RÉSUMÉS DES INTERVENTIONS

Olivier SAUTORY

Dans la **première conférence spéciale, Benoît QUENNEVILLE** (Statistique Canada) a présenté les différents aspects des travaux réalisés par la Division des séries chronologiques - recherche et analyse, de Statistique Canada :

- les pratiques actuelles de désaisonnalisation, avec la méthode d'ajustement saisonnier X11-ARIMA ;

- les pratiques statistiques intégrées dans la chaîne de production d'informations : raccordement des segments d'une même série, étalonnage et interpolation des séries, "calendrialisation" des données ;

- les développements théoriques récents en série chronologiques ;

- les projets en cours : étude de nouveaux filtres d'estimation de la tendance-cycle, estimation de la variance des séries désaisonnalisées, estimation de modèles FARMA pour la prévision de séries caractérisées par une mémoire longue.

Comment résoudre le dilemme : produire de plus en plus de données, de plus en plus rapidement, tout en maintenant des standards de qualité, avec des budgets réduits ? Dans la **deuxième conférence spéciale, Sylvie MICHAUD** (Statistique Canada) a présenté les nouvelles technologies de collecte des données utilisées à Statistique Canada. Des tests sont en cours de réalisation concernant les ordinateurs portatifs, les ordinateurs à écran sensible, les machines avec reconnaissance de l'écriture, l'utilisation du téléphone "touch-tone".

Très souvent les utilisateurs de données d'enquête réalisent leurs études en se servant de méthodes d'inférence ou d'analyse des données fondées sur l'hypothèse que les observations proviennent d'un échantillon aléatoire simple. Qu'en est-il lorsque ces données sont issues d'un plan d'échantillonnage complexe ? **Gad NATHAN** (Université Hébraïque de Jérusalem), a montré, au cours de la **troisième conférence spéciale**, comment on peut prendre en compte les effets du plan de sondage dans l'analyse des données. Une possibilité est de construire des statistiques qui sont basées sur le plan d'échantillonnage complexe. Une autre possibilité est de modifier les statistiques ou les tests d'hypothèses classiques afin qu'ils conviennent au plan d'échantillonnage. Cette possibilité facilite l'emploi des programmes informatisés standardisés.

Session 1 : traitement de séries chronologiques

L'analyse conjoncturelle est un exercice difficile qui, pour maintenir sa crédibilité, doit tenter de reposer sur des méthodes réplicables dont on peut aisément évaluer la pertinence. S'appuyant sur son système d'enquêtes de conjoncture, l'Insee a développé une méthode de diagnostic conjoncturel, présentée par **Stéphane GRÉGOIR** (Direction générale de l'Insee, département de la conjoncture) : une investigation économétrique des relations stables qui existent entre les soldes d'opinion aux différentes questions qualitatives d'une enquête de conjoncture, et l'évolution quantifiée de variables macro-économiques reliées aux thèmes étudiés dans cette enquête, permettent une prévision de ces variables.

L'étude d'une série temporelle passe souvent par une phase de lissage de la série qui permet d'en estimer la tendance ; c'est le cas même lorsque le lissage n'est pas le but ultime de l'analyse comme, par exemple, en désaisonnalisation (voir les logiciels X11 ou X11-ARIMA). **Dominique LADIRAY** (Ensae, division CGSA) a présenté une procédure intégrée dans la version 6 du logiciel SAS, qui permet à l'utilisateur de construire lui-même le lisseur adapté à sa série, à partir de moyennes mobiles, de médianes mobiles ou de fonctions splines.

Les séries démographiques (nombre de naissances, de mariages ...) présentent des mouvements saisonniers, plus ou moins marqués, qui évoluent avec le temps. Ces phénomènes sont intéressants en eux-mêmes d'un point de vue sociologique. Une analyse aussi précise que possible des variations saisonnières est nécessaire pour apprécier convenablement l'évolution réelle des phénomènes de fond (natalité, nuptialité ...) notamment d'un point de vue conjoncturel. **Jean-Claude LABAT** (Insee-DG, département de la démographie) a présenté les méthodes utilisées pour le traitement des séries mensuelles de mariages.

Session 2 : codification automatique

QUID (QUestionnaire d'IDentification) est un système de chiffrement automatique conçu et développé à l'Insee par **Jacques LORIGNY** et son équipe pour traiter les réponses obtenues dans un langage naturel à partir d'une question ouverte dans un questionnaire. **Lionel VIGLINO** (Ensae, 2^{e} année SEA) a présenté l'utilisation de QUID lors de la codification de la catégorie socioprofessionnelle à partir du bulletin individuel du recensement français dans les départements d'outre-mer. La méthode générale de chiffrement automatique QUID a été complétée par des algorithmes chargés de trouver et de traiter l'information nécessaire au chiffrement parmi les 12 variables du bulletin qui concernent la codification de la CS. Les avantages du système QUID sont l'augmentation de la justesse du chiffrement et la réduction des délais de traitement.

À l'occasion de la rénovation de l'enquête Emploi effectuée en 1990, le mode de traitement de l'enquête a été amélioré : pour diminuer la durée des traitements, l'option prise a consisté à automatiser autant que possible les chiffrements requis dans l'enquête. **Catherine BERGERA** (Insee-DG, département des projets) a présenté les outils de chiffrement automatique de la catégorie socioprofessionnelle, et les performances de ces codifications. Ces outils déchargent les Directions régionales, selon les cas, de 50% à 90% des chiffrements sans perte de qualité par rapport à un chiffrement manuel, tout en assurant une certaine homogénéité.

Le point central du système QUID est le critère, dit "infomax", qui consiste à choisir toujours la question qui possède la plus grande quantité d'information. Or la crainte a été formulée que ce critère puisse se révéler trop rigide dans certains contextes de données. Depuis deux ans, une nouvelle recherche en cours au Département des projets, dite "Multiquid", vise à explorer les voies d'assouplissement du critère infomax. **Jacques LORIGNY** (Insee-DG, département des projets) en a présenté les principaux résultats. La première phase (2^{e} semestre 1990) a montré l'intérêt, pour chiffrer chaque cas individuel, d'explorer conjointement plusieurs quids voisins du critère infomax. La seconde phase (1er semestre 1991) a consisté à "filmer", en quelque sorte, le phénomène de l'apprentissage par acquisition croissante et ses effets sur la productivité et sur la fiabilité. Enfin, la troisième phase (2^{e} semestre 1991) a apporté une solution originale au problème de la génération automatique de règles indépendantes à partir d'une base d'exemples.

Session 3 : l'usage des modèles LOGIT

L'analyse des données qualitatives est très ancienne en statistique et a concerné pendant longtemps, pour des raisons de pratique de calcul, les seuls tableaux de contingence. Le traitement des données individuelles a connu son plein essor dès que les ordinateurs ont permis, en temps réduit, l'estimation du maximum de vraisemblance sur de très larges échantillons. L'application du modèle LOGIT devient de ce fait d'un usage facile pour le praticien économiste, démographe, sociologue, etc. Après la présentation générale des procédures statistiques attachées au modèle LOGIT, ou régression logistique, par **Alain TROGNON** (Insee-DG, division conditions de vie des ménages), la session a porté sur deux exemples d'application.

L'enquête "Actifs financiers" permet une analyse de la transmission des patrimoines des ménages. **Anne LAFERRÈRE** (Insee-DG, division patrimoine des ménages) a étudié les facteurs influençant les pratiques suivantes, à l'aide de régressions logistiques : la pratique testamentaire, la donation entre époux (ou au dernier vivant), la donation aux enfants.

À partir d'enquêtes réalisées par l'Ined en 1978 et 1988, **Laurent TOULEMON** (Ined) a montré comment la pratique contraceptive varie fortement selon la situation "démo-

graphique" : âge, nombre d'enfants, situation de couple. L'orateur a en particulier rappelé les avantages de l'échelle logistique pour la description des proportions, et a proposé une méthode qui permet de transformer les paramètres de la régression logistique en proportions estimées, beaucoup plus concrètes.

Session 4 : l'analyse discriminante et applications

L'analyse discriminante regroupe l'ensemble des techniques statistiques ayant pour but d'identifier et de caractériser des classes d'individus définies *a priori*, en fonction de variables descriptives, numériques ou catégorielles. **Olivier SAUTORY** (Insee-DG, division des méthodes statistiques et des sondages) a donné un aperçu des différentes méthodes d'analyse discriminante : analyse factorielle discriminante, méthodes probabilistes paramétriques ou non paramétriques, méthodes de sélection de variables pas à pas...

Henri MARIOTTE (Ensae, 3e année SEA) a présenté les résultats d'une étude utilisant l'analyse discriminante pour pallier le problème des non-réponses partielles dans les enquêtes auprès des ménages, dans le but d'imputer une valeur vraisemblable. La méthode utilise en général des variables explicatives qualitatives, rendues quantitatives par l'intermédiaire d'une analyse des correspondances multiples. Elle a été comparée à la méthode d'imputation par hot-deck.

La régression logistique et l'analyse discriminante sont des techniques "concurrentes" permettant de décrire les relations entre une variable qualitative "expliquée" et des variables "explicatives" quantitatives ou qualitatives. **Olivier SAUTORY** a mis en parallèle les approches théoriques à la base de ces deux méthodes, afin de voir ce qui les relie et ce qui les différencie. **Chang VONG** (Ensae, 2e année CGSA) a présenté les résultats d'une étude comparative menée sur les données de l'enquête "Budget de famille" réalisée en 1989 auprès des ménages : les deux techniques ont été utilisées pour analyser les relations entre une variable dichotomique (possession d'une carte bancaire, d'une assurance-vie...) et un certain nombre de caractéristiques socio-démographiques du ménage (catégorie socioprofessionnelle, revenu, catégorie de commune...).

Session 5 : contrôle de qualité du recensement

Après le recensement de 1990, l'Insee a réalisé une enquête ayant pour objectif une mesure et une étude des erreurs de dénombrement des personnes dans ce recensement. Au cours de cette enquête (aréolaire), qu'a présentée **Nicole COEFFIC** (Insee-DG, division logement), l'enquêteur dénombrait tous les logements situés sur les aires de l'échantillon, puis tous les occupants de ces logements. La confrontation

des résultats obtenus avec les documents des recensements a permis une estimation du taux d'omission des personnes au recensement (de l'ordre de 2%) et du taux de doubles comptes (proche de 1%).

Jean-Claude DEVILLE (Insee-DG, division des méthodes statistiques et des sondages) a montré comment la technique des sondages a été appliquée au contrôle de la qualité des données provenant du recensement de la population de 1990, à deux stades :

- pour la "saisie de masse", confiée à des entreprises extérieures à l'Insee, la vérification de la qualité du travail, opérée par sondage, s'est réalisée en minimisant le coût de contrôle (coût de manipulation des dossiers de districts et de saisie des bulletins échantillonnés) tout en gardant supérieure à un certain seuil la précision de la mesure du taux d'erreur pour chaque type de bulletin ;

- en ce qui concerne la codification, surtout celle des catégories socioprofessionnelles et des structures familiales, la technique de contrôle par sondage a utilisé une information auxiliaire très riche, provenant de la précodification du bulletin et d'informations tirées de l'essai de recensement : on avait ainsi une idée *a priori* des risques d'erreurs au niveau de chaque bulletin.

Gérard BADEYAN (Insee-DG, département des projets) a exposé les résultats de ces procédures de contrôle. Pour l'exhaustif léger seuls 4 lots (unités de traitement) parmi 242 ont fait l'objet d'un refus. La qualité moyenne de la saisie-chiffrement s'est avérée bonne par rapport aux normes exigées (par exemple, pour les bulletins individuels 1,6% d'enregistrements erronés pour une norme contractuelle de 3%). Pour Colibri la rigueur de la phase d'apprentissage a permis de limiter à deux les contrôles négatifs sur l'ensemble de l'opération (environ 12 mois de contrôles collectifs hebdomadaires ou bimensuels pour chaque établissement). Les taux d'erreurs pour les variables de saisie correspondent aux taux habituels pour ce type de travail (0,4%). Pour les chiffrements complexes, les taux obtenus sont particulièrement bons (de 0,70% pour la PCS à un chiffre, à 2,42% pour la PCS à quatre chiffres) et assez nettement inférieurs aux normes exigées.

Débat : "Communication et Méthodologie"

Un débat, organisé par **Claude SEIBEL** (Insee-DG, direction des statistiques démographiques et sociales), a porté sur les problèmes de communication en matière de méthodologie statistique. Les participants à ce débat, **Jean-François ROYER** (DR de Bourgogne), **Jacques ANTOINE** (CNAM), **Eric BASSI** (AFP), **Philippe TASSI** (Médiamétrie), se sont efforcés de répondre aux questions suivantes :

- comment accompagner la communication de la production statistique des éléments méthodologiques nécessaires à la compréhension de cette production ?

- faut-il dans le dialogue avec les utilisateurs, notamment avec les médias, attirer leur attention sur les marges d'imprécision liées aux techniques statistiques employées ?

- les utilisateurs des données statistiques souhaitent-ils connaître les méthodes mises en **oeuvre ?**

Si tous les intervenants se sont accordés sur la nécessité de clarifier le statut des concepts et des données produites, d'éviter toute langue de bois dans la communication, de publier l'ordre de grandeur de la précision obtenue pour les chiffres produits, ils ont aussi souligné les obstacles qui émanent autant des statisticiens que des utilisateurs. Un travail en commun de longue haleine doit donc s'instaurer. Peut-être serait-il favorisé par des formations conjointes des utilisateurs - tels que les journalistes - et des statisticiens ? Ainsi l'Ensae et le centre de formation des journalistes (CFJ) pourraient-ils engager une collaboration sur ces thèmes ?