concordance de certains renseignements figurant dans l'une et l'autre; nous complétons de plus les renseignements comptables avec des informations sur les productions, l'emploi, la répartition du chiffre d'affaires entre les divers groupes de produits, la décomposition des investissements par nature, etc. : c'est S. U. S. E. le « système unifié de statistiques d'entreprises (1) ».

Je n'ai pas besoin d'insister sur le fait que de telles fusions font apparaître les lacunes et les défauts de chacun des fichiers; elles nous obligent à avoir une conscience bien meilleure de la valeur de nos instruments; elles nous imposent la nécessité d'efforts pour améliorer la qualité de chacune de nos sources. Elles sont donc au total très bénéfiques pour la précision de notre information économique et sociale.

Mais elles aussi exigent énormément de travail de cadres pour concevoir, organiser, vérifier, modifier ... La mise au point d'une base de données fusionnées qui soit propre à l'utilisation exige des mois lorsqu'elle est devenue habituelle, souvent des années quand il s'agit de la première opération du genre dans un domaine particulier.

II — La pratique de l'induction

Quoi qu'il en soit, nous disposons aujourd'hui de fichiers contenant des données individuelles nombreuses. Comment les utilise-t-on pour mieux connaître les phénomènes économiques et sociaux?

1. « *Les chiffres vont parler* »

Le néophyte se croit privilégié s'il a à sa disposition des informations concernant plusieurs milliers d'unités. Il va suffire de regarder, croit-il, pour voir apparaître la réalité des phénomènes et pour départager définitivement les diverses thèses émises par des penseurs ne disposant pas des mêmes informations.

Malheureusement les chiffres ne parlent pas aussi clairement, ou tout au moins il faut un certain art et beaucoup de persévérance pour les faire parler.

Au début, dans les premiers travaux qui portent le plus souvent sur une fraction des données, tout paraît confus. Les dispersions à l'intérieur des sous-groupes d'individus restent énormes. Sauf quand elles ne font que traduire des évidences, les régressions conduisent à des coefficients de corrélation multiple très faibles. Les coefficients de régression eux-mêmes sont souvent peu satisfaisants.

1. Voir F. Pierrugues, « Une fusion de quatre fichiers de données d'entreprises en 1971 : S.U.S.E. », *Économie et statistique*, octobre 1975.

Par exemple la base de données peut concerner les entreprises et rassembler les informations obtenues par les enquêtes sur les investissements, éventuellement après fusion d'enquêtes successives et même fusion avec des résultats provenant d'autres sources. On s'imaginera pouvoir quantifier précisément l'effet de l'aisance de trésorerie sur l'investissement et on sera fort déçu de trouver entre ces deux caractéristiques une corrélation si faible que la liaison supposée est même douteuse.

On s'imagine alors qu'il faut décomposer la population statistique : distinguer par exemple suivant les branches et les régions. Mais loin de préciser les choses, de telles décompositions aggravent souvent l'impression de désordre que les premiers essais avaient dégagée. Ainsi les corrélations restent très faibles à l'intérieur des branches et les coefficients de régression varient d'une branche à l'autre d'une manière qui semble être totalement erratique et n'avoir aucun rapport avec les caractéristiques de ces branches.

Arrivés à ce stade certains arrêtent leurs efforts quant à l'étude des données et se réfugient dans le discours. Cependant il reste le plus souvent possible de dégager des enseignements valables à condition d'organiser patiemment les faits.

Mais alors d'une part il importe d'exploiter à fond les possibilités qu'ouvre le grand nombre des unités observées : une liaison qui reste incertaine avec un échantillon de 200 unités se dégage au contraire clairement sur un effectif de 10 000, même si le coefficient de corrélation reste faible; par exemple, un modèle peut être conçu et traité, qui permette de considérer simultanément toutes les branches et toutes les régions et qui incorpore néanmoins certains effets propres à ces branches ou régions.

D'autre part la recherche n'est fructueuse que si le raisonnement ou la réflexion interviennent pour orienter l'examen des données. Il faut se fixer *a priori* des hypothèses, c'est-à-dire une théorie du phénomène, théorie que l'on testera plus ou moins complètement, que l'on quantifiera plus ou moins précisément. Si par exemple il s'agit du rôle que joue l'aisance de trésorerie sur l'investissement, on reconnaîtra que celui-ci dépend d'abord des besoins d'extension des capacités de production, besoins qui ont été perçus en fonction de la demande s'adressant aux entreprises. Un modèle est nécessaire pour exprimer comment aisance de trésorerie et besoins de capacité sont susceptibles de s'articuler entre eux. Avant de m'expliquer en termes généraux sur ce sujet je dois m'arrêter quelques instants à une objection de principe.

2. « *L'analyse des données* » *n'est pas une panacée*

Certains statisticiens considèrent comme très suspecte toute intervention de l'aprioriisme, toute spécification préalable d'une théorie et d'un modèle. Ils croient pouvoir faire appel à des techniques statistiques qui seraient plus puissantes et moins sujettes à arbitraire que les méthodes classiques procédant le plus souvent des techniques de régression et d'analyse de la variance. Les travaux sur « l'analyse des données », desquelles le professeur D. Dugué nous a parlé (1), auraient dégagé des procédures grâce auxquelles les sciences humaines pourraient progresser selon une voie purement factuelle. Il semble bien que cet espoir soit le plus souvent déçu lorsqu'on essaie de le matérialiser.

1. D. Dugué, « Problèmes actuels de statistique mathématique », *Journal de la Société de Statistique de Paris*, 3e trimestre 1974.

Sans doute le mouvement pour l'analyse des données a-t-il permis de définir des procédures adaptées à des problèmes que l'on n'avait guère étudiés auparavant (analyse des correspondances pour caractériser les tableaux de contingence, analyse arborescente pour construire des nomenclatures, etc.). Mais, pour la recherche en vue de laquelle les méthodes traditionnelles ont été conçues (détermination de lois entre grandeurs quantitatives), il n'existe pas de procédé efficace qui élimine totalement l'apriorisme.

On recommande l'analyse factorielle qui a effectivement sa place au premier stade de l'étude. Mais on doit bien remarquer que cette analyse suppose un choix préalable de l'ensemble des grandeurs sur lequel elle porte et que les résultats dépendent de l'ensemble retenu. Pour l'employer à bon escient, il faut déterminer au mieux l'ensemble en question grâce à une réflexion *a priori* quant aux liaisons dont l'existence est sujette à examen : il faut éviter d'introduire simultanément deux variables nécessairement corrélées entre elles, il faut se garder vis-à-vis des liaisons évidentes dont la présence masquerait les phénomènes que l'on veut explorer; il faut par ailleurs limiter le nombre des variables sans quoi la méthode devient très lourde.

De plus l'analyse factorielle traite les données d'une façon qui diffère peu de celle adoptée lors d'un emploi systématique de régressions. Dans un cas comme dans l'autre les informations, c'est-à-dire les distributions multidimensionnelles de l'ensemble des grandeurs observées, sont d'abord résumées par le calcul des moments du premier et second ordres, les autres caractéristiques des distributions n'étant considérées ni par l'une ni par l'autre des techniques. Celles-ci conduisent alors en pratique à des résultats voisins. Or un problème se pose précisément parce que ces premiers résultats sont généralement décevants et n'épuisent pas les enseignements que l'on peut tirer des données.

3. *Va-et-vient entre calculs et formulation*

En vue de réduire l'influence des idées *a priori* et d'aboutir néanmoins à des résultats assez précis, la meilleure démarche semble devoir aller du général au plus spécifique. De premiers calculs sont conduits par des méthodes aussi purement descriptives qu'il est possible. Puis des formulations plus étroites sont introduites et ajustées.

Par exemple on considérera tout d'abord les corrélations entre taux d'investissement, rythmes de croissance passés des ventes, aisance de trésorerie, etc. Puis on essaiera un modèle simple rendant compte du phénomène d'accélération. On cherchera ensuite à y introduire les effets de branche et à y apprécier le rôle de l'aisance financière, grâce à différentes formulations essayées successivement.

De la sorte on a une certaine connaissance directe de l'ensemble des données avant de chercher à estimer grâce à elles des modèles. Cette démarche consistant à rechercher des formulations de plus en plus étroites, et à retenir celles qui s'avèrent expliquer le mieux les données remplace pour les sciences humaines la démarche expérimentale des sciences physiques. Bien qu'elle offre moins de garantie, elle n'a pas une nature totalement différente.

L'étude de la distribution des données sur l'ensemble de la population et sur diverses sous-populations doit intervenir tout d'abord. A ce stade l'analyse factorielle joue souvent un rôle intéressant. Mais elle n'est pas la seule.

Une des opérations les plus délicates qui intervienne à ce moment consiste dans l'élimination des valeurs aberrantes.

Les sources d'erreur dans les données des gros fichiers sont nombreuses. Certaines erreurs sont très fortes, comme celles provenant du recours à de mauvaises unités de mesure. Il importe de les faire disparaître, d'autant plus que les procédures statistiques usuelles sont basées sur les moments du second ordre des distributions et attribuent donc un grand poids aux valeurs extrêmes. Or la vérification de la base de données permet rarement de repérer et de corriger toutes les erreurs. C'est pourquoi on procède habituellement à une élimination automatique des unités qui apparaissent très anormales soit à l'examen de certains ratios (j'en ai déjà parlé), soit à la lumière des premières régressions ou analyses factorielles.

Au moment où des formulations de plus en plus précises sont essayées, le chercheur peut évidemment tirer parti de toute la richesse des spécifications aléatoires qui ont été imaginées; il peut même en inventer de nouvelles. Les décrire serait donc ici hors de propos. Pour donner une idée de la diversité des situations, je peux néanmoins relever trois particularités qui, se présentant souvent, méritent de retenir l'attention, d'autant plus qu'elles influencent les estimations et les tests.

En premier lieu les unités rassemblées dans les grandes bases de données peuvent habituellement être classées en fonction de critères divers et l'appartenance à telle ou telle catégorie est susceptible d'influencer les phénomènes sur lesquels on porte l'attention. Il convient donc le plus souvent d'expliciter ces classements dans les modèles retenus : des indices particuliers repèrent non seulement le numéro de chaque unité mais aussi son appartenance aux diverses catégories et la possibilité d'effets propres à chaque catégorie est reconnue. On parle alors de modèles à deux ou plusieurs indices.

En second lieu, dans l'étude du phénomène considéré, l'influence de tel facteur peut varier d'une unité à une autre sans relation avec le classement de cette unité et d'une manière qui semble purement aléatoire. On spécifie alors un modèle à coefficients aléatoires qui reconnaît une telle possibilité.

En troisième lieu, si certains des facteurs supposés jouer sont mesurés exactement pour chaque unité, d'autres ne sont pas directement observés et ne sont saisis que par l'intermédiaire de grandeurs supposées être en bonne corrélation avec eux (grandeurs dites « proxy » en anglais). Il y a donc des « erreurs sur les variables », c'est-à-dire des écarts entre les grandeurs observées et celles qui devraient intervenir dans l'explication. Cette particularité doit elle aussi être explicitée dans le modèle.

III — La théorie statistique en face de la pratique

On ne doit pas être surpris de constater que la recherche théorique s'est orientée vers les problèmes nouveaux que faisait apparaître dans ses progrès la pratique de l'induction. En fait l'exposé de ce soir peut être une occasion de constater que certaines des recherches modernes en statistique mathématique sont beaucoup moins gratuites que ne le pensent souvent les praticiens déroutés par la difficulté des articles ou ouvrages rendant compte de ces recherches.

Sans prétendre le moins du monde être complet ou même résumer correctement ceux des travaux que j'évoquerai, je vais me référer aux indications données tout à l'heure quant aux procédures employées pour le traitement des grandes bases de données individuelles.

1. *Estimation avec tests préliminaires. Choix entre modèles. Hypothèses emboîtées*

Tout d'abord que penser du principe même de la démarche procédant par va-et-vient entre calculs et formulations? Elle semble assez étrangère aux spécifications fondamentales retenues par la statistique mathématique. Qu'ils soient classiques ou bayésiens, les logiciens exposent que l'induction procède à partir de deux éléments : un modèle traduisant ce que l'on sait *a priori* du phénomène et un échantillon résultant d'observations nouvelles sur ce phénomène; il s'agit alors de préciser le modèle à l'aide de l'échantillon soit en estimant les paramètres inconnus, soit en testant certaines hypothèses particulières. Il ne semble pas être question de choisir le modèle après des traitements successifs des données.

En réalité il n'y a pas de véritable antinomie entre la démarche des praticiens et le point de départ de la statistique mathématique. Le modèle du théoricien doit être compris comme général et comme englobant tous les modèles particuliers auxquels le praticien peut parvenir après son étude des données. Néanmoins le va-et-vient consiste en des opérations logiques moins simples que l'estimation d'un modèle ou le test d'une hypothèse. Ces opérations avaient été peu étudiées jusqu'à une période récente.

Devant les besoins de la pratique, des recherches théoriques assez nombreuses leur ont été consacrées depuis une dizaine d'années. Le théoricien doit bien entendu s'en tenir à des procédures relativement simples mais typiques de la démarche suivie par les praticiens. Les procédures suivantes retiennent particulièrement l'attention :

- Dans le cadre du modèle général H une hypothèse $H_0 \subset H$ est d'abord testée; si elle est rejetée, la spécification H est estimée sans référence à cette hypothèse particulière, sinon la spécification H_0 est estimée. Parmi les questions que pose une telle procédure, figure celle de savoir quel seuil de signification retenir pour le test préalable en vue d'atteindre l'efficacité la plus grande possible au moment de l'estimation (1),
- Le modèle général est conçu comme étant l'union de deux spécifications H_1 et H_2. On choisit d'abord laquelle des deux s'applique, puis on l'estime (2),
- On considère une série d'hypothèses emboîtées du type ... $H_{n+1} \supset H_n \supset H_{n-1}$ ... et on veut décider laquelle doit être retenue à la lumière de l'échantillon. La question se pose alors de savoir s'il faut considérer les hypothèses dans l'ordre descendant (en allant du plus général au plus particulier) ou dans l'ordre ascendant et quels seuils retenir dans chaque cas (3).

2. *Procédures robustes*

On sait que les tests usuels, les régressions multiples et autres méthodes classiques jouissent de propriétés très intéressantes dans les cas sur lesquels la théorie a depuis longtemps concentré son attention. Parmi les hypothèses habituellement retenues par la théorie figure

1. Voir par exemple à ce sujet T. Sawa et T. Hiromatsu, « Minimax Regret Significance Points for a Preliminary Test in Regression Analysis », *Econometrica*, novembre 1973.
2. Pour réfléchir sur ce problème, le lecteur peut se reporter au cas traité par D. Leech, « Testing the error specification in nonlinear regression », *Econometrica*, juillet 1975.
3. Voir W. J. Kennedy et T. A. Bancroft, « Model Building for Prediction in Regression based upon Repeated Significance Tests », *Annals of Mathematical Statistics*, août 1971.

celle selon laquelle les éléments aléatoires des modèles suivraient des lois de probabilité de Laplace Gauss.

Il s'agit d'une hypothèse commode qui fournit parfois une première approximation valable. L'expérience semble bien montrer cependant qu'elle ne vaut pas pour la plupart des bases de données individuelles, même quand elles ont fait l'objet de vérification, particulièrement minutieuses : il y a une proportion certes faible mais trop élevée d'unités qui dévient très fortement des tendances moyennes autour desquelles se groupent la plupart des autres unités. Les distributions statistiques ont des queues plus épaisses que celle de Laplace Gauss.

L'existence d'unités anormalement déviantes a d'ailleurs amené les praticiens à éliminer plus ou moins automatiquement ces unités quand ils en constatent la présence, ainsi que nous l'avons vu tout à l'heure. Il est remarquable que certaines recherches théoriques récentes justifient cette pratique et recommandent même que l'élimination porte sur une proportion notable des unités (1).

En effet certains mathématiciens ont étudié la « robustesse » des procédures usuelles et de diverses alternatives, telles celle supposant une élimination préalable des données extrêmes. Ils ont cherché à déterminer les propriétés des diverses procédures en cause dans les cas dans lesquels les éléments aléatoires ne sont plus gaussiens, par exemple quand leurs distributions ont des queues épaisses. Ils ont alors trouvé que les méthodes usuelles n'étaient effectivement pas très robustes, contrairement aux procédures alternatives envisagées.

On en arrive ainsi aujourd'hui à un stade où la théorie rejoint des pratiques qui ont cours depuis longtemps mais sur lesquelles on n'osait pas trop insister par crainte de paraître faire un choix parmi les observations.

3. *Des méthodes classiques dans des situations complexes*

D'autres recherches théoriques, suscitées par le traitement des grands échantillons de données individuelles, se situent davantage dans le prolongement de la statistique mathématique classique mais concernent des modèles plus complexes que ceux sur lesquels l'effort s'était concentré. Pour en donner une idée, je peux me référer aux trois particularités que j'ai signalées précédemment comme se rencontrant assez souvent dans les spécifications aléatoires retenues par les praticiens.

Dans le traitement des modèles à deux ou plusieurs indices on n'est pas habituellement intéressé aux effets particuliers qui, dépendant des catégories auxquelles appartiennent les unités, jouent un rôle perturbateur. Le but du traitement est plutôt de quantifier les relations qui s'appliquent à l'ensemble des unités. Deux méthodes simples peuvent être conçues considérant l'une la distribution des unités à l'intérieur de chaque catégorie élémentaire, l'autre la distribution des moyennes calculées sur ces catégories : analyse intraclasse d'une part, analyse interclasse d'autre part. La meilleure méthode combine, on s'en doute, les deux types d'analyse, mais elle dépend des hypothèses que le modèle retient quant aux effets particuliers et sa définition n'est pas évidente *a priori* (2).

1. Pour un exposé général sur les procédures robustes, voir P. J. Huber, « Robust Statistics : A Review », *Annals of Mathematical Statistics*, August 1972. Pour avoir une idée de la proportion optimale des rejets on pourra se reporter à F. C. Leone and E. Moussa-Hamouda, « Relative Efficiency of O. Blue Estimators in Simple Linear Regression », *Journal of the American Statistical Association*, December 1973.

2. Une excellente présentation de ces problèmes a été donnée par P. Mazodier, « L'estimation des modèles à erreurs composées », *Annales de l'INSEE*, mai-août 1971.

Quand un modèle à coefficients aléatoires est retenu, on ne se propose évidemment pas d'estimer les coefficients qui s'appliquent à chacune des unités. Seules importent les valeurs moyennes de ces coefficients sur l'ensemble des unités, ainsi que, accessoirement, les dispersions des coefficients. On imagine sans peine que les régressions usuelles estiment bien les coefficients moyens; en revanche les écarts-types calculés suivant les méthodes usuelles peuvent donner une idée trop favorable de la précision atteinte (1). C'est une des raisons pour lesquelles il convient de ne pas trop s'illusionner quand, à la suite d'une analyse coassique sur données individuelles nombreuses, on voit apparaître de très faibles écarts-types.

A côté du modèle classique de la régression multiple, la statistique mathématique et l'économétrie ont étudié depuis longtemps deux autres modèles généraux : d'une part le modèle à équations multiples dont les coefficients sont soumis à des restrictions traduisant la structure du phénomène considéré, d'autre part le modèle supposant une relation linéaire entre variables sujettes à erreurs. Ce dernier était d'ailleurs peu utilisé, notamment parce que son identification reposait soit sur l'existence d'informations concernant la matrice des variances et covariances des erreurs, soit sur l'existence de variables instrumentales correlées avec les grandeurs du modèle mais non avec les erreurs affectant leur observation.

Les spécifications adaptées aux bases de données individuelles font maintenant apparaître des cas dans lesquels quelques-unes des grandeurs seulement sont considérées comme affectées d'erreurs d'observation importantes et simultanément quelques variables instrumentales assez naturelles peuvent être trouvées. Mais, le plus souvent, le modèle a aussi les particularités qui sont habituelles dans les modèles à équations multiples. Il faut donc appliquer des méthodes d'estimation combinant les principes élaborés pour le traitement des équations simultanées d'une part, des erreurs sur les variables d'autre part (2). On ne sera pas surpris d'apprendre que les calculs sont alors complexes. Mais avec les moyens desquels nous disposons aujourd'hui ils deviennent possibles.

* * *

Cet exposé trop rapide montre, je l'espère, la diversité des travaux et des préoccupations suscités par le traitement des grands échantillons de données individuelles. Des champs d'exploration nouveaux et nombreux sont ouverts aux statisticiens de la jeune génération. Ceci témoigne de la vitalité de notre discipline.

DISCUSSION

M. Gibrat, ancien président. — L'analyse des données pénètre de plus en plus la vis industrielle. Je prendrais un exemple que je travaille beaucoup depuis quelque temps : les réseaux de surveillance et d'alerte de la pollution atmosphérique. Il y en a quelques-uns aujourd'hui, il y en aura trente à la fin de l'année prochaine : 30 postes, une mesure de SO^2 par exemple par quart d'heure soit cent par jour. Le cycle des saisons capital est d'un an : cela fait plus d'un million de données pour un seul réseau. Naturellement on n'a pas

1. Sur ce point et sur des méthodes d'estimation propres aux modèles à coefficients aléatoires voir B. R. Frœhlich, « Some Estimators for a Random Coefficient Regression Model », *Journal of the American Statistical Association*, june 1973.

2. Voir Z. Griliches, « Errors in variables and other inobservables », *Econometrica*, novembre 1974.

encore traité cela, mais une seule mesure par jour en fait dix mille et c'est de toute évidence insuffisant. En face de cela une bonne quinzaine de paramètres au minimum.

On s'est donc lancé dans l'utilisation de méthodes mise en honneur par M. Benzecri. Hélas sur dix études provenant toutes de thésards 3e cycle ou de bureaux d'études spécialisés, sept ou huit sont fort criticables. L'un constate des courbes paraboliques dans les résultats, s'y intéresse et essaye de les interpréter. Il a oublié qu'il s'agit sûrement de l'effet Guttmann. D'autres, les plus fréquents, utilisent des moyennes journalières oubliant que le phénomène est météorologique et à l'échelle de l'heure subit des variations de un à dix. Ne résistent guère que les études de proximité recherchant les profils semblables. On sent que cette nouvelle technologie a été très mal digérée, sans doute trop vite apprise et que le désir marqué de ne pas avoir d'idée *a priori*, de ne pas faire de « modèles » a supprimé dans la plupart des cas tout examen préalable sérieux du phénomène physique.

Il y a dans ces méthodes l'immense intérêt de pouvoir grâce à l'ordinateur traiter facilement des problèmes autrefois intouchables. Il y a certainement beaucoup à gagner en évitant au maximum des idées *a priori*, surtout si elles sont basées sur des acceptations irraisonnées de loi normale à plusieurs variables. *L'avenir de ces méthodes est très grand.* Mais leur utilisation reste très délicate.

Il est très agréable d'avoir devant soi au lieu de dizaines de milliers de données (c'est l'ordre de grandeur traité aujourd'hui), cinq ou six axes principaux détenteurs de presque toute l'information et quelques graphiques plans. Mais leur interprétation est très difficile. Les premières valeurs propres suggèrent des conclusions qui étaient évidentes, les suivants (3 à 6 par ex.) paraissent de peu de valeur. Il y a des pièges partout.

A ne pas mettre entre toutes les mains. Exiger qu'une réflexion physique puisse précéder le choix des données et la mise en route d'un programme. Ne pas se laisser impressionner par la technologie des calculs. Tels sont les conseils à donner aux utilisateurs. Ils paraissent aujourd'hui absolument nécessaires.

M. Volle. — La description du traitement des données individuelles a reflété davantage ce qui devrait être que ce qui est. Qu'il s'agisse de la coordination des sources, de la technologie statistique ou de l'organisation du travail, de grands progrès restent nécessaires pour obtenir une information pertinente.

Il a été par ailleurs abondamment question d'analyse des données. Il ne convient pas de faire porter à cet ensemble de méthodes la responsabilité des maladresses de quelques praticiens peu expérimentés ou peu prudents. Il est clair que l'analyse des données requiert, pour être bien utilisée, une bonne maîtrise du domaine étudié, particulièrement du langage et des concepts mis en œuvre.

Correctement pratiquée, l'analyse des données apparaît comme la systématisation d'un épisode de contemplation des données, par lequel il faut passer avant toute modélisation et toute induction probabiliste.

Qui dit modèle, dit volonté d'action, donc politique et stratégie; car l'on sépare les variables en classes (exogènes et endogènes) selon les nécessités de la politique que l'on désire mener. La relation de l'analyse des données aux modèles est analogue à celle de la contemplation à l'action (analogie dont la portée reste, bien sûr, limitée).

M. Gallais-Hamonno. — Dans sa première partie M. Malinvaud a détaillé un certain nombre de contraintes qui existent sur la collecte et l'exploitation des enquêtes. Je voudrais souligner en tant qu'utilisateur des travaux de l'I. N. S. E. E. que les responsables d'enquêtes sont également soumis à des contraintes de publication. A ce propos, il faudrait que le public soit mieux informé de la possibilité d'obtenir des données non publiées.

Réponse de E. MALINVAUD. — Effectivement, l'I. N. S. E. E. met aujourd'hui à la disposition du public beaucoup plus de données qu'il n'en publie. Ces données, le plus souvent conservées dans des tableaux sur microformes, peuvent être obtenues auprès des observatoires économiques installés dans chaque région. Nous nous efforçons à chaque occasion de faire connaître l'existence de ce service. La presse et vous-même devraient nous y aider.

Après la séance du 6 octobre, Jacques Mairesse m'a signalé qu'il avait écrit à une autre occasion quelques pages concernant le sujet de mon exposé. Avec sa permission, je me permets de reproduire le passage ci-dessous qui complète très utilement la discussion sur des points que j'ai trop peu abordés.

« *L'insuffisance des tests* — L'économètre qui n'a pas acquis l'habitude de travailler sur de grands échantillons doit aussi être averti de difficultés concernant les tests habituels de significativité des variables ou de comparaison d'ajustements (tests de Student, de Fisher). Le grand nombre d'observations fait que la précision des estimations paraît excellente et que les tests usuels conduisent presque systématiquement à des conclusions positives, ce qui réduit beaucoup leur intérêt pratique.

Par ailleurs le carré du coefficient de corrélation multiple R^2 des régressions (égal à la part de la variabilité expliquée par rapport à la variabilité totale) est souvent assez faible, notamment si la variable dépendante choisie est de préférence un ratio. Il ne faut pas s'en étonner : la très grande dispersion des données individuelles traduit la complexité des phénomènes à ce niveau d'analyse (où les simplifications entraînées par la loi des grands nombres ne jouent pas) et la multiplicité des facteurs qui peuvent en rendre compte. Une telle dispersion n'interdit pas l'étude des facteurs dont on ne pense qu'ils sont les plus notables.

L'éventualité d'erreurs de spécification. — Si la précision des résultats ne fait pas problème, elle est cependant conditionnelle aux hypothèses des modèles étudiés et il n'est pas exclu que ces résultats soient en fait biaisés. Il convient donc de s'interroger sur l'éventualité d'erreurs de spécification et sur la robustesse des résultats à ces erreurs. On doit chercher à remettre en cause et à faire varier les hypothèses qui paraissent les moins justifiées ou les plus contestables, que ce soit *a priori* ou au regard de la vraisemblance des résultats obtenus. Là encore la réflexion théorique et les connaissances déjà acquises sont nécessaires, de même qu'une certaine expérience.

Deux types d'erreurs de spécification méritent d'être assez systématiquement envisagés : premièrement l'échantillon auquel un même modèle est supposé s'appliquer peut être de dimension plus ou moins grande et être défini suivant des critères plus ou moins fins (par exemple d'activité économique ou de taille d'entreprises); secondement certaines variables peuvent avoir été incluses à tort et certaines autres variables omises à tort (ou bien encore ne figurer que sous forme linéaire alors qu'elles pourraient intervenir par exemple sous forme quadratique). Ces erreurs de spécification et les biais d'estimation qu'elles sont susceptibles d'entraîner sont facilement étudiés sur les fichiers de données individuelles; c'est un avantage résultant du grand nombre des observations et de la multiplicité des variables suivies simultanément.

La possibilité d'autres types d'erreurs de spécification doit, bien sûr, être également examinée; mais l'application des biais correspondants s'avère à la fois plus compliquée et beaucoup plus délicate. C'est le cas des biais de simultanéité qui découlent de l'estimation par les moindres carrés habituels d'une régression prise isolément, quand la relation correspondante devrait s'intégrer dans un modèle plus large à plusieurs équations. C'est le cas aussi des biais qui peuvent résulter d'erreurs systématiques ou aléatoires sur les variables.

On est amené en général à retenir des hypothèses alternatives qui s'imposent surtout par leur commodité et qui ne sont pas plus vraisemblables que les hypothèses premières dont on cherche à connaître les implications. On peut néanmoins acquérir de la sorte une idée plus ou moins imparfaite de la robustesse ou de la fragilité des résultats.

Les divers essais auxquels nous avons procédé sur les données de recensement industriel nous ont montré que les biais provoqués par les erreurs de mesure sur les variables pouvaient être plus notables que ceux liés au choix des échantillons ou à celui des variables. De même les biais de simultanéité n'ont pas paru graves; il faut dire toutefois que, faute de meilleure référence, nous avons considéré seulement le modèle d'équations simultanées de concurrence parfaite où salaires et prix sont supposés donnés pour les entreprises. Cette expérience concerne l'estimation des principaux paramètres des fonctions de production; elle n'est pas nécessairement généralisable à d'autres domaines ou à d'autres corps de données.

Nous sommes cependant conduit à reconnaître que le nombre des observations et l'abondance des variables ne sauraient suffire aux besoins de l'analyse économétrique. La quantité des données ne saurait suppléer leur qualité. S'il n'est pas possible en général d'avoir d'excellentes données, du moins serait-il utile de pouvoir connaître assez précisément certaines caractéristiques des erreurs qui les affectent. C'est là sans doute une conclusion importante, bien que finalement peu étonnante, de nos investigations » (J. Mairesse).