LES JOURNEES DE METHODOLOGIE STATISTIQUE DE L'INSEE

M. CHRISTINE

INSEE, Chef-adjoint de l'Unité "Méthodes Statistiques"

Organisées à l'initiative de l'Unité « Méthodes statistiques », les VIIèmes Journées de méthodologie statistique de l'Insee (JMS) se sont tenues les 4 et 5 décembre 2000 à Paris, au Centre de conférences Pierre Mendès-France du ministère de l'Économie, des Finances et de l'Industrie[1].

Inaugurées en 1991, ces Journées sont devenues un lieu d'échanges apprécié entre les méthodologues de l'Insee, des ministères ou de divers organismes, publics ou privés, réalisant des travaux statistiques. Ouvertes sur d'amples domaines d'intérêt, elles s'adressent à un public diversifié, tant à l'intérieur du système statistique public qu'à sa périphérie. Elles permettent de faire connaître des travaux novateurs à un large public et d'en assurer la diffusion.

Après une absence de plus de deux ans et demi (la précédente édition avait eu lieu en mars 1998)[2], le succès habituel de cette manifestation s'est confirmé à nouveau : près de 500 inscriptions ont été enregistrées, pour une participation effective cumulée de l'ordre des 3/4. Le secteur public a représenté bien évidemment la grande majorité de l'auditoire : un peu plus de 200 participants de l'Insee, dont 70 « régionaux », 140 représentants des ministères, une cinquantaine pour les établissements publics et les grandes entreprises nationales, 35 pour le monde de l'enseignement ou de la recherche, une vingtaine pour les organismes de sécurité sociale. Le secteur privé était également représenté par une vingtaine de participants, quelques étrangers ou représentants d'institutions internationales complétant l'auditoire.

La logistique sans faille, assurée par l'Unité « Communication externe » de l'Insee, ainsi que les conditions très confortables offertes par le ministère, ont sans nul doute contribué à la réussite du colloque et à la satisfaction de ses auditeurs. Grâce également au travail actif de l'atelier d'impression de l'Insee et à de nombreux

[1] En marge de ces VIIèmes JMS, le Département de l'action régionale de l'Insee a organisé une journée de méthodologie régionale, plus particulièrement destinée à la présentation de travaux régionaux et à la discussion entre statisticiens régionaux. Cette réunion s'est tenue le 6 décembre 2000 à la Direction Générale de l'Insee.

[2] Les Actes complets de ces VIèmes JMS ont été publiés dans *Insee méthodes*, n° 84-85-86, mars 1999.

collaborateurs bénévoles, les participants ont pu recevoir chacun une valisette contenant près d'un millier de pages de documents.

Cinq grands thèmes, quarante communications

Cette septième édition des Journées de méthodologie statistique était organisée autour de cinq grands thèmes d'étude : sondages et échantillonnage ; statistique théorique et techniques d'estimation ; analyse des données, indices et algorithmes ; économétrie des comportements individuels ; techniques d'enquête, modes de collecte, questionnements, qualité. Les trois premiers ont alimenté la première journée, les deux suivants avaient été réservés pour le lendemain.

Pour chacun des thèmes, les sujets traités visaient à respecter autant que possible un équilibre entre communications théoriques et présentations de cas pratiques. Tout en conservant un cœur de cible centré sur les problèmes de sondage et d'estimation, on avait par ailleurs cherché à élargir l'ensemble des contributeurs potentiels afin d'ouvrir les JMS à une large gamme de questions traitées.

Ainsi, près de 40 contributions ont été présentées, dont 25 signées ou cosignées par des auteurs de l'Insee.

Sondages et échantillonnage

Les *grands chantiers méthodologiques actuels de l'Insee* en matière de statistique auprès des ménages, et plus particulièrement sur les questions de sondages, étaient présents dès la première session. En ouverture, Jean-Claude DEVILLE et Yves TILLE ont exposé leurs récents travaux théoriques sur les sondages « équilibrés » et décrit l'algorithme qui réalise des tirages d'échantillons selon cette méthode : la macro « Cube ». Laurent WILMS a présenté des simulations qu'il a effectuées avec cet outil, dans le cadre du tirage des unités primaires de l'échantillon-maître, en cours de constitution à partir des données du Recensement général de la Population de 1999.

Puis Jean DUMAIS et Michel ISNARD, après une présentation générale du recensement rénové de la population, ont mis l'accent sur les questions liées à l'estimation de la population des grandes communes à partir de sondages, ainsi que sur celles relatives à la mise en oeuvre d'un processus « de synthèse » utilisant données d'enquêtes et fichiers administratifs.

Enfin, l'équipe chargée de l'enquête auprès des sans-domicile (Pascal ARDUIN, Emmanuel MASSE, Nancy VIARD) a exposé les différents problèmes méthodologiques (plan de sondage) et pratiques (collecte) auxquels elle avait été confrontée dans le cadre de cette enquête tout à fait innovante.

Deux communications écrites étaient associées à cette session : l'une détaillait la méthodologie de constitution, à partir du recensement de 1999, des échantillons maître et Emploi pour les enquêtes de l'Insee, et, notamment, les outils cartographiques afférents (Georges BOURDALLE, Marc CHRISTINE et Laurent WILMS) ; l'autre, de Gildas ROY et Aurélie VANHEUVERZWYN (Médiamétrie), présentait un algorithme relatif à la réalisation de sondages par quota.

Dans le domaine de la théorie des sondages, enfin, Ruilin REN (Université de SOUTHAMPTON) a présenté ses travaux relatifs à l'estimation d'une fonction de répartition et des fractiles d'une population finie.

Statistique théorique et techniques d'estimation

Les questions d'estimation et de statistique théorique ont constitué également une part importante de la première journée.

Différents *problèmes théoriques d'estimation* ont tout d'abord été abordés. Celui traité par Jean-Louis PHILOCHE (Groupe des Ecoles des Télécommunications) était l'estimation du maximum de vraisemblance dans une analyse en composantes principales insérée dans un modèle gaussien bruité. La communication de Julien DAMON (Médiamétrie), centrée sur les méthodes de prévision de processus autorégressifs hilbertiens appuyées sur des modèles à temps continu, incluait un exemple d'application au taux d'audience de la télévision. Celle de Farid BENINEL et Michel GRUN-REHOMME (Universités de Poitiers et Paris II) était ciblée sur la mesure de la concordance entre les profils de réponse de deux individus formant partie d'un même panel et interrogés à des dates régulières. Marc CHRISTINE et Christian ROBERT ont illustré les liens inédits entre la statistique et la justice pénale à partir d'un fait divers qui les a conduits à construire un intervalle de confiance « exact » permettant de situer la valeur d'une certaine probabilité dans un contexte de dégénérescence des estimateurs habituels[3].

Les trois communications suivantes étaient consacrées aux problèmes des *estimations « continues »*. Jean-Claude DEVILLE a d'abord rappelé la nécessité de bien définir les objets ou grandeurs à mesurer dans un processus continu. Nathalie CARON et Philippe RAVALET ont ensuite traité des techniques d'estimation dans les enquêtes répétées. Une application importante est celle de la future enquête Emploi en continu : la communication de Philippe FEVRIER et Jonathan BOSREDON portait, dans ce cadre, sur la construction théorique d'estimateurs lorsque le phénomène observé peut être modélisé par un processus de séries chronologiques.

[3] Un suspect dans une affaire criminelle prétendait avoir utilisé son téléphone portable à une certaine heure, ce qui était aisément vérifiable, et en un certain lieu, ce qui dans l'absolu n'était pas vérifiable. Aussi le juge d'instruction a-t-il demandé à l'Insee s'il était possible d'estimer la probabilité de la véracité de cette assertion.

Après le *temps*, l'*espace*... Michel HANNOUN a en effet conclu cette deuxième session en dressant un panorama des différentes techniques d'estimation spatiale. Il a en particulier souligné leurs spécificités par rapport aux techniques génériques d'estimation statistique.

Analyse des données, indices et algorithmes

La troisième session était consacrée à l'analyse des données, au travers de questions de cryptage des données et de secret statistique, de classification et de construction d'indices.

Catherine QUANTIN (CHU de Dijon) a présenté une méthode de chaînage de données sensibles permettant de réunir les différentes parties du dossier d'un même patient tout en respectant l'anonymat de ce dernier. Un tel système a été avalisé par la CNIL dans le cadre de la mise en place du PMSI (programme de médicalisation des systèmes d'information) dans les établissements de santé (privés). Sur un thème voisin, mais dans un contexte très différent, Lionel VIGLINO a fait une démonstration du logiciel Argus utilisé pour le cryptage de certaines cases d'un tableau statistique, dans le but d'une mise en conformité avec les règles du secret statistique relatif aux enquêtes « entreprises ».

Les trois exposés suivants traitaient de *techniques de classification.* Marie COTTRELL (Université de Paris I) et Sophie PONTHIEUX ont décrit une technique tout à fait innovante d'analyse des données, fondée sur la classification « neuronale », et l'ont comparée aux méthodes traditionnelles d'analyse pour la description des conditions de vie. Marc CHRISTINE et Michel ISNARD ont ensuite présenté l'algorithme qu'ils ont mis au point pour réaliser l'agrégation d'unités statistiques sous un triple jeu de contraintes : taille des agrégats, contiguïté géographique et homogénéité ou hétérogénéité vis-à-vis de certaines variables d'intérêt. Ce type d'outil pourra être utilisé pour la construction des unités primaires des futurs échantillons-maîtres, mais aussi dans les problématiques de typologie spatiale. Enfin, Brigitte GELEIN-DOUKKALI a montré comment construire une typologie des zones d'emploi à partir de techniques de segmentation.

Deux communications, présentant les fondements théoriques à la construction des indices de prix (François MAGNIEN, Lionel VIGLINO), concluaient cette session.

Économétrie des comportements individuels

L'économétrie des comportements individuels a ouvert la deuxième journée. Plusieurs exposés au contenu très pédagogique ont été entendus : Jean-Marc ROBIN sur le traitement de l'endogénéité et l'utilisation des variables instrumentales,

Bruno CREPON sur les méthodes d'appariement dans l'évaluation des politiques économiques, Daniel COURGEAU (INED) sur les analyses « multi-niveaux ».

Trois sujets plus appliqués illustraient des techniques particulières d'économétrie. Jacques MAIRESSE, en collaboration avec Nathalie GREENAN, a démontré que l'on pouvait enrichir les études économétriques sur les entreprises en combinant les données recueillies sur celles-ci avec des données relatives à leurs salariés, même si l'on en enquête très peu par entreprise. La contribution cosignée de Jean-Loup MADRE (INRETS) portait sur l'estimation de la demande de transport à partir de techniques d'économétrie bayésienne. Enfin, Dominique DESBOIS (INRA) a montré comment évaluer les marges brutes standard des exploitations agricoles à partir d'un modèle économétrique d'estimation des coûts de production.

Techniques d'enquête, modes de collecte, questionnements, qualité

Une bonne statistique n'est pas seulement affaire de bonne théorie : encore faut-il que le processus de collecte, les outils de codification en aval et la qualité des données soient maîtrisés et assurés. C'est sur ces points que portait la cinquième et dernière session des JMS.

Il a d'abord été question de *codification automatique* : comparaison entre les codifications des enquêtes Emploi du temps et Budget de famille, par Sophie DESTANDAU, Frédérique DESCHAMPS et Françoise DUMONTIER, et codification de la profession, par Alain CHENU et Francis GUGLIELMETTI.

Les quatre communications suivantes ont décrit des *techniques de collecte innovantes*. La contribution de Valérie DEROIN et Jean-Marc BEGUIN (SESSI) a porté sur la collecte par Internet des enquêtes de branche. Gildas ROY et Aurélie VANHEUVERZWYN (Médiamétrie) ont mis en évidence le double problème que soulève, pour les sondages par téléphone, la généralisation du portable : perturbations dans la constitution des échantillons sur numéros fixes (l'acquisition d'un portable s'accompagnant souvent, chez certaines catégories d'utilisateurs, de l'abandon du poste fixe), et difficultés propres liées à la réalisation d'une enquête par appel d'un numéro de portable. François BECK (Observatoire français des drogues et toxicomanies) a montré quels protocoles particuliers devaient être mis en oeuvre pour évaluer la consommation de substances psychoactives dans la population. Enfin, Yannick LEMEL a décrit une méthode très originale pour évaluer la position sociale des individus à partir de leurs caractéristiques sociodémographiques (âge, profession, niveau de ressources...).

Dans la deuxième partie de la session, ont été présentés plusieurs exemples *d'exploitation des données*. Pascale BREUIL s'est intéressée aux performances comparées d'un panel d'individus (le panel européen) et d'un panel de logements

(l'enquête Emploi) dans la mesure des trajectoires d'emploi. Fabrice MURAT, Xavier d'HAULTFOEUILLE et Thierry ROCHER se sont attachés à décrire les problèmes d'évaluation du système éducatif vus sous l'angle de la comparabilité internationale.

Des travaux récents de Statistique Canada concluaient la session : développement de l'enquête sur la santé dans les collectivités canadiennes (Yves BELAND), et utilisation de la taxe sur les produits et services dans le remaniement de l'enquête mensuelle du commerce de gros et de détail (Marie BRODEUR).

Conclusion des échanges : l'exigence de la qualité

Une table ronde finale sur la question de la qualité a clôturé ces deux journées particulièrement denses.

Présidée par Michel GLAUDE, Directeur des statistiques démographiques et sociales à l'Insee, elle a réuni des intervenants d'origines diverses ayant des approches différentes de la définition de la qualité, de la façon de la mesurer et des moyens de l'améliorer : l'Insee (Michel BLANC, Raoul DEPOUTOT, Alain DESROSIERES), Eurostat (Roberto BARCELLAN), le monde universitaire (Gilbert SAPORTA, CNAM et Société Française de Statistique), enfin le secteur privé (Yves CARRIOU, SOFRES).

Nul doute que les démarches entreprises par les statisticiens privés, notamment la recherche de la certification, ne constituent effectivement un défi à relever pour la statistique publique ! À ce titre, l'échange de vues était particulièrement éclairant et enrichissant.

La VIII ème session des JMS devrait se tenir en décembre 2002.