PRÉSENTATION

Olivier Sautory
Chef de la Division "Méthodologie d'élaboration et d'analyse des données" de l'Insee

Les quatrièmes "Journées de méthodologie statistique", organisées par l'Unité "Méthodes statistiques" de l'Insee et le Groupe des Ecoles Nationales d'Economie et Statistique, se sont déroulées dans les locaux du GENES les 18 et 19 octobre 1995. Les objectifs de ces Journées, qui sont de favoriser les échanges de savoirs et d'expériences entre statisticiens de l'Insee et du système statistique public, ont été une nouvelle fois atteints : en effet, plus de 450 personnes ont écouté leurs collègues exposer les résultats de leurs travaux en matière de démographie des entreprises ou de techniques d'amélioration des enquêtes. Les participants ont également découvert des méthodes statistiques encore peu répandues dans notre environnement (analyse exploratoire, statistique non paramétrique), accompagnées d'exemples d'applications, ainsi que les "gros" outils informatiques conçus par l'Insee.

Comme c'est la tradition pour ces Journées, deux collègues canadiens nous ont fait part de leurs expériences à Statistique Canada, l'une relative à la refonte de l'enquête sur la population active, l'autre sur la formation des méthodologistes. Cette dernière conférence a servi d'introduction à un débat consacré à l'enseignement de la méthodologie statistique.

Analyse exploratoire des données

Dans la première conférence spéciale, **Eugène Horber** (Université de Genève) a proposé une réflexion sur la visualisation graphique et semi-graphique comme approche alternative de la statistique. Son objectif était de montrer qu'à côté du regard habituel du statisticien (distribution, paramètres, modèles, etc.), il existe un autre regard plus visuel qui tire sa force de l'interaction dynamique entre les outils, les capacités visuelles du cerveau et les moyens informatiques. Cet exposé était accompagné d'une série d'exemples illustrant les différentes familles d'outils graphiques et visuels.

L'analyse exploratoire des données a fait son entrée dans les logiciels statistiques dès le début des années 1980. Il a fallu cependant attendre les récents progrès techniques et la montée en puissance des machines pour qu'elle devienne une offre standard des principaux logiciels. **André Wielki** (Direction Générale de l'Insee (DG), département de l'informatique) a présenté le module d'analyse exploratoire du logiciel SAS, appelé SAS-INSIGHT, à travers plusieurs exemples et démonstrations. Grâce à ce module, l'utilisateur dispose d'une collection d'outils graphiques simples, robustes, et interactifs pour comprendre et explorer ses données sans se mettre a priori dans le carcan d'un modèle statistique.

Les méthodes de la statistique exploratoire peuvent être utilisées sur de gros fichiers, comme l'a montré **Magda Tomasini** (DG, département de la conjoncture), en présentant les résultats d'une étude portant sur la modélisation de la loi des délais de la construction (délai entre l'ouverture du chantier et l'achèvement des travaux) et sur les différents facteurs influençant cette variable. Des outils tels que les Box-plots, QQ plots, median polish,..., ont été utilisés sur un fichier de plusieurs centaines de milliers d'enregistrements, sur lequel il a été possible de chercher et valider des hypothèses sur la forme et les déterminants de la loi.

Certaines techniques de la statistique exploratoire permettent de juger de la stabilité des résultats vis-à-vis d'une légère modification des données. Les méthodes de détection de la multicolinéarité dans le modèle linéaire répondent d'une certaine façon à un objectif similaire, car cette multicolinéarité engendre des problèmes numériques et statistiques qui se traduisent par des difficultés d'estimation potentiellement très graves. **Hélène Erkel-Rousse** (DG-CREST) s'est intéressée aux indicateurs proposés par D.A. Belsley, E. Kuh, et R.E. Welsch : les indices de conditionnement et le tableau de composition des variances, et a montré comment ils doivent être correctement interprétés.

Les "gros" outils de la Statistique Publique

Le codage automatique existe depuis longtemps à l'Insee, grâce au logiciel QUID. Le projet SICORE, lancé en 1993, visait à améliorer et à généraliser les principes initiaux de QUID, afin de créer un système général de chiffrement automatique. **Pascal Rivière** (DG, département des projets) a décrit l'ensemble de l'architecture SICORE, en commençant par définir ce que l'on entend par codage automatique. Puis il a présenté l'outil de chiffrement : séparation entre programmes et connaissances, contenu des bases de connaissances, algorithmes d'apprentissage et de codage, avant de donner des exemples d'utilisation de SICORE en production sur de nombreux types de libellés : nationalités, communes, professions, etc.

L'Insee, créateur, gestionnaire et grand utilisateur de nomenclatures, a développé un outil modulaire, appelé SYNAPSE, pour faciliter leur consultation, la gestion de leurs évolutions, leur diffusion la plus large possible, et leur coordination. **Emile Bruneau** (DG, département des normes statistiques et comptables) a présenté l'organisation générale de SYNAPSE, qui est composé de deux sous-applications : un serveur logique qui regroupe les données structurelles et formelles, toutes les fonctions utilisateurs et de gestion des nomenclatures et un serveur linguistique, outil de génie linguistique, qui regroupe les données textuelles et les fonctions directement liées à cet outil (grammaires, algorithmes de recherche, fonctions de gestion d'un dictionnaire, d'un thesaurus et d'une base d'index).

Michel de Bie a présenté LEDA2, un outil développé par l'Insee permettant d'aider l'utilisateur du logiciel pour la recodification de variables et production de tableaux sur de gros fichiers, avec la possibilité d'opérer simultanément sur plusieurs fichiers (par exemple Immeubles, Logements, Individus) reliés entre eux par des identifiants. Les recodifications sont spécifiées par des formules de calcul, des tables de décision, ou des enchaînements de tels calculs. Les tableaux sont plus généraux que ceux de la procédure Tabulate de SAS (par exemple : possibilité de définir un filtre pour chaque tableau d'une demande LEDA2). Plusieurs recodifications et tableaux peuvent être produits par une seule lecture d'un groupe de fichiers (économie pour de gros fichiers).

Formation à la méthodologie statistique

Jean Dumais (Statistique Canada) a présenté, dans la deuxième conférence spéciale, la formation des méthodologistes à Statistique Canada : le profil du méthodologiste type au recrutement, son plan de carrière, comment le programme de formation et perfectionnement participe à sa progression, et comment la progression des méthodologistes est liée à celle des autres groupes professionnels. Une part importante de la communication a été consacrée à la description du "Cours de base sur les enquêtes" : à travers la conception et la réalisation d'une enquête-ménages, cet enseignement vise l'apprentissage des notions de base en échantillonnage, traitement des données, conception de questionnaires, logistique de collecte, base de données, analyse et diffusion des résultats.

Le débat qui a suivi, dirigé par **Jean-Claude Deville** (DG, unité méthodes statistiques), a tenté de répondre à la question : "Peut-on enseigner la méthodologie statistique ?". **Jean-Jacques Droesbeke** (Université Libre de Bruxelles) a parlé de son expérience de formateur à la statistique, qui est souvent, et paradoxalement, mieux reçue par les non-mathématiciens. **Alain Trognon** (Groupe des Ecoles Nationales d'Economie et Statistique) a présenté les enseignements de l'Ensae et de l'Ensai qui abordent la méthodologie statistique, qui ne figure toutefois pas comme la préoccupation principale de ces écoles. Pour **Jean-Marie Grosbras** (Direction Régionale d'Ile-de-France), les

difficultés principales de l'enseignement de la méthodologie statistique viennent d'une part d'un manque de connaissance de base de la statistique (dès le lycée), d'autre part de l'inadaptation du cadre scolaire pour une telle formation. Le débat a porté sur une opposition entre une théorie et une pratique de la statistique, qui serait difficilement enseignable en dehors du "terrain", bien que cette distinction puisse être remise en cause par les apports récents des sciences sociales. Un consensus est apparu sur la nécessité de constituer une mémoire de la méthodologie statistique dans un institut tel que l'Insee.

Statistique non paramétrique

Les méthodes de la statistique non paramétrique, répandues depuis longtemps dans les centres de recherche en traitement des données, sont appelées à devenir un des outils usuels de l'analyse exploratoire ou mathématique des grands ensembles de chiffres, au même titre que la régression linéaire standard, dans l'analyse des enquêtes Insee. Pour rendre plus accessibles les deux exposés qui ont complété la session, **Michel Delecroix** (Ensai) a introduit deux méthodes de base : l'estimation de densité commune de variables observées, et l'estimation non paramétrique d'une courbe de régression.

Denis Fougère (CNRS-CREST) et **Daniel Verger** (DG-CREST) présenté une étude sur les distributions de revenus en France tirés de plusieurs enquêtes sur les "Revenus fiscaux" réalisées par l'Insee et la Direction Générale des Impôts entre 1970 et 1990. Des fonctions de densité des revenus ont été estimées aux différentes dates par la méthode du noyau, et comparées avec celles obtenues à partir d'hypothèses paramétriques usuelles (distributions log-normales par exemple). Puis différents indicateurs d'inégalité ont été estimés sur la base des résultats de régressions non-paramétriques dans lesquelles les variables explicatives sont les caractéristiques du ménage (par exemple l'âge du chef de ménage).

L'objet du travail présenté par **Michel Simioni** (GREMAQ - Toulouse) était l'estimation d'une fonction de gain à partir de données individuelles extraites de l'enquête *Formation Qualification Professionnelle* réalisée par l'Insee en 1993 : ceci a consisté à réaliser la régression du salaire perçu par un individu sur un ensemble de variables telles que l'ancienneté dans l'entreprise, le nombre d'années d'études initiales, les expériences professionnelles... Les résultats obtenus à partir d'estimations paramétriques et non-paramétriques ont été comparés et commentés en vue de dégager des profils âge-gains selon la formation des individus.

Démographie des entreprises

Marie-Christine Parent (DG, département système statistique d'entreprises) a présenté la constitution d'une base d'analyse longitudinale de données d'entreprises à partir des données annuelles BIC (bénéfices industriels et commerciaux) sur la période 1984-1992, auxquelles ont été ajoutées les dates de création et de cessation de l'entreprise issues du répertoire SIRENE. Cette base permet d'estimer les données manquantes à l'aide d'hypothèses d'interpolation, de reconstituer une information macro-économiquement plus fiable sur le champ des BIC, de tester des simulations de traitements et de corrections tels que corrections des durées d'exercice, décalage entre année civile et année comptable, estimations de données pour l'année de création, ...

La démographie d'entreprises consiste à observer les populations d'entreprises, ainsi que les mouvements (créations et cessations) qui affectent ces populations. Le dispositif actuel, qui s'appuie sur l'exploitation statistique du répertoire SIRENE, se compose principalement de séries de créations d'entreprises et de stocks. L'objectif de l'étude présentée par **Dominique Francoz** (DG, unité répertoire et démographie des entreprises et des établissements) était de mesurer le nombre de cessations survenues dans une période de temps déterminée. Les difficultés rencontrées étaient principalement dues au délai pouvant exister entre la date à laquelle survient la cessation et celle à laquelle elle est enregistrée dans le répertoire, ainsi qu'à l'absence de la date de la cessation.

L'étude de la survie des entreprises se heurte aux mêmes difficultés que celle des cessations. Les modèles de durée semblent appropriés pour résoudre les problèmes liés au retard voire à l'absence d'enregistrement des cessations. Dans ce cas, la durée considérée est la durée de vie de l'entreprise. La modélisation doit prendre en compte des phénomènes de censure : censure à gauche dans le cas où on ne connaît pas la date exacte de cessation mais seulement la date à laquelle elle a été enregistrée dans le fichier et censure à droite dans le cas où l'entreprise est toujours vivante au moment de l'observation. **Amel Gharbi** (DG - Université Paris I) a exposé des travaux réalisés sur une cohorte d'entreprises créées ou reprises en 1987 et suivies jusqu'en 1994.

Conférence spéciale : la refonte de l'échantillon de l'enquête sur la population active canadienne

L'enquête sur les forces de travail canadiennes, ou Enquête sur la Population Active (EPA), est un sondage mensuel auprès des ménages dont le but est d'estimer plusieurs variables du marché du travail comme l'emploi et le chômage. Après chaque recensement décennal, on fait la refonte du plan d'échantillonnage de l'EPA afin de tenir compte des changements dans la répartition géographique et dans les caractéristiques de la population canadienne. **Normand Laniel** (Statistique Canada) a présenté les

principales améliorations, en termes d'efficacité, apportées lors de la refonte des années 90 : remaniement de la stratification pour tenir compte des régions géographiques d'intérêt pour les programmes faisant usage de l'échantillon de l'EPA, utilisation d'une répartition plus proche de la répartition optimale, obtenue en contrôlant la taille moyenne des unités primaires d'échantillonnage dans les régions urbaines.

Amélioration des enquêtes

Les données de l'enquête sur les revenus fiscaux sont une des sources principales de connaissance des revenus en France, mais elles ignorent toutes les prestations non imposables : il est donc nécessaire d'imputer les prestations principales, soit par application des barèmes (allocations familiales...), soit par utilisation de méthodes économétriques. La communication de **Daniel Verger** (DG-CREST) et **Didier Contencin** (DG-DGI) a consisté en la présentation de la principale de ces imputations économétriques, à savoir celle qui concerne les allocations logement. La méthode utilisée commence par une phase d'estimation de la probabilité d'être allocataire, puis du montant de l'allocation, en fonction de variables socio-démographiques : cette phase a été réalisée à partir des données de l'enquête Logement conduite par l'Insee. Ces estimations servent ensuite à déterminer les ménages de l'enquête revenus fiscaux bénéficiaires, puis à leur imputer un montant d'allocation à l'aide d'une technique économétrique.

L'évaluation de l'Enquête Trimestrielle Emploi (E.T.E.), réalisée par l'Insee depuis juin 1992, a conduit l'institut à s'interroger sur certains problèmes de fiabilité des déclarations, dus au mode d'interrogation (téléphone ou face à face), à l'identité du répondant (selon que c'est l'individu concerné ou bien une autre personne du ménage), ou à une certaine volatilité du marché du travail en juin et en septembre pour certaines catégories de population. **Corinne Detour, Christine Lagarenne** (DG, division emploi) et **Pierrette Schuhl** (DG, unité méthodes statistiques) ont présenté les principaux résultats de deux enquêtes de qualité : l'enquête *Protocole*, qui a consisté à doubler l'E.T.E. de juin 1994 par une enquête réalisée en face à face, une semaine plus tard, sur un échantillon de 1200 ménages, et l'enquête *Transitions sur le marché du travail*, qui, a consisté à réinterroger, exactement un mois après, l'E.T.E. de septembre 1994, selon le même mode de collecte, un échantillon de 5600 personnes, auxquelles on demandait en particulier de retracer leur situation semaine par semaine depuis la semaine de référence de l'E.T.E. D'après ces enquêtes, l'effet téléphone et l'effet dû à l'identité du répondant apparaissent non significatifs pour la variable synthétique de l'enquête (activité au sens du BIT) ; par ailleurs, il apparaît que la fenêtre d'observation de l'E.T.E. de septembre ne permet pas de prendre en compte correctement l'entrée des jeunes sur le marché du travail.

Conclusion

Comme l'a souligné en clôture de ces Journées Michel Glaude (Directeur des statistiques démographiques et sociales de l'Insee), l'évolution des effectifs des participants témoigne du succès grandissant de ces Journées, et de la nécessité de faciliter les échanges entre les méthodologues statisticiens de la "statistique officielle". L'Unité "Méthodes statistiques" a donc décidé de lancer une nouvelle série de documents de travail à caractère méthodologique qui devrait faciliter la publication des travaux de méthodologie statistique.